The Stata Journal (2007)
7, Number 2, pp. 245–248

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

William D. Dupont
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville, TN
william.dupont@vanderbilt.edu

Abstract. This article reviews *A Handbook of Statistical Analyses Using Stata*, Fourth Edition, by Sophia Rabe-Hesketh and Brian S. Everitt.

Keywords: gn0037, applied statistics, Stata texts

1 Introduction

Rabe-Hesketh and Everitt (2007) provide an authoritative overview of modern methods of applied statistics. The writing is clear but terse. In only 342 pages they cover the major methods of biostatistics together with explanations of how to perform these analyses using Stata. The authors also provide valuable references to other texts that cover the material in greater detail. As noted by Winter (2004), the official Stata documentation is extensive: 15 volumes containing more than 6,000 pages (StataCorp 2005). This documentation is also organized as a reference work and is arranged in alphabetical order by program name. As a consequence it is not the sort of literature that readers are likely to browse looking for new ideas or techniques that they have not heard of elsewhere. Rabe-Hesketh and Everitt (2007) is the antithesis of the standard Stata documentation. It is a concise overview that will be a valuable guide to researchers who wish to confirm that they are up to date in their knowledge of applied statistics or are looking for important gaps in their understanding.

For the most part, this text covers topics that I think of as biostatistics. Much of this material will also be of use to Stata users who are not working in that field. I write this review from my perspective as a medical statistician.

2 Contents

The book starts with an introductory chapter on Stata. It provides a straightforward discussion of the Stata command interface, the various Stata windows, datasets, log files, do files, pulldown menus, getting help, command syntax, and other features that you might expect in such an introduction. Somewhat unexpected is a discussion of looping through variables or observations, macros, some of the more powerful data management

commands, matrix manipulation, Mata, and Stata programming. All of this is covered in an introductory chapter of only 41 pages.

Later chapters are on data description and simple inference, multiple regression, analysis of variance, logistic regression, generalized linear models, analysis of longitudinal data, random-effects models, generalized estimating equations (GEE), epidemiology, survival analysis, maximum likelihood estimation, principal component analysis, and cluster analysis. (You can find the table of contents for this book by going to http://www.crcpress.com and then searching for rabe.) These methods are illustrated with many helpful examples. The datasets for these examples may be freely downloaded. The authors place a commendable emphasis on graphical displays of data, residual analyses, and other means of checking for model fit.

A complete list of topics outlined in this book would be long, but the following will give some idea of the range of topics covered: The use of first- and second-order interaction terms is discussed in the chapters on analysis of variance. The chapter on logistic regression includes a discussion of the proportional odds model for categorical response data. The chapter on generalized linear models includes a discussion of the Huber–White sandwich estimator as well as the use of bootstrapping to obtain robust standard error estimates. Rabe-Hesketh is particularly well known for her work on longitudinal data analysis, and her chapters on this topic are among the best in this text. There is a nice discussion of response-feature analysis, which is a simple but often effective approach to repeated-measures data. The discussion of random- and mixed-effects models is helpful, as are the contrasts between the `gllamm`, `xtreg`, `xtmixed`, and `xtgee` commands for analyzing these data. The chapter on survival analysis discusses not only the proportional hazards model but also the use of time-dependent covariates. There is an interesting discussion of different methods of generating dendrograms in the chapter on cluster analysis. This is an abridged list of the topics discussed in this text. If you look hard, you will no doubt discover some favorite topic that they have omitted. But the breadth of topics covered in this short text is remarkable.

3 Strengths and weaknesses

This book's greatest strength is the breadth of the material that it covers in a few hundred pages. Its greatest weakness follows from this strength: it lacks the depth that most students will need when first learning about a topic. Nevertheless, this book succeeds at its objectives and will be a valuable resource for anyone with access to more detailed explanations of the topics that it summarizes. The many examples and effective graphics add greatly to the value of this publication.

Achieving a perfect balance in any text is always difficult, particularly a text like this one providing a synopsis of an entire discipline. One topic where I would recommend a little more detail concerns the variance–covariance matrix for the parameter estimates from a GEE model. When this topic was first introduced by Zeger and Liang (1986), they incorporated the Huber–White sandwich estimator as an integral part of their approach. Evidently, Zeger and Liang were unaware of the work of Huber (1967) and

White (1982) and independently rediscovered this technique. As a consequence, there is potential for confusion over whether a GEE analysis implies use of the Huber–White sandwich estimator. This detail is important, as without this correction, error estimates from GEE analyses can be misleading if the working variance–covariance matrix is misspecified. The authors of `xtgee` chose not to use the sandwich estimator by default. Rabe-Hesketh and Everitt do discuss the sandwich estimator in their chapter on generalized linear models, and they do use this estimator in one of their GEE examples. A little more emphasis on the importance of using this robust error estimate in GEE models would have been appropriate.

This book is well written and carefully copyedited. The only typographical error that I found is in the chapter on principal component analysis (sec. 14.2), where the authors use the notation y_{ji} rather than z_{ji} to describe the principal components. This error introduces some confusion between the principal components z_{ji} and the original variables y_{ji} from which the principal components are derived.

4 New material in the fourth edition

The fourth edition of this book is 34 pages longer than the third edition. New material on the `xtmixed` command for mixed models, new material on Mata, and more exercises have been added. However, I believe that most owners of the third edition would benefit more by buying the Rabe-Hesketh and Skrondal (2005) text on longitudinal modeling than by buying the fourth edition of Rabe-Hesketh and Everitt (2007). On the other hand, owners of the first or second edition of this book would benefit from the updated material, including the discussion of the new Stata graphics introduced in version 8.

5 Conclusions

If you are going to read only one book on applied statistics, then this one would not be a good choice. Most readers will find that it covers too much ground too quickly to stand on its own. On the other hand, if you have access to the Stata manuals and a comprehensive library of statistics texts, then you will find Rabe-Hesketh and Everitt (2007) to provide a valuable overview of modern statistical methods. The book gives a good feel for how two senior statisticians go about routine analyses of data. I recommend it for people who wish to pursue their own continuing education. It will be helpful in suggesting minor refinements to techniques that you are already using. For topics that are unfamiliar to the reader, this book identifies areas that are well worth learning. Here the authors refer the readers to more detailed texts. The Stata manuals themselves will also be a valuable source of more information on the topics introduced in this book. I could also see it used as a graduate text for a course on statistical consulting. Such a course would be targeted for students who had already received a firm grounding in biostatistics and would be intended to provide review and practice in the analysis and interpretation of a wide variety of datasets.

6 References

Dupont, W. D. 2002. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. Cambridge: Cambridge University Press.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 221–233. Berkeley, CA: University of California Press.

Rabe-Hesketh, S., and B. Everitt. 2007. *A Handbook of Statistical Analyses Using Stata*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

Rabe-Hesketh, S., and A. Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

StataCorp. 2005. *Stata Statistical Software: Release 9*. College Station, TX: StataCorp.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–26.

Winter, N. 2004. Review of A Handbook of Statistical Analyses Using Stata by Rabe-Hesketh and Everitt. *Stata Journal* 4: 350–353.

Zeger, S. L., and K.-Y. Liang. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42: 121–130.

About the author

William D. Dupont is a professor of Biostatistics and Preventive Medicine at Vanderbilt University School of Medicine. His interests include the epidemiology of benign breast disease, power and sample size calculations, statistical graphics, and teaching intermediate-level biostatistics to physician scientists. He is the author of *Statistical Modeling for Biomedical Researchers* (Cambridge University Press 2002), which uses Stata to teach biostatistics.