Assignment 2
Deadline: 4/6/2023
Instructor: Panagiotis Papastamoulis

Exercise 1 (Big Data Regression: computational techniques)**.** First install the following:

```
install.packages("biglm") # version 0.9-2.1
install.packages("fastmatch")
```

Next, download the specific versions of the packages `bit_1.1-15.2`, `ff_2.2-14.2.tar.gz` and `ffbase_0.12.8`. Go to the relevant pages in CRAN repository, search for the archived versions of each package and download the relevant `*.tar.gz` file. Then, open `R` and run

```
install.packages("path/bit_1.1-15.2.tar.gz", type = "source", repos = NULL)
install.packages("path/ff_2.2-14.2.tar.gz", type = "source", repos = NULL)
install.packages("path/ffbase_0.12.8.tar.gz", type = "source", repos = NULL)
```

by replacing `path` with the path to your download directory.

Open `R` and run the following script

```
set.seed(am) # replace am with your AM
p <- rpois(1, lambda = 120)
n <- 2000000
b <- rt(p, df = 5)
outFile <- "big_data_regression.csv"
zz <- file(outFile, "w")
colNames <- c("y", paste0("x", 1:(p-1)))
colNames <- paste0(colNames, collapse=",")
cat(colNames, "\n", file = zz)
for (i in 1:n){
        x <- matrix(rnorm(p-1), nrow = 1)
        y <- b[1] + x %*% b[-1] + rnorm(1)
        xy <- cbind(y, x)
        cat(paste0(xy, collapse = ","), file = zz, append=TRUE,"\n")
        if( i %% 100000 == 0){
                cat(paste0("write to file ",outFile, " line: ", i), "\n")
        }
}
close(zz)
```

The previous code snippet will create a file: ‘‘`big_data_regression.csv`’’ and will write[1] each line of the synthetic dataset. The task is to estimate a linear regression model based on this dataset, without loading the data into memory[2]. The header of the generated file shows the name of the variables: the response variable is y and the remaining ones are explanatory variables.

1. Use the command `bigglm.ffdf()` of the `ffbase` library to estimate the regression coefficients and report your results. Use the option `sandwich = FALSE`.

2. Compute $X^\top X$ and $X^\top \boldsymbol{y}$ using a recursive approach and then use the `solve()` command in order to obtain the least squares estimate

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}.$$

[1] The file will be ≈ 5 Gb. Make sure that you have at least 10 Gb of free space at your hard disk.

[2] Advice: Do not try to load directly the file into R, unless you want to crash your machine.

For this purpose you may loop through successive chunks of the rows of the `ffdf` object you have already loaded previously. Compare your findings with the ones obtained previously in terms of accuracy (note: the results should be identical) and time.

3. Use a sub-sampling approach in order to estimate the regression coefficients. For each random split of the data derive an estimate and its standard error. Weight the different estimates with the inverse of the variance to report a weighted estimate. Compare your findings to the ones obtained previously.

Exercise 2 (Big Data Regression: airlines dataset)**.** Download the data from the airlines dataset

http://stat-computing.org/dataexpo/2009/the-data.html

The data provides arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. Use the data for a year with the same last digit with your AM. You want to fit a model for arrival delay, in minutes using as covariates

- The month
- The weekday
- The distance
- The departure delay
- The departure time

Describe the model you estimated and write a short report explaining what you see.

Exercise 3 (Communities and Crime Data Set[3])**.** Find the data in

http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized#

including some description about them. The task is to find a model in order to describe the response variable

`murders`: number of murders in 1995

Note that you need some pre-processing of the data to remove some variables that are not useful. Be as detailed as possible so as your report to be self-explained. Explain why you selected the specific model, how good you think it is and any limitations that may apply.

Exercise 4 (Presidential elections in USA, 2016[4])**.** The file `presidential_elections.xlsx` (see the `data` directory at e-class) contains data regarding the elections for the candidates of both parties (Democrats and Republicans) for the presidential election of 2016. The data are given in two sheets. The one file has the socio-economic characteristics of the counties and the second one the votes for the candidates.

The task is to create a model using as response whether Trump got more than 50% of the votes at each county for the Republicans, using as explanatory variables the socio-economic characteristics of the counties. You need to find a reasonable model selecting covariates and being able to use the model to explain the behavior of voters.

Description of the variables is given in separate sheet in the excel. Provide a report with your findings. Be as detailed as possible so as your report to be self-explained. Explain why you selected the specific model, how good you think it is and any limitations that may apply.

[3] Exercise 3 is assigned to students where the last AM digit is $\leqslant 4$

[4] Exercise 4 is assigned to students where the last AM digit is $\geqslant 5$