Analytics 590: Homework # 1

Jiajia Liu

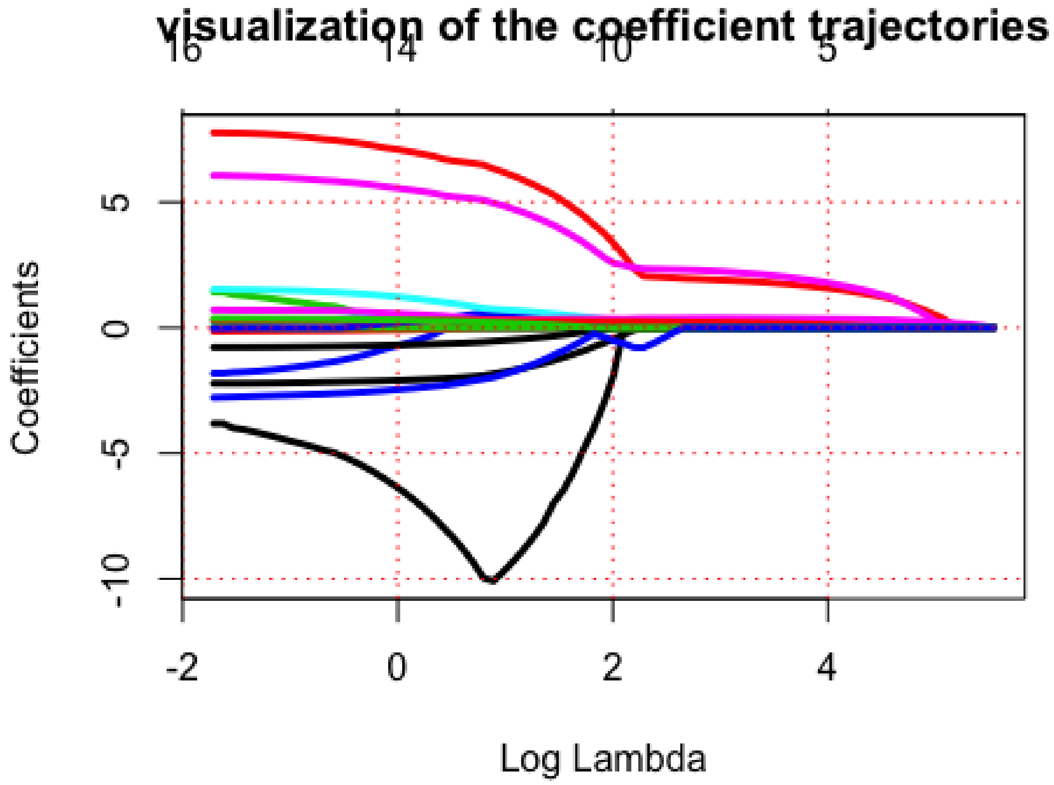
Sept, 15st

This is homework 1.

1

1.1:

```
set.seed(1234)

library(MASS)
library ("glmnet")

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-13

hitters<- read.csv("/Users/kaimouto/Desktop/ANLY 590/assignments/Hitters.csv"
)
hitters = na.omit(hitters)
x<- model.matrix(Salary~ .-X-League-Division-NewLeague-Salary, data=hitters)

noCV.lasso=glmnet(x,hitters$Salary, alpha=1)

plot(noCV.lasso, xvar = "lambda", lwd = 3,main='visualization of the coeffici
ent trajectories')
grid(col = 2,lty = 3)
```



```
coef(noCV.lasso)[,5]

##  (Intercept)  (Intercept)        AtBat         Hits        HmRun
## 444.06785122   0.00000000   0.00000000   0.08064999   0.00000000
##         Runs          RBI        Walks        Years       CAtBat
##   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
##        CHits       CHmRun        CRuns         CRBI       CWalks
##   0.00000000   0.00000000   0.06719193   0.17823025   0.00000000
##      PutOuts      Assists       Errors
##   0.00000000   0.00000000   0.00000000
```

From the table and the plot above, we can see the final three predictors are variable “Hits”, variable “CRuns”, and variable “CRBI”, which their coffieient are not equal to zero.

```
set.seed(1234)

cv.lasso=cv.glmnet(x,hitters$Salary, type.measure = 'mse', alpha=1)

#choose the Lambda

y<- cv.lasso$lambda.min
y

## [1] 2.935124
```

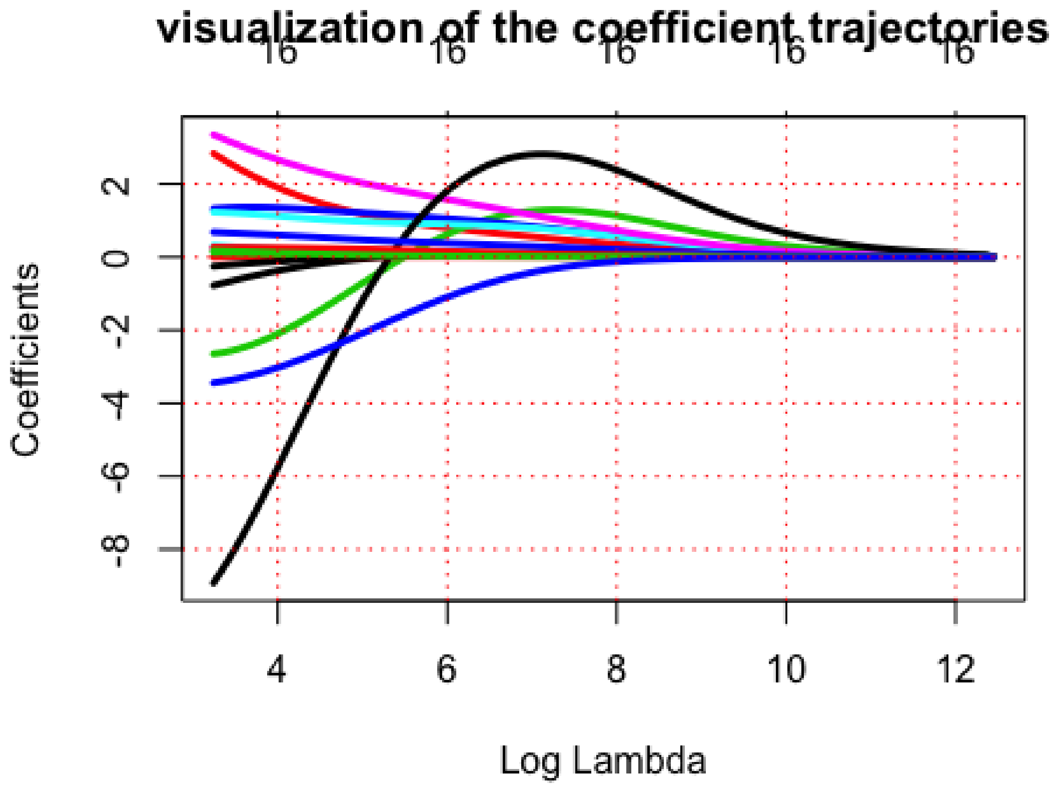
By applying the lasso with cross-validation, we can find the optimal value of the regularization penality is 2.935.

```
cv.lasso_new <- glmnet(x,hitters$Salary,lambda = y,alpha = 1)

coef(cv.lasso_new)

## 18 x 1 sparse Matrix of class "dgCMatrix"
##                     s0
## (Intercept) 75.0316249
## (Intercept)  .
## AtBat       -1.7099436
## Hits         6.0823207
## HmRun        .
## Runs         .
## RBI          0.2602043
## Walks        4.7542167
## Years       -9.0357077
## CAtBat       .
## CHits        .
## CHmRun       0.5745583
## CRuns        0.6973902
## CRBI         0.3033389
## CWalks      -0.4924500
## PutOuts      0.2804673
## Assists      0.1839622
## Errors      -1.7085130
```

By using the lambda to fit, we can get variable “AtBat”, variable “Hits”, variable “RBI”, variable“Walks”, variable “Years”,variable “CHmRun”, variable “CRuns”,variable“CRBI”, variable“CWalks”, variable“PutOuts”, variable“Assists”, and variable“Errors” from the results above. There are 12 non-zero coefficient in total.

1.2:

```
set.seed(1234)

library(MASS)
library ("glmnet")

hitters<- read.csv("/Users/kaimouto/Desktop/ANLY 590/assignments/Hitters.csv"
)
hitters = na.omit(hitters)
x<- model.matrix(Salary~ .-X-League-Division-NewLeague-Salary, data=hitters)

noCV.ridge=glmnet(x,hitters$Salary, alpha=0)

plot(noCV.ridge, xvar = "lambda", lwd = 3,main='visualization of the coeffici
ent trajectories')
grid(col = 2,lty = 3)
```



```
set.seed(1234)

cv.ridge=cv.glmnet(x,hitters$Salary, type.measure = 'mse', alpha=0)

#choose the Lambda

y<- cv.ridge$lambda.min
y

## [1] 28.01718
```

By applying the lasso with cross-validation, we can find the optimal value of the regularization penality is 28.017.

```
cv.ridge_new <- glmnet(x,hitters$Salary,lambda = y,alpha = 0)

coef(cv.ridge_new)

## 18 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept) 27.994797056
## (Intercept)  .
## AtBat       -0.723504487
## Hits         2.704625078
## HmRun       -2.595387303
## Runs         1.351260127
## RBI          1.226757322
## Walks        3.254158746
## Years       -8.485767019
## CAtBat      -0.001692903
## CHits        0.125773172
## CHmRun       0.661490023
## CRuns        0.297444435
## CRBI         0.242310378
## CWalks      -0.228174794
## PutOuts      0.270173394
## Assists      0.169600493
## Errors      -3.413387246
```

By using the lambda to fit, we can see all of the coefficient have non-zero values, which are 16 in total.

2:

In general, variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. On the other hand, bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

As a general rule, bias-variance trade off refers that if we use more flexible methods/models, the variance will increase and the bias will decrease and the model will be more complex with higher dimensions. In this tradeoff, the role of regularization could increase the bias a little bit, but more importantly, it reduces the variance of the model significantly.Overall, it dramatically increases the accuracy of the model. According to 1.1, the model with 12 predictors generating from the lasso regression will have a higher bias but lower variance compared to the model using all the 16 attributes in Hitters. However, compared to the model with the final last three variables, the model with 12 vairables has higher variance but lower bias.