Contents

List of figures xv

List of tables xxiii

List of boxed tips xxv

Preface xxix

Support materials for the book xxxv

1 Getting started 1

1.1 Conventions 1

1.2 Introduction 4

1.3 The Stata screen 7

1.4 Using an existing dataset 9

1.5 An example of a short Stata session 11

1.6 Video aids to learning Stata 18

1.7 Summary 19

1.8 Exercises 19

2 Entering data 21

2.1 Creating a dataset 21

2.2 An example questionnaire 23

2.3 Developing a coding system 24

2.4 Entering data using the Data Editor 29

2.4.1 Value labels 33

2.5 The Variables Manager 34

2.6 The Data Editor (Browse) view 40

2.7 Saving your dataset 41

2.8 Checking the data 43

2.9 Summary 49
2.10 Exercises 49
3 Preparing data for analysis **51**
3.1 Introduction 51
3.2 Planning your work 51
3.3 Creating value labels 57
3.4 Reverse-code variables 60
3.5 Creating and modifying variables 65
3.6 Creating scales 70
3.7 Saving some of your data 73
3.8 Summary 74
3.9 Exercises 75
4 Working with commands, do-files, and results **77**
4.1 Introduction 77
4.2 How Stata commands are constructed 78
4.3 Creating a do-file 82
4.4 Copying your results to a word processor 88
4.5 Logging your command file 89
4.6 Summary 91
4.7 Exercises 92
5 Descriptive statistics and graphs for one variable **93**
5.1 Descriptive statistics and graphs 93
5.2 Where is the center of a distribution? 94
5.3 How dispersed is the distribution? 98
5.4 Statistics and graphs—unordered categories 100
5.5 Statistics and graphs—ordered categories and variables 110
5.6 Statistics and graphs—quantitative variables 112
5.7 Summary 119
5.8 Exercises 120

6 Statistics and graphs for two categorical variables 123

6.1 Relationship between categorical variables 123

6.2 Cross-tabulation 124

6.3 Chi-squared test 127

6.3.1 Degrees of freedom 129

6.3.2 Probability tables 129

6.4 Percentages and measures of association 132

6.5 Odds ratios when dependent variable has two categories 135

6.6 Ordered categorical variables 137

6.7 Interactive tables 140

6.8 Tables—linking categorical and quantitative variables 142

6.9 Power analysis when using a chi-squared test of significance 145

6.10 Summary 147

6.11 Exercises 148

7 Tests for one or two means 151

7.1 Introduction to tests for one or two means 151

7.2 Randomization 154

7.3 Random sampling 156

7.4 Hypotheses 156

7.5 One-sample test of a proportion 157

7.6 Two-sample test of a proportion 159

7.7 One-sample test of means 164

7.8 Two-sample test of group means 166

7.8.1 Testing for unequal variances 171

7.9 Repeated-measures t test 172

7.10 Power analysis 174

7.11 Nonparametric alternatives 182

7.11.1 Mann–Whitney two-sample rank-sum test 182

7.11.2 Nonparametric alternative: Median test 183

7.12 Video tutorial related to this chapter 184

7.13 Summary . . . 184
7.14 Exercises . . . 185

8 Bivariate correlation and regression 189

8.1 Introduction to bivariate correlation and regression . . . 189
8.2 Scattergrams . . . 190
8.3 Plotting the regression line . . . 196
8.4 An alternative to producing a scattergram, binscatter . . . 197
8.5 Correlation . . . 201
8.6 Regression . . . 207
8.7 Spearman's rho: Rank-order correlation for ordinal data . . . 212
8.8 Power analysis with correlation . . . 213
8.9 Summary . . . 215
8.10 Exercises . . . 215

9 Analysis of variance 219

9.1 The logic of one-way analysis of variance . . . 219
9.2 ANOVA example . . . 220
9.3 ANOVA example with nonexperimental data . . . 229
9.4 Power analysis for one-way ANOVA . . . 232
9.5 A nonparametric alternative to ANOVA . . . 234
9.6 Analysis of covariance . . . 237
9.7 Two-way ANOVA . . . 249
9.8 Repeated-measures design . . . 255
9.9 Intraclass correlation—measuring agreement . . . 260
9.10 Power analysis with ANOVA . . . 262
9.10.1 Power analysis for one-way ANOVA . . . 263
9.10.2 Power analysis for two-way ANOVA . . . 265
9.10.3 Power analysis for repeated-measures ANOVA . . . 267
9.10.4 Summary of power analysis for ANOVA . . . 269
9.11 Summary . . . 270
9.12 Exercises . . . 270

10 Multiple regression **273**

10.1 Introduction to multiple regression 273

10.2 What is multiple regression? 274

10.3 The basic multiple regression command 275

10.4 Increment in R-squared: Semipartial correlations 279

10.5 Is the dependent variable normally distributed? 281

10.6 Are the residuals normally distributed? 284

10.7 Regression diagnostic statistics 288

10.7.1 Outliers and influential cases 289

10.7.2 Influential observations: DFbeta 291

10.7.3 Combinations of variables may cause problems 292

10.8 Weighted data 294

10.9 Categorical predictors and hierarchical regression 296

10.10 A shortcut for working with a categorical variable 305

10.11 Fundamentals of interaction 306

10.12 Nonlinear relations 313

10.12.1 Fitting a quadratic model 315

10.12.2 Centering when using a quadratic term 321

10.12.3 Do we need to add a quadratic component? 323

10.13 Power analysis in multiple regression 325

10.14 Summary 328

10.15 Exercises 329

11 Logistic regression **333**

11.1 Introduction to logistic regression 333

11.2 An example 334

11.3 What is an odds ratio and a logit? 338

11.3.1 The odds ratio 340

11.3.2 The logit transformation 340

11.4 Data used in the rest of the chapter 341

11.5 Logistic regression 342

11.6 Hypothesis testing 353
11.6.1 Testing individual coefficients 353
11.6.2 Testing sets of coefficients 354
11.7 More on interpreting results from logistic regression 356
11.8 Nested logistic regressions 360
11.9 Power analysis when doing logistic regression 362
11.10 Next steps for using logistic regression and its extensions 365
11.11 Summary 365
11.12 Exercises 366
12 Measurement, reliability, and validity **369**
12.1 Overview of reliability and validity 369
12.2 Constructing a scale 370
12.2.1 Generating a mean score for each person 371
12.3 Reliability 372
12.3.1 Stability and test–retest reliability 375
12.3.2 Equivalence 376
12.3.3 Split-half and alpha reliability—internal consistency 376
12.3.4 Kuder–Richardson reliability for dichotomous items 379
12.3.5 Rater agreement—kappa (κ) 380
12.4 Validity 383
12.4.1 Expert judgment 383
12.4.2 Criterion-related validity 384
12.4.3 Construct validity 385
12.5 Factor analysis 386
12.6 PCF analysis 391
12.6.1 Orthogonal rotation: Varimax 394
12.6.2 Oblique rotation: Promax 396
12.7 But we wanted one scale, not four scales 397
12.7.1 Scoring our variable 398
12.8 Summary 399

12.9 Exercises 400

13 Working with missing values—multiple imputation 401

13.1 The nature of the problem 401

13.2 Multiple imputation and its assumptions about the mechanism for missingness 403

13.3 What variables do we include when doing imputations? 405

13.4 Multiple imputation 406

13.5 A detailed example 407

13.5.1 Preliminary analysis 408

13.5.2 Setup and multiple-imputation stage 410

13.5.3 The analysis stage 413

13.5.4 For those who want an R^2 and standardized βs 414

13.5.5 When impossible values are imputed 416

13.6 Summary 418

13.7 Exercises 419

14 The sem and gsem commands 421

14.1 Linear regression using sem 421

14.1.1 Using the SEM Builder to fit a basic regression model . . . 423

14.2 A quick way to draw a regression model and a fresh start 430

14.2.1 Using sem without the SEM Builder 433

14.3 The gsem command for logistic regression 433

14.3.1 Fitting the model using the logit command 434

14.3.2 Fitting the model using the gsem command 436

14.4 Path analysis and mediation 442

14.5 Conclusions and what is next for the sem command 446

14.6 Exercises 448

15 An introduction to multilevel analysis 451

15.1 Questions and data for groups of individuals 451

15.2 Questions and data for a longitudinal multilevel application 452

15.3 Fixed-effects regression models 453

15.4 Random-effects regression models 454
15.5 An applied example 456
15.5.1 Research questions 456
15.5.2 Reshaping data to do multilevel analysis 457
15.6 A quick visualization of our data 460
15.7 Random-intercept model 461
15.7.1 Random intercept—linear model 461
15.7.2 Random-intercept model—quadratic term 464
15.7.3 Treating time as a categorical variable 468
15.8 Random-coefficients model 471
15.9 Including a time-invariant covariate 474
15.10 Summary 479
15.11 Exercises 479
16 Item response theory (IRT) 481
16.1 How are IRT measures of variables different from summated scales? . 482
16.2 Overview of three IRT models for dichotomous items 484
16.2.1 The one-parameter logistic (1PL) model 484
16.2.2 The two-parameter logistic (2PL) model 486
16.2.3 The three-parameter logistic (3PL) model 487
16.3 Fitting the 1PL model using Stata 487
16.3.1 The estimation 490
16.3.2 How important is each of the items? 492
16.3.3 An overall evaluation of our scale 494
16.3.4 Estimating the latent score 495
16.4 Fitting a 2PL IRT model 496
16.4.1 Fitting the 2PL model 497
16.5 The graded response model—IRT for Likert-type items 502
16.5.1 The data 502
16.5.2 Fitting our graded response model 504
16.5.3 Estimating a person's score 509

16.6 Reliability of the fitted IRT model 509
16.7 Using the Stata menu system 512
16.8 Extensions of IRT 515
16.9 Exercises 516
A What's next? 519
A.1 Introduction to the appendix 519
A.2 Resources 519
A.2.1 Web resources 520
A.2.2 Books about Stata 522
A.2.3 Short courses 525
A.2.4 Acquiring data 525
A.2.5 Learning from the postestimation methods 526
A.3 Summary 527
References 531
Author index 535
Subject index 537