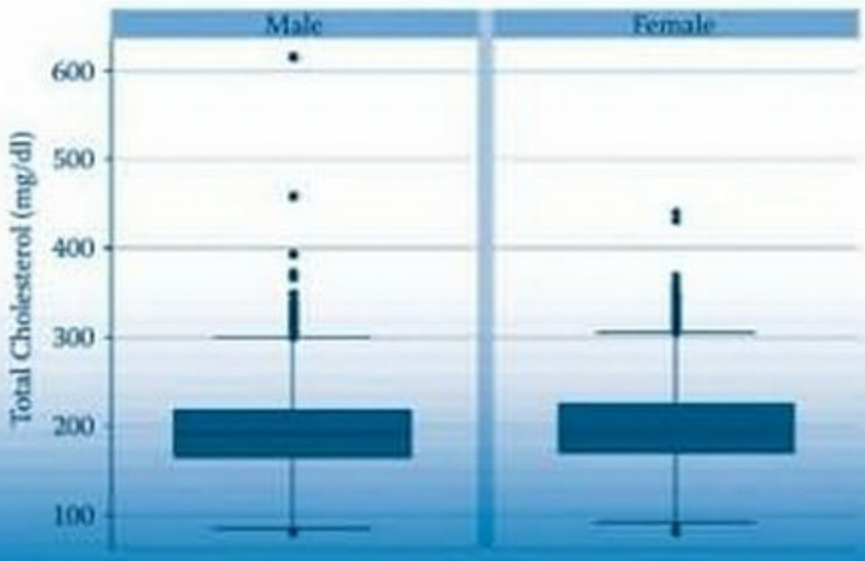
Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Applied Survey Data Analysis



Steven G. Heeringa
Brady T. West
Patricia A. Berglund

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Applied Survey Data Analysis

Chapman & Hall/CRC

Statistics in the Social and Behavioral Sciences Series

Aims and scope

Large and complex datasets are becoming prevalent in the social and behavioral sciences and statistical methods are crucial for the analysis and interpretation of such data. This series aims to capture new developments in statistical methodology with particular relevance to applications in the social and behavioral sciences. It seeks to promote appropriate use of statistical, econometric and psychometric methods in these applied sciences by publishing a broad range of reference works, textbooks and handbooks.

The scope of the series is wide, including applications of statistical methodology in sociology, psychology, economics, education, marketing research, political science, criminology, public policy, demography, survey methodology and official statistics. The titles included in the series are designed to appeal to applied statisticians, as well as students, researchers and practitioners from the above disciplines. The inclusion of real examples and case studies is therefore essential.

Published Titles

Analysis of Multivariate Social Science Data, Second Edition
David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane I. Galbraith

Applied Survey Data Analysis
Steven G. Heeringa, Brady T. West, and Patricia A. Berglund

Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition
Jeff Gill

Foundations of Factor Analysis, Second Edition
Stanley A. Mulaik

Linear Causal Modeling with Structural Equations
Stanley A. Mulaik

Multiple Correspondence Analysis and Related Methods
Michael Greenacre and Jorg Blasius

Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences
Brian S. Everitt

Statistical Test Theory for the Behavioral Sciences
Dato N. M. de Gruijter and Leo J. Th. van der Kamp

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Applied Survey Data Analysis

Steven G. Heeringa
Brady T. West
Patricia A. Berglund

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

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

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-8066-7 (Hardback)

Library of Congress Cataloging-in-Publication Data

Heeringa, Steven, 1953-
Applied survey data analysis / Steven G. Heeringa, Brady West, and Patricia A. Berglund.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4200-8066-7 (alk. paper)
1. Social sciences--Statistics. 2. Social surveys--Statistical methods. I. West, Brady T. II. Berglund, Patricia A. III. Title.

HA29.H428 2010
001.4'22--dc22 2009051730

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Preface

This book is written as a guide to the applied statistical analysis and interpretation of survey data. The motivation for this text lies in years of teaching graduate courses in applied methods for survey data analysis and extensive consultation with social and physical scientists, educators, medical researchers, and public health professionals on best methods for approaching specific analysis questions using survey data. The general outline for this text is based on the syllabus for a course titled "Analysis of Complex Sample Survey Data" that we have taught for over 10 years in the Joint Program in Survey Methodology (JPSM) based at the University of Maryland (College Park) and in the University of Michigan's Program in Survey Methodology (MPSM) and Summer Institute in Survey Research Techniques.

Readers may initially find the topical outline and content choices a bit unorthodox, but our instructional experience has shown it to be effective for teaching this complex subject to students and professionals who have a minimum of a two-semester graduate level course in applied statistics. The practical, everyday relevance of the chosen topics and the emphasis each receives in this text has also been informed by over 60 years of combined experience in consulting on survey data analysis with research colleagues and students under the auspices of the Survey Methodology Program of the Institute for Social Research (ISR) and the University of Michigan Center for Statistical Consultation and Research (CSCAR). For example, the emphasis placed on topics as varied as weighted estimation of population quantities, sampling error calculation models, coding of indicator variables in regression models, and interpretation of results from generalized linear models derives directly from our long-term observation of how often naïve users make critical mistakes in these areas.

This text, like our courses that it will serve, is designed to provide an intermediate-level statistical overview of the analysis of complex sample survey data—emphasizing methods and worked examples while reinforcing the principles and theory that underly those methods. The intended audience includes graduate students, survey practitioners, and research scientists from the wide array of disciplines that use survey data in their work. Students and practitioners in the statistical sciences should also find that this text provides a useful framework for integrating their further, more in-depth studies of the theory and methods for survey data analysis.

Balancing theory and application in any text is no simple matter. The distinguished statistician D. R. Cox begins the outline of his view of applied statistical work by stating, "Any simple recommendation along the lines *in applications one should do so and so* is virtually bound to be wrong in some or, indeed, possibly many contexts. On the other hand, descent into yawning

THEORY BOX P.1 AN EXAMPLE THEORY BOX

Theory boxes are used in this volume to develop or explain a fundamental theoretical concept underlying statistical methods. The content of these "gray-shaded" boxes is intended to stand alone, supplementing the interested reader's knowledge, but not necessary for understanding the general discussion of applied statistical approaches to the analysis of survey data.

vacuous generalities is all too possible" (Cox, 2007). Since the ingredients of each applied survey data analysis problem vary—the aims, the sampling design, the available survey variables—there is no single set of recipes that each analyst can simply follow without additional thought and evaluation on his or her part. On the other hand, a text on applied methods should not leave survey analysts alone, fending for themselves, with only abstract theoretical explanations to guide their way through an applied statistical analysis of survey data.

On balance, the discussion in this book will tilt toward proven recipes where theory and practice have demonstrated the value of a specific approach. In cases where theoretical guidance is less clear, we identify the uncertainty but still aim to provide advice and recommendations based on experience and current thinking on best practices.

The chapters of this book are organized to be read in sequence, each chapter building on material covered in the preceding chapters. Chapter 1 provides important context for the remaining chapters, briefly reviewing historical developments and laying out a step-by-step process for approaching a survey analysis problem. Chapters 2 through 4 will introduce the reader to the fundamental features of **complex sample designs** and demonstrate how design characteristics such as stratification, clustering, and weighting are easily incorporated into the statistical methods and software for survey estimation and inference. Treatment of statistical methods for survey data analysis begins in Chapters 5 and 6 with coverage of univariate (i.e., single-variable) descriptive and simple bivariate (i.e., two-variable) analyses of continuous and categorical variables. Chapter 7 presents the linear regression model for continuous dependent variables. Generalized linear regression modeling methods for survey data are treated in Chapters 8 and 9. Chapter 10 pertains to methods for event-history analysis of survey data, including models such as the Cox proportional hazards model and discrete time models. Chapter 11 introduces methods for handling missing data problems in survey data sets. Finally, the coverage of statistical methods for survey data analysis concludes in Chapter 12 with a discussion of new developments in the area of survey applications of advanced statistical techniques, such as multilevel analysis.

To avoid repetition in the coverage of more general topics such as the recommended steps in a regression analysis or testing hypotheses concerning regression parameters, topics will be introduced as they become relevant to the specific discussion. For example, the iterative series of steps that we recommend analysts follow in regression modeling of survey data is introduced in Chapter 7 (linear regression models for continuous outcomes), but the series applies equally to model specification, estimation, evaluation, and inference for generalized linear regression models (Chapters 8 and 9). By the same token, specific details of the appropriate procedures for each step (e.g., regression model diagnostics) are covered in the chapter on a specific technique. Readers who use this book primarily as a reference volume will find cross-references to earlier chapters useful in locating important background for discussion of specific analysis topics.

There are many quality software choices out there for survey data analysts. We selected Stata® for all book examples due to its ease of use and flexibility for survey data analysis, but examples have been replicated to the greatest extent possible using the SAS®, SPSS®, IVEware, SUDAAN®, R, WesVar®, and Mplus software packages on the book Web site (http://www.isr.umich.edu/src/smp/asda/). Appendix A reviews software procedures that are currently available for the analysis of complex sample survey data in these other major software systems.

Examples based on the analysis of major survey data sets are routinely used in this book to demonstrate statistical methods and software applications. To ensure diversity in sample design and substantive content, example exercises and illustrations are drawn from three major U.S. survey data sets: the 2005–2006 National Health and Nutrition Examination Survey (NHANES); the 2006 Health and Retirement Study (HRS); and the National Comorbidity Survey-Replication (NCS-R). A description of each of these survey data sets is provided in Section 1.3. A series of practical exercises based on these three data sets are included at the end of each chapter on an analysis topic to provide readers and students with examples enabling practice with using statistical software for applied survey data analysis.

Clear and consistent use of statistical notation is important. Table P.1 provides a summary of the general notational conventions used in this book. Special notation and symbol representation will be defined as needed for discussion of specific topics.

The materials and examples presented in the chapters of this book (which we refer to in subsequent chapters as ASDA) are supplemented through a companion Web site (http://www.isr.umich.edu/src/smp/asda/). This Web site provides survey analysts and instructors with additional resources in the following areas: links to new publications and an updated bibliography for the survey analysis topics covered in Chapters 5–12; links to sites for example survey data sets; replication of the command setups and output for the analysis examples in the SAS, SUDAAN, R, SPSS, and Mplus software systems; answers to frequently asked questions (FAQs); short technical

TABLE P.1

Notational Conventions for Applied Survey Data Analysis

Notation	Properties	Explanation of Usage
	Indices and Limits	
N, n	Standard usage	Population size, sample size
M, m	Standard usage	Subpopulation size, subpopulation sample size
h	Subscript	Stratum index (e.g., $\bar{y}_h$)
α	Subscript	Cluster or primary stage unit (PSU) index (e.g., $\bar{y}_{h\alpha}$)
i	Subscript	Element (respondent) index (e.g., $y_{h\alpha i}$)
j, k, l	Subscripts	Used to index vector or matrix elements (e.g., β_j)
	Survey Variables and Variable Values	
y, x	Roman, lowercase, italicized, end of alphabet	Survey variables (e.g., systolic blood pressure, mmHg; weight, kg)
Y_i, X_i	Roman, uppercase, end of alphabet, subscript	True population values of y, x for individual i, with $i = 1,\ldots, N$ comprising the population
y_i, x_i	Roman, lowercase, end of alphabet, subscript	Sample survey observation for individual i (e.g., $y_i = 124.5$ mmHg, $x_i = 80.2$ kg)
$\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Y}, \boldsymbol{X}$	As above, **bold**	Vectors (or matrices) of variables or variable values (e.g., $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$)
	Model Parameters and Estimates	
β_j, γ_j	Greek, lowercase	Regression model parameters, subscripts
$\hat{\beta}_j, \hat{\gamma}_j$	Greek, lowercase, "^" hat	Estimates of regression model parameters
$\boldsymbol{\beta}, \boldsymbol{\gamma}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$	As above, **bold**	Vectors (or matrices) of parameters or estimates (e.g., $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$)
$B_j, b_j, \boldsymbol{B}, \boldsymbol{b}$	Roman, otherwise as above	As above but used to distinguish finite population regression coefficients from probability model parameters and estimates
	Statistics and Estimates	
$\bar{Y}, P, \sigma_y^2, S_y^2, \bar{y}, p, s_y^2$	Standard usage	Population mean, proportion and variance; sample estimates as used in Cochran (1977)
$\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}$	Standard usage	Variance–covariance matrix; sample estimate of variance–covariance matrix
R^2, r, ψ	Standard usage	Multiple-coefficient of determination (R-squared), Pearson product moment correlation, odds ratio
ρ_y	Greek, lowercase	Intraclass correlation for variable y
Z, t, χ^2, F	Standard usage	Probability distributions

reports related to special topics in applied survey data analysis; and reviews of statistical software system updates and any resulting changes to the software commands or output for the analysis examples.

In closing, we must certainly acknowledge the many individuals who contributed directly or indirectly in the production of this book. Gail Arnold provided invaluable technical and organizational assistance throughout the production and review of the manuscript. Rod Perkins provided exceptional support in the final stages of manuscript review and preparation. Deborah Kloska and Lingling Zhang generously gave of their time and statistical expertise to systematically review each chapter as it was prepared. Joe Kazemi and two anonymous reviewers offered helpful comments on earlier versions of the introductory chapters, and SunWoong Kim and Azam Khan also reviewed the more technical material in our chapters for accuracy. We owe a debt to our many students in the JPSM and MPSM programs who over the years have studied with us—we only hope that you learned as much from us as we did from working with you. As lifelong students ourselves, we owe a debt to our mentors and colleagues who over the years have instilled in us a passion for statistical teaching and consultation: Leslie Kish, Irene Hess, Graham Kalton, Morton Brown, Edward Rothman, and Rod Little. Finally, we wish to thank the support staff at Chapman Hall/CRC Press, especially Rob Calver and Sarah Morris, for their continued guidance.

$\bar{y}_h$

Steven G. Heeringa
Brady T. West
Patricia A. Berglund
Ann Arbor, Michigan

1

Applied Survey Data Analysis: Overview

1.1 Introduction

Modern society has adopted the **survey method** as a principal tool for looking at itself—"a telescope on society" in the words of House et al. (2004). The most common application takes the form of the periodic media surveys that measure population attitudes and beliefs on current social and political issues:

> Recent international reports have said with near certainty that human activities are the main cause of global warming since 1950. The poll found that 84 percent of Americans see human activity as at least contributing to warming. (*New York Times,* April 27, 2007).

One step removed from the media limelight is the use of the survey method in the realms of marketing and consumer research to measure the preferences, needs, expectations, and experiences of consumers and to translate these to indices and other statistics that may influence financial markets or determine quality, reliability, or volume ratings for products as diverse as automobiles, hotel services, or TV programming:

> CBS won the overall title with an 8.8 rating/14 share in primetime, ABC finished second at 7.7/12…. (http://www.zap2it.com, January 11, 2008)

> The Index of Consumer Sentiment (see Figure 1.1) fell to 88.4 in the March 2007 survey from 91.3 in February and 96.9 in January, but it was nearly identical with the 88.9 recorded last March. (Reuters, University of Michigan, April 2007)

Also outside the view of most of society is the use of large-scale scientific surveys to measure labor force participation, earnings and expenditures, health and health care, commodity stocks and flows, and many other topics. These larger and longer-term programs of survey research are critically important to social scientists, health professionals, policy makers, and administrators and thus indirectly to society itself.

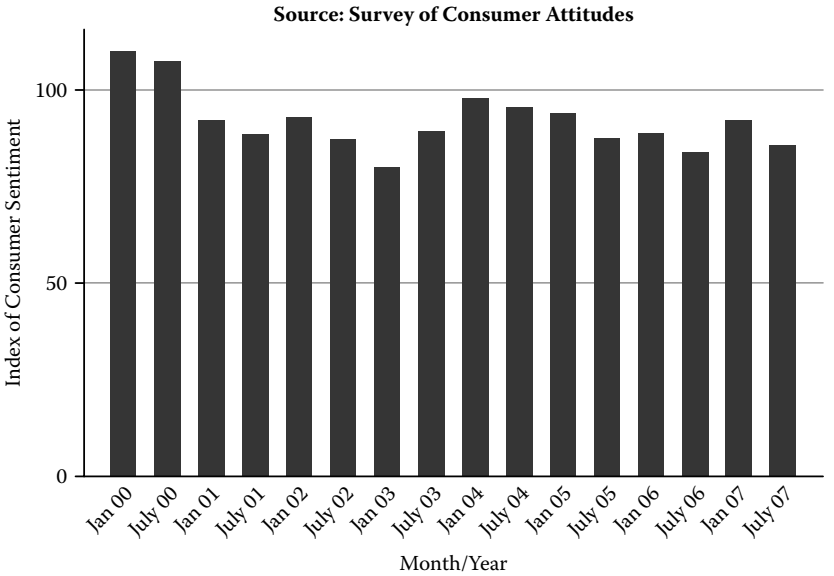


FIGURE 1.1
Index of Consumer Sentiment, January 2000–July 2007.

> Real median household income in the United States rose between 2005 and 2006, for the second consecutive year. Household income increased 0.7 percent, from $47,845 to $48,201. (DeNavas-Walt, Proctor, and Smith, 2007)

> In a series of logistic models that included age and one additional variable (i.e., education, gender, race, or APOE genotype), older age was consistently associated with increased risk of dementia ($p < 0.0001$). In these trivariate models, more years of education was associated with lower risk of dementia ($p < 0.0001$). There was no significant difference in dementia risk between males and females ($p = 0.26$). African Americans were at greater risk for dementia ($p = 0.008$). As expected, the presence of one (Odds Ratio = 2.1; 95% C.I. = 1.45 – 3.07) or two (O.R. = 7.1; 95% C.I. = 2.92 – 17.07) APOE e4 alleles was significantly associated with increased risk of dementia. (Plassman et al., 2007)

The focus of this book will be on analysis of **complex sample survey data** typically seen in large-scale scientific surveys, but the general approach to survey data analysis and specific statistical methods described here should apply to all forms of survey data.

To set the historical context for contemporary methodology, Section 1.2 briefly reviews the history of developments in theory and methods for applied survey data analysis. Section 1.3 provides some needed background on the data sets that will be used for the analysis examples in Chapters 2–12. This short overview chapter concludes in Section 1.4 with

a general review of the sequence of steps required in any applied analysis of survey data.

1.2 A Brief History of Applied Survey Data Analysis

Today's survey data analysts approach a problem armed with substantial background in statistical survey theory, a literature filled with empirical results and high-quality software tools for the task at hand. However, before turning to the best methods currently available for the analysis of survey data, it is useful to look back at how we arrived at where we are today. The brief history described here is certainly a selected interpretation, chosen to emphasize the evolution of probability sampling design and related statistical analysis techniques that are most directly relevant to the material in this book. Readers interested in a comprehensive review of the history and development of survey research in the United States should see Converse (1987). Bulmer (2001) provides a more international perspective on the history of survey research in the social sciences. For the more statistically inclined, Skinner, Holt, and Smith (1989) provide an excellent review of the development of methods for descriptive and analytical treatment of survey data. A comprehensive history of the impacts of sampling theory on survey practice can be found in O'Muircheartaigh and Wong (1981).

1.2.1 Key Theoretical Developments

The science of survey sampling, survey data collection methodology, and the analysis of survey data date back a little more than 100 years. By the end of the 19th century, an open and international debate established the **representative sampling method** as a statistically acceptable basis for the collection of observational data on populations (Kaier, 1895). Over the next 30 years, work by Bowley (1906), Fisher (1925), and other statisticians developed the role of randomization in sample selection and large-sample methods for estimation and statistical inference for simple random sample (SRS) designs.

The early work on the representative method and inference for simple random and stratified random samples culminated in a landmark paper by Jerzy Neyman (1934), which outlined a cohesive framework for estimation and inference based on estimated confidence intervals for population quantities that would be derived from the probability distribution for selected samples over repeated sampling. Following the publication of Neyman's paper, there was a major proliferation of new work on survey sample designs, estimation of population statistics, and variance estimation required to develop confidence intervals for sample-based inference, or what in more recent times has been labeled **design-based inference** (Cochran, 1977; Deming, 1950;

Hansen, Hurwitz, and Madow, 1953; Kish, 1965; Sukatme, 1954; Yates, 1949). House et al. (2004) credit J. Steven Stock (U.S. Department of Agriculture) and Lester Frankel (U.S. Bureau of the Census) with the first applications of area probability sampling methods for household survey data collections. Even today, the primary techniques for sample design, population estimation, and inference developed by these pioneers and published during the period 1945–1975 remain the basis for almost all descriptive analysis of survey data.

The developments of the World War II years firmly established the probability sample survey as a tool for describing population characteristics, beliefs, and attitudes. Based on Neyman's (1934) theory of inference, survey sampling pioneers in the United States, Britain, and India developed optimal methods for sample design, estimators of survey population characteristics, and confidence intervals for population statistics. As early as the late 1940s, social scientists led by sociologist Paul Lazarsfeld of Columbia University began to move beyond using survey data to simply describe populations to using these data to explore relationships among the measured variables (see Kendall and Lazarsfeld, 1950; Klein and Morgan, 1951). Skinner et al. (1989) and others before them labeled these two distinct uses of survey data as *descriptive* and *analytical*. Hyman (1955) used the term *explanatory* to describe scientific surveys whose primary purpose was the analytical investigation of relationships among variables.

During the period 1950–1990, analytical treatments of survey data expanded as new developments in statistical theory and methods were introduced, empirically tested, and refined. Important classes of methods that were introduced during this period included log-linear models and related methods for contingency tables, generalized linear models (e.g., logistic regression), survival analysis models, general linear mixed models (e.g., hierarchical linear models), structural equation models, and latent variable models. Many of these new statistical techniques applied the method of maximum likelihood to estimate model parameters and standard errors of the estimates, assuming that the survey observations were *independent* observations from a known probability distribution (e.g., binomial, multinomial, Poisson, product multinomial, normal). As discussed in Chapter 2, data collected under most contemporary survey designs do not conform to the key assumptions of these methods.

As Skinner et al. (1989) point out, survey statisticians were aware that straightforward applications of these new methods to complex sample survey data could result in underestimates of variances and therefore could result in biased estimates of confidence intervals and test statistics. However, except in limited situations of relatively simple designs, exact determination of the size and nature of the bias (or a potential correction) were difficult to express analytically. Early investigations of such "design effects" were primarily empirical studies, comparing design-adjusted variances for estimates with the variances that would be obtained if the

data were truly identically and independently distributed (equivalent to a simple random sample of equal size). Over time, survey statisticians developed special approaches to estimating these models that enabled the survey analyst to take into account the complex characteristics of the survey sample design (e.g., Binder, 1983; Kish and Frankel, 1974; Koch and Lemeshow, 1972; Pfeffermann et al., 1998; Rao and Scott, 1981). These approaches (and related developments) are described in Chapters 5–12 of this text.

1.2.2 Key Software Developments

Development of the underlying statistical theory and empirical testing of new methods were obviously important, but the survey data analyst needed computational tools to apply these techniques. We can have nothing but respect for the pioneers who in the 1950s fitted multivariate regression models to survey data using only hand computations (e.g., sums, sums of squares, sums of cross-products, matrix inversions) performed on a rotary calculator and possibly a tabulating machine (Klein and Morgan, 1951). The origin of statistical software as we know it today dates back to the 1960s, with the advent of the first mainframe computer systems. Software systems such as BMDP and OSIRIS and later SPSS, SAS, GLIM, S, and GAUSS were developed for mainframe users; however, with limited exceptions, these major software packages did not include programs that were adapted to complex sample survey data.

To fill this void during the 1970s and early 1980s, a number of stand-alone programs, often written in the Fortran language and distributed as compiled objects, were developed by survey statisticians (e.g., OSIRIS: PSALMS and REPERR, CLUSTERS, CARP, SUDAAN, WesVar). By today's standards, these programs had a steep "learning curve," limited data management flexibility, and typically supported only descriptive analysis (means, proportions, totals, ratios, and functions of descriptive statistics) and linear regression modeling of multivariate relationships. A review of the social science literature of this period shows that only a minority of researchers actually employed these special software programs when analyzing complex sample survey data, resorting instead to standard analysis programs with their default assumption that the data originated with a simple random sample of the survey population.

The appearance of microcomputers in the mid-1980s was quickly followed by a transition to personal computer versions of the major statistical software (BMDP, SAS, SPSS) as well as the advent of new statistical analysis software packages (e.g., SYSTAT, Stata, S-Plus). However, with the exception of specialized software systems (WesVar PC, PC CARP, PC SUDAAN, MicroOSIRIS, CLUSTERS for PC, IVEware) that were often designed to read data sets stored in the formats of the larger commercial software packages, the microcomputing revolution still did not put tools for the analysis of complex

sample survey data in the hands of most survey data analysts. Nevertheless, throughout the late 1980s and early 1990s, the scientific and commercial pressures to incorporate programs of this type into the major software systems were building. Beginning with Version 6.12, SAS users had access to PROC SURVEYMEANS and PROC SURVEYREG, two new SAS procedures that permitted simple descriptive analysis and linear regression analysis for complex sample survey data. At about the same time, the Stata system for statistical analysis appeared on the scene, providing complex sample survey data analysts with the "`svy`" versions of the more important analysis programs. SPSS's entry into the world of complex sample survey data analysis came later with the introduction of the Complex Samples add-on module in Version 13. Appendix A of this text covers the capabilities of these different systems in detail.

The survey researcher who sits down today at his or her personal computing work station has access to powerful software systems, high-speed processing, and high-density data storage capabilities that the analysts in the 1970s, 1980s, and even the 1990s could not have visualized. All of these recent advances have brought us to a point at which today's survey analyst can approach both simple and complex problems with the confidence gained through a fundamental understanding of the theory, empirically tested methods for design-based estimation and inference, and software tools that are sophisticated, accurate, and easy to use.

Now that we have had a glimpse at our history, let's begin our study of applied survey data analysis.

1.3 Example Data Sets and Exercises

Examples based on the analysis of major survey data sets are routinely used in this book to demonstrate statistical methods and software applications. To ensure diversity in sample design and substantive content, example exercises and illustrations are drawn from three major U.S. survey data sets.

1.3.1 The National Comorbidity Survey Replication (NCS-R)

The NCS-R is a 2002 study of mental illness in the U.S. household population ages 18 and over. The core content of the NCS-R is based on a lay-administered interview using the World Health Organization (WHO) CIDI (Composite International Diagnostic Interview) diagnostic tool, which is designed to measure primary mental health diagnostic symptoms, symptom severity, and use of mental health services (Kessler et al., 2004). The NCS-R was based on interviews with randomly chosen adults in an equal probability, multistage sample of households selected from the University of Michigan

National Sample master frame. The survey response rate was 70.9%. The survey was administered in two parts: a Part I core diagnostic assessment of all respondents (n = 9,282), followed by a Part II in-depth interview with 5,692 of the 9,282 Part I respondents, including all Part I respondents who reported a lifetime mental health disorder and a probability subsample of the disorder-free respondents in the Part I screening.

The NCS-R was chosen as an example data set for the following reasons: (1) the scientific content and, in particular, its binary measures of mental health status; (2) the multistage design with primary stage stratification and clustering typical of many large-scale public-use survey data sets; and (3) the two-phase aspect of the data collection.

1.3.2 The Health and Retirement Study (HRS)—2006

The Health and Retirement Study (HRS) is a longitudinal study of the American population 50 years of age and older. Beginning in 1992, the HRS has collected data every two years on a longitudinal panel of sample respondents born between the years of 1931 and 1941. Originally, the HRS was designed to follow this probability sample of age-eligible individuals and their spouses or partners as they transitioned from active working status to retirement, measuring aging-related changes in labor force participation, financial status, physical and mental health, and retirement planning. The HRS observation units are age-eligible individuals and "financial units" (couples in which at least one spouse or partner is HRS eligible). Beginning in 1993 and again in 1998 and 2004, the original HRS 1931–1941 birth cohort panel sample was augmented with probability samples of U.S. adults and spouses/partners from (1) pre-1924 (added in 1993); (2) 1924–1930 and 1942–1947 (added in 1998); and (3) 1948–1953 (added in 2004). In 2006, the HRS interviewed over 22,000 eligible sample adults in the composite panel.

The HRS samples were primarily identified through in-person screening of large, multistage area probability samples of U.S. households. For the pre-1931 birth cohorts, the core area probability sample screening was supplemented through sampling of age-eligible individuals from the U.S. Medicare Enrollment Database. Sample inclusion probabilities for HRS respondents vary slightly across birth cohorts and are approximately two times higher for African Americans and Hispanics. Data from the 2006 wave of the HRS panel are used for most of the examples in this text, and we consider a longitudinal analysis of multiple waves of HRS data in Chapter 12.

1.3.3 The National Health and Nutrition Examination Survey (NHANES)—2005, 2006

Sponsored by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC), the NHANES is a survey of the adult, noninstitutionalized population of the United States. The NHANES

is designed to study the prevalence of major disease in the U.S. population and to monitor the change in prevalence over time as well as trends in treatment and major disease risk factors including personal behaviors, environmental exposure, diet, and nutrition. The NHANES survey includes both an in-home medical history interview with sample respondents and a detailed medical examination at a local mobile examination center (MEC). The NHANES surveys were conducted on a periodic basis between 1971 and 1994 (NHANES I, II, III), but beginning in 1999, the study transitioned to a continuous interviewing design. Since 1999, yearly NHANES data collections have been performed in a multistage sample that includes 15 primary stage unit (PSU) locations with new sample PSUs added in each data collection year. Approximately 7,000 probability sample respondents complete the NHANES in-home interview phase each year and roughly 5,000 of these individuals also consent to the detailed MEC examination. To meet specific analysis objectives, the NHANES oversamples low-income persons, adolescents between the ages of 12 and 19, persons age 60 and older, African Americans, and Hispanics of Mexican ancestry. To ensure adequate precision for sample estimates, NCHS recommends pooling data for two or more consecutive years of NHANES data collection. The NHANES example analyses provided in this text are based on the combined data collected in 2005 and 2006. The unweighted response rate for the interview phase of the 2005–2006 NHANES was approximately 81%.

Public use versions of each of these three major survey data sets are available online. The companion Web site for this book provides the most current links to the official public use data archives for each of these example survey data sets.

1.3.4 Steps in Applied Survey Data Analysis

Applied survey data analysis—both in daily practice and here in this book—is a process that requires more of the analyst than simple familiarity and proficiency with statistical software tools. It requires a deeper understanding of the sample design, the survey data, and the interpretation of the results of the statistical methods. Following a more general outline for applied statistical analysis presented by Cox (2007), Figure 1.2 outlines a sequence of six steps that are fundamental to applied survey data analysis, and we describe these steps in more detail in the following sections.

1.3.4.1 Step 1: Definition of the Problem and Statement of the Objectives

The first of the six steps involves a clear specification of the problem to be addressed and formulation of objectives for the analysis exercise. For example, the "problem" may be ambiguity among physicians over whether there should be a lower threshold for prostate biopsy following prostate specific antigen (PSA) screening in African American men (Cooney et al., 2001). The

Step	Activity
1	Definition of the problem and statement of the objectives.
2	Understanding the sample design.
3	Understanding design variables, underlying constructs, and missing data.
4	Analyzing the data.
5	Interpreting and evaluating the results of the analysis.
6	Reporting of estimates and inferences from the survey data.

FIGURE 1.2
Steps in applied survey data analysis.

corresponding objective would be to estimate the 95th percentile and the 95% confidence bounds for this quantity (+/– .2 ng/ml PSA) in a population of African American men. The estimated 95% confidence bounds can in turn be used by medical experts to determine if the biopsy threshold for African American men should be different than for men of other race and ethnic groups.

As previously described, the problems to which survey data analyses may be applied span many disciplines and real-world settings. Likewise, the statistical objectives may vary. Historically, the objectives of most survey data analyses were to describe characteristics of a target population: its average household income, the median blood pressure of men, or the proportion of eligible voters who favor candidate X. But survey data analyses can also be used for decision making. For example, should a pharmaceutical company recall its current products from store shelves due to a perceived threat of contamination? In a population case-control study, does the presence of silicone breast implants significantly increase the odds that a woman will contract a connective tissue disease such as scleroderma (Burns et al., 1996)? In recent decades, the objective of many sample survey data analyses has been to explore and extend the understanding of multivariate relationships among variables in the target population. Sometimes multivariate modeling of survey data is seen simply as a descriptive tool, defining the form of a functional relationship as it exists in a finite population. But it is increasingly common for researchers to use observational data from complex sample surveys to probe causality in the relationships among variables.

1.3.4.2 Step 2: Understanding the Sample Design

The survey data analyst must understand the sample design that was used to collect the data he or she is about to analyze. Without an understanding of key properties of the survey sample design, the analysis may be inefficient,

biased, or otherwise lead to incorrect inference. An experienced researcher who designs and conducts a randomized block experimental design to test the relative effectiveness of new instructional methods should not proceed to analyze the data as a simple factorial design, ignoring the blocking that was built into his or her experiment. Likewise, an economics graduate student who elects to work with the longitudinal HRS data should understand that the nationally representative sample of older adults includes stratification, clustering, and disproportionate sampling (i.e., compensatory population weighting) and that these design features may require special approaches to population estimation and inference.

At this point, we may have discouraged the reader into thinking that an in-depth knowledge of survey sample design is required to work with survey data or that he or she may need to relearn what was studied in general courses on applied statistical methods. This is not the case. Chapters 2 through 4 will introduce the reader to the fundamental features of **complex sample designs** and will demonstrate how design characteristics such as stratification, clustering, and weighting are easily incorporated into the statistical methods and software for survey estimation and inference. Chapters 5–12 will show the reader that relatively simple extensions of his or her current knowledge of applied statistical analysis methods provide the necessary foundation for efficient and accurate analysis of data collected in sample surveys.

1.3.4.3 Step 3: Understanding Design Variables, Underlying Constructs, and Missing Data

The typical scientific survey data set is **multipurpose**, with the final data sets often including hundreds of variables that span many domains of study—income, education, health, family. The sheer volume of available data and the ease by which it can be accessed can cause survey data analysts to become complacent in their attempts to fully understand the properties of the data that are important to their choice of statistical methods and the conclusions that they will ultimately draw from their analysis. Step 2 described the importance of understanding the sample design. In the survey data, the key features of the sample design will be encoded in a series of **design variables**. Before analysis begins, some simple questions need to be put to the candidate data set: What are the empirical distributions of these design variables, and do they conform to the design characteristics outlined in the technical reports and online study documentation? Does the original survey question that generated a variable of interest truly capture the underlying construct of interest? Are the response scales and empirical distributions of responses and independent variables suitable for the intended analysis? What is the distribution of missing data across the cases and variables, and is there a potential impact on the analysis and the conclusions that will be drawn?

Chapter 4 discusses techniques for answering these and other questions before proceeding to statistical analysis of the survey data.

1.3.4.4 Step 4: Analyzing the Data

Finally we arrive at the step to which many researchers rush to enter the process. We are all guilty of wanting to jump ahead. Identifying the problem and objectives seems intuitive. We tell ourselves that formalizing that step wastes time. Understanding the design and performing data management and exploratory analysis to better understand the data structure is boring. After all, the statistical analysis step is where we obtain the results that enable us to describe populations (through confidence intervals), to extend our understanding of relationships (through statistical modeling), and possibly even to test scientific hypotheses.

In fact, the statistical analysis step lies at the heart of the process. Analytic techniques must be carefully chosen to conform to the analysis objectives and the properties of the survey data. Specific methodology and software choices must accommodate the design features that influence estimation and inference. Treatment of statistical methods for survey data analysis begins in Chapters 5 and 6 with coverage of univariate (i.e., single-variable) descriptive and simple bivariate (i.e., two-variable) analyses of continuous and categorical variables. Chapter 7 presents the linear regression model for continuous dependent variables, and generalized linear regression modeling methods for survey data are treated in Chapters 8 and 9. Chapter 10 pertains to methods for event-history analysis of survey data, including models such as the Cox proportional hazard model and discrete time logistic models. Chapter 11 introduces methods for handling missing data problems in survey data sets. Finally, the coverage of statistical methods for survey data analysis concludes with a discussion of new developments in the area of survey applications of advanced statistical techniques, such as multilevel analysis, in Chapter 12.

1.3.4.5 Step 5: Interpreting and Evaluating the Results of the Analysis

Knowledge of statistical methods and software tools is fundamental to success as an applied survey data analyst. However, setting up the data, running the programs, and printing the results are not sufficient to constitute a thorough treatment of the analysis problem. Likewise, scanning a column of p-values in a table of regression model output does not inform us concerning the form of the "final model" or even the pure effect of a single predictor. As described in Step 3, interpretation of the results from an analysis of survey data requires a consideration of the error properties of the data. Variability of sample estimates will be reflected in the **sampling errors** (i.e., confidence intervals, test statistics) estimated in the course of the statistical analysis. **Nonsampling errors**, including potential bias due to survey nonresponse

and item missing data, cannot be estimated from the survey data (Lessler and Kalsbeek, 1992). However, it may be possible to use ancillary data to explore the potential direction and magnitude of such errors. For example, an analyst working for a survey organization may statistically compare survey respondents with nonrespondents in terms of known correlates of key survey variables that are readily available on the sampling frame to assess the possibility of nonresponse bias.

As survey data analysts have pushed further into the realm of multivariate modeling of survey data, care is required in interpreting fitted models. Is the model reasonably identified, and do the data meet the underlying assumptions of the model estimation technique? Are there alternative models that explain the observed data equally well? Is there scientific support for the relationship implied in the modeling results? Are interpretations that imply causality in the modeled relationships supported (Rothman, 1988)?

1.3.4.6 Step 6: Reporting of Estimates and Inferences from the Survey Data

The end products of applied survey data analyses are reports, papers, or presentations designed to communicate the findings to fellow scientists, policy analysts and administrators and decision makers. This text includes discussion of standards and proven methods for effectively presenting the results of applied survey data analyses, including table formatting, statistical contents, and the use of statistical graphics.

With these six steps in mind, we now can now begin our walk through the process of planning, formulating, and conducting analysis of survey data.

2

Getting to Know the Complex Sample Design

2.1 Introduction

The first step in the applied analysis of survey data involves defining the research questions that will be addressed using a given set of survey data. The next step is to study and understand the sampling design that generated the sample of elements (e.g., persons, businesses) from the target population of interest, given that the actual survey data with which the reader will be working were collected from the elements in this sample. This chapter aims to help the readers understand the complex sample designs that they are likely to encounter in practice and identify the features of the designs that have important implications for correct analyses of the survey data.

Although a thorough knowledge of sampling theory and methods can benefit the survey data analyst, it is not a requirement. With a basic understanding of complex sample design features, including stratification, clustering, and weighting, readers will be able to specify the key design parameters required by today's survey data analysis software systems. Readers who are interested in a more in-depth treatment of sampling theory and methods are encouraged to review work by Hansen, Hurwitz, and Madow (1953), Kish (1965), or Cochran (1977). More recent texts that blend basic theory and applications include Levy and Lemeshow (2007) and Lohr (1999). A short monograph by Kalton (1983) provides an excellent summary of survey sample designs.

The sections in this chapter outline the key elements of complex sample designs that analysts need to understand to proceed knowledgeably and confidently to the next step in the analysis process.

2.1.1 Technical Documentation and Supplemental Literature Review

The path to understanding the complex sample design and its importance to the reader's approach to the analysis of the survey data should begin with a review of the technical documentation for the survey and the sample design. A review of the literature, including both supplemental

methodological reports and papers that incorporate actual analysis of the survey data, will be quite beneficial for the reader's understanding of the data to be analyzed.

Technical documentation for the sample design, weighting, and analysis procedures should be part of the "metadata" that are distributed with a survey data set. In the real world of survey data, the quality of this technical documentation can be highly variable, but, at a minimum, the reader should expect to find a summary description of the sample, a discussion of weighting and estimation procedures, and, ideally, specific guidance on how to perform complex sample analysis of the survey data. Readers who plan to analyze a public use survey data set but find that documentation of the design is lacking or inadequate should contact the help desk for the data distribution Web site or inquire with the study team directly to obtain or clarify the basic information needed to correctly specify the sample design when analyzing the survey data.

Before diving into the statistical analysis of a survey data set, time is well spent in reviewing supplemental methodological reports or published scientific papers that used the same data. This review can identify important new information or even guide readers' choice of an analytic approach to the statistical objectives of their research.

2.2 Classification of Sample Designs

As illustrated in Figure 2.1, Hansen, Madow, and Tepping (1983) define a sampling design to include two components: the sampling plan and a method for drawing inferences from the data generated under the sampling plan. The vast majority of survey data sets will be based on sample designs that fall in Cell A of Figure 2.1—that is, designs that include a sampling plan based on probability sample methods and assume that statistical inferences concerning population characteristics and relationships will be derived using the "design-based" theory initially proposed by Neyman

Sampling Plan	Method of Inference	
	Design-based	Model-based
Probability Sample	A	B
Model-dependent Sample	C	D

FIGURE 2.1
Classification of sample designs for survey data.

(1934). Consequently, in this book we will focus almost exclusively on sample designs of this type.

2.2.1 Sampling Plans

Probability sampling plans assign each member of the population a known nonzero probability of inclusion in the sample. A probability sample plan may include features such as stratification and clustering of the population prior to selection—it does not require that the selection probability of one population element be independent of that for another. Likewise, the sample inclusion probabilities for different population elements need not be equal. In a probability sampling of students in a coeducational secondary school, it is perfectly appropriate to sample women at a higher rate than men with the requirement that weights will be needed to derive unbiased estimates for the combined population. Randomization of sample choice is always introduced in probability sampling plans.

Model-dependent sampling plans (Valliant et al., 2000) assume that the variables of interest in the research follow a known probability distribution and optimize the choice of sample elements to maximize the precision of estimation for statistics of interest. For example, a researcher interested in estimating total annual school expenditures, y, for teacher salaries may assume the following relationship of the expenditures to known school enrollments, x: $y_i = \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2 x_i)$. Model-dependent sampling plans may employ stratification and clustering, but strict adherence to randomized selection is not a requirement. Model-dependent sampling plans have received considerable attention in the literature on sampling theory and methods; however, they are not common in survey practice due to a number of factors: the multipurpose nature of most surveys; uncertainty over the model; and lack of high-quality ancillary data (e.g., the x variable in the previous example).

Though not included in Figure 2.1, **quota sampling**, **convenience sampling**, **snowball sampling**, and "**peer nomination**" are nonprobability methods for selecting sample members (Kish, 1965). Practitioners who employ these methods base their choices on assumptions concerning the "representativeness" of the selection process and often analyze the survey data using inferential procedures that are appropriate for probability sample plans. However, these sampling plans do not adhere to the fundamental principles of either probability sample plans or a rigorous probability model-based approach. Is it impossible to draw correct inferences from nonprobability sample data? No, because by chance an arbitrary sample could produce reasonable results; however, the survey analyst is left with no theoretical basis to measure the variability and bias associated with his or her inferences. Survey data analysts who plan to work with data collected under nonprobability sampling plans should carefully evaluate and report the potential selection biases and other survey errors that could affect their final results. Since there is no true

statistical basis of support for the analysis of data collected under these nonprobability designs, they will not be addressed further in this book.

2.2.2 Inference from Survey Data

In the Hansen et al. (1983) framework for sample designs, a sampling plan is paired with an approach for deriving inferences from the data that are collected under the chosen plan. **Statistical inferences** from sample survey data may be "design-based" or "model-based." The natural design pairings of sampling plans and methods of inference are the diagonal cells of Figure 2.1 (A and D); however, the hybrid approach of combining probability sample plans with model-based approaches to inference is not uncommon in survey research. Both approaches to statistical inference use probability models to establish the correct form of confidence intervals or hypothesis tests for the intended statistical inference. Under the "design-based" or "randomization-based" approach formalized by Neyman (1934), the inferences are derived based on the distribution of all possible samples that could have been chosen under the sample design. This approach is sometimes labeled "nonparametric" or "distribution free" because it relies only on the known probability that a given sample was chosen and not on the specific probability distribution for the underlying variables of interest, for example, $y \sim N(\beta_0 + \beta_1 x, \sigma^2_{y\cdot x})$ (see Chapter 3).

As the label implies, model-based inferences are based on a probability distribution for the random variable of interest—not the distribution of probability for the sample selection. Estimators, standard errors, confidence intervals, and hypothesis tests for the parameters of the distribution (e.g., means, regression coefficients, and variances) are typically derived using the method of maximum likelihood or possibly Bayesian models (see Little, 2003).

2.3 Target Populations and Survey Populations

The next step in becoming acquainted with the sample design and its implications for the reader's analysis is to verify the survey designers' intended study population—who, when, and where. Probability sample surveys are designed to describe a **target population**. The target populations for survey designs are **finite populations** that may range from as few as a 100 **population elements** for a survey of special groups to millions and even billions for national population surveys. Regardless of the actual size, each discrete population element ($i = 1, \ldots, N$) could, in theory, be counted in a census or sampled for survey observation.

In contrast to the target population, the **survey population** is defined as the population that is truly eligible for sampling under the survey design

(Groves et al., 2004). In survey sampling practice, there are geographical, political, social, and temporal factors that restrict our ability to identify and access individual elements in the complete target population and the de facto **coverage** of the survey is limited to the survey population. Examples of geographic restrictions on the survey population could include persons living in remote, sparsely populated areas such as islands, deserts, or wilderness areas. Rebellions, civil strife, and governmental restrictions on travel can limit access to populations living in the affected areas. Homelessness, institutionalization, military service, nomadic occupations, physical and mental conditions, and language barriers are social and personal factors that can affect the coverage of households and individuals in the target population. The timing of the survey can also affect the coverage of the target population. The target population definition for a survey assumes that the data are collected as a "snapshot" in time when in fact the data collection may span many months.

The target population for the National Comorbidity Survey Replication (NCS-R) is defined to be age 18 and older adults living in the households in the United States as of July 1, 2002. Here is the exact definition of the survey population for the NCS-R:

> The survey population for the NCS-R included all U.S. adults aged 18+ years residing in households located in the coterminous 48 states. Institutionalized persons including individuals in prisons, jails, nursing homes, and long-term medical or dependent care facilities were excluded from the survey population. Military personnel living in civilian housing were eligible for the study, but due to security restrictions residents of housing located on a military base or military reservation were excluded. Adults who were not able to conduct the NCS-R interview in English were not eligible for the survey. (Heeringa et al., 2004)

Note that among the list of exclusions in this definition, the NCS-R survey population excludes residents of Alaska and Hawaii, institutionalized persons, and non-English speakers. Furthermore, the survey observations were collected over a window of time that spanned several years (February 2001 to April 2003). For populations that remain stable and relatively unchanged during the survey period, the time lapse required to collect the data may not lead to bias for target population estimates. However, if the population is mobile or experiences seasonal effects in terms of the survey variables of interest, considerable change can occur during the window of time that the survey population is being observed.

As the survey data analyst, the reader will also be able to restrict his or her analysis to **subpopulations** of the survey population represented in the survey data set, but the analysis can use only the available data and cannot directly reconstruct coverage of the unrepresented segments of the target population. Therefore, it is important to carefully review the definition of

the survey population and assess the implications of any exclusions for the inferences to be drawn from the analysis.

2.4 Simple Random Sampling: A Simple Model for Design-Based Inference

Simple random sampling with replacement (SRSWR) is the most basic of sampling plans, followed closely in theoretical simplicity by simple random sampling without replacement (SRSWOR, or the short form, SRS, in this text). Survey data analysts are unlikely to encounter true SRS designs in survey practice. Occasionally, SRS may be used to select samples of small localized populations or samples of records from databases or file systems, but this is rare. Even in cases where SRS is practicable, survey statisticians will aim to introduce simple stratification to improve the efficiency of sample estimates (see the next sections). Furthermore, if an SRS is selected but weighting is required to compensate for nonresponse or to apply poststratification adjustments (see Section 2.7), the survey data now include complex features that cannot be ignored in estimation and inference.

2.4.1 Relevance of SRS to Complex Sample Survey Data Analysis

So why is SRS even relevant for the material in this book? There are several reasons:

1. SRS designs produce samples that most closely approximate the assumptions (**i.i.d.**—observations are *independent* and *identical in distribution*) defining the theoretical basis for the estimation and inference procedures found in standard analysis programs in the major statistical software systems. Survey analysts who use the standard programs in Stata, SAS, and SPSS are essentially defaulting to the assumption that their survey data were collected under SRS. In general, the SRS assumption results in underestimation of variances of survey estimates of descriptive statistics and model parameters. Confidence intervals based on computed variances that assume independence of observations will be biased (generally too narrow), and design-based inferences will be affected accordingly. Likewise, test statistics (t, χ^2, F) computed in complex sample survey data analysis using standard programs will tend to be biased upward and overstate the significance of tests of effects.
2. The theoretical simplicity of SRS designs provides a basic framework for design-based estimation and inference on which to build a bridge to the more complicated approaches for complex samples.

3. SRS provides a comparative benchmark that can be used to evaluate the relative efficiency of the more complex designs that are common in survey practice.

Let's examine the second reason more closely, using SRS as a theoretical framework for design-based estimation and inference. In Section 2.5, we will turn to SRS as a benchmark for the efficiency of complex sample designs.

2.4.2 SRS Fundamentals: A Framework for Design-Based Inference

Many students of statistics were introduced to simple random sample designs through the example of an urn containing a population of blue and red balls. To estimate the proportion of blue balls in the urn, the instructor described a sequence of random draws of $i = 1, \ldots, n$ balls from the N balls in the urn. If a drawn ball was returned to the urn before the next draw was made, the sampling was "with replacement" (SRSWR). If a selected ball was not returned to the urn until all n random selections were completed, the sampling was "without replacement" (SRSWOR).

In each case, the SRSWR or SRSWOR sampling procedure assigned each population element an equal probability of sample selection, $f = n/N$. Furthermore, the overall probability that a specific ball was selected to the sample was independent of the probability of selection for any of the remaining $N - 1$ balls in the urn. Obviously, in survey practice random sampling is not typically performed by drawing balls from an urn. Instead, survey sampling uses devices such as tables of random numbers or computerized random number generators to select sample elements from the population.

Let's assume that the objective of the sample design was to estimate the mean of a characteristic, y, in the population:

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{2.1}$$

Under simple random sampling, an unbiased estimate of the population mean is the sample mean:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.2}$$

The important point to note here is that there is a true population parameter of interest, $\bar{Y}$, and the estimate of the parameter, $\bar{y}$, which can be derived

from the sample data. The sample estimate $\bar{y}$ is subject to sampling variability, denoted as $Var(\bar{y})$, from one sample to the next. Another measure of the sampling variability in sample estimates is termed the standard error, or $SE(\bar{y}) = \sqrt{Var(\bar{y})}$. For simple random samples (SRS) selected from large populations, across all possible samples of size n, the standard error for the estimated population proportion is calculated as follows:

$$SE(\bar{y}) = \sqrt{Var(\bar{y})} = \sqrt{(1 - n/N) \cdot \frac{S^2}{n}} \approx \sqrt{\frac{S^2}{n}} \text{ if } N \text{ is large} \tag{2.3}$$

where
$S^2 = \Sigma_{i=1}^{N}(Y_i - \bar{Y})^2 / (N-1)$;
n = SRS sample size; and
N = the population size.

Since we observe only a single sample and not all possible samples of size n from the population of N, the true $SE(\bar{y})$ must be estimated from the data in our chosen sample:

$$se(\bar{y}) = \sqrt{var(\bar{y})} = \sqrt{(1 - n/N) \cdot \frac{s^2}{n}} \approx \sqrt{\frac{s^2}{n}} \text{ if } N \text{ is large} \tag{2.4}$$

where
$s^2 = \Sigma_{i=1}^{n}(y_i - \bar{y})^2 / (n-1)$;
n = SRS sample size; and
N = the population size.

The term $(1 - n/N)$ in the expressions for $SE(\bar{y})$ and $se(\bar{y})$ is the **finite population correction (fpc)**. It applies only where selection of population elements is without replacement (see Theory Box 2.1) and is generally assumed to be equal to 1 in practice if $f = n/N < 0.05$.

If the sample size, n, is large, then under Neyman's (1934) method of design-based inference, a 95% confidence interval for the true population mean, $\bar{Y}$, can be constructed as follows:

$$\bar{y} \pm t_{0.975,n-1} \cdot se(\bar{y}) \tag{2.5}$$

THEORY BOX 2.1 THE FINITE POPULATION CORRECTION (fpc)

The fpc reflects the expected reduction in the sampling variance of a survey statistic due to sampling without replacement (WOR). For an SRS sample design, the fpc factor arises from the algebraic derivation of the expected sampling variance of a survey statistic over all possible WOR samples of size n that could be selected from the population of N elements (see Cochran, 1977).

In most practical survey sampling situations, the population size, N, is very large, and the ratio n/N is so close to zero that the fpc ~ 1.0. As a result, the fpc can be safely ignored in the estimation of the standard error of the sample estimate. Since complex samples may also employ sampling without replacement at one or more stages of selection, in theory, variance estimation for these designs should also include finite population corrections. Where applicable, software systems such as Stata and SUDAAN provide users with the flexibility to input population size information and incorporate the fpc values in variance estimation for complex sample survey data. Again, in most survey designs, the size of the population at each stage of sampling is so large that the fpc factors can be safely ignored.

2.4.3 An Example of Design-Based Inference under SRS

To illustrate the simple steps in design-based inference from a simple random sample, Table 2.1 presents a hypothetical sample data set of n = 32 observations from a very large national adult population (because the sampling fraction, n/N, is small, the fpc will be ignored). Each subject was asked to rate his or her view of the strength of the national economy (y) on a 0–100 scale, with 0 representing the weakest possible rating and 100 the strongest possible rating. The sample observations are drawn from a large population with population mean $\bar{Y} = 40$ and population variance $S_y^2 \cong 12.80^2 = 164$. The individual case identifiers for the sample observations are provided in Column (1) of Table 2.1. For the time being, we can ignore the columns labeled Stratum, Cluster, and Case Weight.

If we assume that the sample was selected by SRS, the sample estimates of the mean, its standard error, and the 95% confidence interval would be calculated as follows:

$$\bar{y} = \sum_{i=1}^{n} y_i / n = \sum_{i=1}^{32} y_i / 32 = 40.77$$

$$se(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{\sum_{i=1}^{32} (y_i - \bar{y})^2 / \ [n \cdot (n-1)]} = 2.41$$

$$95\%\ CI = \bar{y} \pm t_{.975,31} \cdot se(\bar{y}) = (35.87, 45.68)$$

TABLE 2.1

Sample Data Set for Sampling Plan Comparisons

Case No. (1)	Stratum (2)	Cluster (3)	Economy Rating score, y_i (4)	Case Weight, w_i (5)
1	1	1	52.8	1
2	1	1	32.5	2
3	1	1	56.6	2
4	1	1	47.0	1
5	1	2	37.3	1
6	1	2	57.0	1
7	1	2	54.2	2
8	1	2	71.5	2
9	2	3	27.7	1
10	2	3	42.3	2
11	2	3	32.2	2
12	2	3	35.4	1
13	2	4	48.8	1
14	2	4	66.8	1
15	2	4	55.8	2
16	2	4	37.5	2
17	3	5	49.4	2
18	3	5	14.9	1
19	3	5	37.3	1
20	3	5	41.0	2
21	3	6	45.9	2
22	3	6	39.9	2
23	3	6	33.5	1
24	3	6	54.9	1
25	4	7	26.4	2
26	4	7	31.6	2
27	4	7	32.9	1
28	4	7	11.1	1
29	4	8	30.7	2
30	4	8	33.9	1
31	4	8	37.7	1
32	4	8	28.1	2

The fundamental results illustrated here for SRS are that, for a sample of size n, unbiased estimates of the population value and the standard error of the sample estimate can be computed. For samples of reasonably large size, a confidence interval for the population parameter of interest can be derived. As our discussion transitions to more complex samples and more complex statistics we will build on this basic framework for constructing confidence

intervals for population parameters, adapting each step to the features of the sample design and the analysis procedure.

2.5 Complex Sample Design Effects

Most practical sampling plans employed in scientific surveys are not SRS designs. Stratification is introduced to increase the statistical and administrative efficiency of the sample. Sample elements are selected from naturally occurring clusters of elements in multistage designs to reduce travel costs and improve interviewing efficiency. Disproportionate sampling of population elements may be used to increase the sample sizes for subpopulations of special interest, resulting in the need to employ weighting in the descriptive estimation of population statistics. All of these features of more commonly used sampling plans will have effects on the accuracy and precision of survey estimators, and we discuss those effects in this section.

2.5.1 Design Effect Ratio

Relative to SRS, the need to apply weights to complex sample survey data changes the approach to estimation of population statistics or model parameters. Also relative to SRS designs, stratification, clustering, and weighting all influence the size of standard errors for survey estimates. Figure 2.2 illustrates the general effects of these design features on the standard errors of survey estimates. The curve plotted in this figure represents the SRS standard

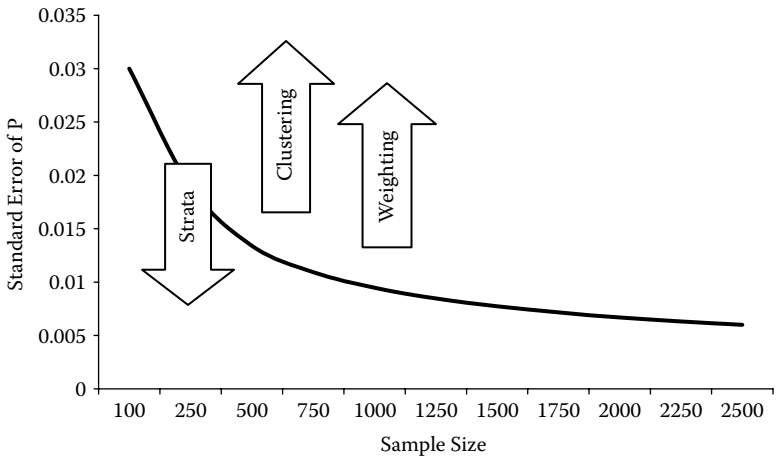


FIGURE 2.2
Complex sample design effects on standard errors.

error of a sample estimate of a proportion P as a function of the sample size n. At any chosen sample size, the effect of sample stratification is generally a reduction in standard errors relative to SRS. Clustering of sample elements and designs that require weighting for unbiased estimation generally tend to yield estimates with larger standard errors than an SRS sample of equal size.

Relative to an SRS of equal size, the complex effects of stratification, clustering, and weighting on the standard errors of estimates are termed the **design effect** and are measured by the following ratio (Kish, 1965):

$$D^2(\hat{\theta}) = \frac{SE(\hat{\theta})^2_{complex}}{SE(\hat{\theta})^2_{srs}} = \frac{Var(\hat{\theta})_{complex}}{Var(\hat{\theta})_{srs}} \tag{2.6}$$

where

$D^2(\hat{\theta})$ = the design effect for the sample estimate, $\hat{\theta}$;
$Var(\hat{\theta})_{complex}$ = the complex sample design variance of $\hat{\theta}$; *and*
$Var(\hat{\theta})_{srs}$ = the simple random sample variance of $\hat{\theta}$.

A somewhat simplistic but practically useful model of design effects that survey statisticians may use to plan a sample survey is

$$D^2(\hat{\theta}) \approx 1 + f(G_{strat}, L_{cluster}, L_{weighting})$$

where

G_{strat} = the relative gain in precision from stratified sampling compared to SRS;
$L_{cluster}$ = the relative loss of precision due to clustered selection of sample elements;
$L_{weighting}$ = the relative loss due to unequal weighting for sample elements.

The value of the design effect for a particular sample design will be the net effect of the combined influences of stratification, clustering, and weighting. In Sections 2.5 to 2.7 we will introduce very simple models that describe the nature and rough magnitude of the effects attributable to stratification, clustering, and weighting. In reality, the relative increase in variance measured by D^2 will be a complex and most likely nonlinear function of the influences of stratification, clustering, and weighting and their interactions. Over the years, there have been a number of attempts to analytically quantify the anticipated design effect for specific complex samples, estimates, and subpopulations (Skinner, Holt, and Smith, 1989). While these more advanced models are instructive, the sheer diversity in real-world survey designs and analysis objectives generally requires the empirical approach of estimating design effects directly from the available survey data:

$$d^2(\hat{\theta}) = \frac{se(\hat{\theta})^2_{complex}}{se(\hat{\theta})^2_{srs}} = \frac{var(\hat{\theta})_{complex}}{var(\hat{\theta})_{srs}} \tag{2.7}$$

where

$d^2(\hat{\theta}) =$ the estimated design effect for the sample estimate, $\hat{\theta}$;
$var(\hat{\theta})_{complex} =$ the estimated complex sample design variance of $\hat{\theta}$; and
$var(\hat{\theta})_{srs} =$ the estimated simple random sample variance of $\hat{\theta}$.

As a statistical tool, the concept of the complex sample design effect is more directly useful to the designer of a survey sample than to the analyst of the survey data. The sample designer can use the concept and its component models to optimize the cost and error properties of specific design alternatives or to adjust simple random sample size computations for the design effect anticipated under a specific sampling plan (Kish, Groves, and Kotki, 1976). Using the methods and software presented in this book, the survey data analyst will compute confidence intervals and test statistics that incorporate the estimates of standard errors corrected for the complex sample design—generally bypassing the need to estimate the design effect ratio.

Nevertheless, knowledge of estimated design effects and the component factors does permit the analyst to gauge the extent to which the sampling plan for his or her data has produced efficiency losses relative to a simple random sampling standard and to identify features such as extreme clustering or weighting influences that might affect the stability of the inferences that he or she will draw from the analysis of the data. In addition, there are several analytical statistics such as the Rao–Scott Pearson χ^2 or likelihood ratio χ^2 where estimated design effects are used directly in adapting conventional hypothesis test statistics for the effects of the complex sample (see Chapter 6).

2.5.2 Generalized Design Effects and Effective Sample Sizes

The design effect statistic permits us to estimate the variance of complex sample estimates relative to the variance for an SRS of equal size:

$$\begin{aligned} var(\hat{\theta})_{complex} &= d^2(\hat{\theta}) \cdot var(\hat{\theta})_{srs}; \quad \text{or} \\ se(\hat{\theta})_{complex} &= \sqrt{d^2(\hat{\theta})} \cdot se(\hat{\theta})_{srs} \end{aligned} \tag{2.8}$$

Under the SRS assumption, the variances of many forms of sample estimates are approximately proportionate to the reciprocal of the sample size, that is, $var(\hat{\theta}) \propto 1/n$.

For example, if we ignore the fpc, the simple random sampling variances of estimates of a population proportion, mean, or simple linear regression coefficient are

$$var(p) = \frac{p(1-p)}{(n-1)}$$

$$var(\bar{y}) = \frac{s^2}{n}$$

$$var(\hat{\beta}) = \frac{\hat{\sigma}^2_{y.x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Before today's software was conveniently available to analysts, many public use survey data sets were released without the detailed stratification and cluster variables that are required for complex sample variance estimation. Instead, users were provided with tables of generalized design effects for key survey estimates that had been computed and summarized by the data producer. Users were instructed to perform analyses of the survey data using standard SAS, Stata, or SPSS programs under simple random sampling assumptions, to obtain SRS sampling variance estimates, and to then apply the design effect factor as shown in Equation 2.8 to approximate the correct complex sample variance estimate and corresponding confidence interval for the sample estimate. Even today, several major public use survey data sets, including the National Longitudinal Survey of Youth (NLSY) and the Monitoring the Future (MTF) Survey, require analysts to use this approach.

Survey designers make extensive use of design effects to translate between the simple analytical computations of sampling variance for SRS designs and the approximate variances expected from a complex design alternative. In working with clients, samplers may discuss design effect ratios, or they may choose a related measure of design efficiency termed the **effective sample size**:

$$n_{eff} = n_{complex} / d^2(\hat{\theta}) \tag{2.9}$$

where

n_{eff} = the effective sample size, or the number of SRS cases required to achieve the same sample precision as the actual complex sample design.

$n_{complex}$ = the actual or "nominal" sample size selected under the complex sample design.

The design effect ratio and effective sample size are, therefore, two means of expressing the precision of a complex sample design relative to an SRS of equal size. For a fixed sample size, the statements "the design effect for the proposed complex sample is 1.5" and "the complex sample of size $n = 1000$ has an effective sample size of $n_{eff} = 667$" are equivalent statements of the precision loss expected from the complex sample design.

2.6 Complex Samples: Clustering and Stratification

We already noted that survey data collections are rarely based on simple random samples. Instead, sample designs for large survey programs often feature stratification, clustering, and disproportionate sampling. Survey organizations use these "complex" design features to optimize the variance/cost ratio of the final design or to meet precision targets for subpopulations of the survey population. The authors' mentor, Leslie Kish (1965, 1987), was fond of creating classification systems for various aspects of the sample design process. One such system was a taxonomy of complex sample designs. Under the original taxonomy there were six binary keys to characterize all complex probability samples. Without loss of generality, we will focus on four of the six keys that are most relevant to the survey data analyst and aim to correctly identify the design features that are most important in applications:

Key 1: Is the sample selected in a single stage or multiple stages?
Key 2: Is clustering of elements used at one or more sample stages?
Key 3: Is stratification employed at one or more sample stages?
Key 4: Are elements selected with equal probabilities?

In the full realm of possible sample approaches this implies that there are at least 2^4 or 16 possible choices of general choices for complex sample designs. In fact, the number of complex sample designs encountered in practice is far fewer, and one complex design—multistage, stratified, cluster sampling with unequal probabilities of selection for elements—is used in most in-person surveys of household populations. Because they are so important in major programs of household population survey research, we will cover these multistage probability sampling plans in detail in Section 2.8. Before we do that, let's take a more basic look at the common complex sample design features of clustering, stratification, and weighting for unequal selection probabilities and nonresponse.

2.6.1 Clustered Sampling Plans

Clustered sampling of elements is a common feature of most complex sample survey data. In fact, to simplify our classification of sample designs it is possible to view population elements as "clusters of size 1." By treating elements as single-unit clusters, the general formulas for estimating statistics and standard errors for clustered samples can be applied to correctly estimate standard errors for the simpler stratified element samples (see Chapter 3).

Survey designers employ sample clustering for several reasons:

- Geographic clustering of elements for household surveys reduces interviewing costs by amortizing travel and related expenditures over a group of observations. By definition, multistage sample designs such as the area probability samples employed in the NCS-R, National Health and Nutrition Examination Survey (NHANES), and the Health and Retirement Study (HRS) incorporate clustering at one or more stages of the sample selection.
- Sample elements may not be individually identified on the available sampling frames but can be linked to aggregate cluster units (e.g., voters at precinct polling stations, students in colleges and universities). The available sampling frame often identifies only the cluster groupings. Identification of the sample elements requires an initial sampling of clusters and on-site work to select the elements for the survey interview.
- One or more stages of the sample are deliberately clustered to enable the estimation of multilevel models and components of variance in variables of interest (e.g., students in classes, classes within schools).

Therefore, while cluster sampling can reduce survey costs or simplify the logistics of the actual survey data collection, the survey data analyst must recognize that clustered selection of elements affects his or her approach to variance estimation and developing inferences from the sample data. In almost all cases, sampling plans that incorporate cluster sampling result in standard errors for survey estimates that are greater than those from an SRS of equal size; further, special approaches are required to estimate the correct standard errors. The SRS variance estimation formulae and approaches incorporated in the standard programs of most statistical software packages no longer apply, because they are based on assumptions of independence of the sample observations and sample observations from within the same cluster generally tend to be correlated (e.g., students within a classroom, or households within a neighborhood).

The appropriate choice of a variance estimator required to correctly reflect the effect of clustering on the standard errors of survey statistics depends on the answers to a number of questions:

1. Are all clusters equal in size?
2. Is the sample stratified?
3. Does the sample include multiple stages of selection?
4. Are units selected with unequal probability?

Fortunately, modern statistical software includes simple conventions that the analyst can use to specify the complex design features. Based on a simple set of user-supplied "design variables," the software selects the appropriate variance estimation formula and computes correct design-based estimates of standard errors.

The general increase in design effects due to either single-stage or multistage clustered sampling is caused by correlations (nonindependence) of observations within sample clusters. Many characteristics measured on sample elements within naturally occurring clusters, such as children in a school classroom or adults living in the same neighborhood, are correlated. Socioeconomic status, access to health care, political attitudes, and even environmental factors such as the weather are all examples of characteristics that individuals in sample clusters may share to a greater or lesser degree. When such group similarity is present, the amount of "statistical information" contained in a clustered sample of n persons is less than in an independently selected simple random sample of the same size. Hence, clustered sampling increases the standard errors of estimates relative to a SRS of equivalent size. A statistic that is frequently used to quantify the amount of homogeneity that exists within sample clusters is the **intraclass correlation**, ρ (Kish, 1965). See Kish et al. (1976) for an in-depth discussion of intraclass correlations observed in the World Fertility Surveys.

A simple example is useful to explain why intraclass correlation influences the amount of "statistical information" in clustered samples. A researcher has designed a study that will select a sample of students within a school district and collect survey measures from the students. The objective of the survey is to estimate characteristics of the full student body and the instructional environment. A probability sample of $n = 1{,}000$ is chosen by randomly selecting 40 classrooms of 25 students each. Two questions are asked of the students:

1. What is your mother's highest level of completed education? Given the degree of socioeconomic clustering that can exist even among schools within a district, it is reasonable to expect that the intraclass correlation for this response variable is positive, possibly as high as $\rho = 0.2$.
2. What is your teacher's highest level of completed education? Assuming students in a sampled class have only one teacher, the intraclass correlation for this measure is 1.0. The researcher need not ask the question of all $n = 1{,}000$ students in the sample. An

identical amount of information could be obtained by asking a single truthful student from each sample classroom. The effective sample size for this question would be 40, or the number of sample classrooms.

When the primary objective of a survey is to estimate proportions or means of population characteristics, the following model can be used to approximate the design effect that is attributable to the clustered sample selection (Kish, 1965):

$$D^2(\bar{y}) = 1 + L_{cluster} \approx 1 + \rho \cdot (B - 1) \tag{2.10}$$

where

ρ = the intraclass correlation for the characteristic of interest;
B = the size of each cluster or primary sampling unit (PSU).

The value of ρ is specific to the population characteristic (e.g., income, low-density cholesterol level, candidate choice) and the size of the clusters (e.g., counties, enumeration areas [EAs], schools, classrooms) over which it is measured. Generally, the value of ρ decreases as the geographical size and scope (i.e., the heterogeneity) of the cluster increases. Typical values of ρ observed for general population characteristics range from .00 to .20 with most between .005 and .100 (Kish et al., 1976).

In practice, it is sometimes valuable to estimate the value of ρ for a survey characteristic y. While it is theoretically possible to estimate ρ as a function of the sample data (y_i, $i = 1, \ldots, n$) and the sample cluster labels ($\alpha = 1, \ldots, a$), in most survey designs such direct estimates tend to be unstable. Kish (1965) suggested a synthetic estimator of ρ, which he labeled the **rate of homogeneity** (***roh***). Using the simple model for the estimated design effect for a sample mean, the expression is rearranged to solve for *roh* as

$$roh(y) = \frac{d^2(\bar{y}) - 1}{\bar{b} - 1} \tag{2.11}$$

where $\bar{b}$ is the average sample size per cluster.

Algebraically, this synthetic estimate of the intraclass correlation is restricted to a range of values from $-1/(\bar{b} - 1)$ to 1.0. Since ρ is almost always positive for most survey variables of interest, this restricted range of the synthetic approximation *roh* is not a problem in practice. As the analyst (and not the designer) of the sample survey, the calculation of *roh* is not essential to the reader's work; however, when the reader begins to work with a new survey data set, it can be valuable to compute the *roh* approximation for a few key survey variables simply to get a feel for the level of intraclass correlation in the variables of interest.

Let's return to the hypothetical sample data set provided in Table 2.1. Previously, we assumed a simple random sample of size $n = 32$ and estimated the mean, the SRS standard error, and a 95% confidence interval for the mean of the economic conditions index. Now, instead of assuming SRS, let's assume that eight clusters, each of size four elements, were selected with equal probability. The cluster codes for the individual cases are provided in Column (3) of Table 2.1. Without digressing to the specific estimators of the mean and variance for this simple clustered sample (see Cochran, 1977), let's look at the correct numerical values for the estimated mean, standard error of the mean, and the 95% CI for $\bar{Y}$:

$$\bar{y}_{cl} = 40.77,\ se_{cl}(\bar{y}) = 3.65,\ CI(\bar{y}_{cl}) = (32.12, 49.49)$$

Comparing these results to those for the SRS example based on the same data (Section 2.4.3), the first thing we observe is that the estimate of $\bar{y}$ is identical. The change that occurs when clustering is introduced to the example is that the standard error of the mean is increased (due to the presence of the intraclass correlation that was used in developing the clusters for this example). The width of the 95% CI is also increased, due to the increase in standard error and the reduction in the degrees of freedom for the Student t statistic used to develop the CI limits (degrees of freedom determination for complex sample designs will be covered in Chapter 3). The relative increase in $se(\bar{y})$ for the simple cluster sample compared with an SRS of $n = 32$ and the synthetic estimate of the intraclass correlation are computed as follows:

$$d(\bar{y}) = \sqrt{d^2(\bar{y})} = \frac{se_{cl}(\bar{y})}{se_{srs}(\bar{y})} = \frac{3.65}{2.41} = 1.51$$

$$roh(y) \approx \frac{1.51^2 - 1}{4 - 1} = 0.43$$

2.6.2 Stratification

Strata are nonoverlapping, homogeneous groupings of population elements or clusters of elements that are formed by the sample designer prior to the selection of the probability sample. In multi-stage sample designs (see Section 2.8), a different stratification of units can be employed at each separate stage of the sample selection. Stratification can be used to sample elements or clusters of elements. As already noted, if elements are viewed as "clusters of size $B = 1$" then the general formulas for estimating statistics and standard errors for stratified clustered samples can be applied to correctly compute these statistics for the simpler stratified element samples (see Chapter 3). To take this "unification" one step further, if the sample design does not incorporate

explicit stratification, it may be viewed as sampling from a single stratum (i.e., $H = 1$).

In survey practice, stratified sampling serves several purposes:

- Relative to an SRS of equal size, stratified samples that employ **proportional allocation** or **optimal allocation** of the sample size to the individual strata have smaller standard errors for sample estimates (Cochran, 1977).
- Stratification provides the survey statistician with a convenient framework to disproportionately allocate the sample to subpopulations, that is, to oversample specific subpopulations to ensure sufficient sample sizes for analysis.
- Stratification of the probability sample can facilitate the use of different survey methods or procedures in the separate strata (see Chapter 3).

Every stratified sample design involves the following four steps in sample selection and data analysis:

1. Strata ($h = 1, \ldots, H$) of N_h clusters/elements are formed.
2. Probability samples of a_h clusters or $a_h = n_h$ elements are independently selected from each stratum.
3. Separate estimates of the statistic of interest are computed for sample cases in each stratum and then weighted and combined to form the total population estimate.
4. Sampling variances of sample estimates are computed separately for each stratum and then weighted and combined to form the estimate of sampling variance for the total population estimate.

Because stratified sampling selects independent samples from each of the $h = 1, \ldots, H$ explicit strata of relative size $W_h = N_h/N$, any variance attributable to differences among strata is eliminated from the sampling variance of the estimate. Hence, the goal of any stratification designed to increase sample precision is to form strata that are "homogeneous within" and "heterogeneous between": Units assigned to a stratum are like one another and different from those in other strata, in terms of the measurements collected. To illustrate how stratification works to reduce sampling variance, let's consider the case of a simple stratified random sample of size

$$n = \sum_{h=1}^{H} n_h$$

where the stratum sample sizes are proportional to the individual strata: $n_h = n \cdot N_h / N$. This **proportionate allocation** of the sample ensures that each

population element has an equal probability of being included in the sample, denoted by $f_h = n_n / N_h = n / N = f$. Let's compare $var(\bar{y}_{st,pr})$ for this stratified sample with $var(\bar{y}_{srs})$, ignoring the fpc:

$$\begin{aligned}\Delta &= var(\bar{y}_{srs}) - var(\bar{y}_{st,pr}) \\ &= \frac{S^2_{total}}{n} - \frac{S^2_{within}}{n} = \frac{\left[S^2_{within} + S^2_{between}\right]}{n} - \frac{S^2_{within}}{n} \\ &= \frac{S^2_{between}}{n} \\ &= \frac{\sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2}{n}\end{aligned} \tag{2.12}$$

The simple random sample of size n includes the variance of y both within the strata, S^2_{within}, and between the strata, $S^2_{between}$, but the stratified sample has eliminated the between-stratum variance component and consequently yields a sampling variance for the mean that is less than or equal to that for an SRS of equal size. The expected amount of the reduction in the variance of the mean is a weighted function of the squared differences between the stratum means, $\bar{Y}_h$, and the overall population mean, $\bar{Y}$. Hence the importance of forming strata that differ as much as possible in the $h = 1, \ldots, H$ values of the $\bar{Y}_h$.

Let's return to the example data set in Table 2.1 and assume that the $n = 32$ observations were selected from four population strata of equal size, such that $W_h = N_h/N = 0.25$. The stratum identifiers for each sample case are provided in Column (2) of Table 2.1. Under this proportionately allocated, stratified random sample design, we compute the following:

$$\begin{aligned}\bar{y}_{st,pr} &= \sum_{h=1}^{H} W_h \cdot \bar{y}_h \\ &= 0.25 \times 51.1 + 0.25 \times 43.32 + 0.25 \times 39.60 + 0.25 \times 29.04 \\ &= 40.77\end{aligned}$$

$$\begin{aligned}se(\bar{y}_{st,pr}) &= \sqrt{var(\bar{y}_{st,pr})} = \sqrt{\sum_{h=1}^{H} W_h^2 s_h^2 / n_h} \\ &= 2.04\end{aligned}$$

$$d(\bar{y}_{st,pr}) = \frac{se(\bar{y}_{st,pr})}{se(\bar{y}_{srs})} = \frac{2.04}{2.41} = 0.85$$

$$d^2(\bar{y}_{st,pr}) = 0.85^2 = 0.72$$

$$n_{eff} = n / d^2(\bar{y}_{st,pr}) = 32 / 0.72 = 44.44$$

Note that the unbiased sample estimate, $\bar{y}_{st,pr} = 40.77$, is identical to that obtained when we assumed that the 32 observations were sampled using SRS, and also identical to that when we assumed that the 32 observations were obtained from a cluster sample. The estimated standard error, $se(\bar{y}_{st,pr}) = 2.04$, is much smaller, resulting in an estimated design effect ratio of 0.72. For this example, it appears that the stratified sample of size $n = 32$ (eight cases from four strata) yields the sample precision of an SRS of more than 44 cases.

The simple algebraic illustration and previous exercise calculation demonstrate how stratification of the sample can lead to reductions in the variance of sample estimates relative to SRS. In contrast to this simple example, the precision gains actually achieved by stratification, G_{strat}, in complex samples are difficult to determine analytically. The gains will be a function of the stratum means, the variances of y within individual strata, and whether the sampling within strata includes additional stages, clustering, unequal probabilities of selection, or other complex sample design features.

2.6.3 Joint Effects of Sample Stratification and Clustering

To see how the two design features of stratification and clustering contribute to the design effect for a complex sample, let's revisit the sample data in Table 2.1, treating it as a stratified sample from $h = 1, \ldots, 4$ strata. Within each stratum, a proportionately allocated sample of two clusters ($a_h = 2$) of size 4 are selected with equal probability. This is a very simple example of a stratified "**two-per-stratum**" or "paired selection" cluster sample design that is very common is survey practice (see Chapter 4). Table 2.2 compares

TABLE 2.2

Sample Estimates of the Population Mean and Related Measures of Sampling Variance under Four Alternative Sample Designs

Sample Design Scenario	Estimator	$\bar{y}$	$se(\bar{y})$	$d(\bar{y})$	$d^2(\bar{y})$	n_{eff}
SRS	$\bar{y}_{srs}$	40.77	2.41	1.00	1.00	32
Clustered	$\bar{y}_{cl}$	40.77	3.66	1.51	2.31	13.9
Stratified	$\bar{y}_{st}$	40.77	2.04	0.85	0.72	44.4
Stratified, clustered	$\bar{y}_{st,cl}$	40.77	2.76	1.15	1.31	24.4

Source: Table 2.1.

the results for this stratified, cluster sample design scenario with those for the simpler SRS, clustered, and stratified designs considered earlier.

Note that for the stratified, cluster sample design scenario the estimated design effect falls between the high value for the clustered-only scenario and the low value for the scenario where only stratification effects are considered. As illustrated in Figure 2.2, real survey designs result in a "tug of war" between the variance inflation due to clustering and the variance reduction due to stratification. From Table 2.2, the design effect for the stratified, clustered sample is greater than 1, indicating a net loss of precision relative to an SRS of size n—clustering "won" the contest. Estimates of design effects greater than 1 are common for most complex sample designs that combine stratification and clustered sample selection. In population surveys, large stratification gains are difficult to achieve unless the survey designer is able to select the sample from a frame that contains rich ancillary data, $\boldsymbol{x}$, that are highly correlated with the survey variables, $\boldsymbol{y}$. The survey data analyst needs to be aware of these design effects on the precision of sample estimates. Correctly recognizing the stratum and cluster codes for sample respondents provided by a data producer for use by data analysts is essential for correct and unbiased analyses of the survey data. We discuss the importance of identifying these codes in a survey data set in Chapter 4.

2.7 Weighting in Analysis of Survey Data

2.7.1 Introduction to Weighted Analysis of Survey Data

When probability sampling is used in survey practice, it is common to find that sample inclusion probabilities for individual observations vary. Weighting of the survey data is thus required to "map" the sample back to an unbiased representation of the survey population. A simple but useful device for "visualizing" the role of case-specific weights in survey data analysis is to consider the weight as the number (or share) of the population elements that is represented by the sample observation. Observation i, sampled with probability $f_i = 1/10$, represents 10 individuals in the population (herself and nine others). Observation j, selected with probability $f_j = 1/20$, represents 20 population elements. In fact, if each sample case is assigned a weight equal to the reciprocal of its probability of selection, $w_i = 1/f_i$, the sum of the sample weights will in expectation equal the population size:

$$E\left(\sum_{i=1}^{n} w_i\right) = N$$

As a survey data analyst, the reader can use this fact during his or her preanalysis "checklist" to understand the distribution of the sampling weights that the data producer has provided (see Chapter 4).

Generally, the final analysis weights in survey data sets are the product of the sample selection weight, w_{sel}, a nonresponse adjustment factor, w_{nr}, and the poststratification factor, w_{ps}:

$$w_{final,i} = w_{sel,i} \times w_{nr,i} \times w_{ps,i} \tag{2.13}$$

The sample selection weight factor (or **base weight**) is simply the reciprocal of the probability that a population element was chosen to the sample, $w_{sel,i} = 1/f_i$. Under the theory of design-based inference for probability samples, weighted estimation using these "inverse probability" weight factors will yield unbiased (or nearly unbiased) estimates of population statistics. For example,

$$\begin{aligned} \bar{y}_w &= \frac{\sum_{i=1}^{n} w_{sel,i} \cdot y_i}{\sum_{i=1}^{n} w_{sel,i}} \text{ is unbiased for } \bar{Y} \\ s_w^2 &= \frac{\sum_{i=1}^{n} w_{sel,i} \cdot (y_i - \bar{y})^2}{\sum_{i=1}^{n} w_{sel,i} - 1} \text{ is an estimate of } S^2 \end{aligned} \tag{2.14}$$

Note how in each of these estimators, the selection weight "expands" each sample observation's contribution to reflect its share of the population representation. Throughout this book, weighted estimation of population parameters will follow a similar approach, even for procedures as complex as the pseudo-maximum likelihood (PML) estimation of the coefficients in a multivariate logistic regression model.

If we were able to always complete an observation on each selected sample case, the development of the survey analysis weights would be very straightforward. The computation of the weight would require only knowledge of the sample selection probability factors (see Section 2.7.2). In a carefully designed and well-managed survey, these probability factors are known and carefully recorded for each sample case. The computation of w_{sel} is therefore an accounting function, requiring only multiplication of the probabilities of selection at each stage of sampling and then taking the reciprocal of the product of the probabilities.

As we will see in Chapter 4, **survey data producers**, who in most cases have the responsibility for developing the final analysis weights have a more difficult task than the simple arithmetic required to compute w_{sel}. First, weighting by w_{sel} will yield only unbiased estimators of population parameters if all n elements in the original sample are observed. Unfortunately, due to **survey nonresponse**, observations are collected only for r cases of the original probability sample of n elements (where $r <= n$). Therefore, survey data producers must develop statistical models of the conditional probability that a sample element will be an observed case. In general terms, the nonresponse adjustment factor, w_{nr}, in the analysis weight is the reciprocal of the estimated conditional probability that the sample case responds. The objective in applying nonresponse factors in survey weights is to attenuate bias due to differential nonresponse across sample elements. A price that may be paid for the bias reduction through nonresponse weighting takes the form of increases in standard errors for the weighted estimates. The potential magnitude of the increases in standard errors is discussed next in the context of weighting for unequal selection probabilities.

Even after computing w_{sel} and adjusting for nonresponse through the factor w_{nr}, most survey data producers also introduce the poststratification factor w_{ps} into the final weight. As its label implies, the poststratification factor is an attempt to apply stratification corrections to the observed sample "post" or after the survey data have been collected. While not as effective as estimation for samples that were stratified at the time that they were selected, the use of poststratification weight factors can lead to reduced standard errors (variance) for sample estimates. Furthermore, poststratification weighting can attenuate any sampling biases that may have entered the original sample selection due to sample frame noncoverage or omissions that occurred in implementing the sample plan. An example of the latter would be systematic underreporting of young males in the creation of household rosters for the selection of the survey respondent.

In the sections that follow, we describe the development of these three important weighting factors in more detail.

2.7.2 Weighting for Probabilities of Selection

As previously described, the selection weight factor, w_{sel}, is introduced in the calculation of the final analysis weight to account for the probability that a case was selected for the sample. Common reasons for varying probabilities of case selection in sample surveys include the following:

1. Disproportionate sampling within strata to achieve an optimally allocated sample for a specific population estimate (Cochran, 1977).

2. Disproportionate sampling within a stratum or group of strata to deliberately increase the sample size and precision of analysis for certain domains of the survey population.
3. The use of sample screening across the sample strata and clusters to identify and differentially sample subpopulations, such as the NHANES oversampling of household members with disabilities.
4. Unequal probabilities that arise from subsampling of observational units within sample clusters, such as the common procedure of selecting a single random respondent from the eligible members of sample households.
5. Unequal probabilities that reflect information on final sampling probability that can be obtained only in the process of the survey data collection, for example, in a **random digit dialing (RDD)** telephone sample survey, the number of distinct landline telephone numbers that serve a respondent's household.

Due to the multipurpose nature of modern population surveys (many variables, many statistical aims), the use of disproportionate sampling to optimize a sample design for purposes of estimating a single population parameter is rare in survey practice. Far more common are designs in which geographic domains or subpopulations of the survey population are oversampled to boost the sample size and precision for stand-alone or comparative analysis. For example, the HRS employs a roughly two-fold oversampling of age-eligible African Americans and Hispanic individuals to improve precision of cross-sectional and longitudinal analyses of these two population subgroups (Heeringa and Connor, 1995). The NCS-R employs a Part 1 screening interview to obtain profiles of respondents' mental health related symptoms but disproportionately samples persons with positive symptom counts for the more intensive Part 2 questionnaire (Kessler et al., 2004). As these examples illustrate, the selection probability for individual observational units often includes multiple components—for example, a base factor for housing unit selection, a factor for subsampling targeted groups within the full population, and a probability of selecting a single random respondent within the household. For HRS sample respondents, the selection weight factor, w_{sel}, is computed as the product of the reciprocals of three probabilities: (1) $w_{sel,hh}$, the reciprocal of the multistage probability of selecting the respondent's housing unit (HU) from the area frame (see Section 2.8); (2) $w_{sel,sub}$, the reciprocal of the probability of retaining the sample HU under the design objective of a 2:1 oversampling of eligible African American and Hispanic adults; and (3) $w_{sel,resp}$, the reciprocal of the conditional probability of selecting a respondent within the eligible household. This calculation is illustrated as follows:

TABLE 2.3

Illustration of Example w_{sel} Computations Based on the HRS

Description of Sample Case						
Sample ID	Race/Ethnicity of Respondent	Eligible Rs/Household	$w_{sel,hh}$	$w_{sel,sub}$	$w_{sel,resp}$	w_{sel}
A	Black	2	2000	1	2	4000
B	Black	1	2000	1	1	2000
C	Hispanic	2	2000	1	2	4000
D	White	1	2000	2	1	4000

$$f_{sel} = f_{sel,hh} \times f_{sel,sub} \times f_{sel,resp}$$

$$\Rightarrow$$

$$w_{sel} = \frac{1}{f_{sel}} = w_{sel,hh} \times w_{sel,sub} \times w_{sel,resp}$$

Drawing on the general features of the HRS sample selection, Table 2.3 illustrates the sample selection weight calculations for four hypothetical respondents from different race/ethnicity groups who have different numbers of eligible respondents in their sample households.

2.7.3 Nonresponse Adjustment Weights

As described already, the most widely accepted approach to compensate for **unit nonresponse** in surveys is for the data producer to develop and apply a nonresponse adjustment factor to the sample selection weight that is used in analysis. Underlying the weighting adjustment for nonresponse is a model of the **response propensity**—conditional on sample selection, the probability that the unit will cooperate in the survey request. In a sense, the concept of response propensity treats response to the survey as another step in the "sample selection process." But unlike true sample selection in which the sampling statistician predetermines the sampling probability for each unit, an underlying propensity model—for the most part outside the control of a statistician—determines the probability that a sampled case will be observed. The multiplication of the original sample selection weight for each sample unit by the reciprocal of its modeled response propensity creates a new weight, which, if the model is correct, enables unbiased or nearly unbiased estimation of population statistics from the survey data.

Two related methods for estimating response propensity and computing the nonresponse adjustment are commonly used in survey practice: a

simple weighting class adjustment method and a propensity score weighting approach.

2.7.3.1 Weighting Class Approach

The "weighting class method" (Little and Rubin, 2002) assigns all eligible elements of the original sample—survey respondents and nonrespondents—to classes or cells based on categorical variables (e.g., age categories, gender, region, sample stratum) that are predictive of response rates. Original sample cases that were found to be ineligible for the survey are excluded from the nonresponse adjustment calculation. The weighting class method makes the simple assumption that the response propensities for cases within a given weighting class cell are equal (i.e., respondents are MAR, or missing at random, conditional on the categorical variables and interactions implicit in the defined cells). The common response propensity for cases in a cell is estimated by the empirical response rate for the cases assigned to that cell. The weighting class nonresponse adjustment is then computed as the reciprocal of the response rate for the cell $c = 1, ..., C$ to which the case was assigned:

$$w_{nr,wc,i} = \frac{1}{rrate_c} \tag{2.15}$$

where $rrate_c$ = the response rate for weighting class $c = 1, \ldots, C$.

To be effective at reducing potential bias due to differential nonresponse across sample cases, the variables chosen to create the weighting classes must be predictive of survey participation rates. Recent work (Little and Vartivarian, 2005) has shown that reductions in both the bias and variance of an estimate are possible if weighting classes are formed based on variables related to *both* response propensity and the survey variable of interest.

2.7.3.2 Propensity Cell Adjustment Approach

The propensity cell weighting approach also assigns an adjustment factor to each respondent's sample selection weight that is equal to the reciprocal of the estimated probability that they participated in the survey. However, in the propensity adjustment method, the assignment of cases to adjustment cells is based on individual response propensity values estimated (via logit transform) from a logistic regression model:

$$\hat{p}_{resp,i} = prob(respondent = yes \mid X_i) = \left(\frac{e^{X_i\hat{\beta}}}{1 + e^{X_i\hat{\beta}}} \right) \tag{2.16}$$

THEORY BOX 2.2 ESTIMATING RESPONSE PROPENSITIES IN NR ADJUSTMENTS

In our discussion, both the weighting class and propensity cell adjustments for nonresponse require estimation of response rates within defined cells. Traditionally, many survey statisticians have employed the sample selection weight, w_{sel}, to compute weighted estimates of response rates in these cells:

$$rrate_{weighted,cell} = \sum_{i \in cell}^{n_{cell}} w_{sel,i} \cdot y_i \Big/ \sum_{i \in cell}^{n_{cell}} w_{sel,i} \tag{2.18}$$

where

y_i = 1 if sample person i is a respondent, 0 otherwise;
$w_{sel,i}$ = the sample selection weight for case i;
n_{cell} = the total count of eligible sample cases in cell c = 1, …, C.

Recently, Little and Vartivarian (2005) have suggested that the unweighted response rate may be the preferred estimate of the response propensity for the cell, depending on what design variables are used to construct the cells:

$$rrate_{cell} = \sum_{i \in cell}^{n_{cell}} y_i \Big/ n_{cell} \tag{2.19}$$

Note that under the propensity modeling approach, the computed value of the empirical response rate for cases in a modeled propensity quantile need not always fall within the range of scores for the quantile. For example, the range of propensity scores used to define an adjustment cell might be 0.75–0.79, but the corresponding response rate for cases in the cell could be 0.73. This is a reflection of lack of fit in the propensity model. Standard tests may be used to evaluate the goodness of fit of the logistic regression model of response propensity. We recommend including as many theoretical predictors of responding as possible in the model, and especially predictors that are also correlated with the survey variables of primary interest (Little and Vartivarian, 2005).

where
X_i = a vector of values of response predictors for $i = 1,...,n$;
$\hat{\beta}$ = the corresponding vector of estimated logistic regression coefficients.

Adjustment cells are then defined based on quantile ranges—often deciles, such as 0–0.099 or 0.10–0.199—of the distribution of response propensities (Little and Rubin, 2002). The propensity score nonresponse weighting adjustments are then computed as the reciprocal of the response rate in the quantile range cell $d = 1, ..., D$ to which the case was assigned:

$$w_{nr,prop,i} = \frac{1}{rrate_d} \tag{2.17}$$

where $rrate_d$ = the weighted response rate for propensity cell $d = 1, \ldots, D$.

The use of individual estimated response propensities for these adjustments (i.e., $\hat{p}_{resp,i}$ as opposed to the response rates, $rrate_d$, within the quantile range cells, as shown in (2.17)) is often not recommended, as extremely low estimated response propensities (e.g., $\hat{p}_{resp,i} = 0.001$) could result in more variance in the weights. This increased variance in the weights would decrease the precision of weighted estimates based on the survey data (see Section 2.7.5).

To employ either method for nonresponse adjustment, the characteristics used to define the cells or model the response propensities must be known for both respondents and nonrespondents. In cross-sectional sample surveys such as the NCS-R or NHANES, this limits the nonresponse adjustment to characteristics of sample persons or households that are known from the sampling frame or are completely observed in the screening process. In the case of the 1992 HRS baseline sample, a simple weighting class adjustment approach was employed to develop the nonresponse adjustment (Heeringa and Connor, 1995). Region (Northeast, South, Midwest, West), urban/rural status of the primary stage unit (PSU), and race of the respondent (Black, Hispanic, White, and Other) were used to define weighting class adjustment cells. See Table 2.4 for an illustration of typical values of w_{nr} for various cells and how the values of this adjustment factor influence the final composite weight.

2.7.4 Poststratification Weight Factors

The nonresponse adjustment procedures just described have the property that only data available for sampled respondent and nonrespondent cases are used to compute weighting adjustments. Another weighting technique used in practice to improve the quality of sample survey estimates is to incorporate known information on the full survey population—borrowing strength from data sources external to the sample. **Poststratification** is one

TABLE 2.4

Illustration of Example w_{final} Computations Including w_{sel}, w_{nr}, and w_{ps} Based on the HRS

Description of Sample Case						
Sample ID	Nonresponse Adjustment Cell	Poststratum	w_{sel} [a]	w_{nr}	w_{ps}	w_{final}
A	Black, Northeast, urban	Age 50–54, male, Northeast, Black	4000	1.3	1.04	5408
B	Black, South, rural	Age 55–61, female, South, Black	2000	1.15	.96	2208
C	Hispanic, West, urban	Age 50–54, male, West, Hispanic	4000	1.25	1.06	5300
D	White, Midwest, rural	Age 55–61, female, Midwest, White	4000	1.18	.97	4578

[a] From Table 2.3.

such method for using population data in survey estimation. **Raking ratio estimation**, **generalized regression (GREG) estimation**, and **calibration** are other forms of postsurvey weight adjustment that may be employed to improve the precision and accuracy of survey estimates. Here we will focus on poststratification, the most common technique applied in general social and health survey weighting.

Simple poststratification forms $l = 1, \ldots, L$ poststrata of respondent cases (just as the sample designer might form $h = 1, \ldots, H$ design strata prior to sample selection). The criteria used to select variables for forming poststrata include (1) variables such as age, gender, and region that define poststrata for which accurate population control totals are available from external sources; (2) poststratification variables that are highly correlated with key survey variables; and (3) variables that may be predictive of noncoverage in the sample frame. Cross-classifications of these variables are then used to define the $l = 1, \ldots, L$ poststrata. To ensure efficiency in the poststratification, poststrata are generally required to include a minimum of $n_l = 15$ to 25 observations. If the cross-classification of poststratification variables results in cells with fewer than this minimum number, similar poststrata are collapsed to form a larger grouping.

Poststratification weighting involves adjusting the final weights for respondent sample cases so that weighted sample distributions conform to the known population distributions across the $l = 1, \ldots, L$ poststrata. The poststratification factor applied to each respondent weight is computed as follows:

$$w_{ps,l,i} = \frac{N_l}{\sum_{i=1}^{n_l} (w_{sel,i} \times w_{nr,i})} = \frac{N_l}{\hat{N}_l} \tag{2.20}$$

where

$w_{ps,1,i}$ = the post-stratification weight for cases in post-stratum $l = 1, \ldots, L$; and

N_l = the population count in post-stratum l obtained from a recent Census, administrative records, or a large survey with small sampling variance.

In the case of the 1992 HRS baseline sample, poststrata of sample respondents were defined based on age category (50–54, 55–61) for the 1931–1941 birth cohorts, gender, race/ethnicity, and Census region. Population control totals, N_l, for each poststratum were obtained as weighted population totals from the March 1992 demographic supplement to the U.S. Current Population Survey (Heeringa and Connor, 1995). See Table 2.4 for an illustration of typical values of w_{ps} for various poststrata and how the values of this adjustment factor influence the final composite sampling weight (w_{final}).

2.7.5 Design Effects Due to Weighted Analysis

The weights that readers will be using in their survey data analysis are therefore compound products of several adjustments—unequal probabilities of selection, differential patterns of response, and poststratification. Each of these factors can have a different effect on the precision and bias reduction for a design-based population estimate. Along with sample stratification and clustering, weighted estimation contributes to the final design effect for a survey estimate. In Figure 2.2 shown earlier in this chapter, the large arrow representing the net effect of weighting could be split into three smaller arrows: the small arrows for selection weighting and nonresponse adjustment pointing up to increased variance and the small poststratification arrow pointing down to lower variance. It is analytically impossible in most surveys to partition out the size of the contribution of each factor (stratification, clustering, selection weights, nonresponse adjustment, poststratification weighting) to the variance of sample estimates.

Changes in standard errors due to weighting are related to the variance of the weight values assigned to the individual cases and the correlations of the weights with the values and standard deviations of the variables of interest. We generally find in survey practice that the net effect of weighted estimation is inflation in the standard errors of estimates. This reflects the empirical fact that through the sequence of steps (e.g., selection weighting, nonresponse weighting) case weights can be quite variable. Furthermore, most survey weights are at best only moderately correlated with the distributional properties of the survey variables.

In recent years, the survey literature has used the term *weighting loss* to describe the inflation in variances of sample estimates that can be attributed to weighting. A simple approximation used by sampling statisticians to anticipate $L_{weighting}$, the proportional increase in variance of the sample mean due to weighted estimation, is

$$L_{weighting}(\bar{y}) \approx cv^2(w) = \frac{\sigma^2(w)}{\bar{w}^2} = \left\{ \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i \right)^2} \cdot n \right\} - 1 \qquad (2.21)$$

where

$cv^2(w)$ = the relative variance of the sample weights;
$s(w)$ = the standard deviation of the sample weights; and
$\bar{w}$ = the mean of the sample weights.

This simple model of weighting loss was introduced by Kish (1965), and despite its widespread use, it was intended more as a design tool than as a strong model of true weighting effects on variances of sample estimates. The Kish model of weighting loss was originally presented in the context of proportionate stratified sampling and represented the proportional increase in the variance of means (and proportions as means of binary variables) due to arbitrary disproportionate sampling of strata. It assumes that proportionate allocation is the optimal stratified design (i.e., variances of y are approximately equal in all strata) and that the weights are uncorrelated with the values of the random variable y. Little and Vartivarian (2005) show clear examples where this simple model of weighting loss breaks down. As in any model, the quality of the predictions is tied to how closely the data scenario matches the model assumptions. Even within a survey data set, there may be variables for which the weighting actually improves the precision of estimates (due to intended or chance optimality for that variable) and many others for which the variability and randomness of the weights simply produces an increase in the variance of estimates.

To illustrate the application of Kish's (1965) weighting loss model to a sample design problem, consider the following example. A survey is planned for a target population that includes both urban and rural populations of children. A total of 80% of the target population lives in the urban domain, and 20% lives in the rural villages. The agency sponsoring the survey is interested in measuring the mean body mass index (BMI) of these children. The agency would like to have roughly equal precision for mean estimates for urban and rural children and decides to allocate the sample equally (50:50) to the two geographic domains (or strata). They recognize that this will require weighting to obtain unbiased estimates for the *combined* area. Urban cases will need to be weighted up by a factor proportional to 0.80/0.50 = 1.6, and rural cases will need to be downweighted by a factor proportional to 0.20/0.50 = 0.4. To estimate $L_{weighting}$, they compute the relative variance of these weights for a sample that is

50% urban and 50% rural. They determine that $L_{weighting} \approx cv^2(w) = 0.36$. Ignoring any clustering that may be included in the sample plan, this implies that the final sample size for the survey must be $n_{complex} = n_{srs} \times 1.36$, or 36% larger than the SRS sample size required to meet a set precision level for a estimating mean BMI for the combined population of urban and rural children.

The reader should note that $L_{weighting}$ is often nontrivial. Survey data sets that include disproportionate sampling of geographic areas or other subpopulations should be carefully evaluated: Despite what may appear to be large nominal sample sizes for the total survey, the effective sample sizes for pooled analysis may be much smaller than the simple case count suggests.

Fortunately, a detailed understanding of each of these contributions to the final design effect for a complex sample survey is more important to the survey designer than it is to the survey analyst. In the role of survey analyst, the reader will often not have access to the detailed information that the survey producer's statisticians used to develop the final weight. As analysts, we must certainly be able to identify the correct weight variable to use in our analysis, be able to perform checks that the designated weight has been correctly carried forward to the data set that we will use for analysis, and be familiar with the syntax required to perform weighted estimation in our chosen statistical analysis software. Subsequent chapters in this book will provide the explanation and examples needed to become proficient in applying weights in survey data analysis.

2.8 Multistage Area Probability Sample Designs

The applied statistical methods covered in Chapters 5–12 of this volume are intended to cover a wide range of complex sample designs. However, because the majority of large public-use survey data sets that are routinely used in the social and health sciences are based on multistage area probability sampling of households, it is important to describe this particular probability sampling technique in greater detail. The generic illustration presented here is most applicable to national household sampling in the United States and Canada. With minor changes in the number of stages and the choice of sampling units, similar household sample designs are used throughout Central and South America, Africa, Asia, and Europe (Heeringa and O'Muircheartaigh, 2010).

Figure 2.3 illustrates the selection of a multistage, area probability sample of respondents. The design illustrated in this figure conforms closely to the actual procedure used in the selection of the NCS-R, HRS, and NHANES samples of U.S. household populations. Under the general procedure illustrated by the figure, the selection of each study respondent requires a four-

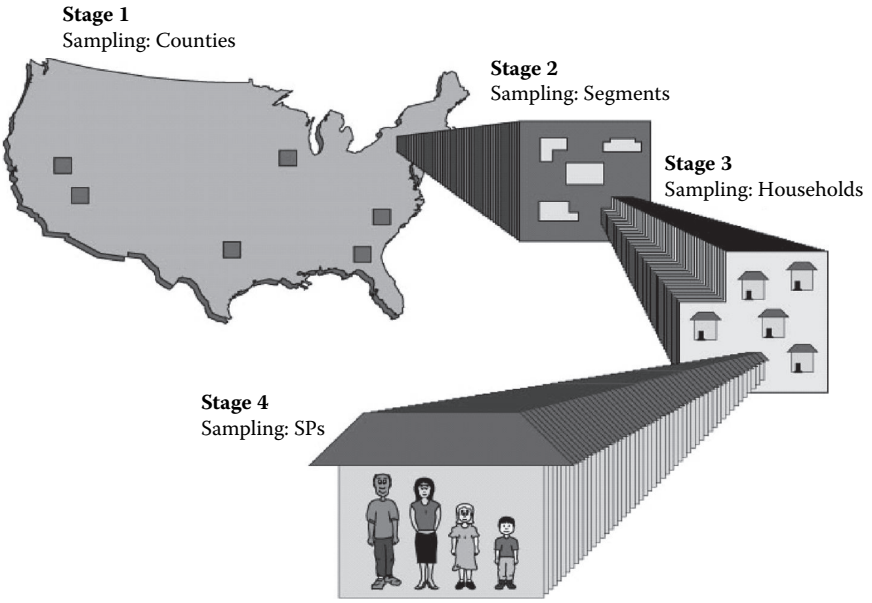


FIGURE 2.3
Schematic illustration of multistage area probability sampling. (From: Mohadjer, L., *The National Assessment of Adult Literacy (NAAL) Sample Design*, Westat, Rockville, MD, 2002. With permission.)

step sampling process: a primary stage sampling of U.S. counties or groups of adjacent counties, followed by a second stage sampling of area segments and a third stage sampling of housing units within the selected area segments, and concluding with the random selection of eligible respondents from the sampled housing units.

2.8.1 Primary Stage Sampling

The **primary stage units** of the multistage sample are single counties or groupings of geographically contiguous counties. The PSUs are therefore the highest-level groupings or "clusters" of sample observations (this will be important to remember when we discuss variance estimation in Chapters 3 and 4). All land area in the target population is divided into PSUs (e.g., the land area of the 50 United States is uniquely divided into roughly 3,100 county and parish [Louisiana] divisions). For comparison, the land area of Chile is uniquely divided into *communa* units, Russia into *raions* (regions), and South Africa and many other countries into Census *enumeration areas*. Ideally, the populations within PSUs will be reasonably heterogeneous to minimize intraclass correlation for survey variables. At the same time, the geographically defined PSUs must be small enough in size to facilitate cost-efficient travel to second stage interview locations. Each designated PSU in the population is assigned to $h = 1, \ldots, H$ sampling strata based on region of

the country, urban/rural status, PSU size, geographic location within regions, and population characteristics. Designs with approximately $H = 100$ primary stage strata are common in multistage samples of U.S. households, but the actual number of strata employed in the primary stage sample can range from less than ten to hundreds depending on available stratification variables and the number of PSUs, a_h, that will be selected from each stratum.

Depending on the study sample design, from 12 to 20 of the primary stage strata contain only a single **self-representing (SR)** metropolitan PSU. Each SR PSU is included in the sample with certainty in the primary stage of selection—for these strata, "true sampling" begins at the second stage of selection. The remaining **nonself-representing (NSR)** primary stage strata in each design contain more than one sample PSU. From each of these NSR strata, one or more PSUs is sampled with **probability proportionate to its size (PPS)** as measured in occupied housing unit counts reported at the most recent census. The number of PSUs selected from each primary stage stratum is decided by the sample allocation. The University of Michigan Survey Research Center (SRC) National Sample Design (Heeringa and Connor, 1995) allocates one primary stage selection per stratum. More commonly, a minimum of two sample PSUs are selected from each primary stage stratum. A "two-per-stratum" design in which exactly two PSUs ($a_h = 2$) are selected with PPS from each stratum maximizes the number of strata that can be formed for selecting a primary stage sample of

$$a = \sum_{h=1}^{H} a_h \text{ PSUs.}$$

The two-per-stratum allocation is also common because a minimum of two PSUs per primary stage stratum is required to estimate sampling variances of estimates from complex samples. One-per-stratum primary stage samples, like that used by the University of Michigan's SRC for HRS and NCS-R, maximize the stratification potential but require a collapsing of design strata to create pseudo-strata for complex sample variance estimation (see Chapter 4).

2.8.2 Secondary Stage Sampling

The designated second-stage sampling units (SSUs) in multistage samples are commonly termed **area segments**. Area segments are formed by linking geographically contiguous Census blocks to form units with a minimum number of occupied housing units (typically 50–100 based on the needs of the study). Within sample PSUs, SSUs may be stratified at the county level by geographic location and race/ethnicity composition of residents' households. Within each sample PSU, the actual probability sampling of SSUs is performed with probabilities proportionate to census counts of the occupied housing units

for the census blocks that comprise the area segment. The number of SSUs that are selected within each sample PSU is determined by survey statisticians to optimize the cost and error properties of the multistage design. The number of SSUs selected in an SR PSU varies in proportion to the total size of the PSU selected with certainty (e.g., HRS and NCS-R use approximately 48 area segments in the New York self-representing PSU). Depending on the total size of the survey and the cost structure for data collection operations, typically from 6 to 24 SSUs are selected from each NSR PSU.

2.8.3 Third and Fourth Stage Sampling of Housing Units and Eligible Respondents

Prior to the selection of the third stage sample of households, field staff from the survey organization visit each sample SSU location and conduct an up-to-date enumeration or "listing" of all housing units located within the physical boundaries of the selected area segments. A third-stage sample of housing units is then selected from the enumerative listing according to a predetermined sampling rate. The third-stage sampling rates for selecting households in the multistage area probability samples are computed using the following "selection equation" (Kish, 1965):

$$\begin{aligned} f &= f_1 \times f_2 \times f_3 \\ &= \frac{MOS_\alpha \times a_h}{MOS_h} \times \frac{b_\alpha \times MOS_{ssu}}{MOS_\alpha} \times \frac{C_h}{MOS_{ssu}} \end{aligned} \tag{2.22}$$

In the final selection equation derived in (2.22), we have the following notation: f = the overall multistage sampling rate for housing units; MOS_α = total population measure of size in the selected PSU α; MOS_h = total population measure of size in the design stratum h; a_h = number of PSUs to be selected from design stratum h; b_α = number of area segments selected in the PSU α; MOS_{ssu} = total household measure of size for the SSU; C_h = a stratum-specific constant = $(f \times MOS_{stratum})/a_h b_\alpha$.

For example, the third-stage sampling rate for selecting an equal probability national sample from the listed housing units for the selected PSUs and area segment SSUs is

$$f_3 = \frac{f}{f_1 \times f_2} = \frac{f \times MOS_h}{a_h \times b_\alpha \times MOS_{ssu}} \tag{2.23}$$

The third-stage sampling rate is computed for each selected SSU in the sample design. This rate is then used to select a random sample of actual housing units from the area segment listing.

Each sample housing unit is then contacted in person by an interviewer. Within each cooperating sample household, the interviewer conducts a short screening interview with a knowledgeable adult to determine if household members meet the study eligibility criteria. If the informant reports that one or more eligible adults live at the sample housing unit address, the interviewer prepares a complete listing of household members and proceeds to randomly select a respondent for the study interview. The random selection of the respondent is often performed using a special adaptation of the objective household roster/selection table method developed by Kish (1949).

Despite the obvious effort and complexity that goes into fielding a multistage area probability sample, relatively simple specifications of primary stage stratum, primary stage cluster (PSU), and final analysis weight variables will be required for analysts desiring appropriate analyses of survey data from this common household sampling design. A detailed discussion of a unified approach to variance estimation for multistage samples is given in Chapter 4.

2.9 Special Types of Sampling Plans Encountered in Surveys

As previously described, most large-scale social, economic, demographic, and health-related surveys are designed to provide the capability to make descriptive inferences to specific survey populations or to analyze multivariate relationships in a population. Although the techniques for applied survey data analysis presented in the following chapters are generally applicable to all forms of probability sample designs suitable for population estimation and inference, some fields of population survey research (e.g., surveys of businesses, hospitals and other nonhousehold units that vary in size and "importance") have developed special methods that will not be covered in detail in this volume. Researchers who are working with survey data for these populations are encouraged to use the survey literature to determine current best practices for these special population surveys.

Survey research on natural populations in environmental (e.g., forestry or fisheries), geological, and some human and animal epidemiological studies is increasingly turning to **adaptive sample designs** to optimize observation, estimation, and inference. If the reader's data are of these types, they will need to use special procedures for estimation and inference. An excellent reference on adaptive sample design can be found in Thompson and Seber (1996).

Increasingly, adaptive sampling procedures are being employed in major population surveys. Groves and Heeringa (2006) apply the term **responsive design** to surveys that adapt sampling, survey measurement, and nonresponse follow-ups to empirical information that is gathered in the survey process. One sampling technique that is critical to the responsive design of

surveys is the use of multiphase sampling, in which sample cases may be subsampled for further contact and interview at a time point, *t*, conditional on the prior disposition (e.g., number of calls, success with contact, resistance to interview) of the case. Presently, the stochastic nature of the sample disposition of each case at time *t* is ignored, and the data are weighted for estimation as though the disposition of cases was a deterministic (fixed) outcome. Current and future research is expected to lead to improved procedures for estimation and inference in multiphase sample designs.

Occasionally, population-based survey methods are employed to perform research that is purely analytical. These include studies that fall in the category of epidemiological case-control designs; randomized population-based experiments including "group randomized trials"; or model-based designs for research on hierarchical or multilevel populations (e.g., research on student, classroom, and school effects). Analysis of "survey" data from these types of analytical research designs requires special approaches. Chapter 12 will explore approaches to several of the more common analytical designs that use survey-like procedures to collect data, but, again, with data of this type, the reader is encouraged to also turn to the statistical literature for an up-to-date and more in-depth description of best practices. See, for example, Burns et al. (1996), Heeringa et al. (2001), and Raudenbush (2000).

3

Foundations and Techniques for Design-Based Estimation and Inference

3.1 Introduction

The fundamental theory and practical procedures for estimation and inference for complex sample survey data have been under development for almost a century. The foundations are present in the work on randomization and the "representative method" (Bowley, 1906; Fisher, 1925) and what is generally agreed to be the breakthrough paper by Neyman (1934) on the theory of design-based inference from probability sample designs. Even today, estimation and inference for survey data remain an evolving field with important new developments in applications of survey data to hierarchical and latent variable modeling (Rabe-Hesketh and Skrondal, 2006), estimation of small-area statistics (Rao, 2003), and Bayesian approaches to model estimation and inference using survey data (Little, 2003). The general concept of design-based inference and its application to survey data analysis was introduced in Section 2.2. The aim of this chapter is to expand the reader's understanding of the key components of design-based approaches to estimation and inference from survey data: consistent estimation of population statistics; robust, distribution-free methods for estimating the sampling variance of estimates; construction and interpretation of confidence intervals; and design-adjusted test statistics for hypothesis testing.

The chapter opens in Section 3.2 with a simple introduction to finite populations and a superpopulation model—two theoretical concepts that have only subtle implications for how the survey data are actually analyzed but are important concepts for interpretation and reporting of the survey findings. Confidence intervals are a primary tool for presenting survey estimates and the corresponding degree of uncertainty due to population sampling. Section 3.3 introduces the confidence interval as the organizing framework for discussion of the three main components of design-based inference: weighted sample estimates of population statistics (Section 3.4); critical values from the sampling distribution of the estimates (Section 3.5); and robust estimates of the standard error of the sample estimate (Section 3.6). Inference

from survey data is not limited to confidence interval construction and interpretation. Tests of specific hypotheses concerning the true population value of a statistic may also be performed based on survey data. Section 3.7 compares approaches to common hypothesis tests under simple random sampling assumptions to the corresponding test procedures for complex sample survey data. The chapter concludes in Section 3.8 with a discussion of the potential impact of sources of total survey error on estimation and inference for survey data.

3.2 Finite Populations and Superpopulation Models

Chapters 1 and 2 have touched on the healthy tension in the historical statistical literature between design-based and model-based approaches to survey inference. There has generally been little debate over the application of design-based inference for **descriptive analysis** of probability samples of survey populations. The objectives of the analysis are clear—to describe with measurable uncertainty due to sampling a population characteristic or process as it exists in a defined finite population. More controversy arises when researchers go beyond descriptive analysis to **analytic uses** of survey data sets—uses that estimate models to explore more universal, multivariate relationships or even to possibly understand causality in relationships among variables. To unify our thinking about finite population statistics versus a more universal model of processes or outcomes, sampling theoreticians have introduced the concept of a **superpopulation model**.

Figure 3.1 illustrates this concept. The figure shows two finite survey populations, each somewhat different yet each generated by the same overarching superpopulation model. Research teams conducting probability sample surveys in the two respective finite populations would estimate their finite population regression parameter as

$$b_{1(1)} = \hat{B}_{1(1)} = \frac{\sum_{i=1}^{n_{(1)}} w_i \cdot y_i \cdot x_i}{\sum_{i=1}^{n_{(1)}} w_i \cdot x_i^2}; \quad b_{1(2)} = \hat{B}_{1(2)} = \frac{\sum_{j=1}^{n_{(2)}} w_j \cdot y_j \cdot x_j}{\sum_{j=1}^{n_{(2)}} w_j \cdot x_j^2} \tag{3.1}$$

In each case, the design-based estimates, $b_{1(1)}$ or $b_{1(2)}$, would be consistent estimates of the corresponding finite population regression coefficients, $B_{1(1)}$ or $B_{1(2)}$. Note that sampling weights (denoted by w_i or w_j) are used to compute the estimates.

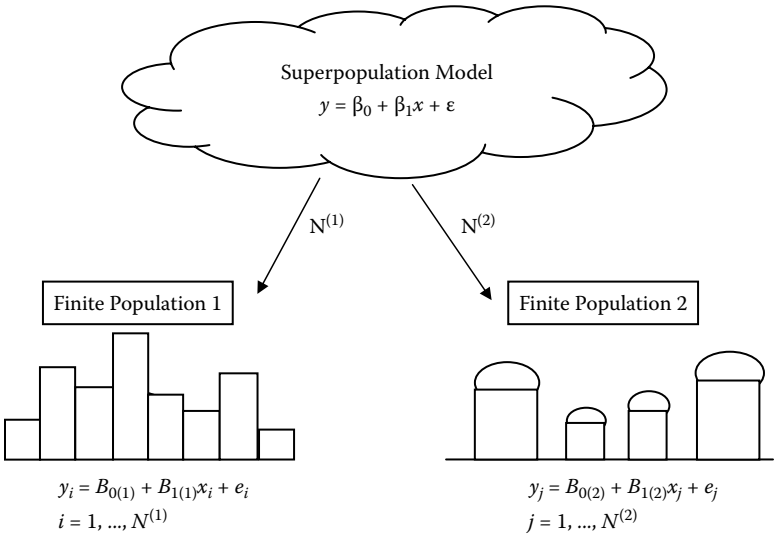


FIGURE 3.1
Superpopulation model and finite populations.

What if the statistical objective of each research team was not to simply estimate the best regression model for its survey population but to infer about the universal model that governs the relationship between y and x? The concept of a superpopulation model provides a bridge to the broader inference by postulating that the two finite populations are simply a size $N^{(1)}$ or $N^{(2)}$ "sample" of pairs of (x,y) observations generated from the superpopulation regression model: $y = \beta_o + \beta_1 \cdot x + \varepsilon$. (See Theory Box 7.1 in Chapter 7 for more details.)

The theoretical linkage of finite population samples to a superpopulation model is not foremost in the mind of an applied survey data analyst as he or she sits down at the computer to begin working with a survey data set. It will also have little bearing on how that analyst applies the design-based statistical analysis techniques described in the second half of this book. Many survey researchers can become careless about reporting conclusions from their data analysis, suggesting universality of their findings when in fact the sample data may have been drawn from a small or very unique survey population. It is important for research analysts reporting analysis results to be clear with their audiences concerning the nature of their inferential statements. Are the inferences targeted to the specific survey population, or is the aim to broaden the scope of the inference to imply more generalizable findings that extend beyond the boundaries of the finite population? If it is the latter, one should expect to provide scientific support for the inferential arguments. Is there a replication of one's findings in surveys of other distinct finite populations? Can one make a convincing theoretical argument for a single universal model for the relationships observed in the finite population?

3.3 Confidence Intervals for Population Parameters

> The form of this solution consists in determining certain intervals which I propose to call confidence intervals…. —Jerzy Neyman, 1934

This quote from Neyman (1934) capped nearly three decades of research and experimentation by leading statisticians of the day (including R. A. Fisher and E. S. Pearson) to establish a theory of estimation for the representative method of sampling (i.e., simple random samples and other probability samples). Since the mid-1930s when the paper was published, Neyman's confidence intervals (CIs) have played a leading role in population estimation and inference for survey data. A generic form for an approximate $100(1 - \alpha)\%$ CI for a population parameter, where α represents a level of significance (e.g., 0.05, corresponding to a 95% confidence interval), is

$$\hat{\theta} \pm t_{1-\alpha/2,df} \cdot se(\hat{\theta}) \tag{3.2}$$

where

$\hat{\theta}$ is an unbiased or consistent estimate of the population parameter, θ;
$t_{1-\alpha/2,df}$ is the value of the Student t distribution with df degrees of freedom;
df is the degrees of freedom for variance estimation under the sample design; and
$se(\hat{\theta})$ is a robust estimate of the standard error of $\hat{\theta}$.

The confidence interval summarizes our uncertainty in inferring a true population value from a single sample. In simple terms, if the identical sample design was used to repeatedly draw samples from the survey population and the 95% CI was created for each sample, 95 in 100 of all such sample CIs would contain the true population value. Note that it is the theoretical distribution of sample estimates over repeated sampling from the population that provides "support" for the statement of uncertainty.

Construction of confidence intervals for population parameters therefore requires three components: a consistent estimate, $\hat{\theta}$; the appropriate critical value from $t_{1-\alpha/2,df}$; and the standard error of the estimate, $se(\hat{\theta})$. The following three sections describe the underlying theory and derivation of each component.

3.4 Weighted Estimation of Population Parameters

Survey weights play a key role in consistent, design-based estimation of finite population parameters. **Consistent estimators** may be **unbiased**

estimators. They may also be estimators such as the ratio mean for stratified, cluster sample data that have a small bias that decreases as the sample size, n, increases and disappears as the sample size approaches the population size, N. Some examples of consistent, weighted estimators of population statistics include

$$\bar{y}_w = \frac{\sum_{i=1}^{n} w_i \cdot y_i}{\sum_{i=1}^{n} w_i} \tag{3.3}$$

to estimate the population mean, $\bar{Y}$

$$b_{1,w} = \frac{\sum_{i=1}^{n} w_i \cdot y_i \cdot x_i}{\sum_{i=1}^{n} w_i \cdot x_i^2} \tag{3.4}$$

to estimate the simple linear regression coefficient, B_1

Weighted estimation is also employed in iterative estimation of parameters for a large class of generalized linear models (GLMs). Estimation of GLMs from survey data involves a "weighted" likelihood function, weighted first derivatives (the score or estimating equations), and then application of the **Newton-Raphson algorithm** to find the coefficient estimates that maximize the weighted **pseudo-likelihood**. To illustrate how weights enter GLM estimation, the weighted pseudo-log-likelihood function used in fitting a logistic regression model to survey data takes the following form:

Pseudo ln(Likelihood):

$$\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_\alpha} w_{h\alpha i} y_{h\alpha i} \cdot \ln(\hat{\pi}(\boldsymbol{x}_{h\alpha i})) + \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_\alpha} w_{h\alpha i}(1 - y_{h\alpha i}) \cdot \ln(1 - \hat{\pi}(\boldsymbol{x}_{h\alpha i})) \tag{3.5}$$

where

$y_{h\alpha i} \in (0,1)$ is the binary dependent variable in the logistic model;

$w_{h\alpha i}$ is the weight value for case i in cluster α and stratum h;

$\boldsymbol{x}_{h\alpha i}$ is the vector of predictor variables in the logistic regression model;

$\hat{\pi}\left(\boldsymbol{x}_{h\alpha i}\right) = e^{\boldsymbol{x}_{h\alpha i}\boldsymbol{b}} \,/\, (1 + e^{\boldsymbol{x}_{h\alpha i}\boldsymbol{b}})$ is the predicted probability that $y_{h\alpha i} = 1$.

These and other weighted approaches to estimating population parameters and model coefficients will be introduced as we cover each of the major procedures for survey data analysis.

Since the early 1950s, survey statisticians, economists, and researchers in other disciplines have debated the role that weights should play in model-based inferences from survey data. Lohr (1999, Section 11.3) provides an excellent review of this debate, including a discussion of contributions from Brewer and Mellor (1973), Hansen, Madow, and Tepping (1983), and DuMouchel and Duncan (1983). A strong argument that emerges from papers by Pfeffermann and Holmes (1985) and Kott (1991) is that the survey weights protect the analyst against unknowingly misspecifying the model by omitting predictor variables or variable interactions that may be associated with the survey weight value (see Section 2.7). Within the past 20 years, leading statisticians who approach survey data from a model-based perspective have endorsed this view that survey weights can serve as a "proxy" for important features of the sample design that have a bearing on the response variable of interest (Little, 1991).

To illustrate this property of weighted estimation, consider the Health and Retirement Study (HRS) sample of black adults. HRS employs a two-fold oversampling of households in U.S. Census blocks with 10% or greater black households to cost effectively increase the size of the black subpopulation sample. Black adults living in the geographic domain of higher-density blocks have twice the selection probability of those who live on a block with fewer than 10% black households. The HRS weights for black respondents adjust for the differential selection probabilities. The HRS public use data set does contain a variable that is coded for the race/ethnicity of the respondent; however, the only source of information about whether the respondent lives in a low- or high-density black neighborhood is encoded in the relative value of the survey weight. If the survey weights were ignored in a regression analysis, the potentially crucial influence of the disproportionate representation of black adults from higher-density (and often lower-income) blocks would be omitted from the model.

Table 3.1 compares two estimated regression models where the dependent variable is the natural log of 2006 household income for HRS black respondents. Each model includes demographic covariates of interest (age, gender of head of household, education level) as well as predictors that control for geographic region and urbanicity of residence. The first estimated model ignores the weights in the computation of model coefficients and their standard errors. The estimated coefficients in the second model correctly incorporate the HRS weights, and the estimated standard errors for the estimated coefficients correctly incorporate design effects for stratification, clustering, and weighting (we discuss determination of the p-values given complex sample designs later in this chapter). Both estimated models point to the importance of education level in determining the income level of HRS black households; however, the estimated effect of a bachelor's degree or even some college

TABLE 3.1

Regression Models of Log-Transformed Household Income for the HRS Black Subpopulation

	Regression Parameter Estimate (Standard Error, *p*-value)	
Independent Variable	**Unweighted**	**Weighted (Design-Based SE)**
Age (continuous)	0.0026 (0.0058, $p = 0.66$)	0.0056 (0.0093, $p = 0.54$)
Gender		
Female	–0.4629 (0.1246, $p < 0.001$)	–0.3034 (0.2199, $p = 0.17$)
Male	Reference category	Reference category
Education		
Grade 0–11	–1.5585 (0.1991, $p < 0.0001$)	–1.9016 (0.2610, $p < 0.0001$)
Grade 12	–1.0304 (0.2011, $p < 0.0001$)	–1.5177 (0.2871, $p < 0.0001$)
Grade 13–15	–0.5145 (0.2152, $p < 0.0001$)	–0.7114 (0.1330, $p < 0.0001$)
Grade 16+	Reference category	Reference Category
Region		
Northeast	0.0804 (0.2743, $p = 0.77$)	0.1462 (0.1680, $p = 0.38$)
Midwest	–0.3331 (0.2635, $p = 0.21$)	–0.2423 (0.2614, $p = 0.36$)
South	–0.2525 (0.2476, $p = 0.31$)	–0.3519 (0.2405, $p = 0.15$)
West	Reference Category	Reference Category
Urbanicity		
Urban	–0.0697 (0.1690, $p = 0.68$)	–0.0553 (0.3751, $p = 0.88$)
Suburban	0.5878 (0.1965, $p = 0.76$)	–0.0262 (0.4764, $p = 0.58$)
Rural	Reference Category	Reference Category

Note: $n = 2{,}465$.

training (13–15 years) is much larger when the survey weights are employed in the estimation. This can be attributed to the fact that a higher proportion of the more highly educated HRS black respondents will reside in the more integrated neighborhoods of the low-density black domain. Consequently, their contribution to the estimated model will be "up-weighted" when the survey weights are applied. (For a second illustration of how survey weights can influence the estimation of a regression model, see the Chapter 7 example of estimation of a model of diastolic blood pressure based on the National Health and Nutrition Examination Survey [NHANES] data.)

There will be situations where the application of survey weights will have no real impact on a survey analysis and may even result in a loss of statistical precision relative to the unweighted alternative. Chambers and Skinner (2003) use a likelihood framework to demonstrate that the benefit of weighted analysis occurs when the sample and corresponding weights reflect an **informative sample design**. In simple terms, an informative design is one in which the design features—stratification, clustering, disproportionate sampling—are associated with the response variable of interest. Korn and Graubard (1999, Sections 4.3 and 4.4) also provide examples of situations

where the application of survey weights does not have a substantial impact on inferences and discuss methods for computing the inefficiency in survey estimates that can arise when sampling weights are used unnecessarily.

The general approach that we adopt in this text is to begin with the assumption that the features of the complex sample design and the associated survey weights are in fact related to the response variable of interest. Consequently, the design-based analyses illustrated in Chapters 5 through 12 will employ weighted estimation of population parameters. However, we agree with Lohr (1999) on the value of comparing models estimated both with and without the survey weights. If the estimated model parameters are very different, the analyst should be able to explain the difference based on documented knowledge about how the weights were constructed (e.g., oversampling of low-income households, persons with disability, older adults).

3.5 Probability Distributions and Design-Based Inference

> The whole procedure consists really in solving the problems which Professor Bowley termed direct problems: given a hypothetical population, to find the distribution of certain characters in repeated samples. If this problem is solved, then the solution of the other problem, which takes the place of the problem of inverse probability, can be shown to follow. —Jerzy Neyman, 1934

3.5.1 Sampling Distributions of Survey Estimates

Neyman's (1934) "distribution of certain characters in repeated samples" is termed the **sampling distribution** of a sample estimate. The theoretical sampling distribution is based on all possible samples of size n that could be selected from a finite population of N elements. Using a single sampling plan, if all possible samples of size n from the N population elements were drawn in sequence, sample estimates were computed for each selected sample, and a histogram of the estimated values was plotted, the shape of the sampling distribution would emerge. Provided that the sample size, n, was sufficiently large, the distribution that would begin to appear as each new sample estimate was added to the histogram would be the familiar bell-shaped curve of a Normal distribution.

Figure 3.2 illustrates a set of nine simulated sampling distributions for sample estimates of the population mean. Each individual graph in this figure represents the histogram of sample estimates, $\bar{y}$, computed from 5,000 independent samples from a large finite population with known mean $\bar{Y} = 25$. The nine simulated sampling distributions displayed in this figure represent nine different probability sampling plans—three levels of sample size (n =

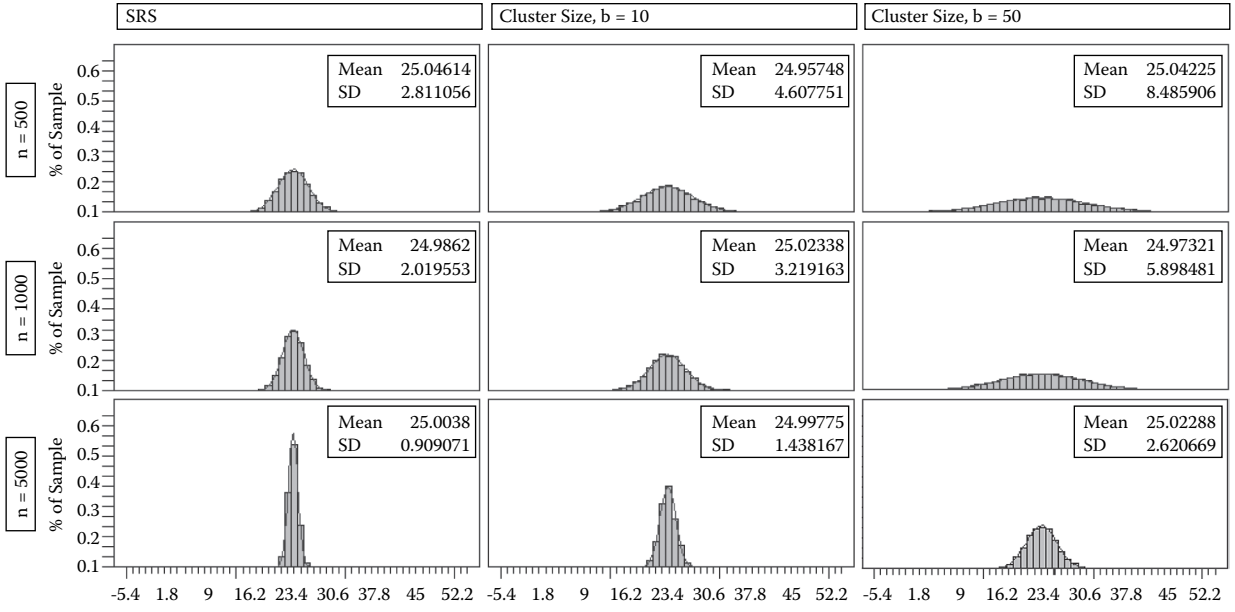


FIGURE 3.2
Sampling distributions for a survey estimate (n = 5,000 simulated samples, $\bar{Y}$ = 25).

500, n = 1,000, n = 5,000) and three levels of clustering (no clustering and clusters of size B = 10 and B = 50). Since $\bar{y}$ is an unbiased estimator regardless of the sampling plan or the sample size, each sampling distribution is centered at the population mean, $\bar{Y} = 25$. As the sample size decreases or the size of sample clusters increases, the dispersion of sample estimates about the population mean value increases. The degree of dispersion of the sample estimates about the mean of the sampling distribution is the **sampling variance** associated with the sample design, which can be written as

$$Var(\bar{y}) = \sum_{s=1}^{S} p(s) \cdot \left(\bar{y}_s - E(\bar{y}_s)\right)^2 \tag{3.6}$$

where

s = 1, …, S indexes all possible samples of size n under the design;
$p(s)$ = the probability that sample s was chosen from the set of S possibilities;
$\bar{y}_s$ = the estimate for sample s.

The square root of the sampling variance is the **standard error** of a probability sample estimate, denoted by $SE(\bar{y})$. Or equivalently, the standard error of a design-based estimate is simply the **standard deviation** of the sampling distribution.

In real-world survey samples, a single sample is observed. It is never practically feasible to observe the full sampling distribution of an estimate, its mean, its variance, or its distributional shape. So how is it possible to make inferential statements based on a sampling distribution that is never observed? Briefly, statistical theory shows that if the sample size n is sufficiently large (e.g., 100 cases) and if $\hat{\theta}$ is an unbiased or otherwise consistent estimator of the population value θ, the sampling distribution converges to an approximately Normal distribution: $f(\hat{\theta}) \sim N(\theta, Var(\hat{\theta}))$. Consequently, the test statistic

$$t = \frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

where both $\hat{\theta}$ and $se(\hat{\theta})$ are *estimated* from the survey sample data, follows the Student t probability distribution with df degrees of freedom (to be defined in the following section). This test statistic can be "inverted" to derive the following probability statement:

$$P_s\left\{\hat{\theta} - t_{1-\alpha/2,df} \cdot se(\hat{\theta}) \le \theta \le \hat{\theta} + t_{1-\alpha/2,df} \cdot se(\hat{\theta})\right\} \cong 1 - \alpha$$

This reexpression of the test statistic as a range of values for θ is the basis for the 100(1 – α)% confidence interval presented in Equation 3.2.

3.5.2 Degrees of Freedom for *t* under Complex Sample Designs

Probability distributions such as the Student t, χ^2, and F play a critical role in the construction of confidence intervals for population values or as the reference distributions for formal tests of hypotheses concerning population parameters. Included in the quantities that define the shape of these distributions are **degrees of freedom** (*df*) parameters. The degrees of freedom are indices of how precisely the true variance parameters of the reference distribution have been estimated from the sample design. Sample designs with large numbers of degrees of freedom for variance estimation enable more precise estimation of the true variance parameters of the reference distribution. Conversely, the smaller the degrees of freedom afforded by the sample design, the less precisely these variance parameters are estimated. Consider the $(1 - \alpha/2 = 0.975)$ critical values for the Student t distribution with varying degrees of freedom: $t_{.975,1} = 12.706$; $t_{.975,20} = 2.0860$; $t_{.975,40} = 2.0211$; $t_{.975,\infty} = Z_{.975} = 1.9600$. Whenever an analyst (or his or her computer software) derives a confidence interval or test statistic from sample data, variance parameters that define the appropriate t, χ^2, or F reference distribution must be estimated from the sample data.

Precise determination of the degrees of freedom for variance estimation available under complex sample designs used in practice is difficult. Currently, computer software programs for the analysis of complex sample survey data employ a **fixed degrees of freedom rule** to determine the degrees of freedom for the reference distribution used to construct a confidence interval (e.g., $t_{1-\alpha/2,df}$) or a p-value for a hypothesis test (e.g., $P(F > F_{k,d})$):

$$df_{des} = \sum_{h=1}^{H} \left(a_h - 1\right) = \sum_{h=1}^{H} a_h - H = \#\,clusters - \#\,strata \tag{3.7}$$

Interested readers are referred to Theory Box 3.1 for a more in-depth (yet not strictly theoretical) explanation of the basis for this rule.

The fixed rule for determining degrees of freedom for complex sample designs is applied by software procedures designed for the analysis of survey data whenever the full survey sample is being analyzed. However, programs may use different rules for determining degrees of freedom for subpopulation analyses. For subpopulations, improved confidence interval coverage of the true population value is obtained using a "variable" degrees of freedom calculation method (Korn and Graubard, 1999):

$$df_{\text{var}} = \sum_{h=1}^{H} I_h \cdot (a_h - 1) \tag{3.8}$$

THEORY BOX 3.1 DEGREES OF FREEDOM FOR VARIANCE ESTIMATION

Consider the pivotal t statistic, $t = (\bar{y} - \bar{Y}_0) / se(\bar{y})$. Under simple random sampling,

$$t_{n-1,SRS} = \frac{(\bar{y} - \bar{Y}_0)}{se(\bar{y})} = \frac{(\bar{y} - \bar{Y}_0)}{\sqrt{s^2 / n}} = \frac{(\bar{y} - \bar{Y}_0)}{\sqrt{\sum_{i=1}^{n} [(y_i - \bar{y})^2 / (n-1)] / n}}$$

A total of $n - 1$ independent contrasts (squared differences) contribute to the SRS estimate of S^2 (once the mean is known, there are only $n - 1$ unique pieces of information for estimating the variance). Consequently, for an SRS design, the t statistic is referred to a Student t distribution with $n - 1$ degrees of freedom.

Now, consider this same test statistic under a more complex stratified cluster sample design:

$$t_{df,complex} = \frac{(\bar{y}_{st,cl} - \bar{Y}_0)}{\sqrt{\text{var}(\bar{y}_{st,cl})_{complex}}} = \frac{(\bar{y}_{st,cl} - \bar{Y}_0)}{se(\bar{y}_{st,cl})_{complex}}$$

$$\text{var}(\bar{y}_{st,cl}) = \sum_{h=1}^{H} W_h^2 \frac{1}{a_h} \left[\frac{1}{(a_h - 1)} \cdot \frac{1}{b_h^2} \left\{ \sum_{\alpha=1}^{a_h} y_{h\alpha}^2 - \frac{y_h^2}{a_h} \right\} \right]$$

$$= \sum_{h=1}^{H} \frac{W_h^2}{a_h b_h^2} \left[\frac{1}{(a_h - 1)} \left\{ \sum_{\alpha=1}^{a_h} (y_{h\alpha} - \frac{y_h}{a_h})^2 \right\} \right]$$

Under a design with $h = 1, \ldots, H$ strata and $\alpha = 1, \ldots, a_h$ equal-sized PSUs per stratum (i.e., the sample size in each PSU in stratum h is b_h), each stratum contributes $a_h - 1$ independent contrasts to the estimate of the $\text{var}(\bar{y}_{st,cl})$. The t-statistic for the complex sample design is no longer referred to a Student t distribution with $n - 1$ degrees of freedom. Instead, the correct degrees of freedom for variance estimation under this complex sample design are

$$design\ df_{fixed} = \sum_{h=1}^{H} (a_h - 1) = \sum_{h=1}^{H} a_h - H = \#\,clusters - \#\,strata$$

The variance estimation technique known as Taylor series linearization (TSL) in particular (see Section 3.6.2) involves approximating

complex nonlinear statistics with simpler linear functions of linear statistics and then estimating the variances of the linear statistics (enabling the use of known variance estimation formulas for complex designs, as shown previously). This gives rise to the use of the fixed degrees of freedom rule when using this variance estimation procedure.

For additional theoretical details on the determination of degrees of freedom in variance estimation under complex sample designs, we refer readers to Korn and Graubard (1999, Section 5.2).

where

I_h = 1 if stratum h has 1 or more subpopulation cases, 0 otherwise;
a_h = the number of clusters (PSUs) sampled in stratum h = 1, …, H.

The variable degrees of freedom are determined as the total number of clusters in strata with 1+ subpopulation cases minus the number of strata with at least one subpopulation observation. Rust and Rao (1996) suggest the same rule for calculating degrees of freedom for test statistics when replicated variance estimation methods are used to develop standard errors for subpopulation estimates. See Section 4.5 for a more in-depth discussion of survey estimation for subpopulations.

3.6 Variance Estimation

Survey analysis that employs confidence intervals (or hypothesis test statistics and *p*-values) requires estimation of the sampling variability of the sample estimates. For simple statistics computed from data collected under equal-probability SRS designs or stratified random sample designs, exact expressions for the estimator of the sampling variance may be derived (Cochran, 1977). Survey analysis programs in Stata or SUDAAN and many other software packages allow the user to choose the exact formula for these simple sample designs.

Consider estimating the population mean under a stratified, *st*, random sample, *ran*, design:

$$\bar{y}_{st,ran} = \sum_{h=1}^{H} W_h \cdot \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \bar{y}_h \tag{3.9}$$

An unbiased estimate of the sampling variance of the mean estimate is computed using the following analytical formula:

$$\text{var}(\bar{y}_{st,ran}) = \sum_{h=1}^{H} W_h^2 \cdot \text{var}(\bar{y}_h) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \cdot W_h^2 \cdot \frac{s_h^2}{n_h} \tag{3.10}$$

Note that this exact expression includes the finite population correction factor for each stratum. The variance expression for this most basic of stratified sample designs illustrates an important fact. For stratified samples, the sampling variance of an estimator is simply the sum of the variance contributions from each stratum. There is no between-stratum component of variance. This simple rule applies equally to more complex designs that use cluster sampling or unequal probability sampling within strata. This fact provides the rationale for using stratification in sample designs and for maximizing the between-stratum component of variance (and thus minimizing the within-stratum component).

When survey data are collected using a complex sample design with unequal size clusters or when weights are used in estimation, most statistics of interest will not be simple linear functions of the observed data, as in the case of Equation 3.9, and alternative methods for estimating the variances of the more complex statistics are required. Over the past 50 years, advances in survey sampling theory have guided the development of a number of methods for correctly estimating variances from complex sample survey data sets. The two most common approaches to the estimation of sampling errors for estimates computed from complex sample survey data sets are (1) through the use of a Taylor series linearization (TSL) of the estimator (and corresponding approximation to its variance) or (2) through the use of resampling variance estimation procedures such as balanced repeated replication (BRR) or jackknife repeated replication (JRR) (Rust, 1985; Wolter, 2007). BRR, JRR, and the bootstrap technique comprise a second class of nonparametric methods for conducting estimation and inference from complex sample survey data. As suggested by the generic label for this class of methods, BRR, JRR, and the bootstrap use replicated subsampling of the sample database to develop sampling variance estimates for linear and nonlinear statistics.

The following sections describe the basic features of the Taylor series linearization technique and the replication methods for estimating the variance of sample estimates from complex sample survey data.

3.6.1 Simplifying Assumptions Employed in Complex Sample Variance Estimation

Before turning to the computational methods, it is important to consider two simplifying assumptions that are routinely used in TSL, JRR, and BRR variance estimation programs available in today's statistical software packages:

1. Primary stage units in multistage sample designs are considered to be selected with replacement from the primary stage strata. Any finite population correction for the primary stage sample is ignored. The resulting estimates of sampling variance will be slight overestimates.
2. Multistage sampling within selected PSUs results in a single **ultimate cluster** of observations for that PSU. Variance estimation methods based on the ultimate clusters "roll up" the components of variance for the multiple stages into a single stage formula that requires only knowledge of the primary stage strata and PSU identifiers to perform the final calculations. All sources of variability nested within the PSU are captured in the composite variance estimate.

Figure 3.3 illustrates the concept of an ultimate cluster for a single sample PSU. Within the sample PSU, two additional stages of sampling identify

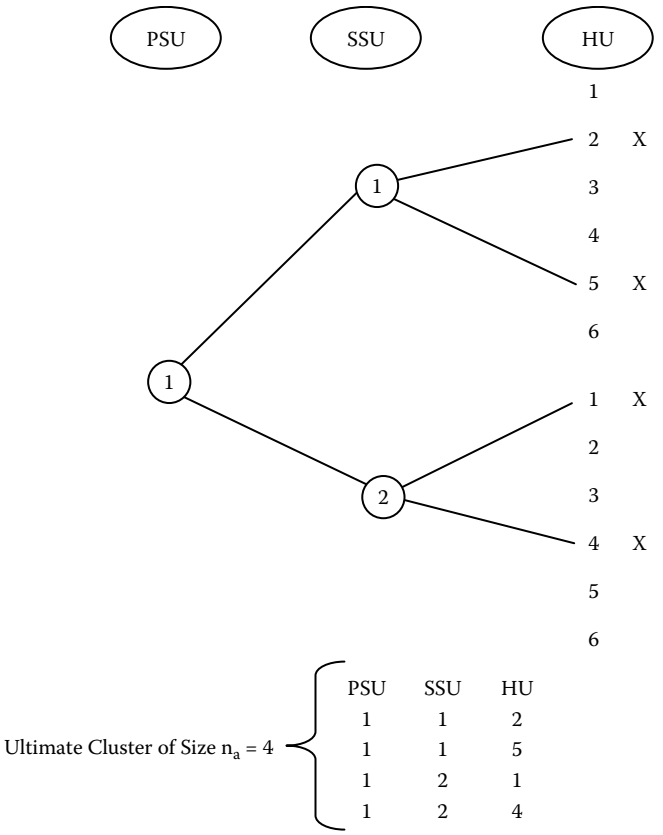


FIGURE 3.3
Ultimate cluster concept.

a unique cluster of four elements (units 2 and 5 within sample secondary stage unit [SSU] 1 and units 1 and 4 within sample SSU 2). Under assumption 1, PSUs are initially selected with replacement within first-stage sampling strata. Following this with-replacement selection of PSUs, all possible ultimate clusters that could arise from the multistage sampling procedure could be enumerated (in theory) within each selected PSU. (The multistage selection process emulates this process of enumerating and sampling a single ultimate cluster of elements to represent the PSU, but at substantially reduced cost and effort.) Then, the survey statistician could (in theory) select a without-replacement sample of these ultimate clusters from all of the PSUs. In practice, one ultimate cluster is generally selected within each PSU, as shown in Figure 3.3. The resulting sample can be thought of as a single-stage without-replacement selection of ultimate clusters, where all elements within each selected ultimate cluster are sampled. This greatly simplifies variance estimation through the use of formulae for stratified, with-replacement sampling of ultimate clusters of observations. Since PSUs are typically sampled *without* replacement from within the primary stage strata in practice, use of the simpler variance estimation formulae results in a slight overestimate of the true variance (Kish, 1965, Section 5.3), which is an acceptably conservative approach for making inferences.

Figure 3.4 provides a simple illustration of a survey data set that is ready for analysis. The data set includes $n = 16$ cases. From each of $h = 1, \ldots, 4$ strata, a total of $a_h = 2$ PSUs has been selected with replacement. An ultimate cluster of $b = 2$ observations has been selected from each sample PSU. The data set contains a unique stratum code, a PSU (or cluster) code within each stratum, and a case-specific value for the weight variable, w_i. As discussed in more detail in Section 4.3, the stratum, cluster, and weight variables represent the minimum set of design variables that are required to perform analyses of sampling variability for complex sample designs. (Note: As an alternative to releasing the detailed stratum and cluster codes, survey data sets may include "replicate weights" for use by software that supports replicated variance estimation—see Section 4.2.1.) This simple example data set will be used in the following sections to illustrate the computational steps in the TSL, JRR, and BRR variance estimation methods.

3.6.2 The Taylor Series Linearization Method

Taylor series approximations of complex sample variances for weighted sample estimates of finite population means, proportions, and linear regression coefficients have been available since the 1950s (Hansen, Hurwitz, and Madow, 1953; Kish and Hess, 1959). Woodruff (1971) summarized the general application of the TSL methods to a broader class of survey statistics. Binder (1983) advanced the application of the TSL method to variance estimation for analysis techniques such as logistic regression or other generalized linear

4 Strata, 2 PSUs per stratum, ultimate clusters of 2 elements per PSU.				
Stratum	**PSU (Cluster)**	**Case**	y_i	w_i
1	1	1	.58	1
1	1	2	.48	2
1	2	1	.42	2
1	2	2	.57	2
2	1	1	.39	1
2	1	2	.46	2
2	2	1	.50	2
2	2	2	.21	1
3	1	1	.39	1
3	1	2	.47	2
3	2	1	.44	1
3	2	2	.43	1
4	1	1	.64	1
4	1	2	.55	1
4	2	1	.47	2
4	2	2	.50	2

FIGURE 3.4
Data set for example TSL, JRR, and BRR sampling variance calculations.

models. The TSL approach to variance estimation involves a noniterative process of five steps:

3.6.2.1 TSL Step 1

The estimator of interest is written as a function of weighted sample totals. Consider a weighted, combined ratio estimator of the population mean of the variable y (Kish, 1965):

$$\bar{y}_w = \frac{\sum_h \sum_\alpha \sum_i w_{h\alpha i} y_{h\alpha i}}{\sum_h \sum_\alpha \sum_i w_{h\alpha i}} = \frac{\sum_h \sum_\alpha \sum_i u_{h\alpha i}}{\sum_h \sum_\alpha \sum_i v_{h\alpha i}} = \frac{u}{v} \tag{3.11}$$

Notice that the estimator of the ratio mean can be expressed as a ratio of two weighted totals, u and v, which are sums over design strata, PSUs, and individual cases of the constructed variables $u_{h\alpha i} = w_{h\alpha i} \cdot y_{h\alpha i}$ and $v_{h\alpha i} = w_{h\alpha i}$. The concept of a sample total is not limited to sums of single variables or sums of weighted values. For reasons that will be explained more fully in later sections, the individual case-level variates used to construct the sample totals for complex sample survey data may be more complex functions involving many variates and functional forms. For example, under a stratified, cluster sample design, the following estimated totals are employed in TSL variance estimation for simple linear and simple logistic regression coefficients:

$$u = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} w_{h\alpha i} \cdot y_{h\alpha i} \cdot x_{h\alpha i}$$

$$v = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} w_{h\alpha i} \cdot x_{h\alpha i}^2; \text{ or even}$$

$$\text{SC} = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} \left[w_{h\alpha i} \cdot x_{h\alpha i} \cdot y_{h\alpha i} - w_{h\alpha i} \cdot x_{h\alpha i} \cdot \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{h\alpha i}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{h\alpha i}}} \right]$$

3.6.2.2 TSL Step 2

Like many other survey estimators for complex sample survey data, the weighted estimator of the population mean is a nonlinear function of the two weighted sample totals. Consequently,

$$Var(\bar{y}_w) = Var\left(\frac{u}{v}\right) \neq \frac{Var(u)}{Var(v)}$$

To solve the problem of the nonlinearity of the sample estimator, a standard mathematical tool, the Taylor series expansion, is used to derive an approximation to the estimator of interest, rewriting it as a linear combination of weighted sample totals:

$$\bar{y}_{w,TSL} = \frac{u_0}{v_0} + (u - u_o)\left[\frac{\partial \bar{y}_{w,TSL}}{\partial u}\right]_{u=u_o, v=v_o} + (v - v_o)\left[\frac{\partial \bar{y}_{w,TSL}}{\partial v}\right]_{v=v_o, u=u_o} + remainder$$

$$\bar{y}_{w,TSL} \cong \frac{u_0}{v_0} + (u - u_o)\left[\frac{\partial \bar{y}_{w,TSL}}{\partial u}\right]_{u=u_o, v=v_o} + (v - v_o)\left[\frac{\partial \bar{y}_{w,TSL}}{\partial v}\right]_{v=v_o, u=u_o}$$

$$\bar{y}_{w,TSL} \cong \text{constant} + (u - u_o) \cdot A + (v - v_o) \cdot B$$

where A and B symbolically represent the derivatives with respect to u and v, evaluated at the expected values of the sample estimates u_0 and v_0.

The quadratic, cubic, and higher-order terms in the full Taylor series expansion of $\bar{y}_w$ are dropped (i.e., the *remainder* is assumed to be negligible). Further, consistent (and preferably unbiased) sample estimates are generally used in place of the expected values of the sample estimates.

3.6.2.3 TSL Step 3

A standard statistical result for the variance of a linear combination (sum) is applied to obtain the approximate variance of the "linearized" form of the estimator, $\bar{y}_{w,TSL}$:

$$\begin{aligned}\mathrm{var}(\bar{y}_{w,TSL}) &\cong \mathrm{var}[\mathrm{constant} + (u-u_0)\cdot A + (v-v_0)\cdot B] \\ &\cong 0 + A^2\,\mathrm{var}(u-u_0) + B^2\,\mathrm{var}(v-v_0) + 2AB\,\mathrm{cov}(u-u_0, v-v_0) \\ &\cong A^2\,\mathrm{var}(u) + B^2\,\mathrm{var}(v) + 2AB\,\mathrm{cov}(u,v)\end{aligned}$$

where:

$$A = \frac{\partial \bar{y}_{w,TSL}}{\partial u}\Big|_{u=u_0, v=v_0} = \frac{1}{v_0};\ B = \frac{\partial \bar{y}_{w,TSL}}{\partial v}\Big|_{u=u_0, v=v_0} = -\frac{u_0}{v_0^2};\ \text{and}$$

u_0, v_0 are the weighted sample totals computed from the survey data.

Therefore,

$$\mathrm{var}(\bar{y}_{w,TSL}) \cong \frac{var(u) + \bar{y}^2_{w,TSL}\cdot var(v) - 2\cdot \bar{y}_{w,TSL}\cdot cov(u,v)}{v_0^2}$$

The sampling variance of the nonlinear estimator, $\bar{y}_{w,TSL}$, is thus approximated by a simple algebraic function of quantities that can be readily computed from the complex sample survey data. The sample estimates of the ratio mean, $\bar{y}_{w,TSL}$, and the sample total of the analysis weights, v_0, are computed from the survey data. The estimates of var(u), var(v), and cov(u,v) are computed using the relatively simple computational formulas described in Step 4.

3.6.2.4 TSL Step 4

Under the TSL method, the variance approximation in Step 3 has been derived for most survey estimators of interest, and software systems such as Stata and SUDAAN provide programs that permit TSL variance estimation for virtually all of the analytical methods used by today's survey data analyst. The sampling variances and covariances of individual weighted totals, u or v, are easily estimated using simple formulae (under an assumption of with-

replacement sampling of PSUs within strata at the first stage) that require knowledge only of the subtotals for the primary stage strata and clusters:

$$\text{var}(u) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \cdot \left[\sum_{\alpha=1}^{a_h} u_{h\alpha}^2 - \frac{u_h^2}{a_h} \right]$$

$$\text{var}(v) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \cdot \left[\sum_{\alpha=1}^{a_h} v_{h\alpha}^2 - \frac{v_h^2}{a_h} \right]$$

$$\text{cov}(u, v) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \cdot \left[\sum_{\alpha=1}^{a_h} u_{h\alpha} \cdot v_{h\alpha} - \frac{u_h \cdot v_h}{a_h} \right]$$

where : $u_{h\alpha}, u_h, v_{h\alpha}, v_h$ are defined below.

Provided that the sample strata codes, cluster codes, and weights are included in the survey data set, the **weighted totals** for strata and clusters are easily calculated. These calculations are illustrated here for u:

$$u_{h\alpha} = \sum_{i=1}^{n_\alpha} u_{h\alpha i}; \; u_h = \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_\alpha} u_{h\alpha i}$$

Returning to the example data set in Figure 3.4:

$$\bar{y}_{w,TSL} = \frac{\sum_h \sum_\alpha \sum_i w_{h\alpha i} y_{h\alpha i}}{\sum_h \sum_\alpha \sum_i w_{h\alpha i}} = \frac{\sum_h \sum_\alpha \sum_i u_{h\alpha i}}{\sum_h \sum_\alpha \sum_i v_{h\alpha i}} = \frac{u}{v} = \frac{11.37}{24} = 0.47375$$

$$v_0 = \sum_h \sum_\alpha \sum_i w_{h\alpha i} = 24$$

$$\text{var}(u) = 0.9777; \; \text{var}(v) = 6.0000; \; \text{cov}(u, v) = 2.4000$$

$$var\left(\bar{y}_{w,TSL}\right) \cong \frac{var(u) + \bar{y}_{w,TSL}^2 \cdot var(v) - 2 \cdot \bar{y}_{w,TSL} \cdot cov(u, v)}{v_0^2}$$

$$= \frac{0.9777 + 0.4737^2 \cdot 6.0000 - 2 \cdot 0.4737 \cdot 2.4000}{24^2}$$

$$= 0.00008731$$

$$se(\bar{y}_{w,TSL}) = 0.00934$$

3.6.2.5 TSL Step 5

Confidence intervals (or hypothesis tests) based on estimated statistics, standard errors, and correct degrees of freedom based on the complex sample design are then constructed and reported as output from the TSL variance estimation program. We show this calculation for a 95% confidence interval for the population mean based on the example data set in Figure 3.4 (note that df = 8 clusters – 4 strata = 4):

$$CI(\bar{y}_{w,TSL}) = \bar{y}_{w,TSL} \pm t_{1-\alpha/2,\, df} \cdot \sqrt{\text{var}_{TSL}(\bar{y}_{w,TSL})}$$

$$\text{e.g., } CI(\bar{y}_{w,TSL}) = 0.4737 \pm t_{1-\alpha/2,4} \cdot 0.0093$$

$$= 0.4737 \pm 2.7764 \cdot (0.0093) = (0.4478, 0.4996)$$

Most contemporary software packages employ the TSL approach as the default method of computing sampling variances for complex sample survey data. TSL approximations to sampling variances have been derived for virtually all of the statistical procedures that have important applications in survey data analysis. The following Stata (Version 10) syntax illustrates the command sequence and output for an analysis of the prevalence of at least one lifetime episode of major depression in the National Comorbidity Survey Replication (NCS-R) adult survey population:

```
. svyset seclustr [pweight=ncsrwtsh], strata(sestrat) ///
vce (linearized) singleunit(missing)

      pweight: ncsrwtsh
          VCE: linearized
  Single unit: missing
     Strata 1: sestrat
         SU 1: seclustr
        FPC 1: <zero>

. svy: mean mde
(running mean on estimation sample)

Survey: Mean estimation

Number of strata = 42  Number of obs =   9282
Number of PSUs =   84  Population size = 9282
                       Design df =         42

--------------------------------------------------------------
             |      Linearized
             | Mean      Std. Err. [95% Conf. Interval]
-------------+------------------------------------------------
         mde | .1917112 .0048768    .1818694   .201553
--------------------------------------------------------------
```

Note that Stata explicitly reports the linearized estimate of the standard error (0.0049) of the weighted estimate of the population proportion (0.1917). These Stata commands will be explained in more detail in the upcoming chapters.

3.6.3 Replication Methods for Variance Estimation

JRR, BRR, and the bootstrap form a second class of nonparametric methods for computing sampling variance of survey estimates. As suggested by the generic label for this class of methods, JRR, BRR, and the bootstrap use replicated subsampling of the database of sample observations to develop sampling variance estimates for linear and nonlinear statistics (Rust, 1985; Shao and Tu, 1995; Wolter, 2007).

The precursor to today's JRR, BRR, and bootstrap techniques for replicated variance estimation was a simple yet statistically elegant technique of interpenetrating samples or replicated sampling developed by P. C. Mahalanobis (1946) for agricultural surveys in Bengal, India. Mahalanobis's method requires selecting the complete probability sample of size n as a set of $c = 1, \ldots, C$ independent sample replicates from a common sampling plan. The replicated estimate of θ and the sampling variance of the simple replicated estimate are

$$\hat{\theta}_{rep} = \sum_{c=1}^{C} \hat{\theta}_c / C, \text{ the mean of the replicate sample estimates}$$
$$\text{var}(\hat{\theta}_{rep}) = \sum_{c=1}^{C} (\hat{\theta}_c - \hat{\theta}_{rep})^2 / C \cdot (C-1) \tag{3.12}$$

Despite their simplicity, simple replicated sample designs are rarely used in population survey practice. If a large number of replicate samples are used to obtain adequate degrees of freedom for variance estimation, the requirement that each replicate sample be a "miniature" of the full sample design restricts the design efficiency by limiting numbers of strata that can be employed at each stage of selection. If a highly efficient stratified design is employed, the number of independent replicates and corresponding degrees of freedom for variance estimation may be seriously limited.

During the late 1950s and 1960s, Mahalanobis's (1946) pioneering idea that replicated samples could be used to estimate sampling variances was extended to the BRR, JRR, and bootstrap methods. These techniques draw on the principles of the simple replication method but yield a more efficient procedure for creating the replicates and provide greater efficiency for estimating sampling variances by increasing degrees of freedom for comparable sample sizes and survey costs.

Each of these replication-based methods employs a generic sequence of five steps:

1. Sample replicates ($r = 1, \ldots, R$) of the full survey sample are defined based on assignment rules tailored to the BRR, JRR, and bootstrap techniques.
2. Full sample analysis weights are revised for each replicate to create $r = 1, \ldots, R$ **replicate weights**.
3. Weighted estimates of a population statistic of interest are computed for the full sample and then separately for each replicate (using the replicate weights).
4. A replicated variance estimation formula tailored to the BRR, JRR, or bootstrap method is used to compute standard errors.
5. Confidence intervals (or hypothesis tests) based on the estimated statistics, standard errors, and correct degrees of freedom are constructed.

JRR, BRR, and the bootstrap employ different methods to form the replicates (Step 1), which in turn requires minor modifications to Steps 2–4 of the general replicated variance estimation algorithm. The following sections describe each of the basic steps for the most common JRR and BRR methods.

3.6.3.1 Jackknife Repeated Replication

The JRR method of replicated variance estimation is applicable to a wide range of complex sample designs including designs in which two or more PSUs are selected from each of $h = 1, \ldots, H$ primary stage strata. The JRR method employs the following five-step sequence:

3.6.3.1.1 JRR Step 1

In most JRR applications, each replicate is constructed by deleting one or more PSUs from a single primary stage stratum. Operationally, the statistical software systems that provide JRR capability do not physically remove the cases for the deleted PSUs in creating each replicate sample. Instead, by assigning a zero or missing weight value to each case in the deleted PSU, the "replicate weight" sweeps the deleted cases out of the computation of the replicate estimate. Sample cases in all remaining PSUs (including those in all other strata) constitute the JRR replicate. See Figure 3.5, in which the second PSU in the first stratum is deleted to form a first JRR replicate.

Each stratum will contribute $a_h - 1$ unique JRR replicates, yielding a total of

$$R = \sum_{h=1}^{H} (a_h - 1) = a - H = \#\,clusters - \#\,strata$$

degrees of freedom for estimating standard errors of sample estimates. Following the "delete one" JRR method employed in Stata, SUDAAN, and

Stratum	PSU (Cluster)	Case	y_i	$w_{i,rep}$
1	1	1	.58	1x2
1	1	2	.48	2x2
1	2	1	.	.
1	2	2	.	.
2	1	1	.39	1
2	1	2	.46	2
2	2	1	.50	2
2	2	2	.21	1
3	1	1	.39	1
3	1	2	.47	2
3	2	1	.44	1
3	2	2	.43	1
4	1	1	.64	1
4	1	2	.55	1
4	2	1	.47	2
4	2	2	.50	2

FIGURE 3.5
Formation of a single JRR replicate for the data example.

other software packages, an additional final JRR replicate could be created for each stratum but it would add no additional precision to the final variance estimate. In our simple example with $H = 4$ strata and two PSUs per stratum, one JRR replicate will be created for each stratum. For a discussion of other methods for JRR replicate formation, see RTI (2004), Rust (1985), and Wolter (2007).

3.6.3.1.2 JRR Step 2

A new replicate weight is computed for each of the JRR replicates. Replicate weight values for cases in the deleted PSU are assigned a value of "0" or "missing." The replicate weight values for the retained PSU cases in the "deletion stratum" are formed by multiplying the full sample analysis weights by a factor of $a_h/(a_h - 1)$. This factor equals 2 in our example. Note that for all strata except the deletion stratum, the replicate weight value is the unaltered full sample weight.

The second JRR replicate is formed by dropping the second PSU in stratum 2 and weighting up the first PSU in Stratum 2, with the process continuing until all replicates have been formed.

3.6.3.1.3 JRR Step 3

Using the replicate weights developed for each JRR replicate sample, the weighted estimates of the population statistic are computed for each replicate. The statistic of interest in our example is the weighted estimate of the population mean, so each replicate estimate takes the following form:

$$\hat{q}_r = \bar{y}_r = \frac{\sum_{i \in rep}^{n_{rep}} y_i \cdot w_{i,rep}}{\sum_{i \in rep}^{n_{rep}} w_{i,rep}} \text{ is computed for each of } r = 1,...,4 \text{ replicates}$$

$$\hat{q}_1 = 0.4724;\ \hat{q}_2 = 0.4695;\ \hat{q}_3 = 0.4735;\ \hat{q}_4 = 0.4661$$

The full sample estimate of the mean is also computed:

$$\hat{q} = \bar{y}_w = \frac{\sum_{i=1}^{n} y_i \cdot w_i}{\sum_{i=1}^{n} w_i} = 0.4737$$

3.6.3.1.4 JRR Step 4

Using the $r = 1, \ldots, R$ replicate estimates and the full sample estimate, the sampling variance is estimated using the following simple formula:

$$\text{var}_{JRR}(\hat{q}) = \sum_r \left(\hat{q}_r - \hat{q}\right)^2$$

$$\text{e.g., } \text{var}_{JRR}(\bar{y}_w) = \sum_{r=1}^{4} (\bar{y}_r - \bar{y})^2 = 0.00005790$$

$$se_{JRR}(\bar{y}_w) = \sqrt{0.00007720} = 0.008786$$

3.6.3.1.5 JRR Step 5

A 100(1 – α)% confidence interval for the population parameter is then constructed:

$$CI(q) = \hat{q} \pm t_{1-\alpha/2,\, df} \sqrt{\text{var}_{JRR}\left(\hat{q}\right)}$$

$$\text{e.g., } CI(\bar{y}) = 0.4737 \pm t_{1-\alpha/2,4} \cdot 0.008786$$

$$= 0.4737 \pm 2.7764 \cdot (0.008786) = (0.4493, 0.4981)$$

Note that this variance estimation technique requires only the full sample weight and the R sets of replicate weights to perform the appropriate variance estimation. This is what enables survey organizations concerned about protecting respondent confidentiality to produce public-use complex sample survey data sets that do not include stratum or cluster codes (which might be used, in theory, to identify a respondent) and include only the required weight variables for estimation of population parameters and replicated variance estimation. In contrast, stratum and cluster codes would be required to estimate variances using Taylor series linearization.

3.6.3.2 Balanced Repeated Replication

The BRR method of variance estimation is a "half-sample" method that was developed specifically for estimating sampling variances under two PSU-per-stratum sample designs. The 2005–2006 NHANES, NCS-R, and 2006 HRS data sets used in the example analyses described in this text employ such a sampling error calculation model (see Section 4.3), with two PSUs (clusters) per stratum. The evolution of the BRR method began with the concept of forming replicates by choosing one-half of the sample. For a complex sample design with $h = 1, \ldots, H$ strata and exactly $a_h = 2$ PSUs per stratum, a **half-sample replicate** could be formed by choosing 1 of the 2 PSUs from each stratum (e.g., in Figure 3.4, choose PSU 1 in strata 1, 2, and 3 and PSU 2 in stratum 4). By default, this choice of one PSU per stratum would define a **half-sample complement** (e.g., PSU 2 in Strata 1, 2, and 3 and PSU 1 in Stratum 4). For H strata and 2 PSUs per stratum, there are 2^H possible different half-samples that could be formed—a total of 16 for the simple sample in Figure 3.4, and 4,398,000,000,000 for the $H = 42$ strata in the NCS-R design. Since a complex sample design with H strata and 2 PSUs per stratum provides only H degrees of freedom for variance estimation, the formation of $R > H$ half-sample replicates will not yield additional gains in efficiency for a replicated half-sample variance estimate.

BRR variance estimation proceeds using the following five steps.

3.6.3.2.1 BRR Step 1

So what is the optimal procedure for selecting which half-sample replicates to employ in the variance estimation? McCarthy (1969) introduced the concept of balanced repeated replication in which individual replicates are formed according to a pattern of "+" and "–" symbols that are found in the

BRR Replicate	Stratum (h)			
	1	2	3	4
1	+	+	+	–
2	+	–	–	–
3	–	–	+	–
4	–	+	–	–

FIGURE 3.6
Hadamard matrix used to define BRR replicates for a $H = 4$ strata design.

rows of a **Hadamard matrix.** The optimal efficiency of the BRR method for variance estimation based on half-samples is due to its "balancing"—that is, complete algebraic cancellation of unwanted between-stratum cross-product terms, such as $(y_{h1} - y_{h2}) \cdot (y_{g1} - y_{g2})$, that enter the half-sample variance computation formula. Readers interested in a more complete mathematical development of this idea are referred to Wolter (2007).

Fortunately, contemporary software for survey data analysis makes it easy to apply the BRR variance estimation method. Using only sampling error stratum and cluster codes (see Section 4.3.1) provided with the survey data set, the software will invoke the correct form of the Hadamard matrix to construct the sample replicates. Figure 3.6 illustrates the 4 × 4 Hadamard matrix than can be used to define BRR replicates for the four-strata sample design in Figure 3.4. Each row of the matrix defines one BRR replicate. For BRR replicate 1, the "+" sign in the columns for Strata 1, 2, and 3 indicate that the first PSU in the stratum is assigned to the replicate. The "–" sign in the Stratum 4 column indicates the second PSU is to be included in replicate 1. Likewise, BRR replicate 2 will include PSU 1 from stratum 1 and PSU 2 from strata 2, 3, and 4.

Figure 3.7 illustrates the form of the first BRR replicate for the Figure 3.4 data set.

Hadamard matrices are defined only for dimensions that are multiples of four. Whenever the number of primary strata defined for a complex sample design is a multiple of four, exactly H BRR replicates are defined according to the patterns of "+/–" indicators in the rows and columns of the $H \times H$ Hadamard matrix. The corresponding BRR variance estimates are said to be **fully balanced**. If the number of primary strata in the complex sample design is not a multiple of four, the Hadamard matrix of dimension equal to the next multiple of $4 > H$ is used. For example, a Hadamard matrix of dimension 44 × 44 is used to define the 44 half-sample replicates for the NCS-R, which has $H = 42$ strata for variance estimation. In such cases, the corresponding BRR variance estimates are said to be **partially balanced**.

Stratum	PSU (Cluster)	Case	y_i	$w_{i,rep}$
1	1	1	.58	1x2
1	1	2	.48	2x2
1	2	1	.	.
1	2	2	.	.
2	1	1	.39	1x2
2	1	2	.46	2x2
2	2	1	.	.
2	2	2	.	.
3	1	1	.39	1x2
3	1	2	.47	2x2
3	2	1	.	.
3	2	2	.	.
4	1	1	.	.
4	1	2	.	.
4	2	1	.47	2x2
4	2	2	.50	2x2

FIGURE 3.7
Illustration of BRR replicate 1 for the example data set.

3.6.3.2.2 BRR Step 2

A new replicate weight is then created for each of the h =1, …, H BRR half-sample replicates created in Step 1. Replicate weight values for cases in the complement half-sample PSUs are assigned a value of "0" or "missing." The replicate weight values for the cases in the PSUs retained in the half-sample are formed by multiplying the full sample analysis weights by a factor of 2.

3.6.3.2.3 BRR Step 3

Following the same procedure outlined for JRR and using the replicate weights developed for each BRR replicate sample, the weighted replicate estimates of the population statistic are computed. The full sample estimate of the population statistic is also computed:

$$\hat{q}_r = \bar{y}_r = \frac{\sum_{i \in rep}^{n_{rep}} y_i \cdot w_{i,rep}}{\sum_{i \in rep}^{n_{rep}} w_{i,rep}}$$

$$\hat{q}_1 = 0.4708;\ \hat{q}_2 = 0.4633;\ \hat{q}_3 = 0.4614;\ \hat{q}_4 = 0.4692$$

$$\hat{q} = \bar{y}_w = \frac{\sum_{i=1}^{n} y_i \cdot w_i}{\sum_{i=1}^{n} w_i} = 0.4737$$

3.6.3.2.4 BRR Step 4

The BRR estimate of sampling variance of the sample estimate is computed using one of several simple formulas. Here we illustrate the computation using one of the more common half-sample variance estimation formulae:

$$\text{var}_{BRR}(\bar{y}_w) = \text{var}_{BRR}(\hat{q}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{q}_r - \hat{q})^2$$

$$= \frac{1}{4}\sum_{r=1}^{4}(\bar{y}_r - \bar{y})^2$$

$$= 0.00007248$$

$$se_{BRR}(\bar{y}_w) = \sqrt{0.00007248} = 0.008515$$

Several software packages such as WesVar PC permit users to choose alternative half-sample variance estimation formulae including a method proposed by Fay (Judkins, 1990). Interested users are referred to Rust (1985) or Wolter (2007) for more information on alternative half-sample variance estimation formulae.

3.6.3.2.5 BRR Step 5

A 100(1 – α)% confidence interval for the population parameter is then constructed (recall that in the case of BRR, $df = H$):

$$CI(q) = \hat{q} \pm t_{1-\alpha/2,\, df}\sqrt{\text{var}_{BRR}\left(\hat{q}\right)}$$

$$\text{e.g., } CI(\bar{y}) = 0.4737 \pm t_{1-\alpha/2,4} \cdot 0.0085$$

$$= 0.4737 \pm 2.7764 \cdot (0.0085) = (0.4501, 0.4974)$$

3.6.3.3 The Bootstrap

A third less commonly applied approach to replicated variance estimation for complex sample surveys is the bootstrap method. Comparative simulation studies and empirical investigations have shown that in most survey applications involving reasonably large data sets, the bootstrap offers no advantage over the more robust TSL, JRR, or BRR methods (Kovar, Rao, and Wu, 1988; Rao and Wu, 1988). As Skinner, Holt, and Smith (1989) point out, the bootstrap method does permit direct evaluation of the sampling distribution of survey estimates and does not rely on large sample normality to derive confidence intervals for survey estimates. Consequently, the bootstrap method may have specific applications in analyzing complex samples in which small sample sizes or highly irregular underlying distributions may result in asymmetry in the sampling distribution of the population estimates.

Due to its limited application in current survey practice, we will not present a detailed introduction to the bootstrap method or illustrate its application in later exercises. We refer the interested reader to Rust and Rao (1996) for a more in-depth treatment of the bootstrap method for survey inference. At this writing, we are aware that significant new research and software development are under way in Stata and R for the application of the bootstrap method to variance estimation for complex sample survey data (Kolenikov, 2009). We will track these developments on the ASDA Web site and provide links to related publications and software user guides.

3.6.4 An Example Comparing the Results from the TSL, JRR, and BRR Methods

The previous sections have presented simple illustrations of the TSL, JRR, and BRR variance calculations based on the example data set in Figure 3.4. While these simple example calculations point out that estimated standard errors for the three calculation methods may in fact differ, the size of these differences is magnified by the small sample size of the example data set. In this section, we present a more realistic comparison based on the NCS-R survey data set.

A number of empirical and simulation-based studies have compared the properties of the TSL, JRR, and BRR methods of variance estimation for complex sample designs, including Kish and Frankel (1974), Kovar, Rao, and Wu (1988), Valliant (1990), and Rao and Wu (1985). Across a range of estimators and sample designs, these studies have shown few important differences in results obtained using the three methods. The methods are unbiased and produce identical results in the special case where the estimator of interest is a linear statistic such as a weighted sample total. For nonlinear estimators commonly employed in survey analysis (e.g., regression coefficients), the TSL and JRR methods tend to have slightly lower bias (smaller mean square error [MSE]) for the estimates of sampling variance. However, confidence intervals constructed using BRR or Bootstrap estimates of standard errors

TABLE 3.2

Comparison of the TSL, JRR, and BRR Variance Estimation Methods for NCS-R Estimates of Descriptive Population Parameters

Statistic/Method		Estimate	Standard Error
Major depression (Prevalence)	TSL	19.1711%	0.4877%
	JRR	19.1711%	0.4878%
	BRR	19.1711%	0.4896%
Alcohol dependence (Prevalence)	TSL	5.4065%	0.3248%
	JRR	5.4065%	0.3248%
	BRR	5.4065%	0.3251%
Household income (Mean)	TSL	$59,277.06	$1596.34
	JRR	$59,277.06	$1596.65
	BRR	$59,277.06	$1595.37

provide better nominal coverage (e.g., 95 in 100 for a 95% CI) of the true population value (Wolter, 2007). For nonsmooth functions of the sample data (e.g., sample quantiles), linearization cannot be applied directly (due to the need for smooth, continuous functions to compute derivatives), and the JRR method is known to result in badly biased estimates of the variance. As a result, BRR is often used in these situations to estimate variances (see Chapter 5).

Table 3.2 presents the results of a simple analysis of the NCS-R data that illustrates the general equivalency of the TSL, JRR, and BRR methods for sampling variance estimation. The analysis focuses on three descriptive estimates of characteristics of the NCS-R survey population: (1) % of the population experiencing symptoms of major depression during their lifetime; (2) % of the population experiencing alcohol dependence during their lifetime; and (3) average household income (in U.S. dollars).

Since each method employs the same overall weighted estimate of the population statistic, the point estimate of the population statistic is identical under the TSL, JRR, and BRR methods. To four significant digits, estimated standard errors are virtually identical under the TSL and JRR methods. BRR estimates of standard errors differ very slightly from the TSL or JRR counterparts, but the difference is so small that it would be negligible in any practical survey analysis setting.

3.7 Hypothesis Testing in Survey Data Analysis

Inference in the analysis of complex sample survey data is not limited to the construction of confidence intervals for population parameters. The complex sample survey data may also be used to derive familiar test statistics (t, χ^2, F)

for formal hypothesis tests of the form H_0: $\theta = \theta_0$ vs. H_A: $\theta \neq \theta_0$. In fact, as the discussion in Section 3.5.1 pointed out, the Student t test statistic is the **pivotal statistic** from which the expression for the $100(1 - \alpha)\%$ CI is derived. Table 3.3 compares the test statistics and hypothesis testing approaches for simple random sample (or independently and identically distributed) data to the modified approaches that have been developed for complex sample survey data.

Simple Student t tests based on single estimates of means, proportions, regression coefficients, and differences or linear combinations of these statistics incorporate the complex design by using correct standard errors in the computation of the test statistic. To test the null hypothesis, the test statistic

TABLE 3.3

Comparisons of Hypothesis Testing Procedures for Simple Random Sample Data versus Complex Sample Survey Data

Simple Random Sample (i.i.d.) Data	Complex Sample Survey Data
Student t-tests of hypotheses ($H_0 \mid H_A$) for means, proportions, single model parameters, such as $\bar{y} = \bar{Y}_0;\ \beta_j = 0$	Design-adjusted Student t-test • Correct standard error in denominator • Design-adjusted degrees of freedom
Student t-tests of simple hypotheses concerning differences or linear combinations of means, such as $(\bar{y}_1 - \bar{y}_2) = 0;\ \sum_j a_j p_j = 0$	Design-adjusted Student t-test • Correct standard error in denominator reflecting separate estimates of variance and covariance of component estimates • Design-adjusted degrees of freedom
X^2 test of independence (association) in bivariate and multiway tables: • Pearson X^2, likelihood ratio G^2	Design-adjusted χ^2 and F-tests • Rao–Scott first- and second-order corrections adjust for design effects in $\hat{\Sigma}(p)$ • X^2 transformed to F-test statistic
Full and partial F-tests of hypotheses for linear regression model goodness of fit and full versus reduced model, such as $H_0: \beta = \{\beta_1, \beta_2, ..., \beta_p\} = 0$ $H_0: \beta_{(q)} = \{\beta_{p-q}, ..., \beta_p\} = 0$	Design-adjusted Wald χ^2 or F-test • Correct $\hat{\Sigma}(\hat{\beta})$ under complex design • Adjusted degrees of freedom
F-tests based on expected mean squares for analysis of variance (ANOVA)-type linear models	Linear regression parameterization of the ANOVA model. Design-adjusted Wald χ^2 or F-tests as in the preceding linear regression cases.
Likelihood ratio X^2 tests for maximum likelihood estimates of parameters in generalized linear models, such as $H_{0,MLE}: \beta = \{\beta_1, \beta_2, ..., \beta_p\} = 0$ $H_{0,MLE}: \beta_{(q)} = \{\beta_{p-q}, ..., \beta_p\} = 0$	Design-adjusted Wald χ^2 or F-test • Correct $\hat{\Sigma}(\hat{\beta})$ under complex design • Adjusted degrees of freedom

is referred to a Student t distribution with degrees of freedom adjusted for the complex sample design. Chapter 5 will describe design-adjusted Student t tests for means, proportions, differences of means and proportions, subpopulation estimates, and general linear combinations of estimates. Design-adjusted t tests for single parameters in linear regression or generalized linear models are addressed in Chapters 7, 8, and 9.

For tests of association between two cross-classified categorical variables, the standard Pearson χ^2, likelihood ratio G^2, and Wald χ^2 are replaced by design-adjusted forms of these traditional test statistics. Like their SRS counterparts, under the null hypothesis, the Rao–Scott "adjusted" χ^2 and G^2 statistics and the adjusted Wald χ^2 (computed with a design-based variance-covariance matrix) are expected to follow a central chi-square distribution. Software packages offer the user the option of using the standard χ^2 version of each test statistic or a modified version that under H_0 is referred to an F distribution. More details on these design-adjusted tests of association for cross-classified data with accompanying examples are provided in Chapter 6.

F-tests are used in standard normal linear regression to test the overall fit of the model as well as to perform joint tests concerning the significance of subsets of the model parameters: for example, $H_0 : \boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_p\} = \mathbf{0}; H_0 : \boldsymbol{\beta}_{(q)} = \{\beta_{p-q}, ..., \beta_p\} = \mathbf{0}$. In the analysis of complex sample survey data, the conventional F-tests are replaced with a Wald χ^2 test statistic that is provided both as a chi-square test or as a transformed F-test statistic. Chapter 7 addresses the use of the Wald test statistic for joint hypothesis tests involving multiple parameters in estimated linear regression models.

Standard approaches to analyses involving generalized linear models (e.g., logistic, probit, and Poisson regression models) typically employ maximum likelihood methods to estimate parameters and standard errors. Tests of hypotheses concerning the significance of a nested subset of the full vector of model parameters, that is, $H_0 : \boldsymbol{\beta}_{(q)} = \{\beta_{p-q}, ..., \beta_p\} = \mathbf{0}$, are therefore performed using the standard **likelihood ratio test (LRT)**. When generalized linear models are fit to complex sample survey data, multiparameter tests of this type are conducted using a design-adjusted Wald χ^2 or F-statistic. These design-corrected Wald tests for GLM analyses are described in more detail in Chapters 8 and 9.

3.8 Total Survey Error and Its Impact on Survey Estimation and Inference

Probability sampling provides a theoretical basis for unbiased estimation of population parameters and the associated sampling variance of the sample estimates. However, a successful scientific survey depends not only on

TABLE 3.4

A Taxonomy of Survey Errors

Variable Errors	Biases
Sampling variance	Sample selection bias
Interviewer variance	Frame coverage bias
Response (measurement) variance	Measurement bias
Coding variance	Nonresponse bias

control of sampling variances but also other sources of error. The collection of all sources of error that can influence the precision and accuracy of survey estimates of population parameters is termed **total survey error** (TSE; Groves, 2004; Lessler and Kalsbeek, 1992). The TSE for a survey estimate is measured as the mean square error, $Mean\ Square\ Error = Variance + Bias^2$, or the variance of the estimate plus the square of the bias in the sample estimate. Table 3.4 provides a typical taxonomy of survey errors.

3.8.1 Variable Errors

The sources of error that cause sample estimates to disperse randomly about the true and unknown population value of interest across replications of the survey process are termed **variable errors**.

Sampling variances or standard errors that have been described at length in the preceding sections derive from the statistical fact that only a subset of the full target population is observed in any given sample. As sample sizes increase, the sampling variance decreases and disappears entirely if a complete census of the target population is conducted. Probability sampling theory provides well-defined guidance for estimating sampling variances for survey estimates.

Interviewer variance and **response variance** (Fuller, 1987; Groves, 2004) enter the data during the actual interview and may be attributed to random inaccuracies in the way that interviewers ask survey questions or record the survey answers or the way that the respondents report the responses. In scientific surveys, interviewer and response variance is minimized by carefully training interviewers and designing and pretesting questions so that they are clearly worded and have comprehensive, easy-to-interpret response categories. Readers interested in an in-depth treatment of interviewer variance and other measurement errors are referred to Biemer et al. (1991).

Coding variance is primarily a technical source of random error in the data set. Research staff responsible for coding the survey data and transcribing the information to a computer file for analysis may make errors in a random manner. As the technology for computer-assisted interviewing and data collection has advanced, this source of error has been reduced.

Each of these sources of variance—sampling, interviewer, response, and coding—contributes to the combined total variability in observed survey

data. Short of conducting elaborate experiments, it will not be possible for the data producer, let alone the survey analyst, to decompose the total variance of a survey estimate into components attributable to each of these sources. Nevertheless, it is important to note that the combined influences of these multiple sources of variability will all be captured in the estimated standard errors that are computed from the survey data.

3.8.2 Biases in Survey Data

The English word *bias* has many meanings, but in a statistical context it refers to the tendency of a statistical estimate to deviate systematically from the true value it is intended to estimate across replications of the survey process. Students of statistics, including survey researchers, have long been taught that "unbiasedness"—the absence of bias—is one of the most desirable properties of a statistical estimator or procedure. In survey practice, bias should certainly be avoided; however, elimination of all sources of bias in survey data is probably neither practical nor even efficient when both the costs and benefits of reducing bias are considered. In fact, survey analysts tend to be less interested in purely unbiased estimators and more interested in estimators that have the property of being a *consistent* estimator of the population parameter of interest. **Consistent estimators** converge to the true population value as the sample size n increases to the population size N. Survey biases listed in the survey error taxonomy given in Table 3.4 may be collapsed into two major types: **sampling bias** and **nonsampling bias** (Kish, 1965).

In probability samples, the greatest potential for sampling bias can be attributed to noncoverage of survey population elements. Sample frame noncoverage occurs when population elements are systematically excluded from the population registry or the data sources (maps and census counts) used to develop area probability frames and therefore have no probability of being included in the sample. Sample noncoverage bias may also occur in the process of screening selected dwelling units to identify and select eligible survey respondents. Survey producers minimize sampling bias through careful design and attention to detail in the sample selection process, field testing of screening and respondent selection procedures prior to the data collection period, and rigorous training and on-site supervision of the field staff for the actual survey data collection. After the survey is complete, poststratification weighting may be employed to attenuate any remaining sample bias in the survey data.

Survey data is also vulnerable to nonsampling bias from two primary sources: (1) measurement bias, or systematic bias in the way respondents interpret and respond to questions; and (2) survey nonresponse. Measurement bias may be deliberate on the part of the respondent, or it may be the unconscious result of poor questionnaire construction or interviewer training. Survey respondents who are asked to report their household income may underreport or fail to mention sources of income such as the sale of a parcel of land. Survey questions that ask about participation in elections,

educational activities, or religious observances may be subject to overreporting of participation, a phenomenon termed *social desirability* bias. Poorly worded or "leading" questions or questionnaires that place questions out of context may yield biased measures for the constructs in which the research investigator is truly interested. Across disciplinary areas, survey methodologists work very hard to understand the survey measurement process and to design questions and questionnaire formats that accurately measure the underlying constructs of interest.

Survey nonresponse is another potential source of bias in sample-based estimates of population characteristics. The failure to obtain any data on a sample household or individual is termed *unit nonresponse.* A missing response to one or more individual variable items in an otherwise complete interview questionnaire is termed *item nonresponse* (Little and Rubin, 2002). Nonresponse to voluntary surveys including those conducted by universities and other scientific research organizations has become a major problem in countries in Western Europe and North America. Nonresponse bias in survey estimates of simple statistics such as population means or proportions is a function of the response rate and the difference in the values of the statistic for responding and nonresponding members of the sample. For estimates of a population proportion, for example, the nonresponse bias can be expressed using the following formula:

$$Bias_{NR}(p) = (1 - RRate) \times (P_R - P_{NR}) \tag{3.13}$$

where $RRate$ is the expected value of the population response rate, p_R is the value of the proportion for respondents and p_{NR} is the value of the proportion for nonrespondents. The absolute value of the expected nonresponse bias increases with the product of the nonresponse rate and the difference in p for respondents and nonrespondents. If the response rate is high or proportions for respondents and nonrespondents do not differ, the nonresponse bias will be very small. Unlike sampling variance, which decreases as sample size increases, nonresponse bias is persistent. Its expected value is not a function of sample size but remains unchanged regardless of how large or small the size of the survey sample.

The potential impact of nonresponse bias on the analysis of survey data may be best illustrated through a simple example. Assume a researcher is interested in studying parents' views on the need for increased government spending (hence potential increases in taxes) for elementary science education. Among parents who agree to participate in the survey, the expected proportion that support increased spending on elementary science education is $P_R = 0.6$, while for noncooperating parents $P_{NR} = 0.4$—a major difference between respondents and nonrespondents. If only 50% of the original random sample of n = 1,000 parents agreed to participate, the expected nonresponse bias for the proportion of interest would be

$$Bias_{NR}\left(\hat{P}\right) = \left(1 - RR\right) \times \left(P_R - P_{NR}\right) = (1 - .5) \times (.6 - .4) = 0.10$$

Assuming for simplicity that the original sample of parents was selected using SRS, the researcher would develop the following 95% confidence interval for this estimate:

$$p_R \pm 1.96 \cdot \left[p \cdot (1-p) / \left(n\right)\right]^{1/2} \cong p_R \pm 1.96 \cdot \left[0.6(1-0.6) / \left(0.5 \cdot 1000\right)\right]^{1/2} = p_R \pm 0.043$$

In this case, the size of the expected nonresponse bias is relatively large in comparison with the size of the 95% confidence interval half-width—a result of the low response rate and the major difference in expected proportions for respondent and nonrespondent cases.

The purpose of this example is not to magnify the potential seriousness of nonresponse bias in survey estimation—high response rates or smaller differences in the expected statistics for respondents and nonrespondents would decrease the size of the expected bias. Some recent research has suggested that we may be overly concerned about the seriousness of nonresponse bias for certain types of survey measures, and especially measures of respondent attitudes and expectations (Keeter et al., 2000). However, the example makes the point that we cannot be complacent and ignore the potential for nonresponse bias in the survey estimation process.

As described in Section 2.7, detailed data on respondents and nonrespondents and a model for the nonresponse mechanism are used to develop nonresponse adjustments to the survey analysis weights in actual survey practice. To the extent that the estimated model accurately describes the underlying nonresponse process, these adjustments to the analysis weight serve to attenuate potential nonresponse bias and its impact on estimates and the corresponding inference that the analyst draws from the survey data.

4

Preparation for Complex Sample Survey Data Analysis

4.1 Introduction

This chapter guides the survey analyst through a sequence of steps that are important to properly prepare for the analysis of a complex sample survey data set. The steps in this chapter should be considered only *after* steps 1 and 2 in Chapter 1 (defining the research problem, stating the research objectives, and understanding the sample design). These steps include reviewing the weights and the sampling error calculation model for the survey data set, examining the amount of missing data for key variables, and preparing for the analysis of sample subclasses. The chapter concludes with a short checklist to remind the reader of these critical preliminary steps.

We remind readers about two important terms. The term **data producer** applies to the statistical agency or the research team that is responsible (with the assistance of survey statisticians) for designing a probability sample from a larger population, collecting the survey data, developing sampling weights, and producing information about the complex sample design that can be used by research analysts for statistical analysis of the data. For example, the National Center for Health Statistics (NCHS) along with its contractor, Westat, Inc., was responsible for producing the 2005–2006 National Health and Nutrition Examination Survey (NHANES) data analyzed throughout this book. The term **data user** is applied to the survey analyst or researcher who will use the survey data to develop and test scientific hypotheses concerning the survey population.

In most cases, the data producer has access to far more detailed information about the complex sample design and personal information concerning study respondents than the data user. To maintain confidentiality for survey respondents and to minimize **disclosure risk**, the data producer is responsible for making optimal use of detailed and often confidential information on the individual sample cases to develop sampling error codes and analysis weights for the survey data set. In addition, the data producer is responsible

for providing the data user with a cleaned data set, a codebook for the survey variables, and technical documentation describing the sample and survey procedures. Many survey programs are now providing data users with a guide for statistical analysis of their survey data set. For example, NCHS provides an online document* with detailed instructions for how to correctly analyze the NHANES data sets.

4.2 Analysis Weights: Review by the Data User

Experience as survey statisticians and consultants has taught us that many survey data analysts struggle with the correct use of sampling weights in survey estimation and inference. For whatever reason, many otherwise sophisticated data users wish to place a "black box" around the process of weight development and the application of weights in their analysis. As described in Section 2.7, the final analysis weights provided in survey data sets are generally the product of a sample selection weight, w_{sel}, a nonresponse adjustment factor, w_{nr}, and a poststratification factor, w_{ps}: $w_{final,i} = w_{sel,i} \cdot w_{nr,i} \cdot w_{ps,i}$. For the reasons just outlined, the data producer is responsible for developing individual weights for each sample case and linking the final analysis weight variable to each observational unit in the survey data file.

The analysis weight assigned to each respondent case is a measure of the number of population members represented by that sample case or, alternatively, the relative share of the population that the case represents. When weights are applied in the statistical analysis of survey data, weighted calculations simply expand each sample case's contribution to reflect its representative share of the target population. Because the process of weight development has been discussed extensively in Section 2.7, the aim of this section is to remove any remaining mystique surrounding the analytic weights that are provided with a survey data set.

Although data analysts will not typically be responsible for the actual weight calculations, they should familiarize themselves with the analysis weight variables and how weighted analysis may influence estimation and inference from their survey data. Key steps in this process of verification and familiarization include the following:

- Verifying the variable name for the appropriate weight for the intended analysis.

* http://www.cdc.gov/nchs/data

- Checking and reviewing the scaling and general distribution of the weights.
- Evaluating the impact of the weights on key survey statistics.

The subsequent sections describe these three activities in more detail.

4.2.1 Identification of the Correct Weight Variables for the Analysis

The data user will need to refer to the survey documentation (technical report, codebook) to identify the correct variable name for the analysis weight. Unfortunately, there are no standard naming conventions for weight variables, and we recommend great caution in this step as a result. A number of years ago, a student mistakenly chose a variable labeled WEIGHT and produced a wonderful paper based on the NHANES data in which each respondent's data was weighted by his or her body weight in kilograms. The correct analysis weight variable in the student's data file was stored under a different, less obvious variable label.

Depending on the variables to be analyzed, there may be more than one weight variable provided with the survey data set. The 2006 Health and Retirement Study (HRS) data set includes one weight variable (KWGTHH) for the analysis of financial unit (single adult or couple) variables (e.g., home value or total net worth) and a separate weight variable (KWGTR) for individual-level analysis of variables (e.g., health status or earnings from a job). The 2005–2006 NHANES documentation instructs analysts to use the weight variable WTINT2YR for analyses of the medical history interview variables and another weight variable (WTMEC2YR) for analyses of data collected from the medical examination phase of the study. The larger sample of the National Comorbidity Survey Replication (NCS-R) Part I mental health screening data (n = 9,282) is to be analyzed using one weight variable (NCSRWTSH), while another weight variable (NCSRWTLG) is the correct weight for analyses involving variables measured for only the in-depth Part II subsample (n = 5,692).

Some public-use data sets may contain a large set of weight variables known as **replicate weights**. For example, the 1999–2000 NHANES public-use data file includes the replicate weight variables WTMREP01, WTMREP02, …, WTMREP52. As mentioned in Chapter 3, replicate weights are used in combination with software that employs a replicated method of variance estimation, such as balanced repeated replication (BRR) or jackknife repeated replication (JRR). When a public-use data set includes replicate weights, design variables for variance estimation (stratum and cluster codes) will generally *not* be included (see Section 4.3), and the survey analyst needs to use the program syntax to specify the replicated variance estimation approach (e.g., BRR, JRR, BRR-Fay) and identify the sequence of variables that contain the replicate weight values (see Appendix A for more details on these software options).

4.2.2 Determining the Distribution and Scaling of the Weight Variables

In everyday practice, it is always surprising to learn that an analyst who is struggling with the weighted analysis of survey data has never actually looked at the distribution of the weight variable. This is a critical step in preparing for analysis. Assessing a simple univariate distribution of the analysis weight variable provides information on (1) the scaling of the weights; (2) the variability and skew in the distribution of weights across sample cases; (3) extreme weight values; and (4) (possibly) missing data on the analysis weight. Scaling of the weights is important for interpreting estimates of totals and in older versions of software may affect variance estimates. The variance and distribution of the weights may influence the precision loss for sample estimates (see Section 2.7). Extreme weights, especially when combined with outlier values for variables of interest, may produce instability in estimates and standard errors for complete sample or subclass estimates. Missing data or zero (0) values on weight variables may indicate an error in building the data set or a special feature of the data set. For example, 2005–2006 NHANES cases that completed the medical history interview but did not participate in the mobile examination center (MEC) phase of the study will have a positive, nonzero weight value for WTINT2YR but will have a zero value for WTMEC2YR (see Table 4.1).

TABLE 4.1

Descriptive Statistics for the Sampling Weights in the Data Sets Analyzed in This Book

	NCS-R: NCSRWTLG	NCS-R: NCSRWTSH	NHANES: WTMEC2YR[c]	NHANES: WTINT2YR[c]	HRS: KWGTR	HRS: KWGTHH
n	5,692	9,282	5,563	5,563	18,467	18,467
Sum	5,692	9,282	217,700,496	217,761,911	75,540,674	82,249,285
Mean	1.00[a]	1.00[a]	39,133.65	39,144.69	4,144.73	4,453.85
SD	0.96	0.52	31,965.69	30,461.53	2,973.48	3,002.06
Min	0.11	0.17	0[b]	1,339.05	0[b]	0[b]
Max	10.10	7.14	156,152.20	152,162.40	16,532	15,691
Pctls.						
1%	0.24	0.36	0	2,922.37	0	0
5%	0.32	0.49	2,939.33	4,981.73	0	1,029
25%	0.46	0.69	14,461.86	16,485.70	2,085	2,287
50%	0.64	0.87	27,825.71	28,040.22	3,575	3,755
75%	1.08	1.16	63,171.48	62,731.71	5,075	5,419
95%	2.95	1.85	100,391.70	96,707.20	10,226	10,847
99%	4.71	3.17	116,640.90	113,196.20	12,951	14,126

[a] Suggests that the sampling weights have been normalized to sum to the sample size.

[b] Cases with weights of zero will be dropped from analyses and usually correspond to individuals who were not eligible to be in a particular sample.

[c] 2005–2006, NHANES adults.

Table 4.1 provides simple distributional summaries of the analysis weight variables for the NCS-R, 2006 HRS, and 2005–2006 NHANES data sets. Inspection of these weight distributions quickly identifies that the scale of the weight values is quite different from one study to the next. For example, the sum of the NCS-R Part I weights is

$$\sum_i \text{NCSRWTSH}_i = 9,282$$

while the sum of the 2006 HRS individual weights is

$$\sum_i \text{KWGTR}_i = 75,540,674$$

With the exception of weighted estimates of population totals, weighted estimation of population parameters and standard errors should be invariant to a linear scaling of the weight values, that is, $w_{scale,i} = k \cdot w_{final,i}$, where k is an arbitrary constant. That is, the data producer may choose to multiply or divide the weight values by any constant and with the exception of estimates of population totals, weighted estimates of population parameters and their standard errors should not change.

For many surveys such as the 2005–2006 NHANES and the 2006 HRS, the individual case weights will be population scale weights, and the expected value for the sum of the weights will be the population size:

$$E\left(\sum_{i=1}^{n} w_i\right) = N$$

For other survey data sets, a normalized version of the overall sampling weight is provided with the survey data. To "normalize" the final overall sampling weights, data producers divide the final population scale weight for each sample respondent by the mean final weight for the entire sample:

$$w_{norm,i} = w_i \Big/ \left(\sum_i w_i / n\right) = w_i / \bar{w}$$

Many public-use data sets such as the NCS-R will have normalized weights available as the final overall sampling weights. The resulting normalized weights will have a mean value of $\bar{w}_{norm} = 1.0$, and the normalized weights for all sample cases should add up to the sample size:

$$\sum_i w_{norm,i} = n$$

Normalizing analysis weight values is a practice that has its roots in the past when computer programs for the analysis of survey data often misinterpreted the "sample size" for weighted estimates of variances and covariances required in computations of standard errors, confidence intervals, or test statistics. As illustrated in Section 3.5.2, the degrees of freedom for variance estimation in analyses of complex sample survey data are determined by the sampling features (stratification, clustering) and not the nominal sample size. Also, some data analysts feel more comfortable with weighted frequency counts that closely approximate the nominal sample sizes for the survey. However, there is a false security in assuming that a weighted frequency count of

$$\sum w_i = 1{,}000$$

corresponds to an effective sample size of $n_{eff} = 1{,}000$. As discussed in Section 2.7, the effective sample size for 1,000 nominal cases will be determined in part by the weighting loss, L_w, that arises due to variability in the weights and the correlation of the weights with the values of the survey variables of interest. Fortunately, normalizing weights is *not necessary* when analysts use computer software capable of incorporating any available complex design information for a sample into analyses of the survey data.

4.2.3 Weighting Applications: Sensitivity of Survey Estimates to the Weights

A third step that we recommend survey analysts consider the first time that they work with a new survey data set is to conduct a simple investigation of how the application of the analysis weights affects the estimates and standard errors for several key parameters of interest.

To illustrate this step, we consider data from the NCS-R data set, where the documentation indicates that the overall sampling weight to be used for the *subsample* of respondents responding to both Part I and Part II of the NCS-R survey (n = 5,692) is NCSRWTLG. A univariate analysis of these sampling weights in Stata reveals a mean of 1.00, a standard deviation of 0.96, a minimum of 0.11, and a maximum of 10.10 (see Table 4.1). These values indicate that the weights have been normalized and have moderate variance. In addition, we note that some sampling weight values are below 0.50. Many standard statistical software procedures will round noninteger weights and set the weight to 0 if the normalized weight is less than 0.5—excluding such cases from certain analyses. This is an important consideration that underscores

the need to use specialized software that incorporates the overall sampling weights correctly (and does not round them).

We first consider unweighted estimation of the proportions of NCS-R Part II respondents with lifetime diagnoses of either major depressive episode (MDE), measured by a binary indicator equal to 1 or 0, or alcohol dependence (ALD; also a binary indicator), in Stata:

```
mean mde ald if ncsrwtlg != .
```

Variable	Mean
MDE	0.3155
ALD	0.0778

Note that we explicitly limit the unweighted analysis to Part II respondents (who have a nonmissing value on the Part II sampling weight variable NCSRWTLG). The unweighted estimate of the MDE proportion is 0.316, suggesting that almost 32% of the NCS-R population has had a lifetime diagnosis of MDE. The unweighted estimate of the ALD proportion is 0.078, suggesting that almost 8% of the NCS-R population has a lifetime diagnosis of alcohol dependence.

We then request weighted estimates of these proportions in Stata, first identifying the analysis weight to Stata with the `svyset` command and then requesting weighted estimates by using the `svy: mean` command:

```
svyset [pweight = ncsrwtlg]
svy: mean mde ald
```

Variable	Mean
MDE	0.1918
ALD	0.0541

The weighted estimates of population prevalence of MDE and ALD are 0.192 and 0.054, respectively. The unweighted estimates for MDE and ALD therefore have a positive bias (there would be a big difference in reporting a population estimate of 32% for lifetime MDE versus an estimate of 19%).

In this simple example, the weighted estimates differ significantly from the unweighted means of the sample observations. This is not always the case. Depending on the sample design and the nonresponse factors that contributed to the computation of individual weight values, weighted and unweighted estimates may or may not show significant differences. When this simple comparison of weighted and unweighted estimates of key population parameters shows a significant difference, the survey analyst should aim to understand why this difference occurs. Specifically, what are

the factors contributing to the case-specific weights that would cause the weighted population estimates to differ from an unweighted analysis of the nominal set of sample observations?

Consider the NCS-R example. We know from the Chapter 1 description that, according to the survey protocol, all Part I respondents reporting symptoms of a mental health disorder and a random subsample of symptom-free Part I respondents continued on to complete the Part II in-depth interview. Therefore, the unweighted Part II sample contains an "enriched" sample of persons who qualify for one or more mental health diagnoses. As a consequence, when the corrective population weight is applied to the Part II data, the unbiased weighted estimate of the true population value is substantially lower than the simple unweighted estimate. Likewise, a similar comparison of estimates of the prevalence of physical function limitations using 2005–2006 NHANES data would yield weighted population estimates that are lower than the simple unweighted prevalence estimates for the observed cases. The explanation for that difference lies in the fact that persons who self-report a disability are oversampled for inclusion in the NHANES, and the application of the weights adjusts for this initial oversampling.

Repeating this exercise for a number of key variables should provide the user with confidence that he or she understands both how and why the application of the survey weights will influence estimation and inference for the population parameters to be estimated from the sample survey data. We note that these examples were designed only to illustrate the calculation of weighted sample estimates; standard errors for the weighted estimates were not appropriately estimated to incorporate complex design features of the NCS-R sample. Chapter 5 considers estimation of descriptive statistics and their standard errors in more detail.

4.3 Understanding and Checking the Sampling Error Calculation Model

The next step in preparing to work with a complex sample survey data set is to identify, understand, and verify the **sampling error calculation model** that the data producer has developed and encoded for the survey data set. Information about the sampling error calculation model can often be found in sections of the technical documentation for survey data sets titled *sampling error calculations* or *variance estimation* (to name a couple of possibilities). In this section, we discuss the notion of a sampling error calculation model and how to identify the appropriate variables in a survey data set that represent the model for variance estimation purposes.

A sampling error calculation model is an approximation or "model" for the actual complex sample design that permits practical estimation of

sampling variances for survey statistics. Such models are necessary because in many cases, the most practical, cost-efficient sample designs for survey data collection pose analytical problems for complex sample variance estimation. Examples of sample design features that complicate direct analytic approaches to variance estimation include the following:

- Multistage sampling of survey respondents.
- Sampling units without replacement (WOR) at each stage of sample selection.
- Sampling of a single primary sampling unit (PSU) from nonself-representing (NSR) primary stage strata.
- Small primary stage clusters that are not optimal for subclass analyses or pose an unacceptable disclosure risk.

The data producer has the responsibility of creating a sampling error calculation model that retains as much essential information about the original complex design as possible while eliminating the analytical problems that the original design might pose for variance estimation.

4.3.1 Stratum and Cluster Codes in Complex Sample Survey Data Sets

The specification of the sampling error calculation model for a complex sample design entails the creation of a **sampling error stratum** and a **sampling error cluster** variable.

These **sampling error codes** identify the strata and clusters that the survey respondents belong to, approximating the original sample design as closely as possible while at the same time conforming to the analytical requirements of several methods for estimating variances from complex sample data. Because these codes approximate the stratum and cluster codes that would be assigned to survey respondents based on the original complex sample design, they will not be equal to the original design codes, and the approximate codes are often scrambled (or masked) to prevent the possibility of identifying the original design codes in any way. Sampling error stratum and cluster codes are *essential* for data users who elect to use statistical software programs that employ the Taylor series linearization (TSL) method for variance estimation. Software for replicated variance estimation (BRR or JRR) does not require these codes, provided that the data producer has generated replicate weights. If replicate weights are not available, software enabling replicated variance estimation can use the sampling error stratum and cluster codes to create sample replicates and to develop the corresponding replicate weights.

In some complex sample survey data sets where confidentiality is of utmost importance, variables containing sampling error stratum and cluster codes may not be provided directly to the research analyst (even if they

represent approximations of the original design codes). In their place, the data producer's survey statisticians calculate replicate weights that reflect the sampling error calculation model and can be used for repeated replication methods of variance estimation. We provide examples of analyses using replicate weights on the book Web site, and interested readers can refer to Appendix A to see how replicate weights would be identified in various software procedures.

Unfortunately, as is true for analysis weights, data producers have not adopted standard conventions for naming the sampling error stratum and cluster variables in public-use data sets. The variable containing the stratum code often has a label that includes some reference to stratification. Two examples from this book include the variable SDMVSTRA from the 2005–2006 NHANES data set and the variable SESTRAT from the NCS-R data set.

The variable containing the sampling error cluster codes is sometimes referred to as a **sampling error computation unit (SECU)** variable, or it may be generically called the "cluster" code, the PSU code, or pseudo-primary sampling unit (PPSU) variable. Two examples that will be used throughout the example exercises in this book include the SDMVPSU variable from the 2005–2006 NHANES data set and the SECLUST variable from the NCS-R data set.

As discussed in Section 2.5.2, some large survey data sets are still released without sampling error codes or replicate weight values that are needed for complex sample variance estimation. In such cases, analysts may be able to submit an application to the data producer for restricted access to the sampling error codes, or they may elect to use the generalized design effect that is advocated by these data producers. For a limited number of survey data sets, online, interactive analysis systems such as the SDA system* provide survey analysts with the capability of performing analysis without gaining direct access to the underlying sampling error codes for the survey data set. Analysts who are working with older public-use data sets may find that these data sets were released with the sampling error stratum and cluster codes concatenated into a single code (e.g., 3001 for stratum = 30 and cluster = 01). In these cases, the variable containing the unified code will need to be divided into two variables (stratum and cluster) by the survey data analyst for variance estimation purposes.

4.3.2 Building the NCS-R Sampling Error Calculation Model

We now consider the sample design for the National Comorbidity Survey Replication as an illustration of the primary concepts and procedures for

* The SDA system is available on the Web site of the University of Michigan Inter-University Consortium for Political and Social Research (ICPSR) and is produced by the Computer-Assisted Survey Methods Program at the University of California–Berkeley. Visit http://www.icpsr.umich.edu for more details.

TABLE 4.2

Original Sample Design and Associated Sampling Error Calculation Model for the NCS-R

Original Sample Design	Sampling Error Calculation Model
The sample is selected in multiple stages. Primary stage units (PSUs), secondary stage units (SSUs), and third stage units are selected without replacement (WOR).	The concept of **ultimate clusters** is employed (see Chapter 3). Under the assumption that PSUs (ultimate clusters) are sampled with replacement, only PSU-level statistics (totals, means) are needed to compute estimates of sampling variance. The ultimate clusters are assumed to be **sampled with replacement (SWR)** at the primary stage. Finite population corrections are ignored, and simpler SWR variance formulas may be used for variance estimation.
Sixteen of the primary stage strata are self-representing (SR) and contain a single PSU. True sampling begins with the selection of SSUs within the SR PSU.	**Random groups** of PSUs are formed for sampling error calculation. Each SR PSU becomes a sampling error stratum. Within the SR stratum, SSUs are randomly assigned to a pair of sampling error clusters.
A total of 46 of the primary stage strata are nonself-representing (NSR). A single PSU is selected from each NSR stratum.	**Collapsed strata** are formed for sampling error calculation. Two similar NSR design strata (e.g., Strata A and B) are collapsed to form one sampling error computation stratum. The Stratum A PSU is the first sampling error cluster in the stratum, and the Stratum B PSU forms the second sampling error cluster.

constructing a sampling error calculation model for a complex sample survey data set. Table 4.2 presents a side-by-side comparison of the features of the original NCS-R complex sample design and the procedures employed to create the corresponding sampling error calculation model. Interested readers can refer to Kessler et al. (2004) for more details on the original sample design for the NCS-R.

Note from Table 4.2 that in the NCS-R sampling error calculation model the assumption of with-replacement sampling of ultimate clusters described in Chapter 3 is employed to address the analytic complexities associated with the multistage sampling and the without-replacement selection of NCS-R PSUs. The assumption that ultimate clusters are selected with replacement within primary stage strata ignores the finite population correction factor (see Section 2.4.2) and therefore results in a slight overestimation of true sampling variances for survey estimates.

To introduce several other features of the NCS-R sampling error calculation model, Table 4.3 illustrates the assignment of the sampling error stratum and cluster codes for six of the NCS-R sample design strata. In self-representing

TABLE 4.3

Illustration of NCS-R Sampling Error Code Assignments

	Original Sample Design		Sampling Error Calculation Model	
	Stratum	**PSU[a]**	**Stratum**	**Cluster**
SR	15	1 2 3 4 5 6 7 8 9 10 11 12	15	1 = {1, 3, 5, 7, 9, 11} 2 = {2, 4, 6, 8, 10, 12}
	16	1 2 3 4 5 6 7 8 9 10 11 12	16	1 = {1, 3, 5, 7, 9, 11} 2 = {2, 4, 6, 8, 10, 12}
	.....			
NSR	17	1701	17	1 = 1701
	18	1801		2 = 1801
	19	1901	18	1 = 1901
	20	2001		2 = 2001

[a] Recall from Section 2.8 and Table 4.2 that in self-representing (SR) strata, sampling begins with the selection of the smaller area segment units. Hence, in the NCS-R, the sampled units (coded 1–12) in each SR stratum (serving as its own PSU) are actually secondary sampling units. We include them in the PSU column because this was the first stage of non-certainty sampling in the SR strata. Unlike the SR PSUs, which serve as both strata and PSUs (each SR stratum is a PSU, that is, they are one and the same), the NSR strata can include multiple PSUs. In the NCS-R, one PSU was randomly selected from each NSR stratum (e.g., PSU 1701). NSR strata were then collapsed to form sampling error strata with two PSUs each to facilitate variance estimation.

(SR) design strata 15 and 16, the "area segments" constitute the first actual stage of noncertainty sample selection—hence, they are the ultimate cluster units with these two strata. To build the sampling error calculation model within each of these two SR design strata, the **random groups method** is used to assign the area segment units to two sampling error clusters. This is done to simplify the calculations required for variance estimation. As illustrated, NCS-R nonself-representing strata 17–20 contain a single PSU selection. The single PSU selected from each of these NSR strata constitutes an ultimate cluster selection. Because a minimum of two sampling error clusters per stratum is required for variance estimation, pairs of NSR design strata (e.g., 17 and 18, 19 and 20) are collapsed to create single sampling error strata with two sampling error computation units (PSUs) each.

Randomly grouping PSUs to form a sampling error cluster does not bias the estimates of standard errors that will be computed under the sampling error calculation model. However, forming random clusters by combining units does forfeit degrees of freedom, so the "variance of the variance estimate" may increase slightly.

If the collapsed stratum technique is used to develop the sampling error calculation model, slight overestimation of standard errors occurs because the

collapsed strata ignore the true differences in the design strata that are collapsed to form the single sampling error calculation stratum. The following section provides interested readers with more detail on the combined strata, random groups, and collapsed strata techniques used in building sampling error calculation models for complex sample survey data.

4.3.3 Combining Strata, Randomly Grouping PSUs, and Collapsing Strata

Combining strata in building the sampling error calculation model involves the combination of PSUs from two or more different strata to form a single stratum with larger pooled sampling error clusters. Consider Figure 4.1. The original complex sample design involved paired selection of two PSUs within each primary sampling stratum. For variance estimation, PSUs from the two design strata are combined, with the PSUs being randomly assigned into two larger sampling error clusters.

The technique of combining strata for variance estimation is typically used for one of two reasons: (1) The sample design has large numbers of primary stage strata and small numbers of observations per PSU, which could lead to variance estimation problems (especially when subclasses are being analyzed); or (2) the data producer wishes to group design PSUs to mask individual PSU clusterings as part of a disclosure protection plan. The sampling error calculation model for the NHANES has traditionally employed combined strata to mask the identity of individual PSUs.

The random groups method randomly combines multiple clusters from a single design stratum to create two or more clusters for sampling error estimation. This is the technique illustrated for NCS-R self-representing design strata 15 and 16 in Table 4.3. Figure 4.2 presents an illustration of forming random groups of clusters for variance estimation purposes. In this illustration,

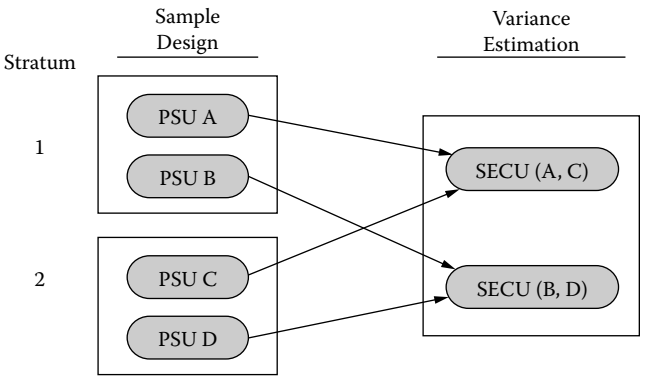


FIGURE 4.1
An example of combining strata.

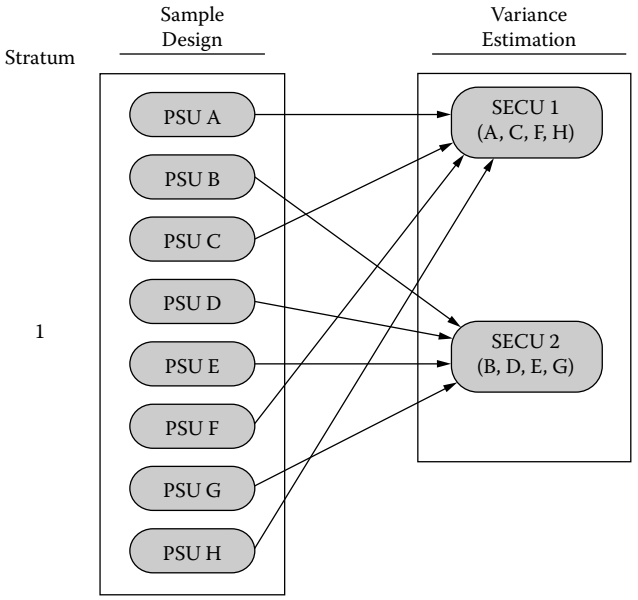


FIGURE 4.2
An illustration of the random groups method.

there are eight area segment selections within an original self-representing design stratum. (The New York metropolitan statistical area [MSA] is a common example of a geographic area that forms a self-representing stratum in national area probability samples of the U.S. population.) Figure 4.2 shows how the eight area segment clusters are randomly assigned into two larger sampling error clusters containing four of the original segments each.

The random groups method is typically used when a large number of small design PSUs have been selected within a single stratum and the data producer wishes to simplify the calculation of variance for that stratum or to minimize disclosure risk that could result from releasing the original stratification and cluster coding. This technique is commonly used to create sampling error clusters in self-representing primary stage strata of multistage sample designs.

Collapsing strata is another technique that data producers may employ in building a sampling error calculation model for a complex sample design. This technique is generally used when a single primary sampling unit is selected from each primary stage design stratum (which precludes direct estimation of sampling variance for the stratum). The technique involves "collapsing" or "erasing" the boundary between *adjacent* strata, so that similar strata become one larger pseudo- or collapsed stratum with multiple primary sampling units. In contrast to the combined stratum method, the original design PSUs remain distinct as sampling error clusters in the collapsed sampling error stratum. We illustrate the idea of collapsing strata in Figure 4.3.

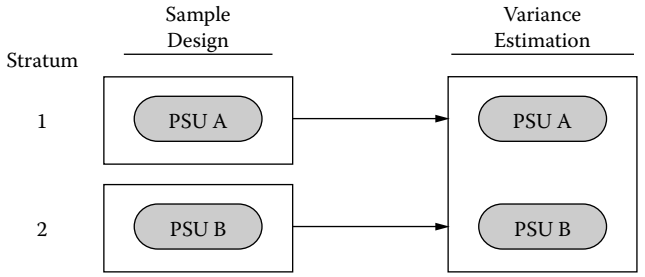


FIGURE 4.3
An example of collapsing strata.

4.3.4 Checking the Sampling Error Calculation Model for the Survey Data Set

As discussed previously in this chapter, identification of the variables identifying the sampling error strata and clusters is an essential step in the preparation for an analysis of a complex sample survey data set. When these variables have been identified, a useful first step is to cross-tabulate these variables to get a sense of the sample sizes in each stratum and cluster.

The Stata software system provides the `svyset` and `svydes` commands that enable users to define these key sample design variables (including sampling weights) for a data set and then describe the distributions of the cases to sampling error strata and clusters. These commands need to be submitted only once in a given Stata session, unlike other software procedures that require these variables to be identified for each individual analysis (e.g., SAS PROC SURVEYMEANS, SUDAAN PROC DESCRIPT); readers can refer to Appendix A for details.

We illustrate the use of the `svyset` and `svydes` commands to show how these key design variables are identified and described in Stata for the NCS-R. First, the correct design variables were determined based on the documentation for the survey data set. Next, the following two Stata commands (`svyset` and `svydes`) were submitted, which produced the output that follows the commands:

```
svyset seclustr [pweight=ncsrwtlg], strata(sestrat) ///
vce(linearized) singleunit(missing)
svydes
```

Stratum	Number of Units Included	Number of Units Omitted	Observations with Complete Data	Observations with Missing Data	Observations per Included Unit		
					Min	Mean	Max
1	2	0	44	0	22	22.0	22
2	2	0	53	0	26	26.5	27

Stratum	Number of Units Included	Number of Units Omitted	Observations with Complete Data	Observations with Missing Data	Observations per Included Unit		
					Min	Mean	Max
3	2	0	77	0	31	38.5	46
4	2	0	87	0	42	43.5	45
5	2	0	72	0	32	36.0	40
6	2	0	68	0	31	34.0	37
7	2	0	127	0	61	63.5	66
8	2	0	84	0	38	42.0	46
9	2	0	89	0	44	44.5	45
10	2	0	78	0	23	39.0	55
11	2	0	57	0	28	28.5	29
12	2	0	61	0	24	30.5	37
13	2	0	63	0	26	31.5	37
14	2	0	45	0	20	22.5	25
15	2	0	73	0	36	36.5	37
16	2	0	66	0	30	33.0	36
17	2	0	41	0	18	20.5	23
18	2	0	55	0	26	27.5	29
19	2	0	62	0	29	31.0	33
20	2	0	159	0	70	79.5	89
21	2	0	187	0	75	93.5	112
22	2	0	178	0	82	89.0	96
23	2	0	186	0	84	93.0	102
24	2	0	197	0	94	98.5	103
25	2	0	264	0	124	132.0	140
26	2	0	155	0	58	77.5	97
27	2	0	172	0	79	86.0	93
28	2	0	117	0	45	58.5	72
29	2	0	199	0	99	99.5	100
30	2	0	121	0	57	60.5	64
31	2	0	267	0	131	133.5	136
32	2	0	191	0	86	95.5	105
33	2	0	82	0	37	41.0	45
34	2	0	236	0	108	118.0	128
35	2	0	236	0	94	118.0	142
36	2	0	203	0	100	101.5	103
37	2	0	218	0	106	109.0	112
38	2	0	197	0	84	98.5	113
39	2	0	215	0	76	107.5	139
40	2	0	164	0	66	82.0	98
41	2	0	211	0	94	105.5	117
42	2	0	235	0	116	117.5	119

Stratum	Number of Units Included	Number of Units Omitted	Observations with Complete Data	Observations with Missing Data	Observations per Included Unit		
					Min	Mean	Max
42	84	0	5,692	0	18	67.8	142
			5,692				
			3,590[a]				
			9,282				

[a] Number of observations with missing values in the survey characteristics.

This Stata output illustrates the frequency distributions of the design variables that are produced by running the `svyset` and `svydes` commands for the NCS-R data set. The sampling error calculation model for the NCS-R sample design recognizes stratification of the population, and the variable containing the stratum codes for sampling error calculations is SESTRAT. The sampling error cluster (SECLUSTR) and overall analysis weight (NCSRWTLG) variables are also identified, in addition to the method of variance estimation (Taylor series linearization). We also indicate what Stata (Version 10+) should do if strata containing a single sampling error cluster are encountered in an analysis (variance estimates should be set to missing; this is the default in earlier versions of Stata). Additional options for the `svyset` command are described in Appendix A. We note that if Stata users wish to reset the design variables declared in the `svyset` command, the `svyset, clear` command can be submitted.

The output from the previous `svydes` command indicates that the NCS-R sampling error calculation model has been specified as a paired selection design, with two sampling error clusters (# Units included) in each of 42 sampling error strata (for 84 sampling error clusters total). Sample sizes in the clusters (denoted in the Stata output by "# Obs per included Unit") range from 18 to 142, with an average of 67.8 sample respondents per ultimate cluster. This "two-per-stratum" sampling error coding is common in many public-use survey data sets and enables survey analysts to employ any of the variance estimation techniques (Taylor series, BRR, JRR) discussed in Chapter 3. Stata also reports that 3,590 observations have missing data on at least one key design variable. This does not come as a surprise, because the Part II sampling weight variable for NCS-R cases (NCSRWTLG) takes on missing values for cases that were not selected or subsampled for Part II of the NCS-R survey.

These simple initial tabulations help to familiarize data users with the sampling error calculation models that will be used for variance estimation and can be useful for describing complex sample design information and methods of variance estimation in scientific and technical publications. Analysts using other statistical software packages (e.g., SAS, SPSS, and SUDAAN) can easily perform these simple descriptive analyses using standard procedures for cross-tabulation.

4.4 Addressing Item Missing Data in Analysis Variables

Many survey data analysts, initially excited to begin the analysis of a new survey data set, are disappointed to find that many of the key survey variables they hope to analyze suffer from problems of missing data. As they move from simple univariate analysis to multivariate analysis, the missing data problem is compounded due to the deletion from analysis by statistical software of any cases with missing values on one or more variables.

The first fact that all data users should recognize is that no survey data set is immune to some form of missing data problem. Therefore, addressing missing data should be as commonplace an analytic activity as choosing the form of a regression model to apply to a dependent variable of interest. Section 2.7 has already introduced weighting adjustments for unit nonresponse—one major source of missing data in sample surveys.

Here we will focus briefly on the implications of ignoring item-missing data in analysis and then will describe a convenient method for investigating the rates and patterns of missing data in a chosen set of analysis variables. Chapter 11 will examine methods for statistical imputation of item-missing data using today's powerful new software tools.

4.4.1 Potential Bias Due to Ignoring Missing Data

Many analysts simply choose to ignore the missing data problem when performing analyses. If the rates of missing data on key analysis variables are very low (say < 1–2% of cases), the penalty for not taking active steps (i.e., weighting or imputation) to address the missing data is probably small. Consider a univariate analysis to estimate the population mean of a variable y. Under a simple deterministic assumption that the population is composed of "responders" and "nonresponders," the expected bias in the respondent mean of the variable y is defined as follows:

$$Bias(\bar{Y}_R) = \bar{Y}_R - \bar{Y} = P_{NR} \times (\bar{Y}_R - \bar{Y}_{NR}) \tag{4.1}$$

where $\bar{Y}$ is the true population mean, $\bar{Y}_R$ is the population mean for responders in the population, $\bar{Y}_{NR}$ is the population mean for nonresponders in the population, and P_{NR} is the expected proportion of nonrespondents in a given sample. For the bias to be large, the rate of missing data must be sizeable, and respondents must differ in their characteristics from nonrespondents. In statistical terms, the potential bias due to missing data depends on both the missing data pattern and the missing data mechanism. Chapter 11 will provide a more detailed review of missing data patterns and mechanisms.

4.4.2 Exploring Rates and Patterns of Missing Data Prior to Analysis

Stata provides data analysts with the `mvpatterns` command to display the patterns and rates of missing data across a set of variables that will be included in an analysis. The following example uses this command to explore the missing data for seven 2005–2006 NHANES variables that will be included in the multiple imputation regression example presented in Section 11.7.2. The variables are diastolic blood pressure (BPXDI1_1), marital status (MARCAT), gender (RIAGENDR), race/ethnicity (RIDRETH1), age (AGEC and AGECSQ), body mass index (BMXBMI), and family poverty index (INDFMPIR). The command syntax to request the missing data summary for this set of seven variables is as follows:

```
mvpatterns bpxdi1_1 marcat riagendr ridreth1 agec agecsq ///
bmxbmi indfmpir
```

The actual Stata output produced by the `mvpatterns` command is provided next. The first portion of the output lists the number of observed and missing values for each variable that has at least one missing value. Since 2005–2006 NHANES has no missing data for age, gender, and race/ethnicity, these variables are not listed in the output. The second portion of the output summarizes the frequencies of various patterns of missing data across the four variables with missing data, using a coding of "+" for observed and "." for missing. For example, 4,308 observations have no missing data for these four variables, and a total of 49 observations have missing data only for the BMXBMI (body mass index) variable.

Variables with No mv's: riagendr ridreth1 agec agecsq

Variable	type	obs	mv	variable label
bpxdi1_1	int	4581	753	diastolic bp
marcat	byte	5329	5	1=married 2=prev married 3=never married
bmxbmi	float	5237	97	body mass index (kg/m**2)
indfmpir	float	5066	268	family pir

Patterns of Missing Values

_pattern	_mv	_freq
++++	0	4308
.+++	1	666
+++.	1	217
++.+	1	49

_pattern	_mv	_freq
.++.	2	41
.+.+	2	39
.+..	3	5
++..	2	4
+.++	1	3
..++	2	1
..+.	3	1

With recent theoretical advances in the theory of statistical analysis with missing data (Little and Rubin, 2002) and today's improved software (e.g., Carlin, Galati, and Royston, 2008; Raghunathan et al., 2001), data producers or data users often consider different methods for the **imputation** (or prediction) of missing data values. Depending on the patterns of missing data and the underlying process that generated the missingness, statistically sound imputation strategies can produce data sets with all complete cases for analysis.

In many large survey programs, the data producer may perform imputation for key variables before the survey data set is released for general use. Typically, methods for **single imputation** are employed using techniques such as **hot deck imputation, regression imputation,** and **predictive mean matching**. Some survey programs may choose to provide data users with **multiply imputed** data sets (Kennickell, 1998; Schafer, 1996). When data producers perform imputation of item-missing data, a best practice in data dissemination is to provide data users with both the imputed version of the variable (e.g., I_INCOME_AMT) and an indicator variable (e.g., I_INCOME_FLG) that identifies the values of the variables that have been imputed. Data users should expect to find general documentation for the imputations in the technical report for the survey. The survey codebook should clearly identify the imputed variables and the imputation "flag" variables.

When imputations are not provided by the data producer (or the analyst chooses to employ his or her own method and model), the task of imputing item missing data will fall to the data user. Details on practical methods for user imputation of item-missing data are provided later in Chapter 11.

4.5 Preparing to Analyze Data for Sample Subpopulations

Analysis of complex sample survey data sets often involves separate estimation and inference for subpopulations or **subclasses** of the full population.

For example, an analyst may wish to use NHANES data to estimate the prevalence of diabetes separately for male and female adults or compute HRS estimates of retirement expectations only for the population of the U.S. Census South Region. An NCS-R analyst may choose to estimate a separate logistic regression model for predicting the probability of past-year depression status for African American women only. When analyses are focused in this way on these specific subclasses of the survey population, special care must be taken to correctly prepare the data and specify the subclass analysis in the command input to software programs. Proper analysis methods for subclasses of survey data have been well established in the survey methodology literature (Cochran, 1977; Fuller et al., 1989; Kish, 1965; Korn and Graubard, 1999; Lohr, 1999; Rao, 2003), and interested readers can consult these references for more general information on estimation of survey statistics and related variance estimation techniques for subclasses. A short summary of the theory underlying subclass analysis of survey data is also provided in Theory Box 4.2.

4.5.1 Subpopulation Distributions across Sample Design Units

Although most current survey data analysis software is programmed to correctly account for subclasses in analysis, a useful step in preparing for data analysis is to examine the distribution of the targeted subpopulation sample with respect to the sampling error strata and clusters that have been defined under the sampling error calculation model. Figure 4.4 illustrates the different distributional patterns that might be observed in practice.

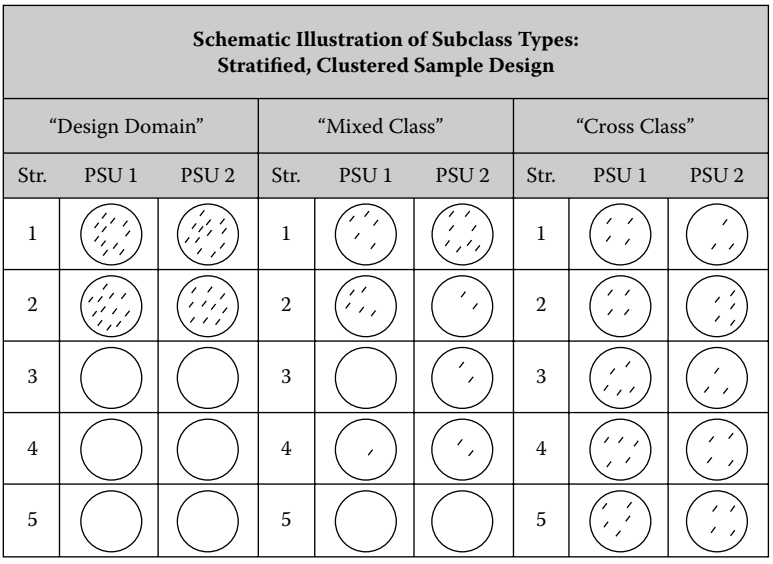


FIGURE 4.4
Schematic illustration of subclass types for a stratified, clustered design.

THEORY BOX 4.2 THEORETICAL MOTIVATION FOR UNCONDITIONAL SUBCLASS ANALYSES

To illustrate the importance of following the unconditional subclass analysis approach mathematically, we consider the variance of a sample total (the essential building block for variance estimation based on Taylor series linearization). We denote design strata by h ($h = 1, 2, \ldots, H$), first-stage PSUs within strata by α ($\alpha = 1, 2, \ldots, a_h$), and sample elements within PSUs by i ($i = 1, 2, \ldots, n_{h\alpha}$). The weight for element i, taking into account factors such as unequal probability of selection, nonresponse, and possibly poststratification, is denoted by $w_{h\alpha i}$. We refer to specific subclasses using the notation S. An estimate of the total for a variable Y in a subclass denoted by S is computed as follows (Cochran, 1977):

$$\hat{Y}_S = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I_{S,h\alpha i} y_{h\alpha i} \tag{4.2}$$

In this notation, I represents an indicator variable, equal to 1 if sample element i belongs to subclass S and 0 otherwise. The closed-form analytic formula for the variance of this subclass total can be written as follows:

$$Var(\hat{Y}_S) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \left[\sum_{\alpha=1}^{a_h} \left(\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I_{S,h\alpha i} y_{h\alpha i} \right)^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I_{S,h\alpha i} y_{h\alpha i} \right)^2}{a_h} \right] \tag{4.3}$$

This formula shows that the variance of the subclass total is calculated by summing the between-cluster variance in the subclass totals within strata, across the H sample strata. The formula also shows how the indicator variable is used to ensure that all sample elements (and their design strata and PSUs) are recognized in the variance calculation; this emphasizes the need for the software to recognize all of the original design strata and PSUs. Analysts should note that if all $n_{h\alpha}$ elements within a given stratum denoted by h and PSU denoted by α do not belong to the subclass S (although elements from that subclass theoretically *could* belong to that PSU in any given sample), that PSU will still contribute to the variance estimation: The PSU helps to

define the total number of PSUs within stratum h (a_h) and contributes a value of 0 to the sums in the variance estimation formula. In this way, sample-to-sample variability in the estimation of the total due to the fact that the subclass sample size is a random variable is captured in the variance calculation.

Design domains, or subclasses that are restricted to only a subset of the primary stage strata (e.g., adults in the Census South Region, or residents of urban counties), constitute a broad category of analysis subclasses. In general, analysis of design domain subclasses should not be problematic in most contemporary survey analysis software. Analysts should recognize that sampling errors estimated for domain subclasses will be based on fewer degrees of freedom than estimates for the full sample or cross-classes, given that design domains are generally restricted to specific sampling strata.

A second pattern that may be observed is a mixed class, or a population subclass that is not isolated in a subset of the sample design strata but is unevenly and possibly sparsely distributed across the strata and clusters of the sampling error calculation model. Experience has shown that many survey analysts often "push the limits" of survey design, focusing on rare or highly concentrated subclasses of the population (e.g., Hispanic adults with asthma in the HRS). While such analyses are by no means precluded by the survey design, survey analysts are advised to exercise care in approaching subclass analyses of this type for several reasons: (1) Nominal sample sizes for mixed classes may be small but design effects due to weighting and clustering may still be substantial; (2) a highly uneven distribution of cases to strata and clusters will introduce instability or bias in the variance estimates (especially those based on the Taylor series linearization method); and (3) software approximations to the complex sample design degrees of freedom (df = # clusters – # strata) may significantly overstate the true precision of the estimated sampling errors and confidence intervals. Survey analysts who intend to analyze data for mixed classes are encouraged to consult with a survey statistician, but we would recommend that these subclasses be handled using unconditional subclass analysis approaches (see Section 4.5.2).

The third pattern we will label a cross-class (Kish, 1987). Cross-classes are subclasses of the survey population that are broadly distributed across all strata and clusters of a complex sample design. Examples of cross-classes in a national area probability sample survey of adults might include males, or individuals age 40 and above. Properly identified to the software, subclasses that are true cross-classes present very few problems in survey data analysis—sampling error estimation and degrees of freedom determination for confidence intervals and hypothesis tests should be straightforward.

We now consider two alternative approaches to subclass analysis that analysts of complex sample survey data sets could take in practice and discuss applications where one approach might be preferred over another.

4.5.2 The Unconditional Approach for Subclass Analysis

Recall from Chapter 3 that the estimated standard error of an estimated survey statistic expresses the degree of variability in the statistic that we would expect from one hypothetical sample to another around the true population parameter of interest. If the subclass analysis were restricted only to sample cases that belonged to the subclass (sometimes referred to as the **conditional subclass analysis** approach, because it conditions subclass inference on the observed sample), sampling error estimates would in effect assume that each hypothetical sample would have the *same* fixed number of subclass members (say m) and that the distribution of the fixed subclass sample size, m, across design strata and clusters would remain unchanged from sample to sample. When analyzing complex sample survey data sets in practice, this would be the case only when analyzing design domains (as previously defined) with sample sizes that are fixed by design.

In all other applications, the subclass sample size is a random variable—varying both in its size and its distribution across the design strata and clusters. By performing an **unconditional subclass analysis** (i.e., one that does not condition on the observed distribution of subclass cases to particular strata and clusters of the full sample design, as described in West, Berglund, and Heeringa, 2008), we take into account these added sources of true sampling variability. Point estimates of population parameters will be identical under both subclass analysis approaches. However, because a conditional analysis does *not* incorporate this latter source of variance, it tends to result in *underestimates* of standard errors, which lead analysts to overstate the precision of estimated survey statistics for subclasses. Therefore, simply using the data management capabilities of statistical software to delete or filter out those cases that do not fall into the subclass of interest can produce biased inferences. This requires some additional preparation on the part of the data analyst.

4.5.3 Preparation for Subclass Analyses

To properly prepare for a correct unconditional subclass analysis of a complex sample survey data set, data users should generate indicator variables for each of the subclasses that they are interested in analyzing:

$$I_{S,i} = \begin{cases} 1 & \text{if case } i \text{ is a member of the subclass of interest} \\ 0 & \text{if case } i \text{ is not a member of the subclass of interest} \end{cases}$$

Each case in the complete sample should be assigned a value of 1 or 0 for this indicator variable. This indicator variable will be used to identify the subclass members to the analysis software. Once these indicator variables have been created for subclasses of interest, data users need to use appropriate software options for unconditional subclass analyses. For example, Stata procedures offer the `subpop()` option, where the subclass indicator variable can be specified in the parentheses; SUDAAN offers the SUBPOPN keyword; SPSS (in the Complex Samples Module) offers the SUBPOP and DOMAIN subcommands; while SAS offers the DOMAIN keyword for PROC SURVEYMEANS, PROC SURVEYREG and PROC SURVEYLOGISTIC specifically and allows users to indicate domains using the first variable listed in the TABLES statement of PROC SURVEYFREQ.

4.6 A Final Checklist for Data Users

We conclude this chapter with a summary of the important preparation steps that data users should follow prior to beginning the analysis of a complex sample survey data set:

1. Review the documentation for the data set provided by the data producer, specifically focusing on sections discussing the development of sampling weights and sampling error (standard error) estimation. Contact the data producer if any questions arise.
2. Identify the correct weight variable for the analysis, keeping in mind that many survey data sets include separate weights for different types of analyses. Perform simple descriptive analyses of the weight variable noting the general distribution of the weights, whether the weights have been normalized or whether there are missing or 0 weight values for some cases. Select a few key variables from the survey data set, and compare weighted and unweighted estimates of parameters for these variables.
3. Identify the variables in the data set containing the sampling error calculation codes (for strata and clusters) that define the sampling error calculation model. Examine how many clusters were selected from each sampling stratum (according to the sampling error calculation model) and whether particular clusters have small sample sizes. If only a single sampling error cluster is identified in a sampling stratum, contact the data producer, or consult the documentation for the data set for guidance on recommended variance estimation methods. Determine whether replicate sampling weights are present if sampling error calculation codes are not available, and

make sure that the statistical software is capable of accommodating replicate weights (see Section 4.2.1).

4. Create a final analysis data set containing only the analysis variables of interest (including the analysis weights, sampling error calculation variables, and case identifiers). Examine univariate and bivariate summaries for the key analysis variables to determine possible problems with missing data or unusual values on the individual variables.
5. Review the documentation provided by the data producer to understand the procedure (typically nonresponse adjustment) used to address unit nonresponse or nonresponse to a wave or phase of the survey data collection. Analyze the rates and patterns of item-missing data for all variables that will be included in the analysis. Investigate the potential missing data mechanism by defining indicator variables flagging missing data for the analysis variables of interest. Use statistical tests (e.g., chi-square tests, two-sample *t*-tests) to see if there are any systematic differences between respondents providing complete responses and respondents failing to provide complete responses on important analysis variables (e.g., demographics). Choose an appropriate strategy for addressing missing data using the guidance provided in Section 4.4 and Chapter 11.
6. Define indicator variables for important analysis subclasses. *Do not delete cases that are not a part of the primary analysis subclass*. Assess a cross-tabulation of the stratum and cluster sampling error calculation codes for the subclass cases to identify the distribution of the subclass to the strata and clusters of the sample design. Consult a survey statistician prior to analysis of subclasses that exhibit the "mixed class" characteristic illustrated in Figure 4.4.

Based on many years of consulting on the analysis of survey data, the authors firmly believe that following these steps before performing any analyses of complex sample survey data sets will eliminate the most common mistakes that result in wasted time and effort or (even worse) unknowingly result in biased, incorrect analyses of the survey data.

5

Descriptive Analysis for Continuous Variables

> Nonfarm payroll employment fell sharply (–533,000) in November, and the unemployment rate rose from 6.5 to 6.7 percent, the Bureau of Labor Statistics of the U.S. Department of Labor reported today.... Job losses were large and widespread across the major industry sectors in November. (U.S. Bureau of Labor Statistics, November 2008, http://www.bls.gov)

5.1 Introduction

Estimation of totals, means, variances, and distribution percentiles for survey variables may be the primary objective of an analysis plan or possibly an exploratory step on a path to a more multivariate treatment of the survey data. Major reports of survey results can be filled with detailed tabulations of descriptive estimates. Scientific papers and publications that may emphasize a more analytical multivariate treatment of the data typically include initial descriptive estimation of population characteristics as part of the overall analysis plan.

This chapter describes techniques for generating population estimates of descriptive parameters for continuous survey variables. Section 5.2 examines features that distinguish descriptive analysis of complex sample survey data from more standard descriptive analyses of simple random samples or convenience samples. Basic methods and examples of estimating important descriptive parameters for univariate, continuous distributions are described in Section 5.3, followed in Section 5.4 by methods for studying simple bivariate relationships between two continuous survey variables. The chapter concludes with coverage of methods for descriptive estimation for subpopulations (Section 5.5) and estimation of simple linear functions of distributional statistics such as differences of means (Section 5.6).

5.2 Special Considerations in Descriptive Analysis of Complex Sample Survey Data

Descriptive analysis of complex sample survey data shares much in common with the standard statistical methods that are taught in any introductory statistics course, but there are several subtle differences that are important to keep in mind.

5.2.1 Weighted Estimation

Descriptive analysis of both continuous and discrete survey measures is intended to characterize the distributions of variables over the full survey population. If population elements have differing probabilities of inclusion in the sample (deliberately introduced by the sampling statistician or induced by a nonresponse mechanism), unbiased estimation for the population distribution requires weighting of the sample data. Depending on the distribution of the weights and the correlations between the weights and the survey variables of interest, unweighted sample estimates may yield a very biased picture of the true population distribution.

It is informative to plot the value of the survey weight against the survey variable of interest. Figure 5.1 illustrates a plot of this type, with the 2006 Health and Retirement Study (HRS) survey weight value on the vertical axis and the value of 2006 HRS total household assets on the horizontal axis. Diagnostic plots of this type inform the user of the following important

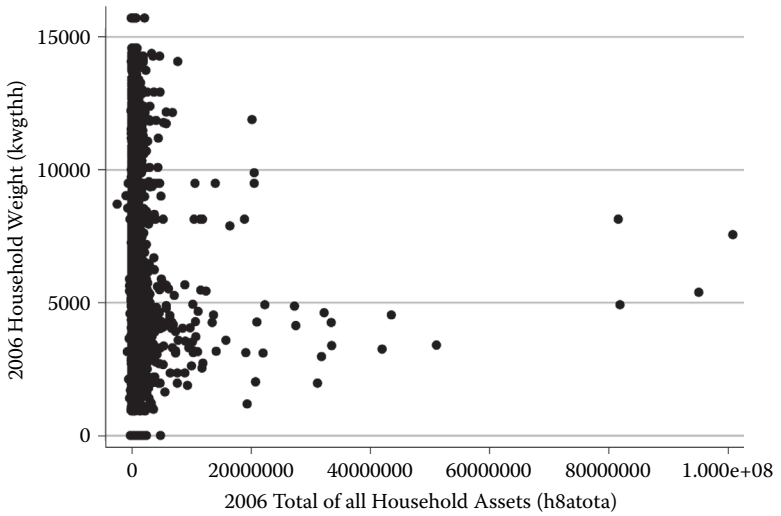


FIGURE 5.1
Examining the relationship of a survey variable and the survey weight in the 2006 HRS data set.

properties of the weights and their relationship to the variable of interest: (1) the variability of the weights, which can affect the sampling variability of descriptive estimates (see Section 2.7.5); (2) any strong functional relationship between individual weight values and survey observations, signaling that the weights are highly "informative" for the estimation of population parameters; and (3) the potential for "weighted outliers"—points representing large values on both the survey variable and the analysis weight that in combination will have a major influence on the estimated distributional statistics and their standard errors. For example, Figure 5.1 suggests that there is considerable variability in the 2006 HRS household weights, but there is little evidence of a functional (e.g., linear, quadratic) relationship of the weights to the measures of total household assets. Therefore, for the total household assets variable, weighted estimation will result in increased sampling variances for estimates, but the weighting may not have major impacts on the estimates of population parameters (see Section 2.7). The plot also identifies four HRS cases with exceptionally large values for total household assets. The analysis weights for these cases are near the midpoint of the weight range. Nevertheless, these four points may warrant investigation, because they may have significant influence on the estimates of descriptive parameters in addition to substantial leverage in the estimation of multivariate models.

5.2.2 Design Effects for Descriptive Statistics

Estimates of descriptive parameters (or descriptive statistics) derived from complex sample survey data can be subject to substantial design effects, due to the stratification, clustering, and weighting associated with the design. Empirical research has consistently demonstrated that complex sample design effects are largest for weighted estimates of population means and totals for single survey variables, smaller for estimates of subpopulation means and proportions, and substantially reduced for regression coefficients and other statistics that are functions of two or more variables (Skinner, Holt, and Smith, 1989; Kish, Groves, and Kotki, 1976). Throughout this chapter, the example analyses will include estimation of design effect ratios to provide the reader with a sense of the magnitude of the design effects and their variability across survey variables and estimated statistics.

5.2.3 Matching the Method to the Variable Type

Survey responses are recorded on a variety of scales, including simple binary choices (yes, no), nominal categories (Democrat, Republican, Independent), ordinal scales (1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor), ordinal counts (years of education, weeks not employed), interval scale variables (age in years), fully continuous variables (systolic blood pressure in mmHg), grouped or "interval censored" continuous variables (income categories, e.g., $10,000–$24,999), and censored or "semicontinuous" variables (the value of

household assets in stocks and bonds). With the obvious exception of nominal categorical variables and interval censored measures, it is common for survey analysts to apply descriptive analysis techniques appropriate for continuous data to survey variables measured on scales that are not strictly continuous.

The descriptive analytic methods described in this chapter may certainly be applied to ordinal, interval scale and semicontinuous survey measures; however, analysis results should be presented in a way that acknowledges the "noncontinuous" character of the underlying data. The classical interpretation of ordinal scales as a measure of an underlying continuous latent variable recognizes that the mapping of that construct to the ordinal metric may not be linear—moving from poor to fair on a health rating scale may not reflect the same degree of health improvement as moving from fair to good. For example, the estimated 75th percentile of the distribution of household mortgage debt for HRS households may be a misleading statistic if the percent of households with no mortgage is not also reported. In short, as is true in all statistical summarization, it is the survey analyst's responsibility to ensure that the reporting and interpretation of even the simplest descriptive analyses are accurate and transparent to the intended audience.

5.3 Simple Statistics for Univariate Continuous Distributions

This section outlines methods for estimating simple statistics that describe the population distributions of survey variables, including graphical tools and statistical estimators for population totals, means, variances, and percentiles. Weighted estimators of the most important distributional parameters for continuous survey variables are described in this section, and, with a few exceptions, most of the major statistical software packages support weighted estimation for these parameters. When available, estimation of these univariate descriptive parameters and the corresponding standard errors is illustrated in Stata using the National Comorbidity Survey Replication (NCS-R), National Health and Nutrition Examination Survey (NHANES), and HRS data sets. Examples of command syntax for estimating these same descriptive statistics using other software systems can be found on the book Web site.

5.3.1 Graphical Tools for Descriptive Analysis of Survey Data

The authors' experience suggests that statistical graphics are currently underused as tools for the exploratory analysis of survey data (Tukey, 1977) or the effective visual display of estimated population distributions of single variables or the population relationships among two or more independent variables (Cleveland, 1993; Tufte, 1983). A thorough treatment of the many

graphical tools available to today's analysts would require a separate volume and, as a general subject, is well covered in any number of existing publications (Maindonald and Braun, 2007; Mitchell, 2008; SAS Institute Inc., 2009). In this chapter, we will use selected tools in Stata to illustrate the concepts of graphical presentation and exploration of survey data.

The key requirement in any graphical treatment of complex sample survey data is that the method and the software must enable the analyst to incorporate the influences of the survey weights in the construction of the final display. Just as unweighted estimation of a mean may be biased for the true population value, a statistical graphic developed from unweighted survey data may present a misleading image of the true survey population that the analysis is intended to represent. The graphics programs in today's statistical software packages differ in their capability to incorporate weights in graphical analyses. Stata is one package that can incorporate the survey weights in the development of statistical graphics.

Stata graphics include a number of options for displaying the estimated form of the population distribution for a single variable, including histograms, boxplots, spikeplots, and kernel density (kdensity) plots. The following two examples demonstrate Stata's capability to build histogram and boxplot graphics that incorporate survey weights.

Example 5.1: A Weighted Histogram of Total Cholesterol Using the 2005–2006 NHANES Data

After first opening the NHANES data set in Stata, we generate an indicator variable for adults (respondents with age greater than or equal to 18), for use in the analyses:

```
gen age18p = 1 if age >= 18 & age != .
replace age18p = 0 if age < 18
```

The Stata software considers system missing values as being larger than the largest available (i.e., nonmissing) value for a variable in a data set. This is the reason for the condition that the value on the AGE variable must be both greater than or equal to 18 and not missing for the indicator variable to be equal to 1. System missing values are important to consider in *all* software packages when manually creating indicator variables so that missing values are not mistakenly coded into either a 1 or 0.

The following Stata commands then generate the weighted histogram of NHANES total cholesterol measures presented in Figure 5.2:

```
generate int_wtmec2yr = int(wtmec2yr)
histogram lbxtc if age18p [fweight=int_wtmec2yr]
```

Note that the Stata command requires that the weight be specified as an integer (frequency weight) value. Since these frequency weights must be integer values, a new weight is created by truncating the decimal places from the original values

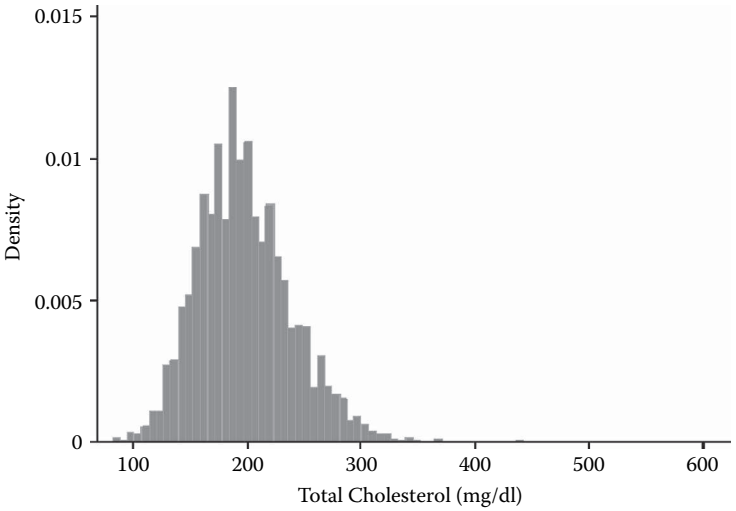


FIGURE 5.2
A weighted histogram of total cholesterol, based on 2005–2006 NHANES data.

of the 2005–2006 NHANES mobile examination center (MEC) analysis weight, WTMEC2YR. Since WTMEC2YR is a population scale weight, trimming the decimal places will have no significant effect on the graphical representation of the estimated distribution of total cholesterol.

Figure 5.2 displays the histogram output from Stata graphics. Since NHANES sample data inputs have been weighted, the distributional form of the histogram is representative of the estimated distribution for the NHANES survey population of U.S. adults age 18 and older.

Example 5.2: Weighted Boxplots of Total Cholesterol for U.S. Adult Men and Women Using the 2005–2006 NHANES Data

For scientific reports and papers, boxplots are a very useful tool for graphical presentation of the estimated population (weighted) distributions of single survey variables. The following Stata command requests a pair of boxplots that represent the estimated gender-specific population distributions of total serum cholesterol in U.S. adults:

```
graph box lbxtc [pweight=wtmec2yr] if age18p==1, by(female)
```

Note that the Stata `graph box` command permits the specification of the analysis weight as a `pweight` (or probability weight) variable—no conversion to integer values is needed. As we will see in upcoming examples, Stata's commands designed for the analysis of (complex sample) survey data all expect users to initially identify a *probability* sampling weight variable, which may not necessarily take on integer values.

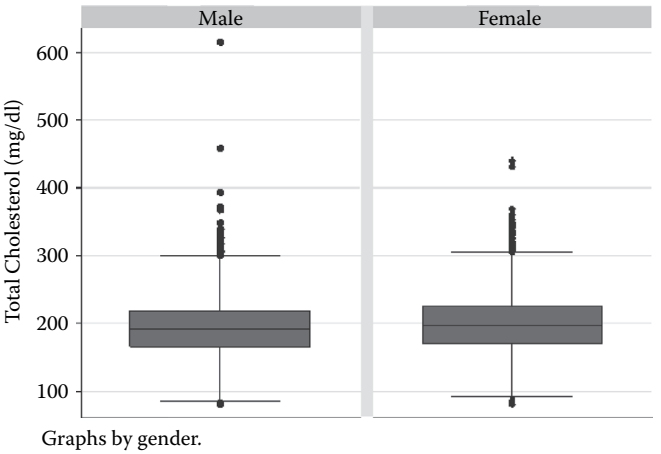


FIGURE 5.3
Boxplots of the gender-specific weighted distributions of total serum cholesterol for U.S. adults. (From 2005–2006 NHANES data.)

The pair of boxplots generated by the Stata command is displayed in Figure 5.3. These boxplots suggest that male and female adults in the NHANES survey population have very similar distributions on this particular continuous variable.

5.3.2 Estimation of Population Totals

The estimation of a population total and its sampling variance has played a central role in the development of probability sampling theory. In survey practice, optimal methods for estimation of population totals are extremely important in government and academic surveys that focus on agriculture (e.g., acres in production), business (e.g., total employment), and organizational research (e.g., hospital costs). Acres of corn planted per year, total natural gas production for a 30-day period, and total expenditures attributable to a prospective change in benefit eligibility are all research questions that require efficient estimation of population totals. In those agencies and disciplines where estimation of population totals plays a central role, advanced "model-assisted" estimation procedures and specialized software are the norm. Techniques such as the generalized regression (GREG) estimator or **calibration estimators** integrate the survey data with known population controls on the distribution of the population weighting factors to produce efficient weighted estimates of population totals (DeVille and Särndal, 1992; Valliant, Dorfman, and Royall, 2000).

Most statistical software packages do not currently support advanced techniques for estimating population totals such the GREG or calibration

methods. Stata and other software systems that support complex sample survey data analysis do provide the capability to compute simple weighted or **expansion estimates** of finite population totals and also include a limited set of options for including population controls in the form of post-stratified estimation (see Theory Box 5.1).

In the case of a complex sample design including stratification (with strata indexed by $h = 1, \ldots, H$) and clustering (with clusters within stratum h indexed by $\alpha = 1, 2, \ldots, a_h$), the simple weighted estimator for the population total can be written as follows:

$$\hat{Y}_w = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} \tag{5.1}$$

A closed-form, unbiased estimate of the variance of this estimator is

$$\text{var}(\hat{Y}_w) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \left[\sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} (w_{h\alpha i} y_{h\alpha i})^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} \right)^2}{a_h} \right] \tag{5.2}$$

(Recall from Section 3.6.2 that this simple estimator for the variance of a weighted total plays an important role in the computation of Taylor series linearization (TSL) estimates of sampling variances for more complex estimators that can be approximated as linear functions of estimated totals.)

This simple weighted estimator of the finite population total is often labeled the **Horvitz–Thompson or H–T estimator** (Horvitz and Thompson, 1952). Practically speaking, this labeling is convenient, but the estimator in Equation 5.1 and the variance estimator in Equation 5.2 make additional assumptions beyond those that are explicit in Horvitz and Thompson's original derivation. Theory Box 5.1 provides interested readers with a short summary of the theory underlying the H–T estimator.

Two major classes of total statistics can be estimated using Equation 5.1. If $y_{h\alpha i}$ is a binary indicator for an attribute (e.g., 1 = has the disease, 0 = disease free), the result is an estimate of the size of the subpopulation that shares that attribute:

$$\hat{Y}_w = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} = \hat{M} \leq \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} \cdot 1 = \hat{N}$$

THEORY BOX 5.1 THE HORVITZ–THOMPSON ESTIMATOR OF A POPULATION TOTAL

The **Horvitz–Thompson estimator** (Horvitz and Thompson, 1952) of the population total for a variable Y is written as follows:

$$\hat{Y} = \sum_{i=1}^{N} \frac{\delta_i Y_i}{p_i} = \sum_{i=1}^{n} \frac{y_i}{p_i} \tag{5.3}$$

In Equation 5.3, $\delta_i = 1$ if element i is included in the sample and 0 otherwise, and p_i is the probability of inclusion in the sample for element i. The H–T estimator is an unbiased estimator for the population total Y, because the only random variable defined in the estimator is the indicator of inclusion in the sample (the y_i and p_i values are fixed in the population):

$$E(\hat{Y}) = E\sum_{i=1}^{N} \frac{\delta_i Y_i}{p_i} = \sum_{i=1}^{N} \frac{E(\delta_i) Y_i}{p_i} = \sum_{i=1}^{N} \frac{p_i Y_i}{p_i} = \sum_{i=1}^{N} Y_i = Y \tag{5.4}$$

An unbiased estimator of the sampling variance of the H–T estimator is

$$\text{var}(\hat{Y}) = \sum_{i=1}^{n} y_i^2 \frac{(1-p_i)}{p_i^2} + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{y_i y_j}{p_{ij}} \left(\frac{p_{ij} - p_i p_j}{p_i p_j} \right) \tag{5.5}$$

In this expression, p_{ij} represents the probability that both elements i and j are included in the sample; these joint inclusion probabilities must be supplied to statistical software to compute these variance estimates.

The H–T estimator weights each sample observation inversely proportionate to its sample selection probability, $w_{HT,i} = w_{sel,i} = 1/p_i$, and does not explicitly consider nonresponse adjustment or post-stratification (Section 2.7). In fact, when the analysis weight incorporates all three of these conventional weight factors, the variance estimator in Equation 5.2 does not fully reflect the stochastic sample-to-sample variability associated with the nonresponse mechanism, nor does it capture true gains in precision that may have been achieved through the poststratification of the weights to external population controls. Because survey nonresponse is a stochastic process that operates on the selected sample, the variance estimator could (in theory) explicitly capture this added component of sample-to-sample variability (Valliant, 2004). This method assumes that the data user can access the individual components of the survey weight. Stata does provide the capability to directly account for

the reduction in sampling variance due to the poststratification using the `poststrata()` and `postweight()` options on the `svyset` command.

The effect of nonresponse and poststratification weighting on the sampling variance of estimated population totals and other descriptive statistics may also be captured through the use of replicate weights, in which the nonresponse adjustment and the poststratification controls are separately developed for each balanced repeated replication (BRR) or jackknife repeated replication (JRR) replicate sample of cases.

Alternatively, if y is a continuous measure of an attribute of the sample case (e.g., acres of corn, monthly income, annual medical expenses), the result is an estimate of the population total of y,

$$\hat{Y}_w = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} = \hat{Y}$$

Example 5.3 will illustrate the estimation of a subpopulation total, and Example 5.4 will illustrate the estimation of a population total.

Example 5.3: Using the NCS-R Data to Estimate the Total Count of U.S. Adults with Lifetime Major Depressive Episodes (MDE)

The MDE variable in the NCS-R data set is a binary indicator (1 = yes, 0 = no) of whether an NCS-R respondent reported a major depressive episode at any point in his or her lifetime. The aim of this example analysis is to estimate the total number of individuals who have experienced a lifetime major depressive episode along with the standard error of the estimate (and the 95% confidence interval).

For this analysis, the NCS-R survey weight variable NCSRWTSH is selected to analyze all respondents completing the Part I survey (n = 9,282), where the lifetime diagnosis of MDE was assessed. Because the NCS-R data producers normalized the values of the NCSRWTSH variable so that the weights would sum to the sample size, the weight values must be expanded back to the population scale to obtain an unbiased estimate of the population total. This is accomplished by multiplying the Part I weight for each case by the ratio of the NCS-R survey population total (N = 209,128,094 U.S. adults age 18+) divided by the count of sample observations (n = 9,282). The SECLUSTR variable contains the codes representing NCS-R sampling error clusters while SESTRAT is the sampling error stratum variable:

```
gen ncsrwtsh_pop = ncsrwtsh * (209128094 / 9282)
svyset seclustr [pweight = ncsrwtsh_pop], strata(sestrat)
```

Once the complex design features of the NCS-R sample have been identified using the `svyset` command, the `svy: total` command is issued to obtain an

unbiased weighted estimate of the population total along with a standard error for the estimate. The Stata `estat effects` command is then used to compute an estimate of the design effect for this estimated total:

```
svy: total mde
estat effects
```

n	df	$\hat{Y}_w$	$se(\hat{Y}_w)$	$CI_{.95}(\hat{Y}_w)$	$d^2(\hat{Y}_w)$
9,282	42	40,092,206	2,567,488	(34,900,000, 45,300,000)	9.03

The resulting Stata output indicates that 9,282 observations have been analyzed and that there are 42 design-based degrees of freedom. The weighted estimate of the total population of U.S. adults who have experienced an episode of major depression in their lifetime is $\hat{Y} = 40,092,206$. The estimated value of the design effect for the weighted estimate of the population total is $d^2(\hat{Y}_w) = 9.03$, suggesting that the NCS-R variance of the estimated total is approximately nine times greater than that expected for a simple random sample of the same size.

Weighted estimates of population totals can also be computed for subpopulations. Consider subpopulations of NCS-R adults classified by marital status (married, separated/widowed/divorced, and never married). Under the complex NCS-R sample design, correct unconditional subpopulation analyses (see Section 4.5.2) can be specified in Stata by adding the `over()` option to the `svy: total` command:

```
svy: total mde, over(mar3cat)
estat effects
```

Subpopulation	n	Estimated Total Lifetime MDE	Standard Error	95% Confidence Interval	$d^2(\hat{Y})$
Married	5322	20,304,190	1,584,109	(17,100,000, 23,500,000)	6.07
Sep./Wid./Div.	2017	10,360,671	702,601	(8,942,723, 11,800,000)	2.22
Never Married	1943	9,427,345	773,137	(7,867,091, 11,000,000)	2.95

Note that the MAR3CAT variable is included in parentheses to request estimates for each subpopulation defined by the levels of that variable.

Example 5.4: Using the HRS Data to Estimate Total Household Assets

Next, consider the example problem of estimating the total value of household assets for the HRS target population (U.S. households with adults born prior to 1954). We first identify the HRS variables containing the sampling error computation units, or ultimate clusters (SECU) and the sampling error stratum codes (STRATUM). We also specify the KWGTHH variable as the survey weight variable for the analysis, because we are performing an analysis at the level of the HRS household financial unit. The HRS data set includes an indicator variable (KFINR for 2006) that identifies the individual respondent who is the financial reporter for each HRS sample household. This variable is used to create a subpopulation indicator (FINR) that restricts the estimation to only sample members who are financial

reporters for their HRS household unit. We then apply the `svy: total` command to the H8ATOTA variable, measuring the total value of household assets:

```
gen finr=1
replace finr=0 if kfinr !=1
svyset secu [pweight=kwgthh], strata(stratum)
svy, subpop(finr): total h8atota
```

n	df	$\hat{Y}_w$	$se(\hat{Y}_w)$	$CI_{.95}(\hat{Y}_w)$
11,942	56	$\$2.84 \times 10^{13}$	$\$1.60 \times 10^{12}$	$(2.52 \times 10^{13}, 3.16 \times 10^{13})$

The Stata output indicates that the 2006 HRS target population includes approximately 53,853,000 households (not shown). In 2006, these 53.9 million estimated households owned household assets valued at an estimated $\hat{Y}_w = \$2.84 \times 10^{13}$, with a 95% confidence interval (CI) of ($\$2.52 \times 10^{13}$, $\$3.16 \times 10^{13}$).

5.3.3 Means of Continuous, Binary, or Interval Scale Data

The estimation of the population mean, $\bar{Y}$, for a continuous, binary or interval scale variable is a very common task for survey researchers seeking to describe the central tendencies of these variables in populations of interest. An estimator of the population mean, $\bar{Y}$, can be written as a nonlinear ratio of two estimated finite population totals:

$$\bar{y}_w = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}} = \frac{\hat{Y}}{\hat{N}} \tag{5.6}$$

Note that if y is a binary variable coded 1 or 0, the weighted mean estimates the population proportion or prevalence, P, of "1s" in the population (see Chapter 6):

$$\bar{y}_w = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}} = \frac{\hat{M}}{\hat{N}} = p \tag{5.7}$$

Since $\bar{y}_w$ is not a linear statistic, a closed form formula for the variance of this estimator does not exist. As a result, variance estimation methods such as TSL, BRR, or JRR (see Chapter 3) must be used to compute an estimate of

the sampling variance of this estimator of the population mean. An application of TSL to the weighted mean in Equation 5.6 results in the following TSL variance estimator for the ratio mean:

$$\text{var}(\bar{y}_w) \doteq \frac{Var(\hat{Y}) + \bar{y}_w^2 \times Var(\hat{N}) - 2 \times \bar{y}_w \times Cov(\hat{Y}, \hat{N})}{\hat{N}^2} \tag{5.8}$$

Closed-form formulas are available to estimate the variances and covariances of the estimators of the population totals included in this TSL expression (see Section 3.6.2), and BRR and JRR replicated variance estimation methods can also be used to estimate the standard errors of the estimated means (see Section 3.6.3).

Example 5.5: Estimating the Mean Value of Household Income using the NCS-R Data

In this example, we analyze the mean of the household income (HHINC) variable, which was collected in Part II of the NCS-R survey. Because this measure was collected in the second part of the NCS-R survey, we apply the survey weight variable specifically computed by NCS-R staff for the Part II respondents (NCSRWTLG):

```
svyset seclustr [pweight=ncsrwtlg], strata(sestrat)
```

The `svy: mean command` is then submitted in Stata to compute the weighted estimate of the mean household income in the NCS-R population:

```
svy: mean hhinc
estat effects
```

n	df	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{.95}(\bar{y}_w)$	$d^2(\bar{y}_w)$
5692	42	$59,277	$1,596	($56,055, $62,498)	6.09

We see that the estimated mean of total household income is $59,277, with an associated 95% confidence interval of ($56,055, $62,498). As observed in the examples estimating population totals, the design effect is fairly large when estimating a descriptive population parameter for the entire target population. The estimated sampling variance of the estimated mean is about six times as large as it would be for a simple random sample of the same size.

Example 5.6: Estimating Mean Systolic Blood Pressure Using the NHANES Data

In this example, the 2005–2006 NHANES data are used to estimate the mean systolic blood pressure (mmHg) for the U.S. adult population age 18 and older. The survey weight is the NHANES medical examination weight (WTMEC2YR). The following Stata command sequence generates the subpopulation indicator

(AGE18P) and computes the weighted estimate of the mean systolic blood pressure, the linearized estimate of the standard error of the weighted mean (based on an appropriate unconditional subclass analysis), a design-based 95% confidence interval for the mean, and a design effect for the estimate:

```
gen age18p = 1 if age >= 18 and age != .
replace age18p = 0 if age < 18
svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra)
svy, subpop(age18p): mean bpxsy1
estat effects
```

n	df	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{.95}(\bar{y}_w)$	$d^2(\bar{y}_w)$
4,615	15	123.11	0.54	(121.96, 124.27)	5.79

Based on the sample of n = 4,615 individual observations, the 2005–2006 NHANES estimate of mean systolic blood pressure for U.S. adults age 18 and older is $\bar{y}_w$ = 123.11 mmHg.

Example 5.7: Estimating the Mean Value of Total Household Assets Using the HRS Data

Next, the 2006 HRS data are used to compute an estimate of the mean of total household assets for U.S. households with at least one adult born prior to 1954. The HRS constructed variable that measures total assets for HRS household financial units is H8ATOTA. As in Example 5.4, the estimate is restricted to the subpopulation of HRS respondents who are the financial reporters for their household unit. The Stata commands required to perform this analysis are as follows:

```
svyset secu [pweight=kwgthh], strata(stratum)
svy, subpop(finr): mean h8atota
estat effects
```

n	df	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{.95}(\bar{y}_w)$	$d^2(\bar{y}_w)$
11,942	56	$527,313	$28,012	($471,196, $583,429)	1.52

Based on this analysis of reports from n = 11,942 HRS financial units, the mean of 2006 household assets in the HRS survey population is $\bar{y}_w$ = $527,313. The estimated design effect for the estimated mean, $d^2(\bar{y}_w)$ =1.52, indicates that while there is still a loss of precision in the HRS estimate of mean household assets relative to simple random sampling, the loss in effective sample size is not as severe as in the earlier examples based on the NCS-R and the NHANES.

5.3.4 Standard Deviations of Continuous Variables

Although experience suggests that it is not a common task, survey analysts may wish to compute an unbiased estimate of the population standard

deviation of a continuous variable. Just as weights are required to obtain unbiased (or nearly unbiased) estimates of the population mean, weights must also be employed to obtain consistent estimates of the standard deviation of a random variable in a designated target population. A weighted estimator of the population standard deviation of a variable y can be written as follows:

$$s_y = \sqrt{\frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_i (y_i - \bar{y}_w)^2}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_i - 1}} \tag{5.9}$$

In Equation 5.9, the estimate of the population mean for the variable y is calculated as in Equation 5.6. Stata currently does not include an explicit command for estimation of these population standard deviations. Users of the SAS software can use PROC UNIVARIATE (with the VARDEF-DF option) or PROC MEANS with a WEIGHT statement to generate this weighted estimate of the population standard deviation S_y.

5.3.5 Estimation of Percentiles and Medians of Population Distributions

Estimation of quantiles, such as the median (Q_{50}) or the 95th percentile (Q_{95}) of the population distribution of a continuous variable, can play an important role in analyses of survey data. A sociologist may wish to compare sample estimates of percentiles of household income for a regional survey population to nationally defined poverty criteria. An epidemiologist may wish to estimate the 95th percentile of prostate-specific antigen (PSA) levels in a metropolitan sample of men over the age of 40.

The **ungrouped method** of quantile estimation (Loomis, Richardson, and Elliott, 2005) builds on results related to the weighted estimator for totals presented earlier in this chapter (see Section 5.3.2), employing a weighted sample estimate of the population cumulative distribution function (CDF) of a survey variable. Specifically, the CDF for a variable y in a given population of size N is defined as follows:

$$F(x) = \frac{\sum_{i=1}^{N} I(y_i \le x)}{N} \tag{5.10}$$

In Equation 5.10, $I(y_i \le x)$ is an indicator variable equal to 1 if y is less than or equal to a specified value of x, and 0 otherwise. The weighted estimator of the CDF from a complex sample of size n from this population is then written as follows:

$$\hat{F}(x) = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I(y_{h\alpha i} \le x)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}} \tag{5.11}$$

The q-th quantile (e.g., $q = 0, 0.25, 0.50, 0.75, 1$) for a variable y is the smallest value of y such that the population CDF is greater than or equal to q. For example, the median would be the smallest value of y at which the CDF is greater than or equal to 0.5. The ungrouped method of estimating a quantile first considers the order statistics (the sample values of y ordered from smallest to largest), denoted by $x_1, \ldots, x_n$, and finds the value of j ($j = 1, \ldots, n$) such that

$$\hat{F}(x_j) \le q < \hat{F}(x_{j+1}) \tag{5.12}$$

Then, the estimate of the q-th population quantile X_q is calculated as follows:

$$\hat{X}_q = x_j + \frac{q - \hat{F}(x_j)}{\hat{F}(x_{j+1}) - \hat{F}(x_j)}(x_{j+1} - x_j) \tag{5.13}$$

Kovar, Rao, and Wu (1988) report results of a simulation study suggesting that BRR performs well for variance estimation and construction of confidence intervals when working with estimators of nonsmooth functions like quantiles. The WesVar PC software currently implements the BRR variance estimation approach for quantiles, and we recommend the use of this variance estimation approach for estimated quantiles in practice. Variance estimates for the estimated quantile in Equation 5.13 can also be computed using Taylor series linearization (Binder, 1991). The SUDAAN software currently uses the linearized variance estimator. The JRR approach to variance estimation is known to be badly biased for these types of estimators (Miller, 1974), but modifications to the jackknife approach addressing this problem have been developed (Shao and Wu, 1989).

Example 5.8: Estimating Population Quantiles for Total Household Assets Using the HRS Data

This example considers the total household assets variable collected from the 2006 HRS sample and aims to estimate the 0.25 quantile, the median, and the 0.75 quantile of household assets in the HRS target population. The SUDAAN software is used in this example because Stata (Version 10) does not currently support procedures specifically dedicated to the estimation of quantiles (and their standard errors) in complex sample survey data sets. The following SUDAAN code

TABLE 5.1

Estimation of Percentiles of the Distribution of 2006 HRS Total Household Assets

	SUDAAN (TSL)		WesVar PC (BRR)	
Percentile	$\hat{Q}_p$	$se(\hat{Q}_p)$	$\hat{Q}_p$	$se(\hat{Q}_p)$
Q_{25}	$39,852	$3,167	$39,852	$3,249
Q_{50} (Median)	$183,309	$10,233	$183,309	$9,978
Q_{75}	$495,931	$17,993	$495,931	$17,460

generates the quantile estimates and standard errors, using an unconditional subclass analysis approach:

```
proc descript ;
nest stratum secu ;
weight kwgthh ;
subpopn finr = 1 ;
var h8atota ;
percentiles 25 75 / median ;
setenv decwidth = 1 ;
run ;
```

Table 5.1 summarizes the results provided in the SUDAAN output and compares the estimates with those generated using the BRR approach to variance estimation in the WesVar PC software. The estimated median of the total household assets for the HRS target population is $183,309. The estimate of the mean total household assets from Example 5.7 was $527,313, suggesting that the distribution of total household assets is highly skewed to the higher dollar value ranges.

The analysis of quantiles of the distribution of total household assets for HRS households was repeated using the WesVar PC software (readers are referred to the ASDA Web site for the menu steps needed to perform this analysis in WesVar). WesVar allows for the use of BRR to estimate the standard errors of estimated quantiles. From the side-by-side comparison in Table 5.1, WesVar's weighted estimates of the quantiles agree exactly with those reported by SUDAAN. However, WesVar's BRR estimates of the corresponding standard errors differ slightly from the TSL standard errors computed by SUDAAN, as expected. The resulting inferences about the population quantiles would not differ substantially in this example as a result.

The DESCRIPT procedure in the SUDAAN software can also be used to estimate quantiles for subpopulations of interest. For example, the same quantiles could be estimated for the subpopulation of adults age 75 and older in the HRS population using the following syntax:

```
proc descript ;
nest stratum secu ;
weight kwgthh ;
subpopn kage > 74 & finr = 1 ;
```

```
var h8atota ;
percentiles 25 75 / median ;
setenv decwidth = 1 ;
run ;
```

Note the use of the SUBPOPN statement to identify that the estimate is based on respondents who are 75 years of age and older and are the financial reporter for their HRS household unit. The estimated quantile values and standard errors generated by this subpopulation analysis are $\hat{Q}_{25,75+}$ = \$40,329.4 (\$4,434.8); $\hat{Q}_{50,75+}$ = \$177,781.3 (\$11,142.9); and $\hat{Q}_{75,75+}$ = \$461,308.3 (\$27,478.0).

5.4 Bivariate Relationships between Two Continuous Variables

Four basic analytic approaches can be used to examine bivariate relationships between two continuous survey variables: (1) generation of a scatterplot; (2) computation of a pair-wise correlation coefficient, *r*; (3) estimation of the ratio of two continuous variables, $\hat{R} = \hat{Y} / \hat{X}$; and (4) estimation of the coefficients of the simple linear regression of one variable on another, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$. The first three of these techniques are reviewed in this section. Chapter 7 will address the estimation of linear regression models for continuous dependent variables in detail.

5.4.1 X–Y Scatterplots

A comparison of the 2005–2006 NHANES MEC measures of high-density lipoprotein (HDL; the *y* variable) and total serum cholesterol (the *x* variable) illustrates how a simple scatterplot can be used to gain insight into the bivariate relationship of two survey variables. Figure 5.4 presents an unweighted scatterplot comparing the two variables. The figure suggests a positive relationship between HDL and total serum cholesterol but also illustrates considerable variability in the relationship and the presence of a few points that stand out as potential outliers. A drawback to the simple X–Y scatterplot summary is that there is no practical way to incorporate the population weights and the corresponding population frequency associated with each point in the two-dimensional X–Y space. Stata does provide a unique display in which the area of the dot representing an X–Y point is proportionate to its survey weight. Unfortunately, in survey data sets with hundreds and even thousands of data points, it is not possible to obtain the resolution needed to evaluate single points or to detect patterns in the weighted scatterplot display. One technique for introducing information on the effect of the weights

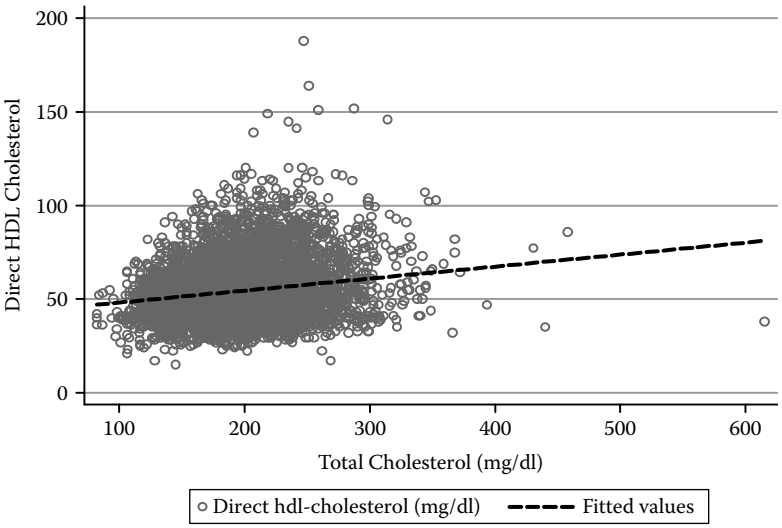


FIGURE 5.4
HDL versus total serum cholesterol (mg/dl) in U.S. adults (unweighted points, with the fit of a weighted regression line included).

in the plotted X–Y relationship is to overlay on the scatterplot the line representing the weighted regression of y on x, $\hat{y}_{wls} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$.

To do this in the Stata software, we use the `twoway` graphing command, where the first command in parentheses defines the scatterplot and the second command overlays the weighted estimate of the simple linear regression model relating total cholesterol to HDL (note that the probability weight variable is defined in square brackets for the second command):

```
twoway (scatter lbdhdd lbxtc if age18p==1) (lfit lbdhdd ///
lbxtc if age18p==1 [pweight=wtmec2yr])
```

5.4.2 Product Moment Correlation Statistic (*r*)

The product moment correlation (r) is a standard measure of the association between two continuous variables. Unfortunately, few current software systems provide the capability to estimate single correlations (or weighted correlation matrices) and confidence intervals for r that account for complex sample design features. A reasonable alternative would be to first standardize the two variables for which a correlation is desired and then compute a weighted estimate of the slope parameter in a simple linear regression model relating the two variables (see Chapter 7).

A weighted estimator of the product moment correlation statistic is written as follows:

$$r_w = \frac{s_{xy,w}}{s_{x,w} \cdot s_{y,w}} =$$

$$\frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}(y_{h\alpha i} - \bar{y}_w)(x_{h\alpha i} - \bar{x}_w)}{\sqrt{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}(y_{h\alpha i} - \bar{y}_w)^2} \cdot \sqrt{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}(x_{h\alpha i} - \bar{x}_w)^2}} \quad (5.14)$$

Kish and Frankel (1974) included the pair-wise correlation statistic in their simulation studies and found that TSL, BRR, and JRR performed similarly in the estimation of standard errors for estimates of pair-wise correlation statistics.

5.4.3 Ratios of Two Continuous Variables

Occasionally, survey analysts need to estimate the ratio of two continuous survey variables (e.g., the ratio of HDL cholesterol level to total cholesterol). The ratio estimator of the population mean (Equation 5.6) of a single variable can be generalized to an estimator of the population ratio of two survey variables:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{\alpha_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{\alpha_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} x_{h\alpha i}} \quad (5.15)$$

The following TSL approximation provides estimates of the sampling variance of ratio estimates:

$$\text{var}(\hat{R}) \doteq \frac{\text{var}(\hat{Y}) + \hat{R}^2 \cdot \text{var}(\hat{X}) - 2 \cdot \hat{R} \cdot \text{cov}(\hat{Y}, \hat{X})}{\hat{X}^2} \quad (5.16)$$

BRR and JRR estimation options now available in the major statistical software packages provide an appropriate alternative to TSL for estimating standard errors of $\hat{R}$.

Example 5.9: Estimating the Population Ratio of High-Density to Total Cholesterol for U.S. Adults

A weighted estimate of this ratio based on 2005–2006 NHANES respondents age 18+ is obtained in Stata using the `svy: ratio` command. Note that we

THEORY BOX 5.2 RATIO ESTIMATORS OF POPULATION TOTALS

If two survey variables x and y are highly correlated and a control total for the variable x, X_{pop}, is known from an auxiliary data source, then the ratio estimator of the population total, $\hat{Y}_R = \hat{R} \cdot X_{pop}$, may provide a more precise estimate than the simple weighted estimator of the population total, $\hat{Y}_w$, described in Section 5.3.2. Because X_{pop} is assumed to be free of sampling variability, the standard error of the ratio estimator of the population total is: $se(\hat{Y}_R) = se(\hat{R}) \cdot X_{pop}$.

once again define the subpopulation indicator (AGE18P) first before running the analysis.

```
gen age18p = 1 if age >= 18 and age != .
replace age18p = 0 if age < 18
svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra)
svy, subpop(age18p): ratio (lbdhdd/lbxtc)
```

n	df	$\hat{R}$	$se(\hat{R})$	$CI_{.95}(\hat{R})$
4996	15	0.275	0.002	(0.271,0.280)

Note that the two variables defining the numerator and denominator of the ratio are indicated in parentheses in the `svy: ratio` command, with the numerator variable listed before a forward slash (/) and the denominator variable listed after the slash.

5.5 Descriptive Statistics for Subpopulations

Section 4.5.3 discussed the important analytical differences between *conditional* and *unconditional* subpopulation analyses. Section 5.3.2 presented examples of unconditional subpopulation analyses in the estimation of totals. Stata's `svy` commands provide two options for correctly specifying subpopulation analyses. The `over()` option requests unconditional subclass analyses for subpopulations defined based on *all* levels of a categorical variable. This option is only available in Stata's descriptive analysis procedures. The more general `subpop()` option can be used with all `svy` commands in Stata. This option requests analysis for a specific subpopulation identified by an indicator variable, coded 1 if the case belongs to the subpopulation of interest and 0 otherwise. Procedures for survey data analysis in other software packages will have similar command options available for these subpopulation analyses (see the book Web site for examples).

Example 5.10: Estimating the Proportions of Males and Females Age > 70 with Diabetes Using the HRS Data

This first subpopulation analysis example uses the 2006 HRS data set to estimate the prevalence of diabetes for U.S. males and females age 70 years and older. Because the HRS variable DIABETES is equal to 1 for respondents with diabetes and 0 otherwise, estimating the mean of this binary variable will result in the prevalence estimates of interest. The following example command uses a logical condition in the `subpop()` option to specify the age range and then the `over()` option to perform separate subpopulation analyses for males and females:

```
svy: mean diabetes, subpop(if kage > 70) over(gender)
```

Gender	df	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{95}(\bar{y}_w)$
Male	56	0.235	0.008	(0.219,0.252)
Female	56	0.184	0.009	(0.167,0.201)

We see that more elderly men (23.5%) are estimated to have diabetes compared with elderly women (18.4%).

It might be tempting for an analyst to make the mistake of taking a *conditional* approach to these subpopulation analyses, using a Stata command like the following:

```
svy: mean diabetes if age > 70
```

Use of the subsetting `if` modifier to define a subpopulation for stratified samples is inappropriate because all cases not satisfying the condition are temporarily deleted from the analysis, along with their sample design information. This effectively fixes the strata sample sizes for variance estimation calculations (when in fact the strata subpopulation sample sizes should be treated as random variables). See Chapter 4 for more details.

Example 5.11: Estimating Mean Systolic Blood Pressure for Males and Females Age > 45 Using the NHANES Data

Consider an example based on NHANES of estimating the mean systolic blood pressure for male and female adults over the age of 45. We once again illustrate a combination of options in Stata to set up this analysis, using a logical condition in the `subpop()` option to specify the age range and then the `over()` option to perform subpopulation analyses for males and females in this age range separately. We also request design effects for each subpopulation estimate using the `estat effects` postestimation command:

```
svy, subpop(if age > 45): mean bpxsy1, over(gender)
estat effects
```

Gender	*df*	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{95}(\bar{y}_w)$	$d^2(\bar{y}_w)$
Male	15	128.96	0.76	(127.35,130.58)	2.60
Female	15	132.09	1.06	(129.82,134.36)	3.62

Survey analysts should be aware that restricting analysis to a subpopulation of the full sample may result in a reduction in the effective degrees of freedom for variance estimation. Since Stata employs the variable degrees of freedom method discussed in Section 3.5.2, its programs ignore any original design strata that do not contain one or more observations from the subpopulation of interest. Stata signals that complete strata have been dropped by including a note in the output indicating that one or more design strata were "omitted because they contain no subpopulation members." Approaches to this issue are not currently uniform across the different major statistical software packages.

The greatest reductions in effective degrees of freedom for variance estimation can occur when survey analysts are interested in estimation for rare subpopulations that comprise only a small percent of the survey population or when the subpopulation of interest is defined by a domain of sample strata, such as a single census region. (Refer to Figure 4.4 for an illustration of how subpopulations may distribute across the strata and clusters of a complex sample design.)

For example, the following Stata `svy: mean` command requests estimates of mean systolic blood pressure for four education groupings of African Americans age 80 and older:

```
svy: mean bpxsy1, subpop(if age >80 & black==1) over(edcat)
```

Stata reports (results not shown) that for one or more of these detailed subpopulation estimates, only 12 of the 15 design strata and 24 of the 30 clusters in the 2005–2006 NHANES design contain one or more eligible respondents from the target subpopulation. Consequently, 12 degrees of freedom are assumed in developing confidence intervals or evaluating test statistics from this analysis.

Procedures for subpopulation analyses focused on estimation of percentiles/quantiles are currently available in the SUDAAN, WesVar PC, and SAS (Version 9.2 and higher) software (see Example 5.8). Examples of subpopulation analyses using these other software systems are available on the book Web site.

5.6 Linear Functions of Descriptive Estimates and Differences of Means

The capability to estimate *functions* of descriptive statistics, especially differences of means or proportions, is an important feature of many survey analyses. In general, many important functions of descriptive statistics can be written as **linear combinations** of the descriptive statistics of interest. Examples of such

linear combinations of estimates include differences of means, weighted sums of means used to build economic indices, or computation of a moving average of three monthly survey means, such as the following:

$$\begin{aligned}\hat{\Delta} &= \bar{y}_1 - \bar{y}_2 \\ \hat{I} &= .25 \cdot \bar{y}_1 + .40 \cdot \bar{y}_2 + 1.5 \cdot \bar{y}_3 + 1.0 \cdot \bar{y}_4 \\ \bar{y}_{moving} &= 1/3 \cdot \bar{y}_{t1} + 1/3 \cdot \bar{y}_{t2} + 1/3 \cdot \bar{y}_{t3}\end{aligned}$$

Consider the general form of a linear combination of $j = 1, \ldots, J$ descriptive statistics (e.g., estimates of means for J subpopulations):

$$f(\theta_1, ..., \theta_J) = \sum_{j=1}^{J} a_j \theta_j \tag{5.17}$$

In Equation 5.17, θ_j represents the statistic of interest for the j-th subpopulation, while the a_j terms represent constants defined by the analyst. This function is estimated by simply substituting estimates of the descriptive statistics into the expression

$$f(\hat{\theta}_1, ..., \hat{\theta}_J) = \sum_{j=1}^{J} a_j \hat{\theta}_j \tag{5.18}$$

The variance of this estimator would then be calculated as follows:

$$var\left(\sum_{j=1}^{J} a_j \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 var(\hat{\theta}_j) + 2 \times \sum_{j=1}^{J-1} \sum_{k>j}^{J} a_j a_k cov(\hat{\theta}_j, \hat{\theta}_k) \tag{5.19}$$

Note that the variance of the linear combination incorporates the variances of the individual component estimates as well as possible covariances of the estimated statistics.

Covariance between descriptive estimates can arise in several ways. First, the statistics of interest may represent overlapping subpopulations (see Kish, 1987). An example would be a contrast comparing the mean systolic blood pressure for men to that for the total population, such as $\hat{\Delta} = \bar{y}_{total} - \bar{y}_{male}$, or a longitudinal analysis in which the mean blood pressures of a sample panel of individuals are compared at two points in time, $\hat{\Delta} = \bar{y}_{time2} - \bar{y}_{time1}$. Due to the intraclass correlation among the sample elements within sample design clusters of complex sample designs, a degree

of covariance is possible even when estimates are based on nonoverlapping samples, such as $\hat{\Delta} = \bar{y}_{female} - \bar{y}_{male}$.

Under conditions where samples are overlapping or the complex design itself induces a covariance in the sample estimates, statistical software must compute and store the covariances of estimates to correctly compute the variance of a linear combination of means or other descriptive statistics. Stata's `lincom` command is one example of a postestimation command that enables analysts to correctly compute estimates and standard errors for linear combinations of estimates. SUDAAN's `contrast` option is another.

5.6.1 Differences of Means for Two Subpopulations

Analysts of survey data are frequently interested in making inferences about differences of descriptive statistics for two subpopulations. The inference can be based on a 95% confidence interval for the estimated difference of the two means, such as

$$CI = (\bar{y}_{female} - \bar{y}_{male}) \pm t_{df,.975} \cdot se(\bar{y}_{female} - \bar{y}_{male}) \tag{5.20}$$

or can employ a **two-sample *t*-test** based on the Student t statistic,

$$t = (\bar{y}_{female} - \bar{y}_{male}) / se(\bar{y}_{female} - \bar{y}_{male}) \tag{5.21}$$

Applying the general formula for the variance of a linear combination (5.19), the standard error of the difference in the mean of the two subpopulations samples can be expressed as

$$se(\bar{y}_1 - \bar{y}_2) = \sqrt{var(\bar{y}_1) + var(\bar{y}_2) - 2cov(\bar{y}_1, \bar{y}_2)} \tag{5.22}$$

Under simple random sampling, the covariance of estimates for distinct samples is zero; however, for clustered samples or samples that share elements, the covariance of the two sample means may be nonzero and generally positive in value. Example 5.12 estimates the contrast in the mean value of total household assets for two subpopulations of HRS households: one subpopulation in which the household head has less than a high school education and a second subpopulation of households headed by a college-educated adult.

Example 5.12: Estimating Differences in Mean Total Household Assets between HRS Subpopulations Defined by Educational Attainment Level

Testing differences of subpopulation means using the svy: mean procedure in Stata is a two-step process. First, the mean of total household assets is estimated

for subpopulations defined by the levels of the EDCAT variable in the HRS data set:

```
gen finr = 1
replace finr = 0 if kfinr != 1
svyset secu [pweight = kwgthh], strata(stratum) ///
vce(linearized) singleunit(missing)
svy, subpop(finr): mean h8atota, over(edcat)
```

Education of Head	Stata Label	$\bar{y}_w$	$se(\bar{y}_w)$	$CI_{.95}(\bar{y}_w)$
0–11 yrs	1	$178,386	$24,561	($129,184 , $227,588)
12 yrs	2	$328,392	$17,082	($294,171, $362,613)
13–15 yrs	3	$455,457	$27,000	($401,369, $509,545)
16+ yrs	4	$1,107,204	$102,113	($902,646, $1,311,762)

Stata automatically saves these estimated means along with their sampling variances and covariances in memory until the next command is issued. Stata assigns the four subpopulations of household heads with 0–11, 12, 13–15, and 16+ years of education to internal reference labels 1, 2, 3, and 4, respectively.

Following the computation of the subpopulation means, the `lincom` postestimation command is used to estimate the difference in means for household heads with 0–11 (subpopulation 1) versus 16+ years of education (subpopulation 4):

```
lincom [h8atota]1 - [h8atota]4
```

Education of Head	$\bar{y}_{0\text{-}11} - \bar{y}_{16+}$	$se(\bar{y}_{0-11} - \bar{y}_{16+})$	$CI_{.95}(\bar{y}_{0-11} - \bar{y}_{16+})$
0–11 vs. 16+	–$928,818	$108,250	(–$1,145,669, –$711,967)

After this postestimation command is submitted, Stata outputs the estimated difference of the two subpopulation means ($\hat{\Delta} = \bar{y}_{0-11} - \bar{y}_{16+} = -\$928{,}818$). The 95% confidence interval for the population difference does not include 0, suggesting that households headed by a college graduate have a significantly higher mean of total household assets compared with households in which the head does not have a high school diploma.

To display the estimated variance–covariance matrix for the subpopulation estimates of mean household assets, the Stata postestimation command `vce` is used:

```
vce
```

The output produced by this command is shown in the symmetric 4 × 4 matrix in Table 5.2. Note that in this example, the covariance of the estimated means for the 0–11 and 16+ household heads is small and negative (-3.438×10^8).

When applying the formula for the standard error of the contrast we get Table 5.2.

TABLE 5.2

Estimated Variance–Covariance Matrix for Subpopulation Means of Total Household Assets

Subpopulation	1	2	3	4
1	6.032×10^8	0.714×10^8	-1.794×10^8	-3.438×10^8
2		2.918×10^8	-0.126×10^8	1.209×10^8
3			7.290×10^8	0.679×10^8
4				1.043×10^{10}

Source: Based on the 2006 HRS data.
Note: Estimates by education level of household head.

$$\begin{aligned} se(\bar{y}_{0-11} - \bar{y}_{16+}) &= \sqrt{\text{var}(\bar{y}_{0-11}) + \text{var}(\bar{y}_{16+}) - 2 \cdot \text{cov}(\bar{y}_{0-11}, \bar{y}_{16+})} \\ &= \sqrt{\left[6.032 + 104.3 - 2 \cdot (-3.438)\right] \times 10^8} \\ &\doteq \$108,250 \end{aligned}$$

The result of this direct calculation matches the output from the `lincom` command in Stata.

5.6.2 Comparing Means over Time

Analysts working with longitudinal survey data are often interested in comparing the means of a longitudinal series of survey measures. Chapter 12 will address the various types of longitudinal data and introduce more sophisticated tools for longitudinal analysis of survey data.

When comparing means on single variables measured at two or more "waves" or points in time, an approach similar to that used in Section 5.6.1 can be applied. However, since longitudinal data are often released as a separate file for each time period with distinct weight values for each time point, special data preparation steps may be needed. We now consider an example of this approach using two waves of data from the HRS study.

Example 5.13: Estimating Differences in Mean Total Household Assets from 2004 to 2006 Using Data from the HRS

To estimate the difference in 2004 and 2006 mean household assets for HRS panel households, the first step is to "stack" the 2004 and 2006 data sets, combining them into a single data set. Provided that they responded to the survey in both HRS waves, each panel household has two records in the stacked data set—one for its 2004 observation and another for the 2006 interview. Each pair of household records includes the wave-specific measure of total household assets and the wave-specific sampling weight. For this example, when stacking the data sets, we assigned these values to two new common variables, TOTASSETS

and WEIGHT. Each record also includes the permanently assigned stratum and cluster codes. As described in Example 5.4, the HRS public use data sets include an indicator variable for each wave of data that identifies the respondents who are the household financial reporters (JFINR for 2004; KFINR for 2006). Using the `over(year)` option, estimates of mean total household assets are computed separately for 2004 and 2006. The `subpop(finr0406)` option restricts the estimates to the financial reporters for each of these two data collection years. The postestimation `lincom` statement estimates the difference of means for the two time periods, its linearized standard error, and a 95% confidence interval for the difference:

```
gen weight = jwgthh
replace weight = kwgthh if year == 2006
gen finr04 = 1 if (year==2004 & jfinr==1)
gen finr06 = 1 if (year==2006 & kfinr==1)
gen finr0406 = 1 if finr04==1 | finr06==1
svyset secu [pweight = weight], strata(stratum)
svy, subpop(finr0406): mean totassets, over(year)
lincom [totassets]2004 - [totassets]2006
```

Contrast	$\bar{y}_{2004} - \bar{y}_{2006}$	$se(\bar{y}_{2004} - \bar{y}_{2006})$	$CI_{.95}(\bar{y}_{2004} - \bar{y}_{2006})$
2004 vs. 2006	–$115,526	$20,025	(–$155,642, –$75,411)

Note that the `svyset` command has been used again to specify the recoded sampling weight variable (WEIGHT) in the stacked data set. The `svy: mean` command is then used to request weighted estimates and linearized standard errors (and the covariances of the estimates, which are saved internally) for each subpopulation defined by the YEAR variable. The resulting estimate of the difference of means is $\hat{\Delta} = \bar{y}_{2004} - \bar{y}_{2006} = -\$115{,}526$, with a linearized standard error of $20,025. The analysis provides evidence that the mean total household assets increased significantly from 2004 to 2006 for this population.

5.7 Exercises

1. This exercise serves to illustrate the effects of complex sample designs (i.e., design effects) on the variances of estimated means, due to stratification and clustering in sample selection (see Section 2.6.1 for a review). The following table lists values for an equal probability (self-weighting) sample of $n = 36$ observations.

Observations			**STRATUM**	**CLUSTER**
7.0685	13.7441	7.2293	1	1
13.6760	7.2293	13.7315	1	2

Observations			STRATUM	CLUSTER
13.2310	10.8922	12.3425	1	3
10.9647	11.2793	11.8507	1	4
11.3274	16.4423	11.9133	1	5
17.3248	12.1142	16.7290	1	6
19.7091	12.9173	18.3800	2	7
13.6724	16.2839	14.6646	2	8
15.3685	15.3004	13.5876	2	9
15.9246	14.0902	16.4873	2	10
20.2603	12.0955	18.1224	2	11
12.4546	18.4702	14.6783	2	12

Any software procedures can be used to answer the following four questions.

a. Assume the sample of $n = 36$ observations is a simple random sample from the population. Compute the sample mean, the standard error of the mean, and a 95% confidence interval for the population mean. (Ignore the finite population correction [fpc], stratification, and the clustering in calculating the standard error.)

b. Next, assume that the $n = 36$ observations are selected as a stratified random sample of 18 observations from each of two strata. (Ignore the fpc and the apparent clustering.) Assume that the population size of each of the two strata is equal. Compute the sample mean, the standard error of the mean and a 95% confidence interval for the population mean. (Ignore the fpc and the clustering in calculating the standard error.) What is the estimated DEFT $(\bar{y})$ for the standard error of the sample mean (i.e., the square root of the design effect)?

c. Assume now that the $n = 36$ observations are selected as an equal probability sample of 12 clusters with exactly three observations from each cluster. Compute the sample mean, the standard error of the mean, a 95% confidence interval for the population mean, and the estimate of DEFT $(\bar{y})$. Ignore the fpc and the stratification in calculating the standard error. Use the simple model of the design effect for sample means to derive an estimate of roh (2.11), the synthetic intraclass correlation (this may take a negative value for this "synthetic" data set).

d. Finally, assume that the $n = 36$ observations are selected as an equal probability stratified cluster sample of observations. (Two strata, six clusters per stratum, three observations per cluster.) Compute the sample mean, the standard error of the mean, the 95% CI, and estimates of DEFT $(\bar{y})$ and roh.

2. Using the NCS-R data and a statistical software procedure of your choice, compute a weighted estimate of the total number of U.S. adults that has ever been diagnosed with alcohol dependence (ALD) along with a 95% confidence interval for the total. Make sure to incorporate the complex design when computing the estimate and the confidence interval. Compute a second 95% confidence interval using an alternative variance estimation technique, and compare the two resulting confidence intervals. Would your inferences change at all depending on the variance estimation approach?
3. (Requires SUDAAN or SAS Version 9.2+) Using the SUDAAN or SAS software and the 2005–2006 NHANES data set, estimate the 25th percentile, the median, and the 75th percentile of systolic blood pressure (BPXSY1) for U.S. adults over the age of 50. You will need to create a subpopulation indicator of those aged 51 and older for this analysis. Remember to perform an appropriate subpopulation analysis for this population subclass. Compute 95% confidence intervals for each percentile.
4. Download the NCS-R data set from the book Web site and consider the following questions. For this exercise, the SESTRAT variable identifies the stratum codes for computation of sampling errors, the SECLUSTR variable identifies the sampling error computation units, and the NCSRWTSH variable contains the final sampling weights for Part 1 of the survey for each sampled individual.
   a. How many sampling error calculation strata are specified for the NCS-R sampling error calculation model?
   b. How many SECUs (or clusters) are there in total?
   c. How many degrees of freedom for variance estimation does the NCS-R provide?
   d. What is the expected loss, L_w, due to random weighting in survey estimation for total population estimates? *Hint*: $L_w = CV^2(\text{weight})$; see Section 2.7.5.
   e. What is the average cluster size, b, for total sample analyses of variables with no item-missing data?
5. Using the statistical software procedure of your choice, estimate the proportion of persons in the 2006 HRS target population with arthritis (ARTHRITIS = 1). Use Taylor series linearization to estimate a standard error for this proportion. Then, answer the following questions:
   a. What is the design effect for the total sample estimate of the proportion of persons with arthritis (ARTHRITIS = 1)? What is the design effect for the estimated proportion of respondents age 70 and older (AGE70 = 1)? *Hint*: Use the standard variance formula,

var(p) = $p \times (1 - p)/(n - 1)$, to obtain the variance of the proportion p under SRS. Use the weighted estimate of p provided by the software procedure to compute the SRS variance. Don't confuse standard errors and variances (squares of standard errors).

b. Construct a 95% confidence interval for the mean of DIABETES. Based on this confidence interval, would you say the proportion of individuals with diabetes in the 2006 HRS target population is significantly different from 0.25?

6. Examine the CONTENTS listing for the NCS-R data set on the book Web site. Choose a dependent variable of interest (e.g., MDE: 1 = Yes, 0 = No). Develop a one-sample hypothesis (e.g., the prevalence of lifetime major depressive episodes in the U.S. adult population is p = 0.20). Write down your hypothesis before you actually look at the sample estimates. Perform the required analysis using a software procedure of your choosing, computing the weighted sample-based estimate of the population parameter and a 95% confidence interval for the desired parameter. Test your hypothesis using the confidence interval. Write a one-paragraph statement of your hypothesis and a summary of the results of your sample-based estimation and your inference/conclusion based on the 95% CI.

7. Two subclasses are defined for NCS-R respondents based on their response to a question on diagnosis of a major depressive episode (MDE) (1 = Yes, 0 = No). For these two subclasses, use the software procedure of your choice to estimate the difference of means and standard error of the difference for body mass index (BMI). Make sure to use the *unconditional* approach to subclass analysis in this case, given that these subclasses can be thought of as *cross-classes* (see Section 4.5). Use the output from this analysis to replicate the following summary table.

Subclass	**Variable**	$\bar{y}$,[$se(\bar{y})$]
MDE = 1 (Yes)	BMI	27.59 (.131)
MDE = 0 (No)	BMI	26.89 (.102)
Difference in Means (1–0)	BMI	.693 (.103)

6

Categorical Data Analysis

> If the same group of individuals is classified in two or more different ways, as persons may be classified as inoculated and not inoculated, and also may be attacked and not attacked by disease, then we may require to know if the two classifications are independent.
>
> **—R. A. Fisher (1925)**

6.1 Introduction

A casual perusal of the codebook and variable descriptions for most public-use survey data sets quickly leads to the observation that the responses to the majority of survey questions in the social sciences and public health and related fields of research are measured as a binary choice (e.g., yes, no), a selection from multinomial response categories (e.g., ethnicity), a choice from an ordinal scale (e.g., strongly agree to strongly disagree), or possibly a discrete count of events. This chapter covers procedures for simple univariate, bivariate, and selected multivariate analyses for such categorical survey responses, focusing on the adaptation of established analytic techniques to complex sample survey data. For readers interested in a fuller discussion of these basic techniques of categorical data analysis, we recommend Agresti (2002).

The outline of this chapter parallels that of Chapter 5. Section 6.2 highlights several important considerations in categorical data analysis for complex sample surveys. Basic methods for analyzing a single categorical variable are described in Section 6.3, including estimation of category proportions, goodness-of-fit (GOF) tests to compare the sample estimates with hypothesized population values, and graphical display of the estimated population distribution across the K categories of the single variable. Section 6.4 extends the discussion to bivariate analyses including statistics that measure association and tests of hypotheses concerning independence of two categorical variables. The chapter concludes in Section 6.5 with coverage of two techniques for multivariate categorical data: the Cochran–Mantel–Haenszel (CMH) test, and a brief discussion of simple log-linear models for cross-tabulated data. Multivariate regression modeling and related methods for categorical data will be introduced later in Chapters 8 and 9.

6.2 A Framework for Analysis of Categorical Survey Data

We begin this chapter by introducing a framework for the analysis of categorical data collected in complex sample surveys. We introduce methods for accommodating complex sample design features and important considerations for both univariate and bivariate analyses of categorical data. Sections 6.3 and 6.4 go into more detail about these analysis approaches.

6.2.1 Incorporating the Complex Design and Pseudo-Maximum Likelihood

The simplicity of categorical survey responses can belie the range of sophistication of categorical data analysis techniques—techniques that range from the simplest estimation of category proportions to complex multivariate and even multilevel regression models. The majority of the standard estimators and test statistics for categorical data analysis are derived under the method of **maximum likelihood** and assume that the data are independent and identically distributed (i.i.d.) according to a discrete probability distribution. Under simple random sampling assumptions, categorical variables are assumed to follow one of several known probability distributions or **sampling models** (Agresti, 2002)—that is, the binomial, the multinomial, the Poisson, the product-multinomial, or, more rarely, a hypergeometric model. Unfortunately, due to sample weighting, clustering, and stratification, the true likelihood of the sample survey data is generally difficult to specify analytically. Therefore, the simple elegance of maximum likelihood methods for estimation and inference does not easily transfer to complex sample survey data. This chapter will introduce the methods and software that survey statisticians have developed to adjust standard analyses for the complex sample design effects, including weighted estimates of proportions, design-based estimates of sampling variance, and generalized design effect corrections for key test statistics. Later, in Chapters 8 and 9, different forms of the generalized linear model and **pseudo-maximum likelihood** techniques will be discussed for regression modeling of categorical data that follow an approximate binomial, multinomial, Poisson, or negative binomial distribution.

6.2.2 Proportions and Percentages

In Section 6.3 we discuss estimation of proportions and their standard errors. In this chapter, we denote estimates of population proportions as p and population proportions as π. Many software packages choose to output estimates of percentages and standard errors on the percentage scale. Translation between estimates and standard errors for proportions and percentages simply involves the following multiplicative scaling:

$$percent = 100 \cdot p, \quad se(percent) = 100 \cdot se(p), \quad \text{var}(percent) = 100^2 \cdot \text{var}(p)$$

While these relationships may be obvious to experienced analysts, our experience suggests that it is easy for mistakes to be made, especially in translating standard errors from the percentage to the proportion scale.

6.2.3 Cross-Tabulations, Contingency Tables, and Weighted Frequencies

Categorical data analysis becomes more complex (and also more interesting) when the analysis is expanded from a single variable to include two or more categorical variables. With two categorical variables, the preferred summary is typically a two-way data display with $r = 1, \ldots, R$ rows and $c = 1, \ldots, C$ columns, often referred to in statistical texts and literature as a **cross-tabulation** or a **contingency table**. Cross-tabs and contingency tables are not limited to two dimensions but may include a third (or higher) dimension, that is, $l = 1, \ldots, L$ layers or subtables based on categories of a third variable. For simplicity of presentation, the majority of the examples in this chapter will be based on the cross-tabulation of two categorical variables (see Section 6.4).

Consider the simple $R = 2$ by $C = 2$ (2×2) tabular display of observed sample frequencies shown in Figure 6.1.

Under simple random sampling (SRS), one-, two-, three-, or higher-dimension arrays of unweighted sample frequencies like the 2×2 array illustrated in Figure 6.1 can be used directly to estimate statistics of interest such as the row proportion in category 1, $p_{1|1} = n_{11} / n_{1+}$, or derive tests of hypotheses for the relationships between categorical variables, such as the Pearson chi-square (χ^2) test. However, because individuals can be selected to a survey sample with varying probabilities, estimates and test statistics computed from the unweighted sample frequencies may be biased for the true properties of the survey population. Therefore, it is necessary to translate from unweighted sample counts to weighted frequencies as shown in Figure 6.2,

Variable 2	Variable 1		Row Margin
	0	1	
0	n_{00}	n_{01}	n_{0+}
1	n_{10}	n_{11}	n_{1+}
Column Margin	n_{+0}	n_{+1}	n_{++}

FIGURE 6.1
Bivariate distribution of observed sample frequencies.

Variable 2	Variable 1		Row Margin
	0	1	
0	$\hat{N}_{00}$	$\hat{N}_{01}$	$\hat{N}_{0+}$
1	$\hat{N}_{10}$	$\hat{N}_{11}$	$\hat{N}_{1+}$
Column Margin	$\hat{N}_{+0}$	$\hat{N}_{+1}$	$\hat{N}_{++}$

FIGURE 6.2
Bivariate distribution of weighted sample frequencies.

where for example, the weighted frequency (or estimated population count) in cell (0,1) is

$$\hat{N}_{01} = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i \in (0,1)} w_{h\alpha i}$$

The weighted proportions estimated from these weighted sample frequencies, such as $p_{rc} = \hat{N}_{rc} / \hat{N}_{++}$, will reflect the relative size of the total population in the corresponding cell, row margin or column margin of the cross-tabulation.

We discuss bivariate analyses in more detail in Section 6.4.

6.3 Univariate Analysis of Categorical Data

Simple descriptive analyses of a single categorical variable in a survey data set can take a number of forms, including estimation of a simple population proportion π for binary responses; estimation of the population proportion, π_k, for each of the $k = 1, \ldots, K$ categories of a multicategory nominal or ordinal response variable; and, with care, estimation of the means for ordinally scaled responses (see Section 5.3.3). To draw inferences about the population parameters being estimated, analysts can construct $100(1 - \alpha)\%$ confidence intervals (CIs) for the parameters or perform Student t or simple χ^2 hypothesis tests. Properly weighted graphical displays of the frequency distribution of the categorical variable are also very effective tools for presentation of results.

6.3.1 Estimation of Proportions for Binary Variables

Estimation of a single population proportion, π, for a binary response variable requires only a straightforward extension of the ratio estimator (Section

5.3.3) for the population mean of a continuous random variable. By recoding the original response categories to a single indicator variable y_i with possible values 1 and 0 (e.g., yes = 1, no = 0), the ratio mean estimator estimates the proportion or **prevalence**, π, of "1s" in the population:

$$p = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I(y_i = 1)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}} = \frac{\hat{N}_1}{\hat{N}} \tag{6.1}$$

Most software procedures for survey data analysis provide the user with at least two alternatives for calculating ratio estimates of proportions. The first alternative is to use the single variable option in procedures such as Stata `svy: prop` and `svy: tab` or SAS PROC SURVEYFREQ. These programs are designed to estimate the univariate proportions for the discrete categories of a single variable as well as total, row, and column proportions for cross-tabulations of two or more categorical variables. The second alternative for estimating single proportions is to first generate an indicator variable for the category of interest (with possible values 1 if a case falls into the category of interest, or 0 if a case does not fall into the category of interest) and then to apply a procedure designed for estimation of means (e.g., Stata `svy: mean` or SAS PROC SURVEYMEANS). Both approaches will yield identical estimates of the population proportion and the standard error, but, as explained next, estimated confidence intervals developed by the different procedures may not agree exactly due to differences in the methods used to derive the lower and upper confidence limits.

An application of Taylor series linearization (TSL) to the ratio estimator of π in Equation 6.1 results in the following TSL variance estimator for the ratio estimate of a simple proportion:

$$v(p) \doteq \frac{V(\hat{N}_1) + p^2 \cdot V(\hat{N}) - 2 \cdot p \cdot Cov(\hat{N}_1, \hat{N})}{\hat{N}^2} \tag{6.2}$$

Replication techniques (jackknife repeated replication [JRR], balanced repeated replication [BRR]) can also be used to compute estimates of the variances of these estimated proportions and the corresponding design-based confidence intervals and test statistics.

If the analyst estimates a proportion as the mean of an indicator variable (e.g., using Stata's `svy: mean` procedure), a standard design-based confidence interval is constructed for the proportion, $CI(p) = p \pm t_{1-\alpha/2,df} \cdot se(p)$. One complication that arises when proportions are viewed simply as the mean of a binary variable is that the true proportion, π, is constrained to lie

in the interval (0,1). When the estimated proportion of interest is extreme (i.e., close to 0 or close to 1), the standard design-based confidence limits may be less than 0 or exceed 1. To address this problem, alternative computations of design-based confidence intervals for proportions have been proposed, including the logit transformation procedure and the modified Wilson procedure (Hsu and Rust, 2007).

Stata's `svy: tab` command uses the logit transformation procedure by default. Implementation of this procedure for constructing a confidence interval is a two-step process. First, using the weighted estimate of the proportion, one constructs a 95% CI for the logit transform of p:

$$\text{CI}\,[\text{logit}(p)] = \{\text{A,B}\} = \left\{\ln\left(\frac{p}{1-p}\right) - \frac{t_{1-\alpha/2,df}\cdot se(p)}{p\cdot(1-p)}, \ln\left(\frac{p}{1-p}\right) + \frac{t_{1-\alpha/2,df}\cdot se(p)}{p\cdot(1-p)}\right\} \tag{6.3}$$

where

p = the weighted estimate of the proportion of interest;
$se(p)$ = the design-based Taylor Series approximation to the standard error of the estimated proportion.

Next, the two confidence limits on the logit scale, A and B, are transformed back to the original (0,1) scale:

$$CI(p) = \left\{\frac{e^A}{1+e^A}, \frac{e^B}{1+e^B}\right\} \tag{6.4}$$

Although procedures such as Stata `svy: tab` and SUDAAN PROC CROSSTAB default to the logit transformation formula for estimating CI(p) for all values of p, the adjusted CIs generally do not differ from the standard symmetric CIs unless $p < 0.10$ or $p > 0.90$. Interested readers can find a description of the modified Wilson Procedure in Theory Box 6.1.

Example 6.1: Estimating the Proportion of the U.S. Adult Population with an Irregular Heart Beat

In this example, data from the 2005–2006 NHANES are used to estimate the proportion of U.S. adults with an irregular heartbeat (defined by BPXPULS = 2 in the NHANES data set). To enable a direct comparison of the results from analyses performed using Stata `svy: tab`, `svy: prop`, and `svy: mean`, we generate a binary indicator, equal to 1 for respondents reporting an irregular heartbeat, and 0 otherwise. The recoding of the original response categories for BPXPULS would not be necessary for estimation performed using either `svy: tab` or `svy:`

`prop`. Note that as in the Chapter 5 analysis examples, we create an indicator for the adult subpopulation age 18 and older, enabling appropriate unconditional subpopulation analyses (see Section 4.5.2). As in previous examples, during the recoding steps, missing data on the original variable are explicitly assigned to a missing data code on the recoded variable:

```
gen irregular = .
replace irregular = 1 if bpxpuls == 2
replace irregular = 0 if bpxpuls == 1
gen age18p = 1 if age >= 18 & age != .
replace age18p = 0 if age < 18
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra) ///
vce(linearized) singleunit(missing)
svy, subpop(age18p): tab irregular, se ci col deff
svy, subpop(age18p): proportion irregular
svy, subpop(age18p): mean irregular
```

Table 6.1 provides the weighted estimate of the proportion, the estimated standard error, the 95% confidence interval, and the estimated design effect produced by each of these three command alternatives. (Note that `svy: tab` and `svy: prop` produce estimated proportions and standard errors for each category; only the results for IRREGULAR = 1 are shown.)

The results in Table 6.1 suggest that 3% of U.S. adults are estimated to have irregular heartbeats, with a 95% CI of (1.8%, 4.8%) using the logit transform

THEORY BOX 6.1 THE MODIFIED WILSON PROCEDURE FOR CI(*P*)

The modified Wilson procedure (Korn and Graubard, 1999; Kott and Carr, 1997) for complex samples constructs the confidence limits for the proportion using the expression

$$CI(p)_{Wilson} = \frac{(2n^{*}p + t^{2}) \pm (t\sqrt{(t^{2} + 4p(1-p)n^{*})})}{2(n^{*} + t^{2})}$$

where

p is the weighted estimate of the population proportion, π;

$t = t_{1-\alpha/2,df}$ is the $(1 - \alpha/2)$ critical value of the Student t distribution with df design degrees of freedom; and

$n^{*} = p \cdot (1 - p)/\text{var}_{des}(p) = n/d^{2}(p)$ is the effective sample size of the estimate of π.

The modified 95% Wilson confidence interval for the proportion of U.S. adults with irregular heartbeat is equal to (0.018, 0.049), which is quite similar to the design-based confidence interval in Table 6.1 computed by Stata using the logit transformation procedure.

TABLE 6.1

Estimated Proportions of Adults in the U.S. Population with Irregular Heartbeats

Variable	*n*	Design *df*	Estimated Proportion	Linearized SE	95% CI	Design Effect
Stata `svy: tab`, CI Based on Logit Transform Technique						
Irregular	5,121	15	0.030	0.007	(0.018, 0.048)	11.33
Stata `svy: prop`, CI Computed Using the Standard Symmetric Interval						
Irregular	5,121	15	0.030	0.007	(0.015, 0.044)	11.33
Stata `svy: mean`, CI Computed Using the Standard Symmetric Interval						
Irregular	5,121	15	0.030	0.007	(0.015, 0.044)	11.33

method in the `svy: tab` command and a 95% CI of (1.5%, 4.4%) from the `svy: mean` and `svy: prop` commands. Due to the small value of p, the logit transform 95% confidence interval for the proportion of adults with an irregular heartbeat is not symmetric about the point estimate of 3% prevalence.

As a brief aside, the large design effect for the sampling variance of the estimated proportion is due to the large cluster sizes in the primary stage of the 2005–2006 NHANES sample. However, the NHANES samples are not primarily designed to generate precise *overall* estimates of these types of proportions; they are designed to generate precise estimates for subpopulations of the U.S. population defined by demographic characteristics such as ethnicity and gender (Mohadjer and Curtin, 2008). NHANES design effects for estimates of subpopulation proportions are considerably smaller.

6.3.2 Estimation of Category Proportions for Multinomial Variables

Estimation of multinomial proportions and standard errors for multicategory survey response variables (e.g., ethnicity, measured by 1 = Mexican, 2 = Other Hispanic, 3 = White, 4 = Black, 5 = Other) is a direct extension of the method for estimating a single proportion. The ratio estimator in Equation 6.1 is simply applied to indicator variables for each distinct category:

$$p_k = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} I(y_i = k)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i}} = \frac{\hat{N}_k}{\hat{N}} \tag{6.5}$$

The result is a vector of estimated population proportions for each category: $\boldsymbol{p} = \{p_1, ..., p_K\} = \{\hat{N}_1 / \hat{N}, ..., \hat{N}_K / \hat{N}\}$. Standard errors for the weighted estimates of the individual category proportions can be computed using

TABLE 6.2

Estimated Proportions of Adults in the U.S. Population by Race/Ethnicity

Race/Ethnicity	*n*	Estimated Proportion	Linearized SE	95% CI	Design Effect
Mexican	1,185	0.081	0.010	(0.059,0.102)	10.28
Other Hispanic	171	0.034	0.007	(0.018,0.050)	12.74
White	2,633	0.714	0.028	(0.655,0.773)	28.39
Black	1,341	0.117	0.020	(0.075,0.160)	28.75
Other	233	0.054	0.006	(0.042,0.066)	5.01

Source: Analysis based on the 2005–2006 NHANES data.

Taylor series linearization (as presented in Equation 6.2) or by BRR or JRR (replication) methods. The most convenient method of simultaneously estimating the *K* category proportions for a multinomial categorical variable is through the use of the single variable option in programs such as Stata `svy: tab` or `svy: prop` or SAS PROC SURVEYFREQ.

Example 6.2: Estimating the Proportion of U.S. Adults by Race and Ethnicity

The following Stata commands estimate the proportion of the 2005–2006 NHANES adult survey population by self-reported race and ethnicity:

```
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra)
svy, subpop(age18p): prop ridreth1
estat effects
```

Table 6.2 presents the estimated proportions of U.S. adults for the five NHANES race/ethnicity categories along with the standard errors of the estimates, the 95% CIs, and design effects for the estimated proportions.

Example 6.3: Estimating the Proportions of U.S. Adults by Blood Pressure Status

A common technique in survey data analysis is to classify the population into discrete categories based on the values of one or more continuous survey variables. This example illustrates the variable recoding steps and the Stata `svy: tab` command to estimate the proportions of U.S. adults falling into each of the following four categories of blood pressure status: (1) normal (systolic blood pressure [SBP] < 120 and diastolic blood pressure [DBP] < 80); (2) prehypertension (SBP 120–139, or DBP 80–89); (3) Stage 1 hypertension (SBP 140–159, or DBP 90–99); and (4) Stage 2 hypertension (SBP 160+, or DBP 100+).

The analysis begins with the data recoding steps. First, four NHANES replicate measurements of blood pressure are averaged to create single mean values for

TABLE 6.3

Estimated Proportions of U.S. Adults by Blood Pressure Status

Blood Pressure Category	*n*	Estimated Proportion	Linearized SE	95% CI	Design Effect
Normal	2,441	0.471	0.011	(0.448, 0.495)	3.58
Prehypertension	1,988	0.419	0.012	(0.394, 0.444)	4.15
Stage 1 Hypertension	470	0.086	0.006	(0.074, 0.101)	3.54
Stage 2 Hypertension	158	0.024	0.002	(0.019, 0.030)	1.79

Source: Analysis based on the 2005–2006 NHANES data.

the diastolic and systolic blood pressure. Based on these mean values of diastolic and systolic pressure, respondents are assigned to one of the four blood pressure categories.

```
egen meanbpxsy = rowmean(bpxsy1 bpxsy2 bpxsy3 bpxsy4)
egen meanbpxdi = rowmean(bpxdi1 bpxdi2 bpxdi3 bpxdi4)
gen bp_cat = .
replace bp_cat = 1 if (meanbpxsy < 120 & meanbpxdi < 80)
replace bp_cat = 2 if ((meanbpxsy >= 120 & meanbpxsy < 140) ///
| (meanbpxdi >= 80 & meanbpxdi < 90))
replace bp_cat = 3 if ((meanbpxsy >= 140 & meanbpxsy < 160) ///
| (meanbpxdi >= 90 & meanbpxdi < 100))
replace bp_cat = 4 if ((meanbpxsy >= 160 & meanbpxsy != .) ///
| (meanbpxdi >= 100 & meanbpxdi != .))
```

The NHANES sample design is identified through the `svyset` command and the `svy: tab` command is then applied to generate weighted estimates of the proportions of the adult population (AGE18P = 1) falling into each of these four categories:

```
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra)
svy, subpop(age18p): tab bp_cat, obs se ci col
svy, subpop(age18p): tab bp_cat, deff
```

The output generated by these commands is summarized in Table 6.3.

These results suggest that an estimated 47% of the adult population has "normal" blood pressure (95% CI = 44.8%, 49.5%), while roughly 42% of the adult population is at the prehypertension stage (95% CI = 39.4%, 44.4%). Approximately 11% of the adult population is estimated to have either Stage 1 or Stage 2 hypertension. As in Example 6.1, the asymmetry of the 95% CIs for the smaller estimated proportions is due to the use of the logit transform method.

6.3.3 Testing Hypotheses Concerning a Vector of Population Proportions

To test a null hypothesis of the form H_0: {$\pi_1 = 0.5$, $\pi_2 = 0.3$, $\pi_3 = 0.15$, and $\pi_4 = 0.05$}, the standard chi-square **goodness-of-fit** test statistic,

$$X^2 = n_{++} \cdot \sum_k (p_k - \pi_k)^2 / \pi_k$$

(with degrees of freedom equal to the number of categories minus 1), can be considerably biased when applied to complex sample survey data. Jann (2008) extended the GOF test to the complex sample survey data setting, and the procedure has been implemented in Stata's `mgof` command (we omit details of the approach here; interested readers can review the article).

Example 6.4: A Goodness-of-Fit Test for Blood Pressure Status Category Proportions

After installing the `mgof` command (Web-aware Stata users can type `findit mgof` for help finding the command), a new variable, PI, is created using the `recode` command that contains the null hypothesis values for each category of the BP_CAT variable created in the data preparation commands for Example 6.3:

```
recode bp_cat (1=.5) (2=.3) (3=.15) (4=.05), generate(pi)
```

Next, the mgof command is submitted to test the null hypothesis:

```
mgof bp_cat = pi if age18p == 1, svy
```

Note that the analysis is restricted to the adult subpopulation. Submitting this command produces the following test statistics and *p*-values for the goodness-of-fit hypothesis test:

Goodness-of-Fit Statistic	X^2	$P(\chi_3^2 > X^2)$
Rao–Scott $X^2_{Pearson}$	450.24	$p < 0.0001$
$X^2_{LR, Jann}$	465.48	$p < 0.0001$

Interpreting the results of these tests, the null hypothesis is rejected by both the design-adjusted Rao–Scott version of Pearson's chi-square test statistic (see Section 6.4.4) and the likelihood ratio chi-square statistic developed by Jann (2008). The weighted estimates computed using the collected sample data therefore do not seem to "fit" the hypothesized population distribution of category proportions.

6.3.4 Graphical Display for a Single Categorical Variable

Categorical survey variables naturally lend themselves to graphical display and analytic interpretation. As discussed in Chapter 5, it is important that the software used to generate graphics for analysis or publication has the capability to incorporate the survey weights for the display to accurately reflect the estimated distribution of the variable in the survey population.

Example 6.5: Pie Charts and Vertical Bar Charts of the Estimated Blood Pressure Status Classification for U.S. Adults from the 2005–2006 NHANES Data

Using Stata graphics, simple weighted pie charts (Figure 6.3) or weighted vertical bar charts (Figure 6.4) produce an effective display of the estimated population distribution across categories of a single categorical variable. The Stata command syntax used to generate these two figures is as follows:

```
* Pie Chart (one long command).
graph pie bp_cat1 bp_cat2 bp_cat3 bp_cat4 ///
[pweight=wtmec2yr] if age18p==1, plabel(_all percent, ///
format(%9.1f)) scheme(s2mono) ///
legend (label /// (1 "Normal") label /// (2 "Pre-Hypertensive")
/// label ///(3 "Stage 1 Hypertensive") label (4 "Stage 2 ///
Hypertensive"))
* Vertical Bar Chart (one long command).
graph bar (mean) bp_cat1 bp_cat2 bp_cat3 bp_cat4 ///
[pweight=wtmec2yr] if age18p==1, blabel(bar, ///
format(%9.1f) color(none)) ///
bar(1,color(gs12)) bar(2,color(gs4)) bar(3,color(gs8)) ///
bar(4,color(black)) ///
bargap(7) scheme(s2mono) over(riagendr) percentages ///
legend (label(1 "Normal")label(2 "Pre-Hypertensive") ///
label(3 "Stage 1 Hypertensive") label (4 "Stage 2 ///
Hypertensive")) ///
ytitle ("Percentage")
```

6.4 Bivariate Analysis of Categorical Data

Bivariate analysis of categorical data may take a number of forms, ranging from estimation of proportions for the joint distribution of the two variables (total proportions) or estimation of "conditional" proportions based on levels of the second categorical variable (i.e., row proportions or column proportions) to techniques for measuring and testing bivariate association between two categorical variables. Bivariate categorical data analyses and reporting are important in their own right, and they are also important as exploratory tools in the development of more complex multivariate models (see the regression model building steps in Section 8.3).

6.4.1 Response and Factor Variables

Unlike simple linear regression for continuous survey variables, there is no requirement to differentiate variables by type (dependent or independent) or postulate a cause–effect relationship between two or more categorical variables that are being analyzed simultaneously. An example where assigning

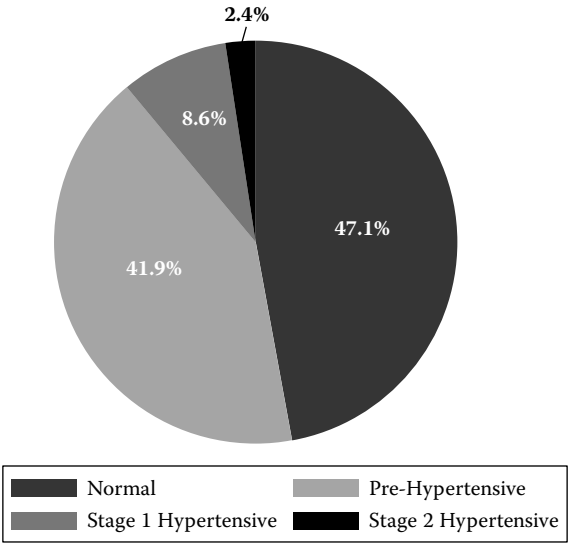


FIGURE 6.3
Pie chart of the estimated distribution of blood pressure status of U.S. adults.

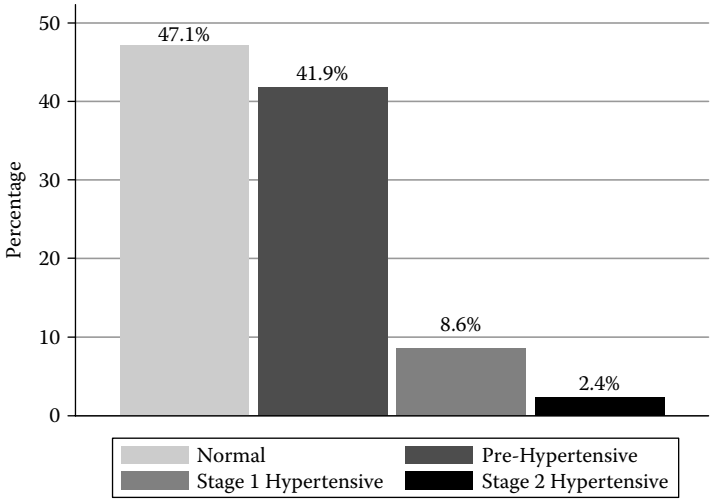


FIGURE 6.4
Bar chart of the estimated distribution of the blood pressure status of U.S. adults. (Modified from the 2005–2006 NHANES data.)

the variables to dependent or independent status is not necessary would be a bivariate analysis of the population distribution by education levels and census region. Nevertheless, in many categorical data analysis problems it is natural to think of one categorical variable as the **response variable** and the others as **factor variables** that may be associated with the discrete outcome for the response variable. This section will introduce examples centered around the analysis of the joint distribution of U.S. adults' experience with a lifetime episode of major depression (yes, no) and their gender (male, female), based on the National Comorbidity Survey Replication (NCS-R) data set. In these examples, it is convenient to label the NCS-R indicator variable for a lifetime major depressive episode (MDE) as the response variable and the respondent's gender (SEX) as the factor. Assigning response and factor labels to the categorical variables also sets up the transition to later discussion of regression modeling of categorical data where dependent and independent variables are clearly specified.

6.4.2 Estimation of Total, Row, and Column Proportions for Two-Way Tables

Based on the weighted frequencies illustrated in Figure 6.2, estimates of population proportions can be computed as the ratio of the weighted sample frequency for the cell to the appropriate weighted total or marginal frequency value. For example, Figure 6.5 illustrates estimation of the **total proportions** of the population in each cell and margin of the table. Note that the numerator of each estimated proportion is the weighted total frequency for the cell, such as $\hat{N}_{A1}$, and the denominator is the weighted total population frequency, $\hat{N}_{++}$. Statistical software also enables the user to condition estimates of population proportions on the sample in particular rows or columns of the crosstabulation. Figure 6.6 illustrates the calculations of weighted estimates of **row proportions** for the estimated population distribution in Figure 6.2.

The following example uses the NCS-R data on lifetime major depressive episode (MDE) and gender (SEX) to illustrate the Stata syntax to estimate total proportions, row proportions, standard errors, and confidence intervals.

	Response		
Factor	0	1	
A	$p_{A0} = \dfrac{\hat{N}_{A0}}{\hat{N}_{++}}$	$p_{A1} = \dfrac{\hat{N}_{A1}}{\hat{N}_{++}}$	$p_{A+} = p_{A0} + p_{A1}$
B	$p_{B0} = \dfrac{\hat{N}_{B0}}{\hat{N}_{++}}$	$p_{B1} = \dfrac{\hat{N}_{B1}}{\hat{N}_{++}}$	$p_{B+} = p_{B0} + p_{B1}$
	$p_{+0} = p_{A0} + p_{B0}$	$p_{+1} = p_{A1} + p_{B1}$	$p_{++} = 1.0$

FIGURE 6.5
Estimation of overall (total) population proportions (multinomial sampling model).

Factor	Response		
	0	1	
A	$p_{0\|A} = \frac{\hat{N}_{A0}}{\hat{N}_{A+}}$	$p_{1\|A} = \frac{\hat{N}_{A1}}{\hat{N}_{A+}}$	$p_{A+} = 1.0$
B	$p_{0\|B} = \frac{\hat{N}_{B0}}{\hat{N}_{B+}}$	$p_{1\|B} = \frac{\hat{N}_{B1}}{\hat{N}_{B+}}$	$p_{B+} = 1.0$

FIGURE 6.6
Estimation of row population proportions (product multinomial sampling model).

Example 6.6 Estimation of Total and Row Proportions for the Cross-Tabulation of Gender and Lifetime Major Depression Status Using the NCS-R Data

The first of two `svy: tab` commands in Stata requests as follows the default estimates of the total proportions for the SEX × MDE crosstabulation along with the corresponding standard errors, 95% CIs, and design effects. The `row` option in the second `svy: tab` command specifies that the estimates, standard errors, CIs, and design effects will be for the row proportions:

```
svyset seclustr [pweight = ncsrwtsh], strata(sestrat)
svy: tab sex mde, se ci deff
svy: tab sex mde, row se ci deff
```

The estimated proportions, standard errors, 95% CIs, and design effect output from these two commands are summarized in Table 6.4.

6.4.3 Estimating and Testing Differences in Subpopulation Proportions

Estimates of row proportions in two-way tables (e.g., $\hat{p}_{1|B} = 0.230$ in Table 6.4) are in fact subpopulation estimates in which the subpopulation is defined by the levels of the factor variable. Analysts interested in testing differences in response category proportions between two levels of a factor variable can use methodology similar to that discussed in Section 5.6 for comparison of subpopulation means.

Example 6.7: Comparing the Proportions of U.S. Adult Men and Women with Lifetime Major Depression

This example uses data from the NCS-R to test a null hypothesis that there is no difference in the proportions of U.S. adult men and women with a lifetime diagnosis of a major depressive episode. To compare male and female row proportions for MDE = 1, the `svy: prop` command with the `over()` option is used to estimate the vector of row proportions (see Table 6.4) and their design-based

TABLE 6.4

Estimated Proportions of U.S. Adults by Gender and Lifetime Major Depression Status

Description	Parameter	Estimated Proportion	Linearized SE	95% CI	Design Effect
		Total Proportions			
Male, no MDE	π_{A0}	0.406	0.007	(0.393, 0.421)	1.87
Male, MDE	π_{A1}	0.072	0.003	(0.066, 0.080)	1.64
Female, no MDE	π_{B0}	0.402	0.005	(0.391, 0.413)	1.11
Female, MDE	π_{B0}	0.120	0.003	(0.114, 0.126)	0.81
		Row Proportions			
No MDE | Male	$\pi_{0\mid A}$	0.849	0.008	(0.833, 0.864)	2.08
MDE | Male	$\pi_{1\mid A}$	0.151	0.008	(0.136, 0.167)	2.08
No MDE|Female	$\pi_{0\mid B}$	0.770	0.006	(0.759, 0.782)	0.87
MDE | Female	$\pi_{1\mid B}$	0.230	0.006	(0.218, 0.241)	0.87

Source: Analysis based on NCS-R data.

variance–covariance matrix. Internally, Stata labels the estimated row proportions for MDE = 1 as `_prop_2`. The `lincom` command is then executed to estimate the contrast of the male and female proportions and the standard error of this difference. The relevant Stata commands are as follows:

```
svyset seclustr [pweight=ncsrwtsh], strata(sestrat) ///
vce(linearized) singleunit(missing)
svy: proportion mde, over(sex)
lincom [_prop_2]Male - [_prop_2]Female
```

The output of the `lincom` command provides the following estimate of the male–female difference, its standard error, and a 95% CI for the contrast of proportions.

$\hat{\Delta} = p_{male} - p_{female}$	$se(\hat{\Delta})$	$CI_{95}(\Delta)$
–0.079	0.010	(–0.098, –0.059)

Because the design-based 95% CI for the difference in proportions does not include 0, the data suggest that the rate of lifetime major depressive episodes for women is significantly higher than that for men.

6.4.4 Chi-Square Tests of Independence of Rows and Columns

For a 2 × 2 table, the contrast of estimated subpopulation proportions examined in Example 6.7 is equivalent to a test of whether the response variable (MDE) is independent of the factor variable (SEX). More generally, under SRS, two categorical variables are independent of each other if the following

relationship of the expected proportion in row *r*, column *c* of the cross-tabulation holds:

$$\hat{\pi}_{rc} = \text{Expected proportion in row } r, \text{ column } c = \frac{n_{r+}}{n_{++}} \cdot \frac{n_{+c}}{n_{++}} = p_{r+} \cdot p_{+c} \quad (6.6)$$

Under SRS, formal testing of the hypothesis that two categorical variables are independent can be conducted by comparing the expected cell proportion under the independence assumption, $\hat{\pi}_{rc}$, to the observed proportion from the survey data, p_{rc}. Intuitively, if the differences $(\hat{\pi}_{rc} - p_{rc})$ are large there is evidence that the independence assumption does not hold and that there is an association between the row and column variables. In standard practice, two statistics are commonly used to test the hypothesis of independence in two-way tables:

Pearson's chi-square test statistic:

$$X^2_{Pearson} = n_{++} \cdot \sum_r \sum_c (p_{rc} - \hat{\pi}_{rc})^2 / \hat{\pi}_{rc} \quad (6.7)$$

Likelihood ratio test statistic:

$$G^2 = 2 \cdot n_{++} \cdot \sum_r \sum_c p_{rc} \times \ln\left(\frac{p_{rc}}{\hat{\pi}_{rc}}\right) \quad (6.8)$$

Under the null hypothesis of independence for rows and columns of a two-way table, these test statistics both follow a central χ^2 distribution with $(R - 1) \times (C - 1)$ degrees of freedom.

Ignoring the complex sample design that underlies most survey data sets can introduce bias in the estimated values of these test statistics. To correct the bias in the estimates of the population proportions used to construct the test statistic, weighted estimates of the cell, row, and column proportions are substituted, for example, $p_{rc} = \hat{N}_{rc} / \hat{N}_{++}$. To correct for the design effects on the sampling variances of these proportions, two general approaches have been introduced in the statistical literature. Both approaches involve scaling the standard $X^2_{Pearson}$ and G^2 test statistics by dividing them by an estimate of a **generalized design effect factor** (GDEFF). Theory Box 6.2 provides a mathematical explanation of how the generalized design effect adjustments are computed.

Fellegi (1980) was the first to propose such a correction based on a generalized design effect. Rao and Scott (1984) and later Thomas and Rao (1987) extended the theory of generalized design effect corrections for these test statistics. The Rao–Scott method requires the computation of **generalized**

THEORY BOX 6.2 FIRST- AND SECOND-ORDER DESIGN EFFECT CORRECTIONS

The Fellegi (1980) method for a generalized design effect correction to the chi-square test statistic is best summarized as a three-step process. First, the average of the design effects for the $R \times C$ (unweighted) proportions involved in the computation of the chi-square statistic is computed. The standard Pearson or likelihood ratio chi-square statistic computed under simple random sampling is then divided by the average design effect. The resulting adjusted chi-square test statistic is referred to a χ^2 distribution with degrees of freedom equal to $(R - 1) \times (C - 1)$ to test the null hypothesis of independence.

Rao and Scott (1984) built on this method by advocating the use of weighted estimates of the proportions in the construction of the standard chi-square statistics. Under the Rao–Scott method, the **generalized design effect** is defined as the mean of the eigenvalues of the following matrix, $\boldsymbol{D}$:

$$\boldsymbol{D} = \boldsymbol{V}_{Design}\boldsymbol{V}_{SRS}^{-1} \tag{6.9}$$

In Equation 6.9, $\boldsymbol{V}_{Design}$ is the matrix of design-based (e.g., linearized) variances and covariances for the $R \times C$ vector of estimated proportions used to construct the chi-square test statistic, and $\boldsymbol{V}_{SRS}$ is the matrix of variance and covariances for the estimated proportions given a simple random sample of the same size. The Rao–Scott generalized design effect factor, *GDEFF,* for two-way tables can then be written as follows:

$$GDEFF = \frac{\sum_r \sum_c (1 - p_{rc}) \cdot d^2(p_{rc}) - \sum_r (1 - p_{r+}) \cdot d^2(p_{r+}) - \sum_c (1 - p_{+c}) \cdot d^2(p_{+c})}{(R-1)(C-1)} \tag{6.10}$$

The design-adjusted test statistics introduced in Equation 6.11 are computed based on **first-order design corrections** of this type.

Thomas and Rao (1987) derived **second-order design corrections** to the test statistics, which incorporate variability in the eigenvalues of the $\boldsymbol{D}$ matrix. These second-order design corrections can be implemented by dividing the adjusted test statistic based on the first-order correction (e.g., $X^2_{R-S} = X^2_{Pearson} / GDEFF$) by the quantity $(1 + a^2)$, where a represents the coefficient of variation of the eigenvalues of the $\boldsymbol{D}$ matrix. The *F*-transformed version of this second-order design-corrected version of the Pearson chi-square statistic is currently the default test statistic reported by Stata's `svy: tab` command for analyses of two-way tables.

Thomas and Rao (1987) used simulations to show that this second-order correction controls Type I error rates much better when there is substantial variance in the eigenvalues of $\boldsymbol{D}$.

design effects that are analytically more complicated than the Fellegi approach. The Rao–Scott procedures are now the standard in procedures for the analysis of categorical survey data in software systems such as Stata and SAS. The design-adjusted Rao–Scott Pearson and likelihood ratio chi-square test statistics are computed as follows:

$$\begin{aligned} X^2_{R-S} &= X^2_{Pearson} / GDEFF, \\ G^2_{R-S} &= G^2 / GDEFF \end{aligned} \tag{6.11}$$

Under the null hypothesis of independence of rows and columns, both of these adjusted test statistics can be referred to a χ^2 distribution with $(R-1) \times (C-1)$ degrees of freedom. Thomas and Rao (1987) showed that a transformation of the design-adjusted $X^2_{R\text{-}S}$ and $G^2_{R\text{-}S}$ values produced a more stable test statistic that under the null hypothesis closely approximated an F distribution. Table 6.5 defines the F-transformed version of these two chi-square test statistics and the corresponding F reference distribution to be used in testing the independence of rows and columns.

A third form of the chi-square test statistic that may be used to test the null hypothesis of independence of rows and columns in a cross-tabulation of two categorical variables is the **Wald chi-square test statistic** (see Theory Box 6.3). We will see in later chapters that Wald statistics play an important role in hypothesis testing for linear and generalized linear models. However, simulation studies have shown that the standard Pearson chi-square test statistic and its design-adjusted forms proposed by Rao and Scott (1984) and Rao and Thomas (1988) perform best for both sparse and nonsparse tables

TABLE 6.5

F-Transformations of the Rao–Scott Chi-Square Test Statistics

***F*-Transformed Test Statistics**	***F* Reference Distribution under H_0**
$F_{R-S,Pearson} = X^2_{R-S}/(R-1)(C-1)$	$F_{(R-1)(C-1),(R-1)(C-1)\cdot df}$
$F_{R-S,LRT} = G^2_{R-S}/(R-1)(C-1)$	$F_{(R-1)(C-1),(R-1)(C-1)df}$

where R is the number of rows, C is the number of columns in the crosstab, and df is the design degrees of freedom

Note: Stata employs a special procdure involving a Satterthwaite correction in deriving these F statistics. This can result in non-integer degrees of freedom (Stater, 2008). See Table 6.6.

THEORY BOX 6.3 THE WALD CHI-SQUARE TEST OF INDEPENDENCE FOR CATEGORICAL VARIABLES

The Wald chi-square test statistic for the null hypothesis of independence of rows and columns in a two-way table is defined as follows:

$$Q_{Wald} = \hat{\boldsymbol{Y}}'(\boldsymbol{H}\hat{\boldsymbol{V}}(\hat{\boldsymbol{N}})\boldsymbol{H}')^{-1}\hat{\boldsymbol{Y}}, \tag{6.12}$$

where

$$\hat{\boldsymbol{Y}} = (\hat{\boldsymbol{N}} - \boldsymbol{E}) \tag{6.13}$$

is a vector of $R \times C$ differences between the observed and expected cell counts, for example, $\hat{N}_{rc} - E_{rc}$ where under the independence hypothesis, $E_{rc} = \hat{N}_{r+} \cdot \hat{N}_{+c} / \hat{N}_{++}$. The matrix term $\boldsymbol{H}\hat{\boldsymbol{V}}(\hat{\boldsymbol{N}})\boldsymbol{H}'$ represents the estimated variance–covariance matrix for the vector of differences. In the case of a complex sample design, the variance–covariance matrix of the weighted frequency counts, $\hat{\boldsymbol{V}}(\hat{\boldsymbol{N}})$, is estimated using a TSL, BRR, or JRR approach that captures the effects of stratification, clustering, and weighting.

Under the null hypothesis of independence, Q_{Wald} follows a χ^2 distribution with $(R - 1) \times (C - 1)$ degrees of freedom. An F-transform of the Wald chi-square test statistic reported in SUDAAN and other software programs is

$$F_{Wald} = Q_{Wald} \times \frac{df - (R-1)(C-1) + 1}{(R-1)(C-1)df} \sim F_{(R\text{-}1)(C\text{-}1),df\text{-}(R\text{-}1)(C\text{-}1)+1} \text{ under } H_0. \tag{6.14}$$

(Sribney, 1998) and that these tests are more powerful than the Wald test statistic, especially for larger tables. As a result, Stata makes the $F_{R\text{-}S,Pearson}$ test statistic (with a second-order design correction incorporated) the default test statistic reported by its `svy: tab` procedure, and this is also the default test statistic reported by SAS PROC SURVEYFREQ (with a first-order design correction incorporated).

Example 6.8: Testing the Independence of Alcohol Dependence and Education Level in Young Adults (Ages 18–28) Using the NCS-R Data

This example uses the Stata `svy: tab` command to compute the Rao–Scott F-statistics and test the independence of two categorical variables that are available in the NCS-R data set (for Part II respondents): ALD, an indicator of receiving a diagnosis of alcohol dependence in the lifetime; and ED4CAT, a categorical variable measuring educational attainment (1 = less than high school, 2 = high

TABLE 6.6

Design-Based Analysis of the Association between NCS-R Alcohol Dependence and Education Level for Young Adults Aged 18–28

Education Level (Grades)	Alcohol Dependence Row Proportions (Linearized SE)		
	0 = No	1 = Yes	Total
0–11	0.909 (0.029)	0.091 (0.029)	1.000
12	0.951 (0.014)	0.049 (0.014)	1.000
13–15	0.951 (0.010)	0.049 (0.010)	1.000
16+	0.931 (0.014)	0.069 (0.014)	1.000
Total	0.940 (0.009)	0.060 (0.009)	1.000
Tests of Independence			
Unadjusted X^2	$P\,(\chi^2_{(3)} > X^2_{Pearson})$	Rao–Scott F	$P\,(F_{2.75,\,115.53} > F_{R\text{-}S})$
$X^2_{Pearson} = 27.21$	$p < 0.0001$	$F_{R\text{-}S,Pearson} = 1.64$	$p = 0.18$
Parameters of the Rao–Scott Design-Adjusted Test			
$n_{18\text{-}28} = 1{,}275$	Design $df = 42$	$GDEFF = 6.62$	$a = 0.56$

school, 3 = some college, 4 = college and above). The analysis is restricted to the subpopulation of NCS-R Part II respondents 18–28 years of age. After identifying the complex design features to Stata, we request the cross-tabulation analysis and any related design-adjusted test statistics by using the `svy: tab` command:

```
svyset seclustr [pweight = ncsrwtlg], strata(sestrat)
svy, subpop(if 18<=age<29): tab ed4cat ald, row se ci deff
```

ED4CAT is specified as the row (factor) variable and ALD as the column (response) variable. Weighted estimates of the row proportions are requested using the `row` option. Table 6.6 summarizes the estimated `row` proportions and standard errors for the ALD × ED4CAT crosstabulation along with the Rao–Scott F-test of independence.

An estimated 9.1% of young adults in the lowest education group have been diagnosed with alcohol dependence at some point in their lifetime (95% CI = 4.7%, 17.0%), while an estimated 6.9% of young adults in the highest education group have been diagnosed with alcohol dependence (95% CI = 4.6%, 10.2%). By default, Stata reports the standard uncorrected Pearson chi-square test statistic ($X^2_{Pearson} = 27.21$, $p < 0.0001$) and then reports the (second-order) design-adjusted Rao–Scott F-test statistic ($F_{R\text{-}S,Pearson} = 1.64$, $p = 0.18$) (see Table 6.5). The standard Pearson X^2 test rejects the null hypothesis of independence at $\alpha = 0.05$; however, when the corrections for the complex sample design are introduced, the Rao–Scott design-adjusted test statistic fails to reject a null hypothesis of independence between education and a lifetime diagnosis of alcohol dependence in this younger population. The appropriate inference in this case would thus be that there is no evidence of a bivariate association between these two categorical factors in this subpopulation. Multivariate analyses examining additional potential predictors of alcohol dependence could certainly be examined at this point (see Chapter 8 for examples).

We remind readers that Stata is using a second-order design correction for the test statistic, which is why the results of these analyses may differ from those found using other software packages (note the decimal degrees of freedom for the design-adjusted *F*-statistic in Table 6.6, due to the second-order correction). If a user specifies the `deff` option, Stata also reports both the mean generalized design effect (*GDEFF* = 6.63) used in the first-order correction and the coefficient of variation of the generalized design effects (*a* = 0.56) used in the second-order correction.

Additional test statistics, including design-adjusted likelihood ratio and Wald test statistics, can be requested in Stata by using the `lr` and `wald` options for the `svy: tab` command. These options do not lead to substantially different conclusions in this illustration and will generally not lead to different inferences about associations between two categorical variables. As mentioned previously, Stata developers advocate the use of the second-order design-adjusted Pearson chi-square statistic (or the Rao–Scott chi-square statistic and its *F*-transformed version) in all situations involving crosstabulations of two categorical variables measured in complex sample surveys (Sribney, 1998).

6.4.5 Odds Ratios and Relative Risks

The **odds ratio**, which we denote by ψ, can be used to quantify the association between the levels of a **response variable** and a categorical factor. Figure 6.7 displays NCS-R weighted estimates (row proportions) of the prevalence of one or more lifetime episodes of major depression by gender.

The odds ratio compares the odds that the response variable takes a specific value across two levels of the factor variable. If the response variable is truly independent of the chosen factor, then $\psi = 1.0$. For example, from Figure 6.7, the estimated male (A)/female (B) odds ratio for MDE is

$$\hat{\psi} = \frac{Odds(MDE = 1 \mid Male)}{Odds(MDE = 1 \mid Female)} = \frac{p_{1|A} / (1 - p_{1|A})}{p_{1|B} / (1 - p_{1|B})} = \frac{p_{1|A} / p_{0|A}}{p_{1|B} / p_{0|B}} = \frac{0.151 / 0.849}{0.230 / 0.770} = 0.595$$

SEX	MDE		
	0	1	
A—Male	$p_{0\|A} = \frac{\hat{N}_{A0}}{\hat{N}_{A+}} = 0.849$	$p_{1\|A} = \frac{\hat{N}_{A1}}{\hat{N}_{A+}} = 0.151$	$p_{A+} = 1.0$
B—Female	$p_{0\|B} = \frac{\hat{N}_{B0}}{\hat{N}_{B+}} = 0.770$	$p_{1\|B} = \frac{\hat{N}_{B1}}{\hat{N}_{B+}} = 0.230$	$p_{B+} = 1.0$

FIGURE 6.7
Estimates of row proportions for MDE by gender.

Note that although this estimate of ψ is computed using the estimated row proportions for the SEX × MDE table, the same estimate would be obtained if the estimated total proportions had been used (Table 6.4):

$$\hat{\psi} = \frac{p_{A1} / p_{A0}}{p_{B1} / p_{B0}} = \frac{0.072 / 0.407}{0.120 / 0.402} = 0.595$$

Since this odds ratio is estimated with no additional controls for other factors such as age or education, it is labeled as an **unadjusted odds ratio**. Note that a correct description of this result is the following: "The *odds* that adult men experience major depression in their lifetime are estimated to be only 59.5% as large as the odds for women." A common mistake in reporting results for estimated odds ratios is to make a statement like the following: "The *probability* that a man experiences an episode of major depression in their lifetime is 59% of that for women."

The latter statement is confusing the odds ratio statistic with a related, yet different, comparative measure, the **relative risk** (computed here using the estimates in Table 6.4):

$$\hat{RR} = \frac{\text{Prob}(MDE = 1 \mid Male)}{\text{Prob}(MDE = 1 \mid Female)} = \frac{p_{1|A}}{p_{1|B}} = \frac{0.151}{0.230} = 0.656$$

The relative risk is the ratio of two conditional probabilities: the probability of MDE for males and the probability of MDE for females. Although both the odds ratio and the relative risk measure the association of a categorical response and a factor variable, they should be distinguished. Only in instances where the prevalence of the response of interest is very small for all levels of the factor (i.e., $p_{1|A}$ and $p_{1|B} < 0.01$) will the odds ratio and relative risk statistics converge to similar numerical values.

If the response and factor variables are independent, then $\psi = 1.0$ (and RR = 1.0). Therefore, to test if categorical response and factor variables are independent, it would be reasonable to construct a confidence interval of the form $\hat{\psi} \pm t_{1-\alpha/2,df} \cdot se(\hat{\psi})$, and establish whether the null value of ψ = 1 is contained within the interval. Although a TSL approximation to $se(\hat{\psi})$ can be derived directly, a CI for ψ is generally obtained from the technique of simple logistic regression.

6.4.6 Simple Logistic Regression to Estimate the Odds Ratio

Logistic regression for binary dependent variables will be covered in depth in Chapter 8. Here, the logit function and simple logistic regression models are briefly introduced to demonstrate their application to estimation of the unadjusted odds ratio and its confidence interval.

The natural logarithm of the odds is termed a **logit function**. Again, using the NCS-R MDE example in Table 6.4, the logits of the probabilities of MDE for the male and female factor levels are

$$\text{logit}(p_{1|A}) = \ln(Odds(MDE = 1 \mid Male)) = \ln\left(\frac{p_{1|A}}{1 - p_{1|A}}\right) = \ln\left(\frac{0.151}{0.849}\right) = -1.727$$

$$\text{logit}(p_{1|B}) = \ln(Odds(MDE = 1 \mid Female) = \ln\left(\frac{p_{1|B}}{1 - p_{1|B}}\right) = \ln\left(\frac{0.230}{0.770}\right) = -1.208$$

Consider a single indicator variable, I_{male}, coded 1 = male and 0 = female, that distinguishes the two levels of SEX. The outcome MDE is coded 1 = yes, 0 = no. A simple logistic regression model for these data is written as follows:

$$\hat{\psi} = \frac{p_{1|A} / (1 - p_{1|A})}{p_{1|B} / (1 - p_{1|B})} = \frac{\exp(\text{logit}(p_{1|A}))}{\exp(\text{logit}(p_{1|B}))} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0)} = \exp(\hat{\beta}_1)$$

Then, we can derive the following result:

$$CI(\psi) = (\exp(\hat{\beta}_1 - t_{1-\alpha/2,df} \cdot se(\hat{\beta}_1)), \exp(\hat{\beta}_1 + t_{1-\alpha/2,df} \cdot se(\hat{\beta}_1)))$$

The resulting confidence interval is not symmetric about the estimated odds ratio but has been shown to provide more accurate coverage of the true population value for a specified level of Type I error (α).

Example 6.9: Simple Logistic Regression to Estimate the NCS-R Male/Female Odds Ratio for Lifetime Major Depressive Episode

As mentioned previously, logistic regression will be covered in detail in later chapters. Here, a simple logistic regression of the NCS-R MDE variable on the indicator of male gender (SEXM) is used to illustrate the technique for estimating the unadjusted Male/Female odds ratio for MDE and a 95% CI for that odds ratio:

```
svyset seclustr [pweight = ncsrwtsh], strata(sestrat)
svy: logistic mde sexm
```

From the output provided by the `svy: logistic` command, the estimated odds ratio and a 95% CI for the population odds ratio are as follows:

$\hat{\psi}_{MDE}$ (SE)	$CI_{.95}(\psi)$
0.597 (0.041)	(0.520, 0.685)

Based on this analysis, the odds that an adult male has experienced a lifetime MDE are only 59.7% as large as the odds of MDE for adult females, which agrees (allowing for some rounding error) with the simple direct calculation. Since the 95% CI does not include $\psi = 1$, we would reject the null hypothesis that MDE status is independent of gender.

6.4.7 Bivariate Graphical Analysis

Graphical displays also are useful tools to describe the bivariate distribution of two categorical variables. The following Stata graphics command generates gender-specific vertical bar charts for the BP_CAT variable generated in Example 6.3 (note that the `pweight` option is used to specify the survey weights, and the `over()` option is used to generate a plot for each level of gender). The output is shown in Figure 6.8.

```
graph bar (mean) bp_cat1 bp_cat2 bp_cat3 bp_cat4 ///
[pweight=wtmec2yr] if age18p==1, blabel(bar, format(%9.1f) ///
color(none)) bar(1,color(gs12)) bar(2,color(gs4)) ///
bar(3,color(gs8)) bar(4,color(black)) ///
bargap(7) scheme(s2mono) over(riagendr) percentages ///
legend (label(1 "Normal") label(2 "Pre-Hypertensive") ///
label(3 "Stage 1 Hypertensive") label(4 "Stage 2 ///
Hypertensive")) ytitle("Percentage")
```

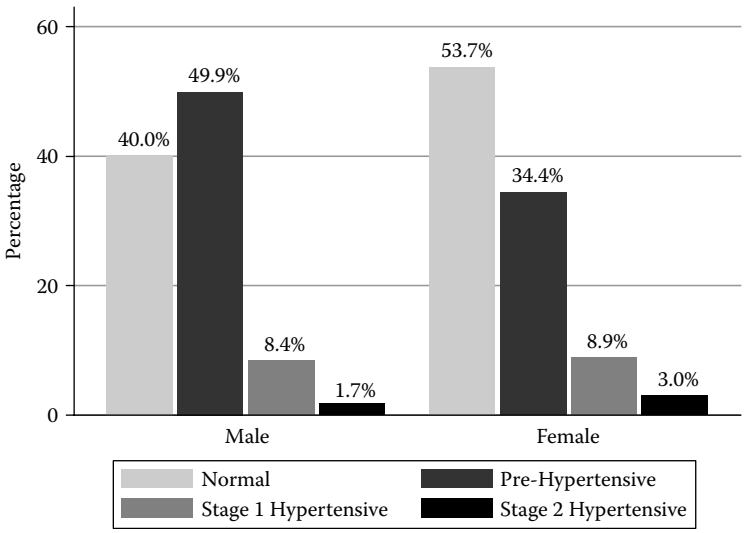


FIGURE 6.8
Bar chart of the estimated distribution of blood pressure status of U.S. adult men and women. (Modified from the 2005–2006 NHANES data.)

6.5 Analysis of Multivariate Categorical Data

During the past 25 years, multivariate analysis involving three (or more) categorical variables has increasingly shifted to regression-based methods for generalized linear models (Chapters 8 and 9). The regression framework provides the flexibility to estimate the association of categorical responses and factors as well as the ability to control for continuous covariates. In this section, we briefly review the adaptation of two long-standing techniques for the analysis of multivariate categorical data to complex sample survey data: (1) the Cochran–Mantel–Haenszel test; and (2) simple log-linear modeling of the expected proportions of counts in multiway tables defined by cross-classifications of categorical variables.

6.5.1 The Cochran–Mantel–Haenszel Test

Commonly used in epidemiology and related health sciences, the CMH test permits tests of association between two categorical variables while controlling for the categorical levels of a third variable. For example, an analyst may be interested in testing the association between a lifetime diagnosis of major depressive episode and gender while controlling for age categories. Although not widely available in software systems that support complex sample survey data analysis, design-based versions of the CMH test are available in SUDAAN's CROSSTAB procedure.

SUDAAN PROC CROSSTAB supports two alternative methods for estimating the adjusted or **common odds ratio** and **common relative risk** statistics: (1) the Mantel–Haenszel (M–H) method; and (2) the logit method. Both methods adjust for the complex sample design and generally result in very similar estimates of common odds ratios and relative risks.

Example 6.10: Using the NCS-R Data to Estimate and Test the Association between Gender and Depression in the U.S. Adult Population When Controlling for Age

In Examples 6.7 to 6.9, we found evidence of a significant overall association between gender and a diagnosis of lifetime depression when analyzing the NCS-R data, where females had greater odds of receiving a diagnosis of depression at some point in their lives. This example is designed to test whether this association holds in the U.S. adult population when controlling for age. Given that the CMH test can be applied when the control variable is a categorical variable, a four-category age variable named AGECAT is constructed: 1 = ages 18–29, 2 = ages 30–39, 3 = ages 40–49, and 4 = ages 50+. The SUDAAN CROSSTAB procedure is then run to derive the CMH test and to use both the Mantel–Haenszel and logit methods to estimate the common age-adjusted male/female odds ratio and relative risk for MDE:

```
proc crosstab ;
nest sestrat seclustr ;
weight ncsrwtsh ;
class agecat sexm mde ;
tables agecat*sexm*mde ;
risk MHOR MHRR1 LOR LRR1 ;
TEST cmh chisq;
print nsum wsum rowper serow colper secol / tests=all
adjrisk=all cmhtest=all ;
run ;
```

Note that the sampling error codes (NEST statement) and the sampling weights (WEIGHT statement) are identified first. Next, the CLASS statement specifies that all three variables are categorical. In the TABLES statement, the AGECAT variable is identified first, defining it as the categorical control variable for this analysis. We define SEXM as the row variable and MDE as the column variable. The RISK statement then requests estimates of the Mantel–Haenszel common odds ratio and relative risk ratio and the logit-based common odds and relative risk ratios. Finally, the TEST statement requests the overall design-adjusted CMH test in addition to the Wald chi-square tests, which will be performed for each age stratum. Table 6.7 summarizes the key elements of the SUDAAN output.

The value of the design-adjusted CMH test statistic in this case is $X^2_{CMH} = 92.46$, which has $p < 0.0001$ on one degree of freedom, suggesting that there is still a strong association between gender and lifetime depression after adjusting for age. This is supported by the design-adjusted Wald chi-square statistics produced by SUDAAN for each age group (see Equation 6.12). In each age stratum, there is evidence of a significant association of gender with lifetime depression. After adjusting for respondent's age, both the M-H and logit estimates of the common male/female odds ratio are about $\hat{\psi} \approx 0.60$. Males appear to have roughly 40% lower odds of having a lifetime diagnosis of depression compared with females

TABLE 6.7

SUDAAN Output for the Cochran–Mantel–Haenszel Test of MDE versus SEX, Controlling for AGECAT

Cochran-Mantel-Haenszel Test Results

$X^2_{CMH} = 94.26$	$df = 1$	$p < 0.0001$

Age Category-Specific Wald Tests of Independence for MDE and SEX

Age Category	18–29	30–39	40–49	50+
Q_{Wald}	23.47	23.22	9.12	38.28
$P(Q_{Wald} > \chi_1^2)$	$p < 0.0001$	$p < 0.0001$	$p < 0.0043$	$p < 0.0001$

Age-Adjusted Estimates of Common Male/Female Odds Ratios and Relative Risks

Statistic	ψ_{M-H}	ψ_{Logit}	RR_{M-H}	RR_{Logit}
Point estimate	0.59	0.60	0.91	0.91
95% CI	(0.51, 0.67)	(0.52, 0.68)	(0.88, 0.93)	(0.89, 0.93)

Source: Analysis based on the NCS-R data.

when adjusting for age. The M–H and logit method estimates of the common risk ratios also suggest a significant difference in the *probability* of depression for the two groups, with the expected probability about 10% lower for males when adjusting for age.

6.5.2 Log-Linear Models for Contingency Tables

A text on the analysis of survey data would not be complete without a mention of log-linear models for multiway contingency tables (Agresti, 2002; Bishop, Feinberg, and Holland, 1975; Stokes, Davis, and Koch, 2002). Log-linear models permit analysts to study the **association structure** among categorical variables. In a sense, a log-linear model for categorical data is analogous to an analysis of variance (ANOVA) model for the cell means of a continuous dependent variable. The dependent variable in the log-linear model is the natural logarithm of the expected counts for cells of a multiway contingency table. The model parameters are estimated effects associated with the categorical variables and their interactions. For example, the following is the log-linear model under the null hypothesis of independence for three categorical variables *X*, *Y*, and *Z*:

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \tag{6.15}$$

where m_{ijk} is the expected cell count under the model.

A model that includes a first-order interaction between the *X* and *Y* variables would be written as

$$\log(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} \tag{6.16}$$

Under simple random sampling assumptions, the cell counts are assumed to follow a Poisson distribution and the model parameters are estimated using the method of maximum likelihood or iterative procedures such **iterative proportional fitting (IPF)**. Tests of nested models (note that the model in Equation 6.15 is nested in the model in Equation 6.16) are performed using the **likelihood ratio test**.

Log-linear models for SRS data can be analyzed in virtually every major software package (e.g., SAS PROC CATMOD). Presently, the major software packages such as Stata, SAS, and SPSS that include programs for analysis of complex sample survey data do not include a program to perform the traditional log-linear modeling. There are likely two explanations for this omission. The first is that the general structure of the input data (grouped cell counts of individual observations) does not lend itself readily to design-based estimation and inference. In Skinner, Holt, and Smith (1989), Rao and Thomas discuss the extension of their design-based adjustment for chi-square test statistics to the conventional likelihood ratio tests for

log-linear models, but this technique would take on considerable programming complexity as the dimension and number of variations on the association structure in the model increases. Grizzle, Starmer, and Koch (GSK; 1969) introduced the weighted least squares method of estimating log-linear and other categorical data models. This generalized technique was programmed in the GENCAT software (Landis et al., 1976) and requires the user to input a design-based variance–covariance matrix for the vector of cell proportions in the full cross-tabular array. Under the GSK method, tests of hypotheses are performed using Wald statistics.

A second explanation for the scarcity of log-linear modeling software for complex sample survey data is that log-linear models for expected cell counts can be reparameterized as logistic regression models (Agresti, 2002) and all of the major software systems have more advanced programs for fitting logistic and other generalized linear models to complex sample survey data. These models will be considered in detail in Chapters 8 and 9.

6.6 Exercises

1. Using the software procedure of your choice and the NCS-R data set, estimate the row proportions in a two-way table where race/ethnicity (RACECAT: 1 = Asian/Other, 2 = Hispanic, 3 = Black, 4 = White) is the factor variable (or row variable) and major depressive episode is the response variable (or column variable). Recall from the Chapter 5 exercises that the sampling error stratum, sampling error cluster, and final sampling weight variables in the NCS-R data set are SESTRAT, SECLUSTR, and NCSRWTSH (Part 1 weight), respectively. Then, answer the following questions:
   a. What is the value of the Rao–Scott F-statistic for the overall test of the null hypothesis that race category and major depressive episode are not associated? Don't forget to report the degrees of freedom for this design-adjusted F-statistic.
   b. Under this null hypothesis, what is the p-value for the F reference distribution?
   c. Based on this test, what is your statistical decision regarding the independence of race and MDE status in the NCS-R survey population?
2. Repeat the analysis from Exercise 1, performing unconditional subclass analyses for men and women separately (SEX: 1 = Male 2 = Female). Do your inferences change when restricting the target population to only men or only women?

3. What proportion of white females in the NCS-R survey population is estimated to have MDE? Compute a 95% CI for this proportion that has been appropriately adjusted for the complex design.
4. Conduct a similar analysis of the association between U.S. REGION (REGION: 1 = Northeast, 2 = North central, 3 = South and 4 = West) and MDE status, estimating the row proportions in a two-way contingency table. Then, answer the following questions:
   a. Is there a significant association between REGION of residence and MDE status in the NCS-R target population? Provide the Rao–Scott design-adjusted *F*-statistic (including the appropriate degrees of freedom) and a *p*-value for the test statistic to support your decision.
   b. What proportion of the NCS-R survey population in the North Central region has a diagnosis of MDE? Provide a point estimate and 95% confidence interval for the proportion. What are the corresponding estimates of the proportion with MDE and the 95% CI for the NCS-R population that resides in the South region?
5. Extend the analysis in Exercise 4 by conducting an analysis of the association between REGION and MDE separately for each of the four race groups (defined by the NCS-R variable RACECAT). Then, answer these questions:
   a. When the race groups are analyzed separately, does the association (or lack thereof) between REGION and MDE continue to hold? Provide the design-adjusted *F*-statistic and *p*-value for each of the race-specific analyses to support your answer.
   b. If the answer to part a is no, how do you explain this pattern of results?

7

Linear Regression Models

7.1 Introduction

> Study regression. All of statistics is regression.

This quote came as a recommendation from a favorite professor to one of the authors while he was in the process of choosing a concentration topic for his comprehensive exam. The broader interpretation of the quote requires placing the descriptor in quotes, "regression," but ask individuals with backgrounds as varied as social science graduate students or quality control officers in a paper mill to decipher the statement and they will think first of the linear regression model. Given the importance of the linear regression model in the history of statistical analysis, the emphasis that it receives in applied statistical training and its importance in real-world statistical applications, the narrower interpretation is quite understandable.

This chapter introduces linear regression modeling for complex sample survey data—its similarities to and how it differs (theoretically and procedurally) from standard ordinary least squares (OLS) regression analysis. We assume that the reader is familiar with the basic theory and methods for simple (single-predictor) and multiple (multiple-predictor) linear regression analysis for continuous dependent variables. Readers interested in a comprehensive reference on the topic of linear regression are referred to Draper and Smith (1981), Kleinbaum, Kupper, and Muller (1988), Neter et al. (1996), DeMaris (2004), Faraway (2005), Fox (2008), or many other excellent texts on the subject.

Focusing on practical approaches for complex sample survey data, we emphasize "aggregated" design-based approaches to the linear regression analysis of survey data (sometimes referred to as population-averaged modeling), where design-based variance estimates for weighted estimates of regression parameters in finite populations are computed using nonparametric methods such as the Taylor series linearization (TSL) method, balanced repeated replication (BRR), or jackknife repeated replication (JRR). Model-based approaches to the linear regression analysis of complex sample survey data, which may explicitly include stratification or clustering effects

in the regression models and may or may not use the sampling weights (e.g., Skinner, Holt, and Smith, 1989; Pfefferman et al., 1998), are introduced in Chapter 12. Over the years, there have been many contributions to the survey methodology literature comparing and contrasting these two approaches to the regression analysis of survey data, including papers by DuMouchel and Duncan (1983), Hansen, Madow, and Tepping (1983), and Kott (1991).

We present a brief history of important statistical developments in linear regression analysis of complex sample survey data to begin this chapter. Kish and Frankel (1974) were two of the first to empirically study and discuss the impact of complex sample designs on inferences related to regression coefficients. Fuller (1975) derived a linearization-based variance estimator for multiple regression models with unequal weighting of observations and introduced variance estimators for estimated regression parameters under stratified and two-stage sampling designs. Shah, Holt, and Folsom (1977) further discussed the violations of standard linear model assumptions when fitting linear regression models to complex sample survey data, discussed appropriate methods for making inferences about linear regression parameters estimated using survey data, and presented an empirical evaluation of the performance of variance estimators based on Taylor series linearization.

Binder (1983) focused on the sampling distributions of estimators for regression parameters in finite populations and defined related variance estimators. Skinner et al. (1989, Sections 3.3.4 and 3.4.2) summarized estimators of the variances for regression coefficients that allowed for complex designs (including linearization estimators) and recommended the use of linearization methods or other robust methods (e.g., JRR) for variance estimation. Kott (1991) further discussed the advantages of using variance estimators based on Taylor series linearization for estimates of linear regression parameters: protection against within-PSU correlation of random errors, protection against possible nonconstant variance of the random errors, and the fact that a within-PSU correlation structure does not need to be identified to have a nearly unbiased estimator. Fuller (2002) provided a modern summary of regression estimation methods for complex sample survey data.

7.2 The Linear Regression Model

Regression analysis is a study of the relationships among variables: a **dependent variable** and one or more **independent variables**. Figure 7.1 illustrates a simple linear regression model of the relationship of a dependent variable, y, and a single independent variable x. The regression relationship among the observed values of y and x is expressed as a regression model, for example, $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the dependent variable, x is the independent

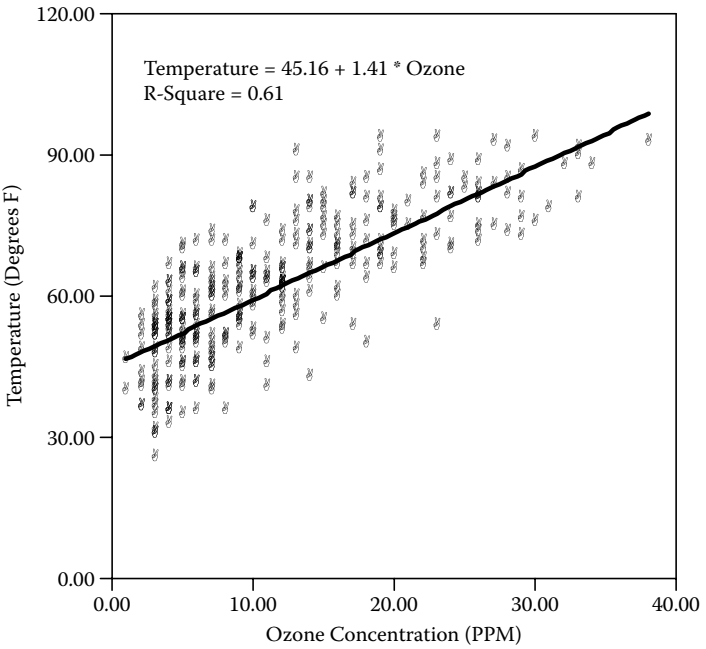


FIGURE 7.1
Linear regression of air temperature on ozone level.

variable, β_0 and β_1 are model parameters, and ε is an error term that reflects the difference between the observed value of y and its conditional expectation under the model, $\varepsilon = y - \hat{y} = y - \beta_o - \beta_1 x$.

In statistical practice, a fitted regression model may be used to simply predict the expected outcome for the dependent variable based on a vector of independent variable measurements $\boldsymbol{x}$, $E(y \mid \boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$, or to explore the functional relationship of y and x. Across the many scientific disciplines that use regression analysis methods, dependent variables may also be referred to as response variables, regressands, outcomes, or even "left-hand-side variables." Independent variables may be labeled as predictors, regressors, covariates, factors, cofactors, explanatory variables or "right-hand-side variables." We primarily refer to response variables and predictor variables in this chapter, but other terms can be used interchangeably.

This chapter will focus on the broad class of regression models known as **linear models**, or models for which the conditional expectation of y given x, $E(y \mid x)$, is a linear function of the unknown parameters. Consider the following three specifications of linear models:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{7.1}$$

Note in this model that the dependent variable, y, is a linear function of the unknown parameters and the independent variable x:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{7.2}$$

In this model (Equation 7.2), the response variable y is still a linear function of the β parameters for x and x^2; however, the linear model defines a *nonlinear* relationship between y and x:

$$\begin{aligned} y &= \boldsymbol{x}\beta + \varepsilon \\ &= \sum_{j=0}^{p} \beta_j x_j + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \end{aligned} \tag{7.3}$$

Here the linear model is first expressed in **vector notation**. Vector notation may be used as an abbreviation to represent a complex model with many parameters and to facilitate computations using the methods of matrix algebra.

When specifying linear regression models, it is useful to be able to reference specific observations on the subjects in a survey data set:

$$y_i = \boldsymbol{x_i}\beta + \varepsilon_i \tag{7.4}$$

where $\boldsymbol{x_i} = [1 \quad x_{1i} \quad \ldots \quad x_{pi}]$ and $\beta^{\boldsymbol{T}} = [\beta_0 \quad \beta_1 \quad \ldots \quad \beta_p]$.
In this notation, i refers to sampled element (or respondent) i in a given survey data set.

7.2.1 The Standard Linear Regression Model

Standard procedures for unbiased estimation and inference for the linear regression model involve the following assumptions:

1. The model for $E(y \mid x)$ is linear in the parameters (see Equation 7.2).
2. Correct model specification—in short, the model includes the true main effects and interaction terms to accurately reflect the true model under which the data were generated.
3. $E(\varepsilon_i \mid \boldsymbol{x_i}) = 0$, or that the expected value of the residuals given a set of values on the predictor variables is equal to 0.
4. Homogeneity of variance: $Var(\varepsilon_i \mid \boldsymbol{x_i}) = \sigma^2_{y \cdot x}$, or that the variance of the residuals given values on the predictor variables is a constant parameter equal to $\sigma^2_{y \cdot x}$.

5. Normality of residuals (and also y): for continuous outcomes, we assume that $\varepsilon_i \mid x_i \sim N(0, \sigma^2_{y \cdot x})$, or that given values on the predictor variables, the residuals are independently and identically distributed (i.i.d.) as normal random variables with mean 0 and constant variance $\sigma^2_{y \cdot x}$.
6. Independence of residuals: As a consequence of the previous point, $Cov(\varepsilon_i, \varepsilon_j \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = 0$, $i \neq j$, or residuals on different subjects are *uncorrelated* given values on their predictor variables.

There are several implications of these standard model assumptions. First, we can write

$$\hat{y} = E(y \mid \boldsymbol{x}) = E(\boldsymbol{x}\beta) + E(\varepsilon) = \boldsymbol{x}\beta + 0 = \boldsymbol{x}\beta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{7.5}$$

This equation for the predicted value of y is the **regression function**, or the expected value of the dependent variable y conditional on a set of values on the predictor variables (of which there are p). Further, we can write

$$Var(y_i \mid \boldsymbol{x_i}) = \sigma^2_{y \cdot x} \tag{7.6}$$

$$Cov(y_i, y_j \mid \boldsymbol{x_i}, \boldsymbol{x_j}) = 0 \tag{7.7}$$

These assumptions, therefore, imply that the dependent variable has constant variance given values on the predictors and that no two values on the dependent variable are correlated given values on the predictors. Putting all of the implications together, we have

$$y_i \sim N(\boldsymbol{x_i}\beta, \sigma^2_{y \cdot x}) \tag{7.8}$$

Values on the dependent variable, y, are therefore assumed to be i.i.d. normally distributed random variables with a mean defined by the linear combination of the parameters and the predictor variables and a constant variance.

7.2.2 Survey Treatment of the Regression Model

Since the late 1940s and early 1950s when economists and sociologists (Kendall and Lazarsfeld, 1950; Klein and Morgan, 1951) first applied regression analysis to complex sample survey data, survey statisticians have sought to relate design-based estimation of regression relationships to the standard linear model. The result was the linked concepts of a finite population and the superpopulation model, which are described in more detail in Chapter 3 (also see Theory Box 7.1).

THEORY BOX 7.1 FINITE POPULATIONS AND SUPERPOPULATION MODELS

In theory, weighted estimation of a linear regression model, $y = \beta_0 + \beta_1 x + \varepsilon$, from complex sample survey data results in unbiased estimates of the regression function, $E(y \mid x) = B_0 + B_1 x$, where $B = [B_0, B_1]$ are **finite population regression parameters**. If instead of observing a sample of size n of the N population elements a complete census had been conducted, the finite population regression parameter B_1 for this simple "line" could be computed algebraically as follows:

$$B_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

The theory suggests that we distinguish between the true regression model parameters, β, and the particular values, ***B***, that characterize the current finite survey population that was sampled for the survey. For example, consider a simple linear regression model of the effect of years of education on adults' earned income. The model can be interpreted as

$$E(Income \mid Education) = \beta_0 + \beta_1 \cdot Education(years)$$

or

$$Income \mid Education \sim N(\beta_0 + \beta_1 \cdot Education(years), \sigma^2_{I \cdot E})$$

In contrast, a strict finite population regression interpretation is that for the particular population that has been sampled, the best prediction of income, given an adult's education level, is found on the line $E(Income|Education) = B_0 + B_1 \cdot Education(years)$. The concept of a **superpopulation** links these two interpretations by introducing the assumption that although the surveyed population is finite (size N), the individual relationship between income and education in that finite population conforms to an underlying superpopulation model. Therefore, the fixed set of pairs of income and education values in the finite population of N elements is itself a sample from an infinite number of possible data pairs that could be generated by a stochastic superpopulation model, denoted by $\zeta: Income = \beta_0 + \beta_1 \cdot Education(years) + \varepsilon$.

Under the superpopulation model, the finite population parameters $\boldsymbol{B}$ will vary about the true model parameters β. However, the bias of making inferences for β based on estimates, $\hat{\boldsymbol{B}}$ (which are unbiased for $\boldsymbol{B}$), is small, and of the order $O(N^{-1/2})$ (meaning that the bias is a function of $N^{-1/2}$, and will thus become small as the population size becomes larger). Therefore, for large populations, an unbiased estimate of $\boldsymbol{B}$ (generally computed using weighted least squares) can serve as an unbiased estimate of β. Model diagnostics can then be used to question the hypothesized superpopulation model or the suitability of $\boldsymbol{B}$ as a summary measure of the relationships.

When survey analysts conduct a regression analysis using complex sample survey data, they choose between two targets—the finite population or a more universal superpopulation model—for their estimation and associated inference. Sometimes this is a conscious choice, and sometimes the choice is less conscious and implicit only in the analyst's presentation of the results and the inferences that are drawn.

In theory, weighted, design-based estimation of a regression equation permits the survey analyst to make inferences concerning values of the regression relationship as it exists within the finite population that corresponds to the geographic, demographic, and temporal definition of the survey population. To extend this inference beyond the survey population requires the analyst to make generalizing assumptions about the relationship of the finite population that has been studied to an overarching superpopulation model that may govern the relationships of the variables of interest. Theory Box 7.1 discusses the relationship of the superpopulation model and the finite population regression function in more detail. In practice, if the survey population is large and the regression model is correctly identified, the survey data analyst may treat these two conceptual approaches as virtually equivalent (Korn and Graubard, 1999; Skinner et al., 1989).

7.3 Four Steps in Linear Regression Analysis

There are four basic steps that analysts should follow when fitting regression models to complex sample survey data: specification, estimation, evaluation (diagnostics), and inference. These steps apply to all types of regression models and not just those discussed in this chapter for continuous response variables. The following sections describe each step, considering the standard approach first followed by the adaptation of the step for complex sample survey data.

7.3.1 Step 1: Specifying and Refining the Model

Survey data are typically observational data. The process of initial specification and subsequent refinement of a regression model for survey data involves multiple iterations of the four-step process. At the beginning of each cycle in this iterative process, it is important for the survey analyst to step back from the "number crunching" and critically evaluate the scientific interpretation and the plausibility of the emerging model.

A model is initially postulated based on subject matter knowledge and empirical investigation of the data. The specific aims of the analysis will often determine the choice of the dependent variable, y, and one or more independent variables of particular interest, $\boldsymbol{x}$. Scientific subject matter knowledge and information gleaned from prior studies and publications can be used to identify additional independent variables (i.e., covariates that are known predictors of the dependent variable) or variables that may **mediate** or **moderate** (DeMaris, 2004) the relationships of the independent variables of primary interest with the dependent variable y. For example, an epidemiologist aiming to model the effect of obesity on systolic blood pressure (BP; in mmHg) decides to include age, gender, and race of the respondent as additional covariates. Based on prior research conducted by a colleague, she has evidence that advanced age moderates the relationship between systolic blood pressure level and obesity. She will test for an interaction between age and the obesity measure (as well as other potential interactions—see Section 7.4).

Empirical results from simple descriptive and graphical statistical analysis of the survey data itself can also be used to identify independent variable candidates for the model. The epidemiologist in our example might conduct an exploratory analysis, plotting systolic BP against respondent age. At older ages, the resulting scatterplot shows a curvilinear relationship to systolic BP, suggesting the addition of a quadratic term to the initial model. See Cleveland (1993) for a good resource on data exploration.

Contemporary survey data sets may contain hundreds of variables, and there is a temptation to bypass the scientific review and empirical investigations and "see what works." Regression programs in statistical software packages often include variable selection algorithms such as **stepwise regression**, **forward selection,** and **backward selection** that are capable of culling a set of significant predictors from a large input of independent variable choices. These algorithms may prove useful in the model exploration and fitting process, but they are numerical tools and should not substitute for the survey analyst's own scientific and empirical assessment of the model and its final form. Careless use of such techniques leaves the analyst exposed to problems of **confounding** or **spurious relationships** that can distort the model and its interpretation.

A variety of variable selection and model-building approaches have been proposed for linear regression models. One such model-building "recipe"

commonly used in our teaching and consulting practice is described in Section 7.4.5.

7.3.2 Step 2: Estimation of Model Parameters

After the survey data analyst has specified a linear regression model (Step 1), the next step in the modeling process involves computation of estimates of the regression parameters in the specified model. This section describes mathematical methods that can be used for estimation of those parameters.

7.3.2.1 Estimation for the Standard Linear Regression Model

By far, the most popular method of estimating unknown parameters in linear regression models is **ordinary least squares estimation**. This method focuses on estimating the unknown set of regression parameters β in a specified model by minimizing the residual sum of squares (or sum of squared errors, SSE) based on the model

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}})^2 \tag{7.9}$$

Once the estimate of β has been obtained analytically (see Equation 7.11), an estimate of the variance of the random errors in the model, $\sigma^2_{y \cdot x}$, is obtained as follows:

$$\hat{\sigma}^2_{y \cdot x} = \frac{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}})^2}{n - (p+1)} \tag{7.10}$$

Here, $p + 1$ is the number of regression parameters in the specified model.

The least squares estimate has several important properties. First, parameter estimates and their variances and covariances are analytically simple to compute, requiring only a single noniterative algebraic or matrix algebra computation:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \\ \text{var}(\hat{\boldsymbol{\beta}}) &= \hat{\Sigma}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \hat{\sigma}^2_{y \cdot x} \end{aligned} \tag{7.11}$$

Second, the estimator is unbiased:

$$E(\hat{\boldsymbol{\beta}} \mid X) = \boldsymbol{\beta} \tag{7.12}$$

Third, the estimator has the lowest variance among all other unbiased estimators that are also linear functions of the response values, making it the **best linear unbiased estimator (BLUE)**. Finally, assuming normally distributed errors, the least squares estimates are equal to estimates derived based on maximum likelihood estimation.

As described in Section 7.2, one key assumption of the standard linear regression model is homogeneity of the error variance:

$$Var(\varepsilon) = Var(y_i \mid \boldsymbol{x_i}) = \sigma^2_{y \cdot x} = \text{constant} \tag{7.13}$$

In practice, it is common to find that the variance of the residuals is heterogeneous—varying over the $i = 1, \ldots, n$ cases with differing values of y and $\boldsymbol{x}$. For OLS estimation, the consequence of heterogeneity of variance is loss of efficiency (larger standard errors) in the estimation of the regression coefficients. **Weighted least squares (WLS) estimation** of the regression coefficients addresses this inefficiency by weighting each sample observation's contribution to the sums of squares by the reciprocal of its residual variance, $W_i = 1 / \sigma^2_{y \cdot x, i}$. In matrix notation, the WLS estimator of the linear regression coefficients is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \tag{7.14}$$

Here, $\boldsymbol{W}$ is an $n \times n$ diagonal matrix (with zeroes off the diagonal and the n values of the inverse variance weights on the diagonal).

The standard linear regression model and the OLS estimator are statistically elegant, but the underlying assumptions are easily violated when analyzing real-world data. To minimize the mean square error of estimation, techniques such as transformation of the dependent variable, WLS estimation, and other approaches such as ridge regression (Hoerl and Kennard, 1970) and robust variance estimation (Fuller et al., 1986; Judge et al., 1985) have been developed to address problems of nonnormality, heterogeneity of variances, collinearity of predictors, and correlated errors.

7.3.2.2 Linear Regression Estimation for Complex Sample Survey Data

Estimation of regression relationships for complex sample survey data alters the standard approach to estimation of coefficients and their standard errors. We first discuss what changes with estimation of the parameters and then address what changes in terms of variance estimation.

7.3.2.2.1 Estimation of Parameters

The observed data from a complex sample survey are typically not "identically distributed." Due to variation in sample selection and sample inclusion

probabilities, survey weights must generally be employed to develop unbiased estimates of the population regression parameters. Recall that in standard methods of regression analysis, WLS estimation incorporates a weight for each sample element inversely proportional to the residual variance. In the context of fitting regression models to complex sample survey data sets, where sampling weights have been calculated to compensate for unequal probability of selection, unit nonresponse, and possibly poststratification (see Section 2.7), the sampling weights can be incorporated into the estimation of the regression parameters via the use of WLS estimation. The contribution of each case to the residual sum of squares is made proportional to its population weight. This results in the following analytic formula for the weighted least squares estimate of the finite population regression parameters:

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y} \tag{7.15}$$

Here, $\boldsymbol{W}$ is an $n \times n$ diagonal matrix (with zeroes off the diagonal and the n values of the sampling weights on the diagonal). Theory Box 7.2 provides a more mathematical motivation for the weighted survey estimator of $\boldsymbol{B}$.

THEORY BOX 7.2 A WLS ESTIMATOR FOR FINITE POPULATION REGRESSION MODELS

When fitting regression models to complex sample survey data collected from a finite population, we choose estimates of the finite population parameters $\boldsymbol{B}$ that minimize the following objective function:

$$f(\boldsymbol{B}) = \sum_{i=1}^{N} (y_i - \boldsymbol{x_i B})^2$$

We can think of this objective function $f(\boldsymbol{B})$ as a finite population "residual" sum of squares, SSE_{pop}. An unbiased sample estimate of this total incorporating the sampling weights can be written as follows:

$$\hat{WSSE}_{pop} = \sum_{h}^{H} \sum_{\alpha}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} (y_{h\alpha i} - \boldsymbol{x_{h\alpha i}} \hat{\boldsymbol{B}})^2$$

In this expression, h is a stratum index, α is a cluster (or primary sampling unit) index, and i is an index for elements within the α-th cluster. Weighted least squares estimation is used to derive estimates of $\boldsymbol{B}$ that minimize this sample estimate of the actual objective function for the finite population.

Therefore, although the motives for conventional WLS estimation (heterogeneity of residual variances) and weighted survey estimation (unbiased population representation) are very different, the WLS computational algorithms that have been built into regression software for several decades serve perfectly well for weighted survey estimation of the finite population regression model. Consequently, analysts who naïvely specified the survey weight as the weight variable in a standard linear regression program (e.g., SAS PROC REG, `regress` in Stata, and SPSS Linear Regression) have obtained the correct design-based estimates of the population regression parameters. Unfortunately, the naïve use of weighted survey estimation in standard linear regression programs generally results in biased estimates of standard errors for the parameter estimates (see next section). Stata explicitly recognizes the differences in weighting concepts and requires the user to declare probability weights, or "pweights." Stata also uses a robust estimator of variance when a pweight is specified.

Correct interpretation of estimated regression parameters is essential for effectively communicating the results of investigations involving regression analyses in scientific publications. In a simple linear regression model for a continuous dependent variable, after obtaining estimates of the regression parameters, we can calculate the following expected values on the response variable, associated with a one-unit change in a given predictor variable x:

$$E(y \mid x = x_0) = \hat{B}_0 + \hat{B}_1 x_0 \tag{7.16}$$

$$E(y \mid x = x_0 + 1) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_1 \tag{7.17}$$

We can therefore write the following about the estimate of the regression parameter β_1:

$$\hat{B}_1 = E(y \mid x = x_0 + 1) - E(y \mid x = x_0) \tag{7.18}$$

The parameter estimate therefore describes, on average, the expected change in the continuous response variable y for a one-unit change in the predictor variable x.

When an additional predictor variable z is added to the model, representing a theoretical control variable or possibly a confounding variable, a portion of the relationship between x and y is attributable to the variable z, and the interpretation of the regression parameter for the predictor variable x requires that the predictor variable z be fixed at a constant value, z_0:

$$E(y \mid x = x_0, z = z_0) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_2 z_0 \tag{7.19}$$

$$E(y \mid x = x_0 + 1, z = z_0) = \hat{B}_0 + \hat{B}_1 x_0 + \hat{B}_1 + \hat{B}_2 z_0 \tag{7.20}$$

$$\hat{B}_1 = E(y \mid x = x_0 + 1, z = z_0) - E(y \mid x = x_0, z = z_0) \tag{7.21}$$

We therefore interpret the estimate of the parameter B_1 in this case as the expected difference in y associated with a one-unit increase in x, holding the value of the predictor variable z constant. When multiple control variables are added to a linear regression model, we hold all of them at fixed values when interpreting the regression parameter for a primary predictor of interest. The interpretation of estimated regression parameters does not change at all when analyzing sample survey data sets collected from finite populations, aside from the fact that the regression parameters are describing relationships in a finite population of interest. We will consider interpretations of regression parameters in detail in all examples presented in this chapter.

7.3.2.2.2 Estimation of Variances of Parameter Estimates

As described in Chapter 3, the complex nature of most sample survey data (stratification, clustering, unequal selection probabilities) precludes the use of conventional variance estimators that can be derived for maximum likelihood estimation (MLE) for data that are presumed to be i.i.d. draws from a probability distribution (i.e., normal, binomial, Poisson). Instead, robust, nonparametric methods based on the TSL of the estimator or replication variance estimation methods (BRR, JRR, bootstrap) are employed.

The general approach to TSL variance estimation for linear regression coefficients can be illustrated for the simple linear regression model (which involves a single predictor variable). Extension of the method to multiple linear regression is straightforward but algebraically more complex. In the case of a simple linear regression model with a single predictor x, an analytic formula for the calculation of the associated finite population regression parameter B (given all data for the finite population with size N) can be written as follows:

$$B = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} x_i^2} = \frac{T_{xy}}{T_{x^2}} \tag{7.22}$$

This formula can be written as a ratio of two totals: T_{xy} and T_{x^2}. We can calculate an estimate of this ratio by applying the survey weights to the observed sample data:

$$\hat{B} = \frac{\sum_{h}\sum_{\alpha}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} y_{h\alpha i} x_{h\alpha i}}{\sum_{h}\sum_{\alpha}\sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} x_{h\alpha i}^{2}} = \frac{t_{xy}}{t_{x^2}} \tag{7.23}$$

Note that the sample estimate $\hat{B}$ is also a ratio of sample totals. Under the TSL approximation method, an estimate of the sampling variance of this ratio of two sample totals can be written as follows:

$$\text{var}(\hat{B}) \cong \frac{\text{var}(t_{xy}) + \hat{B}^2\,\text{var}(t_{x^2}) - 2\hat{B}\,\text{cov}(t_{xy}, t_{x^2})}{(t_{x^2})^2} \tag{7.24}$$

In multiple linear regression, TSL approximation methods require weighted sample totals for the squares and cross-products of all of the y and $\boldsymbol{x} = \{1\ x_1 \ldots x_p\}$ combinations. The computations are more complex, but the approach is a direct extension of the technique shown here for a simple linear regression model. Replication methods like JRR or BRR (see Section 3.6.3) can also be used to estimate sampling variances of estimated regression parameters (Kish and Frankel, 1974).

Statistical software designed for regression analysis of complex sample survey data applies the TSL, BRR, or JRR methods to estimate the sampling variance of each parameter estimate, $\text{var}(\hat{B}_j)$, $j = 0, \ldots, p$, as well the $p(p + 1)/2$ unique covariances, $\text{cov}(\hat{B}_j, \hat{B}_k)$, between the parameter estimates. These estimates of sampling variances and covariances are then assembled into the estimated **variance–covariance matrix** of the parameter estimates:

$$\mathbf{var}(\hat{\boldsymbol{B}}) = \hat{\Sigma}(\hat{\boldsymbol{B}}) = \begin{bmatrix} \text{var}(\hat{B}_0) & \text{cov}(\hat{B}_0, \hat{B}_1) & \cdots & \text{cov}(\hat{B}_0, \hat{B}_p) \\ \text{cov}(\hat{B}_0, \hat{B}_1) & \text{var}(\hat{B}_1) & \cdots & \text{cov}(\hat{B}_1, \hat{B}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\hat{B}_0, \hat{B}_p) & \text{cov}(\hat{B}_1, \hat{B}_p) & \cdots & \text{var}(\hat{B}_p) \end{bmatrix} \tag{7.25}$$

The estimated variances and covariances can then be used to develop Student t-statistics and Wald chi-square or Wald F-statistics required to test hypotheses concerning the population values of the regression parameters (see Section 7.3.4).

A variety of software procedures exist in statistical software packages for fitting linear regression models to survey data using the methods discussed in this section; we consider procedures in some of the more common general-

purpose statistical software packages here. In SAS, PROC SURVEYREG can be used; SPSS offers the general linear model procedure in the Complex Samples Module; Stata offers the `svy: regress` command, which will be considered in this book; and SUDAAN offers PROC REGRESS. See Appendix A for more details on software options for linear regression analysis of survey data.

7.3.3 Step 3: Model Evaluation

The standard linear regression model is an "elegant" statistical tool, but its simplicity and "best" properties hinge on a number of model assumptions (see Section 7.2). Standard texts on linear regression modeling (see Neter et al., 1996) provide detailed coverage of procedures to evaluate the model goodness of fit (GOF), examine how closely the data match the basic model assumptions, and determine if certain observations are unduly influencing the fit of the model. The process does not change substantially when fitting finite population models to complex sample survey data sets. In this section, we briefly consider some of the model diagnostics that can be used to evaluate the model and discuss some current work on model diagnostics adapted for regression models fitted to complex sample survey data.

7.3.3.1 Explained Variance and Goodness of Fit

A standard measure of the "fit" of the regression model to the data is the **coefficient of multiple determination** or the **R^2 statistic**, which is interpreted as the proportion of variance in the dependent variable explained by regression on the independent variables:

$$R^2 = 1 - \frac{SSE}{SST} \tag{7.26}$$

In Equation 7.26, SST refers to the total sum of squares, or the sum of squared differences between the response values and the mean of the response variable, and SSE is given in Equation 7.9. The use of R^2 as a measure of explained variance carries forward to regression modeling of complex sample survey data, although the statistic that is output by the analysis software is actually a *weighted* version, where each squared difference contributing to the sums is weighted by the corresponding sampling weight:

$$R^2_{weighted} = 1 - \frac{WSSE}{WSST} \tag{7.27}$$

Although in theory it could be argued that this weighted R^2 statistic estimates the proportion of population variance explained by the population

regression of y on $\boldsymbol{x}$, in practice it is safe to simply view it as the fraction of explained variance in y attributable to the regression on $\boldsymbol{x}$. Analysts who are new to regression modeling of social science, education, or epidemiological data should not fret if the achieved R^2 values are lower than those seen in their textbook training. Physicists may be disappointed with $R^2 < 0.98$–0.99 and chemists with $R^2 < 0.90$, but social scientists and others who work with human populations will find that their best regression model will often explain only 20%–40% of the variation in the dependent variable.

7.3.3.2 Residual Diagnostics

In the standard regression context, analysis of the distributional properties of the residual terms, $\varepsilon_i = (y_i - \hat{y}_i)$, is used to evaluate how well the assumptions of the normal linear model are met (see Section 7.2). Despite the theoretical distinction between the concept of a finite population regression model and a broader superpopulation model, we recommend using standard residual analysis to evaluate regression models that are fitted to complex sample survey data.

7.3.3.3 Model Specification and Homogeneity of Variance

Two-way scatterplots of the residuals for the estimated model against the predicted values, $\hat{y}$, and the independent variables, x_j, can identify problems with lack of correct functional form (e.g., omitting a squared term for age when modeling blood pressure) or where a moderating variable or interaction has not been correctly included in the model.

These same plots may be used to diagnose a problem with heterogeneity of residual variances. A pattern of residuals that spreads out in a fan-shaped pattern with increasing values for the predicted y or increasing values of an independent variable is a common observation. To address the problem of heterogeneity of variance and to reduce standard errors for the estimated βs, a standard approach in regression analysis is to employ weighted least squares, weighting each observation's contribution to the sum of squared errors inversely proportional to its residual variance, that is, $w_i^{(wls)} = 1/\sigma^2_{y \cdot x_i}$. With complex sample survey data, this becomes complicated, because estimating equations for finite population regression parameters, $\boldsymbol{B}$, already include the survey weight factors, say $w_i^{(survey)}$. While it is possible to create a composite weight, $w_i^* = w_i^{(survey)} \cdot w_i^{(wls)}$, this approach would seem to only further complicate our interpretation of the fitted regression model. In cases of serious heterogeneity of variance, it may be possible to identify a transformation of the dependent or independent variables (see next section) that eliminates much of the residual variance heterogeneity but still permits the use of the survey weights in the estimation of the regression function.

7.3.3.4 Normality of the Residual Errors

This assumption can be assessed using standard diagnostic plots for the model-based residuals (e.g., normal quantile–quantile (Q–Q) plots or histograms). The Q–Q plot is a plot of quantiles for the observed residuals against those computed from a theoretical normal distribution having the same mean and variance as the distribution of observed residuals. A straight 45° line in this plot would therefore suggest that normality is a reasonable assumption for the random errors in the model. We present examples of these Q–Q plots in the application later in this chapter. In large survey data sets, formal tests of the normality of the residuals (e.g., the Kolmogorov–Smirnov and Shapiro–Wilk tests) tend to be extremely "powerful," and the null hypothesis of normality will be rejected due to the slightest deviation from normality. Our recommendation is to use the more informal visual methods illustrated in Section 7.5. Empirical research has provided evidence that as long as the residuals display symmetry about $E(e_i) = 0$ the regression estimates are quite robust against failure of strict normality. If the residual distribution is highly skewed or irregular (e.g., bimodal) the analyst should first determine that the model has been correctly specified (see Section 7.4.1).

Transformation of the dependent variable is a common method to address serious problems of nonnormality of residuals and also heterogeneity of residual variances. Analysts should be careful when making transformations, however, because they can destroy the straightforward interpretation of the parameters previously discussed. A common transformation that is often used when violations of normality are apparent and residual distributions appear to be right-skewed is the natural (base e) log transformation of the response variable:

$$\ln(y) = B_0 + B_1 x + e \tag{7.28}$$

When this particular transformation is used, the parameters still have a somewhat straightforward interpretation:

$$\frac{E(y \mid x = x_0 + 1)}{E(y \mid x = x_0)} = \frac{e^{(B_0 + B_1 x_0 + B_1)}}{e^{(B_0 + B_1 x_0)}} = e^{B_1} \tag{7.29}$$

That is, a one-unit change in a given predictor variable will *multiply* the expected response by $\exp(B_1)$. The important issue to keep in mind when transforming the dependent variable is that predicted values on the response variable need to be back-transformed to the original scale of the response. For example, square root transformations are often used to stabilize variance of the residuals, and predicted values based on this type of model would need to be squared to return to the original scale of the response.

7.3.3.5 Outliers and Influence Statistics

Finally, analysts should determine if the sample data include **outlier values** (observations poorly fitted by a given model) or **influential points** (observations that have a strong impact on the fit of a model). Tools such as standard residual plots, **studentized residuals**, **hat statistics** (measures of leverage), and **Cook's distance (D) statistics** have been developed for this purpose. Regression models can be fitted with and without potential outliers and influential points to assess whether the fit of the model (i.e., the significance of regression parameters) is changing substantially depending on the points in question. For more detail on these diagnostic methods in the standard linear regression case, we refer readers to Neter et al. (1996) or Faraway (2005).

Influential points should still be investigated for their impact on the fit of a model when fitting linear regression models to complex sample survey data sets. What does change is the way that sampling weights are incorporated when calculating the various statistics (e.g., Cook's D statistics) used for residual diagnostics. Li and Valliant (2006) discussed identification of influential points when sampling weights are involved in the estimation of linear regression models, and work by Li (2007) discussed extensions of the adjustments to diagnostic statistics for samples involving stratification and clustering. Li and Valliant (2009) provide an extensive review of the literature on methods for incorporating sampling weights into the calculation of hat matrices and leverage statistics, and current literature suggests that sampling weights are most important for assessing the influence of individual observations on model fit. Unfortunately, to date, these promising methods for producing weighted versions of commonly used diagnostic statistics have not been incorporated into any general-purpose software packages containing procedures for linear regression analysis of complex sample survey data.

It is our hope that the very near future will bring these advances in the software, and we will provide readers with updates in this area on the book's Web site. Until that occurs, we recommend using available software tools for regression diagnostics when fitting models to complex sample survey data, and we will consider examples of such diagnostics in the application presented at the end of the chapter.

This evaluation step of regression analysis may lead to a modification of the model structure. One then cycles back through the model fitting and model diagnosis process, with careful consideration of model structure and parsimony.

7.3.4 Step 4: Inference

After following the previous three steps, the final step in the model-building process is making inferences about the regression parameters in the finite population of interest. This section describes that process.

THEORY BOX 7.3 MODEL EVALUATION FOR COMPLEX SAMPLE SURVEY DATA

When fitting regression models to complex sample survey data collected from a finite population, outlier statistics used for model evaluation are simply adapted by replacing unweighted values with weighted values. This theory box presents mathematical expressions for some of these adaptations, which are not yet widely implemented in general purpose statistical software packages. To keep the expressions as general as possible, we use notation from generalized linear model theory. Generalized linear models will be discussed in more detail in Chapters 8 and 9.

First, we consider computation of **Pearson residuals**:

$$r_{p_i} = (y_i - \mu_i(\hat{\boldsymbol{B}}_w))\sqrt{\frac{w_i}{V(\mu_i)}}$$

where μ_i is the expected value of the outcome, y, for sampled case i, computed as a function of the weighted estimates of the regression parameters; w_i is the sampling weight for the i-th case based on the complex design; and $V(\mu_i)$ is the **variance function** for the outcome, which partly defines the variance of the outcome variable in a generalized linear model as a function of the expected value of the outcome.

The **hat matrix** is computed as

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}$$

where

$$\boldsymbol{W} = diag\left\{\frac{w_1}{V(\mu_1)[g'(\mu_1)]^2}, \ldots, \frac{w_n}{V(\mu_n)[g'(\mu_n)]^2}\right\}$$

This matrix is used for the computation of various diagnostic statistics and, specifically, measures of leverage. This matrix is an $n \times n$ diagonal matrix, with zeroes off the diagonal and diagonal elements defined by the sampling weights divided by a term that is a function of the variance function for the observation and the derivative of the **link function** g for the specific generalized linear model (in a linear regression model, the link function is the identity function). Note that the derivative of the link function is computed for the diagonal elements and is evaluated as a function of the expected value of the outcome according to the model. If a **canonical link** is used to define the

generalized linear model (which is often the case in practice), the diagonal elements simplify to

$$\boldsymbol{W} = diag(w_1 V(\mu_1), ..., w_n V(\mu_n)).$$

Diagonal elements of the hat matrix (denoted by h_{ii}) updated with the sampling weights are used to identify influential cases with high leverage, and a common rule of thumb is to identify diagonal elements larger than $2(p + 1)/n$ or to find large gaps in leverage values. Removing cases with high leverage will generally have little effect on estimates of regression parameters but will have a large impact on the uncertainty of the estimates (i.e., standard errors).

Next, we consider computation of **Cook's Distance statistic** in the complex sample design setting. Cook's D statistic can be useful for identifying observations that have a large impact on estimates of the regression parameters when they are removed from the estimation, taking the precision of the estimates into account. Cook's D statistic is computed for an individual sample observation i as follows:

$$c_i = \frac{w_i^* w_i e_i^2}{p\phi V(\hat{\mu}_i)(1-h_{ii})^2} \boldsymbol{x}_i' [\hat{Var}(U_w(\hat{\boldsymbol{B}}_w))]^{-1} \boldsymbol{x}_i$$

where

w_i^* = sampling weight
w_i = remainder of the diagonal element in the Hat matrix (e.g., $V(\hat{\mu}_i)$ for a canonical link)
e_i = residual
p = number of parameters in the regression model
ϕ = dispersion parameter in the generalized linear model
$\hat{Var}(U_w(\hat{\boldsymbol{B}}_w))$ = linearized variance estimate of the **score equation**, which is used for pseudo maximum likelihood estimation in generalized linear models fitted to complex sample survey data (see Chapter 9)

Once the value of Cook's D has been determined for an individual sample element, a simple transformation of the modified Cook's statistic can be shown to follow an F distribution:

$$\frac{(df - p + 1) \cdot c_i}{df} \sim F(p, df - p)$$

where df = # of PSUs – # of strata (design-based degrees of freedom).

This test statistic is approximately *F*-distributed and can be used to identify any unusual elements that are having a significant impact on the estimates of the regression parameters when taking the complex design features into account.

When computation of these various diagnostic statistics becomes available in the statistical software packages discussed in this book, we will provide updates on the book Web site.

7.3.4.1 Inference Concerning Model Parameters

After rigorously evaluating the fit of a model, an analyst can use the estimated parameters and their standard errors to characterize or infer about the conditional distribution of y given the predictor variables x. The analyst can perform a variety of hypothesis tests concerning the parameters being estimated, ranging from tests for a single regression parameter to tests for multiple regression parameters. We begin this section by considering tests for single regression parameters.

In the standard linear regression context, when the residuals follow a normal distribution, hypothesis tests for a single regression parameter associated with predictor variable k employ a t-test statistic:

$$t = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim t_{n-p} \tag{7.30}$$

Therefore, when performing a test of the null hypothesis H_0: $\beta_k = 0$ versus the alternative hypothesis H_A: $\beta_k \neq 0$, one can calculate a t-statistic by dividing the estimate of the regression parameter β_k by its standard error. For reasonably large samples, when H_0 is true, this test statistic is distributed as a variate from a Student t distribution with $n - (p + 1)$ degrees of freedom, where $p + 1$ is the number of parameters being estimated in the regression model (including the intercept). The analyst (or his or her software) refers the computed t-statistic to a Student t distribution with $n - (p + 1)$ degrees of freedom. If the absolute value of the test statistic, t, exceeds a critical value of the Student t distribution (e.g., $t_{1-\alpha/2,\, n-(p+1)}$ for a two-sided test), H_0 is rejected with a Type I error probability of α. Pivoting on the value of t for the two-sided test, a $100(1 - \alpha)\%$ confidence interval for the true model parameter can be constructed as $\hat{\beta}_k \pm t_{1-\alpha/2,df} \cdot se(\hat{\beta}_k)$.

When constructing the confidence interval (or the pivotal hypothesis test statistic) for a single regression parameter estimated from complex sample survey data, two aspects of the inferential process change: (1) $se(\hat{B})$, or the correct standard error of the estimated regression parameter, is *estimated* using a nonparametric technique like TSL, BRR, or JRR; and (2) the degrees of freedom for the Student t reference distribution must be adjusted to reflect the

reduced degrees of freedom for the complex sample estimate of $se(\hat{B})$. Recall from Chapter 3 that the design degrees of freedom are approximated as

$$df = \sum_h a_h - H$$

or the number of primary stage clusters minus the number of primary stage strata. For example, in the National Comorbidity Survey Replication (NCS-R) data set, the approximation to the design degrees of freedom is 84 – 42 = 42. The correct estimate of the standard error and degrees of freedom for the Student *t* reference distribution can be used to develop a design-based 100(1 – α)% confidence interval (corresponding to a Type I error rate of α) for the regression parameter of interest, as follows:

$$\hat{B} \pm t_{1-\alpha/2,\, df} \cdot se(\hat{B}) \tag{7.31}$$

The *t*-statistic for the comparable two-sided hypothesis test of H_0: $B = 0$ can be developed as $t = \hat{B} / se(\hat{B})$. This is the form of the *t*-test statistic that is routinely printed in tables of regression model output. The "*p*-values" generally printed alongside the test statistics are the probability that $t_{df} \geq |t|$.

In standard linear regression, *F*-tests are often used to test hypotheses about multiple parameters in the model. The **overall *F*-test** typically reported in the analysis of variance (ANOVA) table output generated by software procedures for fitting regression models using standard OLS methods tests the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, that is, that the fitted model predicts $E(y|\boldsymbol{x})$ no better than a model that includes only the intercept ($\beta_0 = \bar{Y}$). **Partial *F*-tests** can be used to test whether selected subsets of parameters in the model are not significantly different from 0. In this case, a "full" model is compared with a nested "reduced" model that contains a subset of the predictor variables in the "full" model. This type of hypothesis test is essentially a test of the null hypothesis that multiple parameters (the parameters omitted in the "reduced" model) are all equal to 0. This multiparameter test can be extremely useful for testing hypotheses about categorical predictor variables represented by several indicator variables in a regression model (see Section 7.4.2). More formally, we can use the following notation to indicate the $p = p_1 + p_2$ predictor variables of interest:

$$\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$$

$$\boldsymbol{x_1} = p_1 \text{ predictors}$$

$$\boldsymbol{x_2} = p_2 \text{ predictors}$$

Then, we can write the full model as follows:

$$full: \quad y = \boldsymbol{x\beta} + \varepsilon \tag{7.32}$$

The reduced model then omits the p_2 predictor variables:

$$reduced: \quad y = \boldsymbol{x_1\beta_1} + \varepsilon \tag{7.33}$$

Then, to test the null hypothesis that the regression parameters associated with the p_2 predictor variables are all equal to 0, that is, H_0: $\boldsymbol{\beta_2 = 0}$, we can calculate the following partial *F*-test statistic:

$$F = \frac{\dfrac{SSE_{reduced} - SSE_{full}}{(n-p_1)-(n-p_1-p_2)}}{\dfrac{SSE_{full}}{n-p_1-p_2}} = \frac{\dfrac{SSE_{reduced} - SSE_{full}}{p_2}}{\dfrac{SSE_{full}}{n-p}} \tag{7.34}$$

Under the null hypothesis and assuming normally distributed residuals, this *F*-statistic follows an *F* distribution with numerator degrees of freedom equal to p_2 and denominator degrees of freedom equal to $n - (p + 1)$. This general result allows one to perform a variety of multiparameter tests when comparing nested linear regression models, at least in the simple random sample setting. This *F*-statistic is also fairly robust to slight deviations from normality in the residuals.

In the complex sample survey data setting, these multiparameter tests must be adapted to the complex design features of the sample. **Wald test statistics** (Judge et al., 1985) replace the overall *F*-test and the partial *F*-test. The equivalent multiparameter Wald test statistics can be calculated as follows:

$$Overall: \ Modified\ X^2_{W,overall} = \frac{\hat{\boldsymbol{B}}_2^T \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}}_2)^{-1} \hat{\boldsymbol{B}}}{p} = F_{W,overall} \tag{7.35}$$

$$Partial: \ Modified\ X^2_{W,partial} = \frac{\hat{\boldsymbol{B}}_2^T \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}}_2)^{-1} \hat{\boldsymbol{B}}_2}{p_2} = F_{W,partial}$$

where $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{B}}_2$ are vectors of estimated regression parameters; and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}})$, $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}}_2)$ are the estimated variance–covariance matrices.

Under the null hypothesis H_0: $\boldsymbol{B = 0}$, the overall modified Wald test statistic, $F_{W,overall}$, follows an *F* distribution with numerator degrees of freedom equal to *p* and denominator degrees of freedom equal to the design degrees of freedom (*df*). Likewise, to test H_0: $\boldsymbol{B_2 = 0}$, or the null hypothesis that the p_2 parameters are all equal to 0 in the nested model, the modified Wald partial test statistic is referred to the critical value of the *F* distribution with p_2 and *df* degrees of freedom.

Wald tests can also be used to test more general hypotheses regarding linear combinations of regression model parameters. Consider the null hypothesis H_0: $\boldsymbol{CB} = \mathbf{0}$, where $\boldsymbol{C}$ is a matrix that defines specific linear combinations of the regression parameters in the vector $\boldsymbol{B}$. In this case, a version of the Wald test statistic that follows a chi-square distribution with degrees of freedom equal to the rank of the matrix $\boldsymbol{C}$ under the specified null hypothesis can be written as follows:

$$X^2_W = [\boldsymbol{C}\hat{\boldsymbol{B}}]'[\boldsymbol{C}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{B}})\boldsymbol{C}']^{-1}[\boldsymbol{C}\hat{\boldsymbol{B}}] \approx \frac{\text{contrast "squared"}}{\text{variance of contrast}} \tag{7.36}$$

We consider one example of this more general type of hypothesis test for linear combinations of multiple parameters. First, suppose that a specified linear regression model includes three parameters of interest, that is, $\boldsymbol{B}' = [B_1, B_2, B_3]$. Using this more general framework for the Wald test statistic, if one wished to test the null hypothesis H_0: $B_2 - B_3 = 0$ (or equivalently H_0: $B_2 = B_3$), the $\boldsymbol{C}$ matrix would take the following form:

$$\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

Then, the Wald test statistic specified in Equation 7.36 would follow a chi-square distribution with one degree of freedom (because the rank of the $\boldsymbol{C}$ matrix is 1). These more general hypothesis tests are available in a variety of software packages that can fit regression models to complex sample survey data sets, including SAS and Stata. Dividing these more general chi-square test statistics by the number of parameters being tested will result in test statistics that follow F distributions, similar to Equation 7.35. We consider examples of how to specify these tests for multiple parameters in Section 7.5.

7.3.4.2 Prediction Intervals

Predicting expected outcomes for populations and single individuals is an important scientific application of regression modeling (Neter et al., 1996). Even in the social and health sciences, where regression models are more often used to explore relationships among dependent and independent variables, prediction from fitted regression models still has a role. In the standard linear regression case, given a set of predictor values, $\boldsymbol{x}_{obs,i}$, the estimated regression model can be used to calculate a predicted value for y, in addition to **confidence intervals** and **prediction intervals** for the predicted value. First, we consider the expected value of the response variable y given an estimated model and a known vector of values on the predictor variables, $\boldsymbol{x}_{obs,i}$:

$$E(y_i \mid \boldsymbol{x}_{obs,i}) = \boldsymbol{x}_{obs,i}\hat{\beta} \tag{7.37}$$

Given this expected value, we can calculate a confidence interval for the expected value once the variance of the expected value is calculated:

$$\text{var}(E(y_i \mid \boldsymbol{x}_{obs,i})) = \boldsymbol{x}'_{obs,i}\,\text{cov}(\hat{\beta})\boldsymbol{x}_{obs,i} \tag{7.38}$$

A 100(1 – α)% confidence interval for the expected value of y given $\boldsymbol{x}_{obs,i}$ (i.e., the average expected outcome for a population of cases with covariates $\boldsymbol{x}_{obs,i}$) can then be calculated as follows:

$$\boldsymbol{x}_{obs,i}\hat{\beta} \pm t_{1-\alpha/2,n-p}\sqrt{\text{var}(E(y_i \mid \boldsymbol{x}_{obs,i}))} \tag{7.39}$$

Note that the previous confidence interval does not take into account the variance of the random errors that are also a part of the linear regression model. A prediction interval for a single future value of y does take this estimated variance ($\hat{\sigma}^2_{y\cdot x}$) into account:

$$\boldsymbol{x}_{obs,i}\hat{\beta} \pm t_{1-\alpha/2,n-p}\sqrt{\text{var}(\hat{E}(y_i \mid \boldsymbol{x}_{obs,i})) + \hat{\sigma}^2_{y\cdot x}} \tag{7.40}$$

Prediction intervals, therefore, are wider than more standard confidence intervals for the expected value because they also include variance in the prediction due to random error. Both intervals can give analysts a notion of the precision of the predicted values based on the fitted model if prediction of future values is an important objective of the modeling process.

Confidence intervals for predicted values can also be computed in the context of regression models for complex sample survey data. Standard errors for the predicted values can be computed using the **delta method**, which is a technique that can accommodate a wide variety of general predictions, based on fixed values of the predictors and the estimated regression parameters. For computational details on this technique, which is an approximate method based on large samples, interested readers can refer to the Stata Version 10 Reference Manual (P Manual, p. 611, StataCorp, 2008). The intervals change in that the degrees of freedom used to calculate the critical t-statistic are now based on the design degrees of freedom, and the variance–covariance matrix of the parameter estimates is estimated using approximate methods like TSL or replicated methods like JRR and BRR. Correct computation of these confidence intervals for predicted values when fitting regression models to complex sample survey data is currently implemented in Stata's `predictnl` postestimation command.

7.4 Some Practical Considerations and Tools

In this section, we discuss important considerations for data analysts fitting regression models to complex sample survey data sets in practice and provide practical guidance on steps to avoid potential pitfalls when fitting regression models and making inferences based on the fitted models.

7.4.1 Distribution of the Dependent Variable

We specifically focus our discussion in this chapter on linear regression models for **continuous dependent variables** (e.g., weight in kilograms, blood pressure in millimeters of mercury to the second decimal place, weekly household expenditures on food items). Surprisingly, many survey data sets include very few variables that are measured on a truly continuous scale. Many response variables may be **semicontinuous**, **censored,** or **grouped** (or "coarsened") in nature. We do not consider models for semicontinuous, censored, or grouped dependent variables in this chapter; Tobit regression models, Heckman selection models, and other forms of latent variable models might be considered by analysts for these types of response variables (Skrondal and Rabe-Hesketh, 2004). The Stata software (Version 10+) currently provides versions of programs designed to handle dependent variables of this type in the setting of complex sample survey data.

Many survey variables of interest are measured as ordinal scale variables. Examples include age in years and education in years. Survey questions such as, "On a scale of 1 to 5, where 1 is excellent and 5 is poor, please rate your overall health?" produce a response on an ordinal scale. Over the years it has been common practice to fit linear regression models to ordinal scale dependent variables. DeMaris (2004) describes an ordinal scale variable as **approximately continuous** if it meets the following conditions: the number of sample observations, n, is large; measurement is at least on an ordinal scale; the response has at least five ordered levels; and the distribution of responses to the ordered categories is not skewed and ideally is approximately normal in appearance. We agree that there are obvious cases where it may be acceptable to apply linear regression to an ordinal dependent variable; however, analysts will generally have a difficult time satisfying the underlying statistical assumptions for the models discussed in this chapter when working with ordinal outcomes. This could lead to highly inefficient or faulty inferences. Especially for ordinal variables with a small number of levels, a better choice is to choose a regression model that is more appropriate for the measurement scale of the dependent variable. Several regression models appropriate for ordinal and categorical response variables are discussed in detail in Chapter 9.

7.4.2 Parameterization and Scaling for Independent Variables

An extremely important aspect of fitting linear regression models is the treatment and coding of categorical *predictor* variables, which can definitely be considered in the linear regression models discussed in the chapter. When analysts consider nominal categorical predictor variables (e.g., race/ethnicity, region of the country) in linear regression models, they often generate **indicator variables** (a.k.a. "dummy" variables) to represent levels of the categorical predictor variables in the models. Regression parameters associated with these indicator variables (which are equal to 1 for cases falling into a specific category and 0 otherwise) represent changes in the expected value of the continuous outcome for a specific category relative to a **reference category**, which does *not* have an indicator variable included in the model. Alternative dichotomous specifications of these indicator variables are possible (e.g., 1/–1 "effect" coding), but we focus on the (1, 0) coding in this book for ease of interpretation.

Consider a categorical predictor variable measuring race/ethnicity with three possible values in a survey of a human population: 1 = Caucasian; 2 = African American; and 3 = Other Ethnicity. Analysts need to choose one of these three categories to be a reference category, and this choice is generally guided by contrasts of interest and research objectives (e.g., comparing African Americans and other ethnic groups with Caucasians). In cases where the choice of the reference category is not clear, choosing the most prevalent group in the sample data will suffice. In the case of the ethnicity variable, an analyst choosing "Caucasian" to be the reference category would need to create two indicator variables to include as predictors in a regression model: a variable indicating African Americans (1 = African American, 0 = Caucasian/Other Ethnicity); and a variable indicating Other Ethnic Group (1 = Other Ethnicity, 0 = Caucasian/African American). Most modern statistical software capable of fitting regression models to survey data will perform this "dummy" coding automatically for categorical predictors, requiring the analyst to simply choose the reference category for the analysis.

Continuing with the ethnicity example, suppose that a regression model was fitted to a continuous response variable y, where ethnicity was the only predictor variable. The previously described dummy coding would lead to the following regression function:

$$E(y \mid x) = B_0 + B_1 x_1 + B_2 x_2 \tag{7.41}$$

In this model, $x_1 = 1$ for African Americans and 0 otherwise, while $x_2 = 1$ for other ethnic groups and 0 otherwise. The expected value on the continuous outcome variable y for African American respondents would therefore be calculated as

$$E(y \mid x) = B_0 + B_1 \times 1 + B_2 \times 0 = B_0 + B_1 \tag{7.42}$$

and the expected value for respondents having other ethnicities would be calculated as

$$E(y \mid x) = B_0 + B_1 \times 0 + B_2 \times 1 = B_0 + B_2 \tag{7.43}$$

Because both indicator variables would be equal to 0 for Caucasians, the expected value on the outcome variable for Caucasians would simply be B_0. The regression parameters B_1 and B_2 therefore represent differences in the expected outcomes between African Americans or Other Ethnicities and Caucasians, and hypothesis tests about differences between the groups could therefore be conducted by testing whether these parameters are equal to 0. Similarly, the difference in expected outcome values between the nonreference groups (African Americans and Other Ethnicities) would be equal to $B_1 - B_2$, or the difference in expected outcomes between these two groups.

When using statistical software to fit regression models to complex sample survey data, analysts have two choices: They can create indicator variables manually or use special options in the different procedures to have the software automatically create indicator variables for the levels of categorical variables, given knowledge about a reference category. For example, users of the Stata software can define reference categories using the command:

```
char varname[omit] value
```

where `varname` refers to a categorical variable, and `value` refers to the specific value that is to be set as the reference category. Then, when fitting the regression model, Stata users can use the `xi:` function *before* specifying the regression command to indicate that categorical variables will be included in the model, and include `i.` before the names of any categorical predictor variables. The following code presents an example of this process in Stata, once again considering the ethnicity example (where OUTCOME is the continuous outcome variable and ETHNIC is the ethnicity variable):

```
char ethnic[omit] 1
xi: svy: regress outcome i.ethnic
```

Stata 11 introduced factor variables as a convenient method of handling categorical variables, and we provide examples using factor variable coding on the book Web site.

Analysts should also exercise caution when interpreting intercept parameters (β_0) in linear regression models. The intercept, or the expected value of the response variable y when all of the predictor variables are fixed at value 0, is often not of much interest, because it may represent an expectation that

THEORY BOX 7.4 ANOVA AND ANCOVA AS LINEAR REGRESSION ANALYSIS

Statistical texts on linear models often include separate chapters for linear regression models for continuous outcomes, ANOVA models (where all predictors are categorical) and analysis of covariance (ANCOVA) models (involving a mix of categorical and continuous predictors). ANOVA and ANCOVA models for normal data are generally discussed in the context of experimental designs (e.g., full factorial, randomized block), and experimental hypotheses are tested using *F*-statistics and multiple comparisons that are functions of expected mean squares. Historically, the distinction among ANOVA, ANCOVA, and linear regression analysis was reinforced in statistical software systems that included separate programs adapted for ANOVA and standard linear regression modeling. Other programs such as SAS PROC GLM integrated these two analyses in a linear model framework.

In fact, ANOVA- and ANCOVA-type analyses can be performed through proper specification of a linear regression model (e.g., main effects, interactions, nesting); see Neter et al. (1996). Analysts who wish to apply ANOVA-type procedures to complex sample survey data can do so using regression analysis programs with indicator variable parameterization of categorical independent variables and interactions appropriate to the ANOVA-type model that they wish to fit to the data (e.g., indicator variables for levels of the main effects and interactions for a full factorial model).

is well outside the range of the collected data on the predictor variables. As a result, "**centering**" of the continuous predictor variables in the model can make the intercept more interpretable:

$$y = \beta_0^* + \beta_1(x - \bar{x}) + \beta_2(z - \bar{z}) + \varepsilon \tag{7.44}$$

By subtracting the means of the predictor variables x and z from the observed values on each variable and then using the "centered" predictors in the model, the reformulated intercept (denoted by an asterisk) now has the interpretation of the expected value on the response variable y when the predictors are equal to their *means*. One can think of this term as representing a *centercept*, or the overall average value on the response variable y.

We note that the choice of omitting the intercept in a regression model forces the expected value of y to be 0 when all of the predictor variables in the model are set to 0. This might be done when preliminary knowledge of the subject matter being studied dictates that the expected value of the response be 0 when all predictor variables are set to 0.

7.4.3 Standardization of the Dependent and Independent Variables

Analysts can also **standardize** all of the variables being considered in a linear regression model (including indicator variables). One can standardize any variable by subtracting the mean for the variable, similar to centering, and then also by dividing by the standard deviation of the variable:

$$x_{i,std.} = \frac{x_i - \bar{x}}{sd(x)} \tag{7.45}$$

This is often done in practice when predictor variables are measured on very different scales (e.g., income and grade point average), and rescales all of the variables so that they are on the same scale (changes of one unit in standardized variables correspond to changes of one standard deviation in the variables). The estimates of the regression parameters in a model where *all* variables have been standardized are often referred to as **standardized regression coefficients**, and these can be used to determine which predictor variable has the largest relative impact on the expected value of the response variable. Nothing changes about this process when analyzing complex sample survey data, but an analyst should note whether weighted estimates or unweighted sample estimates of the mean and standard deviation were used to standardize the variables.

7.4.4 Specification and Interpretation of Interactions and Nonlinear Relationships

Analysts should also exercise caution when interpreting regression parameters associated with predictor variables that define nonlinear relationships between predictors and the outcome, or predictor variables that define interactions between two or more predictors. Consider the following linear regression model defining a nonlinear (and specifically quadratic) relationship between x and y:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{7.46}$$

In this model, the regression parameters β_1 and β_2 are *linked*, because they involve different transformations of the same predictor x, and β_1 alone in the presence of β_2 has no interpretation. Instead, the parameter β_2 measures the extent of the nonlinearity in the relationship between x and y. If the estimate of β_2 suggests that the parameter is not different from 0, one can assume (unless higher-order polynomial terms are significant) that the relationship between x and y is linear.

The same issue arises when considering linear regression models that involve interactions between predictors. Consider the following linear regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon \tag{7.47}$$

In this model, the three regression parameters β_1, β_2, and β_3 are once again *linked*, and one cannot interpret the regression parameter β_1 as being the relationship of the predictor variable x with the response variable y. The full relationship of x with y depends on the values of z (and therefore the value of β_3), and different values of z will result in different relationships of x with y.

In general, interactions between two or more predictor variables are computed for entry into a regression model by saving the product of two or more variables as a new variable. Interactions can be easily computed for two or more continuous predictor variables, for two or more categorical predictor variables, or for combinations of both types of variables. When working with categorical predictor variables, relevant products of *all* dummy variables to be included in the model for a given categorical variable must be computed for all other predictor variables that are specified to interact with the categorical predictor. For example, to include the interaction of a three-category predictor with a continuous variable, two product terms must be computed and included in the regression model: the product of the continuous variable with each of the indicators for the two nonreference categories of the categorical predictor. Most software procedures will perform these tasks automatically.

Plotting predicted values in linear regression models with significant interactions can be helpful when attempting to interpret significant regression parameters associated with the interactions. Social scientists often think of one of the variables involved in an interaction (e.g., z in the previous example) as a **moderator variable**, because that variable moderates the relationship of the other variables involved in the interaction with the response variable (i.e., the relationship of x with the response variable depends on the value of z). The plot in Figure 7.2 illustrates the predicted values of a continuous response variable y based on the parameter estimates in two different regression models, showing two possible interactions between a continuous predictor variable x and a binary moderator variable z, taking on values of 1 and 0 for two different groups.

In the left panel of Figure 7.2, there appears to be an interaction of x with z; the relationship of x with the continuous outcome y clearly depends on the value of z. The right panel shows no interaction, because the relationship of x with y in both groups defined by z is essentially the same. Analysts often make the mistake of interpreting regression parameters associated with single predictors in models that include interactions between that predictor and other predictors as "main effects." For example, in the first model fitted in Figure 7.2 (left panel), an analyst may be tempted to interpret the regression parameter associated with the predictor x as being the "main effect" of x. In truth, the relationship of x depends on the value of z, even if the interaction

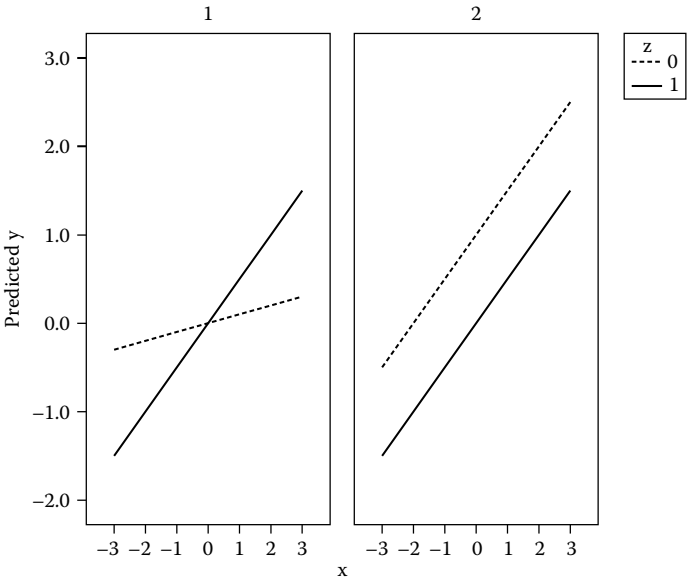


FIGURE 7.2
Hypothetical fits of two linear regression models, one with a significant interaction between x and z (left panel) and one without (right panel).

between x and z *is not significant*, so there is no "main effect" of x. This problem can be eliminated by eliminating regression parameters representing interactions from the model if they are not significantly different from zero, which will be discussed in the next section. Nothing about the interpretation of interactions changes when fitting models to complex sample survey data.

7.4.5 Model-Building Strategies

A universal set of model-building steps that is widely acknowledged and will always lead to the best-fitting linear regression model does not exist. Many statisticians have proposed practical guidelines for model fitting, keeping in mind the four steps in regression modeling discussed in Section 7.3. Out of many quality choices, we consider the model-building steps proposed by Hosmer and Lemeshow (2000). These steps are summarized as follows:

1. Conduct exploratory bivariate analyses (e.g., two-sample t-tests, chi-square tests, tests of correlations, one-way analysis of variance) to get a sense of candidate predictors that appear to have a significant relationship with the response variable.
2. Include those predictor variables that are *scientifically relevant* and have a bivariate relationship of significance $p < 0.25$ with the response variable in the initial multivariate model, and possibly consider

variable selection techniques (e.g., backward selection), with discretion. Be wary of **multicollinearity**, which could be introduced by including strongly correlated predictor variables in the same model and has the potential to inflate the standard errors of parameter estimates (see Faraway, 2005 for more details).

3. Verify the importance of the predictor variables retained in the model, using *t*-tests for individual coefficients and Wald tests for multiple coefficients (see Section 7.3.4), and assess whether the coefficients of *all* of the predictor variables change substantially in the multivariate model (relative to the bivariate case); this represents the preliminary "main" model.
4. Examine the forms of the predictor variables: If they are categorical, are sample sizes in each category large enough to use the categories as they are in the model? If they are continuous, do they have linear relationships with the response variable? Or do the relationships appear to be nonlinear? Residual diagnostics are useful as a part of this step.
5. Consider adding *scientifically relevant* interactions between the predictor variables to the model, one at a time, and do not retain them if they are not significant.
6. If any continuous or ordinal predictor variable has a large number of zeroes, include an indicator variable that is equal to 1 for nonzero values and 0 for zero values in the model, in addition to the predictor variable in question, and see if the fit of the model has been improved.

The idea behind this last step (6) is that if we have a semicontinuous predictor with many zero values and also continuous values, it is unlikely that we would have a true linear relationship passing through the mass of points at 0 and continuing through the range of nonzero values. If we create the indicator variable and also include the predictor variable in the model, this introduces a discontinuity, modeling the effect of being zero or nonzero and then for nonzero values the linear relationship of the predictor with the response variable.

We remind readers that many possible model-fitting strategies can be followed; in particular, Harrell (2001) also provides a comprehensive critique of alternative strategies.

7.5 Application: Modeling Diastolic Blood Pressure with the NHANES Data

In this practical application of linear regression analysis for complex sample survey data, we consider building a predictive model of diastolic

blood pressure (a continuous response variable) based on the sample of data collected from the U.S. adult population (ages >= 18) in the 2005–2006 National Health and Nutrition Examination Survey (NHANES). After exploring the bivariate relationships of the predictors of interest with diastolic blood pressure, we perform a naïve linear regression analysis that completely ignores the complex design features of the NHANES sample. Next, we perform a weighted regression analysis that ignores the stratification and clustering of the NHANES sample design. Finally, we take all of the important design features of the NHANES sample (stratification, clustering, and weighting for unequal probability of selection, nonresponse, and poststratification) into account at each step of the model-building process.

Specifically, the design variables that the documentation for the 2005–2006 NHANES data set states should be used for variance estimation* include SDMVPSU (which contains masked versions, or approximations, of the true primary sampling unit codes for each respondent for the purposes of variance estimation; see Section 4.3.1) and SDMVSTRA (which contains the "approximate" sampling stratum codes for each respondent, for variance estimation purposes). In addition, the appropriate sampling weight to be used to generate finite population estimates of the regression parameters for the U.S. adult population for the years of 2005 and 2006 is WTMEC2YR. This sampling weight variable was selected for analysis purposes instead of WTINT2YR because variables that will be used in the regression analyses were collected as a part of the physical examination, and the NHANES physical examination was performed on a *subsample* of all respondents (which required adjustments to the analysis weights to account for the subsampling and nonresponse to the mobile examination center [MEC] follow-up phase of the NHANES data collection).

7.5.1 Exploring the Bivariate Relationships

In this application, we follow the regression modeling strategies recommended by Hosmer and Lemeshow (2000) to build a model for diastolic blood pressure (see Section 7.4.5). We will describe each of the steps explicitly as a part of the example. First, we consider a set of predictors of diastolic blood pressure that are scientifically relevant: age, gender, ethnicity, and marital status. We begin by identifying the relevant design variables for the NHANES sample in Stata, requesting Taylor series linearization for variance estimation.

```
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra) ///
vce (linearized) singleunit (missing)
```

* http://www.cdc.gov/nchs

An initial descriptive summary of the diastolic blood pressure variable in the NHANES data set (BPXDI1) revealed several values of 0, and we set these values to missing in Stata before proceeding with the analysis:

```
gen bpxdi1_1 = bpxdi1
replace bpxdi1_1 = . if bpxdi1 == 0
```

We also generate an indicator variable for the subpopulation of adults (respondents with age greater than or equal to 18), for use in the analyses:

```
gen age18p = 1 if age >= 18 & age != .
replace age18p = 0 if age < 18
```

With the subclass indicator defined, we now consider a series of simple bivariate regression analyses to get an initial exploratory sense of the relationships of the candidate predictor variables with diastolic blood pressure. We make use of the `svy: regress` command to take the sampling weights, stratification codes, and clustering codes into account when fitting these simple initial regression models so that parameter estimates will be unbiased and variance estimates will reflect the complex design features of the NHANES sample. We first compute a weighted estimate of the mean age for the adult subclass and then center the AGE variable at the weighted mean age based on the NHANES sample (45.60):

```
svy: mean age [pweight = wtmec2yr] if age18p == 1
generate agec = age – 45.60
```

Next, the continuous dependent variable BPXDI1_1 is regressed separately on each of the candidate predictors. For the categorical predictor variables (i.e., race/RIDRETH1, gender/RIAGENDR, and marital status/MARCAT), we consider multiparameter Wald tests in Stata (see Section 7.3.4) to assess the significance of the bivariate relationships. The Stata software allows users to perform these multiparameter Wald tests by using `test` commands immediately after the models have been estimated:

```
xi: svy, subpop(age18p): regress bpxdi1_1 i.ridreth1
test _Iridreth1_2 _Iridreth1_3 _Iridreth1_4 _Iridreth1_5
xi: svy, subpop(age18p): regress bpxdi1 i.marcat
test _Imarcat_2 _Imarcat_3
xi: svy, subpop(age18p): regress bpxdi1_1 i.riagendr
test _Iriagendr_2
svy, subpop(age18p): regress bpxdi1_1 agec
test agec
```

Note in these Stata commands how the indicator for the adult subclass (AGE18P) is explicitly specified for the analysis, via the use of the `subpop()`

TABLE 7.1

Initial Design-Based Bivariate Regression Analysis Results Assessing Potential Predictors of Diastolic Blood Pressure for the 2005–2006 NHANES Adult Sample

Predictor Variable	Parameter Estimate (Linearized SE)	Test Statistic	*p*-value
Ethnicity (n = 4,581)		Wald $F(4,12) = 6.23$	< 0.01
Mexican American	--[a]	--	--
Other Hispanic	1.59 (1.11)	$t(15) = 1.44$	0.17
Non-Hispanic white	2.43 (0.55)	$t(15) = 4.38$	< 0.01
Non-Hispanic black	3.73 (0.75)	$t(15) = 4.95$	< 0.01
Other race	1.78 (1.03)	$t(15) = 1.73$	0.10
Age (Cent.) (n = 4,581)	0.06 (0.02)	$t(15) = 2.77$	0.01
Gender (n = 4,581)		Wald $F(1,15) = 56.43$	< 0.01
Male	--	--	--
Female	–2.84 (0.38)	$t(15) = –7.51$	< 0.01
Marital status (n = 4,578)		Wald $F(2,14) = 37.48$	< 0.01
Married	--	--	--
Previously married	–0.07 (0.68)	$t(15) = –0.11$	0.92
Never married	–4.39 (0.57)	$t(15) = –7.65$	< 0.01

[a] -- denotes reference category.

option. This ensures that Stata will perform an unconditional subclass analysis, treating the adult subclass sample size as a random variable and taking the full complex design of the NHANES sample into account.

We also make use of the `xi:` modifiers in Stata, to have Stata automatically generate indicator variables for the different levels of the categorical predictors to be included in the simple regression models (Stata will by default leave out the indicator for the lowest-valued category as the reference category). Note that after the dependent variable has been specified first following the `regress` command, the categorical predictor variables in the regressions are identified with the `i.` prefix; this will work only in conjunction with the `xi:` modifier, and we use this syntax throughout the remainder of this example. The variables indicated in the previous `test` commands are the indicator variables automatically generated by Stata and saved in the data set; the `test` commands are used to test hypotheses about the regression parameters associated with these indicator variables in each simple model. Table 7.1 presents the results of these initial bivariate analyses.

Stata presents **adjusted Wald tests** for the parameters in each of these models by default, where the standard Wald F-statistic (Section 7.3.4) is multiplied by $(df – k + 1)/df$, where df is the design-based degrees of freedom, and k is the number of parameters being tested (Korn and Graubard, 1990). The resulting test statistic follows an F-distribution with k and $df – k + 1$ degrees of freedom. For example, in the Wald test for the ethnicity predictor, there are $k = 4$

parameters being tested, and the design-based degrees of freedom are equal to 30 (ultimate clusters) minus 15 (strata), or 15. The denominator degrees of freedom for the adjusted test statistic are therefore 15 – 4 + 1 = 12.

Note the different subclass sample sizes in Table 7.1; three of the adult cases appear to have missing data on the marital status variable. The design-based multiparameter Wald tests and *t*-tests for the single parameters suggest that all of the potential predictor variables have potentially significant relationships with the response variable (diastolic blood pressure). Specifically, investigating the weighted parameter estimates in these simple models, males, non-Hispanic blacks, elderly people, and married people appear to have the highest diastolic blood pressures at first glance. Following the guidelines of Hosmer and Lemeshow, we therefore include all of these predictors in an initial model for the response variable measuring diastolic blood pressure.

7.5.2 Naïve Analysis: Ignoring Sample Design Features

In the first regression analysis, we *ignore* the sample weights, stratification, and clustering inherent to the NHANES sample design, do not consider any interactions between the predictors, and use standard ordinary least squares estimation to calculate the parameter estimates for the adult subclass:

```
xi: regress bpxdi1_1 i.ridreth1 i.marcat i.riagendr agec ///
if age18p == 1
```

When fitting regression models in Stata, the first variable listed after the main command is the response variable (BPXDI1_1), and the variables listed after the response variable represent the predictor variables in the model. The variable list is then generally followed by options (after a comma). In this example, we do not include any options; however, we do restrict the analysis conditionally to those subjects with age >= 18 by using the `if` modifier. We also once again use the `xi:` and `i.` modifiers to have Stata automatically generate indicator variables for selected levels of the categorical predictor variables (Stata, by default, treats the lowest-valued level of a categorical predictor as the reference category; see Section 7.4.2 for syntax to manually choose the reference category). Table 7.2 presents OLS estimates of the regression parameters in this preliminary model, along with their standard errors and associated test statistics.

These initial parameter estimates suggest that age has a positive linear relationship with diastolic blood pressure, while being married tends to increase diastolic blood pressure relative to never being married. In addition, females tend to have significantly lower diastolic blood pressure, and Mexican American respondents tend to have the lowest blood pressures (significantly lower than whites, blacks, and other ethnicities). These parameter estimates may be biased, however, because the NHANES sampling weights for respondents given a physical examination were not used to calculate

TABLE 7.2

Unweighted OLS Estimates of the Regression Parameters in the Initial Diastolic Blood Pressure Model

Predictor	Parameter Estimate	Standard Error	*t*-Statistic (*df*)	*p*-Value	95% CI
Intercept	69.672	0.464	150.04 (4569)	<0.001	(68.762, 70.582)
Ethnicity					
Other Hispanic	1.898	1.125	1.69 (4569)	0.092	(–0.308, 4.105)
White	1.672	0.491	3.40 (4569)	0.001	(0.708, 2.635)
Black	4.508	0.563	8.00 (4569)	<0.001	(3.403, 5.613)
Other	2.312	1.005	2.30 (4569)	0.021	(0.343, 4.281)
Mexican	--[a]	--	--	--	--
Marital Status					
Previously married	0.327	0.522	0.63 (4569)	0.531	(–0.697, 1.351)
Never married	–4.216	0.510	–8.27 (4569)	<0.001	(–5.216, –3.216)
Married	--	--	--	--	--
Gender					
Female	–3.402	0.375	–9.08 (4569)	<0.001	(–4.136, –2.667)
Male	--	--	--	--	--
Age (Centered)	0.039	0.011	3.40 (4569)	0.001	(0.017, 0.061)

Source: Analysis based on the 2005–2006 NHANES data.

Notes: $n = 4{,}578$, $R^2 = 0.060$, F-test of null hypothesis that all parameters are 0: $F(8, 4569) = 36.38$, $p < 0.001$.

[a] -- denotes the reference category.

nationally representative finite population estimates. In addition, the standard errors are likely understated, because the weights and the stratified, clustered design of the NHANES sample were not taken into account. We therefore consider these results only for illustration purposes.

7.5.3 Weighted Regression Analysis

Next, we consider weighted least squares estimation for calculating the parameter estimates in the initial model. Note that we explicitly indicate in the Stata command (with the `pweight` option) that the NHANES sampling weights for respondents given a physical examination (WTMEC2YR) should be included in the estimation to calculate estimates of the regression parameters:

```
xi: regress bpxdi1_1 i.ridreth1 i.marcat i.riagendr agec ///
if age18p [pweight=wtmec2yr]
```

Table 7.3 presents weighted estimates of the regression parameters, in addition to **robust standard errors** automatically calculated by Stata's standard regression command (`regress`) when sampling weights are explicitly specified with the `pweight` option. These standard errors are "sandwich-type"

TABLE 7.3

Weighted Least Squares (WLS) Estimates of the Regression Parameters in the Initial Diastolic Blood Pressure Model

Predictor	Parameter Estimate	Robust Standard Error	*t*-Statistic (*df*)	*p*-Value	95% CI
Intercept	70.678	0.489	144.57 (4569)	<0.001	(69.720, 71.637)
Ethnicity					
Other Hispanic	1.787	1.308	1.37 (4569)	0.172	(–0.778, 4.351)
White	2.192	0.519	4.22 (4569)	<0.001	(1.175, 3.209)
Black	4.409	0.612	7.21 (4569)	<0.001	(3.210, 5.608)
Other	1.958	1.040	1.88 (4569)	0.060	(–0.080, 3.997)
Mexican	--[a]	--	--	--	--
Marital Status					
Previously married	0.017	0.663	0.03 (4569)	0.979	(–1.282, 1.316)
Never married	–4.356	0.635	–6.86 (4569)	<0.001	(–5.602, –3.110)
Married	--	--	--	--	--
Gender					
Female	–2.997	0.440	–6.80 (4569)	<0.001	(–3.861, –2.134)
Male	--	--	--	--	--
Age (Centered)	0.017	0.015	1.14 (4569)	0.254	(–0.012, 0.046)

Source: Analysis based on the 2005–2006 NHANES data.
Notes: $n = 4{,}578$, $R^2 = 0.039$, F-test of null that all parameters are 0: $F(8, 4569) = 21.59$, $p < 0.001$.
[a] -- denotes the reference category

standard errors (see Freedman, 2006, for an introduction) that are considered "robust" to possible misspecification of the correlation structure of the observations. In this part of the example, there is some misspecification involved because we have once again *ignored* the stratification and clustering inherent to the NHANES sample design when calculating the standard errors, meaning that they will likely be understated. Stata's automatic calculation of robust standard errors for the parameter estimates in the presence of sampling weights is therefore an effective type of "safeguard" against this failure to incorporate the sample design features in the analysis (meaning that standard errors will not be understated), but we do not recommend following this approach in practice. Readers should be aware that not all software packages capable of survey data analysis perform this type of calculation automatically when standard regression commands are used with sampling weights specified.

In Table 7.3, we note fairly large differences in the parameter estimates relative to the OLS case (Table 7.2), especially in terms of the ethnicity parameters and the centered age parameter. When failing to incorporate the sampling weights (Table 7.2), the linear relationship of age with diastolic blood pressure was being overstated (the parameter is no longer significantly different from zero!), and the differences between the ethnic groups were being overstated as well (note that the difference between other Hispanics

and Mexicans, for example, is no longer approaching significance at the 0.05 level). The estimates in Table 7.3 represent nationally representative parameter estimates, and incorrect use of the estimates in Table 7.2 would have painted an incorrect picture of the relationships of these variables with diastolic blood pressure. We also note that the robust standard errors tend to be larger than the understated standard errors from Table 7.2, where no adjustments to the standard errors were made to account for the complex design features of the NHANES sample.

To emphasize the differences that analysts might see when specifying the sampling weights but failing to specify the sampling error codes (stratum and cluster codes) correctly in specialized software procedures for regression analysis of survey data, we include output from a similar analysis using SAS PROC REG with a WEIGHT statement:

Variable	DF	Parameter Estimate	Standard Error	*t*-Value	Pr > *t*
Intercept	1	70.67812	0.66677	106.00	<.0001
Othhis	1	1.78651	1.16011	1.54	0.1236
White	1	2.19191	0.67357	3.25	0.0011
Black	1	4.40863	0.84061	5.24	<.0001
Other	1	1.95845	1.00650	1.95	0.0517
Prevmar	1	0.01725	0.50332	0.03	0.9727
Nevmar	1	–4.35623	0.52403	–8.31	<.0001
Female	1	–2.99734	0.36059	–8.31	<.0001
Agecent	1	0.01703	0.01200	1.42	0.1558

Readers should note in this SAS output that the weighted parameter estimates are identical to those found in Stata but that most of the standard errors are understated. A more appropriate approach for the SAS users would be to use PROC SURVEYREG and specify the NHANES stratum and cluster variables, enabling appropriate variance estimation.

7.5.4 Appropriate Analysis: Incorporating All Sample Design Features

We now use the `svy: regress` command in Stata to fit the initial finite population regression model to the adult subclass and take *all* of the NHANES complex design features into account, calculating weighted estimates of the regression parameters and linearized estimates of the standard errors for the parameter estimates (incorporating the stratification and clustering of the NHANES sample). Note how an unconditional subclass analysis is requested by specifying the binary AGE18P indicator in the `subpop()` option, similar to the bivariate analyses performed previously:

```
svyset sdmvpsu [pweight = wtmec2yr], strata(sdmvstra) ///
vce(linearized) singleunit(missing)
```

```
xi: svy, subpop(age18p): regress bpxdi1_1 i.ridreth1 ///
i.marcat i.riagendr agec
estat effects, deff
```

We also use the postestimation command `estat effects, deff` to request calculation of design effects for the estimated regression parameters. Table 7.4 presents the estimated parameters in this initial "main" model.

The estimated parameters and tests of significance presented in Table 7.4 confirm most of the simple relationships observed in the initial design-based bivariate analyses and suggest that the relationships remain similar when taking other predictor variables into account in a multivariate analysis (with the exception of the linear relationship of age with diastolic blood pressure). When holding the other predictor variables in this model fixed, non-Hispanic whites and blacks have significantly higher expected diastolic blood pressure values than Mexican Americans; never-married respondents have significantly lower diastolic blood pressure than married respondents; females have significantly lower diastolic blood pressure than males; and, interestingly,

TABLE 7.4

Design-Based Estimates of the Regression Parameters in the Initial "Main" Model for Diastolic Blood Pressure, Linearized Standard Errors for the Estimates, Design-Adjusted Test Statistics and Confidence Intervals for the Parameters, and Design Effects for the Parameter Estimates

Predictor	Est.	Linearized SE	*t*-Statistic (*df*)	*p*-Value	95% CI	$d^2(\hat{B})$
Intercept	70.678	0.501	141.10 (15)	< 0.001	(69.611, 71.745)	0.95
Ethnicity						
Other Hispanic	1.787	1.142	1.56 (15)	0.139	(–0.648, 4.221)	1.57
White	2.192	0.605	3.62 (15)	0.002	(0.903, 3.481)	1.36
Black	4.409	0.761	5.79 (15)	< 0.001	(2.786, 6.031)	1.27
Other	1.958	0.988	1.98 (15)	0.066	(–0.148, 4.064)	1.58
Mexican	--[a]	--	--	--	--	--
Marital Status						
Previously married	0.017	0.718	0.02 (15)	0.981	(–1.513, 1.547)	2.67
Never married	–4.356	0.565	–7.71 (15)	< 0.001	(–5.560, –3.152)	1.69
Married	--	--	--	--	--	--
Gender						
Female	–2.997	0.331	–9.05 (15)	< 0.001	(–3.703, –2.292)	1.29
Male	--	--	--	--	--	--
Age (Centered)	0.017	0.022	0.78 (15)	0.448	(–0.030, 0.064)	3.95

Source: Based on the 2005–2006 NHANES data.

Notes: Subclass $n = 4{,}578$, $R^2 = 0.039$, adjusted Wald test for all parameters: $F(8,8) = 12.66$, $v < 0.001$.

[a] -- denotes the reference category.

age does not appear to have a significant linear relationship with diastolic blood pressure. Age is, therefore, the only predictor that does not appear to be important, but we have considered only a linear relationship thus far. None of the sample sizes for the groups defined by the categorical variables appear to be extremely small, so we do not consider further recoding of these variables. Readers should note that the weighted parameter estimates in Table 7.4 are exactly equal to those in Table 7.3; differences arise in how the estimated standard errors for the parameter estimates are being calculated.

There are several important observations regarding the test statistics for the regression parameters in Table 7.4. First, the degrees of freedom for the *t*-statistics based on the complex sample design of the NHANES (15) are calculated by subtracting the number of strata (15) from the number of sampling error computation units or ultimate clusters (30). These degrees of freedom are substantially different from those noted in Tables 7.2 and 7.3 ($df = 4569$), where the complex design was not taken into account when performing the estimation; this shows how the primary sampling units (rather than the unique elements) are providing the independent contributions to the estimation of distributional variance when one accounts for the complex sample design. In addition, Stata presents an adjusted Wald test for all of the parameters in the model (see the discussion of the Table 7.1 results). The numerator degrees of freedom for this adjusted statistic are equal to k (8 in this example, because eight parameters are being tested; the "null" or "reduced" model still contains the intercept parameter), and the denominator degrees of freedom are calculated as $df - k + 1$ ($15 - 8 + 1 = 8$ in this example). This adjusted Wald test definitely suggests that a null hypothesis that all of the regression parameters are equal to 0 would be strongly rejected.

The design effects presented in Table 7.4 (DEFF) are for the most part greater than 1, suggesting that the complex design of the NHANES sample is generally resulting in a decrease in the precision of the parameter estimates relative to the precision that would have been achieved under a simple random sampling design with the same sample size (see Section 2.4). The losses in precision due to the complex design are not severe (we actually see a gain in precision for some of the parameter estimates), but the effects of the complex design on the standard errors are apparent. When options for obtaining design effects are available in software packages, readers should note the design effects because they can be helpful for future power calculations and sampling designs (see Section 2.5).

We now consider some initial model diagnostics to assess the fit of this preliminary model. We start by saving the residuals in a new variable (RESIDS) in Stata and then by plotting the residuals against the values on the continuous mean-centered age (AGEC) variable. The left-hand panel of Figure 7.3 presents this plot:

```
predict resids, resid
scatter resids agec
```

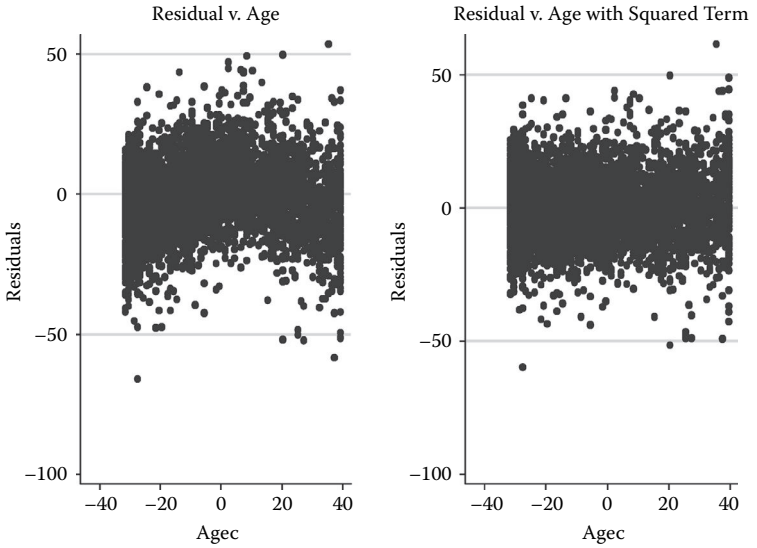


FIGURE 7.3
Plots of residuals versus AGEC for the diastolic blood pressure application before and after the addition of the squared AGEC variable to the model. (Modified from the 2005–2006 NHANES data.)

The first plot in Figure 7.3 indicates a fairly well-defined curvilinear pattern of the residuals as a function of age, suggesting that the structure of the model has been misspecified; there is evidence that age actually has a quadratic relationship with diastolic blood pressure that has not been adequately captured by including a linear relationship of age with the response variable. We therefore add a squared version of age (AGECSQ) to the model to capture this relationship:

```
gen agecsq = agec * agec
xi: svy, subpop(age18p): regress bpxdi1_1 i.ridreth1 ///
i.marcat i.riagendr agec agecsq
```

In the new model (see Table 7.5), the regression parameters for both the centered age predictor and the squared version of the age predictor are significantly different from 0 ($p < 0.001$), confirming that the relationship of age with diastolic blood pressure is in fact nonlinear and quadratic in nature. The *R*-squared of the new model becomes 0.134, suggesting an improved fit by allowing the relationship of age with diastolic blood pressure to be nonlinear. Further, after adding the squared term, the marital status differences observed previously no longer seem to be significant. The right-hand panel of Figure 7.3 shows the improved distribution of the residuals as a function of age after adding the squared term, where there is no pattern evident in the residuals as a function of age.

Now, we consider testing specific interactions of interest, one at a time: the interactions between age (both predictors) and ethnicity, and the interactions between age (both predictors) and gender. This step essentially allows for testing whether the nonlinear relationship of age with diastolic blood pressure tends to be moderated by these two demographic factors; for example, is the quadratic trend in diastolic blood pressure as a function of age flatter (i.e., more stable) for certain ethnic groups than others? We first add the interactions between age and ethnicity to the model and investigate an adjusted Wald test:

```
xi: svy, subpop(age18p): regress bpxdi1_1 i.marcat ///
i.riagendr i.ridreth1*agec i.ridreth1*agecsq
test _IridXagec_2 _IridXagec_3 _IridXagec_4 _IridXagec_5 ///
_IridXagecs_2 _IridXagecs_3 _IridXagecs_4 _IridXagecs_5
```

Note in the `svy: regress` command how the interactions are specified: when the `xi:` modifier is used, specifying the term `i.ridreth1*agec` will include indicators for the levels of the categorical RIDRETH1 variable, the centered age variable, *and* the relevant two-way interactions between the indicators and age in the model. There is no need to specify the individual RIDRETH1 and AGEC variables in the list of predictors when specifying interactions like this; they will be included automatically.

We remind readers that when using the `test` commands in conjunction with survey regression commands, Stata performs an adjusted Wald test by default. The `nosvyadjust` option can be added to a `test` command if a user does not desire the additional adjustment to the test statistic. The multiparameter Wald test for all of the newly added interaction parameters essentially amounts to a design-based test of **change in *R*-squared** for comparing nested models (where in this case, one model includes the interactions and one does not). In the `test` command that follows the `svy: regress` command, note how all eight of the variables created by Stata representing products of the nonreference indicator variables for the ethnic groups and the two age variables are listed. This represents a Wald test of the null hypothesis that all eight of these regression parameters associated with the interactions are equal to zero. The adjusted Wald test performed by Stata actually indicates that we do not have enough evidence to reject this null ($F(8,8) = 0.98$, $p = 0.51$), which suggests that adding the interactions between both age terms and ethnicity is not significantly improving the fit of the model. We therefore proceed without including these interactions in the model.

Next, we add the two-way interactions between the two age terms and gender (RIAGENDR) to the model and again test the associated parameters using a Wald test:

```
xi: svy, subpop(age18p): regress bpxdi1_1 i.marcat ///
i.ridreth1 i.riagendr*agec i.riagendr*agecsq
test _IriaXagec_2 _IriaXagecs_2
```

The Wald test once again suggests that the two regression parameters associated with the interactions between the age terms and gender are not significantly different from zero ($F(2,14) = 1.73$, $p = 0.21$), so we have evidence in favor of the model excluding these interactions. Readers can use similar methods to test interactions between two (or more) categorical predictors.

After determining that the two-way interactions of interest are not significantly improving the fit of the model (i.e., the nonlinear relationship of age with diastolic blood pressure does not appear to be moderated by ethnicity or gender), we refit the "final" model, save variables containing residuals (EHAT1) and predicted values (YHAT1) based on the fit of the "final" model, and generate a series of diagnostic plots to assess the assumptions underlying the model:

```
xi: svy, subpop(age18p): regress bpxdi1_1 i.marcat ///
i.riagendr i.ridreth1 agec agecsq
predict ehat1, resid
symplot ehat1, name(sym_ehat1_1, replace) title(Symplot of ///
Residuals)
histogram ehat1, normal name(h_ehat1, replace) ///
title(Histogram of Residuals)
qnorm ehat1, name(qnorm_ehat1, replace) title(QNorm of ///
Residuals)
predict yhat1, xb
scatter ehat1 yhat1, name(ehat1xyhat1, replace) ///
title(Residuals v. Predicted Y)
graph combine sym_ehat1_1 h_ehat1 qnorm_ehat1 ehat1xyhat1, ///
rows(2)
```

We introduce one additional diagnostic plot that can be helpful for assessing model fit. The `symplot` command in Stata produces a symmetry plot of values on the response variable in the model, which is useful for determining whether the distribution of values on a given continuous variable appears to be symmetric in nature. Specifically, the distance below the median of the first ordered value in the distribution is plotted against the distance above the median of the last ordered value; the distance below the median of the second ordered value is plotted against the distance above the median of the second-to-last ordered value; and so forth. If the values in the symmetry plot lie on the straight diagonal line, there is evidence that the distribution is symmetric. The resulting plot in the upper-left panel of Figure 7.4 suggests that the distribution of the residuals based on the final fitted model is definitely symmetric around 0, alleviating possible concerns about slight deviations from normality.

Collectively, the diagnostic plots presented in Figure 7.4 suggest that the residuals follow a symmetric distribution and that assumptions of normality and constant variance for the residuals definitely seem reasonable. Given any apparent violations of normality in the distribution of the residuals, the

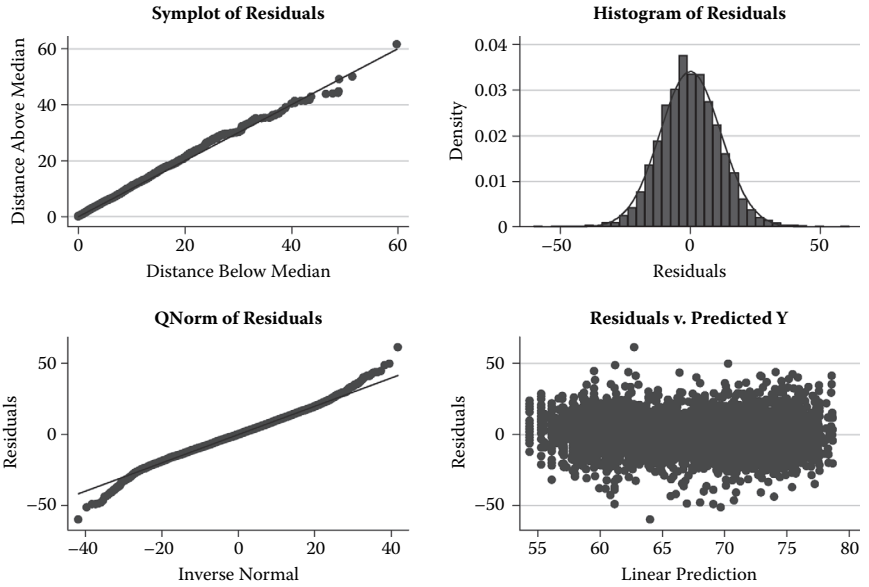


FIGURE 7.4
Diagnostic plots for the "final" regression model fitted to the diastolic blood pressure response variable in the 2005–2006 NHANES data set.

symmetry of the residuals around the expected value of 0 gives us confidence in the inferences that we are making. We are therefore confident with the parameter estimates and tests of significance based on this "final" model. Table 7.5 presents the parameter estimates and associated tests of significance in the "final" model for this example.

The estimates in Table 7.5 provide strong evidence of the quadratic relationship of age with diastolic blood pressure when adjusting for other sociodemographic features and also show how there are still strong ethnicity and gender effects on blood pressure. The investigation presented here did not find any evidence of the relationship of age being moderated by gender or ethnicity, although additional tests would certainly be possible at this point. We once again note the substantial improvement in the model *R*-squared values (from 0.039 to 0.134) due to the inclusion of the squared age term.

7.6 Exercises

1. Fit two linear regression models to the 2005–2006 NHANES data, assuming simple random sampling (i.e., ignoring the weighting, clustering, and stratification) and focusing on the subpopulation of

TABLE 7.5

Estimates of the Regression Parameters in the "Final" Model for the Diastolic Blood Pressure Response Variable

Predictor	Est.	Linearized SE	*t*-Statistic (*df*)	*p*-Value	95% CI	DEFF
Intercept	73.859	0.455	162.37 (15)	< 0.001	(72.889, 74.829)	0.83
Ethnicity						
Other Hispanic	1.189	1.087	1.09 (15)	0.291	(–1.127, 3.505)	1.56
White	1.781	0.631	2.82 (15)	0.013	(0.436, 3.125)	1.67
Black	3.465	0.779	4.45 (15)	< 0.001	(1.804, 5.126)	1.48
Other	1.189	0.934	1.27 (15)	0.223	(–0.803, 3.180)	1.54
Mexican	--[a]	--	--	--	--	--
Marital Status						
Previously married	1.040	0.622	1.67 (15)	0.115	(–0.285, 2.366)	2.21
Never married	–0.343	0.582	–0.59 (15)	0.564	(–1.583, 0.897)	1.83
Married	--	--	--	--	--	--
Gender						
Female	–2.721	0.338	–8.06 (15)	< 0.001	(–3.441, –2.002)	1.49
Male	--	--	--	--	--	--
Age (Centered)	0.125	0.015	8.45 (15)	< 0.001	(0.094, 0.157)	2.03
Age (Centered) Squared	–0.012	0.001	–16.34 (15)	< 0.001	(–0.014, –0.011)	2.26

Source: Analysis based on the 2005–2006 NHANES data.
Notes: $n = 4{,}578$, $R^2 = 0.134$, adjusted Wald test for all parameters: $F(9,7) = 87.12$, $p < 0.001$.
[a] -- denotes the reference category.

those aged 18 and older (AGE18P = 1): one where the dependent variable, systolic blood pressure measurement number one (BPXSY1), is regressed on the predictor BMI (BMXBMI) (Model 1), and one (Model 2) where BPXSY1 is regressed on the predictor variables BMI, age of the respondent (RIDAGEYR), GENDER (RIAGENDR: 1 = Male, 2 = Female), RACE (RIDRETH1: 1 = Mexican, 2 = Other Hispanic, 3 = White, 4 = Black, 5 = Other), and poverty index (INDFMPIR). Make sure that the categorical predictor variables are handled appropriately in Model 2, using indicator variables for the non-reference levels of the predictors. Based on the results of these analyses, answer the following questions:

a. Based on the Model 1 analysis, construct a 95% CI for the regression parameter for BMI.

b. Based on the Model 1 and Model 2 analyses, perform the joint *F*-test of the null hypothesis that the regression coefficients *B*(RIDAGEYR), *B*(RIAGENDR (omit Female)), *B*(RIDRETH1(omit

Other)), and *B*(INDFMPIR) are zero (i.e., not significant). Model 2 is the "full" model and Model 1 is the "reduced" model. Based on the result of this test, do we have evidence of these socio-demographic predictors improving the fit of the model?

2. Answer these questions based on the estimates generated by fitting Model 2:
   a. All else being equal, what is the expected difference in systolic blood pressure for a black as opposed to a white individual?
   b. All else being equal, what is the expected difference in systolic blood pressure for a black as opposed to an "other" individual?
   c. All else being equal, what is the estimated difference in your expected systolic blood pressure 10 years from now compared with today?
   d. What is the expected difference in systolic blood pressure for a male patient with the following characteristics (body mass index [BMI] = 30, Age = 45, Gender = Male, Race = White, PovIndex = 4) compared with his twin sister (BMI = 26, Age = 45, Gender = Female, Race = White, PovIndex = 2)?
3. Replicate the regression analyses performed in problems 1 and 2, correctly accounting for the complex design features of the 2005–2006 NHANES data set (sampling error computation units = SDMVPSU, sampling error strata = SDMVSTRA, final sampling weight for the Interviewed and Examined NHANES sample = WTMEC2YR).
   a. Write a brief paragraph (as you might for a research article) explaining the methods that you used to estimate the model parameters and estimate the standard errors of the estimated parameters in a way that accounts for the complex sample design of the NHANES.
   b. Prepare a table that compares the estimated regression coefficients and standard errors in Models 1 and 2 from the "simple random sample" analysis to the weighted estimates and design-adjusted standard errors from the design-based analysis.
   c. Comment on whether any of the results or inferences from problems 1 and 2 would change when taking the complex design of the NHANES sample into account in the analyses.
4. Why would you not choose to use all three variables representing physical measurements—HEIGHT, WEIGHT, and BMI—in a single regression model?
5. Following the model-building strategies discussed in this chapter, attempt to identify additional predictors of systolic blood pressure (including additional physical measurements), and see whether it is possible to improve the R^2 value of the model. Be careful not to include

predictors that are highly correlated with each other, and consider the possibility of interactions between the predictor variables. Make sure to examine appropriate residual diagnostics once you have arrived at a "final" model, and transform any variables if necessary.

6. Write a short (one-page) summary of the process used to specify and fit your "final" model, including residual analysis and your conclusions concerning the relationship of systolic blood pressure to the covariates that you considered in developing your "final" model.

8

Logistic Regression and Generalized Linear Models for Binary Survey Variables

8.1 Introduction

Perusal of the electronic codebook for any major survey data set quickly points to the fact that many key survey questions require only a simple "yes" or "no" answer. Consider these questions from the three data sets used for the examples in this text:

> **HRS 2006 N001:** "Are you currently covered by Medicare health insurance?"
>
> **NHANES DIQ010:** "Have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?"
>
> **NCSR SC26_4:** "Did you ever use drugs or alcohol so often that it interfered with your responsibilities at work, at school or at home?

The responses to such questions are coded as **binary variables** (or **dichotomous variables**). This chapter is devoted to **generalized linear models** (GLMs) for a binary survey variable—focusing primarily on the application of **logistic regression** analysis to complex sample survey data. There are many excellent texts on standard simple random sample (SRS) methods for logistic regression analysis including comprehensive introductions to the theory, methods, and applications provided in texts by Hosmer and Lemeshow (2000), Agresti (2002), and Allison (1999). Readers wanting a full mathematical treatment of GLMs are referred to McCullagh and Nelder (1989) and McCulloch and Searle (2001).

The general aims of this chapter are to (1) introduce the fundamental concepts of regression for categorical data based on GLMs that are important to understand for effective application of the techniques to survey data; (2) provide a systematic review of the stages in fitting the logistic regression model to complex sample survey data; and (3) present example logistic regression analyses based on the Health and Retirement Study (HRS) and

National Comorbidity Survey Replication (NCS-R) data that illustrate typical applications of the model-building steps to actual survey data sets.

Section 8.2 will touch on the important underlying concepts of GLMs as they apply to **logistic** and **probit regression** but will leave much of the theory and detailed development to specialized texts on the topic. Sections 8.3 through 8.6 describe the theory and methods that are important in the four stages of building and interpreting a logistic regression model: specification, estimation, evaluation, and interpretation/inference for complex sample survey data. Section 8.7 takes the reader through an example analysis in which a multivariate logistic regression model is fit to the binary indicator of a lifetime major depressive episode (MDE) in the NCS-R data set. The chapter concludes in Section 8.8 with a comparative application of the logistic, probit, and complementary log–log (CLL) regression techniques to the problem of modeling the probability of alcohol dependency in U.S. adults.

8.2 Generalized Linear Models for Binary Survey Responses

In principle, generalized linear models for a binary dependent variable and linear regression models for a continuous variable (Chapter 7) share a common aim: to estimate a regression equation that relates the expected value of the dependent variable y to one or more predictor variables, denoted by $\boldsymbol{x}$. In linear regression for a continuous dependent variable, the expected value of y is the conditional mean of y given a vector of covariates, $\boldsymbol{x}$, and is estimated by an equation that is linear in the regression parameters:

$$\hat{Y} = E(y \mid \boldsymbol{x}) = B_0 + B_1 x_1 + \cdots + B_p x_p \tag{8.1}$$

When y is a binary variable with possible values 0 and 1 ($y = \{0,1\}$), $E(y \mid \boldsymbol{x}) = \pi(\boldsymbol{x})$ is the conditional probability that $y = 1$ given the covariate vector $\boldsymbol{x}$. Throughout this chapter, the notation $\pi(\boldsymbol{x})$ will be used to represent this probability that $y = 1$ conditional on a vector of observed predictors, $\boldsymbol{x}$.

A naïve approach to regression analysis of a binary dependent variable is to model the $\pi(\boldsymbol{x})$ as a linear function of $\boldsymbol{x}$. There are several problems with this approach. First, as we will see in Section 8.4, the sample of observations on the dependent variable y is assumed to follow a binomial distribution—a severe violation of the normality and homogeneity of variances assumption required for efficient least squares estimation of the coefficients of the linear regression model. Second, as shown in Figure 8.1, a naïve linear regression model for $\pi(\boldsymbol{x})$ does not accurately capture the relationship between y and

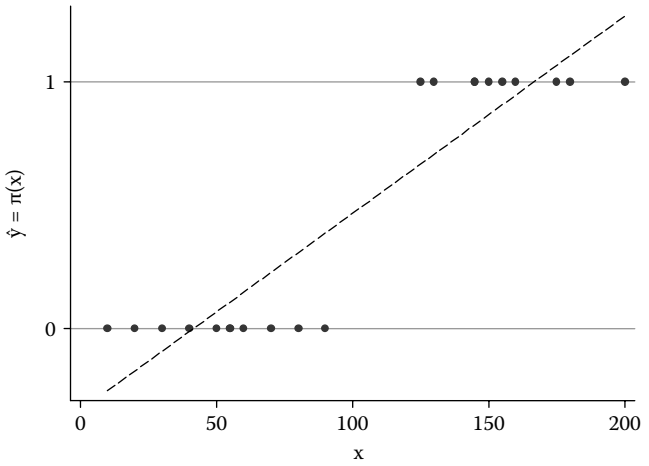


FIGURE 8.1
Naïve use of linear regression for a binary dependent variable.

$\boldsymbol{x}$—and it may even produce predicted values for $\pi(\boldsymbol{x})$ that are outside the permissible range of 0 to 1.

The alternative would be to identify a nonlinear function of $\pi(\boldsymbol{x})$, say $g(\pi(\boldsymbol{x}))$, that yields a fitted regression model that is linear in the coefficients for the model covariates, $\boldsymbol{x}$**.** Ideally, the estimated function, $g(\pi(\boldsymbol{x}))$, should also be chosen so that when it is transformed back the resulting values of the estimated $\pi(\boldsymbol{x})$ will fall in the range between 0 and 1.

In the terminology of generalized linear models, functions like $g(\pi(\mathbf{x}))$ are termed **link functions.** The two most common link functions used to model binary survey variables are the logit and the probit. The functional relationships of the logit and probit to the values of $\pi(\boldsymbol{x})$ are illustrated in Figure 8.2, where one can see the value of $\pi(\boldsymbol{x})$ that would result from a given value of the logit or probit. Theory Box 8.1 presents a latent variable interpretation of the logit and probit and a mechanism for relating these linear functions in $\boldsymbol{x}$ to the $\pi(\boldsymbol{x})$ that are the quantities of true interest in the model estimation.

8.2.1 The Logistic Regression Model

For a logistic regression model, the link function is the **logit**:

$$g(\pi(\boldsymbol{x})) = \text{logit}(\pi(\boldsymbol{x})) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = B_0 + B_1 x_1 + \cdots + B_p x_p \qquad (8.2)$$

The logit is nonlinear in $\pi(\boldsymbol{x})$ but is presumed to be linear in the regression parameters, $\boldsymbol{B} = \{B_0, B_1, \ldots, B_p\}$. Based on the fitted regression model for the

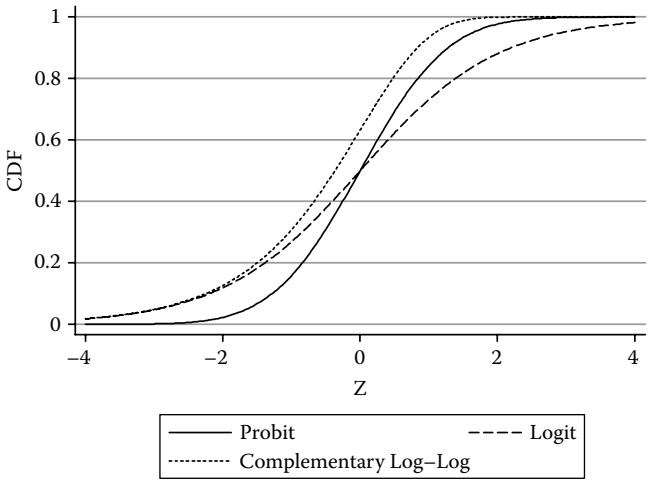


FIGURE 8.2
Cumulative distribution functions (CDFs) for the logit, probit, and CLL links.

logit, the estimated value of $\pi(\boldsymbol{x})$ can be recovered by applying the inverse logit function:

$$\hat{\pi}(\boldsymbol{x}) = g^{-1}(g(\hat{\pi}(\boldsymbol{x}))) = \frac{\exp(\hat{B}_0 + \hat{B}_1 x_1 + \ldots + \hat{B}_p x_p)}{1 + \exp(\hat{B}_0 + \hat{B}_1 x_1 + \ldots + \hat{B}_p x_p)} \tag{8.3}$$

The inverse function $g^{-1}(.)$ is the cumulative distribution function (CDF) for the logistic probability distribution. Under the logistic model in Equation 8.2, $\hat{\pi}(\boldsymbol{x})$ is the logistic CDF evaluated at the estimated logit corresponding to the covariate vector $\boldsymbol{x}$.

Unlike the linear regression model for normally distributed y, there is no direct solution such as the method of least squares to estimate the regression coefficients in the logit model. Instead, an iterative estimation procedure such as the Newton–Raphson or Fisher Scoring algorithm (Agresti, 2002) is used to determine the values of estimated coefficients that maximize the following weighted pseudo-likelihood function:

$$PL(\boldsymbol{B} \mid X) = \prod_{i=1}^{n} \{\pi(\boldsymbol{x}_i)^{y_i} \cdot [1 - \pi(\boldsymbol{x}_i)]^{1-y_i}\}^{w_i} \tag{8.4}$$

It is important to note that the link function is not an explicit transformation of the variable y but rather a transformation of the $E(y \mid \boldsymbol{x}) = \pi(\boldsymbol{x})$. Also

THEORY BOX 8.1 A LATENT VARIABLE INTERPRETATION OF THE LOGIT MODEL

In specifying a logistic (or probit) regression model for a binary outcome variable, the observed response on the binary outcome variable y can be viewed as determined by the value of a *latent* (or unobserved) continuous variable z. If the true value of the link function, z_i, for a survey respondent is less than or equal to some threshold value Z, the respondent answers with a $y = 1$; if the value of $z_i > Z$, the respondent is expected to answer with $y = 0$. Since z_i is a latent variable and never directly observed, its value for each respondent must be modeled as a function of observed survey variables.

The intercept parameter B_0 may be interpreted as a "cutpoint" along the distribution of the logit (or probit) values. Changes in the predictor variables shift this cutpoint, therefore changing the probability of a response of 1; thinking about the logistic regression model in this way will be useful for understanding ordinal logistic regression models in Section 9.3.

note that evaluation of the pseudo-likelihood in Equation 8.4 requires both the original observations, y_i, the modeled values of $\hat{\pi}(\boldsymbol{x}_i)$, and in the case of complex sample survey data, the sampling weights w_i. We discuss this function in more detail in Section 8.4.

Each cycle in the iterative estimation of the logistic regression model requires four operations:

1. The value of the logit in Equation 8.2 is estimated for each survey respondent i based on the current iteration values of the estimated parameters: $z_i = \hat{B}_0 + \hat{B}_1 x_{1i} + \cdots + \hat{B}_p x_{pi}$.
2. The logit for each case is then transformed back to the probability scale by evaluating the logistic CDF at the value of z_i as illustrated in (8.3).
3. The value of the likelihood (Equation 8.4) is then evaluated at the $i = 1, \ldots, n$ values of the estimated $\hat{\pi}(\boldsymbol{x}_i)$ and observed y_i values.
4. The algorithm then adjusts the values of the individual parameter estimates $\hat{B}_j$ to maximize the likelihood function, and returns to repeat the cycle.

The iterative algorithm stops when the change in the estimates of the vector of $\boldsymbol{B}$ values no longer increases the value of the likelihood function. The values of the individual $\hat{B}_j$ at the final iteration are the final estimates.

8.2.2 The Probit Regression Model

The probit regression model is an alternative to logistic regression for modeling a binary dependent variable. Probit regression models are also generalized linear models, and the procedures for estimating the models parallel those illustrated in Section 8.2.1 for logistic regression. The difference in the two approaches is that the link function changes to the **probit** (inverse Normal):

$$g(\pi(\boldsymbol{x})) = \Phi^{-1}(\pi(\boldsymbol{x})) = z = B_0 + B_1 x_1 + \cdots + B_p x_p \tag{8.5}$$

In contrast to logistic regression, where z is assumed to follow the logistic probability distribution, the probit is assumed to follow a standard normal distribution. Again, as in logistic regression, transforming back from the probit scale to the quantities of interest $\pi(\boldsymbol{x})$ requires an evaluation of the standard Normal CDF at the estimated value of the probit, $z_i = \hat{B}_0 + \hat{B}_1 x_{1i} + \cdots + \hat{B}_p x_{pi}$:

$$\hat{\pi}(\boldsymbol{x}_i)_{probit} = \Phi(z_i) = \text{Prob}(Z \le z_i) = \int_{-\infty}^{\hat{B}_0 + \hat{B}_1 x_{1i} + \cdots + \hat{B}_p x_{pi}} \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{Z^2\right\} dZ \tag{8.6}$$

With the probit link (Equation 8.5) replacing the logit and the normal probability transform (Equation 8.6) replacing the logistic CDF transform (Equation 8.3), the steps in estimating the probit regression model are identical to those previously outlined for logistic regression. In general, inferences derived under logistic and probit regressions do not differ significantly.

8.2.3 The Complementary Log–Log Model

A third less commonly used link function in regression analysis of a binary dependent variable is the CLL link. Most software packages for survey data analysis that support logistic and probit regression modeling also permit the user the option to choose the CLL link. The CLL link function is related to the Gompertz distribution and the link may be referred to as the "gompit." The primary distinction for the CLL link is that unlike the logit or probit the CDF is not required to be symmetric about the midpoint $\pi(\boldsymbol{x}) = 0.5$ (see Figure 8.2). Its most common application is in situations where $\pi(\boldsymbol{x})$ is either very close to 0 or close to 1 (at one or the other extreme). Interested readers are referred to Allison (1999) for applications of the CLL link function to both binary regression models and survival analysis.

Section 8.8 will provide a parallel comparison of a model fitted under all three links: the logit, the probit, and the CLL.

8.3 Building the Logistic Regression Model: Stage 1, Model Specification

The four stages in logistic regression modeling of survey data are identical to those presented in Chapter 7 for linear regression: (1) model specification; (2) estimation of model parameters and their standard errors; (3) model evaluation and diagnostics; and (4) interpretation of results and inference based on the final model. As in all statistical model building processes, stages 1–3 define an iterative process designed to sequentially refine and test the model. Several cycles of the model specification, estimation, and evaluation sequence are usually required before a final model is identified and inferences can be made concerning the modeled relationship in the larger survey population.

The logistic regression model is a flexible modeling tool that can simultaneously accommodate both categorical and continuous predictors as well as terms for the interactions among the predictors. As in all regression modeling, identification of a best logistic regression model for the survey data should follow a systematic, scientifically governed process whereby important candidate predictors are identified and evaluated, first individually and then in the multivariate context of other potentially important explanatory variables. Our recommendation is that survey analysts follow Hosmer and Lemeshow's (2000) incremental process for specifying the initial model, refining the set of predictors and then determining the final form of the logistic regression model:

- Perform initial bivariate analyses of the relationship of y to individual predictor variable candidates.
- Select the predictors that have a bivariate association with y at significance $p < 0.25$ as candidates for main effects in a multivariate logistic regression model.
- Evaluate the contribution of each predictor to the multivariate model using the Wald test.
- Check the linearity assumption for continuous predictors.
- Check for scientifically justified interactions among predictors.

Hosmer and Lemeshow (2000) advocate a final step involving the application of polynomial functions and smoothing splines to test that the logistic model is truly linear in the logit. Interested readers are referred to that text

for a description of this final step in the evaluation of the fit of a logistic regression model.

8.4 Building the Logistic Regression Model: Stage 2, Estimation of Model Parameters and Standard Errors

Given the specification of a logistic regression model of the form $\text{logit}(\pi(\boldsymbol{x}))=B_0 + B_1 x_1 + \cdots + B_p x_p$, the second step in the model-building process is to compute estimates of the regression parameters in the model along with their standard errors. For simple random samples, the logistic regression model parameters and standard errors can be estimated using the method of maximum likelihood. The likelihood function for an SRS of n observations on a binary variable y with possible values 0 and 1 is based on the binomial distribution:

$$L(\boldsymbol{\beta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{1-y_i} \tag{8.7}$$

where $\pi(\boldsymbol{x}_i)$ is linked to the regression model coefficients through the logistic CDF:

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i \boldsymbol{\beta})} \tag{8.8}$$

When the survey data have been collected under a complex sample design, straightforward application of maximum likelihood estimation (MLE) procedures is no longer possible, for several reasons. First, the probabilities of selection (and responding) for the $i = 1, \ldots, n$ sample observations are (generally) no longer equal. Sampling weights are thus required to estimate the finite population values of the logistic regression model parameters. Second, the stratification and clustering of complex sample observations violates the assumption of independence of observations that is crucial to the standard MLE approach to estimating the sampling variances of the model parameters and choosing a reference distribution for the likelihood ratio test statistic.

Two general approaches have been developed to estimate the logistic regression model parameters and standard errors for complex sample survey data. Grizzle, Starmer, and Koch (1969) first formulated an approach based on weighted least squares (WLS) estimation. The WLS estimation method was originally programmed for logistic regression in the GENCAT software

package (Landis et al., 1976) and still remains available as an option in programs such as SAS PROC CATMOD. Later, Binder (1981, 1983) presented a second general framework for fitting logistic regression and other generalized linear models to complex sample survey data. Binder proposed **pseudo-maximum likelihood estimation** (PMLE) as a technique for estimating the model parameters. The PMLE approach to parameter estimation was combined with a linearized estimator of the variance–covariance matrix for the parameter estimates, taking complex sample design features into account. Further development and evaluation of the PMLE approach is presented in Roberts, Rao, and Kumar (1987), Morel (1989), and Skinner, Holt, and Smith (1989). The PMLE approach is now the standard method for logistic regression modeling in all of the major software systems that support analysis of complex sample survey data.

In theory, the finite population regression parameters for a generalized linear model of interest are those values that maximize a likelihood equation for the $i = 1, \ldots, N$ elements in the survey population. For a binary dependent variable y (with possible values 0 or 1), the population likelihood can be defined as

$$L(\boldsymbol{B} \mid \boldsymbol{x}) = \prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{1-y_i} \tag{8.9}$$

where under the logit link, $\pi(\boldsymbol{x}_i)$ is evaluated using the logistic CDF and the parameters in the specified logistic regression model: $\pi(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i\boldsymbol{B}) / [1 + \exp(\boldsymbol{x}_i\boldsymbol{B})]$. Note that as in Chapter 7, the finite population model parameters are denoted by the standard alphabetic $\boldsymbol{B}$ to distinguish them from the superpopulation model parameters denoted by β. Here, as in linear regression, the distinction is primarily a theoretical one.

Estimates of these finite population regression parameters are obtained by maximizing the following estimate of the population likelihood, which is a weighted function of the observed sample data and the $\pi(\boldsymbol{x}_i)$ values:

$$\begin{aligned} PL(\boldsymbol{B} \mid \boldsymbol{x}) &= \prod_{i=1}^{n} \{\pi(\boldsymbol{x}_i)^{y_i} \cdot [1 - \pi(\boldsymbol{x}_i)]^{1-y_i}\}^{w_i} \\ with: \boldsymbol{\pi}(\boldsymbol{x}_i) &= \exp(\boldsymbol{x}_i\boldsymbol{B}) / [1 + \exp(\boldsymbol{x}_i\boldsymbol{B})] \end{aligned} \tag{8.10}$$

Like the standard MLE procedure, this weighted pseudo-likelihood function can be maximized using the iterative Newton–Raphson method, or related algorithms. (See Theory Box 8.2 for more detail.)

The next hurdle in analyzing logistic regression models for complex sample survey data is to estimate the sampling variances and covariances of the

THEORY BOX 8.2 PSEUDO-MAXIMUM LIKELIHOOD ESTIMATION FOR COMPLEX SAMPLE SURVEY DATA

For a binary dependent variable and binomial data likelihood, the pseudo-maximum likelihood approach to the estimation of the logistic regression parameters and their variance–covariance matrix requires the solution to the following vector of estimating equations:

$$S(\boldsymbol{B}) = \sum_h \sum_\alpha \sum_i w_{h\alpha i} \boldsymbol{D}'_{h\alpha i} [(\pi_{h\alpha i}(\boldsymbol{B})) \cdot (1 - \pi_{h\alpha i}(\boldsymbol{B}))]^{-1} (y_{h\alpha i} - \pi_{h\alpha i}(\boldsymbol{B})) = \mathbf{0}$$

where: (8.12)

$\boldsymbol{D}_{h\alpha i}$ is the vector of partial derivatives, $\dfrac{\delta(\pi_{h\alpha i}(\boldsymbol{B}))}{\delta B_j}; j = 0, \ldots, p$

In Equation 8.12, h is a stratum index, α is a cluster (or SECU) index within stratum h, and i is an index for individual observations within cluster α. The term $w_{h\alpha i}$ thus refers to the sampling weight for observation i. The term $\pi_{h\alpha i}(\boldsymbol{B})$ refers to the probability that the outcome variable is equal to 1 as a function of the parameter estimates and the observed data according to the specified logistic regression model. For a logistic regression model of a binary variable, this reduces to a system of $p + 1$ estimating equations (where p is the number of predictor variables, and there is one additional parameter corresponding to the intercept in the model):

$$S(\boldsymbol{B})_{\text{logistic}} = \sum_h \sum_\alpha \sum_i w_{h\alpha i} (y_{h\alpha i} - \pi_{h\alpha i}(\boldsymbol{B}))\, \boldsymbol{x}'_{h\alpha i} = \mathbf{0}$$

where: (8.13)

$x'_{h\alpha i}$ = a column vector of the $p+1$ design matrix elements for case $i = [1 x_{1,h\alpha i} \ldots \chi_{p,h\alpha i}]'$

For the probit regression model, the estimating equations reduce to

$$S(\boldsymbol{B})_{\text{probit}} = \sum_h \sum_\alpha \sum_i w_{h\alpha i} \frac{(y_{h\alpha i} - \pi_{h\alpha i}(\boldsymbol{B})) \cdot \phi(\boldsymbol{x}'_{h\alpha i}\boldsymbol{B})}{\pi_{h\alpha i}(\boldsymbol{B}) \cdot (1 - \pi_{h\alpha i}(\boldsymbol{B}))} \boldsymbol{x}'_{h\alpha i} = \mathbf{0}$$

where: (8.14)

$\phi(\boldsymbol{x}'_{h\alpha i}\boldsymbol{B})$ is the standard normal probability density function evaluated at $\boldsymbol{x}'_{h\alpha i}\boldsymbol{B}$.

The weighted parameter estimates are computed by using the Newton–Raphson method to derive a solution for $S(\boldsymbol{B}) = \mathbf{0}$ (Agresti,

2002). Binder showed that the vector of weighted parameter estimates based on pseudo-maximum likelihood estimation is *consistent* for $\boldsymbol{B}$ even when the sample design is complex—that is, the bias of this estimator is of order $1/n$, so that as the sample size gets larger (which is often the case with complex samples), the bias of the estimator approaches 0.

parameter estimates. Binder (1983) proposed a solution to this problem that applied a multivariate version of Taylor series linearization (TSL). The result is a **sandwich-type variance estimator** of the form

$$\text{var}(\hat{\boldsymbol{B}}) = (\boldsymbol{J}^{-1})\,\text{var}[S(\hat{\boldsymbol{B}})](\boldsymbol{J}^{-1}) \tag{8.11}$$

where $\boldsymbol{J}$ is a matrix of second derivatives with respect to the $\hat{B}_j$ of the pseudo-log-likelihood for the data (derived by applying the natural log function to the likelihood defined in (8.10)) and $\text{var}(S(\hat{\boldsymbol{B}}))$ is the variance–covariance matrix for the sample totals of the weighted "**score function**" for the individual observations used to fit the model. Interested readers can find a more mathematical treatment of this approach in Theory Boxes 8.2 and 8.3. Binder's linearized variance estimator, $\text{var}(\hat{\boldsymbol{B}})$, is the default variance estimator in most major software packages for survey data analysis. However, most systems also provide options to select a jackknife repeated replication (JRR) or balanced repeated replication (BRR) method to estimate the variance–covariance matrix, $\text{var}(\hat{\boldsymbol{B}})_{rep}$, for the estimated model coefficients.

8.5 Building the Logistic Regression Model: Stage 3, Evaluation of the Fitted Model

The next step in building the logistic regression model is to test the contribution of individual model parameters and effects and to evaluate the overall goodness of fit (GOF) of the model.

8.5.1 Wald Tests of Model Parameters

When fitting logistic regression models to data collected from simple random samples, the statistical significance of one or more logistic regression parameters can be evaluated using a likelihood ratio test. Under the null hypotheses H_0: $\beta_j = 0$ (single parameter) or H_0: $\beta_q = \mathbf{0}$ (with q parameters), the following test statistic G follows a chi-square distribution with either 1 (for a single parameter) or q degrees of freedom:

THEORY BOX 8.3 TAYLOR SERIES ESTIMATION OF $VAR(\hat{\boldsymbol{B}})$

The computation of variance estimators for the pseudo-maximum likelihood estimates of the finite population parameters in the logistic regression model makes use of the $\boldsymbol{J}$ matrix of second derivatives:

$$\begin{aligned}\boldsymbol{J} &= -\left[\frac{\delta^2 \ln PL(\boldsymbol{B})}{\delta^2 \boldsymbol{B}}\right] | \boldsymbol{B} = \hat{\boldsymbol{B}} \\ &= \sum_h \sum_\alpha \sum_i \boldsymbol{x}'_{h\alpha i} \boldsymbol{x}_{h\alpha i} w_{h\alpha i} \hat{\pi}_{h\alpha i}(\boldsymbol{B})(1-\hat{\pi}_{h\alpha i}(\boldsymbol{B}))\end{aligned} \tag{8.15}$$

Due to the weighting, stratification and clustering inherent to complex sample survey designs, $\boldsymbol{J}^{-1}$ is not equivalent to the variance–covariance matrix of the pseudo-maximum likelihood parameter estimates, as is the case in the simple random sample setting (see Section 8.4). Instead, a **sandwich-type variance estimator** is used, incorporating the matrix $\boldsymbol{J}$ and the estimated variance–covariance matrix of the weighted score equations from Equation 8.13:

$$\hat{Var}(\hat{\boldsymbol{B}}) = (\boldsymbol{J}^{-1}) \text{var}[S(\hat{\boldsymbol{B}})](\boldsymbol{J}^{-1}) \tag{8.16}$$

The symmetric matrix $\text{var}[S(\hat{\boldsymbol{B}})]$ is the variance–covariance matrix for the $p + 1$ estimating equations in Equation 8.13. Each of these $p + 1$ estimating equations is a summation over strata, clusters, and elements of the individual "scores" for the n survey respondents. Since each estimating equation is a sample total of respondents' scores, standard formulae for stratified sampling of ultimate clusters (Chapter 3) can be used to estimate the variances and covariances of the $p + 1$ sample totals. In vector notation,

$$\text{var}[S(\hat{\boldsymbol{B}})] = \frac{n-1}{n-(p+1)} \sum_{h=1}^{H} \frac{a_h}{(a_h-1)} \sum_{\alpha=1}^{a_h} (\boldsymbol{s}_{h\alpha} - \bar{\boldsymbol{s}}_h)'(\boldsymbol{s}_{h\alpha} - \bar{\boldsymbol{s}}_h)$$

which for n large is:

$$\text{var}[S(\hat{\boldsymbol{B}})] \cong \sum_{h=1}^{H} \frac{a_h}{(a_h-1)} \sum_{\alpha=1}^{a_h} (\boldsymbol{s}_{h\alpha} - \bar{\boldsymbol{s}}_h)'(\boldsymbol{s}_{h\alpha} - \bar{\boldsymbol{s}}_h) \tag{8.17}$$

where for the logistic link:

$$\boldsymbol{s}_{h\alpha} = \sum_{i=1}^{n_\alpha} \boldsymbol{s}_{h\alpha i} = \sum_{i=1}^{n_\alpha} w_{h\alpha i}(y_{h\alpha i} - \hat{\pi}_{h\alpha i}(\boldsymbol{B}))\, \boldsymbol{x}'_{h\alpha i} \text{ ; and}$$

$$\bar{\boldsymbol{s}}_h = \frac{1}{a_h} \sum_{\alpha=1}^{a_h} \boldsymbol{s}_{h\alpha}$$

The estimator of $\text{var}[s(\hat{B})]$ for the probit or CLL link is obtained by substituting the appropriate expressions for the individual score functions in the calculation of the $\boldsymbol{s}_{h\alpha}$. For more details, interested readers should refer to Binder (1983).

$$G = -2 \ln \left[\frac{L(\hat{\boldsymbol{\beta}}_{\text{MLE}})_{reduced}}{L(\hat{\boldsymbol{\beta}}_{\text{MLE}})_{full}} \right] \tag{8.18}$$

where:

$$L(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = \text{ the likelihood under the model evaluated at the maximum likelihood estimates of } \boldsymbol{\beta}.$$

The *reduced* model in this case is the model excluding the q regression parameters to be tested, while the *full* model is the model including the q regression parameters. Both models should be fitted using *exactly* the same set of observations for this type of test to be valid.

As described in Section 7.3.4 for the linear regression model, complex sample designs invalidate the key assumptions that underlie the *F*-tests or likelihood ratio tests used to compare alternative models. Instead, Wald-type tests are used to test hypotheses concerning the parameters of a specified logistic regression model. The default output from software procedures enabling analysts to fit logistic regression models to complex sample survey data provides a table of the estimated model coefficients, the estimated standard errors, and the test statistic and a "*p*-value" for the simple hypothesis test, $H_0: B_j = 0$. (See, for example, Table 8.1.) Different software applications may report the test statistic as a Student *t*-statistic, $t = \hat{B}_j / se(\hat{B}_j)$, or as a Wald X^2. If the former of the two tests is output, the test statistic is referred to a Student *t* distribution with nominal design-based degrees of freedom (*df* = #clusters – #strata) to determine and report the *p*-value. If the output is in the form of

TABLE 8.1

Estimated Logistic Regression Model for Arthritis

Predictor[a]	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{56} > t)$
INTERCEPT	Constant	-2.752	0.137	-20.0	< 0.01
GENDER	Male	-0.595	0.045	-13.34	< 0.01
ED3CAT	<12 yrs	0.456	0.056	8.12	< 0.01
	12 yrs	0.267	0.042	6.41	< 0.01
AGE (years)	Continuous	0.047	0.002	22.11	< 0.01

Source: Analysis based on the 2006 HRS data.

Note: n = 18,374.

[a] Reference categories for categorical predictors are GENDER (female); ED3CAT(>12 yrs).

the Wald X^2, the reference distribution for determining the p-value is χ^2_1, or a central chi-square distribution with one degree of freedom. The two tests are functionally equivalent. In fact, the absolute value of the Student t test statistic is simply the square root of the Wald X^2 test statistic for this single parameter hypothesis test.

More generally, logistic regression software programs for complex sample survey data provide convenient syntax to specify Wald tests for a variety of hypotheses concerning the full vector of regression parameters. The general form of the null hypothesis for a Wald test is defined by H_0: $\boldsymbol{CB} = 0$, where $\boldsymbol{C}$ is a matrix of constants that defines the hypothesis being tested (see Section 7.3.4.1 for examples). The Wald test statistic is computed as

$$X^2_{Wald} = (\boldsymbol{C}\hat{\boldsymbol{B}})'[\boldsymbol{C}(\mathbf{var}(\hat{\boldsymbol{B}}))\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{B}}), \tag{8.19}$$

where $\mathbf{var}(\hat{\boldsymbol{B}})$ is a design-consistent estimate of the variance–covariance matrix of the estimated logistic regression coefficients (see Equation 8.11 for an example based on Taylor series linearization). Under the null hypothesis, this Wald test statistic follows a chi-square distribution with q degrees of freedom, where q is the rank, or number of independent rows, of the matrix $\boldsymbol{C}$. This test statistic can also be converted into an approximate F-statistic by dividing the Wald X^2 test statistic by the degrees of freedom.

In Stata, the `test` postestimation command is used to specify a multiparameter hypothesis test after fitting a model. Section 7.5 has already illustrated the application of the `test` command for constructing hypothesis tests for parameters in the linear regression model. The syntax is identical for constructing hypothesis tests for the parameters of the logistic or other generalized linear models estimated in Stata. For example, after fitting a logistic regression model that includes two indicator variables to represent the effect

of three National Health and Nutrition Examination Survey (NHANES) race/ethnicity categories (1 = White, 2 = Black, 3 = Other), the command

```
test _Irace_2 _Irace_3
```

tests the combined effects of race, i.e. H_0: $\{B_{\text{Black}} = 0,\ B_{\text{Other}} = 0\}$, while the command

```
test _Irace_2 -_Irace_3
```

tests the hypothesis H_0: $\{B_{\text{Black}} - B_{\text{Other}} = 0\}$ or equivalently H_0: $\{B_{\text{Black}} = B_{\text{Other}}\}$.

It is important to emphasize here again that survey analysts must be cautious about interpreting single parameter hypothesis tests when the estimated coefficient applies to an indicator variable for a multicategory predictor (e.g., levels of education) or when the model also includes an interaction term that incorporates the predictor of interest. In the former case, a significant result indicates that the category expectation (for the outcome) is significantly different from that of the reference category, but not necessarily other levels of the multicategory predictor. Tests of parameters for main effects in a model with interactions involving that same variable are confounded and not easily interpreted (see Section 8.6 for an example of interpretation of interaction terms).

8.5.2 Goodness of Fit and Logistic Regression Diagnostics

Summary statistics to measure overall goodness of fit and methods for diagnosing the fit of a logistic regression model for individual cases or specific patterns of covariates have been developed for simple random samples of data. These goodness-of-fit statistics and diagnostic tools have been included in the standard logistic regressions programs in most major statistical software packages. Hosmer and Lemeshow (2000, Chapter 5) review these summary statistics and diagnostic methods in detail.

Included in the summary techniques are (1) two test statistics based on the Pearson and deviance residuals for the fitted model; (2) the Hosmer–Lemeshow goodness-of-fit test; (3) classification tables comparing observed values of y with discrete classifications formed from the model's predicted values, $\hat{\pi}(\boldsymbol{x})$; and (4) the area under the receiver operating characteristic (ROC) curve. Several **pseudo-R^2 measures** have also been proposed as summary measures of the fit of a logistic regression model. However, since these measures tend to be incorrectly confused with the R^2 values (explained variability) in linear regression, we agree with Hosmer and Lemeshow (2000) that while they may be used by the analyst to compare the fits of alternative models they should not be cited as a measure of fit in scientific papers or reports.

Archer and Lemeshow (A-L; 2006) and Archer, Lemeshow, and Hosmer (2007) have extended the standard Hosmer and Lemeshow goodness-of-fit test for application to complex sample survey data. The A-L procedure is a modification of the standard Hosmer–Lemeshow test for goodness of fit that takes the sampling weights and the stratification and clustering features of the complex sample design into account when assessing the residuals (y_i - $\hat{\pi}(\boldsymbol{x}_i)$) based on the fitted model. Archer and Lemeshow's paper should be consulted for more details, but Stata users can download the .ado file implementing the `svylogitgof` postestimation command by submitting this command: `findit svylogitgof`.

At the time of this writing, survey analysis software is still in a phase where summary measures of goodness of fit are being translated or newly developed for application to logistic regression modeling of complex sample survey data. Furthermore, it may be some time before the major software systems routinely incorporate robust goodness-of-fit evaluation procedures in the software procedures for complex sample survey data analysis. In lieu of simply bypassing the goodness-of-fit evaluation entirely, we recommend the following:

1. Applying the Archer and Lemeshow goodness-of-fit test and other available summary goodness-of-fit measures when they are available in the software system.
2. If the logistic regression program for complex sample survey data in the chosen software system does not provide capabilities to generate summary goodness-of-fit measures, reestimate the model using the sampling weights in the system's standard logistic regression program. The weighted estimates of parameters and predicted probabilities will be identical and serious lack of fit should be quantifiable even though the standard program tools do not correctly reflect the variances and covariances of the parameter estimates given the complex sample design.

Summary measures such as the Archer and Lemeshow test statistic have the advantage that they yield a single test of the overall suitability of the fitted model. But even when a summary goodness-of-fit measure suggests that the model provides an acceptable fit to the data, a thorough evaluation of the fit of model may also include examination of the fit for specific patterns of covariates. Does the model fit well for some patterns of covariates x_j but not for others? The number and statistical complexity of diagnostic tools that have been suggested in the literature preclude a detailed discussion here. We encourage survey analysts to consult Hosmer and Lemeshow (2000) for a description of the diagnostic options and guidance on how these computational and graphic methods may be applied using Stata, SAS, and other software systems. Regarding regression diagnostics

for logistic regression models fit to complex sample survey data, we can offer the following recommendations at this writing:

1. Use one or more of the techniques described in Chapter 5 of Hosmer and Lemeshow (2000) to evaluate the fit of the model for individual patterns of covariates. If the complex sample logistic regression modeling program in your chosen software system (e.g., SAS PROC SURVEYLOGISTIC) does not include the full set of diagnostic capabilities of the standard programs, use the standard programs (e.g., SAS PROC LOGISTIC) with a weight specification. As mentioned before, the weighted estimates of parameters and predicted probabilities will be identical and serious breakdowns in the model for specific covariate patterns should be identifiable even though the standard program does not correctly reflect the variances and covariances of the parameter estimates given the complex sample design.
2. Remember that regression diagnostics serve to inform the analyst of specific cases where the model provides a poor fit or cases that exert extreme influence on the overall fit of the model. They are useful in identifying improvements in the model or anomalies that warrant future investigation, but small numbers of "predictive failures" occur in almost all regression modeling and do not necessarily invalidate the entire model.

8.6 Building the Logistic Regression Model: Stage 4, Interpretation and Inference

In logistic regression modeling, one can make inferences concerning the significance and importance of individual predictor variables in several ways. As described in Section 8.5.1, Wald X^2 tests may be employed to test the null hypothesis that a single coefficient is equal to zero, H_0: $B_j = 0$, or more complex hypotheses concerning multiple parameters in the fitted model. Confidence intervals (CIs) for individual model coefficients may also be used to draw inferences concerning the significance of model predictors and to provide information on the potential magnitude and uncertainty associated with the estimated effects of individual predictor variables.

Recall from Section 6.4.5 that an estimated logistic regression coefficient is the natural logarithm of the odds ratio comparing the odds that $y = 1$ for a predictor with value $x + 1$ to the odds that $y = 1$ when that predictor has a value x. In addition, the estimated coefficient for an indicator variable

associated with a level of a categorical predictor is the natural logarithm of the odds ratio comparing the odds that $y = 1$ for the level represented by the indicator to the odds that $y = 1$ for the reference level of the categorical variable. Consequently, the estimated coefficients are often labeled the **log-odds** for the corresponding predictor in the model. A design-based confidence interval for the logistic regression parameter is computed as

$$CI_{1-\alpha}(B_j) = \hat{B}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_j) \tag{8.20}$$

Typically, $\alpha = 0.05$ is used (along with the design-based degrees of freedom, *df*), and the result is a 95% confidence interval for the parameter. In theory, the correct inference to make is that over repeated sampling, 95 of 100 confidence intervals computed in this way are expected to include the true population value of B_j. If the estimated CI includes ln(1) = 0, analysts may choose to infer that H_0: $B_j = 0$ is accepted with a Type I error rate of $\alpha = 0.05$.

Inference concerning the significance/importance of predictors can be performed directly for the $\hat{B}_j s$ (on the log-odds scale). However, to quantify the magnitude of the effect of an individual predictor, it is more useful to transform the inference to a scale that is easily interpreted by scientists, policy makers, and the general public. As discussed in Section 6.4.5, in a logistic regression model with a single predictor, x_1, an estimate of the odds ratio corresponding to a one unit increase in the value of x_1 can be obtained by exponentiating the estimated logistic regression coefficient:

$$\hat{\psi} = \exp(\hat{B}_1) \tag{8.21}$$

If the model contains only a single predictor, the result is an estimate of the **unadjusted odds ratio**. If the fitted logistic regression model includes multiple predictors, that is,

$$\text{logit}(\hat{\pi}(\boldsymbol{x})) = \ln\left[\frac{\hat{\pi}(\boldsymbol{x})}{1\text{-}\hat{\pi}(\boldsymbol{x})}\right] = \hat{B}_0 + \hat{B}_1 x_1 + \cdots + \hat{B}_p x_p \tag{8.22}$$

the result $\hat{\psi}_j \mid \hat{\boldsymbol{B}}_{k \neq j} = \exp(\hat{B}_j)$ is an **adjusted odds ratio**. In general, the adjusted odds ratio represents the *multiplicative* impact of a one-unit increase in the predictor variable x_j on the odds of the outcome variable being equal to 1, holding all other predictor variables constant. Confidence limits can also be computed for adjusted odds ratios:

$$CI(\psi_j) = \exp(\hat{B}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_j)) \tag{8.23}$$

Software procedures for logistic regression analysis of survey data generally offer the analyst the option to output parameter estimates and standard errors on the log-odds scale (the original $\hat{B}_j s$) or transformed estimates of the corresponding adjusted odds ratios and confidence intervals.

The adjusted odds ratios and confidence intervals can be estimated and reported for any form of a predictor variable, including categorical variables, ordinal variables, and continuous variables. To illustrate, consider a simple logistic regression model (based on the 2006 Health and Retirement Study [HRS] data) of the probability that a U.S. adult age 50+ has arthritis. The predictors in this main effects only model are gender, education level (with levels less than high school, high school, and more than high school), and age:

$$\text{logit}(\pi(\boldsymbol{x})) = B_0 + B_1 I_{Male} + B_2 I_{Educ,<HS} + B_3 I_{Educ,HS} + B_4 X_{Age(yrs)}$$

where:

I_{Male} = indicator variable for male gender (female is reference);

$I_{Educ,<HS}, I_{Educ,HS}$ = indicators for education level (>high school is reference);

$X_{Age(yrs)}$ = respondent age in years.

The results from fitting this simple model in Stata (code not shown) are summarized in Tables 8.1 and 8.2.

TABLE 8.2

Estimates of Adjusted Odds Ratios in the Arthritis Model and 95% CIs for the Odds Ratios

Predictor[a]	Category	$\hat{\psi}$	$CI_{.95}(\psi)$
GENDER	Male	0.552	(0.505, 0.603)
ED3CAT	<12 yrs	1.558	(1.410, 1.766)
	12 yrs	1.306	(1.202, 1.420)
AGE	Continuous	1.050	(1.044, 1.053)

Source: Analysis based on the 2006 HRS data.

[a] Reference categories for categorical predictors are: GENDER (female); ED3CAT(>12 yrs).

Interpreting the output from this simple example, we can make the following statements:

- The estimated ratio of odds of arthritis for men relative to women is $\hat{\psi} = 0.55$.
- The estimated odds of arthritis for persons with less than a high school education are $\hat{\psi} = 1.56$ times the odds of arthritis for persons with more than a high school education.
- The estimated odds of arthritis increase by a factor of $\hat{\psi} = 1.05$ for each additional year of age.

Note that for continuous predictors, the increment x to $x + 1$ can be a relatively small step on the full range of x. For this reason, analysts may choose to report odds ratios for continuous predictors for a greater increment in x. A common choice is to report the odds ratio for a one standard deviation increase in x. For example, the standard deviation of 2006 HRS respondents' age is approximately 10 years. The estimated odds ratio and the 95% confidence interval for the odds ratio associated with a one standard deviation increase in age are computed as follows:

$$\hat{\psi}_{10yrs} = e^{\hat{B}_4 \cdot 10} = e^{0.047 \times 10} = \exp(0.047 \times 10) = 1.60$$

$$CI_{.95}(\psi_{10yrs}) = \left(\exp(\hat{B}_4 \times 10 - t_{56,0.975} \times 10 \times se(\hat{B}_4)), \exp(\hat{B}_4 \times 10 + t_{56,0.975} \times 10 \times se(\hat{B}_4))\right)$$

$$= \exp(0.47 \pm 2.003 \times 10 \times 0.002) = (1.54, 1.67)$$

When interactions between predictor variables are included in the specified model, analysts need to carefully consider the interpretation of the parameter estimates. For example, consider an extension of the 2006 HRS model for the logit of the probability of arthritis that includes the first-order interaction of education level and gender. The estimated coefficients and standard errors for this extended model reported by Stata (code not shown) are shown in Table 8.3.

Note that when the interaction of gender and education is introduced in the model, the parameter estimates for age, gender, and less than high school education change slightly but the estimated parameter for high school education is substantially reduced from $\hat{B}_{HS} = 0.267$ to $\hat{B}_{HS} = 0.177$. This is due to the fact that this parameter now represents the contrast in log-odds between high school education and greater than high school education for *females only* (the reference level for gender) and is combined with the parameter for the (12 yrs × Male) product term to define the same contrast in log-odds for males. The parameter associated with the product

TABLE 8.3

Estimated Logistic Regression Model for Arthritis, Including the First-Order Interaction of Education and Gender

Predictor[a]	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{56} > t)$
INTERCEPT	Constant	–2.728	0.135	–20.22	< 0.01
GENDER	Male	–0.659	0.061	–10.81	< 0.01
ED3CAT	<12 yrs	0.454	0.063	7.20	< 0.01
	12 yrs	0.177	0.050	3.56	< 0.01
AGE	Continuous	0.047	0.002	22.11	< 0.01
ED3CAT × GENDER	<12 yrs × Male	0.004	0.102	0.04	0.970
	12 yrs × Male	0.201	0.087	2.20	0.026

Source: Analysis based on the 2006 HRS data.

[a] Reference categories for categorical predictors are GENDER (female); ED3CAT(>12 yrs).

term therefore represents a *change* in this contrast for males relative to females. Apparently linked to this decrease in "main effect" size for high school education is a significant positive interaction between a 12th-grade education and male gender. The estimated change in log-odds for males with 12th-grade education relative to males with greater than 12th-grade education is thus computed as 0.177 + 0.201 = 0.378. (Although the results are not shown, the first-order interaction of GENDER and AGE was tested in a separate model and was not significant.)

To explore the impact of the interaction of GENDER and ED3CAT on the estimated logits and odds ratios, assume that AGE is fixed at 65 years. Consider the patterns of covariates shown in columns 2–4 of Table 8.4.

To estimate the value of the logit for each covariate pattern, the estimated coefficients in Table 8.3 are applied to the corresponding values of the predictor variables:

$$\text{logit}(\pi(\boldsymbol{x})) = -2.728 - 0.659 I_{Male} + 0.454 I_{Educ,<HS} + 0.177 I_{Educ,HS} + 0.047 X_{Age(yrs)}$$
$$+ 0.004(I_{Male} \times I_{Educ,<HS}) + 0.201(I_{Male} \times I_{Educ,HS})$$

Table 8.4 shows the estimated logits for the six unique covariate patterns (with age fixed at 65). To evaluate the ratio of odds of arthritis for men and women of different education levels, a general technique to compare the estimated odds for two different patterns of covariates is used. Consider two "patterns" of covariate values $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Using the example in Table 8.4, $\boldsymbol{x}_1$ might be pattern 1, 65 year old males with <HS education, and $\boldsymbol{x}_2$ might be the reference category for the GENDER × ED3CAT interaction, which

TABLE 8.4

Covariate Patterns, Logits, and Odds Ratios for the 2006 HRS Arthritis Model

Pattern (j)	Gender	Education	Age	Logit: z_j	Odds Ratio[a]
1	Male	<HS	65	0.190	0.82
2	Male	HS	65	0.111	0.75
3	Male	>HS	65	–0.267	0.52
4	Female	<HS	65	0.845	1.57
5	Female	HS	65	0.569	1.19
6	Female	>HS	65	0.392	1.00

[a] Relative to joint reference of Female with > High School Education.

would be 65 year old women with >HS education. To obtain the estimated odds ratio that compares $\boldsymbol{x}_1$ with $\boldsymbol{x}_2$, the following five steps are required:

1. Based on the estimated model, compute the values of the logit function for the two sets of covariates:

$$\text{logit}_1 = \sum_{j=0}^{p} \hat{B}_j x_{1j};\ \text{logit}_2 = \sum_{j=0}^{p} \hat{B}_j x_{2j}$$

   These are shown for the six example covariate patterns in Table 8.4; for example, $\text{logit}_1 = 0.190$ (Pattern 1) and $\text{logit}_2 = 0.392$ (Pattern 6).
2. Compute the difference between the two logits, $\hat{\Delta}_{1:2} = \text{logit}_1 - \text{logit}_2 = 0.190 - 0.392 = -0.202$.
3. Exponentiate the difference in the two logits to estimate the odds ratio comparing $\boldsymbol{x}_1$ with $\boldsymbol{x}_2$, $\hat{\psi}_{1:2} = e^{\hat{\Delta}_{1:2}} = e^{-.202} = 0.817$.
4. To reflect the uncertainty in the estimated odds ratios, the estimates should be accompanied by an estimated confidence interval. The CI of an odds ratio comparing two arbitrary covariate patterns, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, takes the general form of expression 8.23. The standard error estimate requires the algebraically complicated derivation of the standard error of the difference in the two logits:

$$se(\hat{\Delta}_{1:2}) = \sqrt{\sum_{j=0}^{p}(x_{1j} - x_{2j})\text{var}(\hat{B}_j) + 2\sum_{j<k}(x_{1j} - x_{2j})(x_{1k} - x_{2k})\text{cov}(\hat{B}_j, \hat{B}_k)} \quad (8.24)$$

   Note that any common values for x_j in logit_1 and logit_2 can be ignored in evaluating this standard error. The calculation of the

standard error of the difference in logits requires the values of $\text{var}(\hat{B}_j)$ and $\text{cov}(\hat{B}_j, \hat{B}_k)$. These may be obtained in Stata by issuing the `estat vce` command after the model has been estimated.

5. Exponentiate the CI limits for the difference in logits to estimate the odds ratio and its 95% CI:

$$\hat{\psi}_{1:2} = \exp(\hat{\Delta}_{1:2});\ CI(\hat{\psi}_{1:2}) = \exp[\hat{\Delta}_{1:2} \pm t_{df,.975} \cdot se(\hat{\Delta}_{1:2})]$$

The technique just described for estimating odds ratios applies generally to *any* observed patterns of covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The final column in Table 8.4 shows the estimated odds ratios comparing each of the six covariate patterns based on the estimated logistic regression model including the interaction of gender and education level. Holding age constant at 65 years, the coding of gender (female reference), and education level (>High School is the reference) results in 65 year old women with >HS education as the natural reference group. A convenient way to analyze and report odds ratios for predictors that have a significant interaction is to use a graphical display of the type shown for the arthritis example in Figure 8.3.

The X (horizontal) axis in Figure 8.3 represents the three ordinal education categories. The Y (vertical) axis is the value of the estimated odds ratio, using women with >HS education as the comparison group. At each education level, the estimated odds ratios are plotted separately for men and women. Note that compared with the women, the odds of arthritis are lower for men and, consistent with the significant interaction in the estimated model, drop substantially for men in the >HS education group. Confidence bars may also be added to this graphical display to enhance the presentation and visual comparison of odds ratios for important covariate patterns.

We include detailed code for generating Figure 8.3 in Stata on the book's Web site.

8.7 Analysis Application

This section presents an example logistic regression analysis that follows the four general modeling stages described in Sections 8.3 through 8.6.

Example 8.1: Examining Predictors of a Lifetime Major Depressive Episode in the NCS-R Data

The aim of this example is to build a logistic regression model for the probability that a U.S. adult has been diagnosed with major depressive episode (MDE) in their lifetime. The dependent variable is the NCS-R variable MDE, which takes a

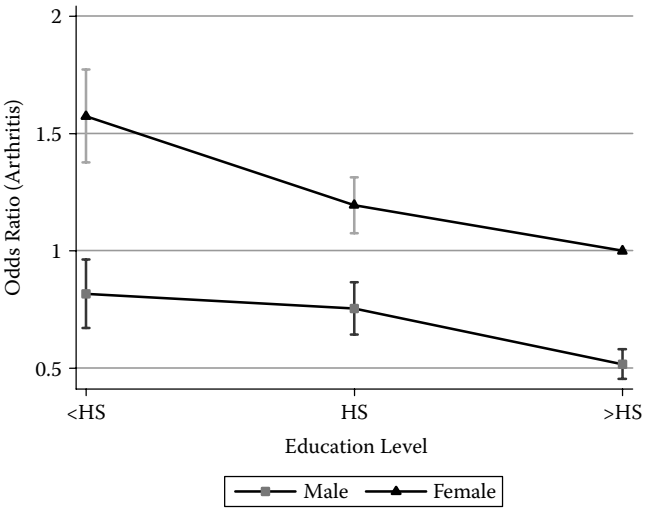


FIGURE 8.3
Plot of estimated odds ratios, showing the interaction between gender and education in the arthritis model. (Modified from the 2006 HRS data.)

value of 1 for persons who meet lifetime criteria for major depression and 0 for all others. The following predictors are considered: AG4CAT (a categorical variable measuring age brackets, including 18–29, 30–44, 45–59, and 60+), SEX (1 = Male, 2 = Female), ALD (an indicator of any lifetime alcohol dependence), ED4CAT (a categorical variable measuring education brackets, including 0–11 years, 12 years, 13–15 years, and 16+ years), and MAR3CAT (a categorical variable measuring marital status, with values 1 = "married," 2 = "separated/widowed/divorced," and 3 = "never married"). The primary research question of analytical interest is whether MDE is related to alcohol dependence after adjusting for the effects of the previously listed demographic factors .

8.7.1 Stage 1: Model Specification

The analysis session begins by specifying the complex design features of the NCS-R sample in the Stata `svyset` command. Note that we specify the "long" or Part 2 NCS-R sampling weight (NCSRWTLG) in the `svyset` command. This is due to the use of the alcohol dependence variable in the analysis, which was measured in Part 2 of the NCS-R survey.

There are 42 sampling error strata and 84 sampling error computation units (two per stratum) in the NCS-R sampling error calculation model, resulting in 42 design-based degrees of freedom.

Following the recommendations of Hosmer and Lemeshow (2000), the model building begins by examining the bivariate associations of MDE with each of the potential predictor variables. Since the candidate predictors are all categorical variables, the bivariate relationship of each predictor with

MDE is analyzed in Stata by using the `svy: tab` command and requesting row percentages (as discussed in Chapter 6).

```
svy: tab ag4cat mde, row
svy: tab sex mde, row
svy: tab ald mde, row
svy: tab ed4cat mde, row
svy: tab mar3cat mde, row
```

Table 8.5 presents the results of these bivariate analyses, including the Rao–Scott *F*-tests of association. The table also presents estimates of the percentages of each predictor category that received the lifetime MDE diagnosis.

Based on these initial tests of association, all of the predictor variables appear to have significant bivariate associations with MDE, including ALD, the indicator of lifetime alcohol dependence, and all of the predictors appear to be good candidates for inclusion in the initial multivariate logistic regression model.

8.7.2 Stage 2: Model Estimation

The next step in the Hosmer–Lemeshow (2000) model building procedure is to fit the "initial" multivariate model, examining the main effects for all five

TABLE 8.5

Initial Bivariate Design-Based Tests of Association Assessing Potential Predictors of Lifetime Major Depressive Episode (MDE) for the NCS-R Adult Sample

Predictor	Rao–Scott *F*-test[a]	Category	% with MDE (SE)
AG4CAT	$F(2.76,115.97) = 26.39$ $p < 0.01$	18–29	18.4 (0.9)
		30–44	22.9 (1.1)
		45–59	22.3 (1.3)
		60+	11.1 (1.0)
SEX	$F(1,42) = 44.83$ $p < 0.01$	Male	15.3 (0.9)
		Female	22.6 (0.7)
ALD	$F(1,42) = 120.03$ $p < 0.01$	Yes	45.2 (0.3)
		No	17.7 (0.7)
ED4CAT	$F(2.90,121.93) = 4.30$ $p < 0.01$	<12 yrs	16.3 (1.2)
		12 yrs	18.6 (0.8)
		13–15 yrs	21.3 (1.0)
		16+ yrs	19.7 (1.1)
MAR3CAT	$F(1.90,79.74) = 11.08$ $p < 0.01$	Married	17.3 (0.7)
		Previously	23.9 (1.5)
		Never	19.4 (1.1)

[a] See Chapter 6 for more details on the derivation of this test statistic, which can be used to test the null hypothesis of no association between the predictor variable and the outcome variable (MDE).

predictors. Stata provides two programs for fitting the multivariate logistic regression model: `svy: logistic` and `svy: logit`. The default output for the `svy: logistic` command is estimates of adjusted odds ratios and 95% confidence intervals for the adjusted odds ratios. Note the use of the `char` command and specification of the omitted reference category for SEX (2, or females) prior to the `svy: logistic` command. This command allows the user to specify custom reference categories rather than accepting the default of having the lowest category omitted:

```
char sex[omit] 2
xi: svy: logistic mde i.ag4cat i.sex ald i.ed4cat i.mar3cat
```

In `svy: logistic`, estimates of the parameters and standard errors for the logit model can be requested by using the `coef` option.

```
xi: svy: logistic mde i.ag4cat i.sex ald i.ed4cat ///
i.mar3cat, coef
```

Stata users who wish to see the estimated logistic regression coefficients and standard errors may also use the companion program, `svy: logit`.

```
xi: svy: logit mde i.ag4cat i.sex ald i.ed4cat i.mar3cat
```

Estimated odds ratios and 95% CIs can be generated in `svy: logit` by adding the `or` option:

```
xi: svy: logit mde i.ag4cat i.sex ald i.ed4cat ///
i.mar3cat, or
```

Other software systems (e.g., SAS PROC SURVEYLOGISTIC) will output both the estimated logistic regression coefficients (and standard errors) and the corresponding odds ratio estimates. (Readers should be aware that the `svy: logit` and `svy: logistic` commands differ slightly in the procedures for calculation of the standard errors for odds ratios.)

In each form of the Stata `svy: logistic` or `svy: logit` command, the dependent variable, MDE, is listed first, followed by the predictor variables (and their interactions, if applicable). Note that the `xi:` modifier is used in conjunction with the `svy: logistic` and `svy: logit` commands to indicate those predictor variables that are categorical in nature: AG4CAT, ED4CAT, SEX, and MAR3CAT. (Because the ALD variable is already coded as either 0 or 1, it does not require program-generated indicator variables.) The `xi:` specification of the model instructs Stata to create a set of indicator variables for each independent variable preceded by the `i.` prefix. Stata will create $K - 1$ indicator variables to represent the K levels of the categorical predictor. By default, Stata will select the lowest valued category as the reference category and only the $K - 1$ indicators for the remaining categories will be included as predictors in the model.

TABLE 8.6

Estimated Logistic Regression Model for the Lifetime MDE Outcome (Output Generated by Using the `svy: logit` Command)

Predictor[a]	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{42} > t)$
INTERCEPT		–1.583	0.121	–13.12	<0.001
AG4CAT	30–44	0.255	0.094	2.71	0.01
	45–59	0.206	0.092	2.26	0.029
	60+	–0.676	0.141	–4.78	< 0.001
SEX	Male	–0.577	0.077	–7.48	< 0.001
ALD	Yes	1.424	0.154	9.24	< 0.001
ED4CAT	12	0.079	0.097	0.82	0.418
	13–15	0.231	0.093	2.48	0.017
	16+	0.163	0.111	1.47	0.148
MAR3CAT	Previously	0.486	0.085	5.69	<0.001
	Never	0.116	0.108	1.07	0.290

Source: Analysis based on the NCS-R data.

Notes: n = 5,692, adjusted Wald test for all parameters: $F(10,33) = 28.07$, $p < 0.001$.

[a] Reference categories for categorical predictors are: AG4CAT (18–29); SEX (Female); ALD (No); ED4CAT (<12 yrs); MAR3CAT (Married).

Tables 8.6 and 8.7 summarize the output generated by fitting the MDE logistic regression model. The initial model includes main effects for the chosen predictor variable candidates but at this point does not include any interactions between the predictors.

8.7.3 Stage 3: Model Evaluation

The adjusted Wald tests in Stata for the AG4CAT, ED4CAT, and MAR3CAT categorical predictors in this initial model are generated by using the `test` command:

```
test _Iag4cat_2 _Iag4cat_3 _Iag4cat_4
test _Imar3cat_2 _Imar3cat_3
test _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
```

Note that we do not request these multiparameter Wald tests for the SEX and ALD predictor variables, because they are represented by single indicator variables in the regression model and the overall Wald test for each predictor is equivalent to the t-test reported for the single estimated parameter for that predictor. Further, note in each of the `test` statements that we include the $K - 1$ indicator variables generated by Stata for each of the categorical predictors (e.g., `_Iag4cat _2`) when the `xi:` modifier is used to identify categorical predictor variables. Stata users can find the names of

TABLE 8.7

Estimates of Adjusted Odds Ratios for the Lifetime MDE Outcome

Predictor[a]	Category	$\hat{\psi}$	95% CI for ψ
AG4CAT	30–44	1.29	(1.067, 1.562)
	45–59	1.23	(1.022, 1.479)
	60+	0.51	(0.383, 0.677)
SEX	Male	0.56	(0.480, 0.656)
ALD	Yes	4.15	(3.042, 5.668)
ED4CAT	12	1.08	(0.890, 1.316)
	13–15	1.26	(1.044, 1.519)
	16+	1.18	(0.941, 1.471)
MAR3CAT	Previously	1.63	(1.369, 1.932)
	Never	1.12	(0.903, 1.396)

Source: Analysis based on the NCS-R data.

Notes: n = 5,692. Adjusted Wald test for all parameters: $F(10,33) = 28.07$. $p < 0.001$.

[a] Reference categories for categorical predictors are: AG4CAT (18–29); SEX (Female); ALD (No); ED4CAT (<12 yrs); MAR3CAT (Married).

these indicators in the Stata Variables window once the model has been fitted. Table 8.8 provides the design-adjusted F-versions of the resulting Wald test statistics and associated p-values.

Two of the three design-adjusted Wald tests are significant at the 0.01 level. The exception is the Wald test for education [$F(3,40) = 2.13$, $p = 0.112$], which suggests that the parameters associated with education in this logistic regression model are not significantly different from zero and that education may not be an important predictor of lifetime MDE when adjusting for the relationships of the other predictor variables with the outcome. If the objective of the model-building process is the construction of a parsimonious model, education could probably be dropped as a predictor at this point. For the purposes of this illustration (and because of the marginal significance), we will retain education in the model moving forward.

8.7.4 Stage 4: Model Interpretation/Inference

Based on the results in Table 8.6 and Table 8.8, it appears that each of the predictors in the multivariate model has a significant (or marginally significant) relationship with the probability of MDE after adjusting for the relationships of the other predictors. Focusing on the primary predictor variable of interest, we see that the odds of having had a major depressive episode at some point in the lifetime are multiplied by 4.15 when a person has had a diagnosis of alcohol dependence at some point in his or her lifetime, when adjusting

TABLE 8.8

Design-Adjusted Wald Tests for the Parameters Associated with the Categorical Predictors in the Initial MDE Logistic Regression Model

Categorical Predictor	*F*-Test Statistic	*P*-value
AG4CAT	$F_{(3,40)} = 19.03$	< 0.001
ED4CAT	$F_{(3,40)} = 2.13$	0.112
MAR3CAT	$F_{(2,40)} = 16.60$	< 0.001

Source: Analysis based on the NCS-R data.

for the relationships of age, sex, education, and marital status. Of course, this model does not allow for any kind of causal inference, given that time ordering of the events is not available in the NCS-R data set; we can, however, conclude that there is strong evidence of an association between the two disorders in this finite population when adjusting for other demographic covariates. We also note that relative to married respondents, respondents who were previously married have significantly higher (63% higher) odds of having had a major depressive episode in their lifetime when adjusting for the other covariates. Further, middle-age respondents have significantly higher odds of lifetime MDE (relative to younger respondents), while older respondents and males have significantly reduced odds of lifetime MDE (again relative to younger respondents and females).

Respondent age is represented in the model as four grouped categories of age. Including grouped categories for age (or recoded categories of any continuous predictor, more generally) in a logistic regression model will result in estimates of the expected contrasts in log-odds for respondents in each of the defined categories, relative to the reference category. Since the model parameters are estimated separately for each defined age group (with age 18–29 as the reference), the model will capture any nonlinearity of effect in the ordered age groupings. Inspecting the estimated coefficients and odds ratios for the grouped age categories in Table 8.6 and Table 8.7, it appears that there is significant nonlinearity in the effect of age on the probability of MDE. Relative to the 18–29-year-old group, the odds of MDE increase by factors of 1.29 (aged 30–44) and 1.23 (aged 45–59) for the middle-age ranges but decrease by a factor of 0.51 in the age 60 and older group. Such nonlinear effects of age are common in models of human disorders and are possibly attributable to normal processes of aging and selective mortality. If the example model was estimated with age (in years) as a continuous predictor variable, at this stage in the model-building process the analyst would reestimate the model including both the linear and quadratic terms for age.

Therefore, at this stage in the model-building process, we have chosen to retain all of the candidate main effects. Next, we apply Archer and Lemeshow's (2006) design-adjusted test to assess the goodness of fit of this initial model (assuming that this procedure has been downloaded and installed):

```
svylogitgof
```

The resulting design-adjusted *F*-statistic reported in the Stata Results window is equal to $F_{A\text{-}L} = 1.229$, with a *p*-value of 0.310. This suggests that the null hypothesis that the model fits the data well is not rejected. We therefore have confidence moving forward that the fit of this initial model is acceptable.

Next, we consider testing some scientifically relevant two-way interactions between the candidate predictor variables. For illustration purposes, we suppose that possible two-way interactions of sex with the other four covariates measuring age, lifetime alcohol dependence, education, and marital status are of interest, if sex is posited by an NCS-R analyst as being a possible moderator of the relationships of these other four covariates with lifetime MDE. We fit a model including these two-way interactions in Stata using the following command:

```
xi: svy: logistic mde i.ag4cat*i.sex i.sex*ald ///
i.ed4cat*i.sex i.mar3cat*i.sex, coef
```

Note how the interactions are specified in this command. When the `xi:` modifier is used for a regression command, the products of the two factors listed after the dependent variable specify that the regression parameters associated with each individual factor should be included in the regression model *in addition to* the parameters associated with the relevant cross-product terms defined by the interaction (e.g., the indicator for AG4CAT = 2 × the indicator for SEX = 1). In other words, listing AG4CAT and SEX in addition to the previous product terms would be redundant, and the main effects are included in the model by default when the interaction terms are specified. Table 8.9 presents the estimates of the regression parameters in this model generated by executing the previous command in Stata.

At this point, the statistical question is whether these two-way interactions are making a significant additional contribution or improvement to the fit of this model to the NCS-R data. That is, are any of the parameters associated with the two-way interaction terms significantly different from 0? We can test this hypothesis by once again using design-adjusted Wald tests. The relevant interaction terms for the regression model are automatically generated by Stata and included in the data set when using the `xi:` modifier, so the cross-product terms in the `test` commands that follow can be easily selected from the Stata Variables window:

```
test _Iag4Xsex_2_1 _Iag4Xsex_3_1 _Iag4Xsex_4_1
test _IsexXald_1
test _Ied4Xsex_2_1 _Ied4Xsex_3_1 _Ied4Xsex_4_1
test _ImarXsex_2_1 _ImarXsex_3_1
```

TABLE 8.9

Estimated Logistic Regression Model for Lifetime MDE, Including First Order Interactions of the Other Predictor Variables with SEX

Predictor[a]	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{42} > t)$
INTERCEPT	Constant	–1.600	0.134	–11.94	<0.001
AG4CAT	30–44	0.220	0.114	1.94	0.059
	45–59	0.214	0.102	2.09	0.042
	60+	–0.646	0.175	–3.68	0.001
SEX	Male	–0.546	0.357	–1.53	0.134
ALD	Yes	1.553	0.211	7.36	<0.001
ED4CAT	12	0.131	0.084	1.56	0.126
	13–15	0.297	0.117	2.54	0.015
	16+	0.242	0.152	1.59	0.118
MAR3CAT	Previously	0.418	0.111	3.78	<0.001
	Never	0.017	0.130	0.13	0.894
AG4CAT × SEX	30–44 × Male	0.097	0.201	0.48	0.633
	45–59 × Male	0.002	0.213	0.01	0.990
	60+ × Male	–0.038	0.302	–0.13	0.901
ALD × SEX	Yes × Male	–0.200	0.242	–0.83	0.413
ED4CAT × SEX	12 × Male	–0.138	0.271	–0.51	0.614
	13–15 × Male	–0.169	0.269	–0.63	0.534
	16+ × Male	–0.194	0.344	–0.56	0.576
MAR3CAT × SEX	Previously × Male	0.182	0.208	0.88	0.385
	Never × Male	0.232	0.212	1.09	0.280

Source: Analysis based on the NCS-R data.

Notes: n = 5,692. Adjusted Wald test for all parameters: $F(19,24) = 17.15$. $p < 0.001$.

[a] Reference categories for categorical predictors are: AG4CAT (18–29 yrs); GENDER (female); ALD (no); ED4CAT(<12 yrs); MAR3CAT (married); SEX(female).

Based on test results presented in Table 8.10, we fail to reject the null hypotheses for all four of the tests, suggesting that these two-way interactions are actually not making a significant contribution to the fit of the model. We therefore do not consider these two-way interactions any further and would proceed with making inferences based on the estimates from the model presented in Table 8.6.

8.8 Comparing the Logistic, Probit, and Complementary Log–Log GLMs for Binary Dependent Variables

This chapter has focused on logistic regression techniques for modeling $\pi(x)$ for a binary dependent variable. As discussed in Section 8.2, alternative

TABLE 8.10

Design-Adjusted Wald Tests of First-Order Interactions of Sex and Other Categorical Predictors in the MDE Logistic Regression Model

Interaction Term	*F*-Test Statistic	$P(\mathcal{F} > F)$
AG4CAT × SEX	$F_{(3,40)} = 0.25$	0.863
ALD × SEX	$F_{(1,42)} = 0.68$	0.413
ED4CAT × SEX	$F_{(3,40)} = 0.13$	0.944
MAR3CAT × SEX	$F_{(2,41)} = 0.77$	0.472

Source: Analysis based on the NCS-R data.

generalized linear models for a binary dependent variable may be estimated using the probit or CLL link function. In discussing these alternative GLMs, we noted that inferences derived from logistic, probit and CLL regression models should generally be consistent.

To illustrate, consider the results in Table 8.11 for a side-by-side comparison of estimated logistic, probit and CLL regression models. The example used for this comparison is a model of the probability that a U.S. adult is alcohol dependent. The data are from the NCS-R long interview (or Part 2 of the survey), and each model includes the same demographic main effects considered in Section 8.7 for the model of MDE: SEX, AG4CAT, ED4CAT, and MAR3CAT. The Stata commands for the estimation of the three models follow (note the use of the `char sex[omit]2` syntax to specify the desired omitted category for sex):

```
char sex[omit] 2
xi: svy: logit ald i.ag4cat i.sex i.ed4cat i.mar3cat
test _Iag4cat_2 _Iag4cat_3 _Iag4cat_4
test _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
test _Imar3cat_2 _Imar3cat_3
xi: svy: probit ald i.ag4cat i.sex i.ed4cat i.mar3cat
test _Iag4cat_2 _Iag4cat_3 _Iag4cat_4
test _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
test _Imar3cat_2 _Imar3cat_3
xi: svy: cloglog ald i.ag4cat i.sex i.ed4cat i.mar3cat
test _Iag4cat_2 _Iag4cat_3 _Iag4cat_4
test _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
test _Imar3cat_2 _Imar3cat_3
```

Table 8.11 presents a summary of the estimated coefficients, standard errors, and *p*-values for simple hypothesis tests of the form H_0: $B_j = 0$. Table 8.12 presents the results for the Wald tests of the overall age, education, and marital status effects. Note that although the coefficients and standard errors for the probit model show the expected difference in scale

TABLE 8.11

Comparison of Logistic, Probit, and CLL Models of Alcohol Dependency in U.S. Adults

		Logistic			Probit			C-L-L		
Predictor[a]	**Category**	$\hat{B}$	$se(\hat{B})$	p	$\hat{B}$	$se(\hat{B})$	p	$\hat{B}$	$se(\hat{B})$	p
Intercept		–3.124	0.225	<0.001	–1.719	0.105	<0.001	–3.140	0.218	<0.001
SEX	Male	0.997	0.119	<0.001	0.471	0.056	<0.001	0.965	0.115	<0.001
AG4CAT	30–44	0.146	0.178	0.416	0.065	0.084	0.444	0.143	0.171	0.408
	45–59	–0.051	0.144	0.726	–0.034	0.067	0.609	–0.045	0.140	0.748
	60+	–1.120	0.212	<0.001	–0.531	0.093	<0.001	–1.083	0.209	<0.001
ED4CAT	12 yrs	–0.268	0.194	0.173	–0.124	0.095	0.200	–0.260	0.185	0.167
	13–15 yrs	–0.264	0.176	0.141	–0.124	0.085	0.152	–0.256	0.169	0.137
	16+ yrs	–0.736	0.197	<0.001	–0.339	0.092	<0.001	–0.713	0.190	<0.001
MAR3CAT	Previously	0.517	0.142	<0.001	0.255	0.069	<0.001	0.494	0.136	<0.001
	Never	0.065	0.169	0.070	0.039	0.077	0.616	0.060	0.164	0.713

Source: Analysis based on the NCS-R data.

Notes: $n = 5{,}692$.

[a] Reference categories for categorical predictors are AG4CAT (18–29 yrs); SEX (female); ED4CAT(<12 yrs); MAR3CAT (Married).

TABLE 8.12

Design-Adjusted Wald Tests of Categorical Predictors in the MDE Models

	Wald F-Test Statistic (p-Value = $P(\mathcal{F} > F)$)		
Categorical Predictor	**Logistic**	**Probit**	**CLL**
AG4CAT	$F_{(3,40)}$ =12.06 (<0.001)	$F_{(3,40)}$ = 15.26 (<0.001)	$F_{(3,40)}$ = 11.52 (<0.001)
ED4CAT	$F_{(3,40)}$ = 4.80 (0.006)	$F_{(3,40)}$ = 4.79 (0.006)	$F_{(3,40)}$ = 4.77 (0.006)
MAR3CAT	$F_{(2,41)}$ = 6.54 (0.003)	$F_{(2,41)}$ = 6.66 (0.003)	$F_{(2,41)}$ = 6.50 (0.004)

Source: Analysis based on the NCS-R data.

from those of the logistic and CLL models, the three models produce similar p-values for the test of each parameter and would not lead to significant differences in inferences concerning the effects of the individual parameters. Given the similarity in these results, we recommend using the logit model for general applications. Faraway (2006, p. 38) points out three advantages of this approach: simpler mathematical formulation of the models, ease of interpretation via odds ratios, and easier analysis of retrospectively sampled data.

8.9 Exercises

1. Using the software procedure of your choice, fit the following simple (binary) logistic regression model to the 2006 HRS data set. Model the binary dependent variable, DIABETES (1 = yes, 0 = no) as a function of the following independent variables: AGE (KAGE), GENDER (1 = Male, 2 = Female), RACE (0 = None Given, 1 = White, 2 = Black, 7 = Other), and ARTHRITIS (1 = yes, 0 = no). Be sure to model the logit(P(Diabetes = 1)); that is, model the probability that a person has diabetes. The predictors GENDER, RACE, and ARTHRITIS should be treated as categorical, and the reference category parameterization should be used. For gender, choose females (2) as the reference category. For race, choose white (1) as the reference category. For arthritis, choose no (0) as the reference category. All analyses should use the 2006 final individual sampling weight, KWGTR, as the analysis weight, and all analyses should incorporate design adjustments for the stratification and clustering of the 2006 HRS sample (the STRATUM and SECU variables). Prepare a table showing the parameter estimates in this model, their design-based standard errors, and 95% confidence intervals for the parameters.
2. Based on the fitted model from Exercise 1, what is the estimated odds ratio comparing men's odds of diabetes with that for women

(holding all other factors constant)? What is a 95% confidence interval for this odds ratio? What would you conclude about the relationship of gender with diabetes based on these results?

3. Based on the fitted model from Exercise 1, if all other variables are held constant, what is the estimated odds ratio for diabetes associated with a 30-year increase in age? Compute the design-adjusted 95% confidence interval for this odds ratio.
4. Perform a joint design-adjusted Wald test of the null hypothesis that B_{male}, B_{black}, and $B_{\text{arthritis}}$ are all not significantly different from zero. Report the test statistic, the degrees of freedom, and a p-value for this test. How would you explain the result of this test to a colleague in plain English?
5. Perform a design-adjusted Wald test of the null hypothesis that $B_{\text{BLACK}} = B_{\text{OTHER}}$. Report the test statistic, the degrees of freedom and a p-value for this test. What does the result of the test mean in plain English?
6. Construct a new variable for the interaction of AGE and GENDER. Refit the original logistic regression model with this age × gender interaction term added. Test and report whether the interaction of age and gender significantly improves the fit of the model. *Hint*: HRS codes GENDER as 1 and 2. You have been asked to use female (2) as the reference category. To create the interaction variable, consider recoding sex to 1 = male, 0 = female. What is your interpretation of the interaction effect?
7. **(Stata Only)** Apply the Archer and Lemeshow (2006, 2007) procedure for testing the goodness of fit of a model of your choosing, and be clear about the model being tested. What is your conclusion about the fit of the model based on this test? Is this fit adequate or not?
8. Prepare a short discussion (two to four paragraphs) describing the results of your analysis of the specified set of potential risk factors for diabetes. To illustrate how one might use your estimated model in practice, include the detailed computation of the predicted probability of having diabetes for someone with a specified set of values on the covariates included in the model.

9

Generalized Linear Models for Multinomial, Ordinal, and Count Variables

9.1 Introduction

Chapter 8 covered generalized linear models (GLMs) for survey variables that are measured on a binary or dichotomous scale. The aim of this chapter is to introduce generalized linear modeling techniques for three other types of dependent variables that are common in survey data sets: **nominal categorical variables**, **ordinal categorical variables,** and **counts** of events or outcomes. Chapter 8 laid the foundation for generalized linear modeling, and this chapter will emphasize specific methods and software applications for three principal methods. Section 9.2 will introduce the "baseline" **multinomial logit regression model** for a survey variable with three or more nominal response categories. The **cumulative logit model** for dependent variables that are measured on an ordinal scale will be covered in Section 9.3. Regression methods for dependent variables that are counts (e.g., number of events, attributes), including **Poisson regression models** and **negative binomial regression models**, are presented in Section 9.4. Stata software will be used to illustrate the applications of these methods, but the reader is encouraged to visit the companion Web site for this text to find each example replicated in the other major software systems that support these advanced modeling procedures.

9.2 Analyzing Survey Data Using Multinomial Logit Regression Models

9.2.1 The Multinomial Logit Regression Model

The multinomial logit regression model is the natural extension of the simple binary logistic regression model to survey responses that have three or more distinct categories. This technique is most appropriate for survey variables with nominal response categories; we present examples of these

(a) NHANES HUQ.040
What kind of place do you go to most often: is it a clinic, doctor's office, emergency room, or some other place?

1. CLINIC OR HEALTH CENTER.....
2. DOCTOR'S OFFICE OR HMO.....
3. HOSPITAL EMERGENCY ROOM.....
4. HOSPITAL OUTPATIENT DEPARTMENT...
5. SOME OTHER PLACE.....
6. REFUSED.....
7. DON'T KNOW.....

(b) NCS-R EM7.1
What about your current employment situation as of today -- are you?

1. EMPLOYED.....
2. SELF-EMPLOYED.....
3. LOOKING FOR WORK; UNEMPLOYED.....
4. TEMPORARILY LAID OFF.....
5. RETIRED.....
6. HOMEMAKER.....
7. STUDENT.....
8. MATERNITY LEAVE.....
9. ILLNESS/SICK LEAVE.....
10. DISABLED.....
11. OTHER (SPECIFY).....

FIGURE 9.1
Survey questions with multinomial response categories.

variables from the 2005–2006 National Health and Nutrition Examination Survey (NHANES) and the National Comorbidity Survey Replication (NCS-R) in Figure 9.1. It is common practice in surveys to use a fairly detailed set of response categories to code the respondent's answer and then recode the multiple categories to a smaller but still scientifically useful set of nominal groupings. For example, the NCS-R public-use data set contains a recoded labor force status variable, WKSTAT3, that combines the 11 questionnaire responses for current work force status into three grouped categories: (1) employed (EMP); (2) unemployed (UN); and (3) not in the labor force (NLF). The multinomial logit regression model is ideally suited for multivariate analysis of dependent variables like WKSTAT3.

Multinomial logit regression may also be applied to survey variables measured on Likert-type scales (e.g., 1 = strongly agree to 5 = strongly disagree) or other ordered categorical response scales (e.g., self-rated health status: 1 = excellent to 5 = poor), but the cumulative logit regression model covered in Section 9.3 may be the more efficient technique for modeling such ordinal dependent variables.

To understand the multinomial logit regression model for a dependent variable y with K nominal categories, assume that category $y = 1$ is chosen as the baseline category. Multinomial logit regression is a method of simultaneously estimating a set of $K - 1$ simple logistic regression models that model

the odds of being in category $y = 2, \ldots, K$ versus the baseline category $y = 1$. Consider the example of the NCS-R recoded variable for labor force status, WKSTAT3, with three nominal categories: 1 = EMP; 2 = UN; 3 = NLF. To fit the multinomial logit regression model to this "trinomial" dependent variable, two generalized logits are needed:

$$\begin{aligned}
\text{logit}(\pi(\text{“UN”} \mid \boldsymbol{x})) &= \text{logit}(\pi_2) = \ln\left(\frac{\pi(y=2 \mid \boldsymbol{x})}{\pi(y=1 \mid \boldsymbol{x})}\right) = B_{2:0} + B_{2:1}x_1 + \cdots + B_{2:p}x_p \\
\text{logit}(\pi(\text{“NLF”} \mid \boldsymbol{x})) &= \text{logit}(\pi_3) = \ln\left(\frac{\pi(y=3 \mid \boldsymbol{x})}{\pi(y=1 \mid \boldsymbol{x})}\right) = B_{3:0} + B_{3:1}x_1 + \cdots + B_{3:p}x_p
\end{aligned} \tag{9.1}$$

A natural question to ask at this point is, "Is it possible to simply estimate the multinomial logit regression model as a series of binary logistic regression models that consider only the response data for two categories at a time?" Strictly speaking, the answer is no. The parameter estimates for what Agresti (2002) labels the "separate-fitting" approach will be similar but not identical to those for simultaneous estimation of the multinomial logits. Standard errors for the former will be greater than those for the simultaneous estimation, and only the latter yields the full variance–covariance matrix needed to test hypotheses concerning the significance or equivalence of parameters across the estimated logits. Fortunately, almost all software systems that support analysis of complex sample survey data now include the capability for the simultaneous estimation of the multinomial logit regression model.

9.2.2 Multinomial Logit Regression Model: Specification Stage

The specification stage of building a multinomial logit model parallels that described in detail in Section 8.3 for specifying a logistic regression model for a binary dependent variable. However, two aspects of the model specification require special emphasis:

1. *Choice of the baseline category.* In the example model formulation for the two distinct logits in Equation 9.1, category $y = 1$ is the selected baseline category. The survey analyst is free to choose which of the K categories he or she prefers to use as the baseline. This choice will not affect the overall fit of the multinomial logit model or overall tests of significance for the parameters associated with predictors included in the model. However, interpretation of the parameter estimates will depend on the selected baseline category, given how the generalized logits are defined. Stata will default to use the lowest numbered category as the baseline category for estimating the logits and corresponding odds ratios. To choose a different category as the baseline for the multinomial logits, the Stata analyst can use the

`baseoutcome(#)` option, where # represents the value of the desired baseline category. In general, when a choice of a baseline category is not clear based on research objectives, we recommend using the most common category (or mode) of the nominal dependent variable.

2. *Parsimony.* Because each of the $K - 1$ logits that form the multinomial logit model will include the identical design vector of covariates, $\boldsymbol{x} = \{1, x_1, ..., x_p\}$, and each estimated logit will have $\boldsymbol{B}_k = \{B_{k:0}, B_{k:1}, ..., B_{k:p}\}$ parameters, the total number of parameter estimates will be $(K - 1) \times (p + 1)$. Consequently, to ensure efficiency in estimation and accuracy of interpretation, the final specification of the model should attempt to minimize the number of predictors that are either not significant or are highly collinear with other significant covariates. Analysts can use design-adjusted multiparameter Wald tests to determine the overall importance of predictors across the $K - 1$ logit functions, and we will consider an example of this in Section 9.2.6.

9.2.3 Multinomial Logit Regression Model: Estimation Stage

The principal difference in estimation for the multinomial logit model versus the simple binary logit model of Chapter 8 is that the pseudo-likelihood function for the data is based on the multinomial distribution (as opposed to the binomial) and the number of parameters and standard errors to be estimated increases from $p + 1$ for the logistic model to $(K - 1) \times (p + 1)$ for the multinomial logit regression model. When survey data are collected from a sample with a complex design, the default in most current software systems is to employ a multinomial version of Binder's (1983) Taylor series linearization (TSL) estimator to derive the estimated variance–covariance matrix of the model parameter estimates. Most software systems also provide a balanced repeated replication (BRR) or jackknife repeated replication (JRR) option to compute replication variance estimates, $\hat{Var}(\hat{\boldsymbol{B}})_{rep}$. Theory Box 9.1 provides a more mathematically oriented summary of the estimation of the multinomial logit regression parameters and their variance–covariance matrix when working with complex sample survey data.

In Stata, the `svy: mlogit` command is used to estimate the multinomial logit regression coefficients and their standard errors. In SAS, analysts employ the standard PROC SURVEYLOGISTIC procedure with the GLOGIT option to perform a multinomial logit regression analysis. Other software options for estimation of multinomial logit regression models are detailed on the book's Web site.

9.2.4 Multinomial Logit Regression Model: Evaluation Stage

Like simple logistic regression and all other forms of generalized linear models, the evaluation stage in building the multinomial logit regression model

THEORY BOX 9.1 ESTIMATION FOR THE MULTINOMIAL LOGIT REGRESSION MODEL

Estimation of the model parameters involves maximizing the following multinomial version of the pseudo-likelihood function:

$$PL_{Mult}(\hat{\boldsymbol{B}} \mid \boldsymbol{X}) = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \hat{\pi}_k(\boldsymbol{x}_i)^{y_i^{(k)}} \right\}^{w_i} \tag{9.2}$$

where

$y_i^{(k)}$ = 1 if $y = k$ for sampled unit i, 0 otherwise;
$\hat{\pi}_k(\boldsymbol{x}_i)$ is the estimated probability that $y_i = k|\boldsymbol{x}_i$; and
w_i is the survey weight for sampled unit i.

The maximization involves application of the Newton-Raphson algorithm to solve the following set of $(K - 1) \times (p + 1)$ estimating equations, assuming a complex sample design with strata indexed by h and clusters within strata indexed by α:

$$S(\boldsymbol{B})_{Mult} = \sum_{h} \sum_{\alpha} \sum_{i} w_{h\alpha i}(y_{h\alpha i}^{(k)} - \pi_k(\boldsymbol{B}))\, \boldsymbol{x}'_{h\alpha i} = \boldsymbol{0} \tag{9.3}$$

where

$y_{h\alpha i}^{(k)} = 1$ if $y = k$ for sampled unit i, 0 otherwise
$\boldsymbol{x}'_{h\alpha i}$ = a column vector of the $p + 1$ design matrix elements for case i $= [1 \quad x_{1,h\alpha i} \quad \cdots \quad x_{p,h\alpha i}]'$;
$\boldsymbol{B} = \{B_{2,0}, ..., B_{2,p}, ..., B_{K,0}, ..., B_{K,p}\}$ is a $(K - 1) \times (p + 1)$ vector of parameters;

$$\pi_k(\boldsymbol{B}) = \frac{\exp(x'_{h\alpha i}\boldsymbol{B}_k)}{1 + \sum_{k=1}^{K} \exp(x'_{h\alpha i}\boldsymbol{B}_k)}$$

with $\boldsymbol{B}_1 = 0$ for $k = 1$ (the baseline).

The variance–covariance matrix of the estimated parameters takes the now familiar sandwich form, based on Binder's (1983) application of Taylor series linearization to estimates derived using pseudo-maximum likelihood estimation:

$$\hat{Var}(\hat{\boldsymbol{B}}) = (\boldsymbol{J}^{-1})\,\mathrm{var}[S(\hat{\boldsymbol{B}})](\boldsymbol{J}^{-1}) \tag{9.4}$$

The matrices $\boldsymbol{J}$ and $\text{var}[S(\hat{\boldsymbol{B}})]$ are derived as illustrated in Theory Box 8.3 for simple logistic regression, with the important change that both are now $(K-1) \times (p+1)$ symmetric matrices, reflecting the full dimension of the parameter vector for the multinomial logit regression model.

begins with Wald tests of hypotheses concerning the model parameters. With $(K-1) \times (p+1)$ parameter estimates, the number of possible hypothesis tests is almost limitless. However, a series of hypothesis tests should be standard practice for evaluating these complex models. Standard *t*-tests for single parameters and Wald tests for multiple parameters should be used to evaluate the significance of the covariate effects in individual logits, that is, H_0: $B_{k:j} = 0$, or across all estimated logits, that is, H_0: $B_{2:j} = \cdots = B_{K:j} = 0$. Example questions that could drive hypothesis tests include the following: Is gender a significant predictor of the odds that a U.S. adult is unemployed versus employed? Is gender a significant predictor in determining the labor force status of U.S. adults regardless of category? Other multiparameter Wald tests can be readily constructed to test custom hypotheses that are relevant for interpretation of a given model. If gender significantly alters the odds that an adult is unemployed or not in the labor force relative to employed, is the gender effect equivalent for unemployment and NLF status? Examples of these general forms of hypothesis tests will be provided in the analytical example in Section 9.2.6.

We note that at the time of this writing, methods for evaluating the goodness of fit of multinomial logit models for complex sample survey data have yet to be developed. Any developments in this area will be reported on the companion Web site for this book.

9.2.5 Multinomial Logit Regression Model: Interpretation Stage

The interpretation of the parameter estimates in a multinomial logit regression model is a natural extension of the interpretation of effects in the simple logistic regression model. Simply exponentiating a parameter estimate results in an adjusted odds ratio, corresponding to the multiplicative impact of a one-unit increase in the predictor variable, x_j, on the odds that the response is equal to k relative to the odds of a response in the baseline category:

$$\begin{aligned} \hat{\psi}_{k:j} &= \exp(\hat{B}_{k:j}) \\ CI(\hat{\psi}_{k:j}) &= \exp[\hat{B}_{k:j} \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_{k:j})] \end{aligned} \tag{9.5}$$

where $\hat{B}_{k:j}$ = the parameter estimate corresponding to predictor j in logit equation k.

If the survey analyst is interested in the impact of a one-unit increase in predictor x_j on the odds of belonging to one of two nonbaseline categories, the following odds ratio estimates the multiplicative effect of a one-unit change in x_j on the odds of being in category k compared with category k':

$$\begin{aligned} \hat{\psi}_{k,k':j} &= \exp(\hat{B}_{k:j} - \hat{B}_{k':j}) \\ CI(\hat{\psi}_{k,k':j}) &= \exp[(\hat{B}_{k:j} - \hat{B}_{k':j}) \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_{k:j} - \hat{B}_{k':j})] \end{aligned} \tag{9.6}$$

where $\hat{B}_{k:j}, \hat{B}_{k':j}$ = the parameter estimates corresponding to predictor j in logit equations k and k'.

More generally, the technique outlined in Section 8.6 can be applied to evaluate the relative odds of being in category k versus the baseline category for any selected multivariate covariate patterns $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ or to compare the odds of being in category k versus k' for any common covariate pattern, $\boldsymbol{x}$. We illustrate these techniques with examples in the following section.

9.2.6 Example: Fitting a Multinomial Logit Regression Model to Complex Sample Survey Data

In this example, we fit a multinomial logit regression model to the NCS-R response variable WKSTAT3, which takes on three values: 1 = EMP; 2 = UN; and 3 = NLF. Figure 9.2 is a simple bar graph that illustrates a weighted estimate of the population distribution for the WKSTAT3 variable.

We consider as predictor variables AG4CAT, SEX, ALD (a lifetime diagnosis of alcohol dependence), MDE (a lifetime diagnosis of a major depressive episode), ED4CAT (1 = 0–11 years, 2 = 12 years, 3 = 13–15 years and 4 = 16+ years), and MAR3CAT (1 = married, 2 = previously married, 3 = never married). This example does not formally follow through with all of the recommended model-fitting steps proposed by Hosmer and Lemeshow (2000); however, Table 9.1 does show the results from a preliminary analysis of the bivariate associations between WKSTAT3 and each of the six categorical predictors that are considered in the initial model specification.

The results for the Wald F-tests (see Chapter 6 for more details on these tests) suggest that the four demographic predictors (AG4CAT, SEX, ED4CAT, and MAR3CAT) all have a significant bivariate association with work force status and that the two diagnosis variables (ALD, MDE) have a somewhat weaker association.

To determine if these marginal associations remain significant when controlling for the other predictors, the following Stata command sequence can

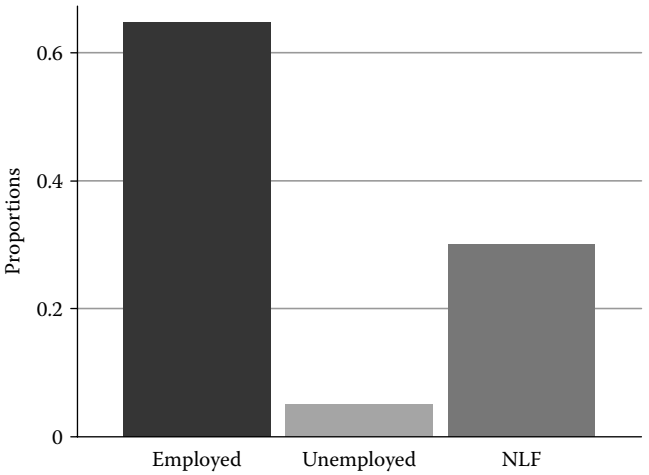


FIGURE 9.2
Weighted distribution of WKSTAT3. (Modified from NCS-R data.)

TABLE 9.1

Initial Bivariate Design-Based Tests Assessing Potential Predictors of WKSTAT3 for the NCS-R Adult Sample

Categorical Predictor	*F*-Test Statistic	$P(\mathcal{F} > F)$
AG4CAT	$F_{4.96,208.51} = 113.49$	< 0.001
SEX	$F_{1.87,78.75} = 27.33$	< 0.001
ALD	$F_{1.72,72.44} = 3.12$	0.057
MDE	$F_{1.73,72.86} = 4.67$	0.016
ED4CAT	$F_{5.15,216.12} = 27.64$	< 0.001
MAR3CAT	$F_{3.20,134.34} = 23.12$	< 0.001

be used to fit the multinomial logit regression model taking the complex design of the NCS-R sample into account (assuming that the variables identifying the NCS-R sampling error codes and the appropriate sampling weight have already been declared using `svyset`):

```
xi: svy: mlogit wkstat3 i.sex ald mde i.ed4cat i.ag4cat ///
i.mar3cat
svy: mlogit, rrr
```

Note that the `svy: mlogit` command is used to fit the model to the WKSTAT3 outcome. The nominal categorical dependent variable WKSTAT3 is listed first, followed by a list of the predictor variables in the model (which could be in any order), with the `xi:` modifier used to generate indicator variables for each predictor that is declared as categorical by the "`i.`"

prefix. The `rrr` option is used in the repetition of the `svy: mlogit` command to request output of the estimated odds ratios (which Stata interprets as *relative risk ratios*) and 95% confidence intervals (CIs). The default baseline category for the multinomial logit regression model in Stata will be the lowest-valued category, which in this example would be 1 = employed. Alternative baseline categories can be identified by using the `baseoutcome(#)` option in the `svy: mlogit` command, where # represents the value of the reference outcome category. For example, consider the following command specification:

```
xi: svy: mlogit wkstat3 i.sex ald mde i.ed4cat i.ag4cat ///
i.mar3cat, baseoutcome(3)
```

This command would fit the multinomial logit model to WKSTAT3 with not in labor force as the baseline category.

Table 9.2 provides the detailed Stata output for the estimated model, including coefficient estimates, linearized standard errors, *t*-statistics, and *p*-values for each of the (two) generalized logits. Table 9.3 presents the estimated coefficients transformed into odds ratios and 95% confidence intervals for the odds ratios. We recommend the more concise display of the estimated odds ratios and confidence intervals in Table 9.3 for reporting results of a multinomial logit regression model in publications (e.g., Kavoussi et al., 2009).

To evaluate the fitted model, we perform multiparameter Wald tests of the overall significance of each of the predictors: AG4CAT, SEX, ALD, MDE, MAR3CAT, and ED4CAT:

```
test _Iag4cat_2 _Iag4cat_3 _Iag4cat_4
test _Isex
test ald
test mde
test _Imar3cat_2 _Imar3cat_3
test _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
```

The Wald tests specified in these test statements are testing the null hypothesis that *all* parameters associated with each individual predictor (e.g., age, education level) in the two logits are not significantly different from zero. Table 9.4 provides the design-adjusted Wald *F*-test statistics and the associated *p*-values for these overall tests of individual effects.

Inspection of these overall test results shows that, as might be expected, the covariates AG4CAT, SEX, ED4CAT, and MAR3CAT are all strongly significant determinants of the relative odds that an adult is employed, unemployed, or not in the labor force. Focusing on the effects of alcohol dependence and major depression, we note an interesting pattern. Controlling for the demographic variables, MDE ($p = 0.330$) does not appear to have a significant effect, while ALD ($p = 0.011$) appears to have a significant relationship with work

TABLE 9.2

Estimated Multinomial Logit Regression Model for WKSTAT3

Logit 2: Unemployed vs. Employed					
Predictor*	**Category**	$\hat{B}_{2:j}$	$se(\hat{B}_{2:j})$	t	$P(t_{42} > \lvert t \rvert)$
INTERCEPT		–0.643	0.296	–2.17	0.035
AG4CAT	30–44	–0.852	0.294	–2.89	0.006
	45–59	–0.838	0.258	–3.25	0.002
	60+	1.828	0.295	6.20	< 0.001
SEX	Male	–1.393	0.198	–7.05	< 0.001
ALD	Yes	–0.164	0.357	–0.46	0.649
MDE	Yes	–0.140	0.157	–0.89	0.379
ED4CAT	12	–0.847	0.235	–3.60	0.001
	13–15	–1.365	0.258	–5.30	< 0.001
	16+	–1.731	0.310	–5.57	< 0.001
MAR3CAT	Previously	–0.589	0.225	–2.62	0.012
	Never	–2.785	0.380	–7.32	< 0.001

Logit 3: Not in Labor Force vs. Employed					
Predictor[a]	**Category**	$\hat{B}_{3:j}$	$se(\hat{B}_{3:j})$	t	$P(t_{42} > \lvert t \rvert)$
INTERCEPT		–3.790	0.173	–2.19	0.034
AG4CAT	30–44	–0.316	0.129	–2.46	0.018
	45–59	0.065	0.171	0.38	0.706
	60+	2.381	0.173	13.78	< 0.001
SEX	Male	–0.640	0.110	–5.82	< 0.001
ALD	Yes	0.333	0.130	2.56	0.014
MDE	Yes	0.098	0.088	1.12	0.269
ED4CAT	12	–0.651	0.141	–4.62	< 0.001
	13–15	–0.917	0.146	–6.26	< 0.001
	16+	–1.229	0.160	–7.70	< 0.001
MAR3CAT	Previously	–0.052	0.105	–0.50	0.621
	Never	0.553	0.132	4.18	< 0.001

Source: Analysis based on the NCS-R data.

Notes: $n = 5{,}692$. Adjusted Wald test for all parameters: $F(22,21) = 73.91$. $p < 0.001$.

[a] Reference categories for categorical predictors are 18–29 (AG4CAT); Female (SEX); No (ALD); No (MDE); <12 yrs (ED4CAT); Married (MAR3CAT).

TABLE 9.3

Estimates of Adjusted Odds Ratios for the Work Force Status Outcome (WKSTAT3)

		Unemployed: Employed		NLF: Employed	
Predictor[a]	**Category**	$\hat{\psi}_{2:j}$	**95% CI for** $\hat{\psi}_{2:j}$	$\hat{\psi}_{3:j}$	**95% CI for** $\hat{\psi}_{3:j}$
AG4CAT	30–44	0.43	(0.24, 0.77)	0.73	(0.56, 0.94)
	45–59	0.43	(0.26, 0.73)	1.07	(0.76, 1.51)
	60+	6.22	(3.43, 11.28)	10.81	(7.62, 15.34)
SEX	Male	0.25	(0.17, 0.37)	0.53	(0.42, 0.66)
ALD	Yes	0.85	(0.41, 1.74)	1.40	(1.07, 1.82)
MDE	Yes	0.87	(0.63, 1.19)	1.10	(0.92, 1.32)
ED4CAT	12	0.43	(0.27, 0.69)	0.52	(0.39, 0.69)
	13–15	0.26	(0.15, 0.43)	0.40	(0.30, 0.54)
	16+	0.18	(0.10, 0.33)	0.29	(0.21, 0.40)
MAR3CAT	Previously	0.55	(0.35, 0.87)	0.95	(0.77, 1.17)
	Never	0.06	(0.03, 0.13)	1.74	(1.33, 2.70)

Source: Analysis based on the NCS-R data.

[a] Reference categories for categorical predictors are 18–29 (AG4CAT); Female (SEX); No (ALD); No (MDE); <12 yrs (ED4CAT); Married (MAR3CAT).

TABLE 9.4

Overall Wald Tests for the Predictors in the Multinomial Model for WKSTAT3

Categorical Predictor	***F*-test Statistic**	$P(\mathcal{F} > F)$
AG4CAT	$F_{(6,37)} = 83.59$	<0.001
SEX	$F_{(2,41)} = 35.75$	<0.001
ALD	$F_{(2,41)} = 5.05$	0.011
MDE	$F_{(2,41)} = 1.14$	0.330
ED4CAT	$F_{(6,37)} = 13.68$	<0.001
MAR3CAT	$F_{(4,39)} = 24.81$	<0.001

Source: Analysis based on the NCS-R data.

force status. Referring back to the *t*-tests of the individual logit parameters in Table 9.2, the pattern for ALD becomes clearer. It appears that ALD significantly affects the odds of being NLF relative to employed but is not significant in explaining unemployment status (relative to being employed).

Consider another example of a design-adjusted postestimation Wald test: a test of the null hypothesis that the three education parameters in logit(2) are equivalent to those in logit(3) (i.e., the effect of education in the two logits is equivalent). We would use the following command in Stata to test this null hypothesis:

```
test [NLF=unemployed]: _Ied4cat_2 _Ied4cat_3 _Ied4cat_4
```

The *F*-test statistic and associated *p*-value for this Wald test are $F_{3,40} = 1.25$, $P(F > F) = 0.30$. The design-adjusted Wald test statistic suggests a failure to reject the null hypothesis that these three pairs of coefficients are equal to each other. Note that we use value labels for the outcome categories in square brackets after the test command; if value labels were not entered in the data set, one would just use the numeric values for the two outcome categories.

As described in Section 9.2.5, the magnitude and direction of significant effects can be interpreted as odds ratios: odds for unemployed versus employed for logit(2) and odds for NLF versus employed for logit(3). Consider first the results for logit(2) comparing unemployed with employed work force status. From Table 9.3, the odds of being unemployed versus employed are significantly lower for middle-aged persons ($\hat{\psi}_{2:30-44} = \hat{\psi}_{2:45-59} = 0.43$) relative to younger persons and men ($\hat{\psi}_{2:male} = 0.25$) relative to women. The odds of unemployment (vs. being employed) decrease with level of education ($\hat{\psi}_{2:12yrs} = 0.43$; $\hat{\psi}_{2:13-15yrs} = 0.26$; $\hat{\psi}_{2:16+yrs} = 0.18$), where less than high school education is the reference category, and are lower for previously married ($\hat{\psi}_{2:prev} = 0.55$) and never married persons ($\hat{\psi}_{2:male} = 0.06$) compared with married persons. All of these interpretations are population estimates of these relationships when holding the other predictor variables in the model at fixed values.

Now, consider the results for the NLF outcome. Relative to young persons aged 18–29 the odds of being NLF decrease significantly for 30–44 year olds ($\hat{\psi}_{3:30-44} = 0.73$) and then rise to a significant increase in NLF odds in the age 60+ retirement years ($\hat{\psi}_{3:60+} = 10.81$). As observed for unemployment, the odds of being NLF versus employed are significantly lower for men ($\hat{\psi}_{3:male} = 0.53$) and decrease with level of education ($\hat{\psi}_{3:12yrs} = 0.52$; $\hat{\psi}_{3:13-15yrs} = 0.40$; $\hat{\psi}_{3:16+yrs} = 0.29$). All else being equal, persons who were never married had higher odds of being NLF ($\hat{\psi}_{3:never} = 1.74$) than their married counterparts. Finally, alcohol dependency is associated with significantly increased odds ($\hat{\psi}_{3:ALD} = 1.40$) of being out of the labor force.

For an example of reporting the results of a multinomial logit regression analysis of complex sample survey data in a scientific publication, we refer readers to Kavoussi et al. (2009).

9.3 Logistic Regression Models for Ordinal Survey Data

Ordinal response questions like the two examples in Figure 9.3 are common in survey practice. The 2006 Health and Retirement Study (HRS) health rating question shown in Figure 9.3 part (a) exemplifies a question type in which the respondent is asked to assign discrete rankings to attributes—their own or those of others. A more subtle form of ordinal response arises in questions like the NHANES activity question shown in part (b) of Figure 9.3. In that question, the ordinality of the response is not explicitly incorporated in a rating-type response scale but the response categories do implicitly capture ever increasing levels of physical activity.

As discussed in Chapter 5, survey analysts have historically treated descriptive analysis of ordinal response data in a number of different ways. The same is true in regression modeling for ordinal categorical data. At one extreme, the ordinal responses are treated as continuous random variables. Many survey analysts do not hesitate to fit a standard linear regression model to ordinal response data. Others completely ignore the natural ordering of the response categories and analyze the ordinal data as though it were nominal in nature (i.e., apply multinomial logit regression as in Section 9.2). In practical terms, neither approach is necessarily wrong. DeMaris (2004) identifies conditions under which he feels that a linear regression treatment leads to robust analysis—enough levels (five or more), large n, and a response distribution that is not highly skewed across the ordinal range. Analysts who ignore the

(a) HRS 2006 KC001
Would you say your health is excellent, very good, good, fair, or poor?
1. EXCELLENT
2. VERY GOOD
3. GOOD
4. FAIR
5. POOR
8. DK (Don't Know); NA (Not Ascertained)
9. RF (Refused)

(b) NHANES 2005-06 PAQ.180
Please tell me which of these four sentences best describes your usual daily activities:
1. You sit during the day and do not walk about very much
2. You stand or walk about quite a lot during the day, but do not have to carry or lift things very often
3. You lift or carry light loads, or have to climb stairs or hills often; or
4. You do heavy work or carry heavy loads.
7. Refused
9. Don't know

FIGURE 9.3
Ordinal response questions from the 2006 HRS and the 2005–2006 NHANES.

ordinality of such responses and apply the general multinomial logit regression technique of Section 9.2 are also certainly not wrong in their approach. However, such models require the estimation of many parameters [recall that one estimates $(K - 1) \times (p + 1)$ parameters when fitting a multinomial logit model] and therefore may not be the most efficient modeling option.

9.3.1 Cumulative Logit Regression Model

A special class of logistic regression models has been developed for regression analysis of ordinal survey variables. These models differ from multinomial logit regression models for nominal categorical response variables in that they acknowledge the ordering of the response categories when estimating the relationships of the predictor variables with the probabilities of having certain responses. Ordinal models involve fewer parameters than multinomial logit regression models and are more parsimonious as a result. Standard statistical texts describe several approaches to specifying (parameterizing) an ordinal logistic regression model (Agresti, 2002; Hosmer and Lemeshow, 2000). Here our focus will be only on the most common form, the cumulative logit model.

A **cumulative logit** is defined for the probability of having an ordinal response less than or equal to k, relative to the probability of having a response greater than k:

$$\begin{aligned} \text{logit}[P(y \le k) \mid \boldsymbol{x}] &= \ln\left[\frac{P(y \le k) \mid \boldsymbol{x}}{P(y > k) \mid \boldsymbol{x})}\right] \\ &= \ln\left[\frac{P(y = 1 \mid \boldsymbol{x}) + \ldots + P(y = k \mid \boldsymbol{x})}{P(y = k+1 \mid \boldsymbol{x}) + \ldots + P(y = K \mid \boldsymbol{x})}\right] \\ &= B_{0(k)} - (B_1 x_1 + B_2 x_2 + \ldots + B_p x_p) \end{aligned} \tag{9.7}$$

For an ordinal variable with K categories, $K - 1$ cumulative logit functions are defined. Each cumulative logit function includes a unique intercept or "cutpoint," $B_{0(k)}$, but all share a common set of regression parameters for the p predictors, $\boldsymbol{B} = (B_1, \ldots, B_p)$. Consequently, a cumulative logit model for an ordinal response variable with K categories and $j = 1, \ldots, p$ predictors requires the estimation of $(K - 1) + p$ parameters—far fewer than the $(K - 1) \times (p + 1)$ parameters for a multinomial logit model.

By using logits defined by cumulative probability, the cumulative logit model captures trends across the adjacent categories of an ordinal response variable. By using a single set of regression parameters for the predictors, the model provides true parsimony in estimating the relationship between predictors and the profile of responses over the ordinal response categories. The concept of a "cumulative logit" is certainly more complex than that of

a baseline category logit that is employed in the parameterization of simple logistic regression or even multinomial logit regression models. In actuality, though, it is simply an alternative way of parameterizing the model for estimating the probability that a response will fall in ordinal category $y = 1, \ldots, K$.

Consider the cumulative logit model in Equation 9.7. If the cumulative logit is estimated for a covariate pattern $\boldsymbol{x} = \{x_1, x_2, \ldots, x_p\}$, then the transform

$$\hat{\varphi}(y \le k \mid \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\hat{\boldsymbol{B}})}{1+\exp(\boldsymbol{x}\hat{\boldsymbol{B}})} = \frac{\exp[\hat{B}_{0(k)} - (\hat{B}_1 x_1 + \hat{B}_1 x_2 + \ldots + \hat{B}_p x_p)]}{1+\exp[\hat{B}_{0(k)} - (\hat{B}_1 x_1 + \hat{B}_1 x_2 + \ldots + \hat{B}_p x_p)]} \tag{9.8}$$

estimates the cumulative probability (denoted here by $\hat{\varphi}(y \le k \mid \boldsymbol{x})$) that the response, y, is less than or equal to ordinal category k. To "recover" estimates of the individual response category probabilities, $\hat{\pi}_k(\boldsymbol{x})$, from the estimated cumulative probabilities simply requires taking the difference in the estimated cumulative probability through response categories k and $k - 1$:

$$\hat{\pi}_k(\boldsymbol{x}) = \hat{\varphi}(y \le k \mid \boldsymbol{x}) - \hat{\varphi}(y \le k-1 \mid \boldsymbol{x}) \tag{9.9}$$

where $\hat{\varphi}(y \le 0 \mid \boldsymbol{x}) = 0$.

9.3.2 Cumulative Logit Regression Model: Specification Stage

The cumulative logit model is a specialized model that is applicable to true ordinal response variables such as a health satisfaction ratings and satisfaction scores, where the categorized response (excellent, very good, good, fair, poor) can be viewed as the observed representation of an underlying continuous measure of a latent attribute, belief, or attitude. The general steps in specifying and building a final cumulative logit regression model (e.g., variable selection, evaluation of potential nonlinearity for continuous predictors, tests of interaction of main effects) are identical to those applied in other regression models.

9.3.3 Cumulative Logit Regression Model: Estimation Stage

Procedures for pseudo-maximum likelihood estimation of the parameters of the cumulative logit regression model are identical to those described in Section 9.2.3 for the multinomial logit regression model. The final parameter estimates are obtained by application of the Newton–Raphson algorithm to maximize the weighted multinomial likelihood given by Equation 9.2. At each iteration, the values of $\hat{\pi}_k(\boldsymbol{x}_i)$ are obtained through the probability calculation sequence similar to that described in Equations 9.8 and 9.9. Like multinomial logit regression, the default variance estimator in most current

software systems is the multinomial version of the Taylor series linearization estimator (Equation 9.4) with most major software packages also providing analysts the options to use a BRR or JRR option to compute replication estimates, $\hat{Var}(\hat{\boldsymbol{B}})_{rep}$.

Analysts should be aware that different software packages use different parameterizations of the cumulative logit model. For example, the Stata `svy: ologit` procedure fits these models using the parameterization described in Equation 9.7, in which the covariate adjustment,

$$\Delta_i = \sum_{j=1}^{p} B_j x_{ij}$$

is *subtracted* from the cutpoint $B_{0(k)}$ for the k-th cumulative logit. Positive values for estimates of the regression parameters in the model will thus signify increases in the probability of higher-valued responses categories for higher values for the corresponding predictor variables. The SURVEYLOGISTIC procedure in the SAS software uses the following parameterization of the model

$$\text{logit}[P(Y \le k) \mid \boldsymbol{x}] = B_{0(k)} + (B_1 x_1 + B_2 x_2 + \ldots + B_p x_p) \tag{9.10}$$

with the covariate adjustment *added* to the cutpoint for cumulative logit k—reversing the sign on the regression parameters compared to the Stata default. Use of the DESCENDING option in the SAS SURVEYLOGISTIC procedure will yield results that are identical to the Stata default. Analysts need to carefully check how these models are being parameterized by the software procedures that they are using prior to interpreting the results from these types of analyses.

9.3.4 Cumulative Logit Regression Model: Evaluation Stage

Procedures (Student t tests, Wald tests) described for testing the significance of individual model parameters or multiparameter predictors in the simple logistic model (Section 8.5.1) apply directly to evaluation of fitted cumulative logit models. Unfortunately, goodness of fit measures and diagnostic tools like those available for multivariate logistic regression (Section 8.5.2) have not been developed for the more complex logistic regression models for ordinal response data and therefore are not yet available in statistical software procedures for complex sample survey data analysis.

Analysts are able to perform one type of diagnostic for the cumulative logit model. The literature occasionally labels the cumulative logit model as the **proportional odds model**. This alternate label derives from a property of the model which results from the fact that each of the $K - 1$ cumulative logits is assumed to share a common set of regression coefficients, $\boldsymbol{B}$, for the

model predictors. When analyzing complex sample survey data, tests of this assumption require the estimation of an alternative **generalized cumulative logit model** (Peterson and Harrell, 1990), where the regression parameters identified in Equation 9.7 are allowed to vary depending on the choice of k (i.e., $B_{k:1}$ instead of B_1). The "proportional odds" or "equal slopes" assumption can be formally tested by fitting this alternative model (also using design-based methods) and performing a design-adjusted Wald test of the null hypothesis that all pair-wise contrasts of the regression parameters for each predictor across the $K-1$ logit functions in the generalized model are equal to zero (e.g., $B_{1:1} - B_{2:1} = 0; B_{1:1} - B_{3:1} = 0$). This design-based test is currently implemented in the user-written Stata command `gologit2` (Stata users can submit the command `findit gologit2` for more details) and the CSORDINAL procedure of the SPSS complex samples module, and additional implementations will be noted on the companion Web site for this book.

In cases where this formal test rejects the null hypothesis of equal slopes, the survey analyst should first extend their model-building investigation to evaluate if the current predictor set must be modified (adding new predictors, interaction terms, or nonlinear effects) for the equal slopes assumption to hold. If this extended investigation fails to resolve the problem of apparent inequality in the slope parameters of the cumulative logit model, the analyst can ignore the ordinality of the response and revert back to a generalized multinomial logit model for categorical response variable (Section 9.2) or consider more parsimonious generalized forms of the cumulative logit model (which are currently implemented in the Stata command `gologit2`). Often, despite the formal rejection of the equal slopes test, the interpretations of the results from the two models—cumulative and generalized logit—may be consistent and the analyst may still choose the cumulative logit model for its simplicity of form and interpretation.

9.3.5 Cumulative Logit Regression Model: Interpretation Stage

Like most regression models, the interpretation of results from a cumulative logit regression model can occur at two different levels. At the evaluation stage (as previously discussed), t-tests of single-parameter predictors or Wald tests of multiparameter predictors will identify those predictors that have a significant relationship with the ordinal response variable. Examination of the estimated coefficients for the cumulative logits can inform the analyst about the directional nature of the relationship of response and predictors. For example, under the Stata default parameterization, positive values for estimates of the regression parameters for a continuous predictor correspond to increased probability of higher-valued response categories as the predictor value itself increases. Likewise, positive coefficients for parameters representing a level of a categorical predictor (e.g., males relative to females) suggest that, relative to the reference category, the distribution of ordinal responses for the predictor category is shifted toward the higher values of the

response distribution. (In SAS, the interpretation would be reversed unless the DESCENDING option is used in PROC SURVEYLOGISTIC.) While estimates of the cutpoint intercept terms of the cumulative logit model are critical to model fit, they typically are of little further interest in the analysis and interpretation of the results. For interpretation and summarization of the model results, survey analysts may choose to go beyond simply establishing the significance and directionality of predictor effects. Following Agresti (2002), a more quantitative presentation of the direction and magnitude of significant effects of predictors in cumulative logit regression models can be based on one of two related sets of statistics.

The first is the set of cumulative odds ratios that can be estimated directly from the fitted model. If the k^{th} cumulative logit is estimated for two covariate patterns, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the following exponential function estimates a **cumulative odds ratio**:

$$\hat{\psi}_{y\le k} = \exp[\hat{\boldsymbol{B}}'(\boldsymbol{x}_1 - \boldsymbol{x}_2)] = \exp[\hat{B}_1(x_{11} - x_{12}) + \cdots + \hat{B}_p(x_{p1} - x_{p2})] \qquad (9.11)$$

The interpretation of the cumulative odds ratio statistic is slightly different from the standard odds ratio. For the given covariate patterns $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the cumulative odds that the ordinal response, y, is less than or equal to category k are $\hat{\psi}_{y\le k}$ times greater for $\boldsymbol{x}_1$ than the odds for $\boldsymbol{x}_2$. Cumulative odds ratios and confidence intervals are readily generated as output from complex sample software programs for cumulative logit modeling. In scientific reports and publications they can be summarized in standard tables (Table 9.6) or using graphical displays of the type illustrated in Figure 8.3.

The second (and related) set of statistics that can be used to quantify the effect of individual predictors are estimates of cumulative probabilities of the form in (9.8). For example, holding all other predictors constant, expression (9.8) could be used to estimate men's and women's cumulative probability of response in category $y = 1, \ldots, K$. Estimated values and standard errors (or CIs) for $\hat{\varphi}(y \le k \mid \boldsymbol{x})$ for men and women could then be compared side by side in a tabular format or in a graph. To compare distributions of the estimated cumulative probabilities across the range of a continuous variable (e.g., age or blood pressure) Agresti (2002) recommends evaluating the cumulative logit for each category at the quartiles of that predictor's distribution (Q_{25}, Q_{50}, Q_{75}).

Readers are encouraged to reference Agresti (2002), Hosmer and Lemeshow (2000), Allison (1999), and Long and Freese (2006) for additional tips on evaluating and interpreting results of cumulative logit regression models.

9.3.6 Example: Fitting a Cumulative Logit Regression Model to Complex Sample Survey Data

This section considers an example of fitting a cumulative logit model to the KC001 variable in the 2006 HRS data set. As seen in Figure 9.3, KC001 captures

individual respondents' self-rating of their general health using a five-point ordinal scale: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor.

The initial data preparation steps require the creation of an analytic variable, SELFRHEALTH, in which numeric missing values on the KC001 variable in the HRS data set (coded 8 and 9) are set to system missing values in Stata:

```
gen selfrhealth = kc001
replace selfrhealth = . if kc001==8 | kc001==9
```

Before beginning the model-building process, analysts should examine simple frequency distributions for ordinal response variables. If the majority of responses on a discrete ordinal outcome are grouped in single categories or highly skewed to the highest or lowest possible values, the cumulative logit model may not be the best choice and simple or multinomial logit regression models would be more appropriate for a recoded version of the ordinal response variable. Figure 9.4 presents a bar chart of the weighted response frequencies for the SELFRHEALTH variable.

We see from this figure that the majority of the HRS-eligible U.S. adult population is estimated to respond with a health rating of fair to very good but significant proportions also report excellent (12.7%) or poor (7.7%) health. This distribution of responses across categories and the implicit continuum of health status that underlies this response scale suggest that this categorical variable would be a reasonable choice for a dependent variable in an ordinal regression model.

The objective in this simple example is to model the health status for U.S. adults age 50+ (recall that the HRS target population is individuals age 50+)

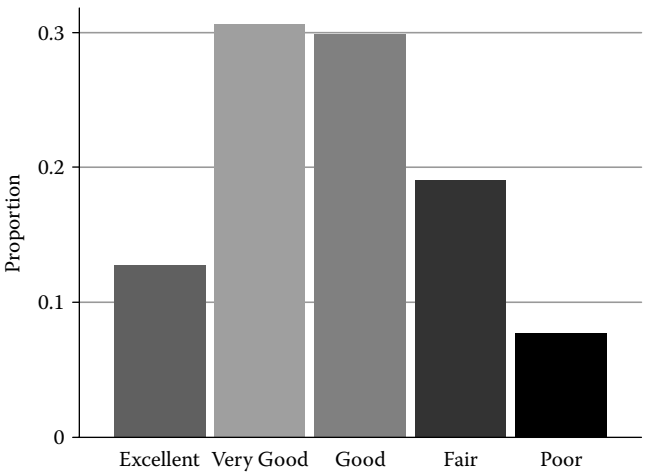


FIGURE 9.4
Bar chart (weighted) of the 2006 HRS variable SELFRHEALTH.

TABLE 9.5

Estimated Cumulative Logit Regression Model for SELFRHEALTH

Predictor	Category	$\hat{B}$	$se(\hat{B})$	t	$P(t_{56} > t)$
INTERCEPT	Cut 1	–0.071	0.153	–0.46	0.645
	Cut 2	1.614	0.153	10.56	<0.001
	Cut 3	2.916	0.158	18.37	<0.001
	Cut 4	4.405	0.165	26.65	<0.001
KAGE		0.029	0.002	13.23	<0.001
GENDER[a]	Male	–0.071	0.032	–2.19	0.033

Source: Analysis based on the HRS data.

Notes: n = 18,442. Adjusted Wald test for all parameters: $F(2,55)$ = 90.21, $p < 0.001$.

[a] Reference category: Female.

as a function of age and gender. We remind readers that the Hosmer and Lemeshow (2000) steps discussed in Chapter 8 are always recommended when building a regression model, but for brevity we consider fitting only a single simple model using these two demographic variables and do not illustrate the full sequence of model-building steps.

The `svy: ologit` command is used to fit the cumulative logit regression model in Stata:

```
xi: svy: ologit selfrhealth kage i.gender
svy: ologit, or
```

Similar to other regression commands in Stata, the response variable (SELFRHEALTH) is listed first, followed by a list of the predictor variables (no interactions are included in this model). The model is reestimated with the `or` option to request that Stata provide output in the form of the estimated cumulative odds ratios and 95% confidence intervals.

Table 9.5 summarizes the estimated "cutpoint" parameters (or $B_{0(k)}$ values from Equation 9.7) for the logistic distribution in this model along with the estimates of the regression parameters for the age and gender predictor variables. As already noted, these cutpoints are not of specific analytic interest but are required to compute predicted probabilities of the ordinal outcomes given input values on the predictor variables. Table 9.6 presents the corresponding estimates of the cumulative odds ratios and design-based 95% confidence intervals for the odds ratios (with standard errors computed using Taylor series linearization).

In Table 9.5, first note the estimates of the four intercept parameters. As previously described, these parameter estimates are rarely of real analytic interest but are used by the analysis software to calculate predicted probabilities of being in one of the five ordered response categories for SELFRHEALTH.

TABLE 9.6

Estimated Cumulative Odds Ratios in the Cumulative Logit Regression Model for SELFRHEALTH

		Cumulative Odds Ratio	
Predictor	**Category**	$\hat{\psi}_{y\le k:j}$	**95% CI for** $\psi_{y\le k:j}$
KAGE	Continuous	1.03	(1.025, 1.034)
GENDER[a]	Male	0.93	(0.873, 0.994)

Source: Analysis based on the HRS data.
[a] Reference category: Female.

Consider next the regression parameter estimates (Table 9.5) and estimated cumulative odds ratios (Table 9.6) for the age and gender predictors. Since each of these demographic factors is represented by a single model parameter, the *t*-test reported in Table 9.5 is equivalent to the Wald test of the significance of each predictor. The *t*-test results suggest that both age and gender are significant predictors in the cumulative logit model for SELFRHEALTH. Not surprisingly, the model coefficient for the KAGE predictor is positive, suggesting that increasing age is positively related to decreasing health status (*higher valued* categories on this rating scale). The estimated cumulative odds ratio is $\hat{\psi}_{y\le k:age} = 1.03$, suggesting that (given Stata's parameterization of this model) the cumulative odds of being in a lower health category ($k + 1, \ldots, K$) relative to a better health category $1, \ldots, k$ (e.g., the odds of being either very good, good, fair, or poor relative to the odds of being excellent) increase by approximately 3% for each additional year of age over age 50 (holding gender fixed). From Table 9.6, the estimated cumulative odds ratio comparing men with women is $\hat{\psi}_{y\le k:gender} = 0.93$. The cumulative odds that a man will be in a lower self-reported health status (poorer health) are 93% of those same odds for a woman of the same age.

To test the proportional odds or "equal slopes" assumption that underlies the cumulative logit model, the model was fitted using the CSORDINAL command in the SPSS Complex Samples module. SPSS prints results for a "Test of Parallel Lines" in the output, and the resulting design-based Wald statistic is $F(6,51) = 3.890$, $p = 0.003$, suggesting that the data do not in fact support the assumption of common regression coefficients for the four cumulative logits. (All other estimates produced by SPSS are exactly identical to those produced by Stata.) A similar test can be performed when fitting this model using the `gologit2` command with the `svy` option in Stata (see `help gologit2` after installation for examples).

Rejection of the null hypothesis of equal slopes leaves us with the difficult question of whether to ignore the apparent violation of this key model assumption in exchange for the simplicity of interpretation offered or to opt instead for a generalized form of the model (Peterson and Harrell, 1990) or other modeling alternative that relaxes the equal slopes assumption but is

far more complex and difficult to interpret. Agresti (2000) points out that for large sample sizes this test is extremely powerful against the null hypothesis that the proportional odds assumption is met. Agresti advises that even when the test rejects the assumption, the parsimony of the cumulative logit model parameterization may still make it a practical choice relative to the much more complex alternatives. We recommend comparing results from alternative models and using more flexible models if there are substantial differences in the resulting inferences.

9.4 Regression Models for Count Outcomes

9.4.1 Survey Count Variables and Regression Modeling Alternatives

Regression models for dependent variables that are discrete counts of events or outcomes are also important in the analysis of survey data. Figure 9.5 provides an example of a survey question from the 2006 HRS that produces a count of falls that respondents age 65 and older experienced in the past two-year period. The example question sequence illustrates a typical convention in survey measurement of counts. Respondents are initially queried as to whether any events occurred during the reference period. If the respondent answers yes, a follow-up question captures a count of events (>0), which is recorded by the interviewer in the boxes. Respondents who

HRS 2006

[If Age 65 or older]

KC079 Have you fallen down in the past 2 years:

1	Yes
2	No
8	Don't Know
9	Refused
Blank	Inapplicable, respondent is <65 years

KC080 How many times have you fallen in the last two years?

☐☐☐	NUMBER OF TIMES
☐	Don't Know
☐	Refused

FIGURE 9.5
Example of a survey question sequence producing a count variable. (From: 2006 HRS.)

answer no to the initial screening question are presumed to have a zero count of events.

It is clear that the question sequence encodes two pieces of information: a count of events (y) and the "exposure" time, t, during which the reported events occurred. The length of the observation period can be fixed (e.g., the two years) or may vary from one sample respondent to another (e.g., since you last saw a doctor). Regression models for these types of counts aim to model the relationship between the count response y and predictor variables of theoretical interest x, or equivalently, the relationship of x with the rate $\lambda = y/t$ at which the event occurred. The regression models described in this section model rates, or event counts per unit of time. Chapter 10 will introduce regression techniques for survival analysis or event history analysis that model the time to an event, where time itself is the dependent variable.

Figure 9.6 illustrates the distribution of counts that results from the 2006 HRS question on number of falls in the past two years. This figure illustrates two common properties of distributions of count-type survey variables. Often, the event of interest occurs rarely in the population, and the distribution of counts is dominated by a very high proportion of zero values. The distribution is also highly skewed with declining frequencies for "1," "2," "3," …, "10+" falls. The survey analyst can choose from a number of alternative approaches to regression modeling of these types of count variables.

Linear regression techniques are frequently used by survey analysts to model count data. Theory Box 9.2 discusses the pros and cons of trying to apply linear regression in the analysis of survey counts. Today, the preferred alternatives for regression modeling of count data are generalized linear models based in full or in part on the Poisson and negative binomial distributions.

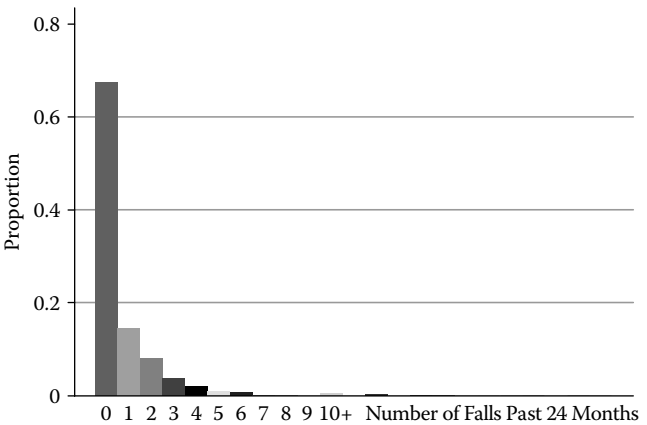


FIGURE 9.6
Histogram (weighted) of 2006 HRS counts of falls in the past 24 months.

THEORY BOX 9.2 LINEAR REGRESSION FOR COUNT VARIABLES

One alternative for modeling a count response might be to fit a standard linear regression model to a log-transformed version of the rate, denoted by $\lambda = y/t$. However, two problems can arise with this approach: the normality assumption for the residuals rarely holds for the counts or transformed rates, and the variance of $\log(\lambda_i)$ is by definition heterogeneous:

$$\text{var}[\log(\lambda_i)] \doteq \frac{\text{var}(\lambda_i)}{\lambda_i^2} = \frac{(1-\lambda_i)}{\lambda_i} \tag{9.12}$$

If the event of interest is fairly common and the distribution of counts is symmetrically (if not normally) distributed over the possible range, then a standard linear regression model of the form $y = B_0 + B_1 x_1 + \cdots + B_p x_p$ may be a practical approach. Initial graphical analyses can be very helpful in making this choice, and residual diagnostics also play an important role (see Chapter 7).

9.4.2 Generalized Linear Models for Count Variables

This section considers four related GLMs for regression modeling of count data: the Poisson regression model; the negative binomial regression model; and "zero-inflated" versions of both the Poisson and negative binomial models. The presentation here is deliberately simplified and will focus on the most important concepts for survey analysts, along with an application to the 2006 HRS data on falls. For a more in-depth mathematical treatment of these models, we encourage readers to see Chapter 8 of Long and Freese (2006) or Hilbe (2007).

9.4.2.1 The Poisson Regression Model

The simplest of the generalized linear models for count data is the **Poisson regression model**. The Poisson distribution, $y_i \sim Poisson(t_i\lambda_i)$, is a natural statistical distribution for describing counts of events, y_i, that randomly occur at some expected rate, λ_i, over a period of time, t_i. In the Poisson regression model, the natural **log link** function is employed to model this rate as a linear function of the predictors, $\boldsymbol{x}$:

$$\log(\lambda_i) = B_0 + B_1 x_{1i} + \cdots + B_p x_{pi} \tag{9.13}$$

A Poisson random variable y_i has the unique property that its mean is equal to its variance:

$$E(y_i \mid \boldsymbol{x}_i) = Var(y_i \mid \boldsymbol{x}_i) = t_i\lambda_i$$

or E(count | $\boldsymbol{x}$) = time × E(rate | $\boldsymbol{x}$). Hence, given the model for the log of the rate in Equation 9.13, we have

$$\begin{aligned} E(y_i \mid \boldsymbol{x}_i) &= t_i\lambda_i \\ &= t_i \exp(B_0 + B_1x_{1i} + \cdots + B_px_{pi}) \\ &= \exp[\log(t_i) + B_0 + B_1x_{1i} + \cdots + B_px_{pi}] \end{aligned} \tag{9.14}$$

When formulating a linear model for the expected value of the response variable, the natural log link transformation can be applied to the expected value for y_i defined in Equation 9.14, to produce a convenient linear combination of predictor variables and regression parameters:

$$g[E(y_i \mid \boldsymbol{x}_i)] = \log[E(y_i \mid \boldsymbol{x}_i)] = \log(t_i) + B_0 + B_1x_{1i} + \cdots + B_px_{pi} \tag{9.15}$$

This is the standard Poisson regression model, incorporating an exposure time **offset variable**, $\log(t_i)$, which represents the observation period for individual *i*.

9.4.2.2 The Negative Binomial Regression Model

A key assumption of the Poisson regression model is that the mean and variance of the observed counts are equal:

$$E(y_i \mid \boldsymbol{x}_i) = Var(y_i \mid \boldsymbol{x}_i) = t_i\lambda_i$$

In practice, the variance of the count variable may differ from the mean. The negative binomial regression model is an extension of the Poisson regression model that relaxes this assumption by introducing a dispersion parameter, α, which allows the variance of the count to differ from the mean by a factor of (1 + α): $Var_{NB}(y_i \mid \boldsymbol{x}_i) = E(y_i \mid \boldsymbol{x}_i) \cdot (1+\alpha) = t_i\lambda_i(1+\alpha)$. If the dispersion parameter is equal to 0, the negative binomial model reduces to the Poisson model.

Both the Poisson and negative binomial GLMs share the natural log link and "log rate" form of the generalized linear model, $\log(\lambda_i) = B_0 + B_1x_{1i} + \cdots + B_px_{pi}$, and an identical expression for the expected count, $E(y_i \mid x_i) = \exp[\log(t_i) + B_0 + B_1x_{1i} + \cdots + B_px_{pi}]$. Pseudo-maximum likelihood estimation of the Poisson regression model will be based on a weighted Poisson likelihood function, while the extended negative binomial model will employ a weighted version of the negative binomial likelihood in the

THEORY BOX 9.3 PSEUDO-LIKELIHOOD FOR THE POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

The weighted pseudo-likelihood for estimating the parameters of the Poisson regression model 9.13 is

$$PL(\boldsymbol{B} \mid y_i, \boldsymbol{x_i}) = \prod_{i=1}^{n} \left\{ \frac{(t_i\lambda_i)^{y_i} \exp(-t_i\lambda_i)}{y_i!} \right\}^{w_i} = \prod_{i=1}^{n} \left\{ \frac{\left[t_i \exp(\boldsymbol{x}_i'\boldsymbol{B})\right]^{y_i} \exp[-t_i \exp(\boldsymbol{x}_i'\boldsymbol{B})]}{y_i!} \right\}^{w_i} \tag{9.18}$$

The pseudo-likelihood that is maximized to estimate regression parameters, $\boldsymbol{B}$, of the negative binomial regression model and the dispersion parameter, α, is

$$PL(\boldsymbol{B}, \alpha \mid y_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})} \cdot \left(\frac{\alpha^{-1}}{\alpha^{-1} + t_i\lambda_i} \right)^{\alpha^{-1}} \cdot \left(\frac{t_i\lambda_i}{\alpha^{-1} + t_i\lambda_i} \right)^{y_i} \right\}^{w_i} = \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + t_i \exp(\boldsymbol{x}_i\boldsymbol{B})} \right)^{\alpha^{-1}} \cdot \left(\frac{t_i \exp(\boldsymbol{x}_i\boldsymbol{B})}{\alpha^{-1} + t_i \exp(\boldsymbol{x}_i\boldsymbol{B})} \right)^{y_i} \right\}^{w_i} \tag{9.19}$$

estimation of the regression model parameters and the additional dispersion parameter, α (see Theory Box 9.3).

In general, the Poisson and negative binomial regression models should yield very similar estimates of the regression parameters and rate ratios. However, if the variance of the count data differs from the mean of the counts by a significant amount ($\alpha \neq 0$), fitting the regression model with the simpler Poisson likelihood will result in biased estimates of standard errors and test statistics.

9.4.2.3 Two-Part Models: Zero-Inflated Poisson and Negative Binomial Regression Models

A **two-part model** is an appropriate alternative if the count variable of interest arises through a mixture of processes. Zero-inflated versions of the Poisson and negative binomial regression model (Long and Freese, 2006) include a Part 1 logistic regression model, such as Equation 8.2, of the probability that

the count is zero and a Part 2 Poisson or negative binomial regression model in Equation 9.13 for the actual counts (See Theory Box 9.3). An example is the analysis of y = number of alcoholic drinks consumed per week, where the model that predicts abstinence ($y = 0$) is very likely to differ from the model that predicts frequency and consumption in that segment of the population that may drink alcohol either occasionally or often, as the case may be.

The two-part model allows the survey analyst to specify one set of predictors, z, for the logistic model that predicts the probability of a zero count and a second set of predictors, $\boldsymbol{x}$, for the Poisson or negative binomial regression function for predicting positive counts, $y > 0$. Stata (Version 10+) offers two programs for fitting these two-part models to count variables from complex sample designs: `svy: zip` (zero-inflated Poisson) and `svy: zinb` (zero-inflated negative binomial). The output from each of these programs will provide two sets of estimated coefficients. The first will be the estimated coefficients, standard errors, and test statistics for the Part 1 logit model of the probability that $y = 0$:

$$\log\left[\frac{P(y=0)}{P(y>0)}\right] = A_0 + A_1 z_1 + \cdots + A_q z_q \tag{9.16}$$

Stata labels this part of the model the "inflation," because it will account for the excess of 0 counts that could not be properly modeled using only a one-part Poisson or negative binomial model. The second set of output produced by Stata for the zero-inflated models is the Part 2 estimates for the Poisson or negative binomial regression of positive counts, $y > 0$:

$$\log[E(y_i \mid \boldsymbol{x}_i)] = \log(t_i) + B_0 + B_1 x_{1i} + \cdots + B_p x_{pi} \tag{9.17}$$

9.4.3 Regression Models for Count Data: Specification Stage

As outlined in Section 9.4.2, the model specification stage in regression analysis of count variables requires the analyst to apply scientific reasoning and quantitative evaluation in choosing between a one- and two-part model. The first question to ask is, "Can the count variable be viewed as resulting from a mixture of processes that creates two groups, one that always responds with a zero and one that sometimes responds with a zero?" If the answer is "No: depending on $\boldsymbol{x}$, events may occur at different rates, but the underlying process (e.g., accidents, bouts of flu) is common across the survey population," then a one-part Poisson (`svy: poisson`) or negative binomial regression model (`svy: nbreg`) is preferred. If the answer is "Yes," then either the zero-inflated Poisson model (`svy: zip`) or the zero-inflated negative binomial model (`svy: zinb`) may provide a better fit to the count data.

The choice between the simpler Poisson regression model and the extended negative binomial regression model will depend on whether the survey data satisfy the Poisson assumption, $E(y_i \mid \boldsymbol{x}_i) = Var(y_i \mid \boldsymbol{x}_i) = t_i\lambda_i$ —an assumption that can be formally evaluated by estimating the dispersion parameter α and its design-based 95% CI (see Section 9.4.4). Once the survey analyst has selected the appropriate GLM for his or her count data, the steps in variable selection, tests of nonlinearity, and adding interaction terms can proceed as described previously for other forms of regression analysis.

9.4.4 Regression Models for Count Data: Estimation Stage

Estimation of the parameters in regression models for count data follows the general method of maximizing a weighted Poisson or negative binomial pseudo-likelihood function (see Theory Box 9.3). Section 8.4 has described the basic steps of the Newton–Raphson maximization algorithm for the example case of simple logistic regression.

The default variance estimator in most current software systems is again a version of the Taylor series linearization estimator (9.4) that is appropriate to the data likelihood (Poisson, negative binomial). As is true for all regression models covered in Chapters 7 through 10, most major software packages also provide analysts the option to use BRR or JRR to compute replication variance estimates, $\hat{Var}(\hat{\boldsymbol{B}})_{rep}$.

In Stata (Version 10+), the following programs are available to estimate the various forms of regression models for count variables: `svy: poisson` (Poisson), `svy: nbreg` (negative binomial), `svy: zip` (zero-inflated Poisson), and `svy: zinb` (zero-inflated negative binomial). Outside of the Stata software package, programs for fitting the zero-inflated regression models to complex sample survey data are not widely available; however, Long and Freese (2006) outline how to fit these and other two-part models using separate programs for logistic regression and the standard Poisson or negative binomial regression program.

9.4.5 Regression Models for Count Data: Evaluation Stage

The evaluation stage in building regression models for count data closely parallels that for other forms of generalized linear models. As described in Section 9.4.1, an important evaluation step in modeling count data is to ascertain whether the Poisson or the negative binomial form of the model best fits the data. When the two models are estimated for the same set of predictors, $\boldsymbol{x}$, the Poisson model is "nested" in the extended negative binomial model. For complex sample survey data, Stata does not provide a likelihood ratio test for α but instead outputs a point estimate of α and the design-based 95% CI for the true value. If the 95% CI for α includes 0, one can use the simpler Poisson regression model.

Once the analyst has chosen the type of model, evaluation of the significance of parameters proceeds in standard fashion. Student *t*-tests for the significance of single parameters are routinely included along with the parameter estimates in the output from Poisson and negative binomial regression procedures in statistical software packages enabling survey data analysis. In Stata, design-adjusted Wald tests for multiple-parameter predictors or custom hypotheses for linear combinations of parameters can be obtained through the use of the `test` postestimation command.

Unfortunately, goodness-of-fit tests available for Poisson and negative binomial regression models fitted to simple random samples of data (e.g., the `estat gof` postestimation command in Stata) have not yet been updated for applications involving complex sample survey data. However, in analysis of complex sample survey data, graphical techniques that compare the modeled distribution of counts (e.g., 0, 1, 2) with the observed distribution from the survey can be extremely useful in gauging the quality of the model fit over the range of responses. Here again, readers are referred to Long and Freese (2006) for examples and Stata program syntax. We aim to provide readers with updates in this area on the book Web site.

9.4.6 Regression Models for Count Data: Interpretation Stage

The interpretation of the results of regression models for counts is generally based on estimates and confidence intervals for a statistic termed the **rate ratio** (RR). Based on the identical log(rate) forms of the Poisson and negative binomial models in Equation 9.15, some simple algebra yields

$$\log[E(y_i / t_i \mid \boldsymbol{x}_i)] = B_0 + B_1 x_{1i} + \cdots + B_p x_{pi} \tag{9.20}$$

This formulation of the Poisson regression model enables exponentiated versions of the regression parameters associated with the predictor variables of interest to be interpreted as rate ratio statistics. Consider the ratio of two expected rates when one predictor, x_j, is increased by one unit:

$$\begin{aligned}
&E(y_i / t_i \mid x_{ji} + 1) = \exp[B_0 + B_1 x_{1i} + B_2 x_{2i} + \cdots + B_j (x_{ji} + 1) + \cdots + B_p x_{pi}] \\
&E(y_i / t_i \mid x_{ji}) = \exp[B_0 + B_1 x_{1i} + B_1 x_{2i} + \cdots + B_j x_{ji} + \cdots + B_p x_{pi}] \\
&\Rightarrow \exp(B_j) = \frac{E(y_i / t_i \mid x_{ji} + 1)}{E(\boldsymbol{y}_i / t_i \mid x_{ji})} \\
&= \frac{\text{Expected rate at } (x_{ji} + 1)}{\text{Expected rate at } x_{ji}} = \hat{R}R_j
\end{aligned} \tag{9.21}$$

Holding all other variables constant, a one-unit increase in the predictor variable x_j will therefore multiply the expected rate at which the event occurs by $RR_j = \exp(B_j)$.

In a similar fashion to odds ratios from logistic regression models, one can estimate $100(1 - \alpha)\%$ confidence intervals for rate ratios as

$$CI_{(1-\alpha)}(\hat{R}R_j) = \exp[\hat{B}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_j)] \tag{9.22}$$

9.4.7 Example: Fitting Poisson and Negative Binomial Regression Models to Complex Sample Survey Data

To illustrate the use of Stata procedures for regression modeling of count survey data, Poisson and negative binomial regression models were fitted to a recoded variable, NUMFALLS24, derived from the 2006 HRS variables KC079 (have you fallen in the past two years? 1 = yes, 5 = no) and KC080 (if you have fallen in the past two years, how many times have you fallen?). Before fitting the regression models, the following Stata code is used to generate the dependent variable, NUMFALLS24, from the original HRS variables:

```
gen numfalls24 = kc080 if kc079 == 1
replace numfalls24 = 0 if kc079 == 5
replace numfalls24 = . if kc080 == 98 | kc080 == 99
```

Since only 2006 HRS respondents age 65 and older were asked about falls, a subpopulation indicator, AGE65P, is also generated:

```
gen age65p = .
replace age65p = 1 if kage >= 65 & kage != .
replace age65p = 0 if kage < 65
```

Additional recoding is also employed to create predictor variables for body weight (BODYWGT), height in inches (TOTHEIGHT), and age category (AGE3CAT):

```
gen bodywgt = kc139
replace bodywgt = . if kc139 == 998 | kc139 == 999
gen heightft = kc141 if kc141 <= 7
replace heightft = . if kc141 == 8 | kc141 == 9
gen heightinc = kc142 if kc142 < 98
replace heightinc = . if kc142 == 98 | kc142 == 99
gen totheight = (heightft*12) + heightinc
gen age3cat = .
replace age3cat = 1 if kage >= 65 & kage <= 74
```

TABLE 9.7

Distribution (Unweighted) of the NUMFALLS24 Variable

Variable	*n*	Median	Mean	SD	Var	Min.	Max.
NUMFALLS24 (All)	11,197	0	0.9	2.5	6.2	0	50
NUMFALLS24 (>0)	3,690	2	2.7	3.7	13.6	1	50

Source: Analysis based on the 2006 HRS data.

```
replace age3cat = 2 if kage >= 75 & kage <= 84
replace age3cat = 3 if kage > 84 & kage != .
```

A final set of data preparation steps creates the time period offset variable, OFFSET24 (measured in units of months, so that estimated rates represent falls per month), and stipulates to Stata that GENDER category 2 (Female) will be used as the reference category:

```
gen offset24 = 24
char gender[omit] 2
```

Note that for the 2006 HRS falls measure, the reference period is fixed at 24 (months), enabling us to model the rate of falls per month. In other applications, the exposure time for individual respondents could vary. In such cases, the value of the exposure variable would be set equal to the exposure time recorded for each case.

Figure 9.6 from Section 9.4.1 presents the weighted distribution of the recoded variable NUMFALLS24. Table 9.7 provides a simple (unweighted) statistical summary of the distribution of this variable—both for all cases and cases where the count > 0.

```
sum numfalls24 if age65p==1, detail
sum numfalls24 if numfalls24 >= 1 & age65p==1, detail
```

The output generated by submitting this command in Stata (summarized in Table 9.7) gives us an initial sense of the distribution of this count outcome. First, we note that 11,197 adults provided a nonmissing response to this question. We also note that the variance is substantially greater than the mean, which suggests that the Poisson distribution may not be the best choice of a distribution for this outcome measure. As a result, we will also fit a negative binomial regression model to this count response in Stata. Because more than two-thirds of the observed counts are 0, a two-part zero-inflated version of the negative binomial regression model will also be estimated.

To keep the illustration simple, respondent reports of the number of falls in the past 24 months are modeled as a function of six predictors: GENDER, AGE, ARTHRITIS, DIABETES, BODYWGT, and TOTHEIGHT. For illustration purposes, we do not take the example analysis through all of the Hosmer

and Lemeshow (2000) steps discussed in previous chapters. But we remind readers that the same steps of identifying key main effects, testing nonlinearity of relationships of continuous predictors and investigating potential first-order interactions (see Section 8.3) would apply in regression modeling of count data.

The following three commands are used to fit the Poisson, negative binomial, and zero-inflated negative binomial models to the generated count variable, NUMFALLS24:

```
svyset secu [pweight=kwqtr], strata (stratum)

xi: svy, subpop(age65p): poisson numfalls24 i.gender ///
i.age3cat arthritis diabetes bodywt totheight, exposure ///
(offset24)

xi: svy, subpop(age65p): nbreg numfalls24 i.gender ///
i.age3cat arthritis diabetes bodywgt totheight, irr ///
exposure(offset24)

xi: svy, subpop(age65p): zinb numfalls24 i.gender ///
i.age3cat arthritis diabetes bodywgt totheight, irr ///
exposure(offset24) ///
inflate (i.age3cat i.gender arthritis diabetes bodywgt ///
totheight)
```

Note that in each model command, the Stata syntax includes two options: `irr` and `exposure( )`. The `irr` option requests that Stata report the results in terms of estimated rate ratios associated with each of the predictors and confidence intervals for the rate ratios. The `exposure( )` option specifies the exposure time offset for each sample case (fixed to 24 months in this case).

In this example, part 1 and part 2 of the zero-inflated negative binomial model use an identical set of predictors, x. If the analyst chooses instead to use a distinct set of predictors z for the part 1 logistic model, those variable names would be provided as arguments of the `inflate( )` keyword option in the `svy: zinb` program statement (see `help svy: zinb` within Stata). Another key feature of the commands is the use of `subpop(age65p)` for an appropriate unconditional analysis of the subpopulation of adults 65 years and older.

The remainder of the command is structured like all other regression commands for survey data in Stata. The recoded count response variable (NUMFALLS24) is listed first, followed by the predictor variables of interest. In standard fashion, the `xi:` and `i.` modifiers request that Stata automatically generate indicator variables for all categories of GENDER and AGE3CAT (65–74, 75–84, 85+) and omit the indicators for the reference categories (either prespecified using `char` or defaulting to the lowest values of the variables) from the model.

TABLE 9.8

Estimated Rate Ratios from the Poisson and Negative Binomial Regression Models for NUMFALLS24

		Poisson		Negative Binomial	
Predictor[a]	**Category**	$\hat{RR}_j$	$CI_{.95}(RR_j)$	$\hat{RR}_j$	$CI_{.95}(RR_j)$
AGE	75-84	1.27	(1.140, 1.413)	1.29	(1.136, 1.459)
	85+	1.79	(1.497, 2.148)	1.87	(1.553, 2.260)
SEX	Male	1.20	(0.968, 1.490)	1.14	(0.906, 1.442)
ARTHRITIS	Yes	1.63	(1.379, 1.920)	1.66	(1.392, 1.987)
DIABETES	Yes	1.30	(1.129, 1.489)	1.30	(1.132, 1.492)
WEIGHT	Continuous	1.00	(0.999, 1.003)	1.00	(0.999,1.003)
HEIGHT	Continuous	0.98	(0.956, 1.000)	0.98	(0.963, 1.009)

Source: Analysis based on the 2006 HRS data.

Notes: Estimated dispersion parameter is $\hat{\alpha} = 3.6$, $CI_{.95}(\alpha) = (3.3, 3.9)$. Adjusted Wald Tests for all parameters. Poisson: $F(7,46) = 15.42$. $p < 0.001$. Negative binomial: $F(7,46) = 14.85$. $p = < 0.001$, $n = 10{,}440$.

[a] Reference categories for categorical predictors are 65–74 (AGE), Female (SEX), No (ARTHRITIS), No (Diabetes).

Table 9.8 compares the estimates and 95% confidence intervals for the rate ratios for the Poisson and negative binomial regression models. The results in Table 9.8 suggest that the one-part Poisson and negative binomial regression models would lead to nearly identical conclusions concerning the significance and nature of the effects of the chosen predictors on the rate of falls in the HRS survey population. The estimated confidence intervals for the rate ratios suggest that adults in older age groups (75–85 and >85) and those with ARTHRITIS or DIABETES experience falls at a significantly higher rate. The estimate of the dispersion parameter is $\hat{\alpha} = 3.6$, $CI_{.95}(\alpha) = (3.3, 3.9)$, indicating that the variance of the observed counts of falls is roughly 4.6 times the mean. This, in turn, implies that the negative binomial is the preferred model for this example.

Table 9.9 contains the Part 2 estimates of these same statistics from the zero-inflated negative binomial model as well as the Part 1 estimates of the odds ratios in the logistic regression model for the probability that the reported count = 0. The results for the Part 1 logistic regression model confirm that relative to the aged 65–74 reference group, older persons aged 75–84 ($\hat{\psi} = 0.46$), and 85+ ($\hat{\psi} = 0.01$) have much smaller odds of zero falls in a two-year (24-month) period. After the Part 1 adjustment for the probability of zero falls, the estimated rate ratios from the Part 2 negative binomial regression model suggest that age does not have further significance in predicting the rate at which falls occur. Interestingly, after controlling for age and health status, men ($\hat{\psi} = 3.45$) have much higher odds of zero falls than women—a result that was not evident in the one-part Poisson and negative binomial regression models where gender appeared to have no significant effect on

TABLE 9.9

Estimated Rate Ratios and Odds Ratios from the Zero-Inflated Negative Binomial Regression Model for NUMFALLS24

		Part 2: NB Regression[a]		Part 1: Logistic Zero Inflation[b]	
Predictor	**Category**	$\hat{RR}_j$	$CI_{.95}(RR_j)$	$\hat{\psi}_j$	$CI_{.95}(\psi)$
AGE	75–84	1.08	(0.91, 1.28)	0.46	(0.29, 0.75)
	85+	1.28	(1.00, 1.64)	0.01	(0.00, 0.02)
SEX	Male	1.46	(1.11, 1.92)	3.45	(1.76, 6.86)
ARTHRITIS	Yes	1.44	(1.13, 1.83)	0.53	(0.34, 0.82)
DIABETES	Yes	1.13	(0.94, 1.36)	0.40	(0.19, 0.81)
WEIGHT	Continuous	0.99	(0.99, 1.00)	0.99	(0.99, 1.00)
HEIGHT	Continuous	0.98	(0.96, 1.01)	1.00	(0.93, 1.06)

Source: Analysis based on the 2006 HRS data.

[a] Estimate of dispersion parameter is $\hat{\alpha} = 2.6$, $CI_{.95}(\alpha) = (2.4, 2.9)$.

[b] Stata output for `svy: zinb` with the `irr` option is in the form of $\hat{B}_j = \ln(\hat{\psi}_j)$ for the Part 1 logistic model. Estimates and CIs shown here are based on the standard transformation in (8.21) and (8.23).

falls. Although women are more likely than men to experience at least one fall, men who are prone to experience falls appear to do so at a significantly higher rate than women of the same age and health status ($\hat{RR} = 1.46$).

The results in Table 9.9 also point out that all else being equal, persons with arthritis have reduced odds of zero falls ($\hat{\psi} = 0.53$) and higher adjusted overall rates of falls ($\hat{RR} = 1.44$). Diabetes shows a mixed pattern of reduced odds of not falling ($\hat{\psi} = 0.40$) but no significant impact on the adjusted rate of falls. Controlling for the other predictors, neither body weight nor height are significant in predicting the odds of falling or the rate of falling.

In the complex sample survey data context, there is currently no statistical test to aid in choosing between the standard negative binomial regression model and the zero-inflated alternative. Based on the results in Tables 9.8 and 9.9, the zero-inflated negative binomial regression appears to offer scientific insights into the complexity of this simple count measure for falls among older adults that are not evident from the simpler one-part model.

9.5 Exercises

1. Using the NCS-R data set, fit a multinomial logistic regression model to the categorical dependent variable marital status (MAR3CAT), which measures marital status (1 = CURRENTLY MARRIED/COHABITATING, 2 = PREVIOUSLY MARRIED,

3 = NEVER MARRIED). Consider as candidate predictors age (AGE), gender (SEX: 1 = Male, 2 = Female), education in categories (ED4CAT) (1 = 0–11 years, 2 = 12 years, 3 = 13–15 years, 4 = 16+ years), and alcohol dependence (ALD 1 = yes, 0 = no). Make sure to account for the complex sample design of the NCS-R (stratum codes = SESTRAT, sampling error computation units = SECLUSTR) when estimating standard errors for the parameter estimates, and make sure to compute unbiased estimates of the regression parameters using the final sampling weight from Part 2 of the survey (NCSRWTLG). (Even though most of the variables are from Part 1 of the NCS-R survey, alcohol dependence comes from Part 2 of the survey, thus the use of the NCSRWTLG weight for the subsample of respondents given the "long" part of the questionnaire). Generate a table presenting the results of the analysis, including weighted parameter estimates, design-adjusted standard errors, and 95% confidence intervals for the parameters. Then, answer the following questions:

a. What baseline category did you select for the outcome variable and why? How many generalized logits were estimated as a part of the analysis? Write out the form of each estimated logit function in detail, including the estimated regression parameters and the predictor variables.

b. Using a design-adjusted Wald test, test the null hypothesis that alcohol dependence status and education are not important predictors of marital status. What is your conclusion? Can both of these predictors be removed from the model? Support your conclusion with the design-adjusted test statistic, its degrees of freedom, and its p-value.

c. Using a design-adjusted Wald test, test the null hypothesis that the parameters for education and alcohol dependence status are identical in the logit functions being fitted. What is your conclusion? Support your conclusion with the design-adjusted test statistic, its degrees of freedom, and its p-value.

d. What are the correct interpretations of the estimated coefficients for alcohol dependence status in the multiple logit equations? Use the estimated odds ratios and their 95% confidence intervals in interpreting the results.

e. Attempt to simplify your model by removing predictors that are not important overall but not before testing possible two-way interactions between the predictors. What is the final model that results? Do you encounter any estimation issues when testing the interactions? Write a brief paragraph that summarizes the results of the model fitting.

2. Generate a weighted bar chart examining the estimated distribution of the OBESE6CA outcome in the NCS-R population. Use the NCSRWTSH survey weight, or the Part 1 NCS-R weight. If an analyst were interested in predicting responses on this outcome for a member of this population, would ordinal (cumulative logit) regression be a reasonable analytic approach?
3. Fit an ordinal (cumulative logit) regression model to the OBESE6CA outcome in the NCS-R data (1 = BMI less than 18.5, 2 = 18.5–24.9, 3 = 25–29.9, 4 = 30–34.9, 5 = 35–39.9, 6 = 40+). Be careful to recognize the parameterization of the model given the software procedure that you choose to fit the model. Make sure to incorporate the complex features of the sample design and the Part 1 NCS-R weight in the analysis, and consider the following candidate predictors: age, gender, education, region, and race. Be careful with how you handle the categorical predictors. Complete the following exercises based on the estimated coefficients in this model:
   a. Write the ordinal logit model that you are estimating in detail, including the regression parameters and the predictor variables, and using a general specification for the cutpoint k.
   b. What is the interpretation of the relationship of age with the ordinal OBESE6CA outcome? Does age appear to have an impact on obesity level? Justify your answer.
   c. What is the interpretation of the relationship of gender with the OBESE6CA outcome? Does gender appear to have an impact on obesity level? Justify your answer.
   d. Which variables appear to have a significant relationship with obesity level? Are the directions of the relationships positive or negative (in terms of the actual labels for the values on the OBESE6CA outcome)?
   e. Based on the design-adjusted 95% confidence intervals for the cutpoints in this model, does it appear that one might be able to collapse certain outcome categories to simplify the model?
   f. Based on the estimated parameters in this model, compute the predicted probability that a 50-year-old white male in the East region with 0–11 years of education will report that his obesity level is < 18.5.
4. This exercise focuses on identifying predictors of number of reported falls over the past 24 months (NUMFALLS24) among those 65 and older in the HRS data set. This variable has been created using the raw variables KC079/KC080 from the 2006 HRS data set and is included in the 2006 HRS sample data set for use with this exercise (see Section 9.4.7 for syntax used). These questions were asked

only of those 65 and older; therefore, the use of an unconditional subpopulation analysis approach is appropriate.

a. Fit a Poisson regression model to the NUMFALLS24 response variable, treating the number of falls as a count variable (range = 0 to 50). Remember to use a subpopulation analysis for those aged 65 and older. Consider the predictor variables KAGE, GENDER, DIABETES, and ARTHRITIS, and include two-way interactions between the demographic variables and the physical health conditions (e.g., KAGE × DIABETES). Make sure to account for the complex sample design of the HRS (stratum codes = STRATUM, sampling error computation units = SECU) when estimating standard errors for the parameter estimates, and make sure to compute unbiased estimates of the regression parameters using the final sampling weights (KWGTR). Generate a table presenting the results of the analysis, including weighted parameter estimates, design-adjusted standard errors, and 95% confidence intervals for the parameters.

b. Based on the estimates of the parameters in this model, how much does the relative risk of having falls change with a five-year change in age for persons with diabetes (with all other predictor variables held constant)? Report a 95% confidence interval for this risk ratio. Hint: See Section 8.6.

c. Are any of the physical health conditions (e.g., diabetes) important predictors of experiencing a fall during the past two years? If so, interpret the relationships of these predictors with the rate of falls. Keep in mind the interactions that are included in the model when making your interpretations.

d. Refit the model using the same design-based approach, only assuming a negative binomial distribution for the count of falls for the pseudo-maximum likelihood estimation. What is the estimate of the dispersion parameter in this model? What does this suggest about the Poisson regression approach?

e. Compare the estimates of the parameters and the estimated risk ratios obtained using each approach, and discuss whether any of your inferences would differ depending on the assumed distribution for the counts.

f. Would a two-part modeling approach make sense for these data? Justify the reason for your answer.

10

Survival Analysis of Event History Survey Data

10.1 Introduction

Survival analysis, or **event history analysis** as it is often labeled in the social sciences, includes statistical methods for analyzing the time at which "failures" or events of interest occur. Common examples of **event times** of interest in the survival analysis literature include light bulb or motor longevity (engineering), time of death, time to disease incidence or recovery (medicine and public health), time to unemployment, time to divorce, or time to retirement (social sciences). Survival analysis is an important statistical topic that when treated in its full depth and breadth fills entire volumes. This chapter will scratch only the surface of this topic, focusing on basic theory (Section 10.2) and application for three survival analysis techniques that are commonly used with complex sample survey data. Section 10.3 will introduce nonparametric Kaplan–Meier (K–M) analysis of the survivorship function. The Cox proportional hazards (CPH) model will be covered in Section 10.4, and Section 10.5 presents a description and application of logit and complementary log–log (CLL) models for discrete time event history data. Readers interested in the general theory and applications of survival analysis methods are referred to classic texts including Kalbfleisch and Prentice (2002), Lee (1992), and Miller (1981). Newer texts on applied methods (Hosmer, Lemeshow, and May, 2008) and several excellent user guides are also available that illustrate procedures for survival analysis in SAS (Allison, 1995) and Stata (Cleves et al., 2008).

10.2 Basic Theory of Survival Analysis

10.2.1 Survey Measurement of Event History Data

In the context of population-based survey research, survival analysis problems arise in three primary ways: (1) through longitudinal observations on

individuals in a panel survey; (2) by administrative record follow-up of survey respondents; and (3) from retrospective survey measurement of events and times that events occurred. The Health and Retirement Study (HRS) and other longitudinal panel studies prospectively follow and reinterview sample members over time. In this **prospective cohort design**, dates of events such as retirement, institutionalization, morbidity, and mortality are captured in the longitudinal data record. Administrative follow-up of a sample of survey participants using vital statistics, medical records, or other record-keeping systems also generates a prospective measurement of events of interest and the dates on which they occurred. The National Health and Nutrition Examination Survey (NHANES) III linked mortality file (http://cdc.gov/nchs/data/datalinkage) is an excellent example where U.S. Vital Statistics records were used to prospectively detect and record death of an NHANES III respondent for a period of years after the survey was conducted.

The third means of generating "survival data" in the survey context is through the use of retrospective measurement of event histories. Figure 10.1 is an illustration of a question sequence from the National Comorbidity Survey Replication (NCS-R) that uses retrospective recall to report and date events over the course of the respondent's lifetime (or other relevant observation period). Survey methodologists are well aware of recall, telescoping, and

*D37. Think of the very first time in your life you had an episode lasting (several days or longer / two-weeks or longer) when most of the day nearly every day you felt (sad/or/discouraged/or/uninterested) and also had some of the other problems (you cited earlier). Can you remember your exact age?
YES.....1
NO.....5 **GO TO *D37b**
DON'T KNOW.....8 **GO TO *D37b**
REFUSED.....9 **GO TO *D37b**

*D37a. (How old were you?)
_____ YEARS OLD **GO TO *D37b.1**
DON'T KNOW.....998 **GO TO *D37b.1**
REFUSED.....999 **GO TO *D37b.1**

*D37b. About how old were you (the first time you had an episode of this sort)?
IF "ALL MY LIFE" OR "AS LONG AS I CAN REMEMBER,"
PROBE: Was it before you first started school?
IF NOT YES, PROBE: Was it before you were a teenager?
_____ YEARS OLD
BEFORE STARTED SCHOOL 4
BEFORE TEENAGER 12
NOT BEFORE TEENAGER.....13
DON'T KNOW.....998
REFUSED.....999

FIGURE 10.1
Example of NCS-R retrospective event recall questions.

nonrandom censoring errors that may be associated with such retrospective measurements (Groves et al., 2004). Techniques such as event history calendars that use key lifetime events and dates are often employed to anchor the retrospective recall measurement of events and times (Belli, 2000).

10.2.2 Data for Event History Models

Time itself is at the heart of survival analysis—time to occurrence of an event, T, or the time at which the observation on the survey subject is censored, C. Only one of these two times is directly measured. T is measured if the event of interest occurs during the survey observation or follow-up period. The censoring time, C, is measured in cases where the respondent is **lost to follow-up** (due to noncontact, refusal, or other reasons) or if the survey observation period has ended and no event has yet occurred. Estimation of the survival analysis model requires assumptions concerning the relationship of T and C—the most common being that time to event and censoring time distributions are independent of each other after controlling for any covariates in the model.

The majority of survival analysis problems deal with time to a single event or the first occurrence of the event. Often the transition is to an **absorbing state** (e.g., mortality). This is the class of applications that we will consider in this chapter. However, readers should be aware that event history methods have been extended to include models of repeated events (e.g., heart attacks, marriages) or **spells** of time between events (e.g., unemployment episodes; see Allison, 1995; Yamaguchi, 1991).

In the literature and in applications, survival models are often referred to as **continuous time** or **discrete time** models. The dependent variables T and C may be measured on a continuous scale (e.g., number of milliseconds until failure) or a discrete scale (e.g., number of years of marriage until divorce, on an integer scale). Discrete survival times often arise in longitudinal survey designs where observations occur on a periodic basis (e.g., every two years for the HRS), and it is difficult to precisely pinpoint the date and time that the event occurred.

Figure 10.2 illustrates the general nature of the observational data that are available for a survival analysis of survey data. The figure illustrates the **observational window** spanning the time period at which survey observations begin to the point in time when the survey observations end. In the figure, the first horizontal "time" arrow illustrates the case where an event occurs and the time to the event, T, is recorded. The second arrow represents a case that is followed from the beginning of the survey observation window but is lost to follow-up (censored) before the study is complete. A second form of censored observation is represented by the third arrow in the figure. This case is successfully monitored from the beginning to the end of the observation window, but the event of interest does not occur. In cases 2 and 3, the time that is recorded in the data set is the censoring time, C; an event

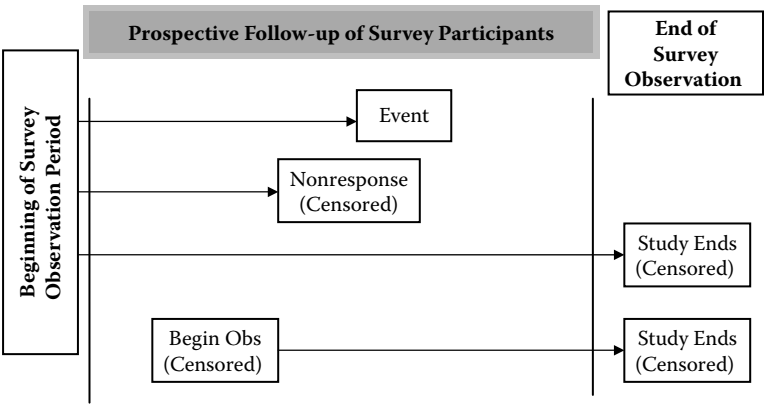


FIGURE 10.2
Prospective view of event history survey data.

time is never observed. The censoring mechanism that is illustrated by cases 2 and 3 in Figure 10.2 is termed **right censoring.** Survey observations may also be **left censored** as illustrated by the fourth and final horizontal arrow in Figure 10.2. Left censoring of survey observations can occur for several reasons including ineligibility in the initial period of survey observation or initial nonresponse.

"Censored" observations such as these are important in survival analysis since they remain representative of the **population at risk** up to the point that further observation is censored. In the analysis examples presented in this chapter, we will assume **Type 1 Censoring**, where any censoring is random over the fixed period of observation.

10.2.3 Important Notation and Definitions

Prior to examining survival analysis models in more detail, several important statistical definitions are required. The **probability density function** for the event time is denoted by $f(t)$ and is defined as the probability of the event at time t (for continuous time) or by $\pi(m)$, denoting the probability of failure in the interval $(m, m + 1)$ for discrete time. The corresponding cumulative distribution functions (CDFs) are defined in the standard fashion:

$$F(t) = \int_0^t f(t)dt \text{ for continuous } t; \text{ or}$$
$$F(m) = \sum_{k \le m} \pi(k) \text{ for } t \text{ measured in discrete intervals of time} \tag{10.1}$$

The CDFs for survival time measure the probability that the event occurs at or before time t (continuous) or before the close of time period m (for discrete time).

The **survivor function** or **survivorship function**, $S(t)$, is the complement to the CDF and is defined as follows:

$$\begin{aligned} &S(t) = 1 - P(T \le t) = 1 - F(t) \text{ for continuous time; or} \\ &S(m) = 1 - F(m) \text{ for discrete time.} \end{aligned} \tag{10.2}$$

The value of the survivor function for an individual is the probability that the event has *not yet occurred* at time t (continuous) or prior to the close of observation period m (discrete time).

The concept of a **hazard** or **hazard function** plays an important role in the interpretation of survival analysis models. A hazard is essentially a conditional probability. For continuous time models, the hazard is $h(t) = f(t) / S(t)$, or the conditional probability that the event will occur at time t given that it has not occurred prior to time t. In discrete time models, this same conditional probability takes the form $h(m) = \pi(m) / S(m-1)$.

10.2.4 Models for Survival Analysis

Survival analysis models are classified into four major types, based on the assumptions that are made concerning the probability distributions of survival times. We briefly discuss these four types of survival analysis models in this section.

Parametric survival models assume a parametric distribution for the probability density function $f(t)$. A very common example is the exponential distribution with rate parameter 1. In this case, $f(t) = e^{-t}$, $F(t) = 1 - e^{-t}$, and $S(t) = e^{-t}$. This distribution is characterized by a constant hazard function. Additional common examples of parametric models for survival times include the Weibull distribution (often used for human mortality), the lognormal distribution, and the gamma distribution. For more details on parametric models for survival data, we refer readers to Kalbfleisch and Prentice (2002) or Miller (1981). Software procedures for fitting parametric survival models to complex sample survey data are not yet widely available. Stata (Version 10+) currently provides the `svy: streg` command for users interested in fitting these models to survey data, but similar procedures have not yet been widely implemented in other statistical packages. As a result, our focus in this chapter will be on nonparametric and semiparametric survival models that can currently be fitted in several major software packages while taking complex sample designs into account.

Nonparametric survival models make no assumptions concerning the probability density functions of survival times and include approaches like **Kaplan–Meier estimation** and **life table methods**. These empirically based methods estimate survivor functions and hazards solely using the observed

survey data and do not rely on any parametric assumptions (i.e., assumed probability models). We present examples of applying these methods to survey data in Section 10.3.

Models such as the popular **Cox proportional hazards model** are labeled as **semiparametric survival models**. The CPH model makes no strong assumptions concerning the underlying probability distribution for event times but does make an assumption of proportionality in the regression parameterization of the individual hazard associated with the covariate vector, $\boldsymbol{x}$. We consider an example of fitting a CPH model to survey data in Section 10.4.

Finally, a special class of survival models exists for **discrete time event history data**. The models and methods previously introduced are largely applicable for survival times that are roughly continuous in nature, taking on many possible values on a continuum. Some survival times may be measured in terms of discrete units of time only; for example, in a panel survey covering five years, a researcher might wish to analyze the number of years until a person loses his or her job (1, 2, 3, 4, or 5) as a function of selected covariates. Because these failure times are measured in terms of a small number of discrete values, alternative methods are needed. Section 10.5 introduces **discrete time logit models** and **complementary log–log models** for discrete time event history data.

In each of the following sections, the example applications will be based on a survival analysis of the age of onset for a major depressive episode (MDE) question from the NCS-R (Figure 10.1). The event of interest for the survival analysis is thus the first MDE respondents experience. The event time is the age they report experiencing their first MDE. Individuals who have never in their life prior to interview had an MDE are assumed to provide right-censored data; that is, we analyze their ages at interview, but the ages are considered to be right censored for the purposes of the survival analysis. In Sections 10.4 and 10.5 we illustrate the fitting of Cox proportional hazards and discrete time logit regression models to the NCS-R data in Stata, considering gender (SEX), age at interview (INTWAGE), marital status (MAR3CAT), education (ED4CAT), and ethnicity (RACECAT) as possible predictors of the hazard of having a first MDE at a given age. Note that in this analysis, each NCS-R respondent's observation window (Figure 10.2) begins at birth and continues until (1) the age at which they first experience an MDE or (2) the age at which they are interviewed for NCS-R.

10.3 (Nonparametric) Kaplan–Meier Estimation of the Survivor Function

The **Kaplan–Meier (K–M)** or **product limit estimator** is a nonparametric estimator of the survivorship function, $S(t)$. It can be applied to event times

that are measured on a continuous time (t) basis or as counts of events measured over discrete time periods, $m = 1, \ldots, M$.

10.3.1 K–M Model Specification and Estimation

A general form of the K–M estimator for complex sample survey data applications can be expressed as

$$\hat{S}(t) = \prod_{t(e) \le t} (1 - \hat{D}_{t(e)} / \hat{N}_{t(e)}) \tag{10.3}$$

where

$t(e)$ = times at which unique events $e = 1, \ldots, E$ are observed;

$$\hat{D}_{t(e)} = \sum_{i=1}^{n} \mathrm{I}\left[t_i = t(e)\right] \cdot \delta_i \cdot w_i; \quad \hat{N}_{t(e)} = \sum_{i=1}^{n} \mathrm{I}\left[t_i \ge t(e)\right] \cdot w_i;$$

$I[\bullet]$ = indicator = 1 if [expression] is true, 0 otherwise;

$\delta_i = 1$ if the observed time for case i is a true event, 0 if censoring time;

w_i = the sampling weight for observation i

The survival at any point in time t is estimated as a product over unique event times where $t(e) \le t$ of the estimated conditional survival rate at each event time $t(e)$, $(1 - \hat{D}_{t(e)} / \hat{N}_{t(e)})$. Note how at each unique event time the rate of events is a ratio of the weighted estimate of total events at $t(e)$ to the estimated population "at risk" at time $t(e)$. Note also that the sampling weights for sampled elements are incorporated into the estimation of the survivorship function.

Unlike the regression-based methods described later in this chapter, the Kaplan–Meier estimator of survivorship does not permit adjustment for individual covariates. However, the Kaplan–Meier estimates of survivorship can be computed separately for subgroups, $h = 1, \ldots, H$, of the survey population (e.g., separately for men and women). At the time of this writing, SUDAAN PROC KAPMEIER provides the capability to conduct a full K-M analysis for complex sample survey data including estimation for groups using the STRHAZ statement. Stata provides the capability to compute and plot weighted estimates of the survivor function, but Version 10 does not permit design based estimation of the standard errors of the estimates or confidence intervals (CIs). SPSS v16 enables design-adjusted Kaplan–Meier estimation using the CSCOXREG command. SAS PROC LIFETEST permits FREQUENCY weighted Kaplan–Meier estimation of the survival function, but this procedure does not support design-adjusted estimates of standard errors and confidence intervals, nor does it allow the use of noninteger weights. Any changes in terms of the capabilities of these software procedures will be reported on the book Web site.

THEORY BOX 10.1 CONFIDENCE INTERVALS USING KAPLAN–MEIER ESTIMATES OF $\hat{S}(t)$

A two-step (transformation, inverse-transformation) procedure can be used to estimate confidence intervals for the survivorship at time t, $S(t)$.

In the first step, the 100(1 – α)% CI is estimated on the log(–log($\hat{S}(t)$) scale:

$$CI_{g(S(t))} = g(\hat{S}(t)) \pm t_{df,1-\alpha/2} \cdot se(g(\hat{S}(t)))$$

$$where:$$

$$g(\hat{S}(t)) = \log(-\log(\hat{S}(t))); \; and \tag{10.4}$$

$$se(g(\hat{S}(t))) = \sqrt{\frac{\text{var}(\hat{S}(t))}{\left[\hat{S}(t) \cdot \log(\hat{S}(t))\right]^2}}$$

For simple random sample data, $\text{var}(\hat{S}(t))$ is based on Greenwood's (1926) estimator. For complex samples, $\text{var}(\hat{S}(t))$ is estimated by Taylor series linearization (RTI, 2004) or through JRR methods.

The inverse transformation, $\exp(-\exp(\bullet))$ is then applied to the lower and upper limits of $CI_{1-\alpha}(g(S(t)))$ to construct the 100(1 – α)% CI for $S(t)$.

Under complex sample designs, SUDAAN employs a Taylor series linearization (TSL), balanced repeated replication (BRR), or jackknife repeated replication (JRR) estimator of the standard error of $\hat{S}(t)$. SUDAAN's TSL estimates of $se(\hat{S}(t))$ require the generation of a large matrix of derivative functions that may lead to computational difficulties. SUDAAN allows users to circumvent this problem by estimating the K-M survivorship and standard errors for a subset of points (e.g., deciles) of the full survivorship distribution. Another alternative that we recommend is to employ the JRR option in SUDAAN to estimate the full survivorship distribution and confidence intervals for the individual $\hat{S}(t)$. Using the TSL or JRR estimates of standard errors, confidence limits for individual values of $\hat{S}(t)$ are then derived using a special two-step transformation/inverse-transformation procedure (see Kalbfleisch and Prentice, 2002 and Theory Box 10.1 for details).

10.3.2 K–M Estimator—Evaluation and Interpretation

Since the K–M estimator is essentially "model free," the evaluation phase of the analysis is limited to careful inspection and display of the estimates

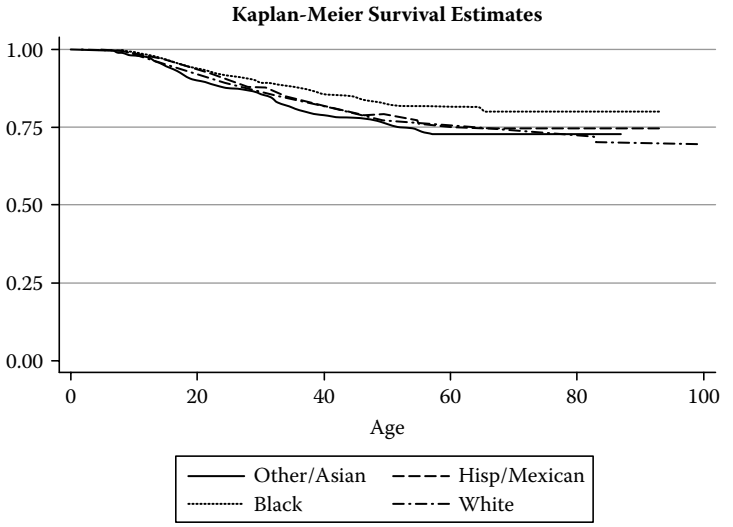


FIGURE 10.3
Weighted Kaplan–Meier Estimates of the survivor functions for lifetime major depressive episode for four ethnic groups. (Modified from the NCS-R data.)

of $\hat{S}(t)$ and the associated standard errors. Interpretation and display of the results are best done using plots of the estimated survivorship functions against time. Such plots are an effective way to display the overall survivorship curves and to compare the survival outcomes for the distinct groups (see Figure 10.3).

Under simple random sampling (SRS) assumptions, two quantitative tests (a Mantel–Haenszel test and a Wilcoxon test) comparing observed and expected deaths over time can be used to evaluate whether the survivorship curves for two groups are equivalent (i.e., test a null hypothesis defined as $H_0 : S_1(t) = S_2(t)$). Analogous Wald-type X^2 test statistics could be developed for complex sample applications, but to the best of our knowledge these have not been implemented in the current software programs that support Kaplan–Meier estimation for complex sample survey data. In this chapter, we will focus on the graphical display of the K–M estimates of the survivorship function, using these plots to inform decisions about predictors to use in CPH and discrete time regression models of survival.

10.3.3 K–M Survival Analysis Example

We begin our analysis example for this chapter by computing and plotting weighted **Kaplan–Meier estimates** of the survivor functions (10.3) for four NCS-R subgroups defined by the race variable (RACECAT). Even if a full K–M analysis is not the goal, this initial step is useful to visually examine

the survivor functions for different groups and to determine whether (1) the functions appear to differ substantially and (2) the survival functions appear to be roughly parallel (one of the assumptions underlying the Cox model of Section 10.4).

We first define a new variable AGEONSETMDE as being equal to the age of first MDE (MDE_OND), if a person has had an MDE in his or her lifetime, or the age of the respondent at the interview (INTWAGE) if a person has not had an MDE in his or her lifetime:

```
gen ageonsetmde = intwage
replace ageonsetmde = mde_ond if mde == 1
```

The recoded variable, AGEONSETMDE, will serve as the time-to-event variable in this example. Next, we declare the structure of the NCS-R time-to-event data in Stata using the `stset` command. This is a required specification step that must be run before invoking a Stata survival analysis command:

```
stset ageonsetmde [pweight = ncsrwtsh], failure(mde==1)
```

With the initial `stset` command, we define AGEONSETMDE as the variable containing the survival times (ages of first MDE onset, or right-censored ages at NCS-R interview for those individuals who had not yet experienced an MDE in their lifetime) and the indicator of whether an event actually occurred or whether the data should be analyzed as right censored. This indicator is simply the indicator of lifetime MDE in the NCS-R data set (MDE), equal to 1 if the individual has ever had an MDE, and 0 otherwise (the censored cases). Note that we also define a "pweight" (or sampling weight) for the time-to-event data with the `[pweight = ncsrwtsh]` addition. Once the `stset` command has been submitted, Stata displays the following output:

```
     failure event:   mde == 1
obs. time interval:   (0, ageonsetmde]
 exit on or before:   failure
           weight:    [pweight=ncsrwtsh]
--------------------------------------------------------------------------
  9282   total obs.
     0   exclusions
--------------------------------------------------------------------------
  9282   obs. remaining, representing
  1829   failures in single record/single failure data
385696   total analysis time at risk, at risk from    t = 0
                              earliest observed entry t = 0
                                   last observed exit t = 99
```

We see that there are 1,829 recorded "failure events" or individuals having experienced an MDE in their lifetime (with MDE = 1). Stata also processes the

total number of observations at risk at each age starting with age 0, and the last observed exit from the risk set (an individual was interviewed at age 99 and reported never having an episode of MDE). Stata saves these variables in memory for any subsequent survival analysis commands, similar to how `svyset` works for identification of complex design variables (strata, clusters, weights).

Stata allows users to plot weighted population estimates of the survivor functions (provided that sampling weights were specified in the `stset` command) for different groups by using the `sts graph` command with a `by()` option, and we now apply this command to the NCS-R data:

```
sts graph, by(racecat) legend(lab(1 "Other/Asian") ///
lab(2 "Hisp/Mexican") lab(3 "Black") lab(4 "White"))
```

This command generates the plot in Figure 10.3.

The resulting plot in Figure 10.3 is very informative. First, by age 90, about 25% of the persons in each race group will have had a major depressive episode. Further, the black race group consistently has fewer persons that have experienced a major depressive episode by each age considered. We also note that there is no strong evidence of distinct crossing of the estimated survival functions across ages; an assumption of parallel lines would seem fairly reasonable for these four groups.

Stata users can also use the `sts list` command to display estimates of the overall survivor function at specified time points (or ages in this example). The following command requests estimates of the survivor function at seven unique time points:

```
sts list, at(10 20 30 40 50 60 70)
```

Submitting this command results in the following Stata output:

```
      failure _d:  mde == 1
analysis time _t:  ageonsetmde
          weight:  [pweight=ncsrwtsh]
```

Time	Beg. Total	Fail	Survivor Function
10	9171.84	139.004	0.9850
20	8169.95	632.016	0.9165
30	6390.66	401.574	0.8659
40	4736.49	327.113	0.8155
50	2997.27	186.745	0.7760
60	1779.61	59.0606	0.7568
70	949.309	22.2811	0.7438

```
Note: survivor function is calculated over full data and
evaluated at indicated times; it is not calculated from
aggregates shown at left.
```

TABLE 10.1

Selected Results from the Kaplan–Meier Analysis of the NCS-R Age of Onset for MDE Data in SUDAAN

Age of Onset	$\hat{S}(t)$	SUDAAN $se(\hat{S}(t))$	SUDAAN $CI_{.95}(S(t))$
10	0.985	0.001	(0.982, 0.988)
20	0.917	0.004	(0.908, 0.924)
30	0.866	0.005	(0.855, 0.876)
40	0.816	0.005	(0.804, 0.826)
50	0.776	0.006	(0.764, 0.787)
60	0.757	0.006	(0.745, 0.769)
70	0.744	0.007	(0.730, 0.757)

We note that design-adjusted standard errors for these estimates of the survivor function are not readily available (any updates in Stata will be indicated on the book Web site). At present, SUDAAN PROC KAPMEIER does provide the capability to compute weighted estimates of the survivor function along with design-based standard errors and 95% confidence intervals for the survival function. Example SUDAAN syntax required to conduct this descriptive analysis for the NCS-R MDE data is as follows:

```
proc kapmeier ;
nest sestrat seclustr ;
weight ncsrwtsh ;
event mde ;
class racecat ;
strhaz racecat ;
time ageonsetmde ;
setenv decwidth=4 ;
output / kapmeier=all filename="c10_km_out" filetype=sas ;
replace ;
run ;
```

Table 10.1 contains an extract from the complete set of results produced by this analysis. The table displays the total population K–M estimates of the survivor function, along with the standard errors and 95% CIs for the values of the survivor function at ages t = 10, 20, 30, …, 70. The full output includes age-specific estimates of the survivor function for the total population and separately for each of the four race/ethnic groups defined by RACECAT. In SUDAAN PROC KAPMEIER, the stratified K–M analysis is invoked by the use of the STRHAZ statement. SUDAAN does not provide the capability to plot the estimated curves and confidence intervals for the survivorship at each age of onset; however, the analysis output can be saved using SUDAAN's OUTPUT statement and exported to a software package that has a general graphics capability to produce the desired plots.

10.4 Cox Proportional Hazards Model

The CPH model (Cox, 1972) is one of the most widely used methods for survival analysis of continuous time event history data. CPH regression provides analysts with an easy-to-use and easy-to-interpret multivariate tool for examining impacts of selected covariates on expected hazard functions. Despite its widespread use and the semiparametric nature of the underlying model, survey analysts must be careful to examine the critical "proportional hazards" assumption and to ensure that the model is fully and correctly identified. Focusing largely on clinical trials, Freedman (2008) cautions against blind use of Cox regression models without performing some initial descriptive analyses of the survival data.

10.4.1 Cox Proportional Hazards Model: Specification

The Cox model for proportional hazards is specified as follows:

$$h(t \mid \boldsymbol{x}_i) = h_0(t)\exp\left(\sum_{j=1}^{P} B_j x_{ij}\right) \tag{10.5}$$

In Equation 10.5, $h(t \mid \boldsymbol{x}_i)$ is the expected hazard function for an individual i with vector of covariates $\boldsymbol{x}_i$ at time t. The model includes a baseline hazard function for time t, $h_0(t)$, that is common to all population members. The individual hazard at time t is the product of the baseline hazard and an individual-specific factor,

$$\exp\left(\sum_{j=1}^{P} B_j x_{ij}\right)$$

a function of the regression parameters, $\boldsymbol{B} = \{B_1,\ldots,B_p\}$ and the individual covariate vector, $\boldsymbol{x}_i = \{x_{i1},\ldots,x_{ip}\}$. Individual hazards are therefore a proportional scaling of the baseline hazard: $h(t \mid \boldsymbol{x}_i) \propto h_0(t)$. There is no separate intercept parameter in the CPH model specification; the baseline expectation (i.e., the hazard when all covariates are equal to 0) is absorbed into $h_0(t)$.

The CPH model is most applicable to continuous or nearly continuous measurement of event and censoring times. Available software for estimating the CPH model anticipates the possibility that multiple events/censoring may occur at any time point t and incorporates rules for handling such ties. For example, Stata assumes that events occur before censoring if there are ties—censored cases at time t remain in the risk set for events at time t. If the time scale for the survey observations is very coarse (e.g., years in a study of

children aged 5–18) or explicitly discrete (e.g., the two-year period between consecutive HRS survey interviews), many events and censorings may occur at time period t. If the measurement of time is coarse and ties are common, analysts might consider the complementary log–log discrete time model of Section 10.5 as an alternative to CPH regression.

The values of covariates in the CPH model may be fixed (e.g., gender, race), or they may be **time-varying covariates**, meaning that the values of the covariates can change depending on the time at which the set of sample units "at risk" is being evaluated. Stata allows users to set up time-varying covariates for use in the `svy: stcox` command for estimating the CPH model. If the desired model includes a large number of time-varying covariates, an alternative to the CPH model is to recode the event and censoring times into discrete time units and apply the discrete time logistic or complementary log-log models described in Section 10.5.

10.4.2 Cox Proportional Hazards Model: Estimation Stage

The CPH regression parameters B_j are estimated using a **partial likelihood** procedure, based on the conditional probability that an event occurred at time t (see Freedman, 2008 for a nice primer or Lee, 1992 for more details). For estimation, the E observed event times are ordered such that $t_{(1)} < t_{(2)} < \ldots < t_{(E)}$. The **risk set** at a given time t, $t = 1, \ldots, E$, is the set of respondents who (1) have not experienced the event prior to time t, and (2) were not randomly censored prior to time t. The probability of an event that occurs at time t for the i-th respondent, conditional on the risk set R_t, is defined as follows:

$$P(t_i = t \mid \boldsymbol{x_i}, R_t) = \frac{\exp\left(\sum_{j=1}^{P} B_j x_{ij,t}\right)}{\sum_{l \in R_t} \exp\left(\sum_{j=1}^{P} B_j x_{lj,t}\right)} \tag{10.6}$$

where R_t = {set of cases still "at risk" at time t}.

This conditional probability is computed for every event that occurs. The partial likelihood function for the survey observations of event times is then defined as the product of these probabilities (see Theory Box 10.2). An iterative mathematical algorithm is then applied to derive estimates of the regression parameters that maximize the partial likelihood function.

Binder (1992) describes the Taylor series linearization approach for estimating $Var(\hat{\boldsymbol{B}})_{TSL}$ that accounts for the complex sample design. BRR and JRR methods can also be used to estimate a replicated alternative, $Var(\hat{\boldsymbol{B}})_{Rep}$. Routines capable of fitting proportional hazards models and computing design-based estimates of standard errors for the parameter estimates are currently

THEORY BOX 10.2 PARTIAL LIKELIHOOD FOR THE COX PROPORTIONAL HAZARDS MODEL

The partial likelihood that forms the basis for estimating the regression parameters of the Cox proportional hazards model is the product of conditional probabilities—one for each distinct event, $e = 1, \ldots, E \leq n$, that is recorded in the survey period. Each conditional probability component of this partial likelihood is the ratio of the time $t(e)$ hazard for the case experiencing the event (case i) to the sum of hazards for all sample cases remaining in the risk set at $t(e)$:

$$L(B) = \prod_{i=1}^{n} \left[\frac{h(t(e)_i \mid \boldsymbol{x}_i)}{\sum_{j=1}^{n} I\left[t(0)_j < t(e)_i \leq t(e)_j \right] \cdot h(t(e)_j \mid \boldsymbol{x}_j)} \right]^{\delta_i \cdot w_i}$$

where

$t(0)_j$ is the observation start time for respondent j;
$t(e)_i$ and $t(e)_j$ are the respective event (censoring) times for respondents i and j;
$I[t(0)_j < t(e)_i \leq t(e)_j]$ is the 0, 1 indicator that respondent j is in the risk set when unit i experiences an event;
$\delta_i = 1$ if unit i experiences the event, 0 if unit i is censored;
w_i = the survey weight for unit i.

In the partial likelihood for complex sample survey applications, the contribution of the i-th case to the partial likelihood is raised to the power of its survey weight, w_i. Censored cases that remain in the risk set contribute to the denominator of the conditional probabilities for events that occur to other cases; however, the direct contributions of censored cases to the product likelihood are factors of 1 (due to the $\delta_i = 0$ exponent).

implemented in the SPSS Complex Samples module, Stata, SUDAAN, R, and IVEware (see Appendix A).

10.4.3 Cox Proportional Hazards Model: Evaluation and Diagnostics

Procedures for constructing confidence intervals and testing hypotheses for the parameters of the Cox proportional hazards model parallel those described in previous chapters for the parameters of other linear and generalized linear regression models (see Section 7.5, for example).

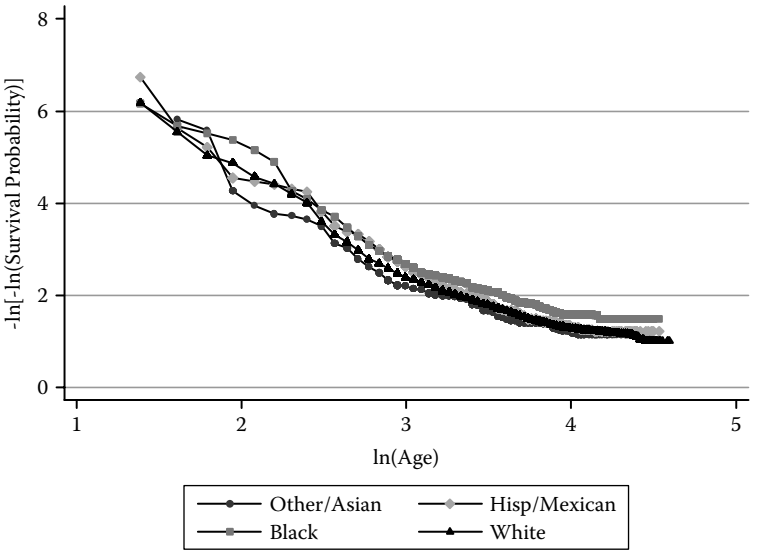


FIGURE 10.4
Testing the proportional hazards assumption for the fitted Cox model, with no adjustment for covariates. (Modified from the NCS-R data.)

Diagnostic toolkits for CPH models fitted to complex sample survey data sets are still in the early stages of development. Inspection of the Kaplan–Meier estimates of the survivor functions for different groups can be helpful as an initial step to see if the survival functions are approximately parallel. A better graphical check of the proportional hazards assumption that is currently implemented in Stata is a plot of $-\ln(-\ln(\hat{S}_h(t)))$ against $\ln(t)$, where $\hat{S}_h(t)$ is a weighted estimate of the survival function for group h (see Figure 10.4). Based on the specified Cox model, the transformed versions of the survival functions should be parallel as a function of $\ln(t)$.

At the time of this writing, Stata and other software systems have not incorporated many of the standard residual diagnostics for CPH models in the CPH model programs for complex sample survey data (Version 10 of SUDAAN and Versions 10+ of Stata provide the most options). For example, it is possible to generate **partial Martingale residuals** for the purpose of checking the functional forms of continuous covariates in the Cox model. These residuals are computed as

$$M_i = \delta_i - \hat{H}_i(t) \tag{10.7}$$

where δ_i represents an indicator variable for whether or not a given sample case i had the event of interest occur (1 = event, 0 = censored), and $\hat{H}_i(t)$ represents a weighted estimate of the **cumulative hazard function** based on the fitted model (or the cumulative sum of all instantaneous probabilities of

failure up until time *t*; one can also think of this function as the "total rate" of experiencing an event up to time *t*). These residuals can be plotted against the values of individual continuous covariates to check for the possibility of nonlinear relationships, which would indicate that the functional forms of continuous covariates may have been misspecified.

Any developments allowing users to generate these residuals and further evaluate the goodness of fit of Cox models will be highlighted on the companion Web site for the book. For additional discussion of fitting Cox models to complex sample survey data sets, we refer interested readers to Cleves et al. (2008, Section 9.5).

10.4.4 Cox Proportional Hazards Model: Interpretation and Presentation of Results

Interpretation of results and population inference from the CPH model is typically based on comparisons of **hazards—**the conditional probabilities that the event will occur at time *t* given that it has not occurred prior to time *t***.** Given estimates of the regression parameters in the model, consider the ratio of estimated hazards if predictor variable x_j is incremented by one unit and all other covariates remain fixed:

$$\begin{aligned} \hat{HR}_j &= \frac{h_0(t)\exp(\hat{B}_1 x_1 + \cdots \hat{B}_j(x_j + 1) + \cdots \hat{B}_p(x_p)}{h_0(t)\exp(\hat{B}_1 x_1 + \cdots \hat{B}_j(x_j) + \cdots \hat{B}_p(x_p)} \\ &= \exp(\hat{B}_j) \end{aligned} \tag{10.8}$$

The one-unit change in x_j will multiply the expected hazard by $\hat{HR}_j = \exp(\hat{B}_j)$. This multiplicative change in the hazard function is termed the **hazard ratio**. If x_j is an indicator variable for a level of a categorical predictor, then $\hat{HR}_j = \exp(\hat{B}_j)$ is the relative hazard compared with the hazard for the reference category used to parameterize the categorical effect.

The procedure for developing a 100(1 – α)% CI for the population hazard ratio parallels that used to build a CI for the odds ratio statistic in simple logistic regression:

$$CI(HR_j) = \exp(\hat{B}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_j)) \tag{10.9}$$

10.4.5 Example: Fitting a Cox Proportional Hazards Model to Complex Sample Survey Data

We now turn to fitting the CPH model to the NCS-R age of major depressive episode onset data. The appropriate data structure for this type of analysis features one row per sampled individual, with that row containing

measures on the survival time, a censoring indicator, covariates of interest, and complex sample design features (stratum and cluster codes and survey weights); this is the common form of most public-use survey data sets arising from cross-sectional samples. Prior to fitting the Cox model to the NCS-R data, we once again declare the variables containing the relevant NCS-R sampling error codes to Stata (note that we use the sampling weight variable NCSRWTSH and request a Taylor series linearization approach [Binder, 1992] for variance estimation):

```
svyset seclustr [pweight=ncsrwtsh], strata(sestrat) ///
vce(linearized) singleunit(missing)
```

Next, we submit the `stset` data preparation command required by Stata for survival analysis. Note that a common `pweight` variable must be specified in the `svyset` and `stset` commands:

```
stset ageonsetmde [pweight = ncsrwtsh], failure(mde==1)
```

After defining the complex design variables and the `stset` coding of the dependent variable AGEONSETMDE, the Cox model is fit using the `svy: stcox` command (note that value 2 [female] for the predictor variable SEX is explicitly set as a reference category first by submitting the `char` command):

```
char sex[omit] 2
xi: svy: stcox intwage i.sex i.mar3cat i.ed4cat i.racecat
```

Note that we use the `xi:` modifier to declare certain predictor variables (SEX, MAR3CAT, ED4CAT, and RACECAT) as categorical, so that Stata includes the appropriate dummy variables in the model. The structure of the `svy: stcox` command is fairly simple and similar to other regression analysis commands for survey data within Stata, provided that the two previous steps (declaring the time-to-event data and the complex design variables) have been followed carefully. What makes the command unique is the fact that a dependent variable (AGEONSETMDE in this case) is *not* specified; the first variable listed in the command, INTWAGE, is a predictor variable in the Cox model. The initial use of the `stset` command directs Stata to set up the appropriate Cox model, and Stata reads the list of variables in the `svy: stcox` command as the predictors to be included in the model.

Submitting this command produces the estimated hazard ratios in Table 10.2.

The design-adjusted *F*-test of the null hypothesis that all of the regression parameters in the Cox model are equal to zero strongly suggests that some of the predictors have a significant impact on the hazard of MDE onset at any given age, and the design-based 95% confidence intervals for the hazard ratios also support this finding. Specifically, Hispanic and black respondents

TABLE 10.2

Estimated Hazard Ratios for the NCS-R Cox Proportional Hazards Model of Age to Onset of Major Depression

Predictor[a]	Estimated Hazard Ratio	Estimated SE	*t*-Statistic (*df*)	*p*-Value	95% CI
Education					
12 Years	0.95	0.063	–0.85 (42)	0.401	(0.825, 1.082)
13–15 Years	1.05	0.061	0.79 (42)	0.436	(0.931, 1.177)
16+ Years	0.91	0.058	–1.42 (42)	0.164	(0.804, 1.039)
Ethnicity					
Hispanic	0.78	0.105	–1.86 (42)	0.070	(0.594, 1.021)
Black	0.62	0.092	–3.22 (42)	0.003	(0.459, 0.837)
White	1.08	0.127	0.66 (42)	0.511	(0.853, 1.370)
Marital Status					
Previously married	1.65	0.099	8.37 (42)	<0.001	(1.464, 1.866)
Never married	1.08	0.096	0.91 (42)	0.369	(0.906, 1.297)
Gender					
Male	0.64	0.040	–7.28 (42)	<0.001	(0.560, 0.720)
Age	0.95	0.002	–20.80 (42)	<0.001	(0.947, 0.956)

Notes: n = 9,282. Design-adjusted F-test of null hypothesis that all parameters are 0: $F(10, 33) = 53.02$, $p < 0.001$.

[a] Reference categories for categorical predictors are <12 yrs (Education); Asian/Other (Ethnicity); Married (Marital Status); Female (Gender).

have marginally ($p < 0.10$) and significantly ($p < 0.01$) reduced hazard of onset of MDE at any given age holding the other covariates fixed, relative to the Asian/Other ethnic group; this agrees with the initial evidence provided by the Kaplan–Meier estimates of the survival functions for the four ethnic groups. Further, with every one-year increase in age at the time of interview (a birth cohort predictor), the hazard of MDE onset is multiplied by 0.95, or decreased by about 5%, holding the other covariates fixed; this suggests that younger individuals at the time of the interview were at higher hazard of MDE onset at any given age than older individuals. Finally, previously married individuals had a hazard of MDE onset that is about 65% higher at any given age compared with currently married individuals at that age, and males have a hazard of MDE onset that was about 36% lower at any given age than females, holding the other covariates fixed.

We now consider the graphical check of the proportional hazards assumptions described in Section 10.4.3. Recall that this graphical check involves a plot of $-\ln(-\ln(\hat{S}_h(t)))$ against $\ln(t)$, where $\hat{S}_h(t)$ is a weighted estimate of the survival function for group h and t represents the time variable. Based on the specified Cox model, the transformed versions of the survival functions should be parallel as a function of $\ln(t)$. Remember that the `stset`

command specification of the dependent variable and the `pweight` value remains in effect. The `stphplot` command is used to generate this special diagnostic plot:

```
stphplot, by(racecat) legend(lab(1 "Other/Asian") ///
lab(2 "Hisp/Mexican") lab(3 "Black") lab(4 "White"))
```

This command generates the plot in Figure 10.4.

From Figure 10.4, we see minimal evidence of the transformed survival curves crossing during the ages when the majority of the data were collected, suggesting that an assumption of proportional hazards would be reasonable for the four race groups. If the transformed survival curves were not approximately parallel (e.g., there is evidence at the highest ages of a possible cross between the White ethnic group and the Hispanic/Mexican and Other/Asian groups), one could fit a model allowing for the effects of ethnicity to vary depending on the age at which a risk set is being formed for estimation. This could be done by including interactions of race with age in this model, which would be time-varying covariates. For more on how to do this when using `svy: stcox`, Stata users can see `help tvc _ note.`

Optionally, Stata users could also plot $\ln(-\ln(\hat{S}_h(t)))$ against t (a variation of the parallel lines check) or could adjust the estimates for mean values of covariates using the following options (the `adj()` option includes dummy variables as covariates, such as SEXF as an indicator for females, that are equivalent to the dummy variables generated by Stata for the model fitting):

```
stphplot, by(racecat) adj(intwage sexf swd nevermarried ///
ed12 ed1315 ed16 ) nonegative nolntime ///
legend(lab(1 "Other/Asian") lab(2 "Hisp/Mexican") ///
lab(3 "Black") lab(4 "White"))
```

The resulting plot in Figure 10.5 shows that when adjusting for the means of the covariates (and using the alternative specification of the plot enabled by the `nonegative` and `nolntime` options), we still do not have convincing evidence against an assumption of proportional hazards.

10.5 Discrete Time Survival Models

Survival models for events measured in discrete time are needed only when events are reported in discrete time units. For a dependent variable representing a discrete survival time, there are $m = 1, \ldots, M$ observed intervals, such as time periods between waves of a longitudinal survey, or time measured in

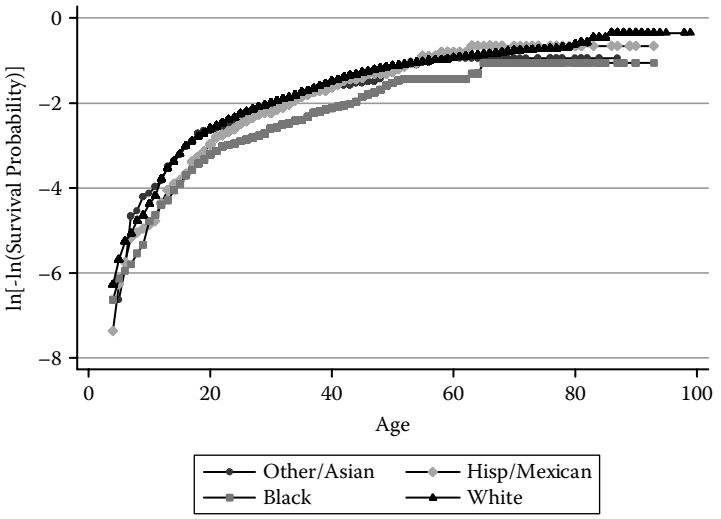


FIGURE 10.5
Testing the proportional hazards assumption for the fitted Cox model, with adjustment for covariates. (Modified from NCS-R data.)

years (e.g., year 1, year 2). In some cases, the "discrete" nature of the survival data may be due to **coarsening** in the dating of events (within the past four weeks, did you…?) or even deliberate grouping of times to provide disclosure protection for individual survey respondents.

10.5.1 The Discrete Time Logistic Model

A popular choice for modeling hazard functions when working with discrete survival outcomes is the **discrete time logit model** (Allison, 1995; Singer and Willett, 1993; Yamaguchi, 1991). This model is defined as follows:

$$\begin{aligned} \ln\left(\frac{h_{i,m}}{1-h_{i,m}}\right) &= B_{0,m} + \boldsymbol{x}_{i,m}\boldsymbol{B} \\ &= B_{0,m} + B_1 x_{1,m} + \cdots + B_p x_{p,m} \end{aligned} \tag{10.10}$$

where $h_{i,m}$ refers to the hazard of failure at discrete time m for respondent i, $B_{0,m}$ is a time-specific intercept term that applies to all individuals at time m, $\mathbf{x}_{i,m}$ is a row vector of values on covariates (possibly time-varying) for respondent i, and $\boldsymbol{B}$ is the vector of regression parameters. It is important to note that Equation 10.10 models the logit of the individual hazard, $h_{i,m}$. To recover the estimated hazard from the estimated logit model requires the inverse logit transformation:

$$\hat{h}_{i,m} = \frac{\exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})}{1+\exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})} \tag{10.11}$$

Once the discrete time logit model has been estimated, estimates of the individual hazards, standard errors and CIs can be readily generated using Stata's `predict` postestimation command.

Allison (1995) describes an alternative discrete time model that is based on the CLL link:

$$\log[-\log(1-h_{i,m})] = \alpha_m + B_1 x_{1,m} + \cdots + B_p x_{p,m} \tag{10.12}$$

The CLL version of the discrete time model is most applicable when the true model for the underlying continuous time process that generates the discrete time measures is the Cox proportional hazard model of Section 10.4. An advantage of the CLL discrete time model is that $\hat{HR}_j = \exp(\hat{B}_j)$. Exponentiating the estimated $\hat{B}_j$ yields an estimate of the hazard ratio corresponding to a one unit change in x_j, which is identical to the interpretation for the CPH model.

The regression parameters from the two discrete time models do have slightly different interpretations—log(odds) for the logit version, log(hazard) for the CLL—however, both models share the following features:

- Identical **person-time** data set formats are required to fit the two models.
- Both models permit direct estimation of the effect of time on the hazard, through the $\hat{B}_{0,m}$ in the logit form and the $\hat{\alpha}_m$ in the CLL model.
- Both models permit the use of time-varying covariates, $\mathbf{x}_{i,m}$.
- For complex sample survey data, pseudo-maximum likelihood estimation is used to estimate the regression parameters in each model (see Section 8.2).

In most cases, the choice of the logit or the CLL form of the discrete time survival model will not affect the survey analyst's interpretation of the significance and nature of the relationship of the model covariates to the hazard function.

10.5.2 Data Preparation for Discrete Time Survival Models

Discrete time survival models require a special data input structure that is labeled a person-time data format. The person-time data input format contains multiple records for each respondent: one record to represent each discrete time interval during which the respondent was observed, *up to and including* the time interval when the event of interest actually occurred or

the observation was censored. A respondent who experiences the event at the third ($m = 3$) discrete time point will have three records in the data set. Another respondent whose observation period is censored at time $m = 5$ will have five person-time data records in the data set.

Each person-time record will contain a value for a binary dependent variable, Y_{im}, taking on value 1 if the event occurred to respondent i at time period m, or value 0 if no event occurred or the respondent was censored at time m. The sequence of person-time records extends from the first discrete time point, $m = 1$, until the time period when an event actually occurred (i.e., until the dependent variable is equal to 1) or until time M (the last possible discrete time period) if no event occurred (all records for this type of censoring will have the dependent variable equal to 0). If a respondent drops out of the sample or leaves the study for some other reason prior to the end of the time period, he or she will have a person-time record for each time point at which he or she was observed plus one additional record for the time point at which he or she was lost. All of the records will have the dependent variable $Y_{im} = 0$.

The discrete time models offer an advantage in that they directly model the effect of time on the hazard of the event. Each person-time record therefore requires one or more variables to indicate the time period. There are several ways to parameterize the effect of time in the discrete time model. The method we recommend is to create a set of indicator variables—one for each time period, $\boldsymbol{I} = \{I_1, \ldots, I_M\}$. Each person-time record will contain all M indicator variables, but only the indicator corresponding to the time period m will have $I_m = 1$. All other time indicators (for times $m' \neq m$) for that record will be set to $I_{m' \neq m} = 0$. When this parameterization is used, the logit or CLL model should be estimated without the overall intercept term. (If the overall intercept is modeled, one time indicator variable should be dropped.) We illustrate the implications of this coding for the specification of a discrete time model in Equation 10.13:

$$\begin{aligned} \ln\left(\frac{h_{i,m}}{1-h_{i,m}}\right) &= \boldsymbol{B}_0 \cdot \boldsymbol{I} + B_1 x_{1,m} + \cdots + B_p x_{p,m} \\ &= B_{0,1} \cdot 0 + \cdots + B_{0,m} \cdot 1 + \cdots + B_{0,M} \cdot 0 + B_1 x_{1,m} + \cdots + B_p x_{p,m} \\ &= B_{0,m} + B_1 x_{1,m} + \cdots + B_p x_{p,m} \end{aligned} \tag{10.13}$$

There are two advantages to this "time indicator" parameterization of the discrete time models. First, no functional relationship (e.g., linear, quadratic) is imposed on the relationship between the hazard and time. Second, if time periods are not of equal length or must be collapsed (if event times are sparse), the individual time period intercepts will adjust for the variability in the size of the time units and the regression parameters for the covariates will not be biased.

An alternative parameterization is to model the sequential discrete time unit indices (m = 1, 2, 3, …, M) as linear or quadratic predictors. For the discrete time logit model:

$$\ln\left(\frac{h_{i,m}}{1-h_{i,m}}\right) = B_0 + \gamma_1 m + \gamma_2 m^2 + B_1 x_{1,m} + \cdots + B_p x_{p,m} \tag{10.14}$$

This is a much more parsimonious parameterization of time and may be preferred if the total number of discrete times is large (say, $M > 15$). (For simplicity of presentation and comparability to the CPH model results in Table 10.2, this parameterization is used in the analysis example in Section 10.5.5). One practical analysis strategy is to first fit the discrete time model using the unconstrained time indicator parameterization (Equation 10.13). If the estimated coefficients show a linear or quadratic trend, the model could be reestimated more parsimoniously using the model in Equation 10.14.

In addition to the binary dependent variable and the indicators representing the time period, each person-time record in the input data set will include the vector of p covariates, $\mathbf{x}_m = \{x_{1,m}, \ldots, x_{p,m}\}$. Individual variables in the covariate vector may be time invariant (e.g., gender) or may be time varying (e.g., body weight in kilograms).

Finally, in survey data sets, an analysis weight and the sampling error stratum and cluster codes are assigned to each person-time record. In most cases, the weight value assigned to each time point will be the baseline weight (m = 1) assigned to the sample cases at the start of the observation period. However, in cases such as longitudinal surveys where there may be substantial panel attrition due to nonresponse over the multiple waves in the observation period, the time t weight for each observed case may provide a better weighted representation of the population risk set at time t. The person-time input structure for discrete time models allows the user this flexibility to update the time-specific weight for prior nonresponse.

The following examples should help to clarify the required person-time data structure. Assume that a sample of individuals is observed for five waves of survey data collection, and a discrete survival time T is recorded for each individual, equal to 1 for events that occur at (or before) Wave 1, equal to 2 for events that occur at (or before) Wave 2, and so forth (T = 1, 2, 3, 4, 5). Table 10.3 illustrates the person-time data inputs for two hypothetical respondents who experience the event of interest: the first at the second wave (m = 2) and the second at the fourth wave (m = 4).

In Table 10.3, note the presence of a time-invariant sampling weight (which may be time varying depending on the design of the sample) and stratum and cluster codes for sampling error estimation (also time invariant). In this hypothetical example, time is represented in the model as a single integer-value covariate, making this input consistent with model parameterization Equation 10.14. If the time-period indicator parameterization of the model in

TABLE 10.3

Appropriate Data Structure for Discrete Time Survival Models for Individuals Experiencing an Event of Interest

ID	Weight	Stratum	Cluster	Time (t)	Y_{im}	X_{im}	Z_i
1	1.56	1	1	1	0	23	1
1	1.56	1	1	2	1	25	1
2	0.82	1	2	1	0	14	0
2	0.82	1	2	2	0	16	0
2	0.82	1	2	3	0	17	0
2	0.82	1	2	4	1	17	0

TABLE 10.4

Appropriate Data Structure for Discrete Time Survival Models for Censored Individuals

ID	Weight	Stratum	Cluster	Time (t)	Y_{im}	X_{im}	Z_i
3	1.25	2	1	1	0	19	1
3	1.25	2	1	2	0	19	1
3	1.25	2	1	3	0	18	1
3	1.25	2	1	4	0	21	1
3	1.25	2	1	5	0	22	1

Equation 10.13 was chosen, each person-time record would include five indicator variables, one for each of the $M = 5$ possible observation times in this hypothetical example. Since the two example cases in Table 10.3 experienced the event of interest, the value of the dependent variable is coded as $Y_{im} = 1$ for the event time, and $Y_{im} = 0$ for all person-year records preceding the event. The X_{im} variable (e.g., body mass index) is a time-varying covariate, and the Z_i variable (e.g., indicator of female gender) is a time-invariant covariate.

Table 10.4 illustrates an example of the person-time data structure for a censored respondent who did not experience the event of interest by the fifth and final wave of data collection. Note in Table 10.4 that the person-time records for this right-censored case include the same input variables as the event cases in Table 10.3; however, the values of the dependent variable including the final time point are coded $Y_{im} = 0$.

10.5.3 Discrete Time Models: Estimation Stage

Mathematically, estimation of the regression parameters for the discrete time logit and CLL models for complex sample survey data is a direct application of the pseudo-maximum likelihood methods described in Section 8.3 for the simple logistic and CLL models for a binary dependent variable. Likewise, procedures for estimation of $Var(\hat{\boldsymbol{B}})_{TSL}$ or alternatively $Var(\hat{\boldsymbol{B}})_{Rep}$ are identical to those for the standard logit and CLL generalized linear models. Provided

THEORY BOX 10.3 THE LACK OF A CLUSTERING EFFECT IN PERSON-TIME DATA

The explanation for absence of a "clustering effect" in the person-time data is that the pseudo likelihood for the complete set of observed event and censoring times can be factored into separate pseudo-likelihoods for the individual event times, $m = 1, \ldots, M$. Consequently, estimation for each component of the factored likelihood can be treated as if its set of person-time records were independent of those used to estimate the hazard probabilities for other discrete time points. See Allison (1982) for a mathematical proof that the full likelihood for the discrete time models can be factored into time-period specific likelihoods.

that the data are structured correctly in the person-time input format previously described, any statistical software procedure capable of fitting logistic or CLL regression models to complex sample survey data can be used to fit these models.

As pointed out by Allison (1995, Chapter 7), some analysts' natural intuition leads them to believe that because the required person-time data structure involves multiple observations for the same person, an extra level of clustering may be introduced in the data that is similar to the clustering that occurs in repeated measures data (see Chapter 12). Fortunately, as explained in Theory Box 10.3, this is actually not a concern in the estimation of sampling variances for the parameters in discrete time models. In modeling event history data, the issue of correlation among the person-time records arises only when working with discrete event time models that permit individuals to have multiple events.

10.5.4 Discrete Time Models: Evaluation and Interpretation

As already described, procedures for confidence interval construction, hypothesis testing, and general model evaluation for the discrete time models follow directly from those described in Chapter 8 for the logit and CLL models for a binary dependent variable:

1. *Discrete time logit model.* The regression coefficients in the discrete time logit model are estimated on the log-odds scale, where the odds in question pertain to the conditional hazard probabilities. Exponentiating the estimated regression coefficients yields the estimated odds ratio: $\hat{\psi}_j = \exp(\hat{B}_j)$. Confidence intervals for these estimated odds ratios are constructed in the now familiar fashion, $CI(\psi_j) = \exp(\hat{B}_j \pm t_{df,1-\alpha/2} \cdot se(\hat{B}_j))$. Given pseudo-maximum likelihood

estimates of the regression parameters and a vector of covariates, $\boldsymbol{x}_{i,m}$, the estimated hazard at time m can be computed as follows:

$$\hat{h}_{i,m} = \frac{\exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})}{1 + \exp(\hat{B}_{0,m} + \boldsymbol{x}_{i,m}\hat{\boldsymbol{B}})} \tag{10.15}$$

Note that this is mathematically identical to the computation of a predicted probability based on a fitted logistic regression model.

2. *Discrete time complementary log–log model.* An advantage of the CLL model for discrete time data is that population inferences and discussion of results use the hazard ratio statistic $\hat{HR}_j = \exp(\hat{B}_j)$. Procedures for constructing confidence intervals and interpreting the hazard ratio statistics generated from the CLL model for discrete time data are identical to those presented for the Cox proportional hazards model in Section 10.4.4.

With this background, we now consider fitting the discrete time logit and discrete time CLL models to the NCS-R data on age of onset for major depression.

10.5.5 Fitting a Discrete Time Model to Complex Sample Survey Data

An important first step in fitting the discrete time hazard model is to transform the NCS-R input data into the appropriate person-time format. The following Stata commands "expand" the NCS-R data set from a one-record-per-person file to a multiple-record-per-person file, with records from year = 1 to year at age of interview (INTWAGE). The use of the following "`gen pyr = _n`" command creates a new variable called PYR, which represents person years of life ranging from 1 to year at age of interview (INTWAGE):

```
expand intwage
sort caseid
gen pyr = _n
```

The NCS-R data set now includes one "person-year" time record for each year of life for a given individual (e.g., 45 rows for a 45-year-old). Next, the indicator variable MDETV is generated for each person-year record. The variable MDETV = 1 if the value of time (PYR) corresponding to the person-year record is equal to the age of onset of MDE, and MDETV = 0 otherwise:

```
gen mdetv = 1 if pyr == mde_ond
replace mdetv = 0 if pyr != mde_ond
```

The variable MDETV will be the dependent variable when fitting the logit and CLL regression models to these data. The following example Stata list

output for the *first respondent* in the NCS-R data set shows the person-year data for this discrete time modeling example:

```
list caseid intwage ncsrwtsh sestrat seclustr pyr mdetv ///
ageonsetmde if caseid == 1
```

	caseid	intwage	ncsrwtsh	sestrat	seclustr	pyr	mdetv	ageonsetmde
1.	1	41	2.02426	1	2	1	0	34
2.	1	41	2.02426	1	2	2	0	34
3.	1	41	2.02426	1	2	3	0	34
4.	1	41	2.02426	1	2	4	0	34
5.	1	41	2.02426	1	2	5	0	34
6.	1	41	2.02426	1	2	6	0	34
7.	1	41	2.02426	1	2	7	0	34
8.	1	41	2.02426	1	2	8	0	34
9.	1	41	2.02426	1	2	9	0	34
10.	1	41	2.02426	1	2	10	0	34
11.	1	41	2.02426	1	2	11	0	34
12.	1	41	2.02426	1	2	12	0	34
13.	1	41	2.02426	1	2	13	0	34
14.	1	41	2.02426	1	2	14	0	34
15.	1	41	2.02426	1	2	15	0	34
16.	1	41	2.02426	1	2	16	0	34
17.	1	41	2.02426	1	2	17	0	34
18.	1	41	2.02426	1	2	18	0	34
19.	1	41	2.02426	1	2	19	0	34
20.	1	41	2.02426	1	2	20	0	34
21.	1	41	2.02426	1	2	21	0	34
22.	1	41	2.02426	1	2	22	0	34
23.	1	41	2.02426	1	2	23	0	34
24.	1	41	2.02426	1	2	24	0	34
25.	1	41	2.02426	1	2	25	0	34
26.	1	41	2.02426	1	2	26	0	34
27.	1	41	2.02426	1	2	27	0	34
28.	1	41	2.02426	1	2	28	0	34
29.	1	41	2.02426	1	2	29	0	34
30.	1	41	2.02426	1	2	30	0	34
31.	1	41	2.02426	1	2	31	0	34
32.	1	41	2.02426	1	2	32	0	34
33.	1	41	2.02426	1	2	33	0	34
34.	1	41	2.02426	1	2	34	1	34

Note that the sampling error stratum and cluster codes and sampling weight values are included on each person-year record for the respondent.

This respondent was 41 years old when interviewed and first had an MDE at age 34. Although there are more data records for this individual (aged 35–41), only the records from ages 1 to 34 (the event time) should be analyzed when we fit the model.

Next, we fit the standard design-based logit regression model (Chapter 8) to the person-year data. The `svy: logit` model command assumes that the complex design features of the NCS-R data set have already been declared to Stata (see the `svyset` command in Section 10.4.5). However, unlike the Cox proportional hazards model, Stata's `stset` command is *not* required. The following predictors of the event are included in the discrete time model: PYR (person-year), INTWAGE (age at interview), SEX, ED4CAT, RACECAT, and MAR3CAT. We note that including INTWAGE in the model serves to introduce a cohort effect in the model; in other words, do persons of different ages at the time of the interview have different hazards?

```
char sex[omit] 2
xi: svy: logistic mdetv pyr intwage i.sex i.ed4cat ///
i.racecat i.mar3cat if pyr <= ageonsetmde
```

The use of the `if` modifier is especially important in the model commands; we want to analyze only those person-years for times equal to or less than the age of MDE onset (or age of interview for right-censored individuals never having experienced an MDE). Fitting the logit version of the model results in the estimates of odds ratios presented in Table 10.5.

Focusing on the results for the discrete time logit model, we see that the estimated odds ratios in this discrete time hazard model are remarkably similar to the estimated hazard ratios found for the Cox proportional hazards model fitted in Section 10.4.5, which suggests that the two approaches result in similar estimates of the impacts of these sociodemographic predictors on the risk of MDE onset at any given age. For example, the discrete time hazard model suggests that holding all other covariates fixed (including person-year, or the age at which a person might be at risk), males have 36% lower odds of having an MDE for the first time than females.

Given that the discrete time modeling approach can be implemented using standard design-based logistic regression modeling techniques, any of the diagnostics discussed in Chapter 8 could be applied when evaluating the fit of this model.

Finally, the comparable discrete time complementary log–log model is fit to the person-time data using Stata's `svy: cloglog` command:

```
xi: svy: cloglog mdetv pyr intwage i.sex i.ed4cat ///
i.racecat i.mar3cat if pyr <= ageonsetmde, eform
```

Table 10.6 presents the results for the CLL model in the form of estimates and confidence intervals for the hazard ratios (note the use of the `eform` option in the previous command).

TABLE 10.5

Estimated Odds Ratios for the NCS-R Discrete Time Logit Model of Age to Onset of Major Depression

Predictor[a]	Estimated Odds Ratio	95% CI
Education		
12 Years	0.98	(0.858, 1.120)
13–15 Years	1.11	(0.977, 1.232)
16+ Years	0.98	(0.863, 1.114)
Ethnicity		
Hispanic	0.78	(0.594, 1.024)
Black	0.63	(0.468, 0.857)
White	1.08	(0.848, 1.367)
Marital Status		
Previously married	1.64	(1.449, 1.854)
Never married	0.97	(0.808, 1.153)
Gender		
Male	0.64	(0.566, 0.727)
Person-year (time)	1.03	(1.029, 1.038)
Interview age	0.94	(0.939, 0.948)

Notes: Design-adjusted F-test of null hypothesis that all parameters are 0: $F(11, 32) = 53.63$, $p < 0.001$.

[a] Reference categories for categorical predictors are <12 yrs (Education); Asian/Other (Ethnicity); Married (Marital Status); Female (Gender).

Compared with the results in Table 10.4 for the Cox proportional hazards model (hazard ratios) and the discrete time logit model results in Table 10.5 (odds ratios), we again see a similar pattern of results in the estimated hazard ratios for the CLL discrete time model for age of onset of major depression. Controlling for the other predictors and a linear time effect, education level does not appear to have a significant impact on the hazard of MDE. Black respondents have a significantly lower hazard for MDE relative to persons of Asian and other race/ethnicities. Controlling for the other predictors in the model, previously married persons have a greater hazard of MDE than married adults, and the estimated hazard for men is about 64% of that for women (all else being equal).

TABLE 10.6

Estimated Hazard Ratios for the NCS-R Discrete Time C-L-L Model of Age to Onset of Major Depression

Predictor[a]	Estimated Hazard Ratio	95% CI
Education		
12 Years	0.98	(0.858, 1.119)
13–15 Years	1.10	(0.977, 1.231)
16+ Years	0.98	(0.863, 1.114)
Ethnicity		
Hispanic	0.78	(0.595, 1.024)
Black	0.63	(0.469, 0.858)
White	1.08	(0.849, 1.366)
Marital Status		
Previously married	1.64	(1.448, 1.851)
Never married	0.97	(0.809, 1.152)
Gender		
Male	0.64	(0.567, 0.727)
Person-year (time)	1.03	(1.029, 1.038)
Interview age	0.94	(0.939, 0.948)

Notes: Design-adjusted F-test of null hypothesis that all parameters are 0: $F(11, 32) = 53.65$, $p < 0.001$.

[a] Reference categories for categorical predictors are <12 yrs (Education); Asian/Other (Ethnicity); Married (Marital Status); Female (Gender).

10.6 Exercises

1. This exercise considers the data from the NCS-R. The objective of this analysis is to model the age-specific hazard of first onset of general anxiety disorder (GAD); in other words, have different age cohorts in the NCS-R population had different hazards of onset of general anxiety disorder as a function of age? Begin the analysis by generating an age of first onset of GAD variable, equal to age of onset of GAD (GAD_OND) if DSM_GAD is equal to 1 (or Yes) or equal to age at interview (INTWAGE) if DSM_GAD equals 5 (or No) and GAD_OND equals missing. Note that cases not qualifying for GAD are *right-censored* cases in terms of age of GAD onset, so you should also compute an indicator variable for the censored cases. Present the code that you used to compute these variables, which will serve

as the "time-to-event" variable and the right-censoring variable in this analysis.

2. Generate weighted (NCS-R Part 1 sampling weight = NCSRWTSH) Kaplan–Meier estimates of the survival curves for the four age groups defined by the categorical age variable AG4CAT (1 = age 18–29, 2 = 30–44, 3 = 45–59, 4 = 60+). What do the survival curves suggest about the hazards of initial GAD diagnosis as a function of age for these four age cohorts?
3. Fit a Cox proportional hazards model to the age of onset of GAD, recognizing the right censoring and the complex sample design features of the NCS-R (stratum codes = SESTRAT, cluster codes = SECLUSTR, NCS-R part 1 sampling weight = NCSRWTSH). Consider as predictors of the hazard function the categorical age variable (AG4CAT) and the indicator variable for females (SEXF). Based on the weighted estimates of the regression parameters in this model and their estimated standard errors, is there evidence of significant differences between the age cohorts in terms of the hazard functions? What are the differences? Compute the estimated hazard ratio of initial GAD onset for females compared with males, holding age cohort constant.
4. **(Stata Only)** Test the assumption of proportional hazards for the four age groups using the graphical methods discussed in this chapter. Does this assumption seem justified? What would be an alternative analytic approach if this assumption seems violated?
5. Using the data management methods discussed in this chapter, construct an expanded version of the NCS-R data set that could be used to fit a discrete time logit model to the age of onset of GAD. Generate a table showing the data structure for the first two respondents in the NCS-R data set.
6. Fit a discrete time logit model to the expanded data set considering the same predictors of first onset of GAD used to fit the Cox model (AG4CAT and SEXF). Make sure to recognize the complex design features when fitting the model, and don't forget to include age as a predictor in the model (see Section 10.5.5). Based on the design-adjusted tests of significance for the parameters in this model, would we make the same inferences about the differences in the hazards for the four age groups and for females compared with males? Generate a table comparing the estimated odds ratios from the discrete time logit approach to the estimated hazard ratios from the Cox modeling approach.

11

Multiple Imputation: Methods and Applications for Survey Analysts

11.1 Introduction

Missing data is a ubiquitous problem in the analysis of survey data. Section 2.7.3 introduced the concept of unit nonresponse and weighting adjustments to compensate for potential bias due to completely missing data for significant fractions (e.g., 10%, 20%, 30%) of the probability sample that was selected to represent the survey population (see Figure 11.1). The example analyses of the National Comorbidity Survey Replication (NCS-R), 2006 Health and Retirement Survey (HRS), and 2005–2006 National Health and Nutrition Examination Survey (NHANES) data sets presented in Chapters 5 through 10 all employed weights that adjusted for unit nonresponse on the part of original sample.

Throughout the preceding discussion of statistical methods and examples in Chapters 5 through 10, the problem of **item-missing data** for otherwise complete cases was allowed to remain in the background. As a case in point, recall the example analysis in Section 9.4 in which a negative binomial regression model was fit to the HRS data for falls. In addition to the dependent variable (count of falls), the model included six predictor variables. Table 11.1 summarizes the item-missing data rates for the count of falls and the six predictors in that model. The interviewed respondents in the 2006 HRS panel included $n = 11{,}731$ adults age 65+ who were eligible to be asked this question. After list-wise deletion, Stata used only $n = 11{,}197$ complete data observations to estimate the regression model parameters and the standard errors of the parameter estimates—a 4.6% reduction in the nominal sample size. The example analysis incorporated the 2006 HRS weighting adjustments for baseline sample unit nonresponse and panel attrition but did not attempt to compensate for the otherwise complete cases that were lost in the analysis due to the presence of item-missing data on one or more of the analysis variables.

The overall aim of this chapter is to bring the problem of item-missing data forward and focus on it, examine its nature, assess its potential impact on estimation and inference, and offer practical solutions for dealing with "item

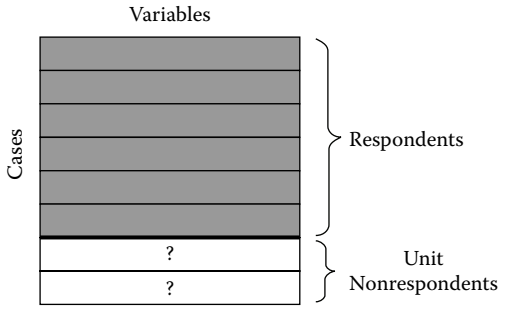


FIGURE 11.1
Unit nonresponse.

TABLE 11.1

Item-Missing Data Rates for Variables Included in the 2006 HRS Falls Model

Variable	Falls	Age	Gender	Arthritis	Diabetes	Weight	Height
%Missing	4.5%	0.0%	0.0%	0.2%	0.1%	1.4%	1.4%

Note: n = 11,731 eligible respondents aged 65 and older.

missingness" in the course of survey data analysis. The chapter begins with an overview of important missing data concepts in Section 11.2. Section 11.3 then provides an overview of the multiple imputation approach to analysis with missing data, outlining the general framework of model specification, imputation, and estimation and inference. This is followed in Section 11.4 by an introduction to the imputation model and its role in multiple imputation analyses of survey data. Approaches and software available for generating the imputations under the selected imputation model are described in Section 11.5. Methods for estimation and inference based on multiply imputed data are covered in Section 11.6. The chapter concludes in Section 11.7 with applications of the Stata `ice` command for multiple imputation and the `mim` modifier for multiple imputation estimation to 2005–2006 NHANES data on the diastolic blood pressure of U.S adults.

11.2 Important Missing Data Concepts

11.2.1 Sources and Patterns of Item-Missing Data in Surveys

Respondents' participation in most surveys is a voluntary decision. To ensure high levels of overall survey cooperation and to guarantee that basic human subjects protections are met, respondents are often read a statement to the

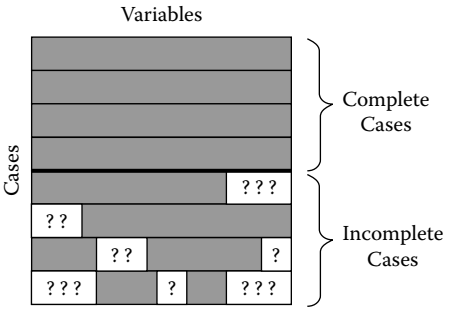


FIGURE 11.2
Generalized pattern of missing data.

following effect at the start of the interview: "Your participation is voluntary. If we come to a question that you do not wish to answer, you may skip it." Categories of "Don't Know" or "Refused" are often explicitly included in the response options for individual survey items. Item-missing data for individual variables can also occur due to mistakes in interview administration, although in today's world, computer-assisted interviewing (CAI) minimizes many of the item-missing data problems associated with incorrect skips or missed-item follow-ups. Many variables are subject to low rates of item-missing data, and possibly a few more sensitive or difficult items (e.g., income measures) have higher missing data rates ranging from 5% to 15%. The arbitrary nature of these response outcomes typically produces a **generalized pattern** of item-missing data in which there is no particular hierarchical or monotonic trend in the missing data structure (see Figure 11.2).

Item-missing data can also occur when a **phase** of a survey data collection activity, such as a blood draw or medical records follow-up, may require special consent of the subject. Failure to obtain cooperation can lead to missing data for variables from that entire phase of the study. For example, approximately 4.1% of adults who participated in the initial 2005–2006 NHANES interview did not agree to participate in a subsequent medical mobile examination center (MEC) assessment. HRS requests special consent of respondents to access U.S. Social Security earnings records and Medicare records (for panel members age 65+). At any given wave, approximately 20% of respondents do not provide consent to this special records linkage. Nonresponse to complete phases of a multiphase survey produces a **monotonic pattern** of item-missing data (Figure 11.3)—core data are present for interviewed cases, but data for subsequent phases may be missing. Although imputation methods may be used to address monotonic patterns of item-missing data including complete phase nonresponse (Schafer et al., 1996), it is more common in survey practice to see nonresponse weighting adjustments applied to repair this problem.

Survey designers may also deliberately decide to use randomized procedures to permit item-missing data on selected variables for subsets of respondents. The technique of **matrix sampling** or "missing by design" sampling

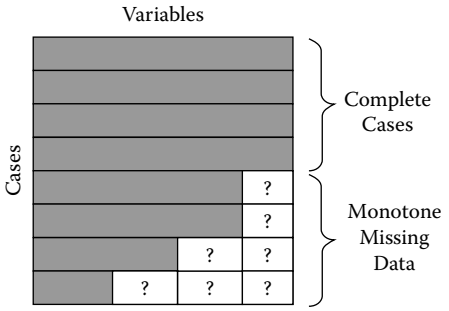


FIGURE 11.3
Monotonic pattern of missing data.

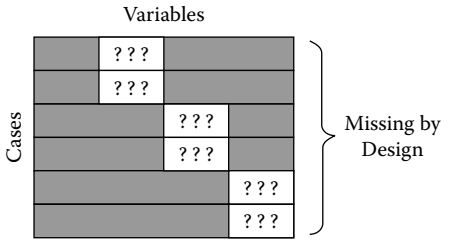


FIGURE 11.4
Matrix sampling (missing by design).

(Thomas et al., 2006) is often employed with modularized sets of survey questions. A battery of core questions is asked of all respondents and more in-depth modularized questionnaire components are randomly assigned to subsamples of survey participants. Matrix sampling designs tend to produce a nonmonotonic missing data structure such as that illustrated in Figure 11.4, and the technique of **multiple imputation** (to be discussed later in this chapter) has typically been used to analyze these data (see Raghunathan and Grizzle, 1995).

11.2.2 Item-Missing Data Mechanisms

In describing the multiple imputation model and techniques, a special notation introduced by Little and Rubin (2002) will be employed. The true underlying values for a vector of $j = 1, \ldots, p$ survey variables will be labeled $\boldsymbol{Y} = \{Y_1, \ldots, Y_p\}$. This underlying set of true values of the variables of interest is decomposed into two subsets of values, $\boldsymbol{Y} = \{\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}\}$, where $\boldsymbol{Y}_{obs}$ are the values that are observed and $\boldsymbol{Y}_{miss}$ are the values that are not observed and replaced by item-missing values (e.g., a "." code in SAS or Stata).

A **missing data mechanism** defines the distribution of the missing data given the underlying data and can be thought of as a probability model for

a response indicator. In other words, why are the missing data arising? Are reasons for missing data related to the study variables in any way? Survey analysts need to have a basic understanding of missing data mechanisms to evaluate the models and methods that are applied in compensating for missing data.

There are generally three recognized missing data mechanisms (Little and Rubin, 2002). Theory Box 11.1 provides statistical definitions for the various missing data mechanisms. More generally, data are defined to be **missing completely at random** if the probability that a respondent does not report an item value is completely independent of the true underlying values of all of the observed and unobserved variables. In other words, cases with missing data represent the equivalent of a simple random subsample of the full sample. Consider a survey measuring the weights and heights of individuals. In the MCAR case, the probability of having a missing value on height has nothing to do with the actual height, the actual weight, or any other possible covariates that might have been measured.

Data are **missing at random** if missingness depends only on the observed values of the variables in the survey. In other words, the probability of a value being recorded as missing depends on *observed values of other variables* but not on the missing values themselves. Following the same previous example, the probability of having a missing value on height might depend on the observed value of weight but *not* on the unobserved value of height. Finally, data are **not missing at random** if missingness depends on the unobserved values of the variables with missing data (in addition to the observed values) in the survey data set. In other words, the probability of having a missing value on height does depend on the actual height and possibly also depends on the weight. Data that are NMAR are difficult to handle analytically.

As described in Section 11.2.3, complete case analysis (list-wise deletion) requires the assumption that the missing data are MCAR. From the standpoint of being able to effectively and practically address item-missing data in a survey data set, a MAR mechanism is the more reasonable assumption. For example, the predictive distribution used to draw imputed (or predicted) values for $\mathbf{Y}_{miss}$ may be a regression model in which the predictors are selected from $\mathbf{Y}_{obs.}$ Most multiple imputation software assumes that the missing data mechanism is in fact MAR.

Certainly, the MAR assumption may not strictly apply for all missing items. It is reasonable to expect that the probability of item-missing data on survey variables such as personal income may depend, at least in part, on the underlying value. In applying imputation methods that assume MAR, the resulting imputations may not completely compensate for an NMAR missing data mechanism. Unfortunately, missing data problems where the mechanism is strongly NMAR may not be easily addressed in a simple and straightforward manner. Techniques for analyzing data subject to nonignorable missingness such as selection models or **pattern mixture models** have

THEORY BOX 11.1 STATISTICAL DEFINITIONS FOR MISSING DATA PATTERNS AND MECHANISMS

Let $\boldsymbol{Y}$ = a data matrix, *if no data were missing* (i.e., the matrix of true, known values, where there are no missing data). This is illustrated as follows for a simple data set of $i = 1, \ldots, 4$ cases and $j = 1, \ldots, 4$ variables. Note in the illustration that values (y_{12}, y_{23}, y_{31}, y_{44}) are boldfaced and italicized to signify that these true underlying values will not be observed in the survey:

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & \boldsymbol{y_{12}} & y_{13} & y_{14} \\ y_{21} & y_{22} & \boldsymbol{y_{23}} & y_{24} \\ \boldsymbol{y_{31}} & y_{32} & y_{33} & y_{34} \\ y_{41} & y_{42} & y_{43} & \boldsymbol{y_{44}} \end{bmatrix}$$

Using the terminology of Little and Rubin (2002), the missing observations are labeled $\boldsymbol{Y}_{miss}$. The remaining observed values of $\boldsymbol{Y}$ are noted as $\boldsymbol{Y}_{obs}$. Readers should note that the "data" matrix $\boldsymbol{Y}$ can include design variables, including stratum and cluster codes, which generally do not have any missing data.

Let $\boldsymbol{M}$ = a missing data indicator matrix of the same dimensions as $\boldsymbol{Y}$, where the value in row i and column j is equal to 1 if the corresponding value in $\boldsymbol{Y}$ is recorded as missing in an actual survey, and 0 if the value is observed. Returning to the simple example of four cases and four variables, the matrix of missing data indicators would look like this:

$$\boldsymbol{M} = \begin{bmatrix} 0 & \boldsymbol{1} & 0 & 0 \\ 0 & 0 & \boldsymbol{1} & 0 \\ \boldsymbol{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{1} \end{bmatrix}$$

A **missing data pattern** describes the distribution of the values in $\boldsymbol{M}$, while a **missing data mechanism** describes the conditional distribution of the values in $\boldsymbol{M}$ *given* the values in $\boldsymbol{Y}$.

Mathematically, the **missing completely at random (MCAR) mechanism** can be written as $P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M})$ for all $\boldsymbol{Y}$; that is, the distribution of the missing data does not depend on the values in $\boldsymbol{Y}$ in any way. This is seldom a reasonable missing data mechanism to assume in practice. The **missing at random (MAR) mechanism** implies that $P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}|\boldsymbol{Y}_{obs})$ for all $\boldsymbol{Y}$; that is, the distribution of item-missing data depends on the observed components of $\boldsymbol{Y}$. Finally, a **nonignorable missing data**

mechanism (also referred to as a **not missing at random mechanism**, or **NMAR**) implies that $P(\boldsymbol{M}|\boldsymbol{Y}) = P(\boldsymbol{M}|\boldsymbol{Y}_{obs},\boldsymbol{Y}_{miss})$; that is, missingness depends on both the observed components of $\boldsymbol{Y}$ and the values of $\boldsymbol{Y}$ that were not observed.

been proposed (Little and Rubin, 2002), but they require either independent assessment of the missing data model or interpretation based on a sensitivity analysis conducted over a set of reasonable models postulated for the NMAR mechanism.

11.2.3 Implications of Item-Missing Data for Survey Data Analysis

The default procedure for handling missing data in statistical software is **list-wise deletion—**any case with missing data for one or more required variables is dropped from the analytic computations. Even when rates of missing data for individual variables are low, the combined loss of cases due to list-wise deletion can be substantial. With list-wise deletion, the number of cases included in the analysis will vary depending on the variables chosen for the analysis and their item-missing data rates. Simply adding or deleting a single predictor can produce substantial changes in the number of cases that are included in the analysis.

The analysis implications of item-missing data are therefore both practical and statistical. From a practical standpoint, list-wise deletion permits the size and composition of the analysis sample to vary depending on the variables that are included in the analysis, making it difficult to maintain a standardized set of inputs across a range of analytic methods. Statistically, list-wise deletion of cases due to item-missing data reduces effective sample size and precision regardless of the missing data mechanism. If the missing data are MAR, simple list-wise deletion of cases can result in biased analysis for the complete cases (because the probability of missing data is a function of other observed variables that could also be correlated with the variables of interest). Unbiased analyses are possible only with list-wise deletion if the cases with missing data represent a simple random sample of the full sample (the MCAR assumption).

Loss of precision and the potential for bias are obvious threats to any analysis of survey data, but how serious is the threat to the quality of inferences obtained if item-missing data are ignored? Schafer et al. (1996) explored the option of imputing all missing data in the MEC phase of NHANES III—both complete unit nonresponse and item-missing data for otherwise complete MEC cases. Jans, Heeringa, and Charest (2008) closely replicated this analysis using the 2003–2004 NHANES data. In both cases, their investigations found only very small and scientifically nonsignificant differences between MEC descriptive estimates that used only complete

cases and those based on a completed data set where missing items were multiply imputed.

While these two empirical studies found little impact of a full imputation of item-missing data on substantive analyses of the NHANES MEC data, we should be careful not to generalize these findings. It is true that relative to unit nonresponse where 20%, 30%, or even more of the observations are missing from the survey data, item-missing data rates of 1%, 5%, or even 10% are not likely to produce major biases for survey estimates based on only the complete cases. As already shown, list-wise deletion of missing data cases from a complex multivariate analysis can result in significant loss of statistical information. A properly conducted multiple imputation analysis can recover the "information" contained in those cases with one or more missing items and yield an analysis that maximizes the use of all of the observed data.

11.2.4 Review of Strategies to Address Item-Missing Data in Surveys

Before turning to the subject of multiple imputation, it is useful to review the options that are available to survey analysts faced with the problem of item-missing data. In each of these options, the data would be analyzed using the assigned survey analysis weights including any adjustments that have been incorporated into these weights to compensate for **unit nonresponse** (all studies), **longitudinal attrition** (e.g., HRS 2006), **phase nonresponse** (e.g., 2005–2006 NHANES MEC), or **missing by design** (e.g., NCS-R Part 2 subsampling). Design-based analysis procedures that incorporate design effects due to stratification and clustering are also assumed:

- *Option 1: Complete case analysis.* The simplest approach to handling item-missing data is to do nothing and analyze only the complete cases, assuming that the missing data are MCAR. Software that defaults to list-wise deletion will automatically strike any case with item-missing data from the analysis. Practically, if the individual and aggregate rates of item-missing data for variables included in the analysis are very low (say, <2%), analysis based on only complete cases should be acceptable. In analyses where missing data rates are low, the list-wise deletion of cases with item-missing data may result in small losses of sample precision, irregular counts of case inputs over different analyses, and potentially small biases if the missing data mechanism is MAR or nonignorable. Of course, if the aggregate rate of missing data or the individual rate for a single key variable (e.g., household income, diastolic blood pressure) is higher (say, > 5% to 10%), the precision losses and instability in case counts for differing analyses and the potential for bias under MAR will rise. In the latter case, it is probably worth an investigation of how sensitive the results of the analysis are to different treatments of the item-missing data. One practical plan of action when missing

data rates rise to moderate or high levels is to conduct the exploratory analyses on complete cases. Once preliminary analyses are in hand, the analysis could be replicated using the multiple imputation approach to handling item-missing data that is described in Sections 11.3 through 11.6.

- *Option 2: Analyze complete cases, but introduce additional weighting adjustments to compensate for item-missing data on a key variable.* The use of weighting to compensate for missing data is generally limited to monotonic missing data patterns in which large numbers of variables are associated with each "step" in the pattern. As described in Section 11.2.1, unit nonresponse, phase nonresponse, and longitudinal attrition in panel surveys produce missing data patterns where a weighting adjustment is the practical choice. There is no theoretical barrier to employing a reweighting approach to compensate for item-missing data on key survey variables. Practically, though, attempting to adjust the base weight variables to address item-missing data on single variables leads to difficulties. If the pattern of missing data is general, different weighting adjustment factors would be needed for each target variable with missing data. Analytically, the weight-by-variable approach would work for univariate analyses, but which of the variable-specific weights would be chosen for a multivariate analysis? Simply put, imputation is the better strategy for addressing generalized patterns of item-missing data.
- *Option 3: Perform a single imputation of missing values, creating a "complete" data set.* Public-use survey data sets are often released with a single imputed value replacing missing data on key survey variables. Ideally, the data producers who performed the imputation also included a companion imputation flag variable that distinguishes actual from imputed values on the data set (see Section 4.4). Data users may also choose to perform their own single imputations using an established stochastic imputation method such as the hot deck (see Theory Box 11.3), regression imputation, or predictive mean matching (Little and Rubin, 2002). Use of deterministic procedures such as single mean, median, or modal value imputation are not encouraged unless the imputation is simply serving to fill in a small handful of missing values of an otherwise nearly complete survey variable. SAS, Stata, and other software systems include programs that allow data users to perform imputations using these techniques. The advantage to a singly imputed data set is that it is "complete," with missing values replaced by analyzable data entries. Provided that the imputation technique is multivariate and retains the stochastic properties in the observed data, a single imputation may address potential bias for a MAR missing data mechanism. The principal shortcoming in the standard design-based analysis of a singly-imputed data

set is that it precludes estimation and inference that fully reflects the variance attributable to the item-missing data imputations. Here again, if rates of item-missing data for individual variables and in the aggregate are low, the resulting slight underestimation of variance may not be a true problem. Rao and Shao (1992) have proposed a technique for estimating variances from singly imputed data sets, but their method has not been widely incorporated into survey data analysis programs in the major software systems.

- *Option 4: Perform multiple imputations of missing values, and use multiple imputation procedures for estimation and inference.* A fourth alternative to address the problems of generalized patterns of missing data in survey analysis is to conduct the analysis in the multiple imputation (MI) framework. The remaining sections in this chapter describe the method of multiple imputation, given its flexibility and ease of application in missing data problems arising from complex sample survey data sets.
- *Option 5: Use full information maximum likelihood (FIML) methods.* This alternative approach generally involves the use of all available information in a given survey data set (including values from cases with incomplete data) and applies maximum likelihood estimation techniques given all of the observed information to compute parameter estimates of interest. The FIML technique assumes that the missing data are MAR and is generally used when fitting mixed-effects models to longitudinal data sets that may be unbalanced in nature due to attrition and wave nonresponse. Applications of the technique for complex sample survey data are less common, primarily due to the lack of a clear methodology for incorporating the sampling weights into the maximum likelihood analysis and the corresponding inferential techniques. Readers should note that FIML is a model-based analytic procedure that deviates from the design-based procedures that have been the primary focus of this book, requiring the analyst to specify models appropriately incorporating complex design features. We thus focus on multiple imputation techniques in this chapter, given that they can easily accommodate design-based analysis techniques for complex sample survey data.
- *Option 6: Use the expectation maximization (EM) algorithm.* This approach to handling missing data, which is described in extensive detail in Little and Rubin (2002), is an iterative procedure defined by a two-step algorithm. The algorithm is initiated by a model of interest that the analyst wishes to fit to a data set that might include missing values on the variables of interest. First, in the expectation (E) step, the EM algorithm computes the expected value of the complete data log-likelihood function (given the model) based on the

cases with *complete* data and the algorithm's estimates of the sufficient statistics for the missing data (given the model and the available data points for the incomplete cases). In the initial E step, no values are imputed for the missing data points. Then, in the maximization (M) step, the algorithm substitutes the expected values for the missing data obtained from the E step and maximizes the likelihood function *as if no data were missing* to obtain new parameter estimates. The new parameter estimates are substituted back into the E step, and a new M step is performed. The procedure iterates through these two steps until some prespecified convergence criteria for the parameter estimates are satisfied (i.e., there is little change in the estimates at subsequent steps). This approach is currently employed by the Missing Values Analysis (MVA) module in the SPSS software but is limited in that standard errors and associated test statistics (e.g., *t*-tests) do not reflect the uncertainty in estimating the missing values. Because of this limitation, multiple imputation methods are generally considered more appropriate for survey data.

11.3 An Introduction to Imputation and the Multiple Imputation Method

11.3.1 A Brief History of Imputation Procedures

Statistical imputation procedures are techniques for assigning analyzable values to item-missing data on survey variables. Prior to the early 1970s, it was rare to see applications of imputation to item-missing data. Further, when imputations methods were applied, the techniques were very ad hoc, and little attention was paid to the theoretical implications of the method for analysis and inference. The U.S. National Academy of Sciences Panel on Incomplete Data (Madow and Olkin, 1983) brought together leading survey statisticians to consider item-missing data as a true statistical problem and to lay the foundation for a more theoretically justified treatment of missing data in survey analysis. Since that time, the statistical science of imputation and other methods of analysis that address the missing data problem have flourished. A taste of the variety of methods that have been proposed can be found in the colourful names of the techniques: mean imputation, predictive mean matching, nearest neighbor, regression imputation, the hot deck, row and column method, and multiple imputation. Kalton and Kasprzyk (1986) and Little and Rubin (2002) review the wide range of techniques, the methods and applications, and the relative strengths and the weaknesses of the techniques.

11.3.2 Why the Multiple Imputation Method?

There is no single imputation method or statistical modeling technique that is optimal for all forms of item-missing data problems. As described in Section 11.2.4, survey analysts have a range of options to consider depending on the missing data pattern, the missing data mechanism, and rates of item missing values. Imputation techniques ranging from the simplest univariate mean substitution to a full application of multiple imputation analysis share some basic attributes—for example, all imputation techniques produce a *completed* data set that can then be analyzed using standard software procedures for the analysis of complex sample survey data. However, in cases where the missing data problem is nontrivial, the goals of the imputation process should be more ambitious. The following is a list of those goals and a description of how each is met by the method of multiple imputation:

- *Goal 1: Model-based.* Imputations should be model-based to ensure the statistical transparency and integrity of the imputation process. A process of arbitrary assignment of imputations, even if based on expert judgment, is not replicable in the scientific sense and may not preserve the statistical properties (means, variances, covariances) of the survey data. The imputation model should be broader than the analysis models that will be analyzed using the imputed data (see Figure 11.6 later in this chapter). The model that underlies the imputation process is often an explicit distributional model, but good results may also be obtained using techniques where the imputation model is implicit (see Theory Box 11.2).
- *Goal 2: Stochastic.* The imputation procedure should be stochastic—based on random draws of the model parameters and error terms from the predictive distribution of Y_{miss}. For example, in linear regression imputation of the missing values of a continuous variable y_k, the predictive function used to impute missing values based on the conditional predictive distribution for y_k may be written as $\hat{y}_{k,miss} = \hat{B}_0 + \hat{B}_{j \neq k} \cdot y_{j \neq k} + e_i$. In forming the imputed values of $y_{k,miss}$, the individual predictions should incorporate random draws of the coefficients and independent draws of the errors (e_i) from their respective estimated distributions. In a hot deck imputation procedure or procedures such as predictive mean or propensity score matching imputation, the donor value for $y_{k,miss}$ is drawn at random from observed values in the same hot deck cell or in a matched "neighborhood" of the missing data case.
- *Goal 3: Multivariate.* To preserve associations among the many variables that may be included in the imputation model, the imputation procedure should be multivariate in nature. The multiple imputation algorithms included in today's major statistical software packages preserve the multivariate properties of the data, through a

THEORY BOX 11.2 EXPLICIT VERSUS IMPLICIT MODELS FOR IMPUTATION OF ITEM-MISSING DATA

Imputations of missing values for survey variables may be based on explicit distributional models or on techniques that "imply" an underlying model or set of assumptions concerning the missing data. In the example described in Section 11.4, the joint distribution for the three continuous variables (diastolic BP, age, and BMI) is assumed to be multivariate normal: $f(\boldsymbol{y}|\boldsymbol{\theta}) = \text{MVN}(\boldsymbol{\mu}, \Sigma)$. This is an example of an explicit imputation model. A modification of this explicit model might be to assume that diastolic BP, age, and BMI follow multivariate normal distributions that are conditioned on gender. This modified model is an example of the general location model (Schafer, 1999).

An example of an imputation procedure that incorporates an implicit imputation model is hot deck imputation, in which cases are grouped into cells based on common values for observed categorical variables (e.g., age category × gender), and within these cells missing values of another variable (e.g., diastolic blood pressure) are imputed by randomly drawing a replacement value from an observed "donor" in the same hot deck cell (Kalton and Kaspryzk, 1986). Although this simple hot deck method does not impute based on an explicit regression model for y, the model that is implied by the method might be the general location model or an analysis of variance (ANOVA) model.

sequence of conditional imputations (given monotonic missing data patterns) or through the use of iterative Markov chain Monte Carlo (MCMC) methods that are designed to simulate draws of missing values from the joint posterior distribution of the multivariate set of survey items.

- *Goal 4: Multiple.* Multiple independent applications of the imputation procedure permit the estimation of the variance that is attributable to imputing missing values. In a fashion similar to simple replicated estimates of variance due to the sampling process (Section 3.6.3), the multiple repetitions of the imputation process enable estimation of the added variance that is due to the imputation process. Multiple imputation inference requires that the imputation process be independently repeated multiple times, $m = 1, \ldots, M$. In theory, MI inference is most efficient at recovering statistical information from incomplete cases when $M = \infty$. Obviously, analysts' tolerance for MI would have faded away quickly if each analysis required an infinite number of independently imputed data sets. Fortunately, research has shown that virtually all of the efficiency possible in an MI analysis can be achieved using as few as $M = 5$ to $M = 10$ independent

repetitions of the imputation process (e.g., Rubin, 1986). As described previously, Rao and Shao (1992) provided an alternative jackknife method that can be used to estimate imputation variance from a singly imputed data set. Given the availability of user-friendly programs for multiple imputation analysis, the MI approach is most accessible to the typical survey analyst and is therefore the method that we cover in this chapter.

- *Goal 5: Robust.* No imputation model or procedure will ever exactly match the true distributional assumptions for the underlying random variables or the assumed missing data mechanism. Imputation procedures applied to survey data sets by data producers or data users should be reasonably robust against modest departures from the underlying assumptions. Fortunately, empirical research has demonstrated that if the more demanding theoretical assumptions underlying MI are relaxed, applications to survey data can produce estimates and inferences that remain valid and robust (Herzog and Rubin, 1983).
- *Goal 6: Usable.* For applied survey data analysis, usability and software support is always a criterion that should be used in selecting a statistical procedure. In its most mathematically rigorous representation, the theory of multiple imputation can be highly complex (Rubin, 1987; Schafer, 1997), involving explicit probability models for the data and the missing data mechanism and what Rubin (1996) labels "Bayesianly relevant and justifiable" methods for obtaining valid statistical inferences for population statistics. For highly specific problems with complex missing data patterns or mechanisms, survey analysts who lack the mathematical sophistication are encouraged to consult with a survey statistician prior to conducting and publishing their analysis. However, for general missing data problems of the type illustrated in Section 11.7, today's software provides a usable platform for conducting a multiple imputation analysis that is both informed and generates correct inferences for the target population.

11.3.3 Overview of Multiple Imputation and MI Phases

The ideas that underlie multiple imputation methods for missing data were formulated by Donald Rubin in the early 1970s (Rubin, 1977), and over the succeeding 30 years MI theory and methods have been advanced by Rubin and his students (Rubin, 1999). Multiple imputation is not simply a technique for imputing missing data. It is also a method for obtaining estimates and correct inferences for statistics ranging from simple descriptive statistics to the population parameters of complex multivariate models.

As illustrated in Figure 11.5, three distinct phases comprise a complete **multiple imputation analysis** of a survey data set. The first phase is the

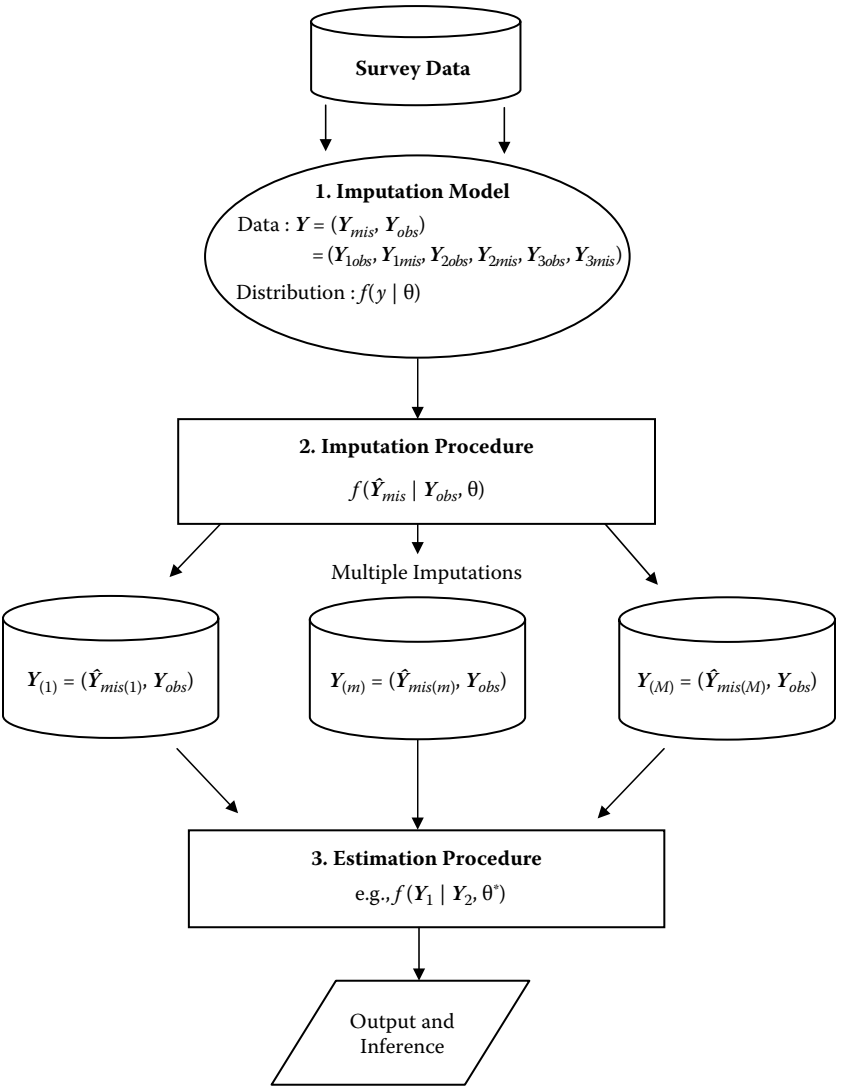


FIGURE 11.5
Multiple imputation analysis of survey data.

definition of the data and the distributional components of the imputation model. The second phase applies specific methods and algorithms to generate the multiple imputations of the missing values. The third and final phase is the calculation of MI estimates and standard errors and the construction of confidence intervals (CIs) and hypothesis test statistics for making inferences regarding population parameters and relationships. The following three sections of this chapter explore each of these phases in more detail.

11.4 Models for Multiply Imputing Missing Data

The actual process of imputing item missing values is governed by the **imputation model**, which is defined by the set of values that are available to the imputation process, $\boldsymbol{y} = \{\boldsymbol{y}_{obs}, \boldsymbol{y}_{miss}\}$, and the distributional assumptions, $f(\boldsymbol{y}|\boldsymbol{\theta})$, for the multivariate relationships among the elements of $\boldsymbol{y}$. Consider a problem in which there are three continuous variables of interest in a 2005–2006 NHANES analysis of adult diastolic blood pressure (BP): y_1 = diastolic BP (mmHg); y_2 = age (years); and y_3 = body mass index (BMI) (kg/m^2). Further, let us assume that the item-missing data rate is 14% for diastolic BP and 2% for BMI and that following rigorous edit checks, cleaning, and follow-up by NHANES staff, age is observed for every case. One possible imputation model would be to include all three variables in the imputation process and to assume that the joint distribution for these three continuous variables is multivariate normal: $f(\boldsymbol{y}|\boldsymbol{\theta}) = \text{MVN}(\boldsymbol{\mu}, \Sigma)$. Under this model, multiple imputations of the missing data for diastolic BP and BMI are easily performed using methods described in Schafer (1997). The imputations performed under this imputation model would serve for univariate MI estimation of the mean diastolic BP, μ_1, or mean BMI, μ_3, and they would also serve for the MI estimation of the regression of diastolic BP on age and BMI, for example, $E(y_1 \mid y_2, y_3) = B_0 + B_1 \cdot y_2 + B_2 \cdot y_3$. Van Buuren and Oudshoorn (1999) illustrate the process of imputation model selection for an MI analysis of blood pressure in a cohort of persons age 85+ from Leiden in the Netherlands.

11.4.1 Choosing the Variables to Include in the Imputation Model

The choice of variables to include in the imputation model should not be limited only to variables that have item-missing data or variables that are expected to be used in a subsequent analysis. As a general rule of thumb, the set of variables included in the **imputation model** for an MI analysis should be much larger and broader in scope than the set of variables required for the **analytic model**. For example, if age and BMI are the

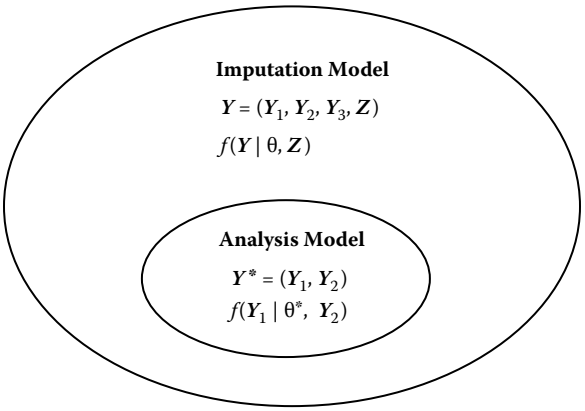


FIGURE 11.6
Relationship of the imputation and analysis model.

chosen predictors in the analytic model for diastolic BP, the imputation of item-missing data for diastolic BP and BMI might include many additional variables such as gender, race/ethnicity, marital status, height, weight, and systolic BP. Figure 11.6 is a schematic representation of the desired relationship between the imputation model and the analytic model. The large outer oval represents the broad set of imputation model variables that can contribute to the derivation of imputed values for missing items. Nested within this broad set of imputation model variables is the smaller set of variables that will be the focus of the analysis. Obviously, it is not feasible to define an imputation model and perform multiple imputations using every possible variable in the survey data set. Based on recommendations from Schafer (1999) and Van Buuren and Oudshoorn (1999), some practical guidelines for choosing which variables to include in the imputation model are the following:

- All analysis variables ($\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ in Figure 11.6).
- Other variables that are correlated or associated with the analysis variables ($\boldsymbol{Y}_3$).
- Variables that predict item-missing data on the analytic variables ($\boldsymbol{Z}$).

Failure to include one or more analysis variables ($\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$) in the imputation model can result in bias in the subsequent MI estimation and inference. Including additional variables that are good predictors of the analytic variables (e.g., $\boldsymbol{Y}_3$) improves the precision and accuracy of the imputation of item-missing data. For example, when analyzing longitudinal data from a panel survey, $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, and $\boldsymbol{Y}_3$ might represent measures of the same characteristic across three waves. An analytic model may be focused on estimating the relationship of $\boldsymbol{Y}_2$ and $\boldsymbol{Y}_3$, but an analyst imputing missing values for

$\boldsymbol{Y}_3$ could certainly use $\boldsymbol{Y}_1$ to improve the imputation model. Finally, under the assumption that item-missing data are MAR, incorporating variables ($\boldsymbol{Z}$) that predict the propensity for response reduces bias associated with the item-missing data mechanism.

Work in this area (Rubin, 1996; Reiter, Raghunathan, and Kinney, 2006) has also emphasized the importance of including variables identifying complex design features (survey weights, indicators for sampling error computation units (clusters) and indicators for sampling error strata) in imputation models. In particular, recent work by Reiter et al. demonstrates that ignoring complex sample design variables in the imputation models can lead to bias in multiple imputation estimates when the design variables are related to the survey variable of interest (which is often the case in practice) and including design variables in the imputation models when they are *not* related to the survey variable of interest can at worst lead to inefficient (and thus conservative) multiple imputation estimates.

11.4.2 Distributional Assumptions for the Imputation Model

Once the set of variables for the imputation model has been chosen, a decision must be made concerning the joint distributional model for the chosen variables. In theoretical discussions of multiple imputation methods, convenient choices of multivariate models for the joint distribution of the broad set of imputation variables might be multivariate normal (continuous), multinomial (categorical), or **general location** (mixed categorical or continuous) models. As described in Section 11.5, many of the current software programs explicitly assume that the variables in the imputation model follow one of these standard multivariate data models.

Because the chosen imputation model may include many variables of different types (continuous, categorical, count, semicontinuous), it may be difficult if not impossible to specify an analytical form for the joint posterior distribution of these variables. In such cases, several authors (Van Buuren et al., 1999; Raghunathan et al., 2001) have proposed using iterative Monte Carlo Markov Chain (MCMC) methods such as the **Gibbs sampler** to approximate imputation draws from the unknown **joint posterior distribution** of the imputation model variables (Schafer, 1997). In such cases, the joint posterior distribution of the data is assumed to exist, but its exact form is never truly observed. A potential vulnerability of this latter technique is that the target distribution that is being simulated in the MCMC process may not in fact exist for some missing data problems. Fortunately, empirical tests and simulation studies suggest that when applied to large survey data sets with moderate rates of item-missing data, these generalized methods of multivariate imputation do in fact yield reasonable results even when the number and diversity of variables included in the imputation model is large.

11.5 Creating the Imputations

The second phase of a multiple imputation analysis is to generate $m = 1, \ldots, M$ **completed data sets** in which the missing values, $\boldsymbol{y}_{miss}$, have been imputed. The $m = 1, \ldots, M$ independently imputed versions of the data set are termed **repetitions**.

A variety of algorithms, often tailored to the specific missing data pattern and mechanism, are used to generate the imputations. Readers are referred to the comprehensive texts by Rubin (1987), Schafer (1997), and Gelman et al. (2003) and articles by Tanner and Wong (1997), Gelfand and Smith (1990), and Gelfand et al. (1990) for more in-depth coverage of specific algorithms for addressing missing data under explicit data models and theoretically exact Bayesian methods.

Here we consider the more typical survey data context where the imputation model is multivariate and includes variables of all types, and there is a generalized pattern of missing data across the variables in the model. In such cases of a "messy" pattern of missing data where exact methods do not strictly apply, the authors of multiple imputation software have generally followed one of three general approaches described in the following sections.

11.5.1 Transforming the Imputation Problem to Monotonic Missing Data

The first approach is to transform the pattern of item-missing data to a monotonic pattern by first using simple imputation methods or an MCMC posterior simulation approach to fill in the missing values for variables in the model that have very low rates of item-missing data. Under a specified model for the data, imputation of a true monotonic pattern of item-missing data is greatly simplified: The imputation of missing data is reduced to a sequence of imputations for single variables. Each imputation requires only a single step. This approach works best when the generalized pattern of missing data is dominated by missing data for one or two variables. Consider the missing data for the HRS falls example in Table 11.1. The generalized pattern of missing data is dominated by the missing data on falls (4.5%), with lesser rates for weight (1.4%) and height (1.4%), virtually no missing data for arthritis (0.2%) and diabetes (0.1%), and complete data for age and gender. Through inspection of the pattern of missing data across individual cases, it would be possible to use simple methods to "fill in" selected missing values for the variables with small amounts of missing data, creating a monotonic pattern of missing data (Little and Rubin, 2002) and simplifying the imputation problem. Solas 3.0 (Statistical Solutions) and SAS PROC MI offer the user this option.

11.5.2 Specifying an Explicit Multivariate Model and Applying Exact Bayesian Posterior Simulation Methods

A second algorithmic approach to generating imputations for a generalized pattern of item-missing data is to declare an explicit probability model for the data, for example, $f(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \text{MVN}(\boldsymbol{\mu}, \Sigma)$ or $f(\boldsymbol{y}\,|\,\boldsymbol{\theta}) =$ general location model (Schafer, 1997). Individual imputations are then generated by random draws from the correct posterior distribution of the missing values $f(\hat{\boldsymbol{Y}}_{miss}\,|\,\boldsymbol{Y}_{obs}, \theta)$ under the explicit probability model.

Early versions of multivariate imputation programs, such as NORM (Schafer, 1997), assumed a multivariate normal distribution for all variables. Actual imputations for the multivariate model employed **data augmentation**, a form of the general class of MCMC posterior simulation algorithms. Later updates such as MIX (Schafer, 1997) incorporated the option of a general location model, more suitable to a mix of continuous and categorical variables. SAS PROC MI is another general purpose MI program that employs the MCMC methods of Schafer (1997) and can accommodate both continuous and categorical variables. For nonmonotone missing data patterns, Solas 3.0 applies linear regression to continuous and ordinal variables and discriminant analysis to assign imputations to missing values for nominal categorical variables.

11.5.3 Sequential Regression or "Chained Regressions"

The third alternative for multiply imputing item-missing data for large mixed sets of continuous, nominal, ordinal, count, and semicontinuous variables is the **sequential regression algorithm**, or the similar technique of "**chained equations**." The sequential regression algorithm forms the basis for the %IMPUTE command in the IVEware software system (Raghunathan et al., 2001). The chained equations algorithm (Van Buuren and Oudshoorn, 1999) has been extended and enhanced (Carlin et al., 2008; Royston, 2005) and is available in the `ice` command of Stata Version 10+. Each of these algorithms is based on an iterative MCMC Gibbs sampler algorithm. Each iteration (t = 1, …, T) of the algorithm moves one by one through the sequence of variables in the imputation model, $\boldsymbol{Y} = \{Y_1, ..., Y_k, ..., Y_p, Z_1, ..., Z_q\} = \{\boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3}, \boldsymbol{Z}\}$. In the very first iteration, the variable with the *smallest* amount of missing data is regressed on variables in the imputation model with complete data, using an appropriate model (e.g., logistic regression for a binary variable). Imputations for the missing values on the first variable are generated by stochastic draws from the predictive distribution defined by the regression model. Next, the variable with the second smallest amount of missing data is regressed on the first imputed variable and all other variables in the imputation model with complete data, and imputed values are generated in the same way. This algorithm proceeds until the variable with the most missing data is regressed on all other variables (some having imputed values) in the imputation model to develop the predictive distribution for the last variable. The first iteration

results in a completed data set after the final set of imputations for the last variable is randomly drawn from the appropriate predictive distribution.

Subsequent iterations use some or all of the variables in the imputation model (depending on the software) to reimpute the values that were originally missing for each of the variables. Based on the current (iteration t) values of the observed and imputed values for the imputation model variables, regression models are once again estimated for each target variable that originally had missing data, and, as mentioned already, the regression models can be tailored to each specific variable. For example, the IVEware software uses linear regression (continuous variables), multinomial logistic regression (nominal/ordinal variables), Poisson regression (count variables), or two-stage logistic/linear regression (semicontinuous variables) to model the conditional predictive distribution of individual variables Y_k, $f(\hat{\boldsymbol{Y}}_k^{(t)} \mid \hat{\boldsymbol{Y}}_{j\neq k}^{(t)}, \boldsymbol{\theta}^{(t)})$. Updated imputations, $\hat{Y}_k^{(t)}$, are then generated by stochastic draws from the predictive distribution defined by the updated regression model. Once the last variable in the sequence has been imputed, the algorithm again cycles through each variable. The iteration of the Gibbs sampler cycles stops when a user defined convergence criterion is met (e.g., $T = 10$ iterations in the Stata `ice` program).

Under an explicitly defined imputation model, $f(\boldsymbol{Y}, \boldsymbol{Z} \mid \theta)$, and a suitable prior distribution, $g(\theta)$, for the distributional parameters, the Gibbs sampler algorithm will, in theory, converge to the Bayesian joint posterior distribution for the data. Since the sequential regression method never explicitly defines $f(\boldsymbol{Y}, \boldsymbol{Z} \mid \theta)$, the assumption must be made that the posterior distribution does exist and that the final imputations generated by the iterative algorithm do represent draws from an actual, albeit unknown, posterior distribution. Although the exact theoretical properties of the resulting imputations remain somewhat of an unknown, in most applications the sequential regression approach with its application of the Gibbs sampler algorithm does converge to a stable joint distribution and the multivariate imputations generated by the algorithm show reasonable distributional properties. Further, empirical studies have shown that this method produces results comparable to those for the EM algorithm and the exact methods of Bayesian posterior simulation (Heeringa, Little, and Raghunathan, 2002).

The example exercise presented in Section 11.7 will demonstrate the application of the Stata `ice` sequential regression program to a multivariate missing data problem.

11.6 Estimation and Inference for Multiply Imputed Data

Once the item-missing data have been multiply imputed under a suitable imputation model and the $m = 1, \ldots, M$ completed data sets have been stored

in an appropriate data format, the next step in a complete MI analysis is to compute multiple imputation estimates of the population parameters and the variances of the MI estimates. The MI estimates and standard errors can then be used to construct confidence intervals for the population quantities of interest to the survey analyst.

Many of the major software systems now include a pair of programs for conducting multiple imputation analysis: one program to perform the multiple imputations of item-missing data, and a second to perform MI estimation and inference based on the multiple files that are produced by the imputation procedure. Examples of such pairs of programs include `ice` and `mim` in Stata Version 10+, `mi impute` and `mi estimate` in Stata Version 11, PROC MI and PROC MIANALYZE in SAS Version 9.2, and %IMPUTE and %DESCRIBE / %REGRESS in IVEware. The newest version of Mplus at the time of this writing (Version 5.21) also has built-in multiple imputation analysis procedures. Although the software supports coordinated processing of both the imputation and estimation/inference phases of an MI analysis, provided that some care is taken in choosing the imputation model, the imputation and estimation phases can be performed separately by different persons (i.e., imputation by data producers, analysis by data users) or by using separate programs (e.g., conducting imputations in IVEware and then estimation and inference in PROC MIANALYZE).

11.6.1 Estimators for Population Parameters and Associated Variance Estimators

Multiple imputation estimates of population parameters (denoted here by θ) are computed using a simple averaging of the parameter estimates obtained by performing standard (possibly design-based) analyses of the $l = 1, \ldots, M$ completed data sets based on the multiple imputation algorithm:

$$\bar{\theta} = \frac{1}{M}\sum_{l=1}^{M}\hat{\theta}_l \tag{11.1}$$

where $\hat{\theta}_l$ = estimate of θ from the completed data set $l = 1, \ldots, M$.

The corresponding multiple imputation variance for the estimate $\bar{\theta}$ is estimated as a function of the average of the estimated sampling variances for each repetition estimate and a between-imputation variance component that captures the imputation variability over the M repetitions of the imputation process:

$$var(\bar{\theta}) = \bar{U} + \left(\frac{M+1}{M}\right)\cdot B \tag{11.2}$$

where

$$\bar{U} = \text{Within-imputation variance} = \frac{1}{M}\sum_{l=1}^{M} U_l = \frac{1}{M}\sum_{l=1}^{M} var(\hat{\theta}_l); \text{ and}$$

$$B = \text{Between-imputation variance} = \frac{1}{M\text{-}1}\sum_{l=1}^{M} (\hat{\theta}_l - \bar{\theta})^2$$

In applications involving complex sample survey data, the estimated sampling variance of each repetition estimate, $var(\hat{\theta}_l)$, is computed using an appropriate Taylor series linearization or replication variance estimator.

11.6.2 Model Evaluation and Inference

Based on the estimates of the within and between components of the multiple imputation variance estimator, the **fraction of missing information** is computed as

$$\hat{\gamma}_{mi} = \left[\frac{\left(\frac{M+1}{M}\right)\cdot B}{\left(\frac{M+1}{M}\right)\cdot B + \bar{U}} \right] = \left[\frac{\left(\frac{M+1}{M}\right)\cdot B}{T} \right] \tag{11.3}$$

If missing data for a single variable y were imputed based only on the distribution of the observed values of that same variable, the fraction of missing information would equal the missing data rate for y. However, more generally, when the model for imputing y conditions on observed values of other related variables (denoted here by x_1 and x_2), such as $\hat{y}_{i,miss} = \hat{B}_0 + \hat{B}_1 \cdot x_{1i} + \hat{B}_2 \cdot x_{2i}$, the imputation borrows strength from the multivariate relationships and the fraction of information lost will be reduced: $0 < \hat{\gamma}_{mi} <$ Missing rate for y. We note that the fraction of missing information is specific to an *estimate* of interest and will tend to vary slightly depending on the estimate.

Confidence intervals for descriptive population parameters or single parameters in superpopulation regression models are constructed from the multiple imputation estimate, its standard error, and a critical value from the Student t distribution (Rubin and Schenker, 1986):

$$CI_{(1-\alpha)}(\theta) = \bar{\theta} \pm t_{\tilde{v}_{mi},1-\alpha/2} \cdot se(\bar{\theta}) \tag{11.4}$$

where $\tilde{v}_{mi}$ = the degrees of freedom for the MI variance estimate (see Theory Box 11.3).

The unique feature of the MI confidence interval for population parameters is the degrees of freedom determination for the Student t distribution.

THEORY BOX 11.3 DEGREES OF FREEDOM FOR REPEATED MI INFERENCES FROM COMPLEX SAMPLE SURVEY DATA

In general, large-sample multiple imputation confidence intervals for population values are based on a Student *t* distribution with the following degrees of freedom:

$$v_{mi} = (M-1)\hat{\gamma}^{-2} \tag{11.5}$$

where $\hat{\gamma}$ = the estimated fraction of missing information (Equation 11.3).

Barnard and Rubin (1999) point out that this large-sample approximation may not be suitable for "small" samples. Even in small simple random samples, this formula can lead to an estimate of degrees of freedom that exceeds the number of complete observations ($df \sim n - 1$). For most complex sample designs, including those that serve as examples in this text, application of this large-sample formula results in values of v_{mi} that greatly exceed the nominal design-based degrees of freedom (#clusters – #strata). Barnard and Rubin propose the following adjustment to establish the degrees of freedom for small samples or designs with limited degrees of freedom:

$$\tilde{v}_{mi} = \left(\frac{1}{v_{mi}} + \frac{1}{\hat{v}_{obs}} \right)^{-1} \tag{11.6}$$

where

v_{mi} = the large sample MI degrees of freedom;

$\hat{v}_{obs}$ = the degrees of freedom for the complete data,

$$= \left(\frac{v_{com}+1}{v_{com}+3} \right) \cdot v_{com} \cdot (1-\hat{\gamma}_{mi})$$

v_{com} = the degrees of freedom for the complete case analysis;

$\hat{\gamma}_{mi}$ = the estimated fraction of missing information.

In the analysis of a small simple random sample, the degrees of freedom for the Student *t* reference distribution would be $v_{com} = n - 1$. The complete data degrees of freedom for an analysis of a complex sample survey data set would be the now familiar $v_{com} = v_{des} \approx \#clusters - \#strata$. Chapter 3 discussed the approximate nature of the current rules for determining design-based degrees of freedom for complex sample survey data. The same or greater level of approximation must hold in any MI inferences from complex sample survey data. Additional research in this area is certainly needed. The Barnard and Rubin (1999)

approximation to the effective degrees of freedom is now incorporated in multiple imputation estimation software such as Stata `mim` and SAS PROC MIANALYZE.

Current software uses the "small sample" method of Barnard and Rubin (1999) to determine the degrees of freedom, v, for constructing the confidence interval. Theory Box 11.3 provides a brief description of this degrees of freedom calculation. The MI confidence intervals are routinely reported for individual descriptive parameters or regression model parameters by the multiple imputation analysis programs in Stata (`mim`) and SAS (PROC MIANALYZE). Simulation studies have demonstrated that for large sample sizes, this MI confidence interval provides true coverage of the population value that is very close to the nominal $100(1 - \alpha)\%$ coverage level.

Li, Raghunathan, and Rubin (1991) developed a multiple imputation F-test statistic that serves the function of a Wald statistic to test multiparameter hypotheses. In Stata, this multiparameter F-test is invoked using the combination of the `mim` modifier and the Stata `testparm` postestimation command.

11.7 Applications to Survey Data

The multiple imputation analysis example presented in this section considers a linear regression model for the 2005–2006 NHANES mobile examination center (MEC) measures of diastolic blood pressure. The example will focus on a multiple imputation treatment of the item-missing data for 2005–2006 NHANES adult respondents who agreed to participate in the MEC phase of the study. The nonresponse adjustment factor included in the NHANES MEC analysis weight, WTMEC2YR, serves as compensation for the 4.1% of NHANES interview respondents who did not consent to participate in the MEC but does not address problems with item-missing data.

11.7.1 Problem Definition

Table 11.2 provides an overview of item-missing data rates for a variety of items from the 2005–2006 NHANES data. The table considers the overall missing data rates only for the 2005–2006 NHANES MEC adult respondents, or those 95.9% of the adult interview respondents who agreed to participate in the MEC data collection phase.

Although actual missing data rates can vary substantially across surveys, the missing data rates displayed in this table show a typical pattern. Missing

TABLE 11.2

Item-Missing Data Rates for Selected 2005–2006 NHANES Variables

Variable	Variable Name	NHANES Adult MEC Respondents	
		n	% Missing
Age	AGEC	5,334	0.0%
Gender	RIAGENDR	5,334	0.0%
Marital status	MARCAT	5,329	0.1%
Race/ethnicity	RIDRETH1	5,334	0.0%
Body mass index	BMXBMI	5,237	1.8%
Poverty index	INDFMPIR	5,066	5.0%
Diastolic blood pressure	BPXDI1_1	4,581	14.1%

data on key demographic variables such as age, gender, and race are generally extremely low. Missing data rates for socioeconomic variables such as education level or marital status are also typically <1% of interviewed cases. Significantly higher rates of missing data occur for variables that measure personal or family income, assets, benefit amounts, or other such financial variables. The NHANES Poverty Index variable is a constructed measure derived from survey measures of household income and family characteristics and holds a 5.0% missing data rate in the 2005–2006 NHANES. Among 2005–2006 NHANES sample persons who consented to the MEC physical exam, examples of missing data rates for physical measures are 1.8% for BMI and 14.1% for diastolic blood pressure.

The aim of this example exercise is to multiply impute the missing values for these seven MEC variables and to perform multiple imputation analyses.

11.7.2 The Imputation Model for the NHANES Blood Pressure Example

The first step in the MI analysis of diastolic blood pressure is to impute the item-missing data using an imputation model that includes the following mixture of continuous and categorical variables:

BPXDI1_1: Diastolic blood pressure
MARCAT: Marital Status
RIAGENDR: Gender
RIDRETH1: Race/Ethnicity Category
AGEC, AGECSQ: Linear and quadratic terms for age (years)
BMXBMI: Body mass index (kg/m^2)
INDFMPIR: Poverty Index (continuous scale with values in the range 1-5)

In this example, the variables included in the imputation model are those that will be used to estimate the linear regression model for diastolic blood pressure. Recall that the 2005–2006 NHANES MEC data have no missing values for the age, gender, and race/ethnicity variables. Therefore only diastolic blood pressure, marital status, body mass index, and the poverty index variables require imputations. The linear and quadratic terms for age and the gender and race/ethnicity predictors are used by the `ice` command in Stata to build the regression models used to impute missing values for BPXDI1_1, MARCAT, BMXBMI and INDFMPIR, but their observed distribution will not be altered across the $M = 5$ multiple imputations. For simplicity in this example, we include the same variables in the imputation model that are included in the analytic model of interest, but, more generally, additional variables could be included in the imputation models.

11.7.3 Imputation of the Item-Missing Data

We illustrate the imputation step here using one general purpose program for imputation of item-missing data: the Stata `ice` procedure. The `ice` (Imputation by Chained Equations) program is capable of multiply imputing generalized patterns of missing data for a multivariate vector of survey variables. Section 4.4 has illustrated the use of the Stata `mvpattern` command to display the individual rates and patterns of missing data for the seven variables that will be included in this analysis. After examining the rates and patterns of item-missing data for the variables included in the imputation model, the next step is to multiply impute the item-missing data for BPXDI1_1, MARCAT, BMXBMI and INDFMPIR. Using the following `ice` command, the missing data are independently imputed $M = 5$ times using the sequential regression technique, and five completed data sets with the observed and imputed values are created.

```
ice bpxdi1_1 marcat m2 m3 r2 r3 r4 r5 g2 agec agecsq bmxbmi ///
indfmpir, saving(f:\applied_analysis\imputed_nhanes) m(5) ///
seed(123) replace passive(m2:marcat==2\ m3:marcat==3) ///
substitute(marcat:m2 m3)
```

Immediately following the `ice` command is the list of variables that will be included in the imputation process. Recall that only the variables BPXDI1_1, MARCAT, BMXBMI and INDFMPIR have missing data. Variables for Age (AGEC, AGECSQ), Gender (G2), and Race (R2, R3, R4, R5) are included to be used as predictors in the regression models that `ice` will use to predict (impute) missing values of the other four variables. From this syntax, gender and race are represented as indicator variables with a reference category parameterization (e.g., R2 is an indicator for RACE = 2, and RACE = 1 is the reference category of RACE). MARCAT is a three-category variable with a small amount of item-missing data. It is listed as MARCAT, but marital

status is also represented on this input line by the indicator variables M2 and M3. Stata will default to impute MARCAT (and all other categorical variables) through a multinomial logit regression; however, when marital status enters as a predictor in the regression models for imputing other variables, the `replace passive` and `substitute` keywords inform Stata to use the indicators M2 and M3 to represent the effects of marital status.

The initial output from the `ice` command is displayed as follows. The first output table summarizes the distribution of cases by the counts of missing values. Conforming to the previous `mvpattern` results, a total of 4,308 cases have 0 missing items; 935 have missing data for one of the seven imputation model variables; 85 have two variables missing; and six cases are missing a trio of values. The second portion of this output is a summary identifying the exact form of the regression model and the independent variables used in the regression models for each variable. Although the option is not used in this example, Stata `ice` does provide the user the option to control the exact form of the regression model that will be used to impute missing data for individual variables.

#missing values	Freq.	Percent	Cum.
0	4,308	80.76	80.76
1	935	17.53	98.29
2	85	1.59	99.89
3	6	0.11	100.00
Total	**5,334**	**100.00**	

Variable	Command	Prediction equation
bpxdi1_1	regress	m2 m3 r2 r3 r4 r5 g2 agec agecsq bmxbmi indfmpir
marcat	mlogit	bpxdi1_1 r2 r3 r4 r5 g2 agec agecsq bmxbmi indfmpir
m2		[Passively imputed from marcat==2]
m3		[Passively imputed from marcat==3]
r2		[No missing data in estimation sample]
r3		[No missing data in estimation sample]
r4		[No missing data in estimation sample]
r5		[No missing data in estimation sample]
g2		[No missing data in estimation sample]
agec		[No missing data in estimation sample]
agecsq		[No missing data in estimation sample]
bmxbmi	regress	bpxdi1_1 m2 m3 r2 r3 r4 r5 g2 agec agecsq indfmpir
indfmpir	regress	bpxdi1_1 m2 m3 r2 r3 r4 r5 g2 agec agecsq bmxbmi

TABLE 11.3

Weighted Estimates of Means and Percentages for Selected Variables before and after Multiple Imputation in Stata (among Adults Who Completed the 2005–2006 NHANES Interview and MEC Surveys)

		Estimates before Imputation		Multiple Imputation Repetition Estimates (n = 5334)				
Variable	**Statistic**	n_{obs}	**Value**	m = 1	m = 2	m = 3	m = 4	m = 5
Marital status	%Married	5,329	63.92	63.90	63.90	63.94	63.92	63.90
Body mass index	Mean	5,237	28.34	28.44	28.42	28.45	28.42	28.45
Poverty index	Mean	5,066	3.11	3.06	3.06	3.07	3.06	3.07
Diastolic blood pressure	Mean	4,581	70.71	70.49	70.49	70.48	70.51	70.35

The example `ice` command instructs Stata to store the multiply imputed data in a system file labeled `imputed _ nhanes.` This output data set from `ice` includes a concatenation of the data files that result from the independent repetitions of the MI process. The records corresponding to each repetition data set are identified by a system generated variable, `_ mj` (e.g., `_ mj` = 1, …, `_ mj` = 5). Records with `_ mj` = 0 correspond to the original (before imputation) version of the data set.

We strongly recommend that users visually examine and compare the key statistics for the original and imputed versions of variables included in the MI run. To display selected results from the concatenated data set (`imputed _ nhanes`), the following Stata descriptive commands can be used:

```
svy: mean bmxbmi indfmpir bpxdi1_1 [pweight=wtmec2yr], over(_mj)
svy: proportion marcat [pweight=wtmec2yr], over(_mj)
```

Table 11.3 provides a simple comparison for the four variables in the example problem.

11.7.4 Multiple Imputation Estimation and Inference

Once the multiple imputations are complete, the data are ready for analysis. In this example, two MI analyses of the imputed data are performed. The first analysis computes the MI estimate and 95% CI for the mean diastolic blood pressure for U.S. adults who completed the MEC portion of the NHANES survey (n = 5,334). The second analysis uses MI methods to estimate a linear regression model of diastolic blood pressure (BPXDI1_1) on MARCAT (Marital Status), RIAGENDR (Gender), RIDRETH1 (Race/Ethnicity), AGEC and AGECSQ (linear and quadratic terms for age in years), BMXBMI (body mass index (kg/m^2)), and INDFMPIR (Poverty Index).

Each analysis uses the $M = 5$ multiply imputed data sets and applies the general procedures for multiple imputation estimation and inference outlined in Section 11.6. In Stata, including the `mim` modifier invokes a multiple imputation analysis for MI data sets produced by the `ice` command or independently developed MI data sets that include the correct variable coding (i.e., `_mj`) for the M repetitions.

11.7.4.1 Multiple Imputation Analysis 1: Estimation of Mean Diastolic Blood Pressure

Assuming that a single data file containing the five multiply imputed data sets (i.e., the `imputed_nhanes` file produced by `ice` earlier) is open in Stata, with the five imputed data sets indexed by the `_mj` variable, the Stata commands required to generate the MI estimate of the adult population mean of diastolic blood pressure are as follows:

```
svyset sdmvpsu [pweight=wtmec2yr], strata(sdmvstra)
mim: svy: mean bpxdi1_1
```

The active Stata data set for this analysis is the output file generated by the Stata `ice` multiple imputation run. The initial command is the now-familiar `svyset` command in which the stratum, cluster, and analysis weight variables for the 2005–2006 NHANES are specified. The second command in the sequence includes the `mim` modifier that informs Stata that the active data set includes multiple imputations and that multiple imputation computations for estimates, standard errors, degrees of freedom, and 95% CIs are desired. Following the `mim` modifier is the standard Stata `svy: mean` command for design-based estimation of means and their standard errors. Table 11.4 provides the multiple imputation estimates of the mean, its standard error, and a 95% CI for the population mean. For comparison, the table also reports the same statistics for a standard complete case analysis of mean diastolic blood pressure.

Following Equation 11.1, the final multiple imputation estimate of the mean diastolic blood pressure is $\bar{y}_{mi} = 70.46$. We also note the larger standard error compared to the complete case analysis, correctly incorporating between-imputation variance.

TABLE 11.4

2005–2006 NHANES Estimates of Mean Diastolic Blood Pressure of U.S. Adults: Comparison of Complete Case and Multiple Imputation Analyses

Analysis Type	n	$\bar{y}$	$se(\bar{y})$	95% CI	df
Complete case	4,581	70.72	0.229	(70.268, 71.165)	15
Multiple imputation	5,334	70.46	0.311	(69.758, 71.138)	12.5

After the Stata `mim` command, the `matrix` postestimation command may be used to store and display the within-imputation variance, between-imputation variance, and total variance of this MI estimate of the mean of diastolic blood pressure for U.S. adults. The use of the `noisily` option requests output for each value of the `_mj` variable, indicating each imputed file as well as the non-imputed data (`_mj` = 0):

```
mim, noisily storebv: svy: mean bpxdi1_1
matrix list e(MIM_W)
matrix list e(MIM_B)
matrix list e(MIM_T)
```

The resulting estimate of the "within" imputation variance (11.2) is $\bar{U} = 0.09176358$, the "between" component is $B = 0.00417692$, and the total variance is $T = 0.09677588$.

Applying expression (11.2) for the total variance of the multiple imputation confirms the latter result:

$$var(\bar{y}_{mi}) = \bar{U} + \left(\frac{M+1}{M}\right) \cdot B = T$$

$$= 0.09176358 + \frac{6}{5} \cdot 0.00417692 = 0.09677588$$

$$se(\bar{y}_{mi}) = \sqrt{var(\bar{y}_{mi})} = 0.31108823$$

The estimated fraction of missing information is then computed as follows:

$$\hat{\gamma}_{mi} = \left[\frac{\left(\frac{M+1}{M}\right) \cdot B}{\left(\frac{M+1}{M}\right) \cdot B + \bar{U}} \right] = \left[\frac{(6/5) \cdot 0.00417692}{0.09677588} \right] = 0.0518$$

Following Barnard and Rubin (1999), the approximate degrees of freedom (11.6) for the Student *t* distribution used to construct the confidence interval for the MI estimate of the mean diastolic blood pressure is 12.5, which is slightly less than the 15 degrees of freedom assumed for a design-based analysis of the 2005–2006 NHANES complex sample survey data.

11.7.4.2 Multiple Imputation Analysis 2: Estimation of the Linear Regression Model for Diastolic Blood Pressure

The next step in the example exercise is to extend the analysis to a regression model in which the mean of diastolic blood pressure is modeled as a function

TABLE 11.5

Estimated Linear Regression Models for Diastolic Blood Pressure of U.S. Adults (Comparison of Complete Case and Multiple Imputation Analyses).

		Multiple Imputation (n = 5,334)			**Complete Case Analysis (n = 4,305)**		
Predictor[a]	**Category**	$\hat{B}_j$	$se(\hat{B}_j)$	***p*-Value**	$\hat{B}_j$	$se(\hat{B}_j)$	***p*-Value**
Intercept	Constant	68.691	1.221	<0.001	68.287	1.451	<0.001
Race/ethnicity	Other Hispanics	1.176	1.028	0.277	1.653	1.160	0.175
	White	2.109	0.559	0.005	2.145	0.661	0.005
	Black	3.134	0.794	0.003	3.311	0.869	0.002
	Other	1.915	0.931	0.062	1.664	0.923	0.092
Marital status	Previously married	0.799	0.700	0.287	1.034	0.693	0.157
	Never married	–0.409	0.657	0.552	–0.373	0.573	0.525
Gender	Female	–2.779	0.422	<0.001	–2.699	0.389	<0.001
Age (centered)	Continuous	0.121	0.013	<0.001	0.119	0.016	<0.001
Age (centered) squared	Continuous	–0.012	0.001	<0.001	–0.012	0.001	<0.001
Body mass index	Continuous	0.183	0.033	<0.001	0.197	0.039	<0.001
Poverty index	Continuous	–0.127	0.151	0.424	–0.108	0.138	0.446

Source: Analysis based on the 2005–2006 NHANES data.
Note: MI degrees of freedom are determined separately for each estimate.
[a] Reference categories are Mexican (Race/Ethnicity); Married (Marital Status); Male (Gender).

of the selected covariates. The Stata command required to fit the multiple linear regression model to the multiply imputed data sets is as follows:

```
xi: mim: svy: regress bpxdi1_1 i.marcat i.riagendr ///
i.ridreth1 agec agecsq bmxbmi indfmpir
```

The standard Stata `svy: regress` command is preceded by the `mim:` modifier to invoke the multiple imputation analysis of the $M = 5$ multiply imputed data sets. The command also uses the `xi:` modifier to indicate that levels of the categorical predictors (`i.` prefix) will be represented in the model by indicator variables with a reference category parameterization. Table 11.5 provides the MI estimates of the regression parameters and standard errors along with the *p*-values for the *t*-tests of the null hypothesis $H_0 : B_j = 0$ for each parameter indexed by *j*. For comparison, the right half of Table 11.5 provides comparable estimates for a complete case analysis in which there is list-wise deletion of cases with missing values on the dependent variable or one or more predictors.

On first observation of the comparative results in Table 11.5, the estimated coefficients in the two fitted regression models are not identical—an expected outcome given that the imputation procedure has permitted 1,029 additional cases to enter the analysis. The estimated standard errors for the MI estimation of the regression model are generally less than or equal to the standard errors of the coefficients from the analysis of the $n = 4{,}305$ cases that had no missing data on either the dependent variables or the seven predictors. The slightly smaller standard errors from the MI analysis reflect the fact that the imputation has recovered additional statistical information from the incomplete cases that were excluded by list-wise deletion from the analysis of the nonimputed data. Finally, a comparison of the hypothesis test results for the single parameter *t*-tests of the null hypothesis $B_j = 0$ show differences in the *p*-values depending on the analysis approach. However, in this particular model, the differences are not large enough to lead an analyst to reverse his or her decision concerning the statistical significance of the individual predictors in the model.

Finally, to illustrate the MI multiparameter test of Li et al. (1991), consider a joint test of the null hypothesis that the effects of marital status and race on diastolic blood pressure levels are not significantly different from zero. To perform this test following the MI regression analysis, the Stata `mim: testparm` modifier and command are used:

```
mim: testparm _Imarcat_2 _Imarcat_3 _Iridreth1_2 _ ///
Iridreth1_3 _Iridreth1_4 _Iridreth1_5
```

The Stata output generated by this test command is shown as follows:

```
( 1) _Imarcat_2 = 0
( 2) _Imarcat_3 = 0
( 3) _Iridreth1_2 = 0
( 4) _Iridreth1_3 = 0
( 5) _Iridreth1_4 = 0
( 6) _Iridreth1_5 = 0
    F( 6, 301.9) = 4.39
      Prob > F = 0.0003
```

Based on the test results, the hypothesis of no effect of marital status and race on diastolic blood pressure is rejected.

We note that at the time of this writing, Stata Version 11 will soon be released. This newest version of Stata will include a brand new command (`mi`) offering extensive multiple imputation analysis capabilities, and we will include examples of the use of this command on the book Web site as soon as possible.

11.8 Exercises

1. These exercises consider data from the 2005–2006 NHANES. The objective of this set of exercises is to perform a typical imputation and analysis session. The logical steps include examination of missing data, imputation of missing data values using a multiple imputation software tool of choice, and analysis of the imputed data sets using a companion software tool of choice capable of handling multiply imputed data sets. Begin by downloading the following subset of data from the 2005–2006 NHANES: `c11_exercises_nhanes.dta` (available from the book Web site). Note that this data set is limited to adults 18+ years of age and those that completed the NHANES medical examination (n = 5,534). The variables used in the imputation and analysis are gender (RIAGENDR), body mass index (BMXBMI), race/ethnicity (RIDRETH1), age (RIDAGEYR), and systolic blood pressure (BPXSY1). The data set also contains the NHANES complex design variables and probability weight (SDMVSTRA, SDMVPSU, WTMEC2YR).
   a. Examine simple descriptive statistics for these variables (means, proportions, ranges, and counts of missing values) keeping in mind that the full n is 5,334. Use a software tool of choice for this step.
   b. Pay close attention to the types of the variables with missing data (continuous, ordinal, binary, or nominal) and the amount of missing data; that is, what percent are missing on each variable? Prepare a table including the type of each variable in the data set along with the missing data rate for that variable.
2. Using your software of choice and the data from Exercise 1, examine the missing data patterns that exist in the file.
   a. How many unique missing data patterns exist?
   b. Which variables have some missing data?
   c. Which variables have full data?
3. Prepare syntax to impute the missing values using the software of your choosing. Keep in mind what type of model should be used for the imputation, depending on the type of variable to be imputed. (For example, linear regression for continuous variables, ordinal or cumulative logit regression for ordered variables, binary logistic regression for binary variables, and so forth.) If you choose to use the Stata 10 `ice` command for imputation, use the `dryrun` option first to carefully check your imputation syntax and logic. Use M = 5, or prepare five imputed data sets during this step, and be sure to use a seed value so that your results can be replicated at a later time.

4. Execute the imputation commands and produce five imputed data sets for subsequent analysis. Save the imputed data set, and make sure to now use this imputed data set for all subsequent analyses.
   a. What is the name of the variable that identifies the imputation number 1–5? Note that this will depend on the software procedure that you are using.
   b. How many observations does this "concatenated" file contain? If you changed the number of imputations, how would that impact the number of observations in this file?
5. Perform an appropriate design-based descriptive analysis on each imputed variable in each imputed data set (1–5) contained in the concatenated imputed data set. For example, if the imputed variable is continuous, estimate means within each imputed data set using the correct survey weight (WTMEC2YR), and estimate appropriate standard errors for the estimated means using Taylor series linearization or a replication technique. Create a table showing the estimated means and proportions (and estimated standard errors) for each variable in each imputed data set.
6. Next, perform a multiple imputation analysis to estimate the mean BPXSY1, and compute a variance estimate for the estimated mean that incorporates the design-based standard errors from the analysis of each imputed data set in addition to the between-imputation variance in the estimates. (Use a command designed to analyze multiply imputed data sets and incorporate the complex sample variance corrections in addition to the variability introduced by imputing the missing values multiple times. In Stata Version 10+ this command would be `mim`, in SAS it would be PROC MIANALYZE, and in IVEware it would be %DESCRIBE.)
   a. What is the overall estimated mean for systolic blood pressure for all imputed data sets combined?
   b. What is the standard error of the estimated mean, and what has been accounted for in the standard error when using the `mim` modifier with the `svy: mean` command (in Stata) or a similar command in another software package?
   c. Obtain the within- and between-imputation variance components by either calculating these statistics manually or requesting them from the software package of choice (see Section 11.6.1).
7. Finally, fit a regression model with systolic blood pressure as the dependent variable and age, gender, race, and BMI as predictors. Fit the model first using a standard design-based analysis of complete cases only, and then perform a multiple imputation analysis using the software of your choice.

a. Prepare a table comparing estimated regression parameters and standard errors under the two analysis approaches (complete cases only, and multiple imputation).
b. Have the multiple imputations of the item-missing values and subsequent multiple imputation analysis changed any of your conclusions about the significance of the predictor variables?

12

Advanced Topics in the Analysis of Survey Data

12.1 Introduction

Chapters 5 through 10 presented design-based approaches to the most common statistical analyses of complex sample survey data. Models certainly played an important role in defining these analyses, but in each case, the estimation and inferences for population parameters were based on the final survey weights (incorporating unequal probabilities of selection into the sample, nonresponse adjustments, and poststratification adjustments) and the large sample properties associated with the expected sampling distribution of estimates under repeated sampling from the chosen probability sample design. As described in Section 2.2, the "tried and true" design-based techniques for the analysis of complex sample survey data are appealing because they require minimal assumptions, produce consistent estimates of the target population parameters, and provide robust measures of the sampling variability of the estimates.

Most survey practitioners accept these design-based methods as the method of choice for simple descriptive estimation of population characteristics. Moving to more analytical treatments of survey data such as multivariate linear and logistic regression, there has been some controversy over time about the best analytic approaches (Section 3.4); however, as a practical choice, most survey analysts today will follow a design-based approach to estimating multivariate regression models.

As survey analysis problems move away from the safety of simple descriptive analysis objectives and large sample sizes, the role of explicit statistical models becomes more critical, and today, the cutting-edge developments in survey data analysis are focused on *model-driven* approaches to the analysis of survey data. Some common applications in which statistical models must play a role include the following:

- Survey data with multiple levels where the hierarchical relationships and variation are substantively important (multilevel or hierarchical models).

- Longitudinal or repeated measures models.
- Latent variable models such as structural equation models.
- Small-area estimation.

This final chapter will provide a brief overview of important developments in these areas and describe areas of active and future research. Section 12.2 summarizes current work by survey statisticians who are looking more generally at the role Bayesian models and methods can play in the analysis of survey data sets. Over the past 15 years, applications of **generalized linear mixed models** (GLMMs) to "multilevel" modeling of hierarchical data (e.g., schools, classrooms, students) have grown to be increasingly important. These models, along with **generalized estimating equation (GEE)** models, are important in the analysis of longitudinal and other repeated measures forms of survey data. Section 12.3 introduces GLMMs and includes an example application of fitting a GLMM to longitudinal data on household net worth from years 2000 to 2006 of the Health and Retirement Study (HRS). Section 12.4 provides a summary of recent work on the application of latent variable models, such as structural equation models, to complex sample survey data. A brief review of the recent developments in the applications of survey data to **small-area estimation** problems is provided in Section 12.5. Finally, nonparametric methods have always been important statistical tools for data analysts, and recently there has been increased interest in theoretical developments and applied research on the application of nonparametric analyses to complex sample survey data. Section 12.6 reviews this recent work on nonparametric methods.

12.2 Bayesian Analysis of Complex Sample Survey Data

We briefly introduce Bayesian approaches to the analysis of survey data in this section. We assume that readers have some prior knowledge of Bayesian analysis methods and introduce terminology that might seem foreign to some readers without rigorous definitions. We recommend Gelman, Carlin, and Stern (2003) for readers who are interested in learning about Bayesian data analysis techniques.

Many leading survey statisticians see a model-based approach as providing a necessary principled framework for addressing problems in survey inference:

> Bayesian methods provide a unified framework for addressing all of the problems of survey inference, including inferences for descriptive or analytical estimands, small or large sample sizes, inference under

> planned ignorable sample selection methods such as probability sampling, and problems where modeling assumptions play a more central role such as missing data or measurement error. R. J. A. Little (Chambers and Skinner, 2003)

Briefly, Bayesian inference is focused on the posterior distribution of some parameter of interest θ (e.g., a population mean), denoted by $p(\theta|y)$, under an assumed model for the observed data y, denoted by $f(y|\theta)$, and a posited prior distribution for the parameter θ, denoted by $p(\theta)$. A Bayesian analysis of complex sample survey data involves analyzing the data to obtain a posterior distribution of parameters defining a model of interest and then using the model to determine a posterior predictive distribution for the unobserved elements of the population. Once a posterior predictive distribution is available, predictions on survey variables for elements not included in the sample can be computed and inferences can be made to the total population.

Complex sample design effects (i.e., stratification, clustering and weighting due to differential inclusion probabilities/nonresponse adjustment/poststratification adjustment) can be directly modeled or not, depending on the degree to which they are informative for the distributions of the variables of interest in the analysis. Compared with a strictly design-based approach to inference, this is a distinguishing feature of more model-driven approaches. Design-based methods always incorporate weights and design variables such as strata and clusters in the analysis, which ensures robust variance estimation but sometimes leads to losses in efficiency (i.e., larger standard errors). Bayesian methods for survey data analysis do not rely on large-sample asymptotic theory, but, in exchange, they are heavily model driven and require many assumptions, especially concerning choices of prior distributions and distributions for the observed data. In addition to the unifying theoretical framework cited in the opening quote in this section, Little (2003) identifies several features that should make the Bayesian approach appealing to survey analysts:

- Standard design-based inference (confidence intervals, test statistics) can be derived in the Bayesian framework using reasonable assumptions concerning the distribution of the data and the prior distribution for the parameter of interest.
- Survey applications typically assume noninformative priors; however, information gleaned from repeated and past surveys could be used to develop informed priors—for example, information on the tails of the distribution that generates outlier values in a single survey.

When sample sizes are large and diffuse priors apply, Bayesian estimates and posterior probability intervals are similar to design-based estimates and confidence intervals.

As noted already, a Bayesian approach to inference from survey data can, if necessary, directly incorporate the features of the complex sample design in estimation and inference. **Ignorable sampling mechanisms** are important when making inferences based on Bayesian methods. If the sampling mechanism (e.g., stratification variables, measures of size in probability proportionate to size [PPS] sampling, cluster definitions) and the parameter of interest (e.g., mean household net worth) are a priori independent, and the sampling indicator (or the probability of inclusion in the sample) is independent of the variable of interest, then the sampling design is unconfounded or **noninformative**. If the sampling indicator depends only on the observed values of survey variables of interest, then the sampling mechanism is **ignorable**. In other words, as long as the unobserved data are not associated with the probability of inclusion in the sample, the design can be labeled as ignorable. Ignorable sampling mechanisms are closely related to the missing at random (MAR) concept (see Section 11.2.2).

Ignorable mechanisms are important because inferences about the parameter of interest can be made without explicitly modeling the sampling mechanism. So how does one make a complex design "ignorable" when using a Bayesian inference approach to an analysis? Incorporating stratification and clustering is relatively straightforward in Bayesian models and is quite similar to how design effects are handled in multilevel modeling approaches. Effects of sampling strata are treated as fixed (or parameters to be estimated), and effects of sampling clusters are treated as random. Introducing these effects in Bayesian models makes the complex design ignorable. However, many modeling issues arise: There could be a loss of degrees of freedom if there are many strata; one could decide to include only "significant" stratum effects; one could choose to use covariates on which the strata were based rather than indicators for the strata themselves; and one has to posit a distribution for the random cluster effects, including an assumption of whether the random effects are independent of the residual errors. None of these issues are truly unique to Bayesian approaches, and they are all shared with more model-based multilevel modeling approaches (see Section 12.3).

Bayesian analysis methods are mathematically intensive and often require simulations to determine posterior distributions. A popular choice for estimating a posterior distribution is the Gibbs sampling technique (see Gelman et al., 2003, Section 11.3, for an overview).

The hardest aspect of performing Bayesian analyses of complex sample survey data is properly accounting for unequal probabilities of selection via sampling weights (Gelman, 2007). This is currently an active area of research. Current suggestions for maintaining the ignorability assumption when selection probabilities are unequal is to allow model parameters to vary as a function of selection probabilities (i.e., include interactions of predictors with selection probabilities in models) or to stratify by selection probabilities (Holt and Smith, 1979). Zheng and Little (2005) propose an estimation method based on penalized splines for estimating totals from PPS sampling

designs, where selection probabilities are known for all population members, and show promising properties of their method compared with more standard model-based approaches.

Because fully weighted estimators reduce bias at the cost of increasing sampling variance, a compromise is often required, and various **weight-trimming** approaches (Potter, 1990), or methods of reducing extreme weights without introducing severe bias (i.e., keeping the mean squared error relatively small), have been developed for the Bayesian analysis framework. Holt and Smith (1979) proposed a Bayesian approach to estimating means where the means of strata defined by sampling weights (or inverses of probabilities of selection) were treated as random variables (or random effects) and showed that this approach had the effect of "smoothing" the sampling weights (as opposed to weight trimming). Elliott and Little (2000) built on this approach for estimation of means, showing improved performance of a weight smoothing model based on a nonparametric spline function for the underlying means of the strata based on the sampling weights. Elliott (2007) expanded this weight-smoothing method to linear and generalized linear models.

Given the complications already introduced, statistical software for Bayesian analysis of complex sample survey data is still in its developmental stages. The Bayesian Inference Using Gibbs Sampling (BUGS)* software is currently a popular tool for Bayesian data analysis and is capable of fitting some of the models discussed in this section using a multilevel modeling approach (see Gelman and Hill, 2006, for nicely worked examples). Software attempting to implement the previously discussed weight-smoothing approaches is still being developed at the time of this writing.† We will provide links and updates on any developments in terms of statistical software for Bayesian analysis of complex sample survey data on the book Web site.

12.3 Generalized Linear Mixed Models (GLMMs) in Survey Data Analysis

12.3.1 Overview of Generalized Linear Mixed Models

Generalized linear mixed models comprise a broad class of regression-type models that include a number of specific analytic forms such as **hierarchical linear models (HLMs)** or **multilevel models**, models for **repeated measures** on individuals, or **growth curve models** of longitudinal change in individual attributes. We use the adjective *generalized* to acknowledge that linear mixed modeling techniques are now available for continuous, nominal,

* http://www.mrc-bsu.cam.ac.uk/bugs/
† http://www.sph.umich.edu/~mrelliot/trim/trim.htm

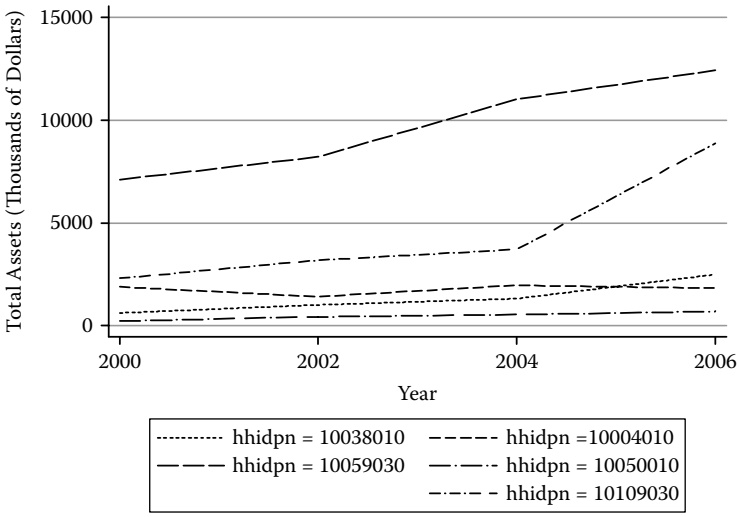


FIGURE 12.1
Multiple line plot showing trends in total household assets from 2000 to 2006 for a small subsample of five HRS 2006 households.

ordinal, and count-type data. Generalized linear mixed models are **subject-specific models** that analyze the influence of both **fixed effects** and **random effects** on individual attributes. The label *subject-specific* refers to the fact that GLMMs are modeling the effect on the individual unit (see Figure 12.1), explicitly controlling for the random effects of randomly sampled individuals or clusters of observations. This is in contrast to the **marginal modeling** or **population averaged** approach employed by the **generalized estimating equations** technique. This modeling technique is an alternative for the analysis of repeated measures or other dependent sets of observations on population units that does *not* explicitly incorporate random subject effects and adjusts standard errors for the clustering present in the observations in a manner similar to Taylor series linearization (TSL; using robust **sandwich estimates** of standard errors).

Following Fitzmaurice, Laird, and Ware (2004), a general expression for the GLMM is

$$g\{E(Y_{it} \mid \boldsymbol{b_i})\} = \eta_{it} = \boldsymbol{X_{it}\beta} + \boldsymbol{Z_{it}b_i} \tag{12.1}$$

where:

g(•) = a known link function, often identity (Normal) but also
log (Poisson) or Logistic (see Chapters 8 and 9);
$\boldsymbol{X}_{it}$ = row vector of $j = 1,\ldots,p$ covariates
for subject i, at observation $t = 1,\ldots,T$;

$\boldsymbol{\beta}$ = vector of $j = 1,\ldots,p$ fixed effects regression parameters;
$\boldsymbol{Z}_{it}$ = row vector of $k = 1,\ldots,q$ covariates
for subject i, at observation $t = 1,\ldots T$;
$\boldsymbol{b}_i$ = vector of $k = 1,\ldots q$ random effects.

From this generic expression, more exact expressions for individual models may be derived. For example, a simple **random intercept model** for change in a continuous random variable, y, over times $t = 1, \ldots, T$ can be written as follows:

$$\begin{aligned} y_{it} &= \boldsymbol{X}_{it}\boldsymbol{\beta} + b_i + e_{it} \\ &= (\beta_0 + b_i) + \beta_1 x_{it,1} + \cdots + \beta_p x_{it,p} + e_{it} \end{aligned} \tag{12.2}$$

where:
$b_i \sim N(0, \tau^2)$ = random effect for individual i;
$e_{it} \sim N(0, \sigma^2)$ = random error term; and
$\text{cov}(b_i, e_{it}) = 0$.

The notation in Equation 12.1 suggests a GLMM for modeling multiple dependent observations on a single individual, such as longitudinal measures in a panel survey (Fitzmaurice et al., 2009). However, by using slightly different subscripts, we can show that the GLMM also represents a model where the dependency is not among repeated observations on a single individual but among units that are clustered in a hierarchical data set. Consider for example a survey such as the National Assessment of Educational Progress (NAEP), where measures on individual students are nested within schools. The random intercept model can be rewritten to model fixed effects on individual student test performance, $y_{i(s)}$, while accounting for the random effects of the school clusters:

$$\begin{aligned} y_{i(s)} &= \boldsymbol{X}_{i(s)}\boldsymbol{\beta} + b_s + e_{i(s)} \\ &= (\beta_0 + b_s) + \beta_1 x_{i(s),1} + \cdots + \beta_p x_{i(s),p} + e_{i(s)} \end{aligned} \tag{12.3}$$

where:
$b_s \sim N(0, \tau^2)$ = random effect for school s;
$e_{i(s)} \sim N(0, \sigma^2)$ = random error term; and
$\text{cov}(b_s, e_{i(s)}) = 0$.

The literature generally refers to models such as Equation 12.3 that incorporate dependence among observations due to hierarchical clustering of the ultimate observational units as **hierarchical linear models** or **multilevel models** (Goldstein, 2003; Raudenbush and Bryk, 2002). Because the

dependence among observations results from hierarchical levels of clustering, HLMs are often described by decomposing the fixed and random components for the overall GLMM into models for the individual levels. For example, consider the simple model (Equation 12.3) for test scores where clusters of students are nested within schools. The level 1 model is the model of the student test outcome, controlling for the effects (which may be school specific) of covariates for that student. Note that the following level 1 model allows the intercept and the effects of the student-level covariates to be unique to the sampled schools:

Level 1:

$$y_{i(s)} = \beta_{0(s)} + \beta_{1(s)} x_{i(s),1} + \cdots + \beta_{p(s)} x_{i(s),p} + e_{i(s)} \tag{12.4}$$

Level 2 of the multilevel specification defines equations for each of the random school-specific coefficients in the level 1 model. In the case of Equation 12.3, only the intercept randomly varies across sampled schools. For example, suppose that the school-specific intercept is defined by the fixed overall intercept, β_0, and the random school effect, b_s, while the effects of the covariates are fixed across schools (i.e., defined by fixed effect parameters only):

Level 2:

$$\begin{aligned} \beta_{0(s)} &= \beta_0 + b_s \\ \beta_{1(s)} &= \beta_1 \\ &\cdots \\ \beta_{p(s)} &= \beta_p \end{aligned} \tag{12.5}$$

A complete treatment of generalized linear mixed models is beyond the scope of this intermediate-level text on applied survey data analysis. Many texts describe the theory and applications of GLMMs in detail, including (in no particular order) Diggle et al. (2002), Gelman and Hill (2007), Fitzmaurice et al. (2004), Verbeke and Molenberghs (2000), Molenberghs and Verbeke (2005), West, Welch, and Galecki (2007), Raudenbush and Bryk (2002), and Goldstein (2003). Our objective here is to acquaint the reader with generalized linear mixed models, to introduce several key issues in the application of these models to complex sample survey data, and to describe current research findings that address these issues. The introduction to these models and the discussion will be developed around a specific GLMM—a model of longitudinal change in the net worth of individual households. Because repeated measures and growth curve models of this type share their GLMM genes with HLMs and multilevel models, the analytic issues and recommended

solutions described in the following sections are easily extended to multilevel forms. Throughout the following discussion, we will emphasize the common roots by referring to individual observations at a specific time as level 1 observations and the individual respondent as the level 2 observational unit.

12.3.2 Generalized Linear Mixed Models and Complex Sample Survey Data

GLMMs are specifically designed to address nonindependence or "dependency" of observations. In HLMs the dependency arises because observational units are hierarchically clustered, such as students within classrooms and classrooms with schools. In repeated measures or longitudinal models, the dependency arises because observations are clustered within individuals—for example, daily diaries of food intake for National Health and Nutrition Examination Survey (NHANES) respondents, or longitudinal measures of household assets for HRS panel respondents.

This problem of lack of independence for observations should sound familiar. The design-based survey data analysis techniques that have been the core subject of this book are constructed to address the intraclass correlation among observations in sample clusters. It is natural to draw the analogy between these two estimation problems. Take, for example, two parallel sets of data. The first data set records scores on a test administered to students who are clustered within classrooms, schools, and school districts. The second data set includes test scores for the same standardized test administered to a probability sample of students in their homes, where households were selected within area segments and primary sampling units (PSUs) of a multistage national sample design. In both cases, there is a hierarchical ordering of units: districts, schools, classes, and students in the first case; and PSUs, area segments, households, and students in the second case.

If the survey analyst were interested only in inferences concerning national student performance on the standardized test, robust inferences could be obtained using the design-based estimates of population parameters (e.g., the mean test performance) and their standard errors. School districts would define the ultimate clusters in the first data set, and PSUs would form the ultimate clusters of the second. However, the survey analyst using the first data set may have broader analytic goals. Specifically, he or she may be interested in estimating the proportion of variance in test scores that is attributable to the student, the class, the school, and the district. The analyst working with the second data set may not be especially interested in the **components of variance** associated with PSUs, secondary sampling units (SSUs) within PSUs, households, and individuals. The first analyst will likely choose an HLM form of the generalized linear mixed model. The second analyst will be satisfied with a standard design-based regression analysis in which the

Taylor series linearization or replication estimates of overall sampling error are used to develop confidence intervals and test statistics.

Consider another problem in repeated measures analysis in which a simple random sample of individuals is asked to complete a daily diary of hay fever symptoms for each of $t = 1, \ldots, 7$ consecutive days. The dependent variable for each daily measurement is whether the subject experienced hay fever symptoms ($y_{it} = 1$) or did not ($y_{it} = 0$). The independent variables in the analysis might include age, gender, daily pollen count, and an indicator of previous allergy diagnosis. To estimate the relationship of hay fever symptoms to pollen count levels, the survey analyst could choose among three approaches that all use variants of logistic regression modeling:

1. A GLMM in which the logit[$P(y_{it} = 1)$] is modeled as a function of **fixed effects** that include the constant effects of age, gender and time t pollen count, and **random effects** of individuals in which the repeated diary measures are clustered.
2. A GEE model in which the logit model relating symptoms to the covariates is estimated using GEE methods and the covariance matrix for the model coefficients is separately estimated using the robust Huber-White **sandwich estimator** (Diggle et al., 2002).
3. A design-based analysis using a program such as Stata `svy: logit` with a single record for each daily report and the individual as the "cluster."

The GLMM analysis is a subject-specific analysis, explicitly controlling for the random subject effects, and would yield estimates of the fixed effects (or regression parameters) associated with the covariates *in addition to* estimates of the variances of the random subject effects. The GEE and `svy: logit` approaches are population-averaged (or marginal) modeling techniques and would provide comparable estimates of robust standard errors for the logistic regression coefficients for the covariates. The GEE and design-based population-averaged approaches would *not* separately estimate the variances of the random subject effects or their contributions to the total sampling variability. This is the key distinction between these alternative approaches to analyzing clustered or longitudinal data: GLMMs enable analysts to make inferences about between-subject (or between-cluster) variance, based on the variances of the subject-specific (or cluster-specific) random effects explicitly included in the models, while GEE and design-based modeling approaches are concerned only with overall estimates of parameters and their total sampling variance.

It is increasingly common to see survey samples designed to be optimal for analysis under a GLMM-type model. For example, a multistage national probability sample of school districts, schools, classrooms, and students could be consistent with a GLMM model that would enable the education researcher

to study the influence of each level in this hierarchy on student outcomes. To achieve data collection efficiency, the sampling statistician designing the sample of individuals for the hay fever study could select a primary stage sample of U.S. counties and then a clustered sample of individuals within each primary stage county. The result would be a three-level data set with days nested within individuals and individuals nested within counties.

However, even when care is taken to design a probability sample to support multilevel or longitudinal analysis using GLMM-type models, several theoretical and practical issues should be addressed in the application of these models to complex sample survey data:

1. *Stratification.* Using the terminology introduced in Section 12.2, if the stratification used in the sample selection (e.g., regions, urban/rural classification, population size) is informative—that is, associated with the survey variables of interest—the stratum identifiers, or at a minimum the major variables used to form the strata, will need to be included as fixed effects in the model.
2. *Clustering.* In the general sample design context, clustering of population elements serves to reduce the costs of data collection. The increase in sampling variance attributable to the intraclass correlation of characteristics within the cluster groupings is considered a "nuisance," inflating standard errors of estimates to no particular analytical benefit. In the context of a specific analysis involving a hierarchical linear (or multilevel) model, the clustering of observations is necessary to obtain stable estimates of the effects and variance components at each level of the hierarchy. Ideally, the sample design clusters may be integrated into the natural hierarchy of the GLMM, and the cluster effects on outcomes can be directly modeled using additional levels of random effects—say, nesting students within classrooms, classrooms within schools, and schools within county PSUs.
3. *Weighting.* Conceptually, one of the more difficult problems in the application of GLMMs to complex sample survey data is how to handle the survey weights. Theoretically, if the weights were noninformative, they could be safely omitted from the model estimation. If the sample design and associated weights are informative for the analysis, the variables used to build the weights or the weight values themselves could be included as fixed effects in the model. But there are other complications with weighting in GLMMs. Attrition adjustments that are often included in longitudinal survey weights can yield different weight values for each measurement time point. Clusters in a hierarchical sample design may not enter the sample with equal probability. Even in a national equal probability multistage sample of students, the most efficient samples require that

counties and schools are selected with PPS. Conditional on a given stage of sampling, the observed units would enter with varying probability.

The following section uses the example of estimating longitudinal change in individual attributes over time to illustrate some of the solutions that have been proposed for incorporating complex sample design effects into a GLMM analysis. Section 12.3.4 presents an example analysis in which longitudinal growth in HRS households' assets from 2000 to 2006 is modeled using a simple random intercept model.

12.3.3 GLMM Approaches to Analyzing Longitudinal Survey Data

A variety of longitudinal survey designs may be used to collect data over time. **Repeated cross-sectional surveys** such as the NHANES, **rotating panel designs** such as the Current Population Survey, and true **panel survey** designs such as the Panel Study of Income Dynamics (PSID; Hill, 1992) and the HRS (Juster and Suzman, 1995) all play an important role in understanding societal- and individual-level change over time. Binder (1998) and Kalton and Citro (1993) provide sound overviews of the key issues associated with longitudinal sample designs, both at the design and analysis stages. Menard (2008) and Lynn (2009) address design, measurement, and analysis for longitudinal surveys.

Panel surveys, with repeated measurements on individuals over multiple waves of data collection, enable longitudinal analysis of individual-level change over time. The simplest type of analysis is the analysis of individual change over two points in time denoted by t and $t - k$, for example,

$$\bar{\delta}_{t-k,t} = \frac{\sum_i w_i \cdot (y_{t,i} - y_{t-k,i})}{\sum_i w_i} \tag{12.6}$$

where w_i is a population weight for sample element i; or,

$$\hat{y}_{t,i} = \hat{\beta}_0 + \hat{\gamma} y_{t-k,i} + \hat{\beta}_1 x$$

where $\hat{\gamma}, \hat{\beta}_1$ are weighted estimates of regression parameters.

We considered an example of analyzing the mean change between two points in time in Example 5.13.

GLMMs can extend these simple "two-point" change models to examine trends in repeated measures or growth curves over multiple observations on the individual sample persons. Our aim in this section is to present a

synthesis of the existing literature on suggested approaches to using GLMMs for longitudinal analysis of complex sample survey data and to end with a practical example using existing statistical software that is capable of fitting some of the proposed models.

We first describe general notation for models that can be fitted to longitudinal survey data arising from complex sample designs. Let y_{it} be the response for individual i at time t, which may vary randomly over time. We assume that randomly sampled individual i has some "permanent" status that partly defines the observed responses, which can be denoted by b_i. These values for the randomly sampled individuals are assumed to arise from a normal distribution with a fixed mean (for the population) of 0 and a variance of τ^2. Then, observed values for y_{it} conditional on a randomly sampled value of b_i are assumed to arise from a normal distribution with mean $\beta_t + b_i$ (where β_t is the population mean of the response at time t) and variance σ^2. More succinctly,

$$\begin{aligned} b_i &\sim N(0, \tau^2) \\ y_{it} \mid b_i &\sim N(\beta_t + b_i, \sigma^2) \end{aligned} \tag{12.7}$$

Thinking in a multilevel modeling context, τ^2 represents between-individual variance, while σ^2 represents within-individual variance over time. The variance of y_{it} is thus partitioned into variance across individuals and transitory variance and can be written as $\tau^2 + \sigma^2$.

A simple model for an outcome y measured over time can then be written as follows:

$$y_{it} = \beta_t + b_i + e_{it} \tag{12.8}$$

In this model, $t = 1, \ldots, T$, where T represents the total number of waves (or panels), $i = 1, \ldots, n$ (persons), β_t is the population mean of the outcome at time t, b_i is a long-term difference from β_t for person i (a random variable corresponding to a random individual effect), and e_{it} is a transitory effect at time t for person i (each individual will randomly vary around his or her permanent state at time t). It is important to note that the e_{it} values will be correlated over time. One possible model expressing this correlation could be the first-order autoregressive [AR(1)] model:

$$e_{it} = \rho e_{i(t-1)} + \varepsilon_{it} \tag{12.9}$$

The model in Equation 12.9 can thus be rewritten as follows, using the AR(1) model for the transitory effects:

$$y_{it} = \beta_t + b_i + \rho e_{i(t-1)} + \varepsilon_{it} \tag{12.10}$$

In this specification, ε_{it} represents random measurement error, and the b_i and ε_{it} are assumed to be independent. We also assume that $E(b_i) = E(\varepsilon_{it}) = 0$, $var(b_i) = \tau^2$, $var(\varepsilon_{it}) = \sigma^2_\varepsilon$, and e_{it} and ε_{it} are mutually independent and stationary. We can then write

$$\text{var}(e_{it}) = \frac{\sigma^2_\varepsilon}{(1-\rho^2)} \tag{12.11}$$

We focus on a GLMM approach that has been developed to fit models of this form to longitudinal data from complex sample surveys (Skinner and Holmes, 2003).

To fit this model using GLMM methods, we must first address the question of how the longitudinal survey weights should be incorporated in the model estimation. Pfeffermann et al. (1998) laid the groundwork for appropriate methods of incorporating sampling weights into the estimation of parameters in multilevel models for complex sample survey data. They also proposed appropriate variance estimators (based on the delta method). The methods proposed by these authors can be used to estimate two-level multilevel models for longitudinal survey data given sampling weights computed for *both* the level 2 units (individuals) and the level 1 units (repeated observations). Theory Box 12.1 describes the procedure for constructing the weights for the individual levels.

Unfortunately, the multilevel modeling approach proposed by Pfeffermann et al. (1998) does not allow for autocorrelation of transitory effects within persons, as described in Equation 12.9 (the approach was described for the case where level 2 units were PSUs and level 1 units were sampled individuals), and our example in this chapter therefore does not consider this possible autocorrelation. Skinner and Holmes (2003) developed a method to circumvent this problem and enable the methods proposed by Pfeffermann et al. to be applied, but this method has yet to be programmed in any general purpose statistical software packages. A description of an alternative covariance structure modeling approach to serially correlated observations is provided in Theory Box 12.2.

Rabe-Hesketh and Skrondal (2006) further investigated methods of fitting GLMMs to complex sample survey data (and specifically binary outcomes). Based on their research, these authors stressed the importance in likelihood PMLE estimation (see Chapter 8) of incorporating appropriately scaled weights at *all* levels of a multilevel study, e.g., school-level, or level 2, weights and student-level, or level 1, weights, and not simply student-level weights). Addressing a limitation of the approaches proposed by Pfefferman et al. (1998) and Skinner and Holmes (2003), these authors also discussed

THEORY BOX 12.1 WEIGHTED ESTIMATION IN TWO-LEVEL GLMMS

Following Pfeffermann et al. (1998), the sampling weights for individuals at level 2 and repeated observations at level 1 are computed as follows. In general, given a longitudinal study design with these two levels, the appropriate weight for the data from individual i at time t, w_{it}, is computed as

$$w_{it} = \frac{1}{\pi_i \pi_{t|i}} \tag{12.12}$$

where π_i represents individual i's original probability of inclusion in the sample, and $\pi_{t|i}$ represents the probability that individual i responds at time t. The $\pi_{t|i}$ values might be estimated using a propensity modeling approach (e.g., Lepkowski and Couper, 2002), where the probability of responding at time t is predicted using a logistic regression model with predictors from previous waves. We assume that some reasonable methodology has been used for estimation of these response probabilities in this discussion.

For purposes of the multilevel modeling approach discussed in this section, define the **level 2 weight** as

$$w_i = \frac{1}{\pi_i} \tag{12.13}$$

and define the **level 1 weight** as

$$w_{t|i} = \frac{1}{\pi_{t|i}} \tag{12.14}$$

Then, we have

$$w_{it} = w_i w_{t|i} \tag{12.15}$$

and

$$w_{t|i} = \frac{w_{it}}{w_i} \tag{12.16}$$

A reasonable approach suggested by Skinner and Holmes (2003) is to let $w_i = w_{i1}$, or to let the level 2 weight for individual i be the inverse of the probability of being selected for the sample *and* responding at the

baseline wave. Then, $w_{1|i} = w_{i1}/w_{i1} = 1$. This procedure will result in less variance in the level 1 weights denoted by $w_{t|i}$.

The computed level 1 weights can then be scaled, as recommended by Pfeffermann et al. (1998) and Skinner and Holmes (2003) to minimize small-sample estimation bias. Scaling weights (e.g., normalizing weights; see Chapter 4) generally does not have an impact on the estimates of parameters in *multivariate* models, but this is not the case when fitting *multilevel* models to complex sample survey data. Skinner and Holmes (p. 212) suggest rescaling the level 1 weights previously described as follows:

$$w^{*}_{t|i} = \frac{t^{*}(i) w_{t|i}}{\sum_{t=1}^{t^{*}(i)} w_{t|i}} \tag{12.17}$$

In this notation, $t^{*}(i)$ represents the last wave at which individual i provides a response. Note that this method ensures that the average rescaled weight for a given individual i is equal to 1. Rabe-Hesketh and Skrondal (2006) outline additional options for scaling the level 1 weights, one of which we consider in the example in this section.

the computation of "sandwich" estimates of standard errors (based on a linearization approach) to incorporate stratification and primary stage clustering of a complex sample into the calculation of design-based standard errors for the parameter estimates derived using the PML approach. Their proposed method is implemented in the Generalized Linear Latent and Mixed Models (gllamm) program that these authors developed for the Stata system (visit http://www.gllamm.org). The underlying assumption of this approach is that the primary stage clusters represent the highest-level units in the multilevel design (e.g., students at level 1 nested within schools at level 2, and schools at level 2 nested with primary stage clusters at level 3). Via simulation, these authors found good coverage for confidence intervals based on the robust standard errors. Thinking about multilevel models for longitudinal survey data, we can apply the same models where the repeated observations across waves define the level 1 observations, the sampled individuals define level 2, and the primary sampling units define level 3 of the data hierarchy. We consider an example of this type later in Section 12.3.4.

Skinner and Vieira (2007) specifically focused on the impacts of cluster sampling on variance estimation when analyzing longitudinal survey data with multilevel models. They concluded that the simple method of including a random cluster effect in a three-level multilevel model for longitudinal

THEORY BOX 12.2 THE COVARIANCE STRUCTURE MODELING APPROACH TO SERIALLY CORRELATED OBSERVATIONS OF SKINNER AND HOLMES (2003)

In this section we briefly describe a covariance structure modeling approach to fitting models to longitudinal data collected from complex sample surveys. Recall the basic model specified in Equation 12.10, using the AR(1) model for the transitory effects:

$$y_{it} = \beta_t + u_i + \rho v_{i(t-1)} + \varepsilon_{it} \tag{12.18}$$

Now, let $y_i = (y_{i1},\ldots, y_{iT})'$ be a vector of observations on the dependent variable of interest collected on the i-th individual, assuming no nonresponse. Then, we have

$$E(y_i) = \beta = (\beta_1,\ldots,\beta_T)' \tag{12.19}$$

and

$$Var(y_i) = \tau^2 J_T + \sigma^2 V_T(\rho) \tag{12.20}$$

where J_T is a $T \times T$ matrix of 1s, and $V_T(\rho)$ is a $T \times T$ matrix with the (t,t') and (t',t) elements defined by $\rho^{(t'-t)}$, for $1 \le t \le t' \le T$.

This model therefore has T parameters for the mean of the outcome, but only three parameters for the variance of the outcome (σ^2_ε, σ^2_u, and ρ). Denote the set of three parameters by φ, and define the sample estimator of the variance–covariance matrix of the vector of observations on a given individual i as

$$\hat{S} = \sum_{i \in s} w_i (y_i - \bar{y})(y_i - \bar{y})' / \sum_{i \in s} w_i \tag{12.21}$$

The mean in Equation 12.21 is actually a vector of weighted estimates of means at the T time points. Then, let $\hat{A} = vech(\hat{S})$ or a vector containing the $T(T + 1)/2$ distinct elements of $\hat{S}$, and let $A(\varphi) = vech(Var(y_i))$, or a vector containing the $T(T + 1)/2$ distinct elements of the variance–covariance matrix for y_i defined in (12.20). Estimation of the parameters defining the variance–covariance matrix then proceeds by obtaining a generalized least squares (GLS) estimator of φ, denoted by $\hat{\varphi}_{GLS}$, as the iterative solution that minimizes

$$\left[\hat{A} - A(\varphi)\right]' \hat{V}^{-1}(\hat{A}) \left[\hat{A} - A(\varphi)\right] \tag{12.22}$$

where $\hat{V}(\hat{A})$ is a design-consistent estimator of the variance–covariance matrix of $\hat{A}$, possibly obtained using linearization or replication methods for variance estimation.

Given the GLS estimate of the set of variance–covariance parameters denoted by φ, a goodness-of-fit statistic is readily obtained as

$$X^2_W = \left[\hat{A} - A(\hat{\varphi}_{GLS})\right]' \hat{V}^{-1}(\hat{A}) \left[\hat{A} - A(\hat{\varphi}_{GLS})\right] \tag{12.23}$$

where, provided that the model is correct and the sample is large enough for $\hat{V}(\hat{A})$ to be a good approximation to the variance–covariance matrix of $\hat{A}$, then the goodness-of-fit statistic in Equation 12.23 will be distributed as a chi-squared statistic with $k - 3$ degrees of freedom, with $k = T(T + 1)/2$. It is important to note that $\hat{V}(\hat{A})$ may be unstable with a small number of primary sampling units in the design, and Skinner and Holmes (2003) present alternatives.

Skinner and Holmes (2003, pp. 209–210) also discuss two possible approaches to dealing with nonresponse due to attrition when using the covariance structure modeling approach. One approach involves considering only the attrition sample who responded at all points up to time T and developing adjusted weights denoted by w_{iT} that incorporate both the probability of selection and the probability of responding until time T. The previously described procedure can then be applied using the adjusted weights. These authors also describe a more sophisticated approach that we do not describe in detail here; interested readers can refer to the text for more details.

The covariance structure modeling approach offers the advantage of easily handling serial correlation, unlike the multilevel modeling approach, and provides users with a goodness-of-fit test. Unfortunately, the covariance structure approach described here has yet to be readily implemented in any general purpose statistical software packages. Further, Skinner and Holmes (2003) report evidence of bias in estimates of variance components when using the previously described GLS approach with the V matrix estimated from the data. We consider an example using the multilevel modeling approach in the example instead for these reasons, and we expect the covariance structure modeling approach to be more readily implemented in the near future.

survey data (where primary sampling units define level 3 of the multilevel study design, individuals define level 2, and the repeated measurements define level 1) has the potential to seriously underestimate effects of clustering on the standard errors of parameter estimates. In a practical application, model-based standard errors for key parameter estimates in the three-level model were nearly identical to those found in a two-level model *excluding* the random cluster effects, clearly suggesting that attempting to incorporate the effects of cluster sampling via random cluster effects and relying on model-based standard errors would be erroneous. These authors showed that calculation of "robust" standard errors (Goldstein, 2003, p. 80) in a three-level multilevel model including random cluster effects at level 3 resulted in standard errors that most closely resembled those computed by applying the classical linearization approach to the computation of design-based standard errors (incorporating clustering, and considered by the authors to be a "gold standard") for parameter estimates in a *two-level* model for the longitudinal data, including random individual effects only (a method similar to that proposed by Rabe-Hesketh and Skrondal, 2006). The example provided by these authors, however, did not consider the effects of weighting or stratification, and the authors acknowledged that additional work was needed (and ongoing) in this area.

Additional work by Vieira and Skinner (2008) found that point estimators for the regression parameters and variance–covariance parameters in a model for longitudinal survey data based on PML estimation, when combined with a variance estimator based on Taylor series linearization (allowing for clustering of the sample), had the best performance overall, providing more support for the work of Rabe-Hesketh and Skrondal (2006). However, their simulations considered only respondents at all possible waves of a panel survey ("complete" respondents, meaning respondents with any wave nonresponse were dropped) and once again ignored sampling weights that might be adjusted for other types of attrition or wave nonresponse. We note that the general PML estimation method and linearization variance estimation method proposed by the authors can accommodate single sampling weights for the individuals, possibly adjusted for other forms of nonresponse, which varies from the multilevel modeling approach accommodating weights at both level 2 (sampled individuals) and level 1 (repeated observations) of the longitudinal hierarchy.

12.3.4 Example: Longitudinal Analysis of the HRS Data

Based on suggestions for practice from this literature, the application presented in this section will use the `gllamm` procedure in Stata to fit a two-level multilevel model to the 2000, 2002, 2004, and 2006 HRS data, with individuals at level 2 and time points (or waves) at level 1, and will compute robust standard errors for the estimated parameters in the model based on the primary sampling units at level 3 of the data hierarchy. Appropriate

sampling weights for the level 2 individuals and the level 1 time points (where the level 1 weights will be rescaled) will be used to estimate the multilevel model. We note that analysts would need access to sampling weights representing the inverses of the probabilities of selection for the PSUs to appropriately incorporate random PSU effects at the third level of a multilevel model, and this information is often not available in panel survey data sets. Further, analysts are not often explicitly interested in estimating the variance between PSUs (and the implied correlation of individuals within PSUs), which is generally considered a nuisance parameter (but should still be accounted for when computing standard errors). For this reason we take the approach of Rabe-Hesketh and Skrondal (2006) and compute robust (or linearization-based) standard errors recognizing the clustering by PSU in the HRS data set.

In this example we consider an analysis of data from four waves of the Health and Retirement Study, collected in 2000, 2002, 2004, and 2006. Identified heads of households 50 years of age and older in the sampled financial units (or households) of the HRS sample that entered the sample prior to 2000 provided information on total assets for the household in each wave, and some wave nonresponse was present. We present a partial listing (observation numbers 5–20) of the necessary "vertical" structure of this type of longitudinal survey data file for analysis using the `gllamm` procedure:

```
list hhidpn stratum secu baseweight weight year ///
totalassets in 5/20
```

	hhidpn	stratum	secu	basewe~t	weight	year	totalassets
5.	10004010	1	2	4287	4733	1	1893656
6.	10004010	1	2	4287	4832	2	1420000
7.	10004010	1	2	4287	5111	3	1973000
8.	10004010	1	2	4287	5422	4	1832000
9.	10038010	2	2	4287	4733	1	625000
10.	10038010	2	2	4287	4832	2	1020000
11.	10038010	2	2	4287	5111	3	1325000
12.	10038010	2	2	4287	5422	4	2500000
13.	10050010	2	2	4923	5384	1	230000
14.	10050010	2	2	4923	5997	2	434000
15.	10050010	2	2	4923	5908	3	544200
16.	10050010	2	2	4923	5564	4	701871.63
17.	10059030	2	2	4613	5287	1	7112737
18.	10059030	2	2	4613	5459	2	8230000
19.	10059030	2	2	4613	5819	3	11030000
20.	10059030	2	2	4613	5426	4	12435000

Note that there are four rows per household in this data structure, containing a year identifier (e.g., 1 = 2000, 2 = 2002), complex design information (STRATUM and SECU), and information on the weights (described below).

For analysis purposes, we recode the YEAR variable to represent number of years since 2000 (YRSSINCE00), taking on values 0, 2, 4, and 6. We also focus our analysis on those HRS sample respondents with the highest education level (EDCAT = 4) and rescale the total assets variable to be measured in thousands of dollars:

```
gen yrssince00 = 0 if year == 1
replace yrssince00 = 2 if year == 2
replace yrssince00 = 4 if year == 3
replace yrssince00 = 6 if year == 4
keep if edcat == 4
gen totassets000 = totalassets / 1000
```

After this initial data management, we can make use of Stata's graphical capabilities for panel data to examine an initial **multiple line plot** for a small subsample of HRS households with the highest possible education classification. We generate the plot in Figure 12.1 using the following two commands (additional editing was performed using Stata's graph editor):

```
xtset hhidpn year
xtline totassets000 if hhidpn <= 10200000, overlay ///
ytitle(Total Assets (Thousands of Dollars)) ttitle(Year)
```

From this small subsample of households, we definitely see substantial between-household variance in the mean values of total assets (suggesting the inclusion of a random intercept in a multilevel model for the data), in addition to evidence of a slightly increasing trend in total assets across these years.

Per the HRS documentation (Heeringa and Connor, 1995), sampling weights representing inverses of the products of the probability of being sampled and the probability of responding at a given wave (see Equations 12.13 through 12.15 for the calculation method) were computed for responding households at each wave to account for wave nonrespondents (this calculation defined the WEIGHT variable in the previous listing). We define the base sampling weight (or level 2 weight) for each individual household to be the household's sampling weight in the first year that he or she entered the HRS panel data file (which we rename to be BASEWEIGHT). We can then compute the level 1 weights for purposes of the multilevel analysis by dividing the wave-specific weights by the base sampling weight, per Equation 12.16:

```
gen l1weight = weight / baseweight
```

Next, we rescale the level 1 weights, using "Method 1" of Rabe-Hesketh and Skrondal (2006):

```
gen sqw = l1weight^2
egen sumsqw = sum(sqw), by(hhidpn)
egen sumw = sum(l1weight), by(hhidpn)
gen l1weight_r = l1weight * sumw/sumsqw
```

The `gllamm` procedure requires a unique PSU identifier for the purpose of computing robust standard errors, and, as is common in many public-use data sets, the sampling error computation units in the HRS data file are coded with a value of 1 or 2 within each stratum. We therefore also compute a new version of the variable containing the sampling error computation units in the HRS data file, combining the stratum information with the SECU information:

```
gen newsecu = stratum * 100 + secu
```

We are now ready to fit the multilevel model of interest using the `gllamm` procedure. We wish to test for the presence of a linear trend in household total assets and to examine between-household variance in the intercepts. We can fit this model by using the following `gllamm` command:

```
gen pwt2 = baseweight
gen pwt1 = l1weight_r
xi: gllamm totassets000 yrssince00, i(hhidpn) pweight(pwt) ///
adapt cluster(newsecu)
```

Note the creation of two new variables that the command can read (PWT2 and PWT1), containing the level 2 and level 1 weights, respectively. The `pweight` option with the PWT argument will cause `gllamm` to look for the specific variables PWT1 and PWT2 to identify the appropriate level 1 and level 2 weights (if a user were to supply the option `pweight(weight)`, `gllamm` would look for WEIGHT1 and WEIGHT2). The `cluster(newsecu)` option implements the robust standard error computation, and the `i(hhidpn)` option identifies households as level 2 units with associated random effects in the multilevel model. The `adapt` option indicates that **adaptive quadrature** should be used for estimation; for more details on this technique, interested readers can refer to the documentation for `gllamm` (Rabe–Hesketh, Skrondal, and Pickles, 2004). Submitting this command results in the following Stata output:

```
Running adaptive quadrature
Iteration 0:   log likelihood = -1.104e+08
Iteration 1:   log likelihood = -1.099e+08
Iteration 2:   log likelihood = -1.092e+08
Iteration 3:   log likelihood = -1.091e+08
Iteration 4:   log likelihood = -1.091e+08
Iteration 5:   log likelihood = -1.091e+08
```

```
Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -1.091e+08
Iteration 1:   log likelihood = -1.091e+08 (backed up)
Iteration 2:   log likelihood = -1.091e+08
Iteration 3:   log likelihood = -1.091e+08
number of level 1 units = 4350
number of level 2 units = 1112
Condition Number = 6410.0376
gllamm model
log likelihood = -1.091e+08
Robust standard errors for clustered data: cluster(newsecu)
totassets000  Coef      Std. Err.  z     P>|z|  [95% Conf.  Interval]
yrssince00    81.73993  13.04773   6.26  0.000  56.16684    107.313
_cons         741.527   78.63583   9.43  0.000  587.4036    895.6503
Variance at level 1
4402531 (1464316.4)
Variances and covariances of random effects
***level 2 (hhidpn)
var(1): 3510574.4 (1524032.8)
```

From this output, we see that the estimate of the parameter associated with the number of years since 2000 is 81.74, suggesting that for those persons in this older adult population *with the highest possible education*, each additional year after 2000 resulted in an expected increase of approximately $82,000 in total assets. The procedure produces simple z-test statistics by dividing the parameter estimates by their robust standard errors (which are adjusted for the clustering of the complex HRS sample design) and refers the test statistic to a standard normal distribution to gauge the significance of the parameter (i.e., test whether the parameter is equal to 0). We therefore have fairly strong evidence in favor of a positive linear trend for this subgroup ($p < 0.001$).

Examining the estimates of the variance components, we see that the estimated variance of the intercepts at level 2 of this model is 3,510,574, with a standard error of 1,524,032; this would suggest that significant variance between households exists in terms of expected values of total assets, which was evident in Figure 12.1. Standard multilevel modeling techniques could be pursued at this point in an effort to explain portions of this unexplained between-household variance; see West et al. (2007) for some possible strategies.

Next, to illustrate an alternative design-based approach to analyzing these data (which is similar to a marginal modeling approach using the GEE technique), we generate a new weight called WGT1_2, which is the product of the level 1 and level 2 weights (per Equation 12.15), and respecify the `svyset` command for an analogous `svy: regress` analysis of these same variables:

```
gen wgt1_2 = pwt1*pwt2
svyset newsecu [pweight = wgt1_2]
svy: regress totassets000 yrssince00
```

Submitting these commands in Stata generates the following output:

```
Survey: Linear Regression
Number of strata = 1                      Number of obs = 4447
Number of PSUs = 104                      Population size = 11793472
                                          Design df = 103
                                          F( 1, 103) = 39.74
                                          Prob > F = 0.0000
                                          R-squared = 0.0043
                    Linearized
totassets000  Coef.      Std. Err.  t     P>|t|  [95% Conf.  Interval]
yrssince00    83.84876   13.30053   6.30  0.000  57.4703     110.2272
_cons         744.0301   79.56743   9.35  0.000  586.2269    901.8333
```

The number of observations displayed in the `gllamm` and `svy: regress` output differ due to the manner in which cases with non-zero weights are counted. Please see the stata `svy: regress` and `gllamm` documentation for details. This example demonstrates that a classical design-based approach to this panel survey analysis, ignoring the clustering of observations within individuals and using Taylor series linearization to compute robust, design-adjusted standard errors based on the primary sampling units only (individuals can be thought of as secondary sampling units in the panel study), yields very similar inferences. The primary advantage of the GLMM approach over this population-averaged approach is the additional inference that is possible regarding between-subject variance at a lower level of the multistage sample design.

We have presented a very simple example of getting started with using the `gllamm` procedure to fit multilevel models to longitudinal survey data sets arising from complex sample designs. Readers interested in a similar example for cross-sectional data from a complex sample survey are referred to Rabe-Hesketh and Skrondal (2006). The `gllamm` documentation (http://www.gllamm.org) describes much more detailed applications and additional features of this very powerful software procedure. The HLM (http://www.ssicentral.com/hlm) and MLwiN (http://www.cmm.bristol.ac.uk) software packages also enable analysts to fit multilevel models to complex sample survey data using sampling weights at multiple levels of the study hierarchy (provided that these weights are available from the data producer). Given the recent rapid advances in the development of statistical software for survey data analysis, we are optimistic that the methodology discussed in the current literature for analysis of longitudinal survey data will be more widely implemented in the near future.

12.3.5 Directions for Future Research

Research on statistical models for longitudinal data collected from complex sample surveys is likely to accelerate in the near future. The overview of current work presented in this chapter has identified several important areas where additional statistical development is likely, including the following:

- Incorporating improved methods of adjusting for patterns of wave nonresponse that deviate from pure attrition (either via enhanced weighting procedures or imputation procedures, e.g., Kalton, 1986) into the modeling approaches.
- Development of enhanced statistical software that makes the theoretical approaches easier for analysts to use.
- Enhancing the proposed modeling approaches to accommodate a wider range of covariance structures for longitudinal data.
- Adjusting standard errors for the *estimation* of covariance parameters in Skinner and Holmes's (2003) two-stage modeling approach.
- Development of methods for assessing model fit and model diagnostics.

12.4 Fitting Structural Equation Models to Complex Sample Survey Data

Analysts of survey data, and particularly social scientists, are often interested in fitting **structural equation models (SEMs)** to the survey data, exploring complicated direct and indirect relationships between variables, and testing complex theoretical models that may be causal in nature. There are a number of complete texts devoted to structural equation modeling, latent variable modeling, and related methods (including **confirmatory factor analysis**, or CFA, and **exploratory factor analysis**, or EFA). For a good practical introduction to these techniques, we recommend Kline (2004) or Schumacker and Lomax (2004). For more of a theoretical background on SEMs, readers can turn to the classic text from Bollen (1989).

Relatively recent statistical and technical developments have enabled analysts to fit structural equation models to complex sample survey data sets using readily available statistical software and to make inferences based on those models that correctly adjust for complex sampling features. Much of the pioneering work in this area has been done by the authors of the **Mplus statistical software** (http://www.statmodel.com). Some of the most fundamental work introducing methods for fitting SEMs to complex sample survey data can be found in Muthén and Satorra (1995), which established a

theoretical framework for fitting design-adjusted models and for making inferences based on those models. The methods introduced in that paper have been implemented in the Mplus software, and interested readers can find multiple examples of using Mplus to fit these models in Chapter 9 of the Mplus documentation (Muthén and Muthén, 1998–2007).

Several technical papers on the topic of fitting SEMs to complex sample survey data are available at http://www.statmodel.com. For readers interested in recently published examples of these types of applications using the Mplus software, we suggest Zajacova, Dowd, and Aiello (2009), Stapleton (2008), or Striegel-Moore et al. (2008). We aim to provide additional examples of using Mplus to fit these models and additional published applications on the book Web site.

12.5 Small Area Estimation and Complex Sample Survey Data

Chapters 4 and 5 introduced estimation and inference for subclasses of the survey population. Included under the general umbrella of subclass analysis is estimation for geographic domains of the survey population. Surveys such as the NCS-R, NHANES, and HRS generally support separate estimation and analysis for large geographic areas such as census regions or divisions; however, the potential to directly estimate statistics for smaller geographic units such as individual states, counties, or municipalities is limited by poor precision due to the size and distribution of the multistage sample design. In cases such as the estimation of unemployment rates where direct state- or metropolitan-level estimates are important, major survey programs such as the U.S. Current Population Survey expand the stratification, primary stage sample allocation, and overall sample size of the survey to enable stand-alone estimation for these critical geographic reporting units.

Over the past 50 years, a class of statistical techniques labeled **small area estimation methods** (Rao, 2003) has been developed to address the problem of estimating population characteristics for small areas. Most of these techniques blend survey data, ancillary population data, and a population model to estimate small area statistics. For larger domains, methods such as generalized regression estimation or calibration methods (DeVille and Särndal, 1992) simply use population data to adjust direct regression estimates computed from the survey data. Other techniques such as synthetic estimation or structure preserving estimation use survey data to model the structure of relationships in the survey population and then combine that structure with administrative or population data for the target small area to generate a small area estimate. Additional techniques such as composite estimation or the James-Stein method create an estimate that is a weighted function of

a direct survey estimate for the small area and a model-based estimate or a direct estimate for a larger domain.

Recent advances in methodology for small area estimation have centered on the application of generalized linear mixed models of the form in Equation 12.1 and Bayesian methods (see Section 12.2). Interested readers are encouraged to examine Rao (2003) for a comprehensive treatment of various area-level and unit-level GLMM models for small-area estimation.

12.6 Nonparametric Methods for Complex Sample Survey Data

Another area that is (at present) wide open for additional research is the application of nonparametric statistical methods to the analysis of complex sample survey data. Breidt and Opsomer (2009) provide an excellent overview of current developments in theory and methods, and we highlight some important recent work in this section.

In Chapter 11 of *Analysis of Survey Data*, Chambers, Dorfman, and Sverchkov (2003) provide a detailed but largely theoretical discussion of possible methods for incorporating complex sample design features into nonparametric regression methods designed for exploratory data analysis, such as localized **smoothing models**. Their focus is on "scatterplot smoothing," or modeling the functional relationship between some dependent variable y and some independent variable x. The authors perform a series of simulations evaluating different methods for incorporating complex sample designs into smoothing methods, but results were not clear in terms of optimal methods or whether sampling weights should be incorporated in the analyses. Further, only the case of stratification was considered, and the authors indicate that extensions of their theory for cluster sampling designs are needed. The authors conclude that their proposed design adjustments to standard smoothing methods can be beneficial when sampling schemes are **informative** (i.e., probability of selection is related to the survey variable of interest) and that there would be little loss of efficiency when using their suggested design adjustments if a sampling scheme was not informative.

Opsomer and Miller (2005) provide additional guidance on the amount of smoothing that should be done in nonparametric regression applications with survey data, and Harms and Duchesne (2009) present an evaluation of design adjustments for the local linear regression estimator.

In terms of **nonparametric density estimation**, which is used for comparing population subgroups in terms of distributions on continuous survey variables, estimating quantiles, and generating smooth estimates of density functions without relying on parametric models, Buskirk and Lohr (2005) examined the finite-sample and asymptotic properties of a modified kernel density estimator incorporating both sampling weights and kernel

weights. These authors presented regularity conditions under which the sample estimator of the density function is consistent and normal under various modes of inference used with sample survey data and also presented methods for incorporating clustering when computing design-based confidence bands for a density function. The authors present applications of their design-adjusted density estimation methods to two large survey data sets (NCVS and NHANES), but unfortunately their methodology has yet to be programmed in any of the more general-purpose statistical software packages.

Our hope is that these promising methods, which to date have largely been evaluated using theoretical simulations and only a handful of applications, will continue to be studied and will be more widely available for analysts of survey data as statistical software for survey data analysis continues to develop. As with other methods discussed in this book, we aim to update readers with any software or methodological developments on the book Web site.

Appendix A: Software Overview

A.1 Introduction

Software for the analysis of complex sample survey data has become increasingly available and sophisticated over the past decade. Most analysts can access procedures for survey data analysis either from within the standard procedures of major statistical software packages such as Stata, SAS/STAT (referred to as SAS from this point forward), SPSS, and SUDAAN (SAS-callable version) or via stand-alone or more specialized software such as SUDAAN (stand-alone version), Mplus, R, WesVar, and IVEware. In the overview and evaluation of software for survey data analysis presented in this appendix, we make a distinction between software packages with integrated survey procedures (Stata) and software that performs a more specialized subset of analysis such as survey data analysis or another analytic function along with survey data analysis (SUDAAN). The reason for this distinction is clear: Programs that cannot perform a full range of data management, analysis, graphing, and other tasks require more work and setup costs for the research analyst to perform survey data analyses correctly. Although this penalty is often minor, research analysts generally prefer to work efficiently and to minimize the amount of time and effort needed for data management and data set up prior to analysis.

For this reason as well as general availability, we focus on four software packages (Stata, SAS, SPSS, and SUDAAN) heavily used by a wide range of analysts working in varied fields. Although SUDAAN is primarily a data analysis package for complex sample survey data, the fact that it functions as both a stand-alone and SAS-callable tool essentially makes it a SAS-like product and therefore, by extension, is an add-on to the SAS suite of procedures and functionality. In addition to the four packages mentioned, we also present a brief summary of the survey data analysis capabilities in other software packages that are perhaps more secondary due to what they are primarily used for. These include IVEware, Mplus, R, and WesVar. We also include code samples for these software packages on our Web site. Strictly for illustration purposes, we present example code for each software procedure in this chapter, with the exception of Stata (which is highlighted in all other chapters). For those readers interested in detailed examples of the use

of procedures in SAS, SPSS, SUDAAN, IVEware, WesVar, R, and Mplus to replicate the examples presented in the book, please refer to the book Web site (http://www.isr.umich.edu/src/smp/asda/).

This appendix presents an overview of modern software procedures that correctly perform survey data analysis and provides guidance on the relative strengths and weaknesses of the software included in the textbook examples and on the accompanying Web site. We choose software based on general applicability for needed tasks, experience in teaching these tools, and overall availability on the market. Naturally, there are other excellent software tools not evaluated here, but for the majority of survey data analysis needs, the programs examined in this overview provide analysts with very good options. Prior to evaluation of current software, we present a short historical overview of software options for survey data analysis during the past 25–30 years. Many of these packages were developed by early researchers who recognized a need for survey data analysis tools and were unable to fulfill that need through general software options of that era.

We consider key features such as overall ability to analyze survey data (e.g., sample design structures accommodated), variance estimation methods available, range of analytic techniques offered, hypothesis testing ability, effectiveness in dealing with common issues such as subpopulation analysis, and overall availability. Although cost of software is an obvious consideration for most users, the wide variability in costs borne by users makes this type of analysis impractical from our perspective. Given anticipated version changes and rapid software development, we plan to regularly update our Web site examples as software developments emerge. Rather than recommend a particular tool as the "best" solution, our goal is to objectively evaluate and consider the pertinent features of the major software in use at this time. We also provide a general rationale for our emphasis on Stata as the main tool used throughout the text examples. Finally, we point out particular areas of survey data analysis in which each software package currently excels or falls short.

A.1.1 Historical Perspective

Prior to the late 1990s, most survey data analysts used specialized software tools for correct estimation of population parameters and variance estimation tasks. Some of the more popular tools were SUDAAN, IVEware, WesVar, and early Stata routines during the 1990s and older packages such as OSIRIS (University of Michigan, 1982), PC, and Super CARP (Fuller et al., 1989) in the 1970s and 1980s.

Only during the past decade have a majority of major statistical software producers incorporated options to correctly analyze survey data as part of their overall package. Prior to these developments, many users of software such as SAS and SPSS were motivated to program custom routines directly

through the use of standard coding or macro language coding. For jackknife repeated replication (JRR) or balanced repeated replication (BRR) routines, the effort required to code the routines is fairly minimal, and a user with a good understanding of the statistical issues and the software syntax could effectively implement a repeated replication method for general use.

The early survey data analysis tools were generally developed for the individual needs of a group of researchers analyzing survey data and were often very well designed. Indeed, each of the major software packages in use today owes a debt to the early software and work of the researchers who developed these tools and in turn promoted the use of correct analytic techniques for survey data analysis. However, most of these early packages required extra data management steps and data transfer from package to package, resulting in additional time demands on the analyst. With the advent of survey procedures included within major software packages, the survey analyst was able to easily do the necessary work without the extra burden of additional software.

A.1.2 Software for Sampling Error Estimation

The software reviewed in detail in this appendix includes four commonly used data analysis tools: SAS/STAT (Version 9.2); Stata (Version 10+); SPSS Complex Samples (Version 16); and SUDAAN (Version 9.0). This list is far from exhaustive but does include commonly used software used by a range of analysts. We limit our review to the PC platform, but many of these tools are also available on UNIX, LINUX, and Mac platforms as well. As mentioned earlier, additional software packages reviewed but in less detail are WesVar (version 4.3), R (version 3.16 of the `survey` package), Mplus (version 5.2), and IVEware. See the text Web site for code examples for all software discussed as well as links to external sites for software and statistical assistance.

For a systematic approach to selection of a survey data analysis tool, consider the software and respective features outlined in Table A.1 through Table A.4. For each software package, these tables present the types of complex sample design structures that can be specified, the available variance estimation methods, analytic procedures included in the package, key options such as hypothesis testing, subpopulation analysis, and whether the software will import multiply imputed data sets. Although these tables are not intended to provide readers with a complete list of every possible analytic technique, they include common tasks routinely performed by survey data analysts. Subsequent sections provide detailed reviews of the features summarized in Tables A.1 through A.4 for each software package, and available features will be updated on the Web site for the text as new developments occur.

TABLE A.1

Ability of Software to Accommodate Various Complex Sample Designs

	Stata	SAS	SUDAAN	SPSS	IVEware	WesVar	MPlus	R
Sample design								
With and without replacement	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Equal and unequal weighting	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nonstratified and stratified	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Single and multistage	Yes	No	Yes	Yes	No	Yes	Yes	Yes

TABLE A.2

Variance Estimation Capabilities and Additional Analysis Features of the Software Packages

	Stata	SAS	SUDAAN	SPSS	IVEware	WesVar	MPlus	R
Variance estimation method								
TSL (Taylor series)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
JRR (delete one)	Yes	Yes	Yes	No	Yes	Yes	No	Yes
JRR (replicate weights)	Yes	Yes	Yes	No	No	Yes	No	Yes
BRR	Yes	Yes	Yes	No	No	Yes	No	Yes
BRR (with Fay's adjustment)	Yes	Yes	Yes	No	No	Yes	No	Yes
Single cluster per stratum	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Features of survey analysis								
Subpopulation analysis	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Finite population correction	Yes	Yes	Yes	Yes	No	Yes	No	Yes

TABLE A.3

Available Analytic Techniques in the Software Packages

	Stata	SAS	SUDAAN	SPSS	IVEware	WesVar	MPlus	R
Analytic technique								
Descriptive								
Means	Yes	Yes	Yes	Yes	Yes	Yes	NA[a]	Yes
Totals	Yes	Yes	Yes	Yes	Yes	Yes	NA	Yes
Ratios	Yes	Yes	Yes	Yes	Yes	Yes	NA	Yes
Percentiles	No	Yes	Yes	No	No	Yes	NA	Yes
Contingency tables	Yes	Yes	Yes	Yes	Yes	Yes	NA	Yes
Regression								
Linear	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Binary logistic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ordinal logistic	Yes	Yes	Yes	Yes	Yes[b]	No	Yes	Yes
Multinomial logistic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Poisson regression	Yes	No	Yes	No	Yes	No	Yes	Yes
Probit	Yes	Yes	No	Yes	Yes[b]	No	Yes	Yes
Cloglog	Yes	Yes	No	Yes	Yes[b]	No	No	Yes
Survival analysis								
Cox proportional hazards model	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Kaplan–Meier estimation	Yes	Yes	Yes	Yes	No	No	No	Yes
Analysis of multiply imputed data sets	Yes	Yes	Yes	No	Yes	Yes[c]	Yes	Yes

[a] Not applicable; this product is designed as a modeling tool.
[b] Can be done using the SASMOD feature of IVEware.
[c] Limited to tables analysis only; not available for regression.

TABLE A.4

Hypothesis Testing Capabilities in the Software Packages

	Stata	SAS	SUDAAN	SPSS	IVEware	Wesvar	Mplus	R
Means, totals, and ratios								
Confidence limits and *t* test[a]	Yes	Yes	Yes	Yes	Yes	Yes	NA	Yes
Contingency tables								
Rao–Scott adjusted *F*[b]	Yes	Yes	Yes	Yes	No	Yes	NA	Yes
Wald and Pearson chi-square[b]	Yes	Yes	Yes	Yes	No	Yes	NA	Yes
Wald and Rao-Scott likelihood ratio[b]	Yes	Yes	Yes	Yes	No	No	NA	Yes
Regression								
Linear								
Linear contrasts[c]	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Wald *F* and adjusted *F*[d]	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Wald chi-square and adjusted chi-square[d]	Yes	Yes	Yes	Yes	No	No	Yes	No
Logistic and GEE (binary, ordinal, multinomial, Poisson, cloglog, probit)								
Linear contrasts[c]	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Wald *F* and adjusted *F*[d]	Yes	Yes	Yes	Yes	No	Yes	No	No

Continued

TABLE A.4 (*Continued*)

Hypothesis Testing Capabilities in the Software Packages

	Stata	SAS	SUDAAN	SPSS	IVEware	Wesvar	Mplus	R
Wald chi-square and adjusted chi-square[d]	Yes	Yes	Yes	Yes	No	No	Yes	Yes
Archer and Lemeshow GOF[e]	Yes	No	No	No	No	No	No	No
Test of parallel lines assumption (ordinal logistic regression)[f]	Yes	No	No	Yes	No	No	No	No
Survival analysis								
Cox PH model								
Linear contrasts[c]	Yes	No	Yes	Yes	No	No	Yes	Yes
Wald *F* and Wald chi-square[d]	Yes	No	Yes	Yes	No	No	Yes	Yes
Survival curves								
Corrected standard errors for estimates	No	No	Yes	Yes	No	No	No	Yes

[a] Test of means significantly different from 0 and test of difference between means (see Sections 5.6 and 5.6.1).
[b] Tests for independence of rows and columns in contingency table analysis or differences in population proportions (see Sections 6.4.3 and 6.4.4).
[c] Test of individual parameter = 0, test of linear contrasts between parameters in regression models, and,
[d] Test of multiple parameters in the model (overall or partial tests; see Section 7.3.4.1 and similar sections of Chapters 8–10).
[e] Test significance for goodness of fit (adjusted for complex sample design) for overall model (see Section 8.5.2).
[f] Tests assumption of proportional odds or "equal slopes" in ordinal logistic regression (adjusted for complex sample design; see Section 9.3.6).

A.2 Overview of Stata® Version 10+

Stata® (http://www.stata.com) is an excellent software tool for the survey data analyst. This package offers the ability to handle numerous sample design structures, a full range of variance estimation methods, extensive survey commands, the ability to perform correct subpopulation analyses for every command, and many theoretically advanced options for significance testing (e.g., second-order design corrections for chi-square statistics; see Chapter 6). Another valuable feature is the ability to analyze multiply imputed data sets.

A key feature in Stata is the convenient method of declaring the "survey" or complex design variables one time prior to analysis. This approach allows the analyst to specify the complex design variables and probability (or sampling) weights through use of the `svyset` command, and once they are declared, these variables will be in effect for the entire survey data analysis session or until changed. The `svyset` command allows users to specify different forms of sampling weights, such as probability or replicate weights, which is another indication of the flexibility and range of this software. In addition, there are options to declare variables for finite population corrections and explicit poststratification adjustments (when available) within the `svyset` command syntax.

In practice, analysts may encounter public release data sets that employ a sampling error calculation model (see Chapter 4) where each stratum includes only a single cluster for sampling error computations, as opposed to more traditional sampling error calculation models specifying two or more sampling error clusters per stratum. This can present problems in some software packages, as most survey data analysis software is designed to recognize multiple sampling error clusters in each stratum for variance estimation, and the "singleton" or "lonely" sampling error cluster can interfere with variance estimation procedures. Fortunately, this problem can be overcome through use of the `singleunit` option in the `svyset` command. This feature allows the Stata user to declare that the sampling error calculation model for the survey data set includes only one sampling error cluster in at least one of the sampling error strata and offers four options for how variance estimation should proceed in this situation (`singleunit(missing)`, `singleunit(certainty)`, `singleunit(scaled)`, and `singleunit centered`)). The `singleunit(missing)` option is the default option and simply causes the program to stop execution without reporting standard errors; `singleunit(certainty)` informs Stata to include the "singleton" primary sampling unit (PSU) with certainty (i.e., the PSU does not contribute to variance estimation), `singleunit(scaled)` uses the average sampling variance among all strata with multiple units for the single unit, and `singleunit(centered)`uses deviations from the grand mean across units for variance contributions from single units. See Chapter 4 and the Stata help function for `svyset` for more details on this topic.

Variance estimation methods available for each of Stata's survey data analysis procedures include the Taylor series linearization method, JRR, and BRR with an option for Fay's adjustment. The ability to use these varied approaches is a distinct advantage in situations that may call for a replicated variance estimation approach, such as JRR, rather than a linearization method. There are numerous reasons for a replication approach. For example, one situation that often arises with public use data files is the publication of "replicate weights" instead of the more typical first-stage stratum and cluster (a.k.a. sampling error computation unit, or SECU) variables in addition to a final sampling weight variable. This is generally due to confidentiality issues and the desire to publish "masked" variables for public use that represent the design structure without jeopardizing confidentiality. Stata can incorporate replicate weights within the `svyset` command, and this feature is one of numerous reasons why Stata is a preferred tool for many survey analysts. See Chapter 3 for more details on variance estimation methods.

The number of Stata survey commands is currently quite extensive. Although a naïve count of procedures does not fully explain the capability of a particular tool, this software offers an extensive list of survey data analysis procedures and options. In fact, every analytic command includes design-based significance testing options if statistically appropriate and tenable. An additional and perhaps less obvious advantage of the Stata software is the consistency of syntax between the specialized survey commands and more standard simple random sample commands, thus allowing a Stata user new to survey data analysis a nearly immediate knowledge of the syntax required to execute the accompanying survey commands. At the simplest level, a Stata user merely needs to add `svy:` in front of a standard analysis command to invoke design-based estimation, provided that the relevant design variables have been identified using the `svyset` command.

The ability to perform "unconditional" subpopulation analyses with correct handling of zero cells in the clusters that can result with a subpopulation analysis (i.e., sampling error clusters where respondents from the subpopulation do not appear in the sample) is yet another common issue for survey data analysts. This issue is important due to the effect that zero cells in the clusters will have on variance estimation and the degrees of freedom used by the program for significance testing. As previously detailed in Chapter 4, this issue can be a "thorn in the side" for analysts, and the manner in which the zero cells are handled differs substantially across the common software packages evaluated in this chapter. All of Stata's survey procedures provide users with an accurate and convenient approach to the analysis of subpopulations via the `subpop()` option within the survey command syntax. An unconditional subpopulation analysis like that invoked via use of the `subpop()` option is almost always the correct approach for a subpopulation analysis, ensuring that (1) the full complex design structure is recognized during the process of variance estimation, and (2) the subpopulation sample size is treated as a random variable. See Section 4.5 for more details.

Given the array of Stata survey procedures and the manner in which the commands are organized (one for each separate type of analysis instead of a single procedure/command that does many things), it would be impractical to list all of the survey commands here. In fact, even the highlights of the various survey commands are extensive. In terms of descriptive analyses (see Chapter 5 and Chapter 6), users can estimate means, proportions, ratios, totals, and either one-way frequency tables or two-way cross-tabulations. In terms of linear regression modeling (see Chapter 7), linear regression for continuous outcomes and a number of related techniques such as constrained linear regression and Tobit regression are available. For binary outcome variables, logistic and probit regression modeling procedures are available, along with a number of convenient options such as skewed logistic regression and complementary log-log regression. For categorical outcomes, ordered logistic, probit regression, multinomial logistic, conditional logistic, and stereotype logistic regression can be performed. In terms of count outcomes, Stata offers Poisson regression, negative binomial regression, and generalized negative binomial regression along with zero-inflated or zero-truncated options as appropriate for these models. Analysts working with time-to-event (or survival) survey data can also fit design-based Cox proportional hazards models or other parametric survival models, which is a nice feature that is not currently widely available across the major statistical software packages (see Chapter 10).

Each of these survey commands has additional analytic options available (many of which are detailed in the chapters of this book). In addition, a variety of **postestimation commands** are also available after executing the commands, such as those that compute design effects (DEFF), misspecification effects (MEFF), subgroup contrasts, design-adjusted Wald tests for multiple regression parameters, subpopulation sizes, and other useful statistics. For example, after fitting a linear regression model, residuals and model diagnostics are available postestimation, and examples of using these commands are presented in Chapter 7. Given that Stata code is presented throughout the text, we refrain from providing detailed examples of Stata commands here. Please see our Web site for sample programs and the Stata Web site (http://www.stata.com) for full documentation.

In summary, Stata offers an extensive range of design-based analysis capabilities, all common variance estimation methods, excellent subpopulation analysis tools, and numerous tools for design-adjusted hypothesis testing. Stata 11 has introduced new multiple imputation capabilities, but we were unable to review these commands in detail prior to publication. Relevant updates in this regard and links to additional materials will be provided on the book Web site, but this brand-new command enabling both multiple imputation and multiple imputation analysis provides a nice addition for analysts of survey data who are faced with missing data problems. We discuss the use of the existing `ice` command and `mim` modifier for missing data problems in Chapter 11.

A.3 Overview of SAS® Version 9.2

The current version of SAS as of publication (Version 9.2, http://www.sas.com) offers five survey procedures composed of four analysis survey procedures and one sample selection procedure. There are numerous sampling/analytic techniques within each procedure rather than the one command per analytic technique approach implemented by Stata. This is essentially a difference in style and structure and does not necessarily represent a loss of capability overall. Although SAS does not offer as many survey procedures or analytic techniques when compared with Stata, it can perform many common complex sample survey analyses using either repeated replication or Taylor series methods for variance estimation.

The five procedures currently offered for sample selection and survey data analysis are PROC SURVEYSELECT, PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, and PROC SURVEYLOGISTIC. PROC SURVEYSELECT is primarily a sampling tool offering the ability to develop probability samples. Because this is not an analysis procedure we do not focus on PROC SURVEYSELECT in this appendix. PROC SURVEYMEANS and PROC SURVEYFREQ offer descriptive analytic capabilities for means, ratios, totals, contingency tables, and other types of descriptive analyses. For regression analysis, the two main procedures are PROC SURVEYREG (for linear regression) and PROC SURVEYLOGISTIC (binary, ordinal, and multinomial logistic regression). Starting in SAS 9.2, each of the four survey analysis procedures offered the option of JRR, BRR, or Taylor series for variance estimation as well as a DOMAIN statement for correct subpopulation analysis (see Chapter 4). SAS software is also capable of analyzing multiply imputed data sets through use of the survey procedures in combination with PROC MI/MIANALYZE. Additionally, a new feature in SAS 9.2 is the NOMCAR option for handling missing data. This option treats missing data as a separate domain or subpopulation rather than excluding the missing cases from the analysis and is available for all survey procedures (see the SAS 9.2 documentation for more details).

By default, the SAS survey analysis procedures do not allow for a single PSU per stratum when performing variance estimation and will collapse all strata with a single PSU for the variance estimation step. Alternatively, the analyst can opt not to collapse all single PSUs into one stratum for variance estimation via use of the "NOCOLLAPSE" option in PROC SURVEYREG only. In this situation, the single PSU contributes a 0 for variance estimation.

The SAS survey analysis procedures can analyze data arising from various types of complex samples using the sampling with replacement, "ultimate cluster" sampling error calculation model, and formulas described in Chapter 3. The design is defined in each procedure using the STRATUM and CLUSTER statements. The sample weight variable is declared through use of the WEIGHT statement. These statements are required for the proper use

of the SAS survey analysis procedures, given that a complex sample design involves stratification, clustering, and weighting. For example, typical PROC SURVEYMEANS syntax is as follows:

```
proc surveymeans data=one;
stratum str;
cluster cluster;
weight weight;
var income;
run;
```

Several useful survey analysis features were introduced for these procedures in SAS 9.2. For example, replicate weights could be accommodated by declaring the replicate weights in the WEIGHT statement. The analyst could also specify the variance estimation method in the procedure statement through use of the `varmethod=jackknife` or `varmethod=brr` options (the default variance estimation method for each analysis procedure is Taylor series linearization). The following example syntax illustrates typical use of SAS PROC SURVEYMEANS with a JRR variance estimation method and use of four replicate weights:

```
proc surveymeans varmethod=jackknife data=one;
repweights rwgt1 rwgt2 rwgt3 rwgt4;
var income;
run;
```

Subpopulation analyses can be correctly performed in all SAS survey analysis procedures with the use of the DOMAIN statement or the "implied domain" statement (for PROC SURVEYFREQ). In the case of PROC SURVEYFREQ, use of the domain variable as the first variable in the TABLES statement functions like a DOMAIN statement. Multiply imputed data sets can be used with all SAS survey procedures with required output read into and analyzed by PROC MIANALYZE. Although PROC MIANALYZE is not considered a "survey" procedure, it is capable of analyzing output from each of the survey procedures outlined. Indeed, it is a key tool for the analysis of multiply imputed survey data. PROC MIANALYZE requires the design-based estimates of the standard errors or the variance–covariance matrix for proper analysis, and with this approach, the analyst can account for both the complex sample and the variability in estimates introduced by multiple imputation. See Chapter 11 for more on multiple imputation analysis for survey data and the SAS documentation for PROC MI/PROC MIANALYZE for more details.

A.3.1 The SAS SURVEY Procedures

The core analytic SAS procedures can be separated into descriptive procedures (PROC SURVEYMEANS and PROC SURVEYFREQ) and regression procedures

(PROC SURVEYREG and PROC SURVEYLOGISTIC). PROC SURVEYMEANS offers the ability to perform survey data analysis for means, means within subpopulations, percentiles, ratios, and totals, all taking into account the complex design variables and the sample weights, thus resulting in correct variance estimation. One common analytic technique is a significance test for a linear combination of means (e.g., a difference between means) including a corrected standard error for the difference. This feature is not currently included with PROC SURVEYMEANS but can be approached either with a SAS Institute submacro (smsub.sas; see http://support.sas.com) or with use of contrasts in PROC SURVEYREG (see the SAS documentation for PROC SURVEYREG for details). The PROC SURVEYMEANS default output includes an estimated mean, design-based standard error, sample size (n), and design-based 95% confidence limits for the mean. When using the ratio option in the procedure, the defaults change to comply with the usual output for a ratio rather than a mean (see the book Web site for examples). The following is basic PROC SURVEYMEANS syntax with a domain statement specified along with stratum, cluster, and weight variables defined (see the SAS documentation for more details and options):

```
proc surveymeans data=one;
stratum str;
cluster cluster;
weight weight;
var income;
domain sex;
run;
```

PROC SURVEYFREQ performs contingency table analyses including estimation of proportions and population totals with design-based standard errors. This procedure can also perform design-corrected bivariate tests of association and significance through use of the Rao–Scott chi-square test and the likelihood ratio test (Rao and Scott, 1984) and the Wald chi-square and log-linear chi-square test (see Chapter 6). The default output from PROC SURVEYFREQ includes weighted and unweighted frequency counts along with a standard error for the weighted frequency, weighted percentages and standard errors. To obtain the previously mentioned Rao–Scott or Wald tests and other options such as "DEFF" for design effects, the user can request these options within the procedure.

The SURVEYFREQ procedure includes an implied and optional domain feature rather than an explicit statement for subpopulation analyses, as illustrated in the following example syntax (where SEX defines the domain variable, EDUCAT defines the row variable of the contingency table, and MARCAT is the column variable):

```
proc surveyfreq data=one;
strata strata;
cluster cluster;
```

```
weight weight;
tables sex*educat*marcat;
run;
```

PROC SURVEYREG is the SAS tool for linear regression with survey data (see Chapter 7). This procedure computes weighted estimates of the parameters in specified linear regression models, along with design-corrected standard errors and variance–covariance matrices computed by the variance estimation method of choice. Subpopulation analyses can be defined via the DOMAIN statement (first possible in SAS 9.2). For hypothesis tests, the analyst can request linear contrasts and tests of model effects. Significance testing can be performed with the use of either the ESTIMATE or CONTRAST statement and custom hypothesis testing of linear combinations of regression parameters is available (see the book Web site for examples). The procedure also includes design-based Wald *F*-tests for regression parameters, design effects, and confidence limits for parameter estimates, along with many other options. See Chapter 7 for more details on variance estimation and significance testing for linear regression analysis of survey data. The following is a simple example of PROC SURVEYREG code:

```
proc surveyreg data=one;
strata strata;
cluster cluster;
weight weight;
model income=sex age educationyrs;
run;
```

The final SAS survey analysis procedure is PROC SURVEYLOGISTIC, which performs logistic regression for discrete categorical outcomes, including binary, ordinal, and nominal dependent variables. This procedure uses the pseudo maximum likelihood method of parameter estimation (Binder, 1983) along with the usual variance estimation options, an optional DOMAIN statement, and significance testing options for survey data analysis. Multiparameter hypothesis tests are available through the use of the TEST or CONTRAST statements, which operate on the design-adjusted variance–covariance matrix of the parameter estimates for tests of hypotheses. See Chapters 8 and 9 for statistical details on fitting these models and relevant adjustments for complex sample survey data. Various extensions of logistic regression modeling are available through use of the "link" option on the model statement. The four types of links are cloglog, glogit, logit, and probit. The following is generic syntax for PROC SURVEYLOGISTIC with the default logit link (more examples are available on the book Web site):

```
proc surveylogistic data=one;
strata strata;
cluster cluster;
```

```
weight weight;
model diabetes (event='1')=sex age educat;
run;
```

A summary evaluation of SAS 9.2 and its survey procedures indicates a good range of designs allowed, a newly expanded range of variance estimation methods, subpopulation analysis options for all procedures, and an option for treating missing data as "not missing at random" when using the Taylor series method. Particularly with the improvements introduced in version 9.2, SAS is a strong survey data analysis tool.

A.4 Overview of SUDAAN® Version 9.0

SUDAAN (http://www.rti.org/sudaan/) is another excellent tool for the survey data analyst. The history of this package is long and impressive. It began as a single-procedure tool during the early 1970s and has since evolved into a key software tool capable of analyzing many types of correlated data sets. The software can be executed as either a SAS-callable or stand-alone product and offers a wide range of features and commands. One important benefit of the SUDAAN package is that the syntax very closely mirrors that of SAS, so the experienced SAS user will find learning and invoking the program relatively easy. On the other hand, knowledge of SAS command syntax is not required to use the program to full effect as it is well organized with excellent documentation containing thorough language and example manuals.

SUDAAN features the ability to specify numerous sample designs for conducting design-based analysis (e.g., with and without replacement selection of sampling units), a good range of weights (probability and replicate weights), and the ability to handle single primary sampling units per stratum (the MISSUNIT option on the NEST statement). The SUDAAN MISSUNIT option informs the software of one primary sampling unit per stratum and invokes the use of deviations of values from "single" sampling error clusters from overall means as the contributions of those clusters to variance calculations (see the SUDAAN documentation for more details).

The three commonly used methods for variance estimation are included: Taylor series linearization, JRR, and BRR. The list of SUDAAN analysis commands is lengthy, with descriptive (PROC DESCRIPT, PROC CROSSTAB, PROC RATIO), regression (PROC REGRESS, PROC LOGIST/RLOGIST, PROC MULTILOG, PROC LOGLINK) and survival analysis (PROC KAPMEIER, PROC SURVIVAL) commands included. Each of the commands listed includes hypothesis testing tools appropriate for the corresponding analytic technique (see the SUDAAN language and example guides for complete

details). SUDAAN can also analyze multiply imputed data sets within any of its procedures.

One potential drawback of SUDAAN is the lack of data management tools, but, then again, this tool is not promoted as a general-purpose statistical software package. In practice, it is often used for the analysis of survey data rather than data set construction, and, given the ability to read in various types of external data sets and seamless use with SAS, this is not a major concern.

Another common task is production of graphics based on survey analyses, and though SUDAAN does not contain a graphics capability, its SAS interface enables easy movement between SUDAAN and SAS/GRAPH or SAS/ODS (Output Delivery System). The analyst need only save the SUDAAN output as a SAS type file and can then use the output data set with the key SAS tools for graphing, reporting, or further data analysis. Alternatively, the analyst could save output in another format such as an ASCII file and use the file with other software as needed.

A.4.1 The SUDAAN Procedures

There are three descriptive procedures in SUDAAN. The DESCRIPT procedure is designed for estimation of means, totals, proportions, percentages, and quantiles (see Chapter 5) with design-based standard errors estimated using the selected variance estimation method. The DESCRIPT procedure also includes the ability to perform linear contrasts of estimates for individual levels of categorical variables (e.g., testing the difference of means for two subpopulations). Complex design variables containing stratum and cluster codes that enable estimation of standard errors are declared in the NEST statement, and the optional design specification for the procedure statement allows the analyst to specify the appropriate sample design. The following example code illustrates SUDAAN syntax using the default of DESIGN=WR (With-Replacement selection of first-stage primary sampling units, or "ultimate clusters"; see Chapter 3), with the use of the FILETYPE=SAS for reading a SAS data set and the SUBPOPN statement for unconditional identification of a specific subpopulation for analysis. Subpopulation analysis is available on every SUDAAN procedure and provides correct subpopulation analyses similar to the Stata `subpop()` or the SAS DOMAIN statements:

```
proc descript data=one filetype=sas;
nest strata cluster;
weight weight;
var income;
subpopn sex=1;
run;
```

The RATIO procedure is designed to estimate weighted ratios of selected observed variables and includes options similar to those of the DESCRIPT

procedure. The PROC RATIO syntax is very similar to the preceding DESCRIPT commands but requires the addition of the numerator and denominator variables defining the ratio of interest:

```
proc ratio data=one;
nest strata cluster;
weight weight;
numer depressedwomen;
denom depressedtotal;
run;
```

The third descriptive procedure, PROC CROSSTAB, is used for estimation of frequency and contingency tables, percentages, odds ratios, and risk ratios, all with design-based estimates of variances and standard errors. The analyst can also obtain complex design adjusted chi-square and Cochran-Mantel-Haenszel (CMH) tests (see Chapter 6). SUDAAN also offers design-corrected CMH tests for trends, a convenient tool for analysts working with categorical variables.

Sample PROC CROSSTAB syntax presented next includes the usual design variable specification in the NEST statement and the tables needed for the analysis among the subpopulation of women (SEX = 1). This example also includes the TEST statement and requests a design-based chi-square test of association. The design-adjusted test statistics available for testing hypotheses about bivariate associations between categorical variables are numerous: Wald *F* and adjusted Wald *F*, Wald chi-square and adjusted chi-square, and a Satterthwaite test (Satterthwaite, 1946) as well (the existing literature does not provide overwhelming support for use of any one of these tests in all situations, but the adjusted tests generally work well; we recommend that analysts consider running all of these tests and comparing results):

```
proc crosstab data=one filetype=sas;
nest strata cluster;
weight weight;
class marcat;
subpopn sex=1;
tables depressed*marcat;
test chisq;
run;
```

The regression procedures include PROC REGRESS (used for linear regression), PROC LOGIST (RLOGIST for SAS-callable SUDAAN, to avoid confusion with SAS PROC LOGISTIC) for logistic regression with binary dependent variables, PROC MULTILOG for logistic regression with ordinal or nominal outcomes, and PROC LOGLINK for count outcomes. Each of the regression procedures of SUDAAN uses generalized estimating equations (GEE) for efficient estimation of robust and complex design corrected variances

and standard errors. (Refer to the SUDAAN documentation for details on the implementation of this technique.) Each regression procedure includes appropriate hypothesis testing capabilities as well as a wide range of printed output and ability to output data sets in various formats for further use.

Linear regression of complex sample survey data is implemented through PROC REGRESS in SUDAAN. This procedure outputs the usual weighted estimates of regression parameters with robust and complex design corrected standard errors as well as a design-corrected variance–covariance matrix, model testing, and parameter testing via the TEST statement. Each SUDAAN modeling procedure includes the ability to estimate linear contrasts of regression parameters by using either the CONTRAST or EFFECTS statements as well as optional tests of groups of parameters via the TEST statement. The REGRESS procedure also includes various printed output options such as design effects as well as output data sets for further use. Although the SUDAAN SAS-callable version runs within the SAS program, it does not interface with the SAS Output Delivery System directly but instead produces SAS-type output files. These files can then be used directly within SAS for further analysis.

One typical use of the SUDAAN output within SAS might be examination of regression diagnostics and associated graphics. Given the wide range of SAS graphics within SAS/GRAPH and the ODS, use of SUDAAN and SAS together can be a powerful combination for SAS/SUDAAN users. Following is syntax for a typical SUDAAN PROC REGRESS example, with a TEST statement (design-adjusted Wald *F*-test) to test the null hypothesis that the regression parameters for both AGE and SEX are equal to 0, and an OUTPUT statement for producing a SAS data file including estimated parameters (beta) and their standard errors:

```
proc regress data=one;
nest strata cluster;
weight weight;
model income=age sex;
test waldf;
effects sex age / name="test effects of age and sex";
output beta sebeta / filename=outestimate filetype=SAS;
run;
```

Logistic regression analysis is performed using either PROC LOGISTIC or PROC RLOGIST (for SAS-callable SUDAAN). This procedure is used for binary dependent variables and any types of predictor variables (binary, categorical, or continuous). Design and weight variables are declared in the usual manner for SUDAAN. Tests of contrasts or "chunk" parameter testing is included through use of the EFFECTS statement or the CONTRAST statement. The user should declare categorical variables through use of the CLASS statement prior to the MODEL statement. Beginning with SUDAAN

Version 9, PROC LOGISTIC/RLOGIST offered the Hosmer–Lemeshow (H–L) goodness-of-fit test (Hosmer and Lemeshow, 2000). Here is example code for SAS-callable SUDAAN using the JRR variance estimation method, an EFFECTS statement for testing the parameters associated with the indicator variables AGE1, AGE2, and AGE3 against 0, and the HLTEST statement for the H-L goodness-of-fit test:

```
proc rlogist data=one method=jrr;
nest str secu;
weight weight;
class maritalcat;
model depressed=maritalcat female age1 age2 age3;
effects age1 age2 age3 / name="age dummy test";
print / hltest=default;
run;
```

Ordinal and multinomial logistic regression is available in PROC MULTILOG. The two model types are based on generalizations of the logistic regression model and differ in the type of dependent variable as well as the estimation method used. SUDAAN uses a cumulative logit modeling approach for ordinal dependent variables and a generalized logit modeling approach for nominal categorical outcomes. See Chapter 9 for details and statistical background on these methods. The user must declare the method in the MODEL statement and the outcome as a categorical variable in the CLASS or SUBGROUP statement of the procedure. For example, the following SUDAAN code uses PROC MULTILOG with an ordinal outcome (self-rated health status, ranging from 1 to 5) and the same predictors as the previous logistic example, along with a test specification for Wald chi-square tests rather than the default Wald *F*-test:

```
proc multilog data=one;
nest str secu;
weight weight;
class maritalcat healthstatus;
model healthstatus=maritalcat female age1 age2 age3 / cumlogit;
test walchi;
effects age1 age2 age3 / name="age dummy test";
run;
```

For a nominal categorical outcome, the syntax would differ in that the categorical variable must be declared in the CLASS statement (marital status) and the use of "/GENLOGIT" is required on the model statement, indicating a generalized logit estimation method rather than the cumulative logit model for ordinal regression. The MULTILOG procedure can read in multiply imputed data sets and account for the variability introduced by imputation along with the complex sample design (as can all other SUDAAN procedures).

For Poisson or count-type outcomes (see Chapter 9), SUDAAN offers PROC LOGLINK, which like PROC LOGIST and PROC MULTILOG uses GEE for estimation and includes relevant design-adjusted hypothesis tests. The dependent variable must be continuous and greater than or equal to 0 for count-type regression models, and an offset or logarithm of the offset can also be specified for fitting rate models. The predictor variables can be either continuous or categorical, and categorical variables are to be declared in the CLASS statement prior to modeling. The LOGLINK procedure offers a full range of tests and output options, like the other SUDAAN modeling procedures. It can also handle multiply imputed data sets and subpopulation analyses. The following example syntax illustrates the use of PROC LOGLINK with a count outcome (the number of incidents of asthma per year), a TEST statement for a Wald *F*-test of the null hypothesis that the regression parameters are equal to 0, and use of the SUBPOPN statement for a subpopulation analysis of black females:

```
proc loglink data=one;
nest str secu;
weight weight;
subpopn female=1 & black=1;
class race;
model numincidents=age race;
test waldf;
run;
```

The SURVIVAL procedure in SUDAAN is used for survival (or time-to-event) analysis with discrete and continuous proportional hazards models that include censored data. This procedure can treat time as either continuous or discrete. For continuous time, SUDAAN uses the Cox proportional hazards model, and for discrete time, the program uses a log-likelihood function able to handle ties for failure or occurrence of the event (see Chapter 10). Estimated hazard ratios and design-adjusted estimates of standard errors are produced along with optional hypothesis tests for regression parameters and linear contrasts developed by the analyst. The SURVIVAL procedure can also handle time-dependent covariates correctly. Any of the major variance estimation methods can be selected along with the full array of sample design types. One important note is that PROC SURVIVAL does not currently handle multiply imputed data sets. Significance testing is similar to other modeling procedures in that either hypothesis testing for regression parameters or linear contrasts are possible. For goodness-of-fit evaluation, Martingale residuals (Therneau, Grambsch, and Fleming, 1990) and normalized Schoenfeld (1982) and score residuals are available output options. Here is an example of typical syntax where the event of interest is onset of depression (MDE), the outcome is years until event or censoring (AGEMDE), and the predictors are all time-invariant and categorical (SEX, RACECAT, and EDUCATION):

```
proc survival data=one;
nest str secu;
weight weight;
event mde;
class sex racecat education;
model agemde=sex racecat education;
run;
```

Hypothesis tests of whether parameter estimates in the model are equal to zero in PROC SURVIVAL include design-adjusted versions of the Wald chi-square, Wald *F*-test, Satterthwaite chi-square, and *F*-test and adjusted versions of each of these tests. Other tests such as whether parameter estimates are equal to each other (e.g., race group 1 = race group 2) are also possible. Additional tests for linear combinations of variables or custom contrasts can also be obtained.

For more descriptive survival analysis not involving multiple covariates, SUDAAN offers the KAPMEIER procedure. This is very similar to the SURVIVAL procedure with the difference that the KAPMEIER procedure is designed to evaluate time to an event without covariates by use of the product-limit estimation method (Kalbfleisch and Prentice, 2002). The procedure will produce weighted point estimates and design-based estimates of standard errors for survival functions and is generally used in conjunction with a graphing approach for visualizing the point estimates. As previously detailed, SUDAAN's easy output of data for further processing in a graphing software such as SAS, Stata, or R renders this a minor point, as plots of the survival estimates can easily be generated outside of SUDAAN. Use of an OUTPUT statement is required with PROC KAPMEIER, and the TIME and EVENT statements are also mandatory. Optional features include unconditional restriction of estimation to a subpopulation via the SUBPOPN statement, alternative methods for variance estimation (JRR, BRR, or Taylor series), and the STRHAZ statement for estimation of survival rates for subpopulations. The following example is of SUDAAN syntax for PROC KAPMEIER. The code illustrates the use of the EVENT, TIME, CLASS, STRHAZ, and SUBPOPN statements along with the required OUTPUT statement:

```
proc kapmeier data=one design=wr;
nest str secu;
weight weight;
subpopn women=1;
event depression;
time onsetage ;
class agegroup;
strhaz agegroup;
output / kapmeier=all filename="c:\km_all_out"
filetype=sas replace ;
run ;
```

In summary, SUDAAN is an excellent tool for survey data analysis. It accommodates a wide range of sample designs and offers a variety of variance estimation methods, numerous analytic tools, and frequent updates as statistical theory emerges. It is relatively easy to learn given the similarity to SAS syntax and produces output files that are ready to use in other more general packages for graphing and reporting or further analysis. See the book Web site for more SUDAAN examples that replicate the examples used in the book chapters.

A.5 Overview of SPSS®

We review SPSS Version 16 (http://www.spss.com) with a focus on the Complex Samples (CS) module. This add-on module can be included with the base package installation and offers many survey data analysis tools as well as the full range of usual SPSS features to perform data management, analysis, and graphing tasks (via commands that are a part of the base SPSS package). An added feature of this software is that it offers a dual option of running the procedures with either a point-and-click approach (and optionally saving the generated syntax) or running the procedures directly via generation of command syntax by the analyst. The online tutorials for each of the CS commands are excellent and provide a step-by-step path through the command options as well as interpretation and examples of model fit and other considerations during the analysis process.

All survey data utility and analysis commands are contained in the Complex Samples module. SPSS offers a number of complex samples routines including sampling plan (CSSAMPLE) and analysis plan (CSPLAN) "wizards" for sample selection and data preparation prior to analysis, along with a number of "Complex Samples" analysis commands: frequencies (CSTABULATE), descriptives (CSDESCRIPTIVES), crosstabs (CSTABULATE), ratios (CSRATIOS), general linear models (CSGLM), logistic regression with binary or multinomial outcomes (CSLOGISTIC), ordinal regression (CSORDINAL), and Cox regression (CSCOXREG) for survival analysis.

The analysis plan wizard allows the analyst to specify a number of complex sample designs, such as With-Replacement selection of ultimate clusters (WR), equal-size clusters Without Replacement (WOR), and unequal-size clusters Without Replacement (WOR) along with the finite population correction (FPC) ability. All analysis commands use the Taylor series linearization method for variance estimation and allow the use of a subpopulation indicator, the ability to perform appropriate design-based hypothesis tests, and additional options such as design effects and other optional statistics. Version 16 of SPSS does not offer the ability to analyze multiply imputed data sets.

A.5.1 The SPSS Complex Samples Commands

Two commands comprise sampling and analysis plan setup utilities: CSSAMPLE and CSPLAN. The CSSAMPLE procedure provides the ability to generate complex samples from existing sets of data (similar to PROC SURVEYSELECT in SAS), while the CSPLAN routine is designed to establish the complex sample "analysis plan" in preparation for analysis. The CSPLAN command asks the analyst to describe the sampling weight and stratum and PSU (or cluster) variables designed for sampling error calculations to the program for correct analysis of complex sample survey data. This command generates a "plan" file that is then saved for use with any of the SPSS CS analysis commands. The following is an example of a typical CSPLAN command, including the setup of the complex design features, and identification of the sampling weight and the stratum and cluster codes for sampling error calculations. The file generated by running this command (csplan1.csaplan) can then be used throughout the analysis session or edited during the session if desired:

```
* Analysis Preparation Wizard.
CSPLAN ANALYSIS
/PLAN FILE='C:\csplan1.csaplan'
/PLANVARS ANALYSISWEIGHT=NCSRWTLG
/SRSESTIMATOR TYPE=WR
/PRINT PLAN
/DESIGN STRATA=SESTRAT CLUSTER=SECLUSTR
/ESTIMATOR TYPE=WR.
```

There are four SPSS "CS" analytic commands each for descriptive tasks and modeling techniques. In the descriptive group, CSFREQUENCIES and CSCROSSTABS are used for analysis of categorical data, CSDESCRIPTIVES for continuous variables, and CSRATIOS for ratios. CSTABULATE drives both CSFREQUENCIES and CSCROSSTABS for one- and two-way table analysis and generates unweighted and weighted frequency counts, linearized standard errors, confidence limits, design effects, and a Wald *F*-test of equal cell proportions along with an option for subpopulation analysis. There is also an option for handling missing data with either listwise or table-by-table deletion. The SPSS program menus list two options that use the CSTABULATE command: frequencies and crosstabs. The following is typical SPSS syntax for this command including options for estimated population cell size, linearized standard errors, confidence limits, design effects, and a test for cell homogeneity for a single categorical variable of obesity categories (OBESE6CA):

```
* COMPLEX SAMPLES FREQUENCIES.
CSTABULATE
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
```

```
/TABLES VARIABLES=OBESE6CA
/CELLS POPSIZE
/STATISTICS SE CIN(95) DEFF
/TEST HOMOGENEITY
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

Note that the analysis command refers to the "plan" file generated earlier, describing the complex design features.

The CSDESCRIPTIVES command performs descriptive analyses of continuous variables and includes similar options to CSTABULATE except for the test of equal cell proportions. The following syntax example illustrates the use of a subpopulation indicator variable for women (SEXF) and use of the square root of the design effect (or DEFT) along with the standard output of standard errors, confidence limits, and a design effect. A *t*-test of the null hypothesis that the mean of AGE is equal to 0 is also requested:

```
* COMPLEX SAMPLES DESCRIPTIVES.
CSDESCRIPTIVES
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
/SUMMARY VARIABLES=AGE
/SUBPOP TABLE=SEXF DISPLAY=LAYERED
/MEAN TTEST=.05
/STATISTICS SE DEFF DEFFSQRT CIN(95)
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

The CSCROSSTABS command is designed for the design-based analysis of two-way or *n*-way contingency tables. This code is actually using a variant of the CSTABULATE command for *n*-way contingency table analysis and includes options for weighted estimates of row and column percentages, linearized standard errors, DEFF, DEFT, confidence limits, and coefficient of variation (for design corrections; see Chapter 6) along with a design-based test for independence of rows and columns. For 2 × 2 tables, the analyst can request odds ratio and risk ratio statistics. The following example illustrates a 2 × 4 cross-tabulation of SEX by REGION with many of the available table options demonstrated:

```
* Complex Samples Crosstabs.
CSTABULATE
/PLAN FILE='C:\csplan1.csaplan'
/TABLES VARIABLES=SEX BY REGION
/CELLS POPSIZE ROWPCT COLPCT
/STATISTICS SE CV CIN(95) DEFF
/TEST INDEPENDENCE
/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

The CSRATIOS command uses the CSDESCRIPTIVES command and performs a ratio analysis (see Chapter 5) by declaring the numerator and

denominator variables along with various options. The following example illustrates a ratio analysis of taxable income (TAXINC) among all income (ALLINC) and requests that SPSS output linearized standard errors and 95% confidence limits using the default missing data option of table-by-table exclusion:

```
* Complex Samples Ratios.
CSDESCRIPTIVES
/PLAN FILE='c:\csplan1.csaplan'
/RATIO NUMERATOR=TAXINC DENOMINATOR=ALLINC
/STATISTICS SE CIN(95)
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

The SPSS regression commands for complex sample survey data include CSGLM for continuous outcomes, CSLOGISTIC for binary or nominal outcomes, CSORDINAL for ordinal logistic regression, and CSCOXREG for survival analysis with the Cox proportional hazards model (Cox, 1972; see Chapter 10). All of these commands include subpopulation analysis options, complex design corrected standard errors, and hypothesis testing appropriate for the procedure used.

The CSGLM command performs regression analyses for continuous dependent variables and offers the usual features such as design-based standard errors for parameter estimates, design-based confidence limits for parameters, design effects and square roots of design effects, subpopulation analysis, and hypothesis testing. Specifically, the hypothesis tests include unadjusted and adjusted Wald *F* and chi-square tests for tests of parameters being equal to 0 and degrees of freedom specifications based on either the sample design or a fixed number of degrees of freedom. There is an optional estimated means technique for marginal means for factors or interactions included in the model as well. The following example shows syntax for a basic linear regression for household income (HHINC) with a number of options specified (confidence limits, design effects, and *F*-tests):

```
* COMPLEX SAMPLES GENERAL LINEAR MODEL.
CSGLM HHINC WITH AGE REGION
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
/MODEL AGE REGION
/INTERCEPT INCLUDE=YES SHOW=YES
/STATISTICS PARAMETER SE CINTERVAL DEFF
/PRINT SUMMARY VARIABLEINFO SAMPLEINFO
/TEST TYPE=F PADJUST=LSD
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA CILEVEL=95
```

See the book Web site for replication of examples presented in Chapter 7 using the CSGLM procedure.

The CSLOGISTIC command is used to fit design-based logistic regression models to binary or multinomial outcomes. This command offers options and hypothesis testing specific to logistic regression modeling and, as such, uses a pseudo maximum likelihood approach by default (see Chapter 8 for more details) for model estimation. The command also includes two options for assessing model fit: a classification approach or calculation of pseudo *R*-square values (Cox and Snell, 1989). Analysts can request that estimates of odds ratios and confidence limits around the ORs (odds ratios) be displayed, in addition to hypothesis tests including the Wald *F* and chi-square tests, both design adjusted and unadjusted. The pseudo *R*-square statistics are based on calculations suggested by Cox and Snell (1989), Nagelkerke (1981), and McFadden (1974) but are not design adjusted in this version of SPSS. Other available options include display of correlations and covariances among parameter estimates, subpopulation analysis, and design effects. The CSLOGISTIC command can also perform pairwise comparisons of predicted probabilities between levels of categorical predictor variables for fixed values of model covariates (see the SPSS documentation for details). The following example is syntax used to fit a logistic regression model to a dichotomous outcome (MDE), with common options of odds ratios, confidence limits, design effects, and specification of the reference category for the outcome variable (see the book Web site for additional examples):

```
* COMPLEX SAMPLES LOGISTIC REGRESSION.
CSLOGISTIC MDE(LOW) WITH SEXF BMI AGECENTERED
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
/MODEL SEXF BMI AGECENTERED
/INTERCEPT INCLUDE=YES SHOW=YES
/STATISTICS PARAMETER EXP SE CINTERVAL DEFF
/TEST TYPE=F PADJUST=LSD
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1E-006 RELATIVE]
LCONVERGE=[0] CHKSEP=20 CILEVEL=95
/PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

The CSORDINAL command is designed for fitting ordinal logistic regression models. There are a number of link options within the CSORDINAL command, including logit, complementary log-log, probit, and other related options along with excellent guidance included in the program's tutorial. The command includes options specific to this type of logistic regression analysis for ordinal outcomes, such as calculation of pseudo *R*-square statistics (non–design-adjusted; Cox and Snell, 1989) for model fit, a design-adjusted test for the parallel lines assumption (see Chapter 9), as well as the usual options for any type of logistic regression (hypothesis tests and output options previously noted). The following is an example of the syntax for the CSORDINAL command:

```
* COMPLEX SAMPLES ORDINAL REGRESSION.
CSORDINAL ED4CAT (ASCENDING) WITH MDE BMI
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
/LINK FUNCTION=LOGIT
/MODEL MDE BMI
/STATISTICS PARAMETER EXP SE CINTERVAL
/NONPARALLEL TEST
/TEST TYPE=F PADJUST=LSD
/MISSING CLASSMISSING=EXCLUDE
/CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1E-006 RELATIVE]
LCONVERGE=[0] METHOD=NEWTON CHKSEP=20 CILEVEL=95
/PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

For design-based survival analyses, SPSS includes the CSCOXREG command. This command models hazard rates using the Cox proportional hazards model for outcomes representing times to events and can handle both time-varying and time-invariant covariates. One key difference between CSCOXREG and the other SPSS CS commands is that the structure of the data set is often organized with multiple and varying numbers of records per individual. This structure is used to provide discrete units between events of interest such as year one of life to onset of an illness or censoring (see Chapter 10 for more on survival analysis and data file structure). The analyst can provide a multiple record file or a one-record-per-individual file or can create the needed data set within the command directly. The CSCOXREG command includes subpopulation analysis, hypothesis testing, control over the degrees of freedom used for significance testing (based either on the sample design or a fixed value), and the ability to perform multiple comparisons of hazard rates for groups defined by categorical variables. Like all SPSS regression commands, predictors that are categorical are defined as "factors," and "covariates" are treated as continuous predictors. Other options include declaration of the "event" of interest, a variable representing time at which the event occurred (for noncensored cases), and a test of the proportional hazards assumption (see Chapter 10 for details). Options for displaying estimated survival plots for models with and without covariates are also built into the command, which is a nice feature, and simple Kaplan-Meier curves can be displayed as well. The following is an example of the use of the CSCOXREG command for modeling time until onset of depression or time to censoring if no event occurred (AGEEVENT):

```
* COMPLEX SAMPLES COX REGRESSION.
CSCOXREG AGEEVENT BY OBESE6CA WITH SEXF
/PLAN FILE='C:\CSPLAN1.CSAPLAN'
/VARIABLES STATUS=MDE(1)
/MODEL OBESE6CA SEXF
/PRINT SAMPLEINFO EVENTINFO
/STATISTICS PARAMETER EXP SE CINTERVAL
/TESTASSUMPTIONS PROPHAZARD=KM
```

```
/PLOT SURVIVAL CI=NO
/TEST TYPE=F PADJUST=LSD
/CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1E-006 RELATIVE]
LCONVERGE=[0] TIES=EFRON CILEVEL=95
/SURVIVALMETHOD BASELINE=EFRON CI=LOG
/MISSING CLASSMISSING=EXCLUDE.
```

As previously mentioned, the CSCOXREG command can also produce Kaplan-Meier survival curves when used as a survival model without predictors. This approach will output a survival curve and baseline survival and cumulative hazards tables along with complex design-corrected standard errors from the CSCOXREG command and offers numerous types of plots such as survival, failure, hazard, and log-log of the survival function. The following code illustrates how to request a K–M survival curve from the CSCOXREG command:

```
* Complex Samples Cox Regression.
CSCOXREG AGEEVENT
/PLAN FILE='c:\ncsr_p1wt.csaplan'
/VARIABLES STATUS=mde(1)
/PRINT SAMPLEINFO EVENTINFO BASELINE
/STATISTICS PARAMETER SE
/TESTASSUMPTIONS PROPHAZARD=KM
/PLOT SURVIVAL CI=YES
/TEST TYPE=F PADJUST=LSD
/CRITERIA MXITER=100 MXSTEP=5 PCONVERGE=[1E-006 RELATIVE]
LCONVERGE=[0] TIES=EFRON CILEVEL=95
/SURVIVALMETHOD BASELINE=EFRON CI=LOG
/MISSING CLASSMISSING=EXCLUDE.
```

A.6 Overview of Additional Software

In this section, we present an abbreviated overview of four additional software tools for complex sample survey data analysis: WesVar (version 4.3), the R `survey` Package (version 3.16), IVEware, and Mplus (version 5.2). For each of these software packages we provide general guidance on overall features and recommended usage but omit syntax examples. Please refer to the book Web site for analysis examples and code samples.

A.6.1 WesVar®

WesVar is an excellent software tool for survey data analysis and is produced by the Westat organization (http://www.westat.com/westat/

statistical_software/wesvar). WesVar is primarily a repeated replication tool for the analysis of survey data (in terms of variance estimation) and is available free of charge. It features a point-and-click interface, organizes projects in "workbooks" for shared use, and is quite flexible in terms of data types accepted. For example, the software is able to read in data sets from SPSS, SAS, Stata, SPlus/R, Excel/Access, ASCII, and relational database products. A key strength of WesVar is its ability to create replicate weights within the program as well as to handle existing replicate and probability weights. It also offers a full range of repeated replication methods for variance estimation, such as JRR for two-per-stratum or *n*-per-stratum cluster samples, BRR, and BRR with Fay's adjustment. Subpopulation analyses can be performed with the use of a subgroup variable statement in each procedure. The analytic procedures of WesVar include the usual descriptive analyses plus the optional estimation of percentiles adjusting for complex design features. In terms of procedures for regression analysis, WesVar can fit linear regression models to continuous outcomes and logistic regression models to dichotomous or unordered outcomes. All analytic techniques include design-adjusted hypothesis tests as well as the ability to use multiply imputed data sets. Westat provides excellent online documentation rich with examples for readers interested in using WesVar.

A.6.2 IVEware (Imputation and Variance Estimation Software)

IVEware is a free software tool produced and maintained by the University of Michigan Survey Methodology Program (http://www.isr.umich.edu/src/smp/ive). IVEware runs either as a stand-alone tool or as a SAS-callable tool (the software was originally based on SAS macros). It offers both multiple imputation capabilities and variance estimation for complex sample survey data analyses through three macros: %IMPUTE, %DESCRIBE, and %REGRESS. In addition, the package offers a fourth module called %SASMOD, which allows the SAS user to perform complex sample survey data analyses for additional SAS procedures not included in the %DESCRIBE or %REGRESS modules (see the user documentation for a list of %SASMOD procedures available).

The %IMPUTE module performs multiple imputation of missing data, while the %DESCRIBE, %REGRESS, and %SASMOD modules are used for survey data analysis. All three analysis modules can read in either multiply imputed or standard survey data sets and analyze the survey data using appropriate design-based techniques for variance estimation. The %DESCRIBE module will perform various descriptive analyses such as estimation of means, contingency tables, and ratios with design-based variances estimated using the Taylor series linearization method. The %REGRESS module provides the user with a variety of regression techniques, including linear, logistic (binary, multinomial, ordinal), Tobit, Poisson, and proportional hazards modeling for

survival analysis. One of the advantages of this program is the flexibility of using the linearization technique for the descriptive procedures and the JRR variance estimation method for regression analysis. This approach provides the benefit of avoiding the empty cell problem often encountered when fitting regression models to subpopulations. Another advantage of IVEware is the ability to impute missing data using the flexible multivariate sequential regression technique of Raghunathan et al. (2001) and then to analyze the multiply imputed data sets using the correct design-based variance estimation methods without the need for multiple procedures or software tools.

A.6.3 Mplus

Mplus Version 5.2 (http://www.statmodel.com/) is a relatively new (initially developed during the late 1990s) and advanced analysis tool designed primarily for complex statistical modeling. Mplus includes complex design corrections for every analytic technique in the package, including advanced structural equation modeling. These techniques are analytically advanced, including single- and multilevel models, observed and latent variables, and approaches for cross-sectional and longitudinal data. Mplus offers the ability to handle multiply imputed data sets or files with missing data and can analyze a wide range of outcomes: continuous, categorical (binary, nominal, ordinal), count, censored, or various combinations of these variables. The analyst can specify latent or observed variables and multiple levels of analysis when using survey data.

Another useful feature of Mplus is the flexibility to approach survey data analysis in two ways. The first approach, commonly called the *design-based approach,* uses variables that represent the stratification and clustering inherent to the sample design and subsequently adjusts variances taking these design features (and sample weighting) into account. The second approach, discussed in Chapter 12, incorporates design features directly within the multiple levels of the model framework (multilevel modeling) and accounts for stratification and clustering within the model specification.

Multiply imputed data sets can be used with all Mplus routines, and subpopulation options are also available. Though Mplus is a very useful and advanced addition to the set of software for survey data analysis, most analysts would perform data management and descriptive analyses in other software because Mplus does not offer a direct or simple way to accomplish these tasks. This is likely due to the specialized nature of the tool but is an important consideration when selecting a general-purpose survey analysis tool.

A.6.4 The R survey Package

The R `survey` package (visit http://www.r-project.org/, where links to CRAN Web sites can be used to download the specific package) is a free software tool that offers a full range of survey data analysis techniques. This

package is one of a number of specialized packages enabling R users to perform specialized statistical analyses. R and the R `survey` package (version 3.16 reviewed here) are free software tools that can analyze survey data using the Taylor series linearization method or the usual repeated replication techniques (JRR or BRR) for design-based variance estimation. It handles multistage cluster designs and unequally weighted sample designs and offers estimation of descriptive statistics, generalized linear models, and pseudo maximum likelihood methods for fitting regression models (Lumley, 2005). All summary and modeling techniques include appropriate hypothesis testing options. Also included are survey-adjusted graphics and options such as subpopulation analysis, raking, calibration, and post-stratification. The R `survey` package is an excellent survey analysis tool provided that the analyst is familiar with the use of R procedures and R language concepts. However, it does not offer extensive point-and-click tools and would be somewhat of a challenge to learn for an inexperienced data analyst.

A.7 Summary

This appendix has presented a brief evaluation and overview of current software tools with the goal of providing practical guidance for the survey data analyst. Each of the reviewed software packages has basic abilities to analyze complex sample survey data, and one can expect variation in terms of ease of use, techniques available, and range of options. All of the reviewed packages can perform common analyses of survey data, and most include pertinent features such as subpopulation options, hypothesis testing, ability to handle strata with single PSUs, multiple imputation analysis, and key survey data analysis features. The ultimate choice of software is a complex consideration that includes data management considerations, but current software options present many excellent choices, and we have attempted to outline popular choices that are currently available in this appendix.

References

Agresti, A., *Categorical Data Analysis*, John Wiley & Sons, New York, 2002.

Allison, P.D., Discrete-time methods for the analysis of event histories, *Sociological Methodology*, 13, 61–98, 1982.

Allison, P.D., *Survival Analysis Using the SAS System*: *A Practical Guide*, SAS Institute, Cary, NC, 1995.

Allison, P.D., *Logistic Regression Using the SAS® System: Theory and Application*, Cary, NC, 1999.

Archer, K.J. and Lemeshow, S., Goodness-of-fit test for a logistic regression model estimated using survey sample data, *Stata Journal*, 6(1), 97–105, 2006.

Archer, K.J., Lemeshow, S., and Hosmer, D.W., Goodness-of-fit tests for logistic regression models when data are collected using a complex sample design, *Computational Statistics and Data Analysis*, 51, 4450–4464, 2007.

Barnard, J. and Rubin, D.B., Small-sample degrees of freedom with multiple imputation, *Biometrika,* 86(4), 948–955, 1999.

Belli, R.F., Computerized event history calendar methods: Facilitating autobiographical recall, *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 471–475, 2000.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (Eds.), *Measurement Errors in Surveys*, John Wiley & Sons, New York, 1991.

Binder, D.A., On the variances of asymptotically normal estimators from complex surveys, *Survey Methodology,* 7, 157–170, 1981.

Binder, D.A., On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review,* 51, 279–292, 1983.

Binder, D.A., Use of estimating functions for interval estimation from complex surveys, presented at *International Statistical Institute Meetings in Cairo*, 1991.

Binder, D.A., Fitting Cox's proportional hazards model from survey data, *Biometrika*, 79, 139–147, 1992.

Binder, D.A., Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodology*, 24(2), 101–108, 1998.

Bishop, Y.M., Feinberg, S.E., and Holland, P.W., *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA, 1975.

Bollen, K.A., *Structural Equations with Latent Variables*, Wiley-Interscience, New York, 1989.

Bowley, A.L., Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science, *Journal of the Royal Statistical Society,* 69, 548–557, 1906.

Breidt, F.J. and Opsomer, J.D., Nonparametric and semiparametric estimation in complex surveys, in C.R. Rao and D. Pfeffermann (Eds.), *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics*, Vol. 29, Elsevier, North Holland, 2009.

Brewer, K.R.W. and Mellor, R.W., The effects of sample structure on analytic surveys, *Australian Journal of Statistics*, 15, 145–152, 1973.

Bulmer, M., History of social survey, in N.J. Smeltser and P.B. Baltes, *International Encyclopedia of the Social and Behavioral Sciences*, vol. 21, 14469–14473, Elsevier, Oxford, 2001.

Burns, C.J., Laing, T.J., Gillespie, B.W., Heeringa, S.G., Alcser, K.H., Mayes, M.D., et al., The epidemiology of scleroderma among women: Assessment of risk from exposure to silicone and silica, *Journal of Rheumatology*, 23(11), 1904–1912, 1996.

Buskirk, T. and Lohr, S., Asymptotic properties of kernel density estimation with complex survey data, *Journal of Statistical Planning and Inference*, 128, 165–190, 2005.

Cameron, A. and Trivedi, P., *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, 1998.

Carlin, J.B., Galati, J.C., and Royston, P., A new framework for managing and analyzing multiply imputed data in Stata, *Stata Journal*, 8(1), 49–67, 2008.

Chambers, R.L., Dorfman, A.H., and Sverchkov, M. Yu., Nonparametric regression with complex sample survey data, in R.L. Chambers and C.J. Skinner, (Eds.),. *Analysis of Survey Data*, John Wiley and Sons, London, 2003.

Chambers, R.L. and Skinner, C.J. (Eds.), *Analysis of Survey Data*, John Wiley & Sons, New York, 2003.

Cleveland, W.S., *Visualizing Data*, Hobart Press, Summit, NJ, 1993.

Cleves, M., Gould, W.W., Gutierrez, R.G., and Marchenko, Y., *An Introduction to Survival Analysis using Stata*, 2d ed., Stata Press, College Station, TX, 2008.

Cochran, W.G., *Sampling Techniques*, 3d ed., John Wiley & Sons, New York, 1977.

Converse, J.M., *Survey Research in the United States: Roots and Emergence*, University of California Press, Berkeley, 1987.

Cooney, K.A., Strawderman, M.S., Wojno, K.J., Doerr, K.M., Taylor, A., Alcser, K.H., et al., Age-specific distribution of serum prostate-specific antigen in a community-based study of African-American men, *Urology*, 57, 91–96, 2001.

Cox, D.R., Regression models and life tables, *Journal of the Royal Statistical Society-B*, 34, 187–220, 1972.

Cox, D.R., Applied Statistics: A Review, *Annals of Applied Statistics*, 1(1), 1–16, 2007, 1.

Cox, D.R. and Snell, E.J., *The Analysis of Binary Data*, 2d ed., Chapman and Hall, London, 1989.

DeMaris, A., *Regression with Social Data*, John Wiley & Sons, New York, 2004.

Deming, W.E., *Some Theory of Sampling*, John Wiley & Sons, New York, 1950.

DeNavas-Walt, C., Proctor, B.D., and Smith, J., Current population reports, P60-233, *Income, Poverty and Health Insurance Coverage in the United States: 2006*, U.S. Government Printing Office, Washington, DC, 2007.

Deville, J.-C. and Särndal, C.-E., Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 376–382, 1992.

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L., *Analysis of Longitudinal Data*, 2d ed., Clarendon Press, Oxford, 2002.

Draper, N.R. and Smith, H., *Applied Regression Analysis*, 2d ed., John Wiley & Sons, New York, 1981.

DuMouchel, W.H. and Duncan, G.S., Using sample survey weights in multiple regression analyses of stratified samples, *Journal of the American Statistical Association*, 78, 535–543, 1983.

Elliott, M.R., Bayesian weight trimming for generalized linear regression models, *Survey Methodology*, 33(1), 23–34, 2007.

Elliott, M.R. and Little, R.J.A., Model-based approaches to weight trimming, *Journal of Official Statistics*, 16, 191–210, 2000.

Ezzatti-Rice, T.M., Khare, M., Rubin, D.B., Little, R.J.A., and Schafer, J.L., A comparison of imputation techniques in the Third National Health and Nutrition Examination Survey, *Proceedings of the American Statistical Association, Survey Research Methods Section*, 303–308, 1993.

Faraway, J.J., *Linear Models with R*, Chapman & Hall, CRC, London, 2005.

Faraway, J.J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC, New York, 2006.

Fellegi, I.P., Approximate tests of independence and goodness of fit based on stratified multistage samples, *Journal of the American Statistical Association*, 75, 261–268, 1980.

Fisher, R.A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh, 1925.

Fitzmaurice, G.M., Davidian, M., Verbeke, G., and Molenberghs, G. (Eds.), *Longitudinal Data Analysis*, John Wiley & Sons, Hoboken, NJ, 2009.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H., *Applied Longitudinal Analysis*, John Wiley & Sons, Hoboken, NJ, 2004.

Fox, J., *Applied Regression Analysis and Generalized Linear Model*, 2d ed., Sage, Thousand Oaks, CA, 2008.

Freedman, D.A., On the so-called "Huber Sandwich Estimator" and "robust standard errors," *American Statistician*, 60(4), 299–302, 2006.

Freedman, D.A., Survival analysis: A primer, *American Statistician*, 62, 110–119, 2008.

Fuller, W.A., Regression analysis for sample survey, *Sankyha, Series C*, 37, 117–132, 1975.

Fuller, W.A., *Measurement Error Models*, John Wiley & Sons, New York, 1987.

Fuller, W.A., Regression estimation for survey samples (with discussion), *Survey Methodology*, 28(1), 5–23, 2002.

Fuller, W.A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H.J., *PC CARP*, Iowa State University, Statistical Laboratory, Ames, 1989.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M., Illustration of Bayesian inference in normal data models using Gibbs sampling, *Journal of the American Statistical Association*, 85, 972–985, 1990.

Gelfand, A.E. and Smith, A.F.M., Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85, 398–409, 1990.

Gelman, A., Struggles with survey weighting and regression modeling, *Statistical Science*, 22(2), 153–164, 2007.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., *Bayesian Data Analysis*, 2d ed., Chapman & Hall / CRC Press, Boca Raton, FL, 2003.

Gelman, A. and Hill, J., *Data Analysis Using Regression and Multilevel / Hierarchical Models*, Cambridge University Press, New York, 2006.

Goldstein, H., *Multilevel Statistical Models*, 3d ed., Arnold, London, 2003.

Greenwood, M., The "error of sampling" of the Survivorship Tables. Reports on public health and medical subjects, No. 33, Appendix 1, H.M. Stationery Office, London, 1926.

Grizzle, J., Starmer, F., and Koch, G., Analysis of categorical data by linear models, *Biometrics*, 25, 489–504, 1969.

Groves, R.M., *Survey Errors and Survey Costs*, 2d ed., John Wiley & Sons, New York, 2004.

Groves, R.M. and Couper, M., *Nonresponse in Household Interview Surveys*, John Wiley & Sons, New York, 1998.

Groves, R.M. and Heeringa, S.G., Responsive design for household surveys: Tools for actively controlling survey errors and costs, *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3), 439–457, 2006.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R., *Survey Methodology*, John Wiley & Sons, New York, 2004.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G., *Sample Survey Methods and Theory, Volumes I and II*, John Wiley & Sons, New York, 1953.

Hansen, M.H., Madow, W.G., and Tepping, B.J., An evaluation of model-dependent and probability-sampling inferences in sample surveys, *Journal of the American Statistical Association*, 78, 776–793, 1983.

Harms, T. and Duchesne, P., On kernel nonparametric regression designed for complex survey data, *Metrika*, published online March 12, 2009 at http://www.springerlink.com/content/b61n117362222pn4/fulltext.pdf.

Harrell, F.E. Jr., *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer-Verlag, New York, 2001.

Heeringa, S. and O'Muircheartaigh, C., Sample design for cross-national, cross-cultural survey programs, in J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, et al. (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, John Wiley & Sons, Hoboken, NJ (in press).

Heeringa, S. and O'Muicheartaigh, C., *Survey Methods in Multinational, Multiregional, and Multicultural Contexts,* 247–263.

Heeringa, S.G., Alcser, K.H., Doerr, K., Strawderman, M., Cooney, K., Medberry, B., et al., Potential selection bias in a community-based study of PSA Levels in African-American men, *Journal of Clinical Epidemiology*, 54(2), 142–148, 2001.

Heeringa, S.G. and Connor, J., 1980 SRC National Sample: Design and Development, Technical report, Survey Research Center, University of Michigan, Ann Arbor, 1986.

Heeringa, S.G. and Connor, J., Technical Description of the Health and Retirement Survey Sample Design, Technical report, Survey Research Center, University of Michigan, Ann Arbor, 1995, accessed June 2009 at http://hrsonline.isr.umich.edu/sitedocs/userg/HRSSAMP.pdf.

Heeringa, S., Little, R.J.A., and Raghunathan, T., Multivariate imputation of coarsened survey data on household wealth, in R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*, John Wiley & Sons, New York, 2002.

Heeringa, S., Wagner, J., Torres, M., Duan, N., Adams, T. and Berglund, P., Sample designs and sampling methods for the Collaborative Psychiatric Epidemiology Studies (CPES), *International Journal of Methods in Psychiatric Research*, 13(4), 221–239, 2004.

Herzog, T. and Rubin, D.B., Using multiple imputations to handle nonresponse in sample surveys, in W.G. Madow, I. Olkin, and D.B. Rubin (Eds.), *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliography*, Academic Press, New York, 1983.

Hilbe, J.M., *Negative Binomial Regression*, Cambridge University Press, Cambridge, 2007.

Hill, M.S., *The Panel Study of Income Dynamics: A User's Guide,* Sage, Beverly Hills, CA, 1992.

Hoerl, A.E. and Kennard, R.W., Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67, 1970.

Holt, D. and Smith, T.M.F., Post stratification, *Journal of the Royal Statistical Society, Series A (General)*, 142(1), 33–46, 1979.

Horvitz, D.G. and Thompson, D.J., A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, 47, 663–685, 1952.

Hosmer, D.W. and Lemeshow, S., *Applied Logistic Regression,* Wiley, New York, 1989.

Hosmer, D.W. and Lemeshow, S., *Applied Logistic Regression*, 2d ed., John Wiley & Sons, New York, 2000.

Hosmer, D.W., Lemeshow, S., and May, S., *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2d ed., John Wiley & Sons, Hoboken, NJ, 2008.

House, J.S., Juster, F.T., Kahn, R.L., Schuman, H., and Singer, E., *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*, University of Michigan Press, Ann Arbor, 2004.

Rao, J.N.K. and Rust, K.F., Variance estimation for complex surveys using replication techniques, *Statistical Methods in Medical Research*, 5, 283–310, 1996.

Hyman, H.H., *Survey Design and Analysis*, Free Press, New York, 1955.

Jann, B., Multinomial goodness of fit: Large-sample tests with survey design correction and exact tests for small samples, *Stata Journal*, 8(2), 147–169, 2008.

Jans, M., Heeringa, S.G., and Charest, A.-S., Imputation for missing physiological and health measurement data: Tests and applications, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2450–2457, 2008.

Judge, G.G., Griffiths, W.E., Hill, R.C., and Lee, T.-C., *The Theory and Practice of Econometrics*, 2d ed., John Wiley & Sons, New York, 1985.

Judkins, D.R., Fay's method for variance estimation, *Journal of Official Statistics*, 6, 223–239, 1990.

Juster, F.T. and Suzman, R., The Health and Retirement Study: An overview, *Journal of Human Resources*, 1995(30 Suppl.), S7–56, 1995.

Kaier, A.N., Observations et experiences concernant des denombrements representatives. Discussion appears in Liv. 1, XCIII-XCVII, *Bulletin of the International Statistical Institute,* 9, Liv. 2, 176–183, 1895.

Kalbfleisch, J.D. and Prentice, R.L., *The Statistical Analysis of Failure Time Data,* Wiley, New York, 1980.

Kalbfleisch, J.D. and Prentice, R.L., *The Statistical Analysis of Failure Time Data*, 2d ed., John Wiley & Sons, New York, 2002.

Kalton, G., *Introduction to Survey Sampling*, Sage, Beverly Hills, CA, 1983.

Kalton, G., Handling wave nonresponse in panel surveys, *Journal of Official Statistics*, 2(3), 303–314, 1986.

Kalton, G. and Citro, C., Panel surveys: Adding the fourth dimension, *Survey Methodology*, 19, 205–215, 1993.

Kalton, G. and Kasprzyk, D., The treatment of missing survey data, *Survey Methodology*, 12(1), 1–16, 1986.

Kavoussi, S.K., West, B.T., Taylor, G.W., and Lebovic, D.I., Periodontal disease and endometriosis: Analysis of the National Health and Nutrition Examination Survey, *Fertility & Sterility*, 91(2), 335–342, 2009.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S., Consequences of reducing nonresponse in a national telephone survey, *Public Opinion Quarterly*, 64, 125–148, 2000.

Kendall, P.L. and Lazarsfeld, P.F., Problems of survey analysis, in R.K. Merton and P.F. Lazaarsfeld (Eds.), *Continuities in Social Research: Studies in the Scope and Method of "The American Soldier,"* Free Press, Chicago, 1950.

Kennickell, A.B., Multiple imputation in the Survey of Consumer Finances, Federal Reserve Board, Paper 78, Washington, DC, September 1998.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., et al., The US National Comorbidity Survey Replication (NCS-R): Design and field procedures, *International Journal of Methods in Psychiatric Research,* 13(2), 69–92, 2004.

Kish, L., A procedure for objective respondent selection within the household, *Journal of the American Statistical Association*, 44, 380–387, 1949.

Kish, L., *Survey Sampling*, John Wiley & Sons, New York, 1965.

Kish, L., *Statistical Design for Research*, New York: John Wiley & Sons, 1987.

Kish, L. and Frankel, M.R., Inference from complex samples, *Journal of the Royal Statistical Society, Series B,* 36, 1–37, 1974.

Kish, L. and Hess, I., On variances of ratios and their differences in multi-stage samples, *Journal of the American Statistical Association*, 54, 416–446, 1959.

Kish, L., Groves, R.M., and Krotki, K., Sampling errors for fertility surveys, *Occasional Papers*, No. 17, World Fertility Survey, 1976.

Klein, L.R. and Morgan, J.N., Results of alternative statistical treatment of sample survey data, *Journal of the American Statistical Association,* 46, 442–460, 1951.

Kleinbaum D., Kupper L., and Muller K., *Applied Regression Analysis and Other Multivariable Methods*, 2d ed., Duxbury Press, Belmont, CA, 1988.

Kline, R.B., *Principles and Practice of Structural Equation Modeling*, 2d ed., Guilford Press, New York, 2004.

Koch, G.G. and Lemeshow, S., An application of multivariate analysis to complex sample survey data, *Journal of the American Statistical Association,* 54, 59–78, 1972.

Kolenikov, S., Resampling variance estimation for complex survey data, *Stata Journal* (in press).

Korn, E.L. and Graubard, B.I., Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics, *American Statistician*, 44, 270–276, 1990.

Korn, E.L. and Graubard, B.I., Scatterplots with survey data, *American Statistician*, 52(1), 58–69, 1998.

Korn, E.L. and Graubard, B.I., *Analysis of Health Surveys*, John Wiley & Sons, New York, 1999.

Kott, P.S., A model-based look at linear regression with survey data, *American Statistician*, 45, 107–112, 1991.

Kott, P.S. and Carr, D.A., Developing an estimation strategy for a pesticide data program, *Journal of Official Statistics*, 13(4), 367–383, 1997.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J., Bootstrap and other methods to measure errors in survey estimates, *Canadian Journal of Statistics,* 16 Suppl., 25–45, 1988.

Landis, R.J., Stanish, W.M., Freeman, J.L., and Koch, G.G., A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT), *Computer Programs in Biomedicine*, 6, 196–231, 1976.

Lee, E.S. and Forthofer, R.N., *Analyzing Complex Survey Data,* 2d ed., Sage, Thousand Oaks, CA, 2006.

Lee, E.T., *Statistical Methods for Survival Analysis*, John Wiley & Sons, New York, 1992.

Lepkowski, J.M. and Couper, M.P., Nonresponse in the second wave of longitudinal household surveys, pp. 259–271 in R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse*, John Wiley & Sons, New York, 2002.

Lessler, J.T. and Kalsbeek, W.D., *Nonsampling Errors in Surveys*, John Wiley & Sons, New York, 1992.

Levy, P.S. and Lemeshow, S., *Sampling of Populations: Methods and Applications*, 4th ed., John Wiley & Sons, New York, 2007.

Li, J., Linear regression diagnostics in cluster samples, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Joint Statistical Meetings, 2007.

Li, J. and Valliant, R., Influence analysis in linear regression with sampling weights, *Proceedings of the Section on Survey Methods Research, American Statistical Association*, 3330, 2006.

Li, J. and Valliant, R., Survey weighted hat matrix and leverages, *Survey Methodology*, 35(1), 15–24, 2009.

Li, K.H., Raghunathan, T.E., and Rubin, D.B., Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution, *Journal of the American Statistical Association*, 86, 1065–1073, 1991.

Little, R.J., The Bayesian approach to sample survey inference, chapter 4 in R. Chambers and C.J. Skinner (Eds.), *Analysis of Survey Data*, John Wiley & Sons, Hoboken, NJ, 2003.

Little, R.J.A., Inference with survey weights, *Journal of Official Statistics*, 7, 405–424, 1991.

Little, R.J.A. and Rubin, D.B., *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, New York, 2002.

Little, R.J. and Vartivarian, S., Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2), 161–168, 2005.

Lohr, S.L., *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove, CA, 1999.

Long, J.S. and Freese, J., *Regression Models for Categorical Dependent Variables Using Stata*, 2nd ed., Stat Press, College Station, Texas, 2006.

Loomis, D., Richardson, D.B., and Elliott, L., Poisson regression analysis of ungrouped data, *Occupational and Environmental Medicine* 62, 325–329, 2005.

Lumley, T., R software from the R Project, http://www.r-project.org/, V2.7 Analysis of complex survey samples, maintained by Thomas Lumley, University of Washington, 2005.

Lynn, P. (Ed.), *Methodology of Longitudinal Surveys*, John Wiley & Sons, New York, 2009.

Madow, W.G. and Olkin, I. (Eds.), *Incomplete Data in Sample Surveys, Volume 3: Proceedings of the Symposium*, Academic Press, New York, 1983.

Mahalanobis, P.C., Recent experiments in statistical sampling in the Indian Statistical Institute, *Journal of the Royal Statistical Society*, 109, 325–370, 1946.

Maindonald, J.H. and Braun, W.J., *Data analysis and graphics using R: an example-based approach*, 2d ed., Cambridge University Press, New York, 2007.

McCabe, S.E., West, B.T., Morales, M., Cranford, J.A., and Boyd, C.J., Does early onset of non-medical use of prescription drugs predict subsequent prescription drug abuse and dependence? Results from a national study, *Addiction*, 102(12), 1920–1930, 2007.

McCarthy, P.J., Pseudoreplication: Half samples, *Review of the International Statistical Institute*, 37, 239–264, 1969.

McCulloch, C.E. and Searle, S.R., *Generalized, Linear and Mixed Models*, John Wiley & Sons, New York, 2001.
McCullagh, P. and Nelder, J.A., *Generalized Linear Models,* 2d ed., Chapman and Hall, London, 1989.
McFadden, D., Conditional logit analysis of qualitative choice behavior, in P. Zarembka (Ed.), *Frontiers in Economics*, Academic Press, New York, 1974.
Menard, S.W. (Ed.), *Handbook of Longitudinal Research,* Academic Press, New York, 2008.
Miller, R., *Survival Analysis*, John Wiley & Sons, New York, 1981.
Miller, R.G., The jackknife—a review, *Biometrika*, 61, 1–15, 1974.
Mitchell, M.N., *A Visual Guide to Stata Graphics*, 2d ed., Stata Press, College Station, TX, 2008.
Molenberghs, G. and Verbeke, G., *Models for Discrete Longitudinal Data,* Springer, New York, 2005.
Mohadjer, L. and Curtin, L.R., NHANES, Balancing sample design goals for the National Health and Nutrition Examination Survey, *Survey Methodology*, 34(1), 119–126, 2008.
Morel, G., Logistic regression under complex survey designs, *Survey Methodology*, 15, 202–223, 1989.
Muthén, B.O. and Satorra, A., Complex sample data in structural equation modeling, *Sociological Methodology*, 25, 267–316, 1995.
Muthén, L.K. and Muthén, B.O., *Mplus User's Guide*, 5th ed., Muthén and Muthén, Los Angeles, CA, 1998–2007.
Nagelkerke, N.J.D., A note on the general definition of the coefficient of determination, *Biometrika*, 78(3), 691–692, 1981.
Neter, J., Kutner, M.H., Wasserman, W., and Nachtsheim, C.J., *Applied Linear Statistical Models*, 4th ed., McGraw-Hill/Irwin, Boston, 1996.
Neyman, J., On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, 558–606, 1934.
O'Muircheartaigh, C. and Wong, S.T., The impact of sampling theory on survey practices: A review, *Bulletin of the International Statistical Institute*, 465–493, 1981.
Opsomer, J.D. and Miller, C.P., Selecting the amount of smoothing in nonparametric regression estimation for complex surveys, *Journal of Nonparametric Statistics*, 17(5), 593–611, 2005.
Peterson, B. and Harrell, F., Partial proportional odds models for ordinal response variables, *Applied Statistics,* 39, 205–217, 1990.
Pfeffermann, D. and Holmes, D.J., Robustness considerations in the choice of method of inference for regression analysis of survey data, *Journal of the Royal Statistical Society, Series A*, 148, 268–278, 1985.
Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J., Weighting for unequal selection probabilities in multilevel models, *Journal of the Royal Statistical Society*, Series B 60(1), 23–40, 1998.
Plassman, B.L., Langa, K.M., Fisher, G.G., Heeringa, S.G., Weir, D.R., Ofstedal, M.B., et al., Prevalence of dementia in the United States: The Aging, Demographics, and Memory Study, *Neuroepidemiology*, 29, 125–132, 2007.
Potter, F., A study of procedures to identify and trim extreme sample weights, *Proceedings of the Survey Research Methods Section*, American Statistical Association, 225–230, 1990.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A., GLLAMM Manual, U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 160, 2004.

Rabe-Hesketh, S. and Skrondal, A., Multilevel modelling of complex survey data, *Journal of the Royal Statistical Society-A*, 169, 805–827, 2006.

Rabe-Hesketh, S. and Skrondal, A., *Multilevel and longitudinal modeling using Stata,* 2d ed., Stata Press, College Station, TX, 2008.

Raghunathan, T.E. and Grizzle, J.E., A split questionnaire survey design, *Journal of the American Statistical Association,* 90(429), 54–63, 1995.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P., A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27(1), 85–95, 2001.

Rao, J.N.K., *Small Area Estimation*, Wiley Series in Survey Methodology, John Wiley & Sons, New York, 2003.

Rao, J.N.K. and Scott, A.J., The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American Statistical Association,* 76, 221–230, 1981.

Rao, J.N.K. and Scott, A.J., On chi-squared test for multiway contingency tables with cell proportions estimated from survey data, *Annals of Statistics*, 12, 46–60, 1984.

Rao, J.N.K. and Shao, J., Jackknife variance estimation with survey data under hot deck imputation, *Biometrika,* 79, 811–822, 1992.

Rao, J.N.K. and Thomas, D.R., The analysis of cross-classified categorical data from complex sample surveys, *Sociological Methodology*, 18, 213–269, 1988.

Rao, J.N.K. and Wu, C.F.J., Inference from stratified samples: Second order analysis of three methods for nonlinear statistics, *Journal of the American Statistical Association*, 80, 620–630, 1985.

Rao, J.N.K. and Wu, C.F.J., Resampling inference with complex survey data, *Journal of the American Statistical Association*, 83, 231–241, 1988.

Raudenbush, S.W., Synthesizing results for NAEP trial state assessment, in D.W. Grissmer and M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement,* National Center for Educational Statistics, Washington, DC, 2000.

Raudenbush, S.W. and Bryk, A.S., *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2d ed., Sage, Newbury Park, CA, 2002.

Reiter, J.P., Raghunathan, T.E., and Kinney, S.K., The importance of modeling the sampling design in multiple imputation for missing data, *Survey Methodology*, 32(2), 143–149, 2006.

Research Triangle Institute (RTI), *SUDAAN 9.0 User's Manual: Software for Statistical Analysis of Correlated Data*, RTI, Research Triangle Park, NC, 2004.

Roberts, G., Rao, J.N.K., and Kumar, S., Logistic regression analysis of sample survey data, *Biometrika,* 74, 1–12, 1987.

Rothman, K.J., *Causal Inference*, Epidemiology Resources, MA, 1988, out of print.

Royston, P., Multiple imputation of missing values, *Stata Technical Journal,* 5(4), 527–536, 2005.

Rubin, D.B., Inference and missing data, *Biometrika,* 63(3), 581–592, 1976.

Rubin, D.B., Basic ideas of multiple imputation for nonresponse, *Survey Methodology*, 12(1), 37–47, 1986.

Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, New York, 1987.

Rubin, D.B., Multiple imputation after 18+ years, *Journal of the American Statistical Association*, 91(434), 473–489, 1996.

Rubin, D.B. and Schenker, N., Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association*, 81, 366–374, 1986.

Rueters/University of Michigan Surveys of Consumers, Accessed May 1, 2008 at http://thomsonreuters.com/products_services/financial/UMichigan_Surveys_of_Consumers, April 2007 Report, 1.

Rust, K., Variance estimation for complex estimators in sample surveys, *Journal of Official Statistics*, 1, 381–397, 1985.

Rust, K. and Hsu, V., Confidence intervals for statistics for categorical variables from complex samples, *Proceedings of the 2007 Joint Statistical Meetings*, Salt Lake City, UT, 2007.

SAS Institute, Inc., *SAS/STAT® User's Guide, Version 9*, SAS Institute, Cary, NC, 2003.

SAS Institute Inc., *SAS/GRAPH® 9.2: Statistical Graphics Procedures Guide*, SAS Institute, Cary, NC, 2009.

Satterthwaite, F.E., An approximate distribution of estimates of variance components, *Biometrics*, 110–114, 1946.

Schafer, J.L., *MIX: Multiple Imputation for Mixed Continuous and Categorical Data*, software library for S-PLUS, 1996, Written in S-PLUS and Fortran-77, at http://www.stat.psu.edu/~jls/.

Schafer, J.L., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997.

Schafer, J.L., *NORM: Multiple Imputation of Incomplete Multivariate Data under a Normal Model, Version 2*, 1999, Software for Windows 95/98/NT, at http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer, J.L., Ezatti-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A., and Rubin, D.B., The NHANES III multiple imputation project, *Proceedings of the Survey Research Methods Section*, American Statistical Association, 696–701, 1996.

Schoenfeld, D., Residuals for the proportional hazards regression model, *Biometrika*, 239–241, 1982.

Schumacker, R.E. and Lomax, R.G., *A Beginner's Guide to Structural Equation Modeling*, 2d ed., Lawrence Erlbaum, 2004, Hillsdale, NJ.

Shah, B.V., Holt, M.M., and Folsom, R.F., Inference about regression models from sample survey data, *Bulletin of the International Statistical Institute*, 41(3), 43–57, 1977.

Shao, J. and Tu, D., *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.

Shao, J. and Wu, C.F.J., A general theory for jackknife variance estimation, *Annals of Statistics*, 17, 1176–1197, 1989.

Singer, J.D. and Willett, J.B., It's about time: Using discrete-time survival analysis to study duration and the timing of events, *Journal of Educational and Behavioral Statistics*, 18, 155–195, 1993.

Skinner, C. and Vieira, M. de T., Variance estimation in the analysis of clustered longitudinal survey data, *Survey Methodology*, 33(1), 3–12, 2007.

Skinner, C.J. and Holmes, D.J., Random effects models for longitudinal survey data, chapter 14 in R.L. Chambers and C.J. Skinner (Eds.), *Analysis of Survey Data*, John Wiley & Sons, London, 2003.

Skinner, C.J., Holt, D., and Smith, T.M.F., *Analysis of Complex Surveys*, John Wiley & Sons, New York, 1989.

Skrondal, A. and Rabe-Hesketh, S., *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall / CRC Press, Boca Raton, FL, 2004.

Sribney, W.M., Two-way contingency tables for survey or clustered data, *Stata Technical Bulletin*, 45, 33–49, 1998.

Stapleton, L.M., Variance estimation using replication methods in structural equation modeling with complex sample data, *Structural Equation Modeling: A Multidisciplinary Journal*, 15(2), 183–210, 2008.

STATA Corp., *Release 10, P Manual*, *STATA Survey Data Manual*, College Station, TX, 2008.

Statistical Solutions, Solas 3.0, at http://www.statsol.ie/html/solas/solas_home.html.

Stiller, J.G. and Dalzell, D.R., Hot-deck imputation with SAS arrays and macros for large surveys, *Proceedings of the Twenty-Third Annual AS Users Group International Conference*, 1378–1383, 1998.

Stokes, M.E., Davis, C.S., and Koch G.G., *Categorical Data Analysis Using the SAS System, Second edition*, SAS Institute Inc., Cary, NC, 2002.

Striegel-Moore, R.H., Franko, D.L., Thompson, D., Affenito, S., and May, A., Exploring the typology of night eating syndrome, *International Journal of Eating Disorders*, 41(5), 411–418, 2008.

Sukatme, P.V., *Sampling Theory of Surveys, with Applications*, Iowa State College Press, Ames, 1954.

Tanner, M. and Wong, W., The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, 82, 528–550, 1997.

Therneau, T.M., Grambsch, P.M., and Fleming, T.R., Martingale-based residuals for survival models, *Biometrika*, 77(1), 147–160, 1990.

Thomas, D.R. and Rao, J.N.K., Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistical Association*, 82, 630–636, 1987.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J., and Johnson, C.L., An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey, *Survey Methodology*, 32(2), 217–231, 2006.

Thompson, S.K. and Seber, G.A.F., *Adaptive Sampling*, John Wiley & Sons, New York, 1996.

Tufte, E.R., *The Visual Display of Information*, Graphics Press, Cheshire, CT, 1983.

Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

University of Michigan, Computer Support Group, OSIRIS VI: Statistical Analysis and Data Management Software System, Survey Research Center, Institute for Social Research, 1982.

Valliant, R., Comparisons of variance estimators in stratified random and systematic sampling, *Journal of Official Statistics*, 6(2), 115–131, 1990.

Valliant, R., The effect of multiple weighting steps on variance estimation, *Journal of Official Statistics*, 20(1), 1–18, 2004.

Valliant, R., Dorfman, A.H., and Royall, R.M., *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, New York, 2000.

Van Buuren, S. and Oudshoorn, C.G.M., *Flexible multivariate imputation by MICE*, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054, 1999.

Verbeke, G. and Molenberghs, G., *Linear Mixed Models for Longitudinal Data*, Springer, New York, 2005.

Vieira, M.D.T. and Skinner, C.J., Estimating models for panel survey data under complex sampling, *Journal of Official Statistics*, 24, 343–364, 2008.

West, B.T., Berglund, P., and Heeringa, S.G., A closer examination of subpopulation analysis of complex-sample survey data, *Stata Journal*, 8(4), 520–531, 2008.

West, B.T., Welch, K.B., and Galecki, A.T., *Linear Mixed Models: A Practical Guide Using Statistical Software*, Chapman & Hall / CRC Press, Boca Raton, FL, 2007.

Westat, Inc., *WesVar 4.0 User's Guide*, Westat, Rockville, MD, 2000.

Wolter, K.M., *Introduction to Variance Estimation*, 2d ed., Springer-Verlag, New York, 2007.

Woodruff, R.S., A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, 66, 411–414, 1971.

Yamaguchi, K., *Event History Analysis*, Sage Publications, Newbury Park, CA, 1991.

Yates, F., *Sampling Methods for Censuses and Surveys*, Griffin, London, 1949 (2d ed., 1953; 3d ed., 1960).

Zajacova, A., Dowd, J.B., and Aiello, A.E., Socioeconomic and race/ethnic patterns in persistent infection burden among U.S. adults, *Journal of Gerontology A: Biological Sciences and Medical Sciences*, 64A(2), 272–279, 2009.

Zheng, H. and Little, R.J.A., Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model, *Journal of Official Statistics*, 21(1), 1–20, 2005.