NMSA407: Linear Regression

General Instructions & Homework Assignment no. 3

Deadline: December 22, 2017

General Instructions

- 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 document created with LaTeX. All content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis). No computer code or originally formatted computer output should appear in the document.
- The document 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 nominal $95\ \%$ coverage.
- Reports can be delivered either electronically as a **PDF file**, or printed on paper. The deadline for the delivery of the reports is Friday, December 22 (23:59). Reports can be sent to one of the following e-mail addresses:
  - `maciak@karlin.mff.cuni.cz` (groups Tuesday), or
  - `nagy@karlin.mff.cuni.cz` (group Friday).

  Alternatively, printed version of the report can be submitted at the beginning of the exercise classes on Tuesday, December 19, or Friday, December 22.

Data & Data Description

Telemonitoring involves remote monitoring of patients who are not at the same location as the health care provider. A patient has a number of monitoring devices at home, which record measurements regarding the patient's health conditions. The recordings are captured and stored automatically.

We are given a data set of measurements of patients with an early-stage Parkinson's disease, recruited to a telemonitoring trial for remote symptom progression monitoring. The recordings consist of several biomedical measurements of patient's voice. For each such observation we are also provided with the patient's gender and age, and two unified Parkinson's disease rating scale (UPDRS) scores that are used to measure the progress of the Parkinson's disease.

Our primary interest is to infer whether the UPDRS scores can be predicted from the voice recordings. In particular, we are interested in the relation between the expected UPDRS scores and the recorded noise-to-harmonics ratio of patient's voices (`NHR`), which is conjectured to impact the UPDRS scores.

- The datafile (*RData* file) is available online and it can be downloaded here: hw3_2017.RData
- Once you download the data into your working directory (check/set your working directory in R using commands `getwd()` and `setwd()`), you can load the data file into the R environment using the following command:

  ```
  > load("hw3_2017.RData")
  ```

  The R variable storing the dataset is called `data`.
- The dataset contains 5875 observations and 8 covariates:
  - **age** - patient's age;
  - **sex** - two level factor: 0 – male; 1 – female;
  - **motor_UPDRS** - patient's motor UPDRS score;
  - **total_UPDRS** - patient's total UPDRS score;
  - **Shimmer** - a measure of variation in the amplitude of the patient's voice;
  - **NHR** - noise-to-harmonics ratio of patient's voice;
  - **fHNR** - three-level factor covariate that corresponds to a characteristic of standardized harmonics-to-noise ratio (1 – low, 2 – medium, 3 – high);
  - **fDFA** - four-level factor covariate that corresponds to the signal fractal scaling exponent of patient's voice (1 for low – 4 for high).
- General theme of this homework is to explore the effect of `NHR` on the UPDRS scores.

Homework 3 Assignments

Part 1:
Create a table of suitable descriptive statistics of variables we are going to analyse.

Part 2:

- ❏ For quantitative variables (`age`, `motor_UPDRS`, `total_UPDRS`, `Shimmer`, `NHR`), create a matrix of scatterplots and comment on it with respect to the proposed modelling of the (expected) UPDRS scores as functions of the remaining quantitative variables.
- ❏ Explore the plots of the dependence of the UPDRS scores given other quantitative covariates, and by suitable approach distinguish different levels of some of the factor covariates. Report your most interesting findings, and comment on the provided plot(s).
- ❏ Calculate the pairwise correlation coefficients of the considered quantitative variables. Comment on possible danger of multicollinearity. Report the correlation coefficients.

Part 3:
As a starting model consider the dependence between the UPDRS scores, and the `NHR` covariate. There are two different UPDRS scores which can be modelled. Consider separate models for these two. In addition, consider also a log-transformation of `NHR`, and fit analogous models again. Which of the four models describes the best the relationship between the UPDRS scores and `NHR`? 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 include the remaining covariates (except for `motor_UPDRS` and `total_UPDRS`) available in the data. Do not include any interaction terms yet. Transform the covariates so that the intercept has a meaningful interpretation. Except from `NHR` 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 transforming the response. Provide a $95\ \%$ confidence interval for parameter λ of the Box-Cox transformation. Does the square-root transformation of the UPDRS score in model `m2` improve the model, or its interpretability? Justify your decision. Denote by `m3` the better of the two models — either `m2`, or the model with the transformed UPDRS score, according to your decision.

Part 6:
Appropriately address the problem of multicollinearity in model `m3`, especially with respect to covariate `NHR`. Does exclusion of some additional covariate(s) appear to improve the model? In which way? Support your decision with numerical characteristics and an appropriate plot. Denote the (possibly) new model by `m4`.

Part 7:
In model `m4`, add all the second-order interactions between the variable that corresponds to `NHR`, and the other variables included in the model. In a 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. Remove from the model interactions that are not significant, and denote this final model by `m5`.

Part 8:
Based on model `m5` explain in detail the effect of `NHR` on the UPDRS score.

Part 9:
Based on model `m5` make all pairwise comparisons among the groups that correspond to the categorical covariates (without the interaction terms) included in the model. Interpret the observed differences in words, and decide about their statistical significance. Do not forget to adjust for multiple testing problems. Provide appropriate confidence intervals for these differences.

Part 10:
Draw basic residual plots for model `m5` and comment on validity of assumptions of a normal linear model. Next to the plots included in function `plotLM` consider also plots of (appropriate) residuals against the covariates. Provide formal tests (one for each point) to evaluate the homoscedasticity, and the normality assumption of the random error terms. Briefly state which elements of the statistical inference might be questionable in model `m5`.

Part 11:
In model `m5` consider the contrast sum parametrizations for all categorical variables included in the model. Denote this model as `m5sum`. Report the estimated parameters with the corresponding standard error terms and p-values, and interpret the estimated parameters.