Introduction to R Markdown and GitHub

1 Introduction to R Markdown

R Markdown is a framework that provides a literate programming format for data science. It can be used to save and execute R code within R Studio and also as a simple formatting syntax for authoring HTML, PDF, ODT, RTF, and MS Word documents as well as seamless transitions between available formats. Click on *R-Markdown* to familiarize with this tool.

In the folder *Practical 2* of the GitHub repository *Model Selection in High Dimensions*, you can find the R-Markdown cheat-sheets which are useful and quick references for R Markdown syntax. Remember also that in *Model Selection in High Dimensions* repository, all the practicals and the projects created by the students are stored for evaluation.

You can have a look at **R Markdown** section of the e-book for further references. Now let's start from some basic manipulations (see exercises in R Markdown section of the e-book), we make use of the *Iris* dataset already present in R:

(a) Create an .rmd file from R Studio classic interface and look at the basic notions explained in the new document.

(b) Create an histogram of the sepal width of *Iris Setosa* without showing both the code and the graph. Then, in another code chunk, show only the graph (without the code) and change the height or width of the histogram as you prefer.

(c) Write the formula, both inline and with a Latex environment (e.g. **equation**), of the conditional probability of observing an *Iris Virginica* given that the sepal width is greater than 3. Display both the code and the conditional probability.

2 Introduction to GitHub

GitHub is a version control platform that is commonly used among programmers and software developers. A version control is a system that records changes to a file or a set of files in order to keep track and possibly revert to or modify those changes over time. As you can understand, it is a well suited platform to start new shared projects as the ones of this course.

First of all we need to enter in the GitHub world (see **GitHub** section of the e-book):

(a) Create a free GitHub account on https://github.com/

(b) Read chapter 3 of the *GitHub Guide* (https://smac-group.github.io/ds/github.html)

(c) Install a version of Git (from https://git-scm.com/downloads) which is compatible with the OS of your computer (e.g. Windows/Mac/Linux/Solaris). Once you have downloaded and installed Git, the first thing you should do is to configure it by setting your username and email address (see point a).

(d) Watch the video in Section 3.3 on the GitHub workflow within R Studio in the *GitHub Guide*

(e) Now you have all the elements to create a new R Studio project: follow the indications of the video at point (c) and take the URL from the GitHub repository of the course *Model Selection in High Dimensions* (https://github.com/CaesarXVII/Model-Selection-in-High-Dimensions).

(f) Modify the file (add the name of practical1.rmd file) as you like (e.g. try to solve an exercise). Then *commit* the changes and *push* it to the remote project *Model Selection in High Dimensions*. Do not forget to click on *pull* every time you access to your R Studio project to retrieve the updated version of all the files of the course repository.

(g) In order to properly execute your *commits*, you need to be added as a collaborator of the project. It is sufficient to send an email to *cesare.miglioli@etu.unige.it* with your GitHub name in it from your unige mail account and you will be set as a collaborator.

3 Data on Malnutrition in Zambia (see e-book)

Consider the dataset *Zambia.SAV* down-loadable at *Course Datasets* containing variables assumed to be potential causes for childhood malnutrition:

- breastfeeding duration (month);
- age of the child (month);
- age of the mother (years);
- Body Mass Index (BMI) of the mother (kg/meter^2);
- height of the mother (meter);
- weight of the mother (kg);
- region of residence (9 levels: Central, Copperbelt, Eastern, Luapula, Lusaka, Northern, Northwestern, Southern and Western);
- mother's highest education level attended (4 levels: No education, Primary, Secondary and Higher);
- wealth index factor score;

- weight of child at birth (kg) ;
- sex of the child;
- interval between the current birth and the previous birth (month); and
- main source of drinking water (8 levels: Piped into dwelling, Piped to yard/plot, Public tap/standpipe, Protected well, Unprotected well, River/dam/lake/ponds/stream/-canal/ irrigation channel, Bottled water, Other).
- Others

You can have a look at the e-book to gather more information on this topic. Then:

(a) Load the dataset and build the variables so that they can be used for a regression analysis.

(b) Associate proper names to each variable (hint: look at the comments in the r chunk).

(c) Perform a linear regression on all the available variables.

(d) Reduce the number of covariates (e.g. using the t-test) and add some interactions. Perform a linear regression on the new dataset.

(e) Analise your chosen estimated model with a residual analysis (e.g. residuals vs fitted plot, normal QQ plot etc.).

4 Data on Gene Expression of Leukemia Patients (see e-book)

Certain factors that can affect a child outlook (prognosis) are called prognostic factors. They help doctors decide whether a child with leukemia should receive standard treatment or more intensive treatment. Prognostic factors seem to be more important in acute lymphocytic leukemia (ALL) than in acute myelogenous leukemia (AML).

See *American Cancer Association* for a detailed explanation.

The *leukemia_big.csv* dataset contains gene expression measurements on 72 leukemia patients: 47 ALL (i.e. acute lymphocytic leukemia) and 25 AML (i.e. acute myelogenous leukemia). These data arise from the landmark of Golub et al. (1999) Science paper and exhibit an important statistical challenge because $p >> n$ as we deal with 72 patients and 7128 measurements.

(a) Load the data from the URL http://web.stanford.edu/ hastie/CASI_files/DATA/leukemia_big.csv

(b) Create the response variable y according to the number of ALL and AML patients. In the same fashion create the matrix X of independent variables.

See https://web.stanford.edu/ hastie/CASI_files/DATA/leukemia.html for further details.

(c) Fit a GLM (choose the correct link) to estimate the relationships between the outcome and the factors with the *leukemia_big.csv* dataset. Comment on the results that you obtain.