6 Exercises (due Thursday 18.01.2024, 8:00am)

There is a total of 7 points to achieve on this exercise sheet. Please upload your solutions in typed or readable (!) handwritten form (scans, photographs), and also your code, on Moodle and don't forget to flag all the problems you were able to solve. **If the runtime of your simulations is very high, you may want to save your results beforehand so we don't have to wait at the homework session until your simulations are finished.** Good luck!

Exercise 6.1 (**2P**). *Fit a linear model to the US crime data ('communities.csv') using all the* 122 *explanatory variables and produce a residual plot and a normal QQ-plot to evaluate model fit. Also visualize the distribution of the response variable* Y *and the distribution of the residuals* $\hat{u}_i$ *by histograms.*

Hint: *When preparing the data, remove the first five variables which are just community identifiers. The last variable 'ViolentCrimesPerPop' is the response variable. Also remove all cases (communities) for which there are some missing values. Implement the residual plot and the QQ-plot by hand, not using a ready-made package!*

Exercise 6.2 (**5P**). *Consider the US crime data set again where* $p = 122$ *is the number of explanatory variables and* $n = 319$ *is the number of complete cases without missing values.*

a) ***(2P)*** *Write a program that generates the design matrix which includes the main effects and all the* $\binom{p}{2}$ *two-way interaction effects, i.e., which has* n *rows and* $p + \binom{p}{2} = 7503$ *columns. If there are columns which are exactly identical, keep only one of them. This means that the final matrix may have less than* 7503 *columns.*

   Hint: *Recall that an interaction effect between two variables is a new variable. The corresponding column of the design matrix is given by the point-wise product of the two columns of the interacting variables.*

b) ***(2P)*** *Write a program that selects the* m *most important variables from all the main and interaction effects by forward stepwise selection.*

c) ***(1P)*** *Also implement an algorithm that whenever an interaction effect is selected also includes both of the corresponding main effects. Make it such that the resulting model is of size ate least* m *and at most* $m + 2$.