Anna Titova, M.Sc.

Data Mining
Home Assignment 1

During this assignment you will work with a set of NFL statistics of the regular season 2016. You will perform backward and forward selection in order to uncover the best linear model for the given data. The goal hereby is to explain every team's gained points using given information on total passing yards, number of quarterback sacks and whether of not a team has won the Super Bowl in the last 10 years.
For those of you with no interest or knowledge about American football, here is some insight in the meaning of the variables:

- **Point**: during a game of football each team can earn points by scoring a touchdown (aka bringing the ball to the end of the opponents half of the field) or by scoring field goals (aka kicking the ball through the goal). A team with the most points at the end of the game wins.
- **Passing Yards**: a quarterback (aka team captain) can start an offensive drive by passing the ball to another player to bring it closer to the end zone, thus raising the probability to score a touchdown. Hence, it should be positively correlated to the number of points.
- **Sacks**: if during a drive the defence team manages to tackle the quarterback to the ground before he throws the ball, the offence team gets a yard penalty, thus is moved away from the touchdown. Hence, more sacks should have a negative effect on earned points.
- **Super Bowl wins**: This is a dummy variable that is equal to 1 if a team has been a league champion in the last 11 years and zero otherwise. The effect is unclear, but supposedly positive.

Your task is to write an `R` script that contains the following parts. But first, download the script template *HA1_yournames.R* and the CSV-file *nfl2016.csv* from OLAT.

1. Import the data from *nfl2016.csv* and save it in a data frame called `nfl2016`.
2. Fit a linear model of the form

$$\text{Points} = \beta_0 + \beta_1 \cdot \text{Passing.Yards} + \beta_2 \cdot \text{Sacks} + \beta_3 \cdot \text{SB.Win.11y} + \epsilon$$

3. Perform forward selection to find the best model starting with zero regressors. Choose an appropriate selection criterion! You should use an automatic procedure to search through the models (e.g. a loop).
4. Perform backward selection to find best model. Choose an appropriate selection criterion!
5. Plot the residuals of the two winning models and check for their normality.

Remarks: Write comments for everything you do. Codes that are not written using the template and/or that return error messages will not be evaluated. If you are working in groups, make sure to note down every participant's name and ID.
Submission: Submit your scripts via email to *atitova[at]stat-econ.uni-kiel.de* until the end of May 24th (until 00:00:00, May 25th)