NMSA407: Linear Regression

Winter Term 2016/2017

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, computer-prepared PDF document created with some appropriate software (e.g. LaTeX, OpenOffice Writer, MS Word, ...). 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:**
  Group Matúš Maciak (Monday): 17/10 (15:40) PRINTED ON PAPER
  Group Matúš Maciak (Tuesday): 18/10 (9:00) PRINTED ON PAPER
  Group Marek Omelka (Tuesday): 18/10 (14:00) 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.

Data Description

- the datafile (a *txt* file) which you need for the first homework assignment can be downloaded by clicking on the following link: NMSA407-1617-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("NMSA407-1617-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/NMSA407-1617-HW1.txt", header=T)
```

- The data file gathers some information about two guys from the Czech Republic hitchhiking in Europe over last couple of years. There 181 hitchhiking trips recorded in the data (each trip represented by one independent observation - a single row in the dataset) and 7 recorded covariates. The covariates description is given below.

  a) *hitchhikers* - two level factor covariate expressing how many hitchhikers were hitchhiking on the trip;

  b) *travelDistance* - four level factor covariate stating what was the distance travelled with hitchhiked car (below 50 km, between 50 and 100 km , between 100 and 500 km and finally, distance 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) and the last level is assigned to the remaining European countries (*other* - other country than previously listed).

  e) *waitingTime* - continuous covariate which stands for the total time needed to finally hitchhike some car (values given in minutes).

  f) *carNumber* - the number of cars passing by before the first car stop to load the hitchhikers in;

  f) *dayNight* - two level factor covariate to indicate in which part of the day the hitchhiking took place (day or night);

Homework 1 Assignments

Part 1:

Consider the waiting time (variable *waitingTime*) needed to hitchhike a car. Can we say that hitchhiking in the Czech Republic is different from hitchhiking in the rest of the Europe? By a suitable quantity (that involves also the random character of data) describe this difference.

Part 2:

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

Part 3:

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