UB#50291566 SIDDIQ SYED

Homework-2

Question – 1:

(10 points) (Exercise 9 modified, ISL) 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.

Test mean square error for the linear model using least squares is 1147256

(b) Fit a ridge regression model on the training set, with λ chosen by cross validation. Report the test error obtained.

Test mean square error for the ridge regression model is 1131495

(c) 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.

Test mean square error for lasso model is 1171037

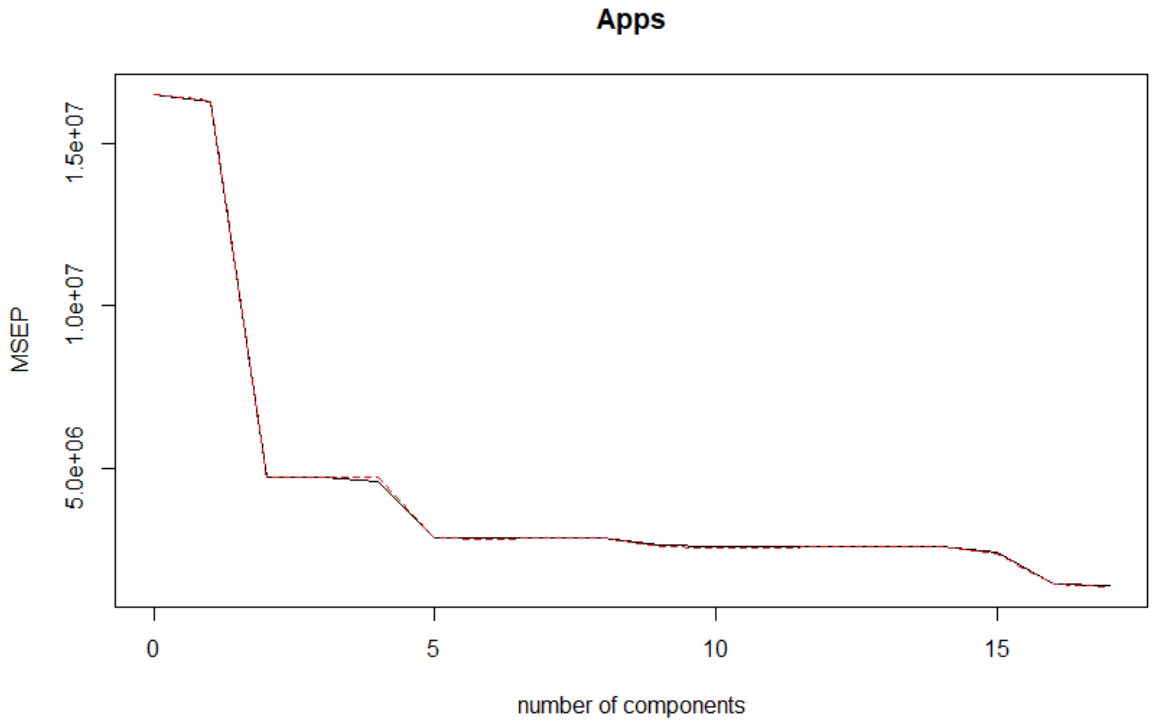
Below are the non-zero coefficient estimates

```
> predict(lasso_fit, s=lasso_bestlambda, type="coefficients")
19 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) -732.00391408
(Intercept)    .
PrivateYes  -260.10475305
Accept         1.40661052
Enroll         .
Top10perc     26.97710806
Top25perc      .
F.Undergrad    .
P.Undergrad    .
Outstate      -0.01791771
Room.Board     0.07821447
Books          .
Personal       .
PhD            .
Terminal      -4.49972646
S.F.Ratio      .
perc.alumni   -1.72122493
Expend         0.04836470
Grad.Rate      1.62754097
```

(d) 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.

Test mean square error for PCR model is 1525762

Below is the plot for MSEP and Number of components.

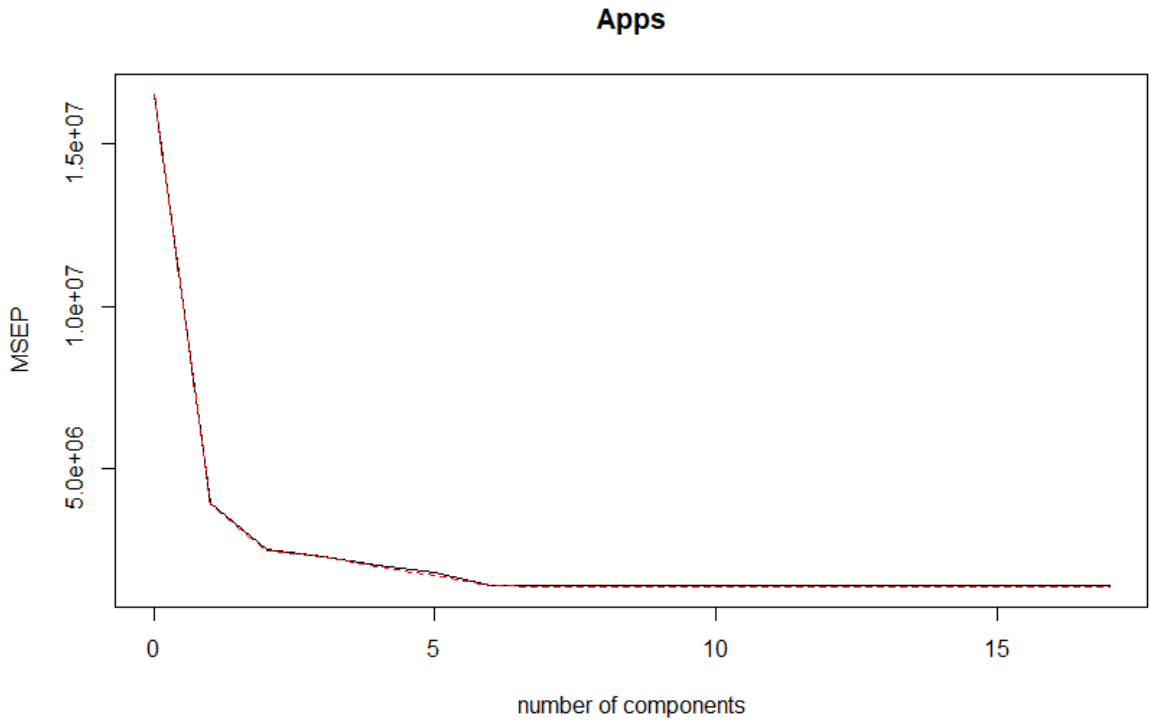


From the above graph we are taking the number of components for prediction to be 10 to test the model fit.

(e) 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.

Test mean square error for the PLS model is 1149496

Below is the plot for MSEP vs Number of components.



From the above graph we are taking the number of components for the prediction to be 10 to test the model fit.

(f) 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?

To compare the above results, computing the R^2 for above 5 models.

```
> R2_LM
[1] 0.8901095
> R2_ridge
[1] 0.8916192
> R2_lasso
[1] 0.8878316
> R2_pcr
[1] 0.8538542
> R2_pls
[1] 0.889895
```

From the R^2 values above it is clear that the all 4 models have close accuracy except for the PCR model with lesser accuracy for the taken training and test data.

Question 2:

(10 points) The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and sociodemographic 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. (The data can be downloaded from https://kdd.ics.uci.edu/databases/tic/tic.html.)

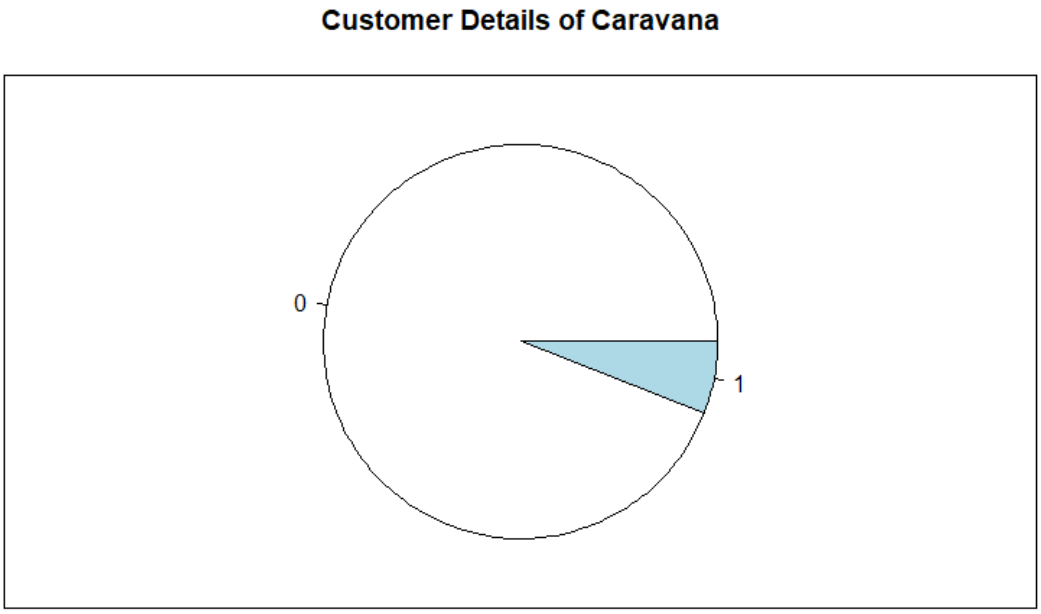
Answer:

The given data set consists of 5822 observations and 86 variables.

In which V86 is the CARAVAN: Number of mobile home policies.

Exploring the Caravan policies bought buy customers.

Below is the pie chart for the number of customers who bought and the Caravan Policy and who did not.

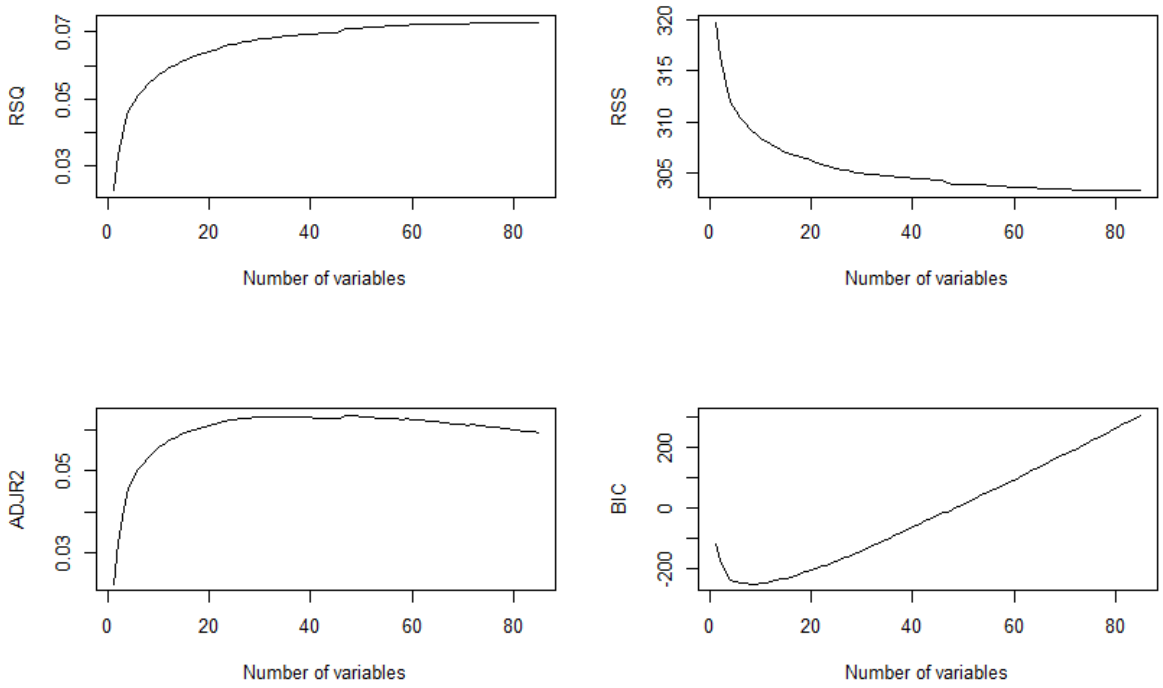


In the above pie chart from 5822 observations of customers only 348 customers finally bought the Caravan policy. Which implies that the models built with OLS, Ridge and lasso will not perform well on the test dataset which include about 85 variables. Lets see the below performances of the models.

The dataset for training is put for Subset selection using forward and backward regressions.

Upon which we can see that for forward selection V47 and then V47,V82 taking the one variable and two variable best fit models for the required outputs. Further subset

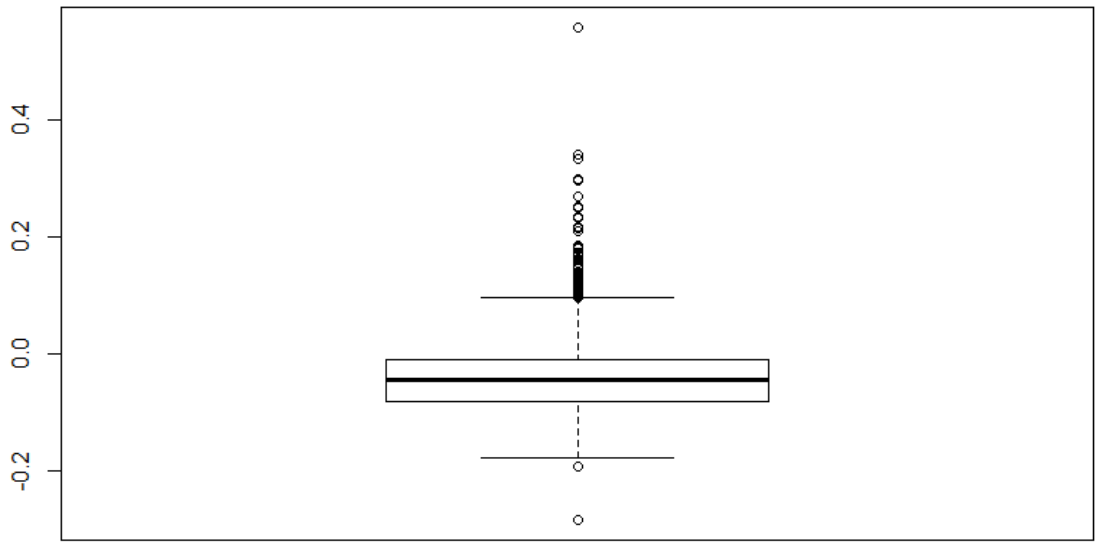
selection can also be seen from the code output. The values of the RSQ, RSS, AdjR^2 and BIC are plotted below with respect to the number of components below.



From the above plots it is inferred that the RSS are very large for the lesser number of variables and also RSQ and AdjR2 are very low for the entire range of variable selections. Which implies that the model is not performing well for the training dataset to select the best combination of variables for best outputs.

Similar trends are seen with respect to the backward subset selection and same cannot be selected for the training of the given dataset.

Training data when put under least square method of predicting the test data, below box plot is obtained for predicted values.



Which implies that the model is predicting very few appropriate predictions as the entire plot is being ranged around 0.

Similar trends are seen with Ridge and lasso models in predicting the given training data with respect to the test dataset predictions.

Hence the prediction of who will buy the policy is very much negligible with the algorithms of Forward, Backward subset selection, Ridge and lasso models.

Also upon further research of the analysis it is observed that using confusion matrix same can be resolved by getting the most dependent variables for the prediction to be made.

Question-3:

(10 points) (Exercise 9 modified, ISL) 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= βX+epsilon

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.

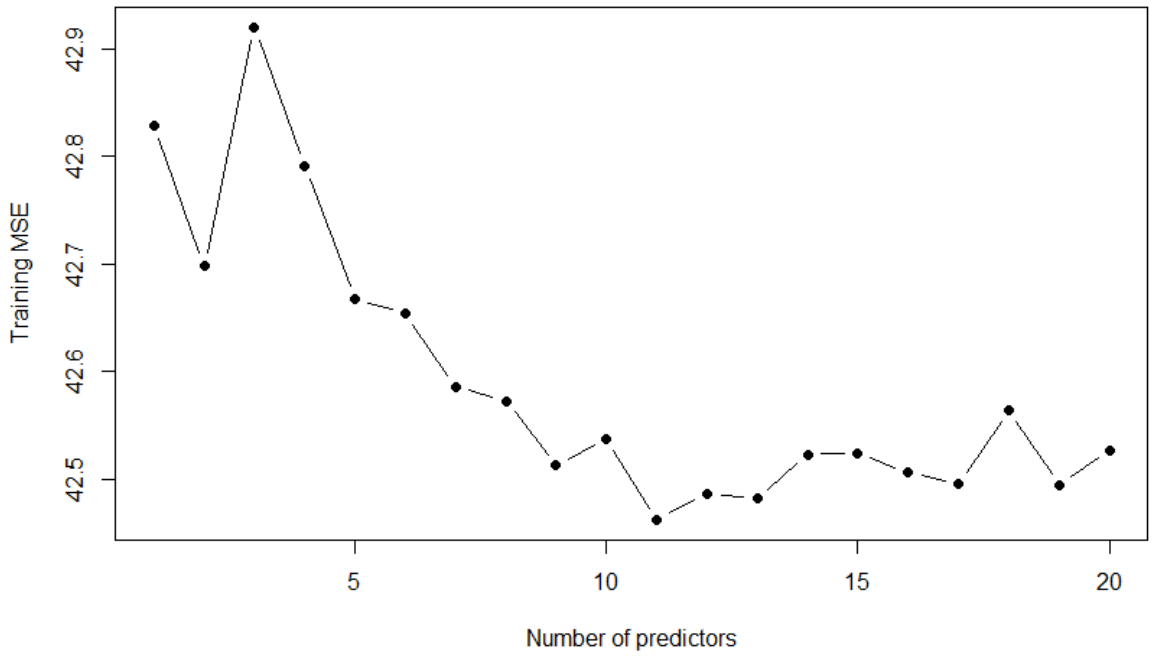
(Note: If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.)

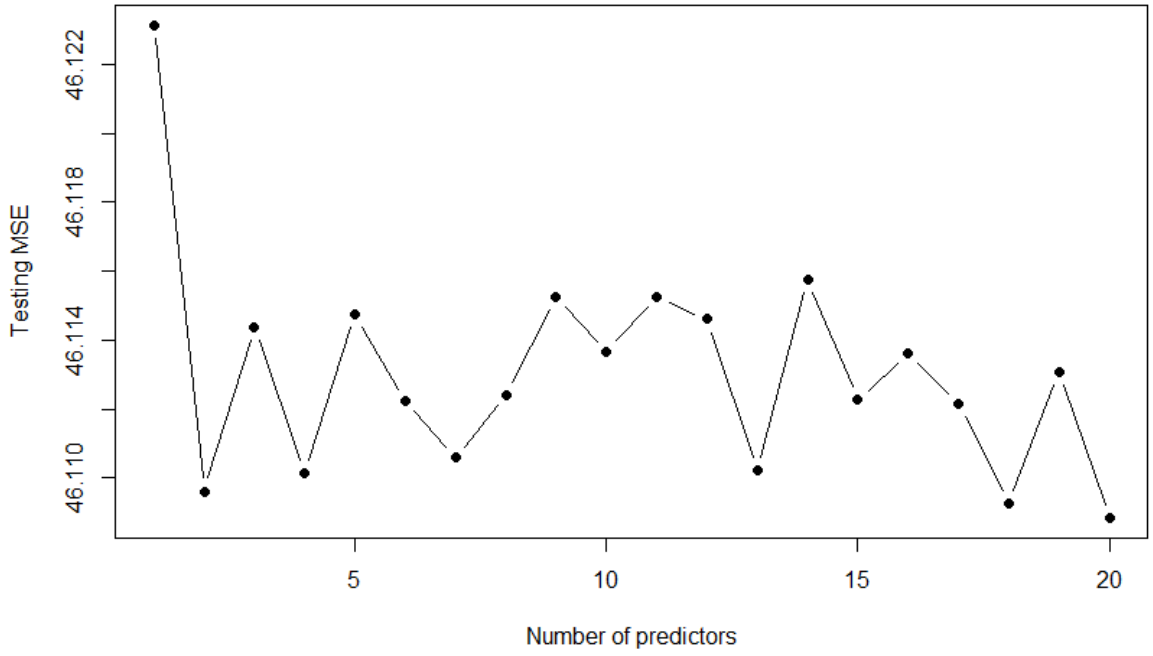
Answer:

The training and test case scenarios for the equation are created as per the code i.e. A matrix with 1000 row and 20 columns is created. Each of the columns has been normalized. Few of the columns are equated to zero for the testing purposes to have improper dataset as mentioned in the question.

Epsilon values have been taken at random with normalized set of numbers.

After forming the dataset and equation in Y for which the training and testing MSE plots are derived which are below.





The least error for the testing set is caught with 20 predictors.

For which the coefficients are obtained in the code which are very close to zeros.

The trend for the Number of predictors and Testing MSE has been deceasing as the predictors increases.