EAS506LEC000 : STATISTICAL DATA MINING

HW ASSIGNMENT – 2

SUBMITTED BY : DITHYA SRIDHARAN

CLASS NUMBER : 48

QUESTION 1

PROBLEM STATEMENT

Consider the College Data Set in the ISLR Package. Predict the number of applications received using other variables in the data set. Fit linear, ridge regression, lasso. PCR and PLS models on the data set and obtain the test error. Comment on the test error obtained in the five cases and compare them.

1. DATA SUMMARY:

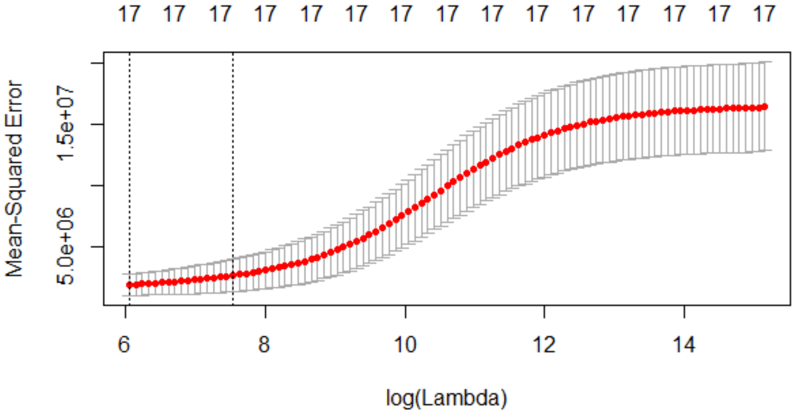
Statistics for a large number of US Colleges from the 1995 issue of US News and World Report. A data frame with 777 observations on the following 18 variables.

2. FITTING LINEAR MODEL:

A linear model was fit on the 18 variables with respect to No. of applications using the lm() function. The linear model using least squares on the training set was fit and the test error obtained was (1.14 * 10^6).

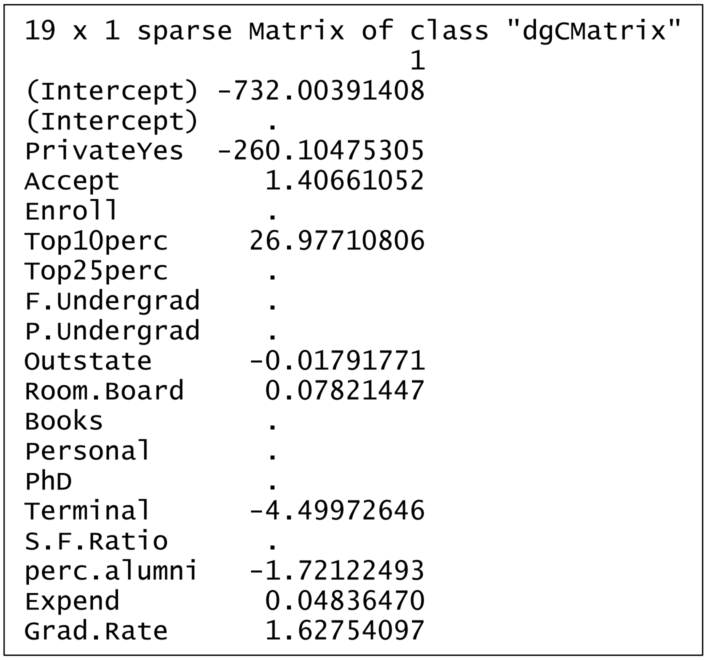
3.FITTING RIDGE REGRESSION MODEL

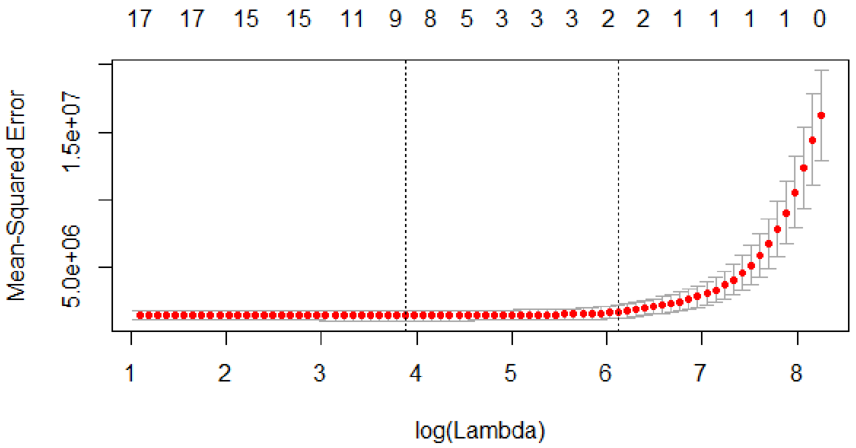
A ridge regression model on the training set, with λ chosen by cross-validation was fit. The best λ obtained by cross validation was found to be 423.304 and the test MSE for ridge was found to be (1.13 * 10^6). It is found that the test MSE for ridge and least squares has very small difference.



4. FITTING A LASSO MODEL

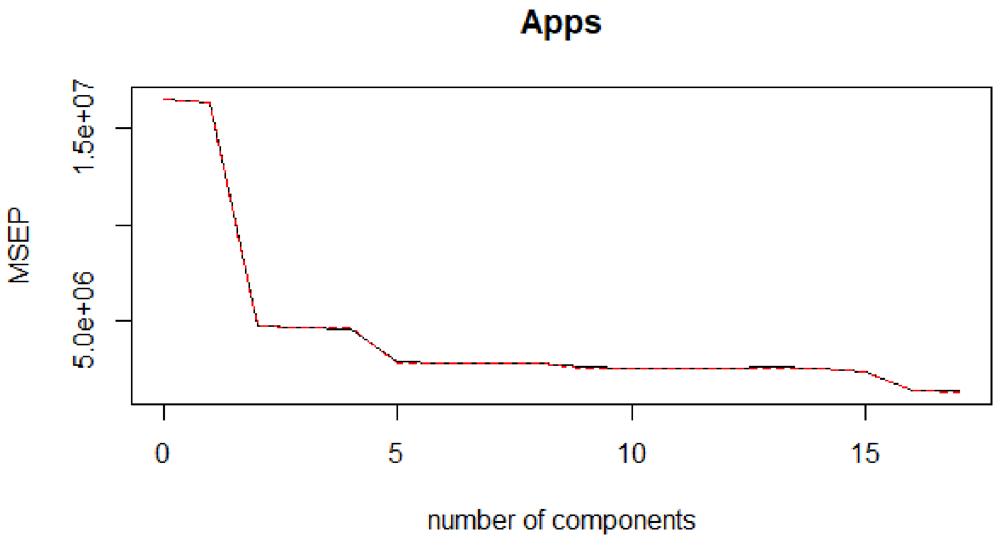
A lasso model on the training set, with λ chosen by cross-validation was fit. The best λ obtained by cross validation was found to be 48.66 and the test MSE for lasso was found to be (1.17 * 10^6).


```
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
```



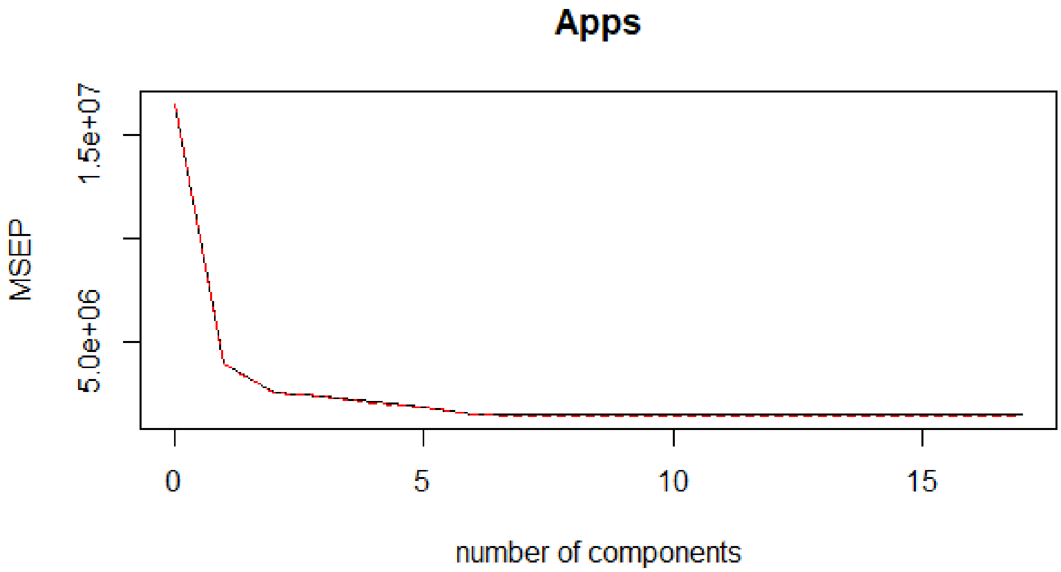
5. FITTING PCR MODEL:

A PCR model on the training set, with *M* chosen by cross-validation was fit. The test MSE for PCR was found to be (1.52 * 10^6). The test MSE for PCR is higher than that obtained by least squares.



6. FITTING PLS MODEL

A PLS model on the training set, with *M* chosen by cross-validation was fit. The test MSE for PLS was found to be (1.144* 10^6). The test MSE for PLS is slightly lower than that obtained by least squares.



7. COMPARING THE RESULTS OBTAINED ABOVE:

To compare the test errors for each model, we compute the R^2 for all the models. Test R^2 is as follows

Model	**R square error**
Linear model	0.8901095
Ridge Regression	0.8916192
Lasso	0.8878316
PCR	0.8538542
PLS	0.889895

8. CONCLUSION

It is found that the R square value for least squares for linear model is 0.8901095, ridge regression is 0.8916192, Lasso is 0.8878316, PCR is 0.8538542, PLS is 0.889895. The test errors obtained by each model is comparable. Since PCR model has the lowest R^2, PCR does not predict number of college applications with high accuracy compared to other models.

QUESTION 2

PROBLEM STATEMENT

Consider the Caravan Insurance Policy Data Set in the ISLR Package. Predict who will be interested in buying a caravan insurance policy. Compute the OLS estimates and compare them with those obtained by variable selection algorithms like forward selection, backward selection, ridge regression and lasso regression. Comment on the estimates obtained.

1. DATA SUMMARY

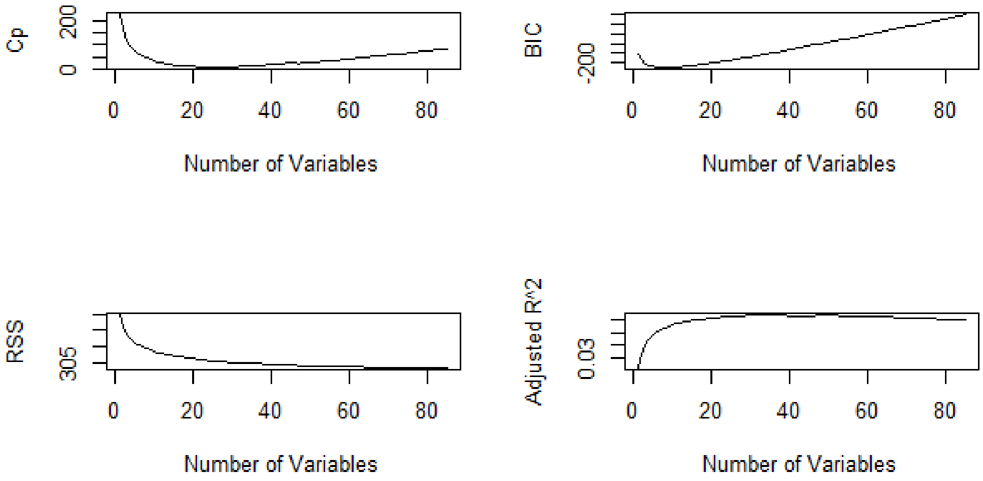
Each real customer record consists of 86 variables, containing sociodemographic data (variables 1-43) and product ownership data (variables 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Variable 86 (Purchase), "CARAVAN: Number of mobile home policies", is the target variable which indicates whether the customer purchase a caravan insurance policy or not.

2. OLS ESTIMATES

A least squares model was fit with respect to "No of mobile home policies". The MSE test error was calculated and was found to be 0.0539. A confusion matrix was obtained to calculate the accuracy of the model. The accuracy of the model was found to be 0.94025.

3. FORWARD SELECTION

The regsubsets() function (part of the leaps library) performs forward selection. Here we use the regsubsets function but specify the `method="forward" option. The summary() command outputs the best set of variables for each model size. V10,V18,V21,V43, V44,V47, V59,V82, V83 V85 are the best subsets by forward. Plots for Cp, BIC, RSS and Adjusted R^2 is as follows.



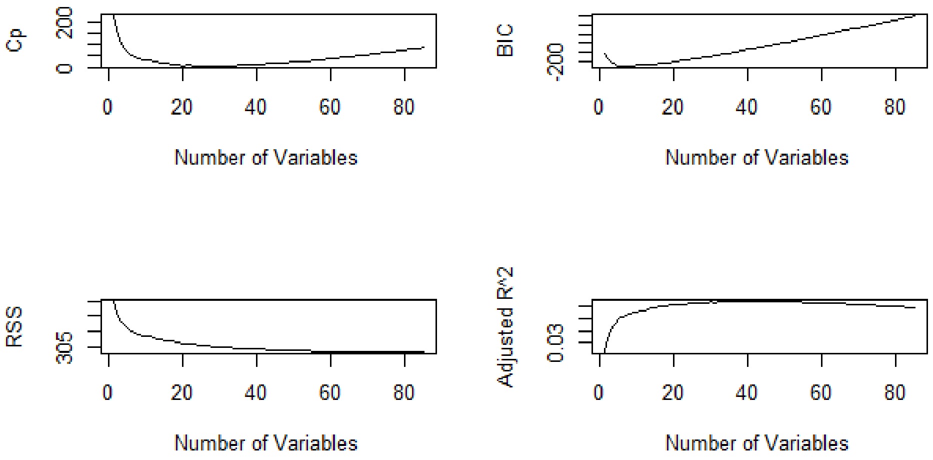
A linear model was fitted with the variables selected by forward method and the MSE test error was calculated. The MSE test error was found to be 0.054008. A confusion matrix was obtained to calculate the accuracy of the model. The accuracy of the model was found to be 0.94.

4.BACKWARD SELECTION

The regsubsets() function (part of the leaps library) performs backward selection. Here we use the regsubsets function but specify the `method="backward" option. The summary() command outputs the best set of variables for each model size. V10,V18,V21, V46,V47, V59,V76, V82 V85 are the best subsets by backward. Plots for Cp, BIC, RSS and Adjusted R^2 is on the next page.

A linear model was fitted with the variables selected by forward method and the MSE test error was calculated. The MSE test error was found to be 0.05407046. A confusion matrix was obtained to calculate the accuracy of the model. The accuracy of the model was found to be 0.94.

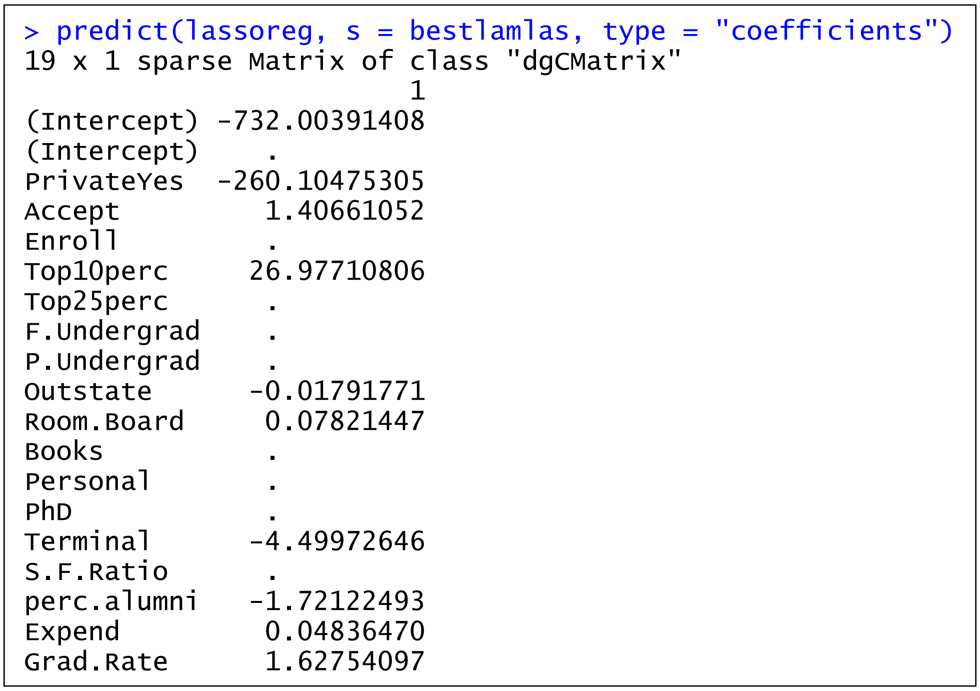
Upto four variables, forward and backward give the same variable selection.



```
> coef(regfit.for, 5)
 (Intercept)         V18         V43         V44         V47         V82
 0.005436302 -0.005756253  0.007686293  0.017944911  0.010949070  0.288184811
> coef(regfit.bac, 5)
 (Intercept)         V10         V18         V47         V59         V82
 0.003176233  0.006904983 -0.008492691  0.011291811  0.009556631  0.286852633
```

5. LASSO REGRESSION

A lasso model on the training set, with λ chosen by cross-validation. The best λ obtained by cross validation was found to be 0.003435912 and the test MSE for lasso was found to be 0.0537716. A confusion matrix was obtained to calculate the accuracy of the model. The accuracy of the model was found to be 0.94025.



```
> predict(lassoreg, s = bestlamlas, 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
```

6. RIDGE REGRESSION

A ridge model on the training set, with λ chosen by cross-validation. The best λ obtained by cross validation was found to be 0.118244 and the test MSE for lasso was found to be 0.0536968. A confusion matrix was obtained to calculate the accuracy of the model. The accuracy of the model was found to be 0.94075.

7. CONCLUSION

It is found that the R square value for least squares for ols model is 0.03528869, forward selection is 0.03343508, backward selection is 0.03376165, lasso is 0.03910229, ridge regression is 0.04043776. The test errors obtained by each model is comparable. Since Ridge model has the highest R^2, Ridge predicts Caravan Insurance with high accuracy compared to other models.

QUESTION 3

PROBLEM STATEMENT

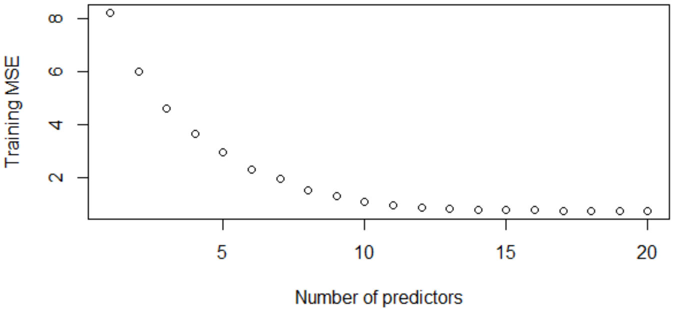
We have seen that as a number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set. Generate a data set with p=20 features, n=1000 observations, and an associated quantitative response vector generated according to the model $Y=X\beta+\epsilon$, where β has some elements that are exactly equal to zero. Perform best subset selection and plot Training and Test MSE.

1.GENERATING DATA WITH 1000 OBSERVATIONS and 20 FEATURES.

Data is generated using rnorm() function to generate random sequence of data. Some beta(b) Is generated for 20 features. A data set $Y=X\beta+\epsilon$ is generated and then split into training and test data where training contains 100 and test contains 900 observations.

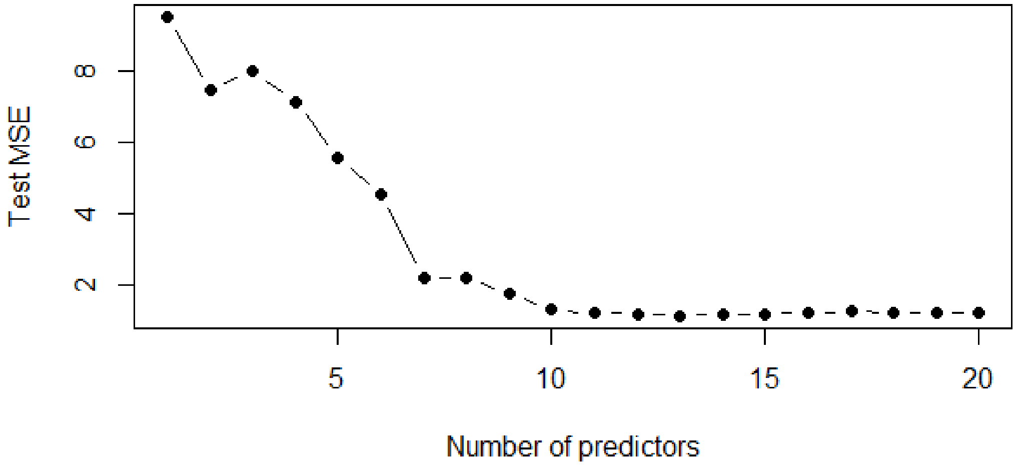
2. BEST SUBSET SELECTION ON THE TRAINING SET

Best subset selection is performed using regsubsets() function and training MSE is plotted for the best model of each size.



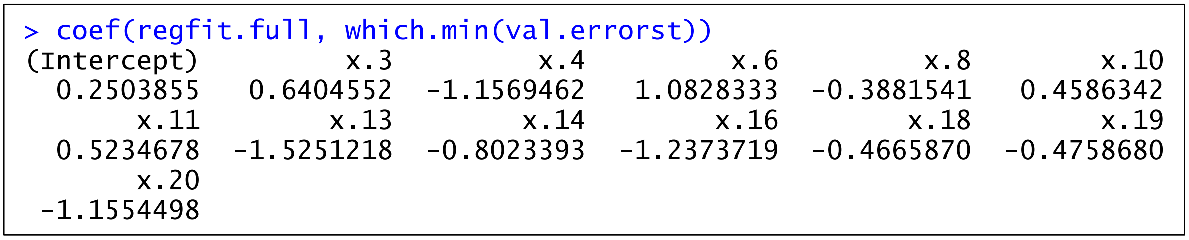
4. BEST SELECTION ON TEST SET

Best subset selection is performed using regsubsets() function and test MSE is plotted for the best model of each size.



We can find that the model size for which test set MSE is minimum is 12. Therefore, 12-variable model has the smallest MSE.

5. MINIMUM TEST MSE MODEL COMPARED TO TRUE MODEL



```
> coef(regfit.full, which.min(val.errorst))
(Intercept)         x.3         x.4         x.6         x.8        x.10 
  0.2503855   0.6404552  -1.1569462   1.0828333  -0.3881541   0.4586342 
       x.11        x.13        x.14        x.16        x.18        x.19 
  0.5234678  -1.5251218  -0.8023393  -1.2373719  -0.4665870  -0.4758680 
       x.20 
 -1.1554498 
```

Here, we can see that the best model is with 12 variables and it has removed all coefficients which are tending to zero. So it has the minimum test MSE.