**Seriman Doumbia**

**ID:202382485**

**Course: Data Science**

**Project: Regression Analysis**

Contents

[Concrete Compressive Strength Regression 2](#_Toc183877455)

[1. Abstract 2](#_Toc183877456)

[2. Data Characteristics 2](#_Toc183877457)

[3. Feature Description 2](#_Toc183877458)

[4. Summary Statistics 2](#_Toc183877459)

[5. Task 3](#_Toc183877460)

[6. Conclusion 18](#_Toc183877461)

# Concrete Compressive Strength Regression

## 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.

## 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.

## 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

## 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

## 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?

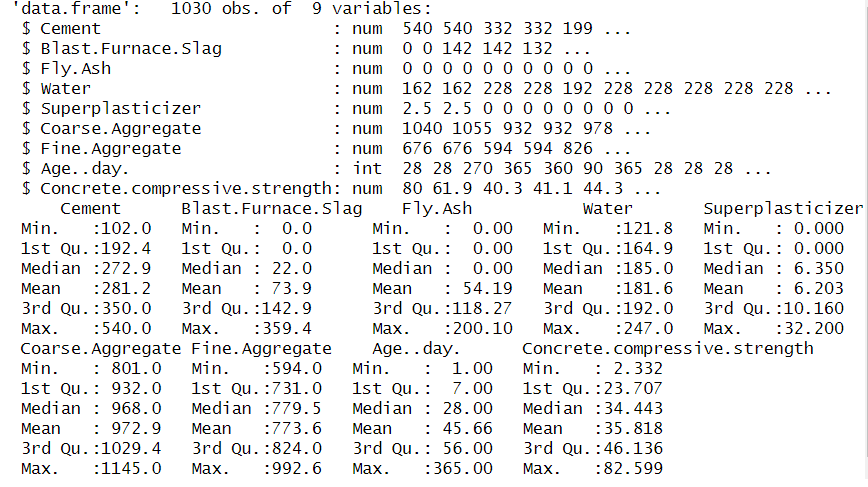
Q3. How large is the effect of each predictor on compressive strength?

Q4. How accurately can I predict compressive strength?

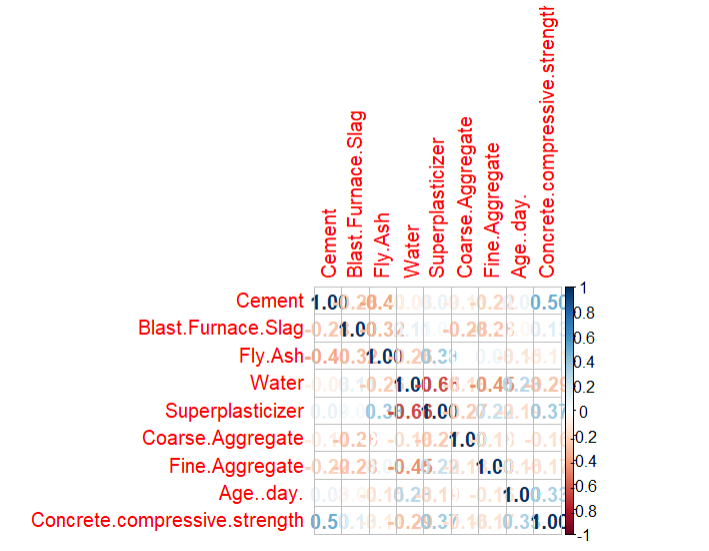
Q5. Is the relationship linear?

Q6. Is there synergy/interaction among the predictors?

* 1. Data Overview

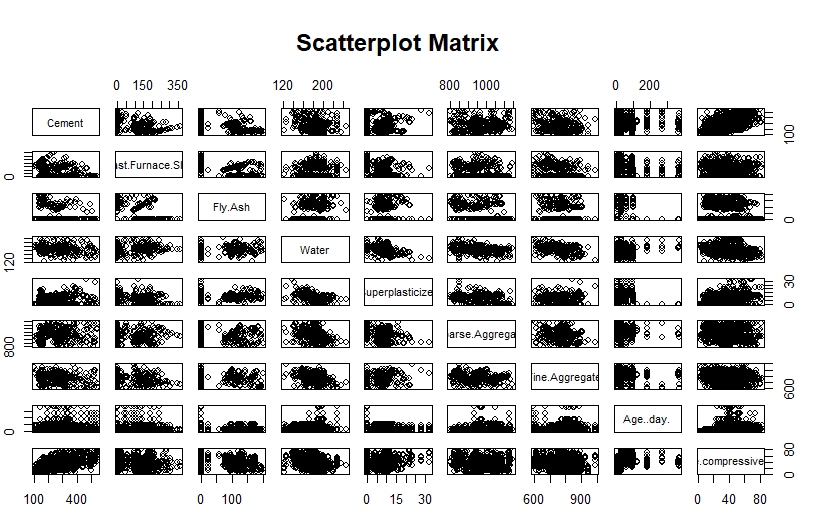


* 1. Correlation analysis

****

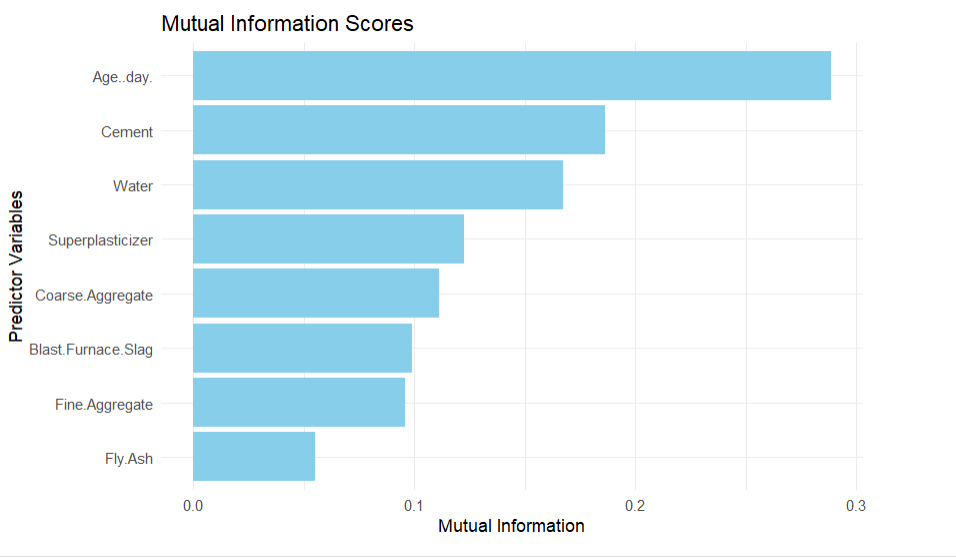
Graph shows high positive relationship between Cement and Concrete compressive strength with 0.5 as correlation value.

* 1. Visualize relationships

****

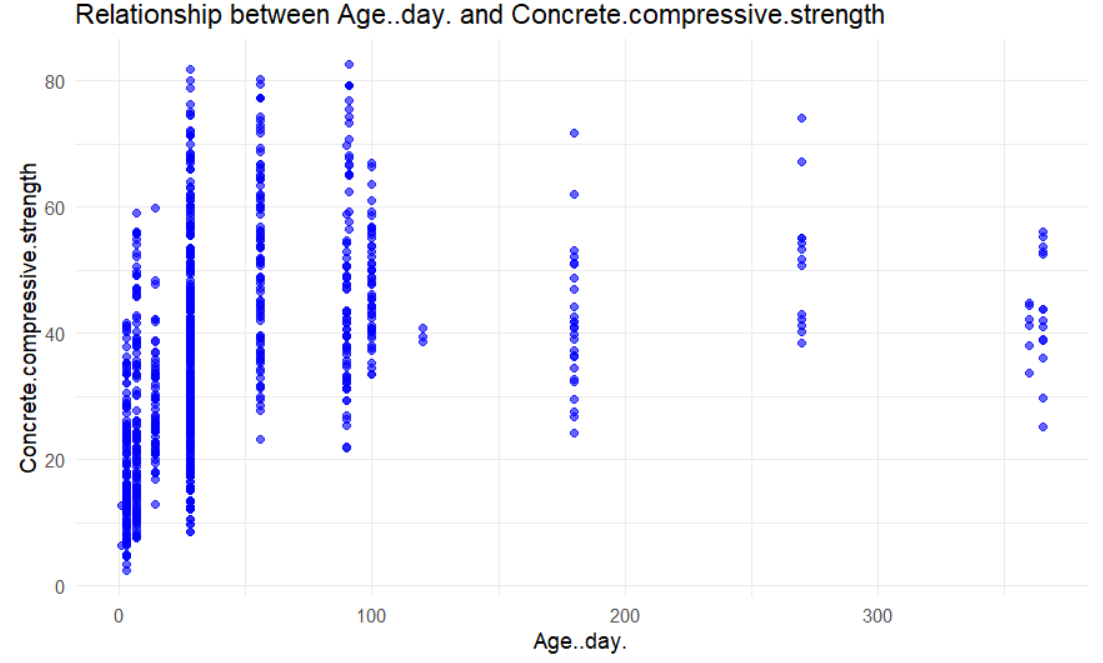
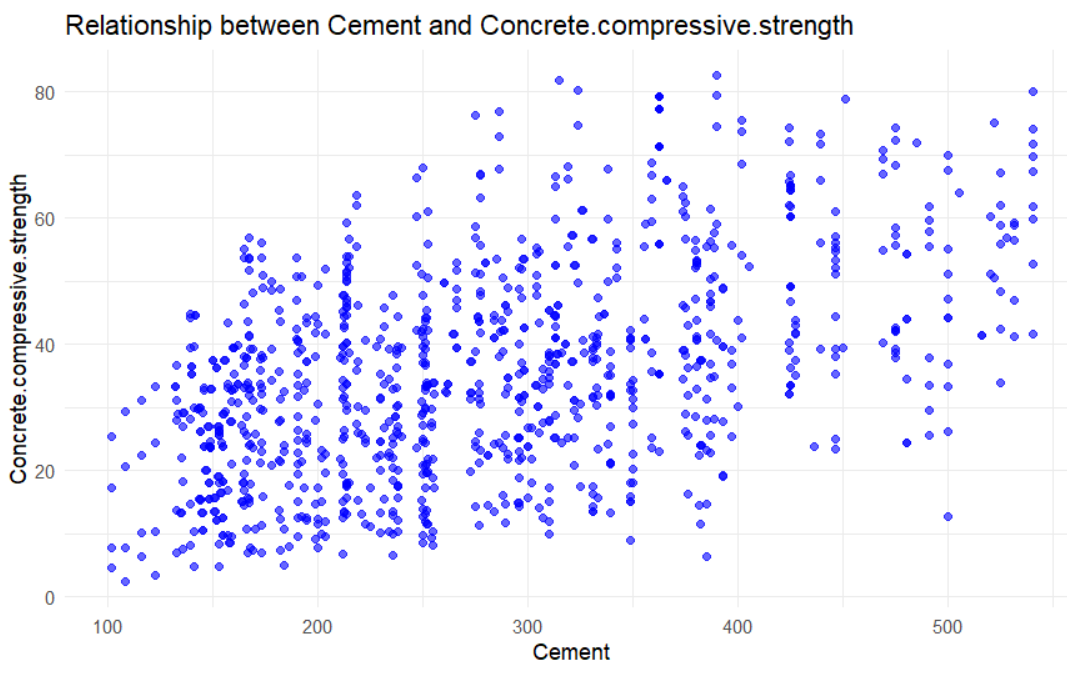
W see from the visual relationship there’s a relationship between some feature of the dataset.

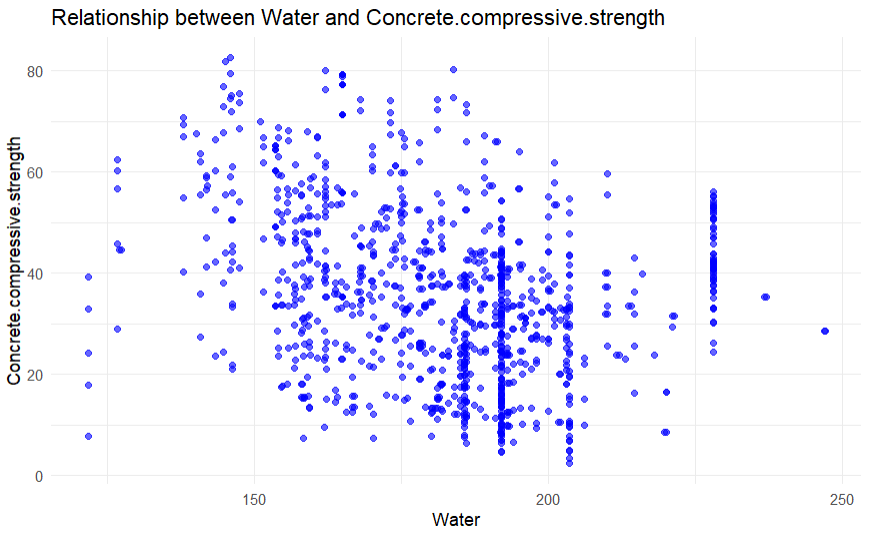
* 1. Mutual Information Scores Graph

****

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.

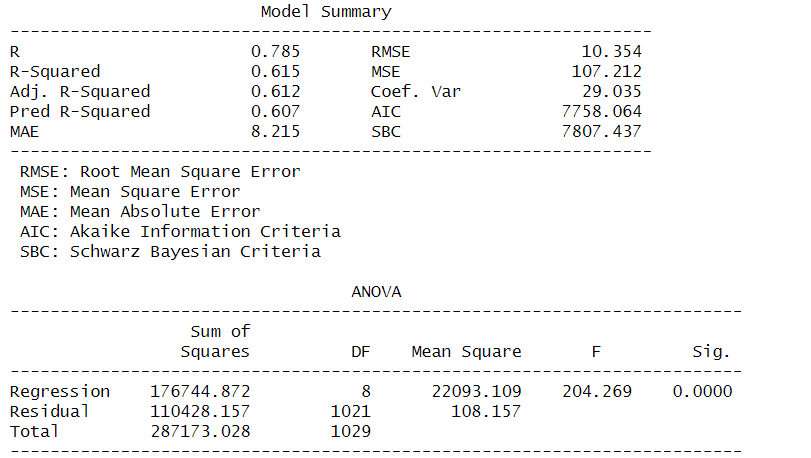
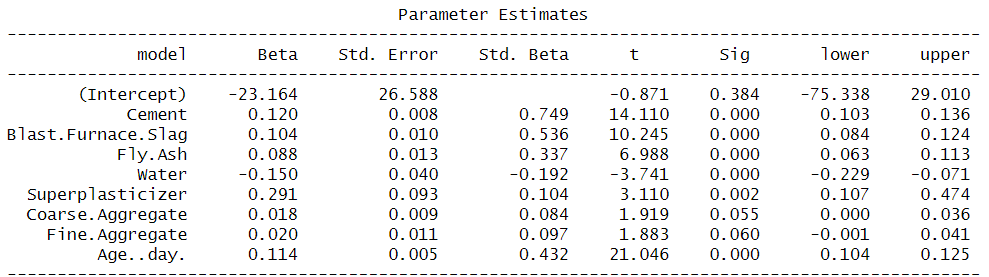
* 1. Top Predictors

**** ****

****

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

* 1. OLS Regression Analysis

**** ****

****

**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.

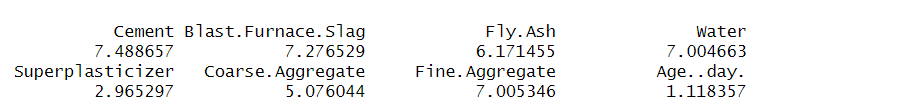
**All are less than 0.05 except coarse and fine aggregates. Therefore, the aggregates do not contribute to compressive strength in this model.**

**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.**

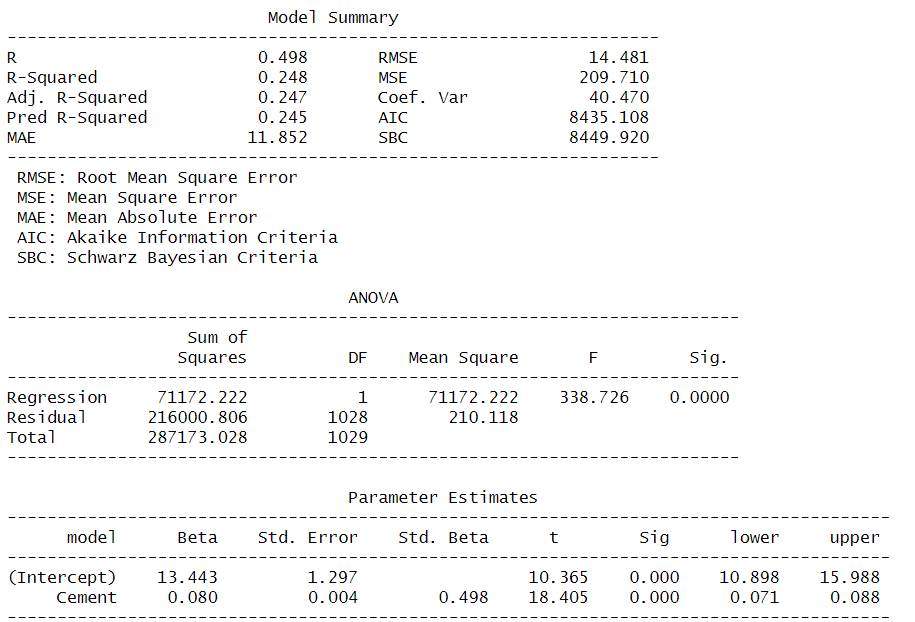
**VIF scores for each feature**

****

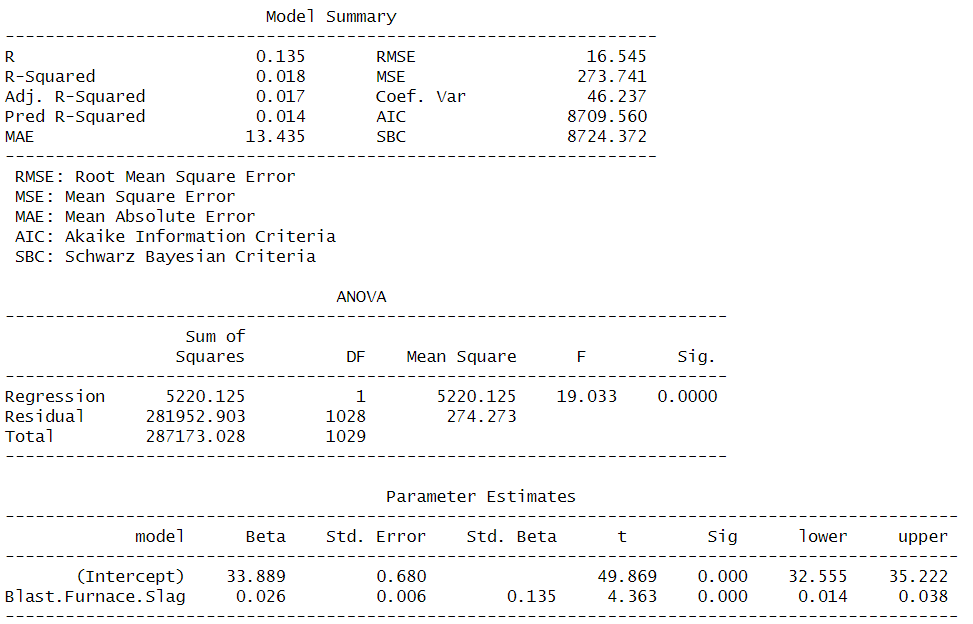
The VIF scores exceeding 5 to 10 indicate collinearity (where 1 is the minimum). The variables Aggregate, Blast Furnace Slag, Water, Fly Ash, and Cement have VIF scores ranging from 5 to 10, indicating potential multicollinearity, particularly if a conservative threshold is applied. Superplasticizer, 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**

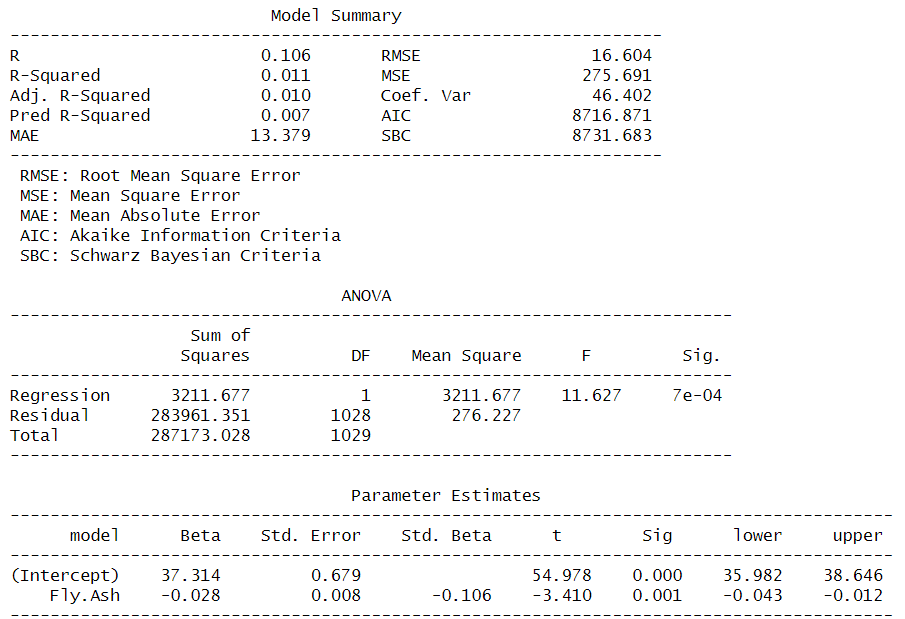
**Cement**

****

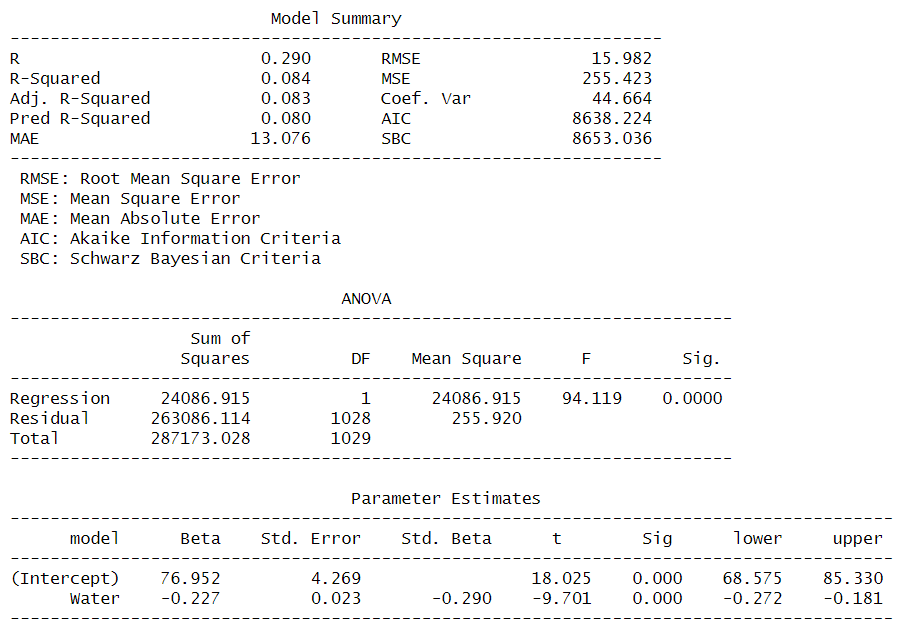
**Blast.Furnace.Slag**

****

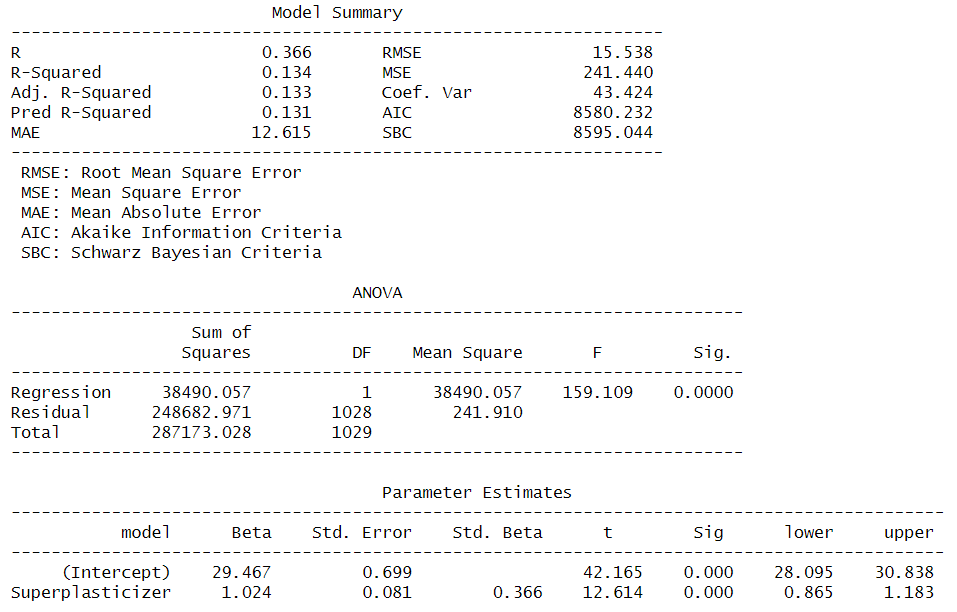
**Fly.Ash**

****

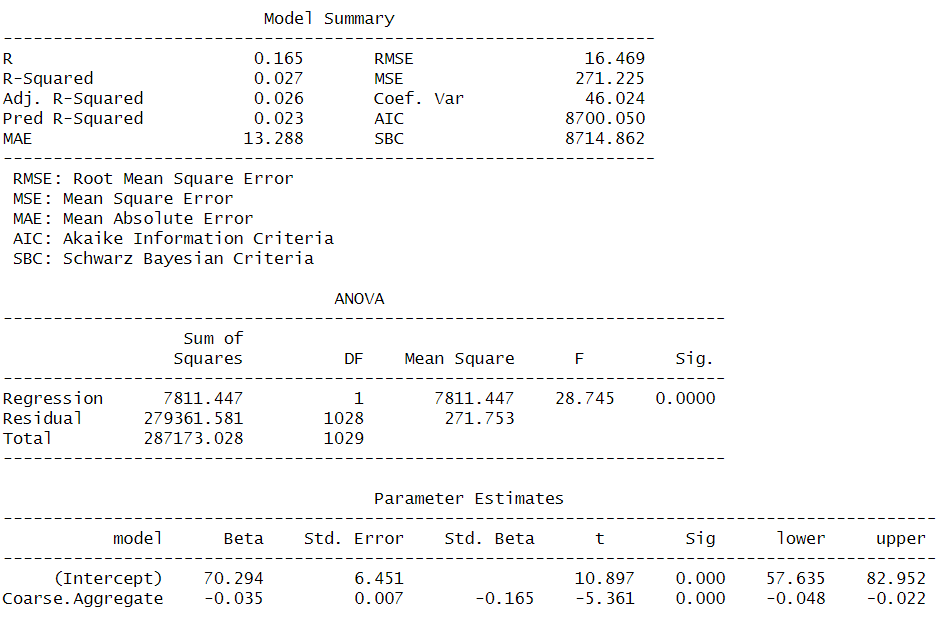
**Water**

****

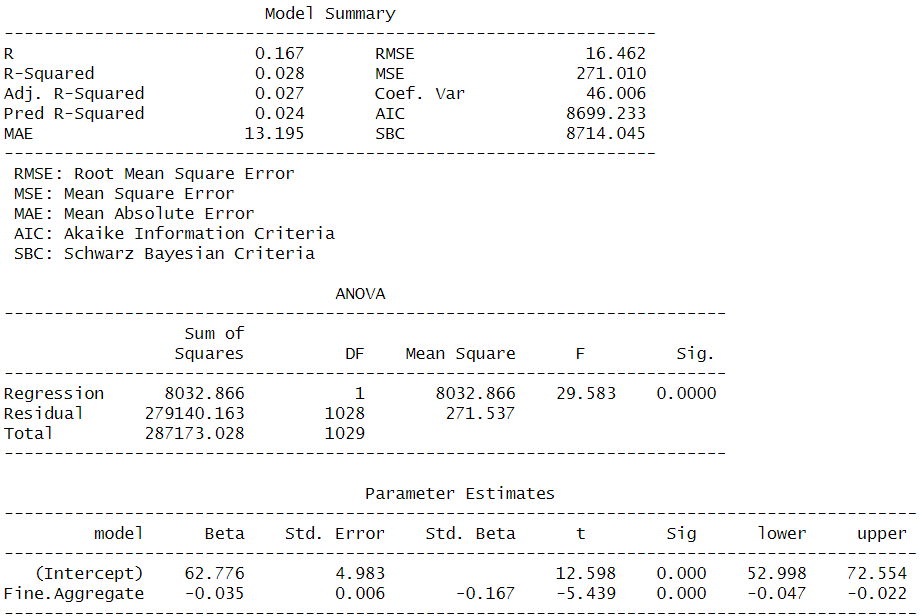
**Superplasticizer**

****

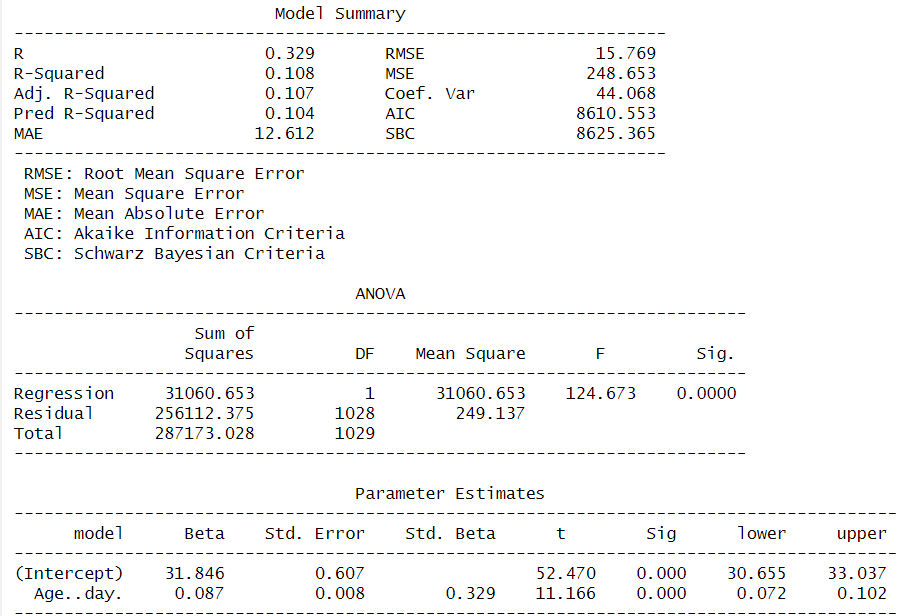
**Coarse.Aggregate**

****

**Fine.Aggregate**

****

**Age..day**

****

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).



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

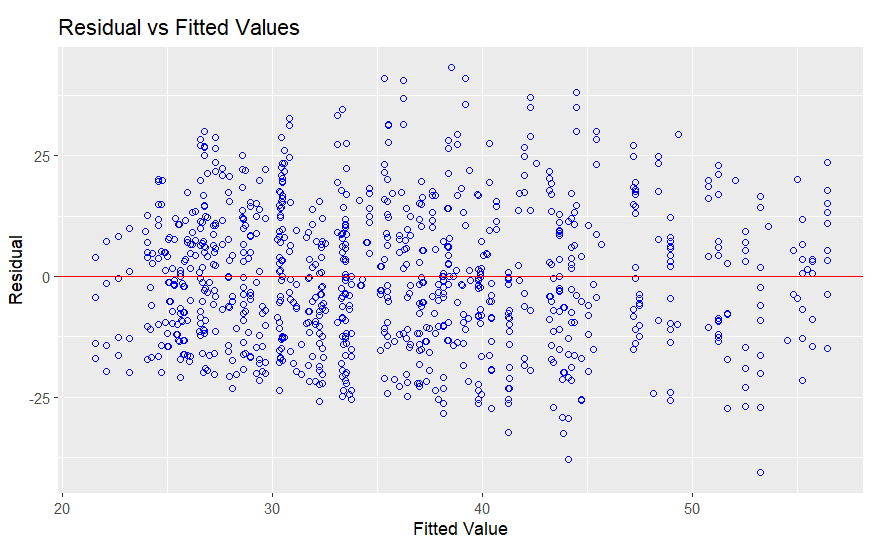
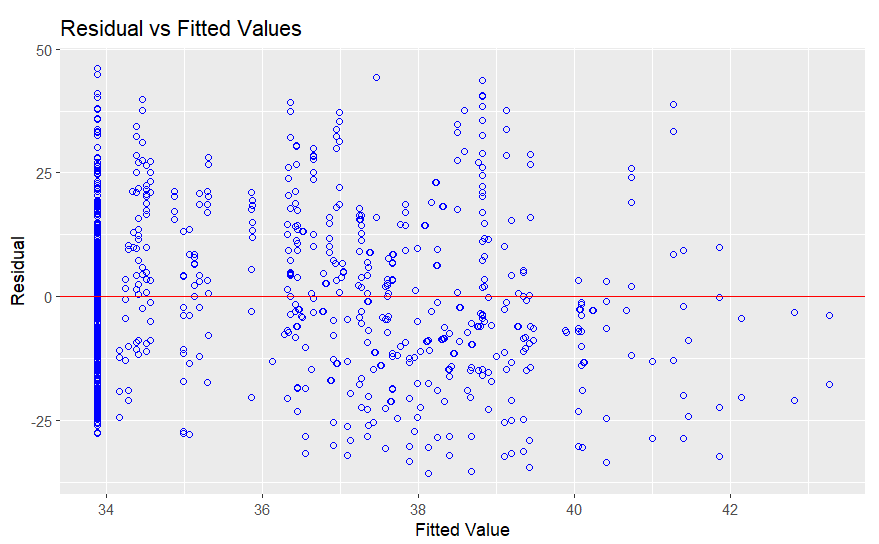
**Q6. Is the relationship linear?**

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.

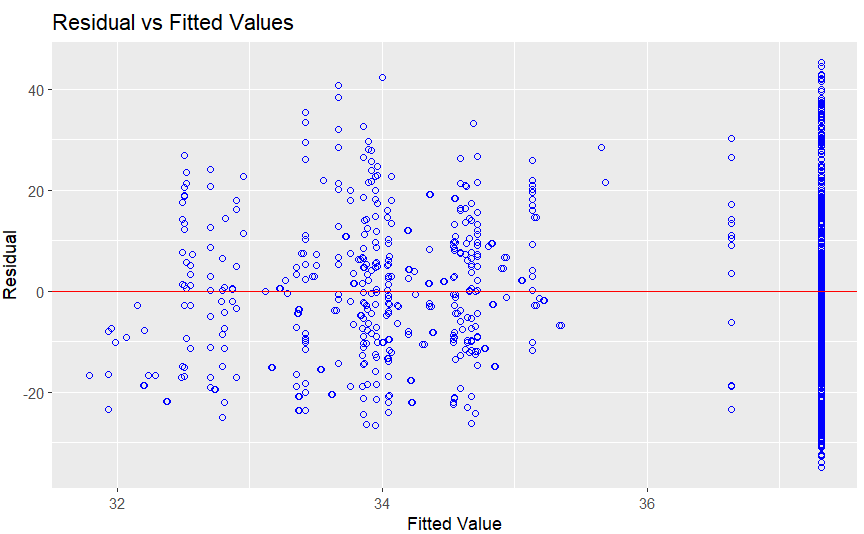
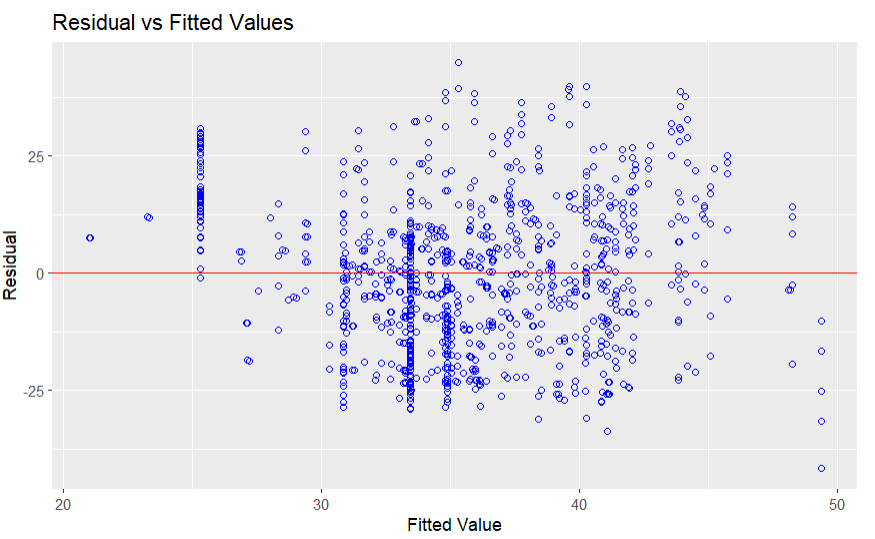
The residual plot with the most non-linear form is for age where for ages 0 to 20, there are negative residuals then the residuals increase from 20 to 100 before decreasing again. Water and fine aggregate have slight non-linear patterns. Transformations of the predictors (e.g., sqrt(X), X^2) could accommodate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

**Residual vs. predicted value plot for each variable**

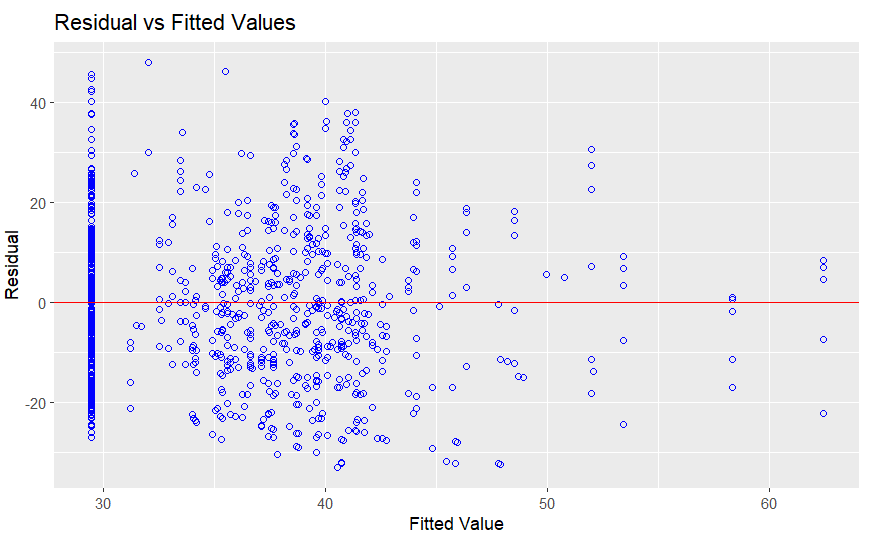
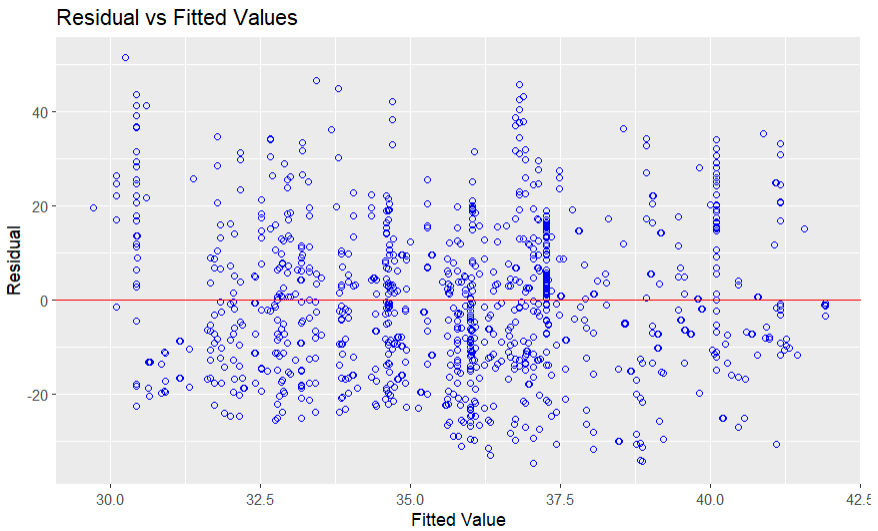
**Cement Blast.Furnace.Slag**

****

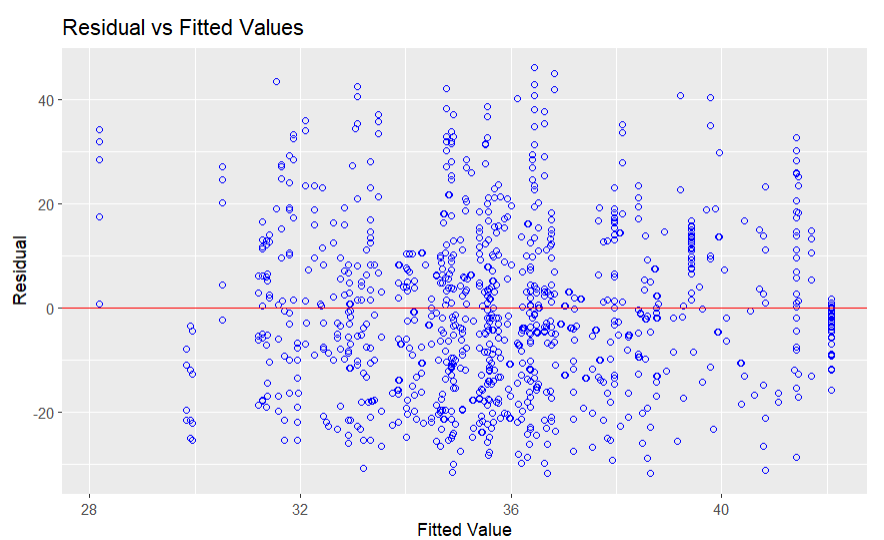
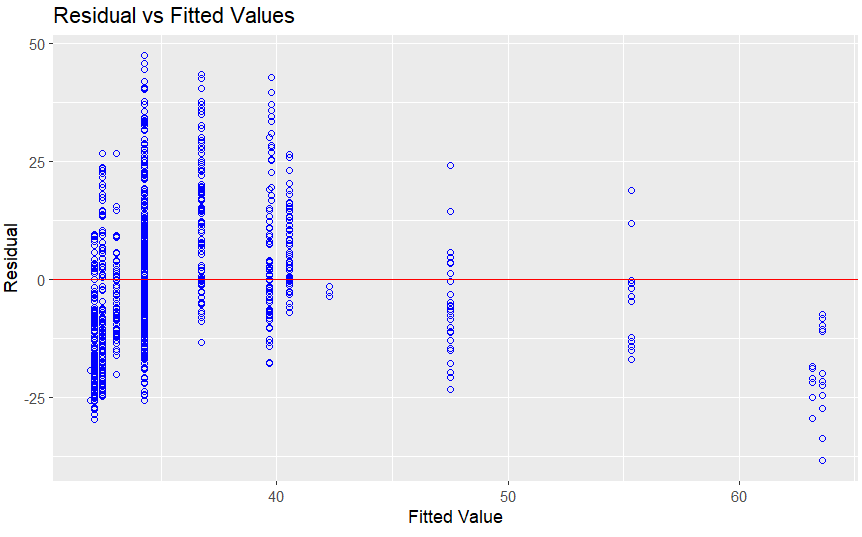
**Fly.Ash Water**

**** ****

**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.

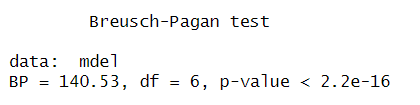
Multiple Linear Regression with 3 predictors combined, is not a good fit for data and it has R^2 value of 48%. The Three combined predictors which give the highest reduction in the uncertainty in predicting response variable is: **Cement+Superplasticizer+Age..day**. So, it will come first in the model.

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.



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

* 1. Feature Engineering with OLS

**Q7. Is there synergy among the predictors?**

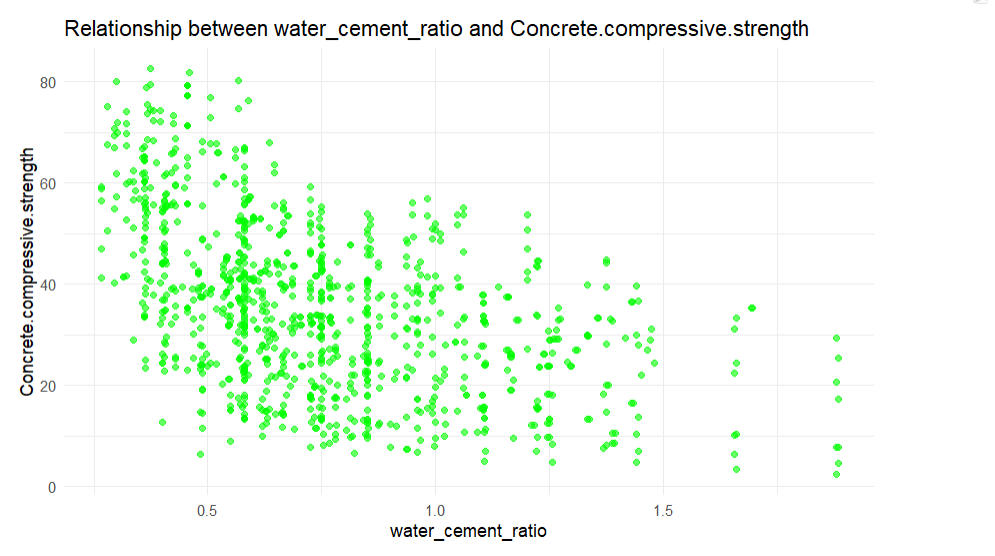
To evaluate the impact of an interaction term that accounts for non-additive relationships, I created a water-to-cement ratio (water:cement) and re-ran an OLS analysis. Including this interaction term resulted in an increase in the R-squared value from 0.615 to 0.618. However, since adding predictors 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.

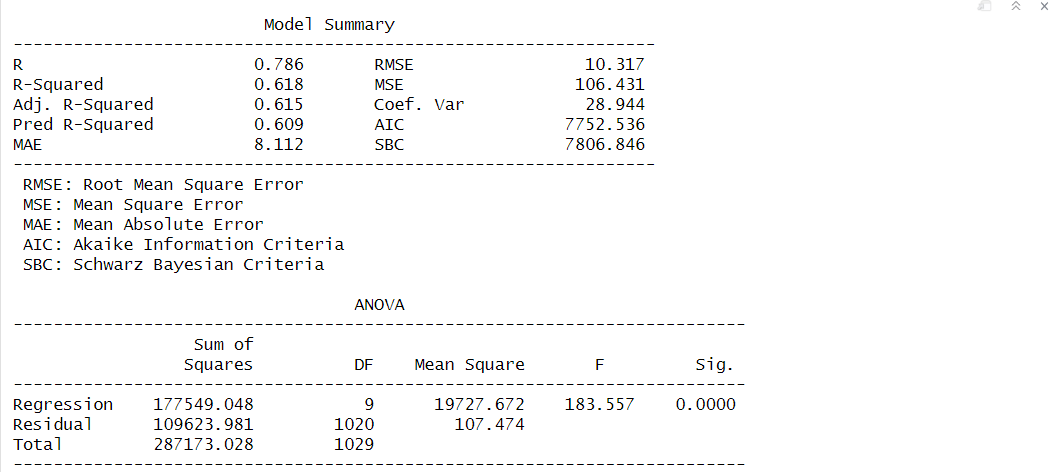
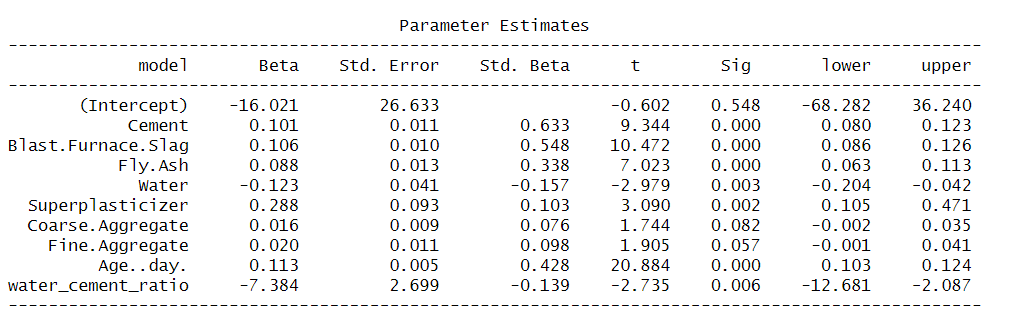
I also tested other interaction terms, including **cement:fine.aggregate**, **cement:coarse.aggregate**, **cement:fine.aggregate:coarse.aggregate**, and **superplasticizer:cement**. However, none of these terms improved the adjusted R-squared, and their p-values for the t-statistics were greater than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

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



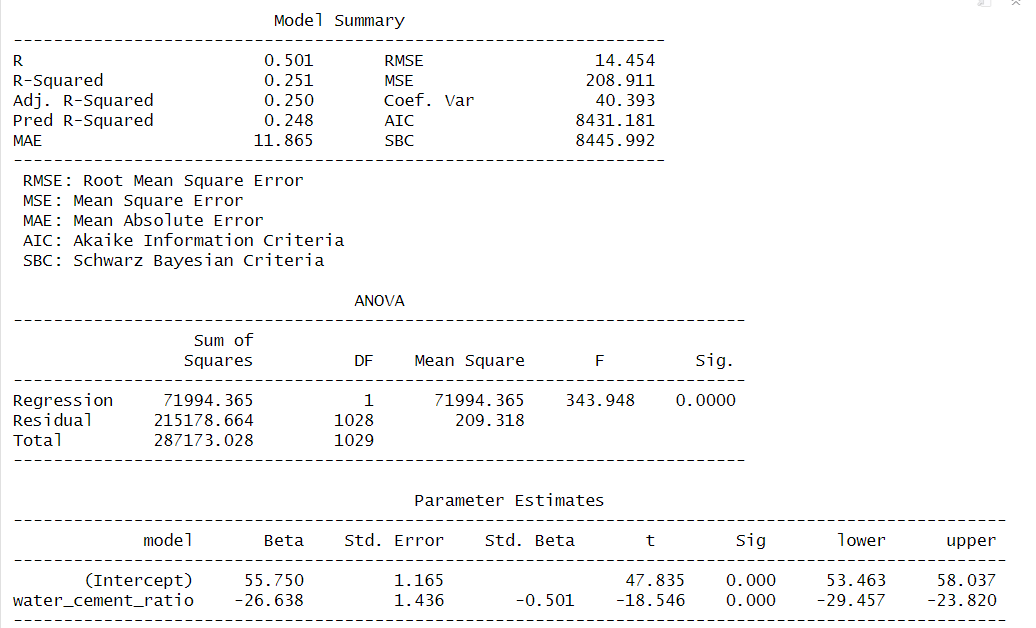
It depicts a negative relationship between **water:cement:ratio** and **compressive strength**

**Generate OLS regression results with water : cement ratio**

**** ****

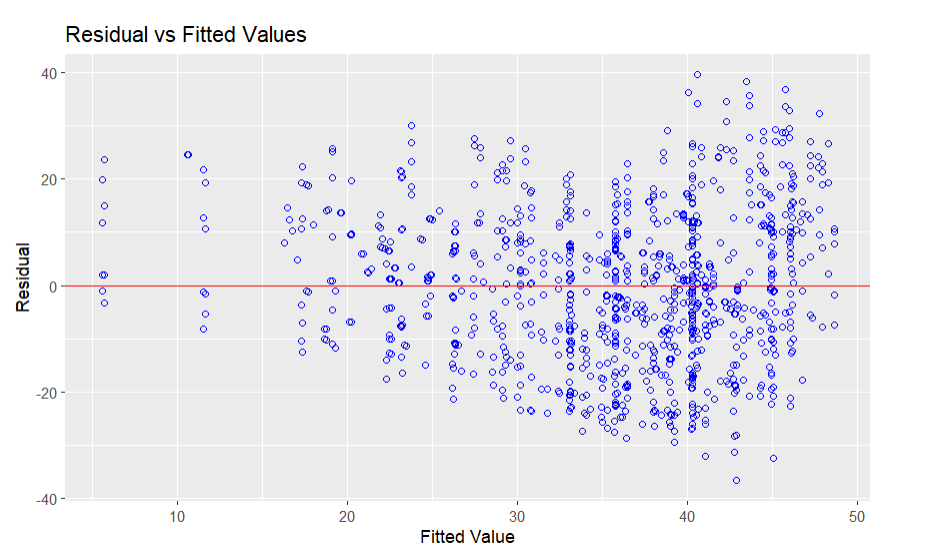
The insertion water:cement ratio in the model provoke an increase of R^2 value from 0.615 to 0.618.

**Generate OLS summary with only water : cement ratio**

****

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.

## 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.