R for Researchers: An Introduction

Prospectus

Tyson S. Barrett, PhD

Project Description

Given the rise of the R programming language in research[1]—including its many benefits to reproducible research—many researchers are interested in benefiting from adopting its use in their analysis workflow. In this book, R will be introduced to individuals not assumed to have any coding experience, approaching R as a valuable tool that any researcher can learn. It will focus on the approaches that provide the most intuitive inputs and outputs, particularly relevant to researchers in the human sciences (e.g. health, behavioral, educational, sociological, and psychological sciences).

Although there are many R books, I have seen most of my students are intimidated by three aspects that nearly all R books have:

1. Most R books are lengthy, providing in-depth discussion about the topics covering a broad range of approaches. If researchers are trying to understand R while also learning statistics, a 300+ page book can be overwhelming.
2. Very few R books are oriented towards researchers the human sciences such as health, behavioral, educational, and psychological sciences. This generally means that these others books discuss approaches not relevant to the researcher and lack discussion on appropriate methods that are relevent to their research.
3. The most intuitive approaches are of fairly recent development. Thus, many R books are outdated and do not provide this intuition that can reduce the cognitive burden of learning the language.

This book will address each aspect. First, this book will *not* be a lengthy user's manual. Instead, it will introduce and demonstrate the important concepts in R necessary for the student to start using it right away in the course. Second, this book specifically addresses the use of R in researching human beings. Although the various fields in the human sciences are different in substantive research, the data situations encountered in these fields are very similar.[2] Third, this book uses the intuitive code-grammar of the `tidyverse` group of packages.[3] This makes the code more readable, as each line of code has a subject, a verb, and possibly adjectives that make it more like reading regular English.

Each chapter will take on a general topic, demonstrate the use of the approaches relevant to the topic, and will provide a zip file that can be downloaded from an online source that provides practice data and code. In addition, I plan on providing online videos walking through the applications.

Unique Value Proposition

The R programming environment has become one of the most widely used statistical tools, in both academia and industry. Because it is a programming language, many capable researchers in health, behavioral, educational, and psychological sciences avoid it altogether. However, recent developments in R have made the tool far more intuitive. But this information is not clear to researchers because many of the available resources to learn R come from a computer science, data science, or biological science background. Further, many of these resources are quite broad with several hundreds of pages. Many of the items discussed are unnecessary, and are potentially overwhelming. This book, instead, addresses the up-to-date uses to get started enough to become literate in the R language without discussing *over* discussing the possible uses of R.

This style of book is important as many undergraduate and graduate programs do not specifically teach R as part of their graduate studies. Instead, many classes introduce it as part of statistics courses. As such, the

[1] see https://forwards.github.io/blog/2017/01/13/mapping-users/, https://doi.org/10.1002/ecs2.2567, and https://journal.r-project.org/archive/2017-2/forwards.pdf

[2] I've seen this firsthand as a data science and statistics consultant at Utah State University. Overall, the data collected is very similar across the fields.

[3] More information at tidyverse.org.

book will be a great tool to use in conjunction with other statistical texts. It will not be overwhelming but, rather, will provide the background in R necessary to work with their own data and do the analyses needed.

Length

The projected length is roughly 45,000 words and 120 book pages; long enough for the information to be taught, but not too much to overwhelm the student.

Pedagogy

Three major approaches to pedagogy will be emphasized: 1) researchers need to see several code examples to become literate in R, 2) researchers need to apply the code to solidify the learning, and 3) visuals help make the abstract functions and data cleaning more concrete.

First, this book contains many code input and output examples. It walks through the pieces of the input that lead to the output using real-life examples from one of the NHANES (National Health and Nutrition Examination Survey) data sets. This helps the code become more intuitive and natural to the researchers.

Second, each chapter ends with an *Apply It* section. These sections provide chapter-specific assignments that guide the researcher through each main topic covered in the chapter, allowing the researcher to apply the material on their own computer. These applications are designed to provide a high likelihood of success to give the researchers experience completing the required tasks. Common areas of error will be discussed in the assignment, providing an explanation of what the error means and how to avoid it. I also want to add "extra" tasks that are more challenging but provide an opportunity to learn the topics at a deeper level.[4]

Third, several visuals will be used to demonstrate the use of various functions and what those functions are actually doing to the data. This is essential because many data transformations and manipulations are quite abstract and can otherwise become confusing. It also tends to help researchers understand how different aspects of the functions control the data manipulations. I intend on making several more for the published version of the book than is currently in the draft version. Beyond the specifics taught by the visuals, it also helps break up the text of the book to keep readers more engaged.

Fourth, review questions are included at the end of the chapters. The review questions include straightforward recall of definitions, meanings, and concepts; other questions are designed to create critical thinking, using the information from the chapter to solve a novel problem.

In addition to these main approaches, the book also uses text boxes (although not fully implemented in the draft) to highlight common errors or areas of confusion relating to the topic. For example, when it comes to importing data, confusion regarding when to use quotes is common. A box would highlight where to use quotes and why. Only a few of these per chapter is planned.

Market

This book is useful in research methods and statistics courses for researchers in the human sciences. The draft version of the book has been used to teach undergraduate, graduate, post-doctoral, and senior researchers in a dedicated R course as well as graduate level statistical courses. A general introduction to quantitative data (in spreadsheets or other forms) is helpful before beginning the book but is not absolutely essential. Also, a general introduction to statistics (t-tests, ANOVA, regression, etc.) can help make some of the middle chapters more meaningful.

I've found that the students most commonly in these courses are those that have little or no experience coding, know some statistical approaches at an applied level, have little interest in general data science, and usually want to be more independent and reliable in their quantitative research. The skills taught in this book allow researchers with little desire to program to work with data independently.

There has also been some stated interest in a spanish version of the book for researchers in Central America.

[4] With the planned additions to the applications, these assignments could become graded assignments for a dedicated R course.

Competition

Several R books exist that can be used to directly teach R or have R taught as part of a statistics course:

- *Discovering Statistics with R* by Andy Field, Jeremy Miles, and Zoë Field, 2012, SAGE Publishing. This is a full text introduction to statistics while showing R examples. It is a great resource but is not directly tied to teaching how to do data work with R. Rather, its focus is on the statistics. This can be useful if an instructor wants to teach from this specific book but can be problematic otherwise.
- *R Cookbook* by Paul Teetor, 2011, O'Reilly Media. This is a full text introduction to using R, using a "recipe" approach. It is lengthy and broadly applied, making it a good introduction to a general R practitioner. However the breadth and length make it difficult for someone to get going quickly and easily.
- *R for Data Science* by Hadley Wickham and Garrett Grolemund, 2017, O'Reilly Media. This book addresses the use of R in data science. It is another great resource but its audience is too broad for it to discuss examples and approaches specific to public health, behavioral and psychological sciences, or education.
- *R for SPSS and SAS Users* by Robert Muenchen, 2011, Springer. This manual thoroughly explains R code in reference to SPSS and SAS. It is, in my experience, overwhelming for students to actually use. It can sometimes be a good resource to look up a specific analysis that the student has done in SPSS or SAS. However, it is somewhat outdated and not visually appealing.
- *Modern Dive* by Chester Ismay and Albert Kim. Much like *R for Data Science*, this book provides an introduction to working with data in R. It, however, instructs on approaches less commonly used in the public health, behavioral and psychological sciences, and education. Although great approaches, this can be distracting to students. In my experience, very few courses teach using the approaches taught in *Modern Dive*.

These five are commonly used in applied statistics courses wherein the instructor wants to either emphasize the use of R or, at the very least, allow its use.

Schedule

I have a first draft prepared, and is available online at: tysonbarrett.com/Rstats. It still, obviously, needs some work but the framework is there. I'd like to get it ready for publication in the next year or so.

Credentials

I have a PhD in Quantitative Psychology, do research regarding methodological advancement, develop R-based research tools, and teach statistics courses each semester for graduate students in public health, behavioral, educational, and psychological sciences. I also teach an introductory R course each fall and spring semester, which currently uses the draft of this book. Across all of my courses, between 60 - 100 students are using the draft book each semester.

Annotated Table of Contents

The book is divided into 3 parts: 1) Introduction to the basics of working and understanding data, 2) Analyses with data, and 3) More advanced data cleaning and understanding. Each chapter ends with a link to practice data and code that walks the reader through using the approaches discussed.

Part I

Chapter 1: The Basics

- Objects
- Data Types
- Functions
- Importing Data

- Saving Data
- Apply It
- Conclusions

This chapter introduces the core concepts of R. This includes the discussion of the various types of objects in R that will be important early on. Because data import is often difficult at first, the chapter also shows basic ways to read in data from different sources. The importing is shown by using the NHANES data set.

Chapter 2: Working with and Cleaning Your Data

- Tidy Methods
- Tidy (Long) Form
- Piping
- Select Variables and Filter Observations
- Mutate Variables
- Grouping and Summarizing
- Reshaping
- Joining (merging)
- Wrap it up
- General Cleaning of the Data Set
- Apply It

This chapter introduces the data manipulation verbs that help data cleaning to be more intuitive. These include *selecting* variables, *filtering* observations, *mutating* variables, among others. These are shown on the NHANES data set imported in Chapter 1.

Chapter 3: Exploring Your Data with Tables and Visuals

- Descriptive Statistics
- Visualizations
- More Advanced Features of `ggplot2`
- Types of Plots
- Color Schemes
- Themes
- Labels and Titles
- Facetting
- Apply It
- Conclusions

This chapter then shows approaches that help the students better understand our cleaned data via summary tables and data visualizations. For the tables, I use the `furniture`, `psych`, and `stargazer` packages. For the visuals, I use `ggplot2` and its extensions. It further covers the flexibility of `ggplot2` by demonstrating the use of different themes, the ability to control font and colors, and the ability to combine plots.

Part II

The chapters in part 2 do not describe any of the analyses in depth. Instead, they show how each method can be done and how assumptions can be checked. These chapters go well with other statistical texts.

Chapter 4: Basic Statistical Analyses

- ANOVA
- Assumptions
- Linear Modeling
- Assumptions

- Comparing Models
- When Assumptions Fail
- Interactions
- Apply It

  Linear models are foundational to statistics in the Health, Behavioral, Educational, and Psychological Sciences. I start with t-tests, ANOVAs, and linear regression. This chapter also shows how to use summary tables to present the results of these models.

Chapter 5: Generalized Linear Models

- Logistic Regression
- Poisson Regression
- Beta Regression
- Apply It

  Another major area of statistics that is used in the fields is Generalized Linear Models, particularly logistic regression. I show how several of these models can be run and checked. Again, I show how the results can be presented reproducibly and succinctly.

Chapter 6: Multilevel Modeling

- GEE
- Mixed Effects
- Apply It
- Conclusions

  This chapter shows how multilevel models can be used in R, showing its use on nested data and generalized to longitudinal data. Both Generalized Estimating Equations and Linear Mixed Effects models are demonstrated.

Chapter 7: Other Modeling Techniques

- Mediation Modeling
- Structural Equation Modeling
- Machine Learning Techniques
- Apply It
- Conclusions

  The other modeling techniques include mediation analysis, structural equation models, and some basic machine learning techniques. This chapter provides some foundations on which a student can start using R to assess these types of models. I would also like to add Directed Acyclic Graphs to the mediation modeling section using `ggdag`.

Part III

Chapter 8: Advanced data manipulation

- Reshaping Your Data
- Repeating Actions (Looping)
- Apply It
- Conclusions

  This chapter extends chapter 2 by discussing reshaping data in more depth, using loops to automate data cleaning and communicating (for loops, apply family, `purrr` package), and writing custom functions. This chapter gives researchers the ability to make R do the work for them with minimal coding.

Chapter 9: Reproducible Workflow with RMarkdown

- R Markdown
- Apply It

  This chapter introduces RMarkdown as a powerful means to produce reproducible reports and articles. It shows how RMarkdown can combine text and R output in a way that reduces errors, increases openness, and can streamline the research reporting process. This chapter uses examples from previous chapters into a full reproducible document.

Chapter 10: Where to go from here

- Common Pitfalls
- Quiz
- Goodbye and Good Luck

  This chapter discusses additional resources that a student can use to continue their learning of R. Given this book is an introduction, it provides resources that can extend their data cleaning abilities, their `ggplot2` abilities, among others. It also includes a practice test that allows the student to review each chapter of the book. The conclusion of this chapter ends the book.