NMSA407: Linear Regression

Winter Term 2018/2019

General Instructions & Homework Assignment no.1

(Submission Deadline: Exercise class no.3)

ℹ General Instructions

- The homework assignment can be carried out in a group of 1 – 3 students (three students per each group are recommended). Different groups can be formed to work on future homework assignments (there will be three assignments during the term).
- Each group is required to submit a well-written PDF document created with LaTeX. All content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis). A computer code or originally formatted computer output should not appear in the document.
- The document can be submitted either in English or Czech/Slovak (Czech and Slovak are allowed to be used within one document). Do not use English and Czech/Slovak in one document. The submitted document must contain the names of all members of the group – the names must be clearly provided in the header on the first page.
- **For each part of this homework assignment do all of the following:**
  1. provide at least one table with descriptive statistics being useful in the context of the problem and comment the values in the table within the context of the given problem;
  2. provide at least one plot being useful in the context of the problem and interpret the figure;
  3. define a probabilistic model that you are about to use and discuss on the model's assumptions;
  4. formulate the set of hypotheses which are tested – explain which test will be used to do so; provide the formula for the test statistic;
  5. state the distribution of the test statistic under the null hypothesis and specify whether this distribution is exact or asymptotic;
  6. provide the value of the test statistic and the corresponding p-value;
  7. formulate your conclusion and interpret the results (understandable for non-statisticians as well);
  8. discuss, which assumptions might not be satisfied;
     (Is it crucial for the validity of the performed test?)
- All statistical tests should be performed at $5\,\%$ significance level, confidence intervals should be all with $95\,\%$ coverage.
- **DEADLINES and FORM OF DELIVERY:**

Monday's session	(*Matúš Maciak*)	Due on 15/10 (09:00)	PRINTED ON PAPER
Tuesday's session	(*Matúš Maciak*)	Due on 16/10 (12:20)	PRINTED ON PAPER
Thursday's session	(*Stanislav Nagy*)	Due on 18/10 (10:40)	PRINTED ON PAPER

- For groups composed of students from different exercise class groups, only one document is required to be delivered to an arbitrarily chosen lecturer within the deadline that applies for the group of this specific lecturer. On the title page, include the names of the authors and provide the exercise group identification to which you want to submit your homework solution.

ℹ Data Description

- the datafile (a `.txt` file) can be downloaded by clicking on the following link: NMSA407-1819-HW1.txt or by downloading from the central webpage of the exercise classes.
- if you download the data into your working directory (check/set your working directory using commands `getwd()` and `setwd()`), you can load them into the `R` environment using the following command:

```
> DATA <- read.table("NMSA407-1819-HW1.txt", header = T)
```

- The data comes from a huge experiment conducted by The European Commission and JRC (Joint Research Centre) a few years ago. The idea of the experiment was to assess the ecological quality of European rivers (in the dataset there are only four Central European countries included). The quality is assessed first with a national assessment method, and later by a European international method. These two assessments methods can be very different (depends on the methodology behind of each member state assessment system). However, they are expected to be highly positively correlated.
  The data consists of 5 variables and 150 independent measurements:

  a) *EU.country* - European country specific code (1: Austria (AT), 2: Romania (RO), 3: Slovenia (SI) and 4: Slovakia (SK)) to identify in which country the specific river quality measurement took place;

  b) *CM.value* - a river quality index given by a European assessment method - Common Metric Value. Technically, it is a standardized value from the interval $[0, 1]$, where 0 stands for the lowest quality and 1 for a high quality status.

  c) *nationalEQR* - a river quality index given by a country specific national assessment method; The idea of the index is the same - it is a standardized value from the interval $[0, 1]$ with 0 for the lowest quality and 1 for high quality sites. However, the methodology behind this index might be very different (and depends on the member state) from the methodology of the European Common Metric value (*CM.value*).

  d) *OK.Status* - a binary variable distinguishing good (value 1) and bad (value 0) quality status: given the European legislative, a river quality below 0.4 of the international Common Metric scale (*CM.value*) is considered to be bad and special actions are required to improve the overall river quality (that takes a lot of time and lot of investments). Values above 0.4 of the Common Metric scale are considered to be OK.

  e) *history* - a binary variable coding for previous experiences of the given country: value one means that the given country already implemented the European assessment method into its national standards and the river quality is assessed using the European assessment method for the national purposes as well; value zero stands for countries which are new to the European assessment methods and are still using the country specific national assessment method to assess the river quality for national purposes;

☞ Homework 1 Assignments

Part 1:

Is the river quality in general different in the four considered EU member states (AT, RO, SI, SK)?
Considered both indices — the European assessment method (*CM.value*), and the country specific national method as well (*nationalEQR*). By a suitable quantity (that involves also the random character of data) describe the possible differences.

Part 2:

Consider a situation where instead of the quality index expressed by *CM.value* you only have an information, whether the value is above $\mathbf{0.4}$ or below (technically it means whether the river is assigned with a bad ecological status or a good one) — variable *OK.status*. Can we say that the quality status of different rivers depends on the member state (AT, RO, SI and SK)?

Part 3:

Consider the national quality assessment method (variable *nationalEQR*) and check whether it depends on the experience of the country with the European assessment method (variable *history*). What could be an appropriate interpretation of the result?
In addition, calculate the point estimate and the interval estimate for an appropriate quantity (state which quantity you consider) which might be used to quantify the dependence being examined. Interpret the calculated confidence interval to a non-statistician.