Contents

List of Tables xvii

List of Figures xix

Preface xxi

1 "The first time" 1

1.1 Starting Stata . . . 1

1.2 Setting up your screen . . . 2

1.3 Your first analysis . . . 2

1.3.1 Inputting commands . . . 2

1.3.2 Files and the working memory . . . 3

1.3.3 Loading data . . . 3

1.3.4 Variables and observations . . . 5

1.3.5 Looking at data . . . 6

1.3.6 Interrupting a command and repeating a command . . . 8

1.3.7 The variable list . . . 8

1.3.8 The in qualifier . . . 8

1.3.9 Summary statistics . . . 9

1.3.10 The if qualifier . . . 11

1.3.11 Define missing values . . . 11

1.3.12 The by prefix . . . 12

1.3.13 Command options . . . 13

1.3.14 Frequency tables . . . 14

1.3.15 Variable labels and value labels . . . 15

1.3.16 Graphs . . . 16

1.3.17 Getting help . . . 16

1.3.18 Recoding of variables 18
1.3.19 Linear regression 19
1.4 Do-files 19
1.5 Exiting Stata 21
1.6 Exercises 22

2 Working with do-files **25**

2.1 From interactive work to working with a do-file 25
2.1.1 Alternative 1 25
2.1.2 Alternative 2 27
2.2 Designing do-files 30
2.2.1 Comments 31
2.2.2 Line breaks 32
2.2.3 Some crucial commands 33
2.3 Organizing your work 35
2.4 Exercises 39

3 The grammar of Stata **41**

3.1 The elements of Stata commands 41
3.1.1 Stata commands 41
3.1.2 The variable list 43
List of variables: Required or optional 43
Abbreviation rules 43
Special listings 45
3.1.3 Options 45
3.1.4 The in qualifier 47
3.1.5 The if qualifier 48
3.1.6 Expressions 50
Operators 51
Functions 52
3.1.7 Lists of numbers 53
3.1.8 Using filenames 53

3.2 Repeating similar commands . . . 55
3.2.1 The by prefix . . . 55
3.2.2 The foreach loop . . . 57
The types of foreach lists . . . 58
Several commands within a foreach loop . . . 59
3.2.3 The forvalues loop . . . 60
3.3 Weights . . . 60
Frequency weights . . . 61
Analytic weights . . . 63
Probability weights . . . 64
3.4 Exercises . . . 65

4 General comments on the statistical commands 67
4.1 Exercises . . . 70

5 Creating and changing variables 73
5.1 The commands generate and replace . . . 73
5.1.1 Variable names . . . 74
5.1.2 Some examples . . . 75
5.1.3 Changing codes with by, _n, and _N . . . 78
5.1.4 Subscripts . . . 82
5.2 Specialized recoding commands . . . 84
5.2.1 The recode command . . . 84
5.2.2 The egen command . . . 84
5.3 More tools for recoding data . . . 86
5.3.1 String functions . . . 86
5.3.2 Date and time functions . . . 90
Dates . . . 91
Time . . . 93
5.4 Commands for dealing with missing values . . . 96
5.5 Labels . . . 98

5.6 Storage types, or the ghost in the machine . . . 100
5.7 Exercises . . . 101
6 Creating and changing graphs **103**
6.1 A primer on graph syntax . . . 103
6.2 Graph types . . . 104
6.2.1 Examples . . . 105
6.2.2 Specialized graphs . . . 107
6.3 Graph elements . . . 107
6.3.1 Appearance of data . . . 109
Choice of marker . . . 111
Marker colors . . . 112
Marker size . . . 113
Lines . . . 113
6.3.2 Graph and plot regions . . . 116
Graph size . . . 117
Plot region . . . 117
Scaling the axes . . . 118
6.3.3 Information inside the plot region . . . 119
Reference lines . . . 120
Labeling inside the plot region . . . 120
6.3.4 Information outside the plot region . . . 124
Labeling the axes . . . 124
Tick lines . . . 127
Axis titles . . . 128
The legend . . . 129
Graph titles . . . 130
6.4 Multiple graphs . . . 131
6.4.1 Overlaying many twoway graphs . . . 131
6.4.2 Option by() . . . 133
6.4.3 Combining graphs . . . 134

Quantile plot . . . 175
Comparing distributions with Q–Q plots . . . 178
7.4 Exercises . . . 179
8 Introduction to linear regression **181**
8.1 Simple linear regression . . . 184
8.1.1 The basic principle . . . 184
8.1.2 Linear regression using Stata . . . 188
The table of coefficients . . . 188
Standard errors . . . 191
The table of ANOVA results . . . 192
The model fit table . . . 195
8.2 Multiple regression . . . 196
8.2.1 Multiple regression using Stata . . . 197
8.2.2 More computations . . . 199
Adjusted R^2 . . . 199
Standardized regression coefficients . . . 200
8.2.3 What does "under control" mean? . . . 201
8.3 Regression diagnostics . . . 203
8.3.1 Violation of $E(\epsilon_i) = 0$ . . . 204
Linearity . . . 206
Influential cases . . . 209
Omitted variables . . . 218
Multicollinearity . . . 218
8.3.2 Violation of $Var(\epsilon_i) = \sigma^2$ . . . 219
8.3.3 Violation of $Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$ . . . 221
8.4 Model extensions . . . 222
8.4.1 Categorical independent variables . . . 222
8.4.2 Interaction terms . . . 225
8.4.3 Regression models using transformed variables . . . 228
Nonlinear relations . . . 229

Eliminating heteroskedasticity 231
8.5 More on standard errors 232
8.5.1 Bootstrap techniques 232
8.5.2 Confidence intervals in cluster samples 234
8.6 Advanced techniques 236
8.6.1 Median regression 236
8.6.2 Regression models for panel data 237
From wide to long format 238
Fixed-effects models 242
8.6.3 Error-components models 245
8.7 Exercises 248
9 Regression models for categorical dependent variables 249
9.1 The linear probability model 250
9.2 Basic concepts 253
9.2.1 Odds, log odds, and odds ratios 253
9.2.2 Excursion: The maximum likelihood principle 258
9.3 Logistic regression with Stata 261
9.3.1 The coefficient table 263
Sign interpretation 264
Interpretation with odds ratios 264
Probability interpretation 265
9.3.2 The iteration block 266
9.3.3 The model fit block 267
Classification tables 268
Pearson chi-squared 271
9.4 Logistic regression diagnostics 272
9.4.1 Linearity 272
9.4.2 Influential cases 276
9.5 Likelihood-ratio test 279
9.6 Refined models 281

9.6.1 Nonlinear relationships . . . 281
9.6.2 Categorical independent variables . . . 282
9.6.3 Interaction effects . . . 284
9.7 Advanced techniques . . . 285
9.7.1 Probit models . . . 285
9.7.2 Multinomial logistic regression . . . 288
9.7.3 Models for ordinal data . . . 292
9.8 Exercises . . . 294
10 Reading and writing data **297**
10.1 The goal: the data matrix . . . 297
10.2 Importing machine-readable data . . . 298
10.2.1 Reading system files from other packages . . . 299
10.2.2 Reading ASCII text files . . . 300
Reading data in spreadsheet format . . . 300
Reading data in free format . . . 302
Reading data in fixed format . . . 304
10.3 Inputting data . . . 307
10.3.1 Input data using the Data Editor . . . 307
10.3.2 The input command . . . 308
10.4 Combining data . . . 312
10.4.1 The GSOEP database . . . 312
10.4.2 The merge command . . . 314
The merge procedure . . . 315
Keeping track of observations . . . 317
Merging more than two files . . . 318
Merging data on different levels . . . 320
10.4.3 The append command . . . 323
10.5 Saving and exporting data . . . 326
10.6 Handling large datasets . . . 327
10.6.1 Rules for handling the working memory . . . 327

10.6.2 Using oversized datasets . . . 329
10.7 Exercises . . . 330
11 Do-files for advanced users and user-written programs **333**
11.1 Two examples of usage . . . 333
11.2 Four programming tools . . . 335
11.2.1 Local macros . . . 335
Calculating with local macros . . . 336
Combining local macros . . . 337
Changing local macros . . . 337
11.2.2 Do-files . . . 339
11.2.3 Programs . . . 339
The problem of redefinition . . . 340
The problem of naming . . . 341
The problem of error checking . . . 341
11.2.4 Programs in do-files and ado-files . . . 342
11.3 User-written Stata commands . . . 345
11.3.1 Parsing variable lists . . . 348
11.3.2 Parsing options . . . 349
11.3.3 Parsing if and in qualifiers . . . 351
11.3.4 Generating an unknown number of variables . . . 352
11.3.5 Default values . . . 354
11.3.6 Extended macro functions . . . 356
11.3.7 Avoiding changes in the dataset . . . 358
11.3.8 Help files . . . 359
11.4 Exercises . . . 360
12 Around Stata **363**
12.1 Resources and information . . . 363
12.2 Taking care of Stata . . . 364
12.3 Additional procedures . . . 366
12.3.1 SJ and STB ado-files . . . 366

12.3.2 SSC ado-files 367
12.3.3 Other ado-files 368
12.4 Exercises 369
References **371**
Author index **377**
Subject index **379**