Practical session - Introduction to R

Mathilde Mougeot, HMVC

April 2025

Goal of the Regression course

- To understand the Ordinary Least Square (OLS) method and the linear model, from a methodological and practical point of view.
- To apply simple or multiple linear models on several data sets using the 'R' language.
- To interprete 'R' outputs of linear model functions.

Warnings and Advices

- The goal of this practical session is not "just to program with R" but more specifically to understand the framework of Modeling, to learn how to developp appropriate models for answering to a given operationnal question and a given data set. This course belongs to the **Data Science courses** and is a preliminary step before using more advanced methods introduced in other 'machine learning' courses. → For each practical session, you should **first understand** the mathematical and statistical backgrounds, **then write your own program with 'R'** to practically answer to the question.

Remarks

- In order to start any run with a clean environment, the two following lines should be always put in the beginning of your code: `rm(list=ls()); graphics.off()`.
- If you need some help to create a markdown file please refer to the web site https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet.

If you are a beginner in `R`

Perform the first exercices of the swirl() package as mentionned in the first lesson.

Some preliminary exercices using `R`

1. **Vector manipulation**. Recall that
$$e^x = \sum_{k \geq 0} \frac{x^k}{k!}.$$
Create a vector called `exp2` storing the 20 first elements of the previous expression. Remove all the values lower than 10^{-8} and compute an approximation of e^2. Finally, compare this result with the value obtained using directly `exp(2)`.

2. **Data simulation**. Use the fonction `rnorm` (`?rnorm` to get some help) to simulate a vector X of size 100 drawn from a Gaussian law $\mathcal{N}(2,1)$, with a mean equaled to 2 and a variance equaled to 1. Compute a second vector Y of the same size by multiplying X with the value 9.8 and by adding a Gaussian noise of zero mean and a standard deviation equaled to 1/10.

3. **Read and write a text file**. Store both X and Y vectors in a data frame (`?data.frame` if necessary) and then store this `data.frame` in a text file using the instruction `write.table` on your hard disk. In a second step, upload the data from the text file into the R environment using the instructionn `read.table`. Compare the values of the data before and after the storage. Conclusions.

4. **Read and write a RData file** . Store both X and Y vectors in a `data.frame`. Store the previous `data.frame` using the instruction `save`. Then, use the instruction `load` to upload the data from the previous file to the R environment. Compare the values of both tables. Conclusion.

5. **Scatter Plot**. Plot the values of Y function of X, first using first the `plot` function then the `ggplot2` function.

6. **Histogram**. Draw the histogram of X. How can you modify the numbers of bins? Compare visually two computed histograms using a small and a large number of bins. Conclusion.

7. **Loop for**. We consider here a random variable which follows a given χ^2 distribution with an unkown degree of freedom. The goal is here to be able to recover the apriori unknow degree of freedom using the computation of its empirical mean using n observations. Write a `R` program to evaluate the accuracy of this estimation for $n = 3$ then $n = 100$. Use a loop `for` to solve this exercice.

8. **Loop for**. Same exercice as previously using the 'sapply' instruction. Conclusion.