Data Management Using Stata: A Practical Handbook

Second Edition

MICHAEL N. MITCHELL

A Stata Press Publication
StataCorp LLC
College Station, Texas

Contents

Acknowledgments v

List of tables xiii

List of figures xv

Preface to the Second Edition xvii

Preface xix

1 Introduction 1

1.1 Using this book . . . 2

1.2 Overview of this book . . . 3

1.3 Listing observations in this book . . . 5

1.4 More online resources . . . 8

2 Reading and importing data files 11

2.1 Introduction . . . 12

2.2 Reading Stata datasets . . . 14

2.3 Importing Excel spreadsheets . . . 16

2.4 Importing SAS files . . . 20

2.4.1 Importing SAS .sas7bdat files . . . 20

2.4.2 Importing SAS XPORT Version 5 files . . . 23

2.4.3 Importing SAS XPORT Version 8 files . . . 24

2.5 Importing SPSS files . . . 25

2.6 Importing dBase files . . . 28

2.7 Importing raw data files . . . 29

2.7.1 Importing comma-separated and tab-separated files . . . 30

2.7.2 Importing space-separated files . . . 33

2.7.3 Importing fixed-column files . . . 35

2.7.4 Importing fixed-column files with multiple lines of raw data per observation . . . 39
2.8 Common errors when reading and importing files . . . 41
2.9 Entering data directly into the Stata Data Editor . . . 43

3 Saving and exporting data files **51**

3.1 Introduction . . . 52
3.2 Saving Stata datasets . . . 52
3.3 Exporting Excel files . . . 55
3.4 Exporting SAS XPORT Version 8 files . . . 59
3.5 Exporting SAS XPORT Version 5 files . . . 60
3.6 Exporting dBase files . . . 61
3.7 Exporting comma-separated and tab-separated files . . . 62
3.8 Exporting space-separated files . . . 65
3.9 Exporting Excel files revisited: Creating reports . . . 66

4 Data cleaning **75**

4.1 Introduction . . . 76
4.2 Double data entry . . . 77
4.3 Checking individual variables . . . 81
4.4 Checking categorical by categorical variables . . . 85
4.5 Checking categorical by continuous variables . . . 87
4.6 Checking continuous by continuous variables . . . 91
4.7 Correcting errors in data . . . 94
4.8 Identifying duplicates . . . 98
4.9 Final thoughts on data cleaning . . . 106

5 Labeling datasets **107**

5.1 Introduction . . . 108
5.2 Describing datasets . . . 108
5.3 Labeling variables . . . 114
5.4 Labeling values . . . 116
5.5 Labeling utilities . . . 122

5.6 Labeling variables and values in different languages 127
5.7 Adding comments to your dataset using notes 132
5.8 Formatting the display of variables 136
5.9 Changing the order of variables in a dataset 140
6 Creating variables **145**
6.1 Introduction 146
6.2 Creating and changing variables 146
6.3 Numeric expressions and functions 150
6.4 String expressions and functions 152
6.5 Recoding 160
6.6 Coding missing values 166
6.7 Dummy variables 168
6.8 Date variables 172
6.9 Date-and-time variables 179
6.10 Computations across variables 186
6.11 Computations across observations 188
6.12 More examples using the egen command 190
6.13 Converting string variables to numeric variables 192
6.14 Converting numeric variables to string variables 199
6.15 Renaming and ordering variables 201
7 Combining datasets **209**
7.1 Introduction 210
7.2 Appending: Appending datasets 210
7.3 Appending: Problems 215
7.4 Merging: One-to-one match merging 225
7.5 Merging: One-to-many match merging 231
7.6 Merging: Merging multiple datasets 236
7.7 Merging: Update merges 240
7.8 Merging: Additional options when merging datasets 243
7.9 Merging: Problems merging datasets 249

7.10 Joining datasets . . . 253
7.11 Crossing datasets . . . 255
8 Processing observations across subgroups 259
8.1 Introduction . . . 260
8.2 Obtaining separate results for subgroups . . . 260
8.3 Computing values separately by subgroups . . . 262
8.4 Computing values within subgroups: Subscripting observations . . . 266
8.5 Computing values within subgroups: Computations across observations . . . 272
8.6 Computing values within subgroups: Running sums . . . 273
8.7 Computing values within subgroups: More examples . . . 276
8.8 Comparing the by and tsset commands . . . 282
9 Changing the shape of your data 287
9.1 Introduction . . . 288
9.2 Wide and long datasets . . . 288
9.3 Introduction to reshaping long to wide . . . 298
9.4 Reshaping long to wide: Problems . . . 301
9.5 Introduction to reshaping wide to long . . . 303
9.6 Reshaping wide to long: Problems . . . 306
9.7 Multilevel datasets . . . 311
9.8 Collapsing datasets . . . 314
10 Programming for data management: Part I 317
10.1 Introduction . . . 317
10.2 Tips on long-term goals in data management . . . 318
10.3 Executing do-files and making log files . . . 322
10.4 Automating data checking . . . 328
10.5 Combining do-files . . . 332
10.6 Introducing Stata macros . . . 336
10.7 Manipulating Stata macros . . . 341
10.8 Repeating commands by looping over variables . . . 344

10.9 Repeating commands by looping over numbers 351
10.10 Repeating commands by looping over anything 353
10.11 Accessing results stored from Stata commands 355

11 Programming for data management: Part II **359**

11.1 Writing Stata programs for data management 359
11.2 Program 1: hello 364
11.3 Where to save your Stata programs 375
11.4 Program 2: Multilevel counting 377
11.5 Program 3: Tabulations in list format 387
11.6 Program 4: Scoring the simple depression scale 394
11.7 Program 5: Standardizing variables 402
11.8 Program 6: Checking variable labels 410
11.9 Program 7: Checking value labels 416
11.10 Program 8: Customized describe command 418
11.11 Program 9: Customized summarize command 432
11.12 Program 10: Checking for unlabeled values 439
11.13 Tips on debugging Stata programs 445
11.14 Final thoughts: Writing Stata programs for data management 450

A Common elements **453**

A.1 Introduction 454
A.2 Overview of Stata syntax 454
A.3 Working across groups of observations with by 456
A.4 Comments 458
A.5 Data types 459
A.6 Logical expressions 469
A.7 Functions 474
A.8 Subsetting observations with if and in 477
A.9 Subsetting observations and variables with keep and drop 480
A.10 Missing values 483
A.11 Referring to variable lists 487

A.12 Frames 490
A.12.1 Frames example 1: Can I interrupt you for a quick question? 491
A.12.2 Frames example 2: Juggling related tasks 493
A.12.3 Frames example 3: Checking double data entry 496
Subject index **503**