INTRODUCTION:

SUMMARY: High-performance concrete (HPC) is a new terminology used in the concrete construction industry. In addition to the three basic ingredients in conventional concrete, i.e., Portland cement, fine and coarse aggregates, and water, the making of HPC needs to incorporate supplementary cementitious materials, such as fly ash and blast furnace slag, and chemical admixture, such as superplasticizer. Compressive strength is an important discussion when we plan on construction. The relationship of the ingredients and the strength is estimated with the help of regression analysis.

DATA INTEGRATION: The dataset was collected from UCI machine learning repository website. Test data from 17 different sources was used to check the reliability of the strength model. In about 1031 concrete samples were evaluated. The regressors in this problem are:

1. Cement (kg/m3) –X1
2. Fly ash (kg/m3) – X2
3. Blast furnace slag (kg/m3) –X3
4. Water (kg/m3) –X4
5. Superplasticizer (kg/m3) –X5
6. Coarse aggregate (kg/m3) –X6
7. Fine aggregate (kg/m3) –X7
8. Age of testing (days) -X8

The data set comprises of 1031 observations and 8 regressors mentioned above to estimate the response variable “Concrete compressive strength” (Y) (MPa).

PURPOSE: The purpose of this project is aimed at providing significant parameters influencing the response. The analysis includes fitting an initial model followed by checking on **multicollinearity, suitable data transformation, significance test of the model, variable selection and model validation.** Final model obtained provides good fit with good predicting capability of future observations.

DATA ANALYSIS:

I. ORIGINAL MODEL: Regression analysis was performed on a full model with all eight regressors and the results obtained from R and Minitab are displayed.

```
Coefficients:
             Estimate     Std. Error t value  Pr(>|t|)
(Intercept) -23.331214   26.585504  -0.878   0.380372
X1            0.119804   0.008489   14.113   < 2e-16 ***
X2            0.103866   0.010136   10.247    < 2e-16 ***
X3            0.087934    0.012583   6.988   5.02e-12 ***
X4           -0.149918   0.040177  -3.731   0.000201 ***
X5            0.292225   0.093424   3.128   0.001810 **
X6            0.018086   0.009392   1.926   0.054425 .
X7            0.020190   0.010702   1.887   0.059491 .
X8            0.114222   0.005427  21.046   < 2e-16 ***
 ---
 Residual standard error: 10.4 on 1021 degrees of freedom
 Multiple R-squared:  0.6155,    Adjusted R-squared:  0.6125
 F-statistic: 204.3 on 8 and 1021 DF, p-value: < 2.2e-16
 PRESS: 112899
 print(vif(lm.concrete))
      X1       X2       X3       X4       X5       X6       X7       X8
7.488944 7.276963 6.170634 7.003957 2.963776 5.074617 7.005081 1.11836
```

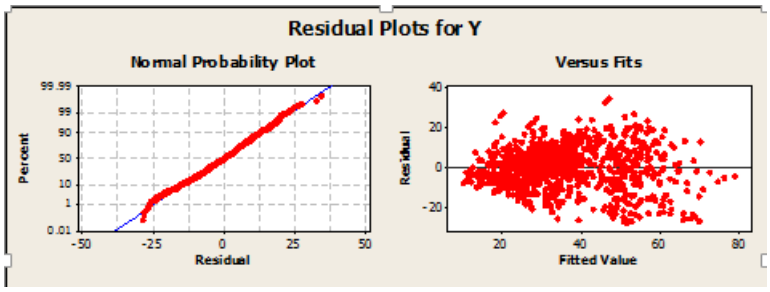

Figure 1: Residual plot for model 1

- CONCLUSION: From the inspection of residual plots, we can see that there is no problem with normal probability plot as most of the residuals fall along the straight line. An **outward opening funnel** is recognized in residual vs. fitted value plot revealing **non-constant variance**. Suitable data transformation should be applied in this case. R squared and Adjusted R squared values are 61.6% and 61.3% respectively which shows the model does not effectively explain the variability in the response. From the table, **predictors X6(Coarse Aggregate) and X7(Fine Aggregate)** are not significant according to their

high "p" value in this analysis. Also we can observe there is moderate multicollinearity present in the model with **maximum VIF of 7.489**.

II. MODEL 2

- **Mutlicollinearity and Data transformation**: **Model re-specification** is done based on the subject knowledge to reduce the effect of multicollinearity. A new variable WC which is defined as Water to cement ratio is taken into account. The generally accepted Abrams rule is a formulation of the observation that an increase in the w/c decreases the concrete strength, whereas a decrease in the w/c ratio increases the strength. Scatter plots of response vs. all regressors was plotted to check for the linear relationship between them. Variable X8(Age) shows a non-linear pattern with the response variable while there was no problem with rest of the variables. Also, no significant curvature was observed, hence there is no need for quadratic terms. **Suitable data transformations applied on the response variable and X8 are Y=√Y and X8=log(X8) respectively.**

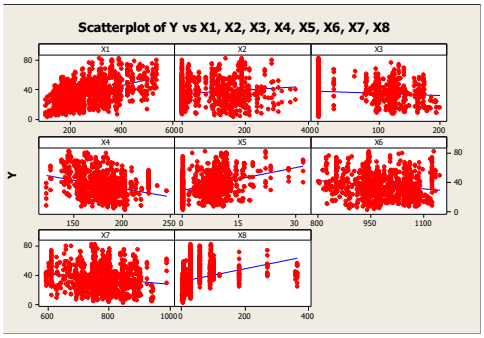

Figure 2: Scatter plot of response vs. all regressors

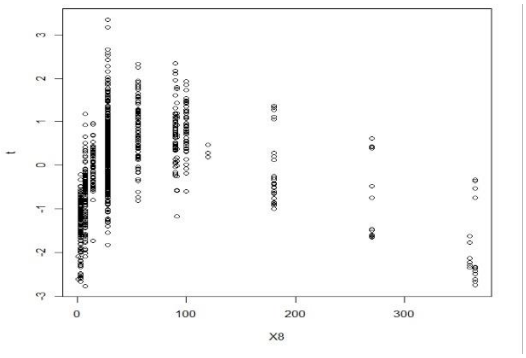

Figure 4: Plot of r-student vs. X8

- New model with transformed data and model respecification is fit and the regression analysis for the 2nd MODEL is as follows.

```
Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)  4.752e+00  4.472e-01  10.626  < 2e-16 ***
WC          -3.141e+00  1.015e-01 -30.931  < 2e-16 ***
X2           7.162e-03  3.976e-04  18.015  < 2e-16 ***
X3           3.997e-03  5.059e-04   7.901 7.14e-15 ***
X5           3.535e-02  5.035e-03   7.021 4.02e-12 ***
X6          -2.622e-05  3.022e-04  -0.087    0.931
X7           1.216e-04  3.275e-04   0.371    0.710
log.X8.      1.725e+00  4.100e-02  42.082  < 2e-16 ***

Multiple R-squared:  0.7885,    Adjusted R-squared:  0.787
```

```
PRESS: 465.728
> print(vif(lm.concrete))
      WC       X2       X3       X5       X6       X7   log.X8.
2.341280 2.709150 2.413293 2.083373 1.271102 1.587460 1.036343
```

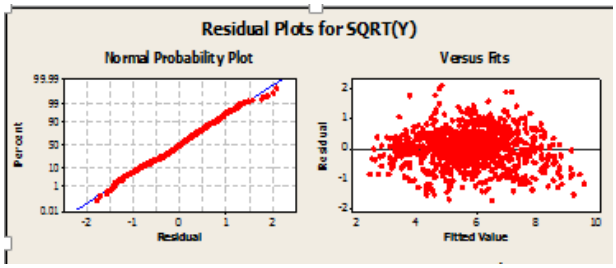

Figure 3: Residual plot of model 2

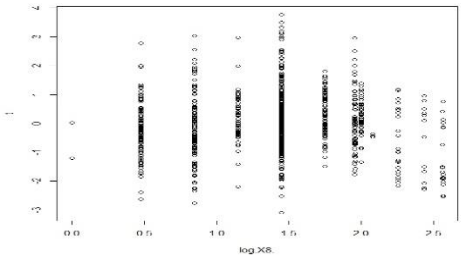

Figure 6: Plot of r-student vs. log X8

- CONCLUSION: **R-squared and Adjusted R-squared** are now 78.85 and 78.7 respectively and the **PRESS** statistic has reduced considerably to 465.728 which shows a good improvement in the model. The plot of Residuals vs. fitted value is much better than the previous model **overcoming the non-constant variance problem.** Multicollinearity problem is also solved as VIF's are less than 5 in this model. The plot of residuals vs. Log(X8) shows a better variance than the initial plot. Plot of residuals vs. rest of the variables were also checked and there was no problem detected.
- RECOMMENDATION: The current model is adequate as it showed improvement in the areas of constant variance and also as **a better predictor than the previous model.** Diagnosis for Leverage and Influential points is carried out in next step.

III. DIAGNOSIS FOR LEVERAGE AND INFLUENTIAL POINTS:

- DESCRIPTION: This analysis consists of identifying the outliers, leverage and influential points. An influential point can be detected by a large cook's D value (D >=1) or $|DFFITS| > 2\sqrt{(p/n)}$. Minitab output showed the observations with large standardized residuals, large DFFIT and high leverage points. The question arises whether to discard these points or to keep them in the model. We can't simply remove these points as they may be valid observations. We may take the readings at the same regressors again in final predicted model and check if the readings are close.

IV. MODEL 3: VARIABLE SELECTION AND MODEL BUILDING:

The next step is to reduce the number of variables in the model by implementing **all possible regressions in Minitab.**

Best Subsets Regression: SQRT(Y) versus X2, X3, X5, X6, X7, log(X8), WC

Response is SQRT(Y)

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	X2	X3	X5	X6	X7	log X8)	WC
1	33.9	33.9	2167.2	1.1782						X	
1	25.2	25.1	2589.2	1.2536							X
2	60.6	60.5	880.8	0.91041						X	X
2	48.6	48.5	1458.0	1.0392		X				X	
3	72.0	71.9	332.2	0.76800	X					X	X
3	68.2	68.1	515.5	0.81834		X				X	X
4	77.5	77.4	68.3	0.68883	X	X				X	X
4	77.4	77.3	73.0	0.69030	X		X			X	X
5	78.8	78.7	4.2	0.66790	X	X	X			X	X
5	77.8	77.7	53.6	0.68384	X	X			X	X	X
6	78.8	78.7	6.0	0.66817	X	X	X		X	X	X
6	78.8	78.7	6.1	0.66821	X	X	X	X		X	X
7	78.8	78.7	8.0	0.66849	X	X	X	X	X	X	X

- CONCLUSION: The highlighted rows shows three best subset models that can be considered. Since the Adjusted R-square seems to be constant for all the three subsets, subset model with **low Mallows Cp 4.2** and **low standard deviation of 0.6679** is being considered as the **final model with five regressors namely X2(Fly ash), X3(Blast furnace slag), X5(super-plasticizer), logX8 (age) and WC (water/cement).**This final model is subjected to thorough analysis and its details using R is shown below.

```
Call:
lm(formula = SQRT.Y. ~ WC + X2 + X3 + X5 + log.X8.)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.8135652  0.0857045  56.165  < 2e-16 ***
WC          -3.1243689  0.0906289 -34.474  < 2e-16 ***
X2           0.0070949  0.0003220  22.032  < 2e-16 ***
X3           0.0039250  0.0004661   8.421  < 2e-16 ***
X5           0.0363245  0.0044627   8.140 1.14e-15 ***
log.X8.      1.7232765  0.0403013  42.760  < 2e-16 ***

Residual standard error: 0.6679 on 1024 degrees of freedom
Multiple R-squared:  0.7885,     Adjusted R-squared:  0.7874
F-statistic: 763.3 on 5 and 1024 DF,  p-value: < 2.2e-16
PRESS: 463.412

> print(vif(lm.concrete))
      WC       X2       X3       X5  log.X8.
1.868116 1.780699 2.052189 1.639420 1.003124
```

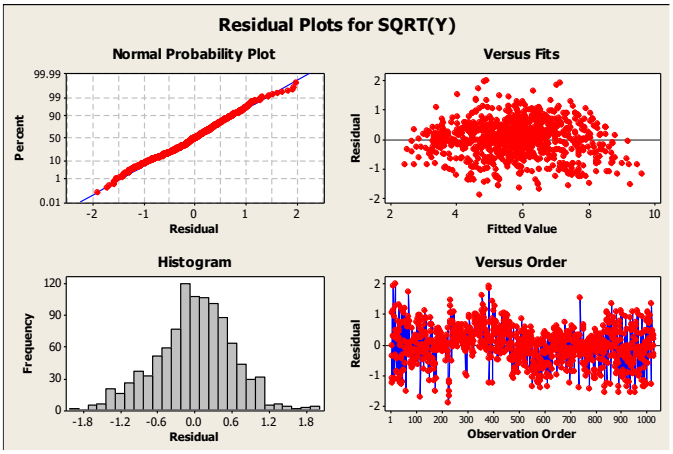

Figure 4: Residual plots for model 3

- CONCLUSION: The final model is a better predictor than the previous model as its **PRESS statistic** is reduced to 463.412 which is a good sign. All the variables in the reduced model are significant and have a **better coefficient estimate** than the previous models. **Multicollinearity is further reduced** making the least square analysis effective for the final model. There is no problem with the plot of residuals vs. fitted value. Thus we can select this as our **final model with significant variables**.

V. MODEL VALIDATION:

- DESCRIPTION: Regression models are used extensively for prediction or estimation, data description, parameter estimation and control. Frequently the user of the regression model is a different individual from the model developer. Therefore it is important to validate this model before it is released for the user. The model validation technique involves checking the regressor coefficients and the predicting capacity of the model.
- ANALYZING THE MODEL COEFFICIENTS: Studying the coefficients in the final regression model is a way to determine if they are **stable** and if their **signs and magnitudes** are reasonable. **Coefficient of WC (water to cement ratio) is negative and valid**. Because addition of water above certain limit decreases the strength. As the amount of water increases, its porosity increases and thereby decreases the strength. This confirms to Abrahams water to cement ratio pronouncement that as WC increases, strength decreases and vice versa. **The coefficient of log X8 is positive and valid**. As the age increases, the strength increases and after a certain period of time the strength

remains constant. **Coefficient of X5(super-plasticizer) is positive and valid**. Super-plasticizer helps in reduction of W/C ratio by reducing the amount of water by 12% to 30%.**Coefficient of X2(Blast furnace slag) is positive and is valid**. Increasing CaO content of the slag results in raised slag **basicity,** thereby increasing the compressive strength. **Coefficient of X3(Fly ash) is positive and valid**. Properly cured concrete made with fly ash creates a denser product because the size of the pores are reduced. This increases the strength and reduces permeability of water. **The amount of coarse and fine aggregate** will add little strength to the model which is what the model estimated by displaying them as non-significant.

- PREDICTION PERFORMANCE: **The data splitting technique** was implemented for validation of the model. 700 observations out of 1031 were used for initial analysis for model estimates and the rest of the data is used for prediction performance. A random sample of 50 observations were taken from test run data and the table below shows few of their results. The closeness between them was significant even though there were few exceptions where predicted data weren't close to original data. It as shown below:

ORIGINAL DATA	PREDICTED DATA
6.8578	6.3052
6.0373	6.1006
7.4926	7.4065
8.2643	7.2749
2.8635	3.2814
3.8845	5.0998
1.5264	0.9194
8.9437	6.4538
6.4070	8.0880
9.0415	7.0484

VI. CONCLUSIONS AND RECOMMENDATIONS:

- The data was initially fit with simple linear equation, but the data showed abnormality. Multicollinearity in the data was eliminated through model re-specification. Variance Inflation Factor confirmed the absence of multicollinearity. The data was then transformed using suitable techniques and analysis of transformed data yielded R-squared of 78.85% and Adjusted R-squared of 78.5% which is a good result. Insignificant variables in the model was eliminated using all possible regressions technique. The sign and magnitude of variable coefficients validated the final model. The predicting power of the model is also good as the original data set values matched the test run values with a few exceptions. Thus this model can be recommended to the user.
- The above analysis limits the value of Adjusted R squared to 78.5 even after removing insignificant variables. Future analysis will be carried out by performing **Principal Component Regression** to estimate if total variance explained by the model will be better than this analysis.

VII. REFERENCES

1. http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
2. I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998).
3. Montgomery, D. C., Peck Elizabeth, A., & Geoffrey Vining, G. Introduction to Linear Regression Analysis, Fifth Edition.
4. http://www.engr.psu.edu/ce/courses/ce584/concrete/library/materials/Admixture/AdmixturesMain.htm
5. https://en.wikipedia.org/wiki/Ground_granulated_blast-furnace_slag