Statistical Analysis and Document Mining

TP1: Multiple linear regression

February 2022

Each team has to upload a report on Teide before the deadline indicated at the course website. The report should contain graphical representations. 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 have to be performed with `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 running instructions. The set of R instructions 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. 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.

Part I and Part II are independent.

Part 1: Multiple regression on simulated data

1. Set the seed of the random generator to 0 (`set.seed(0)`). Simulate $6,000 \times 201 = 1,206,000$ independent random variables with the standard normal distribution. Store them into a matrix, then into a data frame with 6,000 lines and 201 columns. Each of these columns is referred to as a "variable". Useful commands: `rnorm, matrix, data.frame`.

2. Define a Gaussian multiple linear regression model using the last 200 variables to predict the first one. In the report, write a mathematical equation (do not write `R` code!) to define this model. Write a mathematical equation defining the true regression model associated with the data. Compare both models.

3. Estimate the parameters of the linear model using the last 200 variables to predict the first one. Compute the number of coefficients assessed as significantly non-zero at level 5%. Comment the result. Useful commands: `summary(reg)$coefficients`.

4. Simulate a sample of size $n = 1000$ of the following model:

$$\begin{array}{rcl} X_{1,i} & = & \varepsilon_{1,i} \\ X_{2,i} & = & 3X_{1,i} + \varepsilon_{2,i} \\ Y_i & = & X_{2,i} + X_{1,i} + 2 + \varepsilon_{3,i} \end{array}$$

   where $i \in \{1, \ldots, n\}$ and the ε_{ij} are independent $\mathcal{N}(0,1)$ random variables. For a given i, what is the distribution of $(X_{1,i}, X_{2,i})$? Plot the clouds of points of the simulated values of $(X_{1,i}, X_{2,i})_{i=1,\ldots,n}$.

   What is its shape? Why?

5. Let us consider the following 2 models:

$$\begin{array}{rcl} \text{Model 1: } Y_i & = & \beta_1 X_{1,i} + \beta_0 + \tilde{\varepsilon}_{1,i} \\ \text{Model 2: } Y_i & = & \beta_2 X_{2,i} + \beta_0 + \tilde{\varepsilon}_{2,i} \end{array}$$

   where the $\tilde{\varepsilon}_{j,i}$ are independent $\mathcal{N}(0,\sigma^2)$ random variables. For n $= 1000$, check that the estimates of the parameters $\beta_0, \beta_1, \beta_2, \sigma^2$ are close to the true values. Now set the seed to 3 and simulate again $X_{1,i}, X_{2,i}, Y_i$ for $n = 10$. Estimate the parameters. What happens?

6. Let us now consider the model

$$Y_i = \beta_2 X_{2,i} + \beta_1 X_{1,i} + \beta_0 + \varepsilon_i$$

   where $i \in \{1, \ldots, n\}$ and the ε_i are independent $\mathcal{N}(0,\sigma^2)$ random variables. For the previously simulated data with $n = 10$, estimate the parameters $\beta_0, \beta_1, \beta_2, \sigma^2$. What can you say about the effects of X_1 and X_2?

Part 2: Analysis of prostate cancer data

A medical study made on patients with prostate cancer aims to analyze the correlation between the prostate tumor volume and a set of clinical and morphometric variables. These variables include prostate specific antigens, a biomarker for prostate cancer, and a number of clinical measures (age, prostate weight, etc.). The goal of this practical is to build a regression model to predict the severity of cancer, expressed by logarithm of the tumor volume (`lcavol` variable) from the following predictors:

`lpsa`: log of a prostate specific antigen

`lweight`: log of prostate weight

`age`: age of the patient

`lbph`: log of benign prostatic hyperplasia amount

`svi`: seminal vesicle invasion

`lcp`: log of capsular penetration

`gleason`: Gleason score (score on a cancer prognosis test)

`pgg45`: percent of Gleason scores 4 or 5

The file `prostate.data`, available on Chamilo, contains measures of the logarithm of the tumor volume and of the 8 predictors for 97 patients. This file contains also an additional variable, train, which will not be used and has to be removed.

1. **Preliminary analysis of the data**

   (a) Download the file `prostate.data` and store it in your current folder. Read the data in `R` by `prostateCancer <- read.table("./prostate.data", header=T).` Use `attach(prostateCancer)` in order to attach the database to the `R` search path. Build an object `prostateCancer` of class `data.frame` that contains, for each patient, the `lcavol` variable and the values of the 8 predictors. Remove the last column (`train`) of the data frame.
   Help: You can remove columns in data frames by using negative indices to exclude them. With `headers = T in read.table`, the column names are given by `names(prostateCancer).`

   (b) Use the command `pairs` to visualize the correlations between all the variables. pairs plots scatterplots (clouds of points) between all pairs of variables. Analyse the correlations between all the variables and identify the variables which are the most correlated to `lcavol`.

2. **Linear regression**

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

   (b) Give confidence intervals of level 95% for all the coefficients of the predictors with `confint`. Comment the results.

   (c) What can you say about the effect of the `lpsa` variable? Relate your answer to the p-value of a test and a confidence interval.

   (d) Plot the predicted values of `lcavol` as a function of the actual values. Plot the histogram of residuals. Can we admit that the residuals are normally distributed? Compute the residual sum of squares.

   (e) What do you think of the optimality of this model?

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

3. **Best subset selection.** A regression model that uses k predictors is said to be of size k. For instance, `lcavol` $= \beta_1$ `lpsa` $+ \beta_0 + \varepsilon$ and `lcavol` $= \beta_1$ `lweight` $+ \beta_0 + \varepsilon$ are models of size 1. The regression model without any predictor `lcavol` $= \beta_0 + \varepsilon$ is a model of size 0.

   The goal of this question is to select the best model of size k for each value of k in $\{0...8\}$.

   (a) Describe the models implemented in

   ```
   lm(lcavol ~ 1, data=prostateCancer)
   lm(lcavol ~ ., data=prostateCancer[,c(1,4,9)])
   lm(lcavol ~ ., data=prostateCancer[,c(1,2,9)])
   ```

   Compute their residual sums of squares.

   (b) Compute the residual sums of squares for all models of size $k = 2$. What is the best choice of 2 predictors among 8?
   Help: `combn(m,k)` gives all the combinations of k elements among n

(c) For each value of $k \in \{0, \ldots, 8\}$, 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 names of the selected predictors for each value of k.

(d) Do you think that minimizing the residual sum of squares is well suited to select the optimal size for the regression models? Could you suggest another possibility?

4. **Split-validation.** You have now found the best model for each of the nine possible model sizes. In the following, we wish to compare these nine different regression models.

   (a) Give a brief overview of split-validation: how it works? Why it is not subject to the same issues raised in question 3(c)?

   The validation set will be composed of all individuals whose indices are 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 evaluated when using the function `lm(lcavol ~., data=prostateCancer[!valid, c(1, i, j)])`. What is the mean training error for the model ?

   (c) Predict values of `lcavol` on the validation set for the regression model of size two. Compute the mean prediction error and compare it to the mean training error. *Hint*: Use `?predict.lm`. Note that you will have to provide the matrix containing the data of the validation set to the `predict` function, using the `newdata` argument.

   (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.

   (e) What is the main limitation of split-validation ? Illustrate this issue on the cancer dataset. What could you do to address this problem for split-validation? Code such alternative method and comment the result.

5. **Conclusion.** What is your conclusion about the choice of the best model to predict `lcavol` ? Apply the best model and comment the results.