Problem Set 9

Samuel Hunt

April 8, 2025

1 Linear Regression Prediction Models

The original housing data contains just 14 variables while our housing test prepped and housing train prepped data sets have 75 variables each. Our training data has a dimension of 404 observations by 75 variables and thus it has 61 more X variables than the original housing data.

After running our LASSO model, we found that the optimal value of λ, our penalty parameter, is 0.0139. We also found an in-sample RMSE of 1.96 and an out-of-sample RMSE of 1.95.

After running our ridge model, we found that the optimal value of λ, our penalty parameter, is 0.0373. We also found an in-sample RMSE of 1.96 and an out-of-sample RMSE of 1.95.

If we had a data set that had more columns than rows, then we would not be able to use simple linear regression model on it to get accurate results. We would get infinite solutions to our system of equations and the OLS model would not work. Our models both seem to achieve a good medium in the bias-variance tradeoff. First we notice that in both our LASSO and ridge model that the in-sample RMSE and out-of-sample RMSE are extremely close, with just 0.01 difference. This indicates that the models have low variance, as they perform about as well on the training and test sets. Additionally, our errors of 1.95 and 1.96 are not too large in magnitude, which indicates that we also have a relatively low bias, as the models are fairly accurate in their predictions.