Statistical programming with R: examples of exam questions

Instructor: Emmanuel Kemel (kemel@hec.fr)

October 2020

Description

The exam will take the form of MCQs. There will be 3 types of questions.

- questions about R and programming (5p)
- questions about understanding of R scripts (10p)
- questions asking you to use R in order to make calculations (5p)

In addition, there will be a bonus of 1 or 2 points for the prediction competition. A training data set ”houseprices_training” is in the dropbox. Use this data set to make predictions of the price of houses in the prediction data set ”houseprices_topredict”. Solutions with a prediction R^2 between 0.7 and 0.85 will get one additional point. Solutions with a prediction R^2 larger than 0.85 will get two additional points. In order to participate to the prediction competition, add a column ”prediction” to the ”houseprices_topredict” data set (rows in the same order as in the original file) and save it in a .csv file (with a coma separator) under the name ”firstname_lastname.csv” (e.g. emmanuel_kemel.csv) and upload it an the share dropbox folder ”predictions” no later than October 31.

The data-set(s) used in the exercises is in the ”data” folder of the Dropbox.

Questions about R

Question 1

The function read.table returns a:

- a vactor
- a matrix
- a data frame
- an integer

Question 2

The function library()

- opens the help window
- installs a package on your computer
- downloads books online to text mining
- loads a package in the session

Questions about interpretations of R code

Question 3

What does the following chunk do?

```
write.table(data, file="my_data.csv")
```

- if opens a dataset ”my_data.csv”
- it exports a data set called my_data
- it exports a data set called data
- it creates a frequency table of viariables in data

Question 4

Which command would return the same value as the following chunk

```
data[order(data$income),"income"][1]
```

- min(data$income)
- max(data$income)
- data[1,1]
- data[[1]][1]

Questions about production of R code

We consider a data set about volumes of sales of a given product over consecutive weeks t (one week per row), with price and promotion as possible explanatory variables. Promotion is a dummy variable that takes value 1 if a specific promotion campaign has been run on a given week, 0 otherwise. The data set is available in the file ”confood2.txt”.

Question 5

How many rows are there in the data set?

- 520
- 12
- 52
- 5

Question 6

Run an OLS regressions to estimate the model $log(Sales_t) = \beta_0 + \beta_1 log(Price_t) + \epsilon_t$. What is the estimated value of the coefficient β_1 ?

- -5.1
- 4.8
- 0.67
- 0.17