The Stata Journal (2004)
4, Number 3, pp. 350–353

Review of A Handbook of Statistical Analyses Using Stata by Rabe-Hesketh and Everitt

Nicholas Winter
Department of Government
White Hall
Cornell University
Ithaca, NY 14853
nw53@cornell.edu

Abstract. The new edition of the book by Rabe-Hesketh and Everitt (2004) is reviewed.

Keywords: gn0013, introductory, gllamm, Stata texts

1 Introduction

The complete set of Stata manuals fills 13 volumes and weighs just over 26 pounds (almost 12 kg). Gone are the days when one could read the entire documentation in a long day at the beach and attain a relatively complete knowledge of the program. Thus, there is considerable demand for reasonably sized guides that provide a point of entry that is more user friendly and less intimidating than the manuals. *A Handbook of Statistical Analysis Using Stata*, 3rd ed. by Sophia Rabe-Hesketh and Brian Everitt is a newly revised version of one such resource. This book is positioned between the more introductory text by Hamilton (2004) and more advanced, book-length treatments of specific modeling strategies using Stata, such as Long and Freese (2003) on regression modeling with categorical dependent variables or Cleves, Gould, and Gutierrez (2004) on survival analysis.

2 Overview of the book

After giving a brief introduction to the basics of the program, this book covers a broad set of intermediate and advanced techniques in Stata. It includes chapters on descriptive statistics, regression, ANOVA, logistic regression, generalized linear models, descriptive analysis of longitudinal data, random-effects models, generalized estimating equations, basic epidemiology, survival analysis, maximum likelihood, principal components, and cluster analysis—all in less than 300 pages!

Each chapter includes a brief introduction to the relevant statistical model and then illustrates an analysis using that technique with a substantive problem and a real dataset. The book includes complete Stata syntax for each command in the analysis, interspersed with explanations of the Stata features and analytic techniques being employed, along with the output that Stata produces. The datasets are available from the

Internet, so readers can download them and duplicate or extend the analyses as they read. (Oddly, the book does not include directions for loading the datasets directly into Stata from the Internet; the web page simply instructs users to download them as a package to their local computer.)

The chapter on random-effects models is one of the longest and illustrates this approach well. The chapter begins with a very brief discussion of random-effects models, including random-intercept and random-coefficient models, as well as mixed and multilevel models. Using a dataset on postnatal depression, the authors fit a random-intercept model using `xtreg` and illustrate various postestimation tasks, including graphical analysis and the use of `lincom`. Then they turn to `gllamm`, a user-written program that fits generalized linear latent and mixed models. They first refit the equivalent model and then extend it to include random slopes. Each estimation result is followed by a set of further analyses that illustrate ways to use Stata's postestimation facilities, including an analysis of residual normality using `kdensity`, plots of individual-specific response curves, and a plot of predicted values with a confidence interval. The chapter concludes with an analysis, using `xtlogit` and `gllamm`, of a dichotomous dependent variable; this section includes both a brief theoretical discussion and analysis of various types of predicted probabilities.

The major changes in this new edition of the book include the addition of two new chapters (on random effects and cluster analysis) and the updating of the graphics throughout the book to version 8 syntax and output.

3 Strengths

The best characteristic of the book is the way it demonstrates the flow of an analysis in Stata. As the analysis in each chapter unfolds, readers see the ways that Stata's data management, graphics, model estimation, and postestimation facilities flow together, always in the context of a particular statistical model, substantive problem, and dataset.

The book also shows off various tips and tricks for using Stata that the novice or intermediate Stata user might not be aware of or habitually use. A reader working through the analyses will pick up along the way a quick view of a wide range of tools, including several different uses of `foreach` and `forvalues`; the use of `tabstat` as an alternative to `table`; the `preserve`/`restore` sequence; the use of `collapse`; the use of `reshape`; and more. This is where a hands-on book of this sort is at its strongest; using the manuals is a particularly difficult way to learn these sorts of things, if only because one does not know that a particular tool or approach would be helpful and, therefore, does not know to look it up.

Another nice feature (I would have appreciated more of them) is the occasional set of discussions on differences between modeling strategies; for example, `xtreg` and `gllamm` are compared directly, and random effects-style approaches are contrasted with generalized estimating equations for longitudinal data (page 183). For anyone relatively new to longitudinal analysis, for example, these sorts of comparisons could be quite helpful.

4 Limitations

For this sort of book, there is always a trade-off between breadth and depth (and length). Because the book is relatively short and attempts to cover an extremely wide set of techniques, there are definite costs in terms of the depth of coverage it can give to each. The random-effects chapter discussed above is among the longest in the book; at the other extreme is the 11-page chapter on principal-components analysis. This allows for only an extremely cursory introduction to each technique and a rather quick glimpse of Stata's handling of it. This may be a useful introduction, but clearly researchers who want to employ any of these methods for their own research would need to use this book in conjunction with the Stata manuals. They would also be well advised to consult more advanced texts on any of the methods they are not already familiar with—helpfully, the authors give many such references.

This combination of extreme breadth and limited depth makes it difficult to know precisely whom this book is aimed at. It is clearly pitched above the level of the statistical novice, as the introduction to each analytic technique is quite brief; on the other hand, the introductions should be superfluous for anyone well versed in these techniques. Moreover, the data manipulation at the beginning of each chapter may be confusing to Stata newcomers. For example, in the chapter on logistic regression, the data are loaded as a frequency table; `egen, seq()` is used to create the covariates; the data are then manipulated with `reshape` and `expand`; and, finally, the response variable is dichotomized. All this is done in less than two pages, which means that there can be little explanation of what each step accomplishes. Perhaps readers would be better served if the data were provided in a form more suitable for analysis.

In addition, and perhaps inevitably, I took issue with some of the statistical procedures employed. Specifically, the multiple regression chapter employs stepwise regression, a technique notorious for its problems (see, e.g., discussion by Harrell [2001]). While the authors do express caveats about stepwise procedures, their inclusion at least raises questions about the statistical advice they give for other techniques as well. In addition, they make some choices about data management that could get a novice user into trouble. To exclude an outlier in the regression analysis, for example, they `drop` it from the dataset, rather than excluding it from the analysis with an `if` condition. Without a clear explanation that this is what's going on, new users might think that this command means "exclude this from analysis", rather than "drop from the dataset permanently". At the very least, such users might inadvertently `save` the dataset without realizing what they were doing.

5 Conclusion

This book would probably be most useful for two types of Stata users. First, it would be helpful to working researchers or students who have a basic knowledge of Stata but are interested in learning a bit more about its capabilities, especially in the context of several analytic techniques with which they may be unfamiliar. For them, the book can serve as an accessible stepping stone both to the techniques and to modeling in Stata

more generally. Second, for working researchers who are new to Stata and want to make use of several of the techniques listed above, this book can serve as a nice introduction to Stata in general, as well as its implementation of the approaches.

6 References

Cleves, M., W. Gould, and R. Gutierrez. 2004. *An Introduction to Survival Analysis Using Stata*. rev. ed. College Station, TX: Stata Press.

Hamilton, L. C. 2004. *Statistics with Stata (Updated for Version 8)*. Pacific Grove, CA: Duxbury Press.

Harrell, F. E. 2001. *Regression Modeling Strategies*. New York: Springer.

Long, J. S. and J. Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*. rev. ed. College Station, TX: Stata Press.

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

Nicholas Winter is an Assistant Professor in the Department of Government at Cornell University. His interests include American politics; public opinion, race and gender, and politics; and methodology. He is a frequent contributor to Statalist and has written several Stata packages, which may be downloaded from the SSC archive.