

# Preface

This book was written with a particular reader in mind. This reader is learning social statistics and needs to learn Stata but has no prior experience with other statistical software packages. When I learned Stata, I found there were no books written explicitly for this type of reader. There are certainly excellent books on Stata, but they assume extensive prior experience with other packages, such as SAS or IBM SPSS Statistics; they also assume a fairly advanced working knowledge of statistics. These books moved quickly to advanced topics and left my intended reader in the dust. Readers who have more background in statistical software and statistics will be able to read chapters quickly and even skip sections. The goal is to move the true beginner to a level of competence using Stata.

With this target reader in mind, I make far more use of the Stata interface than do any other books about Stata. Advanced users may not see the value in using the interface, and the more people learn about Stata, the less they will rely on the interface. Also, even when you are using the interface, it is still important to save a record of the sequence of commands you run. Although I rely on the commands much more than the dialog boxes in the interface in my own work, I still find value in the interface. The dialog boxes in the interface include many options that I might not have known or might have forgotten.

To illustrate the interface as well as graphics, I have included more than 80 figures, many of which show dialog boxes. I present many tables and extensive Stata “results” as they appear on the screen. I interpret these results substantively in the belief that beginning Stata users need to learn more than just how to produce the results—users also need to be able to interpret them.

I have tried to use real data. There are a few examples where it is much easier to illustrate a point with hypothetical data, but for the most part, I use data that are in the public domain. For example, I use the General Social Surveys for 2002 and 2006 in many chapters, as well as the National Survey of Youth, 1997. I have simplified the files by dropping many of the variables in the original datasets, but I have kept all the observations. I have tried to use examples from several social-science fields, and I have included a few extra variables in several datasets so that instructors, as well as readers, can make additional examples and exercises that are tailored to their disciplines. People who are used to working with statistics books that have contrived data with just a few observations, presumably so work can be done by hand, may be surprised to see more than 1,000 observations in this book’s datasets. Working with these files provides better

experience for other real-world data analysis. If you have your own data and the dataset has a variety of variables, you may want to use your data instead of the data provided with this book.

The exercises use the same datasets as the rest of the book. Several of the exercises require some data management prior to fitting a model because I believe that learning data management requires practice and cannot be isolated in a single chapter or single set of exercises.

This book takes the student through much of what is done in introductory and intermediate statistics courses. It covers descriptive statistics, charts, graphs, tests of significance for simple tables, tests for one and two variables, correlation and regression, analysis of variance, multiple regression, logistic regression, reliability, and factor analysis. There are chapters on constructing scales to measure variables and on using multiple imputation for working with missing values.

By combining this coverage with an introduction to creating and managing a dataset, the book will prepare students to go even further on their own or with additional resources. More advanced statistical analysis using Stata is often even simpler from a programming point of view than what we will cover here. If an intermediate course goes beyond what we do with logistic regression to multinomial logistic regression, for example, the programming is simple enough. The `logit` command can simply be replaced with the `mlogit` command. The added complexity of these advanced statistics is the statistics themselves and not the Stata commands that implement them. Therefore, although more advanced statistics are not included in this book, the reader who learns these statistics will be more than able to learn the corresponding Stata commands from the Stata documentation and help system.

I would like to point out the use of punctuation after quotes in this book. While the standard U.S. style of punctuation calls for periods and commas at the end of a quote to always be enclosed within the quotation marks, Stata Press follows a style typically used in mathematics books and British literature. In this style, any punctuation mark at the end of a quote is included within the quotation marks only if it is part of the quote. For instance, the pleased Stata user said she thought that Stata was a “very powerful program”. Another user simply said, “I love Stata.”

I assume that the reader is running Stata 12, or a later version, on a Windows-based PC. Stata works equally as well on Mac and on Unix systems. Readers who are running Stata on one of those systems will have to make a few minor adjustments to some of the examples in this book. I will note some Mac-specific differences when they are important. In preparing this book, I have used both a Windows-based PC and a Mac.