Series 4

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). In this study 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`. Read in the data with
`sexab <- read.csv("http://stat.ethz.ch/Teaching/Datasets/abuse.csv",header=TRUE)`

- **a)** Read in the data and look at it, do you see any problems? Make sure that all the variables have the correct R data type.
- **b)** Use scatter plots and box plots to display the variable `ptsd` against the variables `csa` and `cpa`.
- **c)** Create a scatter plot of `ptsd` against `cpa`. Use different symbols for abused and non-abused women. R-hint:

```
plot(sexab$cpa, sexab$ptsd, type="n")
text(cpa, tsd, labels=substring(csa,1,1))
```

- **d)** Carry out a test in order to see if sexually abused women have a higher PTSD-score. Why this test does not give you a complete conclusion of the statistical dependence between `ptsd` and the predictors `cpa` and `csa`? Hint: Look at the scatter plot from part c.).
- **e)** Fit a regression model to the data with both predictors and their interaction. What do the resulting coefficients mean?

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. Read in the data with
`mortality <- read.csv("http://stat.ethz.ch/Teaching/Datasets/mortality.csv",header=TRUE).`

- **a)** Get an overview of the data and account for possible problems. Which of the variables need to be transformed?
- **b)** Carry out a multiple linear regression containing all variables. Does the model fit well? Check the residuals.
- **c)** Now take all the non-significant variables out of the model and compute the regression again. Compare your results to part b.).
- **d)** Start with the full multiple linear model. Remove now step by step the variable with the biggest p-value as long as it is over 0.05. Compare the result to part c.). R-hint: Use the R-function `update()`.
- **e)** Again starting from the full model, carry out partial F-tests, in order to answer the question if
  - all meteo-variables
  - all air pollution-variables and
  - all demographic-variables

  can be removed from the model. Use the R-function `anova()`.

Preliminary discussion: Monday, October 29.

Deadline: Monday, November 05.