IEE 578: REGRESSION ANALYSIS

"Regression Analysis on Concrete Data"

Submitted to

-Dr. Douglas C. Montgomery.

Submitted by – Sanket Bhale (1214617222)

Contents

Introduction

The capacity or strength of material to withstand or hold against compression load is known as compression strength. It defines how well the material can hold up to compressive pressure loads.

One of the most important properties of cement accounted in the industry is its compression strength. Depending on the compression strength the grade and quality of concrete is defined. To test the compressive strength of the cement, a concrete block of cylinder or cube is used. These cubes are put under compressive pressure by the testing machines. The strength of concrete depends on its mixture and other factors. The compressive strength is non-linear function of its age and ingredients like cement, fly ash, superplasticizer, etc.

It is important to determine and model parameters affecting the compressive strength to ensure better grade of the concrete. We will conduct Regression analysis on the data available from the source to determine what factors affect the modeling and compressive strength of cement.

Data Set Description:

- Cement (component 1) -- quantitative -- kg in a m3 mixture -- Input Variable
- Blast Furnace Slag (component 2) -- quantitative -- kg in a m3 mixture -- Input Variable
- Fly Ash (component 3) -- quantitative -- kg in a m3 mixture -- Input Variable
- Water (component 4) -- quantitative -- kg in a m3 mixture -- Input Variable
- Superplasticizer (component 5) -- quantitative -- kg in a m3 mixture -- Input Variable
- Coarse Aggregate (component 6) -- quantitative -- kg in a m3 mixture -- Input Variable
- Fine Aggregate (component 7) -- quantitative -- kg in a m3 mixture -- Input Variable
- Age -- quantitative -- Day (1~365) -- Input Variable
- Concrete compressive strength -- quantitative -- MPa -- Output Variable

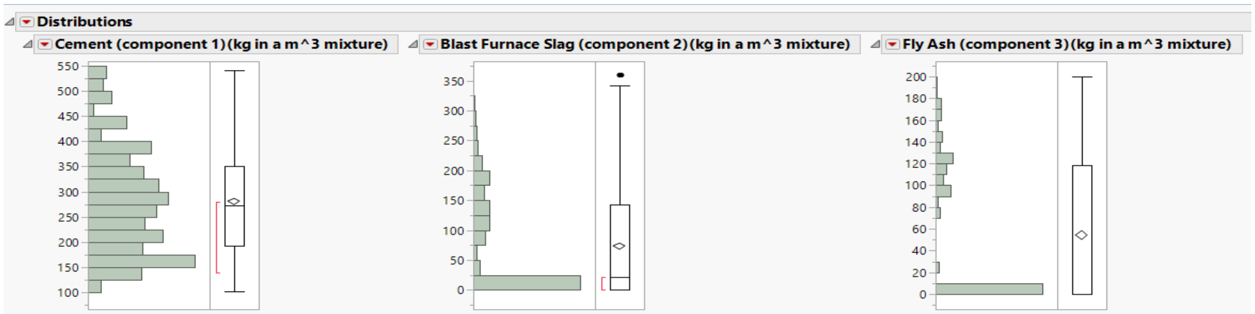
The response variable is the concrete compressive strength.

Source of data: https://www.kaggle.com/pavanraj159/concrete-compressive-strength-data-set

Data Cleaning

The dataset may need cleansing and modifications as it is obtained from open source. Also it has been recommended to clean data before modeling. We check the initial distribution of each parameter and check if there are any outliers:

Initial Distributions:



Cement — Quantiles

100.0%	maximum	540
99.5%		540
97.5%		520
90.0%		425
75.0%	quartile	350
50.0%	median	272.9
25.0%	quartile	192
10.0%		153.15
2.5%		134.9325
0.5%		108.3
0.0%	minimum	102

Cement — Summary Statistics

Mean	281.16786
Std Dev	104.50636
Std Err Mean	3.2562979
Upper 95% Mean	287.55761
Lower 95% Mean	274.77812
N	1030

Blast Furnace Slag — Quantiles

100.0%	maximum	359.4
99.5%		316.1
97.5%		262.2
90.0%		192
75.0%	quartile	143
50.0%	median	22
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Blast Furnace Slag — Summary Statistics

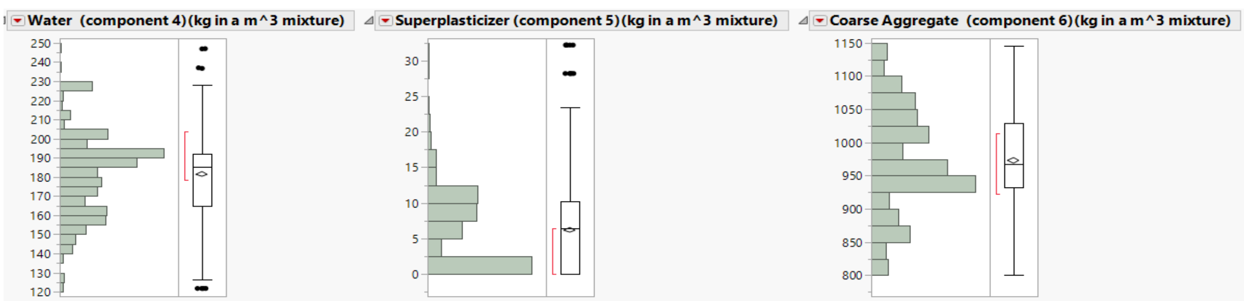
Mean	73.895825
Std Dev	86.279342
Std Err Mean	2.6883649
Upper 95% Mean	79.171129
Lower 95% Mean	68.620522
N	1030

Fly Ash — Quantiles

100.0%	maximum	200.1
99.5%		194.9845
97.5%		174.7
90.0%		141.54
75.0%	quartile	118.3
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Fly Ash — Summary Statistics

Mean	54.18835
Std Dev	63.997004
Std Err Mean	1.9940729
Upper 95% Mean	58.101263
Lower 95% Mean	50.275436
N	1030



Water — Quantiles

100.0%	maximum	247
99.5%		228
97.5%		228
90.0%		203.5
75.0%	quartile	192
50.0%	median	185
25.0%	quartile	164.9
10.0%		154.6
2.5%		141.8
0.5%		122.544
0.0%	minimum	121.8

Water — Summary Statistics

Mean	181.56728
Std Dev	21.354219
Std Err Mean	0.6653728
Upper 95% Mean	182.87292
Lower 95% Mean	180.26164
N	1030

Superplasticizer — Quantiles

100.0%	maximum	32.2
99.5%		31.58
97.5%		18.845
90.0%		12.29
75.0%	quartile	10.2
50.0%	median	6.4
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Superplasticizer — Summary Statistics

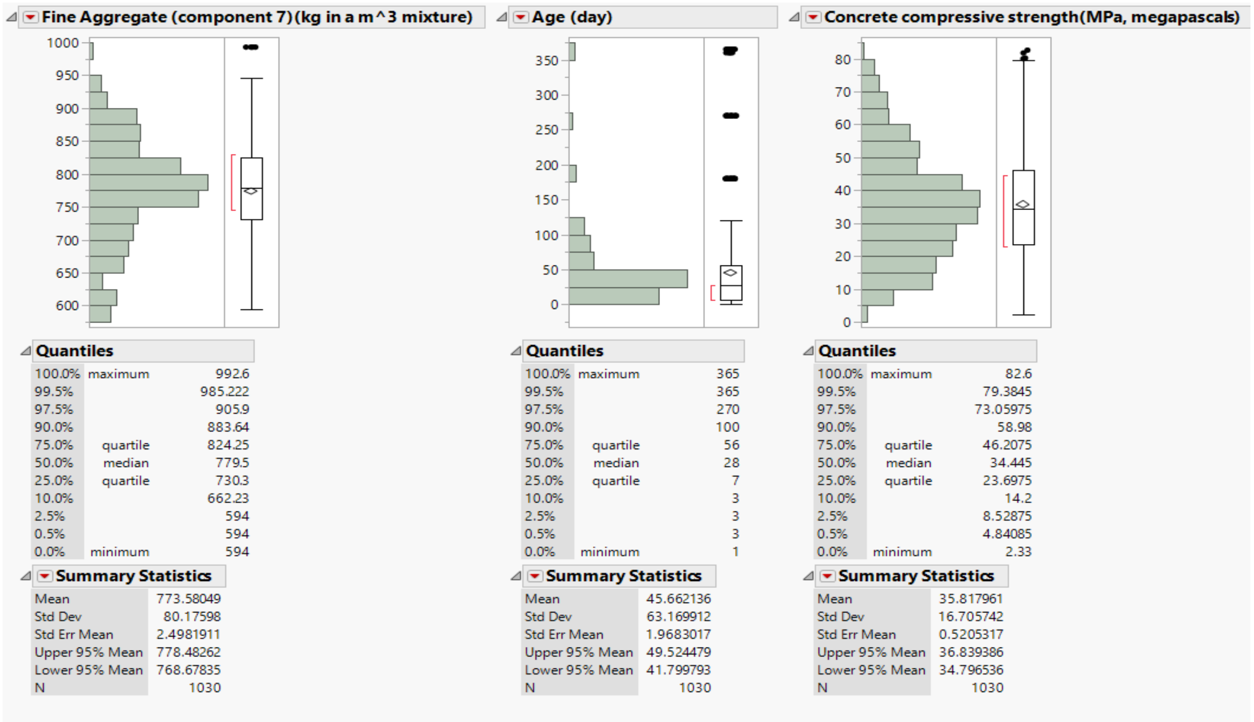
Mean	6.2046602
Std Dev	5.9738414
Std Err Mean	0.186138
Upper 95% Mean	6.5699136
Lower 95% Mean	5.8394068
N	1030

Coarse Aggregate — Quantiles

100.0%	maximum	1145
99.5%		1134.3
97.5%		1125
90.0%		1078.45
75.0%	quartile	1029.4
50.0%	median	968
25.0%	quartile	932
10.0%		852.1
2.5%		822
0.5%		801.1465
0.0%	minimum	801

Coarse Aggregate — Summary Statistics

Mean	972.91893
Std Dev	77.753954
Std Err Mean	2.4227236
Upper 95% Mean	977.67297
Lower 95% Mean	968.16489
N	1030

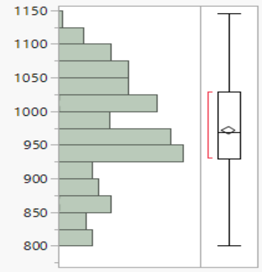


There are too man outliers present in the initial data. This may highly affect the effectiveness of our model.

Removing outliers:

After removing the outliers we obtain the clean file which can now be used to build a model and do the analysis.

Coarse Aggregate (component 6)(kg in a m^3 mixture)



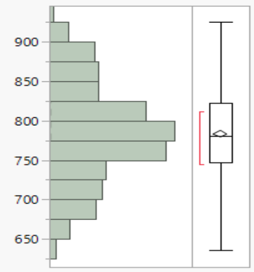
Quantiles

100.0%	maximum	1145
99.5%		1121.325
97.5%		1111
90.0%		1075
75.0%	quartile	1028.65
50.0%	median	969
25.0%	quartile	929.2
10.0%		852.1
2.5%		822
0.5%		801
0.0%	minimum	801

Summary Statistics

Mean	971.4504
Std Dev	77.557441
Std Err Mean	2.8395797
Upper 95% Mean	977.02493
Lower 95% Mean	965.87587
N	746

Fine Aggregate (component 7)(kg in a m^3 mixture)



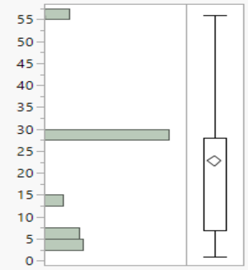
Quantiles

100.0%	maximum	925.7
99.5%		925.7
97.5%		901.8
90.0%		871.86
75.0%	quartile	822
50.0%	median	780.7
25.0%	quartile	746.75
10.0%		699
2.5%		659.9675
0.5%		641
0.0%	minimum	636

Summary Statistics

Mean	783.2929
Std Dev	61.553298
Std Err Mean	2.2536263
Upper 95% Mean	787.71711
Lower 95% Mean	778.86868
N	746

Age (day)



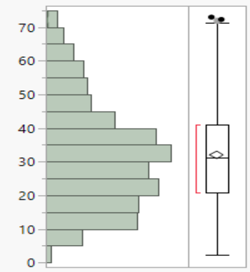
Quantiles

100.0%	maximum	56
99.5%		56
97.5%		56
90.0%		56
75.0%	quartile	28
50.0%	median	28
25.0%	quartile	7
10.0%		3
2.5%		3
0.5%		3
0.0%	minimum	1

Summary Statistics

Mean	22.731903
Std Dev	15.280306
Std Err Mean	0.5594518
Upper 95% Mean	23.830193
Lower 95% Mean	21.633614
N	746

Concrete compressive strength(MPa, megapascals)



Quantiles

100.0%	maximum	72.99
99.5%		72.01915
97.5%		66.321
90.0%		55.215
75.0%	quartile	41.05
50.0%	median	31.225
25.0%	quartile	20.695
10.0%		13.194
2.5%		8.35025
0.5%		4.88145
0.0%	minimum	2.33

Summary Statistics

Mean	32.098016
Std Dev	15.334326
Std Err Mean	0.5614296
Upper 95% Mean	33.200188
Lower 95% Mean	30.995844
N	746

Initial model:

To build the initial model all the regressors were used. They were fitted and following results were obtained:

Summary of Fit

RSquare	0.799113
RSquare Adj	0.796932
Root Mean Square Error	6.910116
Mean of Response	32.09802
Observations (or Sum Wgts)	746

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	139988.93	17498.6	366.4654
Error	737	35191.54	47.7	Prob > F
C. Total	745	175180.47		<.0001*

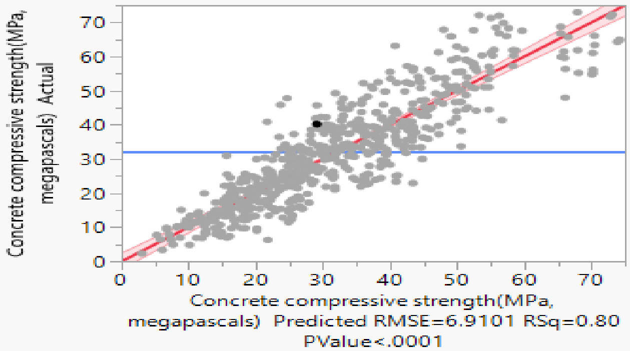
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	18.115993	21.69385	0.84	0.4039	.
Cement (component 1)(kg in a m^3 mixture)	0.1062746	0.006798	15.63	<.0001*	6.7847864
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.0803247	0.008092	9.93	<.0001*	7.3147802
Fly Ash (component 3)(kg in a m^3 mixture)	0.0458505	0.009861	4.65	<.0001*	6.2684273
Water (component 4)(kg in a m^3 mixture)	-0.218982	0.033976	-6.45	<.0001*	5.3186423
Superplasticizer (component 5)(kg in a m^3 mixture)	0.0791701	0.083887	0.94	0.3456	2.9478002
Coarse Aggregate (component 6)(kg in a m^3 mixture)	2.9988e-5	0.007809	0.00	0.9969	5.7233783
Fine Aggregate (component 7)(kg in a m^3 mixture)	0.0038177	0.008606	0.44	0.6574	4.377917
Age (day)	0.5526821	0.01714	32.24	<.0001*	1.0702472

As we can see the values for Super Plasticizer, Coarse aggregate and fine aggregate have Prob> t statistic, therefore these are insignificant and can be ignored while modeling in the next step. Although the VIF's of many parameters are greater or closer to 5, we will see in the next model if there is any need to take action to treat any possible collinearity. Step wise regression analysis is performed.

The Rsq adjusted has value of 79.69%

Actual by Predicted Plot



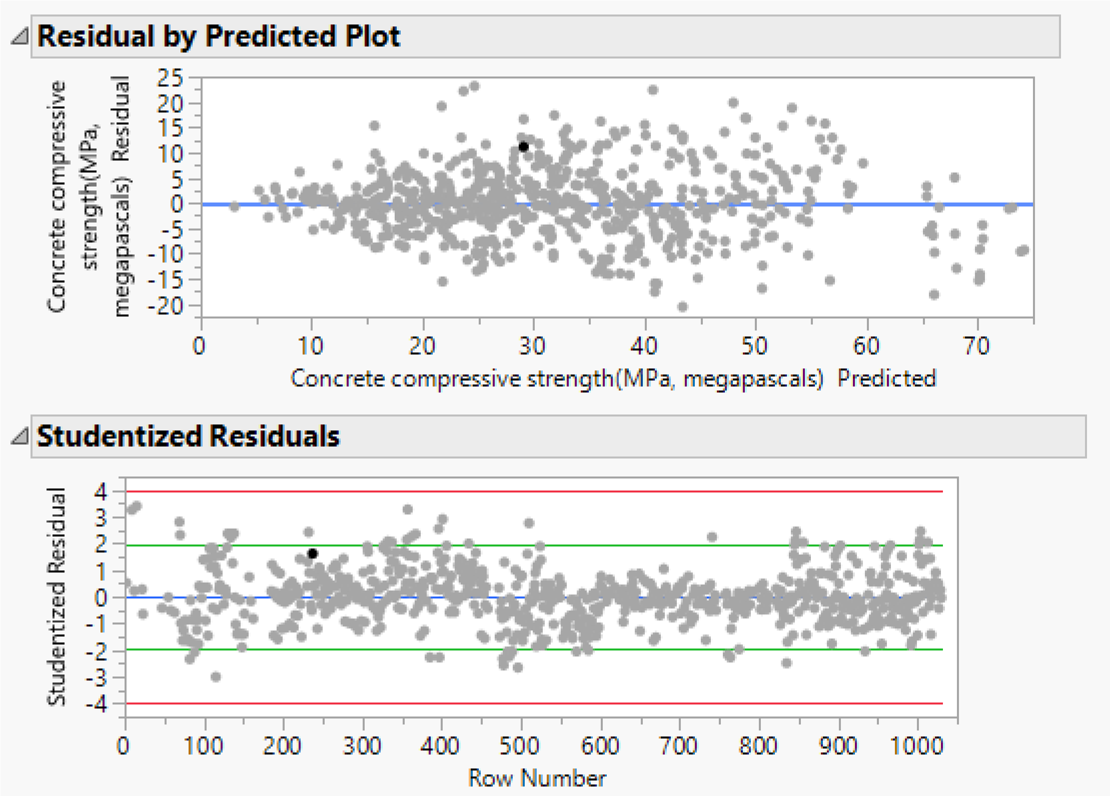
Effect Summary

Source	LogWorth	PValue
Age (day)	142.239	0.00000
Cement (component 1)(kg in a m^3 mixture)	47.063	0.00000
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	21.155	0.00000
Water (component 4)(kg in a m^3 mixture)	9.681	0.00000
Fly Ash (component 3)(kg in a m^3 mixture)	5.405	0.00000
Superplasticizer (component 5)(kg in a m^3 mixture)	0.461	0.34559
Fine Aggregate (component 7)(kg in a m^3 mixture)	0.182	0.65744
Coarse Aggregate (component 6)(kg in a m^3 mixture)	0.001	0.99694

Remove Add Edit FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	707	34058.206	48.1729	1.2752
Pure Error	30	1133.330	37.7777	**Prob > F**
Total Error	737	35191.536		0.2101
				Max RSq
				0.9935



Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

The residual plots look good and there is no need to delete any more outliers or points. All the residuals plotted lie within the permissible limits. It can also be said to have constant variance in the residual plots.

Model 2: (Excluding Superplasticizer, Coarse aggregate and fine aggregate)

The new model is built by excluding the Super Plasticizer, Coarse aggregate and fine aggregate parameters. Following results are obtained after fitting the new model:

Summary of Fit

RSquare	0.798598
RSquare Adj	0.797238
Root Mean Square Error	6.904915
Mean of Response	32.09802
Observations (or Sum Wgts)	746

Analysis of Variance

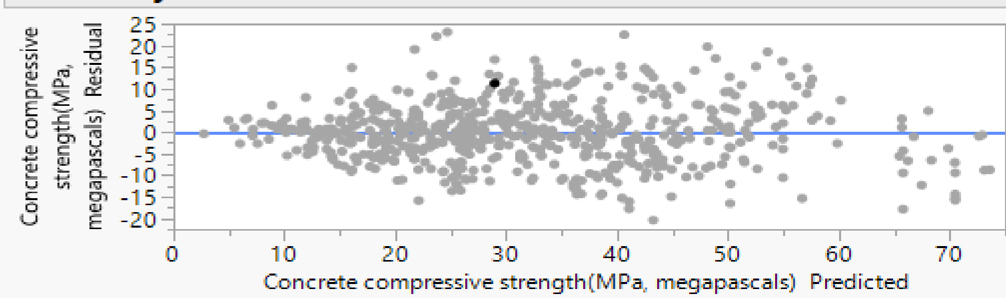
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	139898.86	27979.8	586.8505
Error	740	35281.61	47.7	Prob > F
C. Total	745	175180.47		<.0001*

Parameter Estimates

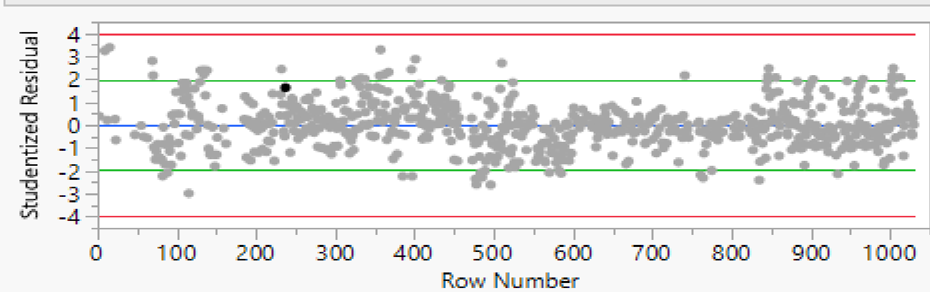
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	24.461523	3.372081	7.25	<.0001*	.
Cement (component 1)(kg in a m^3 mixture)	0.1064127	0.003215	33.10	<.0001*	1.5196718
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.0805974	0.003594	22.42	<.0001*	1.445459
Fly Ash (component 3)(kg in a m^3 mixture)	0.047836	0.005132	9.32	<.0001*	1.7006267
Water (component 4)(kg in a m^3 mixture)	-0.235904	0.015625	-15.10	<.0001*	1.1265273
Age (day)	0.5549727	0.016964	32.72	<.0001*	1.0499083

As seen from the above results, all the regressors present in this model are significant. Also, the VIF's have low values. This opts out the need to do any Box-Cox transformation on the model. Though, The Rsq adjusted value doesn't change much than the previous model, it successfully gets rid of the non-significant variables from the previous model.

Residual by Predicted Plot



Studentized Residuals



Train, validate and test:

Model 2 is now used to train, cross validate and test the prediction capability of the model. Following significant results are obtained:

Summary of Fit

RSquare	0.80487
RSquare Adj	0.802983
Root Mean Square Error	6.935953
Mean of Response	32.44428
Observations (or Sum Wgts)	523

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	102590.01	20518.0	426.5037
Error	517	24871.55	48.1	Prob > F
C. Total	522	127461.56		<.0001*

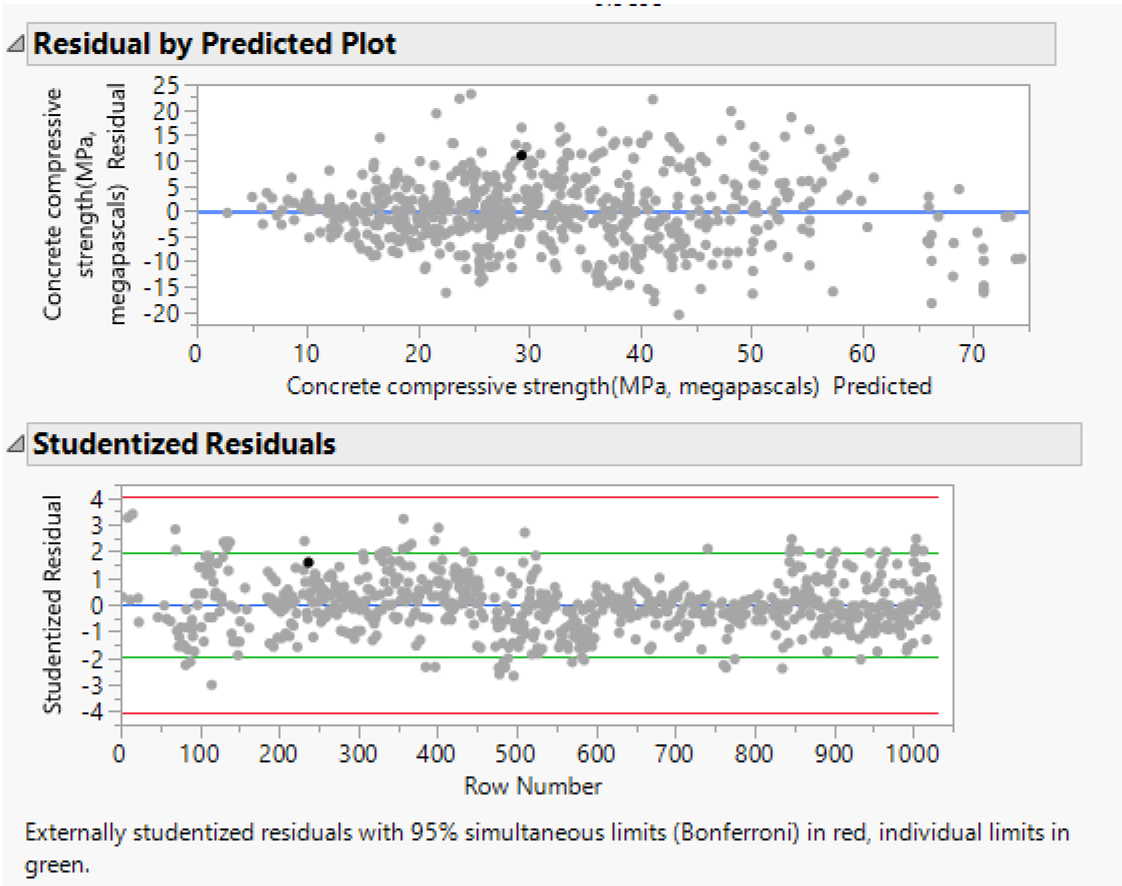
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	26.626556	4.096637	6.50	<.0001*	.
Cement (component 1)(kg in a m^3 mixture)	0.1059229	0.003916	27.05	<.0001*	1.5379511
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.0785094	0.004342	18.08	<.0001*	1.4267209
Fly Ash (component 3)(kg in a m^3 mixture)	0.0434329	0.006306	6.89	<.0001*	1.7618577
Water (component 4)(kg in a m^3 mixture)	-0.244538	0.018811	-13.00	<.0001*	1.1522175
Age (day)	0.5604357	0.020195	27.75	<.0001*	1.0696783

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.8049	6.8961	523
Validation Set	0.7501	7.2404	83
Test Set	0.7946	6.6208	140



Model 3: (Polynomial regression with degree 2)

The previous model performs well. We can still try model a multi linear regression with polynomial regressors up to degree 2. The model is fitted and following results are obtained:

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	664	25948.207	39.0786	1.7031
Pure Error	71	1629.090	22.9449	Prob > F
Total Error	735	27577.297		0.0029*
				Max RSq
				0.9907

Summary of Fit

RSquare	0.842578
RSquare Adj	0.840436
Root Mean Square Error	6.125368
Mean of Response	32.09802
Observations (or Sum Wgts)	746

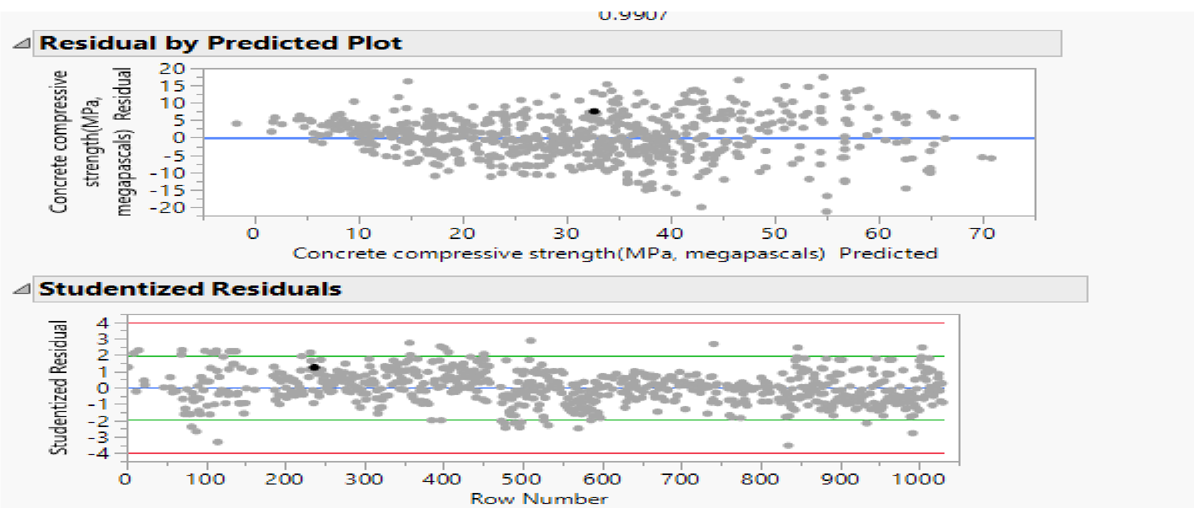
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	147603.17	14760.3	393.3973
Error	735	27577.30	37.5	Prob > F
C. Total	745	175180.47		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28.689609	3.086365	9.30	<.0001*
Cement (component 1)(kg in a m^3 mixture)	0.1120726	0.003358	33.37	<.0001*
(Cement (component 1)(kg in a m^3 mixture)-268.904)*(Cement (component 1)(kg in a m^3 mixture)-268.904)	-0.000129	2.33e-5	-5.54	<.0001*
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.0931377	0.004034	23.09	<.0001*
(Blast Furnace Slag (component 2)(kg in a m^3 mixture)-71.1385)*(Blast Furnace Slag (component 2)(kg in a m^3 mixture)-71.1385)	-0.000186	0.000037	-5.03	<.0001*
Fly Ash (component 3)(kg in a m^3 mixture)	0.0494396	0.004852	10.19	<.0001*
(Fly Ash (component 3)(kg in a m^3 mixture)-63.4177)*(Fly Ash (component 3)(kg in a m^3 mixture)-63.4177)	-0.000201	8.739e-5	-2.29	0.0220*
Water (component 4)(kg in a m^3 mixture)	-0.256691	0.014113	-18.19	<.0001*
(Water (component 4)(kg in a m^3 mixture)-179.569)*(Water (component 4)(kg in a m^3 mixture)-179.569)	0.0029264	0.000635	4.61	<.0001*
Age (day)	0.6197707	0.016492	37.58	<.0001*
(Age (day)-22.7319)*(Age (day)-22.7319)	-0.008467	0.000777	-10.89	<.0001*

The new model performs better than the previous model. There has been high improvement in the value of Rsq adjusted and so we can use this model for further analysis. The parameter interaction and their estimates used in the model can be seen in the above table.



Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

The residual plots are fine and there are no violations.

The prediction equation is as follows:

Prediction Expression

$$
\begin{aligned}
&28.689608754 \\
&+ 0.1120726048 \cdot \text{Cement (component 1)(kg in a m^3 mixture)} \\
&+ \left(\text{Cement (component 1)(kg in a m^3 mixture)} - 268.90402145\right) \cdot \left(\left(\text{Cement (component 1)(kg in a m^3 mixture)} - 268.90402145\right) \cdot -0.000129072\right) \\
&+ 0.0931377198 \cdot \text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} \\
&+ \left(\text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} - 71.13847185\right) \cdot \left(\left(\text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} - 71.13847185\right) \cdot -0.000186379\right) \\
&+ 0.0494396192 \cdot \text{Fly Ash (component 3)(kg in a m^3 mixture)} \\
&+ \left(\text{Fly Ash (component 3)(kg in a m^3 mixture)} - 63.41769437\right) \cdot \left(\left(\text{Fly Ash (component 3)(kg in a m^3 mixture)} - 63.41769437\right) \cdot -0.000200515\right) \\
&+ -0.256690692 \cdot \text{Water (component 4)(kg in a m^3 mixture)} \\
&+ \left(\text{Water (component 4)(kg in a m^3 mixture)} - 179.56903485\right) \cdot \left(\left(\text{Water (component 4)(kg in a m^3 mixture)} - 179.56903485\right) \cdot 0.0029263516\right) \\
&+ 0.6197707007 \cdot \text{Age (day)} \\
&+ \left(\text{Age (day)} - 22.731903485\right) \cdot \left(\left(\text{Age (day)} - 22.731903485\right) \cdot -0.008466844\right)
\end{aligned}
$$

Train validate test Model 3:

Model 3 is trained cross validated and tested. The Rsq adjusted values is little less than the original model. Therefore, we decide to stop here. This regression model explains the variability of the residual of the model better that previous constructed models.

Summary of Fit

RSquare	0.837889
RSquare Adj	0.835131
Root Mean Square Error	6.142661
Mean of Response	31.91161
Observations (or Sum Wgts)	539

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	103167.11	11463.0	303.7986
Error	529	19960.38	37.7	Prob > F
C. Total	538	123127.49		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.664279	3.542859	7.53	<.0001*
Cement (component 1)(kg in a m^3 mixture)	0.1135302	0.003778	30.05	<.0001*
(Cement (component 1)(kg in a m^3 mixture)-268.904)*(Cement (component 1)(kg in a m^3 mixture)-268.904)	-0.000145	2.713e-5	-5.33	<.0001*
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.095579	0.004782	19.99	<.0001*
(Blast Furnace Slag (component 2)(kg in a m^3 mixture)-71.1385)*(Blast Furnace Slag (component 2)(kg in a m^3 mixture)-71.1385)	-0.000203	4.515e-5	-4.50	<.0001*
Fly Ash (component 3)(kg in a m^3 mixture)	0.0460714	0.00553	8.33	<.0001*
Water (component 4)(kg in a m^3 mixture)	-0.250004	0.016402	-15.24	<.0001*
(Water (component 4)(kg in a m^3 mixture)-179.569)*(Water (component 4)(kg in a m^3 mixture)-179.569)	0.002578	0.000717	3.59	0.0004*
Age (day)	0.6146433	0.019277	31.88	<.0001*
(Age (day)-22.7319)*(Age (day)-22.7319)	-0.008437	0.000906	-9.31	<.0001*

Effect Tests

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.8379	6.0854	539
Validation Set	0.8372	6.3352	77
Test Set	0.8529	6.0713	130

Prediction Expression

$$
\begin{aligned}
&26.664278866 \\
&+ 0.1135301927 \cdot \text{Cement (component 1)(kg in a m^3 mixture)} \\
&+ \left(\text{Cement (component 1)(kg in a m^3 mixture)} - 268.90402145\right) \cdot \left(\left(\text{Cement (component 1)(kg in a m^3 mixture)} - 268.90402145\right) \cdot -0.000144641\right) \\
&+ 0.0955789773 \cdot \text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} \\
&+ \left(\text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} - 71.13847185\right) \cdot \left(\left(\text{Blast Furnace Slag (component 2)(kg in a m^3 mixture)} - 71.13847185\right) \cdot -0.000203055\right) \\
&+ 0.0460713612 \cdot \text{Fly Ash (component 3)(kg in a m^3 mixture)} \\
&+ -0.25000388 \cdot \text{Water (component 4)(kg in a m^3 mixture)} \\
&+ \left(\text{Water (component 4)(kg in a m^3 mixture)} - 179.56903485\right) \cdot \left(\left(\text{Water (component 4)(kg in a m^3 mixture)} - 179.56903485\right) \cdot 0.0025780117\right) \\
&+ 0.6146432923 \cdot \text{Age (day)} \\
&+ \left(\text{Age (day)} - 22.731903485\right) \cdot \left(\left(\text{Age (day)} - 22.731903485\right) \cdot -0.008436849\right)
\end{aligned}
$$

Results and Conclusion:

After performing the Regression analysis, the final model 3 is the best among all the models constructed. It explains variability better than other models. All the VIF's are less than 10 and no more violations are observed.