**Linear Models**

(Multiple Linear Regression)

Name: Sharlene Gary Burboz

**Business Problem:**

Concrete is the most widely used material in construction and civil engineering. Our roads, parking lots, buildings all are laid with concrete. The compressive strength of concrete depends on various raw ingredients like cement, water, coarse aggregate, age of cement etc. This project studies the effect of various ingredients on the compressive strength of concrete by using a multiple regression model. Here the aim is to fit a regression model and study the effect of various ingredients of concrete, which act as input variables on the compressive strength of concrete

**Data Characteristics:**

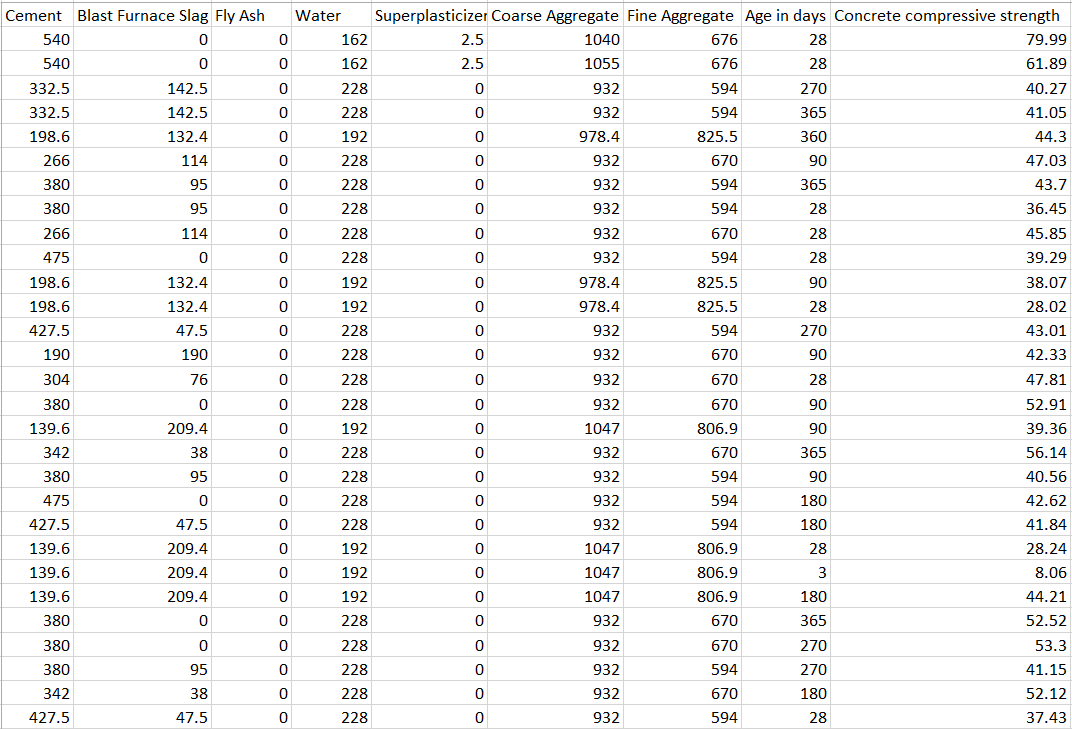
Data from data.world was used for this analysis

There are 1030 observations. There are 9 variables in total. Out of these 9 variables, 8 variables are independent variables and one variable is dependent variable. There are no missing observations.

The input variables include Cement, Blast Furnace Slag, Fly Ash, Water, Age in days, Coarse Aggregate, Super Plasticizer and Fine Aggregate.

The output variable is Concrete Compressive Strength.

**Glimpse of Dataset:**

****

**Analysis:**

The first step involves importing the data and checking for missing values. The following step is to determine the data type of each variable. This is done in R by using the *str(data)* command. After that, descriptive statistics and summary of data is found. The correlation between all variables is found by using cor(data) command. A correlation matrix is obtained. A multiple linear regression model is fitted, taking into account only the variables having a positive correlation with the dependent variable. The next model is fitted, considering all four variables having positive correlations and two variables having negative correlations, namely, Fine Aggregate and Water. The next model is fitted by using all independent variables. The adjusted R square value of all 3 models is found, and the model with the highest value of the adjusted R square is chosen. The next step involves plotting the model and checking for the assumptions of linear regression. A few tests are also done to check for those conditions. We finally conclude whether the model is a good fit or not.

**Interpretation:**

Correlation between all variables is found out. Multicollinearity is absent since no value of correlation matrix is greater than 0.8

3 models of multiple linear regression are fitted. The first model uses the independent variables having positive correlation with the dependent variable. The R square value comes out to be **54%**. The second model uses all four variables having positive correlations and two variables having negative correlations, namely, Fine Aggregate and Water. This model gives Adjusted R square value of **58%** .The next model is fitted by using all independent variables. The Adjusted R square value comes out to be **61%**. Since the value of Adjusted R square value of model 3 is greater, it is a better model than other two models. Hence 61% error can be explained by the model.

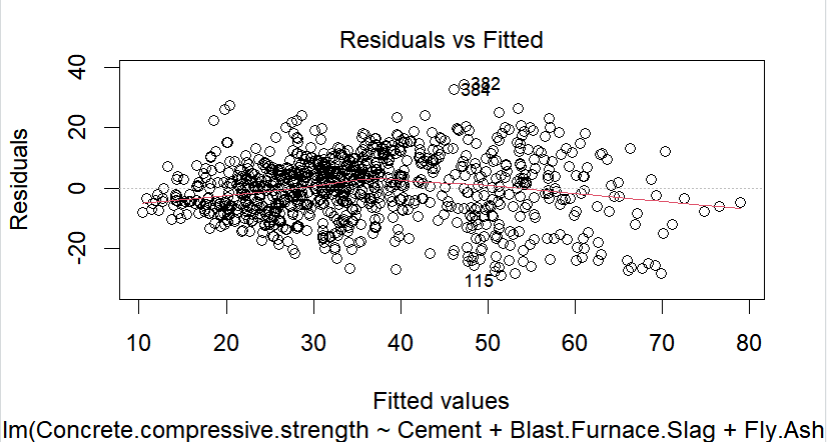
The regression equation obtained is as follows:

*Y= -23.33+ 0.12X1+ 0.10X2+0.09X3 -0.15X4+ 0.29X5+ 0.02X6+0.02X7+0.11X8*

The assumptions of residual analysis are checked by using the *plot(model)* command. This command gives us Residual Plots

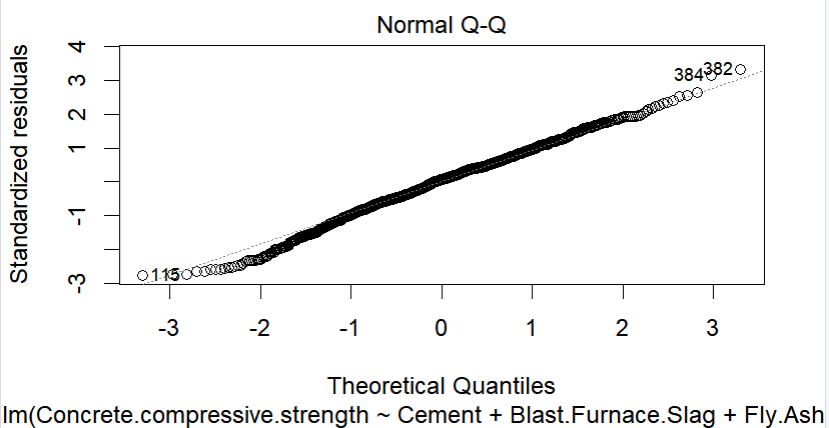
1. **Linearity**

The residual vs fitted plot is used to check linearity. From Residual VS Fitted plot we can observe an almost straight horizontal line with equally spaced residuals. Hence it is safe to assume that linearity assumption is satisfied.

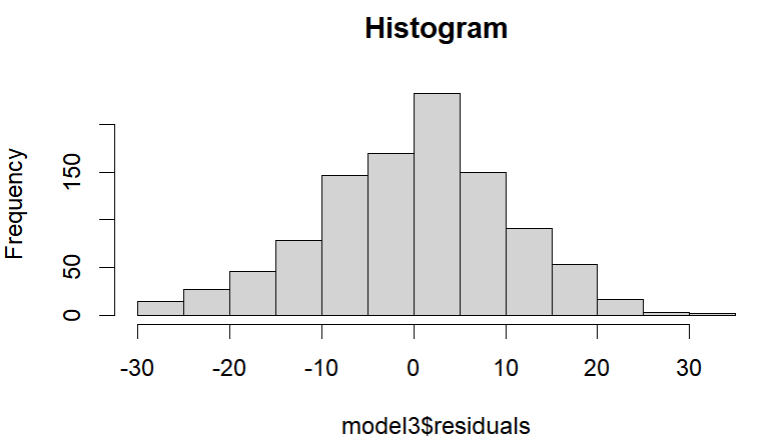


1. **Normality**

Normal QQ plot is used to check the condition of normality. From Normal QQ Plot we can observe that most observations lie on line. Hence normality assumption is satisfied.

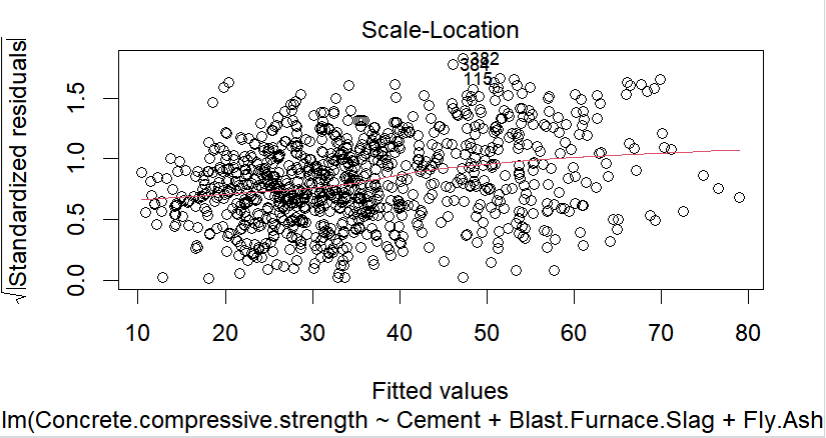


The condition of normality is also checked by plotting a histogram. From histogram it is clear that normality assumption is satisfied.



1. **Homoscedasticity**

Scale location plot is used to determine homoscedasticity. From Scale Location plot we can observe an almost straight horizontal line. However the residuals are not equally spaced. Hence the assumption of homoscedasticity is not satisfied.



Homoscedasticity is also checked by using Breusch Pagan test. The p-value is less than 0.05 indicating that heteroscedasticity is present. A weighted regression model is fitted in order to eliminate heteroscedasticity and Breusch Pagan test is performed again. The p-value comes out to be 0.866 indicating that heteroscedasticity was removed. This indicates that the weighted least squares model is able to explain more of the variance as compared to the multiple linear regression model. Log transformation can also be used in order to eliminate heteroscedasticity.

The regression equation after using weighted least squares is as follows:

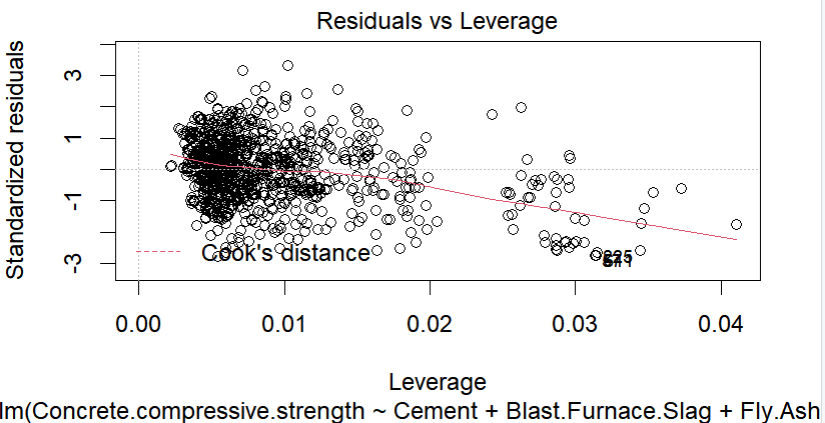
*Y= -32.52+* *0.12X1+* *0.10X2+* *0.08X3+* *-0.11X4+ 0.44X5+* *0.02X6+* *0.02X7+* *0.15X8*

1. **Independence**

Autocorrelation is checked by using Durbin-Watson test. The p-value is 0 indicating there is positive autocorrelation.

**Outliers’ detection**

From Residual VS Leverage plot, we can observe 2 observations 611 and 225 lie beyond Cooke's distance and exceed -3 standard deviation



**Conclusion:** This model can be improved by eliminating autocorrelation and outliers. The Adjusted R square of weighted least squares model is greater than the Adjusted R square of multiple linear regression model indicating that the weighted least squares model offers a better fit to the data compared to the multiple linear regression model.

**R Codes and Output:**

