NMSA407: Linear Regression

Winter Term 2019/2020

General Instructions & Homework Assignment no. 1

(Submission Deadline: Exercise class no. 3)

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 homework assignments all together during the term).

- Each group is required to submit a well-written `pdf` document created in LaTeX. All content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis). Computer codes or originally formatted computer outputs 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 language use the plot labels and figures should correspond with the language used in the main document. The submitted document must contain the names of all members of the group and their exercise group identification 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 framework 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 the theoretical assumptions behind this model;
  4. formulate the set of hypotheses which are tested and explain which test is used;
  5. provide the explicit formula for the test statistic, 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 (the interpretation must be understandable for non-statisticians as well);
  8. discuss which assumptions might not be satisfied and, also, how much crucial for the validity of the performed test these assumptions are.

- All statistical tests are performed at $5\,\%$ significance level, confidence intervals should be all with the 95% coverage.

- **DEADLINES and FORMS OF DELIVERY:**

Tuesday's session	(*Matúš Maciak*)	Due on 15/10 (13:50)	PRINTED ON PAPER
Tuesday's session	(*Stanislav Nagy*)	Due on 15/10 (14:00)	PRINTED ON PAPER
Thursday's session	(*Matúš Maciak*)	Due on 17/10 (10:30)	PRINTED ON PAPER

- For groups composed of students from different exercise sessions, 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.

☞ Data Description

❑ The datafile (a `txt` file) can be downloaded by clicking on the link NMSA407-1920-HW1.txt or by downloading the corresponding file from the central webpage of the NMSA 407 exercise classes.

❑ The downloaded data file can be loaded from the working directory (check/set your working directory using the commands `getwd()` and `setwd()`) into the `R` working environment by using the following command:

```
> DATA <- read.table("NMSA407-1920-HW1.txt", header = T)
```

❑ Alternatively, the data can be also loaded into the `R` working environment directly (if the computer is connected to the internet) by using the command:

```
> DATA <- read.table("http://www.karlin.mff.cuni.cz/˜maciak/NMSA407/NMSA407-1920-HW1.txt", header=T)
```

❑ The data file contains some information about two guys from the Czech Republic hitchhiking in Europe over last couple of years. There are 180 hitchhiking trips recorded in the data (each trip is represented as one independent observation — a single row in the dataset) with 7 available covariates. The covariates description is given below.

a) *hitchhikers* - a two level factor covariate expressing how many hitchhikers were hitchhiking on the trip (“two guys” or a “single guy”);

b) *travelDistance* - a four level factor covariate recording the overall distance traveled with the hitchhiked car (four distinguished categories are: below 50 km, between 50 and 100 km, between 100 and 500 km, and, finally, over 500 km);

c) *driverGender* - the hitchhiked car driver’s gender (male or female);

d) *country* - a ten level factor covariate to indicate in which country the hitchhiking took place. There are 9 European countries recorded separately (*CZ* - Czech Republic, *D* - Germany, *ES* - Spain, *EST* - Estonia, *F* - France, *FIN* - Finland, *P* - Portugal, *PL* - Poland, and *RUS* - Russia). The last level is assigned to the remaining European countries denoted as *other* which are not listed separately listed above;

e) *waitingTime* - a continuous covariate which reflects the total waiting time needed to finally hitchhike some car (the corresponding values are given in minutes);

f) *carNumber* - the number of cars passing by until the first car stopped to pick the hitchhikers up;

g) *dayNight* - a two level factor covariate indicating which part of the day the hitchhiking took place (day or night).

✍ Homework 1 Assignments

Part 1:

Consider the waiting time (the covariate “*waitingTime*”) needed to hitchhike a car. Can we say that hitchhiking in the Czech Republic is different from hitchhiking in other countries? Use a suitable quantity (that involves also the random character of data) to describe this difference.

Part 2:

Does the waiting time depend on the distance the hitchhikers intend to travel?

Part 3:

Instead of the waiting time information consider only a reduced information whether the hitchhiker(s) needed to wait more than an hour or not. Is there the same chance to hitchhike a car within an hour no matter what is the distance the hitchhiker(s) want to travel?