SDA 2023 — Assignment 7

For these exercises the standard *R*-functions `lm`, `hatvalues` and `cooks.distance` can be used. `lm` is needed to fit linear models. The data to be analyzed should be in a `data.frame` format, see the first assignment. `hatvalues` and `cooks.distance` require the output of `lm` as argument. Furthermore, the function `lm.norm.test` and the following functions for collinearity measures are available on Canvas:[1] `varianceinflation`, `conditionindices`, `vardecomposition` and `determinationcoef`.

Make a concise report of *all* your answers in *one single PDF file*, with only *relevant R* code *in an appendix*. It is important to make clear in your answers how you have solved the questions. Graphs should look neat (label the axes, give titles, use correct dimensions etc.). Multiple graphs can be put into one figure using the command `par(mfrow=c(k,l))`, see `help(par)`. Sometimes there might be additional information on what exactly has to be handed in. **Read the file AssignmentFormat.pdf on Canvas carefully**. Be concise, yet complete! Do not write write a too lengthy report but make sure that all information for the required motivation of your answers is provided.

Exercise 7.1 Aerial survey methods are used to estimate the number of snow geese in their summer range areas west of Hudson's Bay in Canada. To obtain the estimates, small aircrafts fly over the range and, when a flock of snow geese is spotted, an experienced observer estimates the number of geese in the flock.

In order to check the reliability of a new, photograph-based estimation method, an experiment was conducted: An airplane carrying two observers flew over 45 flocks. A photograph of the flock is taken and each of both experienced observers independently estimate the number of geese in the flock. The data are contained in the file `geese.txt`.

a. **(Class)** For each observer, draw a scatter plot of the photo count (Y) versus the observer count (x). Do these graphs suggest a simple linear regression model might be appropriate?

b. Perform the linear regression for the two observers separately. Fit the parameters and test the hypothesis: $\beta_1 = 0$ against the alternative: $\beta_1 \neq 0$ with significance level 0.05 in each model.

c. Investigate the residuals by plotting residuals against Y for each model (you can add the line $y = 0$ using the function `abline`). What do these graphs tell you about the model assumptions?

d. Investigate the normality of the errors with one or more appropriate plot. For testing the normality use the function `lm.norm.test`. Note that the residuals are not independent. Read Example 5.5 from the syllabus carefully and have a close look at the code of the function `lm.norm.test`, before you apply it.

e. Repeat all steps in parts a through d while using the log transformation of all counts. Does this transformation stabilize the variance of the error variables?

f. Compare all 4 models that you have fitted; which models do you trust (most): the ones based on the original data, or the ones based on the transformed data? Explain your answer.

g. Write a few sentences about the question: How well does the photo count reflect the observer counts of the number of geese?

Hand in: relevant plots and answers to all questions (except part a, which will be discussed in class).

[1] You can find these functions in the file `functions_Ch8.txt`.

Exercise 7.2 The data in `steamtable.txt` is about a steam engine that produces glycerine: the column `Steam` contains the used amount of steam in pounds per month and the remaining columns contain values of 9 variables that possibly influence the used amount of steam. In this exercise you will set up a multiple linear regression model with the used amount of steam as response variable.

a. **(Class)** Make plots of the response variable against all possible explanatory variables (e.g. using `pairs` and compute the 9 pairwise correlations between the response variable and the explanatory variables. Then perform (only) the *first* step of the step-up method. Comment on your findings.

b. Find a suitable multiple linear regression model. Use diagnostic plots to set up and/or check your model. Give at least one added variable plot and comment on it.

c. Check your model in part b for possible influence points and collinearity. In case you find influence points, fit the model of part b also without these influence points.

d. Investigate the residuals of the selected model for normality.

e. Do you judge the selected model to be appropriate for the data? Motivate your answer.

Hand in: relevant plots and your analyses in part b, the results of parts c, d, and e.

Exercise 7.3 The data in `expensescrime.txt` were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on criminal activities in $1000), `bad` (number of persons under criminal supervision), `crime` (crime rate per 100000), `lawyers` (number of lawyers in the state), `employ` (number of persons gainfully employed by and performing services for a government) and `pop` (population of the state in 1000).
Perform a regression analysis (including variable selection and finally finding a suitable model) using `expend` as response variable and `bad`, `crime`, `lawyers`, `employ` and `pop` as independent variables. Your analysis should at least include:

a. investigation of leverage (potential) and influence points

b. investigation of problems due to multi-collinearity (groups of collinear variables)

c. investigation of residuals.

You may use all global and diagnostic techniques mentioned in the syllabus. State clearly all the choices you make during the regression analysis, including arguments for all your choices. (Note that there are several strategies possible!)