Contents

Preface xv

1. **Applied Survey Data Analysis: Overview** 1
   - 1.1 Introduction 1
   - 1.2 A Brief History of Applied Survey Data Analysis 3
     - 1.2.1 Key Theoretical Developments 3
     - 1.2.2 Key Software Developments 5
   - 1.3 Example Data Sets and Exercises 6
     - 1.3.1 The National Comorbidity Survey Replication (NCS-R) 6
     - 1.3.2 The Health and Retirement Study (HRS)—2006 7
     - 1.3.3 The National Health and Nutrition Examination Survey (NHANES)—2005, 2006 7
     - 1.3.4 Steps in Applied Survey Data Analysis 8
       - 1.3.4.1 Step 1: Definition of the Problem and Statement of the Objectives 8
       - 1.3.4.2 Step 2: Understanding the Sample Design 9
       - 1.3.4.3 Step 3: Understanding Design Variables, Underlying Constructs, and Missing Data 10
       - 1.3.4.4 Step 4: Analyzing the Data 11
       - 1.3.4.5 Step 5: Interpreting and Evaluating the Results of the Analysis 11
       - 1.3.4.6 Step 6: Reporting of Estimates and Inferences from the Survey Data 12

2. **Getting to Know the Complex Sample Design** 13
   - 2.1 Introduction 13
     - 2.1.1 Technical Documentation and Supplemental Literature Review 13
   - 2.2 Classification of Sample Designs 14
     - 2.2.1 Sampling Plans 15
     - 2.2.2 Inference from Survey Data 16
   - 2.3 Target Populations and Survey Populations 16
   - 2.4 Simple Random Sampling: A Simple Model for Design-Based Inference 18
     - 2.4.1 Relevance of SRS to Complex Sample Survey Data Analysis 18
     - 2.4.2 SRS Fundamentals: A Framework for Design-Based Inference 19
     - 2.4.3 An Example of Design-Based Inference under SRS 21

2.5 Complex Sample Design Effects ... 23
 2.5.1 Design Effect Ratio ... 23
 2.5.2 Generalized Design Effects and Effective Sample Sizes ... 25
2.6 Complex Samples: Clustering and Stratification ... 27
 2.6.1 Clustered Sampling Plans ... 28
 2.6.2 Stratification ... 31
 2.6.3 Joint Effects of Sample Stratification and Clustering ... 34
2.7 Weighting in Analysis of Survey Data ... 35
 2.7.1 Introduction to Weighted Analysis of Survey Data ... 35
 2.7.2 Weighting for Probabilities of Selection ... 37
 2.7.3 Nonresponse Adjustment Weights ... 39
  2.7.3.1 Weighting Class Approach ... 40
  2.7.3.2 Propensity Cell Adjustment Approach ... 40
 2.7.4 Poststratification Weight Factors ... 42
 2.7.5 Design Effects Due to Weighted Analysis ... 44
2.8 Multistage Area Probability Sample Designs ... 46
 2.8.1 Primary Stage Sampling ... 47
 2.8.2 Secondary Stage Sampling ... 48
 2.8.3 Third and Fourth Stage Sampling of Housing Units and Eligible Respondents ... 49
2.9 Special Types of Sampling Plans Encountered in Surveys ... 50

3. Foundations and Techniques for Design-Based Estimation and Inference ... 53
3.1 Introduction ... 53
3.2 Finite Populations and Superpopulation Models ... 54
3.3 Confidence Intervals for Population Parameters ... 56
3.4 Weighted Estimation of Population Parameters ... 56
3.5 Probability Distributions and Design-Based Inference ... 60
 3.5.1 Sampling Distributions of Survey Estimates ... 60
 3.5.2 Degrees of Freedom for *t* under Complex Sample Designs ... 63
3.6 Variance Estimation ... 65
 3.6.1 Simplifying Assumptions Employed in Complex Sample Variance Estimation ... 66
 3.6.2 The Taylor Series Linearization Method ... 68
  3.6.2.1 TSL Step 1 ... 69
  3.6.2.2 TSL Step 2 ... 70
  3.6.2.3 TSL Step 3 ... 71
  3.6.2.4 TSL Step 4 ... 71
  3.6.2.5 TSL Step 5 ... 73
 3.6.3 Replication Methods for Variance Estimation ... 74
  3.6.3.1 Jackknife Repeated Replication ... 75

3.6.3.2 Balanced Repeated Replication 78
3.6.3.3 The Bootstrap 82
3.6.4 An Example Comparing the Results from the TSL, JRR, and BRR Methods 82
3.7 Hypothesis Testing in Survey Data Analysis 83
3.8 Total Survey Error and Its Impact on Survey Estimation and Inference 85
3.8.1 Variable Errors 86
3.8.2 Biases in Survey Data 87

4. Preparation for Complex Sample Survey Data Analysis 91
4.1 Introduction 91
4.2 Analysis Weights: Review by the Data User 92
4.2.1 Identification of the Correct Weight Variables for the Analysis 93
4.2.2 Determining the Distribution and Scaling of the Weight Variables 94
4.2.3 Weighting Applications: Sensitivity of Survey Estimates to the Weights 96
4.3 Understanding and Checking the Sampling Error Calculation Model 98
4.3.1 Stratum and Cluster Codes in Complex Sample Survey Data Sets 99
4.3.2 Building the NCS-R Sampling Error Calculation Model 100
4.3.3 Combining Strata, Randomly Grouping PSUs, and Collapsing Strata 103
4.3.4 Checking the Sampling Error Calculation Model for the Survey Data Set 105
4.4 Addressing Item Missing Data in Analysis Variables 108
4.4.1 Potential Bias Due to Ignoring Missing Data 108
4.4.2 Exploring Rates and Patterns of Missing Data Prior to Analysis 109
4.5 Preparing to Analyze Data for Sample Subpopulations 110
4.5.1 Subpopulation Distributions across Sample Design Units 111
4.5.2 The Unconditional Approach for Subclass Analysis 114
4.5.3 Preparation for Subclass Analyses 114
4.6 A Final Checklist for Data Users 115

5. Descriptive Analysis for Continuous Variables 117
5.1 Introduction 117
5.2 Special Considerations in Descriptive Analysis of Complex Sample Survey Data 118
5.2.1 Weighted Estimation 118

5.2.2 Design Effects for Descriptive Statistics 119
5.2.3 Matching the Method to the Variable Type 119
5.3 Simple Statistics for Univariate Continuous Distributions 120
5.3.1 Graphical Tools for Descriptive Analysis of Survey Data 120
5.3.2 Estimation of Population Totals 123
5.3.3 Means of Continuous, Binary, or Interval Scale Data 128
5.3.4 Standard Deviations of Continuous Variables 130
5.3.5 Estimation of Percentiles and Medians of Population Distributions 131
5.4 Bivariate Relationships between Two Continuous Variables 134
5.4.1 X–Y Scatterplots 134
5.4.2 Product Moment Correlation Statistic (*r*) 135
5.4.3 Ratios of Two Continuous Variables 136
5.5 Descriptive Statistics for Subpopulations 137
5.6 Linear Functions of Descriptive Estimates and Differences of Means 139
5.6.1 Differences of Means for Two Subpopulations 141
5.6.2 Comparing Means over Time 143
5.7 Exercises 144

6. Categorical Data Analysis 149
6.1 Introduction 149
6.2 A Framework for Analysis of Categorical Survey Data 150
6.2.1 Incorporating the Complex Design and Pseudo-Maximum Likelihood 150
6.2.2 Proportions and Percentages 150
6.2.3 Cross-Tabulations, Contingency Tables, and Weighted Frequencies 151
6.3 Univariate Analysis of Categorical Data 152
6.3.1 Estimation of Proportions for Binary Variables 152
6.3.2 Estimation of Category Proportions for Multinomial Variables 156
6.3.3 Testing Hypotheses Concerning a Vector of Population Proportions 158
6.3.4 Graphical Display for a Single Categorical Variable 159
6.4 Bivariate Analysis of Categorical Data 160
6.4.1 Response and Factor Variables 160
6.4.2 Estimation of Total, Row, and Column Proportions for Two-Way Tables 162
6.4.3 Estimating and Testing Differences in Subpopulation Proportions 163
6.4.4 Chi-Square Tests of Independence of Rows and Columns 164
6.4.5 Odds Ratios and Relative Risks 170

6.4.6 Simple Logistic Regression to Estimate the Odds Ratio ... 171
6.4.7 Bivariate Graphical Analysis ... 173
6.5 Analysis of Multivariate Categorical Data ... 174
6.5.1 The Cochran–Mantel–Haenszel Test ... 174
6.5.2 Log-Linear Models for Contingency Tables ... 176
6.6 Exercises ... 177

7. Linear Regression Models ... 179
7.1 Introduction ... 179
7.2 The Linear Regression Model ... 180
7.2.1 The Standard Linear Regression Model ... 182
7.2.2 Survey Treatment of the Regression Model ... 183
7.3 Four Steps in Linear Regression Analysis ... 185
7.3.1 Step 1: Specifying and Refining the Model ... 186
7.3.2 Step 2: Estimation of Model Parameters ... 187
7.3.2.1 Estimation for the Standard Linear Regression Model ... 187
7.3.2.2 Linear Regression Estimation for Complex Sample Survey Data ... 188
7.3.3 Step 3: Model Evaluation ... 193
7.3.3.1 Explained Variance and Goodness of Fit ... 193
7.3.3.2 Residual Diagnostics ... 194
7.3.3.3 Model Specification and Homogeneity of Variance ... 194
7.3.3.4 Normality of the Residual Errors ... 195
7.3.3.5 Outliers and Influence Statistics ... 196
7.3.4 Step 4: Inference ... 196
7.3.4.1 Inference Concerning Model Parameters ... 199
7.3.4.2 Prediction Intervals ... 202
7.4 Some Practical Considerations and Tools ... 204
7.4.1 Distribution of the Dependent Variable ... 204
7.4.2 Parameterization and Scaling for Independent Variables ... 205
7.4.3 Standardization of the Dependent and Independent Variables ... 208
7.4.4 Specification and Interpretation of Interactions and Nonlinear Relationships ... 208
7.4.5 Model-Building Strategies ... 210
7.5 Application: Modeling Diastolic Blood Pressure with the NHANES Data ... 211
7.5.1 Exploring the Bivariate Relationships ... 212
7.5.2 Naïve Analysis: Ignoring Sample Design Features ... 215
7.5.3 Weighted Regression Analysis ... 216

7.5.4 Appropriate Analysis: Incorporating All Sample Design Features 218
7.6 Exercises 224

8. Logistic Regression and Generalized Linear Models for Binary Survey Variables 229
8.1 Introduction 229
8.2 Generalized Linear Models for Binary Survey Responses 230
8.2.1 The Logistic Regression Model 231
8.2.2 The Probit Regression Model 234
8.2.3 The Complementary Log–Log Model 234
8.3 Building the Logistic Regression Model: Stage 1, Model Specification 235
8.4 Building the Logistic Regression Model: Stage 2, Estimation of Model Parameters and Standard Errors 236
8.5 Building the Logistic Regression Model: Stage 3, Evaluation of the Fitted Model 239
8.5.1 Wald Tests of Model Parameters 239
8.5.2 Goodness of Fit and Logistic Regression Diagnostics 243
8.6 Building the Logistic Regression Model: Stage 4, Interpretation and Inference 245
8.7 Analysis Application 251
8.7.1 Stage 1: Model Specification 252
8.7.2 Stage 2: Model Estimation 253
8.7.3 Stage 3: Model Evaluation 255
8.7.4 Stage 4: Model Interpretation/Inference 256
8.8 Comparing the Logistic, Probit, and Complementary Log–Log GLMs for Binary Dependent Variables 259
8.9 Exercises 262

9. Generalized Linear Models for Multinomial, Ordinal, and Count Variables 265
9.1 Introduction 265
9.2 Analyzing Survey Data Using Multinomial Logit Regression Models 265
9.2.1 The Multinomial Logit Regression Model 265
9.2.2 Multinomial Logit Regression Model: Specification Stage 267
9.2.3 Multinomial Logit Regression Model: Estimation Stage 268
9.2.4 Multinomial Logit Regression Model: Evaluation Stage 268

9.2.5 Multinomial Logit Regression Model: Interpretation Stage 270

9.2.6 Example: Fitting a Multinomial Logit Regression Model to Complex Sample Survey Data 271

9.3 Logistic Regression Models for Ordinal Survey Data 277

9.3.1 Cumulative Logit Regression Model 278

9.3.2 Cumulative Logit Regression Model: Specification Stage 279

9.3.3 Cumulative Logit Regression Model: Estimation Stage 279

9.3.4 Cumulative Logit Regression Model: Evaluation Stage 280

9.3.5 Cumulative Logit Regression Model: Interpretation Stage 281

9.3.6 Example: Fitting a Cumulative Logit Regression Model to Complex Sample Survey Data 282

9.4 Regression Models for Count Outcomes 286

9.4.1 Survey Count Variables and Regression Modeling Alternatives 286

9.4.2 Generalized Linear Models for Count Variables 288

9.4.2.1 The Poisson Regression Model 288

9.4.2.2 The Negative Binomial Regression Model 289

9.4.2.3 Two-Part Models: Zero-Inflated Poisson and Negative Binomial Regression Models 290

9.4.3 Regression Models for Count Data: Specification Stage 291

9.4.4 Regression Models for Count Data: Estimation Stage 292

9.4.5 Regression Models for Count Data: Evaluation Stage 292

9.4.6 Regression Models for Count Data: Interpretation Stage 293

9.4.7 Example: Fitting Poisson and Negative Binomial Regression Models to Complex Sample Survey Data 294

9.5 Exercises 298

10. Survival Analysis of Event History Survey Data 303

10.1 Introduction 303

10.2 Basic Theory of Survival Analysis 303

10.2.1 Survey Measurement of Event History Data 303

10.2.2 Data for Event History Models 305

10.2.3 Important Notation and Definitions 306

10.2.4 Models for Survival Analysis 307

10.3 (Nonparametric) Kaplan–Meier Estimation of the Survivor Function 308
10.3.1 K–M Model Specification and Estimation 309
10.3.2 K–M Estimator—Evaluation and Interpretation 310
10.3.3 K–M Survival Analysis Example 311
10.4 Cox Proportional Hazards Model 315
10.4.1 Cox Proportional Hazards Model: Specification 315
10.4.2 Cox Proportional Hazards Model: Estimation Stage 316
10.4.3 Cox Proportional Hazards Model: Evaluation and Diagnostics 317
10.4.4 Cox Proportional Hazards Model: Interpretation and Presentation of Results 319
10.4.5 Example: Fitting a Cox Proportional Hazards Model to Complex Sample Survey Data 319
10.5 Discrete Time Survival Models 322
10.5.1 The Discrete Time Logistic Model 323
10.5.2 Data Preparation for Discrete Time Survival Models 324
10.5.3 Discrete Time Models: Estimation Stage 327
10.5.4 Discrete Time Models: Evaluation and Interpretation 328
10.5.5 Fitting a Discrete Time Model to Complex Sample Survey Data 329
10.6 Exercises 333

11. Multiple Imputation: Methods and Applications for Survey Analysts 335
11.1 Introduction 335
11.2 Important Missing Data Concepts 336
11.2.1 Sources and Patterns of Item-Missing Data in Surveys 336
11.2.2 Item-Missing Data Mechanisms 338
11.2.3 Implications of Item-Missing Data for Survey Data Analysis 341
11.2.4 Review of Strategies to Address Item-Missing Data in Surveys 342
11.3 An Introduction to Imputation and the Multiple Imputation Method 345
11.3.1 A Brief History of Imputation Procedures 345
11.3.2 Why the Multiple Imputation Method? 346
11.3.3 Overview of Multiple Imputation and MI Phases 348
11.4 Models for Multiply Imputing Missing Data 350
11.4.1 Choosing the Variables to Include in the Imputation Model 350

9.2.5 Multinomial Logit Regression Model: Interpretation Stage 270
9.2.6 Example: Fitting a Multinomial Logit Regression Model to Complex Sample Survey Data 271
9.3 Logistic Regression Models for Ordinal Survey Data 277
9.3.1 Cumulative Logit Regression Model 278
9.3.2 Cumulative Logit Regression Model: Specification Stage 279
9.3.3 Cumulative Logit Regression Model: Estimation Stage 279
9.3.4 Cumulative Logit Regression Model: Evaluation Stage 280
9.3.5 Cumulative Logit Regression Model: Interpretation Stage 281
9.3.6 Example: Fitting a Cumulative Logit Regression Model to Complex Sample Survey Data 282
9.4 Regression Models for Count Outcomes 286
9.4.1 Survey Count Variables and Regression Modeling Alternatives 286
9.4.2 Generalized Linear Models for Count Variables 288
9.4.2.1 The Poisson Regression Model 288
9.4.2.2 The Negative Binomial Regression Model 289
9.4.2.3 Two-Part Models: Zero-Inflated Poisson and Negative Binomial Regression Models 290
9.4.3 Regression Models for Count Data: Specification Stage 291
9.4.4 Regression Models for Count Data: Estimation Stage 292
9.4.5 Regression Models for Count Data: Evaluation Stage 292
9.4.6 Regression Models for Count Data: Interpretation Stage 293
9.4.7 Example: Fitting Poisson and Negative Binomial Regression Models to Complex Sample Survey Data 294
9.5 Exercises 298

10. Survival Analysis of Event History Survey Data 303
10.1 Introduction 303
10.2 Basic Theory of Survival Analysis 303
10.2.1 Survey Measurement of Event History Data 303
10.2.2 Data for Event History Models 305
10.2.3 Important Notation and Definitions 306
10.2.4 Models for Survival Analysis 307

A.1.1 Historical Perspective....................400
A.1.2 Software for Sampling Error Estimation....................401
A.2 Overview of Stata® Version 10+....................407
A.3 Overview of SAS® Version 9.2....................410
A.3.1 The SAS SURVEY Procedures....................411
A.4 Overview of SUDAAN® Version 9.0....................414
A.4.1 The SUDAAN Procedures....................415
A.5 Overview of SPSS®....................421
A.5.1 The SPSS Complex Samples Commands....................422
A.6 Overview of Additional Software....................427
A.6.1 WesVar®....................427
A.6.2 IVEware (Imputation and Variance Estimation Software)....................428
A.6.3 Mplus....................429
A.6.4 The R survey Package....................429
A.7 Summary....................430

References....................431

Index....................443