Assignment 2

Question 1:

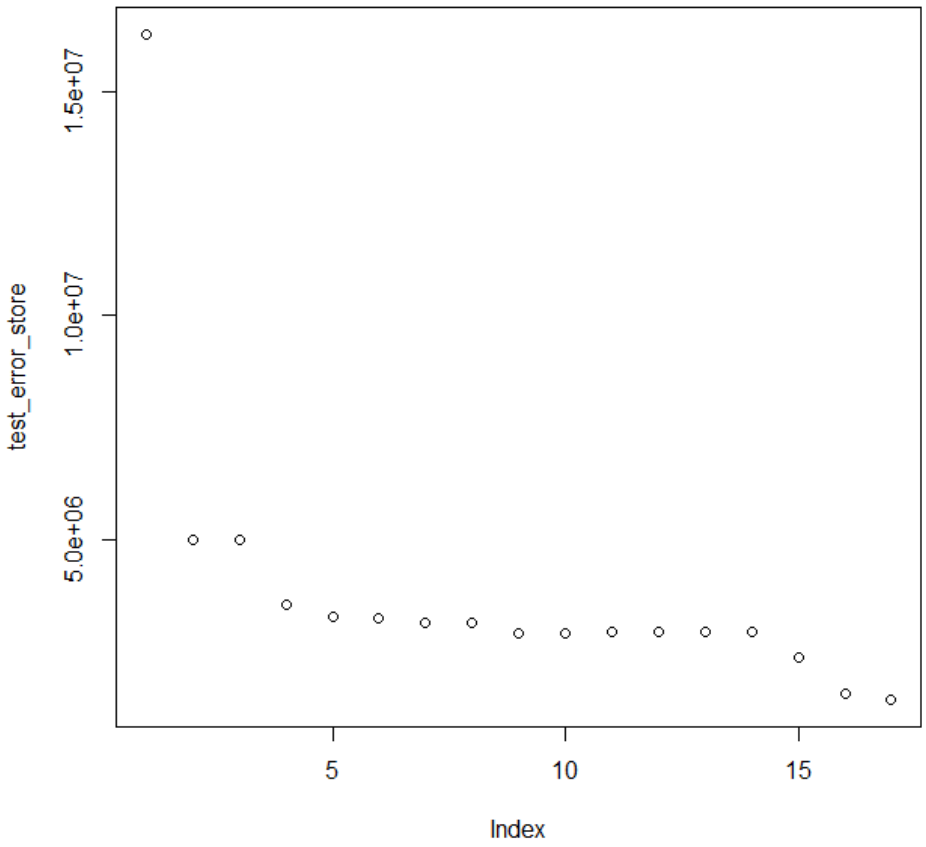
We need to predict Apps variable using remaining 17 predictors. Divided data into train and test set in the ratio 1:1.

First we have applied linear regression on the dataset. From that we obtain ordinary least square for test set. It is around 1439386. It is mean of squared difference between expected and actual values.

In ridge regression, we add 2 norm of the coefficients I the OLS values. It is applies using glmnet function. Cross validated lambda values using cv.glmnet function. Ridge operation is done on the matrices. As class of the variable private is factor, it is converted into numeric for shrinkage methods. Model is trained using train set and applied to test set to estimate error. It is slightly more than the OLS in this case. (actually it should be less)

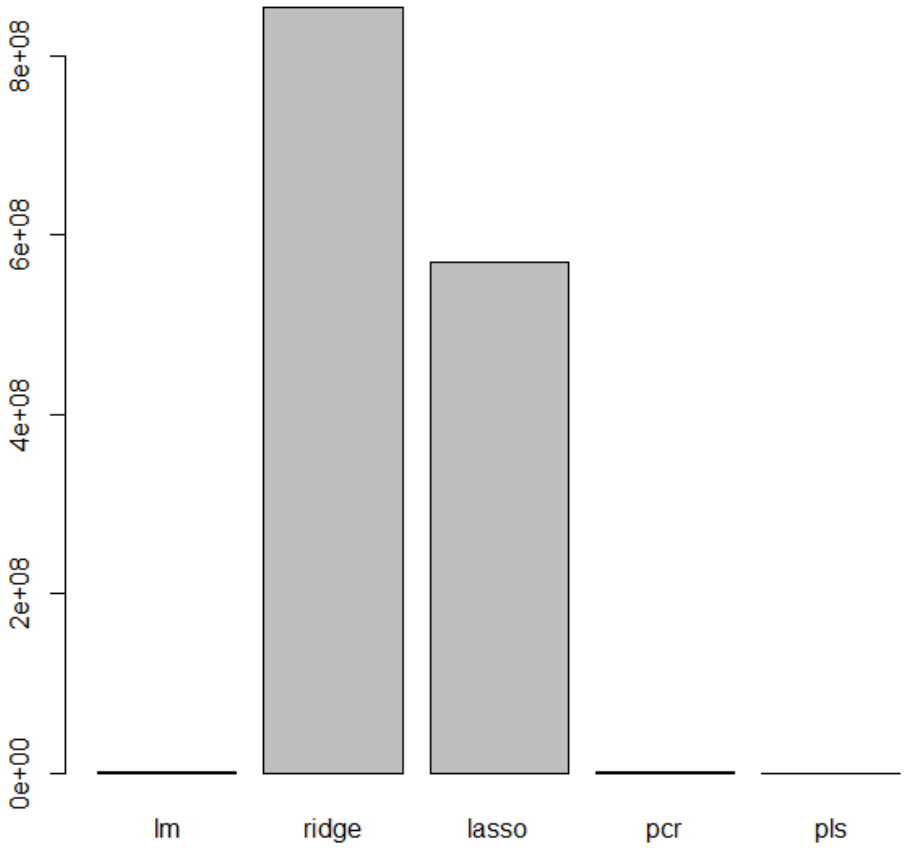
In lasso, we add 1 norm of the coefficients. So some of the irrelevant coefficients might be reduced to zero. We apply lasso regression for best lambda values. Test error is less than the ridge.

We have 17 vaeiables so 17 principal components. We obtain min test error for 17 variables. For PCR we get very less test error as compared to other.



In partial least square, we obtain least test error.

Assignment 2



Comparison of test errors for OLS, ridge, lasso, pcr and pls

```
        OLS        ridge        lasso          pcr          pls
1.439386e+06 8.533880e+08 5.694409e+08 1.439386e+06 5.239496e-01
```

Question 2:

In this we have 85 predictors and need to predict if users will take caravan insurance or not using 1/0. We are using different methods like forward backward subset selection, shrinkage techniques like ridge and lasso and ols.

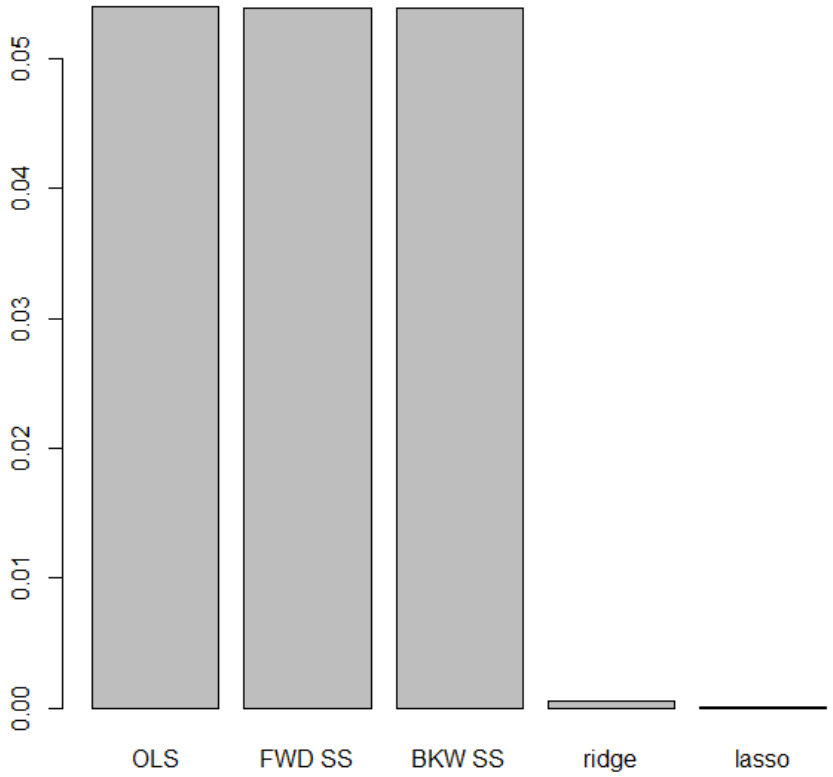
```
which.min(bwss.te)
[1] 38
```

So according to backward subset selection minimum error is for 38 variables.

```
which.min(fwss.te)
[1] 27
```

Assignment 2

According to forward SS, minimum error value is for 27 best variables.



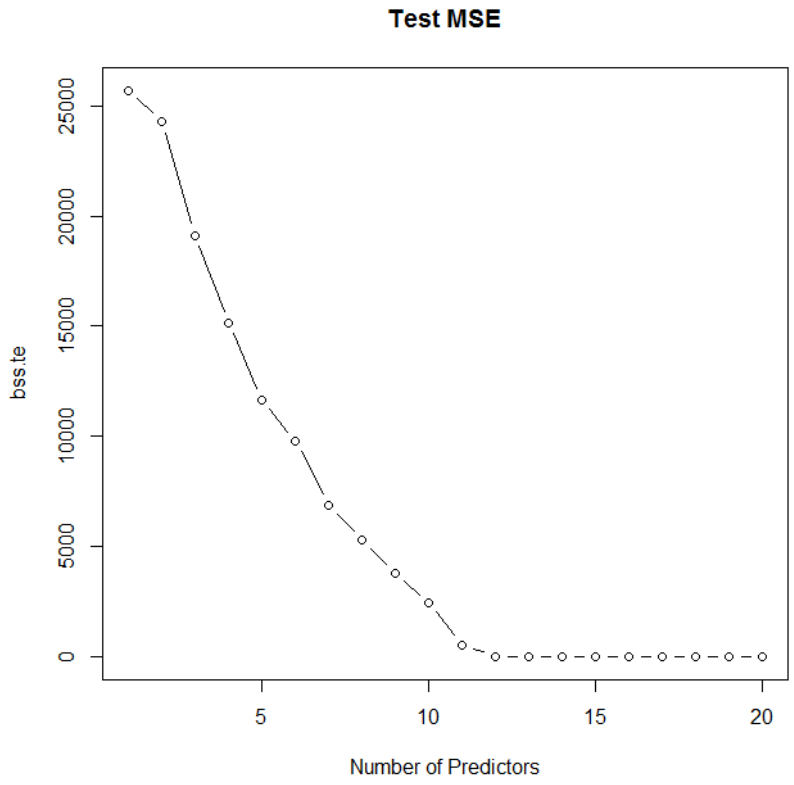

```
         OLS       FWD SS       BKW SS        ridge        lasso
5.398500e-02 5.385551e-02 5.383966e-02 5.911216e-04 5.727662e-05
```

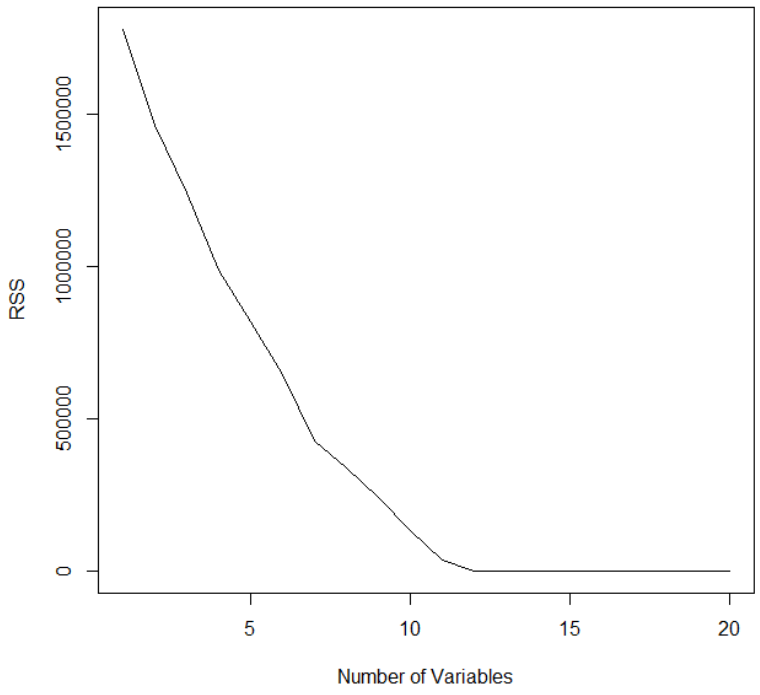
According to above results, we can conclude that we have least test error with lasso and greate st for ordinary least square. For forward subset regression, we get only 27 best predictors but st ill test error is more than the backward ss which displays best 28 predictors.

Question 3

We define our own linear model. We are selecting values of coefficients, 20 predictors and erro r at a random. And calculating response variable using formula y=beta*x+error.
Now using this dataset, we are predicting response variable using best subset selection.

Assignment 2





Training error

Assignment 2

```
Selection Algorithm: exhaustive
          V1  V2  V3  V4  V5  V6  V7  V8  V9  V10 V11 V12 V13 V14 V15 V16 V17
V18 V19 V20
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 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " "*" "*" "*" "*" "*"
"*" "*" "*"
18  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" " " " " " " " " " " " " "*" "*" "*" "*" "*"
"*" "*" "*"
19  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*" "*" "*" "*" "*" "*" "*"
"*" "*" "*"
20  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
"*" "*" "*"
```

The chart above shows best n variables from 1 to 20.

```
> which.min(bss.te)
[1] 13
> which(sum.bss$cp == min(sum.bss$cp))
[1] 13
which(sum.bss$bic == min(sum.bss$bic))
[1] 13
```

Assignment 2

Both cp and bic agree for the 13 best variables. Also for predicted model, least value of error is for 13 best variables for test set.

```
> coef(bestss, id = 13)
(Intercept)          v2          v4          v5          v6          v7
v8          v9
   3.829639    1.001677   13.031623    7.998675   12.011139    8.995053
9.953972    5.061117
        v13         v14         v15         v16         v18         v20
  12.935812    8.972464    7.978423   14.960147   13.971417    9.045056
```

The chart above shows 13 best variables and their beta coefficients generated using best subset selection regression.

```
> betas
      [,1]
 [1,]    0[2,]    1 [3,]    0[4,]   13 [5,]    8 [6,]   12 [7,]    9[8,]   10
 [9,]    5[10,]    0[11,]    0[12,]    0[13,]   13[14,]    9[15,]    8
[16,]   15[17,]    0[18,]   14[19,]    0[20,]    9
```

This above chart shows randomly generated beta values. So we can infer that randomly generated values are closer to the values predicted by the best ss. So response variable doesn't depent much on the predictor numbers 1,3,10,11,12,17,19.