PS9-Onishi

Saryu Onishi

April 2023

1 housing data

dim(housing) = 506 14
dim(housing_train) = 404 14
Hence, there are an equal number of X variables. However, the housing_train_prepped data has 61 more X variables.

2 LASSO model

$\lambda = 0.00139$
In-sample RMSE = 0.1703
Out-of-sample RMSE = 23.3683

3 Ridge regression model

$\lambda = 0.0373$
In-sample RMSE = 0.1734
Out-of-sample RMSE = 23.3700

4 question 10

You cannot estimate a simple linear regression model if a data set has more columns than rows because you would be faced with over-fitting. This means the bias will be very low, but the variance will be very high. Based on the RMSE values of the models, I think the models are over-fitted. This is because the in-sample RMSE can be considered low, but the out-of-sample RMSE is quite high. In fact, the out-of-sample RMSE represent a range that covers almost half of the actual range of the original data set.
In terms of bias-variance trade-off, these models have very low bias, but very high variance.