Problem Set 9

JaeSeok Oh

April 8, 2024

Q.7-9

- The dimension of the training data is 74, including 6th polynomial for each variables. Since the original data only contains 14 columns, the training data includes 6 times more variables.

LASSO model

- Under 6-fold cross validation, λ is 0.00139. The out-of-sample RMSE is 0.188, while the in-sample RMSE is 0.137.

Ridge Regression Model

- Under 6-fold cross validation, λ is 0.0373. The out-of-sample RMSE is 0.180, while the in-sample RMSE is 0.140.

Q.10

- L1(LASSO) model is penalizing the object function by the sum of absolute value of parameters, while L2(Ridge) model is penalizing by the sum of square of parameters. Thus, the panelty is larger in the Ridge model.
- Previous model is set with 74 variables and 102 observations in test data. If I set 7 or more polynomial, it would be much more complicated model, which causes overfitting and high bias. To support this, I run a simple linear regression with 74 variables with 102 observations and find that the RMSE is 21.51 which is much higher than the regularized models in this problem set. Therefore, the simple regression with the data set having more rows than columns should cause problem in terms of bias and overfitting.