

Reproducible Research in R

Dries Debeer & Benjamin Becker

29. and 30. September 2022

FDZ Autumn Academy

Introduction

Who are we?

Dries Debeer

Statistical Consultant at Ghent University (FPPW)

scDIFtest, permimp, eatATA, mstDIF

dries.debeer@ugent.be

Benjamin Becker

Researcher at IQB (Verbund Forschungsdaten)

eatGADS, eatDB, eatATA, pisaRT

b.becker@iqb.hu-berlin.de

Who are you?

1. Occupation, employer?
2. Previous knowledge and experience
 - with reproducible research
 - with R?
 - with other statistical software?
 - with other programming languages?
3. Specific interest/motivation for this workshop?

Why care about reproducible research?

Agenda

- Conceptual things
- Writing Reproducible R Code
- RMarkdown
- Version Control/git

Reproducible Research

RStudio setup

RStudio setup

1. Save the course content to a directory on your machine
2. Open RStudio
3. Choose File < New Project ...
4. Choose Existing Directory
5. Browse to the directory on your machine where you saved the course content and select the “[R-programming](#)” folder as the Project working directory
6. Click Open in new session
7. Click Create Project

RStudio setup - optional

1. Choose Tools < Global options
2. Under General
 - DON'T Restore .RData into workspace at startup
 - NEVER Save workspace to .Rdata on exit:
 - Save the code instead!
 - Use `saveRDS()` and `readRDS()` for objects that require a long time to compute
3. Further personalize RStudio

Writing Reproducible R Scripts

RMarkdown

Version Control via Git and Github

Good programming practices

“Write code for humans, not for machines!”

Invest time in writing readable R-code.

- It will make collaborations easier
- It will make debugging easier
- It will make your analyses more reproducible

There is a complete *tidyverse* style-guide

<https://style.tidyverse.org/>.

Go easy on your eyes

- with spaces before and after: `- + / * = <- < == >`
- always use `<-` for assignments
- only use `=` in function calls
- use indentation (largely automatic in RStudio)
- `CamelCaseNames` vs `snake_case_names`
- be consistent!
- wrap long lines at column 70-80 (Rstudio)

White space

```
new_var=(var1*var2/2)-5/(var3+var4)
```

```
# versus
```

```
new_var <- (var1 * var2 / 2) - 5 / (var3 + var4)
```

Indentation

```
for(name in names){formula=as.formula(paste0("y~.",name))
fit<-lm(formula,data=my_data)
coefs[["name"]]=coef(fit)
print(name)
print(summary(fit))}
```

versus

```
for(name in names){
  formula <- as.formula(paste0("y~.", name))
  fit <- lm(formula, data = my_data)
  coefs[["name"]] <- coef(fit)
  print(name)
  print(summary(fit))
}
```

Wrap long lines

```
final_results <- data.frame(first_variable =  
  sqrt(results$mean_squared_error), second_variable =  
  paste0(results$condition, results$class, sep = ":"),  
  third_variable = results$bias)
```

versus

```
final_results <- data.frame(  
  first_variable = sqrt(results$mean_squared_error),  
  second_variable = paste0(results$condition,  
                           results$class, sep = ":"),  
  third_variable = results$bias)
```

Go easy on your mind

- use meaningful names: “self-explainable”
- always write the formal arguments in function calls (except the first)
- benefit from autocompletion (`<tab>`) => embrace longer names
- use `TRUE` and `FALSE` not `T` and `F`
- comment, comment, comment
 - NOT what (should be clear from the code)
 - but why
 - explain the reasoning, not the code

Use meaningful names

```
V <- myFun(m1_B)
```

```
# versus
```

```
RMSE_age_gender <- get_RMSE(lm_age_gender)
```

Programming advice

Use verbs for functions and nouns for other objects.

Write formal arguments

Benefit from auto completion using tab

```
m1_B <- lm(outcome ~ age*gender,  
           exp1, condition_1, freq)
```

versus

```
lm_age_gender <- lm(outcome ~ age * gender,  
                   data = exp1,  
                   subset = condition_1,  
                   weights = freq)
```

Comment, comment, comment

```
## Start every Rscript with a comment that explains
##  what the code in the script does, why it does
##  this, and to which project it belongs.
##  Your future self will be very thankful!
##
## Mention which packages you are using in this Rscript.

## Use sections to separate chunks -----

## Maybe even subsections =====

## Recode variables so that missings are coded as "NA"
dat[dat %in% c(99, 999)] <- NA # missings coded 99 or 999
```


Keep your code slim

Try to limit your *package-dependencies*.

Only load `library()` the packages that you absolutely need. If you are only using `dplyr`, it does not make sense to load the complete `tidyverse`.

Controversial: when possible, use the `::` operator (and consider not loading the package). `<package>::<function>`

- explicit dependencies
- less name conflicts

Never Attach

Forget about `attach()`!

Don't use it, unless you completely understand what happens (see `?attach`).

Use `with(data.frame, expression)` instead.

```
# using with()
n <- 2e+4
data <- data.frame(x = runif(n),
                   y = runif(n),
                   z = seq_len(n))
result <- with(data, exp(x) / log(z) + 5 * sqrt(y))
```

Writing code is error prone. Incorporate tests and checks in your workflow.

- minimal examples
- write tests and checks
- helpful packages: `testthat`, `RUnit`, `testit`, ...

Computing speed can become an issue. Avoid common pitfalls:

- don't grow, but replace
- vectorize where possible
- check the computing speed

?system.time, microbenchmark or profiling tools

```
n <- 2e+4
data <- data.frame(x = runif(n),
                  y = runif(n),
                  z = seq_len(n))
```

Speed

Don't grow!

```
system.time({  
  new_data <- NULL  
  
  for(row_nr in seq_len(NROW(data))){  
    new_data <- cbind(  
      data[row_nr,],  
      result = exp(data$x[row_nr]) /  
        log(data$z[row_nr]) +  
        5 * sqrt(data$y[row_nr]))  
  }  
})
```

```
>      user  system elapsed  
>    2.20    0.00    2.25
```

Replace!

```
system.time({  
  n_rows <- dim(data)[1]  
  data$result <- rep(NA, n_rows)  
  
  for(row_nr in seq_len(n_rows)){  
    data$result[row_nr] <- exp(data$x[row_nr]) /  
      log(data$z[row_nr]) +  
      5 * sqrt(data$y[row_nr])  
  }  
})  
  
>    user  system elapsed  
> 0.29    0.01    0.33
```

Vectorize!

```
system.time({  
  data$result <- exp(data$x) / log(data$z) +  
    5 * sqrt(data$y)  
})
```

```
>      user  system elapsed  
>         0         0         0
```

Speed

Compare the speed of different implementations using:

`microbenchmark::microbenchmark`

```
get_mean1 <- function(x){  
  weight <- 1/length(x)  
  out <- 0  
  for(i in seq_along(x)){  
    out <- out + x[i] * weight  
  }  
  return(out)  
}
```

```
get_mean2 <- function(x){  
  sum(x)/length(x)  
}
```


Speed

Compare the speed of different implementations using:

```
microbenchmark::microbenchmark
```

```
x <- rnorm(500)
microbenchmark::microbenchmark(
  mean(x), get_mean1(x), get_mean2(x))

> Unit: nanoseconds
>      expr    min     lq   mean  median    uq     max neval
>   mean(x)   2100   2200   3202    2300   2400   88200    100
> get_mean1(x) 13500 13700 43205   13800 14000 2931700    100
> get_mean2(x)   600    700 10165     700    800   935400    100
```

Programming advice

Don't worry about speed before it becomes an issue.

Wrap Up

General Advice

- Investing time in learning R pays off
- It's a steady learning curve
- Learn from masters
- Rewrite important code - the first attempt is usually not the best approach

General R Advice

- Document well
- Use a consistent style
- Write functions
- Split long functions in smaller ones
- Write wrappers
- Use Iteration (don't copy paste)
- Use matrix operations and vectorized functions instead of loops
- Use git

Literature Recommendations

R Resources

- Advanced R Ed. 1 (<http://adv-r.had.co.nz/>)
- Advanced R Ed. 2 (<https://adv-r.hadley.nz/>)
- R Inferno (https://www.burns-stat.com/pages/Tutor/R_inferno.pdf)
- R Packages (<https://r-pkgs.org/>)
- Clean Code (https://mooc.aplikom.or.id/pluginfile.php/1174/mod_resource/content/1/Clean%20Code_%20A%20Handbook%20of%20Agile%20Software%20C%20-%20Robert%20C.%20Martin.pdf)

Thank you for your attention!

Thank you for your attention!

Questions? Remarks?