Series 6

1. The file `fitness.rda` contains measurements of a fitness test for 31 patients. The target variable `oxy` is the rate of oxygen consumption which was measured with a complicated and expensive procedure. Predictors are `age`, `weight`, `runtime` (running time), `rstpulse` (resting pulse), `runpulse` (running/active pulse) and `maxpulse` (maximal pulse).

a) Analyze the data. What transformations are necessary? Are there any other problems? What can you say about the pairwise correlations between the predictors?

b) Fit a model containing all predictors after applying potentially necessary transformations.

c) Perform a residual analysis and ensure that the model does not show a systematic error or any other model violation. Using partial residual plots you can check whether all predictors have been included in the correct form.

d) Check whether there is high multicollinearity by computing the VIFs.
Hint: `library(faraway); vif(fit)`

e) Alleviate the multicollinearity problem by using different methods:

(i) Amputation, i.e. leave out redundant variables.

(ii) Create new variables that are not collinear.

(iii) Perform a ridge regression with an appropriate value for λ.

Save the fitted values for each of these variants and perform a pairwise comparison of these. What do you notice?

f) Use the model from subproblem **e)** (ii) and perform a backward elimination using the p-value as criterion in order to reduce the set of predictors to those that are mandatory to include.

g) The goal of the study is to substitute the expensive and complicated procedure to measure the rate of oxygen consumption by a set of predictors that are inexpensive to obtain. Can we conclude whether this is possible? If so, how should it be done?

2. 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`	average infection risk (in percent)
`cult`	number of bacteorological tests per asymptomatic patient x 100
`xray`	number of X-rays per asymptomatic patient x 100
`beds`	number of beds
`school`	university hospital 1=yes 2=no
`region`	geographical region 1=NE 2=N 3=S 4=W
`pat`	average number of patients a day
`nurs`	number of full-employed, trained nurses
`serv`	percentage of available services from a fixed list of 35 references

After reading in the data turn the variables `school` and `region` into factor variables.

Using the variables `age`, `inf`, `region`, `beds`, `pat`, `nurs` as predictors and `length` as response variable, perform a linear regression analysis and find an optimal model by following the next instructions:

a) Check the correlations between these variables. Which of them are problematic and why? Is there an intuitive explanation of this problem? Combine some of the predictors to improve the situation.

b) Perform the necessary transformations on the predictors and response.

- **c)** Fit a linear regression using the transformed variables. Then, use this model as your starting equation to do backward elimination (using p-values).
- **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. Use the same function as before. Check also the diagnostic plots and comment on the differences with **c)** and **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.

3. So-called "Funds of Hedge Funds" (FoHF), i.e. portfolios of hedge funds, have different investment strategies with specific returns and risk properties. When such a product is evaluated it is important for the investor to choose the investment style that fits his needs. One approach to assess the investment strategy of a FoHF as an outsider is to perform a style analysis based on the returns. Using a regression model (also called multi-factor model in the financial industry) one aims to explain the returns of the FoHF with the returns of the so-called subindices of hedge funds (Long Short Equity, Fixed Income Arbitrage, Global Macro, etc.). The estimated parameters are indications for the chosen investment strategy. Note that not all investment strategies are present due to the construction of FoHFs. The file `FoHF.rda` contains the monthly returns of one FoHF and the hedge fund subindices of EDHEC from January 1997 until December 2004. The meaning of the individual predictors is as follows:

`RV`	Relative value
`CA`	Convertible Arbitrage
`FIA`	Fixed Income Arbitrage
`EMN`	Equity Market Neutral
`ED`	Event Driven Multistrategy
`DS`	Distressed Securities
`MA`	Merger Arbitrage
`LSE`	Long Short Equity
`GM`	Global Macro
`EM`	Emerging Markets
`CTA`	CTA / Managed Futures
`SS`	Short Selling

Fit the following model:

`FoHF` $\sim$ `RV + CA + FIA + EMN + ED + DS + MA + LSE + GM + EM + CTA + SS`

- **a)** Look at the output of `summary()`. What conclusion can you draw with respect to the investment strategy of this FoHF when you consider the estimated coefficients, the p-values, the global F-test and the multiple R-squared?
- **b)** Check whether this model is valid or whether any assumptions are violated. Also test whether there are problems with respect to multicollinearity and whether all predictors have been entered into the model in the correct form.
- **c)** If you have solved the previous subproblem correctly, you will have found some issues. Formulate a strategy how those can be fixed in order to obtain a valid and interpretable result.
  Hint: Creating new predictors is not helpful.
- **d)** Perform variable selection using the BIC criterion. Implement the following search strategies, identify the best/final model and compare:
  - (i) Stepwise variable selection, starting with the full model.
  - (ii) Stepwise variable selection, starting with the empty model.
  - (iii) All Subsets variable selection.
- **e)** **Optional:** For this dataset the Lasso is well suited. Fit the model and generate the Lasso traces which allow to identify important predictors. Choose an appropriate value for λ and retrieve the final model as well as its coefficients.

Preliminary discussion: Monday, November 16.

Deadline: Monday, November 23.