R for Researchers: An Introduction

Prospectus

Tyson S. Barrett, PhD

Table of Contents

Detailed Chapter Outline

The book is divided into 3 parts to help orient the reader to the three main topics of the book: 1) Introduction to the basics of working and understanding data, 2) Statistical analyses with data, and 3) More advanced data cleaning and reproducible workflows. Each chapter ends with a link to practice data and code that walks the reader through using the

approaches discussed. As discussed in other sections, these chapters are oriented towards researchers in the human and behavioral sciences.

Part I

Chapter 1: The Basics

This chapter introduces the core concepts of R, using analogies to the physical world to illustrate how R works. This includes the discussion of the various types of objects in R that will be important early on (e.g. data frames and vectors). This chapter also contains information on working with those important R objects.

After the introduction to R objects, the chapter focuses on basic ways to read in data from different sources, including Microsoft Excel files, SPSS files, Stata files, among others. Because data import can be the first challenge faced by a researcher beginning in R, this is a focus of the first chapter. The importing in this chapter is demonstrated using a publicly-available NHANES (National Health and Nutrition Examination Survey) data set.

Chapter 2: Working with and Cleaning Your Data

This chapter introduces the data manipulation verbs that help data cleaning to be more intuitive. These verbs include *selecting* variables, *filtering* observations, *mutating* variables, and *reshaping* data. Each aspect is shown using the NHANES data set imported in Chapter 1. This chapter also highlights how these verbs work together to allow a researcher to read in their master data file, clean it in such a way that exploratory and confirmatory analyses can be performed on them, and reproduce these steps.

The chapter also begins the emphasis throughout the remainder of the book regarding focusing on the readability of the code. This is done by using pipes from the `tidyverse` packages. It is argued that this makes several aspects of

research easier: replication, error catching, and future use of the code.

Chapter 3: Understanding Your Data

This chapter shows approaches that help the students better understand our cleaned data (as demonstrated in Chapter 2) via summary tables and data visualizations. For the tables, I use the `furniture`, `psych`, and `stargazer` packages that each provide reproducible, publication quality tables with minimal coding.

For the visuals, I use `ggplot2` and its extensions. It shows how this package makes proper data visualization the default and how the framework is comprehensive enough to visualize all common visualization needs. 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 II 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 four chapters will go particularly well with other statistical texts.

Chapter 4: Basic Statistical Analyses

This chapter introduces how to run linear models in R. To allow this chapter to be used in conjunction with other statistical texts, the chapter starts with what most statistical textbooks start with: t-tests, ANOVAs, and linear regression. The chapter shows how to run the models, examine assumptions, and interpret the output. Again, no in-depth discussion of the conceptual aspects of the methods is attempted.

This chapter also shows how to use summary tables to present the results of these models in publishable ways.

Chapter 5: Generalized Linear Models.

Another major area of statistics that is used in research on humans and human behavior is Generalized Linear Models, particularly logistic regression. In this chapter, I show how several of these models can be run, focusing on logistic regression at the beginning. Much like chapter 4, running the models, examining the assumptions, and interpreting the output is emphasized.

This chapter also demonstrates how the results of the GLM approaches can be presented reproducibly and succinctly with minimal coding.

Chapter 6: Multilevel Modeling

This chapter shows how multilevel models can be used in R. Both nested data and longitudinal data are common in the study of humans and, therefore, both are demonstrated. The demonstration covers both Generalized Estimating Equations and Linear Mixed Effects models using `gee`, `lme4`, and `lmerTest` packages.

This chapter also highlights the necessary format of the data and how data reshaping is an essential step to be able to model multilevel data. And, as before, ways to present these models clearly and concisely are shown.

Chapter 7: Other Modeling Techniques

This chapter covers other common modeling techniques, focusing on mediation/path analysis, structural equation models, some basic machine learning techniques, and Bayesian statistics. Each is presented with the corresponding top packages that allow for the methods to be used more simply. These packages include `lavaan`, `caret`, and `brms`.

Because this chapter discusses topics that are not as prevalent yet, readers are pointed toward resources that provide more in-depth coverage of these topics in R. As such, this chapter is only designed to introduce that the approaches are possible and that the coding aspect of using the methods is not much different

than what they have already learned in the book.

Part III

Chapter 8: Advanced data manipulation

This chapter extends Chapter 2 by discussing reshaping data in more depth. It also demonstrates the use of loops to automate data cleaning and communicating (for loops, the apply family of functions, and the `purrr` package). The chapter also shows how writing custom functions works in R and how that can increase their ability to use the loops to automate tasks.

It shows how the use of loops can make it possible to clean large data sets, run analyses, and check assumptions with minimal coding.

Chapter 9: Reproducible Workflow with RMarkdown

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 to show how to create a fully-reproducible document. It also briefly shows the various extensions to RMarkdown for producing journal-formatted documents.

Chapter 10: Where to go from here

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.

Although not currently implemented, I want to include practice questions for

each chapter as review and understanding checks. These review questions would be included throughout the chapter, not just at the end of the chapter.

Introduction to R for Researchers

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 much (or 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).

The problem for researchers to adopt R in their work is not the lack of resources available to them. There are several comprehensive books. Even free resources via online forums are plentiful. The problem is the difficulty of navigating these resources. Researchers are often full-time employees or full-time students and do not have time, nor energy, to orient themselves to where these resources can help them get started. As such, this book is designed to get researchers started in a clear way, without overwhelming them with irrelevant topics. It is an *introduction* to R for researchers so that they can get started.

Because R is a programming environment, it can be used to do nearly anything with data. But most of this functionality is not relevant to a researcher, especially as they are starting out. This book stays within the most important, common data and analysis situations for the human sciences, showing concrete examples of why and how tasks can be completed. These specific topics were chosen based on the common statistical and research methods courses for researchers. It makes this book a great companion text for any statistical or research methods course (in addition to being the main text for an introduction to R class). In fact, the draft of this book has been used in both situations

[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

with great reviews from the students.

To use R well, one must begin to understand the *R language* and the some of the *language of data*. To help in this learning, the book emphasizes learning by using: 1) many coding examples, and 2) visuals to help make the abstract functions and data processes 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 the NHANES data set mentioned earlier. This helps the code become more intuitive and natural to the researchers. In addition, each chapter will provide a zip file that can be downloaded from an online source that provides practice data and code. Although not yet implemented, I plan on providing online videos demonstrating the tasks found in the applications.

Second, 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.

Learning throughout the book is further enhanced by using the most up-to-date, readable, fast, concise coding techniques in R. These state-of-the-art techniques allow a researcher to learn the language quicker, to understand data manipulation more intuitively, and to work with small to large (even Big) data. This enhancement happens as the function names and design correspond well with the actual work being done on the data.

To some extent, each subsequent chapter builds on the preceding one, in a way that models the general workflow of a research project. That is, the book starts with importing data, goes to cleaning data, exploratory and descriptive analyses, and statistical analyses. It finishes with ways to improve efficiency and reproducibility. This makes it possible for

them to use the book, chapter-by-chapter, to perform the analyses. As they get more comfortable, the last chapters will become more and more useful. Once the researcher is feeling comfortable with the material, Chapter 10 will recommend other resources that will expand their skills to more specific research areas.

At the end of using the book, a researcher should be able to do basic data work in R and be able to use (free) resources available to them to solve situation specific problems they encounter. The book ends with a list of resources that they can use next to continue their learning in more specific areas.

Major Selling Points

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 relevant 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.

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.

These four are commonly used in applied statistics courses wherein the instructor

[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.

wants to either emphasize the use of R or, at the very least, allow its use.

This style of book is important as many undergraduate and graduate programs do not specifically teach R as part of their graduate studies. Instead, some classes introduce it as part of statistics courses. Others do not even offer R as part of statistical training. As such, the book will be a great tool to use in conjunction with other statistical texts, whether officially by the professor or unofficially by the student. 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.

In addition, 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. Therefore, the terms and approaches covered are not always understood nor relevant. This book, instead, addresses the up-to-date uses to get started enough to become literate in the R language without *over* discussing the new uses of R.

Intended Audience

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.

Book Details

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. Several figures and some photographs will be used to highlight important concepts and functions. Further, the R code will need to be formatted in such a way that it is clearly code and, if possible, in a colored, clean way. Other than that, no unusual formatting will exist.

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.