SDA 2019 — Assignment 13

For these exercises the standard *R*-functions `lm`, `hatvalues` and `cooks.distance` can be used. The latter two require the output of `lm` as argument, e.g. `cooks.distance(crimelm)`. For the collinearity measures, you can use the functions: `varianceinflation`, `conditionindices`, `vardecomposition` and `determinationcoef`, 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 13.1 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. 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.

Exercise 13.2 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) 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!)

[1] See the file `functions_Ch8.txt` in the previous assignment.