The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 1, pp. 206–211

Review of Data Analysis Using Stata, Third Edition, by Kohler and Kreuter

L. Philip Schumm
Department of Health Studies
University of Chicago
pschumm@uchicago.edu

Abstract. This article reviews *Data Analysis Using Stata*, Third Edition, by Ulrich Kohler and Frauke Kreuter (2012 [Stata Press]).

Keywords: gn0056, data analysis, introductory, teaching, German Socio-Economic Panel

1 Introduction

In the latest edition of their introductory text, Kohler and Kreuter (2012) offer a substantial update covering new statistical material, new features introduced in Stata 11 and 12, and updated datasets used throughout the examples and exercises. I reviewed the first American edition of this book (Schumm 2005), and much of what I wrote in that review applies equally to this new edition. Yet rather than repeat my original comments, I shall highlight the new material and try to make some fresh remarks.

For those unfamiliar with the book, it is organized around the premise that data analysis is an activity best learned by doing—a premise few (if any) statisticians or data analysts would question. The authors have spent considerable effort developing a package of real datasets and do-files to accompany the book (these are easily installed from within Stata by using `net get` or downloaded by using a web browser), and much of the book is written as a tutorial assuming that the reader is following along on his or her computer. Despite this, most of the book may also be read productively without following along, as might be done by someone who already has some experience with Stata or data analysis but is looking to pick up some new skills.

The data used in the book are drawn primarily from the 2009 German Socio-Economic Panel, with a few exceptions including the well-known dataset on survival among passengers of the Titanic. Although the examples are intrinsically interesting (for example, an exploration of the wage gap between men and women), the choice of datasets and material to cover (for example, complex surveys and `svy`) gives the book a distinctly social science feel. Thus, while most of the Stata-related and statistical content is universally applicable, students from other fields may need to work a bit harder to apply what they are learning to their own data and analytic questions.

Many who use Stata regularly do so because they believe that it offers unique advantages over other software for analyzing data. Although this book covers many of these features, it does so without identifying them as such and without any comparisons to other software. Thus teachers who assign the book may wish to explain up front why they chose Stata and what its strengths are to help motivate the students. Readers new to Stata who are considering this book for self-study can find such information summarized on StataCorp's website.

2 Overview

This book covers three distinct, albeit related, topics: an introduction to Stata, an introduction to the practice of data management and analysis, and an introduction to statistical inference and modeling. These topics are interwoven such that the book may be worked through linearly by readers initially unfamiliar with all three. At the same time, those wishing to concentrate on a subset of these topics can do so easily by skipping some of the chapters and referring to them only as necessary (when a discussion depends on material from another chapter, this is usually well marked). In the preface, the authors provide suggestions for using the book to teach introductory courses in data analysis, regression, and the analysis of categorical data.

Introducing Stata receives the most comprehensive treatment. Chapter 1 (appropriately titled "The first time") takes a new Stata user by the hand and dives right into Stata's command-line interface (the command-line interface is used throughout the book to facilitate the discussion of examples and the use of do-files). This involves working through a Stata session involving loading and looking at a dataset; making some simple changes (recoding or labeling a variable); generating a few familiar summaries (mean, range, and standard deviation), tables, and a graph; and even fitting a linear regression model. This chapter does a good job of orienting a new Stata user and is written in a way that makes it accessible even to those with modest computer skills (as is true of the entire book).

Seven additional chapters are devoted primarily to instruction in the use of Stata. Chapter 3 provides an in-depth explanation of Stata's command syntax, including (but not limited to) the use of different types of weights, Stata expressions, operators, and functions, and even the somewhat more advanced topics of the `by` prefix and the use of `foreach` loops. Chapter 4 provides an overview of how Stata's statistical and estimation commands work and, in particular, how the results from such commands may be accessed programmatically for subsequent use. Chapter 5 covers the major issues in manipulating variables, including handling dates and times, missing values, and a brief discussion of storage types and precision. Of particular note, this chapter also shows how to use the underscore variables `_n` and `_N` together with `by`, which is one of Stata's more powerful features. Chapter 6 provides a self-contained introduction to Stata's graphics, covering several of the most commonly used graph types and general techniques for modifying and manipulating graphs (including use of the graph editor). Chapter 12 briefly covers the more advanced topics of macro usage and how to write

Stata programs (both those defined in do-files and those defined in ado-files). Finally, chapter 13 provides information on keeping Stata current and on the various online resources available for learning more about Stata and for obtaining user-written commands (the *Stata Journal*, Statalist, and Statistical Software Components). In sum, a novice user who masters the material in these chapters will have attained sufficient proficiency to work effectively in Stata.

Data management and analysis are primarily covered in chapter 2 ("Working with do-files"), chapters 5 and 6 (described above), chapter 7 ("Describing and comparing distributions"), and chapter 11 ("Reading and writing data"). These chapters, perhaps with occasional reference to chapters 3 and 4, could be used by someone with some experience in Stata who wants to increase his or her skills in manipulating and summarizing data. Chapter 2 sets the tone perfectly by emphasizing the use of do-files (scripts containing Stata code) and providing guidance on organizing them within a project. These do-files then become the official record of one's work and may be rerun at any time to recreate intermediate datasets or analytic results. Although written for Stata users, the ideas in this chapter are relevant for anyone who wants to make his or her data management and analyses more efficient and reproducible.

Chapter 7 uses some of the tools presented in previous chapters to examine the distributions of both discrete and continuous variables. The emphasis is on exploratory techniques without reference to formal statistical notation or analyses. Included are several ways to generate tables, the effective use of dot charts, box plots, histograms, kernel density estimates, quantile plots, and Q–Q plots. By encouraging the reader to look at and think about his or her data before moving to a more formal analysis, the book teaches that data analysis is a back-and-forth, investigative process as opposed to the mechanical application of a fixed set of techniques (this emphasis is then carried through the next three chapters on statistical inference and modeling).

Few analysts will be lucky enough to deal exclusively with Stata-format files. In chapter 11, the authors show how to read data from the three main types of text files (spreadsheet, free format, and fixed format), how to read data from various binary formats (Excel, SAS), and how to enter data by hand. This chapter also deals with combining data, including merging and appending, and with handling large datasets. Once again, although written for Stata, several of the ideas in this chapter are general concepts that would apply in some form regardless of the software package being used.

Lastly, chapters 8, 9, and 10 provide a nonmathematical introduction to statistical inference and modeling. Each chapter is relatively self-contained and can be read on its own. Chapter 8, titled "Statistical inference", is new in this edition and is discussed below. Chapter 9 is a long chapter (87 pages) covering linear regression. This chapter begins with an intuitive example of the regression principle (modeling home size as a function of income) and then goes through the results available following a simple linear regression (coefficients, ANOVA table, F test, and R^2). A section on multiple regression comes next, which includes an explanation of standardized coefficients and of the effects of adding a covariate to the model based on an added-variable plot. This is followed by sections on regression diagnostics, model-building techniques such as the

use of categorical covariates, interaction terms and transformations, and methods for exploring and reporting regression results. The chapter ends with a whirlwind tour of median regression and regression models for panel data (including fixed-effects, random-effects, and population-averaged models).

Chapter 10 deals with regression models for discrete variables, focusing primarily on logistic regression. It begins with an exploration of the linear probability model as a way to motivate the logistic model and spends ample time explaining how to interpret the coefficients from a logistic regression. The chapter also introduces the maximum likelihood principle for fitting a model and the likelihood-ratio test for comparing nested models. Similarly to chapter 9, methods for assessing model fit and for checking model assumptions are discussed, as are building models with transformed covariates and interaction terms. Finally, the alternative probit model, multinomial logistic regression, and models for ordinal data (the stereotype and proportional odds models) are briefly introduced.

Throughout chapters 9 and 10, the emphasis is on model specification (that is, which covariates to include and in what way), checking model assumptions, and exploring the response surface as a function of the covariates. This is entirely appropriate for students first learning to analyze data. Teachers who wish to increase the emphasis on hypothesis testing and model-based inference can easily do so by using the ideas presented in chapter 8 or additional materials of their own.

3 New material

Among the new Stata features covered by the book, four are especially noteworthy. Factor variables (added in Stata 11) are introduced and demonstrated repeatedly, including their use in constructing quadratic and complicated interaction terms. Reading data directly from Excel files (added in Stata 12) is also demonstrated, which will be of particular interest to researchers in biological fields where raw data are often provided in this format. Multiple imputation using the `mi` suite of commands (first added in Stata 11 and substantially enhanced in Stata 12) features prominently in chapter 8 (see below). Finally, `margins` (added in Stata 11) and `marginsplot` (added in Stata 12) are now used as one of the main tools for interpreting and presenting the results of regression models. By making it easy to construct conditional-effects plots even in the presence of interaction or nonlinear terms, these commands encourage researchers to spend time exploring the models they fit (as opposed simply to looking at the significance of the coefficients). Because this is exactly the approach emphasized throughout the text, the addition of these commands strengthens the book considerably.

The other major addition is a new chapter (chapter 8) devoted to statistical inference. All the discussion of inference (that is, standard errors and confidence intervals) has been relocated to this chapter, and readers are referred back to it from chapters 9 and 10 when discussing the estimated coefficients from those models. In addition to presenting the idea of a sampling distribution, the chapter includes a discussion of inference from complex samples (using Stata's `svy` command), the use of poststratification and multiple imputation for handling nonresponse, and a brief discussion of causal inference.

In keeping with the rest of the book, random variables and their distributions are presented nonmathematically via the use of simulation. This is, of course, an excellent device for giving an intuitive sense of what bias is, of how the sampling distribution of a statistic changes as the sample size increases, and of how the central limit theorem works. It is also an excellent way of explaining confidence intervals, and the chapter includes an informative demonstration in this regard. Importantly, the authors do not merely present simulation results but provide the reader with the tools necessary to perform his or her own simulations (via the use of Stata's random number functions and techniques for drawing repeated subsamples from an existing dataset).

At the same time, I believe that the authors may have overreached a bit with this chapter. It is organized around the distinction between *descriptive inference* (making inferences about a fixed population from which you have obtained a sample) and *causal inference* (making inferences about an underlying data-generating mechanism) and even introduces the concept of counterfactuals. Yet although this is undoubtedly an important distinction, I'm afraid that the relatively short treatment provided here may raise more questions than it answers, especially among students who may not be working with survey data (for example, data from a scientific experiment or randomized clinical trial). In fairness, the authors themselves acknowledge that they had the same concern and have addressed this by providing numerous excellent references for those who wish to explore these issues further. Those teaching with the book may also want to provide additional materials and support for this chapter.

4 Final thoughts

With each new release of Stata, those who have been using it for a long time relish in the new features; at the same time, the task faced by new users seems ever more daunting. Thus books like this that guide a new user through the initial learning process are arguably becoming even more valuable. *Data Analysis Using Stata* provides a broad introduction to the Stata software—one that does not assume any prior experience with statistical software or programming.

Selecting an introductory book for a technical subject can be difficult because many contain information that is misleading or in some cases downright incorrect. With books on data analysis and statistics, this often takes the form of reducing the subject to a series of cookbook-like steps, discouraging readers from exploring their own data and giving them a false sense of confidence. This book avoids these pitfalls and instead provides an accurate picture of how real data analysis should be done. In fact, by

choosing to cover a broad range of material while still including many of the technical details, the authors open up the subject and encourage the motivated reader to pursue further study (the text is rich with carefully selected references to help with this).

5 References

Kohler, U., and F. Kreuter. 2012. *Data Analysis Using Stata*. 3rd ed. College Station, TX: Stata Press.

Schumm, L. P. 2005. Review of Data Analysis Using Stata by Kohler and Kreuter. *Stata Journal* 5: 594–600.

About the author

Phil Schumm is a statistical consultant, an assistant director of the Biostatistics Consulting Laboratory, and the director of the Research Computing Group in the Department of Health Studies at the University of Chicago.