PS9 - ECON 5253

Hannah Bermudez

April 2025

1 Q7

- What is the dimension of your training data?
  - The training data has 74 rows and 404 columns.
- How many more X variables do you have than in the original housing data?
  - The original housing dataset had 14 variables, 13 of which were predictors (X variables). This suggests that the training data has 61 more predictor variables than the original dataset.

2 Q8 - Estimating LASSO

- What is the optimal value of lambda?
  - The optimal value of lambda is: 0.00222
- What is the in-sample RMSE?
  - The in-sample RMSE is: 0.138
- What is the out-of-sample RMSE?
  - The out-of-sample RMSE is: 0.184

3 Q9 - Estimating Ridge

- What is the optimal value of lambda?
  - The optimal value of lambda is: 0.0000000001
- What is the in-sample RMSE?
  - The in-sample RMSE is: 0.140
- What is the out-of-sample RMSE?
  - The out-of-sample RMSE is: 0.181

4 Q10

- Would you be able to estimate a simple linear regression model on a data set that had more columns than rows?
  - You cannot estimate a simple linear regression model (using OLS) when there are more columns than rows. This is due to multicollinearity in the X matrix.

- Using the RMSE values of each of the tuned models in the previous two questions, comment on where your model stands in terms of the bias-variance trade-off.

  – For the LASSO model, the in-sample RMSE was 0.138 and the out-of-sample RMSE was 0.184, while the Ridge model had an in-sample RMSE of 0.140 and an out-of-sample RMSE of 0.181. The LASSO model fits the training data slightly better (lower in-sample RMSE) but performs slightly worse on the test data (higher out-of-sample RMSE), indicating it may have higher variance. Conversely, the Ridge model sacrifices a bit of in-sample fit for improved generalization, suggesting it introduces slightly more bias but achieves lower variance overall. Therefore, the Ridge model may be better balanced in terms of the bias-variance trade-off in this case, as its smaller out-of-sample error indicates better predictive performance on new data.