NMSA407: Linear Regression

Winter Term 2018/2019

General Instructions & Homework Assignment no. 3

(Submission Deadline: December 30, 2018)

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 can be formed to work on each homework assignment (there will be three assignments during the term).
- 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). Computer code or originally formatted computer output should not appear in the document.
- 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.
- Deliver your report electronically as a `.PDF` file to one of the e-mail addresses:
  - Groups Monday/Tuesday: `maciak@karlin.mff.cuni.cz`,
  - Group Thursday: `nagy@karlin.mff.cuni.cz`.

  On the title page specify ONE e-mail address of a person (one group member) who is to be contacted regarding evaluation of the homework (see below). **Deadline** for the delivery of the report is

  December 30, 2018 (23:59).

ℹ Data Description

We focus on evaluating the concentration of phosphorus in certain types of soil after adding one of the two available ash-based fertilizers. We consider four different doses of the fertilizers. Some additional (chemical and physical) soil characteristics measured before the fertilizers were applied are also available, and should be used for the evaluation.

- ❏ The datafile (an `.RData` file) can be downloaded from NMSA407-1819-HW3.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-HW3.RData")
  ```

  The `R` variable with the dataset is called `soilData`.
- ❏ The dataset contains 384 observations and 11 covariates. Detailed description of all covariates is given below.

Covariate	Description
`P`	phosphorus concentration measured in μg per kg after adding an ash-based fertilizer;
`AshType`	binary variable for two different types of ash fertilizers;
`SoilType`	factor variable with four levels for four different soil types;
`Rate`	ash fertilizer dose given as the percentage (0 %, 1 %, 2.5 % and 5 %) of the total soil volume;
`SoilCa`	calcium concentration measured in μg per kg;
`SoilK`	potassium concentration measured in μg per kg;
`SoilP`	initial phosphorus concentration measured in μg per kg;
`SoilMg`	magnesium concentration measured in μg per kg;
`SoilpH`	pH acidity of the soil sample;
`SoilSilt`	percentage of silt in the soil sample;
`SoilClay`	percentage of clay in the soil sample.

We are interested whether the final concentration of phosphorus (variable `P`) is affected by the ash fertilizers (`AshType`), and the soil type (`SoilType`), and how is this relationship modified by the remaining covariates (with the exception of `P` and `SoilP`). We expect that higher initial phosphorus concentration (`SoilP`) before applying an ash fertilizer, results in higher final phosphorus concentration (`P`). Therefore, as a response variable consider the ratio of the final and the initial phosphorus concentration (i.e. `P`/`SoilP`).

☞ Homework 3 Assignments

Part 1:

Create a table of suitable descriptive statistics of the dataset. For numerical variables, provide descriptive statistics of the whole dataset, and also categorized by a specific soil type (`SoilType`).

Part 2:

- ❏ For quantitative variables (where appropriate), create a matrix of scatterplots and comment on it with respect to the proposed modeling of the ratio `P`/`SoilP` as a function of the remaining quantitative variables.
- ❏ Create suitable plots that can help to characterize the relationship between the ratio `P`/`SoilP` and the remaining variables.
- ❏ For quantitative regressors, report their pairwise correlation coefficients. Comment on possible multicollinearity issues.

Part 3:

As a starting model `m1`, consider a linear model where you include all regressors (except from `P` and `SoilP`) together with all two-way interactions between `SoilType` and the remaining covariates. Report the coefficient of determination of this model. Draw basic residual plots and comment on validity of assumptions of a normal linear model. Comment on possible difficulties with respect to model `m1`; if necessary, modify the model appropriately.

Part 4:

Consider an appropriate transformation of the response and fit the same model as in Part 3, now with the transformed response (denote the transformed response as `tP`). Denote this model as `tm1`. Comment on validity of assumptions of a normal linear model. Which model do you prefer and why?

Part 5:

Report on significance of the interaction terms in your preferred model. For each test, provide (i) degrees of freedom, (ii) the corresponding value of the test statistic, and (iii) the p-value. Interpret the tests' results and summarize your findings.

Part 6:

Denote by `m2` the model which you obtain from your preferred model from Part 4 by removing all non-significant two-way interactions. Provide a proper statistical test whether `m2` is significantly better. Rigorously specify the null and alternative hypothesis, give the value of the test statistic, its distribution under the null hypothesis, the p-value, and your final conclusion.

Part 7:

Test whether the model you chose in Part 6 can be reduced to model `m3` obtained from your model by removing appropriate non-significant main effects. Rigorously specify the null and alternative hypothesis, give the value of the test statistic, its distribution under the null hypothesis, the p-value, and your final conclusion.

Part 8:

Take the most suitable model from the previous analysis. If necessary, transform the covariates to obtain an equivalent model in which the regression coefficients have meaningful interpretation. Denote the final model by `mFin`. Create a nicely formatted table which summarizes the most important results derived from model `mFin`.

The table should contain (at least):

1. estimates of regression coefficients, their standard errors, and $95\ \%$ confidence intervals;
2. p-values for the tests on regression coefficients where it makes a good practical sense to perform such test;
3. estimated residual standard deviation;
4. coefficient of determination.

Interpret the estimated regression coefficients (in terms so that even non-statisticians and non-mathematicians will understand what you are talking about).

Part 9:

Based on model `mFin`:

1. In detail describe the effect of the fertilizer (`AshType`) on the ratio `P`/`SoilP`.
2. In detail describe the effect of the soil type (`SoilType`) on the ratio `P`/`SoilP`.

Part 10:

Draw basic residual plots for model `mFin` and comment on validity of the assumptions of a normal linear model. Provide formal tests (one for each point) to infer whether (i) the model is homoscedastic, and whether (ii) it is normal. Comment on possible difficulties.

Part 11:

Parametrize model `mFin` in such a way that one immediately sees the mean effects of the covariates in the model (the mean across the soil type (`SoilType`) and/or type of fertilizer (`AshType`), if soil type and/or type of fertilizer are effect-modifiers for some of the remaining covariates). Provide estimates (including standard errors) and 95 % confidence intervals for these effects. Explain the meaning of the obtained numbers in words understandable to non-statisticians as well.

Part 12:

Provide all pairwise comparisons between groups specified by the significant factor covariate(s) in model `mFin`. Correct for possible multiple comparisons appropriately.