NMSA407: Linear Regression

Winter Term 2018/2019

General Instructions & Homework Assignment no. 2

(Submission Deadline: Exercise class no. 9)

ℹ General Instructions

- The homework assignment can be carried out in a group of 1 – 3 students (three students per each group is recommended). Different groups than those you used for the first homework assignment can be formed to work on this second assignment.
- Each group is required to submit a well-written `.PDF` document created with LaTeX. Its content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis). This includes a brief introduction with the some problem description, sections which address the homework task and, at the end, some brief conclusion.

  Computer code or originally formatted computer output should not appear in the document. Figures and tables should always contain labels with sufficient description. Moreover, tables must be carefully designed and formated (same number of decimal digits within each column, centered columns aligned to the decimal point, reasonable and intuitive visual representation, etc.).
- The document must contain the names of the members of the group in the header on the first page. It should be written either in English or Czech/Slovak (Czech and Slovak are allowed to be mixed within one document). Please, do not mix English and Czech/Slovak in one document.
- All statistical tests should be performed at the $5\ \%$ significance level, confidence intervals should be all constructed with the $95\ \%$ coverage probability.
- **DEADLINES and FORM OF DELIVERY:**

Monday's session	(*Matúš Maciak*)	Due on 26/11 (09:00)	PRINTED ON PAPER
Tuesday's session	(*Matúš Maciak*)	Due on 27/11 (12:20)	PRINTED ON PAPER
Thursday's session	(*Stanislav Nagy*)	Due on 29/11 (10:40)	PRINTED ON PAPER

- For groups 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, provide the exercise group identification to which you want to submit your homework solution.

ℹ Data Description

The data comes from a small experiment conducted in a few horse ranches in the Czech Republic in 2014. The purpose of the experiment was to measure a horse specific heart rate when the horse is exposed to various disturbing stimuli, while controlling for some additional horse specific covariates (e.g., age, gender, type). Different disturbing stimuli (see data description for more details) were introduced to each horse and the corresponding heart rate response was measured.

- ❏ The datafile (an `.RData` file) can be downloaded from NMSA407-1819-HW2.RData, or from the central webpage of the exercise classes.
- ❏ Once you download the data into your working directory (check/set your working directory in `R` using commands `getwd()` and `setwd()`), you can load them into `R` using

  ```
  > load("NMSA407-1819-HW2.RData")
  ```

  The `R` variable with the dataset is called `horsesData`.
- ❏ The dataset contains 266 observations and 7 covariates. A detailed description of all the covariates is given below.

  a) `stimulus` - a type of a disturbing stimulus presented to a horse before measuring its heart rate. It is a categorical variable with 7 different levels assumed to be ordered with respect to an increasing disturbance effect (louder, more stressful):
     1 - quiet unexpected noise;
     2 - sudden music playing;
     3 - crushing a plastic bottle;
     4 - crushing an aluminium can;
     5 - spraying a spray all of a sudden;
     6 - throwing a ball towards a horse;
     7 - opening an umbrella in front of the horse;

  b) `HR` - heart rate measured after the stimulus was applied;

  c) `age` - horse age given in years;

  d) `gender` - categorical variable: 1 - for a mare; 2 - for a gelding; 3 - for a stallion;

  e) `type` - categorical variables to distinguish for the horse specific type: 1 - cold-blooded; 2 - warm-blooded; 3 - pony;

  e) `outside` - categorical variables expressing where the horse is usually kept in (it can be also interpreted as a measure on how much time a horse spends in its natural environment - countryside): 1 - for 3 hours daily at most; 2 - for 4 to 9 hours a day; 3 - for more than 10 hours a day;

  f) `utilization` - categorical variables to express the main purpose of the horse: 1 - a racing horse; 2 - a horse meant for training; 3 - only recreational purposes;

The general theme is an exploration of the dependence of the horse heart rate given the remaining variables.

☞ Homework 2 Assignments

Part 1:

Create a table of suitable descriptive statistics of all variables in the dataset.

Part 2:

Create appropriate plots of the heart rate against the remaining variables. Provide a few interesting plots in your report, and comment on their relevance with respect to the proposed modelling of the heart rate given the covariates.

Fit a linear model `m1` with the heart rate as a response and all other covariates as explanatory variables. Do not include any interaction terms. Create a nicely formatted table which summarizes the most important results. Such table should contain (at least):

- ❏ estimates of regression coefficients and their standard errors;
- ❏ corresponding $95\ \%$ confidence intervals;
- ❏ p-values for tests on regression coefficients in those situations **where it makes a good practical sense** to perform such tests;
- ❏ estimated residual standard deviation;
- ❏ coefficient of determination.

Part 3:

In words, interpret each regression coefficient (or a group of coefficients if they all describe a similar quantity). Also a non-statistician should then understand the meaning of the model. Discuss how suitable is the proposed model for prediction of the heart rate based on the considered predictors.

Part 4:

Include some basic residual plots for model `m1`. Based on these plots, comment on the validity of the assumptions of a classical normal linear model. Do not perform any formal statistical tests.

Part 5:

You are interested whether a gelding is in general more calm than a mare.

1. Provide an estimate (based on `m1`), including the standard error and a $95\ \%$ confidence interval, of the effect (on the expected heart rate of the horse) of the horse being a gelding when compared to a mare. In your report, explain the effect in words and describe which approach (brief reference to lecture, ...) you used to obtain the final numbers.

2. Run a formal statistical test of an appropriate hypothesis.
   As always, specify (mathematically) the statistical hypothesis, provide the formula for, and the value of the test statistic, p-value and your conclusion expressed in words understandable by a non-statistician.

3. Visualize the effect of gender using a scatterplot of the heart rate versus age based on a subsets of horses where you distinguish (by different symbols, colours, ...) different options for the `gender` covariate. Add to the plot fitted regression lines showing the model-based estimated dependence of (mean) heart rate on age for considered options of `gender`.
   Comment on the values of the remaining covariates that were used in the plot.

4. Is it possible to say that, in general, male horses (geldings and stallions) are calmer than female horses (mares)? Provide a point estimate, and a $95\ \%$ confidence interval for the effect of gender categorized as male/female on the expected heart rate. If necessary, modify model `m1` appropriately (in the new model it is not needed to report the estimates etc.). Add the fitted regression line for male horses into the plot from part 3., and comment on possible differences from the analysis above (the mare/gelding/stallion situation).

Part 6:

Assume you have a cold-blooded gelding which is mostly stalled inside, his age is 10 years and he is used only for recreational purposes. You are interested in its actual heart rate when an umbrella is opened next to the horse. Find an estimate (based on `m1`) for the actual horse's heart rate. Provide a point estimate including the $95\ \%$ confidence interval for the heart rate.

Return to the residual plots (Assignment 4), and discuss how much reliable is this confidence interval.

Part 7:

Estimate the expected difference in the heart rate between two horses: a young horse at the age of 4 years being kept mostly inside and an old horse at the age of 15 being kept mostly outside when both are exposed to the same stimulus.

Provide an estimate (based on `m1`), including the $95\ \%$ confidence interval. By a suitable statistical test evaluate whether the first horse is expected to have a higher heart rate result than the second horse.

Part 8:

Extend model `m1` so that the variable `outside` possibly modifies the effect of the age on the heart rate. Denote this model as `m2` and suppose that this is a useful model.

1. Specify model `m2` in your report. It is not necessary to include the estimates etc.
2. In detail describe the effect of the age on the heart rate as estimated from `m2`.
3. Test whether the variable `outside` modifies the effect of the age on the heart rate.

Part 9:

Decide whether model `m2` can be simplified (in terms of the number of parameters) by considering the variable `stimulus` as a (suitably parametrized) numeric variable. If yes, describe the effect of the variable `stimulus` (taken as numeric) on the heart rate.