The Stata Journal (2014)
14, Number 3, pp. 697–700

Review of An Introduction to Stata for Health Researchers, Fourth Edition, by Juul and Frydenberg

Ariel Linden
Linden Consulting Group, LLC
Ann Arbor, MI
alinden@lindenconsulting.org

Abstract. In this article, I review *An Introduction to Stata for Health Researchers, Fourth Edition*, by Svend Juul and Morten Frydenberg (2014 [Stata Press]).

Keywords: gn0061, introduction to Stata, data management, statistical analysis, health research

1 Introduction

For instructors of measurement and evaluation and individuals seeking methodological guidance, it is difficult to find a book that both covers key analytic concepts and provides clear direction on how to perform the associated analyses in a given statistical software package. The fourth edition of *An Introduction to Stata for Health Researchers*, by Svend Juul and Morten Frydenberg, fills this need. It does an excellent job of covering a wide range of measurement and evaluation topics while providing a gentle introduction to Stata for those unfamiliar with the software. In fact, though the title suggests the book is for health researchers, it is readily generalizable to many disciplines that implement the same methods.

Many improvements have been made to the book since John Carlin's review of the inaugural edition in 2006 (Carlin 2006), including a reorganization of chapters to more closely mirror the typical flow of a research project, an increase in the number of practice exercises, and a more focused treatment of statistical issues. Additionally, this fourth edition has been updated for Stata 13. On the whole, Juul and Frydenberg have prepared a very accessible book for readers with varied levels of proficiency in statistics or Stata, or both.

2 Overview

Section I includes four chapters (called "the basics") that introduce the reader to Stata. These chapters cover such issues as installing the program, getting help, understanding file types, and using command syntax. While a novice could go directly to the Stata user's manual (in particular, *Getting Started with Stata* and the *Stata User's Guide*), this book offers a more user-friendly introduction. Combined, these 35 pages are more than sufficient to get a Stata novice up and running.

Section II includes six chapters dealing with issues pertaining to data management, such as variable types (numeric, dates and strings) and their manipulation and storage (chapter 5); importing and exporting data (chapter 6); applying labels (chapter 7); generating and replacing values and performing basic calculations (chapter 8); and changing data structure, such as appending, merging, reshaping, and collapsing data (chapter 9). Chapter 10 provides excellent advice on creating documentation (via do-files and logs, etc.) to ensure reproducibility of data management and analytic steps. While creating documentation is seemingly intuitive, not all researchers consistently follow these steps.

Section III includes five chapters focusing on the types of data analyses most widely used in health-related research.

Chapter 11 starts with basic descriptive analytics and then continues on to analyses using epidemiologic tables for binary variables (including the addition of stratified variables). This naturally progresses to analyses of continuous variables, and the chapter demonstrates some visual displays of the data (histograms, Q–Q plots, and kernel density plots) and methods of tabulation. The chapter then ventures into more formal basic statistical analyses, such as t tests, one-way analysis of variance, and nonparametric techniques (`ranksum`).

Chapter 12 presents ordinary least-squares and logistic regression, with a fair amount of exposition on the use of `lincom` for postestimation.

Chapter 13 describes time-to-event analyses, starting with simple curves and tables, and then moves into progressively more complex Cox regression models (without and with time-varying covariates). Next it introduces Poisson models to examine more complex models for rates. Finally, it includes a brief discussion on indirect and direct standardization.

Chapter 14 is titled "Measurement and diagnosis", and it describes graphical plots and statistical tests for assessing measurement variation at one time point, and then again over multiple measurements, for dependent samples. This transitions into methods used for assessing accuracy of diagnostic tests (that is, sensitivity, specificity, area under the curve, etc.).

Chapter 15—"Miscellaneous"—includes topics such as random sampling, sample-size calculations (including a nice example using simulation to estimate power for a noninferiority study), error trapping, and log files.

Section IV includes one comprehensive chapter on graphs (44 pages). The chapter begins by plotting a basic graph and describing the various elements, and it progresses with increasing sophistication. It ends with some important tips on saving the code in do-files so that graphs can be reproduced or enhanced later.

The final section, section V, is composed of a single chapter titled "Advanced topics" and discusses storing and using results after estimation and defining macros and scalars. It then discusses looping through data using `foreach`, `forvalues`, and if/then statements. The chapter ends with a brief overview of creating user-written commands.

3 Comments

The book is well organized, following the logical step-by-step approach that investigators apply to their research: data acquisition and management, analysis, and presentation of results. The many brief examples are useful and generalizable, and the footnotes are helpful additions. When a topic is briefly touched upon, the authors refer the reader to the relevant help resource in Stata for more details. They also provide helpful recommendations for resolving issues that may have multiple solutions.

Another strength of the book is that it contains many important but often overlooked details (even for advanced Stata users), such as why a value may appear differently when formatted as float versus double (pages 45–46) and how this precision may impact comparisons. Other examples include the use of `numlabel` to display both the value and the value label of a variable (page 67), the use of `egen cut()` to easily recode continuous variables into categories (page 75), and setting `showbaselevels` to display a line for the reference level in regression output (page 153). Of arguably greatest value is the fact that the authors continually emphasize the importance of developing good habits in documenting the work process (using do-files and logs) so that all output can be replicated, errors can be tracked down, and time-consuming procedures can be performed repeatedly and efficiently.

There is very little that I would change about this book, and my suggestions all relate to what the authors could consider for future editions. First, the authors use `lincom` and `testparm` extensively in the chapters on regression and time-to-event analyses. Readers would benefit from seeing examples using `margins` (followed by `marginsplot`). `margins` is an extremely flexible command that allows the user to perform various analyses after running regression models, mostly with little additional specification. The authors currently provide only a footnote (page 150) pointing interested readers to the excellent book written by Michael N. Mitchell (2012). Second, some mention of parametric regression models for survival analysis would be valuable (using `streg`), because readers in certain disciplines may prefer these models over Cox regression models (using `stcox`).

Finally, while Stata 13 introduced a new set of commands to estimate treatment effects using propensity score-based matching and weighting techniques, the only mention of such approaches is in appendix A, where the authors briefly describe the *Stata Treatment-Effects Reference Manual* by saying this: "Despite its title, it does not correspond to the methods of analysis that are mainstream in health research". This statement left me somewhat perplexed, given that graduate programs in public health in the United States have a required course in program evaluation that likely covers these methods in at least some detail. Furthermore, there is a growing body of health research literature where using these methods has become commonplace (see, for example, Austin [2007; 2008]). Readers would benefit from an introduction to these techniques, perhaps as a final chapter in which some of the datasets analyzed in previous chapters using regression are reanalyzed using one of these approaches and the results compared. The *Stata Treatment-Effects Reference Manual* offers an excellent

introduction to the methods implemented in Stata, and Stuart (2010) provides a more comprehensive discussion of treatment-effects estimation using an array of approaches.

In summary, I strongly recommend this book both for students in introductory measurement and evaluation courses and for more seasoned health researchers who would like to avoid a steep learning curve when trying to conduct analyses in Stata.

4 References

Austin, P. C. 2007. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *Journal of Thoracic and Cardiovascular Surgery* 134: 1128–1135.

———. 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 27: 2037–2049.

Carlin, J. 2006. Review of An Introduction to Stata for Health Researchers by Juul. *Stata Journal* 6: 580–583.

Juul, S., and M. Frydenberg. 2014. *An Introduction to Stata for Health Researchers*. 4th ed. College Station, TX: Stata Press.

Mitchell, M. N. 2012. *Interpreting and Visualizing Regression Models Using Stata*. College Station, TX: Stata Press.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25: 1–21.

About the author

Ariel Linden is a health services researcher specializing in the evaluation of health care interventions. He is both an independent consultant and an adjunct associate professor at the University of Michigan in the department of Health Management and Policy, where he teaches program evaluation.