Series 6

1. **Collinearity and variable selection:** In a study about infection risk controlling in US hospitals a random sample from 113 hospitals contains the following variables:

id	randomly assigned ID of the hospital
length	average duration of hospital stay (in days)
age	average age of patients (in years)
inf	averaged infection risk (in percent)
cult	number of cultures per non-symptomatic patient x 100
xray	number of X-rays per non-symptomatic patient x 100
beds	number of beds
school	university hospital 1=yes 0=no
region	geographical region 1=NE 2=N 3=S 4=W
pat mittl.	average number of patients a day
nurs mittl.	number of full-employed, trained nurses
serv	percentage of available services from a fixed list of 35 references

Read in the data from: `http://stat.ethz.ch/Teaching/Datasets/senic.dat`. Since some observations span more than a single line, you have to use scan() to read the file into R:

```
senic <-scan("http://stat.ethz.ch/Teaching/Datasets/senic.dat",
  what=list(id=0,length=0,age=0,inf=0,cult=0,xray=0,beds=0,school=0,
  region=0,pat=0,nurs=0,serv=0))}
```

Using `senic <- data.frame(senic)`; `senic <- senic[ ,-1]` you turn the object into a user friendly data frame structure. Turn the variables school and region into so-called factor variables.

Perform a regression analysis on the following model:

`length` $\sim$ `age + inf + region + beds + pat + nurs`

The goal is to find the optimal model.

a) Check the correlation between these (not transformed) variables. Which variables are problematic and why? Suggest a combination of variables to improve the situation.

b) Perform the necessary transformations on the predictors and the response. Will there transformations be necessary for the above combinations as well?

c) Find a good model! To that end, analyze the residuals, identify possible problematic observations. Decide also upon which variables to use in the model and which to remove.
Perform a variable selection. Use as a starting model:
`log(length)` $\sim$ `age + inf + region + log(pat) + pat.bed + pat.nurs`
and remove observations 47, 102 and 112 from the data. The variables pat.bed and pat.nurs are coefficients of patient numbers with number of beds and number of nurses respectively.

d) Perform a backward elimination using the AIC criterion. Use the function `step()`. Check the final model with the usual diagnostic plots.

e) Now perform a forward selection using the AIC criterion. Thus, start with the empty model, i.e.:
`fit.for <- lm(log(length)` $\sim$ `1, data=...)`
Use the same function as before. Check also the diagnostic plots and comment on the differences to **d)**.

f) **Optional:** Perform a stepwise selection. Start with the full model as well as with empty model and compare the results. Check the help file of `step()` on how to perform a stepwise selection.

2. Cross validation: The goal of this exercise is to make you acquainted with the cross-validation technique. Use the data set `data(houseprices)` from the package `library(DAAG)`.

```
> head(houseprices)

   area bedrooms sale.price
9   694        4      192.0
10  905        4      215.0
11  802        4      215.0
12 1366        4      274.0
13  716        4      112.7
14  963        4      185.0
```

a) Perform a leave-one-out cross validation for the model containing both predictors as main effects: `sale.price` $\sim$ `area + bedrooms`
Is there a better model to predict the sale price? What other models are possible anyway? R hint: Use the R-function `CVlm()` from `library(DAAG)`.

b) Optional exercise for advanced users: Instead of using the function `CVlm(data, formula, fold.number, ...)` you could also perform the cross validation "by hand" using a `for`-loop.

Preliminary discussion: Monday, November 28.

Deadline: Monday, December 05.