Preface

This book is about methods that allow you to work efficiently and accurately when you analyze data. Although it does not deal with specific statistical techniques, it discusses the steps that you go through with any type of data analysis. These steps include planning your work, documenting your activities, creating and verifying variables, generating and presenting statistical analyses, replicating findings, and archiving what you have done. These combined issues are what I refer to as the *workflow of data analysis*. A good workflow is essential for replication of your work, and replication is essential for good science.

My decision to write this book grew out of my teaching, researching, consulting, and collaborating. I increasingly saw that people were *drowning* in their data. With cheap computing and storage, it is easier to create files and variables than it is to keep track of them. As datasets have become more complicated, the process of managing data has become more challenging. When consulting, much of my time was spent on issues of data management and figuring out what had been done to generate a particular set of results. In collaborative projects, I found that problems with workflow were multiplied. Another motivation came from my work with Jeremy Freese on the package of Stata programs known as SPost (Long and Freese 2006). These programs were downloaded more than 20,000 times last year, and we were contacted by hundreds of users. Responding to these questions showed me how researchers from many disciplines organize their data analysis and the ways in which this organization can break down. When helping someone with what appeared to be a problem with an SPost command, I often discovered that the problem was related to some aspect of the user's workflow. When people asked if there was something they could read about this, I had nothing to suggest.

A final impetus for writing the book came from Bruce Fraser's *Real World Camera Raw with Adobe Photoshop CS2* (2005). A much touted advantage of digital photography is that you can take a lot of pictures. The catch is keeping track of thousands of pictures. Imaging experts have been aware of this issue for a long time and refer to it as *workflow*—keeping track of your work as it flows through the many stages to the final product. As the amount of time I spent looking for a particular picture became greater than the time I spent taking pictures, it was clear that I needed to take Fraser's advice and develop a workflow for digital imaging. Fraser's book got me thinking about data analysis in terms of the concept of a workflow.

After years of gestation, the book took two years to write. When I started, I thought my workflow was very good and that it was simply a matter of recording what I did. As writing proceeded, I discovered gaps, inefficiencies, and inconsistencies in what I did.

Sometimes these involved procedures that I knew were awkward, but where I never took the time to find a better approach. Some problems were due to oversights where I had not realized the consequences of the things I did or failed to do. In other instances, I found that I used multiple approaches for the same task, never choosing one as the best practice. Writing this book forced me to be more consistent and efficient. The advantages of my improved workflow became clear when revising two papers that were accepted for publication. The analyses for one paper were completed before I started the workflow project, whereas the analyses for the other were completed after much of the book had been drafted. I was pleased by how much easier it was to revise the analyses in the paper that used the procedures from the book. Part of the improvement was due to having better ways of doing things. Equally important was that I had a consistent and documented way of doing things.

I have no illusions that the methods I recommend are the best or only way of doing things. Indeed, I look forward to hearing from readers who have suggestions for a better workflow. Your suggestions will be added to the book's web site. However, the methods I present work well and avoid many pitfalls. An important aspect of an efficient workflow is to find one way of doing things and sticking with it. Uniform procedures allow you to work faster when you initially do the work, and they help you to understand your earlier work if you need to return to it at a later time. Uniformity also makes working in research teams easier because collaborators can more easily follow what others have done. There is a lot to be said in favor of having established procedures that are documented and working with others who use the same procedures. I hope you find that this book provides such procedures.

Although this book should be useful for anyone who analyzes data, it is written within several constraints. First, Stata is the primary computing language because I find Stata to be the best, general-purpose software for data management and statistical analysis. Although nearly everything I do with Stata can be done in other software, I do not include examples from other packages. Second, most examples use data from the social sciences, because that is the field in which I work. The principles I discuss, however, apply broadly to other fields. Finally, I work primarily in Windows. This is not because I think Windows is a better operating system than Mac or Linux, but because Windows is the primary operating system where I work. Just about everything I suggest works equally well in other operating systems, and I have tried to note when there are differences.

I want to thank the many people who commented on drafts or answered questions about some aspect of workflow. I particularly thank Tait Runfeldt Medina, Curtis Child, Nadine Reibling, and Shawna L. Rohrman whose detailed comments greatly improved the book. I also thank Alan Acock, Myron Gutmann, Patricia McManus, Jack Thomas, Leah VanWey, Rich Watson, Terry White, and Rich Williams for talking with me about workflow. Many people at StataCorp helped in many ways. I particularly want to thank Lisa Gilmore for producing the book, Jennifer Neve for editing, and Annette Fett for designing the cover. David M. Drukker at StataCorp answered many of my questions. His feedback made it a better book and his friendship made it more fun to write.

Some of the material in this book grew out of research funded by NIH Grant Number R01TW006374 from the Fogarty International Center, the National Institute of Mental Health, and the Office of Behavioral and Social Science Research to Indiana University–Bloomington. Other work was supported by an anonymous foundation and The Bayer Group. I gratefully acknowledge support provided by the College of Arts and Sciences at Indiana University.

Without the unintended encouragement from my dear friend Fred, I would not have started the book. Without the support of my dear wife Valerie, I would not have completed it. Long overdue, this book is dedicated to her.

Bloomington, Indiana
October 2008

Scott Long