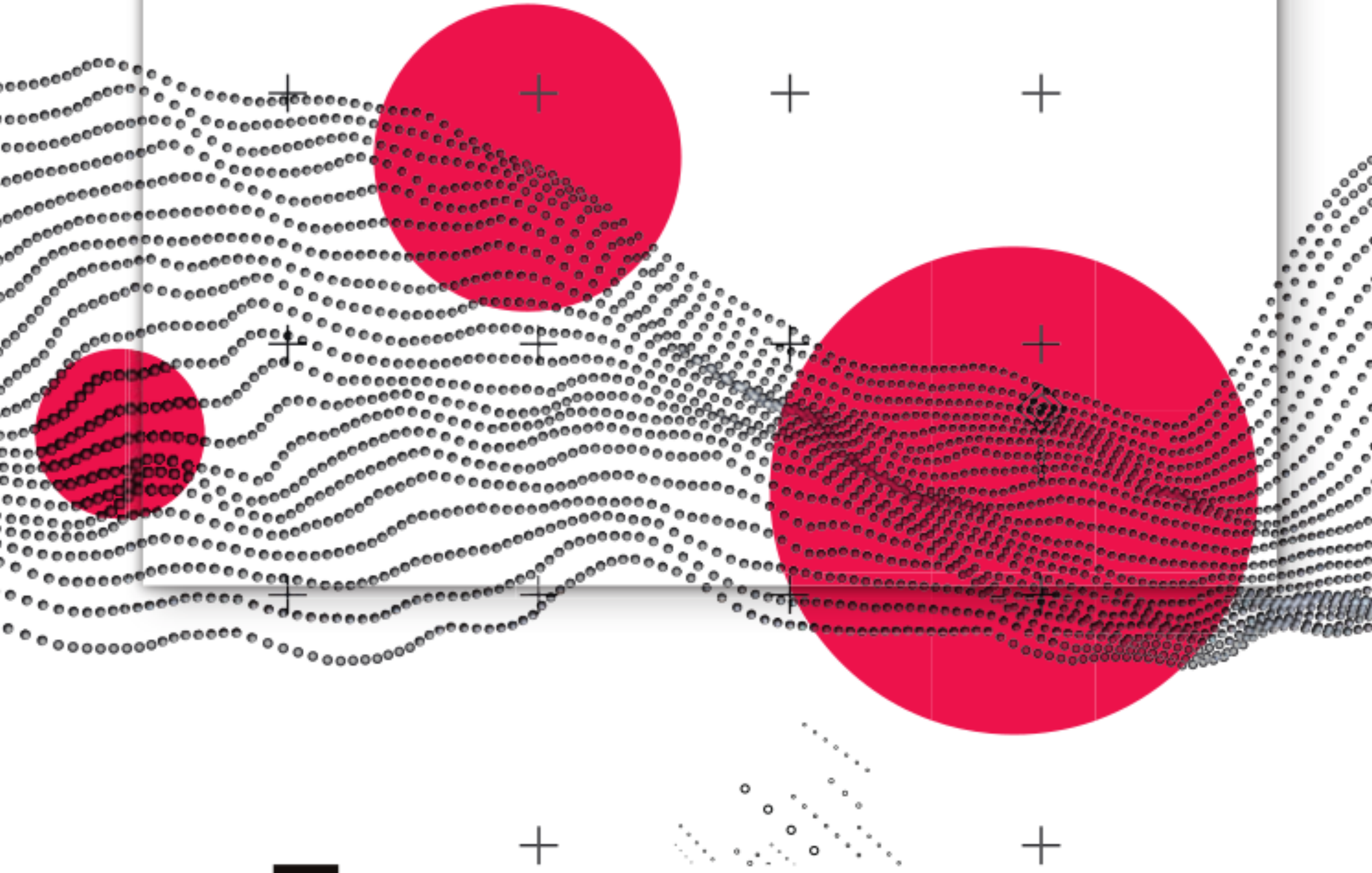
DE GRUYTER
OLDENBOURG

Edgar J. Treischl

PRACTICE R

AN INTERACTIVE TEXTBOOK



Edgar J. Treischl

Practice R

An interactive textbook

DE GRUYTER

Author
Dr. Edgar J. Treischl
Researcher at the Chair of Empirical Economic Sociology
School of Business and Economics
Friedrich-Alexander-University Erlangen-Nuremberg
Findelgasse 7/9
90402 Nuremberg
Germany

ISBN 978-3-11-070496-9
e-ISBN (PDF) 978-3-11-070497-6
e-ISBN (EPUB) 978-3-11-070508-9

Library of Congress Control Number: 2022949290

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the internet at http://dnb.dnb.de.

Cover image: StudioM1 / iStock / Getty Images Plus
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

Part I: **The first steps**

Part II: **The basics**

Part III: **Beyond the basics**

Part I: **The first steps**

1 Introduction

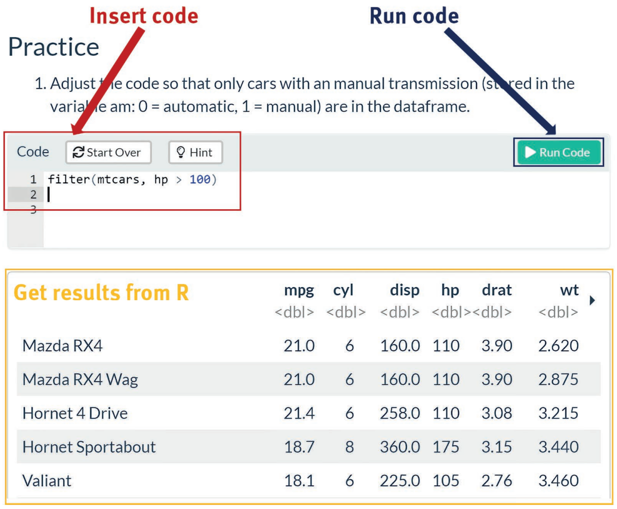
R is a programming language and a powerful tool to analyze data, but R has a lot more to offer than statistics. To mention just a few options, R has many capabilities to visualize data, to collect data (e.g., from a website), or even to create interactive dashboards. From this perspective it is no wonder why R has a huge fan base. Unfortunately, learning R can be though. People who struggle may say that the data handling is complicated, some complain that R lacks a graphical interface, and probably all agree that beginners face a rather steep learning curve. Regardless of our perception, the best way to learn R is by means of practice. For this reason, this book introduces R, focuses on the most important steps for applied empirical research, and explains how to use R in practice. After reading and working on the materials in this book, you will be able to *prepare* and *analyze* data, make *visualizations*, and *communicate* key research insights.

Who should read this book? Overall, the book introduces R and is written for people with no prior knowledge about it. However, Practice R is a textbook for the social sciences, and it is assumed that the reader has prior knowledge in statistics and quantitative methods. Practice R might not be the first choice if you have yet to learn what a *standard deviation*, *Pearson's r*, or a *t-test* is. The same applies for topics of quantitative empirical research. I presume that the reader has knowledge about research designs, is familiar with the difference between cross-sectional and longitudinal data, and other aspects that intermingle with statistics, seeing that quantitative methods are a substantial part of the social science curriculum. Of course, this does not mean that only (social science) students can profit from reading the book. A diverse audience – holding the assumed prior knowledge – may use Practice R to become a proficient R user.

To support you, the book is accompanied by an R package. An R package is a software add-on and extends the capabilities of R. In our case, the `PracticeR` package gives you access to tutorials to practice the discussed content, it provides the code of this book, and also further materials (e.g., a template to create reports) that are supposed to boost your skills. We will learn how to install R packages in the next chapter, but keep in mind that all materials of the book become available once the `PracticeR` package is installed.

Let me outline the idea of the tutorials and how they are related to the content of the book. The tutorials summarize the content and aim to familiarize you with the core concepts. The interactive tutorials are integrated in R and run on your computer. By clicking on the Run button, R code will be executed, and the tutorial shows the results. Don't mind if something goes wrong, you can reload and start over at the click of a button. As an illustration, Figure 1.1 shows a screenshot of the Basics of Data Manipulation (Chapter 4) tutorial. It summarizes how to filter, arrange, and select data. Irrespective of the topic, each tutorial probes you to apply the discussed content. The exercises in the tutorials aim to increase your coding skills and they are ordered as-

https://doi.org/10.1515/9783110704976-001



Practice

1. Adjust the code so that only cars with an manual transmission (stored in the variable am: 0 = automatic, 1 = manual) are in the dataframe.

```
filter(mtcars, hp > 100)
```

```
                     mpg   cyl   disp   hp  drat     wt
                   <dbl> <dbl>  <dbl><dbl><dbl>  <dbl>
Mazda RX4           21.0     6  160.0  110  3.90  2.620
Mazda RX4 Wag       21.0     6  160.0  110  3.90  2.875
Hornet 4 Drive      21.4     6  258.0  110  3.08  3.215
Hornet Sportabout   18.7     8  360.0  175  3.15  3.440
Valiant             18.1     6  225.0  105  2.76  3.460
```

Fig. 1.1: Example tutorial

cendingly by difficulty. Sometimes I'll ask you to adjust the R code, which gives you a better understanding of how the code works. In most instances I will challenge you with typical data analyzing problems. In the more advanced steps, you are supposed to transfer the discussed content to a similar or a new concept. Don't worry, hints are provided to solve the exercises and the tutorials include the solutions. Now that the scope is set, we can divulge the content of Practice R.

The content

Part I lays the foundation and outlines the first steps to work with R:

- Chapter 2 introduces R and RStudio, which is an integrated development environment to work with R. The chapter contains the most important steps to understand how R behaves and outlines in depth how RStudio substantially helps us to increase our R skills. We install both software packages and we discover some of the cool features of RStudio. Next, I give a concise introduction of base R – the programming language – which is essential for subsequent steps. Moreover, the chapter makes you familiar with data types and structures.
- In Chapter 3 we start to explore data. We examine variables, we calculate and visualize descriptive statistics, and we explore how variables are related. We estimate the correlation between two variables, visualize the effect, and interpret the effect size. Data exploration is crucial when we start to work with data. For this reason, this chapter also highlights packages and ways to get a quick overview of new and unfamiliar data. For example, some packages implement graphs to examine several variables at once; others can generate a PDF report with summary statistics

for all variables of a particular data set. Thus, we explore variables, and we get in touch with packages that help us to discover unfamiliar data.

- Chapter 4 focuses on data manipulation steps and introduces the `dplyr` package (Wickham, François, et al., 2022). The latter is the Swiss pocketknife for manipulating data. I introduce the main functions of the package and we will focus on typical steps to prepare data for an analysis. Before we can dive into this topic in the second part, we should take one step back. The last part of this chapter highlights strategies to increase the workflow and, consequently, the efficiency of our work. For example, you may wonder how much R code you need to remember to become an efficient R user. The last section outlines in detail why there is no need to memorize code and introduces strategies to handle (complicated) code.

Part II introduces the basics to analyze data, visualize results, and create reports:

- Chapter 5 outlines the data preparation steps required before we can start to analyze data. We learn how to import data and how to cope with problems that may occur. Depending on the data, the import step may induce errors, but the same may apply during the data cleaning steps, and we should consider the concerns of missing (and implausible) values. Finally, I introduce the main functions from the `forcats` package (Wickham, 2022a). The package is made for *categorical variables* and is a good supplement to our data manipulation skills since categorical variables are often used in social sciences.
- We analyze data in Chapter 6. There is a broad range of possibilities to analyze data with R, however, we apply a linear regression analysis, because it is the workhorse of social science research. First, I give an non-technical introduction for people with a different educational background. Next, we run an example analysis that we will improve step by step. We learn how to develop the model, we examine interaction effects, and we compare the performance of the estimated models. To compare models and to examine the assumption of a linear regression analysis, we also focus on visualization techniques.
- To visualize research findings, Chapter 7 concentrates on the `ggplot2` package (Wickham, Chang, et al., 2022). The package can be quite demanding in the beginning, but we will learn that creating a graph without much customization is far from rocket science. We first focus on typical steps to create and adjust a graph (e.g., adjust a title). Next, we increase the theoretical knowledge by exploring how `ggplot2` works behind the curtain. Ultimately, there are a lot of packages that extend the possibilities of `ggplot2`. The last section highlights some of these possibilities.
- Chapter 8 focuses on reporting. After the analysis and the visualization step, we need to summarize the findings in a document and the `rmarkdown` package makes it possible to create text documents with R (Allaire, Xie, McPherson, et al., 2022). An `rmarkdown` file contains text, graphs, or tables, just like any other text document. However, it is code-based and also contains output from R. Thus, we create tables and graphs with R and include them in the `rmarkdown` document. Using code to

create the report increases the reproducibility of the work and we avoid introducing errors, because we eliminated the need to transfer output from R into a word processing software.

Part III completes the basics and focuses on topics that – at first glance – seem less related to applied empirical research, but that will add to your skill set:

- Chapter 9 introduces Git, a version control system for code, and GitHub, a host for Git-based projects. Think of Git/GitHub as a sort of cloud for code. Suppose you changed a code, but you made a mistake. A version control system lets us travel back in time to find out where the error occurred. GitHub marks changes of the code and forces us to explain – in a few words – what happens to the code when we make an update. GitHub has more advantages, but I guess the example makes clear that a version control system is very valuable if you work with code on a regular basis. Chapter 9 gives a short introduction, we learn the basics to send (receive) code to (from) GitHub, and we connect RStudio with your GitHub account.
- Chapter 10 outlines the advantages of dynamic reports and highlights that whenever possible we are not supposed to repeat ourselves, instead we can automate the boring manual stuff. Say we made a report with R, but the data for the report gets an update. There is no need to manually re-estimate the results, re-create graphs, or tables – create a dynamic report and let R recreate and update the document. Chapter 10 introduces dynamic reports and discusses further steps to automate the reporting process (e.g., to send reports automatically via email).
- Chapter 11 demonstrates that we can collect data with R. Consider you work with a data set that lacks an important variable. Maybe you find this information on a website or in a PDF report; or suppose you want to retrieve data from a social media platform – your R skills help you in all those instances. The last chapter highlights the possibilities of collecting data and underlines the main steps to retrieve data from a PDF file, a website, and a web server.
- Finally, Chapter 12 outlines possible next steps and demonstrates that there are many cool packages and features to discover. This chapter introduces topics, packages, and frameworks, that would otherwise not find a place in an introductory book and we explore the next steps in connection to data preparation, analysis, visualization, and reporting.

Practice R contains only a selection of the possibilities that R offers. Maybe you have to prepare or analyze data, beyond what is covered in the book. Fortunately, R has a large and helpful community and you can find a lot of information on the web. This book introduces R and focuses on the main aspects of applied research, which is why I skip some of the more sophisticated topics. Using info boxes, the book covers additional topics and guidelines on where to find more information. Irrespective of the content, Practice R was written with several guiding principles in mind.