Seriman Doumbia

ID:202382485

Course: Data Science

Project: Regression Analysis

Contents

Concrete Compressive Strength Regression

1. Abstract

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.

2. Data Characteristics

Given is the variable name, variable type, the measurement unit and a brief description. The concrete compressive strength is the regression problem. The order of the listing corresponds to the order of numerals along the rows of the database.

3. Feature Description

Columns Name	**Data Type**	**Description**	**Measurement**
Cement	Quantitative	Input Variable	kg
Blast Furnace Slag	Quantitative	Input Variable	kg
Fly Ash	Quantitative	Input Variable	kg
Water	Quantitative	Input Variable	kg
Superplasticizer	Quantitative	Input Variable	kg
Coarse Aggregate	Quantitative	Input Variable	kg
Fine Aggregate	Quantitative	Input Variable	kg
Age	Quantitative	Input Variable	Date
Concrete compressive strength	Quantitative	Output Variable	MPa

Note: kg for kilogram

4. Summary Statistics

Number of instances (observations): 1030

Number of Attributes: 9

Attribute breakdown: 8 quantitative input variables, and 1 quantitative output variable

Missing Attribute Values: None

5. Task

Is there a relationship between the predictors (age and ingredients) and the response variable (compressive strength)?

Given there is a relationship

Q1. How strong is it?

Q2. Which predictors contribute to compressive strength?

Q3. How large is the effect of each predictor on compressive strength?

Q4. How accurately can I predict compressive strength?

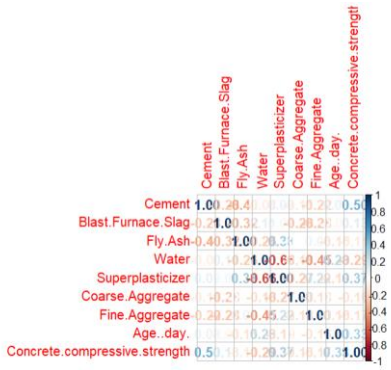
Q5. Is the relationship linear?

Q6. Is there synergy/interaction among the predictors?

5.1 Data Overview

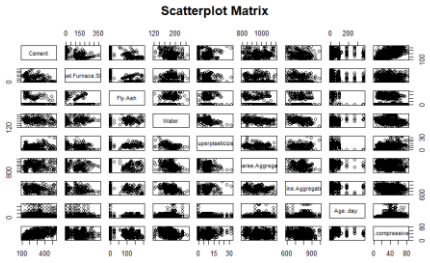
```
'data.frame':   1030 obs. of  9 variables:
 $ Cement                       : num  540 540 332 332 199 ...
 $ Blast.Furnace.Slag           : num  0 0 142 142 132 ...
 $ Fly.Ash                      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Water                        : num  162 162 228 228 192 228 228 228 228 228 ...
 $ Superplasticizer             : num  2.5 2.5 0 0 0 0 0 0 0 0 ...
 $ Coarse.Aggregate             : num  1040 1055 932 932 978 ...
 $ Fine.Aggregate               : num  676 676 594 594 826 ...
 $ Age..day.                    : int  28 28 270 365 360 90 365 28 28 28 ...
 $ Concrete.compressive.strength: num  80 61.9 40.3 41.1 44.3 ...
    Cement      Blast.Furnace.Slag    Fly.Ash           Water       Superplasticizer
Min.   :102.0   Min.   :  0.0      Min.   :  0.00   Min.   :121.8   Min.   : 0.000
1st Qu.:192.4   1st Qu.:  0.0      1st Qu.:  0.00   1st Qu.:164.9   1st Qu.: 0.000
Median :272.9   Median : 22.0      Median :  0.00   Median :185.0   Median : 6.350
Mean   :281.2   Mean   : 73.9      Mean   : 54.19   Mean   :181.6   Mean   : 6.203
3rd Qu.:350.0   3rd Qu.:142.9      3rd Qu.:118.27   3rd Qu.:192.0   3rd Qu.:10.160
Max.   :540.0   Max.   :359.4      Max.   :200.10   Max.   :247.0   Max.   :32.200
Coarse.Aggregate Fine.Aggregate    Age..day.      Concrete.compressive.strength
Min.   : 801.0   Min.   :594.0   Min.   :  1.00   Min.   : 2.332
1st Qu.: 932.0   1st Qu.:731.0   1st Qu.:  7.00   1st Qu.:23.707
Median : 968.0   Median :779.5   Median : 28.00   Median :34.443
Mean   : 972.9   Mean   :773.6   Mean   : 45.66   Mean   :35.818
3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00   3rd Qu.:46.136
Max.   :1145.0   Max.   :992.6   Max.   :365.00   Max.   :82.599
```

5.2 Correlation analysis



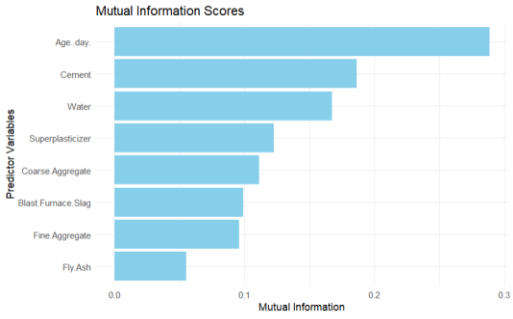
Graph shows high positive relationship between Cement and Concrete compressive strength with 0.5 as correlation value.

5.3 Visualize relationships



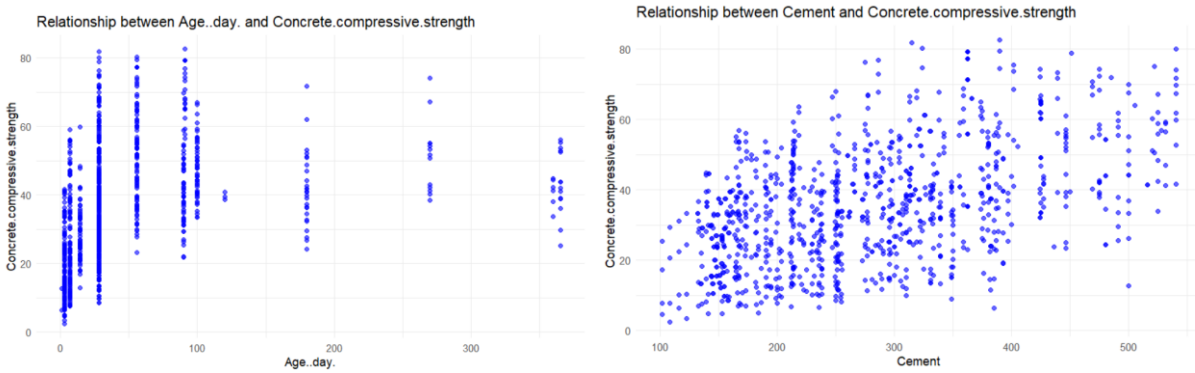
W see from the visual relationship there's a relationship between some feature of the dataset.

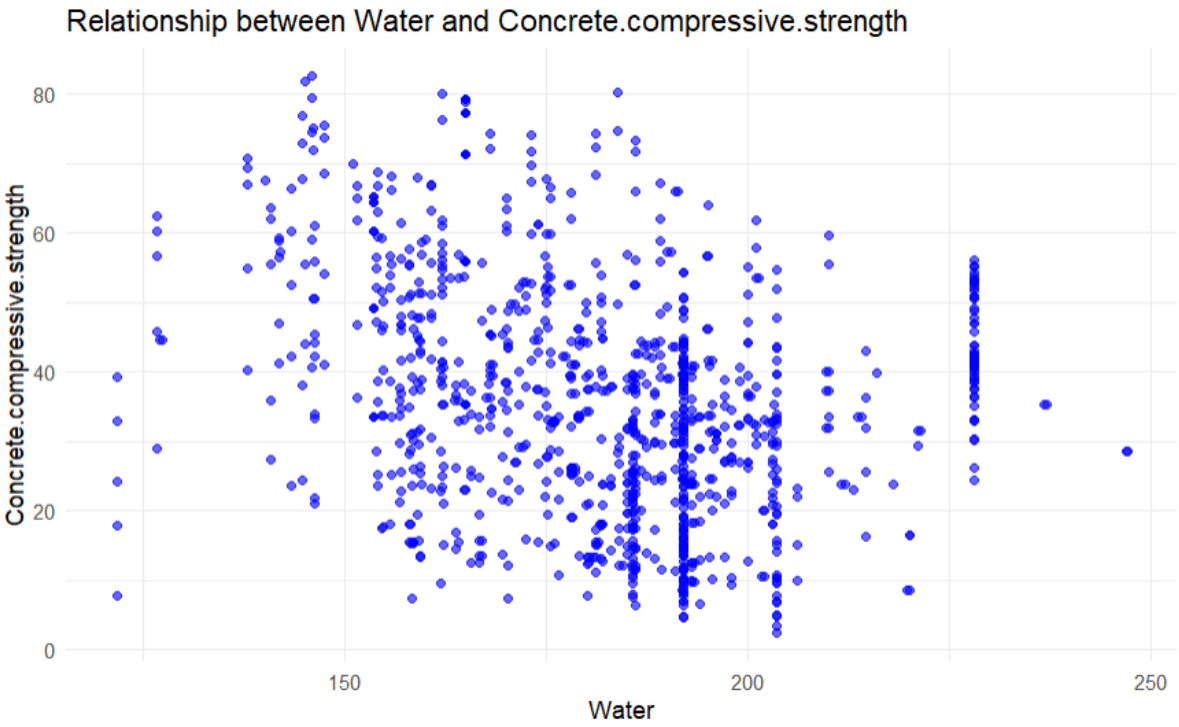
5.4 Mutual Information Scores Graph



The top 3 predictors are: Age, day., Cement, Water, which show that those 3 predictors are the most influential factor for predicting Concrete compressive strength.

5.5 Top Predictors





Overall those 3 predictors show some significant relationship with the response variable.

5.6 OLS Regression Analysis

Model Summary

R	0.785	RMSE	10.354
R-Squared	0.615	MSE	107.212
Adj. R-Squared	0.612	Coef. Var	29.035
Pred R-Squared	0.607	AIC	7758.064
MAE	8.215	SBC	7807.437

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	176744.872	8	22093.109	204.269	0.0000
Residual	110428.157	1021	108.157		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-23.164	26.588		-0.871	0.384	-75.338	29.010
Cement	0.120	0.008	0.749	14.110	0.000	0.103	0.136
Blast.Furnace.Slag	0.104	0.010	0.536	10.245	0.000	0.084	0.124
Fly.Ash	0.088	0.013	0.337	6.988	0.000	0.063	0.113
Water	-0.150	0.040	-0.192	-3.741	0.000	-0.229	-0.071
Superplasticizer	0.291	0.093	0.104	3.110	0.002	0.107	0.474
Coarse.Aggregate	0.018	0.009	0.084	1.919	0.055	0.000	0.036
Fine.Aggregate	0.020	0.011	0.097	1.883	0.060	-0.001	0.041
Age..day.	0.114	0.005	0.432	21.046	0.000	0.104	0.125

```
[1] 3.850583
F0 =  204.269 > F =  3.850583
```

Q1. Is there a relationship between the predictors (age and ingredients) and the response variable (compressive strength)?

Null hypothesis: coefficients for each predictor is zero.

F0 = 204.3 >> 1 (suggests at least one of the predictors is related to compressive strength)

F-statistic = 3.850583 << F0 (associated to the probability that the null hypothesis is true)

Therefore, there is a relationship between the predictors and the response variable.

Q2. How strong is the relationship?

R-squared = 0.616 (61.6% of variance is explained by the model)

Q3. Which predictors contribute to compressive strength?

Look at the p-values for each t-statistic for each predictor where p-values are the probability of t-statistic given the null hypothesis is true. A probability less than (0.05) is considered sufficient to reject the null hypothesis.

All are less than 0.05 except coarse and fine aggregates. Therefore, the aggregates do not contribute to compressive strength in this model.

Q4. How large is the effect of each predictor on compressive strength?

The only predictor confidence interval to include zero is coarse aggregate. The rest are considered to be statistically significant.

To test whether collinearity is the reason why the confidence interval includes 0 for coarse aggregate, the VIF scores are calculated.

VIF scores for each feature

Cement	Blast.Furnace.Slag	Fly.Ash	Water
7.488657	7.276529	6.171455	7.004663
Superplasticizer	**Coarse.Aggregate**	**Fine.Aggregate**	**Age..day.**
2.965297	5.076044	7.005346	1.118357

The VIF scores exceeding 5 to 10 indicate collinearity (where 1 is the minimum). The variables Aggregate, Blast Furnace Slag, Water, Fly Ash, and Cement have VIF scores ranging from 5 to 10, indicating potential multicollinearity, particularly if a conservative threshold is applied. Superplasticizer, with a VIF score below 5, has the widest confidence interval, which might also suggest the presence of multicollinearity. Consequently, we cannot definitively determine whether Coarse Aggregate is statistically significant, as the inclusion of 0 in confidence interval may be influenced by multicollinearity.

To assess association of each predictor, separate OLS for each predictor is performed

Cement

Model Summary

R	0.498	RMSE	14.481
R-Squared	0.248	MSE	209.710
Adj. R-Squared	0.247	Coef. Var	40.470
Pred R-Squared	0.245	AIC	8435.108
MAE	11.852	SBC	8449.920

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	71172.222	1	71172.222	338.726	0.0000
Residual	216000.806	1028	210.118		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	13.443	1.297		10.365	0.000	10.898	15.988
Cement	0.080	0.004	0.498	18.405	0.000	0.071	0.088

Blast.Furnace.Slag

Model Summary

R	0.135	RMSE	16.545
R-Squared	0.018	MSE	273.741
Adj. R-Squared	0.017	Coef. Var	46.237
Pred R-Squared	0.014	AIC	8709.560
MAE	13.435	SBC	8724.372

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	5220.125	1	5220.125	19.033	0.0000
Residual	281952.903	1028	274.273		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	33.889	0.680		49.869	0.000	32.555	35.222
Blast.Furnace.Slag	0.026	0.006	0.135	4.363	0.000	0.014	0.038

Fly.Ash

Model Summary

R	0.106	RMSE	16.604
R-Squared	0.011	MSE	275.691
Adj. R-Squared	0.010	Coef. Var	46.402
Pred R-Squared	0.007	AIC	8716.871
MAE	13.379	SBC	8731.683

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3211.677	1	3211.677	11.627	7e-04
Residual	283961.351	1028	276.227		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	37.314	0.679		54.978	0.000	35.982	38.646
Fly.Ash	-0.028	0.008	-0.106	-3.410	0.001	-0.043	-0.012

Water

Model Summary

R	0.290	RMSE	15.982
R-Squared	0.084	MSE	255.423
Adj. R-Squared	0.083	Coef. Var	44.664
Pred R-Squared	0.080	AIC	8638.224
MAE	13.076	SBC	8653.036

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	24086.915	1	24086.915	94.119	0.0000
Residual	263086.114	1028	255.920		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	76.952	4.269		18.025	0.000	68.575	85.330
Water	-0.227	0.023	-0.290	-9.701	0.000	-0.272	-0.181

Superplasticizer

Model Summary

R	0.366	RMSE	15.538
R-Squared	0.134	MSE	241.440
Adj. R-Squared	0.133	Coef. Var	43.424
Pred R-Squared	0.131	AIC	8580.232
MAE	12.615	SBC	8595.044

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	38490.057	1	38490.057	159.109	0.0000
Residual	248682.971	1028	241.910		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	29.467	0.699		42.165	0.000	28.095	30.838
Superplasticizer	1.024	0.081	0.366	12.614	0.000	0.865	1.183

Coarse.Aggregate

Model Summary

R	0.165	RMSE	16.469
R-Squared	0.027	MSE	271.225
Adj. R-Squared	0.026	Coef. Var	46.024
Pred R-Squared	0.023	AIC	8700.050
MAE	13.288	SBC	8714.862

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	7811.447	1	7811.447	28.745	0.0000
Residual	279361.581	1028	271.753		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	70.294	6.451		10.897	0.000	57.635	82.952
Coarse.Aggregate	-0.035	0.007	-0.165	-5.361	0.000	-0.048	-0.022

Fine.Aggregate

Model Summary

R	0.167	RMSE	16.462
R-Squared	0.028	MSE	271.010
Adj. R-Squared	0.027	Coef. Var	46.006
Pred R-Squared	0.024	AIC	8699.233
MAE	13.195	SBC	8714.045

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	8032.866	1	8032.866	29.583	0.0000
Residual	279140.163	1028	271.537		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	62.776	4.983		12.598	0.000	52.998	72.554
Fine.Aggregate	-0.035	0.006	-0.167	-5.439	0.000	-0.047	-0.022

Age..day

Model Summary

R	0.329	RMSE	15.769
R-Squared	0.108	MSE	248.653
Adj. R-Squared	0.107	Coef. Var	44.068
Pred R-Squared	0.104	AIC	8610.553
MAE	12.612	SBC	8625.365

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	31060.653	1	31060.653	124.673	0.0000
Residual	256112.375	1028	249.137		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	31.846	0.607		52.470	0.000	30.655	33.037
Age..day.	0.087	0.008	0.329	11.166	0.000	0.072	0.102

Looking at the p-value of the t-statistic: all variables have a strong association with compressive strength where **fly ash** has the largest value of 0.001.

Q5. How accurately can this model predict compressive strength?

The accuracy depends on what type of prediction: Individual response (Y = f(X) + ep), the prediction interval is used. Average response (f(X)), the confidence interval is used. Prediction intervals are wider than confidence intervals because the account for the uncertainty associated with the irreducible error (ep).

```
[1] "confidence interval"
       fit      lwr      upr
1 36.29962 33.31548 39.28376
[1] "prediction interval"
       fit      lwr      upr
1 36.29962 15.67507 56.92417
```

Confidence Interval (33, 39): The range where the average compressive strength for the given predictors is expected to lie.

Prediction Interval (15, 56): The range where an individual observation of compressive strength is expected to lie, accounting for random error (ei).

1. The Prediction Interval is wider than the Confidence Interval because it accounts for additional variability in individual responses.
2. The width of the intervals depends on:
    - The variability of the data (lamba^2).
    - Sample size (n).
    - Distance of the predictors from the mean (further predictors result in wider intervals).

Assessing Model Accuracy

```
R-squared:  0.6154647
[1] "R-MSE"
[1] 10.35431
[1] "MAE"
[1] 8.214899
```

R_square shows that **62%** of the variance in the dependent variable is explained by the predictors.

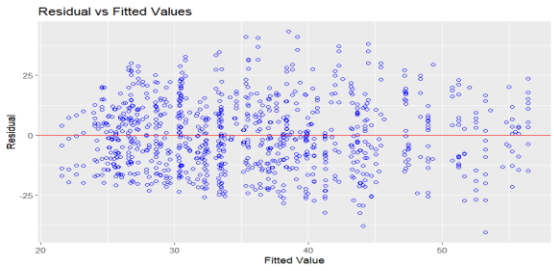
Q6. Is the relationship linear?

Non-linearity can be determined from residual vs. predicted value plot for each variable (top right plots below). When linearity exists, there should be no clear pattern.

The residual plot with the most non-linear form is for age where for ages 0 to 20, there are negative residuals then the residuals increase from 20 to 100 before decreasing again. Water and fine aggregate have slight non-linear patterns. Transformations of the predictors (e.g., sqrt(X), X^2) could accommodate the nonlinearities.

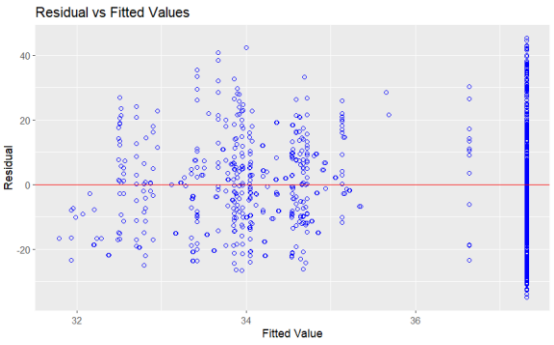
Residual vs. predicted value plot for each variable

Cement

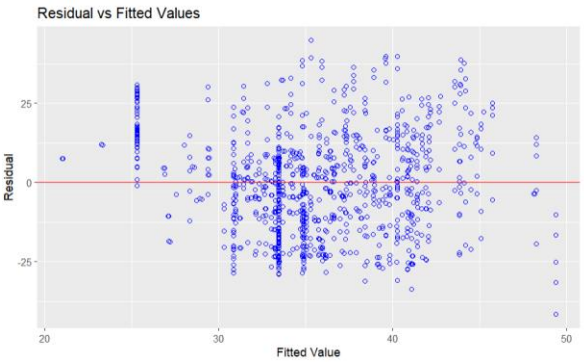


Blast.Furnace.Slag

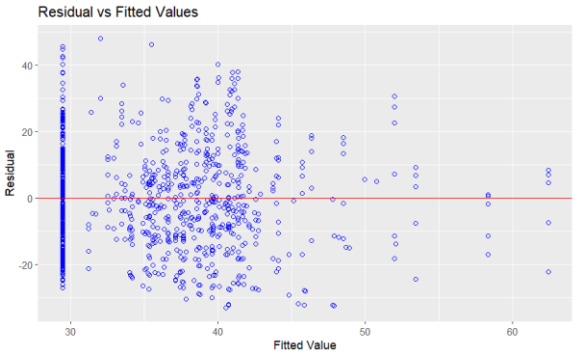
Fly.Ash



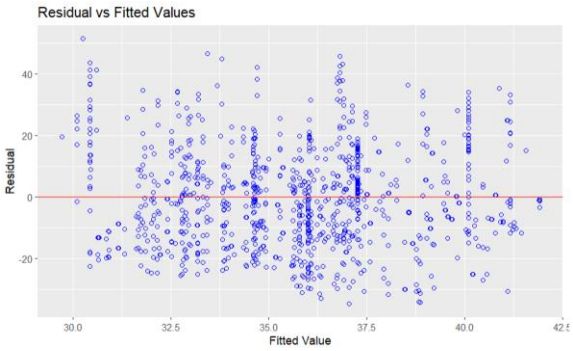
Water



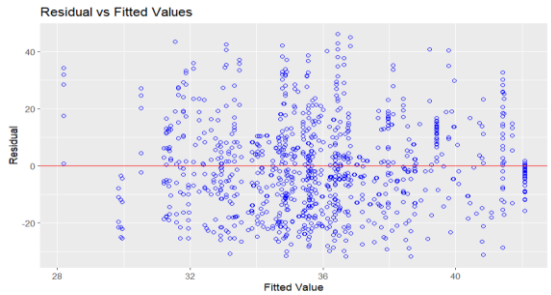
Superplasticizer



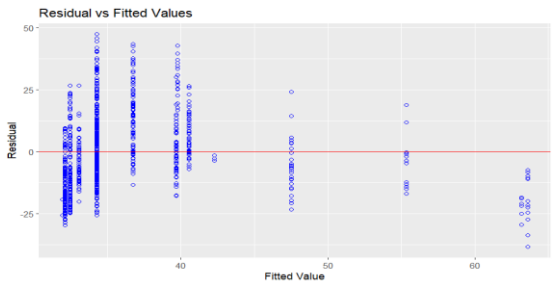
Coarse.Aggregate



Fine.Aggregate



Age..day



Let's conduct Breusch-Pagan test to see if constant variance error assumption is valid or not

Predictor	BP	df	p-value
Cement	**13.782**	**1**	**0.0002052**
Blast.Furnace.Slag	**17.107**	**1**	**3.53e^(-05)**
Fly.Ash	**28.463**	**1**	**9.55e(-08)**
Water	**6.119**	**1**	**0.01**
Superplasticizer	**5.3428**	**1**	**0.02**
Coarse.Aggregate	**0.22026**	**1**	**0.63**
Fine.Aggregate	**5.2906**	**1**	**0.02**
Age..day	**3.3442**	**1**	**0.06**

All predictor has p-value < 0.05 except Coarse.Aggregate Age..day so the assumption of constant variance error is valid for these two predictors. As Coarse.Aggregate and Age..day validate constant variance error assumption, we can apply a transformation function to validate the assumption to the remaining features.

Let's conduct lack of fit test to see if SLR is a good fit of the model or not

Perform Lack-of-Fit Test

Predictor	F0	p-value
Cement	139.44	< 2.2e-16
Blast.Furnace.Slag	226.56	< 2.2e-16
Fly.Ash	229.21	< 2.2e-16
Water	201.64	< 2.2e-16
Superplasticizer	182.61	< 2.2e-16
Coarse.Aggregate	223.13	< 2.2e-16
Fine.Aggregate	222.84	< 2.2e-16
Age..day	192.42	< 2.2e-16

With only one predictor which is Simple Linear Regression (SLR), the null hypothesis is not valid means SLR is not a good fit for the data.

Let's combined predictors to check for model selection

The predictor which gives the highest reduction in the uncertainty in predicting response variable is: **Cement, Superplasticizer, Age..day., Water, Fine.Aggregate, Coarse.Aggregate, Blast.Furnace.Slag, Fly.Ash** respectively. So, we'll insert first Cement in the model.

Multiple Linear Regression with 2 predictors combined, is not a good fit for data and it has R^2 value of 35%. The Two combined predictors which give the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer**. Hence it will be inserted first in the model.

Multiple Linear Regression with 3 predictors combined, is not a good fit for data and it has R^2 value of 48%. The Three combined predictors which give the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer+Age..day**. So, it will come first in the model.

Multiple Linear Regression with 4 predictors combined, is not a good fit for data and it has R^2 value of 55%. The fourth combined predictor which gives the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer+Age..day.+Blast.Furnace.Slag**. So, it will come first in the model.

Multiple Linear Regression with 5 predictors combined, is not a good fit for data and it has R^2 value of 58%. The fifth combined predictor which gives the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer+Age..day.+Blast.Furnace.Slag+Water**. So, it will come first in the model.

Multiple Linear Regression with 6 predictors combined, is a good fit for data and it has R^2 value R^2 of 61%. The sixth combined predictors which give the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer+Age..day.+Blast.Furnace.Slag+Water+Fly.Ash**.

As these 6 predictors combined together is a good fit for data with an acceptable R^2 value, so it can be the selected model. Let's conduct Breusch-Pagan to check the constancy variance assumption of the present selected model.

```
        Breusch-Pagan test

data:  mdel
BP = 140.53, df = 6, p-value < 2.2e-16
```

Looking at the p-value under Breusch-Pagan test: the non-constancy variance assumption is valid for the selected model.

5.7 Feature Engineering with OLS

Q7. Is there synergy among the predictors?

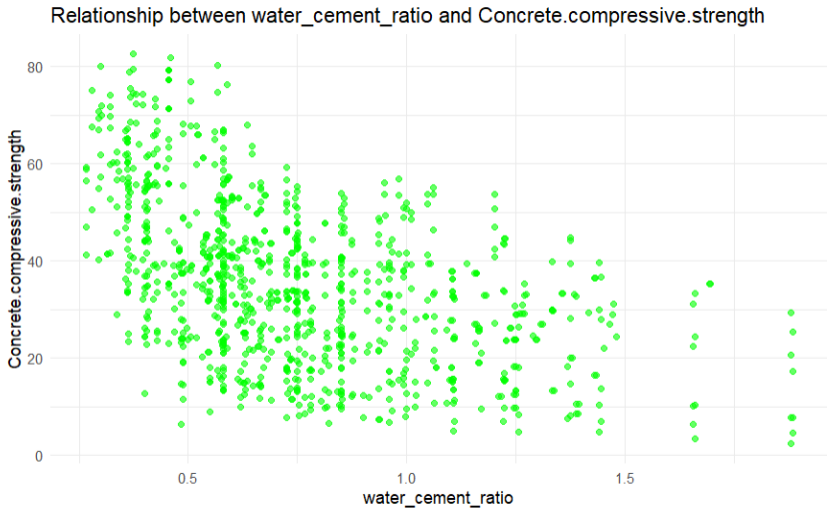
To evaluate the impact of an interaction term that accounts for non-additive relationships, I created a water-to-cement ratio (water:cement) and re-ran an OLS analysis. Including this interaction term resulted in an increase in the R-squared value from 0.615 to 0.618. However, since adding predictors naturally increases R-squared, the improvement of 0.003 is minimal. Adjusted R-squared, which penalizes for additional predictors, is a better measure in this case. It increased from 0.612 to 0.615, suggesting that some synergy exists between these predictors.

Similarly, AIC and SBC, which penalize models for additional complexity, provide further justification for including the interaction term. The AIC decreased from 7758 to 7752, while the SBC remained constant at 7807. This indicates that the added predictor improves the model without overfitting.

While cross-validation would be the best approach to assess the test set performance, the nonlinearity of compressive strength relative to the predictors suggests that linear regression may not be the most suitable model. More complex non-linear models would likely yield better predictive performance. For inference purposes, however, metrics like adjusted R-squared, SBC, and AIC are sufficient for evaluating the inclusion of this interaction term.

I also tested other interaction terms, including **cement:fine.aggregate**, **cement:coarse.aggregate**, **cement:fine.aggregate:coarse.aggregate**, and **superplasticizer:cement**. However, none of these terms improved the adjusted R-squared, and their p-values for the t-statistics were greater than 0.05, indicating they were not statistically significant.

Added interaction term water:cement:ratio plot against compressive strength



It depicts a negative relationship between **water:cement:ratio** and **compressive strength**

Generate OLS regression results with water : cement ratio

Model Summary

R	0.786	RMSE	10.317
R-Squared	0.618	MSE	106.431
Adj. R-Squared	0.615	Coef. Var	28.944
Pred R-Squared	0.609	AIC	7752.536
MAE	8.112	SBC	7806.846

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	177549.048	9	19727.672	183.557	0.0000
Residual	109623.981	1020	107.474		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-16.021	26.633		-0.602	0.548	-68.282	36.240
Cement	0.101	0.011	0.633	9.344	0.000	0.080	0.123
Blast.Furnace.Slag	0.106	0.010	0.548	10.472	0.000	0.086	0.126
Fly.Ash	0.088	0.013	0.338	7.023	0.000	0.063	0.113
Water	-0.123	0.041	-0.157	-2.979	0.003	-0.204	-0.042
Superplasticizer	0.288	0.093	0.103	3.090	0.002	0.105	0.471
Coarse.Aggregate	0.016	0.009	0.076	1.744	0.082	-0.002	0.035
Fine.Aggregate	0.020	0.011	0.098	1.905	0.057	-0.001	0.041
Age..day.	0.113	0.005	0.428	20.884	0.000	0.103	0.124
water_cement_ratio	-7.384	2.699	-0.139	-2.735	0.006	-12.681	-2.087

The insertion water:cement ratio in the model provoke an increase of R^2 value from 0.615 to 0.618.

Generate OLS summary with only water : cement ratio

Model Summary

R	0.501	RMSE	14.454
R-Squared	0.251	MSE	208.911
Adj. R-Squared	0.250	Coef. Var	40.393
Pred R-Squared	0.248	AIC	8431.181
MAE	11.865	SBC	8445.992

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

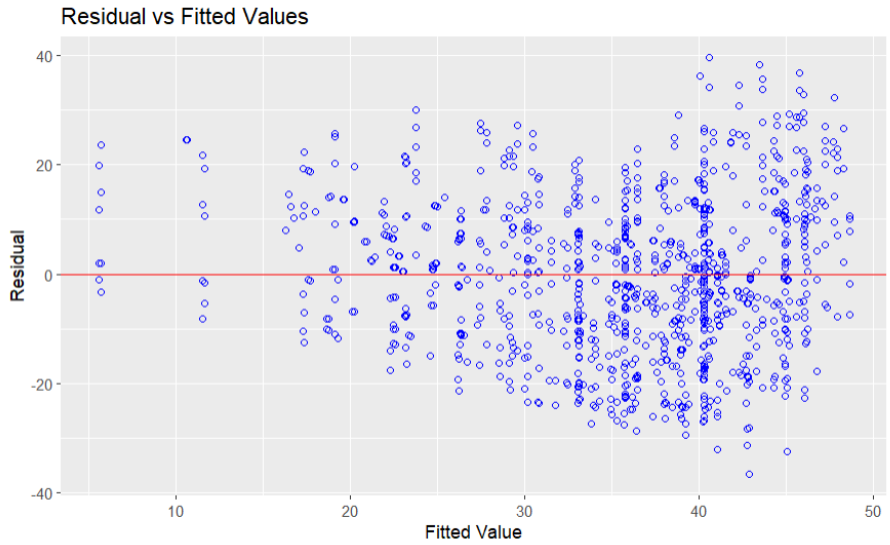
	Sum of Squares	DF	Mean Square	F	Sig.
Regression	71994.365	1	71994.365	343.948	0.0000
Residual	215178.664	1028	209.318		
Total	287173.028	1029			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	55.750	1.165		47.835	0.000	53.463	58.037
water_cement_ratio	-26.638	1.436	-0.501	-18.546	0.000	-29.457	-23.820

Looking at p-value of the t-statistic: water:cement:ratio has a strong association with compressive strength.

Non-linearity can be determined from residual vs. predicted value plot for water_cement_ratio variable



Water:cement:ratio residuals exhibits a near linear relationship.

6. Conclusion

The regression model successfully predicted the compressive strength of concrete based on input variables such as cement, water-cement ratio, Age..day., and other mix components.

Performance metrics such as R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE) indicate a better predictive capability with R^2 of 65%.

The most significant predictors of compressive strength were identified as Cement, Superplasticizer, Age of curing, and water-cement ratio.

This aligns with the theoretical understanding of concrete mechanics, where higher cement content and prolonged curing typically result in stronger concrete.

The relationship between some predictors and compressive strength was found to be non-linear, particularly for variables water_cement_ratio and age, indicating the need for polynomial or interaction terms to capture their effects.

Certain features, such as aggregate, showed limited impact on the prediction and can be excluded to simplify the model.

The model is based on Concrete Compressive Strength Regression dataset, and its generalizability may be limited for other concrete formulations or environmental conditions.

Advanced machine learning models (e.g., Random Forest, XGBoost) can be explored to capture complex interactions and non-linearities in the data.