HOMEWORK 2

STATISTICAL DATA MINING

EAS 506

NAME : SHUBHAM SHARMA

PERSON # : 50290293

CLASS # : 44

PROFESSOR : DR. RACHAEL H. BLAIR

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

(a) Split the data set into a training set and a test set. Fit a linear model using least squares on the training set, and report the test error obtained.

(b) Fit a ridge regression model on the training set, with λ chosen by cross- validation. Report the test error obtained.
(d) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

(e) Fit a PCR model on the training set, with k chosen by cross-validation. Report the test error obtained, along with the value of k selected by cross-validation.
(f) Fit a PLS model on the training set, with k chosen by cross validation.
Report the test error obtained, along with the value of k selected by cross-validation. (g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?

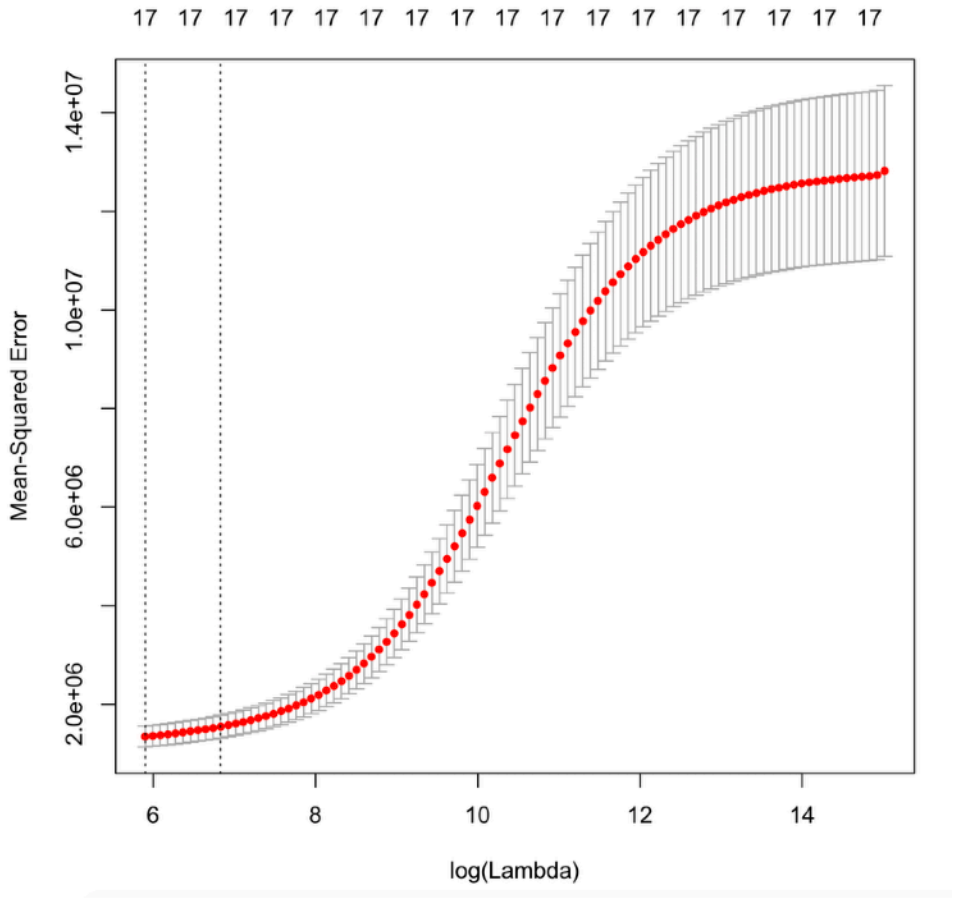
Linear Regression

The given dataset has 777 observations and 18 variables. The variable Private is categorical while rest are quantitative variables. The response variable is the number of applications received, i.e. Apps. The given data set is converted into training and test data with the training set having 388 entries while the test having 389 entries. Linear Regression was performed on the training data and the obtained model was used to predict the response on the test data. The **test set MSE** was found to be **1235823**.

Ridge Regression

While performing ridge regression we notice that none of the coefficients are set to 0 for any lambda. We also notice that if we try to predict coefficients using predict function and setting s (i.e. λ value) to 100 (which is smaller than the 100th λ value i.e. 332.1998) the predicted coefficients are pretty much the same as what we got for the latter. On comparing these coefficients with the ones obtained from the linear model, we observe that the coefficients are somewhat close to the predicted coefficients.

In the graph there are 17 degrees of freedom which are constant while performing ridge since nothing is being shrunk to 0. The **best λ** obtained was **364.5889** which is corresponding to the first dashed line in the above graph. The **test set MSE** for the ridge regression model using the best λ value was found to be **2016168** which is more than the one obtained using OLS.

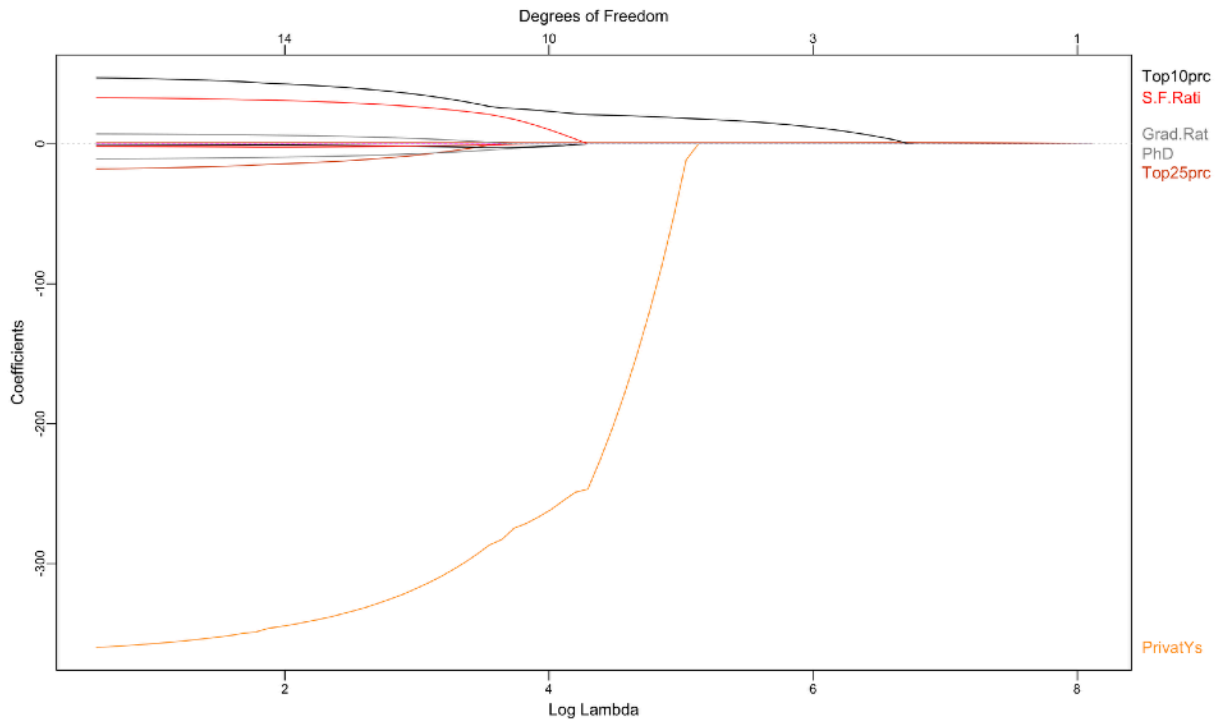


Plot after cross-validation was performed on the train data for Ridge

LASSO

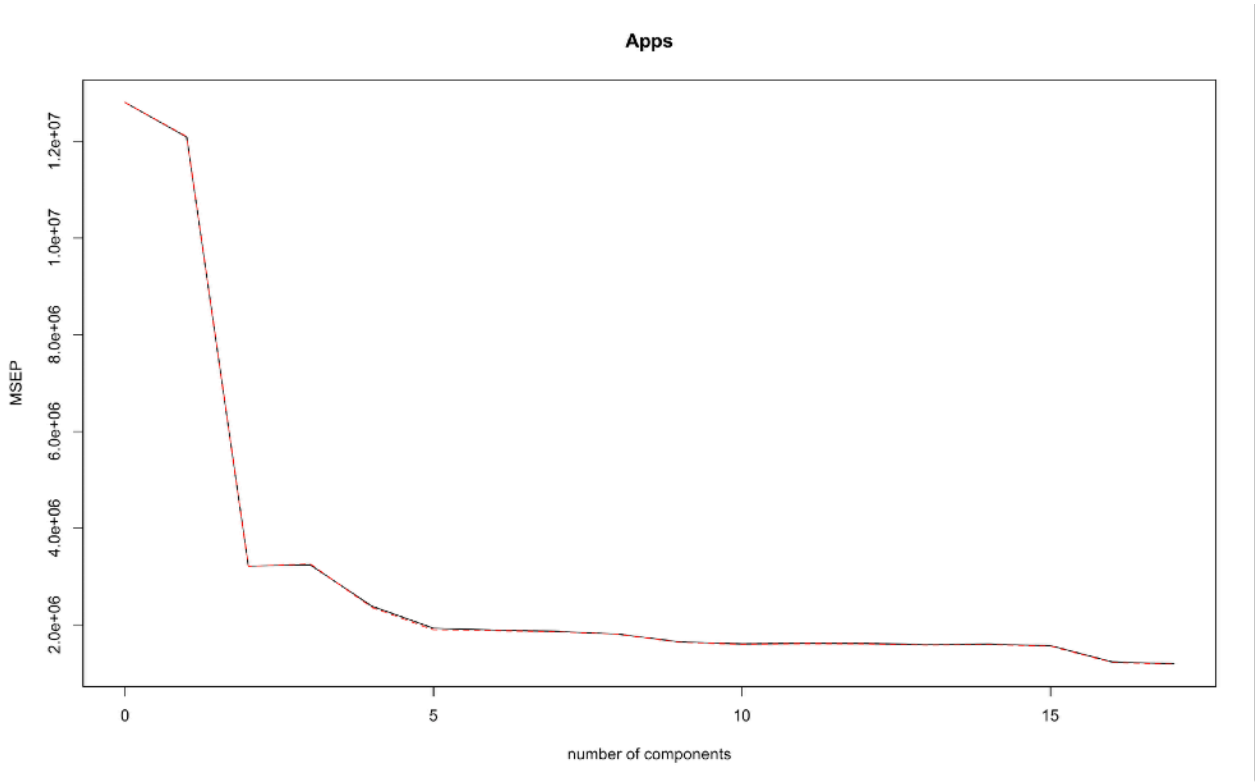
On fitting the lasso model on the training set, we get the **best λ** as **13.72654**.

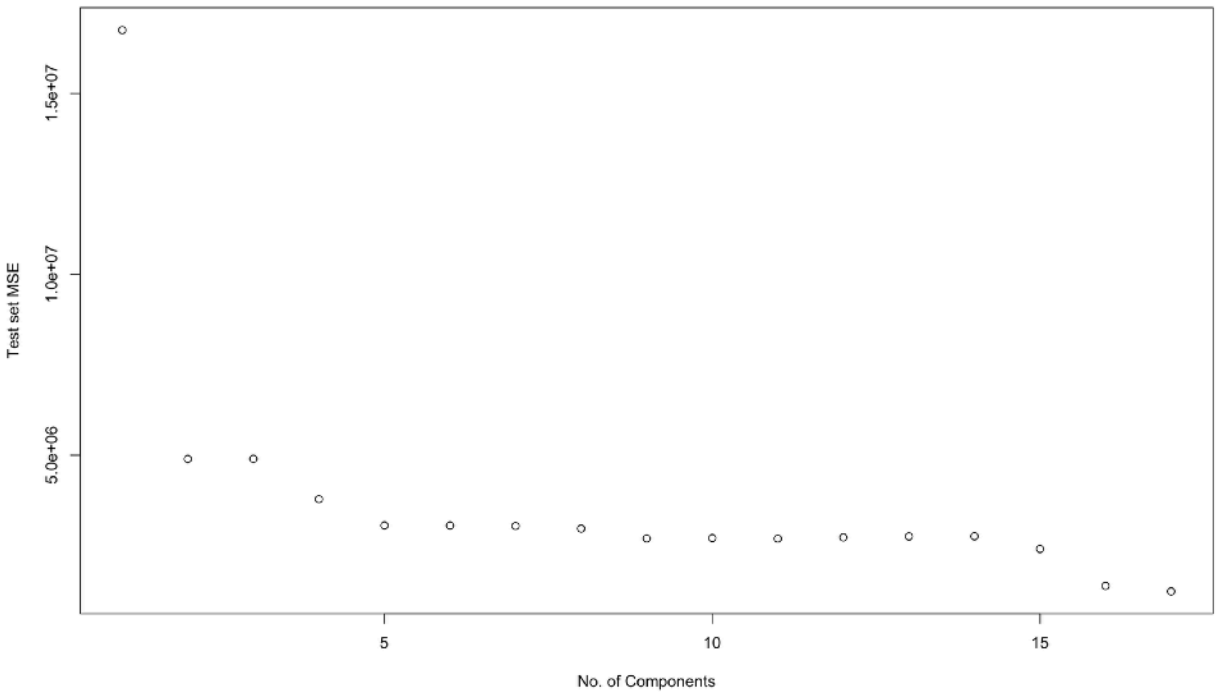
We observe that some of the variables are lost (Enroll, Books, Personal) and their coefficients are shrunk to 0 in our prediction model which comes from LASSO as is apparent in the graph. The test set MSE comes to be 1284445 which is slightly higher than the one obtained in OLS. The number of non-zero coefficient estimates are 14 excluding the intercept.



Plot showing shrinkage of variables in Lasso

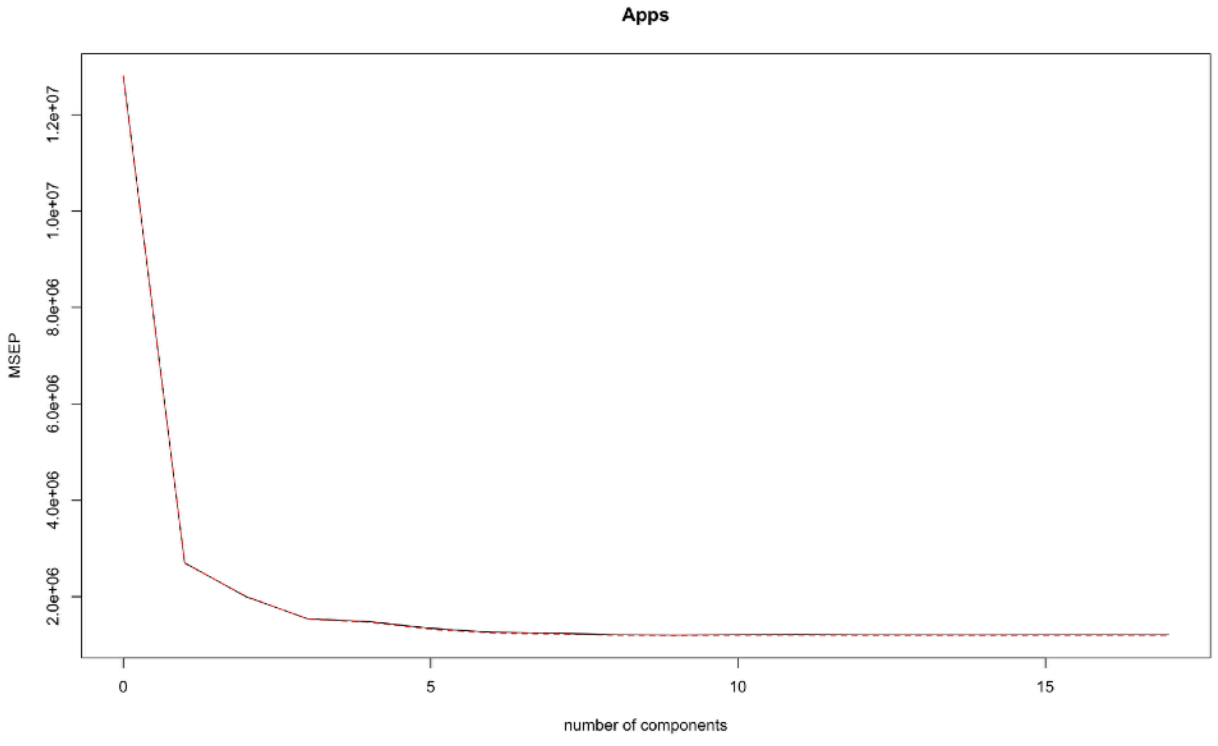
Principal Component Regression

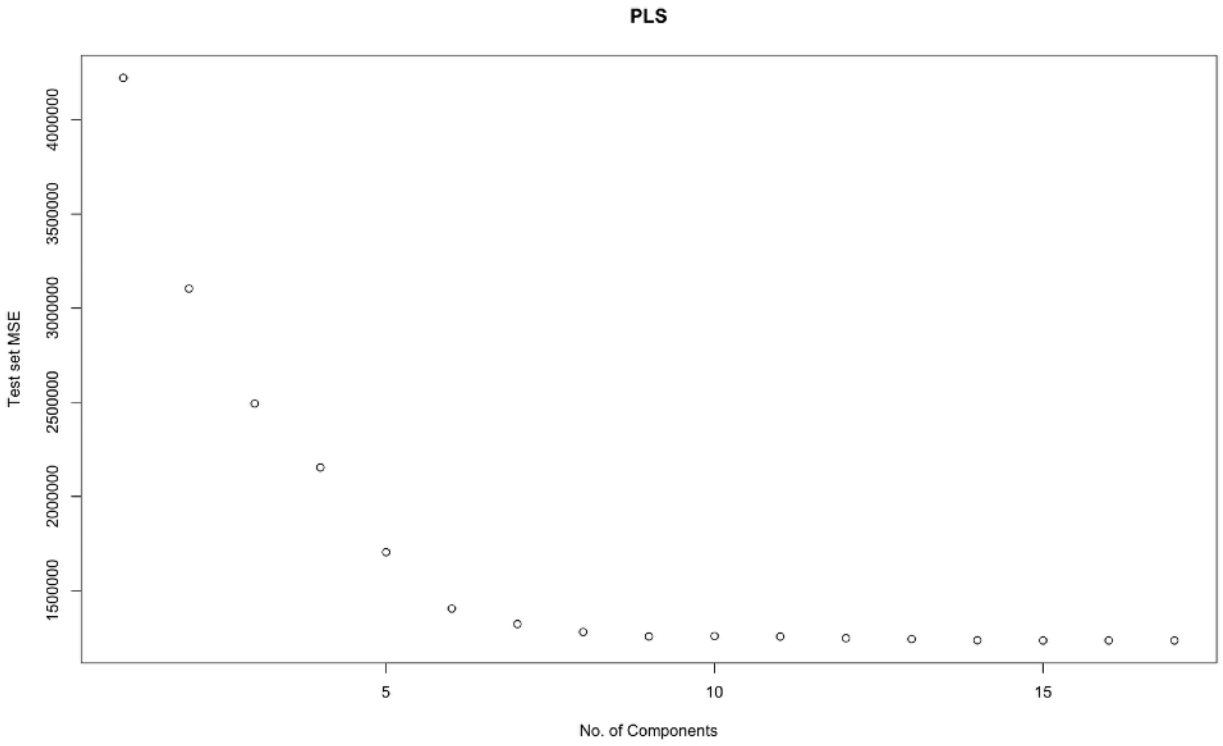




As is clear from the above graph, the test error decreases as the number of components increases but after 10 components, the decrease in error is not significant until 15 components. Thus, to keep the model simple we can take 10 components and this will reasonably give a good model fit using PCR. The **test set MSE** obtained for 10 components is **2709616.**

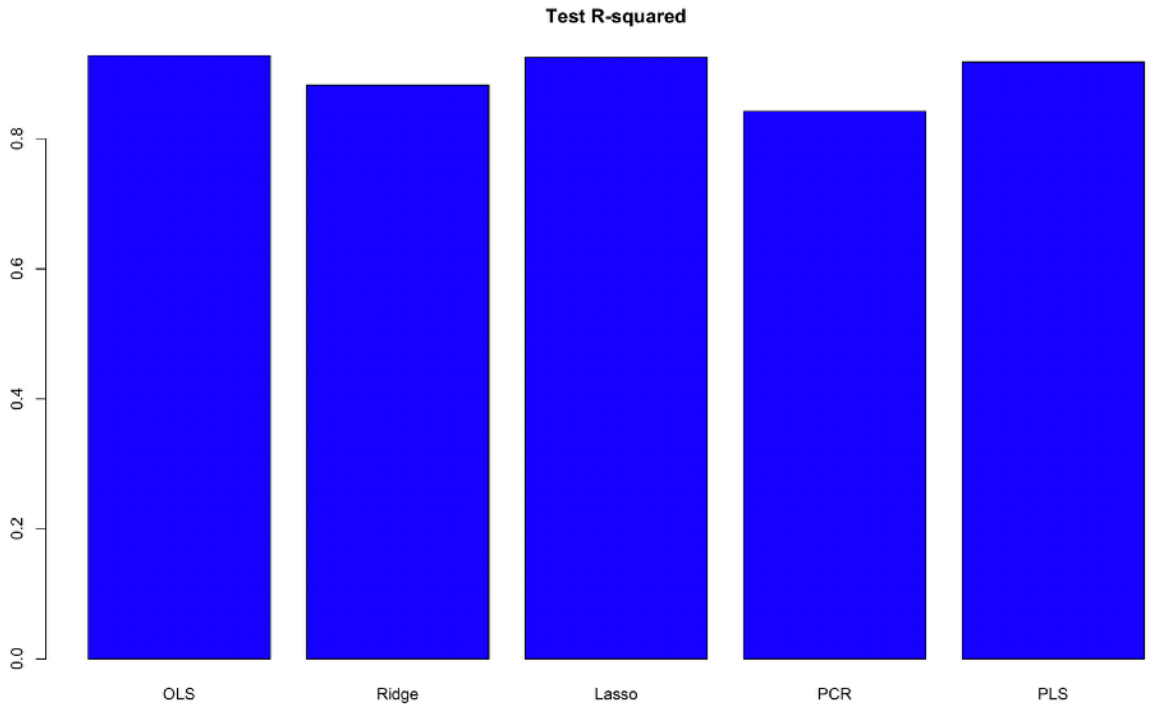
PARTIAL LEAST SQUARES





In the PLS model as well, we can see after 6 components there is no significant reduction in the test set MSE. Thus, we can reasonably assume that **6 components** can fit the model well and the **test set MSE** obtained is **1405180**.

We compute the R^2 for all the cases since it is independent of the scale of Y and would lie between 0 and 1 which is shown in the following plot.



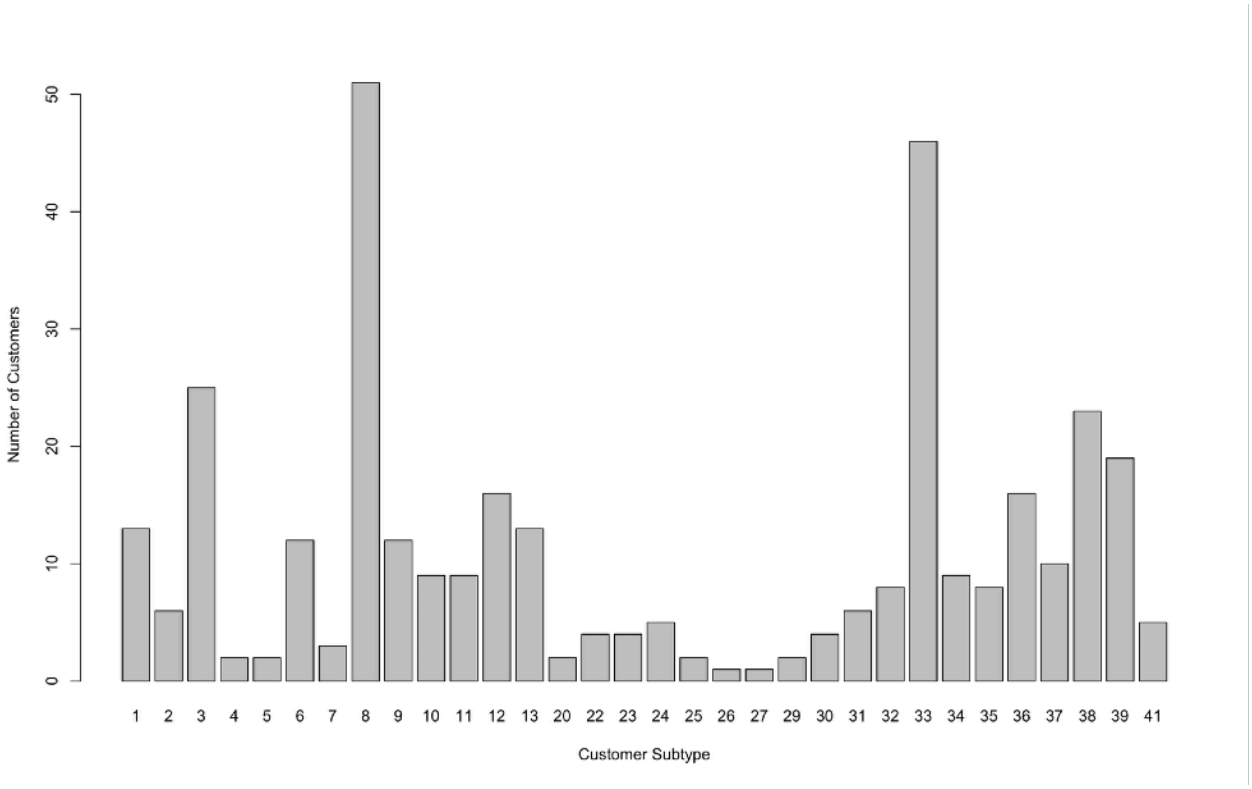
It is clear from the above graph that we get the lowest R^2 for the PCR model and thus we can say that the accuracy for that model will not be as good as the other model. The models for which we get R^2 closest to 1 are OLS, Lasso and PLS with ridge also not far behind thus we can say that these models will be able to predict the number of college applications received with reasonable accuracy.

2) The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and socio- demographic data derived from zip area codes. There are 5,822 customers in the training set and another 4,000 in the test set. The data were collected to answer the following questions: Can you predict who will be interested in buying a caravan insurance policy and give an explanation why? Compute the OLS estimates and compare them with those obtained from the following variable- selection algorithms: Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression. Support your answer.

We observe from the training data that the number of customers who have purchased the Caravan policy is 348. Following observations were made after analyzing the training data:

a. Out of the customers who purchased the Caravan policy, majority of them did not purchase the boat policy(335) and the social security insurance policy(332). Thus, customers who do not purchase these policies have more chances to purchase the Caravan policy.

b. If we do a similar comparison with the contribution to car policies, life insurance and family accident insurance policies: we find that the customers who belong to the 6th category in L4(i.e. pay a premium from $1000 to $4999) for car policies and who do not have life and family accidents insurance policies would have more chances to buy the Caravan policy.

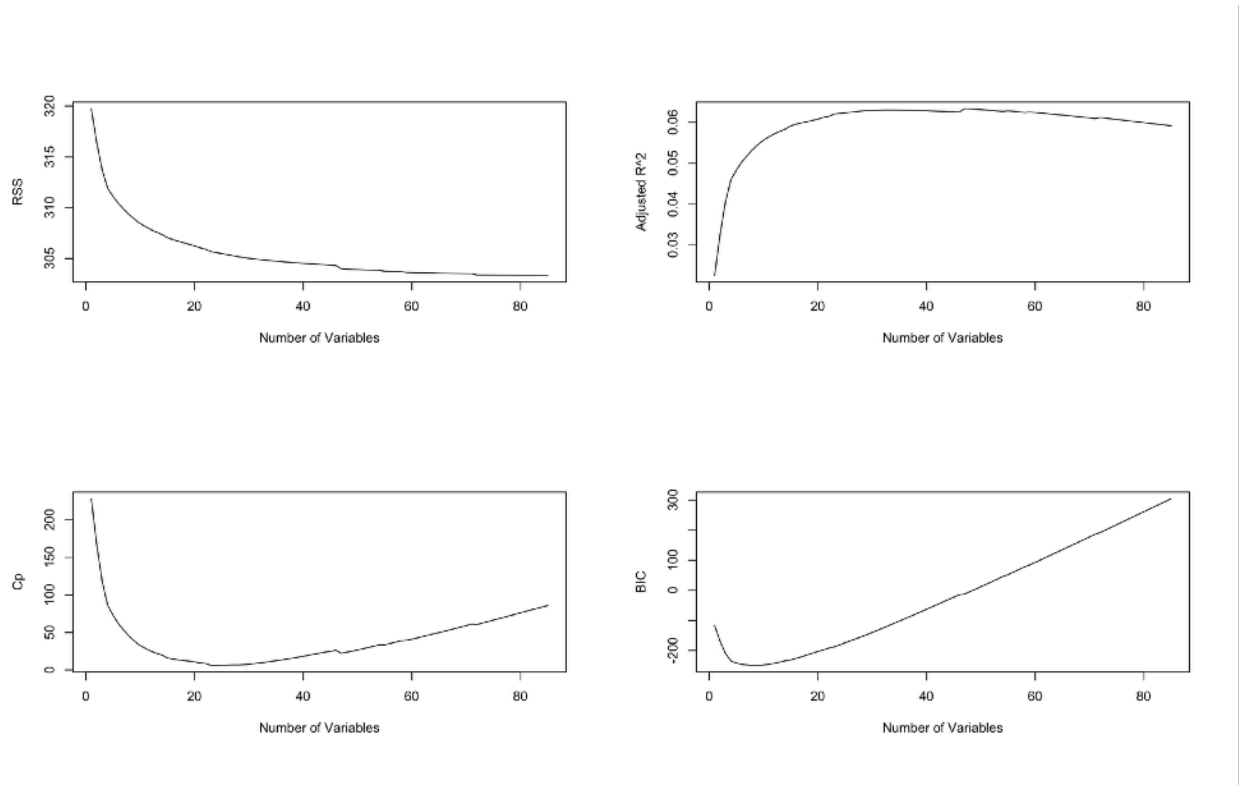


c. From the above graph it is clear that, customers who belong to value label 8 in L0 (i.e. Middle Class families) and label 33 (i.e. Lower class with large families) are more in number in the above distribution which considers the customers which have purchased the Caravan policy. This shows that customers belonging to these categories would more likely purchase the Caravan policy.

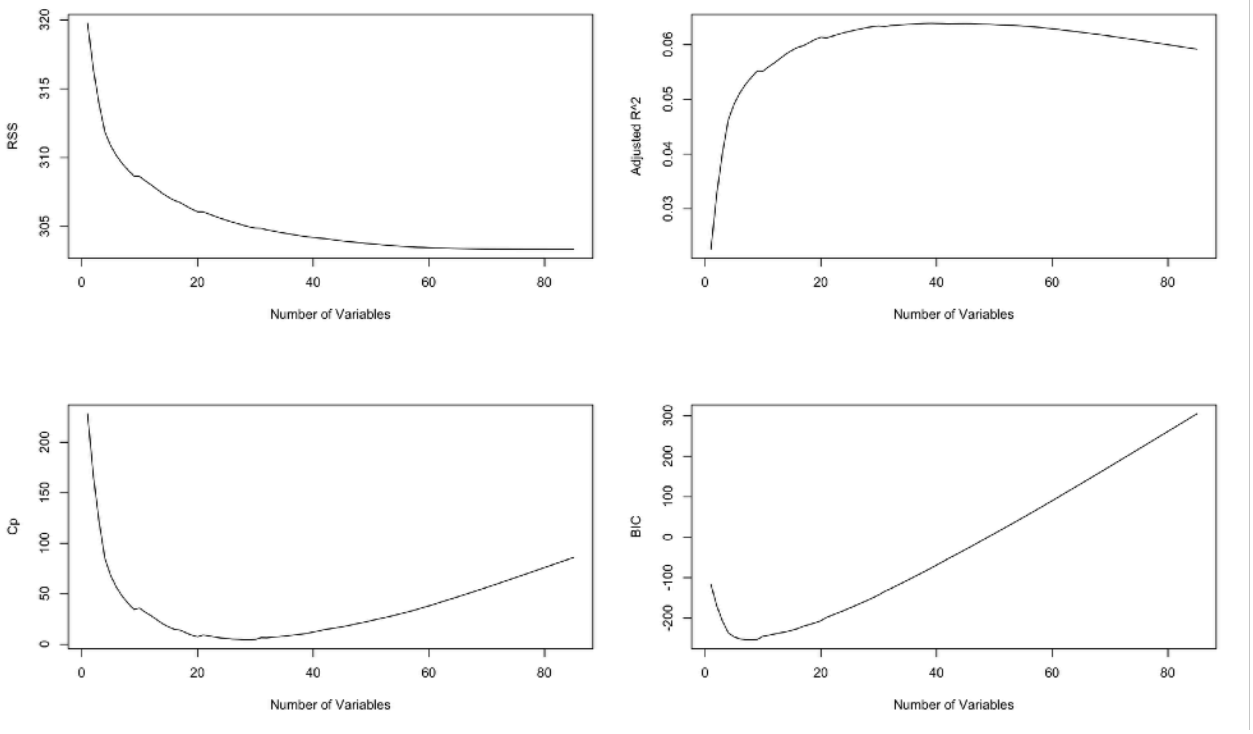
d. On observing the average age of the customers who have bought the Caravan policy, we find that majority of them belong to the age group of 40 to 50.

e. For Average Income, we find the the majority of Caravan policy buyers lie in the average income range of $200 to $999 and especially the ones who lie in the range of $200 to $499 have the most chance to buy the policy.

Below is the graph for forward subset selection:



It is clear that the as the number of variables are increasing in the model the RSS is decreasing while the adjusted R^2 is increasing. The model for minimum Cp comes out to be the one with 23 variables for forward and 29 variables for backward selection and for minimum BIC it is one with 8 variables for both.



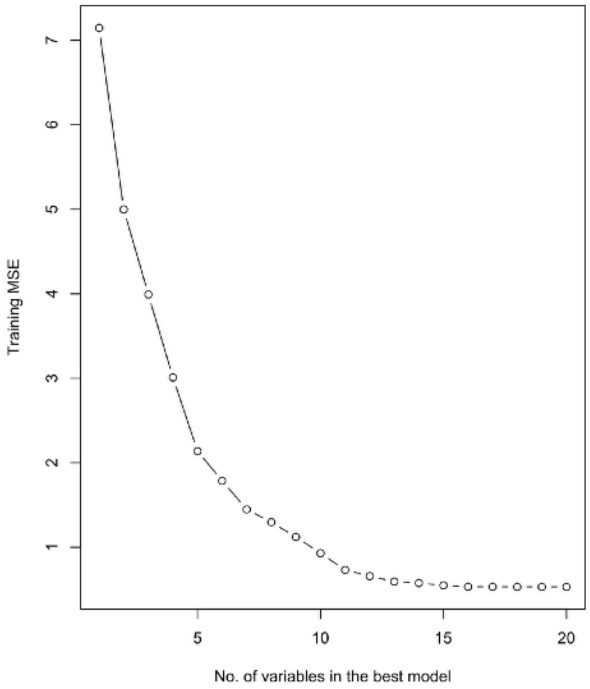
The test set MSE for OLS, Ridge and Lasso are 0.05975, 238 and 214.9875. We observe that both the forward and the backward selection methods have included almost the same variables. We note that the training data provided to us has very less positive responses to the response variable in comparison to the negative response. Thus we don't have a proper representation of the data to make useful prediction which is clear in the disagreement of the mip model size of BIC and Cp. Also, we note that all the coefficients are shrunk in our ridge and lasso model for the same reason. Thus, it is difficult to make predictions using the given models from the given data set.

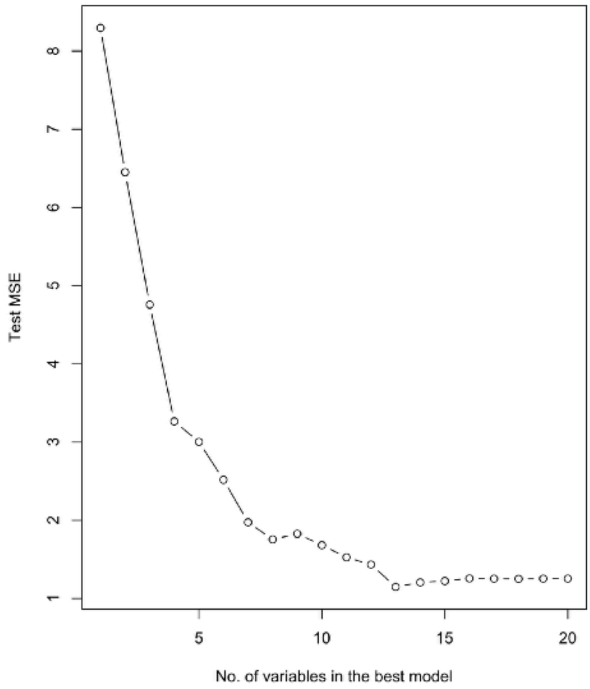
3) We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.
Generate a data set with p = 20 features, n = 1, 000 observations, and an associated quantitative response vector generated according to the model
$Y = X\beta + \varepsilon$
where β has some elements that are exactly equal to zero. Split your data set into a training set containing 100 observations and a test set containing 900 observations.

Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size. Plot the test set MSE associated with the best model of each size.

For which model size does the test set MSE take on its minimum value? Comment on your results. How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

The data was generated as random deviates with mean 0 and standard deviation 1 and filled in a matrix of 1000*20 row-wise. The coefficients and the error term was also generated in a similar manner. Thereafter, the 1st, 6th, 11th, 15th, 16th and 19th coefficients were set to 0. The associated quantitative response vector was generated and the data was split into 100 training observations and 900 test observations.





The test set MSE take on its minimum value for the model having **13 variables**. We observe that as the number of variables are increasing the training set MSE is continuously decreasing while the test set MSE picks up at some places and then decreases. These models show such behavior as in this case the relationships are overfitted and the model is working more than is required to fit the training data. Due to this the relationship which the model is finding in the training data is not found in the test data and thus we get a larger test error.

On looking at the coefficients of the 13 variable model we find that the coefficients which we have not included in the original model have also not been included according to the prediction of our best subset selection method(which we should have expected). Thus the best subset selection method is able to predict the true relationship between our predictors and response with reasonable accuracy.