NMSA407: Linear Regression

Winter Term 2017/2018

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 is 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 document submitted 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 give a suitable interpretation of the figure;
  3. define a probabilistic model that you are about to use and provide at least a brief discussion on the model's assumptions;
  4. formulate the set of hypotheses which are tested – explain which test will be used to do so; and provide the right 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 provide an interpretation of 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:**

Tuesday's session at 10:40	(*Matúš Maciak*)	Due to 17/10 (13:50)	PRINTED ON PAPER
Tuesday's session at 12:20	(*Matúš Maciak*)	Due to 17/10 (13:50)	PRINTED ON PAPER
Friday's session at 9:00	(*Stanislav Nagy*)	Due to 20/10 (10:40)	PRINTED ON PAPER

- For groups which are 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) which you need for the first homework assignment can be downloaded by clicking on the following link: NMSA407-1718-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("NMSA407-1718-HW1.txt", header = T)
```

- data can be also loaded online (if your computer is connected to the internet) using command:

```
DATA <- read.table("http://www.karlin.mff.cuni.cz/~maciak/NMSA407/NMSA407-1718-HW1.txt", header=T)
```

- The data file gathers information about the overall crime rates in 622 US cities in years 1991 and 1999. The rates represent the total number of crimes per 100 000 inhabitants in each city in the given year. Each city is represented by one independent observation (a single row in the dataset) with the two crime rate variables and 2 additional covariates. A detailed description of the data is given below.

  a) *City* - the name of the city;
  b) *State* - factor covariate with 48 levels which represents the abbreviation of the US state where the city is located;
  c) *Region* - a four level factor covariate representing the US region where the city is located (NE northeast, MW midwest, S south, W west);
  d) *Crime91* - total number of crimes per 100 000 inhabitants (rounded to the nearest integer) in the year 1991, according to the US Census Bureau;
  e) *Crime99* - total number of crimes per 100 000 inhabitants (rounded to the nearest integer) in the year 1999, according to the US Census Bureau.

☞ Homework 1 Assignments

Part 1:

Consider the change in the crime level (variable *Crime99 - Crime91*) in each city. Can we say that this change is the same for each US region? By a suitable quantity (that involves also the random character of data) describe the possible differences.

Part 2:

Does the change in the crime depend on the overall level of the crime in 1991 (variable *Crime91*)? How?

Part 3:

Instead of the overall change in the crime level (quantity *Crime99 - Crime91*) consider only an information about whether the number of crimes per 100 000 inhabitants in the city increased between 1991 and 1999, or not. Is this simplified variable different across the US regions?