Mariia Okuneva, M.Sc.

Data Mining
Home Assignment 1

In this home assignmnet you will use data on house sales in King County. 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 house prices using given information on some house features.

Description of the variables:

- **price**: price of the house (prediction target)
- **bedrooms**: number of bedrooms
- **sqft_living**: square footage of the house
- **floors**: total floors in house
- **grade**: overall grade given to the housing unit, based on King County grading system
- **condition**: how good the condition is (overall)

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 *house_data.csv* from OLAT.

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

$$\text{price} = \beta_0 + \beta_1 \cdot \text{bedrooms} + \beta_2 \cdot \text{sqft\_living} + \beta_3 \cdot \text{floors} + \beta_4 \cdot \text{grade} + \beta_3 \cdot \text{condition} + \epsilon$$

3. Write a function that computes the least squares estimator. You need to write your own function, you are not allowed to use any existing function (e.g. lm()). Please name your function "OLS". Your input arguments are *x, y, intercept*, where *x,y* are independent and dependent variables respectively, and *intercept* indicates whether you want to estimate including an intercept. Default is with intercept. The outputs of the function are estimated β parameters as well as the residual sum of squares. Make sure that estimated β coefficients from your function and lm() function coincide.
4. 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). You are not allowed to use any existing function for model selection (e.g., step()), but you can use lm() function.
5. Perform backward selection to find best model. Choose an appropriate selection criterion! You are not allowed to use any existing function for model selection (e.g., step()), but you can use lm() function.

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 (not more than 5 students in one group), make sure to note down every participant's name and ID.
Submission: Submit your scripts via email to *mokuneva[at]stat-econ.uni-kiel.de* until the end of May 16th (until 00:00:00, May 17th)