The Stata Journal (2003)
3, Number 1, pp. 100–104

Review of A Short Introduction to Stata For Biostatistics by Hills and De Stavola

John McGready
Johns Hopkins University
jmcgread@jhsph.edu

Abstract. The new book by Hills and De Stavola (2002) is reviewed.

Keywords: gn0005, biostatistics

1 Introduction

A Short Introduction to Stata for Biostatistics is a welcome addition to the growing body of literature that teaches statistical computing via an example and context-driven approach. While a reader needs to have both some statistical background and some computer literacy to use this book, the text can simultaneously assist in solidifying the ability to interpret the results of various statistical analyses while teaching how to produce these results. Stata provides some of the best documentation I have seen for statistical software, but this documentation is not very helpful for getting a user up and running, and a book of this nature can save a lot of time and spare a lot of angst. The authors state that the book is "written with biostatisticians in mind". While not denying the utility of a book like this to biostatistical professionals, I would argue that the resulting book is more appropriate for those with less grounding in statistical sciences and computing. I will follow-up this claim with specifics later in this review.

The text is written in an interactive fashion, in a style that I dub "call and response". By this, I mean that the book should be read in front of the computer. The book comes with a CD-ROM of datasets providing much fodder for examples. Each description or explanation is reinforced by a set of corresponding command statements, which can be performed by the reader. Hence, the entire text from beginning to end is an interactive journey between text-based explanations (the call) and user participation with Stata (the response). This is an excellent way to present the material, as it provides the user with instant gratification in regard to each concept and corresponding set of commands. It also familiarizes the reader with how Stata "thinks", operates, and presents results.

The book begins with a short introduction to Stata, a description of the on-screen layout, and the Stata-specific menu bars. This section also provides the reader with instructions on how to upload the datasets for the book, which are provided in an accompanying CD-ROM. While this introductory section is very useful, it is rather bare-bones basic. One of the most frequent difficulties I witness when teaching with Stata is the lack of students' understanding about the concept of a working directory in Stata. If this is not cleared up at the beginning of the learning process, it inevitably causes many delays down the road: datasets "can't be found", log files are "lost" on the hard drive, and so on. While it could be argued that directory structures and the corresponding working directories concept should be covered in literature specific to each operating system, a short treatment of this at the beginning of the text would be a very welcome addition. The book does ultimately detail the organization of Stata program-specific directories in the final chapter, Chapter 15, but it still does not explore this working directory concept. This shortcoming is not specific to this text. Another very well-written book on Stata, Lawrence Hamilton's *Statistics with Stata 7* (2002), gives this no mention either.

2 Data management

The data management chapters sandwich the rest of the material discussed in the book. Chapters 1–3 give a very nice, concise introduction to the most commonly needed data management techniques and should be required reading for anyone planning to do data analysis with Stata. More advanced topics related to data management are sprinkled throughout Chapter 5 and covered in Chapters 12–15, the final three chapters in the book.

Chapter 1 on 'Some Basic Commands' gets things started, illustrating how to open a Stata dataset and get a feel for what it contains. The `describe`, `codebook`, `list`, `summarize`, `format`, `tabulate`, `sort`, and `generate` commands are explored in detail, as well as command-specific options. Also detailed in this chapter are how to use Stata as a calculator, Stata command syntax diagrams, and Stata's help utilities.

Chapter 2 on 'Housekeeping' answers the question "Now that I've got my dataset up and running on Stata, and know what it contains, how do I make it work for me?". The chapter details the basics of labeling a dataset, variables and values of a variable, recoding variables, how to deal with dates, and missing values in Stata. Also introduced in this Chapter are log files and do-files. While the chapter is brief, it is chock-full of exercises and highlights two very important points that if overlooked can cause a data analyst much grief:

1. Stata considers any numeric missing value (coded as `.`) to assume an arbitrarily large value and treats it as a large number in logical operations: hence, caution is necessary when applying variable-specific operations. If a user types `list if age > 10`, this command will list all data records in which `age` is larger than 10, including those with a missing value.

2. The pros and cons of labeling variable values: sometimes `label`ing values of a variable taking on a discrete number of levels (gender, race, education category, etc.) can cause confusion if the variable is later `recode`d.

While the section on do-files is helpful, it fails to highlight a nice, built-in feature in Stata, the ability to generate a do-file via the command review window, which can save a lot of typing, cutting and pasting, and toggling back and forth between the main Stata screen and the do-file editor.

Chapter 3 shatters the illusion that all existing data are ready for reading into Stata but offers plenty of helpful tips and exercises on importing data to Stata from other sources (spreadsheets, word processors, tab-delimited files). Chapter 13 deals with data management issues related to repeated measures data and the concepts of 'wide' and 'long' data formats. Nice examples are given on switching between formats, graphing repeated measures data, and merging files in the long format. Chapter 14 introduces ado-files and steps the reader through the construction of several ado-files designed to do useful data management and graphics tasks. Chapter 15 is a critical chapter and should also be required reading for any Stata user: here are the secrets of updating Stata via the Internet and downloading user-written ado-files. Also given is the web address for the Stata Journal, which of course will allow readers who subscribe to access this review!

3 Graphics

After three chapters on the basics of data management, Chapter 4 tackles the basics of graphing in Stata. Stata is a mixed blessing when it comes to graphics: while it is rather simple to get graphical output with Stata, making this output aesthetically pleasing can be frustrating. Furthermore, Stata's `help` menu entries on graphing are cumbersome and confusing to read. Chapter 4's treatment of Stata's graphical capabilities is quite nice. It gives the basics of making histograms, boxplots, cumulative distributions, and scatterplots. The chapter also details useful aesthetic enhancements: adding titles, labels, and lines to graphs, and selecting various symbols to mark data points. Also covered are saving, combining, and printing graphs.

The first 4 chapters constitute a perfect sequencing of the basics of commencing a successful data analysis. After covering only 4 chapters, readers can get their data into Stata, explore it, modify it for their needs, and then look at it!

4 Supplied programs

The CD-ROM that comes with the book includes some downloadable, menu-driven programs. There is a program for constructing tables, `tabmenu1`, and a program for performing multiple regressions, `effmenu1`. Chapter 6 gives a detailed account of the uses of `tabmenu1`. This is a very useful command, which essentially duplicates the capabilities of the `table` command in Stata. However, `tabmenu1` is much easier to use than `table` and includes more options. In addition to the display of variable means, standard deviations and frequencies, `tabmenu1` also allows the user to display confidence intervals, the odds of a binary outcome, and rates as computed as count per exposure time. These quantities can be broken out by levels of up to two explanatory variables. What a powerful tool for exploratory data analysis this feature is!

I wish I could say the same about the `effmenu1` program. The name derives from the word effect, as the program is designed to estimate the 'effect' of an explanatory variable on an outcome variable, allowing adjustment for other factors. In other words, this is a menu-driven program designed to make multiple regression easy to do. While I do not wish to split semantic hairs, I would rather the program be called something along the lines of `assocmenu` in honor of the word *association* (unfortunately, the prefix does not lend itself to further abbreviation), as *effect* is a very strong word to describe statistical relationships. When I mentioned earlier that this book may ultimately not be targeted to biostatisticians, it is partly because of the reliance on `effmenu1` in Chapters 7 and 8.

Chapters 7 and 8 give detailed examples of how to employ `effmenu1` to perform multivariate linear, logistic, and Poisson models accounting for confounders and exploring statistical interaction. No doubt `effmenu1` is a very handy tool for producing estimates of adjusted associations. However, the only information given in the output relates the outcome variable to the primary exposure of interest: no information is given about the coefficient values or confidence intervals for the controlling variables. In many observational and case-control studies, the substantive focus is on the adjusted relationship between an outcome and a primary exposure. The adjusted associations between the outcome and each controlling variable are in comparison of little interest. However, a lot can be learned about a dataset by monitoring all adjusted associations in a regression model and how these associations change (or do not change) in the presence of other controlling variables. This type of approach can not be employed via `effmenu1`, as it only reports the estimates related to the primary 'effect' variable specified.

Chapters 9 and 10 cover the same subject matter as Chapters 7 and 8 but employ Stata regression commands (`regress`, `logit`, `poisson`) to estimate outcome-predictor associations. These commands are more flexible than `effmenu1`, and Stata has some really nice features that complement its suite of regression commands. The book does a good job of detailing how to use `xi` when including categorical predictors in a regression model and of highlighting the built-in prediction commands. Chapter 10 gives a very concise, useful explanation of modeling and testing statistical interaction in regression. However, many additional regression-related features and commands in Stata are omitted in these two chapters. Many of Stata's features for assessing the validity of a linear regression model are ignored. Similarly, there is no mention of robust regression methods, generalized estimating equations approaches to regression on correlated outcomes, or bootstrapping methods. No mention is made of commands for performing ordinal or multinomial regression. While the chapters do a great job of explaining and illustrating the material they cover, the breadth of topics would not satisfy a professional biostatistician.

5 Survival analysis

Stata has a great collection of commands and functions for doing survival analysis. Chapter 11 highlights and gives examples of many of these features. The chapter gives adequate attention to creating a survival dataset, given a failure-time variable and a censoring indicator, with `stset`. The chapter covers constructing Kaplan–Meier curves, computing time-varying event rates, and both parametric and Cox regression models for failure-time data. Interestingly enough, no mention is made of either the log-rank test or the Gehan two-sample test, nor is there any discussion of incorporating time-varying covariates into the Cox model. The latter information is missed because Stata is so adept for this problem.

6 Summary

In general, this is an excellent introductory level book, which will allow its readers to be functional with Stata within a short time period. I have mentioned what I consider to be a few key omissions from the text. One other curious omission is reference to the StataQuest menu system. This user-friendly, free download from Stata would be the perfect accompaniment to the authors' included `tabmenu1` and `effmenu1` commands.

One of the nicest things about reviewing the book is how much I learned about Stata's capabilities in the areas of data management. It's amazing how set in my ways I had become, sometimes using inefficient methods simply because that's all I knew, and I wasn't able to improve on them using Stata's documentation. I am especially grateful to the authors for introducing me to the `compress` command for optimizing variable storage, and the `egen, cut()` function for categorizing continuous measures. The interactive nature of the book enriches the learning experience and makes the information real in the hands of the reader. Although it would not stand by itself as an appropriate text or reference for individuals interested in advanced biostatistical applications in Stata, this book can serve as an excellent accompanying text for introductory epidemiology and biostatistics courses. Furthermore, it is useful as a readable, searchable reference for a data analyst or biostatistician.

7 References

Hamilton, L. C. 2002. *Statistics with Stata: Updated for Version 7.* Pacific Grove, CA: Duxbury Press.

Hills, M. and B. De Stavola. 2002. *A Short Introduction to Stata for Biostatistics.* London: Timberlake Consultants Press.

About the Author

John McGready is a research associate in the Johns Hopkins University Bloomberg School of Public Health Department of Biostatistics.