

Preface

Data Analysis Using Stata goes beyond simply teaching Stata commands—it guides you through all the essential steps of data analysis using practical, real-world examples. We explore both public topics, like income differences between men and women and elections, and more personal issues such as rent and living conditions. By focusing on common sense rather than diving into social science theories, we make the material accessible to a wider audience. Of course, these examples are stand-ins for the scientific theory behind data analysis, but this approach helps make the learning process smoother across different disciplines.

Whether you are a biometrician, econometrician, psychometrician, or some other kind of “metrician”, this book is meant for anyone with an interest in data analysis. While we do not aim to cover every Stata command or statistical technique, we focus on providing a solid understanding of the essentials. By the end, you’ll be equipped to tackle a wide range of data analysis challenges—not just in Stata but also in other languages.

We recommend that both beginners and more experienced readers start with the preface and the first chapter, “The first time”. These sections offer helpful guidance for navigating the book. Beginners will benefit from working through the chapters sequentially, trying out the examples on their computers as they go. More advanced users can explore the index for quick tips or even challenge themselves to write their own commands. And for those who do not have access to Stata, or prefer a different software package, the data analysis methods discussed here can easily be adapted to other platforms.

Structure

Chapter 1, “The first time”, 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. They will find many cross-references within this chapter, which can therefore be viewed as an extended table of contents.

The rest of the book is divided into three parts:

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 more than just the dataset. Chapter 9 introduces linear regression. It explains in general terms the technique itself, how to apply it in Stata, and why we need it. Afterward, we discuss how to test the statistical assumptions of the model. We conclude the chapter with a discussion of sophisticated regression models for causal inference. 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 and 12 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.

Changes in this edition

It has been quite some time since the third edition. Stata has undergone significant transformations since then, continuously evolving to meet the demands of modern data analysis. Each new version has introduced tools and functionalities aimed at improving user experience and analytical capabilities. We aim to reflect these in this new fourth edition. We are excited to announce the addition of a new coauthor, Caro Haensch, to this edition. Her expertise and fresh perspective have enriched the content, ensuring it is comprehensive and up to date with the latest developments in data analysis.

In this edition, we discuss several new features that enhance the practical application of data analysis:

Integration of Python. We illustrate how to incorporate Python scripts into your workflow for data analysis tasks. We show the interface between Stata and Python and discuss classic cases such as the integration of advanced machine learning methods, interactive graphs, and web scraping.

Large language models for coding. We explore the changes and pitfalls associated with using large language models for coding, providing practical advice on their effective use and understanding their limitations.

Reproducibility. We have added a section on dynamic documents (**dyndoc**) to enhance the reproducibility and documentation of data analysis workflows.

Missing values and Unicode. We now offer a more systematic approach to handling missing values and have included a new section on Unicode in the chapter on reading and writing data.

Causal inference, parsimonious description, and prediction. The regression chapter is restructured and extended to differentiate between different cases, such as causal inference, parsimonious description, and prediction.

Restructured book sections. The section on additional features is now consolidated in chapter 1, covering resources and information, updating procedures, and providing additional tools like the *Stata Journal* and Statistical Software Components Archive ado-files.

General updates. We updated and improved various sections where necessary.

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.

- 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/daus4
. cd c:/data/daus4
. net from https://www.stata-press.com/data/daus4/
. net install data
. net get data
```

These commands will install the files needed for the entire book except section 11.5. Readers of this section will need another large data package, namely, an extract of the database of the German Socio-Economic Panel (GSOEP). Because there are many files, you must download them in two steps. You can do that now, or later on, with

```
. mkdir c:/data/daus4/kkhsoep
. cd c:/data/daus4/kkhsoep
. net from https://www.stata-press.com/data/daus4/
. net get kkhsoep1
. net get kkhsoep2
. cd ..
```

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

- The files are also stored as a zip archive, which you can download by pointing your browser to <https://www.stata-press.com/data/daus4/daus4.zip>.

To extract the file `daus4.zip`, create a new folder, `c:/data/daus4`, and copy `daus4.zip` into this folder. Unzip the file `daus4.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 using the internet. Make sure to preserve the `kkhsoep` 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 check where your current working directory is. Make sure that you do not replace our files with a modified version of the same file; 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 shown because they build on each other within a chapter. Some commands will work only if 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.¹ They can be used within Stata using the command `webuse filename`. However, this command assumes that your computer is connected to the internet. Otherwise, you have to download the respective files manually from a different computer.

This book contains a lot of graphs, which are almost always generated with Stata. In most cases, the Stata command that generates the graph is printed above the graph,

1. <https://www.stata-press.com/data/r19/>.

but the more complicated graphs were produced by a Stata do-file. We have included all these do-files in our file package so that you can study these files in case 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 PDF. We refer to the manuals in the same way as the Stata **help** command: [R] **summarize** 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 to a keyword in one of the *Reference* manuals, you can use the online help (see section 1.3.15) to get information on that keyword.

Teaching with this book

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 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 session, the students should save their commands in a log file, and 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, whereas you should make time for a detailed discussion of sections 5.4.3 and 5.4.4 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 session should suffice for an overview of the commands and some interactive practice in the graphs chapter.

Two sessions can be scheduled for chapter 7. A reasonable discussion of statistical inference will take two sessions. The material provided in chapter 8 shows the 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 sessions 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 session each. We recommend that you let the students calculate the regressions of the Anscombe data (see page 324) as a homework assignment or an in-class activity before you start the session on regression diagnostics.

We recommend that toward the end of the course, you spend two sessions on chapter 11, introducing data entry, management, and the like.

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 the online help.