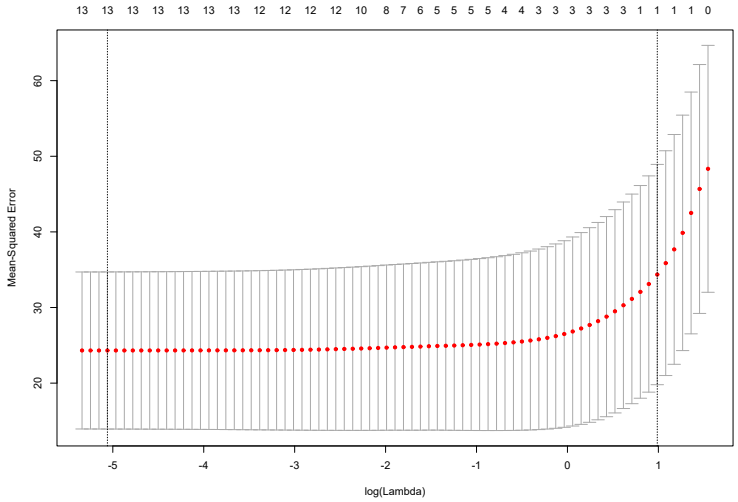
High-Dimensional Data Analysis Exercise 11

Wu, Bo-Kuan (105225010)

June 22, 2018

Use the lasso regression method to analyze the Boston data set. We will try to predict per capita crime rate in the Boston data set. In order to choose best lambda in this model, we will use cross validation to estimate the error rate on test data.

- First separate the Boston data set to two part. One is training data. Another is testing data. The training data containing 404 observations and a testing data containing 102.




```
> coef(cv.out)
14 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 1.0160886
zn          .
indus       .
chas        .
nox         .
rm          .
age         .
dis         .
rad         0.2317245
tax         .
ptratio     .
black       .
lstat       .
medv        .
```

Hence, we can find that the full model is lowest mean square error. Top of above figure indicates number of variables in the model. The above plot indcates that for high lambda error is very high, and the coefficients are restricted to be too small. This indicates that the full model is good.

- We can find the best lambda is 0.006349189.

```
> bestlam=cv.out$lambda.min
> bestlam
[1] 0.006349189
```

- The testing mean square error is 108.3341.

```
> lasso.pred=predict(lasso.mod,s=bestlam,newx=x[test,])
> mse= mean((lasso.pred-y.test)^2)
> mse
[1] 108.3341
```

- Hence, we refit the full model by using the best lambda. The testing mean square error in full model is 509.0434.

```
> lasso.coef
 (Intercept)           zn        indus         chas          nox           rm
16.143029931  0.042830653 -0.066811636 -0.709664263 -9.661563036  0.396794411
         age          dis          rad          tax      ptratio        black
 0.000508987 -0.951639599  0.570438469 -0.002811919 -0.255048311 -0.007535703
       lstat         medv
 0.127259085 -0.190643652
> error=oby-lasso.pred
> mse=mean(error^2)
> mse
[1] 509.0434
```

The full model is

$$\hat{Y} = 16.143029931 + 0.042830653 \times \text{zn} - 0.066811636 \times \text{indus} - 0.709664263 \times \text{chas} - 9.661563036 \times \text{nox} + 0.396794411 \times \text{rm} + 0.000508987 \times \text{age} - 0.951639599 \times \text{dis} + 0.570438469 \times \text{rad} - 0.002811919 \times \text{tax} - 0.255048311 \times \text{ptratio} - 0.007535703 \times \text{black} + 0.127259085 \times \text{lstat} - 0.190643652 \times \text{medv}.$$