The Stata Journal (2009)
9, Number 1, pp. 158–160

Review of The Workflow of Data Analysis Using Stata, by J. Scott Long

Alan C. Acock
Department of Human Development and Family Sciences
Oregon State University
Corvallis, OR
alan.acock@oregonstate.edu

Abstract. This article reviews *The Workflow of Data Analysis Using Stata*, by J. Scott Long.

Keywords: gn0044, data analysis, workflow, data management, wdaus

1 Introduction

The Workflow of Data Analysis Using Stata (Long 2008) is a must read for every Stata user. The book defies a simple description. It is not a substitute for Stata's own reference manuals on data management or programming, but most readers will keep *Workflow* handy as a reference book for when they cannot remember how to do something. It is a great supplemental book for statistics courses precisely because such courses rarely spend much time on the topics this book covers—topics that students need to know to be effective Stata users. *Workflow* is a reflection of what a senior scholar has learned about what happens as we move from raw data to a published report.

Workflow presents many guidelines and few readers will agree with all of them. Many of the guidelines may seem obvious to experienced users, but this is only because they were learned the hard way. For example, on page 143 Long writes, "Never change a variable unless you give it a new name." Those of us who have, on occasion, ignored this advice learned its importance when a pivotal variable was irreparably destroyed. I fear that some beginning Stata users who read such guidelines are likely to ignore them. Long minimizes this risk by showing the likely consequences of ignoring many of his guidelines. Seeing the damage that can be done by ignoring his advice encourages closer attention to all the guidelines.

How the book is best used depends on the reader. My beginning graduate students will be told to read the entire book and to reread the relevant sections before asking me to solve a problem. (Think of how often you have had to explain a loop to a beginning student, one-on-one.) Whoever has the responsibility of organizing, cleaning, and managing datasets for a project should read the entire book and, hopefully, follow many of the guidelines. Data analysts who are working on a project that has a well-managed dataset can read the book selectively but will learn to appreciate the critical role of the data manager and will learn why they should not always do what is quickest for their own particular purpose. Experienced Stata programmers and data managers

should scan through the book carefully enough to see where they agree with Long and where they disagree with him. StataCorp programmers should read it carefully because Long offers a number of "work arounds" and tools that could be implemented into the next version of Stata.

2 Chapters

The first two chapters aim to show readers the importance of planning, organizing, and documenting. I found these to be the least engaging chapters because I wanted to get to the "how do I do it" parts of the book. Chapter 3 focuses on do-files. This chapter is most useful for beginners, but the systematic presentation of how to make do-files robust and easy to debug is also useful for intermediate Stata users and serves as a great reminder for advanced Stata users.

Chapter 4 covers ways to automate your work, delivering a clear explanation of loops and describing how to make simple ado-files. There is sufficient detail to make a beginning or intermediate user much more efficient; experienced Stata programmers can scan this chapter. Chapter 5 may be the most proscriptive chapter. It describes Long's recommendations for naming datasets, variables, and values. Stata's output was designed to work with fairly short variable names, variable labels, and value labels. Long wants the user to be responsive to these design decisions. After reading 70 pages on naming conventions, I am convinced that, for a complex project, I need to pay more up-front attention to these decisions than I have before. This attention to detail would have saved me many hours in the long run.

Chapter 6 is on cleaning data and includes tips on data management. The lesson is that there are many things that can go wrong and only an organized, systematic, and meticulously documented approach is appropriate. Anything less up-front can lead to a disaster and huge inefficiency in the analysis stage of a project. Long covers the numerous Stata commands that make it a great program for data cleaning. His treatment of the `egen` command was too limited, but I was impressed otherwise.

Chapter 7 is on analyzing and presenting results. This is not a replacement for a book on statistics using Stata; only simple procedures are illustrated. Instead, it focuses on efficient analysis. Long did an excellent job describing how to work with the statistics that Stata holds in memory—how to retrieve, use, organize, and report them. Chapter 8, the final chapter, provides guidelines for archiving and protecting data.

3 Conclusion

The book is written assuming that the reader is using Windows. There is an occasional mention of differences for Macintosh and Unix. The Windows user will be fully satisfied; Macintosh or Unix users will have to figure out a few things on their own. I particularly liked the appendix, which offered many tips on how to customize Stata and how to work on a network.

I have consulted on several complex projects where too little of what Long explains was done up-front, before the analysis began. It is almost always necessary to correct data-management mistakes when doing an analysis, but the less correcting the statistical analyst needs to do, the more efficient and effective his or her efforts will be. For example, it is very useful to have a documented, project-wide standard for labeling types of missing values (not just `.a`, `.b`, `.c`, etc.).

The chapters in *Workflow* are largely self-contained for readers with some knowledge of Stata. This results in some redundancies across chapters. Some of the books from Stata Press have brief and unhelpful indexes. By contrast, *Workflow* has a very detailed index, and it is highly usable. This means that the book can be as effective as a reference manual as it is as a textbook. For example, if you want to know how to use `foreach`, you can look it up or, if you cannot remember the exact name, you could look up loops. Long provides a personal web page for the book at

http://www.indiana.edu/~jslsoc/web_workflow/wf_index.htm

It has numerous supplements to the book and will evolve. All do-files and datasets are also available by typing `findit workflow` in Stata's Command window.

Why did I start by describing *Workflow* as a must read? Because it changed what I teach, how I advise others when consulting, and how I manage my own workflow. I am confident that most readers will have the same experience.

4 Reference

Long, J. S. 2008. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.

About the author

Alan C. Acock is a University Distinguished Professor and Barbara Knudson Chair for Family Research and Policy in the Department of Human Development and Family Sciences at Oregon State University.