Problem Set 9

Collin DeVore

April 2023

Question 7

My training data consists of 404 observations of 14 variables (404 by 14 dataframe). The original housing dataframe is 506 observations of 14 variables (506 by 14 dataframe). Though the values have clearly changed, I am not seeing any evidence that I have gained or lost variables.

Question 8

The best lambda parameter shown by the bestrmselasso variable is 0.00356, while the best rmse value is 0.0687, as provided by the mean of the toprmselasso variable.

Question 9

The best lambda parameter shown by the bestrmseridge variable is , which is 0.0373. This is much higher than the previous value. The best rmse value is 0.0679, which is almost the same as the lasso model. This suggests that the two models can give different, but comparable, results.

Question 10

Running a simple OLS model with more columns than rows would be pretty insane and heavily overfit, since the number of variables would exceed the number of observations. This means that the regression would seem to fit the data perfectly, but the results and the model would have almost no generalizability. My lasso model has a lambda parameter of 0.00356, meaning it has fairly low bias and high variance, whereas my ridge model has a lambda value of 0.0373, which is a significantly higher bias with a lower variance. This can easily be

seen because higher lambdas imply higher bias and lower variance.