NMSA407: Linear Regression

Winter Term 2021/2022

General Instructions & Homework Assignment No. 2

(Submission Deadline: December 31st, 2021)

ℹ General Instructions

- ❏ The homework assignment can be carried out in a group of 1 – 2 students. The groups are not required to be the same as those in which you elaborated the first homework assignments.
- ❏ 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 used in 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.
- ❏ **`R` script**

  As part of the solution, please provide also a working and well commented `R` script of the analysis that you performed. The `R` script is only complementary to the report, and will typically not be checked completely. All results of the study have to be fully described in the `pdf` of the report.
- ❏ **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 **December 31, 2021 (23:59 CET)**.
- ❏ **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 $95\ \%$ coverage. With each formal test performed, it is necessary to specify (mathematically) the statistical hypothesis, provide the formula for, and the value of the test statistic, specify the distribution of the test statistic and the p-value, and your conclusion expressed in words understandable to a nonstatistician.

☞ Data Description

For the second homework assignment a dataset of 108 neurodegenerative dementia patients (from Mayo Clinic, Rochester, US) is considered. There are three types of dementia patients included in the dataset: patients suffering from the Alzheimer disease, patients with a frontotemporal lobar degeneration, and patients with a Lewy bodies dementia. In addition, there is also a control group consisting of patients with no dementia. One of many effects of the neurodegenerative dementia disease is a progressive decrease of the volume of specific parts of the patient's brain. In the dataset we consider the volume of the hippocampus part.

We want to estimate the expected volume of the patient's hippocampus given some additional patient's specific information provided in the data.

- ❏ The datafile (an `RData` file) can be downloaded by clicking on the link NMSA407-2122-HW2.RData.
- ❏ The dataset contains 108 independent observations (patients) and 8 covariates.

  a) `diagnosis` - four level factor distinguishing four groups of patients: `NC` - control group; `AD` - Alzheimer patients; `FTLD` - frontotemporal lobar degeneration patients; `DLB` - dementia with Lewy bodies;

  b) `gender` - two level factor: 0 – female; 1 – male;

  c) `age` - patient's age;

  d) `mmse` - a dementia screening test score (0 for a minimum gain and 30 for a maximum gain; it is generally assumed that any score below 24 is an indicator of dementia);

  e) `apoe4` - indicator, whether there is an APOE gene (a gene known as a dementia predisposition) present in the patient's gene pool (0 – no; 1 – yes);

  f) `TIV` - the overall patient's brain volume;

  g) `eTIV` - adjusted overall patient's brain volume — brain volume estimated by a method different from that used for `TIV`;

  h) `hippo` - hippocampus volume.

The general theme of this homework is the exploration of the dependence of the hippocampus volume (variable `hippo`) on the remaining covariates. Certainly, the higher the overall volume of the patient's brain, the higher the volume of the hippocampus is expected. In medical studies such as this one, it is always expected that the patient's age and gender do influence the response.

✍ Homework 2 Assignments

Part 1:
Describe the data—present tables of descriptive statistics and suitable figures, and discuss the particularities of the dataset. Comment especially with respect to the hippocampus volume as a function of the remaining variables.

Part 2:
Analyze the given covariates and comment on possible dangers of multicollinearity.

Part 3:
Find a reasonable model for the dependence of the volume of hippocampus and the overall brain volume (*TIV* and/or *eTIV*). Do not include other covariates yet. But, take into consideration possible transformations of both the regressor(s) and the response (Box-Cox). Provide a $95\ \%$ confidence interval for parameter λ of the Box-Cox transformation. Denote the model you prefer by `m1`, comment on why do you choose this starting model, and if possible, support your claims by suitable numerical characteristics. Comment on the validity of the assumptions of a normal linear model.

Part 4:
Include additional covariates in model `m1`, without interactions so far. Consider also sensible transformations of the covariates. If possible, remove from this model covariates that do not appear to be important with respect to the hippocampus volume. Denote this model by `m2`. Categorical covariates should be included into the model using a parametrization which makes most sense. Report the estimated parameters with the corresponding standard error terms and p-values, and interpret the point estimates.

Part 5:
Consider also pairwise interactions. Report the significance of each interaction term you consider by using a proper Anova table. Comment the table and for each test provide (i) degrees of freedom, (ii) the value of the test statistic, and (iii) the p-value. The summary should include, among others, whether gender and/or age are significant modifiers of any effect of the remaining covariates. Remove from the model interactions that are not significant and denote the final model by `m3`.

Part 6:
In model `m3` explain in detail the effect of age and gender on the response. Provide the confidence bands for (around) for the regression line given age and gender. Use appropriate values for the remaining covariates in the model.

Part 7:
Based on `m3` perform all pairwise comparisons among the four groups specified by the variable `diagnosis`. Interpret the observed differences in words and decide about the statistical significance of the differences. Do not forget to adjust for multiple testing. Provide appropriate confidence intervals for the differences.

Part 8:
Draw basic diagnostic plots for model `m3` and comment on validity of the assumptions of a normal linear model. Provide formal tests (one per assumption) to evaluate the homoscedasticity issue and the normality assumption of the error terms. Discuss possible difficulties.

Part 9:
Consider the contrast sum parametrization for all categorical covariates in `m3`. Discuss the differences between this model and `m3`. Report, and interpret in words, the estimated parameters with the corresponding standard error terms and p-values. Describe the effect of age and gender on the response.