NMSA407: Linear Regression

Winter Term 2020/2021

General Instructions & Homework Assignment no. 2

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 (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 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 of the solution, please provide also a working and well commented `R` script of the whole 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 mian `pdf` file.

❑ **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, 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

A telemonitoring involves a remote tracking of certain patients—usually those who can not be present at the same location as their health care provider. Each patient has a number of monitoring devices which record measurements regarding the patient's health conditions. The recordings are captured, stored, and transmitted automatically.

In this homework assignment, we are interested in some subset of a telemetry data on patients with an early-stage Parkinson's disease. The patients were recruited to a trial for remote symptom progression monitoring. Each patient's record consists of several biomedical measurements of the patient's voice and the voice analysis is used to determine the progress of the disease—measured by two unified Parkinson's disease rating scores (UPDRS). Moreover, for each observation there are some additional patient's specific data provided, such as the patient's gender and age.

Our primary interest is to infer whether the UPDRS scores can be (somehow) predicted from the voice recordings and the patient's specific data. In particular, we are primarily interested in the relationship between the expected ratio of the two UPDRS scores and the recorded noise-to-harmonics ratio of the patient's voice (NHR).

❑ The datafile (*RData* file) is available online and it can be downloaded here: NMSA407-2021-HW2.RData

❑ The dataset contains 850 observations and 7 covariates:

   ❑ `age` - patient's age;

   ❑ `sex` - two level factor variable: 0 – male; 1 – female;

   ❑ `motor_UPDRS` - patient's motor UPDRS score;

   ❑ `total_UPDRS` - patient's total UPDRS score;

   ❑ `Shimmer` - variation measure for the amplitude of the patient's voice;

   ❑ `NHR` - noise-to-harmonics ratio of the patient's voice;

   ❑ `fDFA` - four-level factor variable for the signal fractal scaling exponent;
   (1 for low scaling – 4 for high scaling)

❑ A general theme of this homework is to explore the effect of `NHR` on the proportion of the motor UPDRS within the total UPDRS (**which we model as a ratio of `motor_UPDRS` and `total_UPDRS`**).

❑ **For each part 1 – 3 of this homework assignment do all of the following:**

(a) provide at least one table (e.g., with descriptive statistics) and at least one plot being useful in the context of the problem and comment them within the framework of the given problem;

(b) always define a probabilistic model that you are using (e.g., normal linear regression model, hypothesis tests) and properly discuss the theoretical assumptions behind this model;

(c) for each statistical test provide an 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;

(d) formulate your conclusions and interpret the results (the interpretation must be understandable for non-statisticians as well and it must be specific to the problem);

✍ Homework 1 Assignments

Part 1: Model building
Start with a simple linear regression model where the expected ratio between the two UPDRS scores is modeled with respect to the noise-to-harmonics ratio (NHR).

(a) Consider an optional logarithmic transformation for both, the response variable and, also, the noise-to-harmonic ratio (NHR). **Choose one model only** and clearly explain your choice. Briefly discuss the model qualities in terms of the imposed assumptions.

(b) Use the additional covariates provided in the data and extend the model from the previous step (using two-way interactions at most) such that the output (summary or some anova table) of the **model can be directly used to answer the following questions** imposed by medical doctors:

- ❑ Does the effect of the noise-to-harmonics ratio (`NHR`) depend on the signal scaling component (`fDFA`)?
- ❑ Are the effects of the patient's age and the amplitude variation (`Shimmer`) generally different for male and female patients?
- ❑ Is the signal scaling component (`fDFA`) a significant modifier of the effect of the patient's age?

For each question above specify the corresponding statistical test based on the underlying model (including all necessary details). Interpret the test results in terms of the given questions (such that non-statisticians, i.e., medical doctors will understand).

(c) Finally, modify the underlying model in a way that **only significant effects** will be included and all variables will have a **reasonable interpretation**. Using the final model, interpret the estimated parameters (resp. the groups of similar parameters) and discuss the model limitations—its pros and cons.

Part 2: Model assumptions
Consider the final model from Part 1(c) and **discuss the model in terms of the imposed assumptions**. In particular, focus on the following:

(a) Address homoscedasticity/heteroscedasticity issues in the model. Provide a formal statistical test to assess the homoscedasticity assumption and interpret the results. Discuss possible difficulties.

(b) Verify the normality property and discuss the effects of possibly outlying observations.

(c) Appropriately address the problem of multicollinearity in the model, especially with respect to the noise-to-harmonics ratio (`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.

Part 3: Model inference
Use the model from Part 1(c) to properly address the following problems/questions.

(a) Provide all pairwise comparisons between male and female patients for all groups being defined by the categorical variables in the model. For possible interactions with the continuous covariates consider one or more specific values which give a reasonable interpretation. Interpret the observed differences and decide about their statistical significance. Provide appropriate confidence intervals for these differences.

(b) Use the underlying model which you possibly modified by adding new parameters to answer the following:

❏ Is the effect of the noise-to-harmonics ratio (NHR) different for male and female patients given high fractal scaling exponent (`fDFA` = 4)? What can be said, in general, regarding the patient's age playing the role of a modifier of the effect of NHR?

❏ Can we say, that the effect of NHR is the same for all patients with he scaling exponent `fDFA = 2` and `fDFA = 3`?

❏ Given the model, what is the expected ratio of the UPDRS scores for a 65 years old male patient with a relatively standard noise-to-harmonic ratio (thus, `NHR = 0.02`) and a rather high scaling factor (`fDFA = 4`)? Use reasonable values for the the variables which are left unspecified. What is the corresponding prediction interval?

Always specify the model and the statistical test you are using. Provide a suitable quantity with the corresponding confidence interval. Interpret each test in a way that non-statisticians can also understand.

(c) Use the contrast sum parametrization for all categorical covariates in the model and interpret the corresponding parameter estimates. Draw a plot to visualize the dependence of the response on the noise-to-harmonic ratio for different scaling factors—at least those that you proclaimed to be statistically different. Provide also the confidence band for the regression function(s) and the corresponding prediction band(s).