Biostatistics I: Introduction to R

Introduction

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

Introduction to R

Statistics have flourished in the recent years mainly due to the possibility of doing complex analysis using computers

- Many statistical software exist to do simple and specialized analysis

The **programming language R** is popular for data scientists

- Analysts must not only learn how to use the software but also the ideas behind it
- Learning statistical modelling and algorithm is more important than learning a programming language.

The most valuable tool of a modern quantitative researcher is his/her personal computer

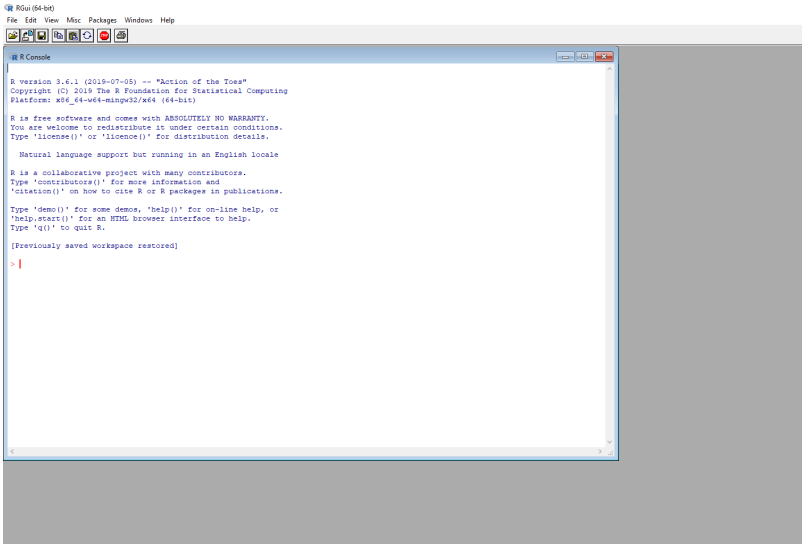
References

- More books that use R (or S) can be found at:

http://www.r-project.org/doc/bib/R-books.html, or
http://www.r-project.org/doc/bib/R-jabref.html

- R ships with a number of helpful manuals
- Other manuals and helpful material are available on-line via CRAN:
  http://cran.r-project.org/other-docs.html

What does R look like ?



What is R

- **R** is a software environment for statistical computing and graphics
  - extensive catalog of statistical and graphical methods
- **R** is mainly used in academia. However, many large companies also use **R** programming language, including healthcare industries but also Uber, Google, Airbnb, Facebook and so on

A brief history of R

- **1993**: University of Auckland, New Zealand by Ross Ihaka and Robert Gentleman
- **1997**: R core Team was formed (20 members)
- **2000**: R 1.0.0 released
- **2004**: First international user conference in Vienna
- **2013**: 5026 packages available
- **2017**: 10875 packages available
- **Now**: `nrow(available.packages())`

Why learn R ?

- **R** is a free software environment for statistical computing and graphics
- It compiles and runs on LINUX, Windows and MacOS
- Open source language
- Users are allowed to modify and redistribute the code
- Advanced statistical language
- Supports extensions
- Related to other languages
- **Flexible and fun!**

Where do I get R ?

- http://cran.r-project.org
- choose your platform, e.g., Windows, Linux

How does R work ?

- Packages built for specific tasks
- Download R packages from the CRAN web site → within R
  - Packages
  - Install package(s)*
  - make you choice(s)
  - load the package using `library()` (**note**: install does not mean load)

* use alternatively `install.packages()`

How to get help in R

- Within R
  - `help.search("topic")` or `??"topic"` (depends on the installed packages)
  - `RSiteSearch("topic")` (requires internet connection)
  - `help()` or `?` invoke the on-line help file for the specified function
  - checking the FAQ
- Online
  - R-help (https://stat.ethz.ch/mailman/listinfo/r-help – mailing list)
  - R-seek (http://www.rseek.org – Google-like searched engine)
  - CRAN Task Views (http://cran.r-project.org/web/views/ – categorization of packages)
  - Crantastic (http://crantastic.org/ – categorization of packages + reviews)
  - R4stats (http://www.r4stats.com/ – examples of basic R programs)
  - R related Blogs (http://www.r-bloggers.com/ – many useful illustrations of R and R packages)
  - Open community for developers (https://stackoverflow.com/ – ask/answer a question)

Disadvantages of R

- Appears intimidating to the first-time user
- Output is not so nice looking (but there are some alternatives)
- Exporting output is more difficult
- Cannot easily handle very big data sets (depends on the installed RAM)
- A lot of things are available but it is sometimes hard to find your way
- The quality of the available packages is greatly varying
- Has been criticized for using only one CPU at a time (but the parallel packages helps you perform tasks in different cores)

Summary

- **R** is a great tool to explore and investigate the data
- Several statistical methods can be performed with **R**
- It is important to understand the methods before applying them in **R**

How to use

R uses packages that perform specific tasks

- Install package only once
- Load package every time you open **R**