PAUL M Girdler
CLASS 15
paulgird@buffalo.edu.au

Homework Assignment 2
EAS506 - Statistical Data Mining 1

2018-10-09

1 Exercise 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.

1.1 Part 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.

The college dataset contains 777 data points with 16 predictors and 1 response variable, **Apps**,the number of applications received by a university. The data was randomly split into training and test with a ratio of 3:1 respectively. A linear model was fitted on the training data.

Summary of the Linear Model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-3156.4  -420.6   -25.4   284.4  6995.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.118e+03  4.360e+02  -2.563  0.01064 *
Accept       1.250e+00  5.481e-02  22.810  < 2e-16 ***
Enroll      -1.532e-01  2.324e-01  -0.659  0.51014
Top10perc    4.773e+01  6.458e+00   7.390 5.73e-13 ***
Top25perc   -1.511e+01  5.203e+00  -2.904  0.00384 **
F.Undergrad  6.236e-02  3.988e-02   1.564  0.11847
P.Undergrad  5.334e-02  3.534e-02   1.509  0.13189
Outstate    -8.946e-02  2.085e-02  -4.290 2.12e-05 ***
Room.Board   1.397e-01  5.702e-02   2.450  0.01459 *
Books        1.160e-01  2.931e-01   0.396  0.69230
Personal    -1.919e-04  7.430e-02  -0.003  0.99794
PhD         -6.722e+00  5.784e+00  -1.162  0.24573
Terminal     1.410e+00  5.969e+00   0.236  0.81334
S.F.Ratio    1.644e+01  1.508e+01   1.090  0.27613
perc.alumni -1.002e+01  4.799e+00  -2.089  0.03718 *
Expend       9.015e-02  1.419e-02   6.353 4.55e-10 ***
Grad.Rate    1.072e+01  3.421e+00   3.134  0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 998.9 on 531 degrees of freedom
Multiple R-squared:  0.924,     Adjusted R-squared:  0.9217
F-statistic: 403.7 on 16 and 531 DF,  p-value: < 2.2e-16
```

!!

The test error, RSS, was calculated for the training and test data.

$$RSS_1 = 529849177 \quad \text{training}$$
$$RSS_2 = 400655474 \quad \text{test}$$

1.2 Part B

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

Ridge regression was performed with cross validation (CV) on 10 folds. The result of CV was plotted.

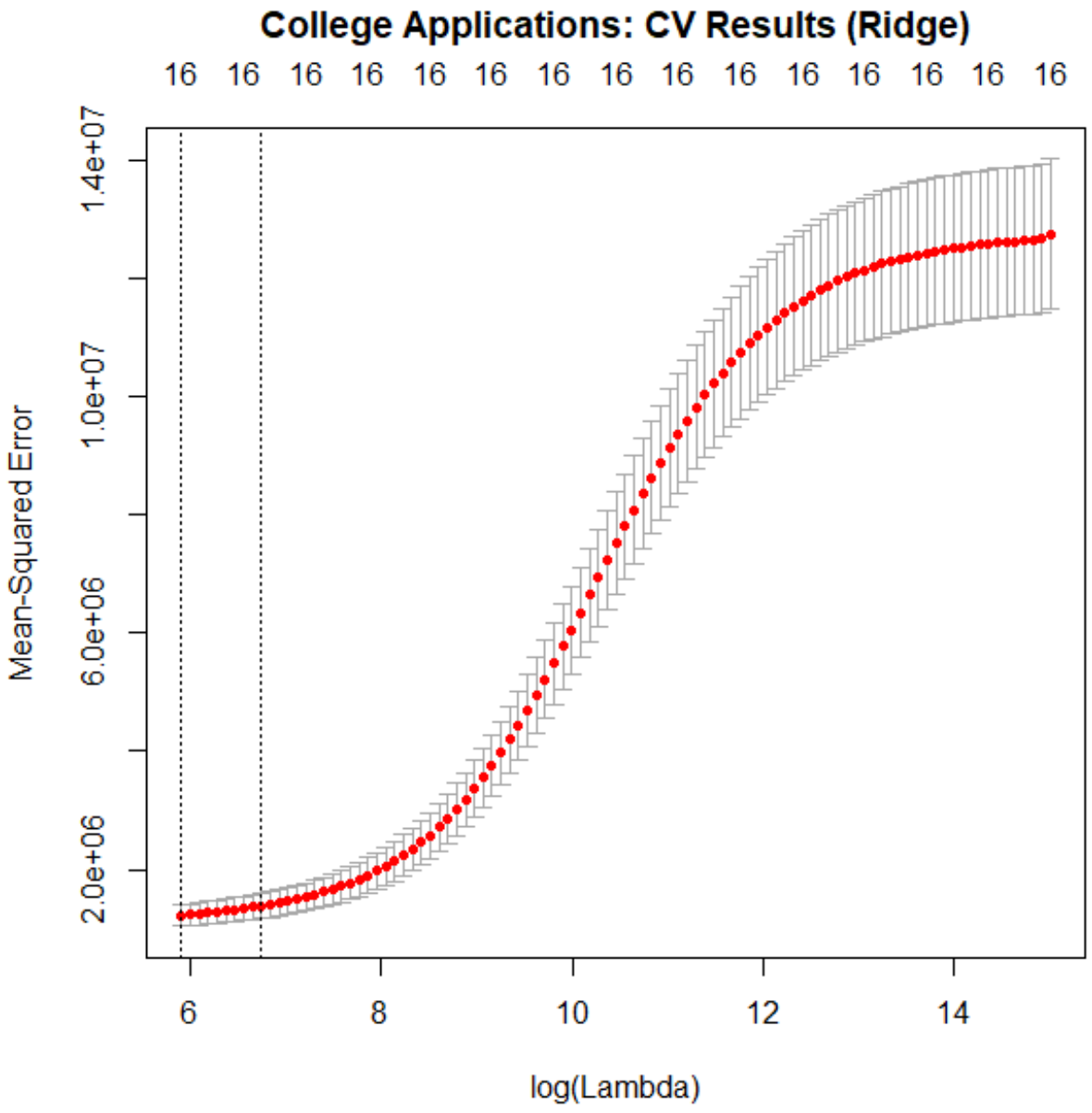
The value, $\lambda_b = 365.91$, was found to minimise training error, and as such was selected from the CV results. A model was built using this value of λ_b.

Summary of the Ridge Regression Model:

```
    (Intercept)         Accept         Enroll      Top10perc      Top25perc    F.Undergrad
  -1.888374e+03   7.700399e-01   7.283180e-01   2.597855e+01  -4.158581e-02   1.030959e-01
   P.Undergrad      Outstate    Room.Board          Books       Personal            PhD
   3.902888e-02  -2.868547e-02   1.504527e-01   2.309803e-01  -3.719083e-02  -1.122479e+00
      Terminal     S.F.Ratio   perc.alumni         Expend     Grad.Rate
!!-1.358437e+00   1.294235e+01  -1.640912e+01   8.448055e-02   1.114906e+01
```

The test error, RSS, was calculated for the training and test data.

$$RSS_1 = 620147652 \quad \text{training}$$
$$RSS_2 = 710408276 \quad \text{test}$$



1.3 Part 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.

Lasso regression was performed with cross validation (CV) on 10 folds. The result of CV was plotted.

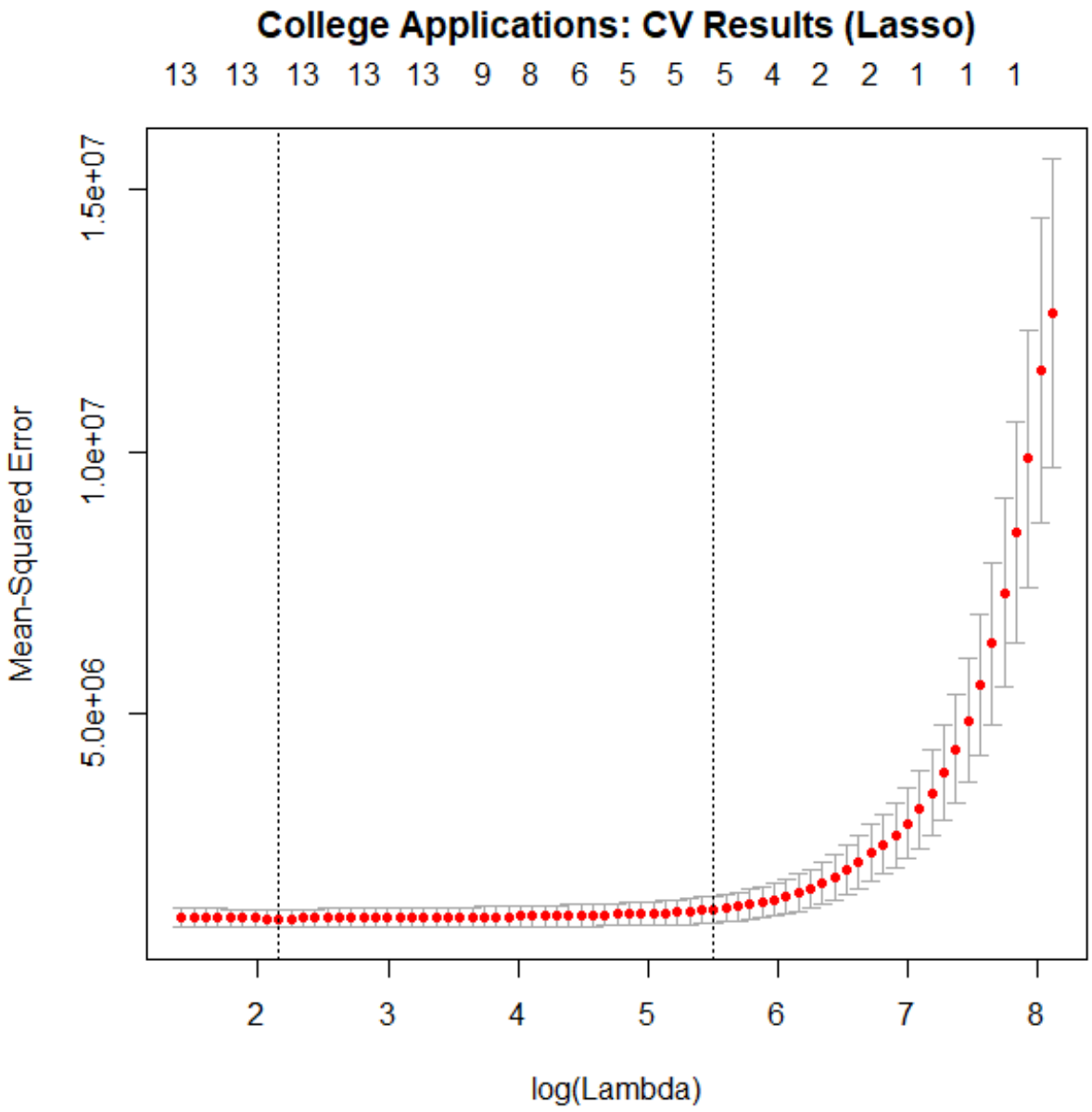
The value, $\lambda_b = 5.963396$, was found to minimise training error, and as such was selected from the CV results. A model was built using this value of λ_b.

Summary of the Lasso Regression Model:

```
   (Intercept)        Accept        Enroll     Top10perc     Top25perc   F.Undergrad
-1.131875e+03  1.227815e+00  0.000000e+00  4.300494e+01 -1.127382e+01  4.233681e-02
  P.Undergrad     Outstate   Room.Board         Books      Personal           PhD
 5.164360e-02 -8.138616e-02  1.283570e-01  9.931926e-02  0.000000e+00 -4.590199e+00
     Terminal     S.F.Ratio   perc.alumni        Expend     Grad.Rate
!! 0.000000e+00  1.359951e+01 -9.923447e+00  8.782101e-02  9.549181e+00
```

The test error, RSS, was calculated for the training and test data.

$$RSS_1 = 531164806 \quad \text{training}$$
$$RSS_2 = 411049325 \quad \text{test}$$

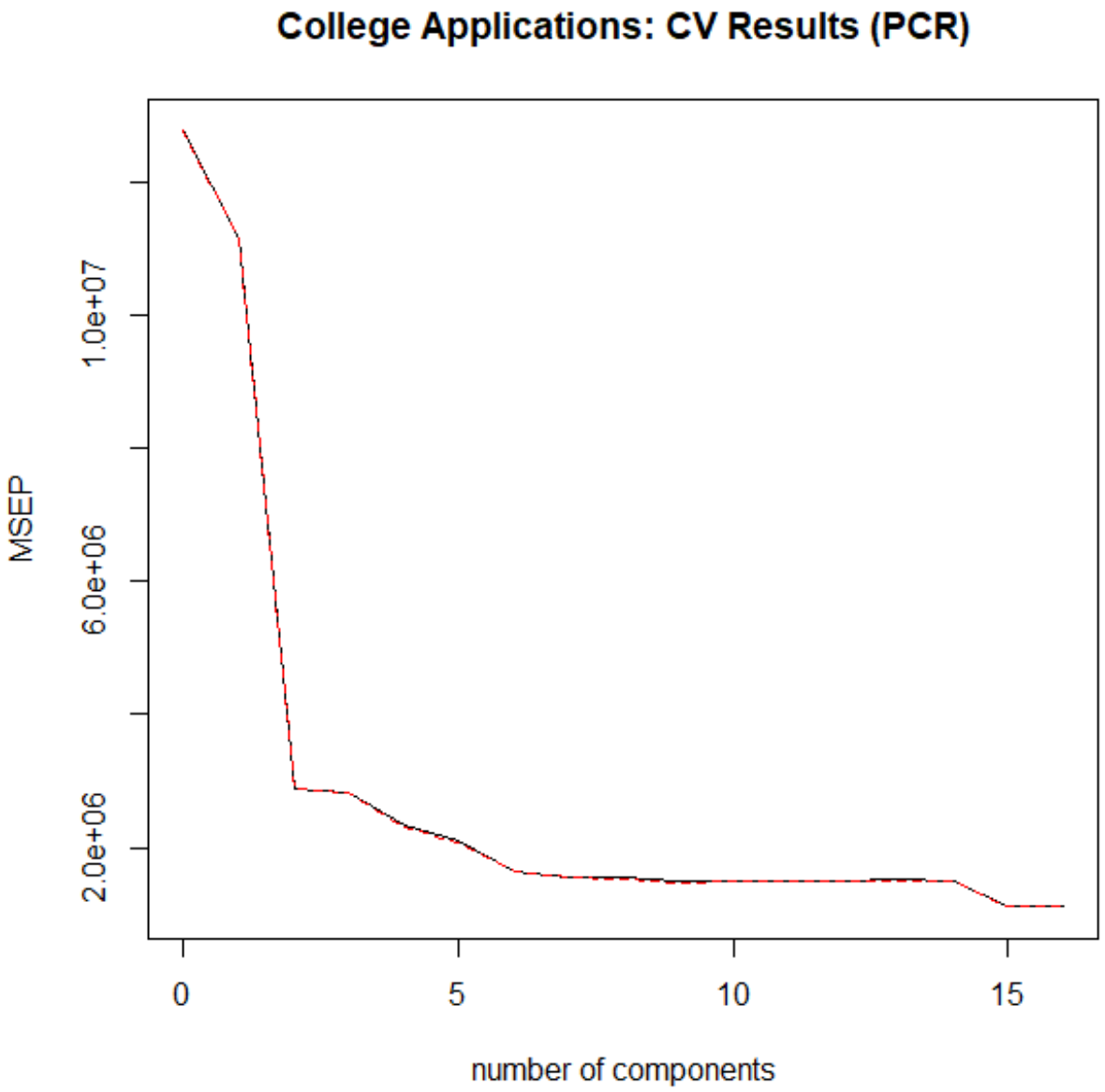


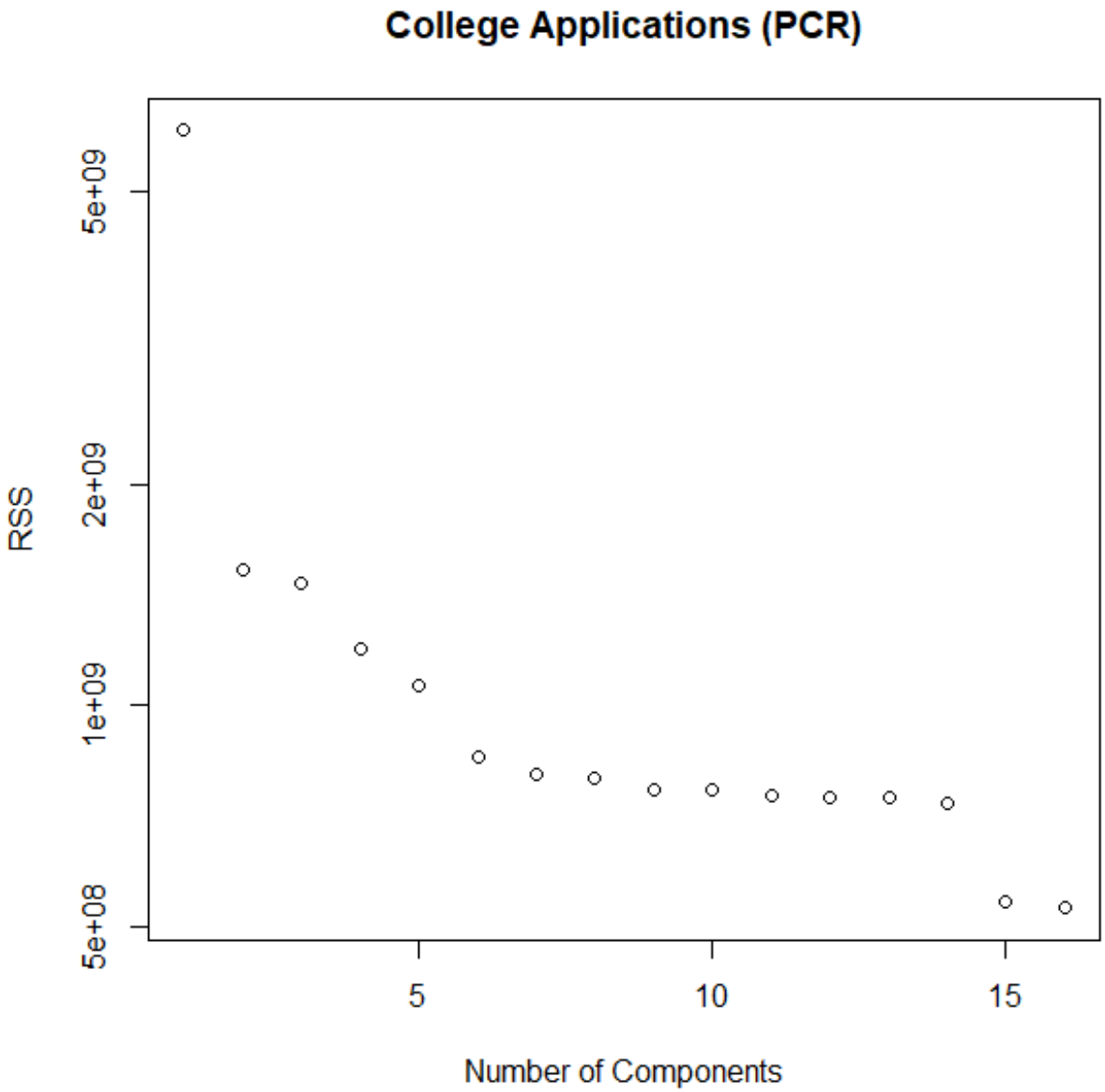
1.4 Part D

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. Principal Components Regression (PCR) was performed with cross validation (CV) on 10 folds. The result of CV was plotted.

The value, $k_b = 6$, was found to be a good trade off between model complexity and error (MSE), as such was selected from the CV results. The test error, RSS, was calculated for the training and test data, using this value of k_b.

$RSS_1 = 851608699$	training
$RSS_2 = 958264707$	test



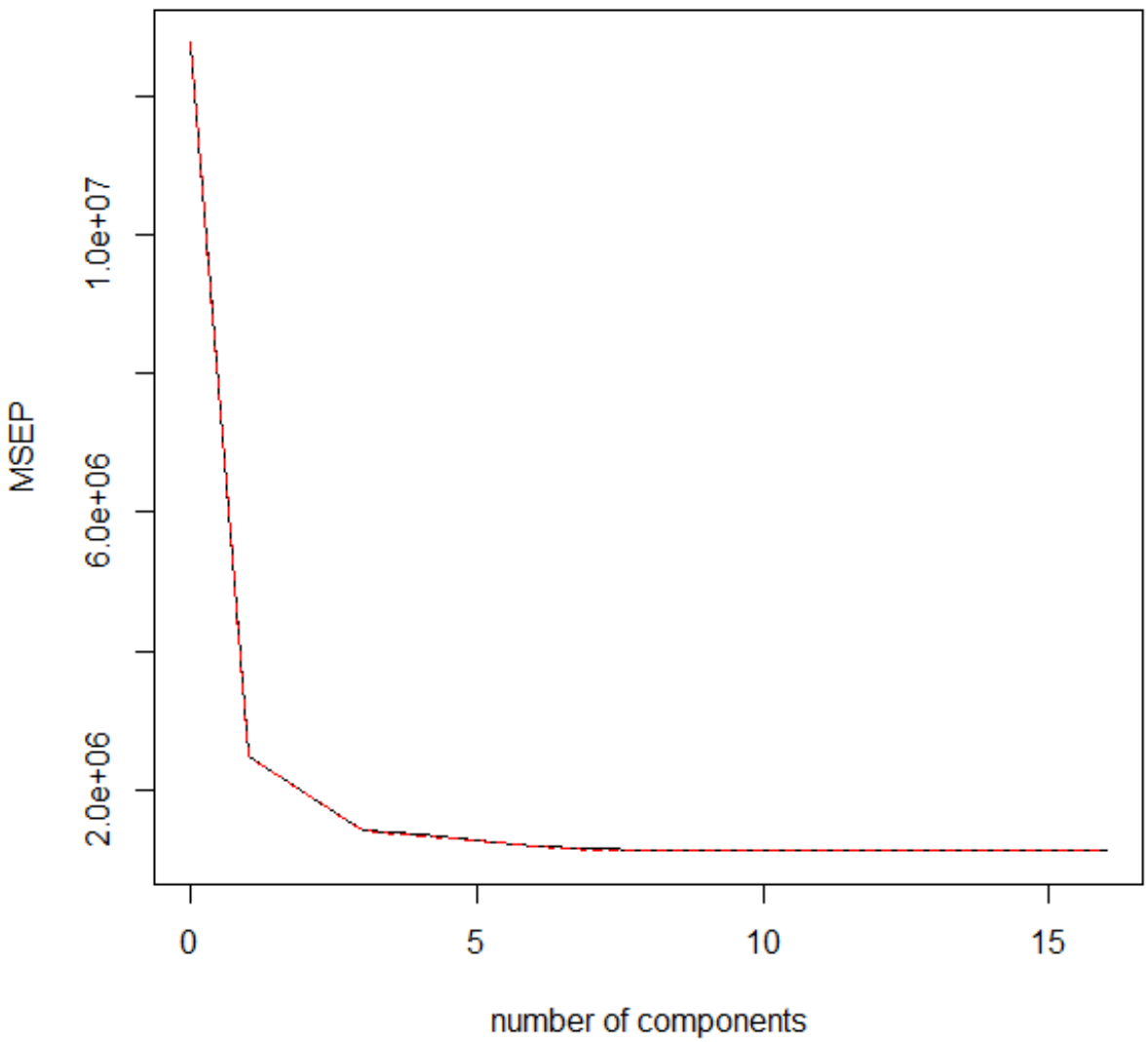


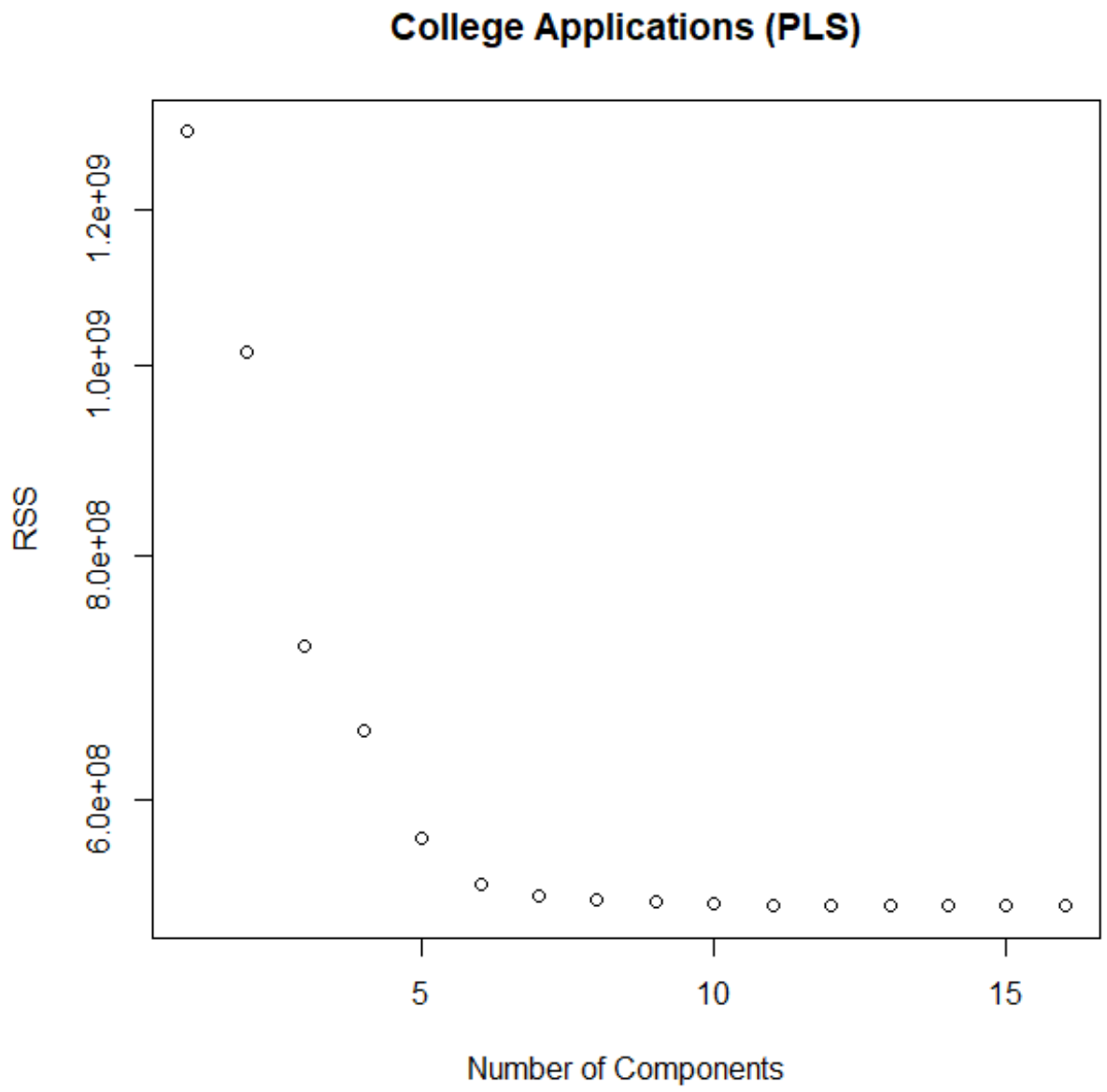
1.5 Part E

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.

Partial Least Square Regression (PLS) was performed with cross validation (CV) on 10 folds. The result of CV was plotted.
The value, $k_b = 6$, was found to be a good trade off between model complexity and error (MSE), as such was selected from the CV results. The test error, RSS, was calculated for the training and test data, using this value of k_b.

$RSS_1 = 544240736$	training
$RSS_2 = 460767507$	test


College Applications: CV Results (PLS)
MSEP
2.0e+06
6.0e+06
1.0e+07
0
5
10
15
number of components



1.6 Part 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?

Table: Summary of Results:

Model	Chosen Parameter	RSS Training	R^2 Training
Linear	None	529849177	0.92
Ridge	$\lambda_b = 365.91$	620147652	0.91
Lasso	$\lambda_b = 365.91$	531164806	0.92
PCR	$k_b = 6$	851608699	0.92
PLS	$k_b = 6$	544240736	0.92

Five methods of regression were used. Each of the methods, except conventional linear regression were optimised through cross validation, and careful selecting of paramaters.
The linear and lasso model had the best RSS on the training data. The PCR model had the worst RSS. All of the models had comparable R^2, with a range of 0.91 to 0.92.

Table: Summary of Model Test Error and:

Model	MSE Test	MAE Test	R^2 test
Linear	1,749,587.22	1,322.72	0.91
Ridge	3,102,219.55	1,761.31	0.85
Lasso	1,794,975.22	1,339.77	0.91
PCR	4,184,562.04	2,045.62	0.91
PLS	2,012,085.18	1,418.48	0.91

The Mean Squared Test Error (MSE Test), Mean Absolute Test Error (MAE Test), and R-squared (R^2) are summarised in the table above. The linear and lasso model had the best test error, and the best R^2, making them the best models. Once again, the PCR model had the worst test error (MSE and MAE). The Ridge model had the worst test R^2. All the models except PCR and Ridge predict college applications with high accuracy.

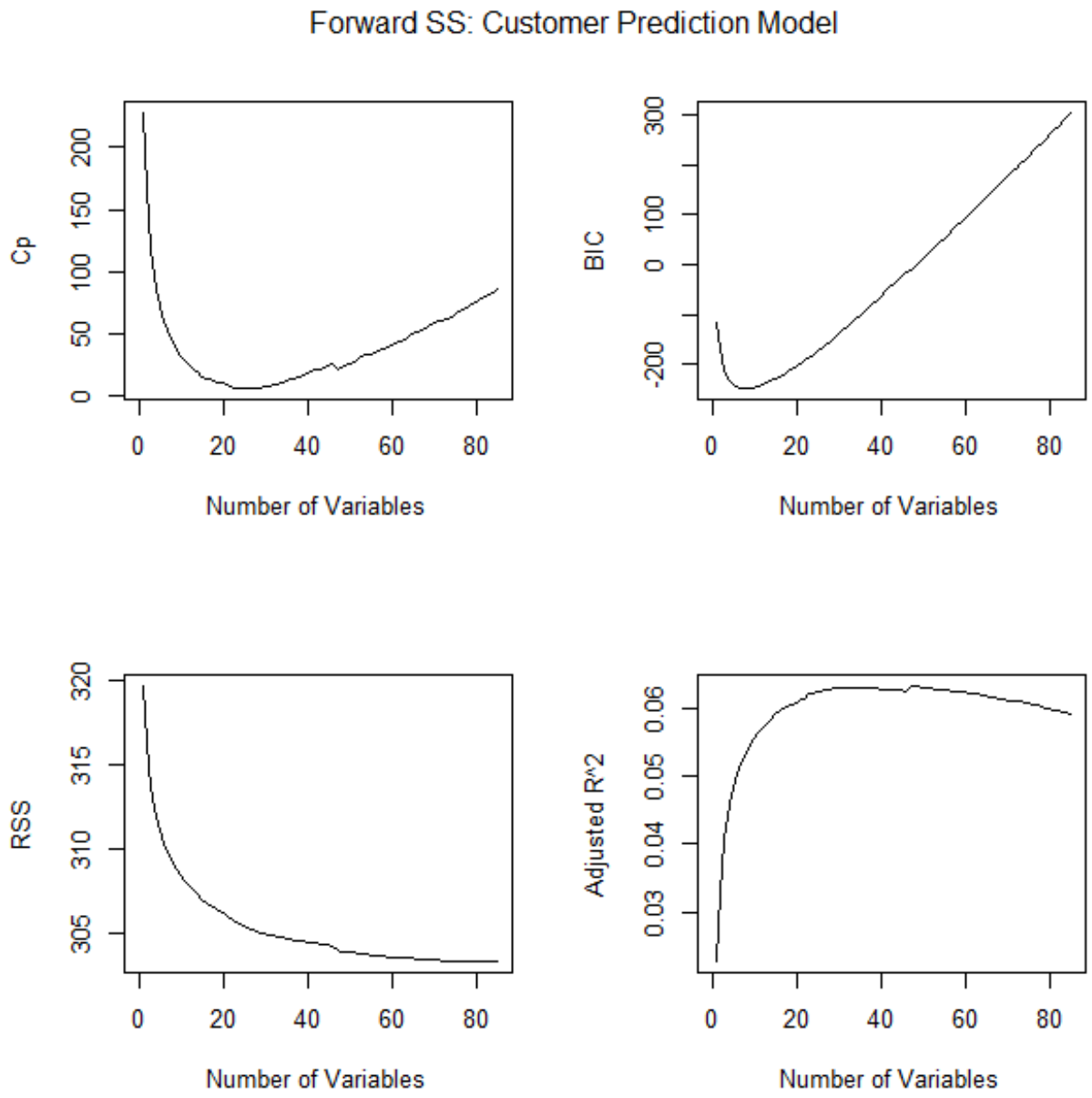
2 Exercise 2

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.

2.1 Part A - Summary

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.

Forward and Backward Selection
Forward and Backward selection was performed using the **regsubsets** function. Plots of Cp, BIC, RSS, and Adj R^2 were compared to select an optimal model based on judgment.



Lasso and Ridge Regression
Lasso and ridge regression were performed using the **glmnet** function. The cross validation plots were examined. In both cases the Λ associated with the minimum training error was selected as the optimal model.

Table: Summary of Model Test Error:

Method	Chosen Parameter	Class Error Training	Class Error Test
Forward	23 variables	5.95%	5.98%
Backward	29 variables	5.95%	5.98%
Lasso	$\lambda_b = 0.003184799$	5.95%	5.98%
Ridge	$\lambda_b = 0.1227271$	5.95%	5.96%

Four methods were used to pick optimal models that minimised RSS or MSE. The first two of the methods used, Forward and Backward Stepwise selection, are subset selection methods. The second two methods are shrinkage methods. Optimal models were chosen via each methodology. Training and test variables were used to predict the targets with rounding used to force the prediction into the binary target groups. Classification error was calculated for each of the models. The preliminary results of the exercise are summarised in the table above.

2.2 Part B

Can you predict who will be interested in buying a caravan insurance policy and give an explanation why?

While the classification error seems good on closer inspection of the raw prediction data it seems the all predictive models rarely predicted the class Customer. Class customer was predicted less than 0.60% for training, and less than 1.30% for the test set.

Table: Confusion Matrix Ridge Regression Training Prediction

	Actual NO	Actual YES
Predicted NO	TN = 5474	FN = 347
Predicted YES	FN = 0	TP = 1

This demonstrates that the models above, despite their reasonable classification error, are not suitable for predicting who will be a potential customer, due to their low sensitivity (less than 1%). This a form of masking and is most probably due to the low support of class customer in the data. Class customer makes up only 5.98% of the training data.
Experimenting with the threshold for rounding, and only examining the Ridge Regression optimal model as an example, we can raise the sensitivity (TP / P) to enhance prediction of customers. Lowering the threshold from 0.5 to 0.1475 we increase sensitivity from 0.29% to 18.68% (a factor of 65), at the expense of the classification error rate, which only increases from 5.96% to 7.54% (a factor of 1.3). Even with this optimising, as is we would at best be only able to predict around $5/20$ potential customers on training data, and $3/20$ on the test data.

Table: Confusion Matrix Ridge Regression Training Prediction

	Actual NO	Actual YES
Predicted NO	TN = 5318	FN = 283
Predicted YES	FN = 156	TP = 65

3 Exercise 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.

3.1 Part A

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

A data set was generated with 20 features and 1000 observations. Five (5) coefficients β were set to equal exactly zero (0). Through experimentation it was found that by using and exponential distribution to estimate the β produced a situation where the test mean square error (MSE) was minimised for an intermediate model size.

Initializing Model Parameters, and Data

```
set.seed(123)
X <- matrix(rnorm(1000 * 20), 1000, 20)
b <- rexp(20, rate = 10)
b[3] <- 0
b[4] <- 0
b[9] <- 0
b[19] <- 0
b[10] <- 0
eps <- rnorm(1000)
Y <- X %*% b + eps
```

3.2 Part B

Split your data set into a training set containing 100 observations and a test set containing 900 observations.

As required for the exercise the data set was randomly split into a training set containing 100 observations and a test set containing 900 observations.

Initializing Model Parameters, and Data

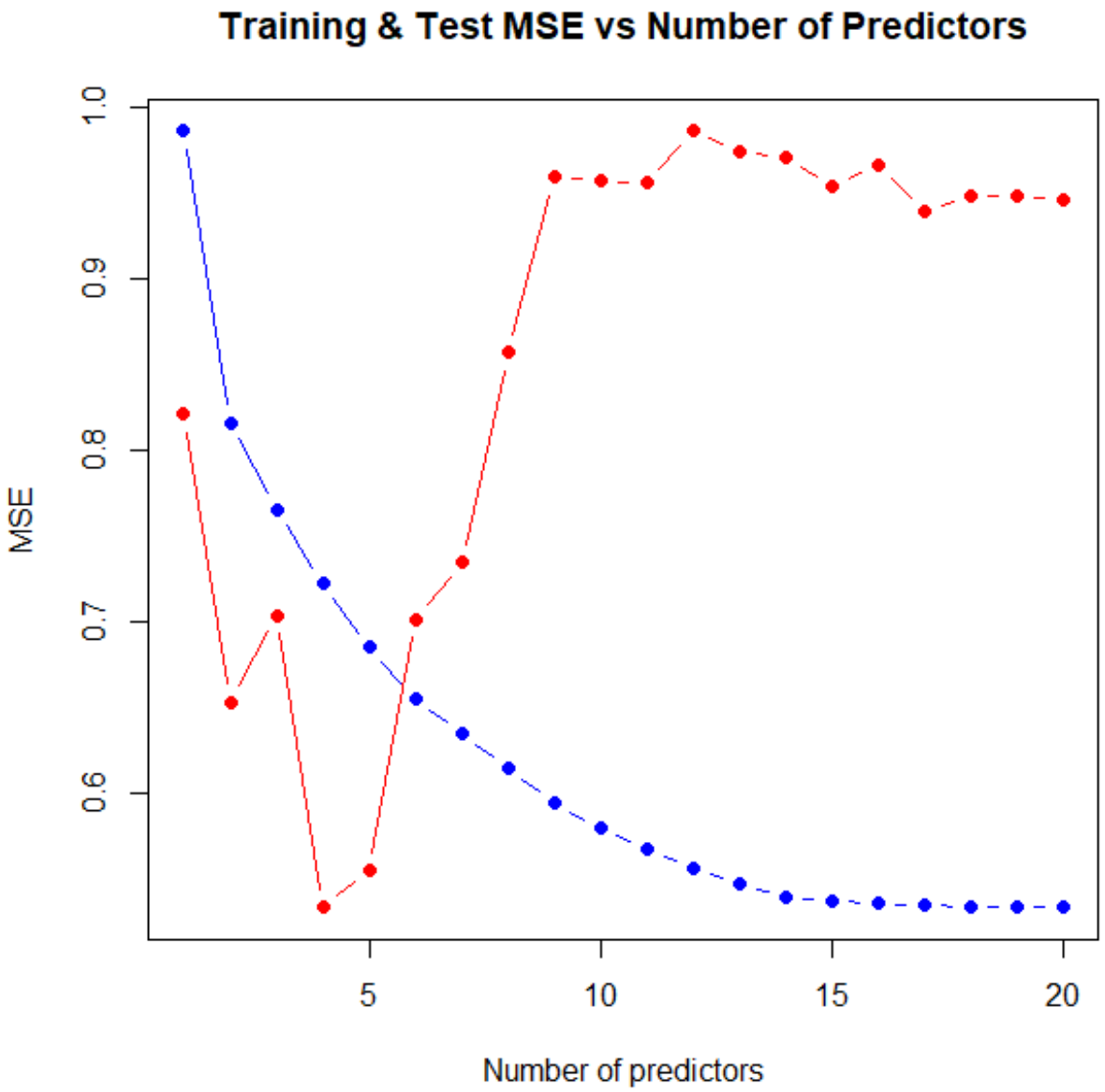
```
train <- sample(seq(1000), 100, replace = FALSE)
test <- -train
X_train <- X[train, ]
X_test <- X[test, ]
Y_train <- Y[train]
Y_test <- Y[test]
```

3.3 Part C

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.

Best subset selection was performed on the training set for each size model, using the **regsubsets** function. Mean Squared Error (MSE) was calculated for the best model, at each size, for both the training and the test data. The results are plotted below.

MSE Training Test



As one would expect, the training error (MSE) steadily decreases as model complexity increases. The test error (MSE), however, decreases to a minimum value and then increases again and levels out. The minimum test error occurs for 4 predictors.

Coefficients: Best model on test (n = 4)

```
(Intercept) -0.01570578
x.1          0.24834808
x.6          0.41674635
x.7          0.24888794
x.13         0.49278354
```

As we can see above, as expected, the model through best subset selection, has excluded all coefficients of $\beta = 0$. The model that performed the best on the test data, however, did not include all 15 of the other non-zero predictors.

Coefficients: Best model on test (n = 15)

```
(Intercept) -0.01459019
x.1          0.18736100
x.2          0.09458982
x.3         -0.15326133
x.6          0.37031497
x.7          0.19422254
x.8          0.11205394
x.10        -0.12397636
x.12         0.04854990
x.13         0.57042817
x.14         0.13379743
x.15         0.17787217
x.16        -0.23160279
x.18        -0.12781347
x.19         0.13465991
x.20         0.22009660
```

Examining the model for 15 predictors, we may expect that the model will only be compromised of the non-zero β's. The model, however, has include the estimated β's x.3, x.10,and x.19, whose true values were set to zero (0). This error in estimating the true β, including zero values, may be due to the abnormal data split training data (n = 100) compared to the test data (n = 900). Increasing the data in the training set may enhance the accuracy of the predicted β's, and possibly more of the zero values would be correctly excluded. For reference the true coefficients are listed below.

Coefficients: True Coefficients of Model

```
1  0.199824626
2  0.001227591
3  0.000000000
4  0.000000000
5  0.078749169
6  0.330521101
7  0.171918534
8  0.023766421
9  0.000000000
10 0.000000000
11 0.032068824
12 0.054017144
13 0.412234182
14 0.008339280
15 0.077219498
16 0.061528507
17 0.031981094
18 0.010562764
19 0.000000000
20 0.089058919
```