The Stata Journal (2005)
5, Number 4, pp. 594–600

Review of Data Analysis Using Stata by Kohler and Kreuter

L. Philip Schumm
Department of Health Studies
University of Chicago

Abstract. The new book by Kohler and Kreuter (2005) is reviewed.

Keywords: gn0030, data analysis, introductory, teaching, GSOEP

1 Introduction

This book is a revised and updated version of a German text published in 2001 by the same authors. As conveyed by the title, its purpose is to introduce Stata and data analysis simultaneously. The authors report having used it for teaching general introductory courses in data analysis and even more specialized courses in regression and categorical data analysis. It could also be used as a guide for self-study by researchers new to Stata.

The book is intentionally written in the style of a tutorial. It begins by instructing readers to download a companion package of datasets and other supporting files (either from within Stata using `net get` or via a web browser) and then encourages them to follow along, first by replicating each example as it appears in the book and then by continuing to explore where the example leaves off. In this way, even readers new to Stata will—if they work through all the examples—acquire a solid, working knowledge of Stata by the end of the book. Unlike some tutorials, the book is also pleasant to read even without following along, as might be done by an experienced Stata user wishing to pick up some new tips and techniques.

Most of the datasets used in the book are taken from the German Socioeconomic Panel (GSOEP), a longitudinal survey study of German households started in 1984 and continuing to the present (the book uses data from 1984–2002). Thus the analyses involve variables such as political party affiliation, income, home size, and life satisfaction (a noteworthy exception is an analysis of the probability of survival based on a well-known dataset on the sinking of the Titanic). This gives the book a distinctly social-scientific flavor, yet the statistical and Stata-related content would be equally relevant to other disciplines.

Although the original book was written during the days of Stata 7, this new version has been brought up to date with Stata 9. Thus the descriptions are consistent with Stata 9's interface, the examples and supporting files all work with Stata 9, and the sections on graphics all make use of the new graphics (first introduced in Stata 8). Note, however, that many of the features introduced in Stata 9 that are rather advanced (e.g., Mata, xtmixed, enhanced survey and multivariate statistics, etc.) are not covered explicitly here.

2 Contents

The book begins with a gentle introduction to Stata's user interface appropriately entitled "The first time". From the outset and throughout the book, everything is done at the command line (the menus are mentioned briefly only once). Important platform-specific details are addressed here (e.g., launching the application, path specification, etc.) and also in subsequent chapters as they arise (e.g., in chapter 6 when discussing graph export). In addition to getting the readers' feet wet, this first chapter introduces them to the help and search commands, essential for any Stata user.

The next five chapters cover in greater detail the basic knowledge needed to use Stata effectively. The first of these is a chapter devoted to the important yet often overlooked topic of working with do-files. More than just a convenient way to store and execute a set of commands, do-files are discussed as a means for structuring the entire process from dataset construction to final analysis so that it can be easily reproduced. Standard headers and footers for do-files are suggested (including commands to set up logging and version control), and good commenting technique is discussed. In addition, the authors present a paradigm for organizing one's work (including revisions and updates) among several different do-files including a *master do-file* for documenting and executing all of the individual do-files in the appropriate order. This accomplishes many of the same objectives that programmers accomplish using advanced tools, such as *makefiles* and code repositories, and chapter 2 does an excellent job of presenting these ideas in an accessible way that can be easily implemented using Stata alone.

Chapter 3 provides a detailed description of Stata's command syntax. All the standard elements are covered (command and variable abbreviation, the if and in qualifiers, weights, command options, expressions and functions, and filename specification), as are a few of the more subtle details such as the many ways of specifying variable and number lists and the use of Boolean expressions. The by prefix and the foreach and forvalues commands are also presented. Although looping can certainly be useful in the data analysis context, my personal view is that one needs to be careful about introducing it so early. For example, I have seen new users waste a lot of time mechanically elaborating on an analysis in combinatorial fashion (e.g., repeating it for many combinations of variables and/or a large number of subgroups), only to be frustrated trying to make sense of the resulting mass of output. In addition, it is common for new users to make the mistake of trying to use loops to construct and/or to modify variables when generate or egen should be used instead.

Despite an unassuming title "Some general comments on the statistical commands" and the fact that it is only four pages long, chapter 4 tackles the important topic of how to retrieve and use saved results stored in r() and e(). Its objective is to steer the reader away from manually transcribing results into subsequent calculations to avoid both errors and having to redo the calculation if the initial analysis needs to be modified. Like the chapter before it, this chapter also makes use of macros—a subject whose syntactic and behavioral subtleties can be confusing at first. For this reason, some readers may find it helpful at this point to look ahead at the section in chapter 11 on macros.

With the reader now familiar with Stata's syntax, the authors are able in chapter 5 to get down to the business of working with data. They begin with the workhorse commands `generate` and `replace`, including a discussion of how the `by` prefix together with the system variables `_n` and `_N` and subscripting can be used to construct a wide range of derived variables (this is one of the most important lessons in learning how to think about data manipulation problems in a Stata-like way). The use of `egen` and `recode` is also discussed, as are some of the most important commands and functions for working with strings and dates, missing values, and labels. The chapter concludes with a brief, nontechnical discussion of storage types, including the use of the `float()` function when testing for equality.

Chapter 6 is devoted to generating graphs and does a marvelous job of distilling Stata's complex graphics into a concise 34-page introduction containing exactly what a new user needs: a broad overview of the types of things that can be done together with enough practical knowledge to start generating useful graphs right away. All the basic graph types are covered, as are methods for customizing a graph's appearance and for combining graphs to generate specialized figures. I would recommend this chapter to anyone getting started with Stata graphics.

In chapters 7–9, the focus shifts from learning to use Stata to the process of data analysis itself. Chapter 7 deals with describing and comparing distributions using both tabular and graphical methods (formal statistical methods for comparison of discrete distributions are mentioned only in passing, and those for comparison of continuous distributions are not mentioned at all). The emphasis throughout is on how to use Stata to generate informative displays relevant to the question at hand. The section on continuous variables includes discussions of box plots, histograms (including an extended discussion of the importance of bin width), kernel density estimation, and quantile and Q–Q plots. Summary statistics, such as the mean, standard deviation, and quantiles, are also discussed, as are methods for displaying such statistics so that they can easily be compared across groups. While I certainly welcome the serious attention paid to what are essentially exploratory techniques, I suspect that in many cases, instructors using the book will want to supplement this chapter with a discussion of a few basic statistical tests (e.g., t test, Wilcoxon rank-sum test, and Pearson's chi-squared test).

Chapter 8 presents a self-contained introduction to linear regression. The chapter presumes little or no statistical background, motivating the approach informally as a means for determining a best-fitting prediction equation from a set of data. Simple linear regression is presented first, followed by a thorough explanation of each piece of the output from `regress`. Multiple regression is then presented, including a discussion of the adjusted R^2, standardized coefficients, and exactly what is meant by "controlling for the effects of covariates". The next section contains a nice discussion of regression diagnostics including checking for nonlinearity and heteroskedasticity using residual plots, identifying influential cases using DFBETAs and Cook's D, and the tradeoff between the problems of omitted variables and multicollinearity. Following this is a discussion of categorical covariates and interaction terms (including the use of conditional-effects plots for interpretation) and a brief introduction to variable transformation, the bootstrap estimate of standard error, and Stata's survey commands. The chapter closes with an introduction to the advanced topics of median regression and regression models for panel data (including error-components models). While it is perhaps helpful to alert the reader to the existence of these topics and to the corresponding Stata commands, readers new to these topics will almost certainly need to consult additional resources before undertaking such analyses themselves (the explanation of `reshape`, however, is generally useful).

Regression models for categorical data are the subject of chapter 9, most of which is spent on logistic regression. Following an explanation of odds and odds ratios and an interesting introduction to the maximum likelihood principle, the chapter largely parallels the preceding chapter, focusing first on interpreting the output of `logit` and then on various measures of model fit and other diagnostics. Probit models, multinomial regression, and ordinal logit models are covered briefly at the end, with emphasis on the interpretation of the corresponding coefficients. As before, I would suspect that some readers will need additional background on these more advanced models before being able to use them effectively. For example, although the authors state that ordinal logistic regression is a generalization of binary logistic regression, they do not fully explain that if the model is correct, the two are actually estimating the same quantity, leading to the obvious and simple diagnostic of fitting and comparing binary logistic regressions for each possible cutpoint.

In chapters 10–12, the authors return to discussing Stata itself. Chapter 10 covers reading and writing data that are not in Stata format together with the `merge` and `append` commands for combining files. Users of large survey (or other) datasets will appreciate the advice on managing memory when working with very large files. Chapter 11 provides additional information on macros and macro-extended functions but is really focused on writing programs and new commands. This is a lot to fit into 28 pages, but the authors do a good job of covering the details most relevant for a beginner. Although I suspect that most Stata users never find occasion to write their own programs, more might if they came across an accessible introduction like that provided in this chapter. At the very least, the information in this chapter will make the reader a more critical consumer of user-written programs.

The final chapter points the reader toward important resources for learning more: the *Stata Journal*, the Statistical Software Components (SSC) archive, the online FAQs, Statalist, and the other books available through the Stata Bookstore and Stata Press. The chapter also explains how to use `update` to keep the Stata software up to date.

3 Strengths and limitations

This book has several strengths. One of these is its emphasis on developing a solid understanding of how Stata works, attacking both data management and analytic tasks from first principles rather than the cookbook-style approach taken by some introductory texts. Another is the obvious care that went into preparing the datasets and supporting files. They transform the book into an interactive experience, making it both more enjoyable and more effective as a learning tool.

A unique and highly valuable feature of the book is its holistic approach to dataset construction and analysis and the strategies it provides for organizing the entire process so that it can be easily reproduced and, when necessary, modified. Chapter 2 lays the groundwork for this approach and would be well worth reading even by experienced researchers using Stata. The potential payoff includes not only avoiding wasted time spent trying to figure out what you did weeks, months, or even years ago, but more importantly, the ability to work more efficiently with less chance of error.

Perhaps the most important strength of the book is the impression of data analysis it conveys, namely, as an activity involving careful thought, a lot of time spent looking at data in tabular or graphical form, and a careful assessment of the adequacy of any model fit to the data. In this respect, the book could serve as a helpful corrective for students in disciplines where the methods courses are overly focused on statistical tests and the mechanical application of "accepted" models.

Of course, covering both Stata and data analysis in a single volume is a difficult task that necessitates painful compromises in the breadth and depth with which certain topics can be covered. The authors have no doubt made these compromises in ways they believe to be appropriate. Nonetheless, there are a few decisions I might have made differently. Many of these have to do with topics I believe could benefit from a more in-depth treatment. For example, the chapter on distributions (chapter 7) presents discrete and continuous distributions in a nonmathematical way without explicit reference to random variables, probability mass functions, density functions, or moments. Even a brief treatment of these concepts woven into the existing content of the chapter would help to make the discussion both here and in subsequent chapters more precise and would prepare the reader for seeing expressions such as $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$ in chapter 8 and the formulas for the binomial distribution and inverse cumulative normal distribution in chapter 9.

Another example of a topic that I believe deserves more attention is the difference between the exact (under normal theory) and asymptotic (under assumptions about the first and second moments) properties of regression coefficients. For instance, on pages 187–188, the authors make the excellent point that in general, a coefficient's confidence interval is more informative than its p-value, yet they also state in rather categorical fashion that "the confidence intervals are ... based on the assumption of normally distributed errors". Instead, it would be very useful to go on to explain that because of the central limit theorem, even in the nonnormal case, the intervals often have excellent coverage probability with moderately sized to large samples.

A final example concerns the use of transformations when fitting regression models. This is an important topic; in fact, there are some fields where, due to the nature of the quantities involved, transformation is almost the rule rather than the exception. Many researchers know that transformation can be helpful, but even experienced researchers often have questions about how to select a transformation in cases where the choice may not be obvious or how to interpret the coefficients from a model fitted using transformed data. Although the authors devote an entire subsection (8.4.3) to this issue, the topic is in my view sufficiently important to merit a longer discussion.

In addition to these few areas that I believe deserve greater emphasis, there is one more area that is barely covered, but it is one of Stata's unique strengths and also very useful for applied work. This is the ability to calculate standard errors in a variety of different ways across most of the basic estimation commands. These include the standard calculation based on the observed or the expected information matrix; the robust or sandwich estimator; the bootstrap; the jackknife; and in cases where the data arise from a complex sample, the linearization/robust method, balanced repeated replication, or a design-appropriate jackknife. Unfortunately, section 8.5 provides only a cursory introduction to the `bootstrap` command and the `svy` prefix, while the `robust` option is mentioned only in passing (page 216) and none is discussed in relation to the analyses in chapter 9. While the derivation of these estimators and their properties is certainly an advanced topic, the basic ideas can be introduced in a simple and straightforward way, and based on this, researchers can begin to investigate for themselves the effects of relaxing both distributional assumptions and the assumption of independence; in this way, I believe that providing more information on this topic would complement nicely the existing material in chapters 8 and 9.

One last remark: The book does not provide exercises—a point that has come up on Statalist. In a response there, Kohler noted that those teaching with the book will typically wish to use exercises from their own disciplines, while those using the book for self-study will likely have their own research questions to which they will want to apply what they have learned. I suspect that as more people use the book, suggestions for exercises will begin to become available (the introduction cites one such set of exercises for use with chapter 7 that is already available via the Internet).

4 Conclusion

Too often, courses in applied statistics and data analysis are taught using a text covering the theory behind the methods but expecting students to pick up the details of a particular software package on their own via hastily assembled or otherwise inadequate resources. This can create unnecessary stumbling blocks and frustration, but perhaps more importantly, it reduces the chance that the student will acquire a set of skills that he or she can apply effectively outside the classroom. The addition of a book like this would be a substantial improvement because of its systematic approach to learning how to manipulate and analyze data using the software and its strategies and advice for organizing the research process.

Although using the book by itself might not be appropriate for a class in a statistics department, it would be adequate for introductory classes in quantitative methods in certain other disciplines. In this case, instructors may wish to supplement some of the statistical content with additional material relevant to their particular disciplines. This would be an effective way of addressing some of the limitations identified above.

In addition to its value in the classroom, this book is also likely to be useful for experienced researchers. Those new to Stata will find that going through the book is one of the best (and quickest) ways to acquire a solid working knowledge of the program. Those who have been using Stata for a while are likely to find several new tips and tricks, which will make them more productive in their work.

5 References

Kohler, U. and F. Kreuter. 2005. *Data Analysis Using Stata*. College Station, TX: Stata Press.

About the Author

Phil Schumm is a statistical consultant and is Assistant Director of the Biostatistics Consulting Laboratory and Director of the Research Computing Group in the Department of Health Studies at the University of Chicago. He is currently involved in the National Social Life, Health, and Aging Project (NSHAP)—a U.S.-wide survey study of social life and health at older ages. He is also involved in developing software to facilitate reproducible and collaborative research.