NMSA407: Linear Regression

Winter Term 2017/2018

General Instructions & Homework Assignment no.2

(Submission Deadline: November 21st, 2017)

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 you formed for the elaboration of the first homework assignment.

- ❑ Each group is required to submit a computer-prepared PDF document created with LaTeX. 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.

- ❑ No computer code or originally formatted computer output should appear in the document. Also, provide only tests which are 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 $5\ \%$ significance level and confidence intervals should be given all with 95% coverage.

- ❑ **DEADLINES and FORM OF DELIVERY:**
  All Groups: 21/11 (12:10 or 13:50 in K4) PRINTED ON PAPER

Data Description

Bike sharing systems are new generation of traditional bike rentals where the whole process of membership, rental and return back has become automatic. Through these systems, a user is able to easily rent a bike from a particular position in the city, and later return it back at any other bike rental location.

Consider a dataset of daily counts of rented bikes in Washington, D.C., USA, in years 2011 and 2012, with the corresponding weather and seasonal information that may affect the number of rented bikes. For each day in the years 2011 and 2012 (731 days in total), the dataset collects the number of bikes rented that day by unregistered users (e.g. tourists, or other visitors of the city) and by registered users (residents). With these two counts, a number of characteristics that may affect the number of rented bikes were recorded. In our analysis, we are interested in the relation of the proportion of bikes rented by tourists to the total number of rented bikes, on the covariates affecting the overall bike rental counts.

❑ the datafile (*RData* file) is available online and it can be downloaded here: hw2_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("hw2_2017.RData")
```

The R variable storing the dataset with the data is called `data`;

❑ The dataset contains 731 observations and 7 covariates. A detailed description of all covariates in the data is below:

a) `dteday` - date;

b) `season` - four level factor covariate that corresponds to the season of the year (1 - winter, 2 - spring, 3 - summer, 4 - autumn);

c) `holiday` - indicator covariate that represents the indicator of whether the day is a public holiday (0 - not a holiday, 1 - holiday);

d) `temp` - numerical covariate that represents the daily mean temperature [°C];

e) `hum` - numerical covariate that represents the daily mean humidity [%];

f) `weekday` - seven level factor covariate the describes the day of the week (1 - Monday, 2 - Tuesday, . . . , 7 - Sunday);

g) `ratio` - numerical covariate that represents the proportion of unregistered users (tourists) to the total number of daily users [%].

The general theme of this homework is the modelling of the dependence of the proportion of tourist rentals to the total number of daily users (variable `ratio`) on the other covariates (`season`, `holiday`, `temp`, `hum`, `weekday`).

Homework 2 Assignments

Part 1:
Create a table of suitable descriptive statistics of all variables we are going to analyze and briefly discuss (interpret) the values in the table within the context of the problem.

Part 2:
For considered quantitative variables (`temp`, `hum` and `ratio`), create a scatterplot and comment on it with respect to the proposed modelling of the ratio of tourists.

Fit a linear model (further referred to as model `m1`) with `ratio` 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 the 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 proportion of tourists based on the considered predictors (temperature, humidity, day of the week, etc.).

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

Part 5:
Consider a week in the winter season with no holidays. At this week the average temperature is about 5°C, and the average humidity is 60%. One may think that in such a week, the average proportion of unregistered users on a weekend day (Saturday and Sunday) is five times the average proportion of unregistered users on a working day (Monday – Friday).

1. Provide an estimate (based on the considered model), including the standard error and a 95 % confidence interval, for the difference between the average ratio of tourist users on a weekend day, and the average ratio of tourist users on a working day (both considered in the winter season, with no holidays, and the given temperature and humidity 5°C and 60 %, respectively). In your report, explain the effect and describe which approach (brief reference to lecture, ...) you are using to arrive at the final numbers.
2. By a suitable statistical test evaluate whether it makes sense to argue about the considered relation of the tourist ratios. 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. Consider a new factor covariate `weekday2` with two levels: 1 - working day (Monday – Friday), 2 - weekend (Saturday and Sunday). Visualize the difference between the ratio of tourist users on working days and on weekend days using a scatterplot of the tourist ratio (variable `ratio`) versus the daily temperature (variable `temp`) based on subsets of data where you distinguish by different options (symbols, colors, etc.) for the `weekday2` covariate. Add to the plot the fitted regression lines showing the model-based estimated dependence of the (mean) tourist ratio on `temp` for the values of the (original) covariate `weekday` corresponding to Friday, Saturday, and Sunday.

Part 6:
We would like to predict the proportion of the tourist users on December 1, 2017. According to the weather prediction, on that day the mean temperature should be about 5°C. What is the estimate for the proportion of tourist users? 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 appropriate diagnostic plots and discuss how much trustworthy is the confidence interval that you have just calculated. Do you see any problem?

Part 7:
Estimate the expected difference in the proportion of tourists between December 1, 2017, and December 2, 2017, given that you assume that the weather conditions on those two days remain the same. Provide a confidence interval for this expected difference.

Part 8:
Modify the model `m1` by considering the logarithmic transformation of `ratio` (instead of `ratio`) and denote this model as `m2`. Compare models `m1` and `m2` in terms of their interpretability, and the validity of assumptions of a classical normal linear model (again, do not perform any statistical tests). Which model do you prefer and why?

Part 9:
Extend the previous model (either `m1` or `m2`, depending on your preference in Part 8) so that the variable `season` possibly modifies the effect of `temp` on `ratio`. Denote this model as `m3` and suppose that this is a useful model.

1. Provide a formal model specification (model formula) of `m3` in your report. It is not necessary to include the estimates in the report.
2. In detail describe the effect of `temp` on `ratio` as estimated by model `m3`.
3. Perform a formal test whether the variable `season` modifies the effect of `temp` on `ratio`.