IEE 578 | REGRESSION ANALYSIS

REGRESSION ANALYSIS FOR PREDICTION OF CONCRETE COMPRESSIVE STRENGTH

INSTRUCTOR: DR. DOUGLAS C MONTGOMERY

December 04, 2019
Arizona State University

- Prakash Sudhakar
1217272901

TABLE OF CONTENTS

DATA PREPARATION:

Background:

Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.

We will start by using all possible regression (best subsets), stepwise regression and then conduct a series of multiple regression analysis that will eventually narrow down the best model predictor of compressive concrete strength to only 6 variables – Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer and Age.

Data Overview:

The dataset consists of 1030 observations and 9 attributes. A data dictionary is also supplied below for reference. The actual concrete compressive strength (MPa) for a given mixture under a specific age was determined from the laboratory. Data is in raw form and is not scaled.

The following dataset has been taken from the UCI repository: https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength

Predictor:

X_1 – Cement (kg/m^3) X_2 – Blast Furnace Slag (kg/m^3) X_3 – Fly Ash (kg/m^3) X_4 – Water (kg/m^3) X_5 – Superplasticizer (kg/m^3) X_6 – Age (days) X_7 – Coarse Aggregate (kg/m^3) X_8 – Fine Aggregate (kg/m^3)

Response: *Y – Concrete Compressive Strength (MPa)*

Sample Data Set Subset:

A sample subset of 75 out of 1030 observations from the data set is shown below for reference:

Cement	***Blast Furnace Slag***	***Fly Ash***	***Water***	***Superplasticizer***	***Coarse Aggregate***	***Fine Aggregate***	***Age***	***Concrete compressive strength***

173.81	93.37	159.9	172.34	9.73	1007.2	746.6	28	37.81086384
190.34	0	125.18	166.61	9.88	1079	798.9	28	24.84871504
250	0	95.69	191.84	5.33	948.9	857.2	28	27.22051248
213.5	0	174.24	159.21	11.66	1043.6	771.9	28	44.63667624
194.68	0	100.52	170.17	7.48	998	901.8	28	37.2661778
251.37	0	118.27	192.94	5.75	1043.6	754.3	28	33.27411176
165	0.02	143.57	163.81	0	1005.6	900.9	56	36.56291228
165	128.5	132.1	175.06	8.08	1005.8	746.6	56	53.72396992
178.03	129.8	118.6	179.94	3.57	1007.3	746.8	56	48.58737372
167.35	129.9	128.62	175.46	7.79	1006.3	746.6	56	51.72448952
172.38	13.61	172.37	156.76	4.14	1006.3	856.4	56	35.852752
173.54	50.05	173.53	164.77	6.47	1006.2	793.5	56	53.77223324
167	75.4	167	164.03	7.91	1007.3	770.1	56	53.46196904
173.81	93.37	159.9	172.34	9.73	1007.2	746.6	56	48.9872698
190.34	0	125.18	166.61	9.88	1079	798.9	56	31.715896
250	0	95.69	191.84	5.33	948.9	857.2	56	39.64487
213.5	0	174.24	159.21	11.66	1043.6	771.9	56	51.25564584
194.68	0	100.52	170.17	7.48	998	901.8	56	43.38872468
251.37	0	118.27	192.94	5.75	1043.6	754.3	56	39.2656582
165	0.02	143.57	163.81	0	1005.6	900.9	100	37.9556538
165	128.5	132.1	175.06	8.08	1005.8	746.6	100	55.0201848
178.03	129.8	118.6	179.94	3.57	1007.3	746.8	100	49.99390476
167.35	129.9	128.62	175.46	7.79	1006.3	746.6	100	53.65502232
172.38	13.61	172.37	156.76	4.14	1006.3	856.4	100	37.6798634

173.54	50.05	173.53	164.77	6.47	1006.2	793.5	100	56.06129356
167	75.4	167	164.03	7.91	1007.3	770.1	100	56.8128224
173.81	93.37	159.9	172.34	9.73	1007.2	746.6	100	50.93848688
190.34	0	125.18	166.61	9.88	1079	798.9	100	33.56369168
250	0	95.69	191.84	5.33	948.9	857.2	100	41.1617172
213.5	0	174.24	159.21	11.66	1043.6	771.9	100	52.95865156
194.68	0	100.52	170.17	7.48	998	901.8	100	44.27814872
251.37	0	118.27	192.94	5.75	1043.6	754.3	100	40.14818748
446	24	79	162	11.61	967	712	28	57.02655996
446	24	79	162	11.64	967	712	28	44.42293868
446	24	79	162	11.64	967	712	28	51.021224
446	24	79	162	10.3	967	712	28	53.38612668
446	24	79	162	11.61	967	712	3	35.36322404
446	24	79	162	11.64	967	712	3	25.02108404
446	24	79	162	11.64	967	712	3	23.34565736
446	24	79	162	11.61	967	712	7	52.00717468
446	24	79	162	11.64	967	712	7	38.01770664
446	24	79	162	11.64	967	712	7	39.300132
446	24	79	162	11.61	967	712	56	61.06688932
446	24	79	162	11.64	967	712	56	56.14403068
446	24	79	162	11.64	967	712	56	55.25460664
446	24	79	162	10.3	967	712	56	54.76507868
387	20	94	157	14.32	938	845	28	50.23522136
387	20	94	157	13.93	938	845	28	46.68441996
387	20	94	157	11.61	938	845	28	46.68441996

387	20	94	157	14.32	938	845	3	22.752708
387	20	94	157	13.93	938	845	3	25.510612
387	20	94	157	11.61	938	845	3	34.77027468
387	20	94	157	14.32	938	845	7	36.83870268
387	20	94	157	13.93	938	845	7	45.89841732
387	20	94	157	11.61	938	845	7	41.66503468
387	20	94	157	14.32	938	845	56	56.33708396
387	20	94	157	13.93	938	845	56	47.96684532
387	20	94	157	11.61	938	845	56	61.45989064
355	19	97	145	13.13	967	871	28	44.02993736
355	19	97	145	12.25	967	871	28	55.45455468
491	26	123	210	3.93	882	699	28	55.55108132
491	26	123	201	3.93	822	699	28	57.915984
491	26	123	210	3.93	882	699	3	25.60910857
491	26	123	210	3.93	882	699	7	33.48883428
491	26	123	210	3.93	882	699	56	59.59042572
491	26	123	201	3.93	822	699	3	29.54897143
491	26	123	201	3.93	822	699	7	37.92118
491	26	123	201	3.93	822	699	56	61.85584685
424	22	132	178	8.48	822	750	28	62.05284
424	22	132	178	8.48	882	750	3	32.01138572
424	22	132	168	8.92	822	750	28	72.09850532
424	22	132	178	8.48	822	750	7	39.00464228
424	22	132	178	8.48	822	750	56	65.69721315
424	22	132	168	8.92	822	750	3	32.10988228
424	22	132	168	8.92	822	750	7	40.28509772

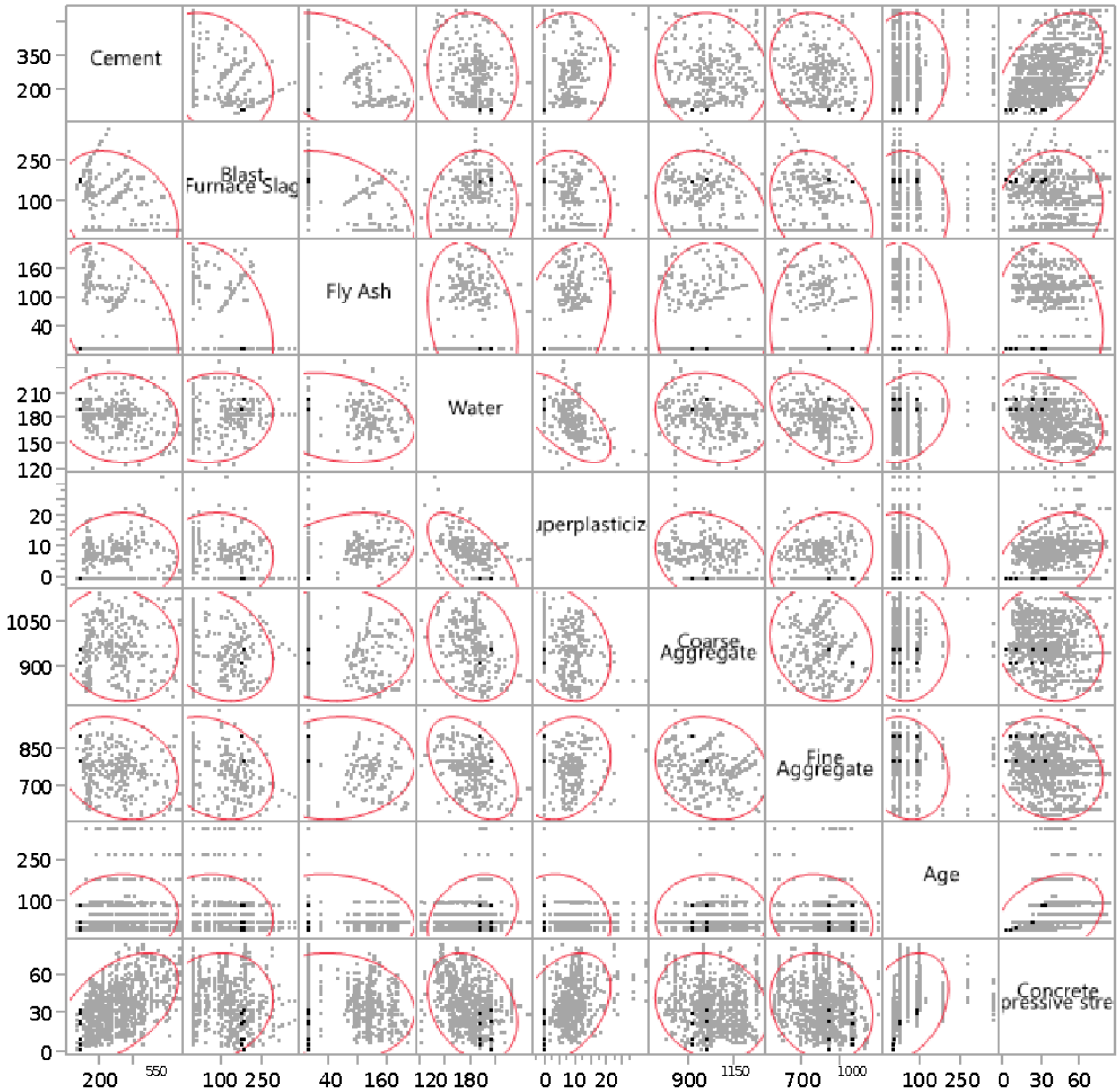
Data Dictionary:

Attributes	***Quantitative / Ca***	***Variable Type***
Cement	Quantitative	Predictor
Blast Furnace Slag	Quantitative	Predictor
Fly Ash	Quantitative	Predictor
Water	Quantitative	Predictor
Superplasticizer	Quantitative	Predictor
Coarse Aggregate	Quantitative	Predictor
Fine Aggregate	Quantitative	Predictor
Age	Quantitative	Predictor
Concrete Compressive Strength	Quantitative	Respons e

GRAPHICAL PLOTS:

Scatterplot Matrix:

A scatterplot matrix for all the included variables of the dataset have been displayed below:



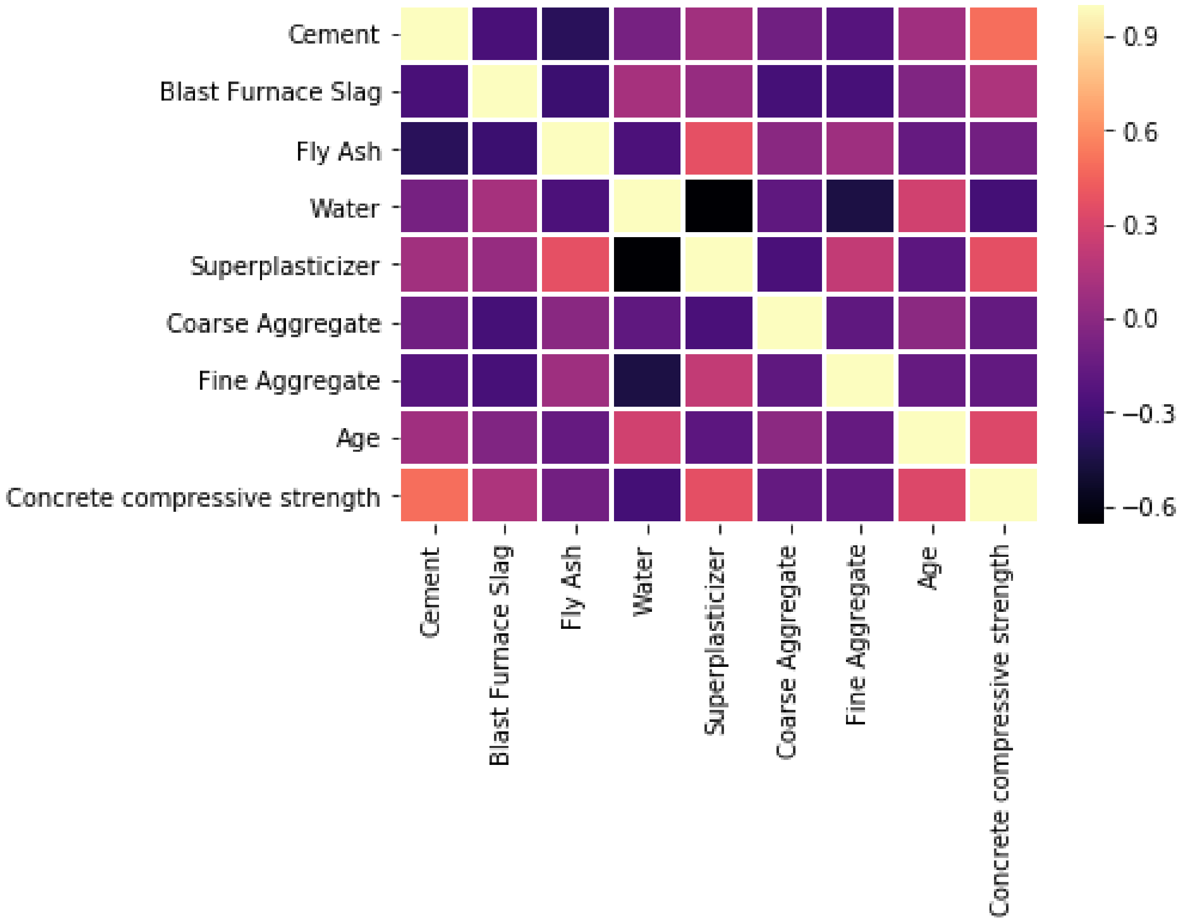
- The scatter plot matrix depicts that there may be slight linear dependencies between the predictor variables, which is evident from the strong correlation between concrete compressive strength and other predictor variables.

- Based on the scatter plot, we perform *Spearman's P correlation* test to understand the correlation between the variables. We find that almost 77 % of the predictors account for negative correlation and remaining accounts for positive correlation.

- The confidence curves in the pairwise plot indicate where the percentage of the data should lie assuming it the model to be bivariate normal distribution. The values outside these regions may be influential.

Correlation Heat Map:

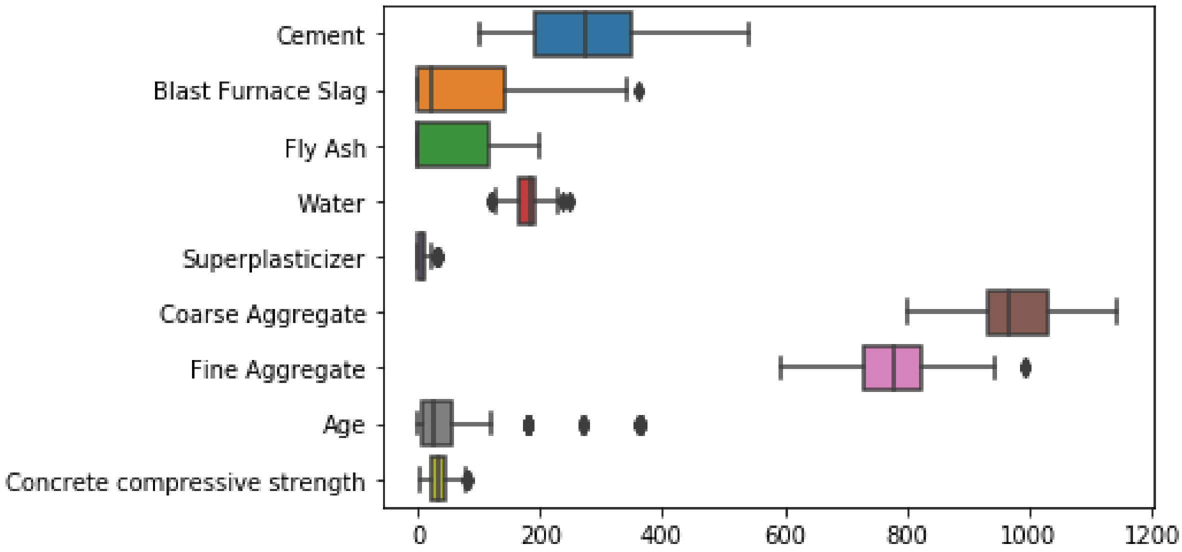
Followed by scatter plots, the analysis of the correlations between various factors affecting concrete strength is shown below:



- From the above correlation plot, it is evident that most of the variables are negatively correlated with mild-moderate effect.

- These plots can be regarded as heat map style displays of multiple correlation statistics.

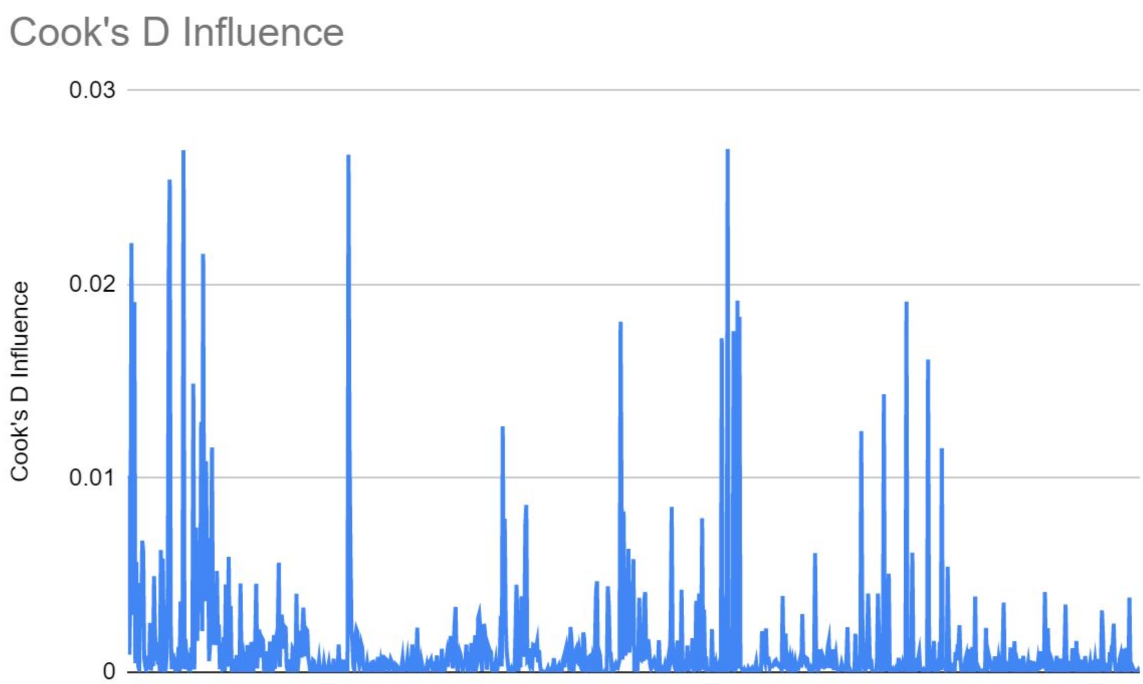
- The number of positive correlations is less when compared to the number of negative correlations present in the model.



Box Plot Inference:

- From the above boxplot, we could see that none of them had the same median or distributions.
- Most of the observations were concentrated on the low end of the scale, which results in a right(positive) skewed distribution.
- It is evident that the variables coarse aggregate, fine aggregate and water are symmetrically distributed, while the remaining variables are positively skewed.
- There are also indications of the outliers present in the data as some of the data points lie outside the whiskers.
- The variables cement, blast furnace slag and fly ash have a better spread compared to water, superplasticizer, age and compressive concrete strength.

Cook's D Influence:



MODEL BUILDING STEPS & HYPOTHESIS TESTING:

Null hypothesis:

The initial assumption is that there is no relation, which is expressed as:
H_0: $ß_1 = ß_2 = ß_3 = ß_4 = ß_5 = ß_6 = ß_7 = ß_8 = 0$

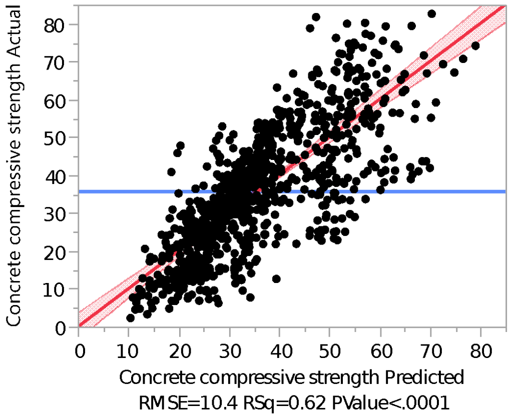
Alternate Hypothesis:

At least one of the independent variables is useful in predicting Y, which is expressed as:
H_1: $ß_1 \neq ß_2 \neq ß_3 \neq ß_4 \neq ß_5 \neq ß_6 \neq ß_7 \neq ß_8 \neq 0$

STEP 1: Initial Model Fit With 8 Predictor Variables:

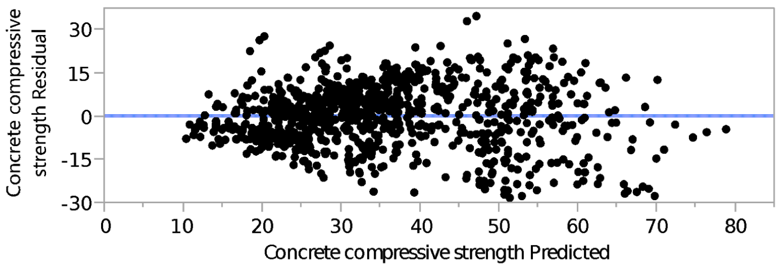
Residual Analysis:

- Actual values Vs Predicted values Plot:



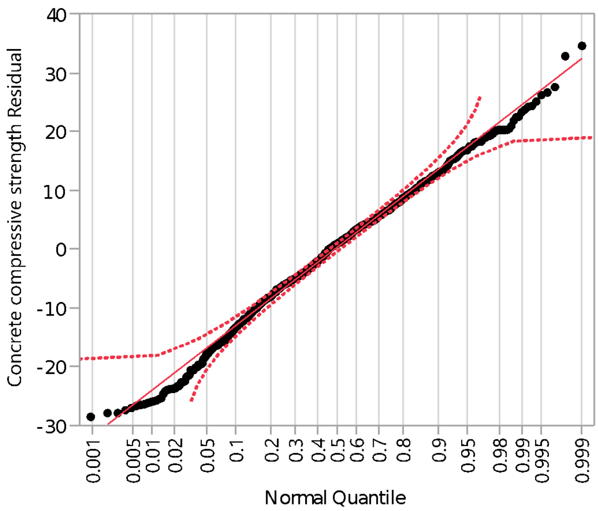
Ideally all over the points should lie closer to the regressed diagonal line, whereas in our case the points are somewhat aligned towards and dispersed away from the regression line. Therefore, the Coefficient of Determination (R^2) – which measures the goodness of fit for the regression line, is 0.62 – which is a bit lousy. There are points at which our model underestimates and overestimates the data.

- Residuals Vs Predicted Plot:



The data points are mostly located similar to the vertical and horizontal ranges, indicating a minimal number of outliers. There are some possibilities for the presence of outliers that could be influential.

- Residual Normal Quantile Plot:



The normal quantile plot above exhibits a linear pattern indicating normality in the data. There is no skewness present but there are unusual values at present at the tails of the plot revealing the possibility of outliers.

Model Summary:

Summary of Fit

RSquare	0.615465
Rsquare Adj	0.612452
Root Mean Square Error	10.39985
Mean of Response	35.81784
Observations (or Sum Wgts)	1030

Press

Press	Press RMSE	Press Rsquare
112914.41613	10.4702267	0.6068

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	176744.87	22093.1	204.2691
Error	1021	110428.16	108.2	**Prob > F**
C. Total	1029	287173.03		<.0001*

AICc	BIC
7758.28	7807.437

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	-23.16376	26.58842	-0.87	0.3839	-75.33795	29.01044	.
Cement	0.1197853	0.008489	14.11	<.0001*	0.1031267	0.1364439	7.4886572
Blast Furnace Slag	0.1038472	0.010136	10.25	<.0001*	0.0839571	0.1237374	7.2765286
Fly Ash	0.0879431	0.012585	6.99	<.0001*	0.0632474	0.1126387	6.1714547
Water	-0.150298	0.040179	-3.74	0.0002*	-0.229141	-0.071455	7.0046632
Superplasticizer	0.2906869	0.09346	3.11	0.0019*	0.1072915	0.4740823	2.9652972
Coarse Aggregate	0.0180302	0.009394	1.92	0.0552	-0.000404	0.0364644	5.0760437
Fine Aggregate	0.0201545	0.010703	1.88	0.0600	-0.000847	0.0411562	7.0053456
Age	0.1142256	0.005427	21.05	<.0001*	0.1035753	0.1248759	1.1183569

Predicted Expression:

$$Y = -23.16375581 + 0.119785255 * X_1 + 0.1038472489 * X_2 + 0.0879430817 * X_3 -0.150297904 * X_4 + 0.2906869435 * X_5 + 0.0180301836 * X_6 + 0.0201544557 * X_7 + 0.1142256201 * X_8$$

Inference:

- The results show that the model is a strong indicator of compressive strength, $F(8,1021) = 204.2691$, $p = 0.0001$.

- The values of the estimates will tell us the relationship between the outcome and the predictor variables. In our model, we have both positive and negative estimates. These estimates will give us an idea of how each predictor will influence the outcome if the effects of the other variables are kept constant. The variables cement, blast furnace slag, fly ash, water, superplasticizer and age make a significant contribution to the model, while coarse aggregate and fine aggregate do not.

- There is a significant relationship between water and concrete strength ($p = 0.0002$), superplasticizer and concrete strength ($p = 0.0019$) and the other factors such as cement, blast furnace slag, fly ash and age are highly significant ($p < 0.0001$).

- The above results indicate that the Variance Inflation Factor (VIF) of all the variables except the coarse aggregate and age is larger than the rest, indicating that there is potential for multicollinearity. Here, the cut-off values for VIF are assumed to be 5, so any value beyond that is considered to confronted with multicollinearity issues.

- The multiple coefficients of determination, R^2 (61.54 %) is not large. In this case we could say that 61.54 % of the variance in the data can be explained by the predictor variables. It also means that 38.46 % of the variation is still unexplained so that the addition of other independent variables could improve the model's fit.

- Some of the important diagnostics are multicollinearity checks and residual analysis such as normality checking. Our main motivation is to reduce multicollinearity and make the variables significant.

STEP 2: All Possible Regression:

Using the All Possible Regression, we found that given the variable number of 6, the C_p is 8.9555. As a result, we excluded 2 variables and arrived at the final predictive model for concrete compressive strength as follows:

Model	*Number*	*RSquare*	*RMSE*	*AICc*	*BIC*	*Cp*
Cement	1	0.2478	14.4954	8435.13	8449.92	971.1068

Cement,Superplasticizer	2	0.3511	13.4705	8285.09	8304.8	698.9999
Cement,Superplasticizer,Age	3	0.4816	12.0451	8055.7	8080.33	354.3021
Cement,Blast Furnace Slag,Water,Age	4	0.5577	11.1313	7894.18	7923.73	154.244
Cement,Blast Furnace Slag,Fly Ash,Water,Age	5	0.611	10.4452	7764.16	7798.61	14.9545
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Age	6	0.614	10.4098	7758.19	7797.55	8.9555
Cement,Blast Furnace Slag,Fly Ash,Water,Fine Aggregate,Age	6	0.6115	10.4427	7764.69	7804.04	15.4428
Cement,Blast Furnace Slag,Fly Ash,Water,Coarse Aggregate,Age	6	0.6111	10.4488	7765.89	7805.25	16.6523
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Coarse Aggregate,Age	7	0.6141	10.4128	7759.81	7804.07	10.5461
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Fine Aggregate,Age	7	0.6141	10.4135	7759.95	7804.21	10.6837
Cement,Blast Furnace Slag,Fly Ash,Water,Coarse Aggregate,Fine Aggregate,Age	7	0.6118	10.4439	7765.95	7810.21	16.6739
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Coarse Aggregate,Fine Aggregate,Age	8	0.6155	10.3998	7758.28	7807.44	9

STEP 3: Selected Testing Models:

Model	***Number***	***RSquare***	***RMSE***	***AICc***	***BIC***	***Cp***
Cement,Blast Furnace Slag,Fly Ash,Water,Age	5	0.611	10.4452	7764.16	7798.61	14.9545
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Age	6	0.614	10.4098	7758.19	7797.55	8.9555
Cement,Blast Furnace Slag,Fly Ash,Water,Superplasticizer,Coarse Aggregate,Age	7	0.6141	10.4128	7759.81	7804.07	10.5461

A multiple linear regression has been performed with the above-mentioned models. The results obtained from these models are displayed below:

- The regression model summary for the model with 5 predictor variables:
  - R – Sq : 61.09 %
  - R – Sq adj : 60.90 %
  - RMSE : 10.44
  - F (5,1024) : 321.62
  - PRESS : 1134888.26
  - R – Sq Pred : 60.48 %
  - Mean Sq Error : 109.1
  - All the variables significantly contribute to the model.
  - The VIFs of all the variables are lesser than 5 indicating the absence of multicollinearity.
  - This model involves a lack of fit with $MS_{Pure\ Error} = 20.86$ and $MS_{Lack\ of\ Fit} = 117.19$.

- The regression model summary for the model with 7 predictor variables:
  - R – Sq : 61.41 %
  - R – Sq adj : 61.14 %
  - RMSE : 10.41
  - F (5,1024) : 232.36
  - PRESS : 113050.62
  - R – Sq Pred : 60.63 %
  - Mean Sq Error : 108.4
  - The VIFs of all the variables are lesser than 5 indicating the absence of multicollinearity.
  - All the variables significantly contribute to the model expect coarse aggregate. Therefore, we do not consider this model because of the contribution of factors.
  - This model involves a lack of fit with $MS_{Pure\ Error} = 24.40$ and $MS_{Lack\ of\ Fit} = 111.40$.

STEP 4: Stepwise Regression:

We undertook stepwise regression in JMP in order to figure out a better predictive model for concrete compressive strength with low VIFs and little change in R^2.

Stepwise Fit for Concrete compressive strength
Stepwise Regression Control

Stopping Rule: Minimum BIC
Direction: Forward

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
110855.97	1023	10.409784	0.6140	0.6117	8.9554673	7	7758.188	7797.545

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"

[x]	[x]	Intercept	29.0302239	1	0	0.000	1
[ ]	[x]	Cement	0.10542749	1	66759.71	616.071	8e-107
[ ]	[x]	Blast Furnace Slag	0.08649363	1	32756.58	302.284	1.6e-59
[ ]	[x]	Fly Ash	0.06870838	1	8547.43	78.877	2.9e-18
[ ]	[x]	Water	-0.2182923	1	11567.41	106.746	7.2e-24
[ ]	[x]	Superplasticizer	0.23900253	1	865.1545	7.984	0.00481
[ ]	[ ]	Coarse Aggregate	0	1	44.27106	0.408	0.52297
[ ]	[ ]	Fine Aggregate	0	1	29.39769	0.271	0.60271
[ ]	[x]	Age	0.11349477	1	47731.47	440.475	1.3e-81

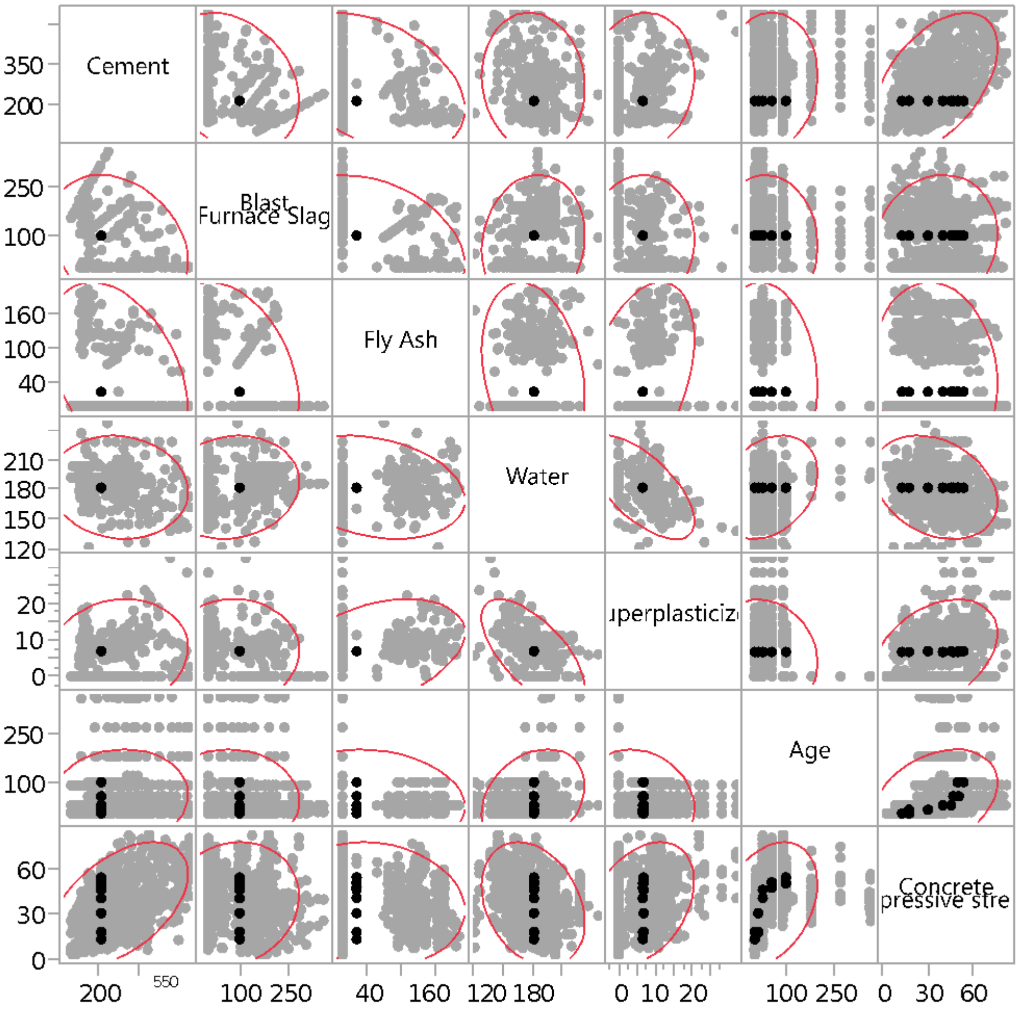
Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Cement	Entered	0.0000	71172.22	0.2478	971.11	2	8435.13	8449.92	( )
2	Superplasticizer	Entered	0.0000	29646.54	0.3511	699	3	8285.09	8304.8	( )
3	Age	Entered	0.0000	37497.74	0.4816	354.3	4	8055.7	8080.33	( )
4	Blast Furnace Slag	Entered	0.0000	19908.47	0.5510	172.23	5	7909.84	7939.38	( )
5	Water	Entered	0.0000	9544.652	0.5842	85.984	6	7832.66	7867.11	( )
6	Fly Ash	Entered	0.0000	8547.43	0.6140	8.9555	7	7758.19	7797.55	( )
7	Coarse Aggregate	Entered	0.5230	44.27106	0.6141	10.546	8	7759.81	7804.07	( )
8	Fine Aggregate	Entered	0.0600	383.5399	0.6155	9	9	7758.28	7807.44	( )
9	Best	Specific	.	.	0.6140	8.9555	7	7758.19	7797.55	(x)

- Any direction of stepwise regression can be performed as it leads us to the same final result of variable selection.

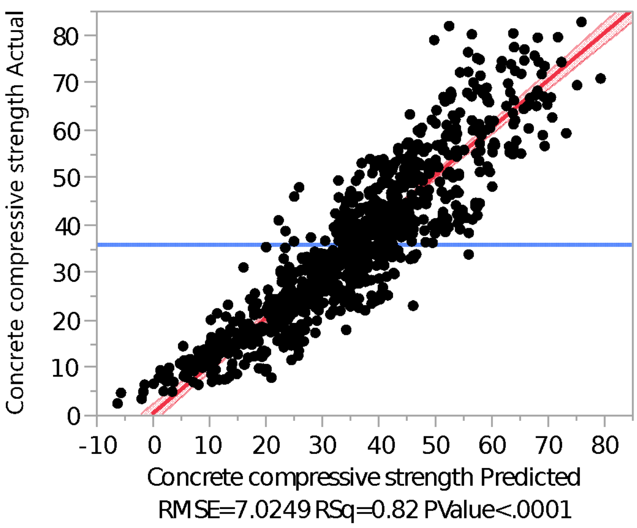
STEP 5: Final Selected Model:

Scatterplot Matrix:



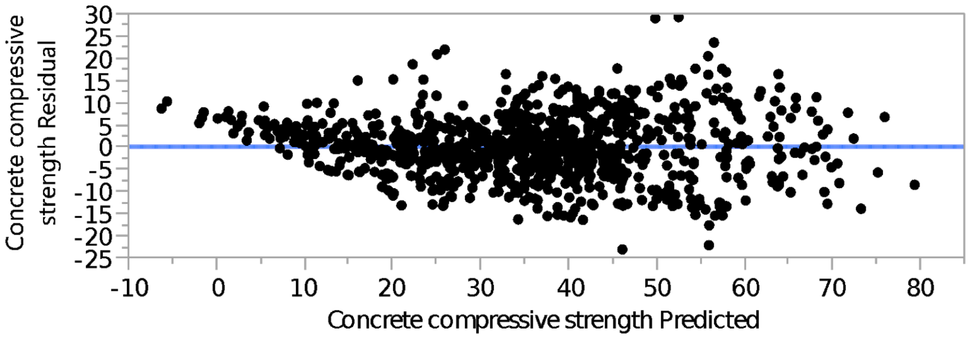
Residual Analysis:

- Actual values Vs Predicted values Plot:



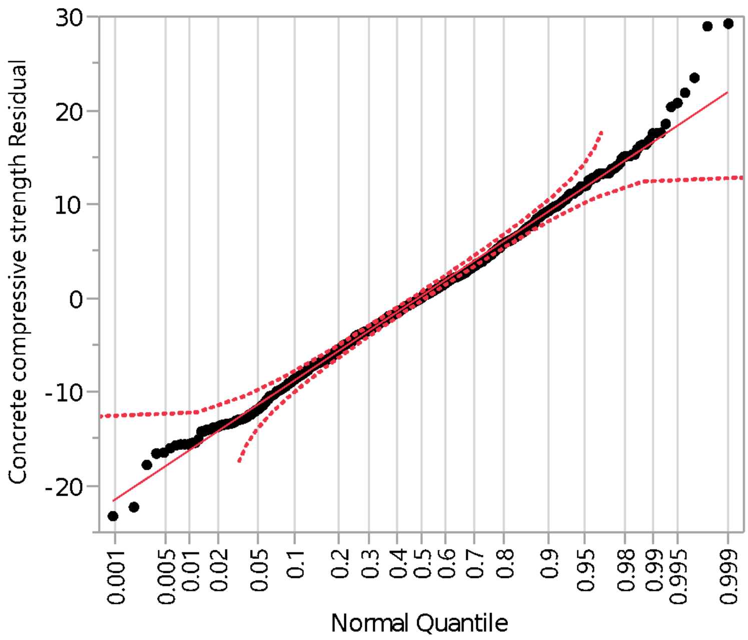
All the data points tend to move closer to the regression line region with slight dispersion leading to a change in the value of R-Sq. We could see that the data points with smaller values lie within the diagonal regressed region, while the points with larger values do not obey this.

- Residual vs Predicted Plot



The data points are similar to the initial model and are contained within the vertical and horizontal bands, indicating reduced outlier possibilities.

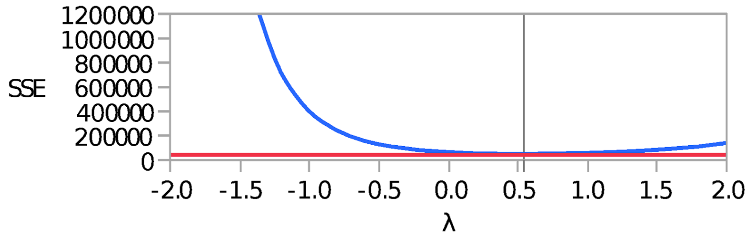
- Residual Normal Quantile Plot:



The normal quantile plot exhibits a linear trend showing the normality of the data. The plot clearly shows the model is skewed which can be inferred from the slight S – shape curve at the tails.

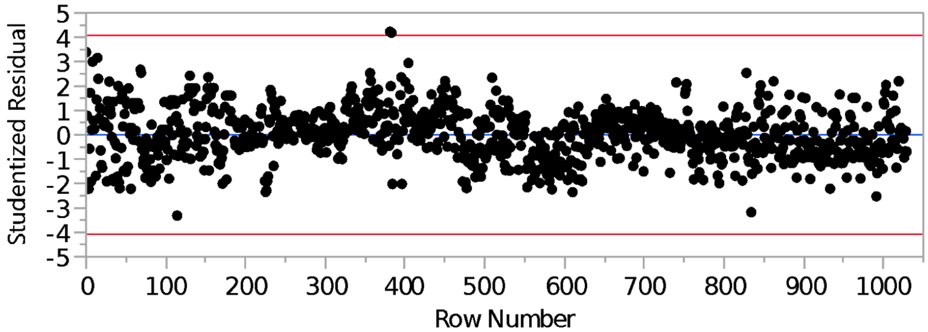
There are some unusual values present at the tails of the plot which show the possibilities of outliers that, if excluded, could have some impact on the final model performance. In order to check this, we have to compare this output with the model values having hat matrix larger than (2p/n).

- **Box Cox Transformations:**



Best λ=0.546

- **Studentized Residuals**:



Model Summary:

Summary of Fit

RSquare	0.824201
RSquare Adj	0.82317
Root Mean Square Error	7.024937
Mean of Response	35.81784
Observations (or Sum Wgts)	1030

AICc	BIC
6948.031	6987.388

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	236688.25	39448.0	799.356
Error	1023	50484.78	49.3	**Prob > F**
C. Total	1029	287173.03		<.0001*

Press

Press	Press RMSE	Press RSquare
51299.736788	7.05730612	0.8214

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	Std Beta	VIF
Intercept	-22.64616	3.102013	-7.30	<.0001*	-28.7332	-16.55913	0	.
Sqrt(Cement)	3.6707251	0.097512	37.64	<.0001*	3.4793788	3.8620715	0.681198	1.9055399
Blast Furnace Slag	0.0856627	0.003549	24.14	<.0001*	0.078698	0.0926274	0.442419	1.9553387
Sqrt(Fly Ash)	0.6493572	0.070663	9.19	<.0001*	0.5106959	0.7880185	0.214217	3.1621736
Water	-0.221366	0.014058	-15.75	<.0001*	-0.248951	-0.193781	-0.28298	1.8792165
Sqrt(Superplasticizer)	0.7512115	0.246136	3.05	0.0023*	0.2682232	1.2341997	0.070692	3.1219128
Log(Age)	8.5973393	0.187481	45.86	<.0001*	8.2294475	8.9652312	0.613163	1.0403944

Predicted Expression:

$$Y = -22.64616 + 3.6707251 * Sqrt(X_1) + 0.0856627 * X_2 + 0.6493572 * Sqrt(X_3) – 0.221366 * X_4 + 0.7512115 * Sqrt(X_5) + 8.5973393 * Log(X_8)$$

Inference:

- The results show that the model is a strong indicator of compressive strength, F (6,1023) = 799.3567, p = 0.0001.

- The F – ratio value for the final model is almost 4 times the value of the initial model which denotes that the variability between the factors is relatively larger than the variability within the factors. The F value combined with a smaller P value indicates that the model is highly significant.

- There is a decrease in the value of RMSE by 30 % indicating a better fit along with the prediction accuracy of the model.

- All the variables are considerably significant with VIFs less than 5 indicating the absence of multicollinearity. The PRESS value is roughly half when compared to the initial model.

STEP 6: Data Splitting & Cross-Validation:

Cross-validation measures the predictive ability of future models to assess the optimal number of components to be retained in your model.

- In order to obtain the model error estimates developed in this project, the dataset is randomly partitioned, with around 80% of all data points used for training data and the remaining 20 % used for validation.

- The training data set will be used to develop and tune the prediction model, while the validation data will be used for final model performance evaluations.

Summary of Fit

RSquare	0.819541
RSquare Adj	0.818224
Root Mean Square Error	7.043373
Mean of Response	36.26056
Observations (or Sum Wgts)	829

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	185193.54	30865.6	622.1760
Error	822	40778.68	49.6	Prob > F
C. Total	828	225972.22		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-24.40292	3.450393	-7.07	<.0001*	-31.17553	-17.6303
Sqrt(Cement)	3.6604633	0.10965	33.38	<.0001*	3.4452373	3.8756893
Blast Furnace Slag	0.0847177	0.00398	21.28	<.0001*	0.0769052	0.0925301
Sqrt(Fly Ash)	0.6301832	0.079297	7.95	<.0001*	0.4745356	0.7858308
Water	-0.211423	0.015597	-13.56	<.0001*	-0.242039	-0.180808
Sqrt(Superplasticizer)	0.9225095	0.274661	3.36	0.0008*	0.3833902	1.4616287
Log(Age)	8.5554532	0.213725	40.03	<.0001*	8.135942	8.9749644

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.8195	7.0136	829
Validation Set	0.8386	6.9632	201

Press

Press	Press RMSE	Press RSquare
51535.770743	7.88455383	0.7719

- The table below displays the data recorded for 10 K-fold cross-validations.

Train Ratio	*R Squared*	*Validation Ratio*	*R Squared*
80	0.8282	20	0.8072
80	0.8219	20	0.8284
80	0.8313	20	0.7956
80	0.8244	20	0.8221
80	0.8255	20	0.8191
80	0.8325	20	0.7842
80	0.8256	20	0.8173
80	0.8203	20	0.8359
80	0.8221	20	0.8295
80	0.8253	20	0.818
Avg =	0.82571	*Avg =*	0.81573

- Every single iteration, the data selected for training and test set differs, we perform 10 iterations and take an average of the R-Sq value to check the consistency.

- It is evident from the table that there is minimal difference in the value of R – Sq for the training and validation, indicating that the model fitted is adequate.

CONCLUSION:

- A multiple linear regression was carried out to investigate whether ingredients and age could significantly predict concrete compressive strength.

- Using all possible and stepwise regression, we were able to detect the variables that play a significant role in the model. Both forward and backward elimination resulted in approximately the same results which excludes coarse aggregate and fine aggregate parameters.

- The results of the regression indicated that the model explained **82.3%** of the **variance** after the application of some necessary transformations and that the model was a significant predictor of compressive strength, **F (6,1023) = 799.3567, p = .001**.

Comparison Table

	R-Sq	*Adj R-Sq*	*F Ratio*	*PRESS*	*Pred R-Sq*	*Mean Sq.Error*	*Cp*	*RMSE*
Initial Model	61.54	61.24	204.2691	112914.41	60.68	108.2	9	10.39
Final Model	82.42	82.31	799.3567	51299.73	82.14	49.3	8.85	7.0249

- The PRESS, Mallow's Cp and RMSE values were major influencing factors for the selection of the final model. ***NOTE:*** *Please refer the inference section of final model for more information.*

- The final predictive model was:

Concrete Compressive Strength = -22.64616 + 3.6707251 * Sqrt(Cement) + 0.0856627 * Blast Furnace Slag + 0.6493572 * Sqrt(Fly Ash) – 0.221366 * Water + 0.7512115 * Sqrt(Superplasticizer) + 8.5973393 * Log(Age)

- The scatterplot of studentized residuals shows that the data met the assumptions of homogeneity of variance & linearity and the residuals are roughly normally distributed. Therefore, no significant variations from these assumptions have been found.

FUTURE SCOPE:

A comparatively large data set with 5000 + measurements will possibly talk more about the prediction of the forecast. Since the current data set used is laboratory produced samples, the prediction accuracy might vary when compared with larger data sets.

Expanding the size of the training and validation set may contribute to some meaningful trends. Any other input parameters that influence the compressive strength could also be considered.

Other statistical techniques, such as classification and neural network methods may be suitable for the study of these information and may also reveal other interesting findings.

REFERENCES:

- 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).

- I-Cheng Yeh, "Modeling Concrete Strength with Augment-Neuron Networks," J. of Materials in Civil Engineering, ASCE, Vol. 10, No. 4, pp. 263-268 (1998).

- I-Cheng Yeh, "Design of High Performance Concrete Mixture Using Neural Networks," J. of Computing in Civil Engineering, ASCE, Vol. 13, No. 1, pp. 36-42 (1999).

- I-Cheng Yeh, "Prediction of Strength of Fly Ash and Slag Concrete By The Use of Artificial Neural Networks," Journal of the Chinese Institute of Civil and Hydraulic Engineering, Vol. 15, No. 4, pp. 659-663 (2003).

- I-Cheng Yeh, "A mix Proportioning Methodology for Fly Ash and Slag Concrete Using Artificial Neural Networks," Chung Hua Journal of Science and Engineering, Vol. 1, No. 1, pp. 77-84 (2003).

- Yeh, I-Cheng, "Analysis of strength of concrete using design of experiments and neural networks," Journal of Materials in Civil Engineering, ASCE, Vol.18, No.4, pp.597-604 (2006).