NMSA407: Linear Regression

General Instructions & Homework Assignment no.2

November 24, 2016

General Instructions

❏ The homework assignment can be 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 you had for the elaboration of the first homework assignment.

❏ Each group is required to submit a computer-prepared PDF document created with some appropriate software (e.g. LaTeX, OpenOffice, MS Word, or others). All content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis) and all questions stated in the assignment need to be carefully addressed.

❏ A computer code or originally formatted computer output should appear in the document. Also, provide only the test which is relevant for the question of interest.

❏ The submitted document must contain all the names of the group members (with the indicated exercise class) – please, provide these names on the title page (if there is one) or in the header in the first page (if there is no title page).

❏ The document must be fully written either in English or Czech/Slovak (Czech and Slovak are allowed to be mixed inside one document, however, do not mix English and Czech/Slovak in one document).

❏ All statistical tests should be performed at the $5\ \%$ significance level and confidence intervals should be given all with the 95% coverage.

❏ **DEADLINES and FORM OF DELIVERY:**
Group Matúš Maciak (Monday): 28/11 (15:40) PRINTED ON PAPER
Group Matúš Maciak (Tuesday): 29/11 (09:00) PRINTED ON PAPER
Group Marek Omelka (Tuesday): 29/11 (14:00) PRINTED ON PAPER

❏ For groups which are composed of students from different exercise classes, only one document is required to be delivered to an arbitrarily chosen lecturer within the deadline that applies for the class of this specific lecturer.

Data Description

For the second homework assignment we will consider the following dataset: 149 urological patients have undertaken a surgery in order to remove kidney stones from their liver. The surgeries took all place in a university hospital in Bánska Bystrica in 2014–2016. From the technological point of view, the surgery can be either invasive (holmium based treatment using a flexible YAG Laser) or noninvasive (an ultrasonic laser PEK). Each patient undertook exactly one surgery while the surgery type was decided for each patient at random. The information for the surgery type is given for each patient in the dataset. In addition, for each patient there is also some additional patient's specific information recorded in the data (e.g. gender, age, surgery time, surgeon who performed the surgery, etc.).

- ❏ the datafile (*RData* file) is available online and it can be downloaded here: hw2_2016.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("hw2_2016.RData")
  ```

  The R variable storing the dataset with the data is called `data`;
- ❏ The dataset contains 149 observations and 8 covariates.

  a) `gender` - patient's gender (`male` or `female`);

  b) `flexPek` - two level factor covariate to distinguish for the noninvasive surgery (pek) or an invasive surgery (flex);

  c) `surgeon` - four value covariate to identify the surgeon who performed the surgery;

  d) `size` - numerical covariate which stands for the overall `size` of the kidney stone(s) given in a diameter in [mm];

  e) `SFR` - indicator covariate to express the surgery result: Stone Free Rate (SFR) equals to one if there were no kidney stones remaining and it is equal to zero otherwise;

  e) `time` - the overall time the surgery took place given in [min];

  f) `intervention` - integer covariate which stands for the number of required interventions during the surgery – if the surgery goes well no interventions are expected;

  g) `age` - patient's age given in years.

A general theme of this homework will be exploration of a model for dependence of the surgery time (variable `time`) on size of the kidney stone (variable `size`) when taken additional covariates (`age`, `flexPek`, `gender`, `SFR`, `surgeon`) into consideration.

Homework 2 Assignments

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

Part 2:
For considered quantitative variables (`time`, `age` and `size`), create a scatterplots and comment it with respect to the proposed modelling of the surgery time.

Fit a linear model (further referred to as model `m1`) with the surgery time as a response and other variables 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;
- ❏ $95\,\%$ confidence intervals;
- ❏ p-values for tests on regression coefficients in those situations where it makes a practical sense to perform such test;
- ❏ 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 be able to understand the meaning of the model. Discuss, whether the model is suitable for predicting the surgery time based on the considered predictors (age, gender, size, etc.).

Part 4:
Include three basic residual plots for model `m1` (result of plotLM function from package mffSM). Based on those plots, comment what you think about validity of assumptions of a classical normal linear model. Do not perform any formal statistical tests.

Part 5:
Suppose that you and your friend disagree on the opinion whether the surgeon no.3 needs more time to perform the surgery than surgeon no. 4.

1. Provide an estimate (based on the considered model), including the standard error and a $95\,\%$ confidence interval, for the difference between the time needed by these two surgeons. In your report, explain the effect and describe which approach (brief reference to lecture, ...) you are using to arrive at final numbers.

2. By a suitable statistical test evaluate whether it makes sense to argue with your partner about the time needed by these two surgeons to perform the surgery. As always, specify (mathematically) the statistical hypothesis, provide the value of the test statistic, p-value (and how it is computed) and your conclusion expressed in words understandable by a non-statistician.

3. Visualize the difference between time needed by surgeons no.3 and no.4 (covariate `surgeon`) using a scatterplot of the surgery time (variable `time`) versus the size of the kidney stone (variable `size`) based on a subsets of data where you distinguish by different options (symbols, colors, etc.) for the `surgeon` covariate. Add to the plot the fitted regression lines showing the model-based estimated dependence of (mean) time on `size` for considered options of the `surgeon` covariate.

Part 6:
Assume that surgeon no.2 is going to perform an invasive type type of the surgery (flex) on a female patient with the overall size of her kidney stones being equal to 22 mm. We are interested in the estimated surgery time needed to complete the surgery. What would be the estimate for the time? Provide an estimate including the 95 % confidence interval for this time.

Try to find a reasonable solution although only values of some of the variables are specified.

Have a look at the appropriate diagnostic plots and discuss how much trustworthy is the confidence interval that you have just calculated. Do you think that there might be some problem here?

Part 7:
Estimate the expected difference in the time needed to complete the surgery between two surgery methods when both applied by the same surgeon on the same patient however, the kidney stones sizes are 10 and 20 millimeters respectively. Provide a confidence interval for this expected difference.

Part 8:
Modify the model `m1` by considering the logarithmic transformation of `size` (instead of the size) and denote this model as `m2`. Compare models `m1` and `m2` in terms of the interpretation of the regression coefficients corresponding to `size` and the logarithm of `size`. Which of the models would you prefer and explain why?

Part 9:
Extend the previous model (either `m1` or `m2`, depending on your preference in Part 9) such that the variable `surgeon` possibly modifies the effect of `size` on `time`. Denote this model as `m3` and suppose that this is a useful model.

1. Provide a formal model specification (model formula) `m3` in your report. It is not necessary to include the estimates in the report.
2. In detail describe the effect of `size` on `time` as estimated by model `m3`.
3. Make a formal test that the variable `surgeon` modifies the effect of `size` on `time`.