# Lab 3: Linear Regression

## Due Date: Sunday January 23rd 2022

1. **Predicting Boston Housing Prices.** The file *BostonHousing.csv* contains

information collected by the US Bureau of the Census concerning housing in the

area of Boston, Massachusetts. The dataset includes information on 506 census

housing tracts in the Boston area. The goal is to predict the median house price in

new tracts based on information such as crime rate, pollution, and number of

rooms. The dataset contains 13 predictors, and the outcome variable is the median

house price (MEDV). Table 6.11 describes each of the predictors and the outcome

variable.

**Table 6.11 Description of Variables for Boston Housing Example**

|  |  |
| --- | --- |
| **CRIM** | Per capita crime rate by town |
| **ZN** | Proportion of residential land zoned  for lots over 25,000 ft2 |
| **INDUS** | Proportion of nonretail business  acres per town |
| **CHAS** | Charles River dummy variable (=1 if  tract bounds river; =0 otherwise) |
| **NOX** | Nitric oxide 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 centers |
| **RAD** | Index of accessibility to radial  highways |
| **TAX** | Full-value property-tax rate per  $10,000 |
| **PTRATIO** | Pupil/teacher ratio by town |
| **LSTAT** | Percentage lower status of the  population |
| **MEDV** | Median value of owner-occupied  homes in $1000s |

a. Why should the data be partitioned into training and validation sets? What

will the training set be used for? What will the validation set be used for?

b. Fit a multiple linear regression model to the median house price (MEDV) as a

function of CRIM, CHAS, and RM. Write the equation for predicting the

median house price from the predictors in the model.

c. Using the estimated regression model, what median house price is predicted

for a tract in the Boston area that does not bound the Charles River, has a

crime rate of 0.1, and where the average number of rooms per house is 6?

d. Reduce the number of predictors:

i. Which predictors are likely to be measuring the same thing among the 13

predictors? Discuss the relationships among INDUS, NOX, and TAX.

ii. Compute the correlation table for the 12 numerical predictors and search

for highly correlated pairs. These have potential redundancy and can

cause multicollinearity. Choose which ones to remove based on this table.

Optional (Not Graded) for Winter 2022

iii. Use two subset selection algorithms: *backward* and *forward* to reduce the remaining predictors. Compute the validation performance for both of the selected models. Compare RMSE, MAPE, and mean error, as well as histograms of the errors. Finally,

describe the best model.