The Stata Journal (2023)
23, Number 4, pp. 1082–1085 DOI: 10.1177/1536867X231212452

An Introduction to Stata for Health Researchers review

Tim Collier
Department of Medical Statistics
London School of Hygiene and Tropical Medicine
London, U.K.
tim.collier@lshtm.ac.uk

Abstract. In this article, I review *An Introduction to Stata for Health Researchers*, by Svend Juul and Morten Frydenberg (2021, Stata Press).

Keywords: gn0099, book review, introduction to Stata, data management, statistical analysis, health research

1 Introduction

This is the fifth edition of Juul and Frydenberg's *Introduction to Stata for Health Researchers*, which has been updated to reflect changes made in Stata versions 14 to 17. The first and fourth editions were reviewed in the *Stata Journal* by Carlin (2006) and Linden (2014), respectively.

The aim of this book, as stated in the preface to the first edition, is to empower health researchers who have little or no experience with Stata to benefit from using Stata in their own research. Although the book is introductory, it covers a wide range of topics and is full of hints and tips, such that even an experienced Stata user may learn something new.

The material is explained with worked examples that the reader can follow, using datasets that can be downloaded from the book's website (https://www.stata-press.com/books/introduction-stata-health-researchers/) or the example datasets that are installed with Stata.

The authors' intention is that the reader should dip into the book as required rather than reading from start to finish. However, there is a sensible ordering of the chapters that roughly follows the flow of a research project, from loading data through data management and analysis to reporting results in tables and figures. There is a healthy emphasis on good practice for data management and analysis, including advice on documentation, data protection, auditing, checking for errors, etc. The book includes a major overhaul of the chapter on taking good care of your data, which the authors think is an important topic (and I fully agree).

2 Content

The book comprises five main parts: Part I, “The basics” (three chapters); Part II, “Data management” (six chapters); Part III, “Analysis” (six chapters); Part IV, “Graphs” (one long chapter); and Part V, “Advanced topics” (one chapter).

Part I, “The basics”, consists of three fairly short chapters aimed at getting the reader up and running with Stata. These chapters cover things such as installing Stata, the Stata interface, ways of getting help, the standard Stata syntax, deciphering of error messages, and more.

I smiled at the opening instruction of chapter 1 on installing Stata, which was “Follow the instructions provided at the time of purchase or see the *Installation Guide* [IG].” This did make me wonder how much of the first few chapters could be replaced by similarly pointing to the *Getting Started* [GS] manual. Even with the caveat that the book is not intended to be read page by page, I was surprised by how quickly things became quite technical. For example, the `sysdir` and `adopath` commands are introduced in section 1.1 without explanation of where or how to submit these commands.

Part II, “Data management”, consists of six chapters focused on issues related to data management. Chapter 4, “Variables”, covers issues dealing with string and numeric variables, date and time variables, and missing values. It also touches on the prickly problem of precision. Then there are two fairly short chapters, one on getting data in and out of Stata and one on labeling. Chapter 7 (somewhat strangely to me called “Calculations”) covers the usual suspects for creating derived variables, for example, `generate`, `replace`, `egen`, and `recode`. It includes good advice on checking newly derived variables for correctness and completeness. Chapter 8 covers commands for changing data structure, including dropping and sorting variables, random sampling, combining datasets, reshaping, and collapsing. Chapter 9, “Taking good care of your data”, emphasizes the importance of good practice for data management to produce accurate and reproducible results. The authors address matters such as organization and storage of files, naming conventions for files and variables, documentation, labeling, data cleaning, and data protection.

Part III, “Analysis”, consists of six chapters focused on statistical methods of analysis. Chapter 10 covers descriptive statistics, basic tables, and simple hypothesis tests for continuous and categorical data. I was pleased to see an example of the use of immediate tables (`tabi`), which can be very helpful when reviewing published results, as well as some of Stata’s tables for epidemiologists. Chapter 11, “Regression analysis”, covers linear (ordinary least-squares) and logistic regression. The focus in this chapter is not on the theory of regression modeling (the authors point to some standard textbooks on regression) but on the technical aspects of how to fit these models in Stata, such as how to include categorical predictor variables and interactions. There is quite a bit on postestimation commands, and the (very useful) `lincom` command is explained in detail, but it is perhaps a pity that there is just one bullet point referring to `margins` and `marginsplot`. Use of robust standard errors and bootstrapping appears at the end of the chapter. Chapter 12, “Time-to-event data”, covers setting up data for survival-time analysis and basic descriptive statistics before going on to Cox proportional hazards and Poisson regression models. The section on Cox proportional hazards models deals with more complex analyses, including time-varying covariates and time-varying coefficients. As with chapter 11, the focus is on how to perform such analyses in Stata rather than on detail of the statistical methods. Chapter 13 is a fairly short chapter on power, precision, and sample size. Some examples are shown of Stata's useful suite of `power` commands followed by an example of how to write a program to address a nonstandard sample-size and power question using simulation (unfortunately, there's a missing returned scalar in the example do-file). Given the importance of power and sample-size calculations in health research, I think this section could be expanded; the examples using the `power` commands are all fairly simple, and a few more complex examples might be useful. Chapter 14 covers methods for assessing agreement, reproducibility, and diagnostic tests. Chapter 15 is titled "Miscellaneous" and covers a fairly random set of topics, including generation of random numbers, random sampling, the suite of commands for the International Classification of Diseases, exporting tables and the `collect` command, and sending graphs to Word or PDF files. Personally, I think the `collect` command is so complex that it does not work well in a broad introductory book like this. Thankfully, the section called "Table 1" can call on `dtable` (a much kinder command) for the sixth edition!

Part IV, "Graphs", consists of one long chapter dedicated to Stata graphs. The authors first explain the anatomy of a Stata graph and the structure of a graph command. This is followed by guidance on general graph options, including schemes and all things related to axes, titles, markers, and lines. Examples are then given of specific graph types, including histograms, other two-way graphs, and bar graphs. The chapter concludes with saving and exporting graphs to different formats. It touches on the issue of journal requirements, but I think the novice (and more experienced) user would find more specific instructions on how to meet size and resolution requirements helpful.

Do-files for each of the graphs in this chapter (and all figures in the book) are included among the online material for the book. One small complaint is that most of the graphs use a scheme called `lean1`, which was created by Svend Juul, the first author. It is not until later in the chapter, in section 16.4, that we are told about schemes and shown how to install new schemes; therefore, any novice trying to reproduce the early graphs will encounter an error message—which might be off-putting.

Finally, part V, "Advanced topics", covers several more advanced topics, including accessing of stored results, macros and scalars, and loops. Some of these are then used in the context of writing a small Stata program, including use of the `syntax` command. The book concludes with a short section on debugging programs.

3 Conclusion

Overall, I am very positive about this book. It is generally easy to read and is full of helpful worked examples that can be followed using the accompanying datasets. I like the emphasis on good practice for data management and statistical analysis, and without mentioning the term explicitly (as far as I could see), the authors cover the vital topic of workflow for data analysis. This is a book on how to do things in Stata rather than how to learn statistical methods, but the reader is pointed throughout to other relevant textbooks.

One of the strengths of the book is the breadth of topics covered. I have been using Stata for 23 years yet still learned some new or alternative ways of doing things (so you can teach an old dog new tricks). However, this also means that most topics are covered with a fairly light touch, and I was left wanting more in some places.

There are some things I would change. I think some of the more technical aspects of chapter 1 (which could easily confuse a novice user) could be removed without any detriment. I found it slightly strange that the section on including graphs in Word and PDF files was included in part III, "Analysis". I think this might be better placed alongside the chapter on graphs in a section on presenting or reporting results, which could also include tables.

However, overall I think this is a very useful book that will help empower health researchers to benefit from using Stata in their research.

4 References

Carlin, J. 2006. Review of An Introduction to Stata for Health Researchers by Juul. *Stata Journal* 6: 580–583. https://doi.org/10.1177/1536867X0600600409.

Juul, S. 2021. *An Introduction to Stata for Health Researchers.* 5th ed. College Station, TX: Stata Press.

Linden, A. 2014. Review of An Introduction to Stata for Health Researchers, Fourth Edition, by Juul and Frydenberg. *Stata Journal* 14: 697–700. https://doi.org/10.1177/1536867X1401400314.

About the author

Tim Collier is a medical statistician with more than 20 years' experience of medical research using Stata. His research focuses mostly on cardiovascular clinical trials, data monitoring, and patient safety in randomized trials. Recently, he has been involved in developing and popularizing the win-ratio method of analysis for composite endpoints in clinical trials. He has developed and delivered Stata training courses in the U.K. and around the world and teaches on the master's in medical statistics at the London School of Hygiene and Tropical Medicine.