NMSA407: Linear Regression

General Instructions & Homework Assignment no.3

Deadline: December, 28 (23:59) or January, 3

General Instructions

- Similarly as both previous homework assignments, this homework assignment can be again carried out in a group of 1 – 3 students (three students per each group is recommended). The groups are not required to be the same as those in the first two homework assignments.
- Each group is required to submit a computer-prepared 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). Any computer code or originally formatted computer output should appear in the document.
- The document which you intend to submit must contain the names of the members of the group in the header on the first page and it should be fully written either in English or Czech/Slovak (Czech and Slovak are allowed to be mixed inside one document). Please, do not mix English and Czech/Slovak in one document.
- All statistical tests should be performed at $5\ \%$ significance level and confidence intervals should be all with the $95\ \%$ coverage.
- There two deadlines for the homework.
  - If you deliver your report before **December 28, (23:59)** then you get the corrected report at the exercise class on January, 3. The report in this case can be delivered electronically as **PDF file** (one file in one e-mail per group) to your group teacher. On the title page, please, specify ONE e-mail address of a person (one group member) who should be contacted after the evaluation of the homework (see below).
  - You can also deliver **a printed version** of your report at the beginning of any of the three exercise classes on Tuesday, **January 3, 2017**. Then you get the corrected report back during the exercise class in the second week of January.
- 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 & Data Description

For the purposes of the third homework assignment we 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 of the Alzheimers 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 disease. 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 only consider the size, respectively the volume, of the hippocampus part.

The idea is to estimate the expected volume of the patient's hippocampus given some information which is provided in the data.

- ❏ the datafile (*RData*) is available online and it can be downloaded from here: nmsa407_hw3_2016.RData
- ❏ once 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:

  ```
  > load("nmsa407_hw3_2016.RData")
  ```

  Alternatively, you can load the data directly using the following command:

  ```
  load(url("http://www.karlin.mff.cuni.cz/~maciak/NMSA407/nmsa407_hw3_2016.RData"))
  ```

  The R variable storing the dataset is called `data`;
- ❏ The dataset contains 108 independent observations (patients) and 9 different covariates. A detailed description of all considered covariates is given below.
  - **diagnosis** - four level factor distinguishing for four groups of patients: NC - control group; AD - Alzheimer patients; FTLD - Frontotemporal lobar degeneration patients; DLB - Dementia with Lewy bodies;
  - **gender** - two level factor: 0 – female; 1 – male;
  - **age** - patient's age;
  - **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);
  - **apoe4** - indicator, whether there is APOE gene (a gene known as a dementia predisposition) present in the patient's gene pool (0 – no; 1 – yes);
  - **TIV** - the overall patient's brain volume;
  - **eTIV** - adjusted overall patient's brain volume;
  - **hippo** - hippocampus volume;
- ❏ A general theme of this homework is to explore the effect of covariates (factor covariates as well as continuous ones) on the hippocampus volume of a patient.
- ❏ It is somehow logical that the higher the overall volume of the patient's brain the higher the volume of the hippocampus part should be also expected. Therefore, for estimating the expected volume of the hippocampus part we will primarily consider the overall volume of the patient's brain.
- ❏ In addition, we will also consider some other covariates we have available in order to improve the final model in a reasonable way.

Homework 3 Assignments

Part 1:
Create a table of suitable descriptive statistics of variables we are going to analyse. For numerical variables, provide descriptive statistics based on the whole dataset and also categorized by a specific diagnosis (`diagnosis`).

Part 2:

- ❏ For quantitative variables (`age, mmse, TIV, eTIV` and `hippo`), create a matrix of scatterplots and comment it with respect to the proposed modelling of the (expected) hippocampus volume (`hippo`) as a function of the remaining quantitative variables.
- ❏ Create also a plot for the dependence of `hippo` given the patient's age and by suitable approach distinguish four different diagnosis and different gender.
- ❏ Since the relationship between the hippocampus volume and the overall brain volume is of the primary interest, create also a separate scatterplot of `hippo` $\sim$ `TIV` and `hippo` $\sim$ `eTIV` where you also additionally distinguish in a suitable way observations from different diagnosis.
- ❏ For considered quantitative regressors, calculate their pairwise correlation coefficients. Comment on possible danger of multicollinearity. Report the correlation coefficients with some reasonable precision.

Part 3:
As a starting model consider the dependence between the hippocampus volume and the overall brain volume. However, there are two different approaches how the overall brain volume is measured (*TIV* or *eTIV* covariate) therefore, consider separate models for both of them. In addition, consider also a logarithm transformation of *TIV* and *eTIV* and fit analogous models again. Which of the four models do you find the best to describe the relationship between the hippocampus volume and the overall brain volume? Explain your decision and support it with some numerical characteristics. Denote the model you choose as model `m1`. Draw basic residual plots for this model and comment on validity of assumptions of a normal linear model.

Part 4:
Consider model `m1` and use the other covariates (except for *TIV* and *eTIV*) which are available in the data. Do not include any interaction terms yet. Make a simple transformations of the covariates so that the intercept has a meaningful interpretation. Except from gender (that is usually important in such studies) remove from the models all covariates that you do not find significant. Denote this model as `m2`. Report the estimated parameters with the corresponding standard error terms and p-values and interpret the estimated parameters.

Part 5:
For model `m2` consider a Box-Cox class of transformations to (possibly) improve the model by using a transformation of the response. Provide a $95\ \%$ confidence interval for parameter λ of the Box-Cox transformation. In the following, use a transformed response variable obtained by taking a suitable (with respect to reasonable interpretability of the linear model) λ from the $95\ \%$ confidence interval. Report, which lambda you took, and state how the transformed response look like. Denote this model as `m3`.

Part 6:
Consider the model `m3` from Part 5. Add all the pairwise interaction between the variables included in `m3`. In a form of a suitable table, report on a significance of each interaction term you are considering. For each test, provide (i) degrees of freedom, (ii) the corresponding value of the test statistic, and (iii) the p-value. Next to the table summarizing your findings. Your summary should include among others if gender and/or age are significant modifiers of any effects of the remaining covariates. Remove from the model interactions that are not significant and denote the final model as `m4`.

Part 7:
Based on model `m4`.

1. Explain in detail the effect of `age` on hippocampus volume (`hippo`).
2. Explain in detail the effect of `gender` on hippocampus volume (`hippo`).

Part 8:
Based on your model `m4` from Part 6 make all pairwise comparisons among four groups specified by the variable `diagnosis`. Interpret the observed differences in words and decide about statistical significance of the differences. Do not forget to adjust for multiple testing problems. Provide also appropriate confidence intervals for these differences.

Part 9:
Draw basic residual plots for the model from Part 6 and comment on validity of assumptions of a normal linear model. Next to the plots included in `plotLM` function consider also plots of (appropriate) residuals against the covariates. Provide also formal tests (one for each point) to evaluate the homoscedasticity issue and the question on the normality assumption of the random error terms. Briefly state which elements of the statistical inference might be somehow questionable if the inference is based on your model.

Part 10:
Take your mode `m4` and consider the contrast sum parametrizations for all categorical variables included in the model. Denote this model as `m4sum`. For this model report the estimated parameters with the corresponding standard error terms and p-values and interpret the estimated parameters.