HW6Test

Jingshi

3/11/2017

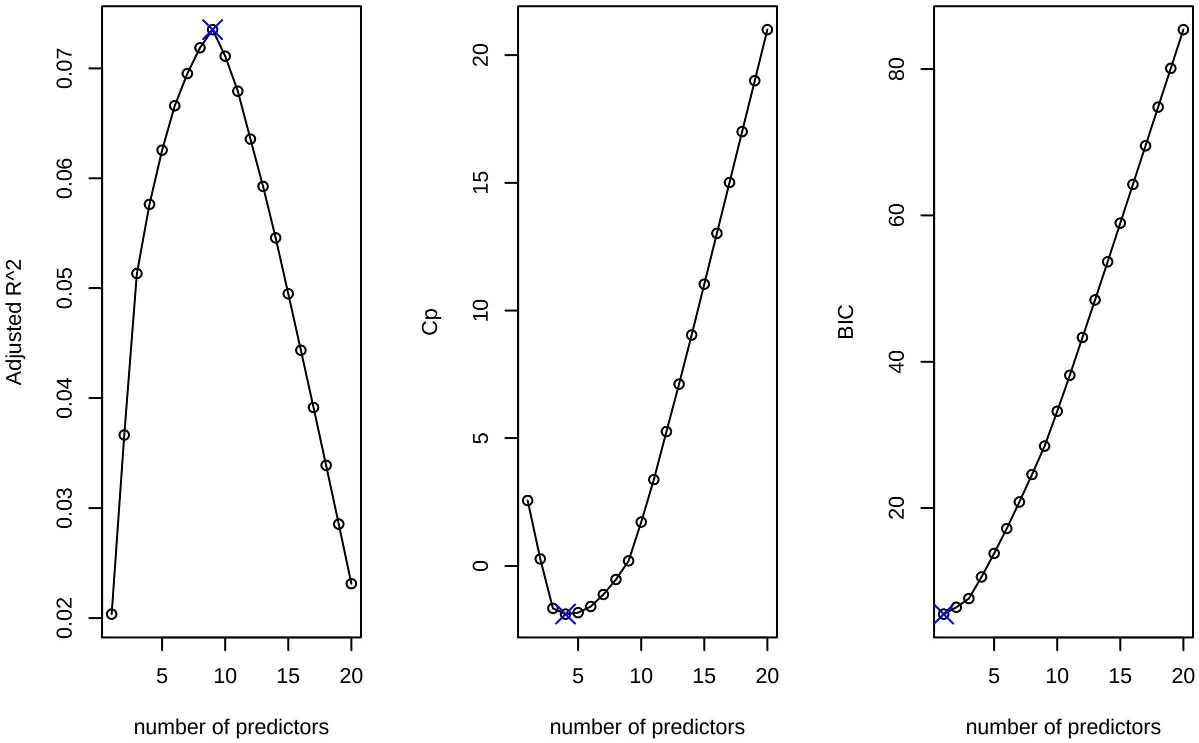
Best subset question

a)

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```r
set.seed(107)
X = as.data.frame(matrix(rnorm(4200), ncol = 21))
names(X)[1] <- "y"
bestSubset <- regsubsets(y~., data = X, nvmax = 20)
mySummary<-summary(bestSubset)
par(mfrow = c(1, 3))
plot(mySummary$adjr2, xlab = "number of predictors", ylab = "Adjusted R^2", type = "o")
points(which.max(mySummary$adjr2), mySummary$adjr2[which.max(mySummary$adjr2)], col = "blue"
       , cex = 2, pch = 4)
plot(mySummary$cp, xlab = "number of predictors", ylab = "Cp", type = "o")
points(which.min(mySummary$cp), mySummary$cp[which.min(mySummary$cp)], col = "blue",
cex = 2, pch = 4)
plot(mySummary$bic, xlab = "number of predictors", ylab = "BIC", type = "o")
points(which.min(mySummary$bic), mySummary$bic[which.min(mySummary$bic)], col = "blue",
cex = 2, pch = 4)
```



As the plots shown, the best model size for adjusted R^2 criteria is 9 because adjusted R^2 is highest at model size 9.

The best model size for Mallows Cp is 4 because Mallows Cp is lowest at model size 4.

The best model size for Bayes Information Criterion is 1 because BIC is lowest at model size 1.

b)

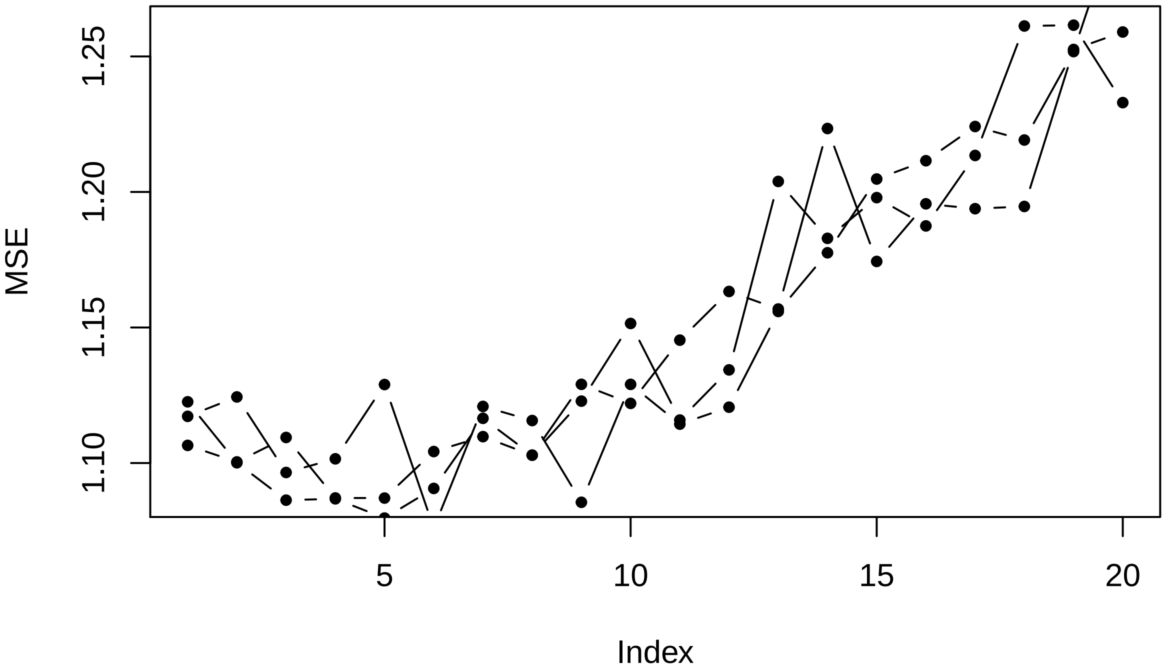
```r
library(boot)
#install.packages("gmp")
#install.packages("HH")
library(HH)
```

```
## Warning: package 'HH' was built under R version 3.3.2

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma

## Loading required package: grid

## Loading required package: latticeExtra

## Loading required package: RColorBrewer

## Loading required package: multcomp

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.3.2

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.3.2

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser

## Loading required package: gridExtra

##
## Attaching package: 'HH'
## The following object is masked from 'package:boot':
##
##     logit
```

```
myDelta<-function(num_variables){
  model<-lm.regsubsets(bestSubset, num_variables)
  r1<-cv.glm(X, glm(model), K = 10)$delta[1]
  r2<-cv.glm(X, glm(model), K = 10)$delta[1]
  r3<-cv.glm(X, glm(model), K = 10)$delta[1]
  range(r1,r2,r3)
}

myDelta1<-function(num_variables){
  model<-lm.regsubsets(bestSubset, num_variables)
  cv.glm(X, glm(model), K = 10)$delta[1]

}

set.seed(11)
for(i in c(1:20)){
  cat ("Error for model size of",i,"is between",myDelta(i)[1],"and",myDelta(i)[2], "\n")
}
```

```
## Error for model size of 1 is between 1.106506 and 1.121885
## Error for model size of 2 is between 1.094966 and 1.110056
## Error for model size of 3 is between 1.092583 and 1.102988
## Error for model size of 4 is between 1.098967 and 1.11681
## Error for model size of 5 is between 1.079332 and 1.106135
## Error for model size of 6 is between 1.077003 and 1.100056
## Error for model size of 7 is between 1.09847 and 1.1405
## Error for model size of 8 is between 1.09937 and 1.122391
## Error for model size of 9 is between 1.105353 and 1.154517
## Error for model size of 10 is between 1.119566 and 1.145204
## Error for model size of 11 is between 1.134963 and 1.143928
## Error for model size of 12 is between 1.135141 and 1.156166
## Error for model size of 13 is between 1.128785 and 1.164939
## Error for model size of 14 is between 1.176015 and 1.181191
## Error for model size of 15 is between 1.143862 and 1.208781
## Error for model size of 16 is between 1.135758 and 1.22054
## Error for model size of 17 is between 1.203654 and 1.222236
## Error for model size of 18 is between 1.199302 and 1.231342
## Error for model size of 19 is between 1.21385 and 1.272799
## Error for model size of 20 is between 1.223498 and 1.274105
```

```
cv.errors <- matrix(NA, 3, 20) # 3 iterations = 3 rows;  20 variables = 20 columns
set.seed(11)
for(i in c(1:20)){
  cv.errors[1,i] = myDelta1(i)
  cv.errors[2,i] = myDelta1(i)
  cv.errors[3,i] = myDelta1(i)
}
plot (cv.errors[1,], pch=20, type="b", ylab = "MSE")
lines (cv.errors[2,], pch=20, type="b")
lines (cv.errors[3,], pch=20, type="b")
```



As the output shows, the model achives lowest MSE between 5 to 9. Models with size less than 9 has comparative lower MSE.

c)

Since it is a linear regression, I would prefer adjusted R^2 given the fact that error is close to the lowest. So the best model size is 9 since it has the highest adjusted R^2.

Compare all four