Modern R in a Corporate Environment

Original materials developed for RADARS

Brian Davis

2018-04-19

Contents

About

Chapter 1

Introduction

> Something that will make life easier in the long-run can be the most difficult thing to do today. For coders, prioritising the long term may involve an overhaul of current practice and the learning of a new skill.

1.1 Course Philosophy

> "The best programs are written so that computing machines can perform them quickly and so that human beings can understand them clearly. A programmer is ideally an essayist who works with traditional aesthetic and literary forms as well as mathematical concepts, to communicate the way that an algorithm works and to convince a reader that the results will be correct." Donald Knuth

1.1.1 Reproducible Research

Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them. There are two basic reasons to be concerned about making your research reproducible. The first is *to show evidence of the correctness of your results.* The second reason to aspire to reproducibility is *to enable others to make use of our methods and results.*

Modern challenges of reproducibility in research, particularly computational reproducibility, have produced a lot of discussion in papers, blogs and videos, some of which are listed here and here.

> Conclusions in experimental psychology often are the result of null hypothesis significance testing. Unfortunately, there is evidence ((from eight major psychology journals published between 1985 and 2013) that roughly half of all published empirical psychology articles contain at least one inconsistent p-value, and around one in eight articles contain a grossly inconsistent p-value that makes a non-significant result seem significant, or vice versa. statscheck and here

> "A key component of scientific communication is sufficient information for other researchers in the field to reproduce published findings. For computational and data-enabled research, this has often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters. Many journals are revising author guidelines to include data and code availability. We chose a random sample of 204 scientific papers published in the journal **Science** after the implementation of their policy in February 2011. We found that were able to

> reproduce the findings for 26%." Proceedings of the National Academy of Sciences of the United States of America
>
> "Starting September 1 2016, JASA ACS will require code and data as a minimum standard for reproducibility of statistical scientific research." JASA

1.1.2 FDA Validation

> "Establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes." -Validation as defined by the FDA in **Validation of Systems for 21 CFR Part 11 Compliance**

1.1.3 The SAS Myth

Contray to what we hear the FDA does not require SAS to be used *EVER*. There are instances that you have to deliver data in XPORT format though which is open and implemented in many programming languages.

> "FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g., in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification. As noted in the FDA guidance, E9 Statistical Principles for Clinical Trials" FDA Statistical Software Clarifying Statement

Good write up with links to several FDA talks on the subject.

1.2 Prerequisites

- We will assume you have minimal experience and knowledge of R
- IT should have installed:
  - R version 3.5
  - RStudio version 1.1
  - MiTeX
  - RTools version 3.4
- We will install other dependencies throughout the course.

1.3 Content

It is impossible to become an expert in R in only one course even a multi-week one. Yet, this course aims at giving a wide understanding on many aspects of R as used in a corporate / production environment. It will roughly be based on R for Data Science. While this is an *excellent* resource it does not cover much of what we will need on a routine basis. Some external resources will be referred to in this book for you to be able to deepen what you would have learned in this course.

This is your course so if you feel we need to hit an area deeper, or add content based on a current need, let me know an we will work to adjust it.

The **rough** topic list of the course:

1. Good programming practices
2. Basics of R Programming
3. Importing Data

4. Tiyding Data
5. Visualizing Data
6. Functions
7. Strings
8. Dates and Time
9. Communicating Results

Making Code Production Ready:

10. Functions (part II)
11. Assertions
12. Unit tests
13. Documentation
14. Communicating Results (part II)

1.4 Structure

My current thoughts are to meet an hour a week and discuss a topic. We will not be going strictly through the R4DS, but will use it as our foundation into the topic at hand. Then give an assignment due for the next week which we go over the solutions. We will incorporate these assignments into a RADARS R package(s?) so we will have a collection of usefull reusable code for the future.

Open to other ideas as we go along. I'm going to try to keep the assignments related to our current work (maybe working through Site Investigator and/or Subscriber Reports) so we can work on the class during work hours.

Chapter 2

Good practices

> "Programs must be written for people to read, and only incidentally for machines to execute." Harold Abelson
>
> "Programming is the art of telling another human being what one wants the computer to do." Donald Knuth
>
> "Let us change our traditional attitude to the construction of programs. Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do." Donald Knuth
>
> "When you write a program, think of it primarily as a work of literature. You're trying to write something that human beings are going to read. Don't think of it primarily as something a computer is going to follow. The more effective you are at making your program readable, the more effective it's going to be: You'll understand it today, you'll understand it next week, and your successors who are going to maintain and modify it will understand it."

2.1 Coding style

Good coding style is like correct punctuation: you can manage without it, butitsuremakesthingseasiertoread. When I answer questions; first, I read the title of the question to see if I can answer the question, secondly, I check the coding style of the question and if the code is too difficult to read, I just move on. Please make your code readable by following e.g. this coding style (most examples below come from this guide).

2.1.1 Comments

In code, use comments to explain the "why" not the "what" or "how". Each line of a comment should begin with the comment symbol and a single space: `#`.

2.1.2 Naming

> There are only two hard things in Computer Science: cache invalidation and naming things. – Phil Karlton

Names are not limited to 8 characters as in some other languages. Be smart with your naming; be descriptive yet concise. Think about how your names will show up in autocomplete.

Throughout the course we will point out some standard naming conventions that are used in R (and other languages). (Ex. `i` and `j` as row and column indicies)

```
# Good
average_height <- mean((feet / 12) + inches)
plot(mtcars$disp, mtcars$mpg)

# Bad
ah<-mean(x/12+y)
plot(mtcars[, 3], mtcars[, 1])
```

2.1.3 Structure

Use commented lines of `-` to create a code outline.

2.1.4 Spacing

Put a space before and after `=` when naming arguments in function calls. Most infix operators (`==`, `+`, `-`, `<-`, etc.) are also surrounded by spaces, except those with relatively high precedence: `^`, `:`, `::`, and `:::`. Always put a space after a comma, and never before (just like in regular English).

```
# Good
average <- mean((feet / 12) + inches, na.rm = TRUE)
sqrt(x^2 + y^2)
x <- 1:10
base::sum

# Bad
average<-mean(feet/12+inches,na.rm=TRUE)
sqrt(x ^ 2 + y ^ 2)
x <- 1 : 10
base :: sum
```

2.1.5 Indenting

Curly braces, `{}`, define the the most important hierarchy of R code. To make this hierarchy easy to see, always indent the code inside `{}` by two spaces.

```
# Good
if (y < 0 && debug) {
  message("y is negative")
}

if (y == 0) {
  if (x > 0) {
    log(x)
  } else {
    message("x is negative or zero")
  }
} else {
  y ^ x
}
```

```
# Bad
if (y < 0 && debug)
message("Y is negative")

if (y == 0)
{
    if (x > 0) {
      log(x)
    } else {
  message("x is negative or zero")
    }
} else { y ^ x }
```

2.1.6 Long lines

Strive to limit your code to 80 characters per line. This fits comfortably on a printed page with a reasonably sized font. If you find yourself running out of room, this is a good indication that you should encapsulate some of the work into a separate function.

If a function call is too long to fit on a single line, use one line each for the function name, each argument, and the closing `)`. This makes the code easier to read and to change later.

```
# Good
do_something_very_complicated(
  something = "that",
  requires  = many,
  arguments = "some of which may be long"
)

# Bad
do_something_very_complicated("that", requires, many, arguments,
                              "some of which may be long"
```

2.1.7 Other

- Use `<-`, not `=`, for assignment. Keep `=` for parameters.

```
# Good
x <- 5
system.time(
  x <- rnorm(1e6)
)

# Bad
x = 5
system.time(
  x = rnorm(1e6)
)
```

- Don't put `;` at the end of a line, and don't use `;` to put multiple commands on one line.
- Only use `return()` for early returns. Otherwise rely on R to return the result of the last evaluated expression.

```r
# Good
add_two <- function(x, y) {
  x + y
}

# Bad
add_two <- function(x, y) {
  return(x + y)
}
```

- Use ", not ', for quoting text. The only exception is when the text already contains double quotes and no single quotes.

```r
# Good
"Text"
'Text with "quotes"'
'<a href="http://style.tidyverse.org">A link</a>'

# Bad
'Text'
'Text with "double" and \'single\' quotes'
```

2.2 Coding practices

2.2.1 Variables

Create variables for values that are likely to change.

2.2.2 *Rule of 3*

Try not to copy code, or copy then modify the code, more than twice.

- If a change requires you to search/replace 3 or more times make a variable.
- If you copy a code chunk 3 or more times *make a function*
- If you copy a function 3 or more times *make your function more generic*
- If you copy a function 3 or more times into a project *make a package*
- If 3 or more people will use the function *make a package*
- If 3 or more projects will use the function *make a package*

Same thing goes for lookup tables and such. The key thing to think about is; if something changes how many touch points will there be? If it is 3 or more places it is time to abstract this code a bit.

2.2.3 Path names

It is better to use relative path names instead of hard coded ones. If you must read from (or write to) paths that are not in your project directory structure create a file name variable at the highest level you can (*always end with the /*) and then use relative paths.
DO NOT EVER USE `setwd()`

```r
# Good
raw_data <- read.csv("./data/mydatafile.csv")
```

```
input_file <- "./data/mydatafile.csv"
raw_data <- read.csv(input_file)

input_path <- "C:/Path/To/Some/other/project/directory/"
input_file <- paste0(input_path, "data/mydatafile.csv")
raw_data <- read.csv(input_file)

# Bad
setwd("C:/Path/To/Some/other/project/directory/data/")
raw_data <- read.csv("mydatafile.csv")
setwd("C:/Path/back/to/my/project/")
```

2.3 RStudio

Download the latest version of RStudio (> 1.1) and use it!

Learn more about new features of RStudio v1.1 there.

RStudio features:

- everything you can expect from a good IDE
- keyboard shortcuts I use frequently
  1. *Ctrl* + *Space* (auto-completion, better than *Tab*)
  2. *Ctrl* + *Up* (command history & search)
  3. *Ctrl* + *Enter* (execute line of code)
  4. *Ctrl* + *Shift* + *A* (reformat code)
  5. *Ctrl* + *Shift* + *C* (comment/uncomment selected lines)
  6. *Ctrl* + *Shift* + */* (reflow comments)
  7. *Ctrl* + *Shift* + *O* (View code outline)
  8. *Ctrl* + *Shift* + *B* (build package, website or book)
  9. *Ctrl* + *Shift* + *M* (pipe)
  10. *Alt* + *Shift* + *K* to see all shortcuts…
- Panels (everything is integrated, including **Git** and a terminal)
- Interactive data importation from files and connections (see this webinar)
- Use code diagnostics:
- **R Projects**:
  - **Meaningful structure** in one folder
  - The working directory automatically switches to the project's folder
  - File tab displays the associated files and folders in the project
  - History of R commands and open files
  - Any settings associated with the project, such as Git settings, are loaded. Note that a *set-up.R* or even a *.Rprofile* file in the project's root directory enable project-specific settings to be loaded each time people work on the project.

> The only two things that make @JennyBryan . Instead use projects + here::here() #rstats pic.twitter.com/GwxnHePL4n
>
> — Hadley Wickham (@hadleywickham) December 11 2017

Read more at https://www.tidyverse.org/articles/2017/12/workflow-vs-script/ and also see chapter *Efficient set-up* of book *Efficient R programming*.

2.4 Getting help

2.4.1 Help yourself, learn how to debug

A basic solution is to print everything, but it usually does not work well on complex problems. A convenient solution to see all the variables' states in your code is to place some `browser()` anywhere you want to check the variables' states.

Learn more with this book chapter, this other book chapter, this webinar and this RStudio article.

2.4.2 External help

Can't remember useful functions? Use cheat sheets.

You can search for specific R stuff on https://rseek.org/. You should also read documentations carefully. If you're using a package, search for vignettes and a GitHub repository.

You can also use Stack Overflow. The most common use of Stack Overflow is when you have an error or a question, you google it, and most of the times the first links are Q/A on Stack Overflow.

You can ask questions on Stack Overflow (using the tag `r`). You need to make a great R reproducible example if you want your question to be answered. Most of the times, while making this reproducible example, you will find the answer to your problem.

If you're confident enough in your R skills, you can go to the next step and answer questions on Stack Overflow. It's a good way to increase your skills, or just to procrastinate while writing a scientific manuscript.

2.5 Keeping up to date

With over 10,000 packages on CRAN it is hard to keep up with the constantly changing landscape. R-Bloggers is an R forcused blog aggregator with dozens of posts per day. Checkit out.

Join the R-help mailing list. Sign up to get the daily digest and scan it for questions that interest you.

2.6 Assignement

1. See these Rstudio Tips & Tricks or these and find one that looks interesting and **practice** it all week.
2. Create an R Project for this class.
3. Create the following directories in your project (tip sheet?)
   - Bonus points if you can do it from R and not RStudio or Windows Explorer
   - Double Bonus points if you can make it a function.
4. Read Chapters 1-3 of the Tidyverse Style Guide
5. Copy one of your R scripts into your R directory. (Bonus points if you can do it from R and not RStudio or Windows Explorer)
6. Apply the style guide to your code.

7. Apply the "Rule of 3"
   - Create variables as needed
   - Identify code that is used 3 or more times to make functions
   - Identify code that would be useful in 3 or more projects to integrate into a package.
8. Read how to make a great R reproducible example

Chapter 3

R Programming Basics

See this vocabulary list for a good starting point on the basics functions in base R and some important libraries.

advr38book

In R there three basic constructs; objects, functions, and environments.

Two most important functions in R `?` and `str`.

3.1 Names

R has strict rules about what constitutes a valid name. A **syntactic** name must consist of letters[1], digits, `.` and `_`, and can't begin with `_`. Additionally, it can not be one of a list of **reserved words** like `TRUE`, `NULL`, `if`, and `function` (see the complete list in `?Reserved`). Names that don't follow these rules are called **non-syntactic** names, and if you try to use them, you'll get an error:

```
_abc <- 1
#> Error: unexpected input in "_"

if <- 10
#> Error: unexpected assignment in "if <-"
```

3.2 Notes

where do factors fit in?

3.3 Atomic Vectors

There are many "atomic" types of data: `logical`, `integer`, `double` and `character` (in this order, see below). There are also `raw` and `complex` but they are rarely used.

You can't mix types in an atomic vector (you can in a list). Coercion will automatically occur if you mix types:

[1]Surprisingly, what constitutes a letter is determined by your current locale. That means that the syntax of R code actually differs from computer to computer, and it's possible for a file that works on one computer to not even parse on another!

```
(a <- FALSE)
```

```
#> [1] FALSE
```

```
typeof(a)
```

```
#> [1] "logical"
```

```
(b <- 1:10)
```

```
#>  [1]  1  2  3  4  5  6  7  8  9 10
```

```
typeof(b)
```

```
#> [1] "integer"
```

```
c(a, b)  ## FALSE is coerced to integer 0
```

```
#>  [1]  0  1  2  3  4  5  6  7  8  9 10
```

```
(c <- 10.5)
```

```
#> [1] 10.5
```

```
typeof(c)
```

```
#> [1] "double"
```

```
(d <- c(b, c))  ## coerced to numeric
```

```
#>  [1]  1.0  2.0  3.0  4.0  5.0  6.0  7.0  8.0  9.0 10.0 10.5
```

```
c(d, "a")  ## coerced to character
```

```
#>  [1] "1"    "2"    "3"    "4"    "5"    "6"    "7"    "8"    "9"    "10"   "10.5" "a"
```

```
c(list(1), "a")
```

```
#> [[1]]
#> [1] 1
#>
#> [[2]]
#> [1] "a"
```

```
50 < "7"
```

```
#> [1] TRUE
```

You can force coercion with `as.logical`, `as.integer`, `as.double`, `as.numeric`, and `as.character`. Most of the time the coercion rules are straight forward, but not always.

```
x <- c(TRUE, FALSE)
typeof(x)
```

```
#> [1] "logical"
```

```
as.integer(x)
```

```
#> [1] 1 0
```

```
as.numeric(x)
```

```
#> [1] 1 0
```

```
as.character(x)
```

```
#> [1] "TRUE"  "FALSE"
```

However, coercion is not associative.

```
x <- c(TRUE, FALSE)

x2 <- as.integer(x)
x3 <- as.numeric(x2)
as.character(x3)
```

```
#> [1] "1" "0"
```

What would you expect this to return?

```
x <- c(TRUE, FALSE)

as.integer(as.character(x))
```

You can test for an "atomic" types of data with: `is.logical`, `is.integer`, `is.double`, `is.numeric`, and `is.character`.

```
x <- c(TRUE, FALSE)

is.logical(x)
```

```
#> [1] TRUE
```

```
is.integer(x)
```

```
#> [1] FALSE
```

What would you expect these to return?

```
x <- 2

is.integer(x)
is.numeric(x)
is.double(x)
```

3.4 Base objects

- "atomic" vector: vector of same type (see above).
- scalar: this doesn't exist, there are only vector of length 1.
- matrices/arrays: just an atomic vector with some dimensions (attribute).

```
vec <- 1:12
vec
```

```
#>  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```
class(vec)
```

```
#> [1] "integer"
```

```
dim(vec) <- c(3, 4)
vec
```

```
#>      [,1] [,2] [,3] [,4]
#> [1,]    1    4    7   10
#> [2,]    2    5    8   11
#> [3,]    3    6    9   12
```

```
class(vec)
```

```
#> [1] "matrix"
```

```
dim(vec) <- c(3, 2, 2)
vec
```

```
#> , , 1
#>
#>      [,1] [,2]
#> [1,]    1    4
#> [2,]    2    5
#> [3,]    3    6
#>
#> , , 2
#>
#>      [,1] [,2]
#> [1,]    7   10
#> [2,]    8   11
#> [3,]    9   12
```

```
class(vec)
```

```
#> [1] "array"
```

- list: vector of elements with possible different types in it.

- data.frame: a list whose elements have the same lengths, and formatted somewhat as a matrix.

Figure out how to make results nicer inside RMarkdown

3.4.1 Vectors

3.4.1.1 Important operators and assignment

3.4.1.2 Comparison

3.4.1.3 Logical, Sets and Missing Values

3.4.2 Control flow

3.4.2.1 Ordering and tabulating

3.4.2.2 Basic math

3.4.3 Matrices

3.4.4 Lists & data.frames

3.4.5 Selecting Values

3.5 Functions

3.5.1 Functional Programming

3.5.2 Functionals

3.5.3 Function operators

3.6 Environments

Scoping