Series 4

1. The file `farm.dat` contains the size A (in acres), the number of cows C and the income I (in $) of 20 farms in the US. Read in the data from the web using
`read.table("http://stat.ethz.ch/Teaching/Datasets/farm.dat", header = TRUE)`.

a) Compute an ordinary linear regression of I versus C. Does the income depend on the number of cows?

b) Give the confidence intervals for the expected income without any cows, with 20 cows, and with C = 8.85 cows. Give also a prediction interval for the income of a farm having no cows and with C = 8.85 cows.

c) Compute an ordinary linear regression of I versus A and also a multiple linear regression of I versus A and C. What do you conclude about the significance of the variables?

d) Use R-function `pairs()` on the data set `farm` to get pairwise plots of all the variables. What can you observe about the relationships between variables? Use this information to interpret the differences in the regression outputs calculated in **a)** and **c)**.

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()`.

3. The data in teengamb.rda comes from a study on teenage gambling in Great Britain. The goal is to fit a multiple regression model, where the gambling expenses (in pounds per year) is the response variable, and sex (0=male, 1=female), status (socio-economic status score, based on the parents employment), income (in pounds per week) and verbal score (number of correctly answered questions from 12 on use of language) are the predictors.

a) Use some visualization methods to gain a first overview on the data. Also decide which transformations are necessary.

b) Make sure that all predictors are from the correct data type in R.

c) Perform a multiple linear regression with all predictors (some of which may be transformed).

d) If all the other predictors remain the same, what is the difference in the predicted gambling expenses between a male and a female? Also give a confidence interval for that difference.

e) Start with an empty model, only containing the intercept. Then add the predictors step by step, one at each time. Use the following sequence: income, sex, verbal, status. After every step, write down the estimated error variance, R-squared and adjusted R-squared. Finally, display these graphically.

4. Which of the following statements are false and why?

a) If a predictor variable takes values from $(-\infty, +\infty)$, the log transformation of that predictor is not recommended.

b) If a predictor variable has a left skewed distribution, the log transformation of that predictor is recommended.

c) In a multiple linear regression non-linear transformations of the predictors change the regression output.

d) Doing many simple linear regressions is not the same as doing a multiple linear regression.

e) When doing model selection you should always choose the model with the largest coefficient of determination.

Preliminary discussion: Monday, October 27.

Deadline: Monday, November 3.