Exercise 1

Advanced Methods for Regression and Classification

October 18, 2018

Load the data `Hitters` from the package `ISLR`. This means that you first need to install the package with

```
install.packages("ISLR")
```

and then load the data with

```
data(Hitters,package="ISLR")
```

Look at `?Hitters` or at `str(Hitters)` for more detailed information. Remove all observations which contain missings by using the command `na.omit()`.

Our goal is to find a model which allows to predict the variable `Salary`, i.e. the salary of the players on the basis of various statistics associated with performance in the previous years.

For the following tasks, split the data randomly into training and test data (about equal halves), build the model with the training data, and evaluate for the test data (using the MSE as a criterion).

Look first at your data. Is any preprocessing necessary or useful?

1. *Full model*: Estimate the full regression model and interpret the results.

2. *Stepwise regression*: Use the function `step()`. Find the optimal model using *forward selection*, *backward selection* and selection in both directions. Compare all the obtained models using ANOVA (see the lecture notes).

3. *Best subset regression*:

   (a) Use *best subset regression* which is implemented in the `library(leaps)` as the function `regsubsets()`, see help. To find the models, examine the best 3 models of each size, for a maximum model size of 8 regressors.

   (b) Plot the results. Which model seems to be the best?

   (c) Save the resulting `summary` as another object. Display the structure `str()` of this object and plot the size of models against BIC values. Which is the best model? Apply `lm()` on the final best model and interpret the results of `summary()`.

Compare all obtained models by calculating the MSE for the test data. Which model shows the best fit to the data?

How would you do the final model selection for repeated splits into training and test data, where you might end up with different "optimal" models for different training data sets?

Save your (successful) R code together with short documentations and interpretations of results in a text file (= R script file), named as *Matrikelnummer_1.R* (no word document, no plots). Submit this file to Exercise 1 of our tuwel course (deadline October 17).