Data Analysis - TP 3

ISEP – November 6th, 2018

Instructions: Prepare a report including the source code and the results. Deposit your report on Moodle and don't forget your binome's name or to make 2 deposits if you did not work alone.

A Step by step linear regression using the "bats" data

A.1 Correlation circles

In this exercise, we will study a small sample from a dataset [Hutcheon et al. 2002], in which different characteristics of bats are studied. The descriptors are the following:

- Species = The name of the observed species
- Diet (1 = phytophage ; 2 = gleaner ; 3 = aerial insectivore ; 4 = vampire)
- BOW = Body mass
- BRW = Brain mass
- AUD = Auditory nuclei volume
- MOB = Main olfactory bulb volume
- HIP = Hippocampus volume

1. Open the file in R using the command **tab=read.table("tabBats.txt")**.
2. Display the content of the variable *tab*. What can you say about the different attributes ?
3. Prompt the classes of the different attributes using the command **str(tab)**. Remove from *tab* all attributes that may not be useful for a correlation analysis.
4. We want to do a quick Principal component analysis of this data set and draw the correlation circle in order to find correlated variables:
   - Use the following commands:
     library(FactoMineR)
     result <- PCA(tab)
   - Comment on the resulting graphs.
   - Browse the content of the variable *result* and comment.
5. According to the previous question, which variables are the most correlated ?

A.2 First linear regression using R

1. We are interested in finding whether there is a link between the body mass of a bat (BOW) and its brain mass (BRW):
   - Use the command **plot(tabBOW,tabBRW)**. What type of graphic is displayed ?
   - Describe the resulting graph. What type of link to you notice between the two variables ? Is this tendency verified for all individuals ?
   - Write down the equation of the regression model that seems suited for this data set.
2. Use the command **mod=lm(tab$BRW∼tab$BOW)** to start the linear regression. Display the variable *mod* to know the regression coefficients.
3. Use the command **plot(mod)**.
   - Explain the significance of each diagram displayed.
   - Based on the diagrams, what can you say on the validity of these regression results and on the data ?
4. Use the command **summary(mod)**.
   - Explain and comment the results on the correlation coefficients.
   - Explain and comment the results on the residuals.
   - Explain what the R^2 coefficient represent in this result.
   - Conclude on the validity of the model.
5. Use the following commands:
   plot(tabBOW,tabBRW)
   abline(coef(mod),col="red")
   Comment the resulting graph.

A.3 Second linear regression

1. Create a new array *tab2* from which you will remove "Pteropus vampyrus".
   - What is the difference between *tab* and *tab2* correlation wise. You can use **plot(tabBOW,tabBRW)** and **plot(tab2$BOW,tab2$BRW)** to compare them, or draw the correlation circle for *tab2*.
   - Why do you think it is better to work with *tab2* instead of *tab* ?
2. Do again questions B-2 to B-5 using *tab2*.
   - Compare the results and comment.
   - Are the result of this second regression better ? Explain why.
3. Use the following commands:
   plot(tabBOW,tabBRW)
   abline(coef(mod),col="red")
   abline(coef(mod2),col="blue")
   Comment the resulting graph.

B Application to the mansize dataset

By using the same methods as in the previous exercises, re-use the "mansize" dataset from last week and do the following analysis:

1. Remind the correlation between the different variable and confirm them by projecting the data into the PCA plan. Use correlation circles to illustrate your results.

2. Run a linear regression to predict the size of an indiviudal based on the size of his femur bone.

3. Comment the regression results by focusing on the different graphic and indexes computed by R.