NMSA407: Linear Regression

Winter Term 2020/2021

General Instructions & Homework Assignment no. 1

ℹ General Instructions

- The homework assignment can be carried out in a group of 1 or 2 students (two students per each group are recommended).
- 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. 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 used for the plot labels and figure captions 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.
- **`R` script:**

  As part the solution, please provide also a working and well commented `R` script of the analysis that you performed. The script is only complementary to the report, and will typically not be checked completely. All results of the study have to be described in the `pdf` file of the report in full.

- **Submissions:**

  Solutions to the homework (`pdf` file and the accompanying `R` script) are to be both uploaded to **SIS**. After logging-in, click on *'Studijní mezivýsledky'* (in Czech) or *'Study group roster'* (in English) and select the corresponding group where you can upload your files. The electronic deadline for the homework delivery is **October 22, 2020 (23:59 CEST)**.

- **Revisions:**

  Revised homework solutions have to be accompanied by a letter with a list of all significant changes made to the originally submitted document (e.g. *"A new paragraph with the description of Figure 1 was added on page 8;"*, *"Formatting of the tables throughout the document was improved as suggested;"*, or *"Section 7 of the document was rewritten completely."*). In case when some of the suggested improvements are not carried out in the revision, reasons for not including them should be explained in the letter.

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

☞ Data Description

We consider the official census data from US by county. The complete dataset contains information on 3148 counties in 51 American states.

- ❏ The `.RData` file containing the data can be downloaded by clicking **here**, or by downloading the corresponding file from the **central webpage** of the NMSA407 exercise classes. In R, the data is loaded into the working environment using the command

```
> (load("NMSA407-2021-HW1.RData"))
```

- ❏ Alternatively, the data can be loaded into the R working environment directly by using the command:

```
> (load("http://www.karlin.mff.cuni.cz/~maciak/NMSA407/NMSA407-2021-HW1.RData"))
```

- ❏ The loaded dataset should be available as a data frame called **US_Data**.

For the 3148 counties, the following variables are available (some identifiers are shortened):

name_16: name of the county;

School: percentage of population under 21-years-old enrolled to educational institutions;

Median.Earnings: median earnings of an adult in the county;

xxx.Population: percentage of county population that identifies as **xxx** (**White.not.Latino** for white – **Latino** for Hispanics and Latin Americans);

State: state in which the county is located;

median_age: median age of the population in the county;

Children: percentage of children raised in a single parent household;

Adult.obesity: percentage of obese adults.

- ❏ **For parts 1–4 of this homework assignment do the following:**

(a) provide at least one table with descriptive statistics useful in the context of the problem and comment the values in the table within the framework of the given problem;

(b) provide a plot being useful in the context of the problem and interpret the figure;

(c) define a probabilistic model that you are about to use and discuss the theoretical assumptions behind this model;

(d) formulate the set of hypotheses which are tested and explain which test is used;

(e) 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;

(f) provide the value of the test statistic and the corresponding p-value;

(g) formulate your conclusions and interpret the results (the interpretation must be understandable for non-statisticians as well);

(h) discuss which assumptions might not be satisfied and, also, how much crucial for the validity of the performed test these assumptions are.

✍ Homework 1 Assignments

Part 1:
We are interested in the median income in the counties of Texas, and its relation to the racial composition of the county.

(a) Is the median income in the counties of Texas where non-Hispanic whites are a majority (that is, more than 50 % of the total population) significantly higher than in other Texas counties?

(b) Provide a suitable confidence interval for the quantity which reflects the difference in the median incomes between both groups from (a).

(c) Can anything interesting be observed about the relation of the racial composition of the county and its median income in the state of Texas?

Part 2:
Consider the four south-western states — Texas, Oklahoma, Arizona, and New Mexico.

(a) Can we say that the obesity rates in these four states are the same?

(b) If the answer to the question from (a) is negative, can we assess which of these states differ significantly from each other, in terms of their obesity prevalence?

(c) Can we use some statistical approach to rank these four states from "the least obese" to the "most obese" one?

Part 3:
Does the percentage of children enrolled to educational institutes in the counties of Texas depend on the percentage of children raised in single parent household? If it does, quantify this relation by a suitable numerical characteristic, derive a confidence interval for it, and comment on its relation to the studied problem.

Part 4:
We say that a county is

- *rich* if the median income in the county is at least USD 30 000;
- *medium* if its median income is at least USD 20 000, but less than USD 30 000;
- *poor* if its median income is less than USD 20 000.

Consider the four southern states — Georgia, Alabama, Mississippi, and Louisiana.

(a) Is the proportion of counties classified as poor/medium/rich in the four southern states the same?

(b) Can we quantify the differences in wealth between the states, in terms of the classification above?

(c) Is it true that the proportion of rich counties in Georgia is higher than the proportion of rich counties in Alabama by at least 5 percentage points?