MASM22/FMSN30(40) – Linear and Logistic Regression (with Data Gathering) 2019

COMPUTER EXERCISE 1: SIMPLE LINEAR REGRESSION

You should first run through *Introduction to RStudio and RStudio projects* and *Basic computations in R*

`http://www.maths.lth.se/matstat/kurser/masm22/lab0_rprojects`

`http://www.maths.lth.se/matstat/kurser/masm22/lab0_basics`

in order to get familiar with the program, then come back here.

We will use R and Rstudio as a computational platform. There are *many* resources freely available on the web to get started with R, see the course home page.

> Warning: in order not to loose your work, create an RStudio project. Then write your codes in a file having extension `.R`, such as `lab1.R`. Save the file in your RStudio project directory.

Write all relevant commands to the script file and run them from there. Add comments and save the file. That way you can easily copy and modify the commands.

A table containing data for statistical analysis is called a `data frame`. R will not consider a data-frame just the same as a mathematical matrix, instead this is *"a two-dimensional array-like structure, in which each column contains measurements on one variable, and each row contains one case."*. Just what you would expect when dealing with some sort of tabled data.

The exercise uses a table with data (see next page). Since it is a small data set, we can create it by running the following (copy/paste will work) from the script file:

```
emission <- data.frame(
  vehicles = c(28, 36, 15, -19, -24, 8, 25, 40, 63, 12, -6, 21),
  pollution = c(22, 26, 15, -18, -21, 7, 21, 31, 52, 8, -7, 20)
  )
```

This creates a data-frame `emission` having columns named `vehicles` and `pollution`. The data for each column have been collected using `c()`.

Look at the data by running `emission`. Also run `summary(emission)`. Check that you have no missing data and that the min, max, and average values seem correct. This is a quick way to find some of your input errors. You can also see the data by clicking on the variable in the environment window in the upper right, or giving the command `View(emission)` (with a capital V). This will open a read-only view of the data.

You can access individual columns by their names using `$`, for example type `emission$pollution`. Otherwise treat the data as a matrix, and type `emission[, 2]` to access the entire second column.

You can get a list of the variable names using `names(emission)`.

A list of useful R-commands for simple linear regression is found on the course home page

`http://www.maths.lth.se/matstat/kurser/masm22/lab1_vt19_useful.pdf`

Exercise 1: Air pollution

The level of pollution because of vehicular emissions in a city varies with the flow of vehicles. The local government has measured the change in flow of vehicles as well as the change in the level of air pollution (both in percentages) on 12 days giving the following results:

Change in flow of vehicles (%)	Change in level of air pollution (%)
28	22
36	26
15	15
−19	−18
−24	−21
8	7
25	21
40	31
63	52
12	8
−6	−7
21	20

(a) From the description above, which variable is the "response variable"?

(b) Obtain a plot of the data with the response variable on the y-axis. Convince yourself that a linear model seems a sensible choice.

(c) Define a simple linear regression model of the response variable on the chosen explanatory variable (covariate). Use R to fit the linear model, then add the estimated model to the plot with data. Does it seem reasonable?

(d) Calculate a (pointwise) confidence interval for the line $E(Y_0) = \beta_0 + \beta_1 X_0$ and add it to the plot. Does it seem reasonable?

(e) Calculate a (pointwise) prediction interval for the observations $Y_{\text{pred},0} = \beta_0 + \beta_1 X_0 + \varepsilon_0$ and add it to the plot. Does it seem reasonable?

(f) Check that the residuals follow the assumptions.

(g) Obtain the estimated regression coefficients and the unbiased estimate of the error variance σ^2.

(h) Test the significance of the null hypothesis that the expected change in air pollution is zero when there is *no change* in the vehicle flow.

(i) Test the significance of the null hypothesis that the expected change in air pollution is the same *regardless* of the change in the vehicle flow.

(j) Construct a 95 % confidence interval for the "vehicles" parameter.

(k) We might expect that a specific change in the vehicle flow would result in an equally large change in the level of air pollution, i.e., $\beta_1 = 1$. Test whether data support this conjecture.