# Homework # 7. AMS 597

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# The homework is due on Monday, May 4, 2020, at 12noon. Solutions will be emailed to you by 1pm. The homework will not be graded and all complete submission will warrant full points. Quiz 9 will be given on Tuesday, May 5, 2020, in class, covering the same materials – for the entire class duration; and will be graded. So please study hard from now on. Please email your completed homework to your TA at: [song.jiecheng@stonybrook.edu](mailto:song.jiecheng@stonybrook.edu)

# Please include (1) R code

# (2) Output from R

# (3) Answers to all the questions asked

# Please refer to the following website website for steps for Penalized Regressions in R:

# <http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/>

#### Penalized Regression with the Boston Housing Data

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive (<http://lib.stat.cmu.edu/datasets/boston>), and has been used extensively throughout the literature to benchmark algorithms. The goal is to model the variable MEDV using the other 13 variables.

There are 14 attributes (variables) in the data set:

CRIM - per capita crime rate by town

ZN - proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS - proportion of non-retail business acres per town.

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX - nitric oxides concentration (parts per 10 million)

RM - average number of rooms per dwelling

AGE - proportion of owner-occupied units built prior to 1940

DIS - weighted distances to five Boston employment centres

RAD - index of accessibility to radial highways

TAX - full-value property-tax rate per $10,000

PTRATIO - pupil-teacher ratio by town

B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT - % lower status of the population

MEDV - Median value of owner-occupied homes in $1000's

1. Please use the random seed 123 to divide the data into 75% training and 25% testing.

1. Please first find the best Ridge Regression model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
2. Please first find the best LASSO model using the training data. Please (a) find the best **λ** valuethrough cross-validation and display this value; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .
3. Please first find the best Elastic Net model using the training data. Please (a) find the best **tuning parameter** valuesthrough cross-validation and display these values; (b) display the coefficients of the fitted model; and (c) make prediction on the testing data, and report the RMSE and the Coefficient of Determination .

