SDA 2020 — 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. 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.

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

Exercise 7.2 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. Throughout this exercise, use the significance level $\alpha = 0.05$ for all hypothesis tests.

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 stepwisely 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. Proceed with this procedure until an appropriate linear regression model is obtained.

c. Estimate the parameters in the full multivariate linear regression model with all explanatory variables in it. Test whether the full model is useful (compared to the model without explanatory variables) via an *overall* analysis, i.e. should at least one of the variables be included in the model?

d. Now stepwisely decrease the full model of part c with the aid of tests of the form $H_0 : \beta_i = 0$. Proceed with this procedure until an appropriate linear regression model is obtained.

e. The parts b and d possibly yielded a different model. At this point, make a pre-selection of your preferred model and motivate your choice. For choosing between the models, take also into account the results of part a.

f. Use diagnostic plots to check and possibly adjust your model. Provide at least one added variable plot and comment on it.

g. Do the t-test in the mean shift outlier model for observation number 4 to check whether this observation is an outlier.

h. Check your model in part f. for possible leverage and influence points and collinearity, and, where appropriate, adjust your model based on these findings. In case you find influence points, fit the model of part f. also without these influence points.

i. Investigate the residuals of the selected model from part h. for normality.

j. Do you judge the resulting final model to be appropriate for the data? Motivate your answer. State that final model and present the estimates of the involved parameters. Also report the estimated variance of the errors and the R^2-value of your final model.

Hand in: relevant plots and your answers to parts a, f, g, and h, the results of parts b, c, and d, your answers to part e and i.

Exercise 7.3 The data in the file `expensescrime.txt` were obtained to determine factors related to state expenditures on fighting criminality (courts, police, etc.). The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on fighting criminality in $1000), `bad` (number of persons under judicial 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 full regression analysis (including variable selection). Use `expend` as the 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!)

Repetition from above: 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.