Homework 3

Due Date 10/25/2018

Question 3

a. Training RSS: Steadily Increase because if s if large enough then it will fall into the budget and will be equivalent to the least squares solutions

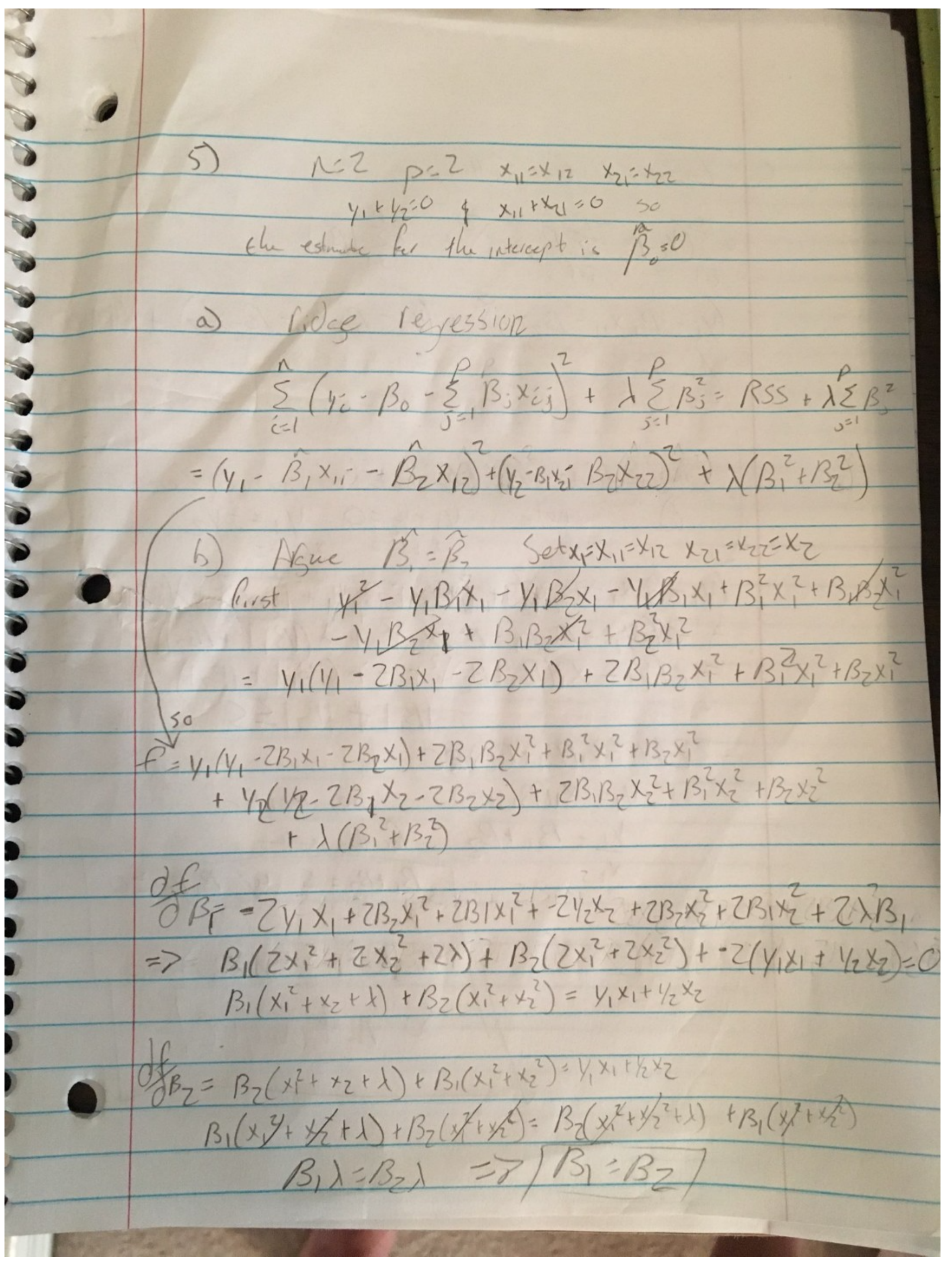
b. Test RSS: Decrease Initially and then eventually start increasing in a U shape because when s is small it is very restrictive or inflexible and so as It gets big there is a decrease in RSS until it hits a point and starts to increase again due to its flexibility.

c. Variance: Steadily Increase due to the relationship with s. As s increase it reduces the restrictive nature on the coefficients and allow for more flexibility and so the with that increase in flexibility so does the variance

d. Bias: Will steadily decrease because the more flexible the model the less bias.

e. Irreducible Error: This will remain constant because it doesnt change with respect to these mentioned variables.

Q2. Chapter 6 Exercise 5



5) $n=2$ $p=2$ $x_{11}=x_{12}$ $x_{21}=x_{22}$

$y_1+y_2=0$ & $x_{11}+x_{21}=0$ so

the estimate for the intercept is $\hat{\beta}_0=0$

a) Ridge regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2$$

$$=(y_1-\hat{\beta}_1x_{11}-\hat{\beta}_2x_{12})^2+(y_2-\beta_1x_{21}-\beta_2x_{22})^2+\lambda(\beta_1^2+\beta_2^2)$$

b) Argue $\hat{\beta}_1=\hat{\beta}_2$ Set $x_1=x_{11}=x_{12}$ $x_{21}=x_{22}=x_2$

first $y_1^2-y_1\beta_1x_1-y_1\beta_2x_1-y_1\beta_1x_1+\beta_1^2x_1^2+\beta_1\beta_2x_1^2$

$-y_1\beta_2x_1+\beta_1\beta_2x_1^2+\beta_2^2x_1^2$

$=y_1(y_1-2\beta_1x_1-2\beta_2x_1)+2\beta_1\beta_2x_1^2+\beta_1^2x_1^2+\beta_2x_1^2$

so

$f=y_1(y_1-2\beta_1x_1-2\beta_2x_1)+2\beta_1\beta_2x_1^2+\beta_1^2x_1^2+\beta_2x_1^2$

$+y_2(y_2-2\beta_1x_2-2\beta_2x_2)+2\beta_1\beta_2x_2^2+\beta_1^2x_2^2+\beta_2x_2^2$

$+\lambda(\beta_1^2+\beta_2^2)$

$\frac{\partial f}{\partial\beta_1}=-2y_1x_1+2\beta_2x_1^2+2\beta_1x_1^2+-2y_2x_2+2\beta_2x_2^2+2\beta_1x_2^2+2\lambda\beta_1$

$\Rightarrow\beta_1(2x_1^2+2x_2^2+2\lambda)+\beta_2(2x_1^2+2x_2^2)+-2(y_1x_1+y_2x_2)=0$

$\beta_1(x_1^2+x_2+\lambda)+\beta_2(x_1^2+x_2^2)=y_1x_1+y_2x_2$

$\frac{\partial f}{\partial\beta_2}=\beta_2(x_1^2+x_2+\lambda)+\beta_1(x_1^2+x_2^2)=y_1x_1+y_2x_2$

$\beta_1(x_1^2+x_2^2+\lambda)+\beta_2(x_1^2+x_2^2)=\beta_2(x_1^2+x_2^2+\lambda)+\beta_1(x_1^2+x_2^2)$

$\beta_1\lambda=\beta_2\lambda\Rightarrow\boxed{\beta_1=\beta_2}$

c) Lasso in this setting

$$0 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(|\beta_1| + |\beta_2|)$$

Argue that β_1 & β_2 are unique in Lasso

d) note $y_1 + y_2 = 0$ $y_1 = -y_2$

$x_1 + x_2 = 0$ $x_1 = -x_2$

$$(y_1 - \beta_1 x_1 - \beta_2 x_1)^2 + (-y_1 + \beta_1 x_1 + \beta_2 x_1)^2 + \lambda(|\beta_1| + |\beta_2|)$$

$$= 2(y_1 - \beta_1 x_1 - \beta_2 x_1)^2 + \lambda(|\beta_1| + |\beta_2|)$$

$$= 2(y_1 - x_1(\beta_1 + \beta_2))^2 = -\lambda(|\beta_1| + |\beta_2|)$$

$$|\beta_1| + |\beta_2| \le s$$

$$2(y_1 - x_1(\beta_1 + \beta_2)) \ge 0$$

$y_1 = x_1(\beta_1 + \beta_2)$ So in order

$\frac{y_1}{x_1} = \beta_1 + \beta_2$ for this setting

$\beta_1 + \beta_2 = s$ w/ $\beta_1, \beta_2 \ge 0$

&

$\beta_1 + \beta_2 = -s$ w/ $\beta_1, \beta_2 \le 0$

So not unique!

Q3. Chapter 6 Exercise 8

#part a

library (leaps)

set.seed(333)

X <- rnorm(100)

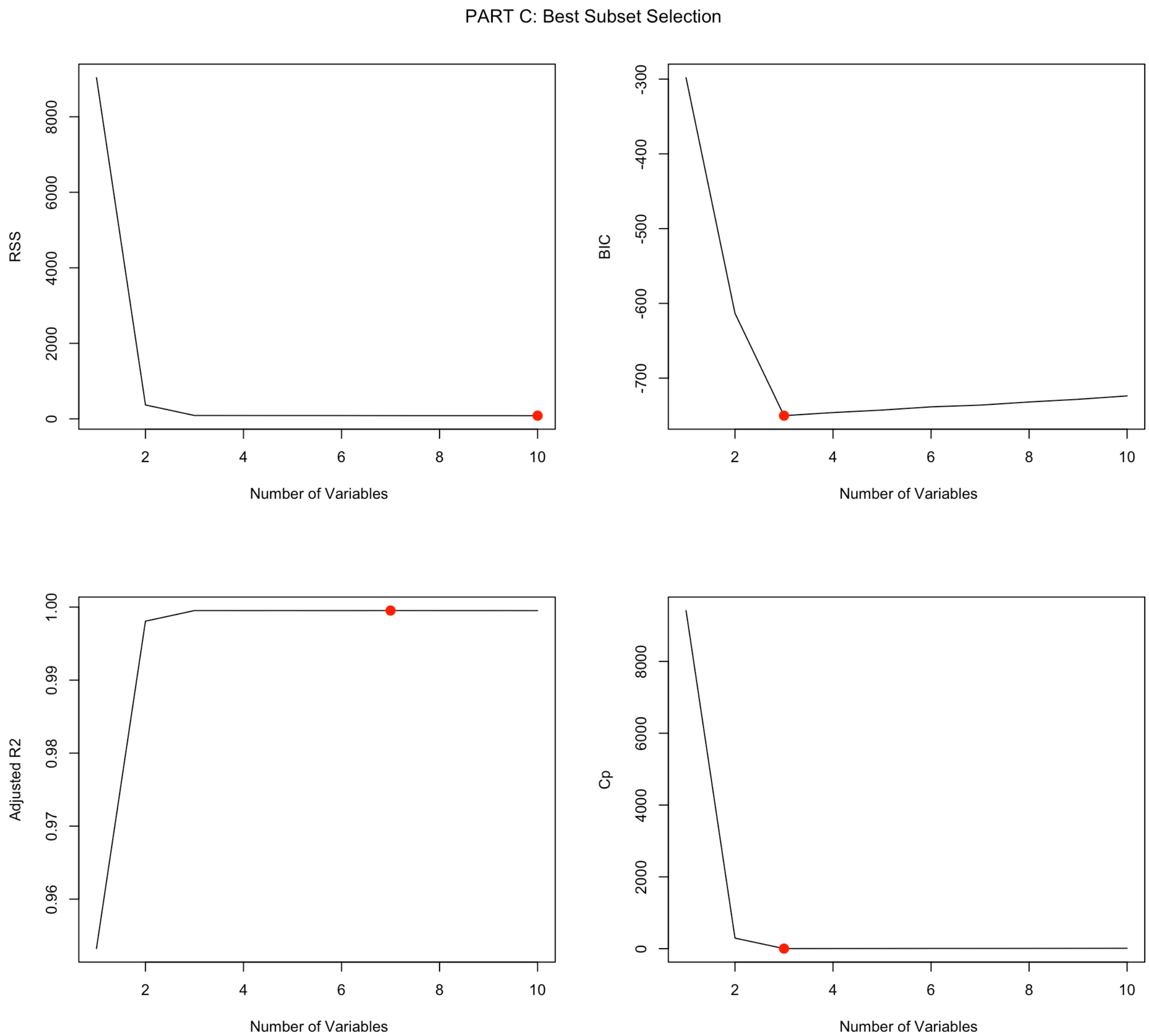
e <- rnorm(100)

```
b0 <- 5

b1 <- 4

b2 <- 35

b3 <- -5

Y <- b0 + b1*X + b2*(X^2) + b3*(X^3)+e

data <- data.frame( y = Y,x = X)

regfit <- regsubsets(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) +
I(X^9) + I(X^10), data = data, nvmax = 10)

reg.sum <- summary(regfit)

par(mfrow = c(2,2))
```

#PARTC

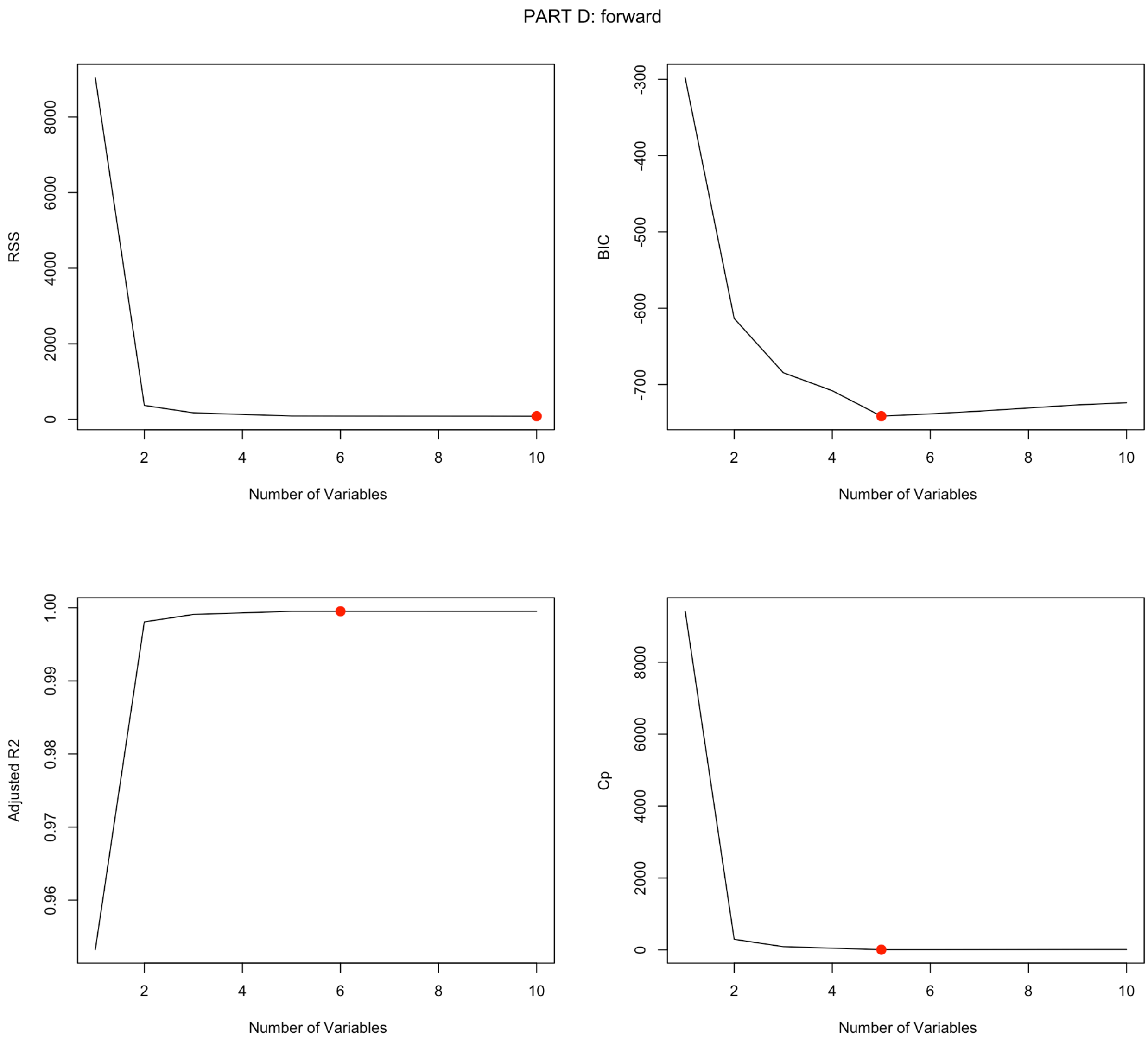
```
plot(reg.sum$rss, xlab = ' Number of Variables' , ylab = 'RSS', type = "l")

points(which.min(reg.sum$rss), reg.sum$rss[which.min(reg.sum$rss)], col = 'red' , cex = 2 , pch
= 20)

plot(reg.sum$bic, xlab = ' Number of Variables' , ylab = 'BIC', type = 'l')

points( which.min(reg.sum$bic), reg.sum$bic[ which.min(reg.sum$bic)], col = 'red' , cex = 2 ,
pch = 20)

plot(reg.sum$adjr2, xlab = ' Number of Variables' , ylab = 'Adjusted R2', type = 'l')

points(which.max(reg.sum$adjr2), reg.sum$adjr2[which.max(reg.sum$adjr2)], col = 'red' , cex =
2 , pch = 20)

plot(reg.sum$cp, xlab = ' Number of Variables' , ylab = 'Cp', type = 'l')

min.cp <- which.min(reg.sum$cp)

points(min.cp, reg.sum$cp[min.cp], col = 'red' , cex = 2 , pch = 20)

mtext('PART C: Best Subset Selection', side = 3 , line = -2, outer = T)
```

 # the number of variables for each value are 10 RSS, 3 BIC, 7 Adjusted R^2, 3 Cp



#part D Forward Stepwise

```
regfit.fwd <- regsubsets(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) + I(X^9) + I(X^10), data = data, nvmax = 10 , method = 'forward')

reg.sum.fwd <- summary(regfit.fwd)

par(mfrow = c(2,2))

plot(reg.sum.fwd$rss, xlab = ' Number of Variables' , ylab = 'RSS', type = "l")

points(which.min(reg.sum.fwd$rss), reg.sum.fwd$rss[which.min(reg.sum.fwd$rss)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.fwd$bic, xlab = ' Number of Variables' , ylab = 'BIC', type = 'l')
```



points(which.min(reg.sum.fwd$bic), reg.sum.fwd$bic[which.min(reg.sum.fwd$bic)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.fwd$adjr2, xlab = ' Number of Variables' , ylab = 'Adjusted R2', type = 'l')

points(which.max(reg.sum.fwd$adjr2), reg.sum.fwd$adjr2[which.max(reg.sum.fwd$adjr2)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.fwd$cp, xlab = ' Number of Variables' , ylab = 'Cp', type = 'l')

points(which.min(reg.sum.fwd$cp), reg.sum.fwd$cp[which.min(reg.sum.fwd$cp)], col = 'red' , cex = 2 , pch = 20)

mtext('PART D: forward', side = 3 , line = -2, outer = T)

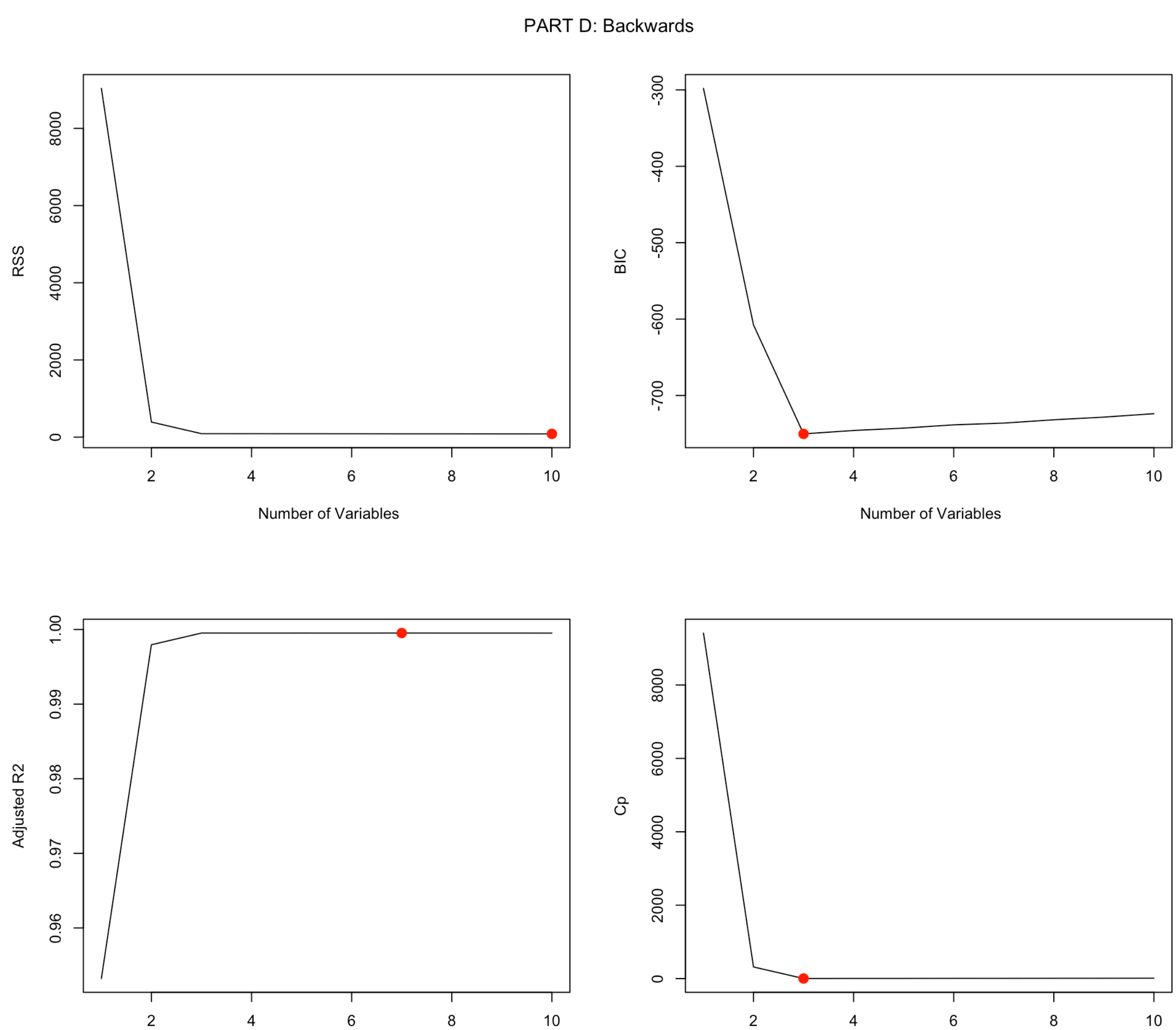
#in part D I found that the number of variables are RSS 10 , BIC 5, Adjusted R^2 6, and Cp 5

```
#backwards

regfit.bwd <- regsubsets(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) + I(X^9) + I(X^10), data = data, nvmax = 10 , method = 'backward')

reg.sum.bwd <- summary(regfit.bwd)

par(mfrow = c(2,2))

plot(reg.sum.bwd$rss, xlab = ' Number of Variables' , ylab = 'RSS', type = "l")

points( which.min(reg.sum.bwd$rss), reg.sum.bwd$rss[ which.min(reg.sum.bwd$rss)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.bwd$bic, xlab = ' Number of Variables' , ylab = 'BIC', type = 'l')

points(which.min(reg.sum.bwd$bic), reg.sum.bwd$bic[which.min(reg.sum.bwd$bic)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.bwd$adjr2, xlab = ' Number of Variables' , ylab = 'Adjusted R2', type = 'l')

points(which.max(reg.sum.bwd$adjr2), reg.sum.bwd$adjr2[which.max(reg.sum.bwd$adjr2)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum.bwd$cp, xlab = ' Number of Variables' , ylab = 'Cp', type = 'l')

points( which.min(reg.sum.bwd$cp), reg.sum.bwd$cp[ which.min(reg.sum.bwd$cp)], col = 'red' , cex = 2 , pch = 20)

mtext('PART D: Backwards', side = 3 , line = -2, outer = T)

#the number of variables for the values are 10 RSS, 3 BIC, 7 adjusted R^2, 3 Cp

coef(regfit, which.min(reg.sum$cp))

coef(regfit.fwd, which.min(reg.sum.fwd$cp))

coef(regfit.bwd, which.min(reg.sum.bwd$cp))
```

Coefficients for the best subset, FWD,BWD

```
(Intercept)           X       I(X^2)       I(X^3)
   5.181436    3.978761    34.992709    -4.983260
> coef(regfit.fwd, which.min(reg.sum.fwd$cp))
 (Intercept)           X       I(X^2)       I(X^3)       I(X^5)       I(X^7)
 5.172125918  4.062689991 35.012337857 -5.019538772 -0.030893994  0.009214946
> coef(regfit.bwd, which.min(reg.sum.bwd$cp))
(Intercept)           X       I(X^2)       I(X^3)
   5.181436    3.978761    34.992709    -4.983260
> points(which.min(reg.sum.bwd$bic), reg.sum.bwd$bic[which.min(reg.sum.bwd$bic)]
```

#All are the same except for the forward selection where the Cp chosen was 5 instead of 3, yet we can see that the first 3 are the same for all. The variable coeffcients are also very similar



library(glmnet)

lasso = model.matrix(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) + I(X^9) + I(X^10), data = data)[, -1]

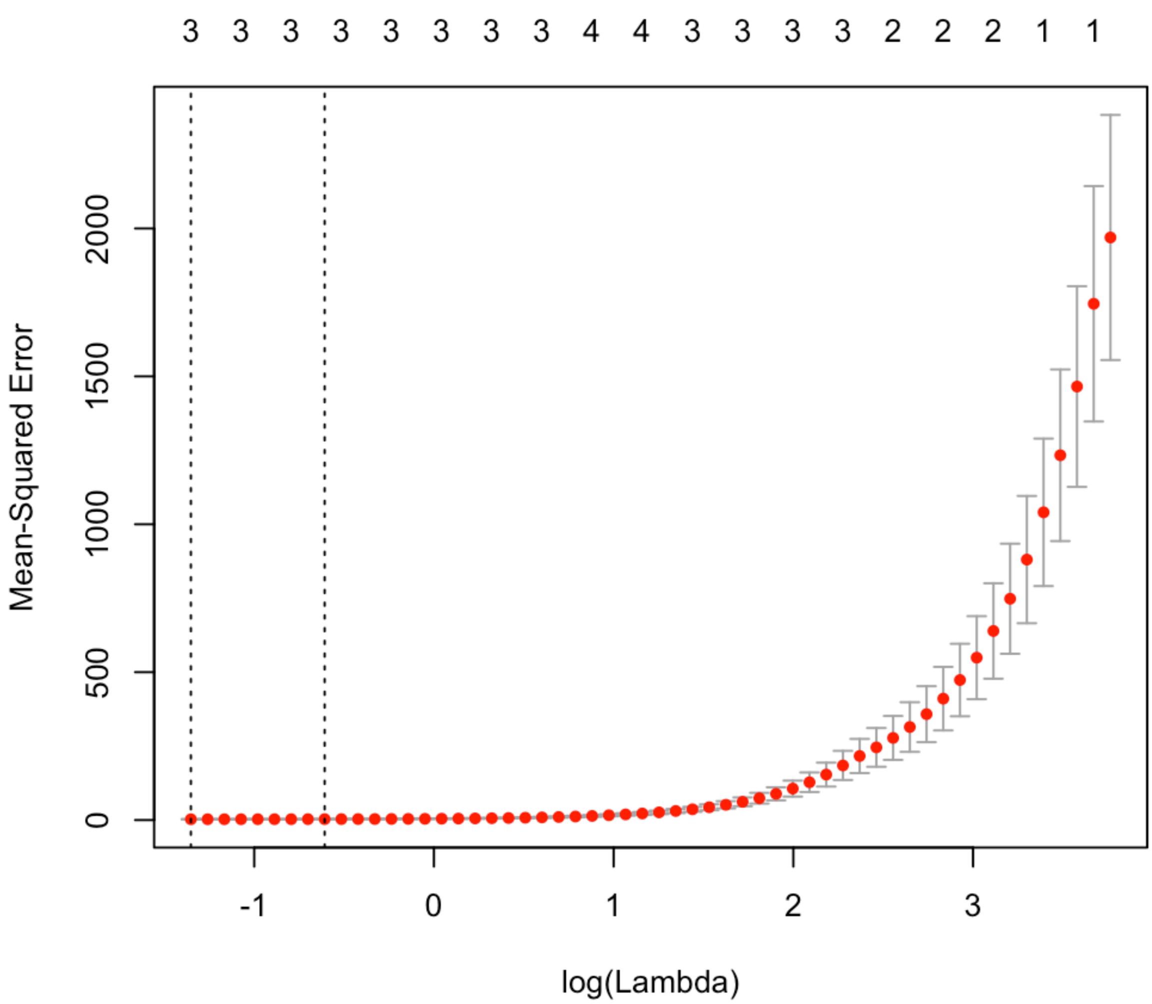
cv.lasso = cv.glmnet(lasso,Y,alpha =1)

plot(cv.lasso)

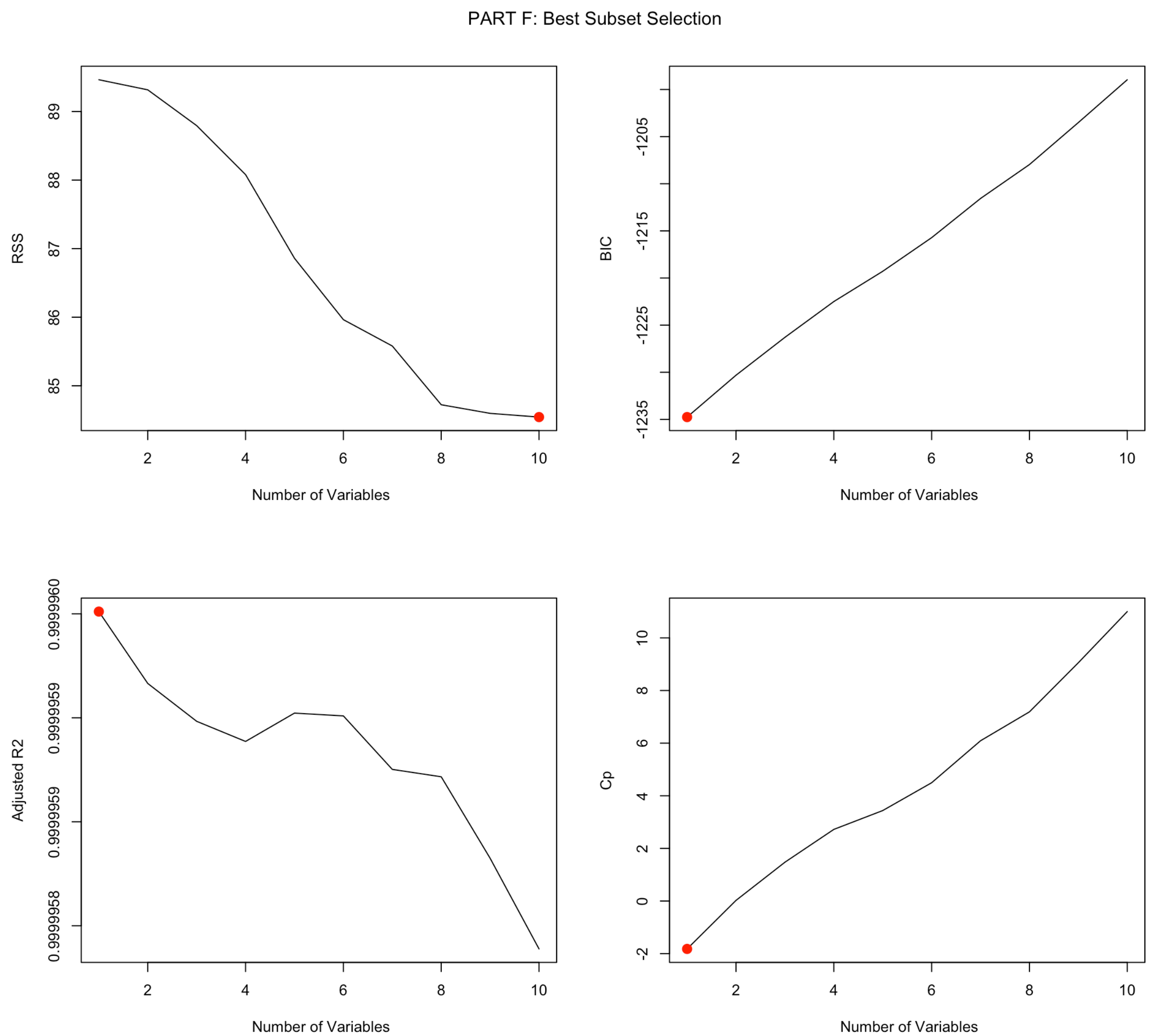
lam = cv.lasso$lambda.min

predict(cv.lasso, s = lam, type = "coefficients")[1:11,]

#the lasso picks the intercept x^2, x^3,x^5 which follows the answers calculated the the previous part except that the X coefficient is not used



```
#part f

b7 = 13

Y= b0 + b7*X^7+e

data <- data.frame( y = Y,x = X)

regfit <- regsubsets(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) +
I(X^9) + I(X^10), data = data, nvmax = 10)

reg.sum <- summary(regfit)

par(mfrow = c(2,2))

plot(reg.sum$rss, xlab = ' Number of Variables' , ylab = 'RSS', type = "l")

points(which.min(reg.sum$rss), reg.sum$rss[which.min(reg.sum$rss)], col = 'red' , cex = 2 , pch
= 20)

plot(reg.sum$bic, xlab = ' Number of Variables' , ylab = 'BIC', type = 'l')
```



points(which.min(reg.sum$bic), reg.sum$bic[which.min(reg.sum$bic)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum$adjr2, xlab = ' Number of Variables' , ylab = 'Adjusted R2', type = 'l')

points(which.max(reg.sum$adjr2), reg.sum$adjr2[which.max(reg.sum$adjr2)], col = 'red' , cex = 2 , pch = 20)

plot(reg.sum$cp, xlab = ' Number of Variables' , ylab = 'Cp', type = 'l')

min.cp <- which.min(reg.sum$cp)

points(min.cp, reg.sum$cp[min.cp], col = 'red' , cex = 2 , pch = 20)

mtext('PART F: Best Subset Selection', side = 3 , line = -2, outer = T)

coef(regfit, which.min(reg.sum$cp))

coef(regfit.fwd, which.min(reg.sum.fwd$cp))

coef(regfit.bwd, which.min(reg.sum.bwd$cp))

```
> coef(regfit, which.min(reg.sum$cp))
(Intercept)      I(X^7)
   5.177755   13.001264
> coef(regfit.fwd, which.min(reg.sum.fwd$cp))
 (Intercept)           X        I(X^2)        I(X^3)        I(X^5)        I(X^7)
 5.172125918  4.062689991 35.012337857 -5.019538772 -0.030893994  0.009214946
> coef(regfit.bwd, which.min(reg.sum.bwd$cp))
(Intercept)           X      I(X^2)      I(X^3)
   5.181436    3.978761   34.992709   -4.983260
```

lasso = model.matrix(Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) + I(X^7) + I(X^8) + I(X^9) + I(X^10), data = data)[, -1]

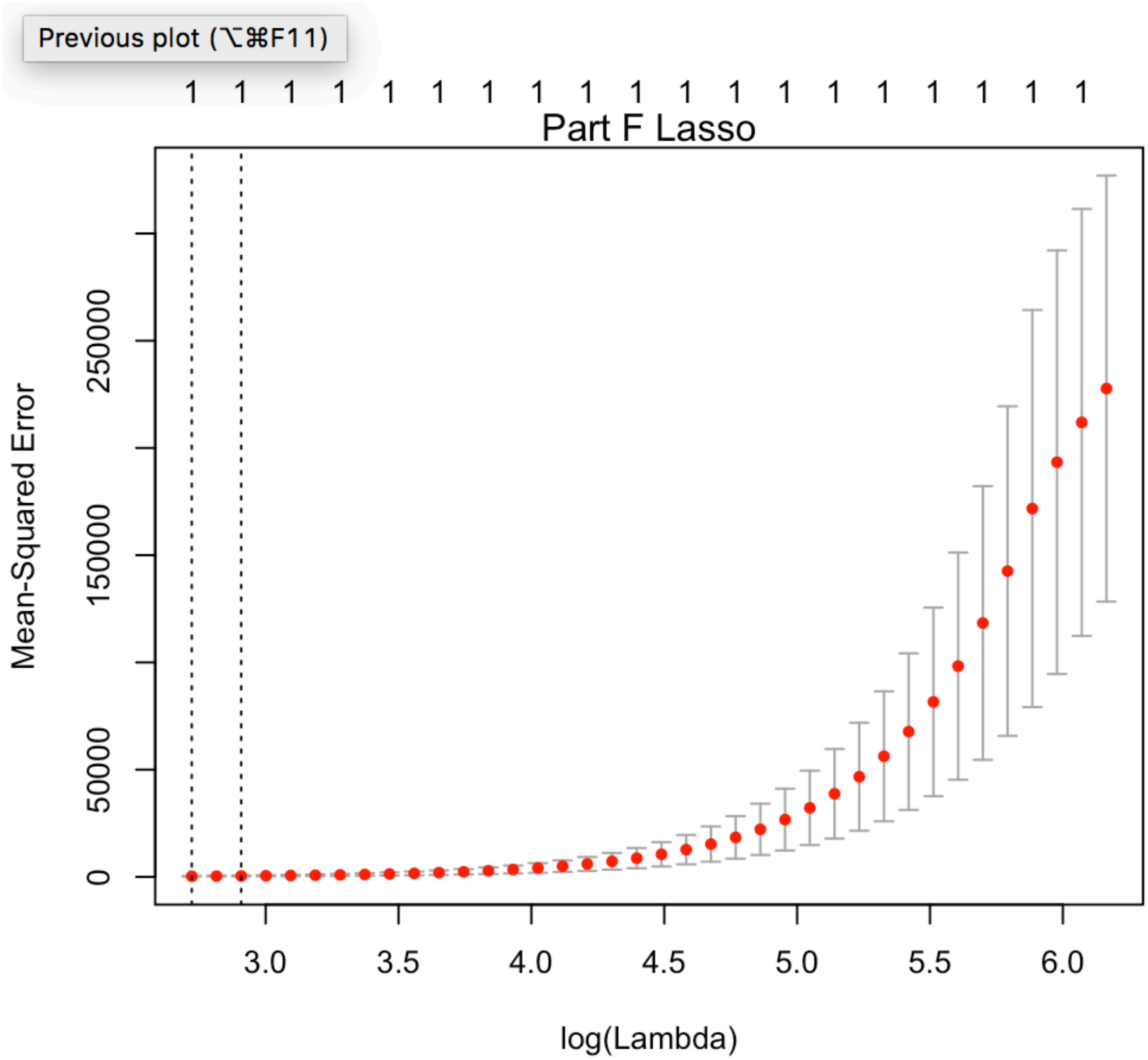
cv.lasso = cv.glmnet(lasso,Y,alpha =1)

plot(cv.lasso)

mtext('Part F Lasso')

lam = cv.lasso$lambda.min

predict(cv.lasso, s = lam, type = "coefficients")[1:11,]



Q4. Chapter 5 Exercise 9

#Excerise 9

library(ISLR)

data(College)

set.seed(33)

train <- sample(1:nrow(College), nrow(College)/2)

test <- -train

College.train <- College[train,]

College.test <- College[test,]

train.lm <- lm(Apps ~ ., data = College.train)

```
predict.lm <-predict(train.lm, College.test)

mean((predict.lm - College.test$Apps)^2)
```

#the test MSE is 1286412

#part c

```
train.rid = model.matrix(Apps~.,data = College.train)

test.rid = model.matrix(Apps~.,data = College.test)

grid <- 10 ^ seq(4, -2, length = 100)

training.ridge <- glmnet(train.rid, College.train$Apps,alpha = 0, lambda = grid, thresh = 1e-12)

cv.ridge <-cv.glmnet(train.rid, College.train$Apps, alpha = 0, lambda = grid, thresh = 1e-12)

lam.ridge <- cv.ridge$lambda.min

lam.ridge

pred.ridge <- predict(training.ridge, s = lam.ridge, newx = test.rid)

mean((pred.ridge - College.test$Apps)^2)
```

#The MSE for the Ridge is 1340660 and is higher than the least squars

#part d

#aplha = 1 for Lasso

```
lasso = glmnet(train.rid, College.train$Apps, alpha = 1 ,lambda = grid, thresh = 1e-12)

lasso.cv = cv.glmnet(train.rid,College.train$Apps, alpha = 1, lambda =  grid, thresh = 1e-12)

lam = cv.lasso$lambda.min

pred.lasso <- predict(lasso,s= lam , newx = test.rid)

mean((pred.lasso - College.test$Apps)^2)
```

#the MSE for the lasso is 1258122

#this MSE is higher than the least squares

```
library(pls)
```

#part e

```
train.pcr <-pcr(Apps ~ ., data = College.train, scale = TRUE, validation = "CV")

validationplot(train.pcr,val.type = "MSEP")
```

mtext(“PCR")

pred.pcr <- predict(train.pcr,College.test,ncomp = 10)

mean((pred.pcr - College.test$Apps)^2)

#the MSE is 2674954

#part F

train.pls <- plsr(Apps ~ ., data = College.train, scale = TRUE, validation = "CV")

validationplot(train.pls,val.type = 'MSEP')

mtext("PLS")

pred.pls <- predict(train.pls,College.test,ncomp = 10)

mean((pred.pls - College.test$Apps)^2)

#the MSE for PLS is 1311462

#part g

test.avg <- mean(College.test$Apps)

lm.r2 <- 1 - mean((predict.lm - College.test$Apps)^2) / mean((test.avg - College.test$Apps)^2)

ridge.r2 <- 1 - mean((pred.ridge - College.test$Apps)^2) / mean((test.avg - College.test$Apps)^2)

lasso.r2 <- 1 - mean((pred.lasso - College.test$Apps)^2) / mean((test.avg - College.test$Apps)^2)

pcr.r2 <- 1 - mean((pred.pcr - College.test$Apps)^2) / mean((test.avg - College.test$Apps)^2)

pls.r2 <- 1 - mean((pred.pls - College.test$Apps)^2) / mean((test.avg - College.test$Apps)^2)

paste('test.avg:',test.avg) #2802.527

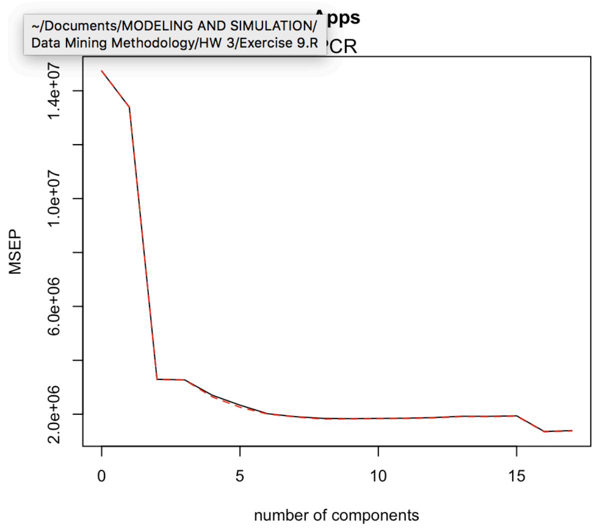
paste('lm.r2:',lm.r2) #.91525

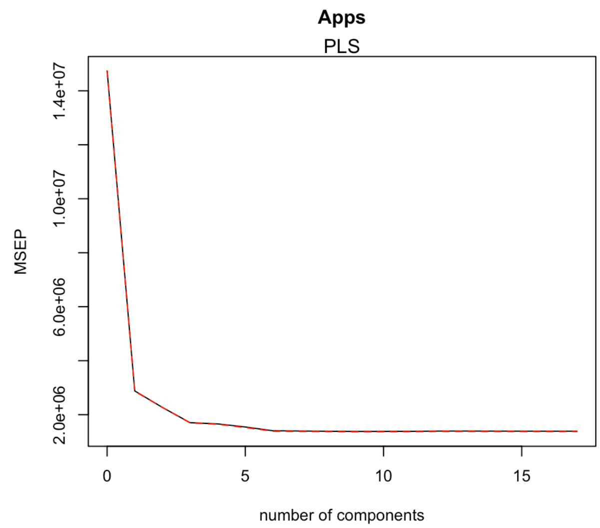
paste('ridge.r2:',ridge.r2) #.9117

paste('lasso.r2:',lasso.r2) #.91711

paste('pcr.r2:',pcr.r2) #.82377

paste('pls.r2:',pls.r2) #.9136





#all models have high accuracy except for the PCR model