Contents

List of figures xv

List of tables xxiii

List of boxed tips xxv

Preface xxix

Support materials for the book xxxv

Glossary of acronyms xxxix

Glossary of mathematical and statistical symbols xli

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 50
2.10 Exercises 50

3 Preparing data for analysis **51**
3.1 Introduction 51
3.2 Planning your work 52
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 133

6.5 Odds ratios when dependent variable has two categories 136

6.6 Ordered categorical variables 138

6.7 Interactive tables 141

6.8 Tables—linking categorical and quantitative variables 143

6.9 Power analysis when using a chi-squared test of significance 146

6.10 Summary 149

6.11 Exercises 149

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 160

7.7 One-sample test of means 164

7.8 Two-sample test of group means 166

7.8.1 Testing for unequal variances 175

7.9 Repeated-measures t test 176

7.10 Power analysis 178

7.11 Nonparametric alternatives 186

7.11.1 Mann–Whitney two-sample rank-sum test 186

7.11.2 Nonparametric alternative: Median test 187

7.12 Video tutorial related to this chapter 188
7.13 Summary 188
7.14 Exercises 189
8 Bivariate correlation and regression 193
8.1 Introduction to bivariate correlation and regression 193
8.2 Scattergrams 194
8.3 Plotting the regression line 199
8.4 An alternative to producing a scattergram, binscatter 201
8.5 Correlation 205
8.6 Regression 210
8.7 Spearman's rho: Rank-order correlation for ordinal data 215
8.8 Power analysis with correlation 216
8.9 Summary 218
8.10 Exercises 218
9 Analysis of variance 221
9.1 The logic of one-way analysis of variance 221
9.2 ANOVA example 222
9.3 ANOVA example with nonexperimental data 231
9.4 Power analysis for one-way ANOVA 234
9.5 A nonparametric alternative to ANOVA 236
9.6 Analysis of covariance 239
9.7 Two-way ANOVA 250
9.8 Repeated-measures design 256
9.9 Intraclass correlation—measuring agreement 261
9.10 Power analysis with ANOVA 263
9.10.1 Power analysis for one-way ANOVA 264
9.10.2 Power analysis for two-way ANOVA 266
9.10.3 Power analysis for repeated-measures ANOVA 268
9.10.4 Summary of power analysis for ANOVA 270
9.11 Summary 271

9.12 Exercises 271

10 Multiple regression 275

10.1 Introduction to multiple regression 275

10.2 What is multiple regression? 276

10.3 The basic multiple regression command 277

10.4 Increment in R-squared: Semipartial correlations 281

10.5 Is the dependent variable normally distributed? 283

10.6 Are the residuals normally distributed? 286

10.7 Regression diagnostic statistics 291

10.7.1 Outliers and influential cases 291

10.7.2 Influential observations: DFbeta 293

10.7.3 Combinations of variables may cause problems 294

10.8 Weighted data 296

10.9 Categorical predictors and hierarchical regression 299

10.10 A shortcut for working with a categorical variable 308

10.11 Fundamentals of interaction 309

10.12 Nonlinear relations 316

10.12.1 Fitting a quadratic model 318

10.12.2 Centering when using a quadratic term 324

10.12.3 Do we need to add a quadratic component? 326

10.13 Power analysis in multiple regression 328

10.14 Summary 333

10.15 Exercises 335

11 Logistic regression 339

11.1 Introduction to logistic regression 339

11.2 An example 340

11.3 What is an odds ratio and a logit? 344

11.3.1 The odds ratio 346

11.3.2 The logit transformation 346

11.4 Data used in the rest of the chapter 347

11.5 Logistic regression 349
11.6 Hypothesis testing 360
11.6.1 Testing individual coefficients 361
11.6.2 Testing sets of coefficients 362
11.7 Margins: More on interpreting results from logistic regression 364
11.8 Nested logistic regressions 372
11.9 Power analysis when doing logistic regression 374
11.10 Next steps for using logistic regression and its extensions 377
11.11 Summary 377
11.12 Exercises 378

12 Measurement, reliability, and validity 381

12.1 Overview of reliability and validity 381
12.2 Constructing a scale 382
12.2.1 Generating a mean score for each person 383
12.3 Reliability 385
12.3.1 Stability and test–retest reliability 387
12.3.2 Equivalence 388
12.3.3 Split-half and alpha reliability—internal consistency 388
12.3.4 Kuder–Richardson reliability for dichotomous items 392
12.3.5 Rater agreement—kappa (κ) 393
12.4 Validity 395
12.4.1 Expert judgment 396
12.4.2 Criterion-related validity 397
12.4.3 Construct validity 397
12.5 Factor analysis 402
12.6 PCF analysis 406
12.6.1 Orthogonal rotation: Varimax 410
12.6.2 Oblique rotation: Promax 412
12.7 But we wanted one scale, not four scales 413
12.7.1 Scoring our variable 414

12.8 Summary 415

12.9 Exercises 416

13 Structural equation and generalized structural equation modeling 417

13.1 Linear regression using sem 417

13.1.1 Using the sem command directly 419

13.1.2 SEM and working with missing values 420

13.1.3 Exploring missing values and auxiliary variables 426

13.1.4 Getting auxiliary variables into your SEM command 428

13.2 A quick way to draw a regression model 429

13.3 The gsem command for logistic regression 432

13.3.1 Fitting the model using the logit command 432

13.3.2 Fitting the model using the gsem command 434

13.4 Path analysis and mediation 440

13.5 Conclusions and what is next for the sem command 444

13.6 Exercises 446

14 Working with missing values—multiple imputation 449

14.1 Working with missing values—multiple imputation 449

14.2 What variables do we include when doing imputations? 450

14.3 The nature of the problem 452

14.4 Multiple imputation and its assumptions about the mechanism for missingness 453

14.5 Multiple imputation 455

14.6 A detailed example 456

14.6.1 Preliminary analysis 457

14.6.2 Setup and multiple-imputation stage 460

14.6.3 The analysis stage 462

14.6.4 For those who want an R^2 and standardized βs 464

14.6.5 When impossible values are imputed 466

14.7 Summary 468

14.8 Exercises 469

15 An introduction to multilevel analysis 471

15.1 Questions and data for groups of individuals 471
15.2 Questions and data for a longitudinal multilevel application 472
15.3 Fixed-effects regression models 473
15.4 Random-effects regression models 474
15.5 An applied example 476
15.5.1 Research questions 476
15.5.2 Reshaping data to do multilevel analysis 477
15.6 A quick visualization of our data 480
15.7 Random-intercept model 481
15.7.1 Random intercept—linear model 481
15.7.2 Random-intercept model—quadratic term 484
15.7.3 Treating time as a categorical variable 488
15.8 Random-coefficients model 491
15.9 Including a time-invariant covariate 494
15.10 Summary 499
15.11 Exercises 500

16 Item response theory (IRT) 501

16.1 How are IRT measures of variables different from summated scales? . 502
16.2 Overview of three IRT models for dichotomous items . . 504
16.2.1 The one-parameter logistic (1PL) model 504
16.2.2 The two-parameter logistic (2PL) model 506
16.2.3 The three-parameter logistic (3PL) model 507
16.3 Fitting the 1PL model using Stata 507
16.3.1 The estimation 510
16.3.2 How important is each of the items? 512
16.3.3 An overall evaluation of our scale 514
16.3.4 Estimating the latent score 515
16.4 Fitting a 2PL IRT model 516
16.4.1 Fitting the 2PL model 517

16.5 The graded response model—IRT for Likert-type items 522
16.5.1 The data 522
16.5.2 Fitting our graded response model 524
16.5.3 Estimating a person's score 529
16.6 Reliability of the fitted IRT model 529
16.7 Using the Stata menu system 532
16.8 Extensions of IRT 535
16.9 Exercises 536
A What's next? 539
A.1 Introduction to the appendix 539
A.2 Resources 539
A.2.1 Web resources 540
A.2.2 Books about Stata 542
A.2.3 Short courses 544
A.2.4 Acquiring data 545
A.2.5 Learning from the postestimation methods 546
A.3 Summary 547
References 551
Author index 555
Subject index 557