STAT448 - Assignment 2

Due by 11:55 pm of May 15, 2020
Assignment should be submitted on Learn as a html Rmarkdown file for the R code and results, properly commented.
Assignments can be done in pairs: both names and student id's should be on the assignment and both students should submit.

1. **(20 marks)** R exercise. On learn, in the folder for the assignment 2 you will find 2 files: the RData file "Residen" and the excel file "Residential-Building-Data-Set.xlsx". Residen is a copy of the dataset in the excel file were the variables names have been changed for use in R. The excel file contains also a description of the variables. More information on the dataset can be obtained on the UCI webpage:

   `https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set`.

   Please show R code and output for all the tasks below:

   (a) Set a seed at the beginning of your code equal to the last 4 numbers of your student id (or one of your student id's if you work in pairs)

   (b) Split the data set into a training set and a test set

   (c) Fit a linear regression model on the training set to explain the "actual sales price" (V104) in terms of the of the other variables excluding the variable "actual construction costs" (V105). Report the test RMSE obtained.

   (d) Fit a linear regression model using stepwise selection on the training set. Report the test RMSE obtained.

   (e) Fit a linear regression model using ridge regression on the training set, with λ chosen by cross validation. Report the test RMSE obtained.

(f) Fit a linear regression model using lasso on the training set, with λ chosen by cross validation. Report the test RMSE obtained along with the number of non-zero coefficient estimates.

(g) Comment on the results obtained. How accurately can we predict the actual sale price? Is there much difference among the test errors resulting from these four approaches?