NMSA407: Linear Regression

Winter Term 2021/2022

General Instructions & Homework Assignment no. 1

Submission Deadline: Ex. Session no. 7 (mid-November 2021)

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. 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 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 `pdf` file of the report in full.
- ❏ **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 **November 18, 2021 (23:59 CET)**.
- ❏ **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 95% coverage. With each formal test performed, it is necessary to specify (mathematically) the statistical hypothesis, provide the formula for, and the value of the test statistic, specify the distribution of the test statistic and the p-value, and your conclusion expressed in words understandable to a non-statistician.

ℹ Data Description

The data comes from an 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 additional covariates (e.g. age, gender, type, etc.).

- ❏ The `.RData` file containing the data can be downloaded by clicking **here**, or by downloading the corresponding file from the **central webpage** of the NMSA407 exercise classes. In `R`, the data is loaded into the working environment using the command

  ```
  > (load("NMSA407-2122-HW1.RData"))
  ```

  The `R` variable with the dataset is called `horsesData`.

- ❏ The dataset contains 266 observations and 7 covariates:

  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 (that is louder, or more stressful) disturbance effect:

  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.

We want to model the dependence of the horse heart rate given the remaining variables.

☞ Homework 1 Assignments

Part 1:

Describe the data. Present tables of descriptive statistics of the variables in the dataset. 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.

Part 2:

Can we say, ignoring for a moment all the covariates except `HR` and `age`, that the heart rate of a horse is independent of its age? Answer the following questions:

(a) Formulate rigorously, and provide a formal test of the previous hypothesis. Describe the test statistic, and provide a suitable confidence interval. For this analysis, do not make use of a linear model.

(b) Fit a simple linear model `m0a` based on only `HR` and `age`, and answer the previous question by analysing `m0a`.

(c) Describe the relation of the two approaches considered in (a) and (b), and explain in detail the differences between the two performed analyses. Which of the two previous approaches (a) and (b) do you prefer, and why?

Part 3:

In this part, we ignore all the covariates except `HR` and `type`. Can we say that the heart rate of a horse is independent of its type? Answer the question in two different ways, analogously as in **Part 2**: (i) Without using a linear model, and (ii) using a suitable linear model called `m0b`. Comment on the similarities and differences of the two procedures, and explain your preference.

Part 4:

Fit a linear model `m1` with the heart rate as a response and all the other covariates as explanatory variables. Do not include any interaction terms. Create a nicely formatted table which summarizes the most important results. Such a 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;
- ❏ the coefficient of determination.

In words, interpret all the values in the table above. In particular, interpret each regression coefficient (or a group of coefficients if they all describe a similar quantity). Also a non-statistician must understand the meaning of the model.

Part 5:

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. Discuss on how much suitable is model `m1` for the prediction of the heart rate based on the considered regressors.

Part 6:

You are interested whether a stallion 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 stallion 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.

3. Visualize the effect of gender using a scatterplot of the heart rate versus age based on 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 (i.e. geldings and stallions together) 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/female horses into the plot from part 3., and comment on the possible differences from the analysis above (the mare/gelding/stallion situation).

Part 7:

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.

Discuss on how much reliable is this confidence interval.

Part 8:

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 9:

Starting from model `m1`, answer each of the two questions raised in **Parts 2** and **3**. This time, we do not ignore the contribution of the other covariates in the dataset. Do you obtain equivalent results as in the analysis based on models `m0a` and `m0b`, respectively? Explain the possible differences in the performed analyses in words. Which of the approaches considered do you prefer, and why?