STA314 Homework 5

student number: 1003942326

Yulin WANG

23/11/2019

Question 1

Since we suppose that:

$$x_{11} = x_{12};\ x_{21} = x_{22};\ x_{11} + x_{21} = 0;\ x_{12} + x_{22} = 0;\ y_1 + y_2 = 0$$

So we have:

$$x_{11} = x_{12} = -x_{21} = -x_{22};\quad y_2 = -y_1$$

(a)

In Ridge Regreesion, We minimize:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 &= (y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&= (y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (-y_1 + \beta_1 x_{11} + \beta_2 x_{11})^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&= 2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(\beta_1^2 + \beta_2^2)
\end{aligned}$$

(b)

Expanding the equation from Part (a) and let it be R:

$$\begin{aligned}
R &= 2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&= 2[y_1^2 + (\beta_1 + \beta_2)^2 x_{11}^2 - 2y_1(\beta_1 + \beta_2)x_{11}] + \lambda(\beta_1^2 + \beta_2^2) \\
&= 2[y_1^2 + \beta_1^2 x_{11}^2 + \beta_2^2 x_{11}^2 + 2\beta_1\beta_2 x_{11}^2 - 2y_1\beta_1 x_{11} - 2y_1\beta_2 x_{11}] + \lambda(\beta_1^2 + \beta_2^2) \\
&= 2y_1^2 + 2\beta_1^2 x_{11}^2 + 2\beta_2^2 x_{11}^2 + 4\beta_1\beta_2 x_{11}^2 - 4y_1\beta_1 x_{11} - 4y_1\beta_2 x_{11} + \lambda\beta_1^2 + \lambda\beta_2^2
\end{aligned}$$

Take partial derivative of R respect to β_1:

$$\begin{aligned}
&\frac{\partial R}{\partial \beta_1} = 4\beta_1 x_{11}^2 + 4\beta_2 x_{11}^2 - 4x_{11}y_1 + 2\lambda\beta_1 \overset{set}{=} 0 \\
&\Rightarrow\ 2\lambda\hat{\beta}_1 = 4x_{11}y_1 - 4\hat{\beta}_1 x_{11}^2 - 4\hat{\beta}_2 x_{11}^2 \\
&\Rightarrow\ \lambda\hat{\beta}_1 = 2x_{11}y_1 - 2(\hat{\beta}_1 + \hat{\beta}_2)x_{11}^2
\end{aligned}$$

Take partial derivative of R respect to β_2:

$$\begin{aligned}
&\frac{\partial R}{\partial \beta_2} = 4\beta_2 x_{11}^2 + 4\beta_1 x_{11}^2 - 4x_{11}y_1 + 2\lambda\beta_2 \overset{set}{=} 0 \\
&\Rightarrow\ 2\lambda\hat{\beta}_2 = 4x_{11}y_1 - 4\hat{\beta}_1 x_{11}^2 - 4\hat{\beta}_2 x_{11}^2 \\
&\Rightarrow\ \lambda\hat{\beta}_2 = 2x_{11}y_1 - 2(\hat{\beta}_1 + \hat{\beta}_2)x_{11}^2
\end{aligned}$$

Thus, the Ridge coefficient estimates satisfy:

$$\lambda\hat{\beta}_1 = \lambda\hat{\beta}_2\ \Rightarrow\ \hat{\beta}_1 = \hat{\beta}_2$$

(c)

In Lasso, We minimize:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| &= (y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(|\beta_1| + |\beta_2|) \\ &= (y_1 - \beta_1 x_{11} - \beta_2 x_{11})^2 + (-y_1 + \beta_1 x_{11} + \beta_2 x_{11})^2 + \lambda(|\beta_1| + |\beta_2|) \\ &= 2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(|\beta_1| + |\beta_2|)\end{aligned}$$

(d)

Expanding the equation from Part (c) and let it be L:

$$\begin{aligned}L &= 2(y_1 - (\beta_1 + \beta_2)x_{11})^2 + \lambda(|\beta_1| + |\beta_2|) \\ &= 2[y_1^2 + (\beta_1 + \beta_2)^2 x_{11}^2 - 2y_1(\beta_1 + \beta_2)x_{11}] + \lambda(|\beta_1| + |\beta_2|) \\ &= 2[y_1^2 + \beta_1^2 x_{11}^2 + \beta_2^2 x_{11}^2 + 2\beta_1\beta_2 x_{11}^2 - 2y_1\beta_1 x_{11} - 2y_1\beta_2 x_{11}] + \lambda(|\beta_1| + |\beta_2|) \\ &= 2y_1^2 + 2\beta_1^2 x_{11}^2 + 2\beta_2^2 x_{11}^2 + 4\beta_1\beta_2 x_{11}^2 - 4y_1\beta_1 x_{11} - 4y_1\beta_2 x_{11} + \lambda|\beta_1| + \lambda|\beta_2|\end{aligned}$$

Take partial derivative of L respect to β_1:

$$\begin{aligned}&\frac{\partial L}{\partial \beta_1} = 4\beta_1 x_{11}^2 + 4\beta_2 x_{11}^2 - 4x_{11}y_1 + \lambda\frac{\partial|\beta_1|}{\partial \beta_1} \overset{set}{=} 0 \\ &\Rightarrow\ 4\hat{\beta}_1 x_{11}^2 + 4\hat{\beta}_2 x_{11}^2 + \lambda\frac{\partial|\hat{\beta}_1|}{\partial \hat{\beta}_1} = 4x_{11}y_1 \\ &\Rightarrow\ \lambda\frac{\partial|\hat{\beta}_1|}{\partial \hat{\beta}_1} = 4x_{11}y_1 - 4(\hat{\beta}_1 + \hat{\beta}_2)x_{11}^2\end{aligned}$$

Take partial derivative of L respect to β_2:

$$\begin{aligned}&\frac{\partial L}{\partial \beta_2} = 4\beta_2 x_{11}^2 + 4\beta_1 x_{11}^2 - 4x_{11}y_1 + \lambda\frac{\partial|\beta_2|}{\partial \beta_2} \overset{set}{=} 0 \\ &\Rightarrow\ 4\hat{\beta}_1 x_{11}^2 + 4\hat{\beta}_2 x_{11}^2 + \lambda\frac{\partial|\hat{\beta}_2|}{\partial \hat{\beta}_2} = 4x_{11}y_1 \\ &\Rightarrow\ \lambda\frac{\partial|\hat{\beta}_2|}{\partial \hat{\beta}_2} = 4x_{11}y_1 - 4(\hat{\beta}_1 + \hat{\beta}_2)x_{11}^2\end{aligned}$$

Thus, the Lasso coefficient estimates satisfy:

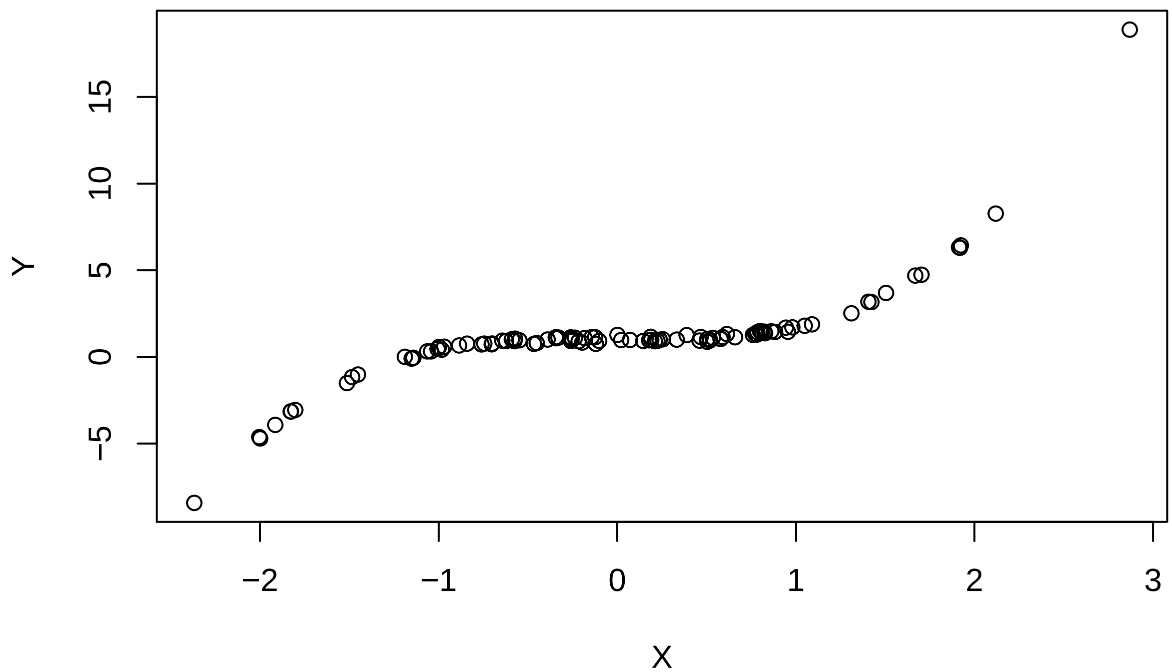
$$\lambda\frac{\partial|\hat{\beta}_1|}{\partial \hat{\beta}_1} = \lambda\frac{\partial|\hat{\beta}_2|}{\partial \hat{\beta}_2} \Rightarrow \frac{\partial|\hat{\beta}_1|}{\partial \hat{\beta}_1} = \frac{\partial|\hat{\beta}_2|}{\partial \hat{\beta}_2}$$

So it shows that the Lasso just requires that β_1 and β_2 are both positive or both negative (ignoring possibility of 0). Thus, there are many possible solutions to the optimization problem for the lasso coefficients.

Question 2

(a)

```
set.seed(19)
X <- rnorm(100)
eps <- 0.1*rnorm(100)
Y <- 1 - 0.1*X + 0.05*X^2 + 0.75*X^3 + eps
plot(X,Y)
```



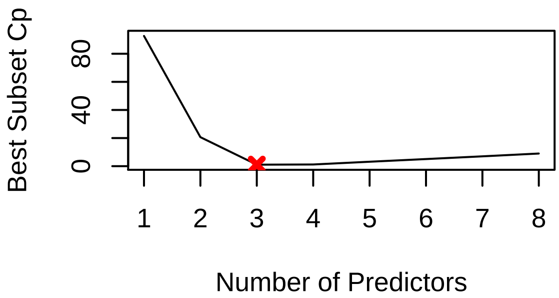
(b)

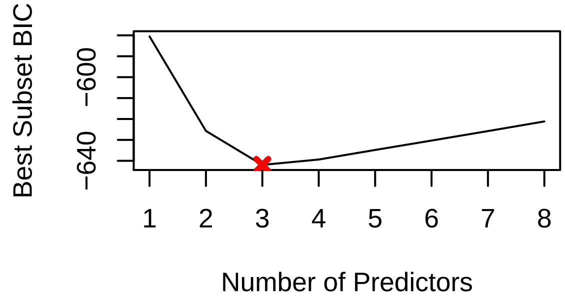
```
#use best subset selection
library(leaps)
df <- data.frame(Y,X,X2=X^2,X3=X^3,X4=X^4,X5=X^5,X6=X^6,X7=X^7,X8=X^8) # or poly(X,8,raw=T)
full_model <- regsubsets(Y~X+X2+X3+X4+X5+X6+X7+X8, data=df, nvmax=8)
full_sum <- summary(full_model)
```

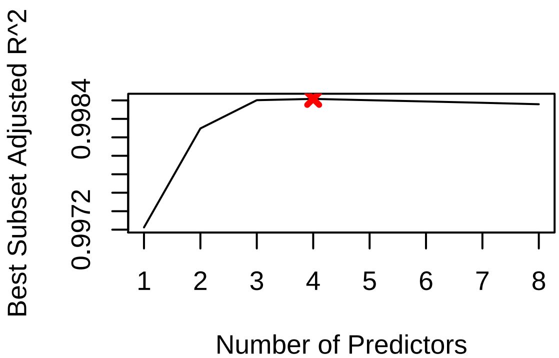
(i)

```
par(mfrow=c(2,2))
#measure Cp
min.cp <- which.min(full_sum$cp)
plot(1:8, full_sum$cp, xlab="Number of Predictors", ylab="Best Subset Cp", type="l")
points(min.cp, full_sum$cp[min.cp], col="red", pch=4, lwd=3)
#measure BIC
min.bic <- which.min(full_sum$bic)
plot(1:8, full_sum$bic, xlab="Number of Predictors", ylab="Best Subset BIC", type="l")
points(min.bic, full_sum$bic[min.bic], col="red", pch=4, lwd=3)
#measure adjusted R^2
max.adjr2 <- which.max(full_sum$adjr2)
```

```r
plot(1:8, full_sum$adjr2, xlab="Number of Predictors", ylab="Best Subset Adjusted R^2", type="l")
points(max.adjr2, full_sum$adjr2[max.adjr2], col="red", pch=4, lwd=3)
```







(ii)

```r
#best model coefficients obtained from Cp
coef(full_model, min.cp)
```

```
## (Intercept)           X          X2          X3
##  1.00661552 -0.08393678  0.05769135  0.74969476
```

```r
#best model coefficients obtained from BIC
coef(full_model, min.bic)
```

```
## (Intercept)           X          X2          X3
##  1.00661552 -0.08393678  0.05769135  0.74969476
```

```r
#best model coefficients obtained from adjusted R^2
coef(full_model, max.adjr2)
```

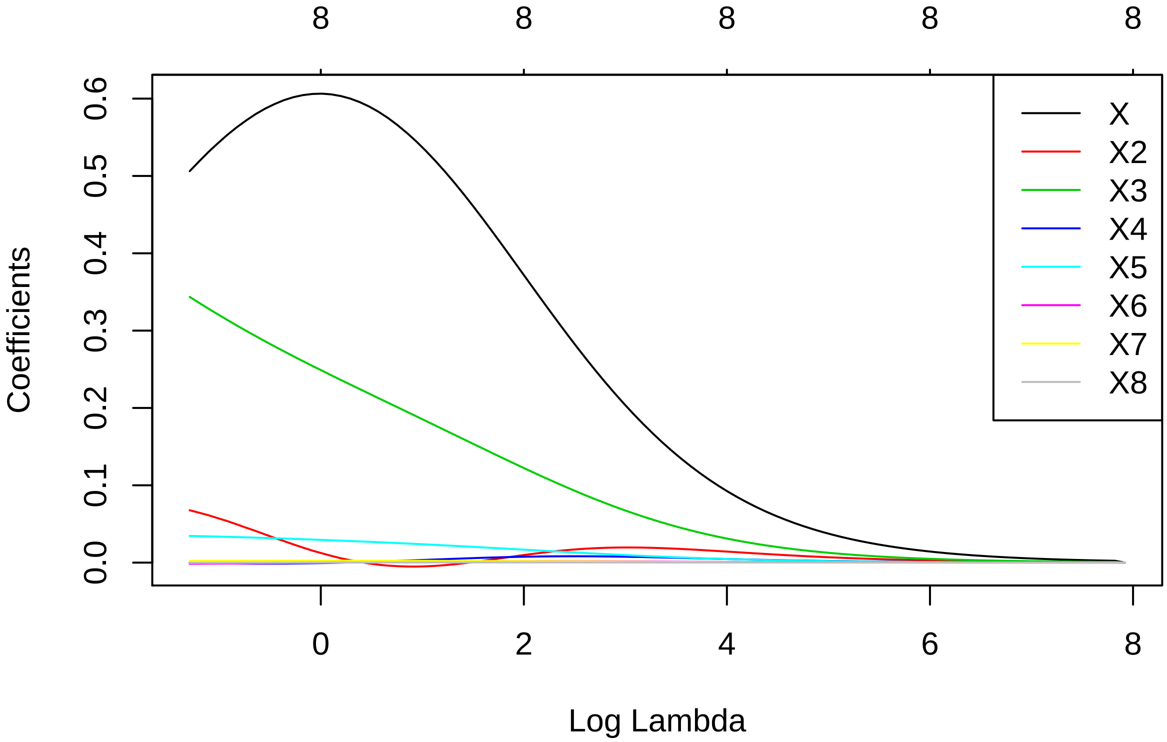
```
##  (Intercept)            X           X2           X3           X5
##  1.003922288 -0.108741496  0.061855586  0.769659485 -0.002612306
```

(c)

```r
#fit the ridge model
library(glmnet)
x_matrix <- as.matrix(df[,-1]) #without Y
ridge_model <- glmnet(x_matrix, Y, alpha = 0, nlambda = 100)
```

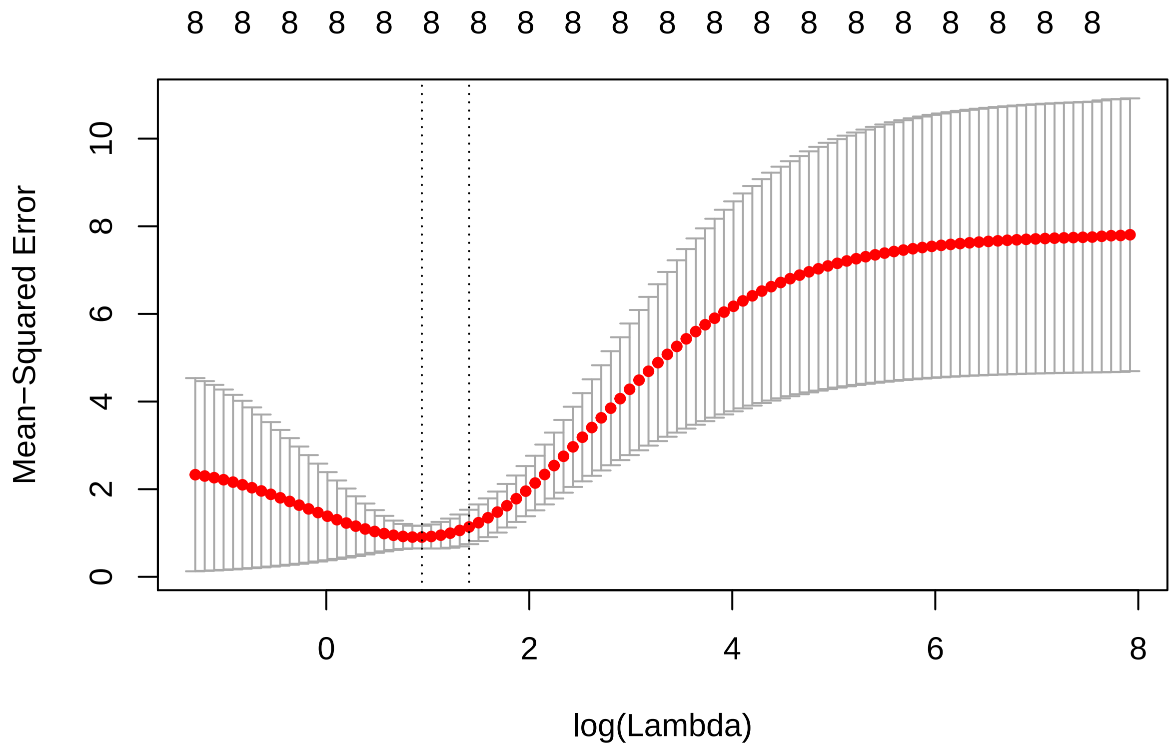
(i)

```
par(mfrow=c(1,1))
plot(ridge_model, xvar = "lambda", col = 1:8)
legend("topright", col = 1:8, legend = row.names(ridge_model$beta), lty = 1)
```



(ii)

```
set.seed(20)
ridge_model_cv <- cv.glmnet(x_matrix, Y, alpha = 0)
par(mfrow=c(1,1))
plot(ridge_model_cv)
```




```r
ridge_lambda_min <- ridge_model_cv$lambda.min
ridge_lambda_min #optimal value of lambda
```

```
## [1] 2.563188
```

```r
log(ridge_lambda_min) #optimal value of log(lambda)
```

```
## [1] 0.9412518
```

(iii)

```r
predict(ridge_model_cv, s=ridge_lambda_min, type="coefficients")
```

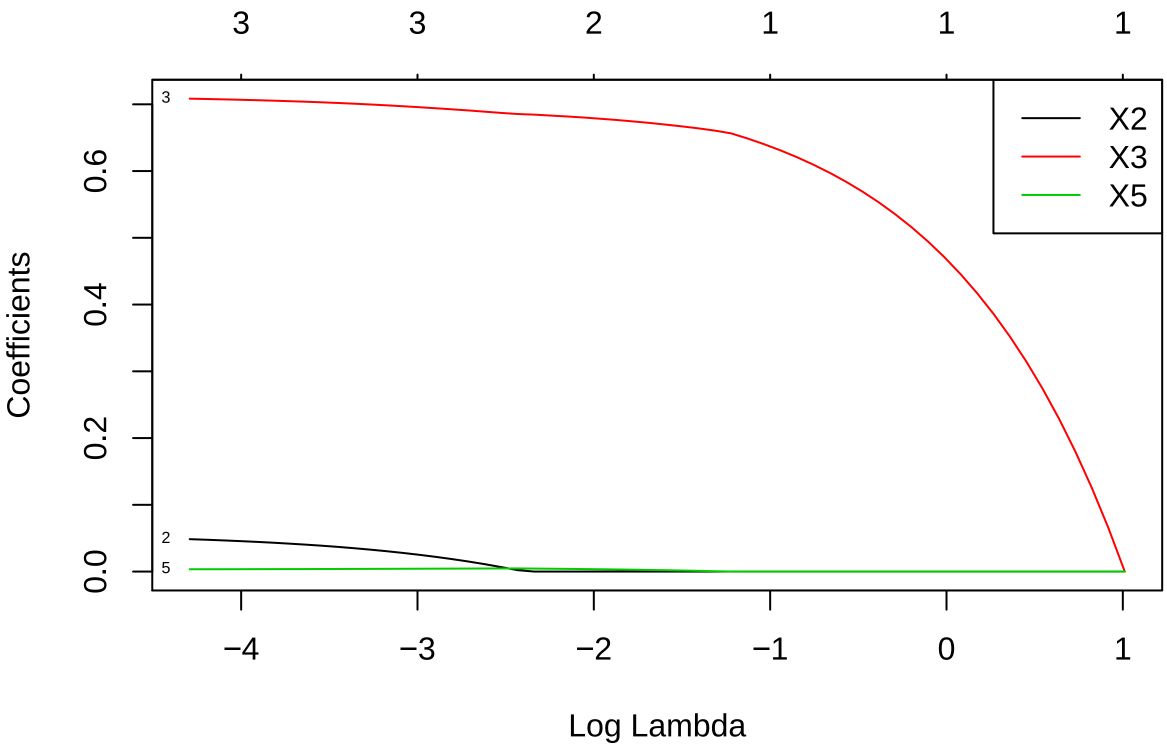
```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  1.0123824359
## X            0.5445251878
## X2          -0.0050818372
## X3           0.1891529314
## X4           0.0031130229
## X5           0.0241446550
## X6           0.0011440555
## X7           0.0023368089
## X8           0.0002151474
```

(d)

```r
#fit the lasso model
lasso_model <- glmnet(x_matrix, Y, alpha = 1, nlambda = 100)
```

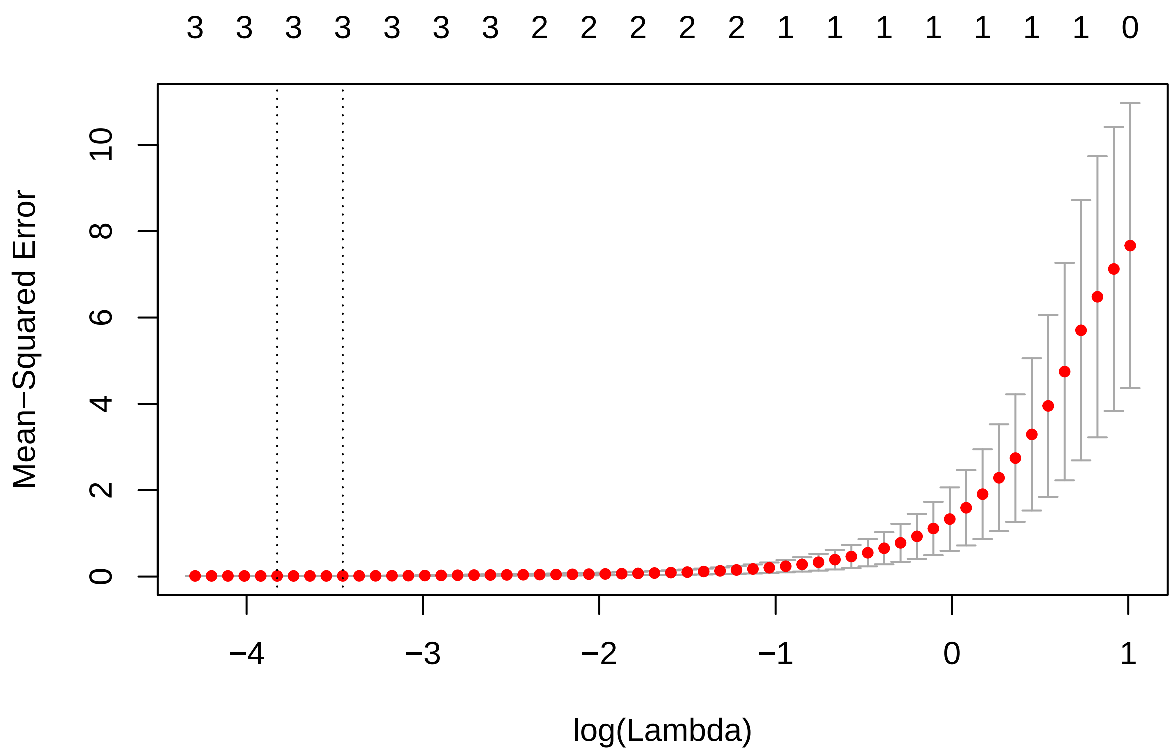
(i)

```
par(mfrow=c(1,1))
plot(lasso_model, xvar = "lambda", label = T)
legend("topright", col = 1:3, legend = c("X2", "X3", "X5"), lty = 1)
```



(ii)

```
set.seed(21)
lasso_model_cv <- cv.glmnet(x_matrix, Y, alpha = 1)
par(mfrow=c(1,1))
plot(lasso_model_cv)
```




```r
lasso_lambda_min <- lasso_model_cv$lambda.min
lasso_lambda_min #optimal value of lambda
```

```
## [1] 0.02178078
```

```r
log(lasso_lambda_min) #optimal value of log(lambda)
```

```
## [1] -3.826727
```

(iii)

```r
predict(lasso_model_cv, s=lasso_lambda_min, type="coefficients")
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) 1.020023170
## X           .
## X2          0.043321050
## X3          0.705545523
## X4          .
## X5          0.003649058
## X6          .
## X7          .
## X8          .
```