Statistical Learning and Data mining

Homework 10

M052040003 鍾冠毅

5.a.

In general, $min \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_j)^2 + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2$. In this case, $\hat{\beta}_0 = 0$ and $n = p = 2$, $min\{(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)\}$.

5.b.

Given $x_{11} = x_{12}$ and $x_{21} = x_{22}$, let $S(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{2}(y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2$ Let $\frac{\partial S}{\partial \hat{\beta}_1} = \frac{\partial S}{\partial \hat{\beta}_2} = 0$ then $\hat{\beta}_1 = \hat{\beta}_2$

5.c.

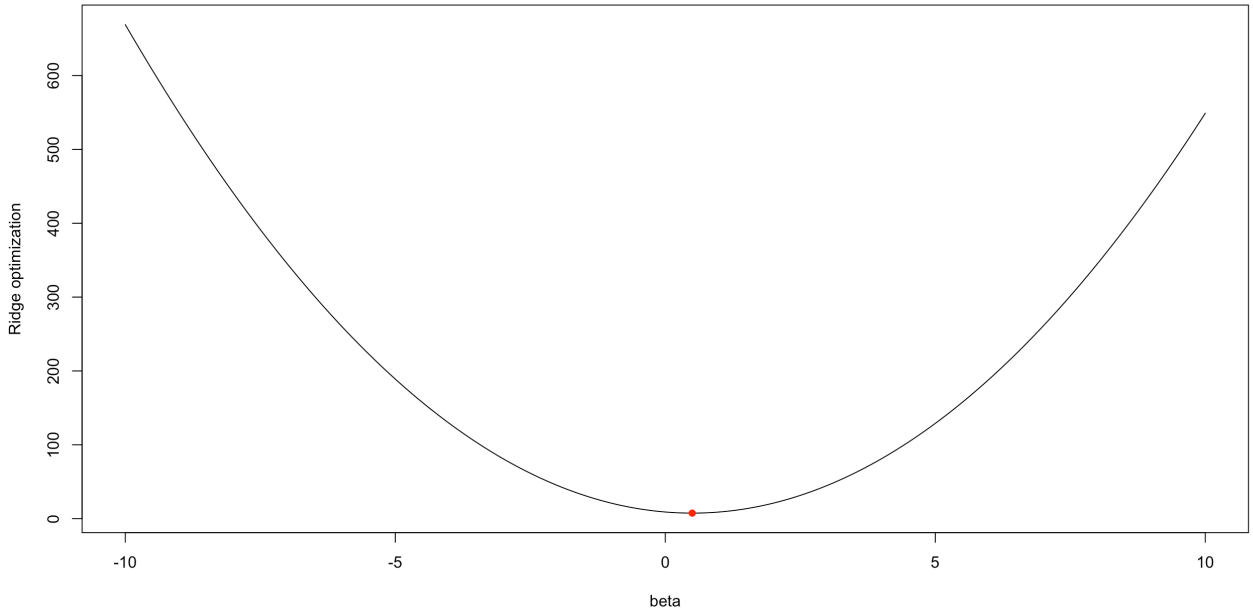
In this case, $\hat{beta}_0 = 0$ and $n = p = 2$, $min\{(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)\}$

5.d.

Under the condition that $|\hat{\beta}_1| + |\hat{\beta}_2| < s$, the restriction area can be drawn as a diamond centered at $(0, 0)$. Consider the squared optimization constraint $(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2$ and known conditions $x_{11} = x_{12}$, $x_{21} = x_{22}$, $x_{11} + x_{21} = 0$, $x_{12} + x_{22} = 0$ and $y_1 + y_2 = 0$. We simplify it to $min\{2(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11})^2\}$ and conclude that $\hat{\beta}_1 + \hat{\beta}_2 = \frac{y_1}{x_{11}}$ which is a line parallel to $\hat{\beta}_1 + \hat{\beta}_2 = s$. By the method of linear programming we learned in high school mathematics, $\hat{\beta}_1 + \hat{\beta}_2 = s$ is a potential solution to Lasso optimization problem. Similarly, $\hat{\beta}_1 + \hat{\beta}_2 = -s$ can be discussed. Thus, the general form of solution is given by two line segment: $\hat{\beta}_1 + \hat{\beta}_2 = s; \hat{\beta}_1 \geq 0;\ \hat{\beta}_2 \geq 0$ and $\hat{\beta}_1 + \hat{\beta}_2 = -s; \hat{\beta}_1 \leq 0;\ \hat{\beta}_2 \leq 0$.

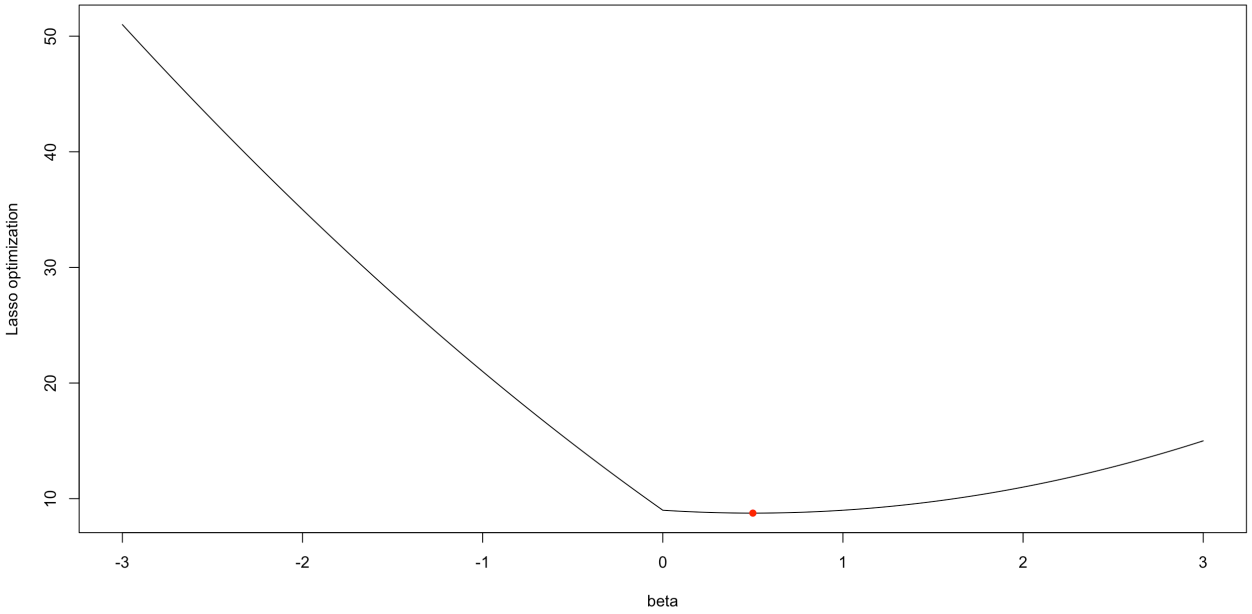
6.a.

Taking the form $(y-\beta)^2+\lambda\beta^2$, it minimizes at $\beta = y/(1+\lambda)$, where $y = 3$ and $\lambda = 5$.



6.b.

Taking the form $(y-\beta)^2+\lambda|\beta|$, it minimizes at $\beta = y-\lambda/2$, where $y = 3$ and $\lambda = 5$.



7.a.

The likelihood for the data is:

$$L(\theta \mid \beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2\right)$$

7.b.

The posterior with double exponential (Laplace Distribution) with mean 0 and common scale parameter b, i.e.$p(\beta) = \frac{1}{2b}\exp(-|\beta|/b)$ is:

$$f(\beta \mid X, Y) \propto f(Y \mid X, \beta)p(\beta \mid X) = f(Y \mid X, \beta)p(\beta)$$

$$f(Y \mid X, \beta)p(\beta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2\right)\left(\frac{1}{2b}\exp(-|\beta|/b)\right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{|\beta|}{b}\right)$$

7.c.

Showing that the Lasso estimate for β is the mode under this posterior distribution is the same thing as showing that the most likely value for β is given by the lasso solution with a certain λ. We can do this by taking our likelihood and posterior and showing that it can be reduced to the canonical Lasso Equation from the book. Let's start by simplifying it by taking the logarithm of both sides:

$$\log f(Y \mid X, \beta)p(\beta)$$

$$= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{|\beta|}{b}\right)\right]$$

$$= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{|\beta|}{b}\right)$$

We want to maximize the posterior, this means:

$$\arg\max_{\beta} \; f(\beta \mid X, Y)$$

$$\propto \arg\max_{\beta} \; \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n}\left(\frac{1}{2b}\right)\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{|\beta|}{b}\right)$$

Since we are taking the difference of two values, the maximum of this value is the equivalent to taking the difference of the second value in terms of β. This results in:

$$= \arg\min_{\beta} \; \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{|\beta|}{b}$$

$$= \arg\min_{\beta} \; \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{1}{b}\sum_{j=1}^{p}|\beta_j|$$

$$= \arg\min_{\beta} \; \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{2\sigma^2}{b}\sum_{j=1}^{p}|\beta_j|\right)$$

By letting $\lambda = 2\sigma^2/b$, we can see that we end up with:

$$= \arg\min_{\beta} \; \sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$= \arg\min_{\beta} \; \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

That is what we know is the Lasso from Equation 6.7 in the book. Thus we know that when the posterior comes from a Laplace distribution with mean zero and common scale parameter b, the mode for β is given by the Lasso solution when $\lambda = 2\sigma^2/b$.

7.d.

The posterior distributed according to Normal distribution with mean 0 and variance c is:

$$f(\beta \mid X, Y) \propto f(Y \mid X, \beta)p(\beta \mid X) = f(Y \mid X, \beta)p(\beta)$$

Our probability distribution function then becomes: $p(\beta) = \prod_{i=1}^{p} p(\beta_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2c\pi}} \exp\left(-\frac{\beta_i^2}{2c}\right) = \left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)$. Substituting our values from (a) and our density function gives us:

$$\begin{aligned}
&f(Y \mid X, \beta)p(\beta) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2\right)\left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)
\end{aligned}$$

7.e.

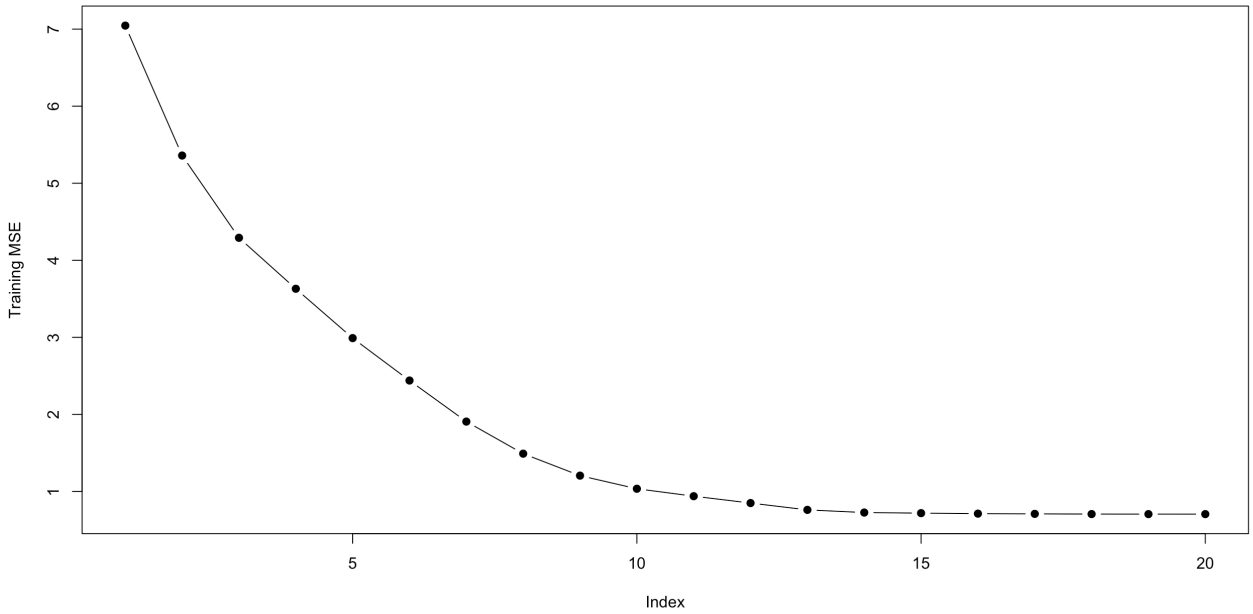
Like from part c, showing that the Ridge Regression estimate for β is the mode and mean under this posterior distribution is the same thing as showing that the most likely value for β is given by the lasso solution with a certain λ. We can do this by taking our likelihood and posterior and showing that it can be reduced to the canonical Ridge Regression Equation 6.5 from the book.We can take the logarithm of both sides to simplify it:

$$\begin{aligned}
&\log f(Y \mid X, \beta)p(\beta) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \\
&= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)
\end{aligned}$$

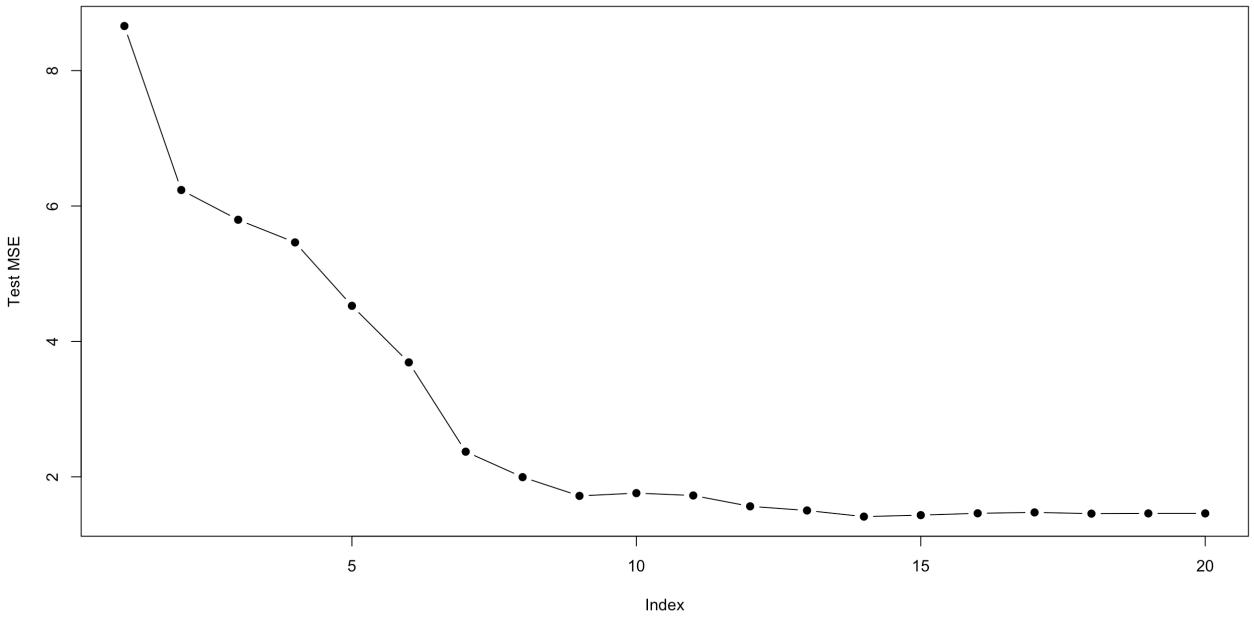
10.a. See the appendix.

10.b. See the appendix.

10.c.



10.d.



10.e.

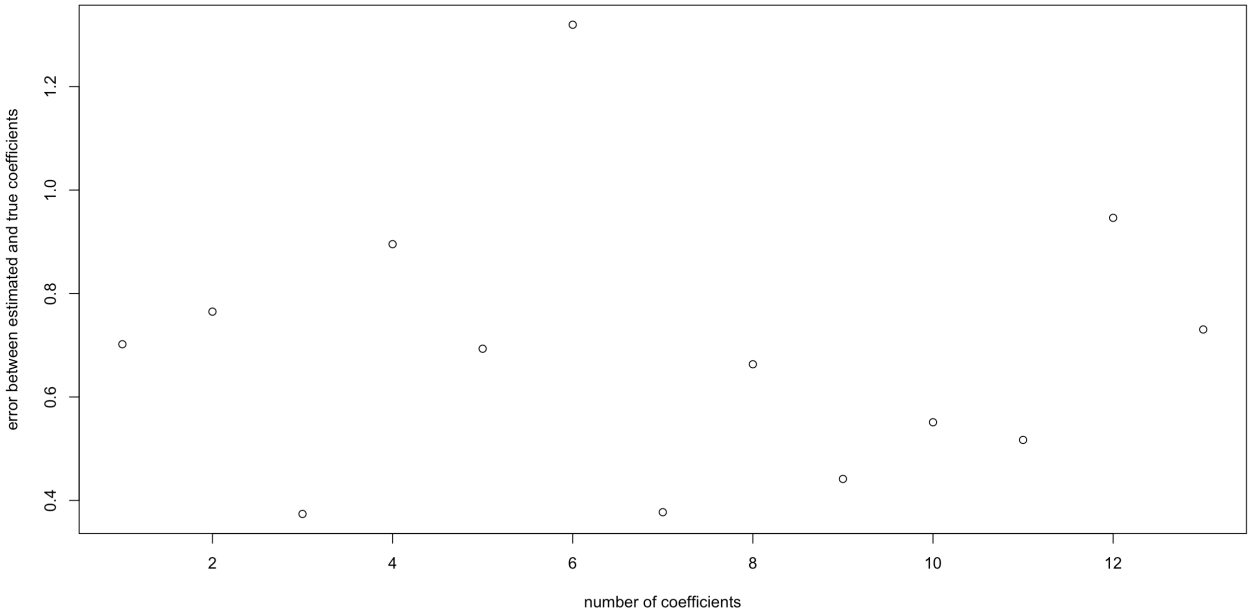
The model with 15 parameters including the intercept reaches the smallest test MSE.

10.f.

```
> coef(regfit.full, id=14)
(Intercept)          x.1          x.2          x.5          x.7          x.8
  0.1630514    0.3707851    0.3176740    1.0424379   -1.2895997    0.8308835
        x.11         x.12         x.13         x.14         x.15         x.16
  0.6919104    0.5638802   -0.3641320   -0.8346409   -0.5667810   -0.1959694
        x.17         x.18         x.20
  0.3128194    1.5567459   -0.7831598
```

Predictors x3, x4, x9, x10 and x19 are recognized that those parameters are 0. Yet, the parameter of x6 is misclassified as 0.

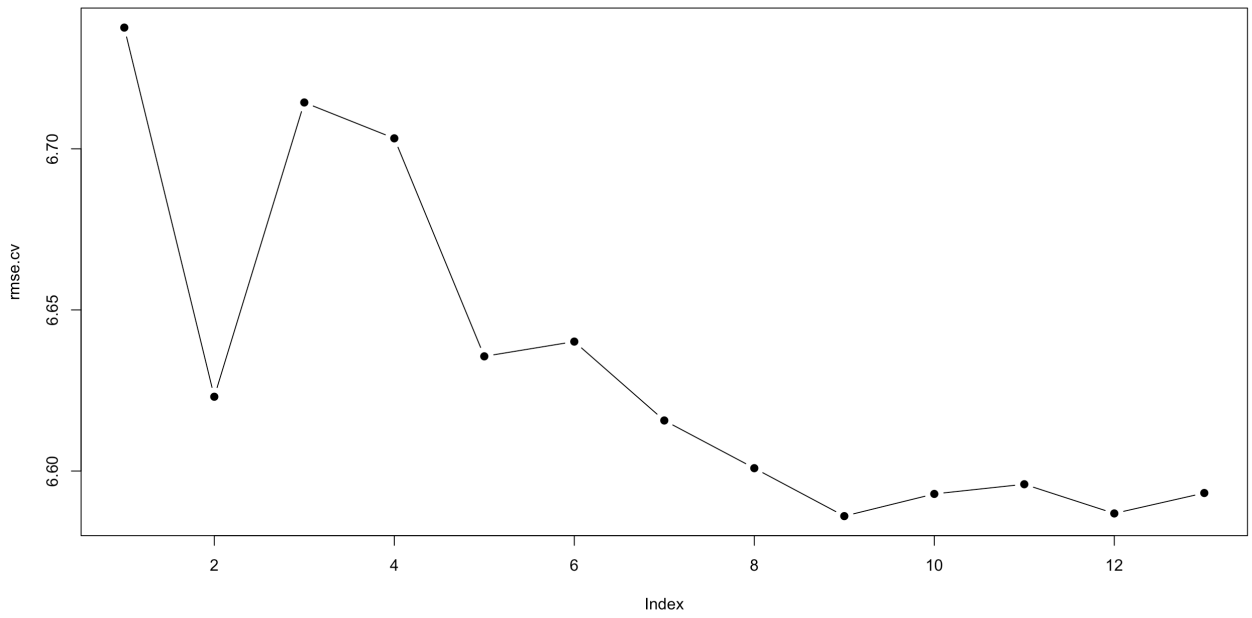
10.g.



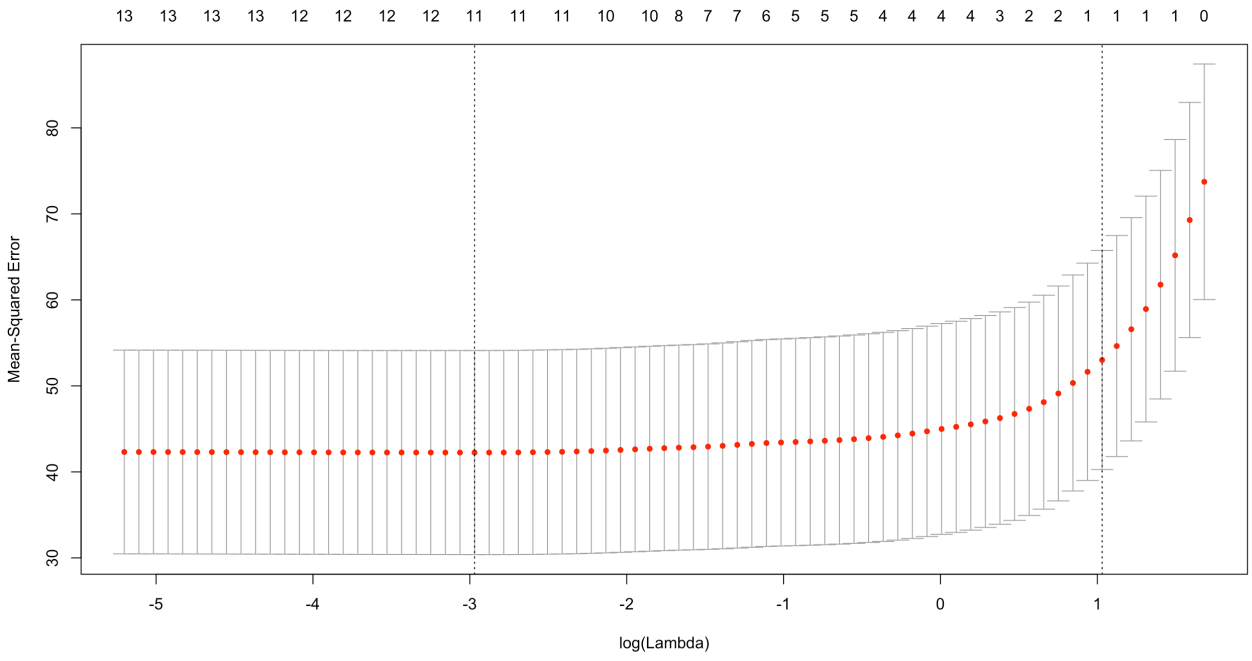
The model with 4 parameters including the intercept minimizes the error between the estimated and true parameter. Test error reaches the minimum with the model containing 15 parameters including the intercept. Thus, the better fit of true coefficients as measured here doesn't mean the model will have a lower test MSE.

11.a.

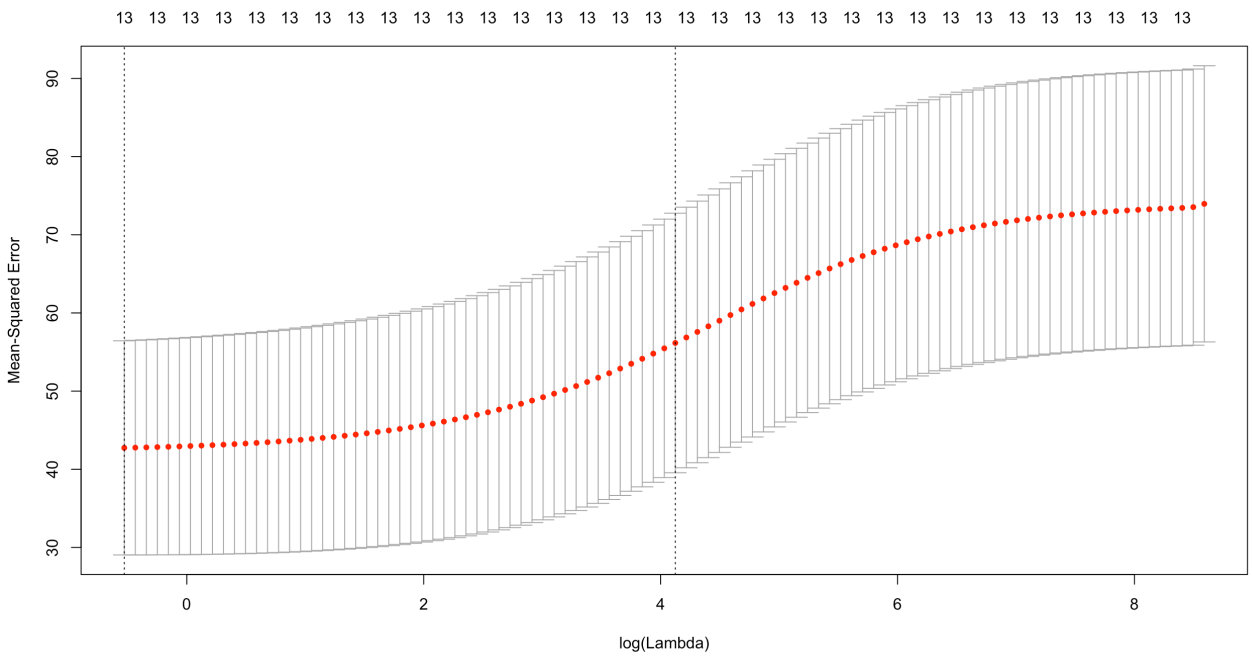
```
> # Best subset selection #
> predict.regsubsets <- function(object,newdata,id,...){
+   form <- as.formula(object$call[[2]])
+   mat <- model.matrix(form,newdata)
+   coefi <- coef(object,id=id)
+   mat[,names(coefi)] %*% coefi
+ }
>
> k <- 10
> p <- ncol(Boston)-1
> folds <- sample(rep(1:k,length=nrow(Boston)))
> cv.errors <- matrix(NA,k,p)
> for(i in 1:k)
+ {
+   best.fit <- regsubsets(crim ~ ., data = Boston[folds!=i, ], nvmax = p)
+   for(j in 1:p)
+   {
+     pred <- predict(best.fit,Boston[folds==i,],id=j)
+     cv.errors[i,j] <- mean((Boston$crim[folds==i]-pred)^2)
+   }
+ }
> rmse.cv <- sqrt(apply(cv.errors,2,mean))
> plot(rmse.cv, pch=19, type="b")
>
> which.min(rmse.cv)
[1] 9
> rmse.cv[which.min(rmse.cv)]
[1] 6.586008
```



```
> #Lasso#
> x <- model.matrix(crim~.-1,data=Boston)
> y <- Boston$crim
> cv.lasso <- cv.glmnet(x,y,type.measure="mse")
> plot(cv.lasso)
>
> coef(cv.lasso)
14 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 0.7894616
zn          .
indus       .
chas        .
nox         .
rm          .
age         .
dis         .
rad         0.2957317
tax         .
ptratio     .
black       .
lstat       .
medv        .
>
> sqrt(cv.lasso$cvm[cv.lasso$lambda == cv.lasso$lambda.1se])
[1] 7.281082
```



```
> #Ridge regression#
> x <- model.matrix(crim~.-1,data=Boston)
> y <- Boston$crim
> cv.ridge <- cv.glmnet(x, y, type.measure="mse", alpha=0)
> plot(cv.ridge)
>
> coef(cv.ridge)
14 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept)  1.146730398
zn          -0.002889389
indus        0.033208330
chas        -0.209288622
nox          2.158437385
rm          -0.158326514
age          0.007072375
dis         -0.109819199
rad          0.056100087
tax          0.002508948
ptratio      0.082673533
black       -0.003147919
lstat        0.042243863
medv        -0.027678989
>
> sqrt(cv.ridge$cvm[cv.ridge$lambda == cv.ridge$lambda.1se])
[1] 7.49396
```



```
> #PCR#
> library(pls)
> pcr.fit <- pcr(crim~., data = Boston, scale = TRUE, validation = "CV")
> summary(pcr.fit)
Data:   X dimension: 506 13
        Y dimension: 506 1
Fit method: svdpc
Number of components considered: 13

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV            8.61    7.204    7.197    6.763    6.744    6.746    6.756
adjCV         8.61    7.201    7.194    6.757    6.738    6.742    6.752
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       6.762    6.625    6.654     6.650     6.652     6.609     6.537
adjCV    6.757    6.619    6.647     6.643     6.643     6.599     6.527

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X       47.70    60.36    69.67    76.45    82.99    88.00    91.14    93.45
crim    30.69    30.87    39.27    39.61    39.61    39.86    40.14    42.47
      9 comps  10 comps  11 comps  12 comps  13 comps
X       95.40     97.04     98.46     99.52     100.0
crim    42.55     42.78     43.04     44.13      45.4
```

The cross-validation MSE of the best subset selection, Lasso and Ridge regression are 6.593, 7.424 and 7.606, respectively. In the PCR method, the model with 13 components reaches its minimum CV/ adjust CV MSE.

11.b.

I propose the PCR model with 13 components for its smallest CV/ adjust CV MSE.

11.c.

The best subset with 9 parameters would be chosen for its more simple model and a “not bad” performance in the CV MSE.