Problem Set 9

William Miller

April 8, 2025

1 Problem Set Questions

1.1 Training Data Dimensions

My housing_test dataframe has 74 features while the original housing_data dataframe had 14. That means we introduced 60 new features with our recipe

1.2 LASSO Model

My best λ is 0.00139, with an in-sample RMSE of 1.96 and an out-of-sample RMSE of 1.95. I have also produced two diagnostic plots attached to this document.

1.3 Ridge Model

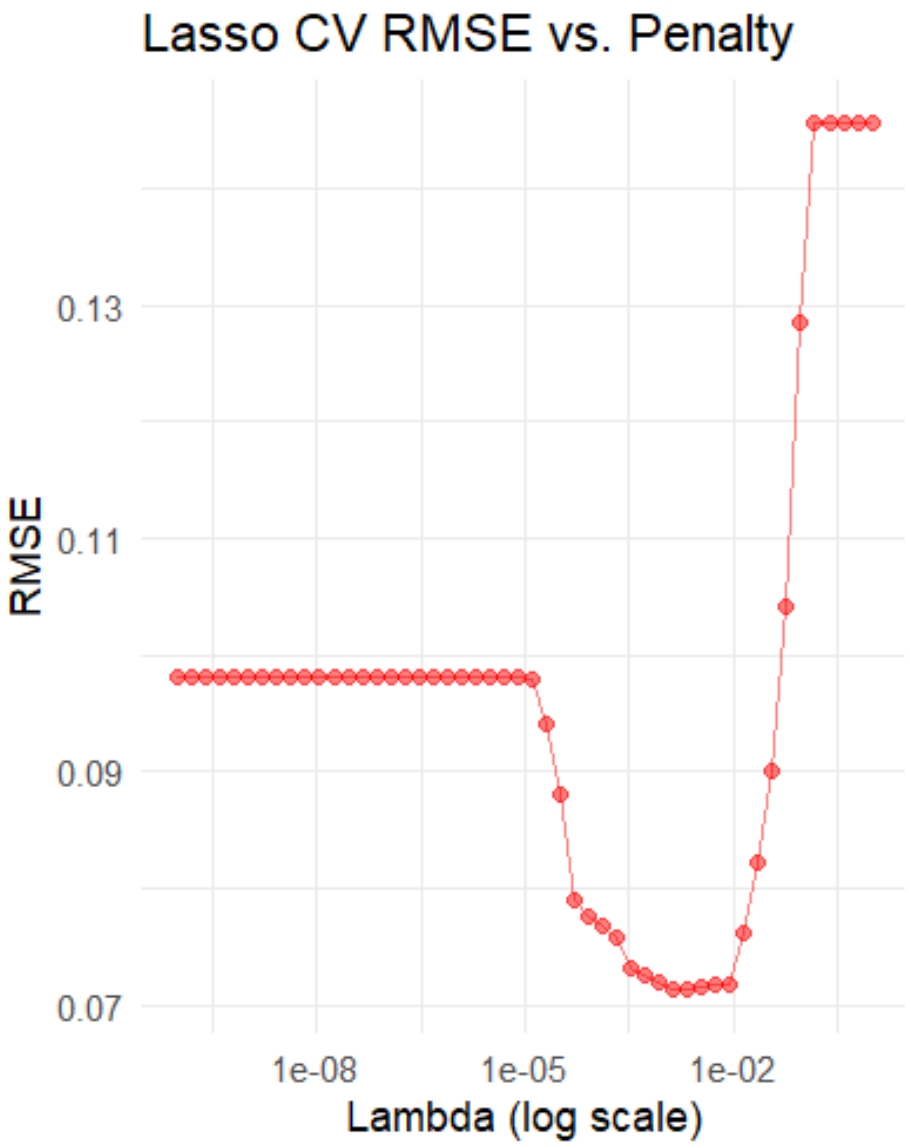
My best λ was 0.0146, with an in-sample RMSE of 1.96 and an out-of-sample RMSE of 1.95. This was very similar to the LASSO model. Diagnostic plots for my ridge model are also available at the end of this document.

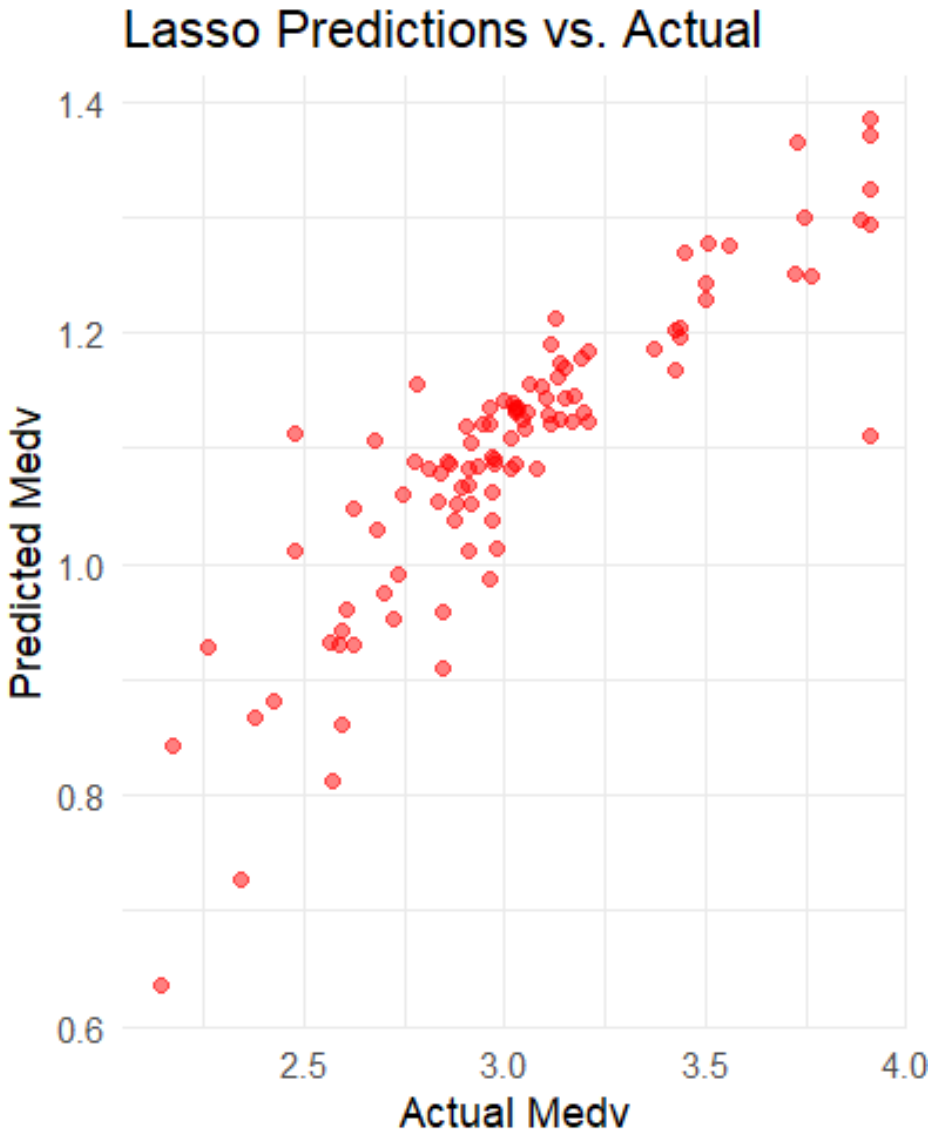
1.4 OLS with k¿n

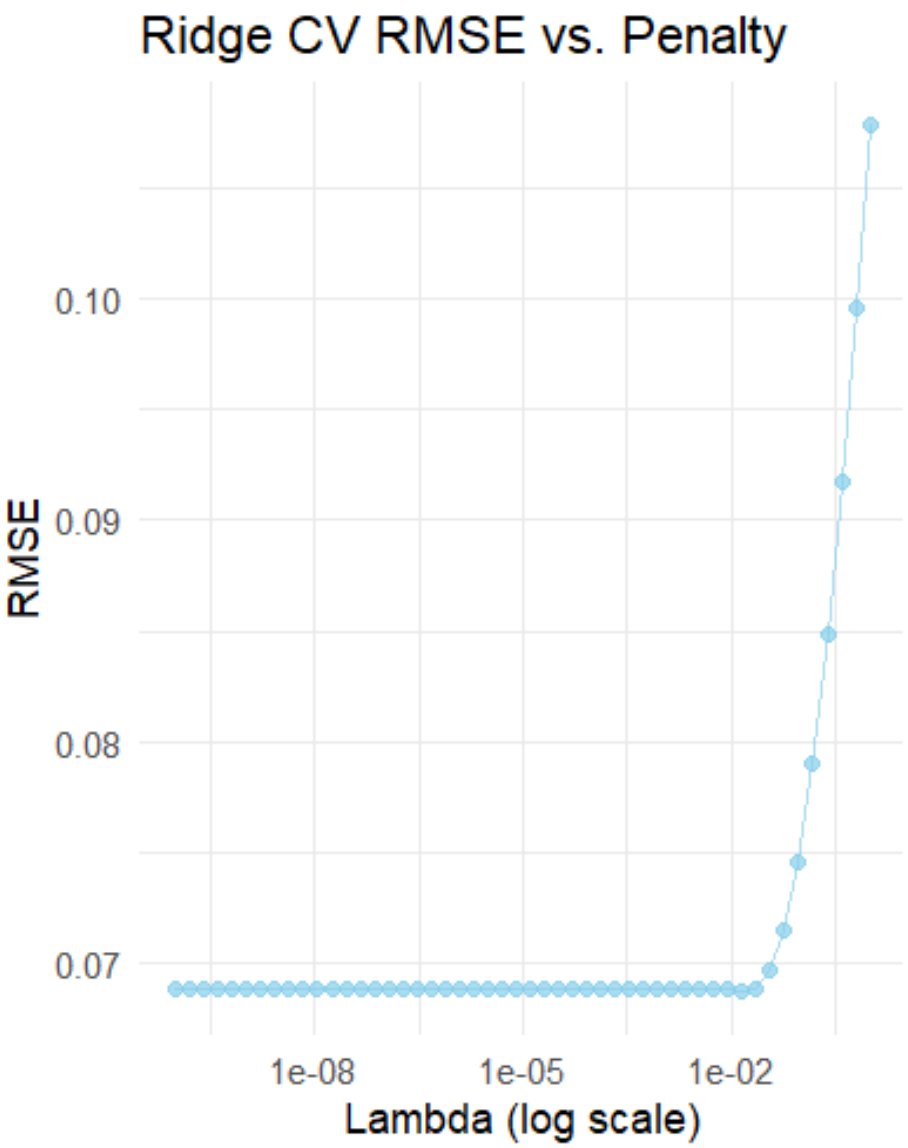
I don't believe OLS will produce a unique solution with more features than observations, which is why tools like LASSO and PCA are so useful for "big data" situations.

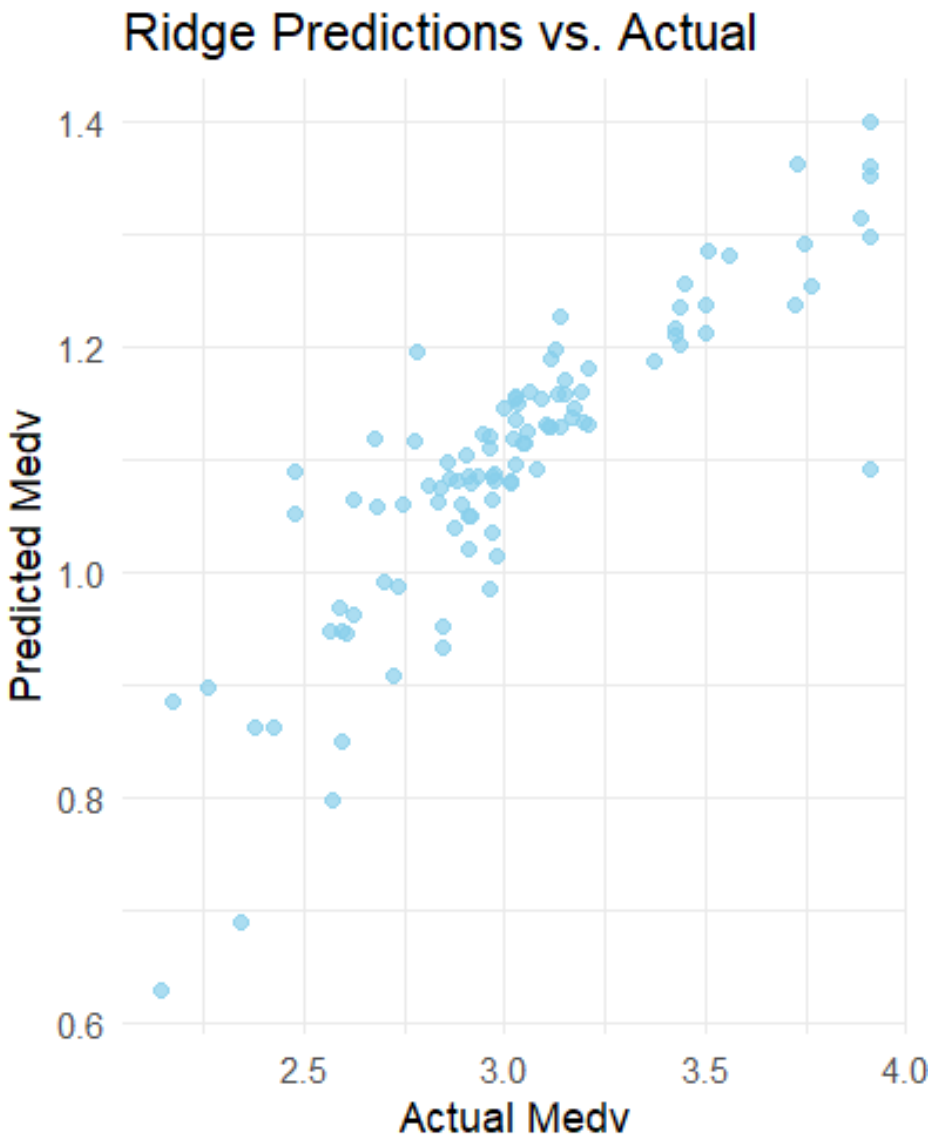
2 Bias-Variance Tradeoff

Because our in-sample and out-of-sample RMSE are small and quite similar for both models, I would say we are effectively balanced bias and variance.


Lasso CV RMSE vs. Penalty
RMSE
0.13
0.11
0.09
0.07
1e-08
1e-05
1e-02
Lambda (log scale)


Lasso Predictions vs. Actual
Predicted Medv
Actual Medv
1.4
1.2
1.0
0.8
0.6
2.5
3.0
3.5
4.0


Ridge CV RMSE vs. Penalty
RMSE
0.10
0.09
0.08
0.07
1e-08
1e-05
1e-02
Lambda (log scale)


Ridge Predictions vs. Actual
Predicted Medv
Actual Medv
1.4
1.2
1.0
0.8
0.6
2.5
3.0
3.5
4.0