NC STATE UNIVERSITY

ST 516

Experimental Statistics for Engineers II

Midterm 1 Project

Written By Team 16 -

Ajinkya Salve

Ansab Jan

John McDonald

Pratyush Prabhat

Instructor:

Dr. Dan Harris

Submitted On:

02/23/2020

Introduction

The construction firm wants to analyze and explore the interaction of behaviors/parameters which impact the compressive strength of a structure. An exhaustive dataset was provided which outlined that 8 predictors had a direct impact on the extrapolation process of the strength. Overall, the firm's senior management team intends to understand and establish the relationship amongst all the predictors and accurately predict the resultant compressive strength given the concrete recipes available. To resolve this issue, our team adopted the approach of applying the dataset with multiple regression models to precisely discern the best fit model in this case. The models explored in this study include PCR, Lasso regression, Ridge Regression and Subset selection. A careful and thorough analysis was performed in order to select a model that best-balanced prediction power and interpretability of strength and concrete recipes respectively.

Executive Summary

The objective of this study is to establish a model capable of accurate prediction of strength of a mixture of concrete, based solely on information about the quantities of constituent elements that produce the concrete mixture. We use Subset Selection, Ridge Regression, the Lasso Method and Principal Components Regression (PCR) to find the most satisfactory model. The results from the analysis detailed in this report point to PCR being the best fit among the models tested, explaining most of the variance in the system, with an R2 value of 0.87, and resulting in the lowest MSE value of 42.66. The model thus fits the primary goal of the undertaking, producing accurate results for the compressive strength of the concrete.

Data

The data was collected from previous studies on concrete strength. Our objective is to predict the compressive strength of the concrete by using the components which are going into the concrete mixture. These components are cement, slag, fly ash, water, super plasticizer, coarse aggregate, fine aggregate, and age. In summary, our dataset has one response variable and 8 predictor variables. We have 1030 observations for these 9 variables. After preliminary analysis of the data we understood that there was no missing data and we didn't need to fill in the missing values. The following is a scatter plot of the grid of predictors against response variable i.e., compressive strength.

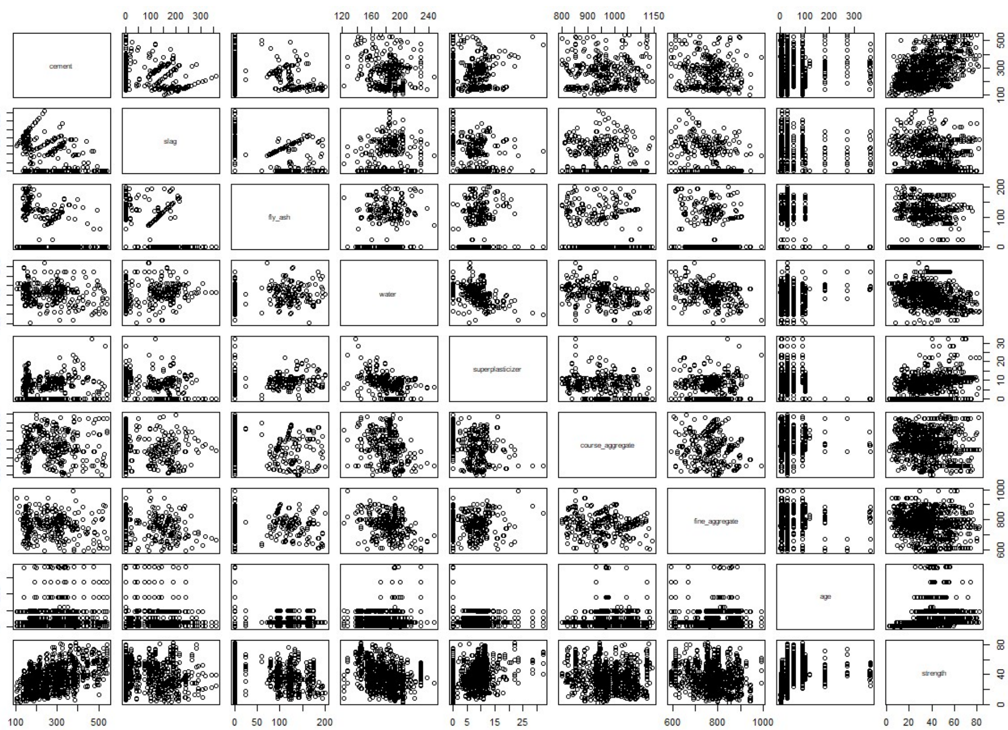

Figure 1- Scatterplot of Grid predictors against response variables

As we fitted our simple linear model to this dataset with 8 predictor variables, we found out that around 61% of the variation in the compressive strength can be explained by these predictor variables. Since we have few predictor variables and large number of observations, we also included square of predictors and their second order interactions to see if we get better relation. After this model, we found that around 81% of the variation in compressive strength can be explained by these additional predictors. Hence, we have decided to include squares of predictors and their second order interactions in our models to get better prediction of the response variable.

Methods

Ridge Regression Model: To explore this regression model we split the given 'Concrete' dataset into the train and test datasets in a 80% - 20% proportion. Post this activity we applied and examined the first order and second order regression models. The MSE for this model is **122.89** and **57.098 for** 8 and 44 predictors respectively. The crucial plots of lambda vs MSE values and the actual vs fitted values for the second order have been illustrated as below in Figure 2.

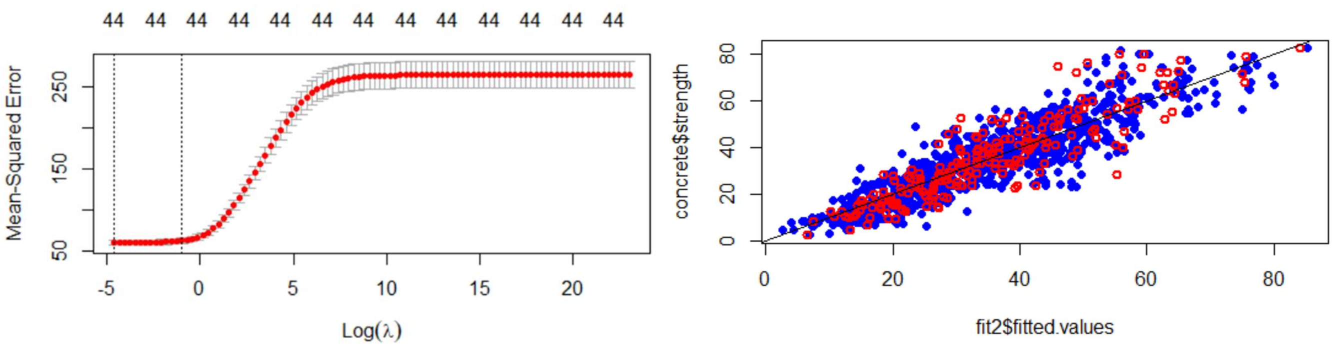

Figure 2- Ridge Regression Model

Lasso Regression Model: Lasso Regression Regularization was applied to the second order model of the dataset to shrink the coefficient estimates. The MSE for this model was noted to be **79.77**. The method was applied to an array of values for λ, ranging from 100 to .001. Cross validation was then used to obtain the optimal value for λ. The method discarded a significant portion of the predictors to give a parsimonious model (27 significant predictors).

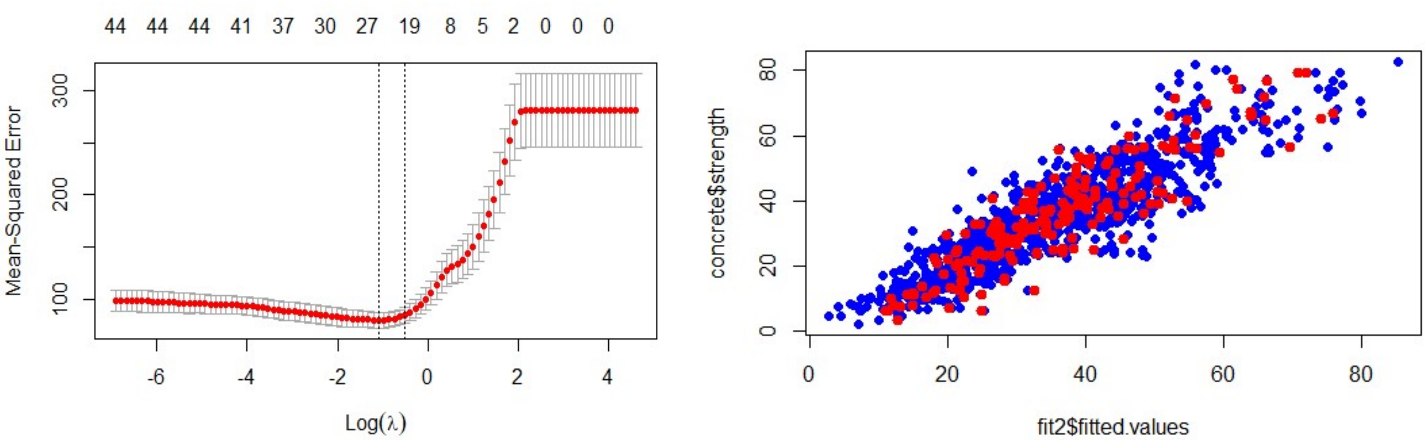

Figure 3 - Lasso Regression Model

Subset Selection: In this model we use all the 44 predictors to find out how many predictors are actually significant out of these 44. As you can see from the following graph, we get 37 significant predictors out of possible 44. For our training dataset which is 80% of our data, we get an MSE of 60.15.

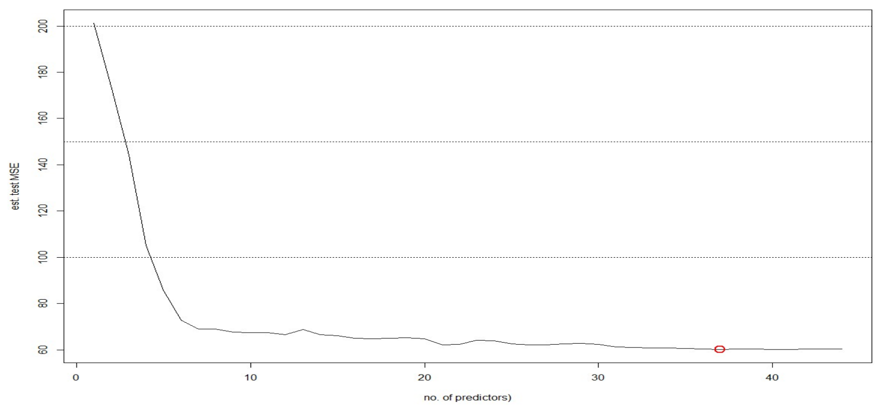

Figure 4- Subset Selection

Pricipal Component Regression: One model solution we chose to pursue is PCR. This form of regression is especially effective at finding latent variables and improving accuracy for predictions. Hence, we chose this method to satisfy the customer's priority goal of accurate predictions for compressive strength of concrete. The graphs below summarize our approach via this method, and its results.

Adding these interactions to the dataset left us with a total of 44 predictors. At which point we generated the principal components for each:

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.4405 2.3974 2.02554 1.81597 1.67997 1.49610 1.38164 1.34015 1.21166 1.18946 1.15823
Proportion of Variance 0.1354 0.1306 0.09325 0.07495 0.06414 0.05087 0.04338 0.04082 0.03337 0.03215 0.03049
Cumulative Proportion  0.1354 0.2660 0.35924 0.43419 0.49833 0.54920 0.59258 0.63340 0.66677 0.69892 0.72941
                          PC12    PC13    PC14    PC15    PC16   PC17   PC18    PC19    PC20    PC21    PC22
Standard deviation     1.09897 1.06622 1.01872 0.97206 0.91650 0.8776 0.8469 0.83262 0.80022 0.74584 0.71395
Proportion of Variance 0.02745 0.02584 0.02359 0.02148 0.01909 0.0175 0.0163 0.01576 0.01455 0.01264 0.01158
Cumulative Proportion  0.75686 0.78270 0.80628 0.82776 0.84685 0.8643 0.8807 0.89641 0.91096 0.92360 0.93519
                          PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30    PC31    PC32
Standard deviation     0.68184 0.61052 0.56731 0.55718 0.50772 0.44863 0.43307 0.38819 0.37465 0.31583
Proportion of Variance 0.01057 0.00847 0.00731 0.00706 0.00586 0.00457 0.00426 0.00342 0.00319 0.00227
Cumulative Proportion  0.94575 0.95423 0.96154 0.96860 0.97445 0.97903 0.98329 0.98672 0.98991 0.99217
                          PC33    PC34    PC35    PC36    PC37    PC38    PC39    PC40    PC41    PC42   PC43
Standard deviation     0.26850 0.25448 0.21097 0.19545 0.17210 0.16098 0.14157 0.13637 0.12702 0.09695 0.0666
Proportion of Variance 0.00164 0.00147 0.00101 0.00087 0.00067 0.00059 0.00046 0.00042 0.00037 0.00021 0.0001
Cumulative Proportion  0.99381 0.99528 0.99629 0.99716 0.99784 0.99842 0.99888 0.99930 0.99967 0.99988 1.0000
                          PC44
Standard deviation     0.02660
Proportion of Variance 0.00002
Cumulative Proportion  1.00000
```

Figure 5 - Principal Components

We note that the cumulative proportion of our 44 predictors is valued at 100%. This is a clear indicator of overfitting. By reducing the PC count, the R-squared value can be marginally reduced while the F-statistic is substantially reduced, meaning our model is less likely to be overfitting. However, to select the optimal PC count, cross validation with MSE test is performed.

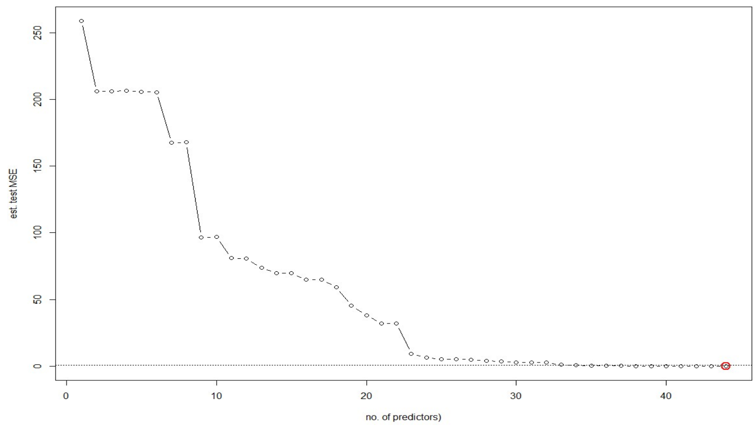

Figure 6 - Finding number of predictors using cross validation

Using all 44 predictors produces the lowest test MSE. We use 44 number of predictors.

```
Residual standard error: 7.307 on 179 degrees of freedom
Multiple R-squared:  0.8717,    Adjusted R-squared:  0.8402
F-statistic: 27.65 on 44 and 179 DF,  p-value: < 2.2e-16
```

Figure 7 - R Squared Value using PCR Model

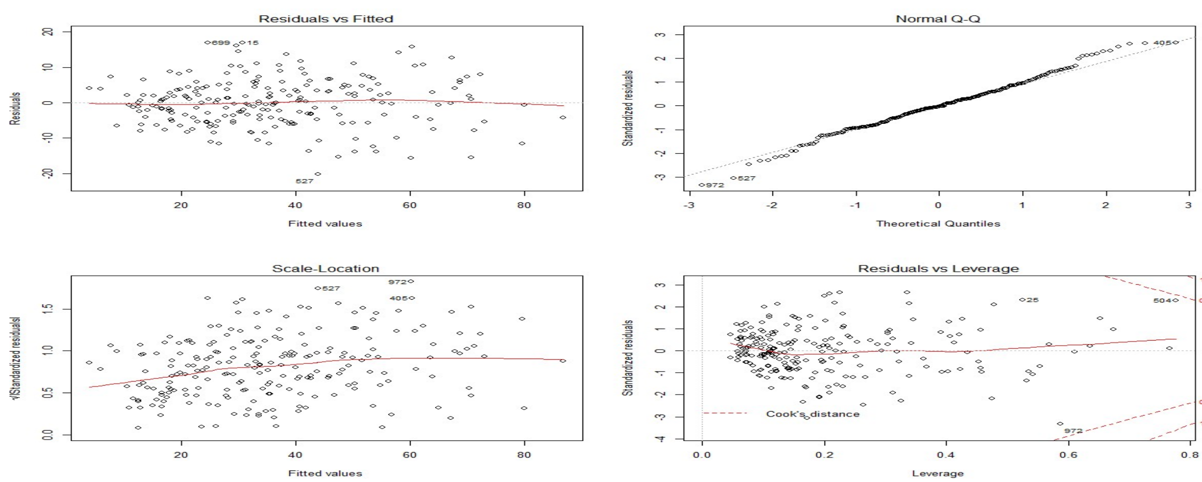

Figure 8 - Residual Diagnostics Plot

For this model we an MSE of 42.66 and an 87% of variation in data is explained by these 44 predictors.

Result:

Prediction Performance and Best Model selection:

Following table gives us the summary of the model fitted to the given data and the their performance as compared to each other.

It can be concluded that Principal Component Regression is the best model for prediction of as it gives the lowest MSE value and highest R-Squared value explaining the variation in response variable i.e., compressive strength. We preferring to use second order model. Since we have 8 predictors variables and 1030 observations of them. We get very high value of MSE and very low value of R-Squared. As we run the second order model by including predictors squared and their interactions as well with each other, we get improved values of MSE as well as R-Squared.

Model Fitted	Predictors	MSE Values	R-Squared Values
Subset Selection	8	120.29	0.63
	44 (squares & interactions)	43	0.87
Ridge Regression	8	122.89	0.6155
	44 (squares & interactions)	57.09	0.81
Lasso Regression	8	115.4	0.61
	44 (squares & interactions)	79.77	0.80
Principal Component Regression	8	120	0.63
	44 (squares & interactions)	42.66	0.87

Statistical Inference and Influential Variables:

After performing cross validation on our data, we find that the significant predictors for this model are all 44 predictors which square of predictors as well as their interactions with other predictors.

```
                      Estimate
(Intercept)      -6.576310e+01
cement            1.369438e-01 c_a_sq             1.332839e-04
slag              1.355981e-01 s_f_sq             1.553829e-03
fly_ash           1.034796e-01 s_w_sq             1.370141e-03
water            -5.651150e-02 s_su_sq           -6.696520e-03
superplasticizer  7.074696e-01 s_c_sq             4.963637e-04
course_aggregate  1.903501e-02 s_fi_sq            4.543584e-04
fine_aggregate    3.674426e-02 s_a_sq             3.501012e-04
age               3.288646e-01 f_w_sq            -3.480014e-03
cement_sq        -2.383738e-04 f_su_sq           -1.674366e-02
slag_sq           6.712197e-04 f_c_sq             1.935089e-04
fly_sq            1.284400e-04 f_fi_sq           -2.005614e-04
water_sq          4.197060e-03 f_a_sq             1.068165e-03
super_sq          6.025990e-03 w_su_sq            1.748461e-02
course_sq        -5.927254e-05 w_c_sq            -8.639674e-04
fine_sq          -7.052083e-04 w_fi_sq           -2.572790e-03
age_sq           -6.691343e-04 w_a_sq             1.505998e-03
c_s_sq            5.102871e-04 su_c_sq           -5.825765e-03
c_f_sq            6.605646e-06 su_f_sq           -8.262732e-03
c_w_sq           -3.128276e-03 su_a_sq            7.258083e-03
c_su_sq          -8.191101e-03 co_f_sq           -4.241777e-04
c_c_sq           -2.589486e-04 co_a_sq            1.519391e-04
c_fi_sq          -8.320499e-04 fi_a_sq            4.898597e-04
```

Figure 9- Coefficients

```
Residual standard error: 7.307 on 179 degrees of freedom
Multiple R-squared:  0.8717,    Adjusted R-squared:  0.8402
F-statistic: 27.65 on 44 and 179 DF,  p-value: < 2.2e-16
```

Figure 10 - R squared value

Also, we get high values of VIF indicating the selected model was the best model for this dataset.

```
         cement            slag         fly_ash           water superplasticizer course_aggregate
      18.510665       20.074697       17.459769       17.199023        14.400362        13.312440
 fine_aggregate             age       cement_sq         slag_sq           fly_sq         water_sq
      16.037957        9.886733       80.478059       69.016430        25.447159        80.370899
       super_sq       course_sq         fine_sq          age_sq           c_s_sq           c_f_sq
      36.767075       50.923871      128.836725       18.724529       153.589277       124.689475
         c_w_sq         c_su_sq          c_c_sq         c_fi_sq           c_a_sq           s_f_sq
     168.069646      146.808099      223.330890      290.019955        15.731325       144.636779
         s_w_sq         s_su_sq          s_c_sq         s_fi_sq           s_a_sq           f_w_sq
     132.829738       81.607769      110.087228      201.393124        18.699785        91.859182
        f_su_sq          f_c_sq         f_fi_sq          f_a_sq          w_su_sq           w_c_sq
      40.765590      121.875599      140.472965       24.203634       138.741097       125.226037
        w_fi_sq          w_a_sq         su_c_sq         su_f_sq          su_a_sq          co_f_sq
     298.580181       61.647169       88.079914      143.062692        21.469837       197.004448
        co_a_sq         fi_a_sq
       5.179896       39.688414
```

Figure 11 - VIF Analysis

In the below graph we compare our PCR results to ordinary least squares. With red representing our new PCR adjusted model, and blue representing ordinary least squares.

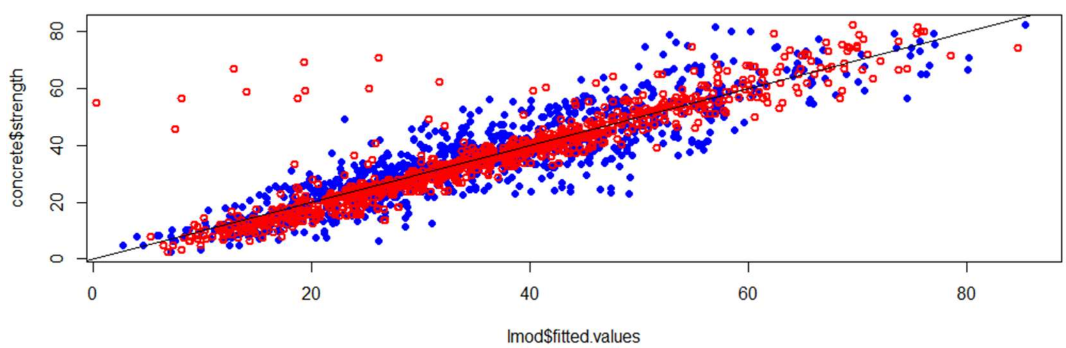

Figure 12 - PCR Adjusted vs Fitted Values

We see a significant improvement in predictions, with some increased variance at lower fitted values. Overall, our PCR approach was effective in improving prediction accuracy, which is our customer's primary goal.

<u>Conclusion</u>

In this report, we analyzed the components going into the concrete and their effect on the compressive strength of the concrete. We divided the data into training and test dataset and fitted the model on the train dataset. After that we calculated the MSE and R-squared values to determine the best model for predicting the compressive strength of the concrete and also determined the major factors affecting the strength of the concrete.

The Principal Component Regression was determined to be the best model for prediction of the compressive strength after model was fitted to the 44 predictors which included 8 original predictors as well as their squares and their interactions with each other. We got the lowest value of MSE as well as highest value of R-squared for this model.

In our model, we had to include squares and interactions of the predictors as there were only 8 predictors and the number of observations were 1030. Also, in the future we recommend using decision trees and compare their performance with model implemented int this report.