EAS 503-Homework 2

Boobalaganesh Ezhilan
UB Number - 50288429

October 9, 2018

1 Problem 1

Knowing the data
In this problem we predict the number of applications received using the other variables in the College data set which is in the ISLR package.Data set has 777 rows and 18 variables from which we need to predict the number of received(APPS)

Splitting the Data
The data set has 777 rows,Out of which I had taken 50 percent of the training data set and other 50 percent of the as test data set.So training data set has 388 observations and test data set has 389 observations.

Fitting the Linear Regression Model
The mean square loss or MSE when using Least Square Method for predicting number of application is given by

$$\boxed{MSE : 966004, RMSE : 983}$$

where MSE is Mean Square Error and RMSE is Root Mean Square Error
The Mean Square Error and Root Mean Square Error for least square model is extremely large in this case.

Fitting the Ridge Regression Model
The Lambda chosen by the cross validation is **0.01149757** .The mean square loss or MSE when using Ridge regression for predicting number of application is given by

$$\boxed{MSE : 965948 RMSE : 983}$$

ridge.PNG

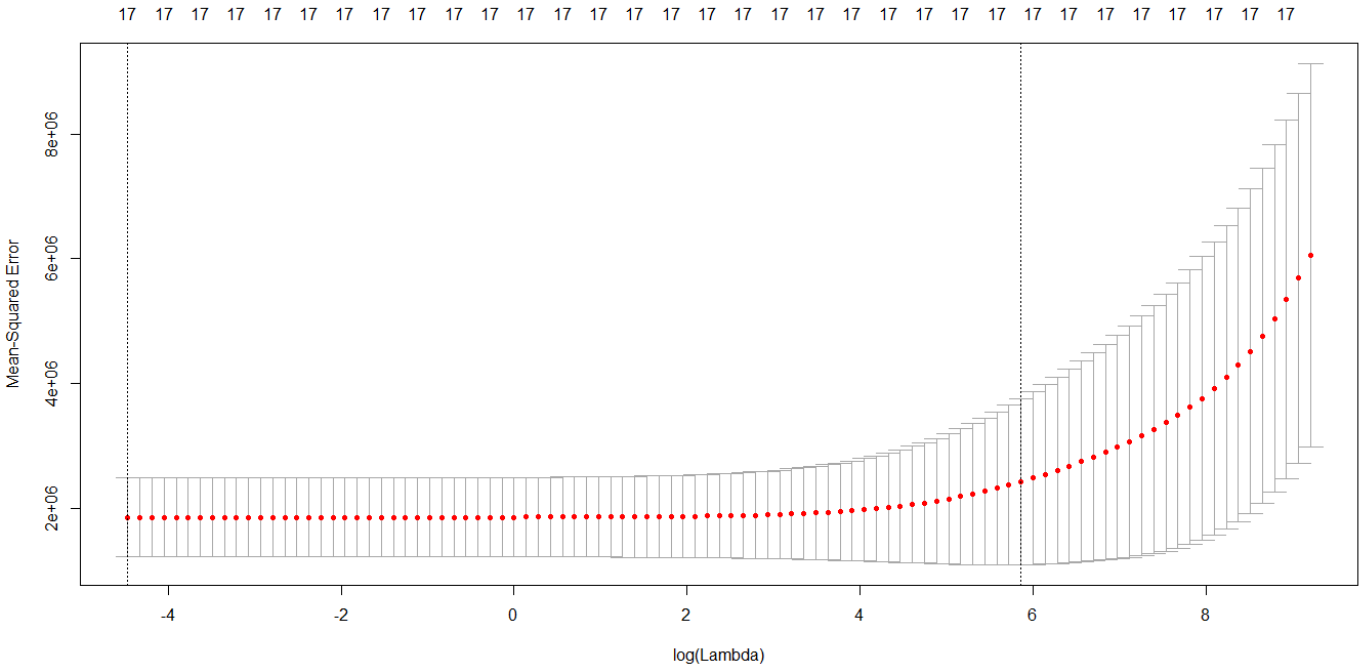


Figure 1: Lambda Vs MSE In Ridge regression

The relation ship between Lambda and the Mean Square error is examined below.From this we can understand that the Lambda is approximately 403.11 where we had found this by cross validation

Fitting the Lasso Regression Model
The Lambda chosen by the cross validation is **28.48036** .The mean square loss or MSE when using Ridge regression for predicting number of application is given by

$$\boxed{MSE : 914232, RMSE : 956}$$

The relation ship between Lambda and the Mean Square error is examined below.From this we can understand that the Lambda is approximately 28 where we had found this by cross validation

lasso.PNG

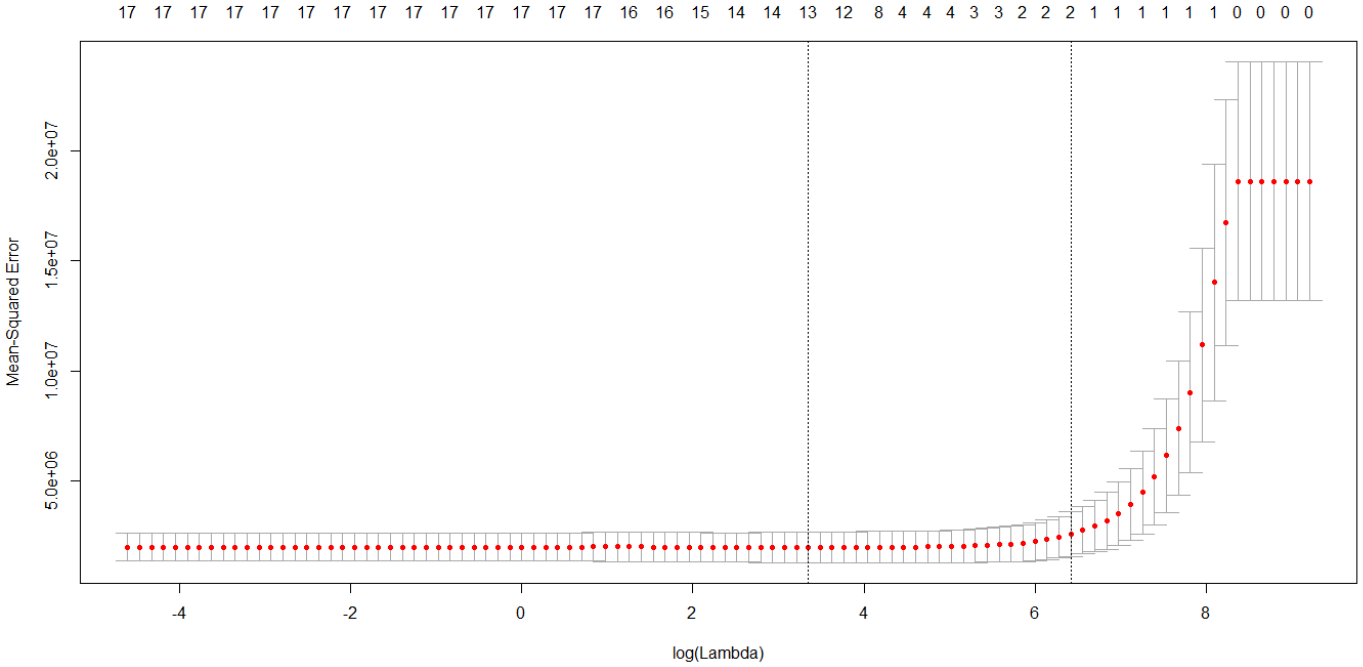


Figure 2: Lambda Vs MSE In Lasso regression

Fitting PCR Model

The **K** chosen by the cross validation is **14** .The mean square loss or MSE when using PCR model for predicting number of application is given by

$$\boxed{MSE : 1260552, RMSE : 1123}$$

variance.PNG

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
SS loadings	360.872	2990410	310.585	1795394	42415.4	224422.3	322493.6	107595.4	201968.3
Proportion Var	360.872	2990410	310.585	1795394	42415.4	224422.3	322493.6	107595.4	201968.3
Cumulative Var	360.872	2990771	2991081.484	4786476	4828891.2	5053313.5	5375807.2	5483402.6	5685371.0

	Comp 10	Comp 11	Comp 12	Comp 13	Comp 14	Comp 15	Comp 16	Comp 17
SS loadings	75356.66	304.384	11321.9	119905.1	8117909	1678697	5988542	2856009
Proportion Var	75356.66	304.384	11321.9	119905.1	8117909	1678697	5988542	2856009
Cumulative Var	5760727.62	5761032.005	5772353.9	5892259.0	14010168	15688865	21677407	24533416

Figure 3: Variance Proportion for different Components

The relation ship between Number of components and the Mean Square error is examined below.From this we can understand that the MSE is minimum around 11,which can also be said with respect to proportion variance which is around 304.384 in this case

Model.PNG

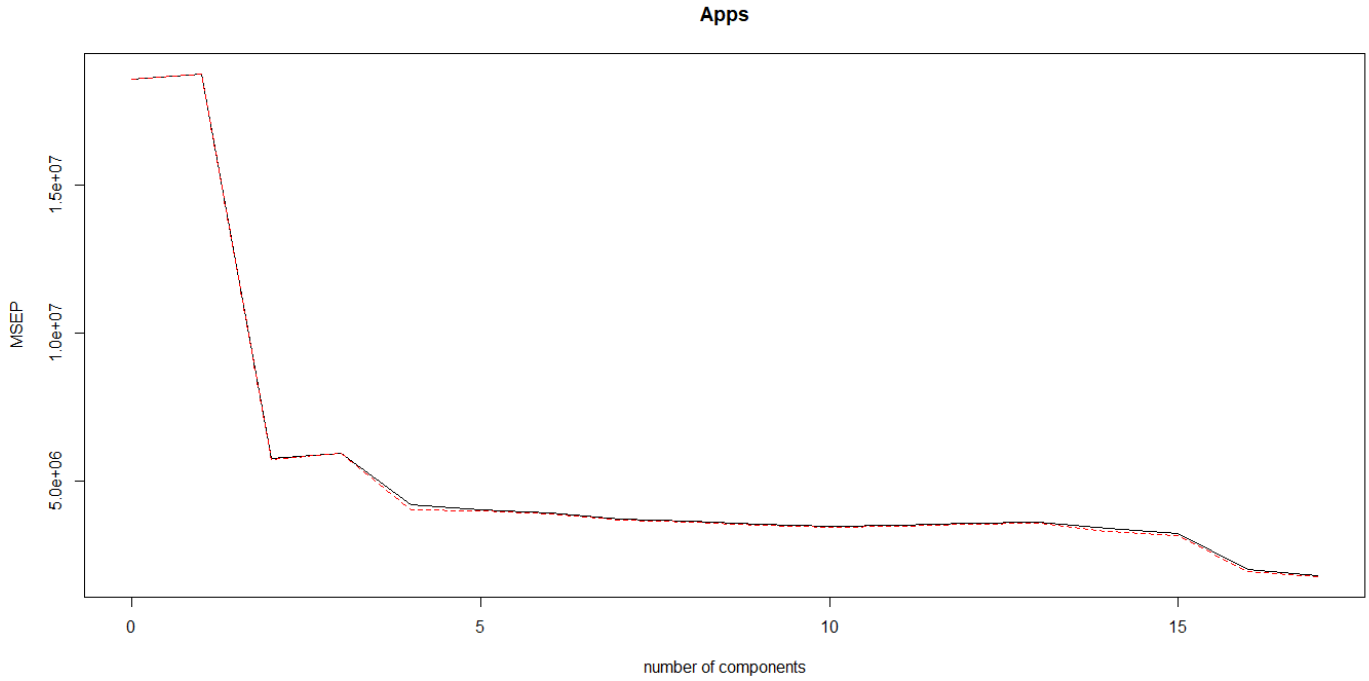


Figure 4: Number of Components vs MSE

Fitting PLS Model

The **K** chosen by the cross validation is **14** .The mean square loss or MSE when using PLS model for predicting number of application is given by

$$\boxed{MSE : 965563, RMSE : 983}$$

The relation ship between Number of components and the Mean Square error is examined below.From this we can understand that the MSE is minimum around 13,which can also be said with respect to proportion variance which is 0.072 around in this case

model.PNG

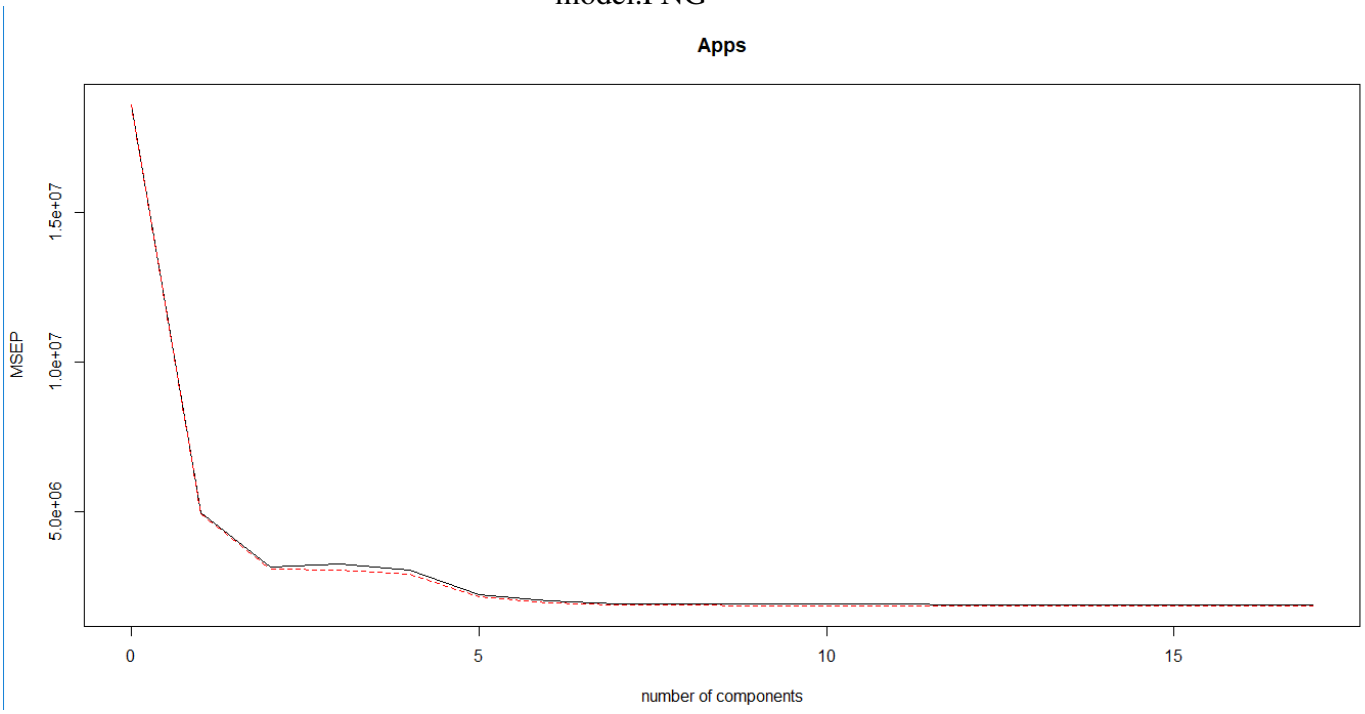


Figure 5: Number of Components vs MSE

Difference among Test errors and accuracy

From the above discussed models of predicting number of application we understood that Ordinary least Squares,ridge regression,Lasso regression,Partial least squares perform more or less same.There MSE for OLS is 966004,ridge is 965948,lasso is 914232 ,PCR is 1260552 and pls is 965563.While looking into lasso model it reduces the variables(i.e shrinks the coefficients) Terminal , Books, and "Enroll" variables to zero and also in PLS model it cut out 4 variables out of 18 variables. It is seems that most of the variables contribute significant information to the college application.By looking into the histogram and output console the mean square error for PCR model is comparatively higher than the other regression models.
While looking into values of R^2 from the histogram,it is 0.9119432 for linear regression model ,0.9073782 for ridge model,0.9107172 for lasso model, 0.8192203 for PCR model and 0.9122182 for PLS model.From the observed model we can infer that value of R^2 is least for PCR model this is because of smallest variation in data of PCR model which describes that the accuracy of the PCR model is very less compared to others.This is visualized with the help of histogram given below.
All models, except PCR, predict college applications with high accuracy.This is also reflected in the test MSE which is the highest for PCR.

errors.PNG

```
> print(paste("MSE_OLS: ", round((rmse.lm)^2), " RMSE_OLS: ", round(rmse.lm)))
[1] "MSE_OLS:  966004  RMSE_OLS:  983"
> print(paste("MSE_Ridge: ", round((rmse.ridge)^2), " RMSE_Ridge: ", round(rmse.ridge)))
[1] "MSE_Ridge:  965948  RMSE_Ridge:  983"
> print(paste("MSE_Lasso: ", round((rmse.lasso)^2), " RMSE_Lasso: ", round(rmse.lasso)))
[1] "MSE_Lasso:  914232  RMSE_Lasso:  956"
> print(paste("MSE_Pcr: ", round((rmse.pcr)^2), " RMSE_Pcr: ", round(rmse.pcr)))
[1] "MSE_Pcr:  1260552  RMSE_Pcr:  1123"
> print(paste("MSE_Pls: ", round((rmse.pls)^2), " RMSE_Pls: ", round(rmse.pls)))
[1] "MSE_Pls:  965563  RMSE_Pls:  983"
```

Figure 6: Test MSE for different models

error bar plot.PNG

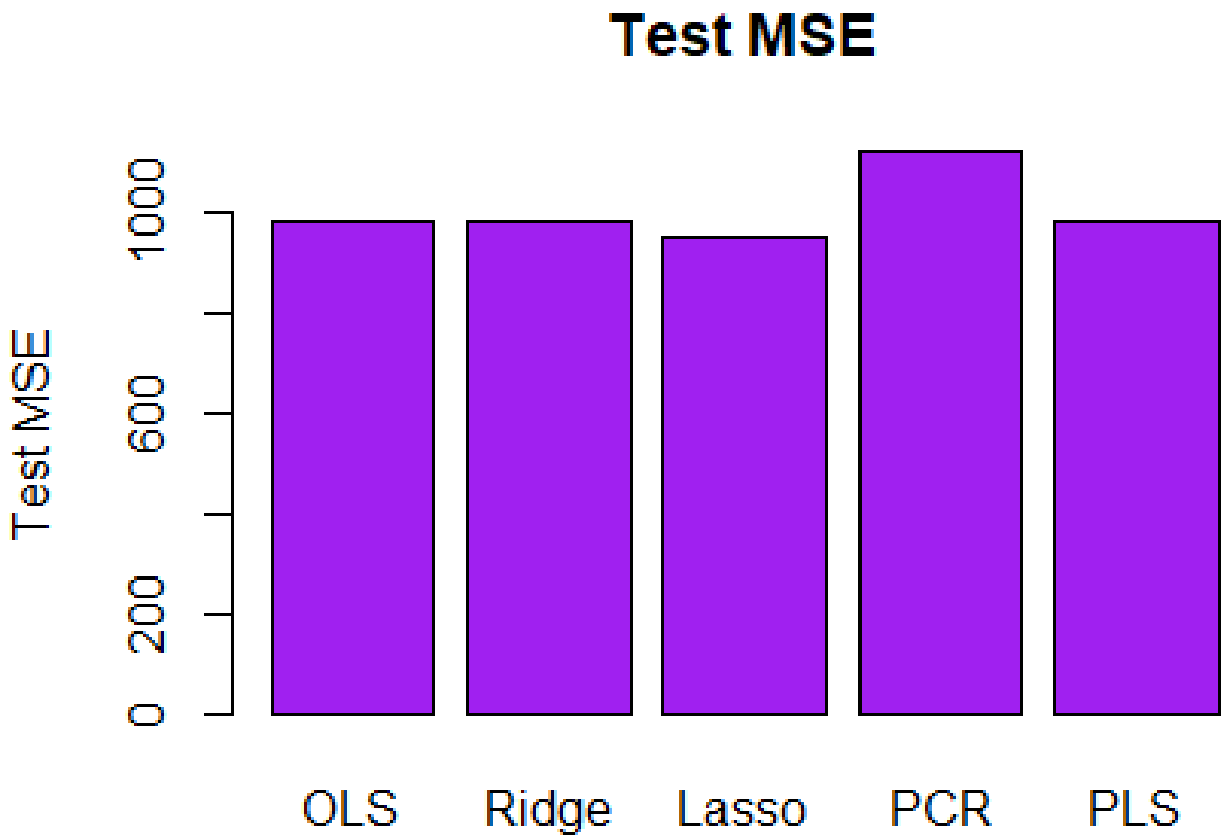
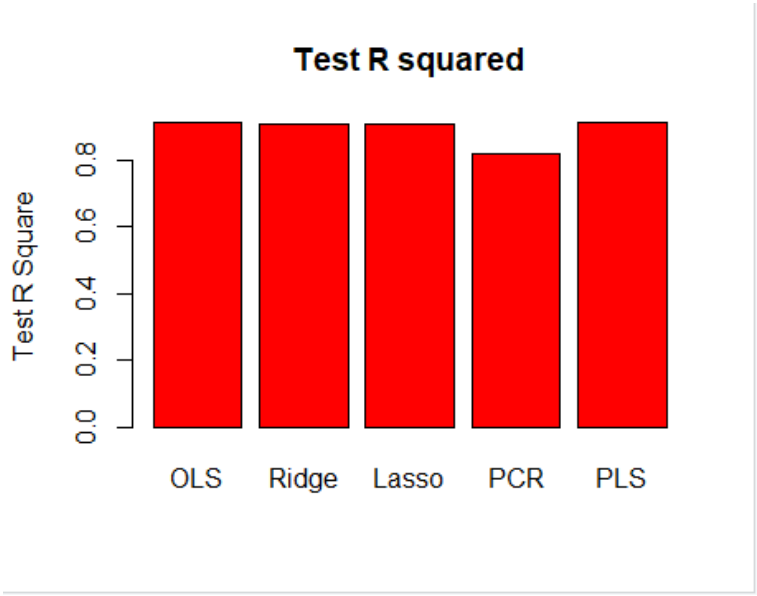


Figure 7: Test MSE for different models



R square loss.PNG

Figure 8: Test R^2 for different models

```
(Intercept) -567.27808449
(Intercept)    .
PrivateYes  -542.31432405
Accept         1.53109383
Enroll        -0.53619295
Top10perc     38.02102131
Top25perc      .
F.Undergrad    .
P.Undergrad    0.04658811
Outstate      -0.06045098
Room.Board     0.06194017
Books          .
Personal       0.06845198
PhD           -9.49657104
Terminal      -0.42003368
S.F.Ratio     14.08436598
perc.alumni    .
Expend         0.06475190
Grad.Rate      5.13255009
> 
```

shrink.PNG

Figure 9: Coefficients after performing lasso

2 Problem 2

In this problem we are given with insurance company benchmark data set from which we need to check whether we can predict caravan insurance policy by computing ordinary lest square method,forward selection ,backward selection,Lasso regression and ridge regression and compare them.

Knowing the data
The data set has three parts of data.The training data with 5822 observations and 86 variables including whether or not they have a caravan insurance policy and test data with 4000 observations and 85 variables and the target for the test data set.

Fitting the linear regression model
I had used the lm() function to fit the data set in linear regression model.Since because of caravan insurance policy is categorical linear model won't be that much efficient to find the output variable.When you see the coefficients of the variable is no variable is greater than 0.5.The highest coefficients value we can get for 0.4958051 V81 which is for AZEILPL Number of surfboard policies variables.The below picture will tell us contribution of each variable to the output variable and I had used round() to predict function and calculated the mean square error.The Mean square error for the linear model is

$$\boxed{MSE : 0.06025}$$

Fitting using Forward selection
I had performed best forward selection for the training data with nvmax = 85 and plotted against BIC and Cp.The model with least Cp error is for 23 and BIC it is 8.In these case no of variable where the Cp is minimum is different from no of variable where the BIC is minimum and calculated .The minimum mean square error which we had found out is around

$$\boxed{MSE : 0.05210329}$$

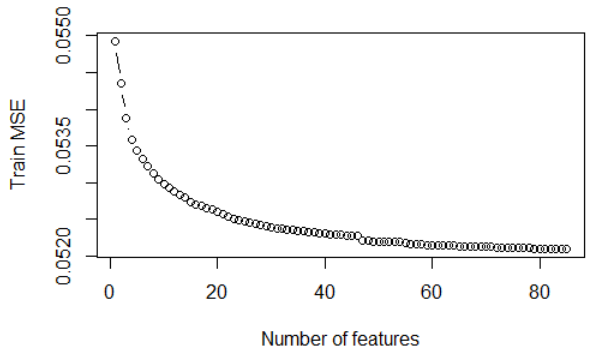
In these case mean square error for forward subset selection is almost same for linear regression model.The test error is also more or less same (i.e it is around 0.05727741) but in case of plotting graph between number of variables with test error and training error it is decreasing in case of training data set but in case of test error it is increasing

Fitting using Backward selection
I had performed best forward selection for the training data with nvmax = 85 and plotted against BIC and Cp.The model with least Cp error is for 29 and BIC it is 8.In these case no of variable where the Cp is minimum is different from no of variable where the BIC is minimum and calculated .The minimum mean square error which we had found out is around

$$\boxed{MSE : 0.05210329}$$

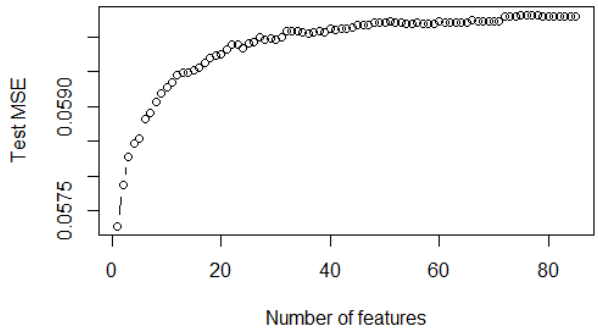
In these case mean square error for forward subset selection is almost same for linear regression model.The test error is also more or less same (i.e it is around 0.05746116) but in case of plotting graph between number of variables with test error and training error it is decreasing in case of training data set but in case of test error it is increasing.This can be seen from graph above

Fitting using Lasso regression model



subs select.PNG

Figure 10: Train error vs No.of.variables



forward subs.PNG

Figure 11: Test error Vs No.of.variables

sub train.PNG

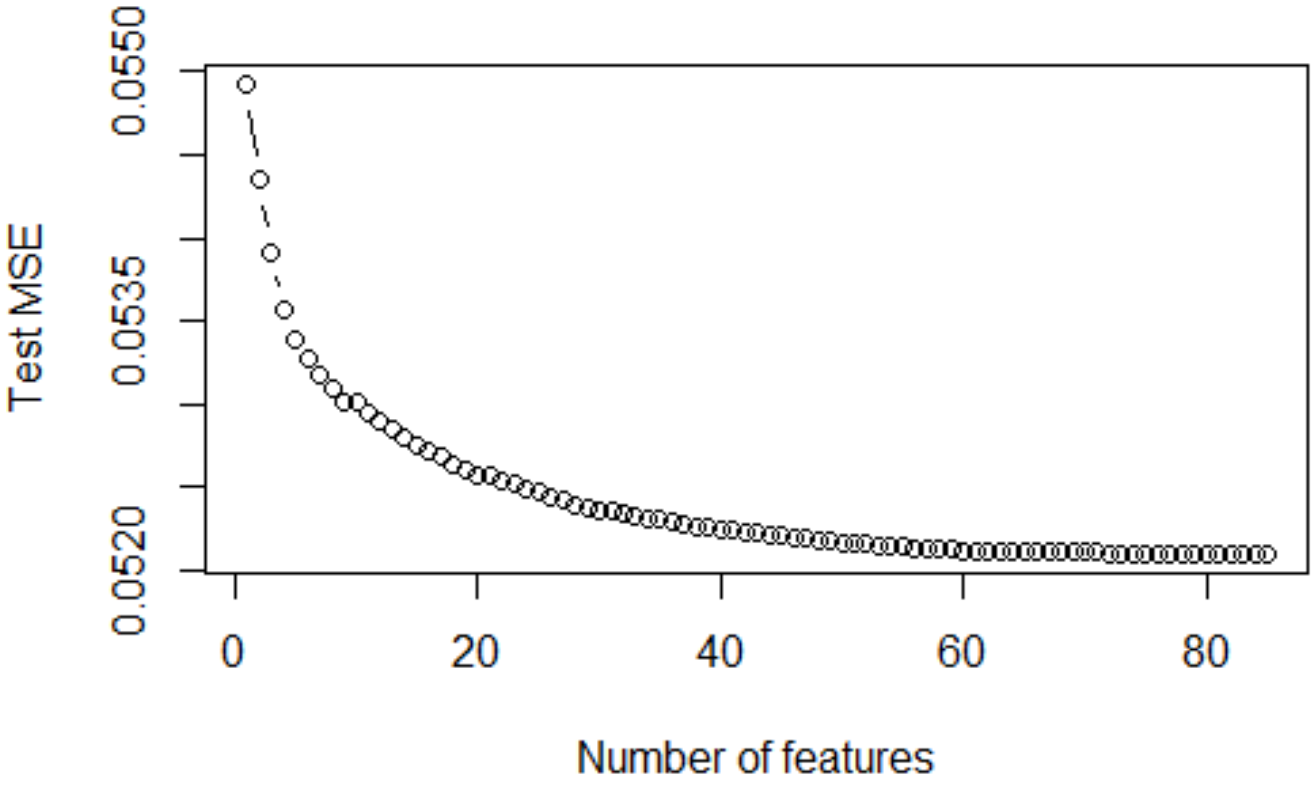
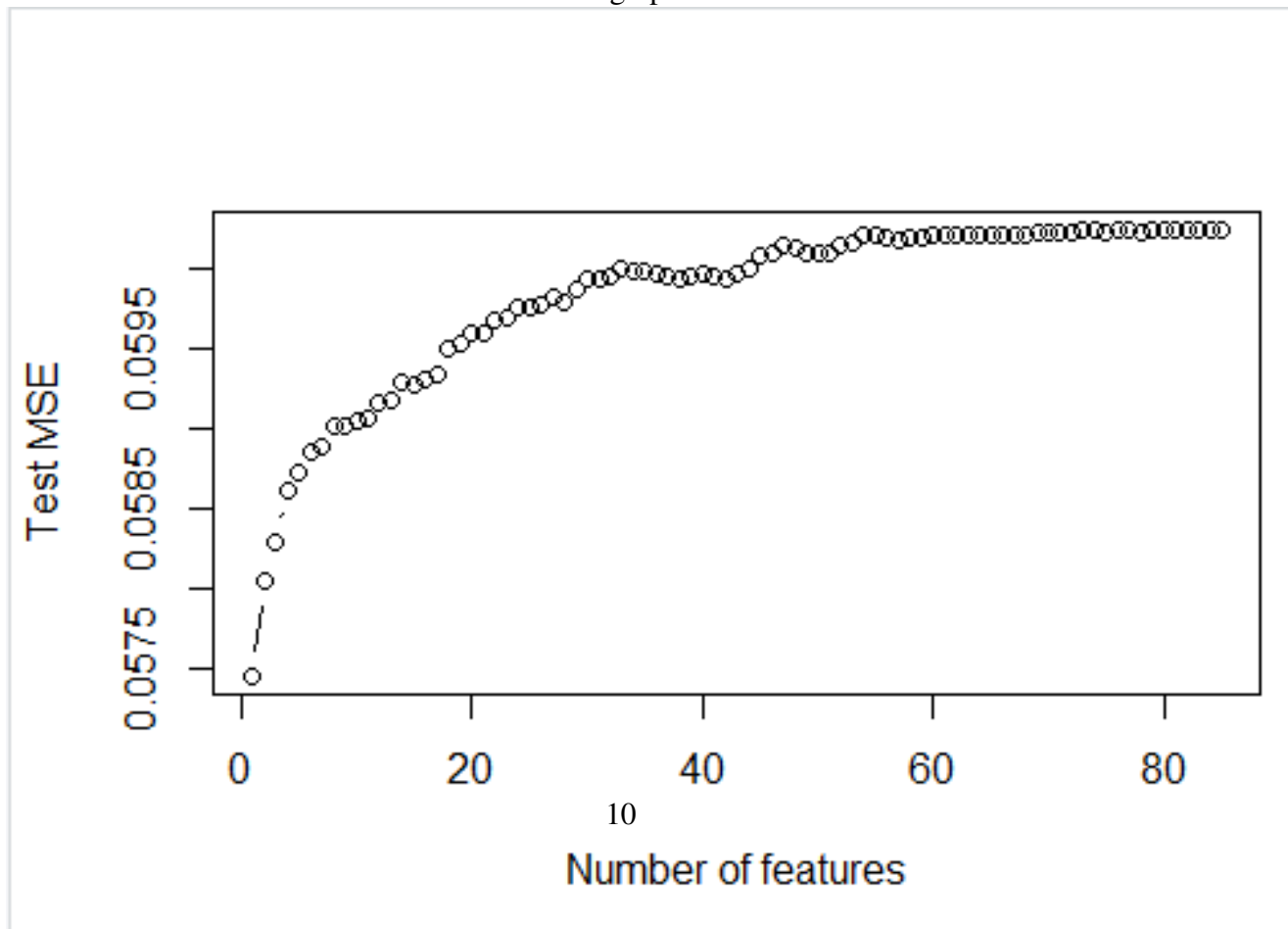


Figure 12: Train error vs No.of.variables

test error graph.PNG



ridge 2.PNG

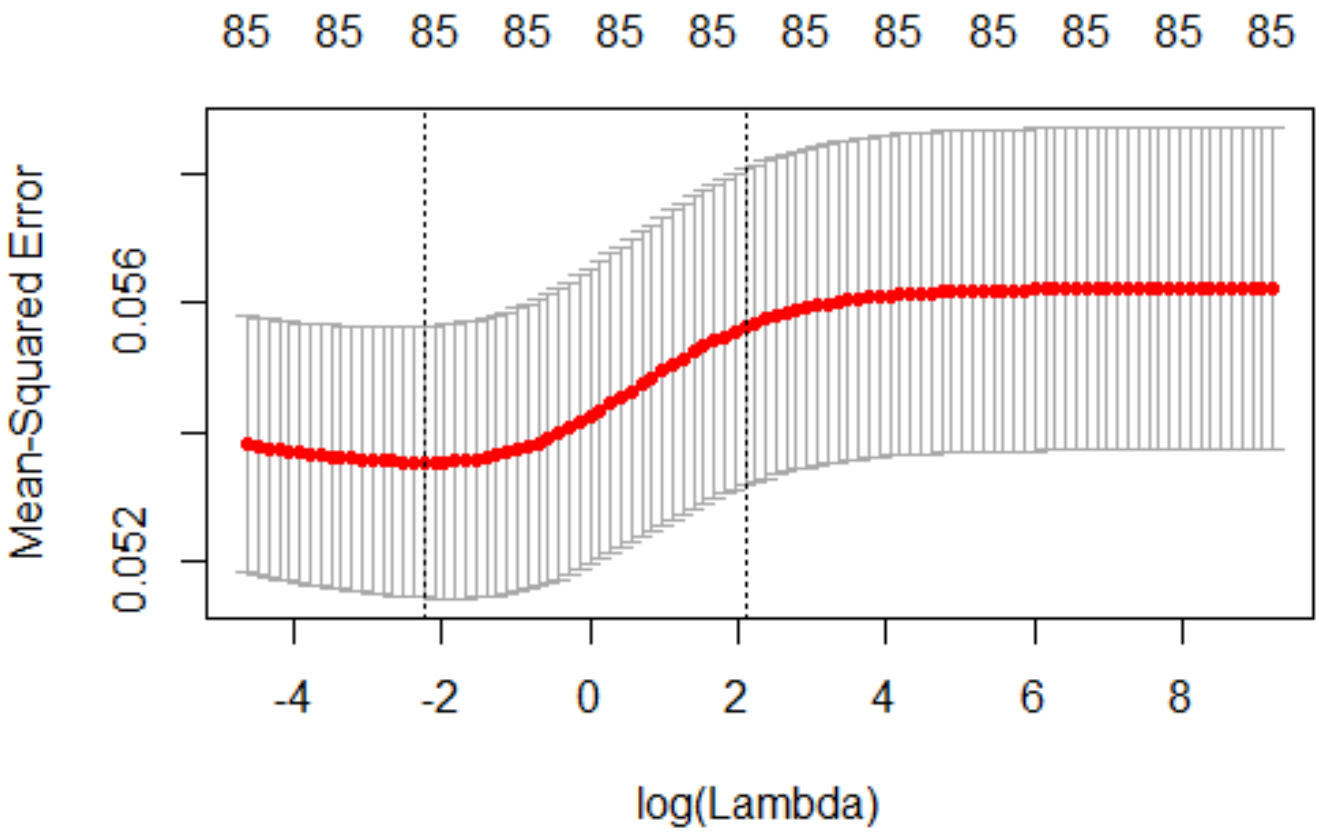


Figure 14: Lambda vs MSE in Ridge Regression

The Lambda chosen by the cross validation is around 0.1072267.The Mean square loss or MSE when using Ridge regression for predicting number of application is given by

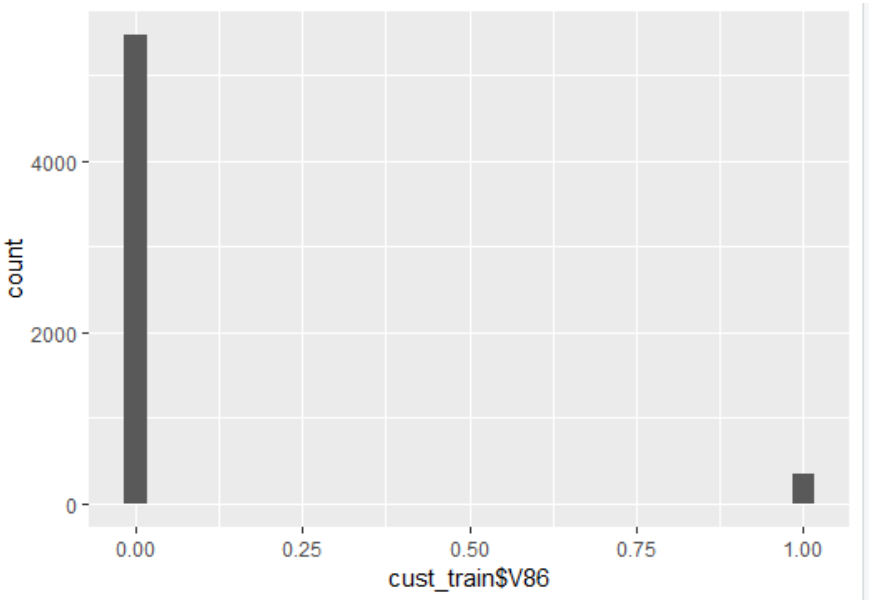
$$\boxed{MSE : 0.0536964176186785 RMSE : 0.0536964176186785}$$

The relation ship between Lambda and mean square error is examined below.From this we can infer that lambda minimum value is around 0.1072267

Fitting using Ridge regression model
The relation ship between Lambda and mean square error is examined below.From this we can infer that lambda minimum value is around 0.1245

$$\boxed{MSE : 0.0541868998326246 RMSE : 0.0536983105439907}$$

Comparison of different models
By applying the training data to different model like OLS ,forward subset selection,backward subset selection,ridge regression and lasso regression we had found almost in all of the models the mean square error is approximately 0.05.we can infer from this is that all model behave similar for this training data set.The error is minimum because of number of zeros in both test and train



and Ones.PNG

Figure 15: Distribution of zeros and ones

are very much higher than number of zeros so model will be predicting who are all not gonna take caravan insurance policy. when we had used linear model to fit the model we could not able to predict the variable that contribute to caravan insurance policy. All of the coefficients are lesser than 0.5.When using the forward subset selection the subset with only 1 variable had given the minimum error,So we can say that one variable is significant contributor for the output variable. Similarly for the backward subset selection also the subset with only 1 variable had given the minimum mean square error.so we can't say that one variable is the best predictor of my output variable.When we apply our training data into lasso and ridge regression the maximum coefficient value for the particular variable is 0.45301.So by this method also we can't able say the significant contributor for the output variable.This is because of the reason that there were more number of zeros in the training data.The models which are used to predict the insurance policy was not that much efficient.So we need to use some other method like zero inflation to predict the output variable.

Why can't we able to predict?
When we look into the output variable of the training data we can infer that number of 1's is very much lesser than number of 0's.This we can find by both graphical and also mathematically From the graph we can say that training data is skewed towards zeros (i.e There were many number of zero compared to one in the target variable) and when we sum the target variables we get 348 from this we can say that there were only 348 ones out of 5822 observation.The probability of getting the one from the target variable is 0.05977327.So it is almost zero.So we can say that Who won't be taking caravan insurance policy ?.But we can't able predict who will be taking caravan insurance policy ?.Since we have large number of zeros we can able be find who won't taking insurance policy.By Applying the given model we can't able be predict the output with given model and given training data.We can use some complex algorithm like c50,Zero R,R part in this case of data.

3 Problem 3

In this problem we need to show that as the number of features increases the corresponding training error will decreases while test error increases by creating random matrix of 1000 observations and

of variables with diff error.PNG

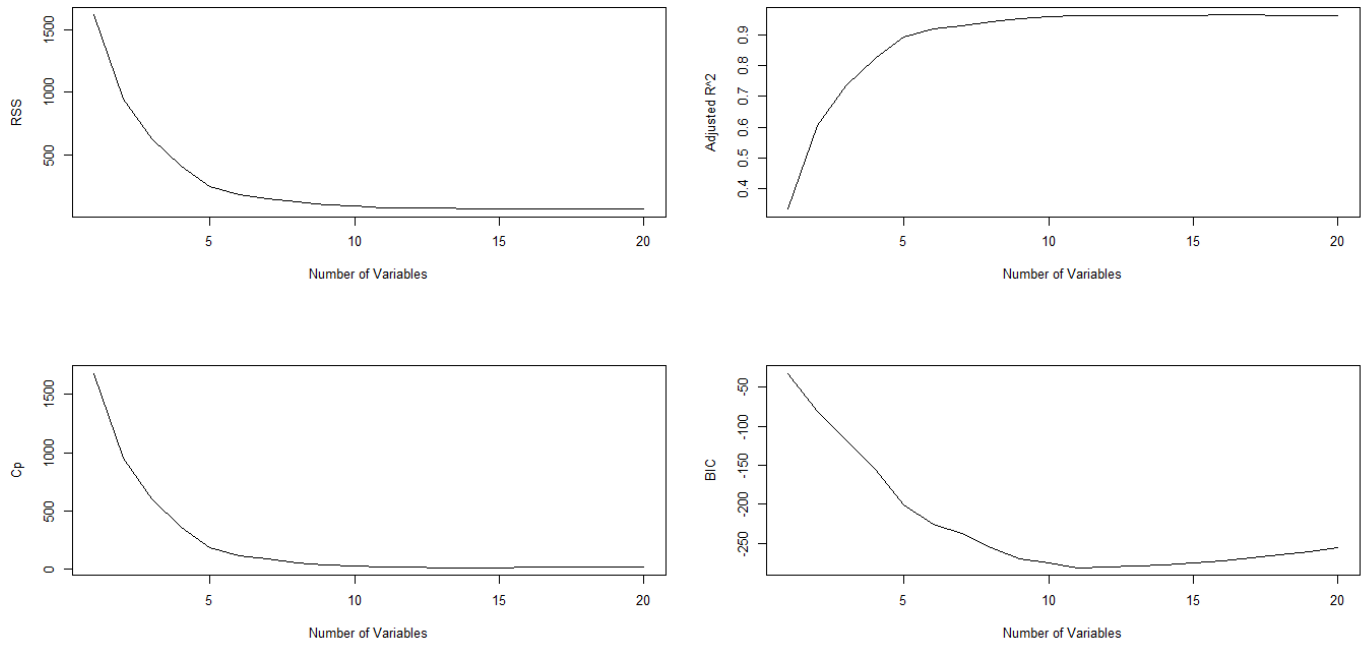


Figure 16: No of variables VS adjusted R^2,Cp,BIC,RSS

20 columns and we need to perform best subset selection on the training set and plot the training set MSE associated with the best model of each size

Knowing the data

I had created random data set which has 1000 observations and 20 features with them and then associated response variable can be found out based upon the formula

$$\boxed{Y = X\beta + \epsilon}$$

so now the data set has 21 variables since response variable is added.Then the data set can be spitted into test and train data for both response variable and features.So now my train data will have 100 observations with 21 variables and test data will have 900 observation with 21 variables.It was mentioned that some beta values(coefficients) will be zero in given question some i had randomly made 6 variables as zero

Performing Best Subset Selection

I had performed best subset selection for the random training data set with nvmax = 20 and plotted No of variables with Cp,BIC,RSS,and Adjusted R^2 and got certain inference.With no of variables against Cp we infer from the graph that it min for 13 and its also agree with the BIC.I had attached the graph below for inference,which can also confirmed with minimum value of Cp and BIC error.

Plotting Training Error against Each Predictors

When plotting no of variables with increasing order of variables we infer that training error is decreasing with increase in number of variables.This can also be seen from the attached picture above.After 13 number of variables we can see that it is almost constant.The graph also linearly decreasing.The Train error is almost zero(i.e 0.80) in case of training error at 20_{th} variable

of best subset.PNG

```
> #They both agree model with 14 variables is the best.
> which(my_sum$cp == min(my_sum$cp))
[1] 13
> which(my_sum$bic == min(my_sum$bic))
[1] 13
>
```

Figure 17: Number of Best Subset

error 20 variables.PNG

```
> print(paste("Test Error: ",training_errors))
 [1] "Test Error:  17.0578423851876"  "Test Error:  12.1998139588"     "Test Error:  8.85690557221261"
 [4] "Test Error:  6.74703602697546"  "Test Error:  4.79544648043705"  "Test Error:  3.18110775871107"
 [7] "Test Error:  2.35298010379192"  "Test Error:  1.56808980646297"  "Test Error:  1.22835674486035"
[10] "Test Error:  1.12751265942131"  "Test Error:  1.03937152052114"  "Test Error:  0.918664355173933"
[13] "Test Error:  0.855087561399191" "Test Error:  0.840669703244891" "Test Error:  0.829380924076383"
[16] "Test Error:  0.823244623995721" "Test Error:  0.819456453788297" "Test Error:  0.817893901178903"
[19] "Test Error:  0.816845451717503" "Test Error:  0.81644560557446"
```

Figure 18: Train error for 20 variables

Plotting Test Error against Each Predictors

When plotting no of variables with increasing order of variables we infer that test error is decreasing with increase in number of variables but not like training error.In case of training error the value is around 0.80 for the 20_{th} variable but in case of the test error the error for the 20_{th} variable is 1.27,so we can say that training error will decrease with increase in number of variables but the test error won't decrease much like training error.Which can be shown in below two pictures

Finding Model size for Minimum MSE in test set

It can be seen from the above graph that test error is minimum around 13 to 14.It can also be said that by minimum of test error, which is attached below

$$\boxed{MinimumTestError : 14}$$

It can be said that 14_{th} variables model has the least test error.Since MSE for the 14_{th} variables is minimum we can say that best model is the one with 14 variables.

Finding the coefficient value for best Model Size

The coefficients for the best model is given that the coefficients of the variable X5,X7,X9,X12,X13 these because of Beta value where the random number generated was 5,7,9,12,13 so these coefficients will become zero.It can be observed from the picture shown that true value coefficients is zero for 5,9,7,12,13 variables and also value of coefficients is almost same for the both predicted and true value.But in case of variable 4 where its value is 0.06976926 in true model,its get neglected in predicted model where its value is almost zero.

train error.PNG

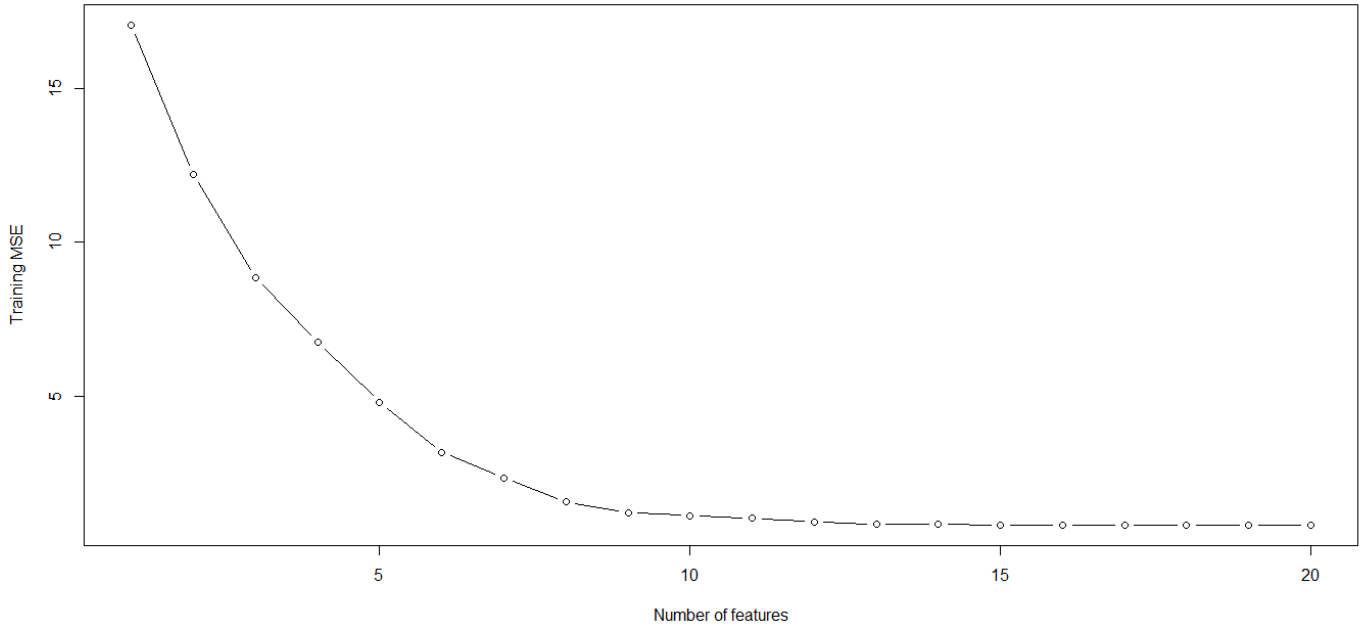


Figure 19: Plot for Train error against different variables

for 20 variables.PNG

```
 [1] "Test Error:  13.8005670772377" "Test Error:  10.7110086333732" "Test Error:  7.2988911655483"
 [4] "Test Error:  4.45185127700651" "Test Error:  2.77241714417196" "Test Error:  2.29066577227981"
 [7] "Test Error:  1.96285695516303" "Test Error:  1.88965121245588" "Test Error:  1.5771840181351"
[10] "Test Error:  1.53114473637691" "Test Error:  1.34106141947725" "Test Error:  1.21322416455985"
[13] "Test Error:  1.24722678670251" "Test Error:  1.18084500177611" "Test Error:  1.20811262653894"
[16] "Test Error:  1.26069287614191" "Test Error:  1.29340943352365" "Test Error:  1.28526863734286"
[19] "Test Error:  1.28244199830924" "Test Error:  1.27940262253444"
```

Figure 20: Test error for 20 variables

graph for 20th variables.PNG

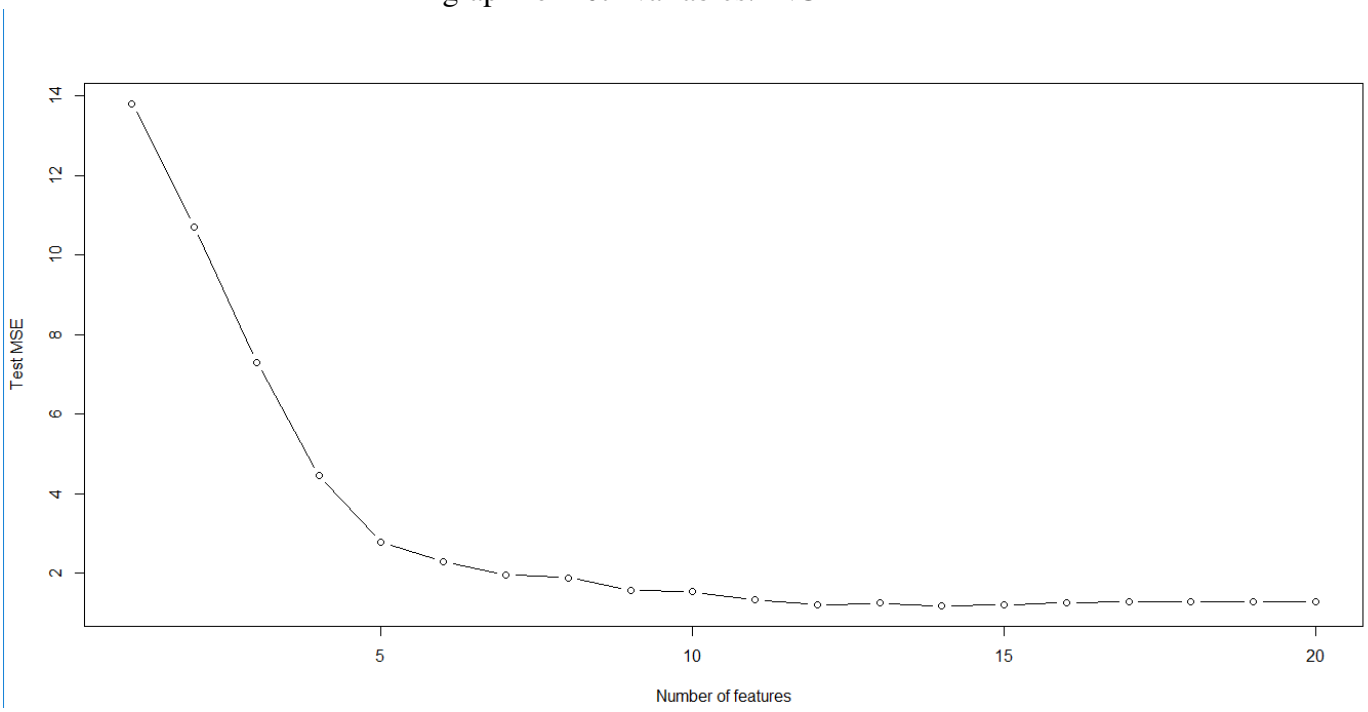


Figure 21: Plot for Train error against different variables

test error.PNG

```
> print(paste("Minimum Test Error: ",which.min(test_errors)))
[1] "Minimum Test Error:  14"
```

Figure 22: Minimum Test Error value

values.PNG

```
> coef(regfit.best, id =which.min(test_errors))
(Intercept)          X1          X2          X3          X6          X8         X10         X11         X14
 0.06246403  1.30812654  0.23616907  1.68616256  2.24844437 -1.21954189 -0.59973884  0.93142627 -2.04414101
        X15         X16         X17         X18         X19         X20
 0.37933449  1.89207877  0.40189547 -1.06847901 -0.14066442  0.37298555
```

Figure 23: Coefficients for best Subset

model coefficient.PNG

```
 [1]  1.29618225  0.13221140  1.68644940  0.06976926  0.00000000  2.32620063  0.00000000 -1.15034663
 [9]  0.00000000 -0.46742720  0.84008947  0.00000000  0.00000000 -1.90933689  0.44854151  2.09251593
[17]  0.36724214 -0.90398370 -0.07502126  0.44925150
```

Figure 24: True Model Coefficients

zero.PNG

```
> beta.zeros
[1] 13  5  9 12  7
```

Figure 25: Beta Zero Values