Data Science
Ruichun Liu
Problem Set 9

1. **Question 5** What is the dimension of your training data (housing.train)?

   The dimension of the training data is 404 by 450.

2. **Question 6** What is the optimal value of λ? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

   The optimal value of λ is 0.00764. The in-sample RMSE is 0.1665363. The out-of-sample RMSE is 0.1628410.

3. **Question 7** What is the optimal value of λ now? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

   The optimal value of λ now is 0.0978. The in-sample RMSE is 0.1644472. The out-of-sample RMSE is 0.1660548.

4. **Question 8** What are the optimal values of λ and α after doing 6-fold cross validation? What is the in-sample RMSE? What is the out-of-sample RMSE? Does the optimal value of α lead you to believe that you should use LASSO or ridge regression for this prediction task?

   The optimal values of λ and α after doing 6-fold cross validation are 0.0167 and 0.268 respectively. The in-sample RMSE is 0.1688519. The out-of-sample RMSE is 0.1714476. The optimal value of α lead me to believe that I should use ridge regression for this prediction task because the value of α is 0.268 which is very small and similar to the case of ridge regression.

5. **Question 9** Explain why you would not be able to estimate a simple linear regression model on the housing.train dataframe. Using the RMSE values of each of the tuned models in the previous three questions, comment on where your model stands in terms of the bias-variance tradeoff.

   From question 5, we know the dimension of the training data is 404 * 450. If we use OLS, then there might be severe overfitting problems without regularization.
   From the previous questions, we can see that the values of RMSE are quite similar and I think these values are kind of high. We can also find the values of λ and α are not too high. I think these models have high bias and low variance.