Homework 2

Freya Genesis D Mello

October 2018

1 Problem Statement 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 LAMBDA chosen by crossvalidation.
(c)Report the test error obtained.
(d) Fit a lasso model on the training set, with LAMBDA chosen by crossvalidation. 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 crossvalidation. 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?

1.1 Understanding the dataset

As a starting step, the necessary libraries are loaded - "leaps", "glmnet", "pls" and "ISLR". The dataset is studied thoroughly to learn that p = 18 and n = 777. The names of the predictors are "Private", "Apps", "Accept", "Enroll", "Top10perc", "Top25perc", "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board", "Books", "Personal", "PhD", "Terminal", "S.F.Ratio", "perc.alumni", "Expend", "Grad.Rate". The function na.omit() is used to get rid of any missing data. The feature "Private" is categorical and others are numerical. The "Private" undergoes data transformation to avoid coercing errors; from "True" to 1 and "False" to 0.

1.2 OLS

A linear model is built with all the predictors: "Private", "Apps", "Accept", "Enroll", "Top10perc", "Top25perc", "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board", "Books", "Personal", "PhD", "Terminal", "S.F.Ratio", "perc.alumni", "Expend", "Grad.Rate". Expend the most significant predictors contributing to the number of application the college receives.

1.3 Ridge Regression Model

The MSE associated with each lambda is plotted after performing cross validation for ridge models. Value of best lambda = 408.6915 1 Coefficients are given in the table. (See fig. 2)

1.4 LASSO

The lasso model 2 is built using the glmnet function setting alpha as 1. The cross validation outputs the value of 12.02 as the minimum value of lambda which is picked up as the best lambda value. The coefficient values are studied and the correlation is learned. The predicted values over the output column are computed and compared to the actual values to result in the errors values.

The number of non-zero coefficient estimates are 14.

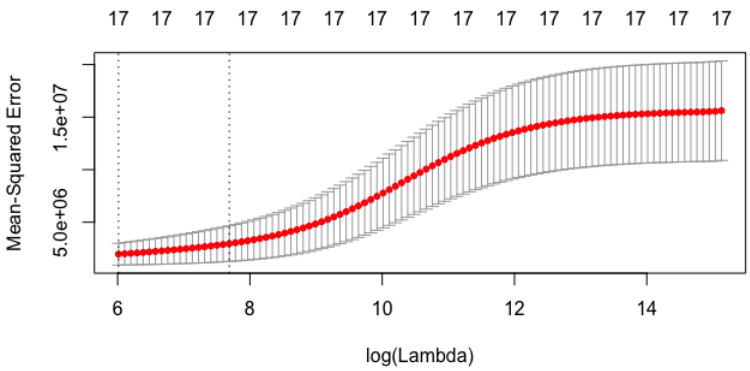

Figure 1: Ridge Regression Model

Variable	Coefficient
Private	-5.296376e+02
Accept	9.724457e-01
Enroll	4.737049e-01
Top10perc	2.480122e+01
Top25perc	1.160183e+00
F.Undergrad	7.752195e-02
P.Undergrad	2.457069e-02
Outstate	-2.081174e-02
Room.Board	1.999760e-01
Books	1.367763e-01
Personal	-9.093341e-03
PhD	-3.714304e+00
Terminal	-4.687648e+00
S.F.Ratio	1.279169e+01
perc.alumni	-8.889593e+00
Expend	7.514164e-02
Grad.Rate	1.138859e+01

Table 1: Ridge Coefficients

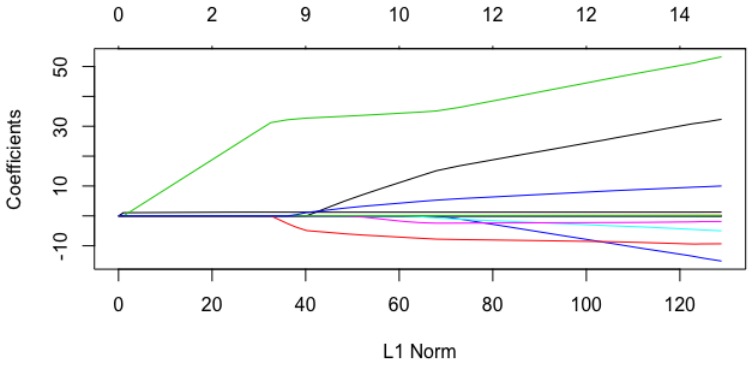

Figure 2: LASSO Model

Variable	Coefficient
Accept	1.62219094
Enroll	-0.46892971
Top10perc	46.47007154
Top25perc	-10.56739822
F.Undergrad	.
P.Undergrad	0.01679260
Outstate	-0.09242571
Room.Board	0.13474990
Books	-0.04899787
Personal	0.03394770
PhD	-2.17660267
Terminal	-5.47617793
S.F.Ratio	24.40522358
perc.alumni	.
Expend	0.06440062
Grad.Rate	5.03621461

Table 2: LASSO Coefficients

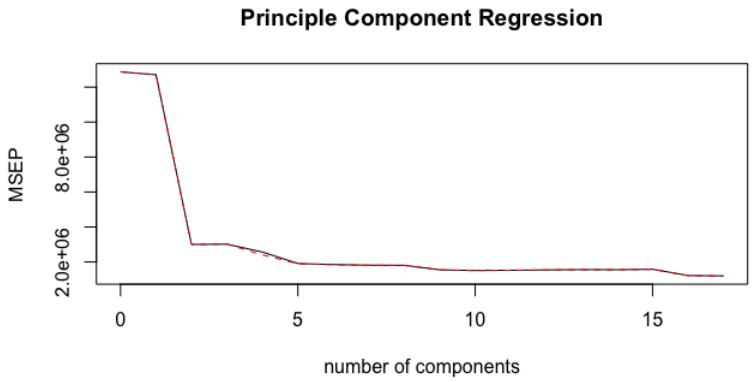

Figure 3: Principle Component Regression Model

1.5 Principle Component Regression

The validation plot 3 describes the number of applications that can be predicted with 4 principal components that have a comparatively lesser MSEP.

1.6 Partial Least Squares

On comparing with Principle Component Regression, the validation plot 4 describes the number of applications that can be predicted with 6 principal components that have a comparatively lesser MSEP.

Since the error rates produced among the models are very high, we cannot accurately predict the the number of college applications received. Among the test errors obtained we see that the PLS model produced the least test error.

2 Problem Statement 2

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

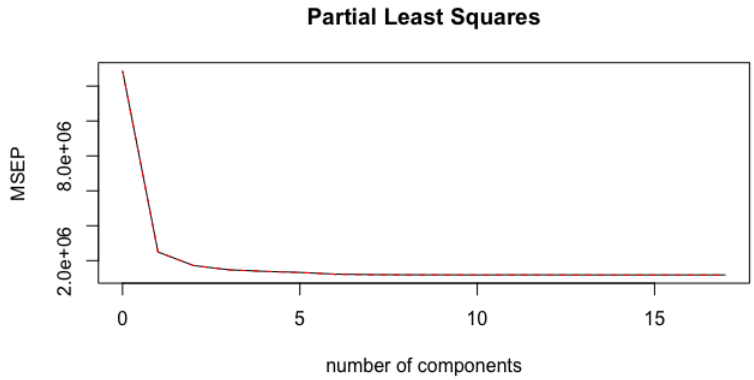

Figure 4: Partial Least Squares Model

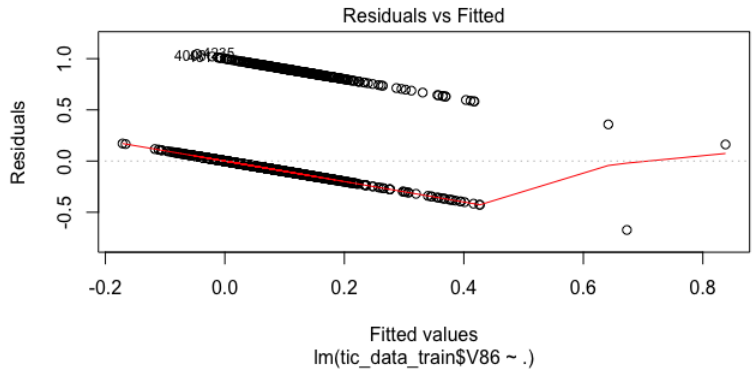

Figure 5: Residuals vs Fitted graph

them with those obtained from the following variable selection algorithms: Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression. Support your answer.

2.1 Inference

The data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The predicted values thus computed by the following linear models are observed and studied. Taking a lambda value of 0.1118244, the ridge model was thus built to have comparatively lesser error than the OLS linear model. The same applied to the LASSO model. The number of people who bought the policy are not enough to be considered to train the data and hence the number predicted was only 3 out of 238 people. The method of classifying that is considered is that if the predicted value is less than 0, then the customer bought the insurance. If its greater than 0 then the customer did not buy the insurance.

2.2 Ordinary Least Squares

A linear regression model is built considering the "V86" feature as the response variable. Train and Test errors are computed by comparing the actual and predicted values from the model. The Train set gives an error of 5.96% whereas the test set gives 5.95%. 5

2.3 Forward Selection

The forward subset selection is performed to reduce the considered features to 23 variables as the graphs suggest 15 to be at the point when the CP value and the BIC value are minimum. 6

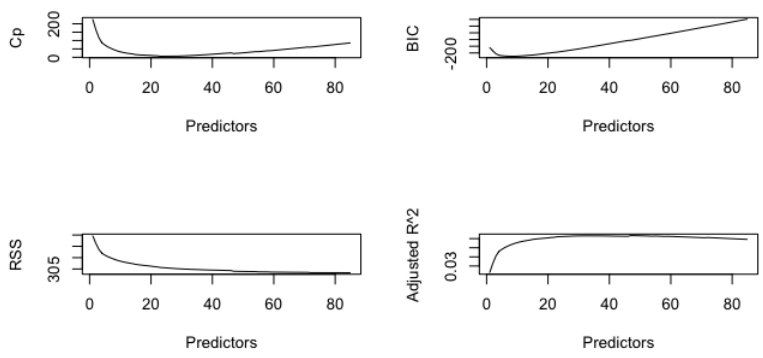

Figure 6: Forward Subset Selection

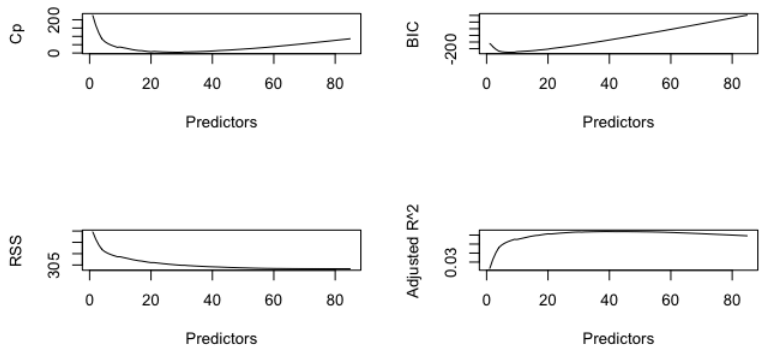

Figure 7: Backward Subset Selection

2.4 Backward Selection

The backward subset selection is performed to reduce the considered features to 23 variables as the graphs suggest 15 to be at the point when the CP value and the BIC value are minimum. 7

2.5 Comparing Forward and Backward Selection

On comparing the backward and forward subset selection techniques and narrowing down the predictors to a value of 23, the train error was computed to be 5.97% and the test error was computed to be 5.975%.

2.6 Ridge

The ridge model is built on the value of lambda based on minimum MSE as 0.1118244. Taking a lambda value of 0.1118244, the ridge model was thus built to have comparatively lesser error than the OLS linear model, a value of 5.369%. 8

2.7 LASSO

The LASSO model is built on the value of lambda based on minimum MSE as 0.00290187. Taking a lambda value of 0.00290187, the LASSO model was thus built to have comparatively lesser error than the OLS linear model, a value of 5.37%. 9

Of all the models thus observed, the LASSO model predicts best if the customer buys the caravan insurance or not with the least error and 91% accuracy.

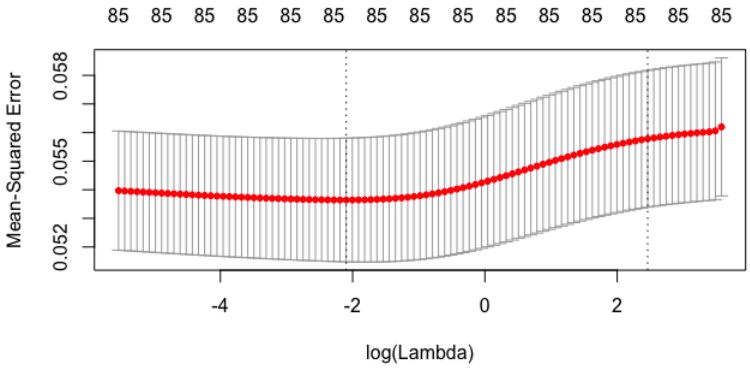

Figure 8: Ridge Model

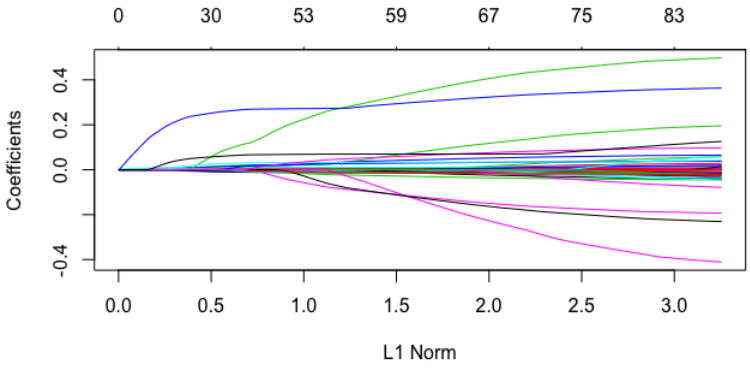

Figure 9: LASSO Model

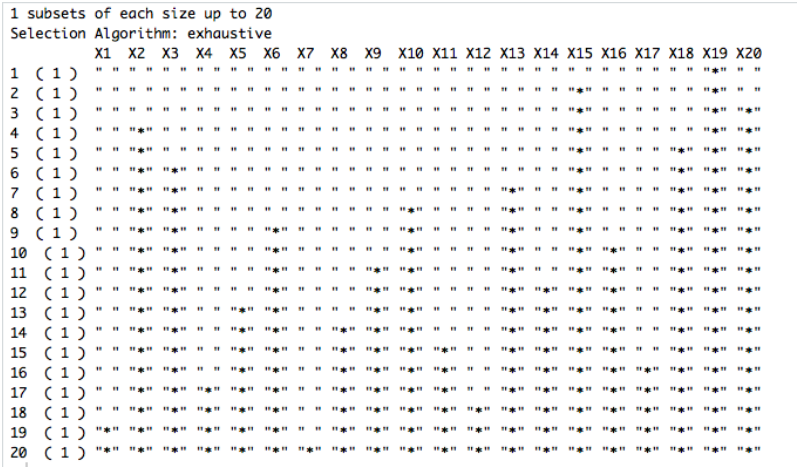

Figure 10: The subsets created by exhaustive method

3 Problem Statement 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 = XB + E where B 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.

3.1 Creating the dataset

Using the formula provided as a basis, the dataset was built using the appropriate Beta coefficients and Bias values. Some beta values were manually set to 0. The dataset has 20 predictor values and a response variable and total number of 1000 records. Its split into training and test data where sample function was used to put 100 records in training and 900 in test data.

3.2 Subset selection

The exhaustive method of subset selection was performed on the data set thus created. 10

The CP, BIC, RSS and Adjusted R squared values were observed across all values of predictors and plotted in 11.

The models thus observed have 8 variables to be the best subsets with minimum value of CP and BIC. Both train and test data are passed into to the model and train and test MSE respectively is computed for each model.

The subsets created with all possible sizes of predictors from exhaustive subset selection are passed into a loop to compute the test and training data mean square errors against the response variable Y and are thus the values are plotted.

For the minimum test mean square error value of 89633.34, the number of predictors fit in the model was 8. We can see in 12 that the training error keeps decreasing and the test error initially decreases and we get a minimum value of test error when we fit 8 predictors in the linear model and then the test error keeps increasing to a certain constant value.

The test set MSE was observed to be 89633.34 where it was minimized as compared to the true model. The other predictors (other than the 8 predictors) had their beta coeffcients reduced to zero and were not included in the model with the minimum test MSE.

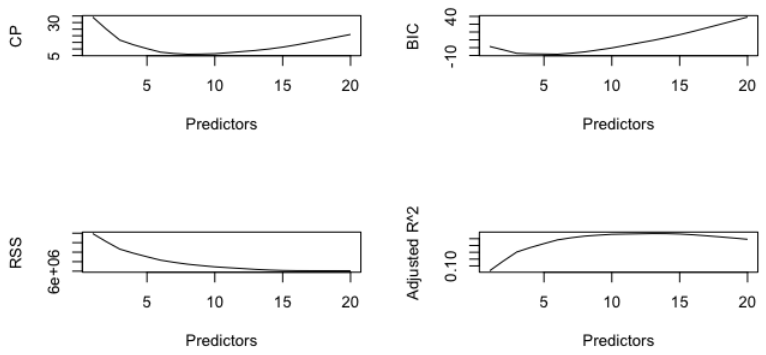

Figure 11: Exhaustive subset selection

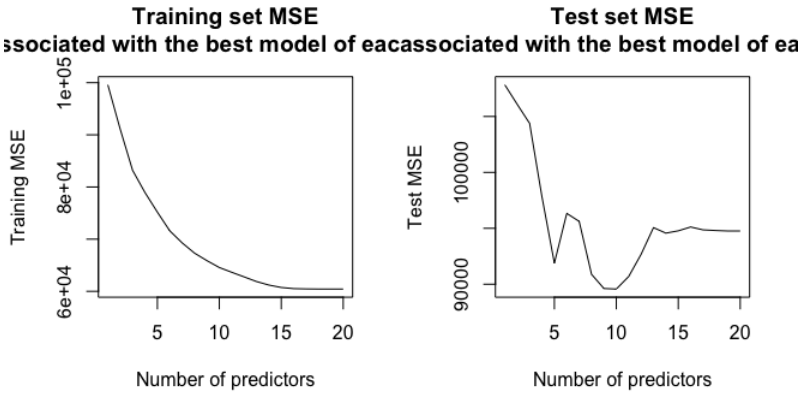

Figure 12: Training and Test set MSE associated with the best model of each size