Series 7

1. The file `CustomerWinBack.rda` provides a dataframe called `cwb`. It contains information about how long could a company hold costumers that cancelled the contract at some point in the past and re-opened their contracts afterwards. There are 295 observations and the following variables:

`duration`	target variable, duration of the customer relationship in days
`offer`	value of the present offered at re-acquisition
`lapse`	time until the customer could be re-acquired
`price`	offered price change in comparison to the first contract
`gender`	gender where 0 = female and 1 = male
`age`	age of the customer

The goal is to find a good model for the duration of the new customer relationship. Since we are primarily interested in an accurate prediction, we use cross validation to evaluate the predictive performance.

a) First, have an overview of the data. Then, fit the following models:
- OLS with all variables. Perform a residual analysis and decide whether transformations are necessary. For the purpose of this exercise, continue working with the untransformed data.
- Model chosen by stepwise model selection using the AIC criterion.
- Model chosen by stepwise model selection using the BIC criterion.
- Ridge regression with optimized λ.
- Lasso regression with optimized λ.

Finally, use a 5-fold cross validation to compare the predictive performance of these models.

b) For the model chosen by AIC model selection, compute the relevance of the predictors using three different approaches: (i) maximal effect, (ii) standardized coefficients, and (iii) LMG criterion. Plot these results against the negative common logarithm (i.e. $-\log_{10}(\cdot)$) of the respective p-value.

2. The file `autodistanz.rda` contains data from 82 men. Their age and their average daily distance traveled by car in 2014 are recorded. All of these men own and drive their own car.

a) Plot the data in a scatter plot and fit a LOESS smoother to visualize the relation between the target variable `distance` and the predictor `age`. Choose a suitable smoothing parameter.

b) Use an appropriate polynomial instead of the LOESS smoother to estimate the relationship between the target variable `distance` and the predictor `age`. Also perform a residual analysis. If the residual analysis suggests that a transformation is necessary, transform the respective variable and repeat the tasks from a) and b).

c) Use a model with cubic B-spline basis functions to estimate the relationship between the target variable `distance` and the predictor `age`. Then, plot the result. Make sure that the model contains as many terms as the polynomial from task b).

3. The data set `no2basel.rda` contains measurements of the amount of N02 in the air in Basel. The predictors are the day of the measurement, the temperature, and the amount of wind.

a) Fit a suitable multiple linear regression model and perform a residual analysis. Do you need to apply a transformation?

b) Model the N02 level with a GAM and perform a residual analysis. Assess whether a transformation is necessary.

c) Which model fits better? Perform a suitable test.

4. In this exercise, the prestige of several professions in Canada is assessed. The dataset is available using the commands `library(car)` and `data(Prestige)`. It is recommended to use the model `gam(prestige ~ s(income) + s(education))`.

a) What happens when you log-transform the variable `income`? Perform a residual analysis to assess whether a transformation of this predictor improves the GAM. Afterwards, repeat again the process, but now use an OLS model.

b) Create a factor variable that allows to differentiate between persons who have more than twelve years of education and those who have less. Include this new variable in the GAM fits from above and perform residual analyses. What is the effect of the factor variable? In comparison to the models fitted in a), which model is preferable?

Preliminary discussion: Monday, November 30.

Deadline: Monday, December 07.

Question hour: January 11, 2016, 3 – 4 pm in HG G 26.3 & January 22, 2016, 3 – 4 pm in HG F 26.3 .