Homework 4

Using Ridge regression and Lasso regression for fitting linear models and for selecting features.

Please use the functions from `tidymodels` package for model fitting and cross-validation. In order to fit Ridge and Lasso regression models, please use the engine `glmnet`. The data for homework is in the file `hw3_studybookno.csv` (where `studybookno` is your study book number), which is attached to the assignment in Moodle. The aim is to fine a good linear model for predicting the variable `y`. For the current assignment use the data as it is (no elimination of possible outliers, no transformations except adding dummies for nominal variables with recipes; fitting method may internally use normalization of variables).

I remind you that you are expected to solve the homework problems by yourself, sharing solutions (and copying parts of other student's solutions) is not allowed.

Please submit your .R or .Rmd file and and html/pdf file which shows you code together with output and comments

Problem 1

Use the command `set.seed(study_book_no)`, where `study_book_no` is the numeric part your study book number, at the beginning of your solution. This should be the only `set.seed()` command in your solution. Set up model evaluation framework by splitting the data to test and train set (80% for training) and further form 10-fold cross-validation data splits from the training data for model selection.

Problem 2

Find the best penalty parameter for the Ridge regression model (considering 10 penalty values from `0.001*sd(y)` to `10*sd(y)` which are uniformly spaced in logarithmic scale) by using cross-validation and compute RMSE of this model on test set. Is the value of the RMSE close to the cross-validation result? Note that one way to get values from a to b which are uniformly spaced in log scale is `exp(seq(log(a),log(b),length.out=n))`.

Problem 3

Find the best penalty parameter for the Lasso regression model (considering 10 different penalty values from $0.001\sigma_y$ to σ_y by using cross-validation and compute RMSE of this model on test set. Is the value of the RMSE close to the cross-validation result?

Problem 4

Use Lasso regression to select the best 10 variables to a linear model (or if exactly 10 is not possible according to the table of degrees of freedom for different penalties, then select largest number of variables not exceeding 10). Use those variables to fit a linear model by standard fitting procedure using the data in the training set and compute it's performance on the test set. Is the performance on the test set better or worse than for the best Lasso model?