DSA211 Statistical Learning with R

Homework 8 Answer

Codes of Q1

```
library(ISLR)
attach(Credit)
dim(Credit)
library(leaps)

# perform Best Subset Selection for the Credt data set
regfit1 <- regsubsets(Balance~.-ID, Credit, nvmax=10)
sum_regfit1 <- summary(regfit1)
plot(sum_regfit1$bic, main="Best Subset Selection procedure with BIC",
    xlab="Number of Variables", ylab="BIC", type="b")
a <- which.min(sum_regfit1$bic)
a
coef(regfit1, a)

# perform Forward Selection for the Credt data set
regfit2 <- regsubsets(Balance~.-ID, Credit, nvmax=10, method="forward")
sum_regfit2 <- summary(regfit2)
plot(sum_regfit2$cp, main="Forward Selection with Cp",
    xlab="Number of Variables", ylab="Cp", type="b")
b <- which.min(sum_regfit2$cp)
b
coef(regfit2, b)

# perform Backward Selection for the Credt data set
regfit3 <- regsubsets(Balance~.-ID, Credit, nvmax=10, method="backward")
```

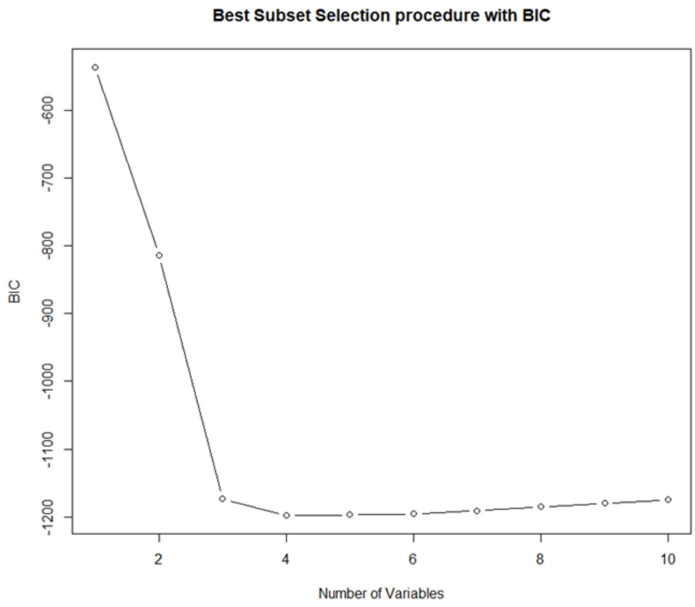
```
sum_regfit3 <- summary(regfit3)
plot(sum_regfit3$adjr2, main="Backward Selection with adjr2",
     xlab="Number of Variables", ylab="Adjusted Rsq", type="b")
c <- which.max(sum_regfit3$adjr2)
c
coef(regfit3, c)


#perform Best Selection for the Credit data set using the Validation set
RNGkind(sample.kind = "Rounding")
set.seed(121)
train <- sample(c(TRUE, FALSE), nrow(Credit), rep=TRUE)
test <- (!train)
regfit4 <- regsubsets(Balance~.-ID, data=Credit[train,], nvmax=10)
test.mat <- model.matrix(Balance~.-ID, data=Credit[test,])
val.error <- rep(NA,10)
for (i in 1:10){
  coefi<- coef(regfit4, id=i)
  pred <- test.mat[,names(coefi)]%*%coefi
  val.error[i] <- mean((Credit$Balance[test]-pred)^2)}
val.error
plot(val.error, main="Best Subset Selection under Validation approach",
     xlab="Number of Variables", ylab="validation mean square error", type="b")
d <- which.min(val.error)
d
# use all the data points to get the estimates
coef(regfit1,d)
```

Output of Q1

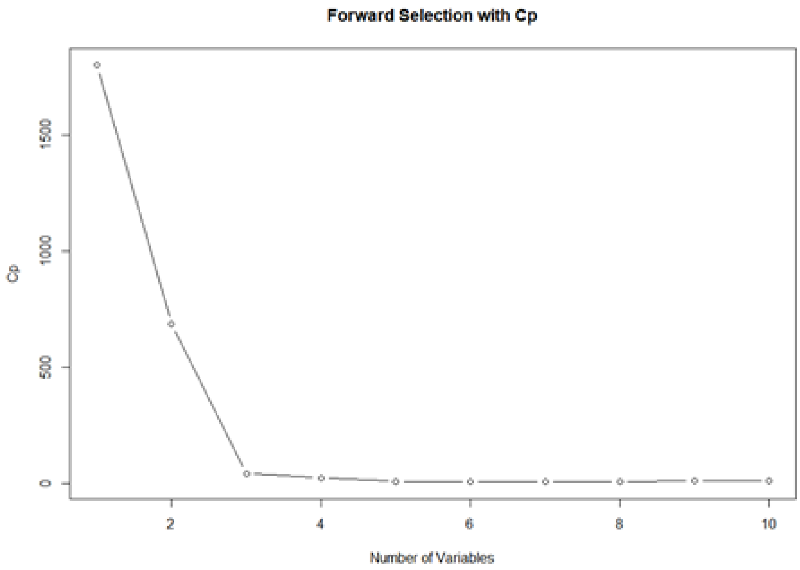
```
> dim(Credit)
[1] 400  12
> library(leaps)
> 
> # perform Best Subset Selection for the Credt data set
> regfit1 <- regsubsets(Balance~.-ID, Credit, nvmax=10)
> sum_regfit1 <- summary(regfit1)
> plot(sum_regfit1$bic, main="Best Subset Selection procedure with BIC",
+      xlab="Number of Variables", ylab="BIC", type="b")
> a <- which.min(sum_regfit1$bic)
> a
[1] 4
> coef(regfit1, a)
 (Intercept)       Income        Limit        Cards   StudentYes 
-499.7272117   -7.8392288    0.2666445   23.1753794  429.6064203 
> 
> # perform Forward Selection for the Credt data set
> regfit2 <- regsubsets(Balance~.-ID, Credit, nvmax=10, method="forward")
> sum_regfit2 <- summary(regfit2)
> plot(sum_regfit2$cp, main="Forward Selection with Cp",
+      xlab="Number of Variables", ylab="Cp", type="b")
> b <- which.min(sum_regfit2$cp)
> b
[1] 6
> coef(regfit2, b)
 (Intercept)       Income        Limit       Rating        Cards          Age   StudentYes 
-493.7341870   -7.7950824    0.1936914    1.0911874   18.2118976   -0.6240560  425.6099369 
> 
> # perform Backward Selection for the Credt data set
> regfit3 <- regsubsets(Balance~.-ID, Credit, nvmax=10, method="backward")
> sum_regfit3 <- summary(regfit3)
> plot(sum_regfit3$adjr2, main="Backward Selection with adjr2",
+      xlab="Number of Variables", ylab="Adjusted Rsq", type="b")
> c <- which.max(sum_regfit3$adjr2)
> c
[1] 7
> coef(regfit3, c)
 (Intercept)       Income        Limit       Rating        Cards          Age GenderFemale 
-488.6158695   -7.8036338    0.1936237    1.0940490   18.1091708   -0.6206538  -10.4531521 
  StudentYes 
 426.5812620 
> 
> #perform Best Selection for the Credit data set using the Validation set
> RNGkind(sample.kind = "Rounding")
> set.seed(121)
> train <- sample(c(TRUE, FALSE), nrow(Credit), rep=TRUE)
> test <- (!train)
> regfit4 <- regsubsets(Balance~.-ID, data=Credit[train,], nvmax=10)
> test.mat <- model.matrix(Balance~.-ID, data=Credit[test,])
> val.error <- rep(NA,10)
> for (i in 1:10){
+   coefi<- coef(regfit4, id=i)
+   pred <- test.mat[,names(coefi)]%*%coefi
+   val.error[i] <- mean((Credit$Balance[test]-pred)^2)}
> val.error
 [1] 59123.33 30136.59 12571.91 12273.74 12641.39 12502.39 12341.33 12395.07 12363.58 12408.11
> plot(val.error, main="Best Subset Selection under Validation approach",
+      xlab="Number of Variables", ylab="validation mean square error", type="b")
> d <- which.min(val.error)
> d
[1] 4
> # use all the data points to get the estimates
> coef(regfit1,d)
 (Intercept)       Income        Limit        Cards   StudentYes 
-499.7272117   -7.8392288    0.2666445   23.1753794  429.6064203 
```

Answers of Q1

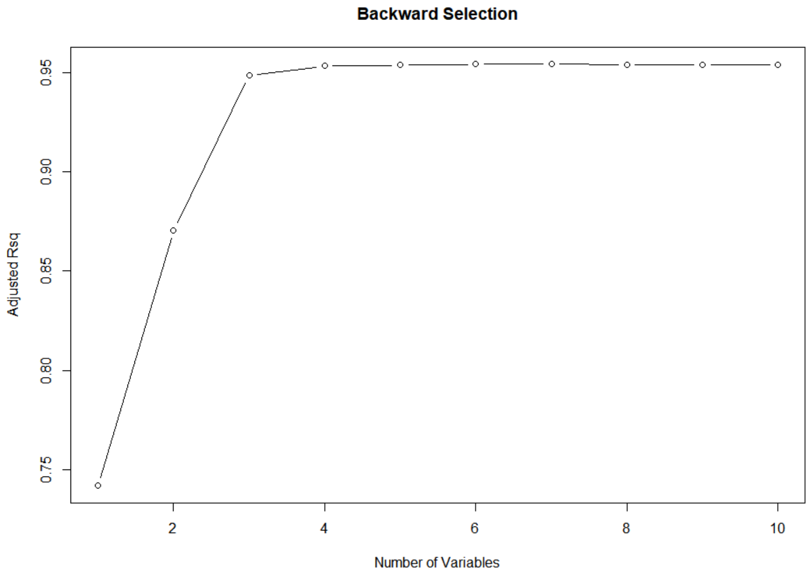
(a) Balance=-499.73-7.8392Incom+0.26664Limit+23.175Cards+429.61StudentYes under the Best Subset Selection



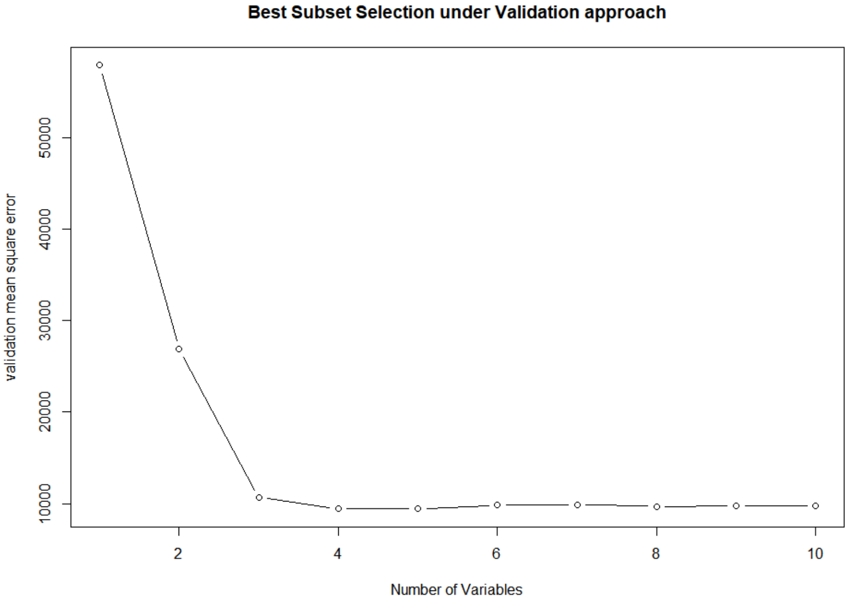
(b) Balance=-493.73-7.7951Incom+0.19369Limit+1.0912Rating+18.212Cards-0.62405Age+425.61StudentYes under the Forward Selection



(c) Balance=-488.62-7.8036Incom+0.19362Limit+1.0940Rating+ 18.109Cards-0.62065Age-10.453GenderFemale+426.58StudentYes under the Backward Selection



(d) Balance=-499.73-7.8392Incom+0.26664Limit+ 23.175Cards+429.61StudentYes under the best Subset Selection with validation approach


Best Subset Selection under Validation approach
validation mean square error
Number of Variables