Preface

As you may have guessed, this book discusses data analysis, especially data analysis using Stata. We intend for this book to be an introduction to Stata; at the same time, the book also explains, for beginners, the techniques used to analyze data.

Data Analysis Using Stata does not merely discuss Stata commands but demonstrates all the steps of data analysis using practical examples. The examples are related to public issues, such as income differences between men and women, and elections, or to personal issues, such as rent and living conditions. This approach allows us to avoid using social science theory in presenting the examples and to rely on common sense. We want to emphasize that these familiar examples are merely standing in for actual scientific theory, without which data analysis is not possible at all. We have found that this procedure makes it easier to teach the subject and use it across disciplines. Thus this book is equally suitable for biometricians, econometricians, psychometricians, and other "metricians"—in short, for all who are interested in analyzing data.

Our discussion of commands, options, and statistical techniques is in no way exhaustive but is intended to provide a fundamental understanding of Stata. Having read this book and solved the problems in it, the reader should be able to solve all further problems to which Stata is applicable.

We strongly recommend to both beginners and advanced readers that they read the preface and the first chapter (entitled *The first time*) attentively. Both serve as a guide throughout the book. Beginners should read the chapters in order while sitting in front of their computers and trying to reproduce our examples. More-advanced users of Stata may benefit from the extensive index and may discover a useful trick or two when they look up a certain command. They may even throw themselves into programming their own commands. Those who do not (yet) have access to Stata are invited to read the chapters that focus on data analysis, to enjoy them, and maybe to translate one or another hint (for example, about diagnostics) into the language of the statistical package to which they do have access.

Structure

The first time (chapter 1) shows what a typical session of analyzing data could look like. To beginners, this chapter conveys a sense of Stata and explains some basic concepts such as variables, observations, and missing values. To advanced users who already have experience in other statistical packages, this chapter offers a quick entry into Stata.

Advanced users will find within this chapter many cross-references, which can therefore be viewed as an extended table of contents. The rest of the book is divided into three parts, described below.

Chapters 2–6 serve as an introduction to the basic tools of Stata. Throughout the subsequent chapters, these tools are used extensively. It is not possible to portray the basic Stata tools, however, without using some of the statistical techniques explained in the second part of the book. The techniques described in chapter 6 may not seem useful until you begin working with your own results, so you may want to skim chapter 6 now and read it more carefully when you need it.

Throughout chapters 7–10, we show examples of data analysis. In chapter 7, we present techniques for describing and comparing distributions. Chapter 8 covers statistical inference and explains whether and how one can transfer judgments made from a statistic obtained in a dataset to something that is more than just the dataset. Chapter 9 introduces linear regression using Stata. It explains in general terms the technique itself and shows how to run a regression analysis using an example file. Afterward, we discuss how to test the statistical assumptions of the model. We conclude the chapter with a discussion of sophisticated regression models and a quick overview of further techniques. Chapter 10, in which we describe regression models for categorical dependent variables, is structured in the same way as the previous chapter to emphasize the similarity between these techniques.

Chapters 11–13 deal with more-advanced Stata topics that beginners may not need. In chapter 11, we explain how to read and write files that are not in the Stata format. At the beginning of chapter 12, we introduce some special tools to aid in writing do-files. You can use these tools to create your own Stata commands and then store them as ado-files, which are explained in the second part of the chapter. It is easy to write Stata commands, so many users have created a wide range of additional Stata commands that can be downloaded from the Internet. In chapter 13, we discuss these user-written commands and other resources.

Using this book: Materials and hints

The only way to learn how to analyze data is to do it. To help you learn by doing, we have provided data files (available on the Internet) that you can use with the commands we discuss in this book. You can access these files from within Stata or by downloading a zip archive.

Please do not hesitate to contact us if you have any trouble obtaining these data files and do-files.[1]

1. The data we provide and all commands we introduce assume that you use Stata 12 or higher. Please contact us if you have an older version of Stata.

- If the machine you are using to run Stata is connected to the Internet, you can download the files from within Stata. To do this, type the following commands in the Stata Command window (see the beginning of chapter 1 for information about using Stata commands).

```
. mkdir c:\data\kk3
. cd c:\data\kk3
. net from http://www.stata-press.com/data/kk3/
. net get data
```

  These commands will install the files needed for all chapters except section 11.4. Readers of this section will need an additional data package. You can download these files now or later on by typing

```
. mkdir c:\data\kk3\kksoep
. cd c:\data\kk3\kksoep
. net from http://www.stata-press.com/data/kk3/
. net get kksoep
. cd ..
```

  If you are using a Mac or Unix system, substitute a suitable directory name in the first two commands, respectively.

- The files are also stored as a zip archive, which you can download by pointing your browser to http://www.stata-press.com/data/kk3/kk3.zip.

  To extract the file `kk3.zip`, create a new folder: `c:\data\kk3`. Copy `kk3.zip` into this folder. Unzip the file `kk3.zip` using any program that can unzip zip archives. Most computers have such a program already installed; if not, you can get one for free over the Internet.[2] Make sure to preserve the `kksoep` subdirectory contained in the zip file.

Throughout the book, we assume that your current working directory (folder) is the directory where you have stored our files. This is important if you want to reproduce our examples. At the beginning of chapter 1, we will explain how you can find your current working directory. Make sure that you do not replace any file of ours with a modified version of the same file; that is, avoid using the command `save, replace` while working with our files.

We cannot say it too often: the only way to learn how to analyze data is to analyze data yourself. We strongly recommend that you reproduce our examples in Stata as you read this book. A line that is written `in this font` and begins with a period (which itself should not be typed by the user) represents a Stata command, and we encourage you to enter that command in Stata. Typing the commands and seeing the results or graphs will help you better understand the text, because we sometimes omit output to save space.

As you follow along with our examples, you must type all commands that are shown, because they build on each other within a chapter. Some commands will only work if

2. For example, "pkzip" is free for private use, developed by the company PKWARE. You can find it at http://pkzip.en.softonic.com/.

you have entered the previous commands. If you do not have time to work through a whole chapter at once, you can type the command

```
. save mydata, replace
```

before you exit Stata. When you get back to your work later, type

```
. use mydata
```

and you will be able to continue where you left off.

The exercises at the end of each chapter use either data from our data package or data used in the Stata manuals. StataCorp provides these datasets online.[3] They can be used within Stata by typing the command `webuse` *filename*. However, this command assumes that your computer is connected to the Internet; if it is not, you have to download the respective files manually from a different computer.

This book contains many graphs, which are almost always generated with Stata. In most cases, the Stata command that generates the graph is printed above the graph, but the more complicated graphs were produced by a Stata do-file. We have included all of these do-files in our file package so that you can study these files if you want to produce a similar graph (the name of the do-file needed for each graph is given in a footnote under the graph).

If you do not understand our explanation of a particular Stata command or just want to learn more about it, use the Stata `help` command, which we explain in chapter 1. Or you can look in the Stata manuals, which are available in printed form and as PDF files. When we refer to the manuals, [R] **summarize**, for example, refers to the entry describing the `summarize` command in the *Stata Base Reference Manual*. [U] **18 Programming Stata** refers to chapter 18 of the *Stata User's Guide*. When you see a reference like these, you can use Stata's online help (see section 1.3.16) to get information on that keyword.

Teaching with this manual

We have found this book to be useful for introductory courses in data analysis, as well as for courses on regression and on the analysis of categorical data. We have used it in courses at universities in Germany and the United States. When developing your own course, you might find it helpful to use the following outline of a course of lectures of 90 minutes each, held in a computer lab.

To teach an introductory course in data analysis using Stata, we recommend that you begin with chapter 1, which is designed to be an introductory lecture of roughly 1.5 hours. You can give this first lecture interactively, asking the students substantive questions about the income difference between men and women. You can then answer them by entering Stata commands, explaining the commands as you go. Usually, the students

3. They are available at http://www.stata-press.com/data/r12/.

name the independent variables used to examine the stability of the income difference between men and women. Thus you can do a stepwise analysis as a question-and-answer game. At the end of the first lecture, the students should save their commands in a log file. As a homework assignment, they should produce a commented do-file (it might be helpful to provide them with a template of a do-file).

The next two lectures should work with chapters 3–5 and can be taught a bit more conventionally than the introduction. It will be clear that your students will need to learn the *language* of a program first. These two lectures need not be taught interactively but can be delivered section by section without interruption. At the end of each section, give the students time to retype the commands and ask questions. If time is limited, you can skip over sections 3.3 and 5.7. You should, however, make time for a detailed discussion of sections 5.1.4 and 5.1.5 and the examples in them; both sections contain concepts that will be unfamiliar to the student but are very powerful tools for users of Stata.

One additional lecture should suffice for an overview of the commands and some interactive practice in the graphs chapter (chapter 6).

Two lectures can be scheduled for chapter 7. One example for a set of exercises to go along with this chapter is given by Donald Bentley and is described on the webpage http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html. The necessary files are included in our file package.

A reasonable discussion of statistical inference will take two lectures. The material provided in chapter 8 shows necessary elements for simulations, which allows for a hands-on discussion of sampling distributions. The section on multiple imputation can be skipped in introductory courses.

Three lectures should be scheduled for chapter 9. According to our experience, even with an introductory class, you can cover sections 9.1, 9.2, and 9.3 in one lecture each. We recommend that you let the students calculate the regressions of the Anscombe data (see page 279) as a homework assignment or an in-class activity before you start the lecture on regression diagnostics.

We recommend that toward the end of the course, you spend two lectures on chapter 11 introducing data entry, management, and the like, before you end the class with chapter 13, which will point the students to further Stata resources.

Many of the instructional ideas we developed for our book have found their way into the small computing lab sessions run at the UCLA Department of Statistics. The resources provided there are useful complements to our book when used for introductory statistics classes. More information can be found at http://www.stat.ucla.edu/labs/, including labs for older versions of Stata.

In addition to using this book for a general introduction to data analysis, you can use it to develop a course on regression analysis (chapter 9) or categorical data analysis (chapter 10). As with the introductory courses, it is helpful to begin with chapter 1, which gives a good overview of working with Stata and solving problems using Stata's online help. Chapter 13 makes a good summary for the last session of either course.