NMSA407: Linear Regression

Winter Term 2019/2020

General Instructions & Homework Assignment no. 3

(Submission Deadline: December 30, 2019)

General Instructions

- The homework assignment can be carried out in a group of 1 – 3 students (three students per each group are recommended). The groups are not required to be the same as those in you had for the elaboration of the previous 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.
- All statistical tests are performed at $5\,\%$ significance level, confidence intervals should be all with the 95% coverage.
- Deliver your report electronically as a `.pdf` file to one of the e-mail addresses:
  - Groups Tuesday 12:20/ Thursday: `maciak@karlin.mff.cuni.cz`,
  - Group Tuesday 14:00: `nagy@karlin.mff.cuni.cz`.

  On the title page specify ONE e-mail address of a person (one group member) who is to be contacted regarding evaluation of the homework. **Deadline** for the delivery of the report is

  December 30, 2019 (23:59).

- **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.

- **R script:**

  As part this homework, please provide also a working `R` script of the analysis 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 report in full.

☞ Data Description

Consider a data set consisting of 108 neurodegenerative dementia patients from Mayo Clinic in Rochester, US. There are three types of dementia patients considered 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 the additional information provided in the data.

- ❏ The datafile (an `RData` file) can be downloaded by clicking on the link hw3_2019.RData or by downloading the corresponding file from the central webpage of the NMSA 407 exercise classes.
- ❏ 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 3 Assignments

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

Part 2:
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`. 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. 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 `diagnosis` on the response.

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 `diagnosis` on the response.