The Workflow of Data Analysis Using Stata

J. SCOTT LONG
Departments of Sociology and Statistics
Indiana University-Bloomington

A Stata Press Publication
StataCorp LP
College Station, Texas

Contents

List of tables xxi

List of figures xxiii

List of examples xxv

Preface xxix

A word about fonts, files, commands, and examples xxxiii

Introduction 1

1.1 Replication: The guiding principle for workflow 2

1.2 Steps in the workflow 3

1.2.1 Cleaning data 4

1.2.2 Running analysis 4

1.2.3 Presenting results 4

1.2.4 Protecting files 4

1.3 Tasks within each step 5

1.3.1 Planning 5

1.3.2 Organization 5

1.3.3 Documentation 5

1.3.4 Execution 6

1.4 Criteria for choosing a workflow 6

1.4.1 Accuracy 6

1.4.2 Efficiency 6

1.4.3 Simplicity 7

1.4.4 Standardization 7

1.4.5 Automation 7

1.4.6 Usability 7

1.4.7 Scalability 8

1.5 Changing your workflow 8

1.6 How the book is organized 8

2 Planning, organizing, and documenting 11

2.1 The cycle of data analysis 13

2.2 Planning 14

2.3 Organization 18

2.3.1 Principles for organization 18

2.3.2 Organizing files and directories 19

2.3.3 Creating your directory structure 21

A directory structure for a small project 21

A directory structure for a large, one-person project 23

Directories for collaborative projects 23

Special-purpose directories 25

Remembering what directories contain 27

Planning your directory structure 29

Naming files 30

Batch files 30

2.3.4 Amoving into a new directory structure (advanced topic) . . 31

Example of moving into a new directory structure 31

2.4 Documentation 34

2.4.1 What should'you document? 36

2.4.2 Levels of documentation 37

2.4.3 Suggestions for writing documentatidn 38

Evaluating your documentation 39

2.4.4 The research log 39

A sample page from a research log 40

A template for research logs 42

2.4.5 Codebooks . . . . 43.

A codebook based on the survey instrument 43

2.4.6 Dataset documentation 44

2.5 Conclusions 45

3 Writing and debugging do-files , 47

3.1 Three ways to execute commands 47

3.1.1 The Command window 48

3.1.2 Dialog boxes 49

3.1.3 Do-files 49

3.2 Writing effective do-files 50

3.2.1 Making do-files robust 51

Make do-files self-contained 51

Use version control 53

Exclude directory information 53

Include seeds for random numbers 55

3.2.2 Making do-files legible 55

Use lots of comments 56

Use alignment and indentation 57

Use short lines 58

Limit your abbreviations 61

Be consistent 63

3.2.3 Templates for do-files 63

Commands that belong in every, do-file 63

A template for simple do-files 66

A more complex do-file template 66

3.3 Debugging do-files 68

3.3.1 Simple errors and how to fix them 68

Log file is open 68

Log file .already exists 68

Incorrect'command name 69

Incorrect variable name 69

Incorrect option 70

Missing comma before options 70

3.3.2 Steps for resolving errors 70

Step 1: Update Stata and user-written programs 70

Step 2: Start with a clean slate 71

Step 3: Try other data o

Step 4: Assume everything could be wrong 72

Step 5: Run the program in steps 72

Step 6: Exclude parts of the do-file 74

Step 7: Starting over 74

Step '8: Sometimes it is not your mistake 75

3.3.3 Example 1: Debugging a subtle syntax error 75

3.3.4 Example 2: Debugging unanticipated results 77

3.3.5 Advanced methods for debugging 81

3.4 How to get help 82

3.5 Conclusions 82

4 Automating your work 83

4.1 Macros 83

4.1.1 Local and global macros 84

Local macros 84

• Global macro's 85

Using double quotes when defining macros 85

Creating long strings 85

4.1.2 Specifying groups of variables and nested models 86

4.1.3 Setting options'with locals 88

4.2 Information, returned by Stata commands 90

Using returned results witlr local macros 92

4.3 Loops: foreach and forvalues 92

The foreach command 94

The forvalues .command 95

4.3.1 Ways to use loops 95
Loop example 1: Listing variable and value labels 96
Loop example 2: Creating interaction variables 97
Loop example 3: Fitting models with alternative measures of education 98
Loop example 4: Recoding multiple variables the same way 98
Loop example 5: Creating a macro that holds accumulated information 99
Loop example 6: Retrieving information returned by Stata . 100
4.3.2 Counters in loops 101
Using loops to save results to a matrix 102
4.3.3 Nested loops 104
4.3.4 Debugging loops 105
4.4 The include command 106
4.4.1 Specifying the analysis sample with an include file 107
4.4.2 • Recoding data using include files 107
4.4.3 Caution when using include files 109
• 4.5 Ado-files 110
4.5.1 A simple program to change directories I1l
4.5.2 Loading and deleting ado-files 112
4.5.3 Listing variable names and labels 113
4.5.4 A general program to change your working directory 117
4.5.5 Words of caution 118
4.6 Help files 119
4.6.1 nmlabel.hlp 119
4.6.2 help me 122
4.7 Conclusions 123
5 Names, notes, and labels 125
5.1 Posting files 125
5.2 The dual workflow of data management and statistical analysis . . . 127
5.3 • Names, notes, and labels 129

5.4 Naming do-files . . . 129

5.4.1 Naming do-files to re-create datasets . . . 130

5.4.2 Naming do-files to reproduce statistical analysis . . . 130

5.4.3 Using master do-files . . . 131

Master: log files . . . 133

5.4.4 A template for naming do-files . . . 134

Using subdirectories for complex analyses . . . 135

5.5 Naming and internally documenting datasets . . . 136

Never name it final! . . . 137

5.5.1 One time only and temporary datasets . . . 137

5.5.2 Datasets for larger projects . . . 138

5.5.3 Labels and notes for datasets . . . 138

5.5.4 The datasignature command . . . 139

A workflow using the datasignature command . . . 140

Changes datasignature does no,t detect . . . 141

5.6 Naming variables . . . 143

5.6.1 The fundamental principle for creating and naming variables 143

5.6.2 Systems for naming variables . . . 144

Sequential naming systems . . . 145

Source naming systems . . . 145

Mnemonic naming systems . . . 146

5.6.3 Planning names . . . 146

5.6.4 Principles for selecting names . . . 147

Anticipate, looking for variables . . . 147

Use simple, unambiguous names . . . 148

Try names before you decide . . . 151

5.7 Labeling variables . . . 151

5.7.1 Listing variable labels and other information . . . 151

Changing the order of variables, in your, dataset . . . 155

5.7.2 Syntax for label variable . . . 155

5.7.3 Principles for variable labels 156

Beware of truncation 156

Test labels before you-post the file 157

5.7.4 Temporarily changing variable labels 157

5.7.5 Creating variable labels that include the variable name 158

5.8 Adding notes to variables 160

5.8.1 Commands for working with notes 161

Listing notes 161

Removing notes 162

Searching notes 162

5.8.2 Using macros and loops with notes 162

5.9 Value labels 163

5.9.1 Creating value labels is a two-step process 164

Step 1: Defining labels 164

Step 2: Assigning labels 164

Why a two-step system? 164

Removing labels 165

5.9.2 Principles for constructing value labels 165

1) Keep labels short 165

2) Include the category number 166

3) Avoid special characters 168

4) Keeping track of where labels are used 169

5.9.3 Cleaning value labels 170

5.9.4 Consistent value labels for missing values 171

5.9.5 Using loops when assigning value labels 171

5.10 Using multiple languages 173

5.10.1 Using label language for different written languages 174

5.10.2 Using label language for short and long labels 174

5.11 A. workflow for names and labels 176

Step 1: Plan the changes 176

Step 2: Archive, clone, and rename 177

Step 3: Revise variable labels 177

Step 4: Revise value, labels 177

Step 5: Verify the changes 178

5.11.1 Step 1-: Check the source data 178

Step la: List the current names and labels 178

Step 1b: Try the current names and. labels 181

5.11.2 Step 2: Create clones and rename variables 182

Step 2a: Create clones 183

Step 2b: Create rename commands 183

Step 2c: Rename variables 184

5.11.3 Step 3: Revise variable labels 185

Step 3a: Create, variables-label commands 185

Step 3b: Revise variable labels 186

5.11.4 Step 4: Revise value labels 187

Step 4a: List the current .labels 188

Step 4b: Create label define commands to edit 189

Step 4c: Revise labels and add them to dataset 193

5.11.5 Step 5:; Check the new names and labels 194

5.12 Conclusions 195

6 *f* Cleaning your data 197

j 6.1 Importing data 198

6.1.1 Data formats 198

ASCII data formats 198

Binary-data .formats 200

6.1.2 Ways to import data 20.1

Stata commands to import data 201

Using other statistical packages to export data 203

Using a data conversion program 20.3

6.1.3 Verifying data conversion 203

Converting the ISSP 2002 data from Russia . . .- 204

6.2 Verifying variables • 210

6.2.1 Values review 211

Values review of data about the scientific career 212

Values review of data on family values 215

6.2.2 Substantive review 216

.What does time to degree measure? . 216

Examining high-frequency values 218

Links among variables . • 220

Changes in survey questions 225

6.2.3 Missing-data review . . . . : . 225

Comparisons and missing values 225

Creating indicators of whether cases are missing 228

Using extended missing values • 228

Verifying and expanding missing-data codes 229

Using include files : 236

6.2.4 Internal consistency review 238

Consistency in data on the scientific career 238

6.2.5 ' Principles for fixing data inconsistencies 241

6.3 Creating variables for analysis • 241

/ 6.3.1 Principles for[1] creating new variables 242

New variables get new names 242

Verify that new variables are correct 243

Document new variables 244

Keep the source variables 244

6.3.2 Core"commands for creating variables 244

The generate command 245

The clonevar command 245

The replace command 246

6.3.3 Creating variables with missing values 247

6.3.4 Additional commands for creating Variables 249

The recode command 249

The egen command 250

The tabulate, generateQ command 252

6.3.5 Labeling variables created by .Stata 253

6.3.6 Verifying that variables are correct 254

Checking the code 255

Listing variables 255

Plotting continuous variables 256

Tabulating variables 258

Constructing variables multiple ways 259

6.4 Saving datasets 260

6.4.1 Selecting, observations 261

Deleting cases versus creating selection variables 261

6.4.2 Dropping variables 262

Selecting variables for the ISSP 2002 Russian data 263

6.4.3 Ordering variables 263

6.4.4 Internal documentation 264

6.4.5 Compressing, variables 264

6.4.6 Running diagnostics 265

The codebpok, problems command 265

Checking for unique ID. variables 267

6.4.7 Adding a data signature 269

6.4.8 Saving the file 270

6.4.9 After a file is saved 271

6.5 Extended example of preparing data for analysis 271

Creating control variables 271

Creating binary indicators-of positive attitudes 274

Creating four.-category scales of positive, attitudes 277

6.6 Merging files 279

6.6.1 Match-merging 280

Sorting the ID variable 281

6.6.2 One-to-one merging 281

Combining unrelated datasets 281

6.6.3 Forgetting to match-merge 283

6.7 Conclusions 285

7 Analyzing data and presenting results 287

7.1 Planning and organizing statistical analysis 287

7.1.1 Planning in the large 288

7.1.2 Planning in the middle . . . . • . . . . - 289

7.1.3 Planning in the small 291

7.2 Organizing do-files 291

7.2.1 Using master do-files 292

7.2.2 What belongs in your do-file? 294

7.3 Documentation for statistical analysis 295

7.3.1 The research log and comments in do-files 295

7.3.2 Documenting the provenance of results 296

Captions on graphs 298

7.4 Analyzing data using automation 298

7.4.1 Locals to define sets of variables 299

7.4.2 Loops for repeated analyses . . . . '. 300

Computing t tests using loops 300

Loops for alternative model specifications . . . . ' . . 302

7.4.3 Matrices to collect and print results 303

Collecting results of t tests 303

Saving results from nested regressions 306

Saving results from different transformations of articles . . . 308

7.4.'4 Creating a graph from a matrix 310

7.4.5 Include files to load data and select your sample 311

7.5 Baseline statistics 312

7.6 Replication 313

7.6.1 Lost or forgotten files 313

7.6.2 Software and version control 314

7.6.3 Unknown seed for random numbers 314

Bootstrap standard errors 314

Letting Stata. set the seed 315

Training and confirmation samples 316

7.6.4 Using a global that is not in your do-file 318

7.7 Presenting results 318

7.7.1 Creating tables 319

Using spreadsheets 319

Regression tables with esttab 321

7.7.2 Creating graphs 323

Colors, black, and white 324

Font size 326

7.7.3 Tips for papers and presentations 326

Papers 326

Presentations 327

7.8 A project checklist 328

7.9 Conclusions 328

8 Protecting your files 331

8.1 Levels of protection and types of files 332

8.2 Causes of data loss and issues in recovering a file 334

8.3 Murphy's law and rules for copying files 337

8.4 A workflow for file protection 338

Part 1: Mirroring active storage 338

Part 2: Offline backups 340

8.5 Archival preservation 343

8.6 Conclusions 345

9 Conclusions 347

A How Stata works 349

A.I How Stata works . . . 349

Stata directories . . . 350

The working directory . . . 350

A.2 Working on a network . . . 351

A.3 Customizing Stata . . . 353

A.3.1 Fonts and window locations . . . 353

A.3.2 Commands to change preferences . . . 353

Options that can be set permanently . . . 353

Options that need to be set each session . . . 355

A.3.3 profile.do . . . 355

Function keys . . . 356

A.4 Additional resources . . . 356

References 359

Author index 363

Subject index 365