

*Christopher Gandrud*

---

# ***Reproducible Research with R and RStudio***

---

## *Author*

---

**Christopher Gandrud**

Yonsei University

Wonju, Republic of Korea

I am a lecturer at Yonsei University (Wonju) in international relations. I research and teach on international political economy, political institutions, and applied social science statistics. Previously, I was a Fellow in Government at the London School of Economics and a research associate at the Hertie School of Governance. In 2012 I completed my PhD in Political Science at the LSE.

I've published articles on political economy and quantitative methods in the Review of International Political Economy and the International Political Science Review.



---

## *Forward*

---

This book would not have been possible without the advice and support of a great many people.

The developer and blogging community has been incredibly important for making this book possible. Foremost among these people is Yihui Xie. He is the developer of the *knitr* package (among others) and also an avid writer and reader of blogs. Without him the ability to do reproducible research would be much harder and the blogging community that spreads knowledge about how to do these things would be poorer. Other great bloggers include Carl Boettiger (who also developed the *knitcitations* package), Markus Gesmann (who developed *GoogleVis*), Jeromy Anglim.

The vibrant and very helpful communities at Stack Overflow <http://stackoverflow.com/> and Stack Exchange <http://stackexchange.com/> are always very helpful for finding answers to problems that plague any coder. Importantly they makes it easy for others to find the answers to questions that others have asked before.

My students at Yonsei University were also an important part of creating this book. One of the reasons that I got interested in using many of the tools covered in this book like using *knitr* in slideshows, was to improve the Introduction to Social Science Data Analysis that I teach. I tested many of the explanations and examples with my students. Their feedback has been very helpful for making the book clearer and more useful.



---

## *Preface*

---

FILL IN



---

## *Stylistic Conventions*

---

I use the following conventions throughout this book to format computer code and actions:

- **Abstract Variables**

Abstract variables, i.e. variables that do not represent specific objects in an example, are in ALL CAPS TYPWRITER TEXT.

- **Clickable Buttons**

Clickable Buttons are in typewriter text.

- **Code**

All code is in typewriter text.

- **Filenames and Directories**

Filenames and directories more generally are printed in *italics*. Camelback is used for file and directory names.

- **Individual variable values**

Individual variable values mentioned in the text are in **bold**.

- **Objects**

Objects are printed in *italics*. Camelback (e.g. CamelBack) is used for object names.

- **Columns**

Columns are printed in *italics*

- **Packages**

R packages are printed in *italics*.

- **Windows**

Open windows are written in **bold** text.

- **Variable Names**

Variable names are printed in *italics*. Camelback is used for individual variable names.





---

## *Required R Packages*

---

This book discusses how to use a number of user-written R packages for reproducible research. These are not included in the default R installation (see Section ??). They need to be installed separately. To install all of the user-written packages discussed in this book use the following code:

```
install.packages("apsrtable",  
                 "devtools",  
                 "ggplot2",  
                 "knitr",  
                 "knitcitations",  
                 "markdown",  
                 "openair",  
                 "texreg",  
                 "tools",  
                 "xtable",  
                 "Zelig")
```



x



---

## *List of Figures*

---



---

## *List of Tables*

---



---

## *Contents*

---







# Part I

## Getting Started



# 1

---

## *Introducing Reproducible Research*

---

### CONTENTS

Research is often encountered as a neat and very abridged package: a conference presentation, journal article, book, or maybe even a website. These presentation documents announce a project’s findings and try to convince us that the results are correct (?). However, the article, slideshow, or book is not the research. Especially in the computational and statistical sciences, these documents are the “advertising”. The research is the “full software environment, code, and data that produced the results” (??, 385). Separating the research from its advertisement makes it difficult for others to complete the very important task of reproducing and validating the research’s findings.

This book gives you the tools to dynamically combine your research with the presentation of your findings. The first tool is a workflow for reproducible research that weaves the principles of reproducibility throughout your entire research project, from data gathering to the statistical analysis and the presentation your results. We will also learn how to use a number of computer tools that make this workflow possible. These tools include:

- the R statistical language which allows us to gather data and analyze it,
- the markup languages L<sup>A</sup>T<sub>E</sub>X and Markdown that allow us to create documents—slideshows, articles, books, and webpages—to present our findings,
- the *knitr* package, which allows us to dynamically combine our data gathering, analysis, presentation documents so that they can be easily reproduced,
- RStudio, a program that brings all of these tools together in one place.

## 1.1 What is reproducible research?

Research results are replicable if there is sufficient information for independent researchers to make the same findings using the same procedures (?, 444). For research that relies on experiments, this can mean a researcher not involved in the original research being able to rerun the experiment and validate that the new results match the original ones. In computational and quantitative empirical sciences results are replicable if independent researchers can obtain the same results as originally presented by following the original procedures used to gather the data and run the computer code. Of course it is sometimes difficult to replicate the original data set because of limited resources.<sup>1</sup> So as a next-best standard we can aim for “really reproducible research” (?, 1226).<sup>2</sup> In computational sciences this means:<sup>3</sup>

the data and code used to make a finding are available and they are sufficient for an independent researcher to make the same findings.

In practice, research needs to be *easy* for independent researchers to reproduce your research (?). If a study is difficult to reproduce it is more likely that no one will reproduce it. It will also be more difficult to tell if errors found during the replication are errors from the original research or errors created by the researcher attempting the verification. In this book we will learn ways to make research easily reproducible to help avoid these problems.

In particular we will learn tools that allow us to dynamically “*knit*”<sup>4</sup> the data, the source code we used for data gathering, and analysis into our presentation documents. Combined with well organized source files and clearly and completely commented code, independent researchers can understand how we obtained our results. This will make our computational research really reproducible.

---

<sup>1</sup>In this book we will actually aim for replicable research. New technologies make it possible to replicate a data set, especially if the data is available over the internet.

<sup>2</sup>The idea of really reproducible computational research was originally thought of and implemented by Jon Claerbout and the Stanford Exploration Project beginning in the 1980s and early 1990s (??). Further seminal advances were made by Jonathan B. Buckheit and David L. Donoho who created the Wavelab library of MatLab routines for their research on wavelets in the mid-1990s (?).

<sup>3</sup>Reproducibility is important for both quantitative and qualitative research (?). However, this book focuses entirely on methods for reproducibility in quantitative computational research.

<sup>4</sup>Much of the reproducible computational research and literate programming literatures have traditionally used the term “weave” to describe the process of combining source code and presentation documents (see ?, 101). In the R community weave is usually used to describe the combination of source code and  $\text{\LaTeX}$  documents. The term “knit” reflects the vocabulary of the *knitr* R package and is used more generally to describe weaving with a variety of markup languages. Because of this, I use the term knit rather than weave in this book.

---

## 1.2 Why should research be reproducible?

Reproducibility research is a key component of science. Working to make your research reproducible from the start of a project also has direct benefits for you as a researcher.

### 1.2.1 For Science

Replicability has been a key part of scientific enquiry from at least the 1200s (??) through the 1660s (?, 2) when the Royal Society began publishing descriptions of experiments and even hosting demonstrations—replications—of experiments and into the present. It has even been called the “demarcation between science and non-science” (?, 2). Why is replication so important for scientific inquiry?

#### *Standard to judge scientific claims*

Replication, or at the least reproducibility, opens claims to scrutiny; allowing us to keep what works and discard what doesn't. Science, according to the American Physical Society, “is the systematic enterprise of gathering knowledge . . . organizing and condensing that knowledge into testable laws and theories.” The “ultimate standard” for evaluating these scientific claims is whether or not the claims can be replicated (??). Research findings cannot even really be considered “genuine contribution[s] to human knowledge” until they have been verified through replication (?, 38). Replication “requires the complete and open exchange of data, procedures, and materials”. Scientific conclusions that are not replicable should be abandoned or modified “when confronted with more complete or reliable . . . evidence”.<sup>5</sup>

#### *Avoiding effort duplication & Cumulative knowledge development*

Not only is reproducibility crucial for evaluating scientific claims, it can also help enable the cumulative growth of future scientific knowledge (??). Reproducible research cuts down on the amount of time scientists have to spend duplicating data or procedures that have already been collected or figured out. Because researchers do not have to discover on their own things that have already been done, they can apply these data and procedures more quickly to building on established findings.

---

<sup>5</sup>See the American Physical Society's website at [http://www.aps.org/policy/statements/99\\_6.cfm](http://www.aps.org/policy/statements/99_6.cfm). See also ?.

### 1.2.2 For You

Working to make your research reproducible does require extra upfront effort. For example, you need to put effort into learning the tools of reproducible research by doing something like reading this book. But beyond the clear benefits for science, why should you make this effort? Making your research reproducible has a number of benefits that make your research process more effective and (hopefully) ultimately easier.

#### *Better work habits*

Making a project reproducible from the start encourages you to use better work habits. It can spur you to more effectively plan and organize your research. It should push you to bring you data and source code up to a higher level of quality than you might if you “thought ‘no one was looking’” (?, 386). This pushes you to root out errors—a ubiquitous part of computational research—earlier in the research process (?, 385). Clear documentation also makes it easier to find errors.<sup>6</sup>

Reproducible research needs to be stored so that other researchers can actually access the data and source code. By taking steps to make you research accessible for others you are also making it easier for you to find it in the future when you want to use the data or methods in other projects. You avoid effort duplication in our own work. You are able to cumulatively build on you work more effectively.

#### *Better teamwork*

A key part of reproducible research is organizing and documenting your work so that it is easy for others to figure out what you are doing. The steps you take to make sure an independent researcher can figure out what you have done also make it easier for your collaborators to understand as well. This applies not only to current collaborators, but also future collaborators. Bringing new members of a research team up to speed on a cumulatively growing research project is faster if they can easily understand what has been done already (?, 386).

#### *Changes are easier*

A third person may or may not actually reproduce your research even if you make it easy to do so. But, it’s almost certain that you will reproduce parts or even all of your research. Often times you will reproduce your work long after you originally worked on it and long since you last remembered the details of how to do it. Almost no actual research process is completely linear. We almost never gather data, run our analyses, and present our results without also going backwards to add variables, make changes to our statistical models, create

---

<sup>6</sup>Of course, reproducibility is “neither necessary nor sufficient to prevent mistakes” (?).

new graphs, alter results tables in light of new findings, and so on. Whether these changes are because of journal reviewers' and conference participants' comments or we discover that new and better data has been made available since beginning the project, designing our research to be reproducible from the start makes it much easier to make these changes.

Changes made to one part of a research project have a way of cascading through the other parts. For example, adding a new variable to a largely completed analysis requires gathering new data and merging it into existing data sets. If we are using data imputation or matching methods this can lead to adjustments to the entire data set. We then have to update our statistical models and the tables and graphs we use to present results. Adding a new variable essentially forces us to reproduce large portions of our research. If we made it easier for others to reproduce our research we also made it easier for us to reproduce it as well. This has been referred to as taking steps to have a "better relationship with our future selves" (?).

#### *Improve research impact*

Reproducible research is more likely to be useful for other researchers. Useful research is cited more frequently (?). Research that is fully reproducible contains more information, i.e. more reasons to use and cite, than research that merely presents findings. Other researchers may use the data or code to look at other often unanticipated questions. For this they will (should) cite your work. Because of this, Vandewalle et al. even argue that "the goal of reproducible research is to have more impact with our research" (?, 1253).

A reason researchers often avoid making their research fully reproducible is that they are afraid other people will use their data and code to compete with them. I'll let Donoho et al. answer this one:

True. But competition means that strangers will read your papers, try to learn from them, cite them, and try to do even better. If you prefer obscurity, why are you publishing? (?, 16)

---

### 1.3 Who should read this book?

This book is intended primarily for professional researchers, upper-level undergraduate, and graduate students working on computational data-driven projects who want to have a systematic workflow that encourages reproducibility and the practical state-of-the-art computer tools to put it into practice. Hopefully, editors at academic publishers will also find the book useful for improving their ability to evaluate and edit reproducible research.

The book will be helpful for people who already have general experience



using technologies such as the R statistical programming language and L<sup>A</sup>T<sub>E</sub>X. The more researchers that start to incorporate the tools of reproducibility the better. So I have also included enough information in the book that even if you have very limited experience with these tools, you will be able to start incorporating them into your workflow right away.

### 1.3.1 Academic Researchers

Hopefully so far in this chapter I've convinced you that reproducible research has benefits for you as a member of the scientific community and personally as a computational researcher. This book is intended to be a practical guide for how to actually make your research reproducible. Even if you already use tools such as R and L<sup>A</sup>T<sub>E</sub>X you may not be leveraging their full potential. This book will teach you useful ways to get the most out of them as part of a coherent reproducible research workflow.

### 1.3.2 Students

Upper-level undergraduate and graduate students conducting original computational research should make their research reproducible for the same reasons that professional researchers do. Forcing yourself to clearly document the steps you took will encourage you to think more clearly about what you are doing and will reinforce what you are learning. It will also hopefully give you a greater appreciation of research accountability and integrity early in your careers (??, 183).

Even if you don't have extensive experience with computer languages, this book will teach you specific habits and tools that you can use throughout your student research and hopefully your careers. Learning these things earlier will save you considerable time and effort later.

### 1.3.3 Teachers

When instructors incorporate the process and tools of reproducible research into their courses they not only build students' understanding of research best practice, but they are also better able to evaluate and provide meaningful feedback on their students' work (?, 183). This book provides a resource that you can use with students to put reproducibility into practice.

If you are teaching computational courses, you may also benefit from making your lecture material dynamically reproducible. They will be easier to update. Making the methods you used to create the material available to students will give them more information. Clearly documenting how you created lecture material can also pass information on to future instructors.

### 1.3.4 Editors

Beyond a lack of reproducible research skills among researchers, an impediment to actually creating reproducible research is a lack of infrastructure to publish it (?). Hopefully, this book will be useful for editors at academic publishers who want to be better at evaluating reproducible research, edit it, and develop systems for making it more widely available. For example, from 2009 the journal *Biostatistics* has had an editor for reproducibility that ensures replication files are available and that results can actually be replicated using these files (?). The more editors there are with the skills to work with reproducible research the more likely it is that researchers will do it.

### 1.3.5 Industry practitioners

Industry practitioners may or may not want to make their work easily reproducible outside of their organization. However, that does not mean that significant benefits cannot be gained from using the methods of reproducible research. First, even if public reproducibility is ruled out to guard proprietary information,<sup>7</sup> making your research reproducible to members of your organization can spread valuable information about how analyses were done and data was collected. This will help build your organization's knowledge and avoid effort duplication. Just as a lack of reproducibility hinders the spread of information in the scientific community, it can hinder it inside of a private organization.

The tools of reproducible research covered in this book enable you to create professional standardized reports that can be easily updated or changed when new information is available. In particular, we learn how to create batch reports.

---

## 1.4 The Tools of Reproducible Research

This book will teach you the tools you need to make your research highly reproducible. Reproducible research involves two broad sets of tools. The first is a **reproducible research environment** that includes the statistical tools you need to run your analyses as well as “the ability to automatically track the provenance of data, analyses, and results and to package them (or pointers to persistent versions of them) for redistribution”. The second set of tools is a **reproducible research publisher**, which prepares dynamic documents that present the results and is easily linked to the reproducible research environment (?).

---

<sup>7</sup>There are ways to enable some public reproducibility without revealing confidential information. See ? for a discussion of one approach.

In this book I will focus on learning how to use the widely available and highly flexible reproducible research environment—R/RStudio. R/RStudio can be linked to numerous reproducible research publishers such as L<sup>A</sup>T<sub>E</sub>X and Markdown with Yihui Xie’s *knitr* package (?). The full list of tools covered in this book include:

- **R:** a programming language primarily for statistics and graphics. It can also be used for data gathering and creating presentation documents.
- ***knitr*:** an R package for literate programming, i.e. it allows us to combine our statistical analysis and the presentation of the results into one document. It works with R and a number of other languages such as Bash, Python, and Ruby.
- **Markup languages:** instructions for how to format a presentation document. In this book we cover L<sup>A</sup>T<sub>E</sub>X and Markdown.
- **RStudio:** an integrated developer environment (IDE) for R that tightly integrates R, *knitr*, and markup languages.
- **Cloud storage & versioning:** Services such as Dropbox and Github that can store data, code, and presentation files, save previous versions of these files, and make this information widely available.
- **Unix-like shell programs:** These tools are useful for setting up and working with large research projects.<sup>8</sup> They also allow us to use command line tools including Pandoc, a program for converting documents from one markup language to another.

---

## 1.5 Why use R, knitr, and RStudio for reproducible research?

### *Why R?*

Why use a statistical programming language like R for reproducible research? R has a very active development community that is constantly expanding what R is capable of. As we will see in this book this enables researchers across a wide range of disciplines to gather data and run statistical analyses. Using the *knitr* package, we can use R to connect our analysis to presentation documents created with markup languages such as L<sup>A</sup>T<sub>E</sub>X, Markdown, and HTML. This allows us to dynamically and reproducibly present our results in articles, slideshows, and webpages.

---

<sup>8</sup>In this book I cover the Bash shell for Linux and Mac as well as Windows PowerShell.

The way we interact with R has benefits for reproducible research. In general we interact with R (or any other programming and markup language) by explicitly writing our steps down as source code. This promotes reproducibility more than our typical interactions with Graphical User Interface (GUI) programs like SPSS<sup>9</sup> and Microsoft Word. When we write R code and embed it in documents using markup languages we are being forced to explicitly express the steps we take to do our research. When we do research by clicking through drop down menus in GUI programs, the steps we take are lost. Or at least documenting them requires considerable extra effort. Also it is generally more difficult to dynamically embed our analysis in presentation documents created by GUI word processing programs in a way that will be accessible to other researchers both now and in the future. I'll come back to these points in Chapter ??.

### *Why knitr?*

Literate programming is a crucial part of reproducible quantitative research.<sup>10</sup> Being able to directly link your analyses, your results, and the code you used to produce the results makes tracing your steps much easier. There are many different literate programming tools for a number of different programming languages. Previously, one of the most common tools for researchers using R and the L<sup>A</sup>T<sub>E</sub>X markup language was the Sweave (?). The package I am going to focus on in this book is newer and is called *knitr* (?). Why are we going to focus on *knitr* in this book and not Sweave or some other tool?

The simple answer is that *knitr* has the same capabilities as Sweave and more. It can work with many more markup languages<sup>11</sup> and can even work with programming languages other than R. It highlights R code in presentation documents making it easier for your readers to follow. It gives you better control over the inclusion of graphics and can cache code chunks. It has the ability to understand Sweave-like syntax, so it will be easy to convert backwards to Sweave if you want to. You can also use much simpler and more straightforward syntax with *knitr* both with L<sup>A</sup>T<sub>E</sub>X and other markup languages.

### *Why RStudio?*

Why use the RStudio integrated development environment for reproducible research? R by itself has the capabilities necessary to gather data, analyse it, and, with a little help from *knitr* and markup languages, present results in a

<sup>9</sup>I know you can write scripts in statistical programs like SPSS, but doing so is not encouraged by the interface and we often have to learn multiple languages just to write scripts that run analyses, create graphics, and deal with matrices.

<sup>10</sup>Donald Knuth coined the term literate programming in the 1970s to refer to a source file that could be both run by a computer and “woven” with a formatted presentation document (?).

<sup>11</sup>It works with L<sup>A</sup>T<sub>E</sub>X, Markdown and reStructuredText. We cover the first two in this book.

way that is highly reproducible. RStudio allows us to do all of these things, but simplifies many of the tasks and navigate them in a more visual way. RStudio is designed for reproducible research. It is a as happy medium between R's text-based interface and a pure GUI.

RStudio is very tightly integrated with technologies such as *knitr*, Markdown, and  $\text{\LaTeX}$  that enable us to present reproducible research. For example, compiling  $\text{\LaTeX}$ PDF documents or HTML webpages in RStudio requires fewer steps than doing the same thing in plain R.

Not only does RStudio do many of the things that R can do but more easily, it is also a very good stand alone editor for writing documents with  $\text{\LaTeX}$ , Markdown, and HTML. For  $\text{\LaTeX}$  documents it can, for example, insert common commands like `\section*` for unnumbered sections or set up lists (see Chapter ??).<sup>12</sup> There are many  $\text{\LaTeX}$  editors available, both open source and paid. But RStudio is currently the best program for creating reproducible  $\text{\LaTeX}$ , Markdown, and HTML documents. It has full syntax highlighting, even for documents with *knitr* code (which it can collapse when you just want to work on the text). It can spell check  $\text{\LaTeX}$  documents. It handles *knitr* code chunks beautifully (see Chapter ??). RStudio makes it easy to navigate through complex documents and run individual chunks.

Finally, RStudio not only has tight integration with various markup languages, it also has capabilities for using other tools such as CSS, JavaScript, and a few other programming languages. It is closely integrated with the version control programs git and SVN. Both of these programs allow you to keep track of the changes you make to your documents (see Chapter ??). This is important for reproducible research since the version control program documents many of your research steps.

### 1.5.1 Installing the Software

Before you read this book you should install the software. All of the software programs covered in detail in this book is open source and can be easily downloaded for free. They are available for Windows, Mac, and Unix. They should run well on most modern computers.

You should install R before installing RStudio. You can download the programs from the following websites:

- **R:** <http://www.r-project.org/>,
- **RStudio:** <http://rstudio.org/download/>.

The download webpages for these programs have comprehensive information on how to install them, so please refer to those pages for more information.

<sup>12</sup>If you are more comfortable with a what-you-see-is-what-you-get (WYSIWYG) word processor like Microsoft Word, you might be interested in exploring Lyx. It is a WYSIWYG-like  $\text{\LaTeX}$  editor that works with *knitr*. It doesn't work with the other markup languages covered in this book. For more information see: <http://www.lyx.org/>.

After installing R and RStudio you will probably also want to install a number of user-written packages that are covered in this book. To install all of these user-written packages, please see page ??.

### *Installing markup languages*

If you are planning on creating L<sup>A</sup>T<sub>E</sub>X documents you need to install a L<sup>A</sup>T<sub>E</sub>X distribution. They are available for Windows, Mac, and Unix. They can be found at: <http://www.latex-project.org/ftp.html>. Please refer to that site for more installation information.

If you want to create markdown documents you will need to install the *markdown* package in R. You can do this the same way that you install any package in R, with the `install.packages` command.<sup>13</sup>

---

## 1.6 Book overview

The purpose of this book is to give you the tools that you will need to do reproducible research with R and RStudio.

### 1.6.1 What this book is not.

This book describes a workflow for reproducible research primarily using R and RStudio. It is designed to give you the necessary tools to use this workflow for your own research. It is not designed to be a complete introduction to R, RStudio, *knitr*, GitHub, the command line, or any other program that is a part of this workflow. Instead it shows you how these tools can fit together to make your research more reproducible. To get the most out of these individual programs I will along the way point you to other resources that cover these programs in more detail.

To that end, I can recommend a number of books that cover more of the nitty-gritty of R and the command line.

- Michael J. Crawley's encyclopaedic R book, appropriately titled, **The R Book** published by Wiley.
- Norman Matloff's tour through the programming language aspects of R called **The Art of R Programming: A Tour of Statistical Design Software** published by No Starch Press.
- For an excellent introduction to the command line in Linux and Mac, though with pretty clear implications for Windows users if they are running PowerShell (see Chapter 2) see William E. Shotts Jr.'s book **The**

---

<sup>13</sup>The exact command is: `install.packages("markdown")`.

**Linux Command Line: A Complete Introduction** also published by No Starch Press.

- The RStudio website (<http://rstudio.org/docs/>) has a number of useful tutorials on how to use *knitr* with L<sup>A</sup>T<sub>E</sub>X and Markdown.

That being said, my goal is for this book to be self-sufficient. A reader without a detailed understanding of these programs will be able to understand and use the commands and procedures I cover in this book. While learning how to use R and the other programs I personally often encountered illustrative examples that included commands, variables, and other things that were not well explained in the texts that I was reading. This caused me to waste many hours trying to figure out, for example, what the `$` is used for (preview: it's the component selector). I hope to save you from this wasted time by either providing a brief explanation of these possibly frustratingly mysterious conventions and/or pointing you in the direction of a good explanation.

### 1.6.2 How to read this book

This book gives you a workflow. It has a beginning, middle, and end. So, unlike a reference book it can and should be read linearly as it takes you through an empirical research processes from an empty folder maybe called *ResearchPaper* to a completed set of documents that reproducibly showcase your findings.

That being said, readers with more experience using tools like R or L<sup>A</sup>T<sub>E</sub>X may want to skip over the nitty-gritty parts of the book that describe how to manipulate data frames or compile L<sup>A</sup>T<sub>E</sub>X documents into PDFs. Please feel free to skip these sections.

If you are experienced with R in particular you may want to skip over the first two sections of Chapter ??: Getting Started with R/RStudio. The latter part of this chapter contains important basic information on the *knitr* package.

### 1.6.3 How this book was written

This book practices what it preaches. It can be reproduced. I wrote the book using the programs and methods that I describe. Full documentation and source files can be found at the Book's GitHub repository. Feel free to read and even use (within reason and with attribution, of course) the Book's source code. You can find it at: <https://github.com/christophergandrud/Rep-Res-Book>. This is especially useful if you want to know how to do something in the book that I don't directly cover in the text.

In the same spirit, I encourage you to make your research files—not just data files, but analysis code and markup—available for other researchers to learn from. Not only does reproducibility help us evaluate past work, but it also pushes forward knowledge in the scientific community.

#### **1.6.4 Contents overview**

The book is broken into four parts. The first part of the book (chapters ??, ??, and ??) gives an overview of the reproducible research workflow as well as the general computer skills that you'll need to use this workflow. Each of the next three parts of the book guide you through the specific skills you will need for each part of the reproducible research process. The second part of the book (chapters ??, ??, and ??) covers the data gathering and file storage process. The third part (chapters ??, ??, and ??) teaches you how to dynamically incorporate your statistical analysis, results figures and tables into your presentation documents. The final part (chapters ??, ??, and ??) covers how to create reproducible presentation documents including  $\text{\LaTeX}$ articles, books, slideshows and batch reports as well as Markdown webpages and slideshows.





# 2

---

## *Getting Started with Reproducible Research*

---

### CONTENTS

Researchers often start thinking about making their research reproducible near the end of the research process when they get start writing up the results. Or maybe later, like when a journal has conditioned an article's acceptance on making its data available. Or maybe even later when another researcher asks if they can use the data. By these points there may be numerous versions of the data set and records of the analyses strewn across multiple folders on the researcher's computer. It can be difficult and time consuming to sift through these files to create an accurate account of how the results were reached. Waiting until near the end of the research process to start thinking about reproducibility can lead to incomplete documentation that does not give us an accurate account of how findings were made. Keeping your eye on reproducibility from the beginning of the research process and continuing to follow a few simple guidelines throughout your research can help solve these problems. Remember "reproducibility is not an afterthought—it is something that must be built into the project from the beginning" (? , 386).

This chapter first gives you a big picture overview of the reproducible research process: a workflow for reproducible research. Then it covers some of the key guidelines that can help make your research more reproducible.

---

### 2.1 The Big Picture: A workflow for reproducible research

To make our research accurately reproducible we should start thinking at the beginning—the data gathering stage—about how other researchers and ourselves will be able to reproduce our results. This book teaches the tools for an integrated workflow for reproducible research. It covers tools that we can use across the three basic stages that most computational researchers go through to answer their research questions:

- data gathering,

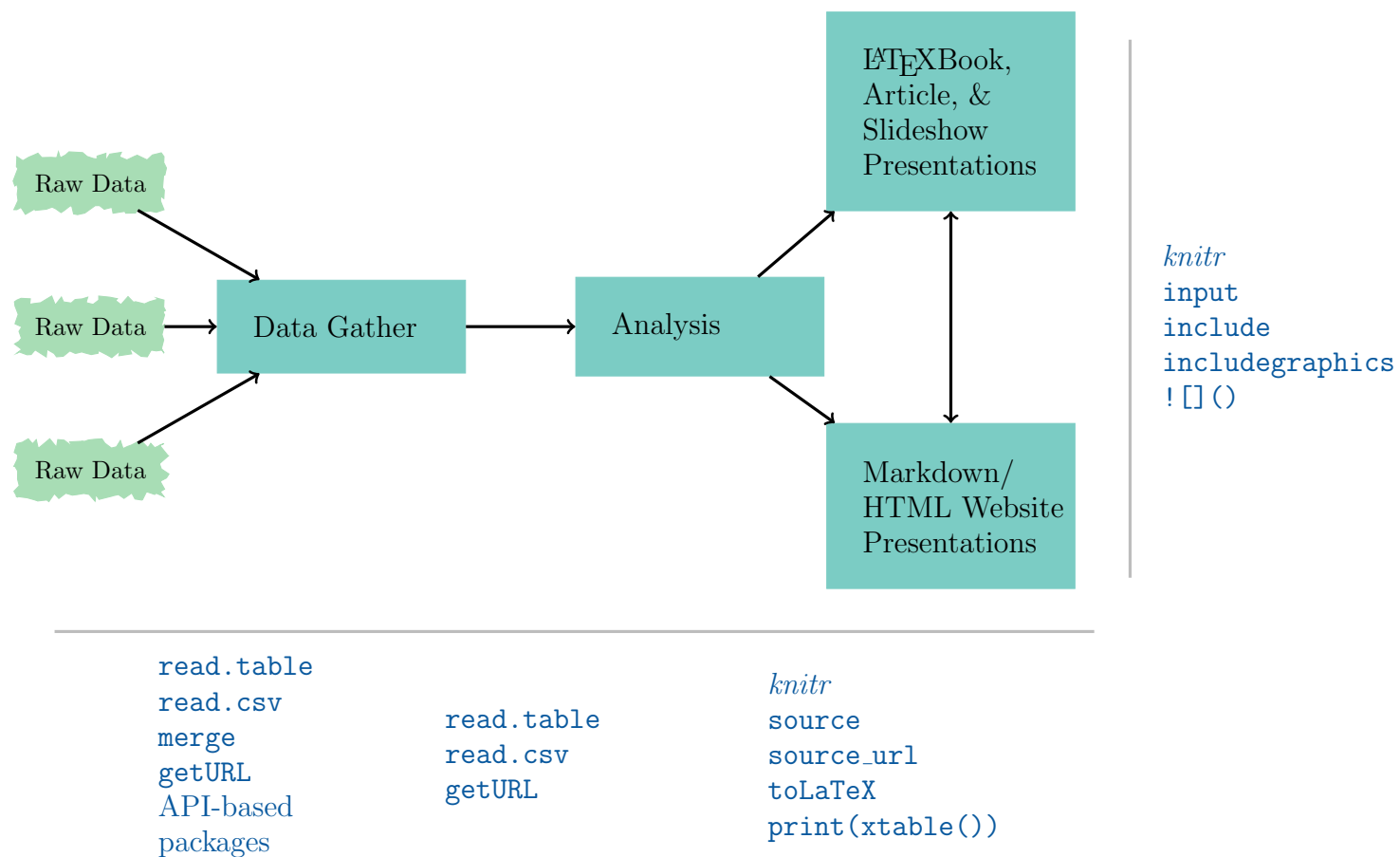
- data analysis,
- results presentation.

Each stage is covered in its own part of the book. Tools for reproducibly gathering data are covered in Part II. Part III teaches tools for tying the data we gathered to our statistical analyses and presenting the results with tables and figures. Part IV discusses how to tie these findings into documents we can use to advertise our findings to others.

If you are like me, you will probably be tempted to start using the individual tools of reproducible research as soon as you find learn them. However, I would recommend stepping back and considering how the stages of reproducible research *tie* together. This will make your reproducible research workflow more coherent from the beginning and save you a lot of backtracking later on. Figure ?? gives you a preview of the workflow and some of the commands we will learn in this book for tying it together. For more information on these tie commands see Table ?. The tips in this chapter should guide you through the reproducible research workflow. The rest of the book will give you the details.

**FIGURE 2.1**

Example Workflow &amp; Commands to Tie it Together



### 2.1.1 Reproducible Theory

An important part of the research process that I do not actively discuss in this book is the theoretical stage. Ideally, if you are using a deductive research design, the bulk of this work would precede and even guide the data gathering process. Because I don't cover this stage of the research process doesn't mean that theory building shouldn't be reproducible. It can in fact be "the easiest part to make reproducible" (? , 1254). For mathematically based theory results clear and complete descriptions of the proofs should be given. Quotes and paraphrases from previous works in the literature obviously need to be fully cited so that others can verify that they accurately reflect the source material.

Just because I don't cover theory replication in depth in this book, doesn't mean that you can't find out information about it here. I do touch on some of the ways to incorporate proofs and citations into your presentation documents. These tools are covered in Part IV.

---

## 2.2 Practical tips for reproducible research

Before we start learning the details of reproducible research with R and RStudio it is useful to cover a few broad tips that will help us organize our research process and put these skills in perspective. The tips are:

1. Document everything!,
2. Everything is a file,
3. All files should be human readable,
4. Reproducible research projects are many files explicitly tied together,
5. Don't put too much in one file
6. Have a plan to organize, store, and make your files available.

Using these tips will help make your computational research really reproducible.

### 2.2.1 Document everything!

In order to reproduce your research others must be able to know what you did. You have to tell them what you did by documenting as much of your research process as possible. Ideally, you should tell your readers how you gathered your data, analyzed it, and presented the results.

One important part of documenting everything with R is to *record your session info*. Many things in R stay the same over time, which makes it easy

for future researchers to recreate what was done in the past. However, things—syntax in particular—do can change from one version of R to another. Also, the way R functions may be handled slightly different on different operating systems. Finally, you may have R set to load packages by default. These packages might be necessary to run your code, but other people might not be able know what packages were loaded from just looking at your source code. The `sessionInfo` command prints a record of all of these things. The information from the session I used to create this book is:

```
sessionInfo()

## R version 2.15.1 (2012-06-22)
## Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)
##
## locale:
## [1] C/en_US.UTF-8/C/C/C/C
##
## attached base packages:
## [1] tools      stats      graphics  grDevices  utils
## [6] datasets  methods   base
##
## other attached packages:
##  [1] reshape_0.8.4      plyr_1.7.1
##  [3] MCMCpack_1.2-4      coda_0.15-2
##  [5] lattice_0.20-10     Zelig_3.5.5
##  [7] boot_1.3-5          MASS_7.3-21
##  [9] xtable_1.7-0        texreg_1.02
## [11] openair_0.7-0       markdown_0.5.2
## [13] knitcitations_0.1-0 bibtex_0.3-2
## [15] knitr_0.8           ggplot2_0.9.2.1
## [17] extrafont_0.11      devtools_0.8
## [19] apsrtable_0.8-8
##
## loaded via a namespace (and not attached):
##  [1] Matrix_1.0-9        RColorBrewer_1.0-5
##  [3] RCurl_1.91-1         Rttf2pt1_1.1
##  [5] XML_3.9-4            cluster_1.14.2
##  [7] codetools_0.2-8      colorspace_1.1-1
##  [9] dichromat_1.2-4      digest_0.5.2
## [11] evaluate_0.4.2       formatR_0.6
## [13] grid_2.15.1          gtable_0.1.1
## [15] httr_0.1.1           labeling_0.1
## [17] memoise_0.1          mgcv_1.7-21
```

```
## [19] munsell_0.4      nlme_3.1-104
## [21] parallel_2.15.1  pkgmaker_0.8
## [23] proto_0.3-9.2    reshape2_1.2.1
## [25] scales_0.2.2     stringr_0.6.1
## [27] whisker_0.3-1
```

It is good practice to include make output of the session info available either in the main document or in a separate readme file. If you want to nicely format the information for a  $\text{\LaTeX}$  document simply use the `toLatex` command.

```
toLatex(sessionInfo())
```

Chapter ?? gives specific details about how to create readme files with dynamically included session information.

### 2.2.2 Everything is a (text) file

Your documentation is stored in files that include data, analysis code, the write up of results, and explanations of these files (e.g. data set codebooks, session info files, and so on). Ideally, you should use the simplest file format possible to store this information. Usually the simplest file format is the humble, but versatile, text file is the simplest file format.<sup>1</sup>

Text files are extremely nimble. They can hold data in, for example, comma-separated values `.csv` files. They can contain our analysis code in `.R` files. And they can be the basis for our presentation documents as markup documents like `.tex` or `.md`, for  $\text{\LaTeX}$  and Markdown files respectively. All of these files can be opened by any program that can read text files—i.e. files with the generic file extension `.txt`.

The main reason reproducible research is best stored in text files is that this helps **future proof** our research. Other common file formats, like Microsoft Word or Excel documents change regularly and may not be compatible with future versions of these programs. Text files, on the other hand, can be opened by a very wide range past and, more likely than not, future programs. Even if future researchers do not have R or a  $\text{\LaTeX}$  distribution, they will still be able to open our text files and, aided by frequent comments, be able to understand how we conducted our research (?, 3).

Text files are also very easy to search and manipulate with a wide range

<sup>1</sup>Depending on the size of your data set it may not be feasible to store it as a text file. Nonetheless, text files can still be used for analysis code and presentation files.

of programs—such as R—that can find and replace text characters as well as merge and separate files. They have a number of clear benefits for reproducible research including enabling versioning and track changes in programs such as Git (see Chapter ??).

### 2.2.3 All files should be human readable

Treat all of your research files as if someone who has not worked on the project will try to read them and understand them. Computer code is a way of communicating with the computer. It is ‘machine readable’ in that the computer is able to use it to understand what we want done.<sup>2</sup> Hopefully, the researcher understands what they are communicating when they write their code. However, there is a very good chance that other people (or you six months in the future) will not understand what is being communicated. So, you need to make your documentation ‘human readable’. To make your documentation accessible to other people you need to **comment frequently** (?, 3) and **format your code using a style guide** (?). For especially important piece of code you should use **literate programming**—where the code and the presentation document text appear in the same document—so that it is very clear to others how you accomplished a piece of research.

#### *Commenting*

In R everything on a line after a # hash (number) character is ignored by R, but is readable to people who open the file. The hash character is a comment declaration character. You can use the # to place comments telling other people what you are doing. Here are some examples:

```
# A complete comment line
2 + 2 # A comment after R code

## [1] 4
```

Because on the first line the # is placed at the very beginning, the entire line is treated as a comment. On the second line the # is placed after the simple equation  $2 + 2$ . R runs the equation as usual and finds the answer 4, but it ignores all of the words after the hash.

According to Nagler (?, 491) to make your code understandable it is good practice to:

- write a comment before a block of code describing what the code does,
- comment on any line of code that is ambiguous.

<sup>2</sup>Of course, if it does not understand it will usually give us an error message.



In this book I follow these guidelines when displaying written code.

Different languages have different comment declaration characters. In  $\text{\LaTeX}$  everything after the % percent sign is treated as a comment and in mark-down/HTML comments are placed inside of `<!-- -->`. The hash character is used for comment declaration in shell scripts.

All of your source code files should begin with a comment header that *at the least* should include:

- a description of what the file does,
- the date it was last updated,
- the name of the file’s creator and any other contributors

You may also want to include other information in the header such as what other files it depends on, what output files it produces, and what version of the programming language you are using.

Here is an example file header for an R source code file that creates the third figure in an article titled “My Article”:

```
#####
# Source code file used to create Figure 3 in 'My Article'
# Created by Christopher Gandrud
# Updated 21 September 2012
#####
```

Feel free to use things like the long series of hash marks above and below the header, white space, and indentations to make your comments more readable.

### *Style guides*

In natural language writing you don’t necessarily need to follow a for things such as punctuation. People could probably figure out what you are saying. But it would be a lot easier for your readers if you use consistent rules. The same is true when writing R code. It’s good to follow consistent rules for formatting your code so that:

- it’s easier for others to understand,
- it’s easier for you to understand.

There are a number of R style guides. Most of them are similar to the Google R Style Guide (<http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html>). Hadley Wickham has a nicely presented style guide. You can find it at <https://github.com/hadley/devtools/wiki/Style>. You can also use the `formatR` package to automatically reformat your code so that it is easier to read.

*Literate programming*

For particularly important pieces of research code it may be useful to not only comment on the source file, but also display code in presentation text. For example, you may want to include key parts of the code you used for your statistical model and an explanation of this code in an appendix following your article. This is commonly referred to as literate programming (?). This book teaches you how to use the *knitr* package, a very useful tool for literate programming.

### 2.2.4 Reproducible research projects are many files explicitly tied together

If everything is just a text file then research projects can be thought of as individual text files that have a relationship with one another. A data file is used as input for an analysis file. The results of an analysis are shown and discussed in a markup file that is used to create a slideshow or PDF document. Researchers often do not explicitly document the relationships between files that they used in their research. For example, the results of an analysis—a table or figure—may be copied and pasted into a presentation document. It will be very difficult for future researchers to trace the results table or figure back to a particular analysis and a particular data set. Therefore, it is important to make the links between your files explicit.

The most dynamic way to do this is to explicitly link your files together using what I'll call **tie commands**. These commands instruct the computer program you are using to gather data, run an analysis, or compile a presentation document to use information from another file. I have compiled a selection of tie commands used in this book in Table ??.

### 2.2.5 Don't put too much in one file

Files that contain many different operations can be very difficult to navigate, even if they have detailed comments. For example, it would be very difficult to find any particular operation in a file that had the code used to gather all of the data and manipulate the data, run all of the statistical models, and create the results figures and tables. If you have a hard time finding things in a file you created, think of the difficulties independent researchers will have!

Because we have so many ways to link files together there is really no need to lump many different operations into one file. So, we can make our operations modular. One source code file should do one task. This will make your code easier to read. By breaking your operations into discrete chunks, it will also make it easier for you and others to find errors (?, 490).

**TABLE 2.1**

A Selection of Commands for Tying Together Your Research Files

Command/Package	Language	Description	Chapters for Further Information
<i>knitr</i>	R	R package with commands for tying commands from R and other programming languages commands into presentation documents	Discussed throughout the book
<code>read.table</code>	R	Reads a table into R. You can use this to import plain-text file formatted data into R.	??
<code>read.csv</code>	R	Same as <code>read.table</code> with default arguments set to import <code>.csv</code> formatted data files.	??
API-based packages	R	Various packages use APIs to gather data from the internet.	??
<code>merge</code>	R	Merges together data frames.	??
<code>source</code>	R	Runs an R source code file	??
<code>source_url</code>	R	From the <i>devtools</i> package. Runs an R source code file from a secure ( <a href="https">https</a> ) url like those used by GitHub	??
<code>print(xtable())</code>	R	Combining the <code>print</code> & <code>xtable</code> commands creates L <sup>A</sup> T <sub>E</sub> X & HTML tables from R objects	??
<code>toLaTeX</code>	R	Converts R objects to L <sup>A</sup> T <sub>E</sub> X	??
<code>input</code>	L <sup>A</sup> T <sub>E</sub> X	Includes L <sup>A</sup> T <sub>E</sub> X files inside of other L <sup>A</sup> T <sub>E</sub> X files	??
<code>include</code>	L <sup>A</sup> T <sub>E</sub> X	Similar to <code>input</code> , but puts page breaks on either side of the included text. Usually this it is used for including chapters chapters.	??
<code>includegraphics</code>	L <sup>A</sup> T <sub>E</sub> X	Inserts a figure into a L <sup>A</sup> T <sub>E</sub> X document.	??
<code>![]()</code>	Markdown	Inserts a figure into a Markdown document.	??

### **2.2.6 Have a plan to organize, store, and make your files available**

Finally, in order for independent researchers to reproduce your research they need to be able actually access the files that instruct them how to do this. Files also need to be organized so that independent researchers can figure out how they fit together. So from the beginning of your research process you should have a plan for organizing your files and a way to make them accessible. Chapter ?? discusses file organization in detail. Chapter ?? teaches you a number of ways to make your files accessible through cloud computing services like Dropbox and GitHub.



# 3

---

## *Getting Started with R, RStudio, and knitr*

---

### CONTENTS

If you have rarely or never used R before the first two sections of this chapter give you enough information to be able to get started and understand the code I use in this book. For more detailed introductions to R please refer to the related resources I mentioned in Chapter ???. Experienced R users might want to skip the first two sections of the chapter. This chapter also gives a brief overview of RStudio. It highlights the key features of main RStudio panel (what appears when you open RStudio) and some of its key features for reproducible research. Finally, I discuss the basics of the *knitr* package and how it is integrated into RStudio.

---

### 3.1 Using R: the basics

As a computer language, R has FILL in

This section covers some of the very basic syntax in R to get you started. If you have little experience with R, reading this section will make it much easier for you to follow along with the examples in the book. I cover the key components of the R language including:

- objects & assignment,
- component selection,
- functions and commands,
- arguments,
- libraries.

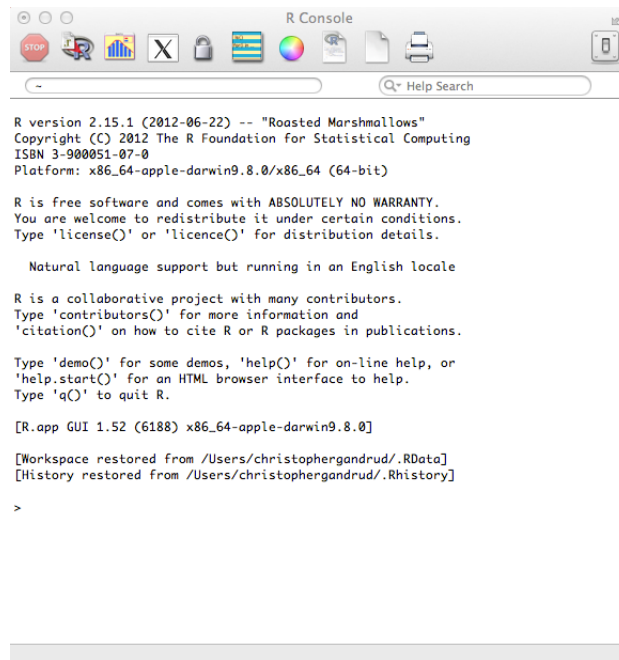
Before discussing each of these components let's open up R and look around.<sup>1</sup> When you open up R you should get a window that looks some-

---

<sup>1</sup>Please see Chapter ?? for instructions on how to install R.

thing like what you see in Figure ??.<sup>2</sup> This window is the **R console**. Under the session information—what version of R you are using, your workspace, and so on—you should see a `>`. This is where you enter R code.<sup>3</sup> To run R code that you have typed into the console type the **Enter** or **Return** key. Now that we have a new R session open we can get started.

**FIGURE 3.1**  
R Startup Console



### 3.1.1 Objects

You will probably have read that ‘R is an object-oriented language’. What are objects? Objects are like the R language’s nouns. They are things, such as a list of numbers, a data set, a word, a table of results from some analysis, and so on. Saying that R is ‘object-oriented’ just means that R is focused

<sup>2</sup>This figure and almost all screenshots in this book are from a computer using the Mac OS 10.8 operating system.

<sup>3</sup>If you are using a Unix-like system such as Ubuntu or Mac OS 10, you can an application called the Terminal. If you have installed R on your computer you can type `r` into the terminal and then the **Enter** or **Return** key it will begin a new R session. You know if a new R session has started if you get the same startup information is printed in the Terminal window.

on doing actions to objects. We will talk about the actions—commands and functions—later in this section. For now let’s create a few objects.

### *Numeric & string objects*

Different types of objects have different data types or modes. Let’s make two basic objects of the numeric and character mode. We can choose almost any name we want for our objects.<sup>4</sup> Let’s call our numeric object *Number*. To put something into the object we use the **assignment operator**: `<-`. Let’s assign the number 10 to our *Number* object.

```
Number <- 10
```

To see the contents of our object, just type its name.

```
Number  
## [1] 10
```

Let’s briefly breakdown this output. 10 is clearly the contents of *Number*. The double hash (`##`) is included in the output by *knitr* to tell you that this is output rather than R code.<sup>5</sup> Finally, `[1]` is the row number of the object that 10 is on. Clearly our object only has one row.

Creating a character object is very similar. The only difference is that we enclose the character string (letters in a word for example) inside of quotation marks (`"`). To create an object called *Words* that contains a character string “Hello World”.

```
Words <- "Hello World"
```

An object’s class is important to keep in mind as it determines what things we can do to it. For example you cannot take the mean of a character mode object like the *Words* object we created earlier:

<sup>4</sup>Objects must begin with an alphabetic character and cannot have spaces.

<sup>5</sup>This makes it easier to copy and past code included in a presentation document by *knitr*.



```
mean(Words)

## Warning: argument is not numeric or logical: returning NA

## [1] NA
```

Trying to find the mean of our *Words* object gave us a warning message and returned the value NA: not applicable. You can also think of NA to mean missing. To find out what class an object has use the `class` command. For example:

```
class(Words)

## [1] "character"
```

### *Vector & data frame objects*

So far we have only looked at objects with a single number or string.<sup>6</sup> Clearly we want to be using objects—data sets—that have many strings and numbers. In R these are usually **data frame** type objects and are roughly equivalent to the type of data structure you would be familiar with from using a program such as Microsoft Excel. We will be using data frames extensively throughout the book. It is also useful to cover other, simpler types of objects, primarily vectors. Vectors are R’s “workhorse”.<sup>7</sup> Knowing how to use vectors will be especially helpful when we clean up raw data in Chapter ?? and make tables in Chapter ??.<sup>7</sup>

### **Vectors:**

Vectors are the “fundamental data type”<sup>8</sup> in R. They are simply an ordered group of numbers, character strings, and so on.<sup>8</sup> It may be useful to think of basically all other data types as complicated forms of vectors. For example, data frames are basically multiple vectors of the same length—i.e. they have the same number of rows—and possibly of different types attached together.

<sup>6</sup>These might be called scalar objects, though in R scalars are just vectors with a length of 1.

<sup>7</sup>If you want information about other types of R objects such as lists and matrices, Chapter 1 of Norman Matloff’s book<sup>7</sup> is a really good place to look.

<sup>8</sup>In a vector every member of the group must be of the same type. Lists are similar to vectors, but allow you to have different types.

Let's create a very simple numeric vector containing the numbers 2.8, 2, and 14.8. To do this we will use the `c` (concatenate) function:

```
NumericVect <- c(2.8, 2, 14.8)

# Show NumericVect's contents
NumericVect

## [1] 2.8 2.0 14.8
```

Vectors of character strings are created in a similar way. The only major difference is that each character string is enclosed in quotation marks like this:

```
CharacterVect <- c("Albania", "Botswana", "Cambodia")

# Show CharacterVect's contents
CharacterVect

## [1] "Albania" "Botswana" "Cambodia"
```

To give you a preview of what we are going to do when we start working with real data sets, let's combine the two vectors *NumericVect* and *CharacterVect* into a new object with the `cbind` function. This function binds the two vectors together as if they were different columns.<sup>9</sup>

```
StringNumObject <- cbind(CharacterVect, NumericVect)

# Show StringNumObject's contents
StringNumObject

##      CharacterVect NumericVect
## [1,] "Albania"      "2.8"
## [2,] "Botswana"    "2"
## [3,] "Cambodia"    "14.8"
```

---

<sup>9</sup>If you want to combine objects as if they were rows of the same column(s) use the `rbind` function.

By binding these two objects together we've created a new matrix object.<sup>10</sup> You can see that the numbers in the *NumericVect* column are between quotation marks. Matrices, like vectors can only have one data type or mode.

**Data frames:** If we want to have an object with rows and columns and allow the columns to contain data with different modes, we need to use data frames. Let's use the `data.frame` command to combine the *NumericVect* and *CharacterVect* objects.

```
StringNumObject <- data.frame(CharacterVect, NumericVect)

# Display contents of StringNumObject data frame
StringNumObject

##   CharacterVect NumericVect
## 1      Albania          2.8
## 2      Botswana          2.0
## 3      Cambodia         14.8
```

There are two important things to notice. The first is that because we used the same name for the data frame object as the previous matrix object, R deleted the matrix object and completely replaced it with the data frame. This is something to keep in mind when you are creating new objects. You will also notice that the strings in the *CharacterVect* object no longer are in quotation marks. This does not mean, however that they are no longer character string type data. To prove this try taking the mean of *CharacterVect* like this:

```
mean(StringNumObject$CharacterVect)

## Warning: argument is not numeric or logical: returning NA

## [1] NA
```

### 3.1.2 Component Selection

The last bit of code will probably be confusing. Why do we have a dollar sign (\$) inbetween the name of our data frame object and the character column? The dollar sign is called the component selector. It basically extracts a part of an object. In the previous example it extracts the *CharacterVect* column

<sup>10</sup>Matrices are vectors with columns as well as rows.

from the *StringNumObject* and feeds this to the `mean` command, which tries (in this case unsuccessfully) to find its mean.

We can of course use the component selector to create new objects with parts of other objects. Imagine that we only have the *StringNumObject*, but want an object with only the information in the numbers column. Let's the following code:

```
NewNumeric <- StringNumObject$NumericVect  
  
# Display contents of NewNumeric  
NewNumeric  
  
## [1] 2.8 2.0 14.8
```

Knowing how to use the component selector will be especially useful when we discuss making tables for presentation documents in Chapter ??.

### 3.1.3 Subscripts

Another way to select parts of an object is to use subscripts. These are denoted with square braces (`[]`). We can use them to select not only columns from data frames but also rows and individual cells. As we began to see in some of the previous output, each part of a data frame as an address captured by its row and column number. We can tell R to find a part of an object by putting the row number, column number/name, or both in square braces. The first part denotes the rows and separated by a comma (,) are the columns.

To give you an idea of how this works let's use the *cars* data set that comes with R. Use the `head` command to get a sense of what this data set looks like.

```
head(cars)  
  
##   speed dist  
## 1     4    2  
## 2     4   10  
## 3     7    4  
## 4     7   22  
## 5     8   16  
## 6     9   10
```

Here we can see a data frame with information on various cars speed (*speed*)

and distance (*dist*). If we want to select only the third through seventh rows we can use the following subscript commands:

```
cars[3:7, ]

##    speed dist
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
```

To select the fourth row of the *dist* column we can type:

```
cars[4, 2]

## [1] 22
```

An equivalent way to do this is:

```
cars[4, "dist"]

## [1] 22
```

### 3.1.4 Functions and commands

If objects are the nouns of the R language functions and commands<sup>11</sup> are the verbs. They do things to objects. Let's use the `mean` command as an example. This command takes the mean of a numeric vector object. Remember our *NumericVect* object from before:

```
# Show contents of NumericVect
NumericVect

## [1]  2.8  2.0 14.8
```

---

<sup>11</sup>For the purposes of this book I treat the two as the same.

To find the mean of this object simply type:

```
mean(x = NumericVect)

## [1] 6.533
```

We use the assignment operator to place a command's output into an object. For example,

```
MeanNumericVect <- mean(x = NumericVect)
```

Notice that we typed the command's name then enclosed the object name in parentheses immediately afterwards. This is the basic syntax that all commands take: `COMMAND(ARGUMENTS)`.

### 3.1.5 Arguments

Arguments modify what commands do. In this example we have given the `mean` command one argument (`x = NumericVect`) telling it that we want to find the mean of `NumericVect`. Arguments use the `ARGUMENT LABEL = VALUE` syntax.<sup>12</sup> To find all of the arguments that an argument can accept look at the *Arguments* section of the commands help file. To access the help file type: `?COMMAND`. For example,

```
?mean
```

The help file will also tell you the default values that the arguments are set at. Clearly, you do not need to explicitly set an argument if you want it's default value.

Let's see how to use multiple arguments for one command with the `round` command. It rounds a vector of numbers. We can use the `digits` option to specify how many decimal places we want the numbers rounded to. For example, to round the object *MeanNumericVect* to one decimal place type:

---

<sup>12</sup>Note: you do not have to put spaces between the argument label and the equals sign or the equals sign and the value. However, having spaces can make your code easier for other people to read.

```
round(x = MeanNumericVect, digits = 1)

## [1] 6.5
```

You can see that arguments are separated by commas.

Arguments for logical arguments must be written as `TRUE` or `FALSE`.<sup>13</sup> Arguments that are character strings should be in quotation marks.

Some arguments do not need to be explicitly labelled. For example we could have written:

```
# Find mean of NumericVect
mean(NumericVect)

## [1] 6.533
```

R will do its best to figure out what you want and will only give up when it can't—producing an error message. However, to avoid any misunderstandings between yourself and R it can be good practice to label all of your arguments. This will also make your code easier for other people to read, i.e. it will be more reproducible.

Finally, you can stack arguments inside of other arguments. To have R find the mean of `NumericVect` and round it to one decimal place use:

```
round(mean(NumericVect), digits = 1)

## [1] 6.5
```

### 3.1.6 Installing new libraries and loading commands

You can install new add-on packages using the `install.packages` command. By default this command downloads and installs the packages from the Comprehensive R Archive Network (CRAN). For the code you will need to install all of the package libraries used in this book see ??.

Commands are stored in R libraries. R automatically loads a number of

---

<sup>13</sup>They can be abbreviated `T` and `F`.

basic libraries by default. One of the great things about R is the many user-created libraries<sup>14</sup> that greatly expand the number of commands we can use. To load a library so that you can use its functions use the `library` command. Use the following code to load the *ggplot2* library that we use in Chapter ?? to create figures.

```
library(ggplot2)
```

Please note that I only specify the library a command is in if the library is not loaded by default when you start an R session.

---

## 3.2 Using RStudio

As I mentioned in Chapter ??, RStudio is an integrated development environment. It provides a centralized and well organized place to do almost anything you want to do with R. As we will see later in this chapter it is especially well integrated with literate programming tools for reproducible research. Right now let's just take a quick tour of the basic RStudio window.

### *The default window*

When you first open RStudio you should get a default window that looks like Figure ?? . In this figure you see three window panes. The large one on the left is the *Console*. This functions exactly the same as the console in regular R. Other panes include the *Workspace/History* panes, usually in the upper right-hand corner. The *Workspace* pane keeps a record of all of the objects in your current workspace. You can click on an object in this pane to see its contents. This is especially useful for quickly looking at a data set in much the same way that you can scan a Microsoft Excel spreadsheet. The *History* pane records all of the commands you have run. It allows you to rerun code and insert it into a source code file.

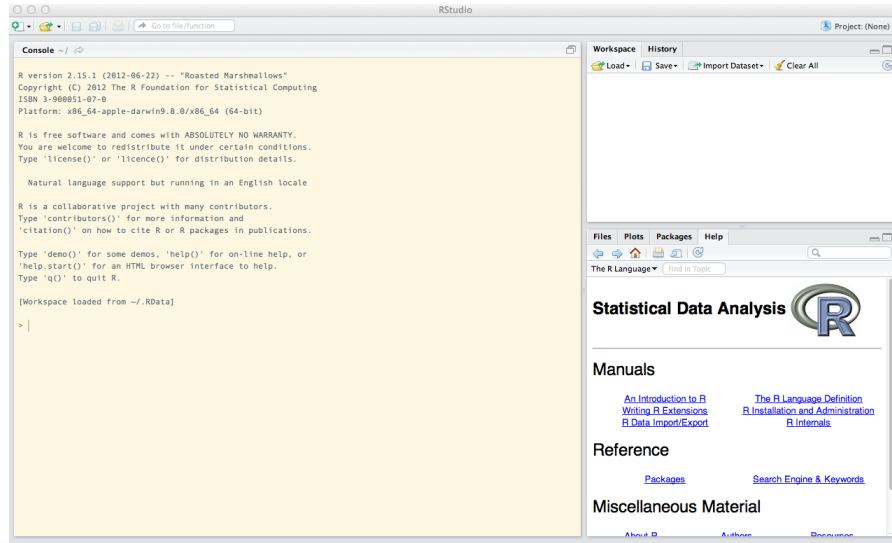
In the lower right-hand corner you will see the *Files/Plots/Packages/Help* pane. We will discuss the *Files* pane in more detail in Chapter ?? . Basically, it allows you to see and organize your files. The *Plots* pane is where figures you create in R will appear. This pane allows you to see all of the figures you have created in a session using the right and left arrow icons. It also lets you save the figures in a variety of formats. The *Packages* pane shows the packages you have installed, allows you to load individual packages by clicking on the

---

<sup>14</sup>For the latest list see: [http://cran.r-project.org/web/packages/available\\_packages\\_by\\_name.html](http://cran.r-project.org/web/packages/available_packages_by_name.html)



**FIGURE 3.2**  
RStudio Startup Panel



dialog box next to them, access their manual files (click on the package name), update the packages, and even install new packages. Finally, the Help pane shows you help files. You can search for help files and search within help files using this pane.

### *The source pane*

There is an important pane that does not show up when you open RStudio for the first time. This is the Source pane. The Source pane is where you create, edit, and run your source code files. It also functions as an editor for your markup files. It is the center of reproducible research in RStudio. Let's first look at how to use the Source pane with regular R files. These have the file extension `.R`.

You can create a new source code document, which will open a new Source pane, by going to **File** → **New**. In this drop down menu you have the option to create a variety of different source code documents. Select the **R Source** option. You should now see a new pane with a bar across the top like the first image in Figure ?? . To run the R code you have in your source code file simply highlight it<sup>15</sup> and click the **Run** icon on the top bar. This sends the code to the console where it is executed. The icon next to the right of **Run** simply runs

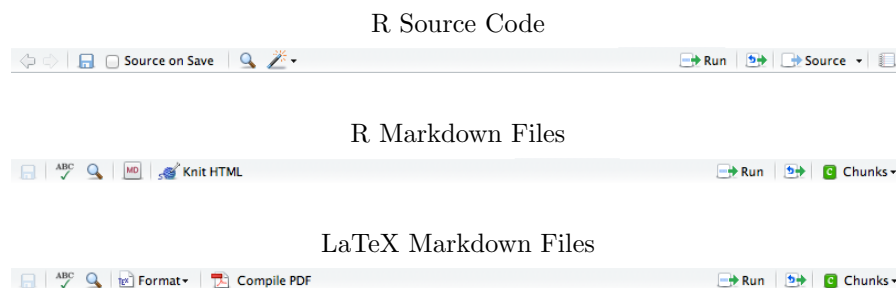
<sup>15</sup>If you are only running one line of code you don't need to highlight the code, you can simply put your cursor on that line.

the code above where you have highlighted. The **Source** icon next to this runs all of the code in file using R's `source` command. The icon next to **Source** is for compiling RStudio notebooks. We will look at creating notebooks later in this chapter.

We will cover how to use the Source pane with literate programming file formats—e.g. R Markdown and R  $\text{\LaTeX}$ —in more detail after first discussing the *knitr* basics.

### FIGURE 3.3

RStudio Source Code Pane Top Bars



---

## 3.3 Using knitr: the basics

To get started with *knitr* in R or RStudio we need to learn some of the basic concepts and syntax. The concepts are the same regardless of the markup language that we are using, but much of the syntax varies.

### 3.3.1 File extensions

When you save a knitable file use a file extension that indicates (a) that it is knitable and (b) what markup language it is using. You can use a number of file extensions for R Markdown files including: `.Rmd` and `.Rmarkdown`.  $\text{\LaTeX}$  documents that include *knitr* code chunks are generally called R Sweave files and have the file extension `.Rnw`. This terminology is a little confusing. It is a holdover from *knitr*'s main literate programming predecessor *Sweave*.<sup>?</sup> You can also use the file extension `.Rtex`, which is less confusing file extension. However, the code chunk syntax for `.Rtex` files is different from that use in `.Rnw` files. We'll look at this in more detail below.

### 3.3.2 Code Chunks

When we want to include R code into our presentation documents we place them in a code chunk. Code chunk syntax differs depending on the markup language we are using to write our documents. Let's see the syntax for R Markdown and R  $\LaTeX$  files.

#### R Markdown

In R Markdown files we begin a code chunk by writing: ````{r}`. A code chunk is closed—ended—simply with: `````. For example:

```
```{r}
# Example of a R Markdown code chunk
StringNumObject <- cbind(CharacterVect, NumericVect)
```
```

#### R $\LaTeX$

There are two different ways to delimit code chunks in R  $\LaTeX$  documents. One way largely emulates the established *Sweave* syntax.<sup>16</sup> *Knitr* now also supports files with the `.Rtex` extension, though the code chunk syntax is different. I will cover both types of syntax for code chunks in  $\LaTeX$  documents, though in throughout the book I use the older and more established *Sweave* style syntax.

#### Sweave-style

Traditional Sweave-style code chunks begin with the following code: `<<>>=`. The code chunk is closed with an at sign (`@`).

#### Rtex-style

Sweave-style code chunks may seem fairly baroque. Another option for  $\LaTeX$  files is the Rtex-style syntax. To begin a code chunk simply type `%begin.rcode`. To close the chunk you use double percent signs: `%%`. Each line in the code chunk needs to begin with a single percent sign. For example:

<sup>16</sup>The syntax originates in a literate programming tool called noweb (??).

```
%% begin.rcode
% # Example of a Rtex-style code chunk
% StringNumObject <- cbind(CharacterVect, NumericVect)
%%
```

### Code chunk labels

Each chunk has a label. When a code chunk creates a plot or is cached (see below for more details) *knitr* uses the chunk label for the resulting file's name. If you do not explicitly give the chunk a label it will be assigned one like: `unnamed-chunk-i`. *i* is the chunk's number.

To explicitly assign chunk labels in R Markdown documents place the label name inside of the braces after the `r`. If we wanted to use the label `ChunkLabel` we would simply type this at the beginning:

```
```{r ChunkLabel}
# Example chunk label
```
```

The same general format applies to the two types of  $\text{\LaTeX}$  chunks. In Sweave-style chunks we would type: `<<ChunkLabel>>=`. In Rtex-style we simply use: `%% begin.rcode ChunkLabel`.

Try not to use spaces or periods in your label names. And all chunk labels *must* be unique.

### Code chunk options

There are many times when we want to change how our code chunks are knitted and presented. Maybe we only want to show the code and not the results or perhaps we don't want to show the code at all but just a figure that it produces. Chunk options can also be used to format the size and placement of this figure. To make these changes, and many others we can specify code chunk options.

Like chunk labels, you specify options in the chunk head. Place them after the chunk label, separated by a comma. Chunk options are written following the same rules as arguments to regular R commands. They have the same `option=value` structure as arguments. The option values must be written in the same way that argument values are. Character strings need to be inside of quotation marks. The logical `TRUE` and `FALSE` operators cannot be written "true" and "false". For example, imagine we have a Markdown code chunk

called `ChunkLabel`. If we only want to have *knitr* include the code in our document, but not actually run it we use the option `eval=FALSE`. This option tells *knitr* not to evaluate (run) the code chunk.

```
```{r ChunkLabel, eval=FALSE}
# Example of a non-evaluated code chunk
StringNumObject <- cbind(CharacterVect, NumericVect)
```
```

Note that all labels and code chunk options must be on the same line. Options are separated by commas. The syntax for *knitr* options is the same regardless of the markup language. Here is the same chunk option in Rtex-style syntax:

```
%% begin.rcode ChunkLabel, eval=FALSE
% # Example of a non-evaluated code chunk
% StringNumObject <- cbind(CharacterVect, NumericVect)
%%
```

Throughout this book we will look at a number of different code chunk options. For the full list of *knitr* options see the *knitr* chunk options page maintained by *knitr*'s creator Yihui Xie: [http://yihui.name/knitr/options#package\\_options](http://yihui.name/knitr/options#package_options).

### 3.3.3 Global Options

So far we have only looked at how to set local options in *knitr* code chunks. If we want an option to apply to all of the chunks in our document we can set **global chunk options**. These options are 'global' in the sense that they apply to the entire document. Setting global chunk options helps us create documents that are formatted consistently without having to repetitively specify the same option every time we create a new code chunk. For example, in this book I center almost all of the figures. Instead of using the `fig.align='center'` option in each code chunk that creates a figure I set the option globally.

To set a global option first create a new code chunk at the beginning of your document<sup>17</sup> You will probably want to set the option `echo=FALSE` so that *knitr* doesn't echo the code. Inside the code chunk use `opts_chunk$set`.

<sup>17</sup>In Markdown, you can put global chunk options at the very top of the document. In  $\text{\LaTeX}$  they should be after the `\begin{document}` command (see Chapter ?? for more information on how  $\text{\LaTeX}$  documents are structured).

You can set any chunk option as an argument to `opts_chunk$set`. The option will be applied across your document, unless you set a different local option.

Here is an example of how we could have all of the figures in a Markdown document created by *knitr* code chunks center aligned. We place the following code at the beginning of the document:

```
```{r GlobalFigOpts, echo=FALSE}
# Center align all knitr figures
opts_chunk$set(fig.align='center')
```
```

### 3.3.4 knitr package options

Chunk options determine how we want to treat code chunks. We can also set package options that affect how the *knitr* package itself runs. For example, the `progress` option can be set as either `TRUE` or `FALSE`<sup>18</sup> depending on whether or not we want a progress bar to be displayed when we knit a code chunk.<sup>19</sup> You can use `base.dir` to set the directory where you want all of your figures to be saved to (see Chapter ??) or the `child.path` option to specify where child documents are located (see Chapter ??).

We set package options in a similar way to global chunk options with `opts_knit$set`. For example, to turn off the progress bar when knitting Markdown documents include this code at the beginning of the document:

```
```{r GlobalFigOpts, echo=FALSE}
# Turn off knitr progress bar
opts_knit$set(progress=FALSE)
```
```

### 3.3.5 knitr & RStudio

RStudio is highly integrated with *knitr* and the markup languages *knitr* works with. Because of this integration it is easier to create and compile *knitr* documents than doing so in plain R. Most of the RStudio/*knitr* features are accessed in the Source pane. The Source pane's appearance and capabilities

<sup>18</sup>It's set as `TRUE` by default.

<sup>19</sup>The *knitr* progress bar looks like this `|>>>>>| 100%` and indicates how much of a code chunk has been run.

changes depending on the type of file you have open in it. RStudio uses the file extension to automatically determine what type of file you have open.<sup>20</sup> We have already seen some of the features the Source pane has for R source code files. Let's now look at how to use *knitr* with R source code files as well as the markup formats we cover in this book: R Markdown, and R  $\text{\LaTeX}$ .

### Compiling R source code notebooks

If you want a quick well formatted—i.e not a copy and pasted version of what is in the console—account of the code that you ran and the results that you got you can use RStudio's "Compile Notebook" capabilities. RStudio uses *knitr* to create a standalone HTML file that includes all code from an R source file as well as the output. This can be useful for recording the steps you took to do an analysis. You can see an example RStudio notebook in Figure ??.

If you want to create a notebook from an open R source code file simply click the **Compile Notebook** icon in the Source pane's top bar (see Figure ??).<sup>21</sup> Then click the **Compile** button in the window that pops up. In Figure ?? you can see near the top center right a small globe icon next to the word "Publish". Clicking this allows you to publish your notebook to RPubS (<http://www.rpubs.com/>). RPubS allows you to share your notebooks over the internet. You can publish not only notebooks, but also any *knitr* Markdown document you compile in RStudio.

### R Markdown

The second image in figure ?? is what the Source pane's top bar looks like when you have an R Markdown file open. You'll notice the familiar **Run** button for running R code. At the far right you can see a new **Chunks** drop down menu. In this menu you can select **Insert Chunk** to insert the basic syntax required for a code chunk. There is also an option to **Run Current Chunk**—i.e. the chunk where your cursor is located—**Run Next Chunk**, and **Run All** chunks. You can navigate to a specific chunk using a drop down menu in the bottom left-hand side of the Source pane (not shown). This can be very useful if you are working with a long document. To knit your file click the **Knit HTML** icon on the left side of the Source pane's top bar. This will create both a knitted HTML file as well as a regular Markdown file with highlighted code, output, and figures in your R Markdown's directory. Other useful features in the R Markdown Source pane's top bar include the **ABC** spell check icon and **MD** icon which gives you a Markdown syntax reference file in the Help pane.

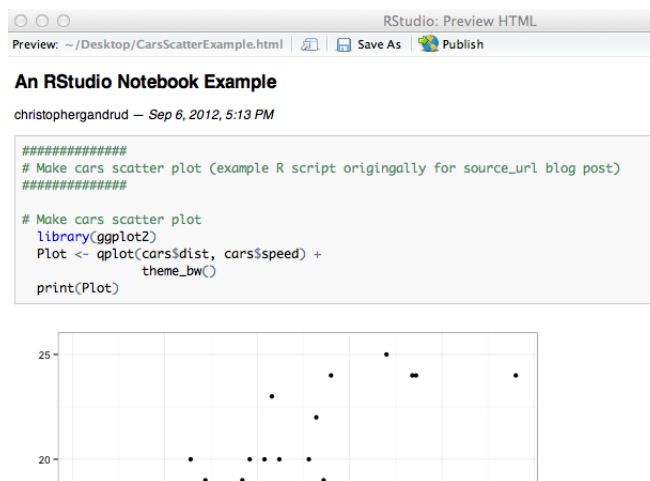
Another useful RStudio *knitr* integration feature is that RStudio can properly highlight both the markup language syntax and the R code. This makes

<sup>20</sup>You can manually set how you want the Source pane to act by selecting the file type using the drop down menu in the lower right-hand corner of the Source pane.

<sup>21</sup>Alternatively, **File** → **Compile Notebook**...

**FIGURE 3.7**

RStudio Notebook Example



your source code much easier to read and navigate. RStudio can also fold code chunks. This makes also navigating through long documents, with long code chunks, much easier. The first image in Figure ?? you can see a small downward facing arrow at line 25. If we click this arrow the code chunk will collapse, like in the second image in Figure ?. To unfold the chunk, just click on the arrow again.

You may also notice that there are a code folding arrows on lines 27 and 34 in the first image. These allow us to fold parts of the code chunk. To do this create a comment line with at least one hash before the comment text and at least four after like this:

```
#### A Comment ####
```

You will be able to fold all of the text after this comment until the next similarly formatted comment (or the end of the chunk).

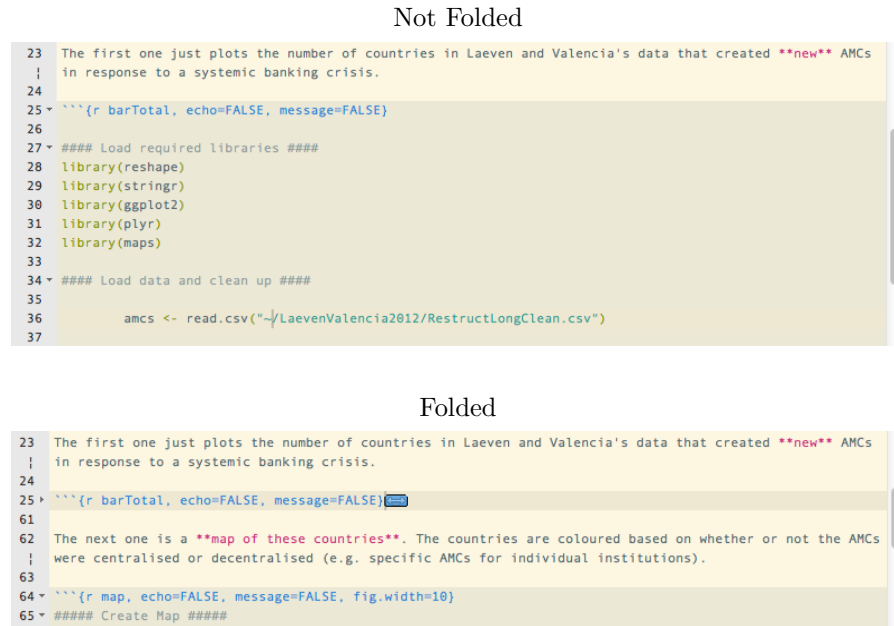
### R $\LaTeX$

You can see in the final image in Figure ?? that many of the options for R  $\LaTeX$ files are the same as R Markdown files. The key differences being that there is a **Compile PDF** icon instead of **Knit HTML**. Clicking this icon knits the file and creates a PDF file in your R  $\LaTeX$ file's directory. There is also a **Format** icon next instead of **MD**. This actually inserts  $\LaTeX$ formatting



**FIGURE 3.8**

Folding Code Chunks in RStudio



commands into your document for things such as section headings and bullet lists.

#### *Change default .Rnw knitter*

By default RStudio is set up to use *Sweave* for compiling  $\text{\LaTeX}$  documents. To use *knitr* instead of *Sweave* when knitting .Rnw files in RStudio you should go to the **Options** window. Click on the **Sweave** button. Select **knitr** from the drop down menu for “Weave files using:”. Finally, click **Apply**. The **Options** window is in different places depending on the version of RStudio you are using:

- **Mac:** RStudio → Preferences,
- **Windows & Linux:** Tools → Options.

### 3.3.6 knitr & R

As an R package, you can of course knit documents in regular R (or using the console in RStudio). All of the syntax in the document you are knitting is the

same as before. But instead of clicking a **Compile PDF** or **knit HTML** button we use is `knit` command. To knit our example Markdown file *Example.Rmd* we first set the working directory (see Chapter ??) to the folder where your *Example.Rmd* file is located with the `setwd` command. In this example I have it on my desktop.<sup>22</sup>

```
setwd("~/Desktop/")
```

Then I knit my file:

```
knit(input = "Example.Rmd", output = "Example.md")
```

Note that if you do not specify the output file *knitr* will determine what the file should be. In this example it would come up with the same name and location.

If you try this example, you find that the *knit* command only created a Markdown file and not an HTML file like clicking the RStudio **knit HTML** did. Likewise, if you use `knit` on a *.Rnw* file you will only end up with a basic *L<sup>A</sup>T<sub>E</sub>X.tex* file and not a compiled PDF. To convert the Markdown file into HTML you need to further run the *.md* file through the `markdownToHTML` command from the *markdown* package, i.e.

```
markdownToHTML(file = "Example.md", output = "Example.html")
```

You can combine the `knit` and `markdownToHTML` steps using *knitr*'s `knit2html` command like this:

```
knit2html(file = "Example.Rmd", output = "Example.html")
```

If we want to compile a *.tex* file in R we run it through the `texi2pdf` command in the *tools* package. This package will run both *L<sup>A</sup>T<sub>E</sub>X* and to create a PDF with a bibliography (see Chapter ?? for more details on using for bibliographies). Here is a `texi2pdf` example:

---

<sup>22</sup>Using the directory name `~/Desktop/` is for Mac computers. Please use alternative syntax discussed in Chapter ?? on other types of systems.

```
# Load tools package
library(tools)
# Compile pdf
texi2pdf(file = "Example.tex")
```

Just like with `knit2html`, you can speed up this process by using the `knit2pdf` command to compile a PDF file from a `.Rnw` or `.Rtex` document.

---

### 3.4 R/RStudio Tips

Finally, here are a few other tips that make using R and RStudio a little easier. In RStudio you can click on the **Help** pane (by default at the lower right, see Figure ??) and enter the command you want help with into the search field.

#### *Autocomplete*

In R and RStudio you do not have to type out every command, argument, object, or even directory name. After you start typing a command/argument/object/directory name you can hit the “tab” key to automatically complete the word you started. The function is particularly good in RStudio. Not only does it give you a list of words to choose from, but it also shows you an abbreviated version of the help file for commands and arguments.

# 4

---

## *Getting Started with File Management*

---

### CONTENTS

Careful file management is crucial for reproducible research. Apart from the fleeting situations where you have an email exchange (or even meet in person) with someone interested in reproducing your research, the main information they can have will be stored across many files: data files, analysis files, and presentation files. If these files are well organized then replication will be easier. File management is also important for you as a researcher, because if your files are well organized you will be able to understand your work more easily. Remember two of the guidelines from Chapter ??:

- Reproducible research projects are many files explicitly tied together,
- Have a plan to organize, store, and make your files available.

Using tools such as R and markup languages like L<sup>A</sup>T<sub>E</sub>X requires a bit more knowledge of how files are located—their **path**—in your computer than you are probably used to. Though more difficult to use at first than the typical point-and-click graphical user interface file handling systems you are probably familiar, R and Unix-like shell programs allow us to control files—creating, deleting, moving them—in powerful and really reproducible ways.

In this chapter we discuss how a reproducible research project may be organized and cover the basics of file path naming conventions in Unix, Mac, and Windows systems. We then learn how to navigate through files in RStudio in the **Files** pane as well as some basic R and Unix-like shell commands for manipulating files. The skills we learn in this chapter will be heavily used in the next Chapter and throughout the book.

In this chapter we work with locally stored files, i.e. files stored on your computer. In the next chapter (Chapter ??) we will discuss various ways to store and access files stored remotely in the cloud.

---

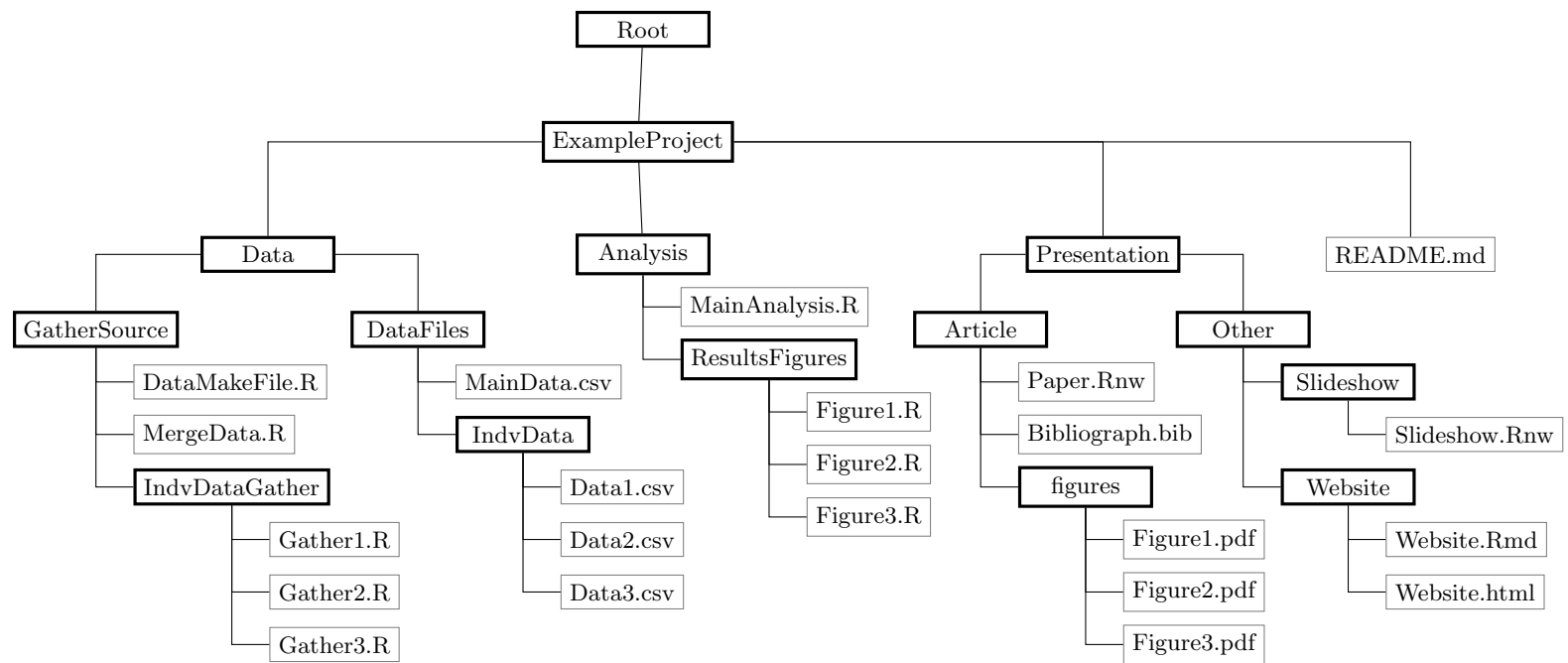
## 4.1 Organizing your research project

Figure ?? gives an example of how the files in a simple reproducible research project could be organized. The project—called *Example Project*—is organized into three main parts: a data gathering section, an analysis section, and a presentation section. The results of the project are presented in an article, slideshow, and website.

In addition to the main folders you will probably notice a file called *README.md*. The *README.md* file gives an overview of all the files in the project. It should briefly describe the project including things such as its title, author(s), and so on. It should also indicate how the folders in the project are organized and how to reproduce the project. This will likely include information on the directory names that you used. The README file should be in the main project folder so that it is easy to find. If you are storing your project as a GitHub repository (see Chapter ??) and the file is called README its contents will automatically be displayed on the repository’s main page. If it is written using Markdown, it will also be properly formatted.

It is good practice to dynamically include the system information for the R session you used to create the project. To do this you can write your readme file with R Markdown. Simply include the `sessionInfo()` command in a code chunk in the R Markdown document. If you knit this file immediately after knitting your presentation document it will record the information for that session.

**FIGURE 4.1**  
Example Research Project File Tree



---

## 4.2 File paths & naming conventions

All of the operating systems covered in this book use organize files in hierarchical directories (or file trees). To a large extent, ‘directories’ can be thought of as the folders you usually see on your Windows or Mac desktop. They are called ‘hierarchical’ because directories are located inside of other directories, like we saw in Figure ??.

### 4.2.1 Root directories

A root directory is the highest directory level in a disk, such as a hard drive.

### 4.2.2 Working directories

---

## 4.3 Operating system-specific naming conventions

*Unix*

*Mac*

*Windows*

---

## 4.4 File navigation in RStudio

The RStudio **Files** pane allows us to navigate and do some basic file manipulations. Figure ?? shows us what this pane looks like.

---

## 4.5 R file manipulation commands

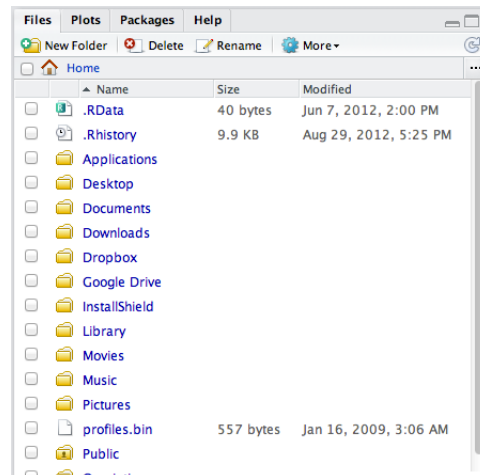
All of the tasks we can accomplish in RStudio’s **Files** pane can also be accomplished using command line R. This allows us to more easily replicate our actions.

**setwd**

The **setwd** command sets the working directory. The working directory is where R (or another computer program) looks for files and where it will place new files. For example, if we are on a Mac or Linux computer we can set the

**FIGURE 4.2**

The RStudio Files Pane



working directory to the *GatherSource* directory in our Example Project (see Table ??) like this

```
setwd("~/ExampleProject/Data/GatherSource")
```

Now R will automatically look in the *GatherSource* folder for files and will save new files into this folder, unless we tell it to do otherwise.

`dir.create`

Sometimes we may want to create a new directory. We can use the `dir.create` command to do this.

`unlink`

You can use the `unlink` command to delete a file, files, or directories.



---

## 4.6 Unix-like shell commands

Though this book is mostly focused on using R for reproducible research it can be useful to use a Unix-like shell program to manipulate files in large projects. A command line shell program is simply a program that allows you to type commands to interact with your computer's operating system.<sup>?</sup> We will especially return to shell commands near the end of the book when we discuss Make files for compiling large documents, and batch reports (Chapter ??). The syntax discussed here is also similar to the used in command line git (Chapter ??) and Pandoc (Chapter ??). We don't have enough space to properly get started with shell programs. For good introductions for Unix and Mac OS 10 computers see William E. Shotts Jr.'s book on the Linux command-line<sup>?</sup> and for Windows users Microsoft maintains a tutorial on Windows PowerShell at <http://technet.microsoft.com/en-us/library/hh848793>.

The one piece of general instruction I will give now is to highlight an important difference in the syntax between R and shell commands. In shell commands you don't need to put parentheses around your arguments. For example if we want to change our working directory to my Mac Desktop in a shell using the `cd` command we simply type:

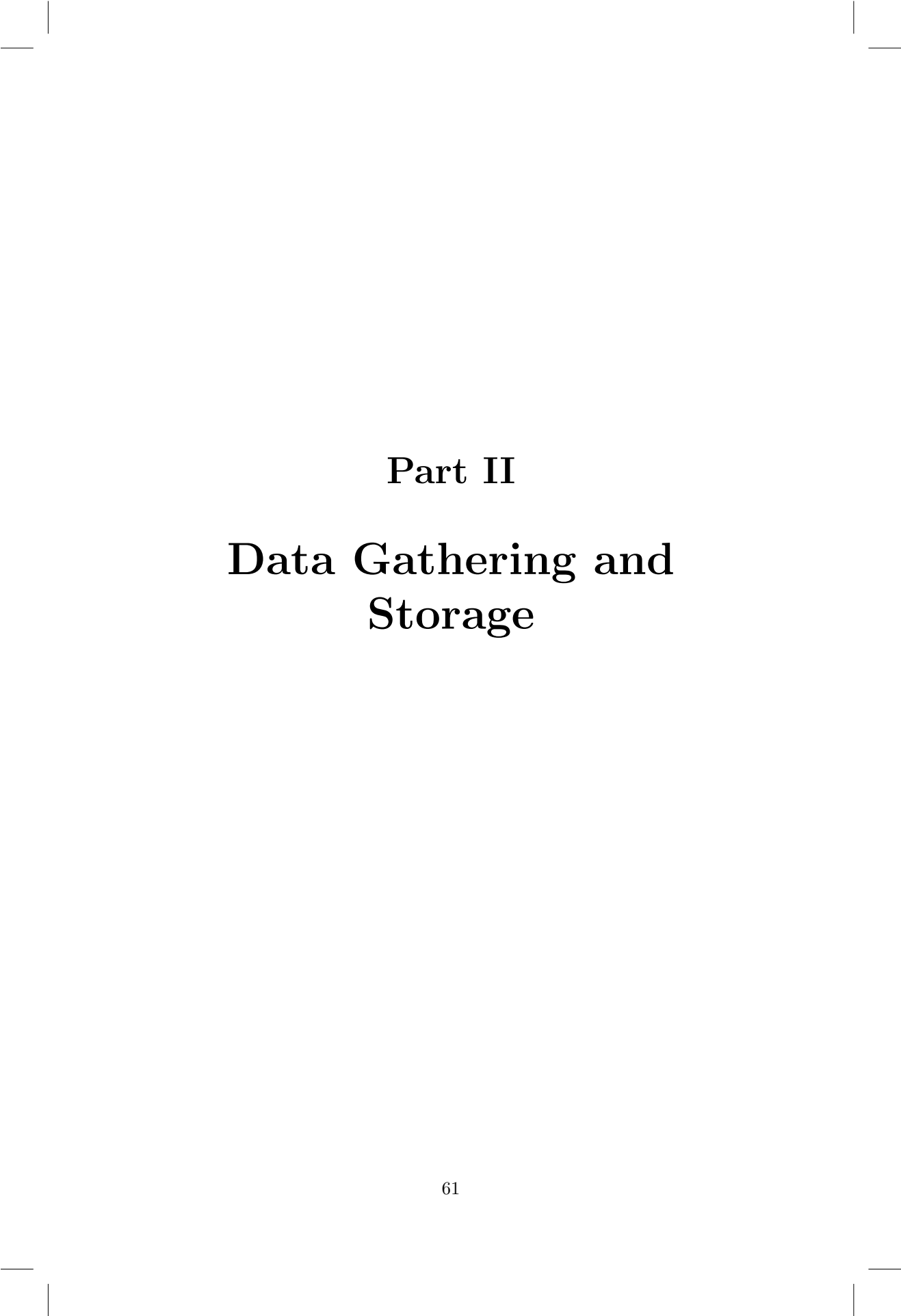
```
cd ~/Desktop
```

`cd`

As we just saw, to change the working directory in the shell can just use the `cd` (change directory) command.

`rm`

The `rm` command is similar to R's `unlink` command. It deletes files or directories.



## Part II

# Data Gathering and Storage



# 5

---

## *Storing, Collaborating, Accessing Files, Versioning*

---

### CONTENTS

A stumbling block to actually reproducing a piece of research is getting a hold of the datasets and the codebooks that describe the data used in an analysis. A useful way to overcome this problem is to keep your research on a cloud storage service (see ?). This chapter describes two different ‘cloud’ storage services—Dropbox and GitHub<sup>1</sup>—that you can use to make your research files easily accessible to others.

Not only do these services enable others to reproduce your research, they also have a number of benefits for your own research workflow. Researchers often face a number of data management issues that, beyond making their research difficult to reproduce, can make doing the initial research difficult.

First, there is the problem of **storing** the data so that it is protected against computer failure—virus infections, spilling coffee on your laptop, and so on. Storing data locally—on your computer—or on a flash drive is generally more prone to loss than on remote servers in the cloud.

Second, we may work on a project with different computers and other devices. For example, we may use a computer at work to run computationally intensive analysis, while editing our presentation document on a tablet computer while riding the train to work. Because of this we need to be able to **access** our files from multiple devices while in different locations. Also, we often need a way for our **collaborators** to access and edit research files.

Finally, we almost never create a data set or write a paper perfectly all at once. We may make changes and then realize that we liked an earlier version, or parts of an earlier version better. This is a particularly important issue in data management where we may transform our data in unintended ways and want to go back to an earlier version. Collaborative projects can have regular incidents of one author accidentally deleting something in a file that another author needed, for example. To deal with these issues we need to store our data in a system that has **version control**. Version control systems keep

---

<sup>1</sup>These services store your data on remote servers

track of changes we make to our files and allow us to access previous versions if we like.

All of these issues are not only issues for the original researchers, but also for independent researchers trying to reproduce the findings.

You can solve all of these problems in a couple of different ways using free or low cost cloud-based storage formats. In this chapter we will learn how to use Dropbox and GitHub for research file:

- storage,
- accessing,
- collaboration,
- version control.

---

## 5.1 Saving data in reproducible formats

Before getting into the details of cloud-based data storage for all of our research files, let's just consider what type of formats you should actually save your data in. A key issue for reproducibility is that others be able to not only get ahold of the exact data you used in your analysis, but be able to understand and use the data not only now, but in the future. Some file formats make this easier than others.

In general, for small to moderately-sized data sets<sup>2</sup> a plain-text format like comma-separated values (`.csv`) or tab-separated values (`.tsv`) can be a good way to store your data. These formats simply store a data set as a text file. A row in the data set is a line in the text file. Data is separated into columns with commas or tabs, respectively. These formats are not dependent on a specific program. Any program that can open text files can open them including a wide variety of statistical programs other than R. This helps future proof your research. Version control systems that track changes to text, like GitHub—are also very effective version control systems with these types of files.

To save data in a plain-text format use the `write.table` command. For example, to save a data frame called *Data* as a csv file type:

---

<sup>2</sup>I don't cover methods for storing and handling very large data sets—with high hundreds of thousands and more observations. For information on large data and R, not just storage, one place to look is this blog post by from RDataMining: <http://rdatamining.wordpress.com/2012/05/06/online-resources-for-handling-big-data-and-parallel-computing-in-r/> (posted 6 May 2012).

```
write.table(x = Data, file = "MyData.csv", sep = ",")
```

The `sep = ","` argument specifies that we want to use a comma to separate the values. For csv files you can use a modified version of this command called `write.csv`.

R is able to save data in a wide variety of other file formats, mostly through the *foreign* package in.

---

## 5.2 Storing your files in the cloud

In this book we'll cover two (largely) free<sup>3</sup> cloud storage services that allow you to store, access, collaborate on, and version control your research files. These services are Dropbox and GitHub. Though they both meet our basic storage needs, they do so in different ways and require different levels of effort to set up.

These two services are certainly not the only way to make your research files available. sResearch oriented services include the SDSC Cloud,<sup>4</sup> the Dataverse Network Project,<sup>5</sup> and RunMyCode.<sup>6</sup> Not all of these services enable us to store, access, collaborate on, and version control our code. Nonetheless, you might find them useful for making your research files available.

### 5.2.1 Dropbox

The easiest types of cloud storage for your research are services like Dropbox and Google Drive. These services typically involve a folder based on your computer's hard drive that is automatically synced with a similar folder on a cloud-based server. Typically you can sign up for the service for free and receive a limited amount of storage space (usually a few gigabytes, which should be plenty if your research is made up of text files.).

Most of these services not only store your data in the cloud, but also provide some way to share files and maybe even includes basic version control. I am going to focus on using Dropbox because it currently offers a complete set of features that allow you to store, version, collaborate, and access your data.

---

<sup>3</sup>Dropbox provides a minimum amount of storage for free, above which a fee is charged. GitHub lets you create publicly accessible repositories for free, but they charge for private repositories.

<sup>4</sup><https://cloud.sdsc.edu/hp/index.php>

<sup>5</sup><http://thedata.org/>

<sup>6</sup><http://www.runmycode.org/>

### 5.2.1.1 Version control

Dropbox has a simple version control system. Every time you save a document on Dropbox a new version is created. One the Dropbox website

### 5.2.1.2 Accessing Data

There are two similar, but importantly different ways to access data stored on Dropbox. All files stored on Dropbox have a URL address through which they can be access from computer connected to the internet. Some of these files can be easily loaded directly into R, while others must be manually (point-and-click) downloaded onto your computer and then loaded into R. The key factor is whether or not the files are located in your Dropbox's *Public* folder. Files in the *Public* folder can be downloaded directly into R. Files not in the *Public* folder have to be downloaded manually.<sup>7</sup>

Either way you find a file's URL address by first right-clicking on the file icon in you Dropbox folder. If the file is stored in the *Public* folder, you go to Dropbox then *Copy Public Link*. This copies the URL into your clipboard from where you can paste it into your R source code (or wherever). Once you have the URL you can load the file directly into R using the `read.table` command for dataframes (see Chapter 5) or the `source` command for source files (see Chapter 8).

If the file is not in your *Public* folder you also go to Dropbox after right-clicking. Then choose *Get Link*. This will open a webpage in your default web browser from where you can download the file. You can copy and paste the page's URL from your browser's address bar.

You can also get these URL links through the online version of your Dropbox. First log into the Dropbox website. When you hover your cursor over a file (or folder) name you will see a chain icon appear on the far right. Clicking on this icon will get you the link.

Storing files in the *Public* folder clearly makes replication easier because the files can be downloaded and run directly in R.

Note that you cannot save files through the URL link. You must save files in the Dropbox folder on your computer.

## 5.2.2 GitHub

Dropbox does a fine job of meeting our four basic criteria for reproducible data storage. GitHub meets these criteria and more.

GitHub was not explicitly designed to host research projects or even data. It was designed to host 'socially coded' computer programs. It built an interface on top of the git version control system that makes it easy relatively easy

---

<sup>7</sup>This is not completely true. It is possible to create a web scraper (see Chapter GET) that could download a data file from a file not in your *Public* folder. However, this is kind of a hassle and not practical, especially since the accessing files from the *Public* folder is so easy.

for a number of collaborators to work together to build a computer program. This seems very far from reproducible research.

However, remember that as reproducible researchers we are just building projects out of interconnected text files. This is exactly the same as computer programming. and like computer programmers, we need ways to store, version control, access, and collaborate on our text files. Because GitHub is very actively used by people with very similar needs (who are also really good programmers), the interface offers many highly developed and robust features for reproducible researchers.

As is usually the case, GitHub's added features mean that it takes a longer time than Dropbox to set up and become familiar. So we need good reasons to want to invest the time needed to learn GitHub rather than just sticking with Dropbox or a similar service. Here is a list of GitHub's key features relative to Dropbox for reproducible research:

- Git is directly integrated into RStudio projects (**RStudio** also supports the subversion version control system, but I don't cover that here).
- Dropbox's version control system only lets you see the file names, the times they were created, who created them, and revert back to specific versions. git tracks every change you make in a way that makes it relatively easy to find the version you want. The GitHub website and GUI programs for Mac and Windows provide nice interfaces for examining specific changes. You can also use the command line to see changes.
- Dropbox creates a new version every time you save a file, which can make it difficult to actually find the version you want. git's version control system only creates a new version when you tell it to.
- Dropbox does not merge conflicting versions of a file together. This can be annoying when you are collaborating on project and more than one author is making changes to documents. GitHub identifies conflicts and lets you reconcile them.
- The GitHub website has an "Issues" area where you can note and discuss issues you have while doing your research. Basically this is an interactive to-do list for your research project.

#### 5.2.2.1 Setting Up GitHub

There are a number of ways to set up GitHub on your computer. I will briefly cover both the command line version (available for Windows, Mac, and Linux) and the Graphical User Interface version currently available only for Windows and Mac.

COMPLETE



### 5.2.2.2 Storage in GitHub

### 5.2.2.3 Accessing in GitHub

### 5.2.2.4 Collaboration in GitHub

Repositories can have official collaborators. Public repositories can have unlimited collaborators. Anyone with a GitHub account can be a collaborator.

#### COMPLETE

Anyone with a GitHub account can make changes to files in a public repository on the repository's website. Simply click the **Edit** button above the file and make edits. If the person making the edits is not the a repository collaborator, their edit will be sent to the repository's owner for approval.<sup>8</sup> This is a useful way for independent researchers to catch errors and directly address them.

#### *Branches*

#### *Syncing repository*

### 5.2.2.5 Version Control in GitHub

GitHub's version control system is much more comprehensive than Dropbox's. However, it also has a steeper learning curve.

#### *Reverting to an old version of a file*

You can use the `git checkout` command to revert to a previous version of a document, because you accidentally deleted something important or made other changes you don't like. To 'checkout' a particular version of a file type:

```
git checkout COMMITREF FILENAME
```

Now the previous version of the file is in your working directory, where you can commit it as usual.

Let's break down the code. **FILENAME** is the name of the file that you want to change<sup>9</sup> and **COMMITREF** is the reference that git gave to the commit you want to revert back to. The reference is easy to find and copy in GitHub. On the file's GitHub page click on the **History** button. This will show you all of the commits. By clicking on **Browse Code** you can see what the file at that commit looks like. Above this button is another with a series of numbers and letters. This is the commit's SHA (Secure Hash Algorithm). For our purposes, it is the commit's reference number. Click on the **Copy SHA** button to the left

<sup>8</sup>This is called a **pull** in git terminology

<sup>9</sup>If it is in a repository's subdirectory you will need to include this in the file name.

of the SHA to copy it. You can then paste it as an argument to your `git checkout` command.

*More Practice with Command Line GitHub*

If you want more practice setting up GitHub in the command line, GitHub and the website Code School have an interactive tutorial that you might find interesting. You can find it at: <http://try.github.com/levels/1/challenges/4>.



# 6

---

## *Gathering Data with R*

---

### CONTENTS

There are many practical issues involved in gathering data that can make replication easier or harder. As with all of the steps in this book: document everything. Replication will be easier if your documentation—source code—can be understood and executed by a computer. Of course there are data gathering situations that simply require manually pointing and clicking, talking with subjects in an experiment, and so on. The best we can do in these situations is just describe our data gathering process in detail CITE. Nonetheless, R's automated data gathering capabilities are extensive and often under utilized. Learning how to take full advantage of them greatly increases replicability and can even save researchers considerable time and effort.

---

### 6.1 Organize Your Data Gathering: Make files

MOVE TO END OF CHAPTER Before getting into the details of using R to automate data gathering, let's start from where all data gathering should start: a plan to organize the process. Clearly organizing your data gathering process from the start of a research project improves the possibility of replicability and can save significant effort over the course of the project.

A key principle of replicable data gathering with R, like replicable research in general is segmenting the process into discrete files that can be run by a common Make file. The Make file's output is the data set(s) that we use in the statistical analyses. There are two types of files that the Make file runs: data clean up files and merging files. Data clean up files bring raw (the rawer the better) individual data sources into R and transform them into something that can be merge with data from the other sources. Some of the R tools for data clean up will be covered in Chapter ?? In this chapter we mostly cover the ways to bring raw data into R. We don't explicitly cover the process of merging data sets together in this book. Merging files are executed by the Make file after it runs the clean up files.

Data gathering Make files usually only need one or two commands `setwd`

and `source`. As we talked about in Chapter ??, `setwd` simply tells R where to look for and place files. `source` tells R to run code in an R source code file.<sup>1</sup> Lets see what a Data make file might look like for our example project (see Figure ??).

```
# Example Make file
setwd("~/ExampleProject/Data/")

# Gather and clean up raw data files.
source("/GatherSource/IndvDataGather/Gather1.R")

source("/GatherSource/IndvDataGather/Gather2.R")

# Merge cleaned data files into object CleanedData
source("GatherSource/MergeData.R")

# Save cleaned & merged Data as MainData.csv
write.csv(CleandedData, file = "/DataFiles/MainData.csv")
```

Here we saved the output data set *CleanData* as a `.csv` formatted file using the `write.csv` command. In our example project, the file *MainData.csv* will be the main file we use for statistical analysis.

You can of course save your data in a wide variety of other formats. To save your data in another plain-text format use the `write.table` command. You can also save all of the objects in your workspace using the `save.image` command.

---

## 6.2 Importing locally stored data sets

Plain text file based data stored on your computer can be loaded into R using the `read.table` command. This command will take the file and turn it into a data frame object. For example, imagine that we have a data set called *Data1.csv* in our root directory. We load it into an R object called *Data* like this:

---

<sup>1</sup>The `source` command is used more in the Chapter ??.

```
Data <- read.table(file = "~/Data1.csv", sep = ",",  
                  header = TRUE)
```

In this code, `file` is clearly where we specify the location and file name of the data set we want to open. `sep` is where we tell R how the data columns are separated. In this case they are separated by commas. Finally, `header = TRUE` tells R that we want to treat the first row of the data set as the variable names.

If you are using RStudio you can do the same thing with drop down menus. To open a plain-text data file click on **Workspace** → **Import Dataset...** → **From Text File...**. In the box that pops up, specify the separator, whether or not you want the first line to be treated as variable labels and other options. This is initially easier than using `read.table`. But it is less reproducible.

To aid reproducibility, locally stored data should include careful documentation of where the data came from and how, if at all, it was transformed before we load it into R.

### 6.2.1 Single files

### 6.2.2 Looping through multiple files

---

## 6.3 Importing data sets from the internet

There are many ways to import data that is stored on the internet directly into R. We have to use different methods depending on where and how the data is stored.

### 6.3.1 Data from non-secure (http) URLs

Importing data into R that is located at a non-secure URL (`http`) is straightforward provided that:

- the data is stored in a simple format, e.g. plain-text,
- the file is not embedded in an HTML website.

We have discussed the first issue in detail. You can determine if the data file is embedded in a website by opening the URL. If you only see the raw plain-text data, you are probably good to go.

Now simply include the URL as the file name in your `read.table` command. Imagine the data we wanted to load was at this URL `http://www.AWebHost.Data1.csv`.

```
Data <- read.table(file = "http://www.AWebHost.Data1.csv",
  sep = ",", header = TRUE)
```

### 6.3.2 Data from secure (https) URLs

### 6.3.3 Compressed data stored online

Sometimes data files can be very large, making them difficult to store and download without compressing them. There are a number of compression methods such as Zip and tar archives. Zip files have the extension `.zip` and tar archives use extensions such as `.tar` and `.gz`. In most cases<sup>2</sup> we can easily download, decompress, and create dataframe objects from these files directly in **R**.

To do this we need to:<sup>3</sup>

- create a temporary file with `tempfile` to store the zipped file which we will remove with the `unlink` command at the end,
- download the file with `download.file`,
- decompress the file with one of the `connections` commands in *base R*,<sup>4</sup>
- read the file with `read.table`.

The reason that we have to go through so many extra steps is that compressed files are more than just a single file, but can contain more than one file as well as metadata.

Let's download a compressed file called `uds_summary.csv` from `?`. It is in a zipped file called `uds_summary.csv.gz`. The file's URL address is `http://www.unified-democracy-scores.org/files/uds_summary.csv.gz`, but I shortened it<sup>5</sup> to `http://bit.ly/S0vxx2` to cut down on the text I have to include in the code.

---

<sup>2</sup>Some formats that require the *foreign* package to open are more difficult. This is because functions such as `read.dta` for opening Stata `.dta` files only accept file names or URLs as arguments, not connections, which we create for unzipped files.

<sup>3</sup>The description of this process is based on a Stack Overflow comment by Dirk Eddelbuettel (see <http://stackoverflow.com/questions/3053833/using-r-to-download-zipped-data-file-extract-and-import-data?answertab=votes\#tab-top>, accessed 16 July 2012).

<sup>4</sup>To find a full list of commands type `?connections` in to the **R** console.

<sup>5</sup>I used the website [bitly.com](http://bitly.com) to shorten the URL.

```
# For simplicity, store the URL in an object called 'url'.
url <- "http://bit.ly/S0vxk2"

# Create a temporary file called 'temp' to put the zip file into.
temp <- tempfile()

# Download the compressed file into the temporary file.
download.file(url, temp)

# Decompress the file and convert it into a dataframe
# class object called 'data'.
data <- read.csv(gzfile(temp, "uds_summary.csv"))

# Delete the temporary file.
unlink(temp)
```

### 6.3.4 Data APIs & feeds

There are growing number of packages that can gather data directly from their sources and import them into R. Needless to say, this is great for reproducible research since it not only makes the data gathering process easier (you don't have to download many of Excel files and fiddle around with them before even getting the data into R, but it also makes replicating the data gathering process much more straightforward. Some examples include:

- The *openair* package, which beyond providing a number of tools for analysing air quality data also has the ability to directly gather data directly from sources such as Kings College London's London Air (<http://www.londonair.org.uk/>) database with the `importKCL` command.

---

## 6.4 Basic web scraping

### 6.4.1 Scraping tables

### 6.4.2 Gathering and parsing text





# 7

---

## *Preparing Data for Analysis*

---

### CONTENTS

Once we have gathered the raw data that we want to include in our statistical analyses we generally need to clean it and merge it into a single data files CAWELY QOUTE ABOUT HOW IT CAN BE BAD TO USE DATA FROM DIFFERENT DATA FRAMES. This chapter covers some of the basics of how to clean data files and merge them using R.

The two main suggestions for cleaning and merging data are to:

- always versions of the original–non-cleaned–data in as raw a state as possible,
- again document everything.

It's a good idea to keep data your original data in as raw a version as possible because it makes reconstructing the steps you took to create your data set easier. Also, while cleaning and merging your data you may transform it in an unintended way, for example, accidentally deleting some observations that you had intended to keep. Having the raw data makes it easy to go back and correct your mistake. Documenting everything also helps you achieve these two goals. Also it makes updating the data set easier if, for example, new data becomes available. MAYBE EXPLAIN MORE.

If you are very familiar with data transformations in R you may want to skip onto the next chapter.

---

## 7.1 Cleaning data for merging

---

## 7.2 Sorting data

---

## 7.3 Merging data sets

---

## 7.4 Subsetting data

**Part III**

**Analysis and Results**



# 8

---

## *Statistical Modelling and knitr*

---

### CONTENTS

---

#### 8.1 Incorporating analyses into the markup

##### 8.1.1 Full code in the main document

###### 8.1.1.1 $\LaTeX$

###### 8.1.1.2 Markdown

##### 8.1.2 Showing code & results inline

Sometimes we want to have some R code or output to show up in the text of our documents. We may want to include stylized code in our text when we discuss how we did an analysis. We may want to report the mean of some variable in our text.

###### 8.1.2.1 LaTeX

###### *Inline static code*

If we just want to include a code snippet in our text we can simply use the  $\LaTeX$  command `\tt`. This sets our text to ‘typewriter’ font, the standard font for inline code in  $\LaTeX$  (I use it in this book, as you have probably noticed).

###### *Inline dynamic code*

If we want to dynamically show the results of some R code in our text we can use the `\Sexpr` command. This is a pseudo  $\LaTeX$  command. Its structure is more like a  $\LaTeX$  command’s structure than `knitr` in that we enclose our R code in curly brackets (`{}`) rather than the usual `<<>>= . . . @` syntax for code chunks.

For example, imagine that we wanted to include the mean—591—in the text of our document. The *rivers* numeric vector, loaded by default in R, has the length of 141 major rivers recorded in miles. We can simply use the `mean`

command to find the mean and the `round` command to round it to the nearest whole number:

```
round(mean(rivers), digits = 0)

## [1] 591
```

To have just the output show up inline with the text of our document we would type something like:

```
The mean length of 141 major rivers in North America
is \Sexpr{round(mean(rivers), digits = 0)} miles.
```

This will produce the sentence:

The mean length of 141 major rivers in North America is 591 miles.

### 8.1.2.2 Markdown

#### *Inline static code*

To include static code inline in an R Markdown (and regular Markdown) document we enclose the code in single backticks (```). For example typing

```
This is example R code: `MeanRiver <- mean(rivers)`.
```

produces:<sup>1</sup>

```
This is example R code: MeanRiver <- mean(rivers) .
```

#### *Inline dynamic code*

Including dynamic code in the body of your R Markdown text is similar to including static code. The only difference is that we put the letter `r` after the first single backtick. For example:

<sup>1</sup>The exact look of the text depends on the CSS style file you are using.

```
`r mean(rivers)`
```

will include the mean value of the *rivers* vector in the text of your Markdown document.

### 8.1.3 Sourcing R code from another file

There are a number of reasons that you might want to have your R source code located in a separate file from your markup even if you plan to compile them together with *knitr*.

First, it can be unwieldy to edit both your markup and long R source code chunks in the same document, even with RStudio's handy *knitr* code collapsing and chunk management options. There are just too many things going on in one document.

Second, you may want to use the same code in multiple documents—an article and presentation for example. It is nice to not have to copy and paste the same code into multiple places, but have multiple documents link to the same source code. Plus if you make changes to the source code, these changes will automatically be made across all of your presentation documents. You don't need to make the same changes multiple times.

Third, other researchers trying to replicate your work might only be interested in specific parts of your analysis. If you have the analysis broken into separate and clearly labeled files it is easier for these researchers to find the specific bits of code that they are interested compared to digging through long markup files.

#### 8.1.3.1 Source from a local file

Usually in the early stages of research you may want to source analysis files located on your computer. Doing this is simple. The *knitr* syntax is the same as above. The only change is that instead of writing all of our code in the chunk we save it to its own file and use the `source` command in *base* R to access it. For example:

#### 8.1.3.2 Source from a non-secure URL ([http](#))

Sourcing from your local computer is fine if you are working alone and do not want others to access your code. Once you start collaborating and generally wanting people to be able to replicate your code, you need to use another method.<sup>2</sup>

---

<sup>2</sup>You can make the replication code accessible for download and either instruct others to change the working directory to the replication file or have them change the directory information as necessary. However, this usually just adds an extra complicating step that



The simplest solution to these issues is to host the replication code in your Dropbox public folder. You can find the file's public URL the same way we did in Chapter 6. Now use the `source` command the same way as before. For example:

### 8.1.3.3 Source from a secure URL (https)

If you are using GitHub or another service that uses secure URLs to host your analysis source code files the steps are generally the same, but you need to use the `source_url` command in the *devtools* package. For GitHub based source code we find the file's URL the same way we did in Chapter ???. Remember to use the URL for the *raw* version of the file. I have a short script hosted on GitHub for creating a scatter plot from data in R's *cars* data set. The script's shortened URL is <http://bit.ly/Ny1n6b>.<sup>3</sup> To run this code and create the scatter plot using `source_url` we simply type:

```
# Load library
library(devtools)

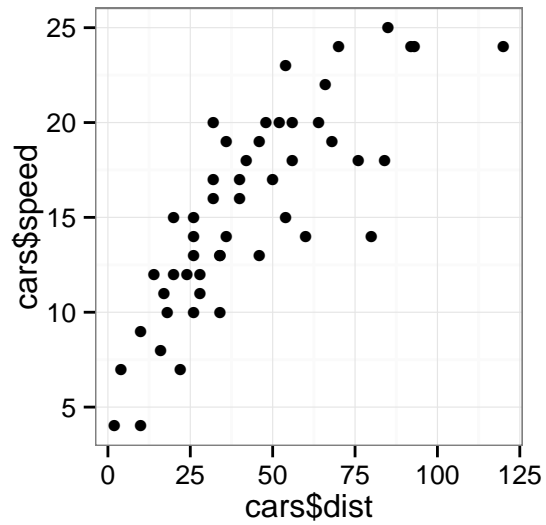
# Create object to hold URL
URL <- "http://bit.ly/Ny1n6b"

# Run the source code to create the scatter plot
source_url(URL)
```

---

makes replication harder. It is also a pain if you are collaborating and each author has to constantly change the directories.

<sup>3</sup>The original URL is at <https://raw.githubusercontent.com/christophergandrud/christophergandrud.github.com/master/SourceCode/CarsScatterExample.R>. This is very long, so I shortened it using bitly (see <http://bitly.com>). You may notice that the shortened URL is not secure. However, it does link to original secure https URL.



---

## 8.2 Saving output objects for future use

---

### 8.3 Literate Programming: Including highlighted syntax in the output

#### 8.3.1 $\text{\LaTeX}$

#### 8.3.2 Markdown/HTML

---

## 8.4 Debugging



# 9

---

## *Showing Results with Tables*

---

### CONTENTS

Graphs and other visual methods, discussed in the next chapter, can often be a more effective way to present results than tables.<sup>1</sup> Nonetheless, tables of results, descriptive statistics, and so on can sometimes be an important part of communicating research findings.

Creating tables by hand can be tedious no matter what program you are using to type up your results. Even more tedious is making changes to hand-created tables when you make changes to your data and models. Creating these tables can actually introduce new errors—post-analysis!—if you incorrectly copy what is in your R output. This is a very real possibility. The mind can go numb doing that sort of work. Also, creating tables by hand is not very reproducible.

Fortunately, we don't actually need to create tables by hand. There are many ways to have R do the work for us. The goal of this chapter is to learn how to how to **automate table creation** for documents produced with both  $\text{\LaTeX}$  and Markdown/HTML. There are a number of ways to turn R objects into tables written in  $\text{\LaTeX}$  or HTML markup. In this chapter I mostly focus on the `xtable` and `texreg` packages. `xtable` can create tables for both of these markup languages. `texreg` only produces output for  $\text{\LaTeX}$ . `knitr` allows us to incorporate these tables directly into our documents.

**Warning:** Automating table creation removes the possibility of adding errors to our analyses by incorrectly copying R output, which is a big potential problem in hand-created tables. Be warned, it is not an error free process. We could easily create inaccurate tables through coding errors. For example, we may incorrectly merge together columns in so that our id variables no longer match the data they are supposed to.

So, as always, it is important to 'eyeball' the output. Does it make sense? If we picked a couple values in the R output do they match what is in our final table? If not, we need to go back to the code and see where things have gone wrong. With that caveat, let's start making tables.

---

<sup>1</sup>This is especially true of the small-print, high-density coefficient estimate tables that are sometimes descriptively called 'train schedule' tables.

---

## 9.1 Table Basics

Before getting into the details of how to create tables from R objects we need to first learn how generic tables are created in  $\text{\LaTeX}$  and Markdown/HTML.

### 9.1.1 Tables in $\text{\LaTeX}$

### 9.1.2 Tables in Markdown/HTML

We can use `xtable` and the `print` command to also create tables for Markdown and HTML documents. Instead of setting the `type` argument to `'latex'` we simply put it to `'html'`.

---

## 9.2 Creating tables from R objects

### 9.2.1 `xtable` & `texreg` basics with supported class objects

#### 9.2.1.1 `xtable` for $\text{\LaTeX}$

#### 9.2.1.2 `xtable` for Markdown

### 9.2.2 `xtable` with non-supported class objects

`xtable` and other commands in similar packages are very convenient for making tables from objects in supported classes.<sup>2</sup> With supported class objects `xtable` knows where to look for the vectors containing the things-coefficient names, standard errors, and so on—that it needs to create the table. With unsupported classes, however, it doesn't know where to look for these things. We need to help it out.

`xtable` does have a way of dealing with `matrix` and `data frame` class objects. The rows of these objects become the rows of the table and the columns become the table columns. So, to create tables with non-supported class objects we need to

1. find and extract the information from the unsupported class object that we want in the table,
2. convert this information into a matrix or data frame where the rows and columns of the object correspond to the rows and columns of the table that we want,
3. use `xtable` with this object to create the table.

---

<sup>2</sup>To see a full list of classes that `xtable` supports type `methods(xtable)` into the R console.

Imagine that we want to create a results table showing the covariate names, coefficient means, and quantiles for marginal posterior distributions from a Bayesian normal linear regression using the `zelig` command<sup>3</sup> and data from the *swiss* data frame.<sup>3</sup> We run our model:

```
# Load required library
library(Zelig)

# Run model
NBModel <- zelig(Examination ~ Education, model = "normal.bayes",
                 data = swiss, cite = FALSE)

# Find NBModel's class
class(NBModel)

## [1] "MCMCZelig"
```

Using the `class` command we found that the model output object is an `MCMCZelig` class object. This class is not supported by `xtable`. If we try to create a summary table called *NBTable* of the results we will get the following error:

```
## Error: no applicable method for 'xtable' applied to an object of
class "MCMCZelig"
```

With unsupported class objects we have to create the summary ourselves and extract the things that we want from it manually. This is where a good knowledge of vectors comes in handy.

First, let's create a summary of our output object *NBModel*:

```
NBModelSum <- summary(NBModel)
```

We created a new object of the class `summary.MCMCZelig`. We're still not there yet as this object contains not just the covariate names and so on but also information we don't want to include in our results table like the formula that we used. The second step is to extract a matrix from inside *NBModelSum*

<sup>3</sup>This data frame is loaded by default.

called *summary* with the component selector (`$`). This matrix is where the things we want in our table are located. I find it easier to work with data frames, so we'll also convert the matrix into a data frame.

```
NBSumDataFrame <- data.frame(NBModelSum$summary)
```

Here is what our model results data frame looks like:

```
##              Mean      SD    X2.5.    X50.    X97.5.
## (Intercept) 10.1397 1.31673  7.5579 10.1566 12.7058
## Education    0.5786 0.09118  0.3963  0.5781  0.7609
## sigma2       34.9703 7.81260 22.9567 33.8782 53.2172
```

Now we have a data frame object that `xtable` can handle. After a little cleaning up (see the chapter's source code for more details) we can use *NBSumdata frame* with `xtable` as before to create the following table:

|             | Mean  | 2.5%  | 50%   | 97.5% |
|-------------|-------|-------|-------|-------|
| (Intercept) | 10.14 | 7.56  | 10.16 | 12.71 |
| Education   | 0.58  | 0.40  | 0.58  | 0.76  |
| sigma2      | 34.97 | 22.96 | 33.88 | 53.22 |

**TABLE 9.1**

Coefficient Estimates Predicting Examination Scores in Swiss Cantons (1888)  
Found Using Bayesian Normal Linear Regression

It may take a bit of hunting to find what you want, but a similar process can be used to create tables from objects of virtually any class.<sup>4</sup> Hunting for what you want is generally easier by clicking on the object in RStudio's workspace pane.

### 9.2.3 Basic knitr syntax for tables

So far we have only looked at how to create  $\text{\LaTeX}$  and HTML tables from R objects. How can we knit these tables into our presentation documents?

The most important `knitr` chunk option for showing the markup created

<sup>4</sup>This process can also be used to create graphics.

by these packages as tables is **results**. The **results** option can have one of three values:

- 'markup',
- 'asis',
- 'hide'.

**hide** clearly hides the results of whatever we have in our code chunk; no results show up. To include tables we create from R objects in our L<sup>A</sup>T<sub>E</sub>X or Markdown output we should set **results='asis'** or **results='markup'**. **asis** simply includes the writes the raw output in the presentation document. **markup** uses an output hook to mark up the results. NEED TO EXPLAIN

---

### 9.3 Tables with apsrtable





# 10

---

## *Showing Results with Figures*

---

### CONTENTS

---

#### 10.1 Including graphics

---

#### 10.2 Basic knitr figure options

---

#### 10.3 Creating figures with plot and ggplot2

---

#### 10.4 Animations

---

#### 10.5 Motion charts and basic maps with GoogleVis

---





# **Part IV**

## **Presentation Documents**



# 11

---

## *Presenting with L<sup>A</sup>T<sub>E</sub>X*

---

### CONTENTS

---

#### 11.1 The Basics

All commands in L<sup>A</sup>T<sub>E</sub>X start with a \

##### 11.1.1 Editors

As I mentioned earlier, RStudio is an fully functional L<sup>A</sup>T<sub>E</sub>X editor as well as an integrated development environment for R. Of course it is oriented towards combining R and L<sup>A</sup>T<sub>E</sub>X. If you want to create a new L<sup>A</sup>T<sub>E</sub>X document you can click **File** → **New** → **R Sweave**.

Remember from Chapter ?? that R Sweave files are basically L<sup>A</sup>T<sub>E</sub>X files that can include *knitr* code chunks. You can compile R Sweave files like regular L<sup>A</sup>T<sub>E</sub>X files in RStudio even if they do not have code chunks. If you use another program to compile them you might need to change the file extension from `.Rnw` to `.tex`.

##### 11.1.2 The header & the body

All L<sup>A</sup>T<sub>E</sub>X documents require a header. The header goes before the body of the document and specifies what type of presentation document you are creating—an article, a book, a slideshow, and so on. L<sup>A</sup>T<sub>E</sub>X refers to these as classes. You also can specify what style it should be formatted in and load any extra packages you may want to use to help you format your document.<sup>1</sup>

The header is followed by the body of your document. You tell L<sup>A</sup>T<sub>E</sub>X where the body of your document starts by typing `\begin{document}`. The very last line of your document is usually `\end{document}`, indicating that your document has ended. When you open a new R Sweave file in RStudio it

---

<sup>1</sup>The command to load a package in L<sup>A</sup>T<sub>E</sub>X is `\usepackage`. For example, if you include `\usepackage{url}` in the header of your document you will be able to specify URL links in the body with the command `\url{SOMEURL}`.

creates an article class document with a very simple header and body like this:

```
\documentclass{article}  
\begin{document}  
\end{document}
```

### 11.1.1.3 Headings

### 11.1.1.4 Footnotes & Bibliographies

#### 11.1.1.4.1 Footnotes

Plain, non-bibliographic footnotes are easy to create in L<sup>A</sup>T<sub>E</sub>X. Simply place `\footnote{` where you would like the footnote number to appear in the text. Then type in the footnote's text and of course remember to close it with a `}`. L<sup>A</sup>T<sub>E</sub>X does the rest, including formatting and numbering.

#### 11.1.1.4.2 Bibliographies

##### *Citing R Packages with BibTeX*

Researchers are pretty good about consistently citing others' articles and data. However, citing the R packages used in an analysis is very inconsistent. This is unfortunate not only because correct attribution is not being given but also because it makes reproducibility harder because it obscures important steps that were taken in the research process. Fortunately, R actually includes the tools to quickly generate citations, including the version of the package you are using. It can also add them directly to an existing bibliography file.

You can automatically create citations for R packages using the `citation` command in *base* R. For example if we want the citation information for the `Zelig` package we would simply type:

```
citation("ggplot2")  
  
## Warning: incomplete final line found on '/Library/Frameworks/R.framework/Versions/2.15/  
  
##  
## To cite ggplot2 in publications, please use:  
##  
## H. Wickham. ggplot2: elegant graphics for data
```

```
## analysis. Springer New York, 2009.
##
## A BibTeX entry for LaTeX users is
##
## @Book{,
##   author = {Hadley Wickham},
##   title = {ggplot2: elegant graphics for data analysis},
##   publisher = {Springer New York},
##   year = {2009},
##   isbn = {978-0-387-98140-6},
##   url = {http://had.co.nz/ggplot2/book},
## }
```

This gives us both the plain citation as well as the BibTeX version for use in  $\text{\LaTeX}$  and MultiMarkdown documents. If you only want the BibTeX version of the citation we can use the `toBibtex` command in the *utils* package.

```
toBibtex(citation("ggplot2"))

## Warning: incomplete final line found on '/Library/Frameworks/R.framework/Versions/2.15/

## @Book{,
##   author = {Hadley Wickham},
##   title = {ggplot2: elegant graphics for data analysis},
##   publisher = {Springer New York},
##   year = {2009},
##   isbn = {978-0-387-98140-6},
##   url = {http://had.co.nz/ggplot2/book},
## }
```

You can append the citation to your existing BibTeX file using the `sink` command in *base R*. This command diverts our output and/or the messages to a file. For example, imagine that our existing BibTeX file is called `biblio.bib`. To add the *Zelig* package citation:

```
# Divert output to biblio.bib
sink(file = "biblio.bib",
      append = TRUE, type = c("output"))
```



```

    )
  toBibtex(citation("ggplot2"))
  sink()

```

This places the citation at the end of our `biblio.bib` file. It is very important to include the argument `append = TRUE`. If you don't you will erase the existing file. The argument `type = c("output")` tells R to include only the output, not the messages.

An even faster way to add citations to a bibliography is with `write.bibtex` command in the *knitcitations* package. To add the *Zelig* citation to our `biblio.bib` file we only need to enter:

```

# Load package
library(knitcitations)

# Write Zelig citation and
# to biblio.bib
write.bibtex(entry = c("ggplot2"),
             file = "bibliography.bib", append = TRUE)

```

Note, you will likely only want to append the citations once. Otherwise your bibliography document will grow with redundant information every time you run this command.

The *knitr* package can also create BibTeX bibliographies for R packages using the `write.bib` command. To use this command you list the packages whose citation details you want to include in a specified file. The command currently does not have the ability to append the citations to an existing file, but instead writes them to a new file.

---

## 11.2 Presentations with Beamer

You can make slideshow presentations with L<sup>A</sup>T<sub>E</sub>X. FILL IN WITH INTRO

*knitr* largely the works the same way in in L<sup>A</sup>T<sub>E</sub>Xslideshows as it does in article or book class documents. There are a few differences to look out for.

### *Slide frames*

A quick way to create each Beamer slide is to use the `frame` command:

```
\frame{  
}
```

If you want to include highlighted *knitr* code chunks on your slides you should add the **fragile** option to the **frame** command.<sup>2</sup> Here is an example:

```
\begin{frame}[fragile]  
  An example fragile frame.  
\end{frame}
```

### Results

By default *knitr* hides the results of a code chunk. If you want to show the results in your slideshow simply set the **results** option to **'asis'**.

---

<sup>2</sup>For a detailed discussion of why you need to use the **fragile** option with the `verbatim` environment that *knitr* uses to display highlighted text in  $\text{\LaTeX}$  documents see this blog post by Pieter Belmans: <http://pbelmans.wordpress.com/2011/02/20/why-latex-beamer-needs-fragile-when-using-verbatim/>.



# 12

---

## *Large L<sup>A</sup>T<sub>E</sub>X Documents: Theses, Books, & Batch Reports*

---

### CONTENTS

In the previous chapter we learned the basics of how to create L<sup>A</sup>T<sub>E</sub>X documents to present our research findings. So far we have only covered basic short documents, like articles. For longer and more complex documents like books we can take advantage of L<sup>A</sup>T<sub>E</sub>X and *knitr* options that allow us to separate our files into manageable pieces. The pieces are usually called , which are combined using a .

These methods can also be used in the creation of : documents that present results for a selected part of a data set. For example, a researcher may want create individual reports of answers to survey questions from interviewees with a specific age. In this chapter we will rely on *knitr* and shell scripts to create batch reports.

---

### 12.1 Planning large documents

Before discussing the specifics of each of these methods, it's worth taking some time to carefully plan the structure of our child and parent documents.

#### 12.1.1 Planning theses and books

Books and theses have a natural parent-child structure, i.e. they are single documents comprise of multiple chapters. They often include other child-like features such as title pages, bibliographies, figures, and appendices. We could include most of these features directly into one markup file. Clearly this file would become very large and unwieldy. It would be difficult to find one part or section to edit. If of your presentation markup are are difficult to find, they are difficult to reproduce.

### 12.1.2 Planning batch reports

COMPLETE

---

## 12.2 Combining Chapters

We will cover three methods for including child documents into our parent documents. The first is very simple and uses the  $\text{\LaTeX}$  command `\input`. The second using *knitr* is slightly more complex, but gives us much more flexibility. The final method is a special case of `\input` that uses the command line program to convert and include child documents written in non- $\text{\LaTeX}$  markup languages.

### 12.2.1 Parent documents

*knitr* global options

*knitr* global chunk options and package options should be set at the beginning of the parent document if you want them to apply to the entire presentation document.

### 12.2.2 Child documents

*Child documents in the same markup language*

*Child documents in a different markup language*

Because *knitr* is able to run not only R code but also bash programs, you can use the command line program to convert child documents that are in a different markup language into the primary markup language you are using for your document. If you have Pandoc installed on your computer,<sup>1</sup> you can call it directly from your parent document including your Pandoc commands in a code chunk with the `engine` option set to either `'bash'` or `'sh'`.<sup>2</sup>

For example, the ?? part of this book is written in Markdown. The file is called *StylisticConventions.md*. It was simply faster to write the list of conventions using the simpler Markdown syntax than  $\text{\LaTeX}$ , which as we saw has a more complicated way of creating lists. However, I want to include this list in my  $\text{\LaTeX}$  produced book. Pandoc can convert the Markdown document into a  $\text{\LaTeX}$  file. This file can then be input into my main document with the  $\text{\LaTeX}$  command `\input`.

Imagine that my parent and *StylisticConventions.md* documents are in

---

<sup>1</sup>Pandoc installation instructions can be found at: <http://johnmacfarlane.net/pandoc/installing.html>.

<sup>2</sup>Alternatively you can run Pandoc in R using the `system` command.

the same directory. In the parent document I add a code chunk with the options `echo=FALSE` and `results='hide'`. In this code chunk I add the following command to convert the Markdown syntax in *StylisticConventions.md* to L<sup>A</sup>T<sub>E</sub>X and save it in a file called *StyleTemp.tex*.

```
pandoc StylisticConventions.md -f markdown \  
-t latex -o StyleTemp.tex
```

The options `-f markdown` and `-t latex` tell Pandoc to convert *StylisticConventions.md* from Markdown to L<sup>A</sup>T<sub>E</sub>X syntax. `-o StyleTemp.tex` instructs Pandoc to save the resulting L<sup>A</sup>T<sub>E</sub>X markup to a new file called *StyleTemp.tex*.

I only need to include a backslash (`\`) at the end of the first line because I wanted to split the code over two lines. The code wouldn't fit on this page otherwise. The backslash tells the shell not to treat the following line as a different line. Unlike in R, the bash shell only reads recognizes a command's arguments if they are on the same line.

After this code chunk we need to tell our parent document to include the converted text. To do this we follow the code chunk with the `input` command like this:

```
\input{StyleTemp.tex}
```

Note that using this method to include a child document that needs to be knit will require extra steps not covered in this book.

---

## 12.3 Creating Batch Reports

### 12.3.1 stich



# 13

---

## *Presenting on the Web and Beyond with Markdown/HTML*

---

### CONTENTS

---

#### 13.1 The Basics

##### 13.1.1 Headings

Headings in Markdown are extremely simple. To create a line in the style of the topmost heading—maybe a title—just place one hash mark (#) at the beginning of the line. The second tier heading just gets two hashes (##) and so on. You can also put the hash mark(s) at the end of the heading, but this is not necessary.



### 13.1.2 Footnotes and bibliographies with MultiMarkdown

### 13.1.3 Math

### 13.1.4 Drawing figures with CSS

---

## 13.2 Simple webpages

### 13.2.1 RPubS

### 13.2.2 Hosting webpages with Dropbox

---

## 13.3 Presentations with Slidify

---

## 13.4 Reproducible websites

### 13.4.1 Blogging with Tumblr

### 13.4.2 Jekyll-Bootstrap and GitHub

see <http://jfisher-usgs.github.com/r/2012/07/03/knitr-jekyll/>

### 13.4.3 Jekyll and Github Pages

---

## 13.5 Using Markdown for non-HTML output with Pandoc

Markdown syntax is very simple. So simple, you may be tempted to write many or all of your presentation documents in Markdown. This presents the obvious problem of how to convert your markdown documents to other markup languages if, for example, you would want to create a  $\text{\LaTeX}$ formatted PDF.

Pandoc can help solve this problem. Pandoc is a command line program that can convert files written in Markdown, HTML,  $\text{\LaTeX}$ , and a number of other markup languages<sup>1</sup> to any of the other formats.

To use Pandoc first install it by following the instructions at <http://johnmacfarlane.net/pandoc/installing.html>. Luckily you do not need to open a shell window in addition to R to run Pandoc. Instead you can run all Pandoc commands in R with the `system` command.

For example,

---

<sup>1</sup>See the Pandoc website for more details: <http://johnmacfarlane.net/pandoc/>

# 14

---

## *Going Beyond the Book*

---

### CONTENTS

#### COMPLETE

---

### 14.1 Licensing Your Reproducible Research

In the United States and many other countries research, including computer code made available via the internet is automatically given copyright protection. However, copyright protection works against the scientific goals of reproducible research, because work derived from the research falls under the original copyright protections (36).

To solve this problem, some authors have suggest placing code under an open source software license like the GNU General Public License (GPL) (36). Being designed to make software more freely available, they are not really adequate for making available the data, code, and other material needed to reproduce research findings in a way that enables scientific validation and knowledge growth (see 36).