Journal of Statistical Software

April 2017, Volume 77, Book Review 3. doi: 10.18637/jss.v077.b03

Reviewer: Oliver Kirchkamp, Hiltrud Niggemann

An Introduction to Stata Programming (2nd Edition)

Christopher F. Baum
Stata Press, College Station, 2016.
ISBN 978-1-59718-150-1. 412 pp. USD 62.00 (P).
http://www.stata.com/bookstore/introduction-stata-programming/

Introduction

Stata is a statistical software package which allows solving a number of standard problems with the help of simple and often straightforward commands. More involved problems may need some programming: Non-standard estimators can often be implemented with the help of small programs. Also, the systematic preparation of data and the efficient presentation of results usually requires writing a program. In these cases, some of Stata's features, which are a blessing for simple tasks, demand at least some creative rethinking. Stata's macros, scalars, matrices, views, and other features are quite different from the concepts found in many other programming languages and often seem intimidating to the uninitiated. A person who wants to use Stata (or any other statistical software package) does not only have to learn a number of commands. Even more important is an understanding of an efficient workflow from data management to statistical estimation.

There are books that specialize on some of the issues mentioned above. E.g. Kohler and Kreuter (2012) provide a general overview of the Stata software. StataCorp's User's Guide (2015) describes the elements of Stata and explains how to write programs. Long (2009) presents a book entirely dedicated to workflow with Stata. Baum's *An Introduction to Stata Programming (2nd Edition)* combines a concise introduction into the Stata programming environment with a large number of applications to workflow, data management and estimation.

Overview

The book can be divided into four parts: A first part (Chapter 1 and 2) provides a general discussion, motivation and introduction into the basic concepts of data handling and programming with Stata. The second, and, in terms of pages, largest part (Chapters 3 to 10) presents essential Stata commands that can be used in do-files (but also in ado-files). In a third part (Chapters 11 and 12) ado-files and the syntax of self-written Stata commands are presented. A last and fourth part (Chapters 13 and 14) introduces the Mata language. Throughout the

book, chapters with odd numbers give a first introduction to a concept. Chapters with even numbers subsequently illustrate the concept with examples.

The first two chapters provide a general discussion of the key aspects of the Stata environment. Chapter 1 highlights the importance of reproducibility. Using a program does not only help solving complex statistical problems. Following a program helps, in particular, structuring and documenting even trivial tasks, thus making the analysis reproducible. Chapter 2 then explains the basic concepts how data is organized in Stata and how and where elements of programs are stored.

Chapters 3 to 10 introduce the main elements of the Stata language as they are used in do-files. Chapter 3 starts with conventions how Stata's commands are structured, how parts of these commands can be stored and manipulated in macros, how subsets of a dataset are selected, how one data type can be converted into another data type, and how standard statistics can be calculated conditionally and unconditionally. Chapter 4 then applies these concepts to standard tasks.

Chapter 5 presents several commands to validate and reorganize, reshape and append data. The chapter also explains how Stata functions return results, how these results can be extracted, and how results can be stored. In line with efficient workflow this chapter presents examples how to translate estimation results into publication ready tables. Chapter 6 again presents several useful applications, in particular for merging and reshaping data, but also for the generation of tables and graphs.

Chapter 7 discusses repetition in various forms. Stata offers a number of concepts to perform the same task in a loop, or repeatedly for different conditions, for different values or with different indicators. This chapter also presents commands that are useful in the context of resampling. Chapter 8 presents applications where statistics are calculated for different countries or individuals, or for a moving window. Furthermore, the chapter illustrates how to collect data from several spreadsheets.

Chapter 9 deals with a restriction in Stata: There is only a single dataset. How can the user organize data from repeated computations? How can the user compare different cases? This chapter presents matrices and external files as one possible solution. Baum also gives further advice how to produce publication ready tables and graphs. Chapter 10 provides examples how statistics for subsets of the data are stored in matrices and how one can display estimation results in a graph if these estimation results are conditional on a variable. The chapter also gives another example for the systematic generation of publication ready tables and explains how data can be extracted from Stata graphs.

Chapters 11 and 12 introduce the reader to writing own programs, i.e., ado-files. Chapter 11 discusses the syntax to define these programs, their options and the variables they refer to. The chapter also explains how to return results and how to write programs so that they follow standard conventions to select subsets of the data and so that these programs can be executed repeatedly for subsets of the data. Some Stata commands require user-written evaluator programs. Baum gives examples how to write such programs so that they can be used in the context of maximum likelihood estimation, nonlinear least-squares, generalized method of moments, or together with resampling commands. Furthermore, the chapter provides some guidelines on programming style. Chapter 12 presents a number of examples which use user-written programs.

Chapters 13 starts with a brief introduction into the fundamentals of the Mata language.

Thereafter, the author explains how to access and modify other Stata objects from within Mata. He then guides the reader through applications where Mata programs, combined with ado-files, are put to practical use. Chapter 14 presents a number of applications, including the reorganisation of data, more involved estimation problems, and improving the presentation of results.

Strengths and limitations

This is, in particular, a book for readers from economics or finance who already have some background using Stata, and who also have some background in econometrics and in data analysis. Readers from other fields will still find a large amount of useful information in this book. However, they might have more difficulties understanding the examples and the problems that are discussed in the book.

Readers who are new to Stata will find the strictly necessary basics which allow them to follow the text. It helps a lot, however, if readers have already some familiarity with Stata. A novice Stata user might prefer more detail and might want to read StataCorp's *User's Guide* first, and Baum's *An Introduction to Stata Programming* thereafter.

The book covers a large number of empirical examples, in particular from economics and finance. The focus on economics and finance will help readers from these fields but for readers from other disciplines this focus could be too specific to allow an easy access to the underlying programming problems. The book also discusses at length practical examples for importing data. Furthermore, the author examines several possibilities to generate informative and reproducible graphs, LaTeX tables, or plain text results with the help of a small program. While LaTeX might be a lingua franca for many readers, some users of other formats might feel left behind.

When it comes to workflow, Baum gives a lot of helpful advice on writing reliable and reproducible code. However, workflow is not the main focus of the book. The author takes a pragmatic perspective, e.g., when he finds on-the-fly work with StatTransfer acceptable instead of exploring how to import foreign data in a reproducible way with the help of a program. Also, some of the programs which are presented in the book contain repetitions of (almost) identical commands which a more stringent approach to workflow might perhaps avoid.

Chapters 11 and 12 are particularly interesting for Stata users. The earlier chapters focus on problems which are relevant issues also in other statistical software packages but which demand a specific approach in the Stata environment. Chapters 11 and 12 discuss a specific feature of Stata: Stata's rather homogeneous family of postestimation commands. In these chapters the focus of the book is not so much on solving a given statistical problem but rather on integrating the user-written programs efficiently and consistently into Stata's postestimation commands.

In terms of structure, the book is based on lots of applications and examples. Baum first introduces a concept and then presents an application of this concept to either the management of data or to an estimation problem. This is neither a book on specific Stata commands nor a book about estimation strategies. As the title says, it is a book about programming techniques for Stata. The author discusses alternative solutions and attacks problems either with a tailored user contributed routine, or with a specialized Stata command or with a creative

combination of several basic commands. More importantly, the author points out common mistakes, dangers and pitfalls.

The book is neither designed to be used as a comprehensive reference for Stata commands, nor as an exhaustive reference for econometric problems. The book can and should be used as a creative inspiration for programming techniques for Stata users. The examples are realistic but sometimes hard to grasp for readers not from economics and finance. Each example serves as a useful illustration for a programming concept but, obviously, the examples cannot exhaust all possible applications. Hence, most readers will read the book from cover to cover. The focus on practical applications also implies that related programming strategies appear and reappear in different chapters. Here an overview could have helped the reader to find where the related concepts appear throughout the book. The book does have an index but this index becomes most helpful once a reader has read the entire volume and is now looking for a specific item.

Conclusion

An Introduction to Stata Programming (2nd Edition) is a well organized book. We find it suitable for any Stata user on an intermediate or advanced level, a user which already has some experience with Stata and who wants go deeper into programming or who wants to extend Stata's built-in commands for estimation and data management. In particular the large number of practical examples, mostly taken from economics and finance, help the reader a lot.

References

Kohler U, Kreuter F (2012). *Data Analysis Using Stata*. Stata Press.

Long JS (2009). *The Workflow of Data Analysis Using Stata*. Stata Press.

StataCorp (2015). *User's Guide*. Stata Press.

Reviewer:

Oliver Kirchkamp
Friedrich-Schiller-Universität Jena
School of Economics
07737 Jena, Germany
E-mail: oliver@kirchkamp.de
URL: https://www.kirchkamp.de/

Hiltrud Niggemann
Schlehendornweg 24
07751 Jena, Germany
E-mail: niggemann@p-wert.de
URL: https://www.p-wert.de/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
April 2017, Volume 77, Book Review 3 *Published:* 2017-04-26
doi:10.18637/jss.v077.b03