Exercise 1

The data in this example comes from a study of the effects of childhood sexual abuse on adult females reported in Rodriguez et al. ("Post-traumatic stress disorder in adult female survivors of childhood sexual abuse: a comparison study", Journal of Consulting and Clinical Psychology, 1997). 45 women who reported childhood sexual abuse (`csa`) were measured for post-traumatic stress disorder (`ptsd`) and childhood physical abuse (`cpa`), both on standardized scales. Additionally, the same quantities were recorded for 31 women who did not experience childhood sexual abuse. The dependent variable is `ptsd`. The data can be downloaded from the website. Read in the data with `read.table(...,sep=",",header=TRUE)`.

(a) Read in the data and investigate it graphically using the R function `pairs()`. Additionally, check if R reads the data correctly (i.e. ptsd and cpa as numerical variables, csa as factor variable).

(b) Investigate the relationship between the variable `ptsd` and `csa` respectively `ptsd` and `cpa` graphically.

(c) Now, create a scatter plot of `ptsd` against `cpa`. Use different colors for abused and non-abused women. What's the problem if we don't separate by abused and non-abused women. (**R-Hint**: First use `plot(...,type="n", pch=16)`. Then use `points(...,pch=16, col=...)` to plot the points for each subset.)

(d) Carry out a test in order to see if sexually abused women have a higher PTSD-score. Why does this test not give you a complete conclusion of the statistical dependence between `ptsd` and the predictors `cpa` and `csa`?

(e) Fit a regression model to the data with both predictors and their interaction. Check the model assumptions using appropriate plots.

(f) Is it appropriate to simplify the model from the previous task, i.e. are there terms that can be left out? If so, again perform a residual analysis of the simpler model.

(g) Draw two plots, one for the model with, one for the model without interaction term. As basis, you can use the plot where you differentiated abused and non abused women by color. Now, draw the regression lines on top of the plots. What's the difference. Do you think the interaction is necessary.

Exercise 2

In a study on the contribution of air pollution to mortality, General Motors collected data from 60 US Standard Metropolitan Statistical Areas (SMSAs). The dependent variable is the age adjusted mortality (called `Mortality` in the data set). The data includes variables

measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. You can download the data from the website and read it with `read.table(...,sep=",", header=TRUE)`

(a) First, set the city names as row names. Then, use histograms to check the distribution of the variables. If necessary, transform them. For right skewed data, use a log-transformation, for percentages, use an arcsin-transformation.

(b) Carry out a multiple linear regression containing all variables. Does the model fit well? Check the residuals. (**R-Hint**: Using "." in `lm(... ~ .)` includes all variables into the model.)

(c) Now take all the non-significant variables out of the model and compute the regression again. Do you think this is a good strategie to simplify the model? Compare your simplified model to the full model using an anova.

(d) Start with the full model. Remove now step by step the variable with the biggest p-value as long as it is over 0.05. Use again an anova to compare the full model to the reduced one. Compare the result to the result of the previous subtask. (**R-Hint**: Use the function `update()`)