The Stata Journal (2006)
**6**, Number 4, pp. 580–583

# Review of An Introduction to Stata for Health Researchers by Juul

John Carlin
Murdoch Childrens Research Institute and
School of Population Health, University of Melbourne
Melbourne, Australia
jbcarlin@unimelb.edu.au

**Abstract.** This article reviews *An Introduction to Stata for Health Researchers* by Svend Juul.

**Keywords:** gn0034, learning Stata, introductory biostatistics, epidemiological statistics, health sciences

## 1 Introduction

Learning how to do effective data analysis is not easy, although well-crafted computer packages such as Stata now bring an extraordinary range of possibilities to an ever-growing number of would-be analysts across many disciplines. This book aims to provide a comprehensive introduction to Stata for health researchers. Svend Juul teaches epidemiology to medical students and has extensive experience in teaching Stata and other computer programs to Ph.D. students in the health sciences. The book is ambitious in attempting not only to provide an introduction but also to cover a wide range of advanced statistical methods. In this regard, it adopts an interesting blend of the style used in the Stata manuals (with content paralleling parts of the *Getting Started* manual, *User's Guide*, and various *Reference* manual entries, including commands such as those used for regression, logistic, and survival-time analysis) and that of more statistically oriented books such as Rabe-Hesketh and Everitt (2004). I think the book is remarkably successful, given such ambitions.

## 2 Summary

The first 100 or so pages of the book provide a comprehensive but compact overview of Stata's essential features and its capabilities for managing and manipulating data. Chapter 1 provides a summary of Stata as used in the Windows environment. As is typical throughout the book, readers do not need to go far to find handy hints, and even experienced users may benefit. For example, chapter 2 covers how to find help and how to best develop one's own capacity to solve problems and to find the specific solutions that might be needed in a particular area. A useful hint found here is that the menus and dialogs can be an effective way to explore the availability of particular methods within the program. These early chapters are generally brief. They also cover

 

Stata file types and names (chapter 3), command syntax (chapter 4), variable types and storage (chapter 5), getting data in and out of Stata (chapter 6), documentation commands (chapter 7), calculations for recoding and transforming variables (chapter 8), and commands for manipulating data structures (chapter 9).

A major feature of the book and of these early chapters in particular is an emphasis on good habits in data analysis. For example, chapter 4 points out that although variable names may be made long, using short but meaningful names is preferable for many reasons (although principally to improve display output). From the beginning of chapter 6, the author emphasizes the dangers of haphazard alteration of datasets and recommends adopting reproducible analysis strategies from the early stages of data cleaning and processing.

In chapter 10, the book progresses from discussion of how to use the program and manage data into a treatment of analytic methods, described with working examples. First, the reader is led through the small number of essential commands that provide the staple of practical data analysis (`describe`, `codebook`, `summarize`, `list`, `browse`). The chapter then covers methods for summarizing categorical variables by using the `tabulate` commands and its variants, as well as for analyzing continuous variables. For users like me whose original Stata education may have been some time ago, there are again some invaluable features revealed in this chapter, such as the `numlabel` command, which enables displaying values along with their labels in table output. The author also draws attention to a useful user-written command, `groups` (downloadable with `ssc`). Not only is this command likely to be valuable in itself for many readers, but in drawing attention to it the author highlights a major feature of Stata: its extensibility and the frequent online publication of new commands by users around the world.

The longest chapter in the book, chapter 11 focuses on graphics. It provides an excellent overview of Stata's now extensive and flexible graphical capabilities. This survey begins with a succinct overview of the architecture of a Stata graph, explaining for example the difference between the plot region and the graph region and showing how simple adjustments can be made to the size and shape of graphs. Many useful tips are crammed into the subsequent exposition of graph details; for example, Juul shows how to provide axis labels and ticks for a log scale and reveals several other useful techniques for axis labeling. He clearly appreciates the importance of good graphical style and displays an impressive attention to the details that can distinguish a really successful graph from one that does not fulfill its potential.

Chapters 12–16 provide a tour (in about 90 pages) of the most commonly used statistical techniques in health research, beginning with stratified analysis of epidemiological studies and moving through regression models (chapter 13), analysis of time-to-event data (chapter 14), techniques for assessing measurement methods and diagnostic tests (chapter 15), and finally a brief treatment of miscellaneous other topics in chapter 16. Although condensing so much statistics into such a short space cannot be done without limiting the background and issues covered, the treatment is still generally hard to fault and will provide useful starting points for many health researchers. In the final chapters, the author briefly examines some of the basics of Stata programming (chapter 17)

and then returns to give more general advice about best practice in data management (chapter 18).

## 3 Assessment

Every reviewer finds a few faults and shortcomings. Although in fact remarkably successful, Juul's attempt to provide such a comprehensive tour of Stata usage does bring some dangers. For example, there is clearly a risk of overloading the reader with details that cannot all be assimilated. In this regard, the reader is well advised to follow the advice given in the preface: treat some of the material as reference matter that can be skipped at first reading and returned to after some experience has been gained. I found the order of presentation of some of the material slightly problematic; for example, I am not sure that providing a detailed introduction of the command syntax (in chapter 4) before discussing datasets and variables (chapters 5 and 6) is the best approach. And sometimes concepts are mentioned before being explained, as with the reference to a do-file at the bottom of page 9. Another example is the discussion of errors and error messages in section 2.4, before command syntax has been introduced.

I noticed only a few issues with the substantive editing of the book. Finding exercises for the reader in chapter 1 but not in later chapters was odd. In a few places, material has been repeated, as with warnings about missing values in logical expressions (p. 37, 46) and an illustration of marker symbols (p. 134, 147). The material on "decimal commas" will be redundant (and even mysterious) for most readers in English-speaking countries, but then those of us outside North America have long had to contend with various Americanisms in the Stata manuals!

More seriously, I found that the author just occasionally strayed from what I would consider an ideal treatment of statistical issues. For example, he tended toward a mechanical approach to hypothesis testing and $p$-values, in which the widespread scourge of "significance" raises its head. (For example, I believe that the discussion of the interaction term on p. 196 appears to erroneously equate $P < 0.05$ with the existence of effect modification at one level of a three-level factor.) When the book shows how statistically based conclusions are drawn, it would be good to see wider adoption of the advice of Kirkwood and Sterne (2003)—among others—to abandon the use of arbitrary significance levels. Another minor concern is that in introducing logistic regression models in section 13.2, the discussion is somewhat confused between the outcome measure (which is the binary indicator of "success" or "failure") and the quantity being modeled, which is a transformed value (the log odds) of the probability or risk parameter. To clarify the analogy with normal linear regression: the outcome in the former is the continuous measure of interest, whereas the quantity being modeled is its mean value (Clayton and Hills 1993).

Finally, although the advice and explanation given throughout this book will empower many readers to do great things with their data, I would have made slightly more prominent the suggestion in the last line of chapter 13: "It might be a good idea to seek expert advice before you begin designing a study." In my experience, most begin-

ning researchers have plenty of scope for benefiting from professional help. However, I heartily recommend this book. It is well written, with an economical and clear style, and will provide an excellent resource for a wide range of Stata users (not just novices) in the health sciences and potentially beyond.

## 4 References

Clayton, D., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Juul, S. 2006. *An Introduction to Stata for Health Researchers*. College Station, TX: Stata Press.

Kirkwood, B. R., and J. A. C. Sterne. 2003. *Essential Medical Statistics*. 2nd ed. Oxford: Blackwell Science.

Rabe-Hesketh, S., and B. Everitt. 2004. *A Handbook of Statistical Analyses Using Stata*. 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.

#### About the author

John Carlin is a statistician by training with more than 15 years' experience in medical and public-health research. He contributes to the design and analysis of a wide range of projects, including work relating to cystic fibrosis, childhood obesity, and early-life determinants of health. His statistical research interests currently focus on methods for handling missing-data problems with multiple imputation. He is a coauthor of a well-known graduate statistics text on Bayesian data analysis.