Data mining and multivariate statistical analysis 2017
1st practical

A report must be uploaded on TEIDE before the deadline.

The report should contain graphical representations that are very important in statistics. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the questions in order and refer to the question number in your report.

Computations and graphics should be performed with the software R.

The report should be written using the Rmarkdown format. It is a file format that allows users to format documents containing text, R instructions and the results provided by R when evaluating instructions. The set of R statements is included in the .rmd document so that it may be possible to replicate your analyzes using the .rmd file. From your .rmd file, you are asked to generate an .html file for the final report. The set of .rmd commands and the procedure to generate .html files is explained in the Rmarkdown cheatsheet. In TEIDE, you are asked to submit both the .rmd and the .html files. In the html file, you should limit the displayed R code to the most important instructions.

1st practical : Prostate cancer.

In a 1989 paper, Stamey and colleagues examined, for patients with prostate cancer, the correlation between prostate cancer volume and a set of clinical and morphometric variables. These variables include prostate specific antigens (PSA), a biomarker for prostate cancer, and a number of clinical measures (age, prostate weight, etc.).

The goal of this practical is to construct a predictive model to predict the (logarithm of) cancer volume (**lcavol** variable) from the following measures (called predictors):

lpsa :	log(prostate specific antigen)
lweight :	log prostate weight
age :	age
lbph :	log of benign prostatic hyperplasia amount
svi :	seminal vesicle invasion
lcp :	log of capsular penetration
gleason :	Gleason score
pgg45 :	percent of Gleason scores 4 or 5

1. **Preliminary analysis of the data**

   Download the data on Chamilo and store the data in your current folder
   http://chamilo2.grenet.fr/inp/courses/ENSIMAG4MMFDAS6/document/Labs/prostate.data

   Create an Rmarkdown file in Rstudio that will contain both R code and textual description of your analysis.

   You will be able to read the data using the following commands. Build an object `pro` of class `data.frame`**,** which will contain for each individual the **lcavol** variable and the values of the 8 predictors. In the following, we use the R command `scale` to center and scale the variables.

```
prostateCancer = read.table("prostate.data",header=T)
pro1 = prostateCancer[,-ncol(prostateCancer)]  # remove the last column

pro = as.data.frame(cbind(scale(pro1[,2:9]),pro1[,1]))
# center and scale the different variables
# create an object of class data.frame with the 8 columns of predictors and the
lcavol variable to predict

names(pro) = c(names(pro1)[2:9],names(pro1)[1])   # keep the names of the column
```

   a) Visualize the data using scatter plots between all pairs of variables. Make sure to define axes in your figures.

Help : `?pairs`

Using these graphs, provide a list of the variables related to **lcavol.**

b) Using scatterplots, can you summarize the main correlations between the 8 different predictors?

2. **Linear regression**

a) Perform a multiple linear regression to build a predictive model for the **lcavol** variable. The variables **gleason** and **svi** should be considered as qualitative variables (`pro[,"gleason"]<-factor(pro$gleason)` and `pro[,"svi"]<-factor(pro$svi)`). Provide the mathematical equation of the regression model and define the different parameters. Use the function `summary` to display the regression table and explain what are the regression coefficients of the lines which names start by gleason.

b) Provide a summary of the main results of the regression.

c) What happens if predictors **lpsa** and **lcp** are removed from the model? Try to explain this new result.

d) Conclude about the effect of the **lcp** variable in the model of question a). Relate your conclusion to a 95% confidence interval you will provide for this regression coefficient.

e) What is the probability distribution of the variable *t* (*T* statistic) under the null hypothesis that should be defined? Provide the parameters of this probability density function.
For which value of *t,* a variable goes from . (0.05<P<0.1) to * (P<0.05). Help : `?qt`
Check that this threshold is compatible with the t-values found for the variables tagged with . and with *.

f) Plot the predicted values of **lcavol** as a function of the actual values. Plot both actual and predicted values as a function of index (after sorting the actual values and applying the same permutation to the predicted values).
Provide the value of the residual sum of squares.

Help : `?lm` In the item `Value`, you will find the different objects returned by the `lm` function. For instance if `obj_lm=lm(...)` then `obj_lm$residuals` contain the vector of residuals.
`?sort.int` to sort the values and keep the permutation of indices.

3. **Effect of the qualitative variables.**

a) By performing a one-way ANOVA, decide if the predictors **svi** et **gleason** affects **lcavol**.

4. **Best Subset selection**

a) Why may the model studied in part 2 a) not be optimal?
A regression model that uses *k* predictors is said to be of size *k*.
For instance, **lcavol**=$\boldsymbol{\beta_0}$+$\boldsymbol{\beta_1}$***lpsa**+ε and **lcavol**=$\boldsymbol{\beta_0}$+$\boldsymbol{\beta_1}$***lweight**+ε are models of size 1.
The regression model without any predictors **lcavol**=$\boldsymbol{\beta_0}$+ε is a model of size 0.
The objective of this part is to select for each value of *k* (*k=0...8*), the best model.

Preliminary question

- Describe the models implemented in `lm(lcavol~1,data=pro)`, `lm(lcavol~.,data=pro[,c(1,4,9)])` `and` `lm(lcavol~.,data=pro[,c(1,2,7)])` ?
- How can you automatically obtain the residual sum of squares of a regression model?
Help : the `sum` function
- How can you automatically perform all the regressions of size *k*=2 ?
Help : `?combn`

b)

- For each value of *k* in {0, ... , 8}, write R code to select the set of predictors that minimizes the residual sum of squares.
- Plot the residual sum of squares as a function of *k*.
- Provide the name of the predictors for each value of *k*.

c) Do you think that minimizing the residual sum of squares is well suited to select the optimal size for the regression models?

5. **Split-validation**

You have now found the best model for each of the 9 possible sizes. In the following, we wish to compare these 9 different regression models.

a) Give a brief overview of split-validation: how it works, why it is not subject to the issues raised in question 4c).

The validation set will be composed of all individuals which indices will be a multiple of 3. Store these indices in a vector called `valid` (use `(1:n)%% 3 == 0` where `n` is the number of individuals).

b) Let us assume that the best model is of size 2 and contains the i*th* and j*th* predictor (replace *i* and *j* by their true values). Describe what is performed when running the function `lm(lcavol~.,data=pro[!valid,c(i,j,9)])`. What is the mean training error for the model ?

c) With the regression model of size 2, predict values of **lcavol** for the validation set.
Hint: `?predict.lm` You have to provide to the `predict` function, using the `newdata` argument, the matrix contain the data of the validation set. Compute the mean prediction error and compare this error to the mean training error.
Hint: You can compare the difference between predicted and actual values using the - operator.

d) Reusing part of the code implemented in questions a)-c), perform split-validation to compare the 9 different models. Plot the training and prediction errors as a function of the size of the regression models. Choose one model, giving the parameter estimates for the model trained on the whole dataset, and explain your choice.
Hint: Make sure that all the different values of `gleason` and `svi` are present in the training dataset.

e) What is the main issue of split-validation ? Illustrate this issue on the cancer dataset.

f) Which method can address the problem of split-validation? Code this alternative method and comment the result.
Hint: `sample` to select indices at random.

6. **Conclusion**

What is your general conclusion about the choice of the best model to predic **lcavol**? Estimate the best model on the whole data set and comment the predictors and parameters. Compare the optimal prediction model to the model in question 2a).