Apurva Shrivastava
925009508
ISEN 613-600

Assignment #5

Problem 1

In this question, we will predict the number of applications received (Apps) using the other variables in the College data set (ISLR package).

```
>
> library(ISLR)
> dim(College)
[1] 777  18
> names(College)
 [1] "Private"     "Apps"        "Accept"      "Enroll"      "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad"
 [9] "Outstate"    "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"    "S.F.Ratio"   "perc.alumni"
[17] "Expend"      "Grad.Rate"
> summary(College)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc      F.Undergrad
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 1707
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
  P.Undergrad         Outstate       Room.Board       Books           Personal         PhD            Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
> |
```

(a) Perform best subset selection to the data. What is the best model obtained according to Cp, BIC and adjusted R 2? Show some plots to provide evidence for your answer, and report the coefficients of the best model.

```
> library(leaps)
> fit.max = regsubsets(Apps~.,College)
> fit.max = regsubsets(Apps~.,College)
> fit.max = regsubsets(Apps~.,data=College,nvmax=17)
> fit.summary = summary(fit.max)
> fit.summary
Subset selection object
Call: regsubsets.formula(Apps ~ ., data = College, nvmax = 17)
17 Variables  (and intercept)
            Forced in Forced out
PrivateYes      FALSE      FALSE
Accept          FALSE      FALSE
Enroll          FALSE      FALSE
Top10perc       FALSE      FALSE
Top25perc       FALSE      FALSE
F.Undergrad     FALSE      FALSE
P.Undergrad     FALSE      FALSE
Outstate        FALSE      FALSE
Room.Board      FALSE      FALSE
Books           FALSE      FALSE
Personal        FALSE      FALSE
PhD             FALSE      FALSE
Terminal        FALSE      FALSE
S.F.Ratio       FALSE      FALSE
perc.alumni     FALSE      FALSE
Expend          FALSE      FALSE
Grad.Rate       FALSE      FALSE
```

Apurva Shrivastava
925009508
ISEN 613-600

```
1 subsets of each size up to 17
Selection Algorithm: exhaustive
          PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal
1  ( 1 )  " "        "*"    " "    " "       " "       " "         " "         " "      " "        " "   " "
2  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "
3  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "
4  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         "*"      " "        " "   " "
5  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      " "        " "   " "
6  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "
7  ( 1 )  " "        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
8  ( 1 )  "*"        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "
9  ( 1 )  "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
10  ( 1 ) "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
11  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         "*"      "*"        " "   " "
12  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
13  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
14  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
15  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   "*"
16  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"
17  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"

          PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1  ( 1 )  " " " "      " "       " "         " "    " "
2  ( 1 )  " " " "      " "       " "         " "    " "
3  ( 1 )  " " " "      " "       " "         "*"    " "
4  ( 1 )  " " " "      " "       " "         "*"    " "
5  ( 1 )  " " " "      " "       " "         "*"    " "
6  ( 1 )  " " " "      " "       " "         "*"    " "
7  ( 1 )  " " " "      " "       " "         "*"    " "
8  ( 1 )  "*" " "      " "       " "         "*"    " "
9  ( 1 )  "*" " "      " "       " "         "*"    " "
10  ( 1 ) "*" " "      " "       " "         "*"    "*"
11  ( 1 ) "*" " "      " "       " "         "*"    "*"
12  ( 1 ) "*" " "      " "       " "         "*"    "*"
13  ( 1 ) "*" " "      "*"       " "         "*"    "*"
14  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
15  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
16  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
17  ( 1 ) "*" "*"      "*"       "*"         "*"    "*"
> names(fit.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
>
```

→ **<u>Finding best models:</u>**

1.

```
> par(mfrow=c(2,2))
> plot(fit.summary$rss,xlab="No. of predictors",ylab="RSS",type="l")
>
```

2.

```
> plot(fit.summary$adjr2,xlab="No. of predictors",ylab="Adjusted R^2",type="l")
> which.max(fit.summary$adjr2)
[1] 13
>
```

This shows that max adjusted R^2 is for model with 13 predictors. Coefficient estimates:

```
> points(13,fit.summary$adjr2[13], col="green",cex=2,pch=20)
>
> coef(fit.max,13)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad   P.Undergrad
-440.74148270 -484.77261885    1.58542302   -0.87824288   50.41461998  -14.63667155    0.05762769    0.04642270
     Outstate    Room.Board           PhD     S.F.Ratio        Expend     Grad.Rate
  -0.08823311    0.14696204  -10.91804823   15.15475056    0.07786425    8.58578735
>
```

3.

Apurva Shrivastava
925009508
ISEN 613-600

```
> plot(fit.summary$cp,xlab="No. of predictors",ylab="Cp",type='l')
> which.min(fit.summary$cp)
[1] 12
>
```

This shows minimum Cp corresponds to model with 12 predictors. Coefficient estimates:

```
> points(12,fit.summary$cp[12],col="green",cex=2,pch=20)
>
> coef(fit.max,12)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad   P.Undergrad
-157.28685883 -511.78760196    1.58691470   -0.88265385   50.41131660  -14.74735373    0.05945481    0.04593068
     Outstate    Room.Board           PhD        Expend     Grad.Rate
  -0.09017643    0.14776586  -10.70502848    0.07246655    8.63961002
>
```

4.

```
> plot(fit.summary$bic,xlab="No. of predictors",ylab="BIC",type='l')
> which.min(fit.summary$bic)
[1] 10
>
```

This shows minimum BIC corresponds to model with 12 predictors. Coefficient estimates:

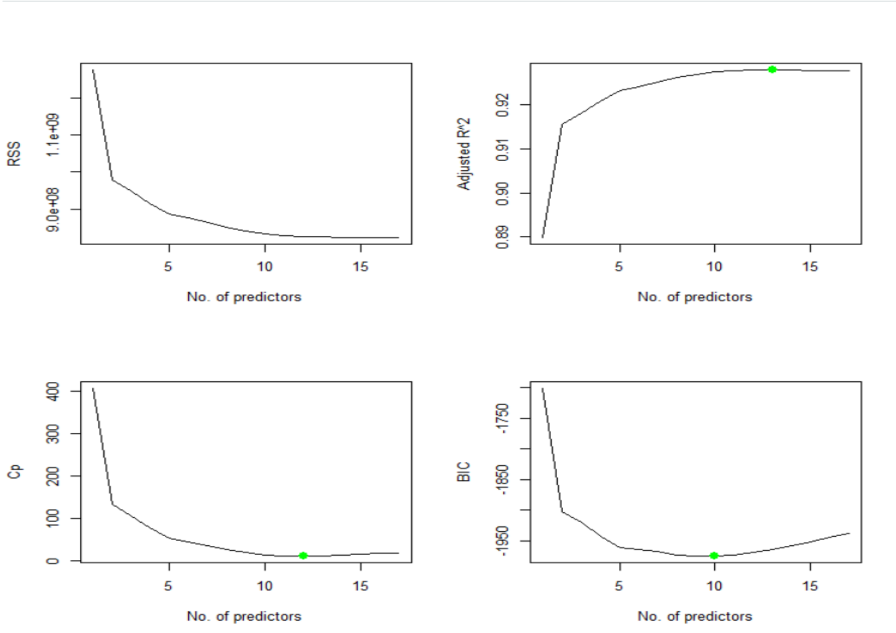
```
> points(10,fit.summary$bic[10],col="green",cex=2,pch=20)
>
> coef(fit.max,10)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc      Outstate    Room.Board
-100.51668243 -575.07060789    1.58421887   -0.56220848   49.13908916  -13.86531103   -0.09466457    0.16373674
          PhD        Expend     Grad.Rate
 -10.01608705    0.07273776    7.33268904
>
```

Summary and Inference with plots:

Model Selection with:	**Criteria**	**Inference**
Adjusted R^2	Select model with maximum value of R^2	Model with 13 predictors has maximum R^2
Cp	Select model with smallest value of Cp	Model with 12 predictors has minimum Cp
BIC	Select model with smallest value of BIC	Model with 10 predictors has the least BIC

Plots with coefficient (pointed in green):

Apurva Shrivastava
925009508
ISEN 613-600



(b) Repeat (a) using forward stepwise selection and backwards stepwise selection. How does your answer compare to the results in (a)?

Forward Stepwise selection:

```
> fit.fwd=regsubsets(Apps~.,data=College,nvmax=17,method="forward")
> summary(fit.fwd)
Subset selection object
Call: regsubsets.formula(Apps ~ ., data = College, nvmax = 17, method = "forward")
17 Variables  (and intercept)
            Forced in Forced out
PrivateYes      FALSE      FALSE
Accept          FALSE      FALSE
Enroll          FALSE      FALSE
Top10perc       FALSE      FALSE
Top25perc       FALSE      FALSE
F.Undergrad     FALSE      FALSE
P.Undergrad     FALSE      FALSE
Outstate        FALSE      FALSE
Room.Board      FALSE      FALSE
Books           FALSE      FALSE
Personal        FALSE      FALSE
PhD             FALSE      FALSE
Terminal        FALSE      FALSE
S.F.Ratio       FALSE      FALSE
perc.alumni     FALSE      FALSE
Expend          FALSE      FALSE
Grad.Rate       FALSE      FALSE
1 subsets of each size up to 17
```

Apurva Shrivastava
925009508
ISEN 613-600

```
Selection Algorithm: forward
          PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal
1  ( 1 )  " "        "*"    " "    " "       " "       " "         " "         " "      " "        " "   " "
2  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "
3  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "
4  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         "*"      " "        " "   " "
5  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      " "        " "   " "
6  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "
7  ( 1 )  " "        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
8  ( 1 )  "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
9  ( 1 )  "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
10  ( 1 ) "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "
11  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         "*"      "*"        " "   " "
12  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
13  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
14  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "
15  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   "*"
16  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"
17  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"
```

```
          PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1  ( 1 )  " " " "      " "       " "         " "    " "
2  ( 1 )  " " " "      " "       " "         " "    " "
3  ( 1 )  " " " "      " "       " "         "*"    " "
4  ( 1 )  " " " "      " "       " "         "*"    " "
5  ( 1 )  " " " "      " "       " "         "*"    " "
6  ( 1 )  " " " "      " "       " "         "*"    " "
7  ( 1 )  " " " "      " "       " "         "*"    " "
8  ( 1 )  " " " "      " "       " "         "*"    " "
9  ( 1 )  "*" " "      " "       " "         "*"    " "
10  ( 1 ) "*" " "      " "       " "         "*"    "*"
11  ( 1 ) "*" " "      " "       " "         "*"    "*"
12  ( 1 ) "*" " "      " "       " "         "*"    "*"
13  ( 1 ) "*" " "      "*"       " "         "*"    "*"
14  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
15  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
16  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
17  ( 1 ) "*" "*"      "*"       "*"         "*"    "*"
>
```

Coefficient estimates of 7 – predictor model:

```
> coef(fit.max,7)
 (Intercept)       Accept       Enroll    Top10perc    Top25perc     Outstate   Room.Board       Expend
-466.5685474    1.5988303   -0.5159191   49.0236357  -14.3277749   -0.1199474    0.1533666    0.0699095
>
```

```
> coef(fit.fwd,7)
 (Intercept)       Accept       Enroll    Top10perc    Top25perc     Outstate   Room.Board       Expend
-466.5685474    1.5988303   -0.5159191   49.0236357  -14.3277749   -0.1199474    0.1533666    0.0699095
>
```

```
> fit.fwd.summary = summary(fit.fwd)
> names(fit.fwd.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
>
```

Finding the best models:

Apurva Shrivastava
925009508
ISEN 613-600

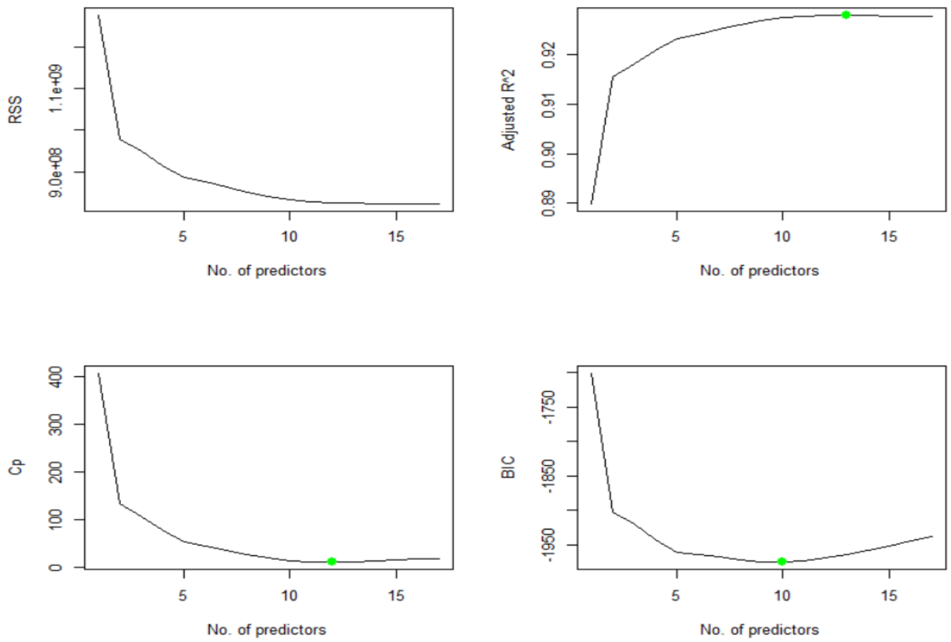
```
> par(mfrow=c(2,2))
> plot(fit.fwd.summary$rss,xlab="No. of predictors",ylab="RSS",type="l")
> plot(fit.fwd.summary$adjr2,xlab="No. of predictors",ylab="Adjusted R^2",type="l")
> which.max(fit.fwd.summary$adjr2)
[1] 13
> points(13,fit.summary$adjr2[13], col="green",cex=2,pch=20)
> coef(fit.fwd,13)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad   P.Undergrad 
-440.74148270 -484.77261885    1.58542302   -0.87824288   50.41461998  -14.63667155    0.05762769    0.04642270 
     Outstate    Room.Board           PhD     S.F.Ratio        Expend     Grad.Rate 
  -0.08823311    0.14696204  -10.91804823   15.15475056    0.07786425    8.58578735 
> 
> plot(fit.fwd.summary$cp,xlab="No. of predictors",ylab="Cp",type='l')
> which.min(fit.fwd.summary$cp)
[1] 12
> points(12,fit.fwd.summary$cp[12],col="green",cex=2,pch=20)
> coef(fit.fwd,12)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad   P.Undergrad 
-157.28685883 -511.78760196    1.58691470   -0.88265385   50.41131660  -14.74735373    0.05945481    0.04593068 
     Outstate    Room.Board           PhD        Expend     Grad.Rate 
  -0.09017643    0.14776586  -10.70502848    0.07246655    8.63961002 
> 
> plot(fit.fwd.summary$bic,xlab="No. of predictors",ylab="BIC",type='l')
> which.min(fit.summary$bic)
[1] 10
> points(10,fit.fwd.summary$bic[10],col="green",cex=2,pch=20)
> coef(fit.fwd,10)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc      Outstate    Room.Board 
-100.51668243 -575.07060789    1.58421887   -0.56220848   49.13908916  -13.86531103   -0.09466457    0.16373674 
          PhD        Expend     Grad.Rate 
 -10.01608705    0.07273776    7.33268904 
> 
```

<u>Summary and Inference with plots:</u>

Model Selection with:	Criteria	Inference
Adjusted R^2	Select model with maximum value of R^2	Model with 13 predictors has maximum R^2
Cp	Select model with smallest value of Cp	Model with 12 predictors has minimum Cp
BIC	Select model with smallest value of BIC	Model with 10 predictors has the least BIC

<u>Plots with coefficient (pointed in green):</u>

Apurva Shrivastava
925009508
ISEN 613-600



Backward Stepwise selection:

```
> fit.bwd=regsubsets(Apps~.,data=College,nvmax=17,method="backward")
> summary(fit.bwd)
Subset selection object
Call: regsubsets.formula(Apps ~ ., data = College, nvmax = 17, method = "backward")
17 Variables  (and intercept)
            Forced in Forced out
PrivateYes      FALSE      FALSE
Accept          FALSE      FALSE
Enroll          FALSE      FALSE
Top10perc       FALSE      FALSE
Top25perc       FALSE      FALSE
F.Undergrad     FALSE      FALSE
P.Undergrad     FALSE      FALSE
Outstate        FALSE      FALSE
Room.Board      FALSE      FALSE
Books           FALSE      FALSE
Personal        FALSE      FALSE
PhD             FALSE      FALSE
Terminal        FALSE      FALSE
S.F.Ratio       FALSE      FALSE
perc.alumni     FALSE      FALSE
Expend          FALSE      FALSE
Grad.Rate       FALSE      FALSE
1 subsets of each size up to 17

Selection Algorithm: backward
          PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal
1  ( 1 )  " "        "*"    " "    " "       " "       " "         " "         " "      " "        " "   " "     
2  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "     
3  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "        " "   " "     
4  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         "*"      " "        " "   " "     
5  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      " "        " "   " "     
6  ( 1 )  " "        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "     
7  ( 1 )  "*"        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "     
8  ( 1 )  "*"        "*"    "*"    "*"       " "       " "         " "         "*"      "*"        " "   " "     
9  ( 1 )  "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "     
10  ( 1 ) "*"        "*"    "*"    "*"       "*"       " "         " "         "*"      "*"        " "   " "     
11  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         "*"      "*"        " "   " "     
12  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "     
13  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "     
14  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   " "     
15  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        " "   "*"     
16  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"     
17  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"        "*"   "*"     
```

Apurva Shrivastava
925009508
ISEN 613-600

```
          PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1  ( 1 )  " " " "      " "       " "         " "    " "
2  ( 1 )  " " " "      " "       " "         " "    " "
3  ( 1 )  " " " "      " "       " "         "*"    " "
4  ( 1 )  " " " "      " "       " "         "*"    " "
5  ( 1 )  " " " "      " "       " "         "*"    " "
6  ( 1 )  " " " "      " "       " "         "*"    " "
7  ( 1 )  " " " "      " "       " "         "*"    " "
8  ( 1 )  "*" " "      " "       " "         "*"    " "
9  ( 1 )  "*" " "      " "       " "         "*"    " "
10  ( 1 ) "*" " "      " "       " "         "*"    "*"
11  ( 1 ) "*" " "      " "       " "         "*"    "*"
12  ( 1 ) "*" " "      " "       " "         "*"    "*"
13  ( 1 ) "*" " "      "*"       " "         "*"    "*"
14  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
15  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
16  ( 1 ) "*" "*"      "*"       " "         "*"    "*"
17  ( 1 ) "*" "*"      "*"       "*"         "*"    "*"
>
```

Coefficient estimates of 7 – predictor model:

```
> coef(fit.max,7)
 (Intercept)       Accept       Enroll    Top10perc    Top25perc     Outstate   Room.Board       Expend
-466.5685474    1.5988303   -0.5159191   49.0236357  -14.3277749   -0.1199474    0.1533666    0.0699095
>
> coef(fit.bwd,7)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc      Outstate    Room.Board        Expend
-734.27312807 -383.62179137    1.57996313   -0.58191135   33.88030869   -0.10169116    0.15500420    0.07574393
>
> fit.bwd.summary = summary(fit.bwd)
> names(fit.bwd.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
>
```

Finding the best models:

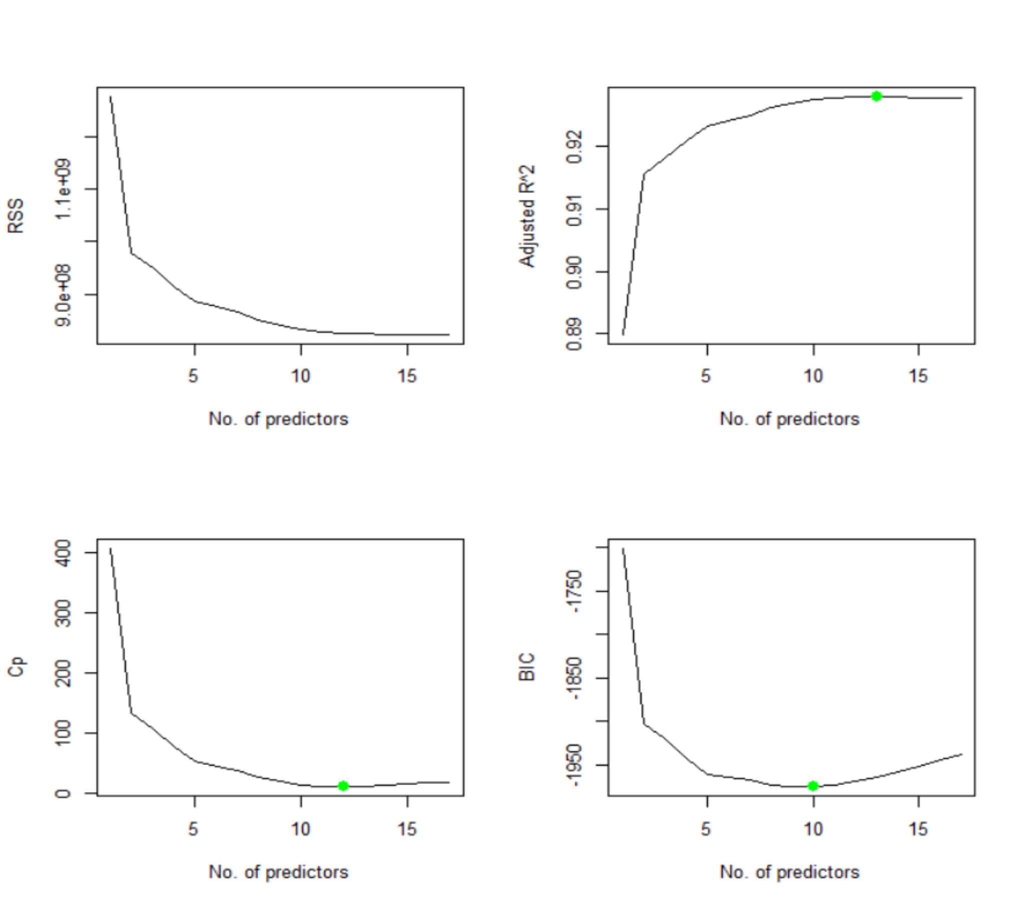
```
> par(mfrow=c(2,2))
>
> plot(fit.bwd.summary$rss,xlab="No. of predictors",ylab="RSS",type="l")
>
> plot(fit.bwd.summary$adjr2,xlab="No. of predictors",ylab="Adjusted R^2",type="l")
> which.max(fit.bwd.summary$adjr2)
[1] 13
> points(13,fit.summary$adjr2[13], col="green",cex=2,pch=20)
> coef(fit.bwd,13)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc    F.Undergrad    P.Undergrad
-440.74148270 -484.77261885    1.58542302   -0.87824288   50.41461998  -14.63667155    0.05762769     0.04642270
     Outstate    Room.Board           PhD     S.F.Ratio        Expend     Grad.Rate
  -0.08823311    0.14696204  -10.91804823   15.15475056    0.07786425    8.58578735
>
> plot(fit.bwd.summary$cp,xlab="No. of predictors",ylab="Cp",type='l')
> which.min(fit.bwd.summary$cp)
[1] 12
> points(12,fit.bwd.summary$cp[12],col="green",cex=2,pch=20)
> coef(fit.bwd,12)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc    F.Undergrad    P.Undergrad
-157.28685883 -511.78760196    1.58691470   -0.88265385   50.41131660  -14.74735373    0.05945481     0.04593068
     Outstate    Room.Board           PhD        Expend     Grad.Rate
  -0.09017643    0.14776586  -10.70502848    0.07246655    8.63961002
>
> plot(fit.bwd.summary$bic,xlab="No. of predictors",ylab="BIC",type='l')
> which.min(fit.summary$bic)
[1] 10
> points(10,fit.bwd.summary$bic[10],col="green",cex=2,pch=20)
> coef(fit.bwd,10)
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc      Outstate     Room.Board
-100.51668243 -575.07060789    1.58421887   -0.56220848   49.13908916  -13.86531103   -0.09466457     0.16373674
          PhD        Expend     Grad.Rate
 -10.01608705    0.07273776    7.33268904
>
```

Apurva Shrivastava
925009508
ISEN 613-600

Summary and Inference with plots:

Model Selection with:	Criteria	Inference
Adjusted R^2	Select model with maximum value of R^2	Model with 13 predictors has maximum R^2
Cp	Select model with smallest value of Cp	Model with 12 predictors has minimum Cp
BIC	Select model with smallest value of BIC	Model with 10 predictors has the least BIC

Plots with coefficient (pointed in green):



Observation:

It is observed that both forward selection and the backward selection model give us the same results with respect to best model basis parameters such as adjusted R^2, BIC and Cp.

Apurva Shrivastava
925009508
ISEN 613-600

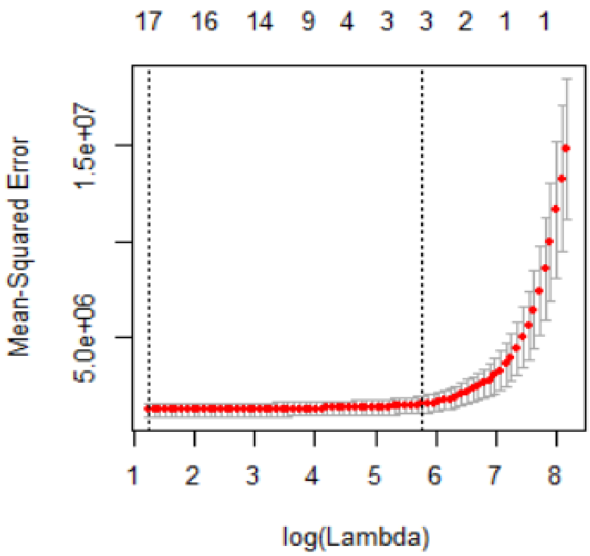
(c) Fit a lasso model on the data. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates.

- Lasso model is built using gelmet() function and setting the value of alpha to 1.
- Cross-validation is performed to obtain the best lambda to minimize the Cross-validation error.
- We predict by fitting a LASSO model to the College data with Apps as the response and all other 17 variables as predictors using the best lambda obtained.

```
> set.seed(1)
> library(glmnet)
Loading required package: Matrix
Loading required package: foreach
foreach: simple, scalable parallel programming from Revolution Analytics
Use Revolution R for scalability, fault tolerance and more.
http://www.revolutionanalytics.com
Loaded glmnet 2.0-13

> x=model.matrix(Apps~.,College)[,-1]
> y=College$Apps
> cv.out=cv.glmnet(x,y,alpha=1)
> plot(cv.out)
> best_lambda=cv.out$lambda.min
> best_lambda
[1] 3.403063
> lasso.best_fit= glmnet(x, y, alpha = 1)
> predict(lasso.best_fit, s = best_lambda, type = "coefficients")
18 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -481.69122766
PrivateYes  -489.47698922
Accept         1.56285991
Enroll        -0.69952897
Top10perc     47.20524294
Top25perc    -12.12210806
F.Undergrad    0.03356097
P.Undergrad    0.04415215
Outstate      -0.08184648
Room.Board     0.14813763
Books          0.01201765
Personal       0.02785918
PhD           -8.24433269
Terminal      -3.21033519
S.F.Ratio     14.04536901
perc.alumni   -0.13535398
Expend         0.07662786
Grad.Rate      8.06878113
> 
```

Apurva Shrivastava
925009508
ISEN 613-600

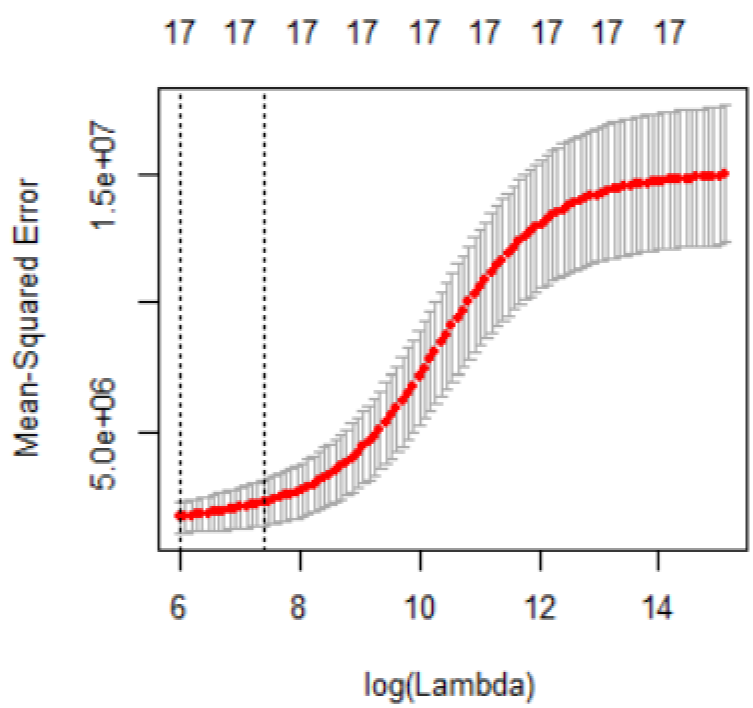


(d) Fit a ridge regression model on the data. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates.

- Ridge regression model is built using glmet() function in R (and alpha=0).
- We perform cross-validation and obtain best value of lambda so that cross-validation error is minimized.
- Using the best lambda value, we predict by fitting a LASSO model Apps as response and all other 17 variables as predictors

Apurva Shrivastava
925009508
ISEN 613-600

```
> cv.out=cv.glmnet(x,y,alpha=0)
> plot(cv.out)
> best_lambda=cv.out$lambda.min
> best_lambda
[1] 400.4766
>
> ridge.best_fit= glmnet(x, y, alpha = 0)
> predict(ridge.best_fit, s = best_lambda, type = "coefficients")
18 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -1.514927e+03
PrivateYes  -5.293325e+02
Accept       9.780751e-01
Enroll       4.666917e-01
Top10perc    2.497314e+01
Top25perc    1.056473e+00
F.Undergrad  7.662859e-02
P.Undergrad  2.445939e-02
Outstate    -2.136542e-02
Room.Board   1.997980e-01
Books        1.352799e-01
Personal    -8.966624e-03
PhD         -3.771159e+00
Terminal    -4.713593e+00
S.F.Ratio    1.282837e+01
perc.alumni -8.831661e+00
Expend       7.527598e-02
Grad.Rate    1.136663e+01
>
```



Apurva Shrivastava
925009508
ISEN 613-600

(e) Now split the data set into a training set and a test set.

i. Fit the best models obtained in the best subset selection (according to Cp, BIC or adjusted R 2) to the training set, and report the test error obtained.

```
> set.seed(1)
> train=sample(c(TRUE,FALSE), nrow(College), rep=TRUE)
> test=(!train)
> fit.best=regsubsets(Apps~.,data=College[train,],nvmax=17)
> summary(fit.best)
Subset selection object
Call: regsubsets.formula(Apps ~ ., data = College[train, ], nvmax = 17)
17 Variables  (and intercept)
            Forced in Forced out
PrivateYes      FALSE      FALSE
Accept          FALSE      FALSE
Enroll          FALSE      FALSE
Top10perc       FALSE      FALSE
Top25perc       FALSE      FALSE
F.Undergrad     FALSE      FALSE
P.Undergrad     FALSE      FALSE
Outstate        FALSE      FALSE
Room.Board      FALSE      FALSE
Books           FALSE      FALSE
Personal        FALSE      FALSE
PhD             FALSE      FALSE
Terminal        FALSE      FALSE
S.F.Ratio       FALSE      FALSE
perc.alumni     FALSE      FALSE
Expend          FALSE      FALSE
Grad.Rate       FALSE      FALSE
1 subsets of each size up to 17
```

```
Selection Algorithm: exhaustive
          PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board
1  ( 1 )  " "        "*"    " "    " "       " "       " "         " "         " "      " "
2  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "
3  ( 1 )  " "        "*"    " "    "*"       " "       " "         " "         " "      " "
4  ( 1 )  "*"        "*"    " "    "*"       " "       " "         " "         " "      " "
5  ( 1 )  "*"        "*"    " "    "*"       "*"       " "         " "         " "      " "
6  ( 1 )  "*"        "*"    " "    "*"       "*"       "*"         " "         " "      " "
7  ( 1 )  "*"        "*"    " "    "*"       "*"       "*"         " "         " "      "*"
8  ( 1 )  "*"        "*"    " "    "*"       "*"       "*"         " "         " "      "*"
9  ( 1 )  "*"        "*"    " "    "*"       "*"       "*"         " "         " "      "*"
10  ( 1 ) "*"        "*"    " "    "*"       "*"       "*"         " "         " "      "*"
11  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         " "      "*"
12  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         "*"      "*"
13  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         " "         "*"      "*"
14  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"
15  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"
16  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"
17  ( 1 ) "*"        "*"    "*"    "*"       "*"       "*"         "*"         "*"      "*"
```

Apurva Shrivastava
925009508
ISEN 613-600

```
          Books Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate
1  ( 1 )  " "   " "      " " " "      " "       " "         " "    " "
2  ( 1 )  " "   " "      " " " "      " "       " "         " "    " "
3  ( 1 )  " "   " "      " " " "      " "       " "         "*"    " "
4  ( 1 )  " "   " "      " " " "      " "       " "         "*"    " "
5  ( 1 )  " "   " "      " " " "      " "       " "         "*"    " "
6  ( 1 )  " "   " "      " " " "      " "       " "         "*"    " "
7  ( 1 )  " "   " "      " " " "      " "       " "         "*"    " "
8  ( 1 )  " "   " "      "*" " "      " "       " "         "*"    " "
9  ( 1 )  " "   "*"      "*" " "      " "       " "         "*"    " "
10  ( 1 ) " "   "*"      "*" " "      " "       " "         "*"    "*"
11  ( 1 ) " "   "*"      "*" " "      " "       " "         "*"    "*"
12  ( 1 ) " "   "*"      "*" " "      " "       " "         "*"    "*"
13  ( 1 ) " "   "*"      "*" " "      " "       "*"         "*"    "*"
14  ( 1 ) " "   "*"      "*" " "      " "       "*"         "*"    "*"
15  ( 1 ) "*"   "*"      "*" " "      " "       "*"         "*"    "*"
16  ( 1 ) "*"   "*"      "*" " "      "*"       "*"         "*"    "*"
17  ( 1 ) "*"   "*"      "*" "*"      "*"       "*"         "*"    "*"
> test.matrix=model.matrix(Apps~.,data=College[test,])
> val.errors=rep(NA,17)
> for(i in 1:17){
+   coeff=coef(fit.best,id=i)
+   pred=test.matrix[,names(coeff)]%*%coeff
+   val.errors[i]=mean((College$Apps[test]-pred)^2)
+ }
> val.errors

 [1] 1714544 1542316 1510924 1512552 1492114 1635683 1645549 1610020 1626584 1616854 1555953 1520681
[13] 1526317 1519996 1522719 1520481 1520331
>
>
> which.min(val.errors)
[1] 5
> coef(fit.best,5)
  (Intercept)    PrivateYes        Accept     Top10perc     Top25perc        Expend
-114.81292106 -617.07313758    1.28203523   50.65442163  -16.79622599    0.05489076
>
>
> fit.best=regsubsets(Apps~.,data=College,nvmax=17)
> coef(fit.best,5)
  (Intercept)        Accept        Enroll     Top10perc      Outstate        Expend
-478.93946837    1.60872093   -0.55385758   33.47143032   -0.10040567    0.08185744
>
```

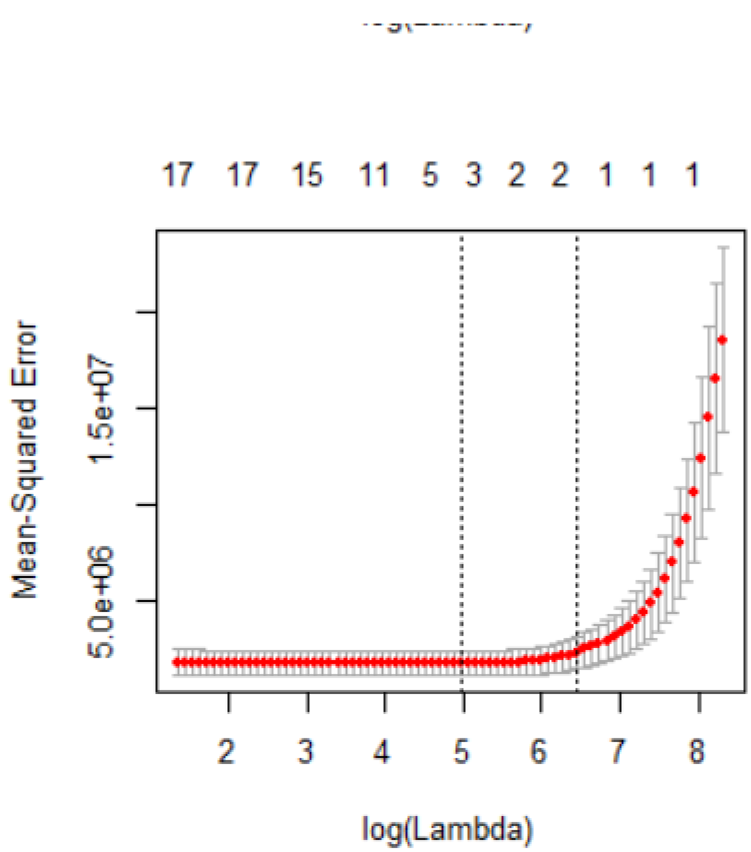
- We select the best model by checking the Validation (Test) Error rates.
- The Validation (Test) Error rates for various models are visible in the screenshots above.
- The test error rate is minimum for the model with 5 predictors combinations:
  - PrivatesYes
  - Accept
  - Top10perc
  - Top25perc
  - Expend
- We use this model and get the coefficients for the same. The coefficients of the best model are also visible in the screenshot above.

ii. Fit a lasso model to the training set, with λ chosen by cross validation. Report the test error obtained.

```
> set.seed(1)
> train=sample(1:nrow(x), nrow(x)/2)
> test=(-train)
> y.test=y[test]
> grid=10^seq(10,-2,length=100)
> lasso.mod=glmnet(x[train,],y[train],alpha=1,lambda=grid)
> cv.out=cv.glmnet(x[train,],y[train],alpha=1)
> plot(cv.out)
> bestlam=cv.out$lambda.min
> bestlam
[1] 144.2049
>
> lasso.pred=predict(lasso.mod,s=bestlam,newx=x[test,])
> mean((lasso.pred-y.test)^2)
[1] 1104498
>
> out=glmnet(x,y,alpha=1)
> lasso.coef=predict(out,type="coefficients",s=bestlam)[1:18,]
> lasso.coef
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad
-708.20363844  -59.24017671    1.38516192    0.00000000   22.63686463    0.00000000    0.00000000
  P.Undergrad      Outstate    Room.Board         Books      Personal           PhD      Terminal
   0.00000000    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000    0.00000000
    S.F.Ratio   perc.alumni        Expend     Grad.Rate
   0.00000000    0.00000000    0.03444127    0.00000000
> lasso.coef[lasso.coef!=0]
  (Intercept)    PrivateYes        Accept     Top10perc        Expend
-708.20363844  -59.24017671    1.38516192   22.63686463    0.03444127
> |
```

Test MSE = 1104498 for best lambda (144.2049) The best model comprises 4 predictors –

1.PrivateYes 2. Accept 3. Top10perc 4. Expend

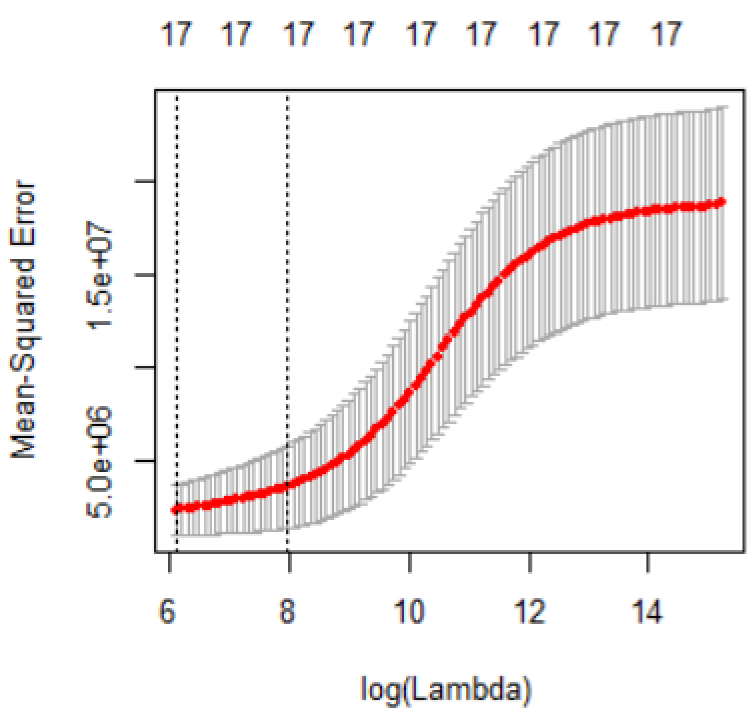


Apurva Shrivastava
925009508
ISEN 613-600

iii. Fit a ridge regression model to the training set, with λ chosen by cross validation. Report the test error obtained.

```
> ridge.mod=glmnet(x[train,],y[train],alpha=0,lambda=grid)
>
>
> set.seed(1)
> cv.out=cv.glmnet(x[train,],y[train],alpha=0)
> plot(cv.out)
> bestlam=cv.out$lambda.min
> bestlam
[1] 450.7435
>
>
> ridge.pred=predict(ridge.mod,s=bestlam,newx=x[test,])
>
> mean((ridge.pred-y.test)^2)
[1] 1036914
>
> out=glmnet(x,y,alpha=0)
> ridge.coef=predict(out,type="coefficients",s=bestlam)[1:18,]
>
> ridge.coef
  (Intercept)    PrivateYes        Accept        Enroll     Top10perc     Top25perc   F.Undergrad 
-1.575123e+03 -5.312141e+02  9.443508e-01  5.084882e-01  2.395085e+01  1.676068e+00  8.195201e-02 
  P.Undergrad      Outstate    Room.Board         Books      Personal           PhD      Terminal 
 2.519290e-02 -1.804998e-02  2.008313e-01  1.443091e-01 -9.669190e-03 -3.428226e+00 -4.551334e+00 
    S.F.Ratio   perc.alumni        Expend     Grad.Rate 
 1.260404e+01 -9.173762e+00  7.444322e-02  1.149378e+01 
> 
```

Test MSE = 1036914 for best lambda (450.7435)



Apurva Shrivastava
925009508
ISEN 613-600

iv. Compare the test errors obtained in the above analysis (i-iii) and determine the optimal model.

Model	**Test MSE**	**No. of predictors**
Best subset selection	1492114	5
LASSO	1104498	4
Ridge regression	1036914	17

The optimal model is the Ridge regression model since it has the lowest Test MSE value.