SDA 2019 — Assignment 12

For these exercises the *R* function `lm` is needed to fit linear models. The data to be analyzed should be in a `data.frame` format, see the first exercise. The function `lm.norm.test` is available on Canvas.[1]

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**.

Exercise 12.1 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. To investigate the reliability of this method, an experiment was conducted. An airplane carrying two observers flew over 45 flocks, and each observer independently estimated the number of geese in the flock. Also, a photograph of the flock was taken so that an exact count of the number in the flock could be obtained. The data are contained in the file `geese.txt`.

a. Draw scatter plots of the observer counts (Y) versus the photo 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 carefully Example 4.5 from the syllabus and have a close look at the code for this function, before you apply it.

e. Repeat the whole procedure in parts a through d using the log transformation of the counts. Does this transformation stabilize the variance?

f. Comparing all 4 models that you have fitted, which models do you trust better? Based on the original data, or based on the transformed data? Explain your answer.

g. Write a few sentences about the two questions: 1. How well do observers count the number of geese? and 2. How do the two observers compare?

Hand in: relevant plots and answers to all questions.

Exercise 12.2 is on the next page.

[1] You can find this and other functions that we will use next time in the file `functions_Ch8.txt`.

Exercise 12.2 This exercise concerns data measured by the Los Angelos Pollution Control District. This agency attempts to construct statistical models to predict pollution levels. The file `airpollution.txt` contains the maximum level of an oxidant (a photochemical pollutant) and the morning averages of four meteorological variables: wind speed, temperature, humidity and insolation (a measure for the amount of sunlight). The data cover 30 days during one summer. Investigate which explanatory variables need to be included into a linear regression model with `oxidant` as the response variable by performing the steps below.

a. Make scatter plots of the four candidate explanatory variables against each other and against the response variable (see the *R*-function `pairs()`). Interpret the plots. Do you judge a linear model to be useful here?

b. Determine for each of the explanatory variables the simple linear regression model. Choose the best among these models, and stepwise extend this model by adding one explanatory variable per step on the basis of the determination coefficient. Use a test to investigate whether the extensions are useful. Determine in this way an appropriate linear regression model for these data.

c. Estimate the parameters in the full multivariate linear regression model with all explanatory variables in it. Test whether the full model is useful via an *overall* analysis, i.e. should at least one of the variables be included in the model?

d. Now stepwise decrease the full model of part c with the aid of tests of the form $H_0 : \beta_i = 0$ with significance level 0.05. Determine in this way an appropriate linear regression model for the data.

e. The parts b and d possibly yielded a different model. If so, which one do you prefer and why? For choosing between the models take also into account the results of part a. Present the estimates of the parameters of the final model of your choice. Also report the estimated variance of the errors and the R^2-value of your final model.

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