Notes for Grad Student Orientation 2018

Charles J. Geyer

Contents

Chapter 1

Introduction

These are some notes for grad student orientation in the School of Statistics, University of Minnesota, Fall 2018.

1.1 License

1.2 R

- The version of R used to make this document is 3.5.1.
- The version of the `bookdown` package used to make this document is 0.7.
- The version of the `rmarkdown` package used to make this document is 1.10.
- The version of the `knitr` package used to make this document is 1.20.
- The version of the `ggplot2` package used to make this document is 3.0.0.

1.3 Other Learning Materials

1.3.1 An Introduction to R

By far the best book on R is free, written by the R core team, always up to date with the current version, and always correct. It is called *An Introduction to R* and can be found in your R distribution. Do

```
help.start()
```

to start browser help and click on the "An Introduction to R" link.

If can also be found on-line at CRAN.

PDF and e-book versions are also available at CRAN.

There are five other manuals that come with R, but they are for experts. You don't want to read them yet.

1.3.2 An R Short Course

Your humble author was one of five instructors for a two-day short course on R. Here are the notes for it.

1.3.3 An R Course (Undergraduate)

Your humble author taught Stat 3701 (undergraduate statistical computing). Here is the web site for that.

Of particular interest are the

- reproducible research examples.
- course notes.

1.3.4 An R Course (PhD Level)

Your humble author taught Stat 8054 (PhD level statistical computing). Here is the web site for that.

This site is a bit out of date. I will redo this when I teach this course again in Spring 2019, including many topics from my Stat 3701 notes linked above, like web scraping, JSON, and SQL databases.

Chapter 2

R Markdown

2.1 Source

The source for this file is https://raw.githubusercontent.com/cjgeyer/Orientation2018/master/01-rmarkdown.Rmd.

To fully understand it you have to compare what you see here (output) to the the source. So open the source in another tab in your browser.

2.2 What is It, and Why do I Want It?

2.2.1 What is It?

R markdown is the latest in a long line of R packages that provide

- literate programming

and

- reproducible research

using R.

It allows you to mix R code that is executed in the production of the document with a document. Of course plain code with comments goes a little way to explaining code, but literate programming is much better.

R markdown can be converted to output formats other than HTML. Among these are PDF, Microsoft Word, and e-book formats. Other output formats are explained in the Rmarkdown documentation.

2.2.2 Newbie Data Analysis

The way most newbies use R or any other statistical package is to dive right in

- typing commands into R,
- typing commands into a file and cut-and-pasting them into R, or
- using RStudio.

None of these actually document what was done because commands get edited a lot.

If you are in the habit of saving your workspace when you leave R or RStudio, can you explain *exactly* how every R object in there was created? *Starting from raw data?* Probably not.

If you work this way, you are never an "expert" even if you have 50 years experience. You are also **not** doing reproducible research.

2.2.3 Expert Data Analysis

The way experts use plain R is to type commands into a file, say `foo.R` and use

```
R CMD BATCH --vanilla foo.R
```

to run R to do the analysis.

There are several ways experts use literate programming R. Type commands with explanations into an R Markdown file, and render it in a clean R environment (empty global environment). Either start R with a clean global environment (with `R --vanilla`) and do

```
library(rmarkdown)
render("foo.Rmd")
```

or start RStudio with a clean global environment (on the "Tools" menu, select "Global Options" and uncheck "Restore .RData into workspace at startup", then close and restart) load the R Markdown file and click "Knit".

Or use an older competitor of R Markdown, such as R function `Sweave` or R package `knitr`.

The important thing is using a clean R environment so all calculations are fully reproducible. Same results every time the analysis is rerun by you or by anybody, anywhere, on any computer that has R.

That's (the computing part of) reproducible research!

2.2.4 No Snarf and Barf

Snarf and barf is a colorful hacker term for cut and paste.

When doing reproducible research you must **never snarf and barf**. It will inevitably get out of date so snarf-and-barfed output does not match the code in the document that purportedly produces it.

In short, snarf and barf inevitably leads to lies (inadvertent, but still lies).

So don't.

With R Markdown (or `Sweave` or `knitr`) you never need to.

2.3 Getting Started

You don't need RStudio to use R Markdown (despite it being created by people who are now all RStudio employees).

If you have an R Markdown file `baz.Rmd` then

```
Rscript -e 'rmarkdown::render("baz.Rmd", "all")'
```

run from the operating system command line renders it into whatever output formats it says it does and

```
Rscript -e 'rmarkdown::render("baz.Rmd", "html_document")'
```

does a specific format.

But for today, we'll use RStudio.

- Start RStudio.
- On the File menu
- select "New File"
- then select "R Markdown"
  - and in the window that comes up fill in a title and author
  - and click the "OK" button

You now should have a toy R Markdown document in the upper left panel of the RStudio app. Now

- Click the button labeled "knit" having a yarn and needles icon.
- and in the window that pops up give it a file name and location for where to save the file.
- and then the rendered document should show up in the lower right panel or in a pop up.

We're in business! We have done R Markdown!

As we go along we can try things.

2.4 Syntax not Involving R

Section 2.3 of the R Markdown book recommends the RStudio cheat sheets.

Section 2.5 of the R Markdown book has its take on the markdown part of R Markdown.

Things to try in your toy document

- italics
- bold face
- monospace font (for code) (uses backticks)
- hyperlinks
- section headings
- lists
- math (if you already know LaTeX or just want to copy the examples in the R Markdown book — we aren't going to try to teach LaTeX here)

2.5 Syntax Involving R

In your toy document RStudio already gives you two code chunks (R that is executed when the document is rendered and the output stuffed in the document). One does a summary and the other a plot.

Here's a trivial code chunk

```
2 + 2
```

```
## [1] 4
```

Note that the result is not in the document source. Every time the document is rendered, R executes the code producing new results. If you change the code, the results also change (unlike what happens if you snarf-and-barf code and results).

The toy document RStudio provides shows a plot using R base graphics, here is another using R package `ggplot2`.

Chapter 3

Parallel Computing in R

Foo!

Chapter 4

Version Control with Git