Ciências ULisboa

ECOLOGIA NUMÉRICA

Exame teórico-prático – 1ª época 2018/2019
22/01/2018 – 3º ano Biologia

This is an exam to be implemented exclusively in R. All the code should be visible in the HTML resulting from compiling the .Rmd that you hand in.

1. (cotação 0.25) Create a folder where you create all the files and objects created during the exam, including any data sets you might use. Create a dynamic report in RMarkdown with a reasonable title, the author identification (name and number), where all code should be shown.
The name of the file should be EPENE1A1819A*****.Rmd. In case you are in a pair, the name should be EPENE1A1819A*****A*****.Rmd, where ***** represent the student(s) number(s). this document and any data sets used must be sent via e-mail (tamarques@fc.ul.pt). Each exercise should be clearly indicated with
# Exercise ?
## Exercise ?.!

(where ? takes the values 1 to 8 and ! from 1 to 5)

2. (cotação 0.25) Using the code below, create a simulated data set for a random variable Y, the ys, where Y represents an index of habitat complexity which can take any real value, which is a function of 6 potential independent variables (x1 to x6). The data were collected in two types of habitat ("hab") and the four seasons of the year ("est") (tip: use copy paste to avoid errors).

```
set.seed(****)
# **** day of the month of the student's birtday
#e.g. Carlos and Maria with birthdays on 1 e 13 May, 0113
b0=rnorm(1,0,0.5)
b1=rnorm(1,0,1)
b2=rnorm(1,0,0.5)
b3=rnorm(1,0,1)
b4=rnorm(1,0,0.5)
b5=rnorm(1,0,1)
b6=rnorm(1,0,2)
n1=rpois(1,30)
n2=2*n1
n=4*n1
hab= sample(x=c("H1", "H2"),size=n,prob=c(0.5,0.5),replace=T)
est= sample(x=c("P", "V", "O", "I"),size=n,prob=rep(0.25,4),replace=T)
x1=c(runif(n2,0,10), runif(n2,10,20))
x2=c(runif(n2,0,10), runif(n2,5,15))
x3= c(rnorm(n1,0,1), rnorm(n1,1,1),rnorm(n1,2,1),rnorm(n1,3,1))
x4= c(rnorm(n1,0,1), rnorm(n1,0,1),rnorm(n1,0,1),rnorm(n1,1,1))
x5=rnorm(n,15,2)
x6=rnorm(n,0,5)
torf=sample(x=0:1,size=6,prob=c(0.2,0.8),replace=T)
ys=b0+b1*x1*torf[1]+b2*x2*torf[2]+b3*x3*torf[3]+b4*x4*torf[4]+b5*x5*torf[5]+b6*x6
*torf[6]+rnorm(n,0,5)
```

2.1 (cotação 0.25) How many observation of ys did you generate?

2.2 (cotação 0.5) Is the sample size the same for the other students, or not? Explain your reasoning.

3. Based on the data generated in the previous exercise, we want to know if the habitat has an influence on the values of the complexity index

3.1 (cotação 0.25) How many observations were collected for each habitat "H1" e "H2", as defined in variable hab

3.2 (cotação 1) Compare with a suitable plot the values of the index in each habitat and interpret the results.

3.3 (cotação 2) Use a statistical test to test if the index is different in the 2 habitats.

4. With the same data created in exercise 2, we want to know if the index values depend n season.

4.1 (cotação 1) Compare with a suitable plot the values of the index in each season and interpret the results.

4.2 (cotação 2) Use a parametric statistical test to test if the index is different in the 4 habitats.

4.3 (cotação 1) If you found significant differences implemented the a posteriori test and interpret the results. If you have not found differences, describe the tests you can use for comparisons a porteriori in a non-parametric context.

5. (cotação 0.25) With the same data created in exercise 2, create a `data.frame` and call it `datays`, suitable to implement a multiple regression analysis to explain the dependent variable `ys` as a function of the independent variables (just `x1` to `x6`, ignore habitat and season).

5.1 (cotação 0.75) Implement the multiple regression that explains ys as a function of x1 to x6.

5.2 (cotação 1) based on your data, what are the variables that seem important to explain the `ys`?

5.3 (cotação 1) What is the expected value of the index for a place where the observed values form `x1` to `x6` are those observed in your sample?

5.4 (cotação 1) What is the observed R-Square value and what can you conclude from your ability to predict the index based on the available covariates.

5.5 (cotação 0.75) In the code above, what was `torf`? Based on this, what are the variables that truly influenced `ys`? Did you make any type I or type II errors? Explain your reasoning.

6. (cotação 0.25) Execute the code

```
set.seed(****)
# **** dias do mês em que o aluno faz anos,
#e.g. Carlos e Maria com anos a 1 e 13 de Maio, 0113
file=ceiling(runif(1,0,100))
```

6.1 (cotação 0.25) What values can `file` take?

6.2 (cotação 0.25) What is the distribution family for the random variable generated?

6.3 (cotação 0.25) What is the probability of observing a value smaller than 50?

6.4 (cotação 0.25) What is the number inside object `file`?

7. (cotação 0.25) Read file “data4EPENg*.txt”, where you replace the * by the number in the object file `file` created in exercise 6. In this data set we have the abundances of 9 species of 3 genera of birds detected in point counts of 10 minutes. The first column contains the corresponding habitat.

7.1 (cotação 0. 5) How many places were present in each habitat?

7.2 (cotação 0.5) Calculate the Euclidean distances between the locations. Identify the two places with the largest difference between them and report that distance value.

7.3 (cotação 1) Implement a cluster analysis using theclustring methods “single” and “complete”.

7.4 (cotação 1) Which of the two analysis provides the best separation per habitat?

8. (cotação 0.25) Read the file “data4EPENdt*.txt” , ”, where you replace the * by the number in the object file `file` created in exercise 6. In this data set we have the environmental variables collected in locations along a river, where the places were ordered from upstream to downstream (or from downstream to upstream). The available variables are depth (prof), altitude (alt), oxygen (O2), pH (pH), salinity (sal), suspended particles (sus), Mercury concentration (Mg) and Lead concentration (Pb).

8.1 (cotação 0.5) Implement a PCA suitable to describe the places as a function of their characteristics

8.2 (cotação 0.5) What is the proportion of the variance explained by the first two axis?

8.3 (cotação 0. 5) How many axis would you recommend retaining for interpretation? Why?

8.4 (cotação 1) Interpret the PCA byplot. Can you identify if locations with small numbers correspond to upstream or downstream?

8.5 (cotação 0.5) If you had to send the police to investigate a factory that might be polluting the river with Mercury, where would you tell them to start looking?

My grade will be:

Cotação total
0.25+
0.25+0.25+0.5+
0.25+1+2+
1+2+1+
0.5+0.5+1+1+1+0.75+
0.25+0.25+0.25+0.25+0.25+
0.25+0.5+0.5+1+1+
0.25+0.5+0.5+0.5+1+0.5
=21