ANLY590AS0

Wupeng Han

2018/9/8

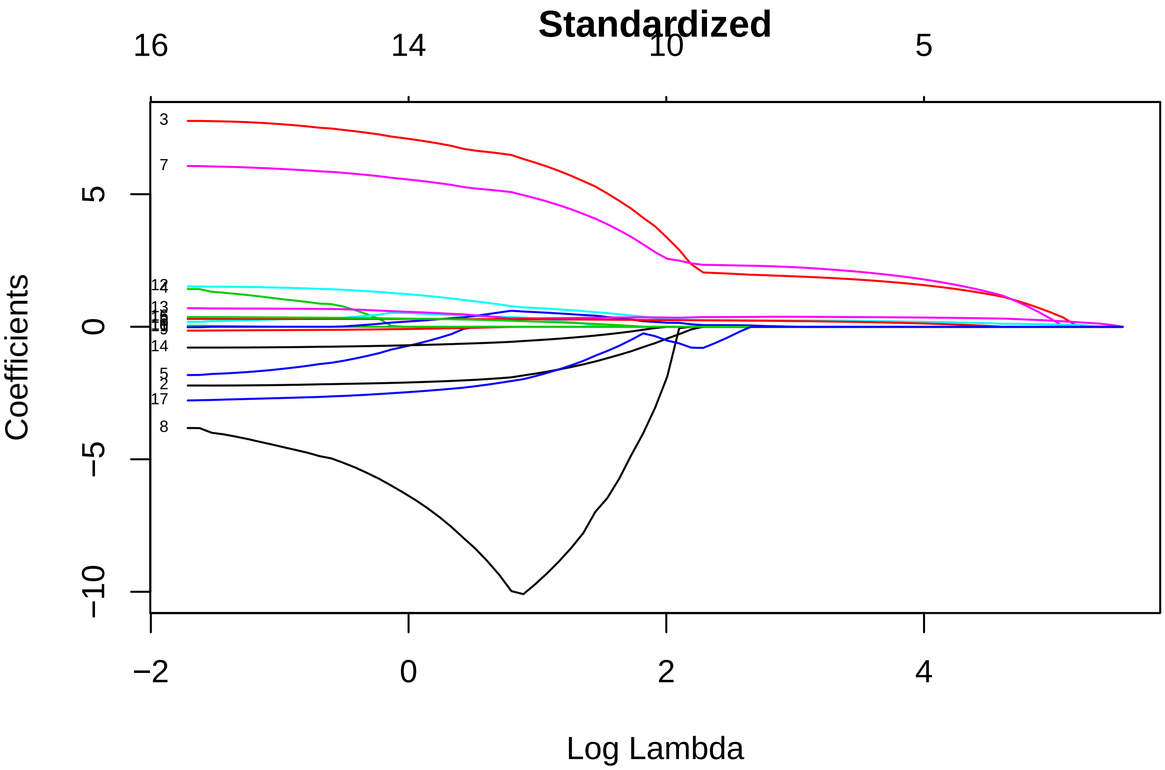
1

1.1

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.4

## Loading required package: Matrix

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 3.4.3

## Loaded glmnet 2.0-16
```

```r
df=read.csv("Hitters.csv")
#select numeric features
df=df[c(2:14,17:20)]
#drop missing value
df<-na.omit(df)

x=model.matrix(Salary~., data= df)
y=df$Salary
fit.lasso<-glmnet(x,y,alpha=1)
#1.1.1
plot(fit.lasso,xvar="lambda",label=TRUE,main="Standardized")
```



```
predict(fit.lasso,s=180, type = "coefficients")
```

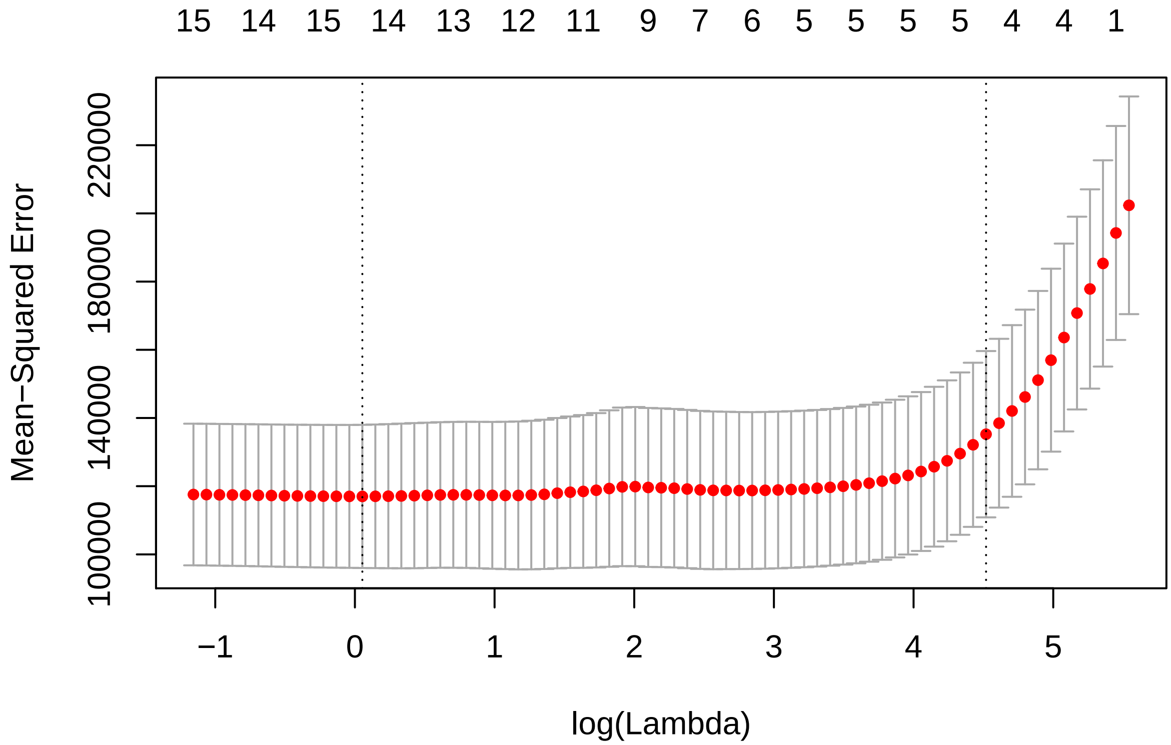
```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept) 450.28669856
## (Intercept)   .
## AtBat         .
## Hits         0.06163890
## HmRun         .
## Runs          .
## RBI           .
## Walks         .
## Years         .
## CAtBat        .
## CHits         .
## CHmRun        .
## CRuns        0.06182082
## CRBI         0.17148502
## CWalks        .
## PutOuts       .
## Assists       .
## Errors        .
```

1.1.2

We can see the final three predictors that remain in the model are “Hits” “CRuns”and “CRBI”, which means those three are the most important features to predict the predict the Hitters’ salary.

1.1.3 & 1.1.4

```
set.seed(123)
cv.lasso<-cv.glmnet(x,y, alpha =1)
plot(cv.lasso)
```



```
#1.1.3
bl<-cv.lasso$lambda.min
print("The optimal value of the regularization penalty:")
```

```
## [1] "The optimal value of the regularization penalty:"
```

```
bl
```

```
## [1] 1.054829
```

```
#1.1.4
predict(fit.lasso,s=bl, type = "coefficients")
```

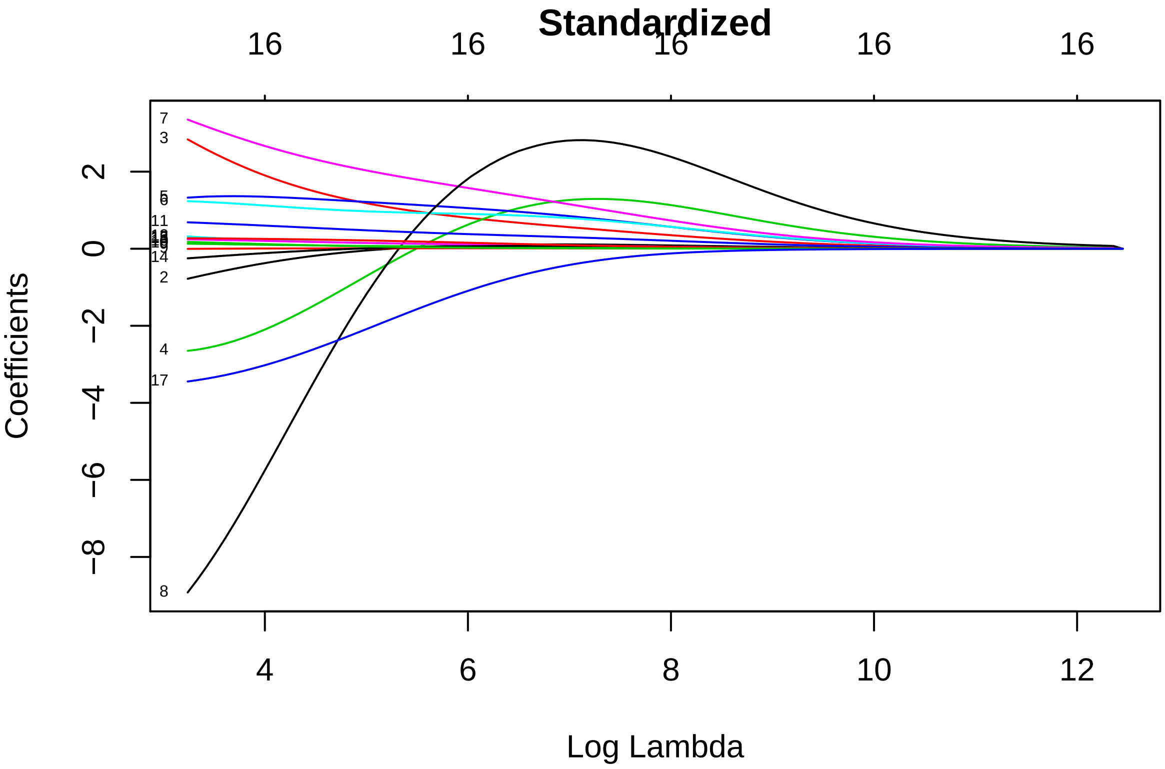
```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) 111.5195886
## (Intercept)   .
## AtBat        -2.0937292
## Hits          7.0542986
## HmRun         .
## Runs         -0.6565960
## RBI           0.5048906
## Walks         5.5274426
## Years        -6.5375570
## CAtBat       -0.0802116
## CHits         .
## CHmRun        0.2155211
## CRuns         1.2046475
## CRBI          0.5528814
## CWalks       -0.6924703
## PutOuts       0.2928267
## Assists       0.3024355
## Errors       -2.4481956
```

There are 13 predictors are left in the model.

1.2

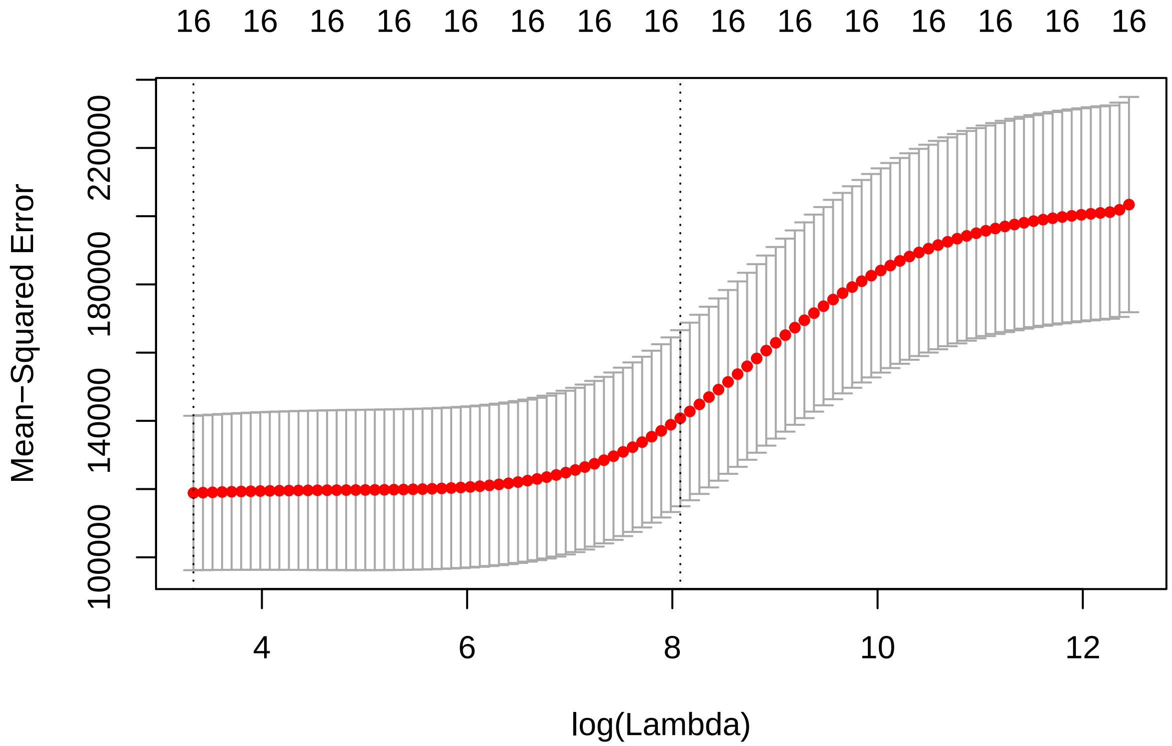
1.2.1

```
fit.ridge=glmnet(x,y,alpha=0)
plot(fit.ridge,xvar="lambda",label=TRUE,main="Standardized")
```



1.2.2

```
set.seed(123)
cv.ridge<-cv.glmnet(x,y, alpha =0)
plot(cv.ridge)
```



```
print("The optimal value of ridge's regularization penalty:")
```

```
## [1] "The optimal value of ridge's regularization penalty:"
```

```
cv.ridge$lambda.min
```

```
## [1] 28.01718
```

2

2.1

The bias-variance tradeoff is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa.

2.2

Regularization methods introduce bias into the regression solution that can reduce variance considerably relative to the ordinary least squares (OLS) solution. Although the OLS solution provides non-biased regression estimates, the lower variance solutions produced by regularization techniques provide superior MSE performance.

2.3

In the ridge and lasso models, we can find the regularization parameter lambda controls the penalty strength.Larger lambda leads to smaller coefficients which means a higher bias.