Exercises

Juan R Gonzalez[*1]

[1]Bioinformatics Research Group in Epidemiology, Barcelona Institute for Global Health, Spain

*juanr.gonzalez@isglobal.org

24 enero 2019

Contents

1 R basics

1. Create a vector `x` with the following entries:

```
3 4 1 1 2 1 4 2 1 1 5 3 1 1 1 2 4 5 5 3
```

2. Check which elements of `x` are equal to 1 (Hint: use "==" operator).
3. Create a variable (e.g vector) `y` having the logarithm of `x`.
4. Create a vector `z` containing the first five elements of `x`.
5. Create a sequence of numbers from 1 to 20 in steps of 0.2 (see help `seq` function).
6. Concatenate `x` and `y` into a vector called `newVec`.
7. Display all objects in the workspace and then remove `newVec` object.
8. Create a `data.frame` called `elasticband` having these two variables.

```
stretch <- c(46,54,48,50,44,42,52)
distance <- c(148,182,173,166,109,141,166)
```

9. Compute the mean of the variable `stretch` from the `elasticbanc` object (Hint: use "$").
10. Which is the class of the variable `distance`?

2 Descriptive analyses

The file `multicentric.txt` contains the data of a multicentric case/control study to discover risk factors of cervical cancer. The file `multicentric.htm` contains a description of the variables included in this file.

1. Import the data in an object called `multi` (Hint: data are in text tab-delimited format: use `read.delim` function).
2. Create anoter variable called `edad1sex.cat` having age at first sexual intercourse (variable `edad1sex`) in three categories: <15, 15-18, 19-25, +25. How many woman had her first sexual intercouse before 15 years old?
3. How many individuals and variables are in this database?
4. Describe the types of variables you have.
5. Create a table describing the number of cases and controls (variable `status`) of each country (variable `pais`) included in the study. Which is the country having more cases? Which is the percentage of cases in Brazil?
6. Create another table describing the number of cases and controls and human papillomavirus (variable `vph`). Do you think those variables are associated? Why?
7. Summarize the duration consuming oral contraceptives (variable `durco`) in cases and controls, in each country and by educative levels (variable `niveledu`), respectively.
8. Categorize the variable number of pregnancies (variable `nembara`) in quartiles. Which is the number of cases in the last quartile of number of pregnancies? and controls?
9. Create a boxplot to compare the age at first pregnancy (variable `edademba`) between cases and controls (variable `status`). Do you think this variable could be a risk factor of cervical cancer?

10. Create an histogram to describe the age when women received their first test of papillomavirus (variable `edad1pap`). Is this variable normally distributed? Could you apply a test to justify your answer? (Hint -> google)

3 Hypothesis testing

1. Using the same study on cervical cancer ... perform a chi-squared test or a t-test to assess which variables are associated with `status` variable (e.g case/control). NOTE: use the appropriate test when analyzing continuous variables.

4 Statistical modelling

1. Fit univariate logistic regression models to investigate which factors (`edad`, `niveledu`, `fumar`, `edad1sex`, `nembara`, `vph`) are associated with cervical cancer (variable `status`).
2. Select the set of variables that better predicts the probability of being diagnosed with cervival cancer. (NOTE: use complete cases 'multi.comp <- multi[complete.cases(multi),]).
3. [**Advanced**]. Compute the area under the roc curve (AUC) of the selected model (Hint: investigate `pROC` package).
4. The file `retinol.doc` describes a study to investigate which are the determinants of retinol observed in plasmatic concentrations since low levels of those micronutrients are associated with some types of tumors. Let's try to decipher whihc are those determinants. To this end:
   - Fit single regresion models to assess association between retinol measured at plasma (variable `retplas`) and variables: `retdiet`, `colest`, `alcohol`, `fibra`, `grasa`, `calorias`, `vitamin` and `fumador`.
   - Select the most significant association and test model assumptions and check linearity.
   - Use an automatic method to select the set of variables that better predict the variable `retplas`. Which is the percentage of variability explained by this model?
5. The file `pulmon.doc` describes a study to investigate which are the most predictive variables of survival in patients diagnosed with cancer. Perform the all the required analyses to answer this scientific question (e.g. Kaplan-Meier, log-rank, Multivariate Cox models).

5 Graphics

1. Load the data available in the file `retinol.txt`. Log-transform the variable `retplas` into another variable called `logretplas`.
2. Create a histogram and a boxplot of `logretplas` variable and plot them side-by-side on the same graphing region. Label the axes accordingly. Save your results as a Jpeg file.
3. Plot `logretplas` (y-axis) versus `retdiet` (X-axis) using an appropriate plotting command. Put a title on the graph and labels on the axes.

4. Fit a linear regresion model between both variables (call this regression `mod`). Add the estimated regression line to the current plot and make it the colour blue.
5. Extract the values of the residuals using `resids <- resid(mod)`. Check that the residuals are normally distributed by creating a Q-Q plot.
6. Create the same plot for each of the three levels of the variable `fumador`. (Hint: Use `coplot`).
7. Build the same plot separated for males and females (variable `sexo`). (Hint: use `xyplot` from package `lattice`).
8. Construct a histogram of `calorias` and overlay the density curve. (Hint: Need `hist`, `lines` and `density`.)
9. Create in one figure (e.g. use `par(mfrow=c(2,1))`) the histogram of variable `calorias` for males and females (variable `sexo`).

6 Reproducible Research

1. Get the R code to answer the questions in the section **R basics** that are available here: (https://github.com/isglobal-brge/R_course/blob/master/Answer_exercises/Exercise_1.R) and create a Markdown document having each task in a section
2. Create the same report for the answers to the exercises of section **Graphics** that are available here (https://github.com/isglobal-brge/R_course/blob/master/Answer_exercises/Exercise_4.R)
3. Load the `multicentric.txt` data into R and create (using `compareGroups` package):
   - A table describing the variables: `edad`, `niveledu`, `fumar`, `edad1sex`, `nembara`, `vph` by cases and controls (variable `status`).
   - A table giving the ORs of developing cervical cancer (variable `status`) for the variables: `edad`, `niveledu`, `fumar`, `edad1sex`, `nembara`, `vph`
4. Load the `pulmon.txt` data into R and create:
   - A table describing the survival (variables `tiempo` and `estado`) for the variables: `sexo`, `edad4`, `estclin`, `ik5`, `cirugia`, `quimio`, `radioter`.
   - A table giving the HR of mortality (variables `tiempo` and `estado`) for the variables `sexo`, `edad4`, `estclin`, `ik5`, `cirugia`, `quimio`, `radioter`.