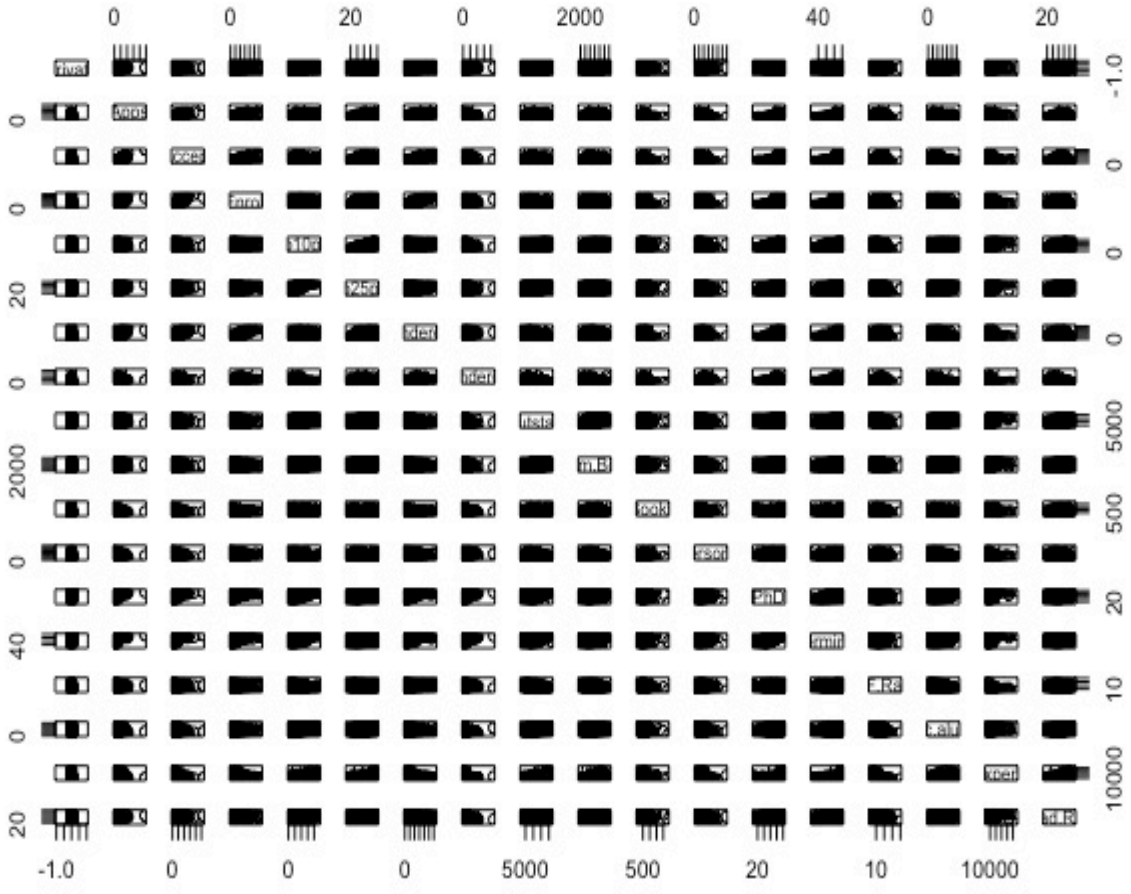
HOMEWORK-2

Solution 1) Used the College dataset from ISLR package to perform linear regression, ridge regression, lasso, PCR and PLS to report and analyze the test errors.

a) Linear Regression:

The dataset contains 18 variables. The predictor Private is a factor (Yes/No) so I am converting it into a Numeric type (1/0). I used 75% off the data for training my model and 25% for testing.
Plotted the pairs graph to observe any visible relation between predictors and response.



Accept and Enroll seem to have an impact on the response variable Apps.
Performed linear regression on the training set to observe an rmse of 959.094.

b) Ridge regression:

Ridge regression is used to reduce collinearity in a model. It does so by shrinking the coefficient estimates of highly correlated variables. The coefficients are shrunk almost to zero but they they still continue to be part of the model.
Here, the glmnet package has been used to perform ridge regression. The rmse observed is 1229.067 for the best lambda value 415.918.

d) Lasso: (Least Absolute Shrinkage Selector Operator)

Lasso is similar to ridge regression in the sense that it shrinks the coefficient estimates of highly correlated variables. However, when the coefficient values are reduced to zero, they are excluded from the model. This goes to say that the lasso model automatically does feature selection. This could lead to some loss of information and lowered accuracy.

The glmnet package has been used to perform lasso regression. The rmse observed is 1069.803 for the best lambda value of 2.219 with no coefficient reduced to zero.

e) PCR: Principal Components Regression

It is based on principal component analysis (PCA). The main idea is to identify the principal components (components that show maximum variance with each other) and use some of them as predictors in a linear regression using least squares procedure. The core assumption of PCR is that the directions of the principal components chosen show the exact directions associated with the response variable.

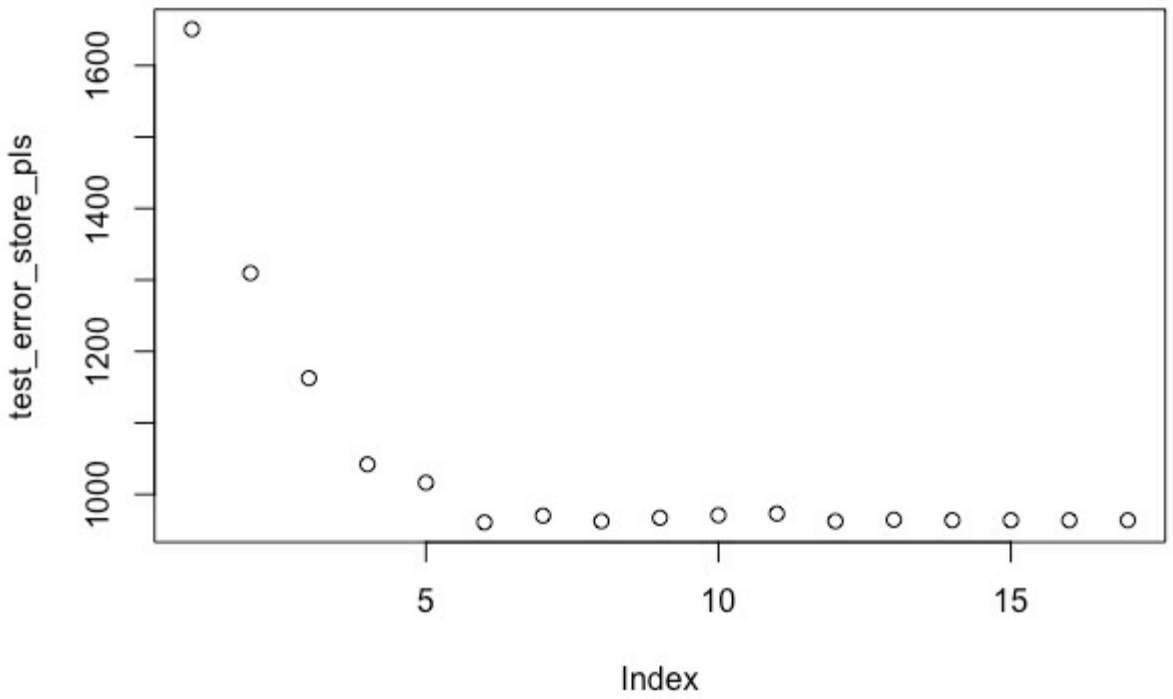
Here, the pls package has been used to perform PCR, following test errors have been observed:

The rmse observed is 963.7846 and is lowest at index 17 (k, chosen by cross-validation).

f) PLS: Partial Least Squares

It is similar to PCA where the principal components where components with maximum variance are desired. In addition, these components need to be highly correlated to the response variable.

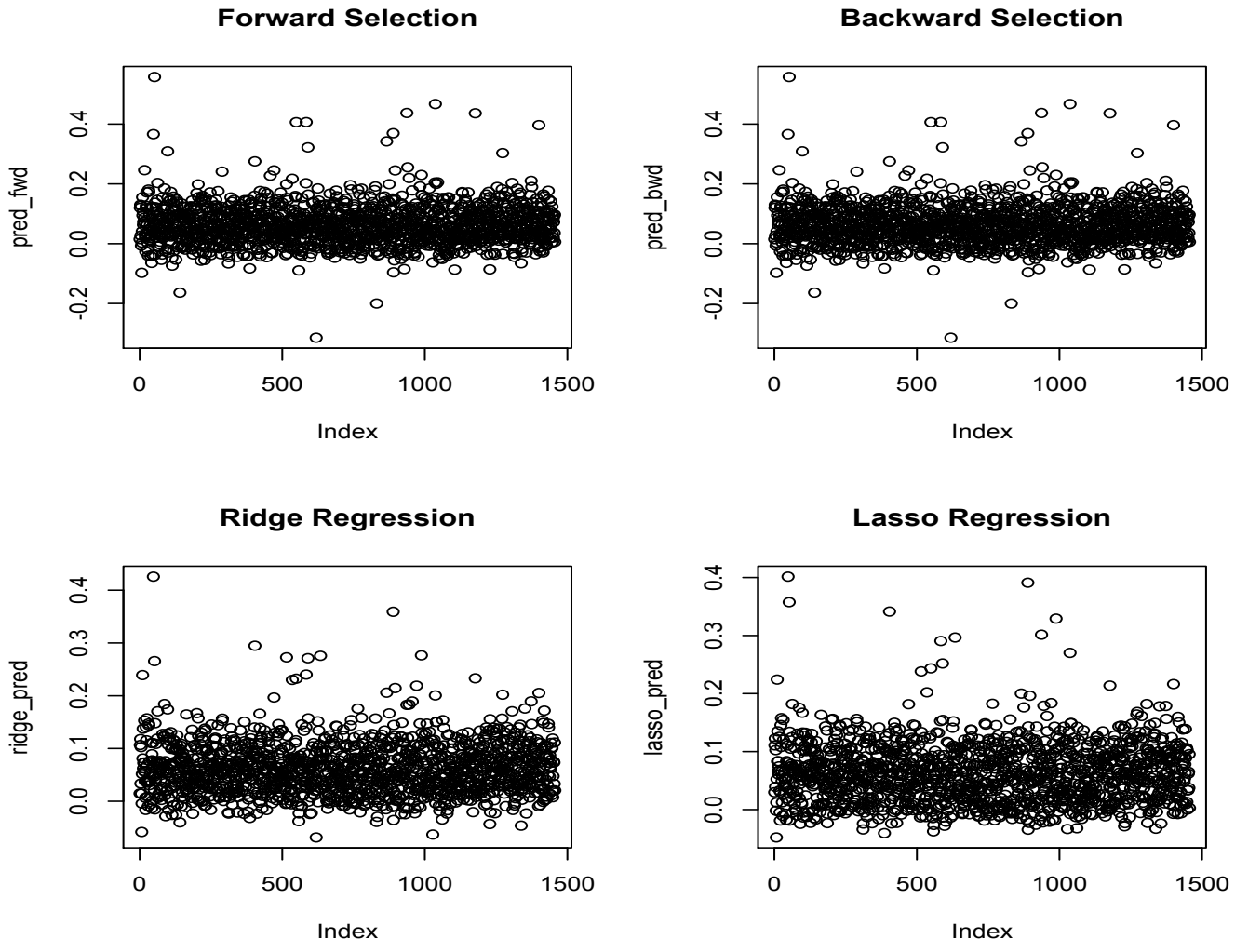
The graph below shows least rmse is 960.784 at index 6 (k, chosen by cross-validation).



g) It can be observed that PLS has the least error while Ridge Regression has the highest error. PCR and OLS have the same error. Since all the models have approximately the same range of error, either one could be used for final prediction. The accuracy appears to be same for all.

Solution 2) To predict who would be interested in buying the caravan insurance policy.
I started with checking the data and response variable (V86). It can be seen that the data is highly unbalanced as there are 5474 entries for those who did not buy the policy against 348 entries of those who did.
The OLS model shows that there are 11 significant variables out of the 85, with rmse as 3.906105.
Forward subset selection gives the best result at index 19 with rmse 0.228753.
Backward subset selection gives the best result at index 15 with rmse 0.2284244.
Lasso identifies 8 variables as significant before reducing them to zero. It has an rmse of 0.2277454 with best λ 0.0033.
Ridge has an rmse of 0.2277318 with best λ 0.1424.

Went ahead to perform various subset selection methods shown in the graph below.

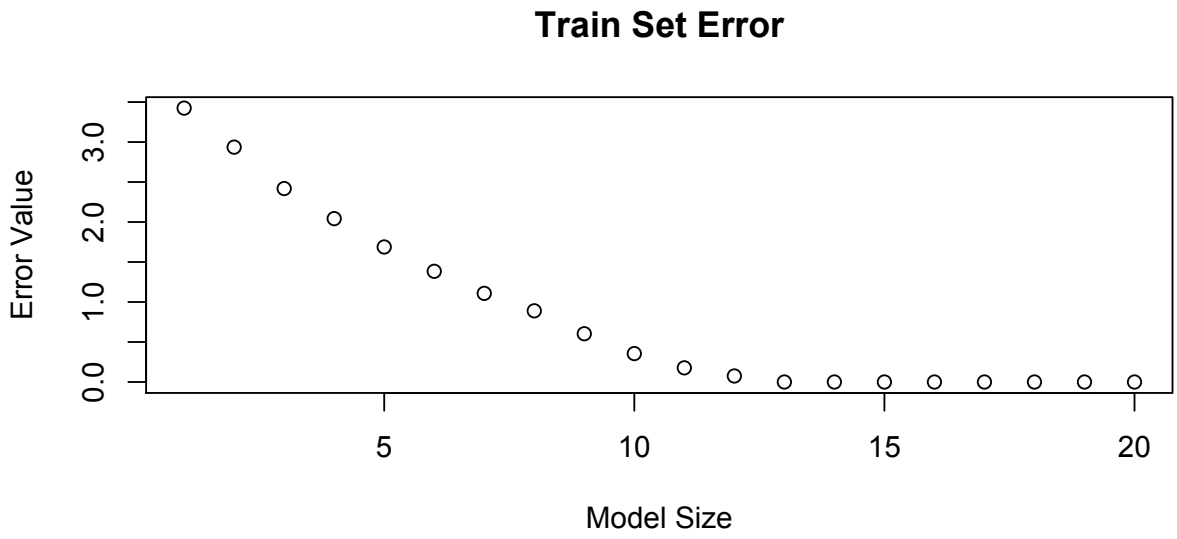


After viewing the prediction result on the test, I see that almost all test data points have been predicted to 0 (not purchase) with minimal predictions for 1 (purchase).

It turns out that since there isn't enough support for the data that purchased caravan insurance policy, the model will always predict that the new data will choose to not purchase the policy.

Solution 3) To prove that as the number of features used in a model increase, the training error will necessarily decrease, but the test error may not.

Generated a data set with p = 20 features, n = 1, 000 observations, and an associated quantitative response vector generated according to the model $\boldsymbol{Y = X\beta + \varepsilon}$ where β has 7 elements that are exactly equal to zero and epsilon as 500.



The test error for the best subset selection model of each size is plotted in the graph above. It can be seen that the test set MSE take on its minimum value for model size 14. The variables considered for this model are V8 to V20, inclusive. This is the same compared to the OLS model. Also, the coefficient estimates for OLS and the minimum error model are approximately equal.

The test error for model size 20 is greater than that of model size 14. This proves that as the number of variables increase, the train error necessarily decreases. However, that may not be the case for test error.