ECON 5253: PS9

William Lorton

April 13, 2021

1 Question 7

The original housing data set has 14 columns: 1 Y (dependent) variable called medv and 13 X (predictor) variables.

The prepped training data set has 75 columns: still 1 Y variable called medv, but now 74 X variables.

We added 61 predictor variables during pre-processing (we went from 13 in the original set to 74 in the training set).

2 Question 8

The LASSO regression with 6-fold cross validation run gave an optimal λ value of 0.00222. The in-sample RMSE was 0.0625. The out-of-sample RMSE was 0.220.

3 Question 9

The Ridge regression with 6-fold cross validation run gave an optimal λ value of 0.0373. The in-sample RMSE was 0.0694. The out-of-sample RMSE was 0.219.

4 Question 10

It is not possible to estimate a linear regression model on a data set that has more columns than rows (i.e., more variables than observations). We would not be able to get meaningful OLS estimates for each of the slope parameter coefficients due to there being a perfect multicollinearity problem (there would be exact linear relationships among the coefficients). Also, we can see from the formula below that the variance of the error term would be negative which makes no sense:

$$\hat{\sigma}^2 = \frac{SSR}{N - K - 1}$$

SSR refers to the sum of squared residuals (the sum of the squared differences between the observed values of the dependent variable and the predicted values), N refers to the sample size (the number of observations), and K refers to the number of predictors. If K is larger than N, the error variance will necessarily be negative.

The out-of-sample prediction errors (measured by the Root Mean Squared Error) for both of the models seem somewhat low. Low prediction error should mean that we are striking a solid balance between model simplicity (high bias; meaning the model doesn't place enough importance on variation in the data) and model complexity (high variance; meaning the model places too much importance on variation in the data to the point that it seeks to attribute meaning to random noise). This should mean that we're neither underfitting nor overfitting the data.