DATA ANALYSIS PROJECT

BY:
SAMUEL ZOWAM
PRABHA PADAMSINGH KANYAL

EXECUTIVE SUMMARY:

The objective of this project was to come up with a statistical learning method for a regression problem that would provide a high prediction accuracy when applied to previously unseen test data. Hence this report is the outcome of the study we undertook to investigate different models and ultimately select the method we deemed to be the best.

Our approach to the problem was to analyze and compare different methods learned in class and to assess the performance of each method based on some approximation of the test mean square error – a measure of the accuracy in predicting the test data. For our analysis, we were guided by the knowledge of the structure of our data which includes the relationships between the independent variables and the response variable as well as the relationship between pairs of the independent variables. We then selected three best models comprising:

1. Model1: Multiple linear regression model with an interaction term and transformation of the output variable
2. Model 2: best subsets selection method
3. Model 3: Lasso regression method.

We found the best subset selection method to be the overall best model because it had the lowest approximation of the test mean square error. We also calculated the true test mean square error given the actual test data.

Furthermore, in the model improvement phase of the project, we revised our decision of our best model. We then sought ways to achieve a better prediction accuracy by analyzing other models not previously considered. Eventually, we concluded that regression trees improved by boosting in fact provided the best prediction accuracy.

MODEL 1: Multiple Linear Regression (MLR) including an interaction term and a transformation of the response variable.

First, after loading the dataset, we make graphs of the independent variables versus each other and independent variables versus the dependent variable to see the relationships between all the variables. The goal will be to have the independent variables correlated strongly with the dependent variable, but not with each other.

From the plot shown below, it becomes immediately obvious that all the independent variables (excluding X6 and X8) appear strongly corelated with the response variable; and we also, in some cases, see that the independent variables appear strongly correlated with each other (X4 on X1, X2, & X5; and X1 and X2). This knowledge is useful and will guide us through our analysis, as we will see.

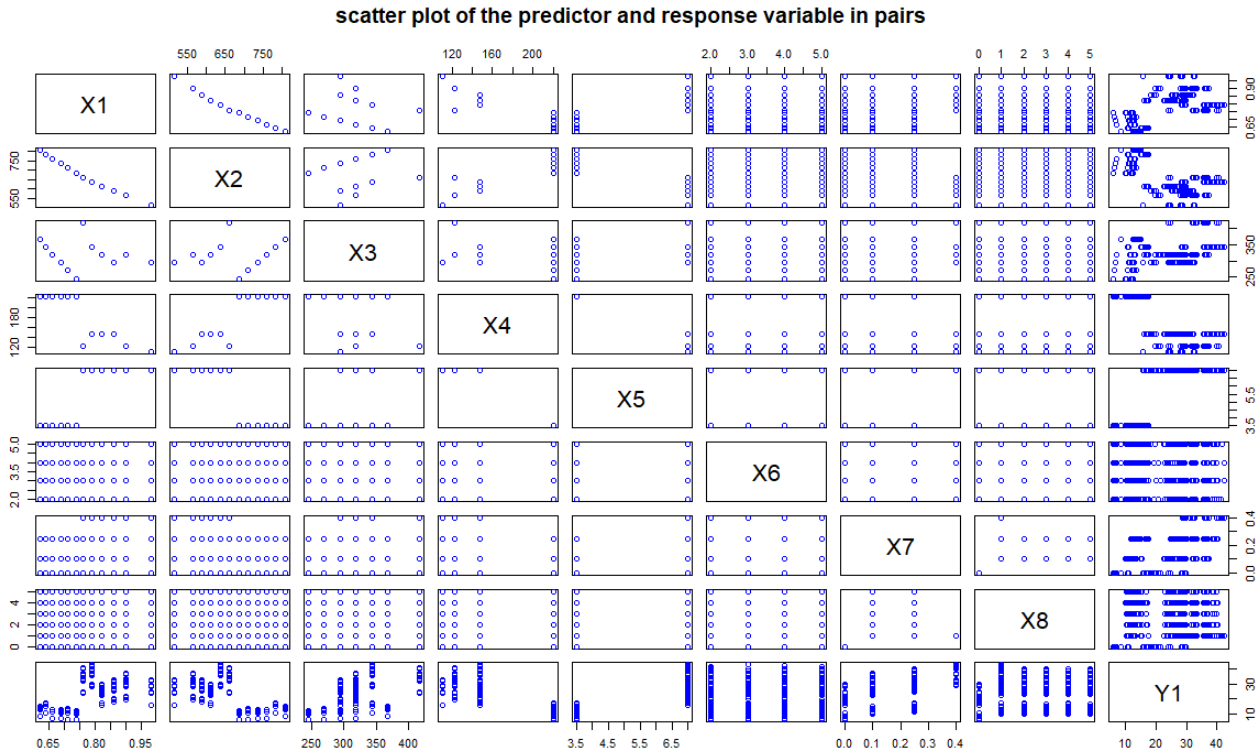


Fig1: scatter plot of the predictors and response variables in pairs

Next, we conduct an MLR of Y1 on all the variables and obtain the results shown in Fig 2a. The coefficients for X4 are undefined (aliased), meaning that X4 is linearly dependent on other variables and causes perfect collinearity - hence it shows up as "NA" in our regression results. We therefore remove it and repeat the MLR of Y1 on other 7 variables and obtain the results shown in Fig 2b.

```
Call:
lm(formula = Y1 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0807 -1.3009 -0.0774  1.3452  8.0654

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.37396   23.04731   3.574 0.000383 ***
X1          -65.63796   12.44078  -5.276 1.91e-07 ***
X2           -0.08271    0.02065  -4.005 7.06e-05 ***
X3            0.05537    0.00791   7.000 7.58e-12 ***
X4                 NA         NA      NA       NA
X5            4.24076    0.40360  10.507  < 2e-16 ***
X6           -0.03053    0.11385  -0.268 0.788706
X7           23.87709    1.32655  17.999  < 2e-16 ***
X8            0.24579    0.08056   3.051 0.002392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.985 on 542 degrees of freedom
Multiple R-squared:  0.9095,    Adjusted R-squared:  0.9083
F-statistic: 777.7 on 7 and 542 DF,  p-value: < 2.2e-16
```

Fig2a: regression results with all variables

```
Call:
lm(formula = Y1 ~ X1 + X2 + X3 + X5 + X6 + X7 + X8)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0807 -1.3009 -0.0774  1.3452  8.0654

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.37396   23.04731   3.574 0.000383 ***
X1          -65.63796   12.44078  -5.276 1.91e-07 ***
X2           -0.08271    0.02065  -4.005 7.06e-05 ***
X3            0.05537    0.00791   7.000 7.58e-12 ***
X5            4.24076    0.40360  10.507  < 2e-16 ***
X6           -0.03053    0.11385  -0.268 0.788706
X7           23.87709    1.32655  17.999  < 2e-16 ***
X8            0.24579    0.08056   3.051 0.002392 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.985 on 542 degrees of freedom
Multiple R-squared:  0.9095,    Adjusted R-squared:  0.9083
F-statistic: 777.7 on 7 and 542 DF,  p-value: < 2.2e-16
```

fig2b: regression results excluding X4

VIF	
X1	107.8
X2	204.7
X3	7.2
X5	30.7
X6	1.0
X7	1.0
X8	1.0

fig2c: VIF

However, there is serious collinearity among the variables (fig 2c) with X2 being most problematic (VIF >>5), and so we drop it as well. Repeating the MLR of Y1 on all the other variables (excluding X2 and X4), we get regression results, but again detect serious collinearity (this time with X5 being the most problematic variable; VIF >5). These results were in fact hinted by our correlation plot (fig1) and so it is not surprising at all!

In our refined model, we now have X1, X3, X6, X7 and X8 as our predictors and Y1 as our response variable. Proceeding from here, we try different interaction terms in our regression model and find the interaction between X7 and X8 to be statistically significant (fig 3a).

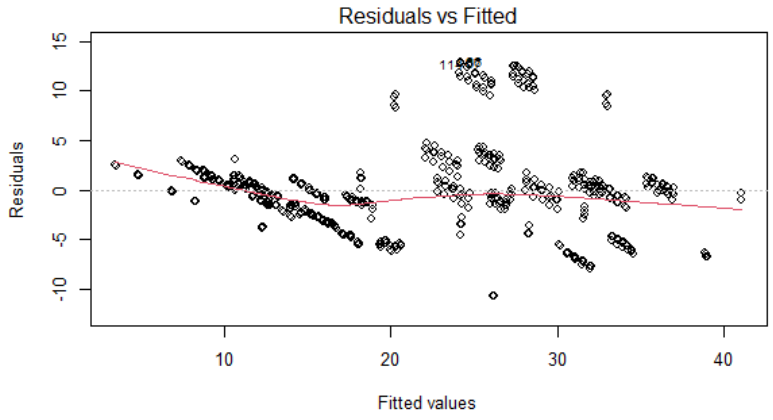
```
Call:
lm(formula = Y1 ~ X1 + X3 + X6 + X7 + X8 + X7:X8)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6959  -2.1774  -0.2687   1.1800  12.9324

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -79.011245   2.279397 -34.663  < 2e-16 ***
X1           66.495230   1.798938  36.964  < 2e-16 ***
X3            0.136559   0.004386  31.132  < 2e-16 ***
X6           -0.028217   0.165972  -0.170    0.865
X7           34.455566   2.893044  11.910  < 2e-16 ***
X8            0.981461   0.212616   4.616 4.88e-06 ***
X7:X8        -5.144469   1.176164  -4.374 1.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.351 on 543 degrees of freedom
Multiple R-squared:  0.8072,    Adjusted R-squared:  0.8051
F-statistic: 378.9 on 6 and 543 DF,  p-value: < 2.2e-16
```

Fig3a: regression results including interaction (excluding X2, X4, X5)



F ig3b: residual plot of the regression in 3a

However, model diagnostics (fig 3b) indicate a potential problem: The assumption of constant variance of OLS is violated, implied by a funnel shape of the residuals versus fitted plot. Doing a BoxCox transformation of the form $Y_1^{\lambda} = X_i + \in$ to find the correct relationship between X and Y, we get that the Maximum Likelihood Estimation for lambda is zero (fig 4a), and hence we transform the response variable using a log transformation.

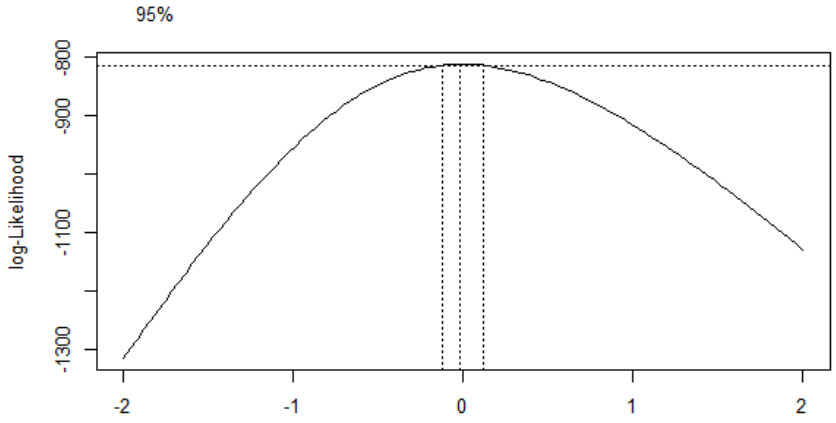


Fig4a: maximum Likelihood estimation of lambda

```
Call:
lm(formula = Y1_T ~ X1 + X3 + X6 + X7 + X7:X8 + X8)

Residuals:
      Min        1Q    Median        3Q       Max
-0.188474 -0.044978 -0.001018  0.031723  0.235471

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.509e-01  4.275e-02 -22.244  < 2e-16 ***
X1           1.495e+00  3.374e-02  44.307  < 2e-16 ***
X3           2.948e-03  8.226e-05  35.835  < 2e-16 ***
X6          -2.288e-04  3.113e-03  -0.074    0.941
X7           7.807e-01  5.426e-02  14.388  < 2e-16 ***
X8           2.861e-02  3.987e-03   7.175 2.39e-12 ***
X7:X8       -1.295e-01  2.206e-02  -5.872 7.51e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08161 on 543 degrees of freedom
Multiple R-squared:  0.8544,    Adjusted R-squared:  0.8528
F-statistic: 530.9 on 6 and 543 DF,  p-value: < 2.2e-16
```

Fig4b: regression results including transformation of Y1

VIF	
X1	1.1
X3	1.0
X6	1.0
X7	2.3
X8	3.4
X7:X8	5.2

Fig3b: VIF of 4b

After the transformation, the residual plot no longer shows a funnel shape or pattern (fig4d), indicating the right transform was used, and none of the predictors have serious collinearity between them (VIF ≤ 5).

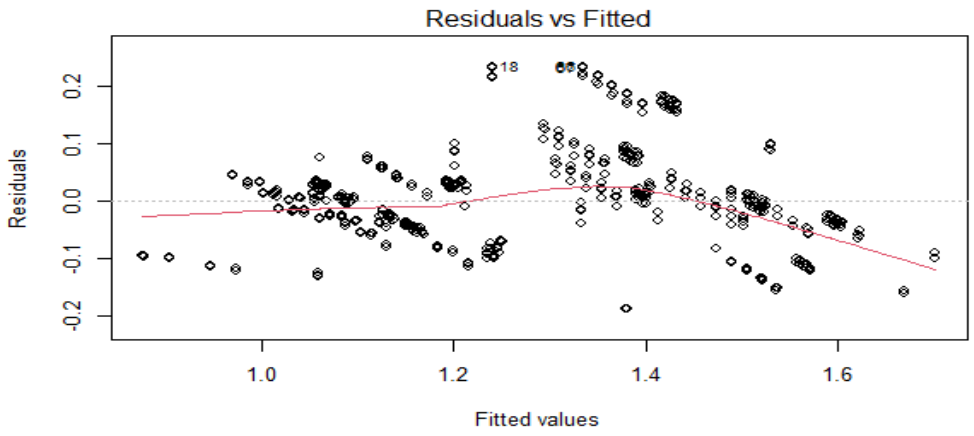


Fig4d: Residual plot of 4b

Model 1: $\log Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_6 + \beta_4 X_7 + \beta_5 X_8 + \beta_6 X_7 : X_8$

$$\log Y_1 = -0.951 + 1.495 X_1 + 0.00295 X_3 + -0.00023 X_6 + 0.781 X_7 + 0.0286 X_8 - 0.1295 X_7 : X_8$$

Hence X1, X3, X8 have positive relationships on Log Y1, while X6 and the interaction between X7 and X8 have negative relationships on Log Y1. X6 is not statistically significant (P value > 0.05, 95% confidence level).

The R^2 (proxy for training error) tells us that 85.4% of the variation in LogY1 is explained by model 1.

MODEL 2: BEST SUBSETS

Here, we want to choose a subset of the predictors such that we get the best accuracy using the linear model with the chosen subset.

However, a linear dependency was found when trying to fit the full model using all the predictors, hence the maximum model size is the model with 7 predictors. Consequently, the best models for each model size, ranging from a model with only one predictor (M1) to a model with 7 predictors (M7) is shown below.

```
8 Variables  (and intercept)
   Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
X4      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X5  X6  X7  X8
1  ( 1 ) " " " " " " " " "*" " " " " " "
2  ( 1 ) " " " " " " " " "*" " " "*" " "
3  ( 1 ) " " " " "*" " " "*" " " "*" " "
4  ( 1 ) "*" " " " " " " "*" "*" " " "*" " "
5  ( 1 ) "*" " " " " " " "*" "*" " " "*" "*"
6  ( 1 ) "*" "*" "*" " " " " "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" " " " " "*" "*" "*" "*"
```

In equation form, the different models are:

$$M1: Y_1 = \beta_0 + \beta_1 X_5$$

$$M2: Y_1 = \beta_0 + \beta_1 X_5 + \beta_2 X_7$$

$$M3: Y_1 = \beta_0 + \beta_1 X_3 + \beta_2 X_5 + \beta_3 X_7$$

$$M4: Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_5 + \beta_4 X_7$$

$$M5: Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_5 + \beta_4 X_7 + \beta_5 X_7$$

$$M6: Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8$$

$$M7: Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_5 + \beta_5 X_6 + \beta_6 X_7 + \beta_7 X_8$$

Fig 5: showing the best 7 models for each case from M1 to M7

Of interesting note is that in searching for the best model (M1 through M7), the selected model M1 with one predictor variable as X5 ($Y_1 = \beta_0 + \beta_1 X_5$) is not surprising at all, since from our correlation plot (fig 1), X5 had the largest correlation with Y1.

To get the best model for each model size, we use residual sum of squares (RSS) and R-squared. As we would expect, with increase in model complexity, RSS would decrease (monotonically) while R-squared would increase monotonically. The RSS and R-squared for the 7 different models are given below:

Table 1.1

	M1	M2	M3	M4	M5	M6	M7
RSS	10900.904	7871.011	5280.595	4941.030	4857.444	4829.791	4829.150
R-squared	0.7956039	0.8524156	0.9009869	0.9073538	0.9089211	0.9094396	0.9094516

However, for selecting the model with the best overall performance, both RSS and R-squared are not suitable (as the model with increasing number of predictors will be better), since both are proxies for the training error which decreases with model complexity. Hence, in comparing models with different numbers of predictors, we use some surrogate of the test error. We compared Adj R^2, CP and BIC for our best model selection (Table 1.2).

Table 1.2

	M1	M2	M3	M4	M5	M6	M7
Adj R^2	0.7952310	0.8518760	0.9004428	0.9066739	0.9080840	0.9084389	0.9082822
Cp	675.206340	337.773498	49.574426	13.533645	6.169659	5.071758	7.000000
BIC	-860.6128	-1033.4155	-1246.6368	-1276.883	-1279.957	-1276.787	-1270.550

Based on Adj R^2 and Cp, the best model is the model with 6 predictors. Using BIC, the best model is the model with 5 predictors. For BIC and CP, smaller values indicate smaller error, while for adjusted R-squared, larger values indicate smaller errors. To present this result visually, we obtain the following plot.

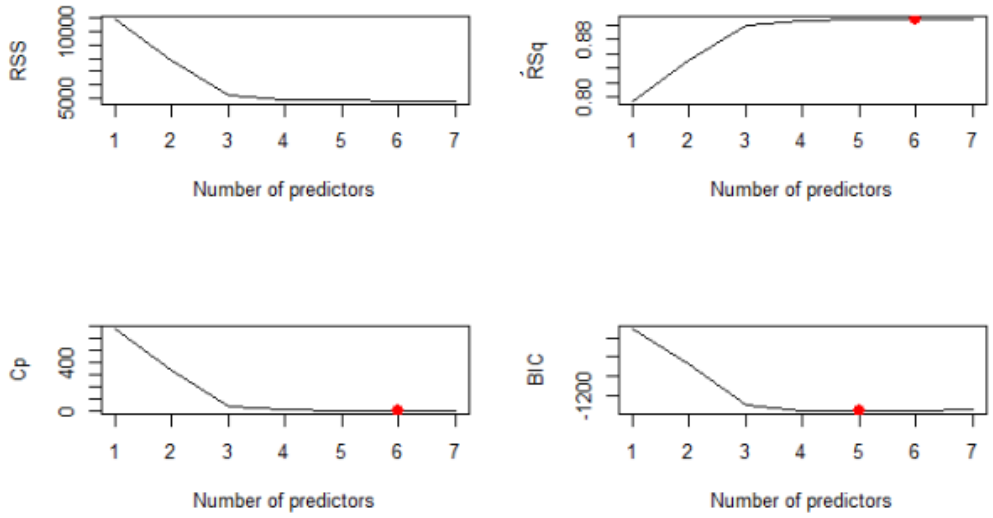


Figure 6: Model selection based on Adj R^2, Cp and BIC.

Hence, while we cannot know which model is best (without knowing the test error- since we only have the surrogates for the test error), we choose to go with the model with 6 predictors, ignoring the sparser model suggested by BIC which penalizes more for having more complexity.

Approximating the test error using the validation set approach (50-50 split), we also get that the best model is the model with 6 predictors (based on lowest test error). This thus gives us more confidence in our model of choice!

Table 1.3 model selection using validation set approach.

Model size	M1	M2	M3	M4	M5	**M6**	M7
Validation set error	22.19528	15.37245	10.37979	9.674985	9.482318	**9.413674**	9.435697

The coefficients of the best model are:

Table 1.4: coefficients of then best model ($R^2 = 0.9094$)

predictors	Intercept	X1	X2	X3	X5	**X7**	X8
Coefficient estimates	82.306	-65.659	-0.083	0.055	4.240	**23.880**	0.246

NB: All coefficients are significant at 95% confidence level.

Our best model thus includes X1, X2, X3, X5, X7 and X8 as our predictor variables.

And the equation for **Model 2** becomes:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8$$

$$Y_1 = 82.306 - 65.659 X_1 - 0.083 X_2 + 0.055 X_3 + 4.240 X_5 + 23.880 X_7 + 0.246 X_8$$

MODEL 3: LASSO REGRESSION

In Model 3, we fit a model containing all the predictors using a technique that shrinks the coefficient estimates towards zero. More so, with Lasso regression, the technique essentially forces some of the coefficients to be exactly zero, and we may expect to have a sparse model which will not include all the variables. Hence our model will do both shrinkage (improving prediction accuracy over OLS) and variable selection (better interpretability).

Next, we fit the lasso regression model and obtain a plot (fig 7) showing the relationship between the coefficient values and lambda (or log lambda). Lambda is simply the tuning parameter which penalizes model flexibility.

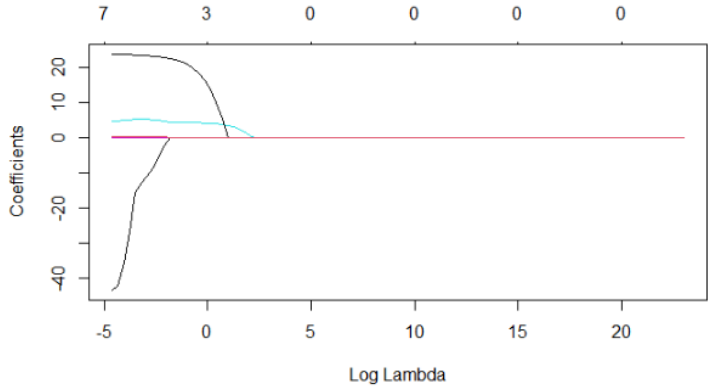


Fig 7: relationship between log lambda and coefficient values

From the above plot we immediately see that as lambda increases, more and more coefficients get shrunk towards zero (each colored line corresponding to a predictor in our dataset). For example, with a log lambda value of 5, all the coefficients equal zero. Therefore, with different (increasing) values of lambda we are reducing the variance and, in a way, making the model less and less flexible while also improving the interpretability of the model.

Hence, using cross validation, we then choose the optimal value of lambda, which is the lambda that minimizes the estimated test error. The best lambda is then chosen to be 0.006831362 or Log (λ) = -2.165.

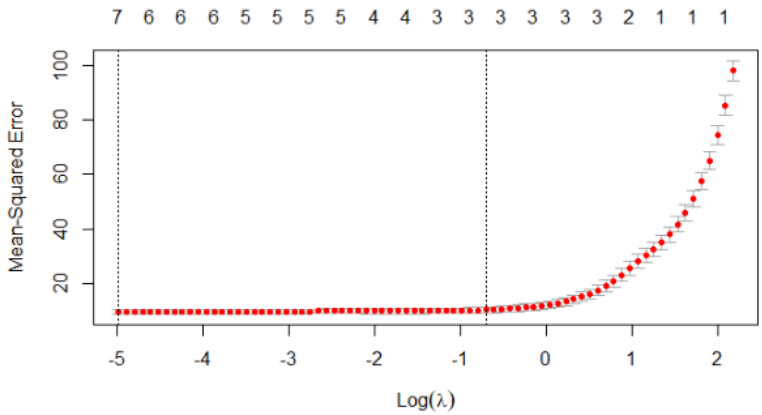


Fig 7: cross validation error as a function of lambda

We therefore output the coefficients associated with the best λ for our model.

```
  (Intercept)            X1            X2            X3            X4            X5
 4.145755e+01 -4.328535e+01 -4.116049e-06  0.000000e+00 -9.391653e-02  4.718440e+00
           X6            X7            X8
-2.220369e-02  2.375288e+01  2.409859e-01
```

Fig 8a: resulting coefficient estimates.

The coefficient of X3 was shrunk all the way to exactly zero. This means it was totally dropped from the model because it was not influential enough. Hence lasso in this case returns a sparse model by shrinking coefficients to zero (based on the optimal lambda value of 0.006831362).

Reporting the coefficient estimates without X3 we have,

```
  (Intercept)            X1            X2            X4            X5            X6
 4.145755e+01 -4.328535e+01 -4.116049e-06 -9.391653e-02  4.718440e+00 -2.220369e-02
           X7            X8
 2.375288e+01  2.409859e-01
```

Fig 8b: resulting coefficient estimates without X3.

Of interesting note is that while the coefficient value for X2 Is not exactly zero, it is indeed very small (almost zero).

And the equation for **Model 3** becomes:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_6 + \beta_6 X_7 + \beta_7 X_8$$

$$Y_1 = 0.415 - 0.433 X_1 - 0.000004116 X_2 - 0.09392 X_4 + 4.718 X_5 - 0.0222 X_6 + 0.2375 X_7 + 0.241 X_8$$

$$R^2 = 0.9092$$

Comparison of all three Models

Model1:

$$\log Y_1 = -0.951 + 1.495X_1 + 0.00295X_3 + -0.00023X_6 + 0.781X_7 + 0.0286X_8 - 0.1295X_7 : X_8$$

Model2:

$$Y_1 = 82.306 - 65.659X_1 - 0.083X_2 + 0.055X_3 + 4.240X_5 + 23.880X_7 + 0.246X_8$$

Model3:

$$Y_1 = 0.415 - 0.433X_1 - 0.000004116X_2 - 0.09392X_4 + 4.718X_5 - 0.0222X_6 + 0.2375X_7 + 0.241X_8$$

Comparing all three models:

Generally, OLS fitting (model 1) does not do very well when the relationship between Y and X is not linear. Hence, in this project we sought to replace the OLS fitting with some alternative fitting procedure: Best subsets selection and Lasso regression. Our primary goal was to improve the prediction accuracy.

Best subsets selection (model 2) approach searches all the possible model (subsets) and then we run a linear regression for each possible combination of the predictors. In this case, the number of possible models was 2^p (or 256 models), which is not too large to make the approach impractical (or highly computationally expensive).

In comparison to Lasso (model 3), both approaches return sparse models with best subset leaving out 2 predictors and Lasso leaving out 1 predictor. In determining the better of the two, we used cross validation (validation set approach at 50-50 split), and found the following results:

Table 1.5

Model	Validation set error
Model 2	9.413674
Model 3	9.893168

Hence based on lower test error, our choice is Best Subset selection model (model 2)

Interpreting the best model. We have that:

The average value of Y1 when all inputs are zero is 82.306.
1 unit increase in X1 causes Y1 to decrease on average by 65.659 units, other predictors held constant.
1 unit increase in X2 causes Y1 to decrease on average by 0.083 units, all other predictors held constant.
1 unit increase in X3 causes Y1 to increase on average by 0.055 units, all other predictors held constant.
1 unit increase in X5 causes Y1 to increase on average by 4.240 units, all other predictors held constant.
1 unit increase in X7 causes Y1 to increase on average by 23.880 units, all other predictors held constant.
1 unit increase in X8 causes Y1 to increase on average by 0.256 units, all other predictors held constant.