# Mathematics for Big Data - Exercise 2a

*Lawrence Adu-Gyamfi (1484610)*

*06/06/2019*

## Contents

# 1 INTRODUCTION

This report presents the code and the results of analysis and predicting the number of applications received using the variables in the College dataset from the ISLR package.

The models considered for the predictions include:

- Linear Model
- Ridge regression
- Lasson regression

```
## corrplot 0.84 loaded
```

```
## Loading required package: Formula
```

```
## Loading required package: plotrix
```

```
## Loading required package: TeachingDemos
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

## 1.1 Explore Dataset

The following section presents some general details about the College dataset.

The variables and the dimensions of the dataset are:

```
names(College)
```

```
##  [1] "Private"     "Apps"        "Accept"      "Enroll"      "Top10perc"
##  [6] "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"    "Room.Board"
## [11] "Books"       "Personal"    "PhD"         "Terminal"    "S.F.Ratio"
## [16] "perc.alumni" "Expend"      "Grad.Rate"
```

```
dim(College)
```

```
## [1] 777  18
```

```
## [1] "There are  0 na values in the dataset"
```

Below is a plot showing the correlation between the variables of the dataset

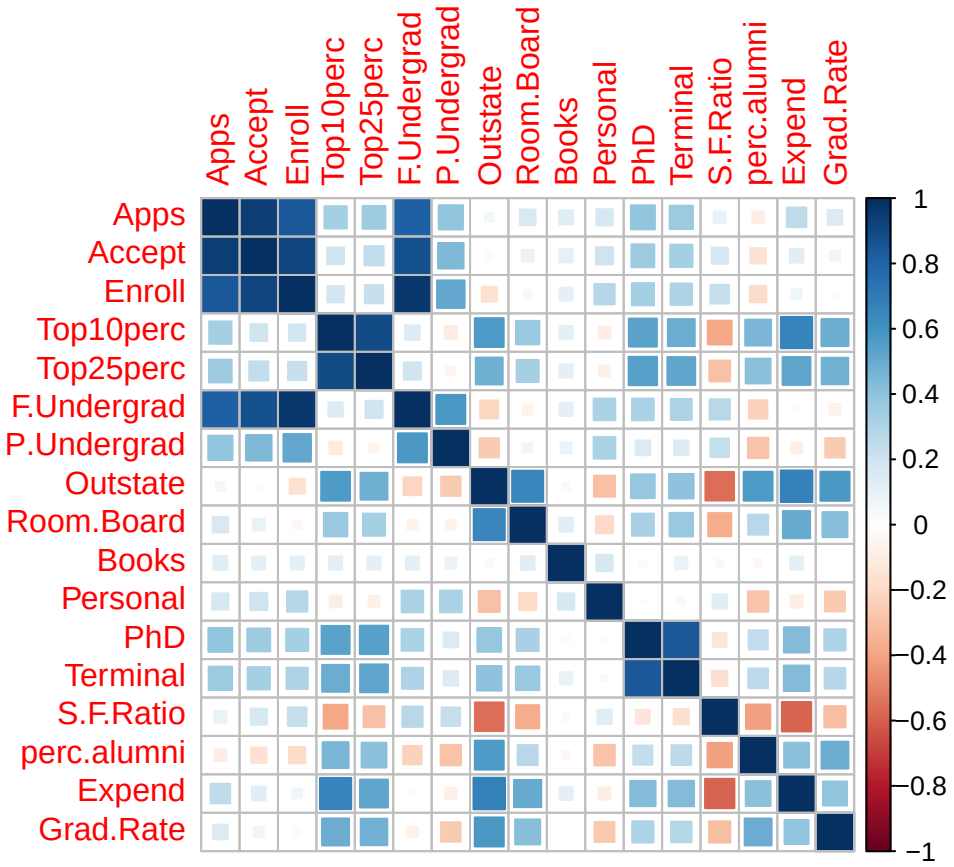


Figure 1: Plot showing correlation of variables in dataset

# 2 MODELLING AND RESULTS

This section details the preparation of the data as well as the modelling process for all the 3 models considered. The results from the models applied to the test dataset for the predictions are also presented.

## 2.1 Data Splitting

The following code splits the data set into a training set and a test set to be used for the analysis and prediction.

```
set.seed(1234)
x = model.matrix(Apps~.,College)
y = College$Apps

train = sample(1:nrow(x), nrow(x)/2)
test = (-train)
```

```
y.test = y[test]
```

## 2.2 Linear Model

**Question: Fit a linear model using least squares on the training set, and report the test error obtained.**

Presented below is the process of model selection using the regsubsets package. The maximum number of variables are considered for the selection.

Plots are presented below as well for the main evaluation criteria and the number of variables that give the optimum value for the corresponding criteria.

```
## Minimum residual obtained is 823059948.125  for model with  17
##  variables
```

```
## [1] "Maximum r-squared obtained is 0.929188695211609  for model with  13 variables"
```

```
## [1] "Minimum Cp obtained is 10.0816057153932  for model with  12 variables"
```

```
## [1] "Minimum BIC is -1974.50195409044  for model with  10 variables"
```

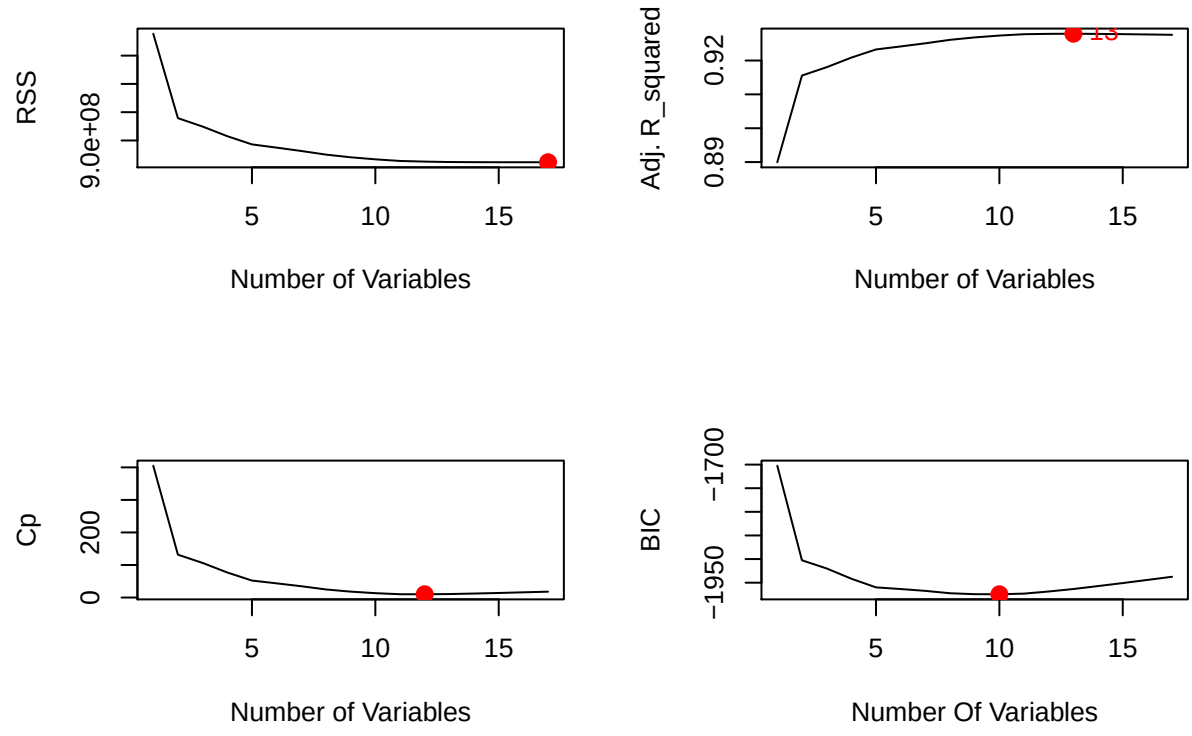


Figure 2: Model evaluation using different criteria

Below, the the linear model is fit on only the train data to make the predictions and check the coefficients. For each model with a different number of variable(x) from 1 to the maximum

in the dataset (17), the mean squared error is evaluated and reported.

```
## Best model is obtained using  9  variables, with a mean squared error of
##  1253366.44366584
```

```
## Below are the coefficients for the variables selected, together with that
## of the intercept
```

```
##  (Intercept)   PrivateYes       Accept       Enroll    Top10perc
## -439.2802468 -370.8340084    1.6480129   -0.6831114   28.6065969
##     Outstate   Room.Board          PhD       Expend    Grad.Rate
##   -0.1207679    0.1255534   -7.9957587    0.1034671    6.1472587
```

Below are the results for performing the same steps as above but using using k-fold cross validation

```
## The means MSE for each of the models are :
```

```
##       1       2       3       4       5       6       7       8       9
## 1714855 1358911 1374875 1384853 1366058 1390442 1368999 1340010 1328096
##      10      11      12      13      14      15      16      17
## 1370767 1329959 1342024 1345768 1346639 1345087 1343742 1343196
```

The following plot shows the MSE for each of the models, indicating the number of variables that gives the lowest error.

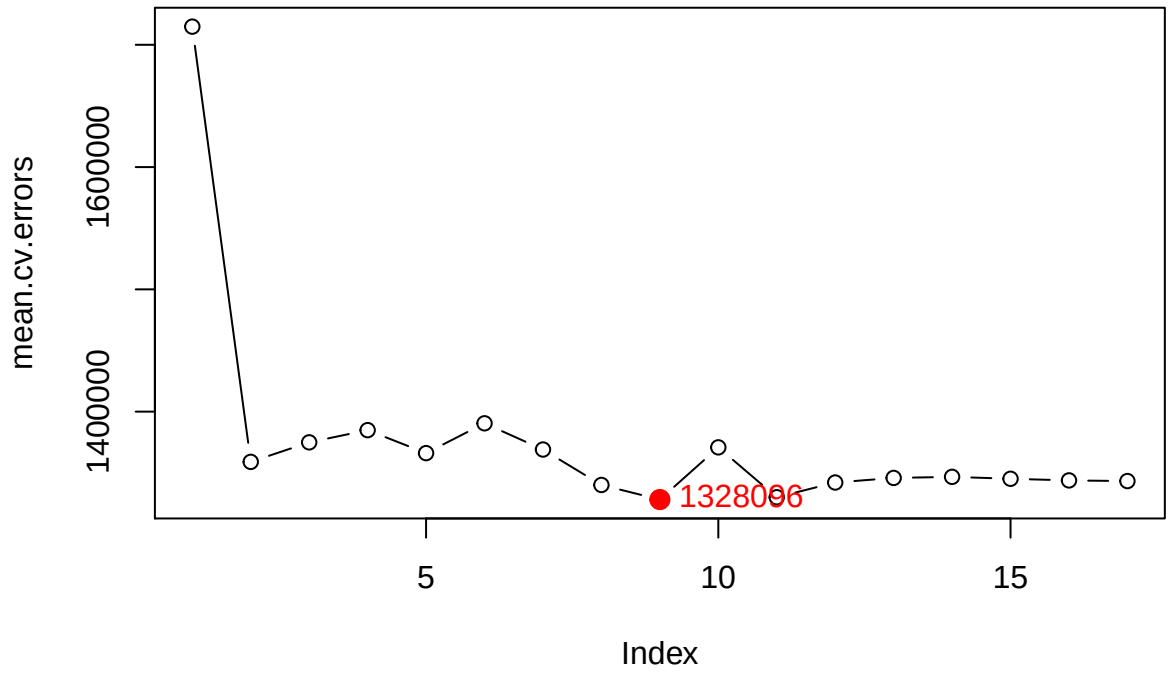


Figure 3: Plot of cross-validation errors for models based on number of variables used.

```
# Fit model on full data to inspect coefficients
reg.best =regsubsets(Apps~.,College, nvmax=17)
coef(reg.best, 11)
```

```
##   (Intercept)    PrivateYes        Accept        Enroll     Top10perc
## -134.72958560 -521.31605948    1.58345760   -0.90094571   49.68836054
##     Top25perc   F.Undergrad      Outstate    Room.Board           PhD
##  -14.71780506    0.07186638   -0.09066672    0.15538909  -10.36195007
##        Expend     Grad.Rate
##    0.07312122    7.99655604
```

## 2.3 Ridge Regression Model

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

```
#### Ridge Regression

grid = 10^seq(10, -4, length=100)
ridge.mod<-glmnet(x[train,], y[train], alpha=0, lambda=grid, thresh=1e-12,
                  standardize = T)
plot(ridge.mod)
```

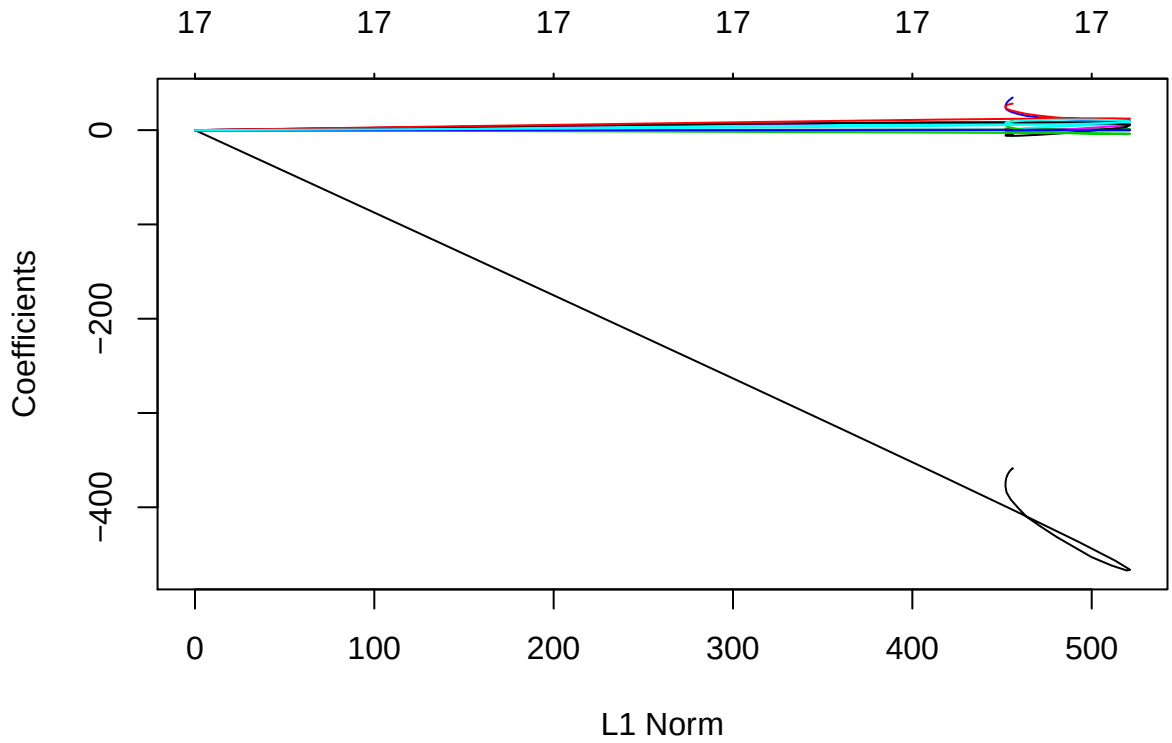


Figure 4: Ridge Regression Coefficients

Below are the results of the evaluation of lambda using cross validation.

```
set.seed(100)
cv.out = cv.glmnet(x[train,], y[train], alpha=0)
plot(cv.out)
```

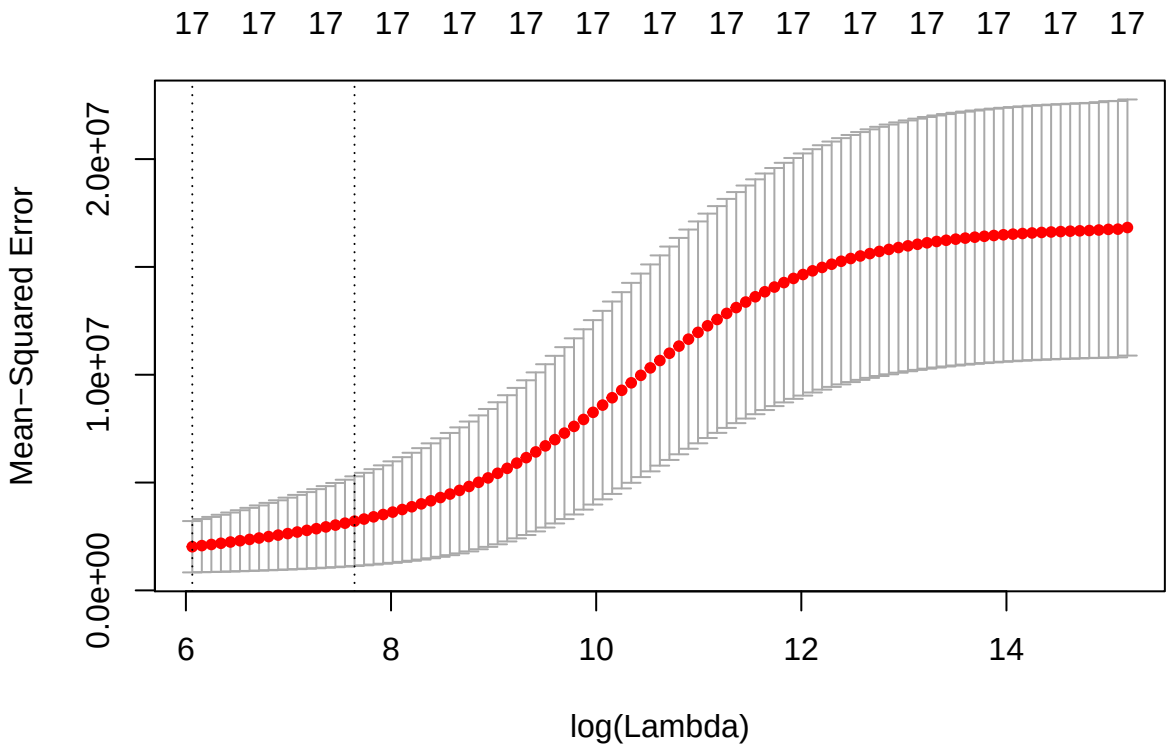


Figure 5: Evaluation of Lambda with Cross-Validation for Ridge Model

```
bestlam = cv.out$lambda.min
print(paste("The best lambda is ", bestlam))
```

```
## [1] "The best lambda is  429.353361877126"
```

```
ridge.pred  = predict(ridge.mod, s=bestlam, newx=x[test,])
ridge.mse = mean((ridge.pred-y.test)^2)
```

```
## [1] "The mean squared error for the predictions using the ridge model is 1291386.15"
```

```
##   (Intercept)   (Intercept)    PrivateYes        Accept        Enroll
## -1.546244e+03  0.000000e+00 -5.309259e+02  9.600152e-01  4.899367e-01
##     Top10perc     Top25perc   F.Undergrad   P.Undergrad      Outstate
##  2.443075e+01  1.396978e+00  7.927307e-02  2.490613e-02 -1.947353e-02
##    Room.Board         Books      Personal           PhD      Terminal
##  2.003817e-01  1.402378e-01 -9.310600e-03 -3.586139e+00 -4.619282e+00
##     S.F.Ratio   perc.alumni        Expend
##  1.264182e+01 -8.996098e+00  7.470976e-02
```

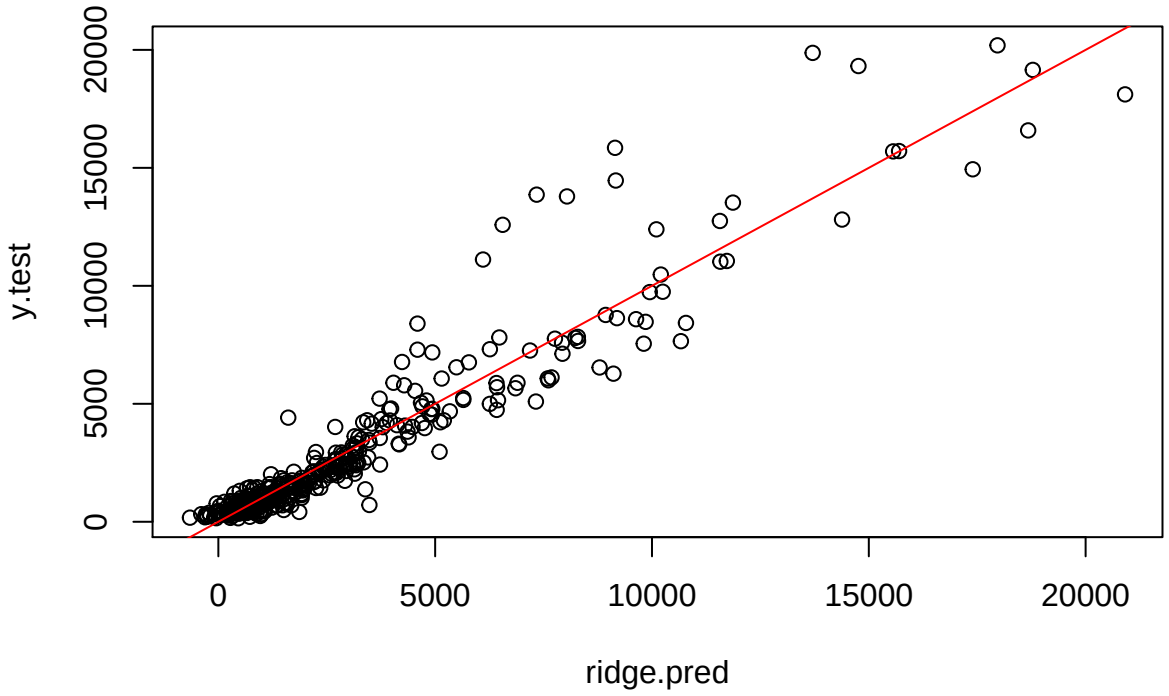


Figure 6: Predictions agains Actuals for Ridge Model

## 2.4 Lasso Regression Model

**Question: Fit a lasso model on the training set, with lambda chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.**

Below are the results after fitting the lasso model on the training set:

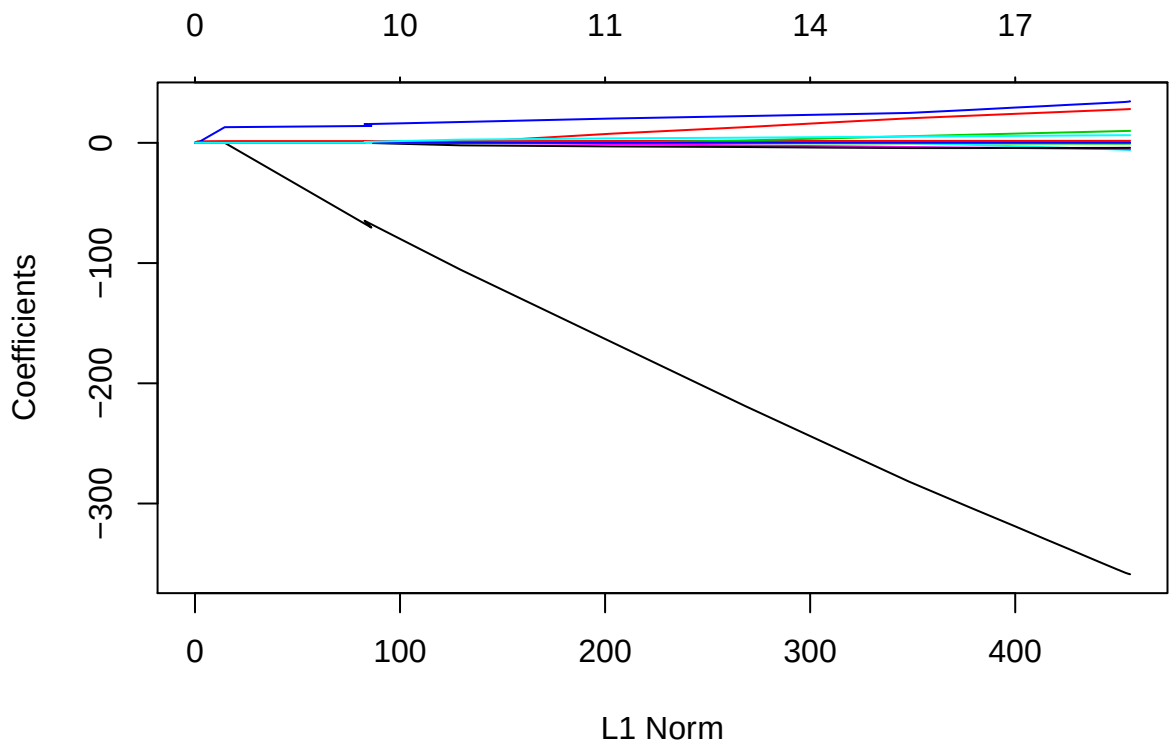


Figure 7: Lasso Model Coefficients against L1 Norm

The plot below shows the cross-validation results for selecting lambda for the lasso model:

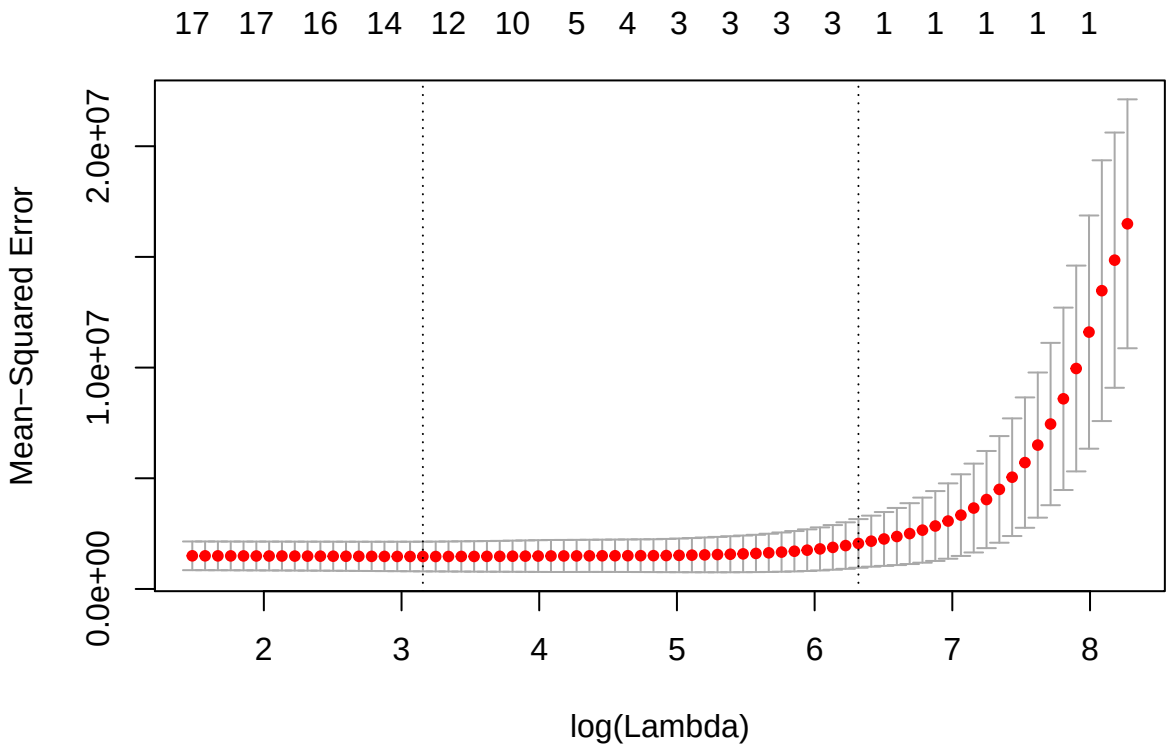


Figure 8: Evaluation of Lambda with Cross-Validation for Lasso Model

```
## [1] "The best lambda is  23.452"
```

```
## [1] "The mean squared error for the predictions using the lasso model is 1269532.22"
```

```
##                lasso.coef
## (Intercept) -6.220017e+02
## PrivateYes   0.000000e+00
## Accept      -4.142321e+02
## Enroll       1.443788e+00
## Top10perc   -1.633480e-01
## Top25perc    3.227667e+01
## F.Undergrad -1.442060e+00
## P.Undergrad  0.000000e+00
## Outstate     1.699106e-02
## Room.Board  -5.522293e-02
## Books        1.223839e-01
## Personal     0.000000e+00
## PhD          1.670811e-04
## Terminal    -5.289167e+00
## S.F.Ratio   -3.360294e+00
## perc.alumni  3.356283e+00
## Expend      -1.007700e+00
## Grad.Rate    6.889514e-02
```

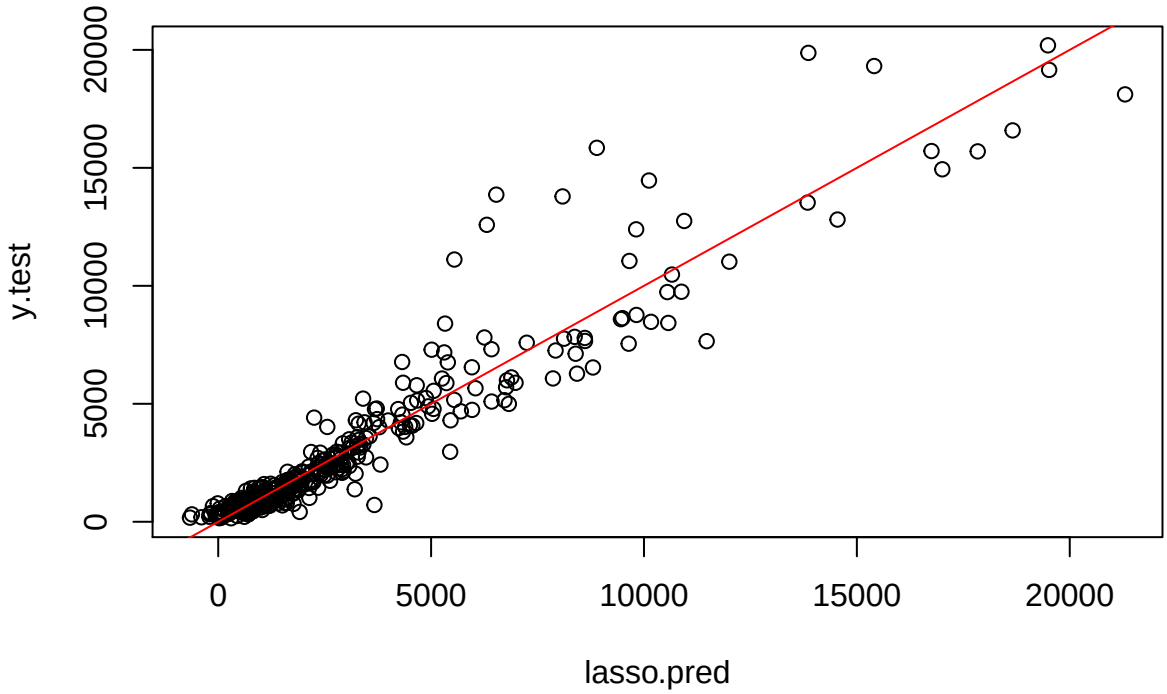


Figure 9: Predictions agains Actuals for Lasso Model