Concrete Compressive Strength Prediction by Linear Regression Model

Submitted by:

Soumya Paul (211391)

Tanu Denwal (201042)

Koustav Saha (200520)

Under the guidance of:

Sandipan Mitra and Rohan Kumar,

Mentors, Stamatics Project 2022

Stamatics, IIT Kanpur

Abstract

With the increase of population in today's world it needs a construction boom to ensure a home for all the people. Also the global climate change is making our surroundings more disaster prone which results in collapsing of various buildings, important architectures. A recent example is the April 2015 Nepal earthquake which push to death around 8964 people and destroys various century old heritage buildings in Nepal. So it is important to improve quality of materials which are used in construction work. Also the material should be cheap so that everyone can afford. As of now our major building material is concrete which comprises of cement, water , Fine aggregate(sand), coarse aggregate. We should only use those kind of concrete which can bear more load and more resistant to various issue like buckling. Now the question is how we can determine compressive strength of concrete? What are the different parameters that effects on the compressive strength of cement? To answer these questions, this project proposes a well-known model to predict whether we should choose a certain sample of concrete based on the information about other parameters are given. Here we take 8 parameters to determine the compressive strength of a concrete sample. We will divide our dataset into train and test dataset and we will determine whether all those parameters are significant in determining the strength of concrete or not . Also we will see based on those significant parameters our multiple linear regression model can predict the compressive strength of a given concrete sample correctly or not.

Acknowledgement

It is our pleasure to present a project on "Concrete Comprehensive Strength Prediction by Multiple Linear Regression". Every accomplishment has constant encouragement and advice from valuable and noble minds to guide us in putting our efforts in the right direction to bring out the project. We want to express our sincere gratitude to our mentors **Sandipan Mitra and Rohan Kumar** for their constant help and support throughout the completion of the project. Without their valuable guidance and motivation, it was nearly impossible to work on this project as a team and understand the practical aspect of the course "Regression Analysis". Also we are thankful to all faculty members and seniors without whose support at various stages, this project would not have materialized. Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses.

Soumya Paul

Tanu Denwal

Koustav Saha

Contents

1 Introduction **5**

2 Objective **6**

3 Methodology **7**

4 Dataset **8**

4.1 Viewing Dataset in Python 8

5 Exploratory data analysis **9**

5.1 Dimension of Data 9

5.2 Basic Information about data 9

5.2.1 Checking if there is any missing value There is no missing value in the data 9

5.3 Data description 10

5.4 Correlation Matrix 11

5.4.1 Pictorial visualization 13

5.5 Box Plot 13

6 Multiple Linear Regression **14**

6.1 OLS Method 15

6.2 Splitting the dataset into Train and Test dataset 15

6.2.1 Model with all the regressors 16

7 Multicollinearity **17**

7.1 VIF Method 17

7.2 Variance Decomposition Method 17

7.3 Fitting model with subset regressors 19

8 Stepwise Selection Method **20**

8.0.1 Model after performing stepwise selection method 21

9 Model diagnostics **22**

9.1 Checking if heteroscedasticity present or not 22

9.2 Checking if any higher order term is needed 22

9.3 Checking of normality assumption 25

10 Predicted response for test dataset and checking the accuracy of our prediction **26**

1 Introduction

The most widely used building material is Concrete. It is used all over the the world because of its its various properties over building other materials. The Compressive Strength of Concrete determines the quality of Concrete. The concrete compressive strength is highly linear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.This is generally determined by a standard crushing test on a concrete cylinder.

According to this test we need to build small cylinders by applying different raw materials to concrete and check the variation in compressive strength. The recommended wait time for testing the cylinder is 28 days which consumes a lot of time, labour and is also prone to human error.

One way of reducing the wait time and reducing the number of combinations is to use prediction models where we can provide information to the computer and make it to try different combinations to predict compressive strength. This way we can reduce our physical efforts and can also reduce the amount of time for experiment. But, to design such model we have to know the relations between all the raw materials and how one material acts in building the strength.

It is possible to derive mathematical equations and run simulations based on these equations, but we cannot expect the relations to be simple in real-world. Also, these tests have been performed for many numbers of times and we have enough real-world data that can be used for predictions. Here, we are going to analyse Concrete Compressive Strength dataset and build Machine Learning model to predict the compressive strength.

2 Objective

Determining the compressive strength using machine which is trained and tested using multiple physical independent datasets such as cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate and also learning to deal with missing data if present. Concrete Compressive Strength is our dependent variable here. We can use different estimation methods. For our case, we are using the ordinary least squares (OLS) regression statistical method.

Estimation should be unbiased and accurate to prediction.That is why we are using the ordinary least square (OLS) method as it most unlikely to be biased and close to our prediction. Also the model used for the estimation is Multiple linear regression model.

With the development of artificial intelligence (AI) in recent years, it is a trend to use machine learning (ML) techniques to predict the concrete compressive strength. ML is a branch of AI, and can be used for several objectives, e.g., classification, regression, clustering, etc. Predicting concrete compressive strength is just one application of the regression function of ML. Compared with other traditional regression method, ML adopts certain algorithms that can learn from the input data themselves and gives highly accurate results for the output data,

3 Methodology

After creating the group of similar concrete records, regression techniques are applied to identify the relationship between concrete compressive strength with other components of concrete.The prediction of concrete strength using water cement ratio, cement/water ratio and generation of a formulae based on observed results is not only erroneous but grossly understated and misleading. Other parameters which includes physical and chemical properties of the constituents of concrete and their influence on concrete strength cannot be overlooked. Therefore, efforts should be concentrated on models taking into account the influence of different constituents parameters on the concrete strength in order to have more reliable and accurate results for the prediction of concrete strength

4 Dataset

For building up the prediction model, we need a huge amount of the experimental data on concrete compressive strength. For the same we have collected 1030 concrete compressive strength tests. In fact, the data set is used for training and testing individual machine learning algorithms. The ordinary Portland cement, water, coarse aggregate, fine aggregate, super-plasticizer, blast-furnace slag and fly ash were the eight parameters of tested concrete. and it was leaved for some time just before the test under normal condition. Then through a typical compressive test procedure of the cylinder specimens with height 150 mm, the compressive strength of the concrete was obtained. Of course, there are total 9 variables in the experimental data sets. The name, unit, min/max values, mean value and standard derivation (SD) of the experimental variables are mentioned in data description section.

4.1 Viewing Dataset in Python

```
Cement (component 1)(kg in a m^3 mixture)  \
0                                    540.0
1                                    540.0
2                                    332.5
3                                    332.5
4                                    198.6

   Blast Furnace Slag (component 2)(kg in a m^3 mixture)  \
0                                                    0.0
1                                                    0.0
2                                                  142.5
3                                                  142.5
4                                                  132.4

   Fly Ash (component 3)(kg in a m^3 mixture)  \
0                                         0.0
1                                         0.0
2                                         0.0
3                                         0.0
4                                         0.0

   Water  (component 4)(kg in a m^3 mixture)  \
0                                    162.0
1                                    162.0
2                                    228.0
3                                    228.0
4                                    192.0

   Superplasticizer (component 5)(kg in a m^3 mixture)  \
0                                                    2.5
1                                                    2.5
2                                                    0.0
3                                                    0.0
4                                                    0.0

   Coarse Aggregate  (component 6)(kg in a m^3 mixture)  \
0                                                 1040.0
1                                                 1055.0
2                                                  932.0
3                                                  932.0
4                                                  978.4

   Fine Aggregate (component 7)(kg in a m^3 mixture)  Age (day)  \
0                                              676.0          28
1                                              676.0          28
2                                              594.0         270
3                                              594.0         365
4                                              825.5         360

   Concrete compressive strength(MPa, megapascals)
0                                        79.986111
1                                        61.887366
2                                        40.269535
```

```
3                                       41.052780
4                                       44.296075
```

5 Exploratory data analysis

EDA is an important step to build any machine learning project. Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

5.1 Dimension of Data

We have 1030 observation across 9 variables.

5.2 Basic Information about data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1030 entries, 0 to 1029
Data columns (total 9 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   Cement (component 1)(kg in a m^3 mixture)             1030 non-null   float64
 1   Blast Furnace Slag (component 2)(kg in a m^3 mixture) 1030 non-null   float64
 2   Fly Ash (component 3)(kg in a m^3 mixture)            1030 non-null   float64
 3   Water  (component 4)(kg in a m^3 mixture)             1030 non-null   float64
 4   Superplasticizer (component 5)(kg in a m^3 mixture)   1030 non-null   float64
 5   Coarse Aggregate  (component 6)(kg in a m^3 mixture)  1030 non-null   float64
 6   Fine Aggregate (component 7)(kg in a m^3 mixture)     1030 non-null   float64
 7   Age (day)                                             1030 non-null   int64
 8   Concrete compressive strength(MPa, megapascals)       1030 non-null   float64
dtypes: float64(8), int64(1)
memory usage: 72.5 KB
```

There are 8 variables which are of float type and 1 variable which is of integer type.

5.2.1 Checking if there is any missing value There is no missing value in the data

```
Variable name
Cement (component 1)(kg in a m^3 mixture)               0
Blast Furnace Slag (component 2)(kg in a m^3 mixture)   0
Fly Ash (component 3)(kg in a m^3 mixture)              0
Water  (component 4)(kg in a m^3 mixture)               0
Superplasticizer (component 5)(kg in a m^3 mixture)     0
Coarse Aggregate  (component 6)(kg in a m^3 mixture)    0
Fine Aggregate (component 7)(kg in a m^3 mixture)       0
Age (day)                                               0
Concrete compressive strength(MPa, megapascals)         0
dtype: int64
```

So, there is no missing values present in our dataset.

5.3 Data description

Here the number of observations, mean, standard deviation, minimum value, 1st quartile, 2nd quartile (median), 3rd quartile and maximum value is given as count, mean, std, min, '25%', '50%', '75%' and max respectively for each of the variables.

```
 Cement (component 1)(kg in a m^3 mixture)  \
count                                1030.000000
mean                                  281.165631
std                                   104.507142
min                                   102.000000
25%                                   192.375000
50%                                   272.900000
75%                                   350.000000
max                                   540.000000

       Blast Furnace Slag (component 2)(kg in a m^3 mixture)  \
count                                          1030.000000
mean                                             73.895485
std                                              86.279104
min                                               0.000000
25%                                               0.000000
50%                                              22.000000
75%                                             142.950000
max                                             359.400000

       Fly Ash (component 3)(kg in a m^3 mixture)  \
count                                   1030.000000
mean                                      54.187136
std                                       63.996469
min                                        0.000000
25%                                        0.000000
50%                                        0.000000
75%                                      118.270000
max                                      200.100000

       Water  (component 4)(kg in a m^3 mixture)  \
count                                1030.000000
mean                                  181.566359
std                                    21.355567
min                                   121.750000
25%                                   164.900000
50%                                   185.000000
75%                                   192.000000
max                                   247.000000

       Superplasticizer (component 5)(kg in a m^3 mixture)  \
count                                          1030.000000
mean                                              6.203112
std                                               5.973492
min                                               0.000000
25%                                               0.000000
50%                                               6.350000
75%                                              10.160000
max                                              32.200000

       Coarse Aggregate  (component 6)(kg in a m^3 mixture)  \
count                                          1030.000000
mean                                            972.918592
std                                              77.753818
min                                             801.000000
25%                                             932.000000
50%                                             968.000000
75%                                            1029.400000
max                                            1145.000000

       Fine Aggregate (component 7)(kg in a m^3 mixture)    Age (day)  \
count                                          1030.000000  1030.000000
mean                                            773.578883    45.662136
std                                              80.175427    63.169912
min                                             594.000000     1.000000
25%                                             730.950000     7.000000
50%                                             779.510000    28.000000
75%                                             824.000000    56.000000
```

```
max                                          992.600000   365.000000

       Concrete compressive strength(MPa, megapascals)
count                                       1030.000000
mean                                          35.817836
std                                           16.705679
min                                            2.331808
25%                                           23.707115
50%                                           34.442774
75%                                           46.136287
max                                           82.599225
```

5.4 Correlation Matrix

It is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

There are three broad reasons for computing a correlation matrix:

1. To summarize a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables highly correlate with each other.

2. To input into other analyses. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise.

3. As a diagnostic when checking other analyses. For example, with linear regression, a high amount of correlations suggests that the linear regression estimates will be unreliable.

In our case correlation matrix looks like -

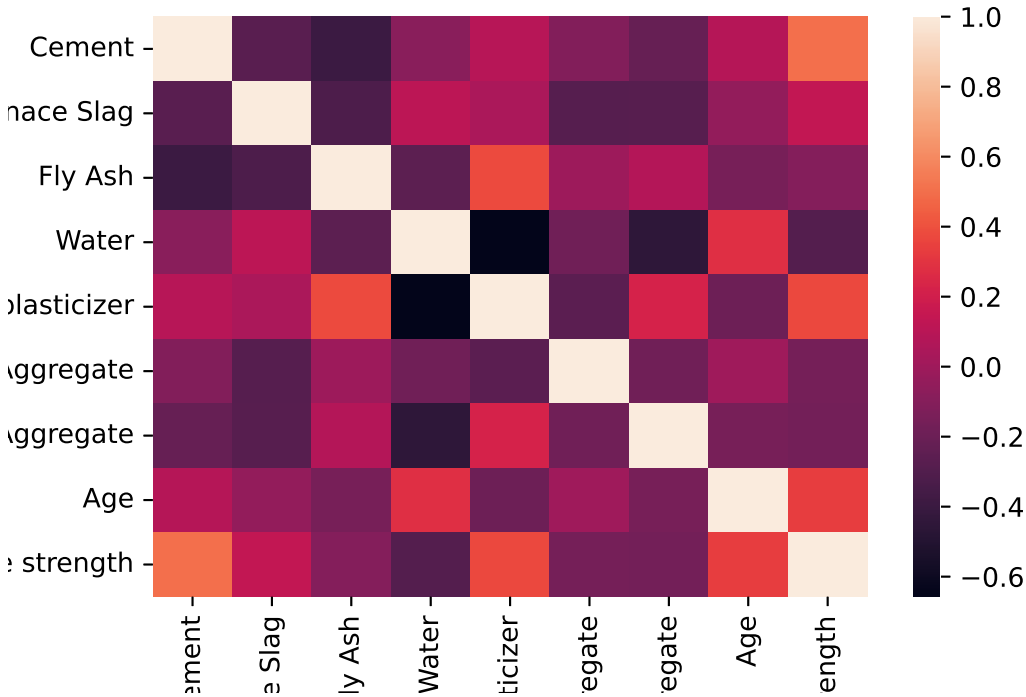
```
                                                      Cement (component 1)(kg in a m^3 mixture)  \
Cement (component 1)(kg in a m^3 mixture)                                                1.000000
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                      -0.275193
Fly Ash (component 3)(kg in a m^3 mixture)                                              -0.397475
Water  (component 4)(kg in a m^3 mixture)                                               -0.081544
Superplasticizer (component 5)(kg in a m^3 mixt...                                       0.092771
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                      -0.109356
Fine Aggregate (component 7)(kg in a m^3 mixture)                                       -0.222720
Age (day)                                                                                0.081947
Concrete compressive strength(MPa, megapascals)                                          0.497833

                                                     Blast Furnace Slag (component 2)(kg in a m^3 mixture)
\
Cement (component 1)(kg in a m^3 mixture)                                                           -0.275193
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                                   1.000000
Fly Ash (component 3)(kg in a m^3 mixture)                                                          -0.323569
Water  (component 4)(kg in a m^3 mixture)                                                            0.107286
Superplasticizer (component 5)(kg in a m^3 mixt...                                                   0.043376
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                                  -0.283998
Fine Aggregate (component 7)(kg in a m^3 mixture)                                                   -0.281593
Age (day)                                                                                           -0.044246
Concrete compressive strength(MPa, megapascals)                                                      0.134824

                                                     Fly Ash (component 3)(kg in a m^3 mixture)  \
Cement (component 1)(kg in a m^3 mixture)                                                -0.397475
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                       -0.323569
Fly Ash (component 3)(kg in a m^3 mixture)                                                1.000000
Water  (component 4)(kg in a m^3 mixture)                                                -0.257044
Superplasticizer (component 5)(kg in a m^3 mixt...                                        0.377340
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                       -0.009977
Fine Aggregate (component 7)(kg in a m^3 mixture)                                         0.079076
Age (day)                                                                                -0.154370
Concrete compressive strength(MPa, megapascals)                                          -0.105753

                                                     Water  (component 4)(kg in a m^3 mixture)  \
```

```
Cement (component 1)(kg in a m^3 mixture)                                   -0.081544
Blast Furnace Slag (component 2)(kg in a m^3 mi...                           0.107286
Fly Ash (component 3)(kg in a m^3 mixture)                                  -0.257044
Water  (component 4)(kg in a m^3 mixture)                                    1.000000
Superplasticizer (component 5)(kg in a m^3 mixt...                          -0.657464
Coarse Aggregate  (component 6)(kg in a m^3 mix...                          -0.182312
Fine Aggregate (component 7)(kg in a m^3 mixture)                           -0.450635
Age (day)                                                                    0.277604
Concrete compressive strength(MPa, megapascals)                             -0.289613

                                                        Superplasticizer (component 5)(kg in a m^3 mixture)
\
Cement (component 1)(kg in a m^3 mixture)                                                           0.092771
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                                  0.043376
Fly Ash (component 3)(kg in a m^3 mixture)                                                          0.377340
Water  (component 4)(kg in a m^3 mixture)                                                          -0.657464
Superplasticizer (component 5)(kg in a m^3 mixt...                                                  1.000000
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                                 -0.266303
Fine Aggregate (component 7)(kg in a m^3 mixture)                                                   0.222501
Age (day)                                                                                          -0.192717
Concrete compressive strength(MPa, megapascals)                                                     0.366102

                                                        Coarse Aggregate  (component 6)(kg in a m^3 mixture)
\
Cement (component 1)(kg in a m^3 mixture)                                                          -0.109356
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                                 -0.283998
Fly Ash (component 3)(kg in a m^3 mixture)                                                         -0.009977
Water  (component 4)(kg in a m^3 mixture)                                                          -0.182312
Superplasticizer (component 5)(kg in a m^3 mixt...                                                 -0.266303
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                                  1.000000
Fine Aggregate (component 7)(kg in a m^3 mixture)                                                  -0.178506
Age (day)                                                                                          -0.003016
Concrete compressive strength(MPa, megapascals)                                                    -0.164928

                                                        Fine Aggregate (component 7)(kg in a m^3 mixture)
\
Cement (component 1)(kg in a m^3 mixture)                                                          -0.222720
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                                 -0.281593
Fly Ash (component 3)(kg in a m^3 mixture)                                                          0.079076
Water  (component 4)(kg in a m^3 mixture)                                                          -0.450635
Superplasticizer (component 5)(kg in a m^3 mixt...                                                  0.222501
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                                 -0.178506
Fine Aggregate (component 7)(kg in a m^3 mixture)                                                   1.000000
Age (day)                                                                                          -0.156094
Concrete compressive strength(MPa, megapascals)                                                    -0.167249

                                                        Age (day)  \
Cement (component 1)(kg in a m^3 mixture)                0.081947
Blast Furnace Slag (component 2)(kg in a m^3 mi...      -0.044246
Fly Ash (component 3)(kg in a m^3 mixture)              -0.154370
Water  (component 4)(kg in a m^3 mixture)                0.277604
Superplasticizer (component 5)(kg in a m^3 mixt...      -0.192717
Coarse Aggregate  (component 6)(kg in a m^3 mix...      -0.003016
Fine Aggregate (component 7)(kg in a m^3 mixture)       -0.156094
Age (day)                                                1.000000
Concrete compressive strength(MPa, megapascals)          0.328877

                                                        Concrete compressive strength(MPa, megapascals)
Cement (component 1)(kg in a m^3 mixture)                                                           0.497833
Blast Furnace Slag (component 2)(kg in a m^3 mi...                                                  0.134824
Fly Ash (component 3)(kg in a m^3 mixture)                                                         -0.105753
Water  (component 4)(kg in a m^3 mixture)                                                          -0.289613
Superplasticizer (component 5)(kg in a m^3 mixt...                                                  0.366102
Coarse Aggregate  (component 6)(kg in a m^3 mix...                                                 -0.164928
Fine Aggregate (component 7)(kg in a m^3 mixture)                                                  -0.167249
Age (day)                                                                                           0.328877
Concrete compressive strength(MPa, megapascals)                                                     1.000000
```

We can say that correlation between 'Concrete compressive strength(MPa, megapascals)' and 'Age (day)' is 0.328877.

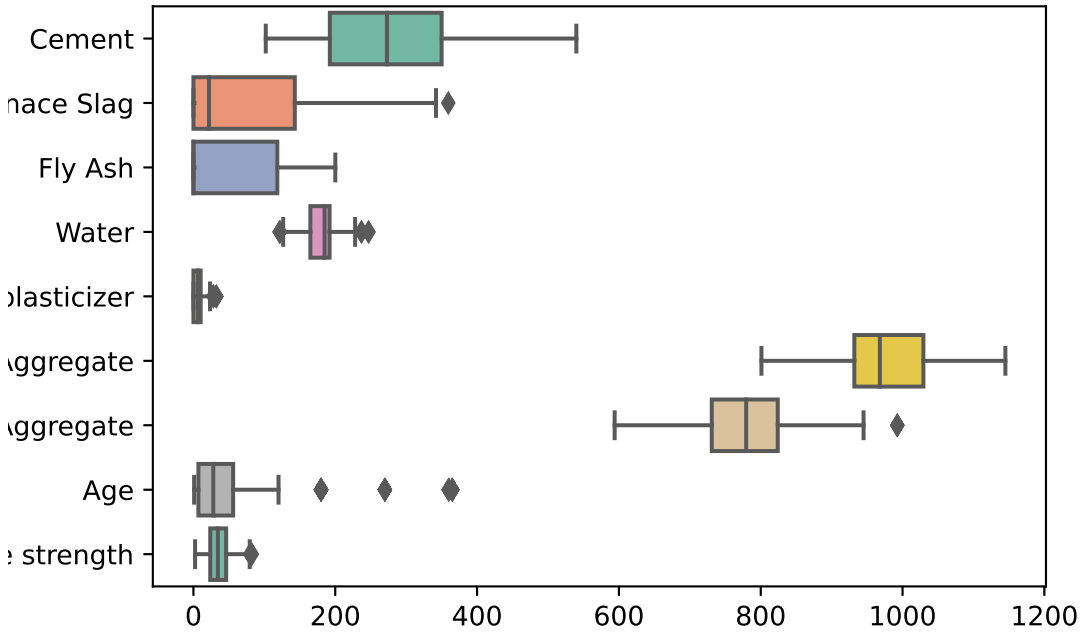
5.4.1 Pictorial visualization

Here is the pictorial visualization of correlation between the variables present in the dataset The more the color lightens, the more the correlation becomes positive. Similarly, The more the color darkens, the more the correlation becomes negative.



5.5 Box Plot

It is a chart that shows data from a five-number summary including one of the measures of central tendency. It does not show the distribution in particular as much as a stem and leaf plot or histogram does. But it is primarily used to indicate a distribution is skewed or not and if there are potential unusual observations (also called outliers) present in the data set. **Box plots** are also very beneficial when large numbers of data sets are involved or compared.



We can see that outilers are present for the variables Blast Furnace Slag (component 2)(kg in a m^3 mixture), Water (component 4)(kg in a m^3 mixture), Superplasticizer (component 5)(kg in a m^3 mixture), Coarse Aggregate (component 6)(kg in a m^3 mixture), Age (day), Concrete compressive strength(MPa, megapascals).

There are some variables like Blast Furnace Slag (component 2)(kg in a m^3 mixture) and Fly Ash (component 3)(kg in a m^3 mixture) which are skewed.

6 Multiple Linear Regression

Multiple linear regression equation looks like:-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon .....(2)$$

(Here , X_i= i th independent variable ,Y=Value of dependent variable, β_i=Coefficient of the ith explanatory variable,ϵ=Random error)

Assumptions:

1. Error has zero mean.
2. Error has constant variance.
3. Errors are uncorrelated.
4. Errors are normally distributed.

Now in our case we have taken 8 independent variables as our predictors and with the help of them we are going to predict the compressive strength of cement. Here our predictors are

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

6.1 OLS Method

Ordinary Least Squares (OLS) regressions are also often just called regressions.It is mostly the preferred method for predicting a model.It is based on minimizing the random part or sum of squared error of our dataset.

Now there are some assumptions about ordinary least square regression.

Assumptions:

1. A linear relationship is assumed between the dependent variable and the independent variables.
2. Regression residuals must be normally distributed and the mean be 0.
3. The residuals are homoscedastic and approximately rectangular-shaped.
4. Absence of Multicollinearity is expected in the model, meaning that independent variables are not too highly correlated.
5. No Autocorrelation of the residuals.

6.2 Splitting the dataset into Train and Test dataset

We have split our dataset into train and test data such that 80% of the observations fall into the train dataset randomly. We will perform our analysis with train dataset only and check how our final model will perform for test dataset.

6.2.1 Model with all the regressors

Now, let's first fit our model with all the regressors and we will see how our analysis deals with this assumptions.

```
OLS Regression Results
==============================================================================================
Dep. Variable:     Concrete compressive strength(MPa, megapascals)    R-squared:                       0.634
Model:                                                         OLS    Adj. R-squared:                  0.630
Method:                                              Least Squares    F-statistic:                     176.2
Date:                                             Sun, 22 May 2022    Prob (F-statistic):          5.33e-172
Time:                                                     16:37:20    Log-Likelihood:                -3080.6
No. Observations:                                              824    AIC:                             6179.
Df Residuals:                                                  815    BIC:                             6222.
Df Model:                                                        8
Covariance Type:                                         nonrobust
==================================================================================================================================
                                                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------------------------
const                                                  -26.4498     28.568     -0.926      0.355     -82.526      29.627
Cement (component 1)(kg in a m^3 mixture)                0.1242      0.009     13.487      0.000       0.106       0.142
Blast Furnace Slag (component 2)(kg in a m^3 mixture)    0.1076      0.011      9.823      0.000       0.086       0.129
Fly Ash (component 3)(kg in a m^3 mixture)               0.0913      0.014      6.670      0.000       0.064       0.118
Water  (component 4)(kg in a m^3 mixture)               -0.1384      0.043     -3.235      0.001      -0.222      -0.054
Superplasticizer (component 5)(kg in a m^3 mixture)      0.3605      0.104      3.468      0.001       0.156       0.565
Coarse Aggregate  (component 6)(kg in a m^3 mixture)     0.0171      0.010      1.690      0.091      -0.003       0.037
Fine Aggregate (component 7)(kg in a m^3 mixture)        0.0198      0.012      1.699      0.090      -0.003       0.043
Age (day)                                                0.1201      0.006     19.601      0.000       0.108       0.132
==============================================================================
Omnibus:                        9.850   Durbin-Watson:                   1.917
Prob(Omnibus):                  0.007   Jarque-Bera (JB):                9.830
Skew:                          -0.256   Prob(JB):                      0.00733
Kurtosis:                       3.156   Cond. No.                     1.04e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model interpretation:-

So we have started with 8 regressors.We fitted our data with Ordinary least square(OLS) method.Now from our python output we get our value of $R^2$is 0.634 and adjusted value of $R^2$is 0.630.Now from these values we can conclude that our model can explain 63.4% variance of actual response accurately. As the value of adjusted $R^2$is moderately high so we can say that our model is moderately good.

7 Multicollinearity

Multicollinearity occurs when independent variables in regression model are correlated. Now if the degree of correlation between variables is high enough,it can cause problems when we fit this model and interpret the result. It generates high variance in coefficient estimates corresponding to interrelated explanatory variables and they do not reflect the actual scenario. Also our model can be highly sensitive to the small changes in dataset.

So there are several test to detect whether multicollinearity arises in our dataset or not.The tests which we have used to test for our model is :-

1. VIF method
2. Variance decomposition method

7.1 VIF Method

It is a way to measure the ammount of multicollinearity in a set of multiple regression variables.Mathematically speaking it is the ratio of overall model variance to the variance of a single independent variable.

Now using VIF method in our model we get:

	variable	VIF
0	Cement (component 1)(kg in a m^3 mixture)	7.285378
1	Blast Furnace Slag (component 2)(kg in a m^3 m...	6.892801
2	Fly Ash (component 3)(kg in a m^3 mixture)	6.084812
3	Water (component 4)(kg in a m^3 mixture)	6.327040
4	Superplasticizer (component 5)(kg in a m^3 mix...	2.851178
5	Coarse Aggregate (component 6)(kg in a m^3 mi...	4.833484
6	Fine Aggregate (component 7)(kg in a m^3 mixture)	6.558222
7	Age (day)	1.083763

As we can see there are some variables where VIF value is > 5. So, as a thumb rule, we can say that there are a set of variables which are involved in multicollinearity.

7.2 Variance Decomposition Method

From R with got this table and the interpretation is given below.

```
  Eigenvalues      CI Cement ‘ Blast Furnace Slag‘ ‘Fly Ash‘
1      5.6752  1.0000 0.0018                0.0043    0.0034
2      0.8799  2.5396 0.0006                0.0630    0.1167
3      0.6852  2.8778 0.0003                0.2137    0.0098
4      0.3937  3.7967 0.0354                0.1468    0.0578
5      0.2990  4.3567 0.0195                0.0541    0.1981
6      0.0505 10.6045 0.6790                0.2486    0.4017
7      0.0090 25.1470 0.1258                0.2049    0.0940
8      0.0075 27.4224 0.1376                0.0646    0.1185
Water Superplasticizer ‘Coarse Aggregate‘ ‘Fine Aggregate‘    Age
  0.0004          0.0036             0.0003           0.0004 0.0075
 0.0001           0.0344             0.0000           0.0000 0.1381
 0.0000           0.0130             0.0000           0.0000 0.3973
 0.0016           0.0495             0.0016           0.0015 0.3904
 0.0018          0.3259             0.0008           0.0003 0.0261
 0.0040           0.1067             0.0164           0.0649 0.0038
 0.9806           0.2838             0.2168           0.1214 0.0357
 0.0115           0.1832             0.7640           0.8116 0.0011
================================
Row 6==> Cement, proportion 0.678991 >= 0.50
Row 7==> Water, proportion 0.980622 >= 0.50
Row 8==> ‘Coarse Aggregate‘, proportion 0.764035 >= 0.50
Row 8==> ‘Fine Aggregate‘, proportion 0.811591 >= 0.50
```

As condition index is > 15 for eigenvalues 0.0090 and 0.0075 and there are two variables - “Coarse Aggregate (component 6)(kg in a m^3 mixture)” and “Fine Aggregate (component 7)(kg in a m^3 mixture)” for which the proportions of variances of the estimators of those two variables contributed by eigenvalue 0.0078 are > 0.5. So, we will drop either of those two variables and go for VIF method again. Here we have dropped the variable “Fine Aggregate (component 7)(kg in a m^3 mixture)”.

Again using VIF Method, here all the VIF’s are < 5. So, we can say that now our regressors are almost free from multicollinearity.

```
                                          variable       VIF
0          Cement (component 1)(kg in a m^3 mixture)  1.926272
1  Blast Furnace Slag (component 2)(kg in a m^3 m...  1.938587
2         Fly Ash (component 3)(kg in a m^3 mixture)  2.402023
3          Water  (component 4)(kg in a m^3 mixture)  2.269540
4  Superplasticizer (component 5)(kg in a m^3 mix...  2.808492
5  Coarse Aggregate  (component 6)(kg in a m^3 mi...  1.596679
6                                          Age (day)  1.081899
```

7.3 Fitting model with subset regressors

```
OLS Regression Results
==============================================================================================
Dep. Variable:     Concrete compressive strength(MPa, megapascals)    R-squared:                       0.632
Model:                                                         OLS    Adj. R-squared:                  0.629
Method:                                              Least Squares    F-statistic:                     200.5
Date:                                             Tue, 24 May 2022    Prob (F-statistic):          1.53e-172
Time:                                                     11:31:06    Log-Likelihood:                -3082.1
No. Observations:                                              824    AIC:                             6180.
Df Residuals:                                                  816    BIC:                             6218.
Df Model:                                                        7
Covariance Type:                                         nonrobust
==============================================================================================================
                                                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                                   19.2458      9.661      1.992      0.047       0.282      38.210
Cement (component 1)(kg in a m^3 mixture)                0.1108      0.005     23.367      0.000       0.101       0.120
Blast Furnace Slag (component 2)(kg in a m^3 mixture)    0.0919      0.006     15.787      0.000       0.080       0.103
Fly Ash (component 3)(kg in a m^3 mixture)               0.0732      0.009      8.503      0.000       0.056       0.090
Water  (component 4)(kg in a m^3 mixture)               -0.1967      0.026     -7.664      0.000      -0.247      -0.146
Superplasticizer (component 5)(kg in a m^3 mixture)      0.3389      0.103      3.281      0.001       0.136       0.542
Coarse Aggregate  (component 6)(kg in a m^3 mixture)     0.0030      0.006      0.520      0.603      -0.008       0.014
Age (day)                                                0.1197      0.006     19.525      0.000       0.108       0.132
==============================================================================
Omnibus:                        9.112   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.011   Jarque-Bera (JB):                9.076
Skew:                          -0.251   Prob(JB):                       0.0107
Kurtosis:                       3.109   Cond. No.                     2.81e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.81e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model interpretation:-

- Adjusted R^2value is 0.629.So after applying variance decomposition method our model can explain 62.9% of actual data accurately.As this value is moderately high so our model is moderately good.
- Now here we can observe for Coarse Aggregate (component 6) p>|t| value comes out to be 0.603 which is significantly higher than 0.05. So we can conclude Coarse Aggregate (component 6) is insignificant to explain our model.

8 Stepwise Selection Method

It is a method which iteratively examines the statistical significance of each independent variable in our linear regression model. There are three techniques - 'forward', 'backward' and 'both'. The forward selection approach starts with no regressors and adds each new variable incrementally, testing for statistical significance. The backward elimination method begins with a full model loaded with all the regressors and then removes one variable each time to test its importance relative to overall results. The main purpose of using stepwise selection method is to reduce the cost of the experiment by reducing the number of independent variables and only taking account of those variables which are most significant, keeping in mind the principal of parsimony. Now to increase the accuracy of our model, we have used stepwise selection method.

```
      feature_idx              cv_scores avg_score  \
1             (0,)  [0.24656017904433702]   0.24656
2           (0, 4)    [0.360860352644516]   0.36086
3        (0, 4, 6)    [0.494192378370946]  0.494192
4     (0, 1, 4, 6)   [0.5711713002261328]  0.571171
5  (0, 1, 3, 4, 6)   [0.5997981317647461]  0.599798

                                       feature_names
      (Cement (component 1)(kg in a m^3 mixture),)
 (Cement (component 1)(kg in a m^3 mixture), Su...
 (Cement (component 1)(kg in a m^3 mixture), Su...
 (Cement (component 1)(kg in a m^3 mixture), Bl...
 (Cement (component 1)(kg in a m^3 mixture), Bl...
```

Thus we have Cement (component 1)(kg in a m^3 mixture), Blast Furnace Slag (component 2)(kg in a m^3 mixture), Water (component 4)(kg in a m^3 mixture), Superplasticizer (component 5)(kg in a m^3 mixture) and Age (day) as our regressors.

8.0.1 Model after performing stepwise selection method

```
OLS Regression Results
==============================================================================================
Dep. Variable:     Concrete compressive strength(MPa, megapascals)    R-squared:                       0.600
Model:                                                         OLS    Adj. R-squared:                  0.597
Method:                                              Least Squares    F-statistic:                     245.2
Date:                                             Tue, 24 May 2022    Prob (F-statistic):          6.25e-160
Time:                                                     11:37:15    Log-Likelihood:                -3117.1
No. Observations:                                              824    AIC:                             6246.
Df Residuals:                                                  818    BIC:                             6274.
Df Model:                                                        5
Covariance Type:                                         nonrobust
==============================================================================================================
                                                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
const                                                   30.5613      4.707      6.492      0.000      21.322      39.801
Cement (component 1)(kg in a m^3 mixture)                0.0854      0.004     22.961      0.000       0.078       0.093
Blast Furnace Slag (component 2)(kg in a m^3 mixture)    0.0627      0.005     13.621      0.000       0.054       0.072
Water  (component 4)(kg in a m^3 mixture)               -0.1817      0.024     -7.649      0.000      -0.228      -0.135
Superplasticizer (component 5)(kg in a m^3 mixture)      0.7290      0.084      8.634      0.000       0.563       0.895
Age (day)                                                0.1140      0.006     17.960      0.000       0.102       0.126
==============================================================================
Omnibus:                        3.545   Durbin-Watson:                   1.938
Prob(Omnibus):                  0.170   Jarque-Bera (JB):                3.447
Skew:                          -0.157   Prob(JB):                        0.178
Kurtosis:                       3.043   Cond. No.                     4.48e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.48e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Clearly our final model is

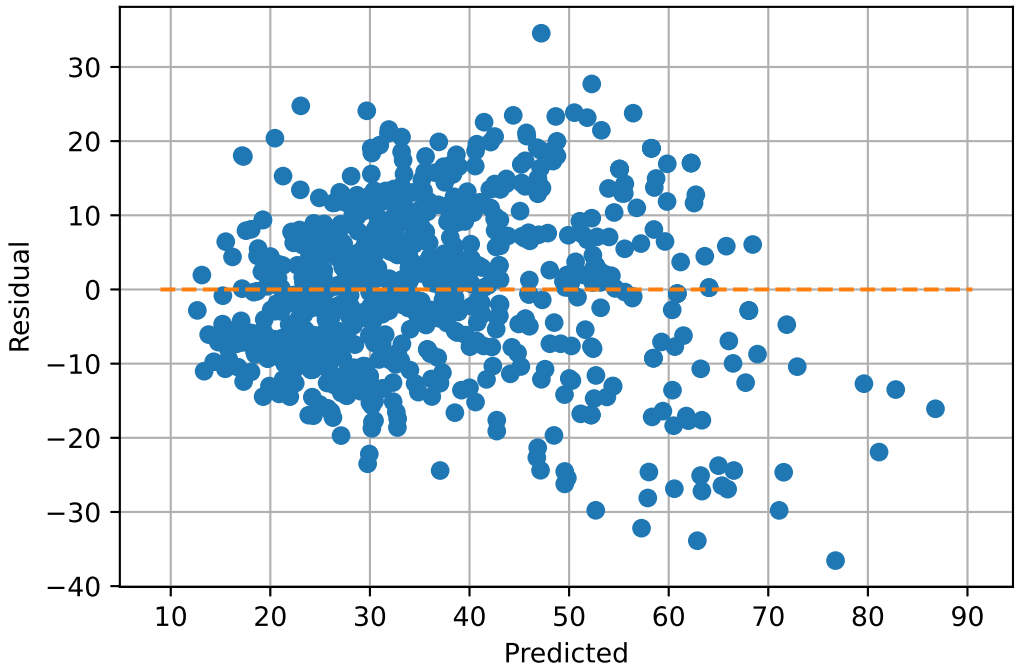
$$\text{Concrete compressive strength(MPa,megapascals)} = 30.5613 + 0.0854 \times \text{Cement(component1)(kg in a } m^3\text{mixture)}$$
$$+0.0627 \times \text{Blast Furnace Slag(component 2)(kg in a } m^3\text{mixture)} - 0.1817 \times \text{Water(component 4)(kg in a } m^3\text{mixture)}$$
$$+0.7290 \times \text{Superplasticizer(component 5)(kg in a } m^3\text{mixture)} + 0.1140 \times \text{Age(day)}$$

This model can explain 60% of the total variation present in response variable.

9 Model diagnostics

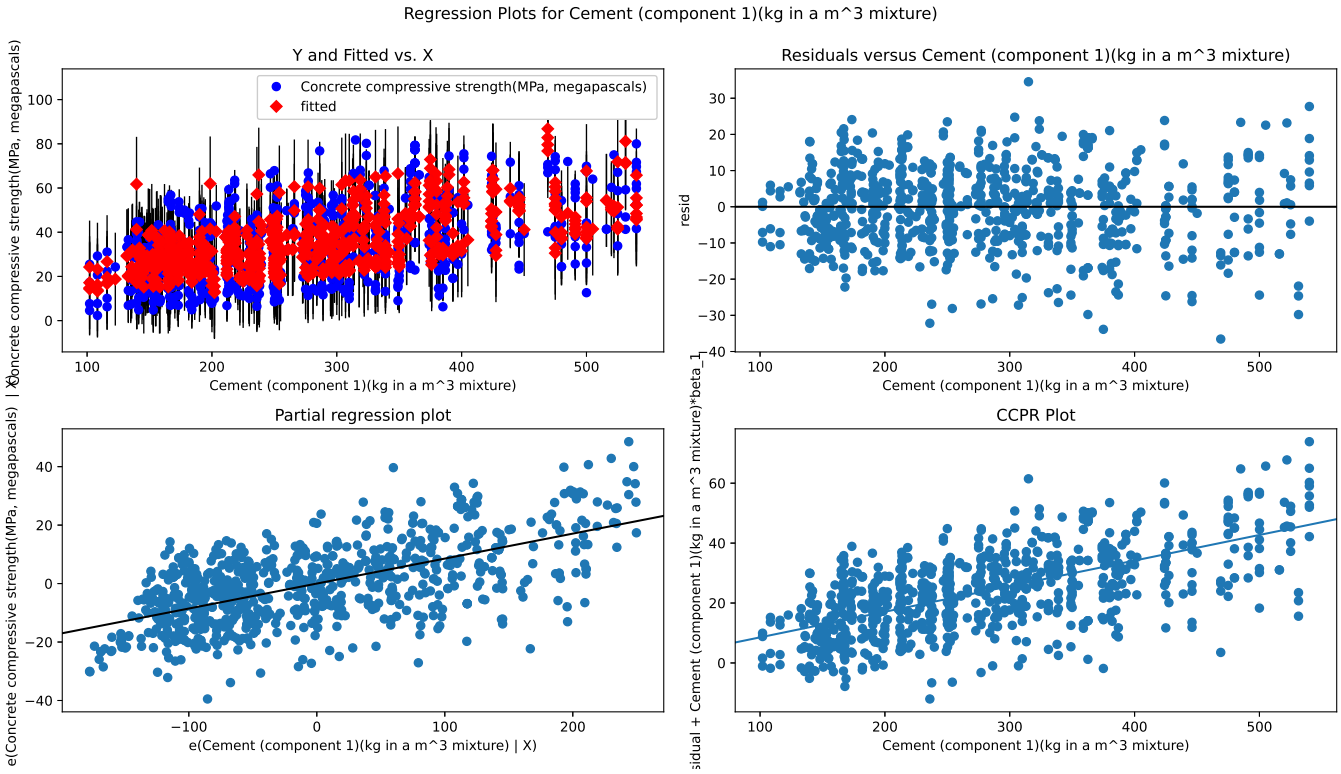
9.1 Checking if heteroscedasticity present or not

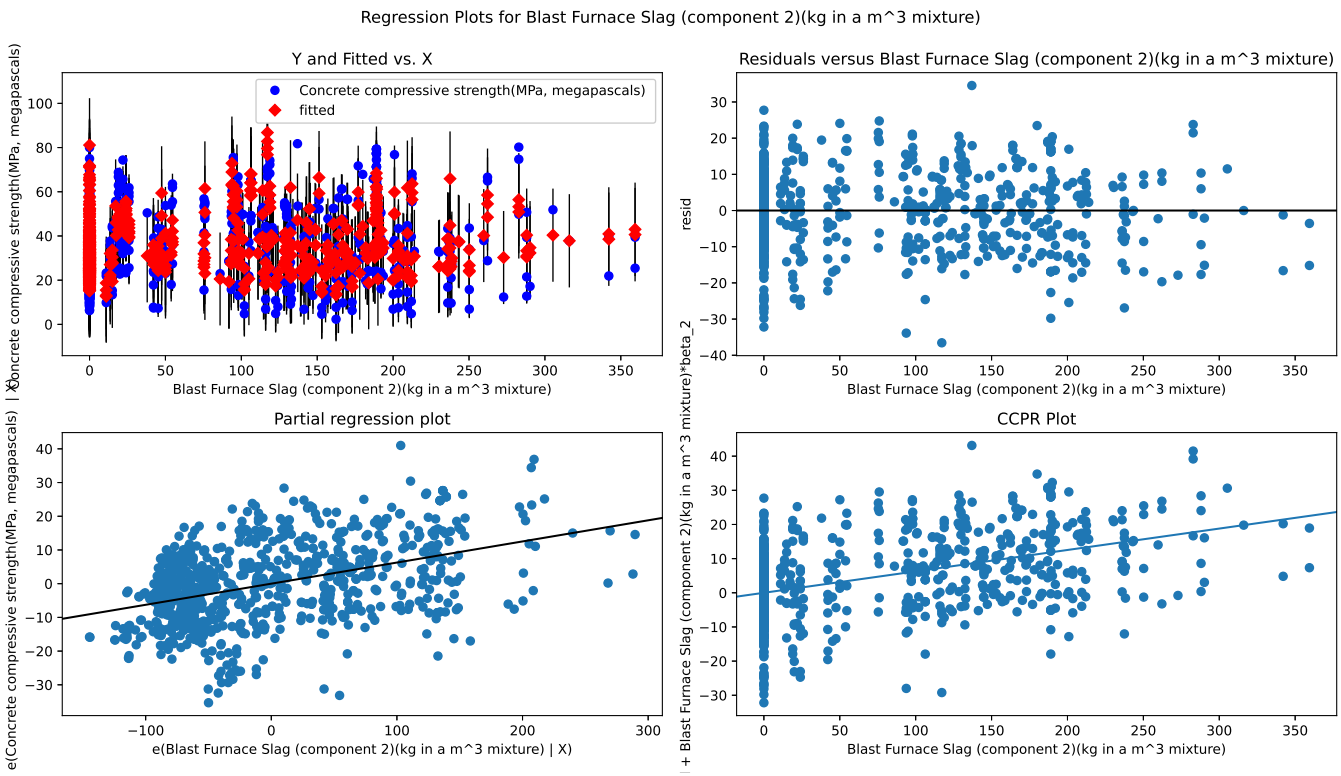
AS the predicted values increase, the residuals doesn't vary significantly in our case. So, we can say that errors are homoscedastic.

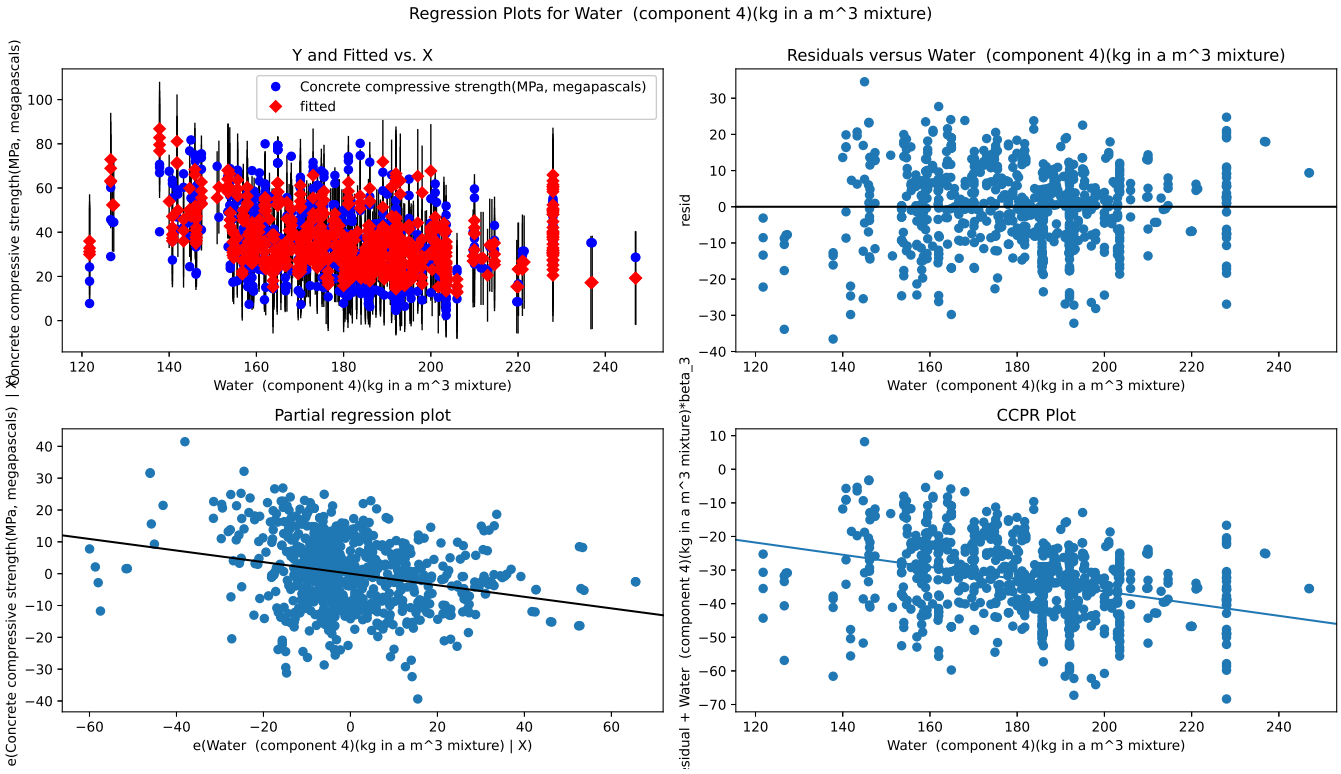


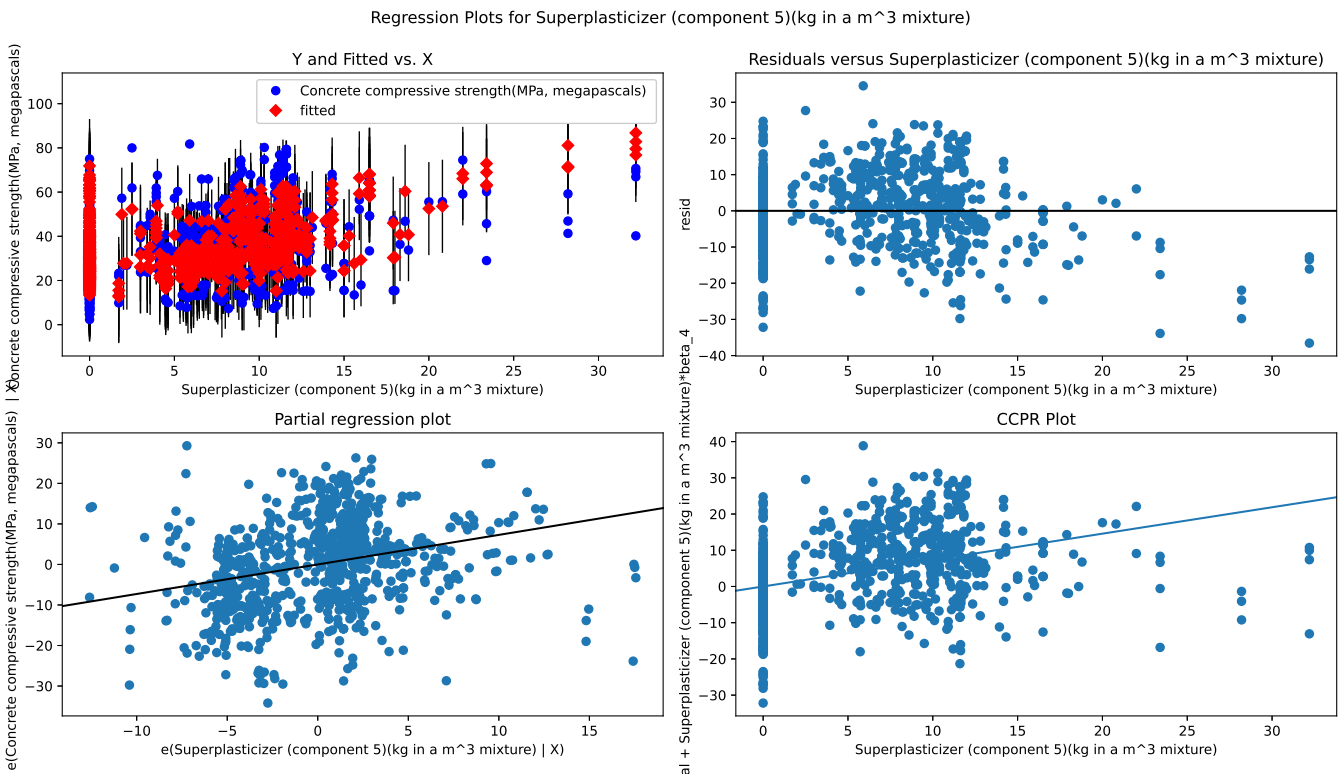
9.2 Checking if any higher order term is needed

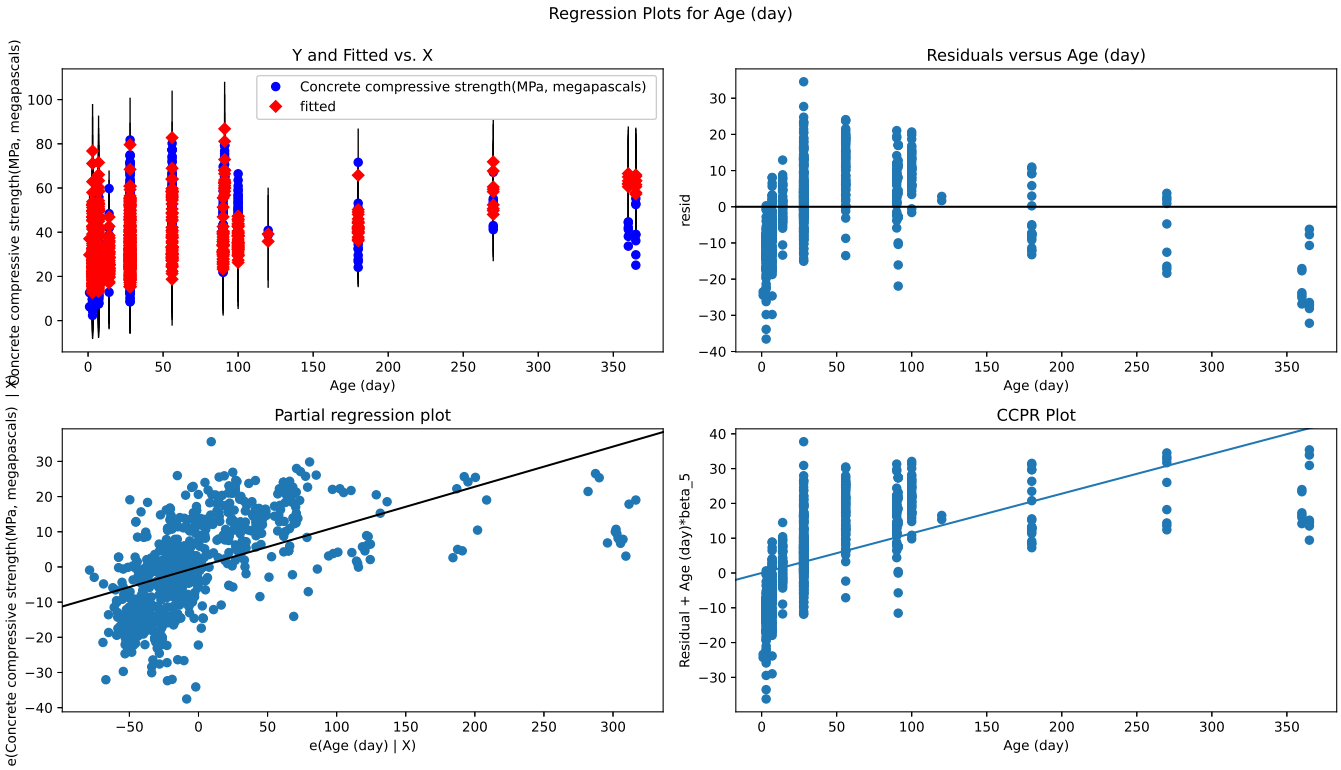
From the plots of "regressors vs Residuals" we can conclude that there is no higher order term involved in the model as there is no siginificant pattern of residual with respect to every regressor present in our model.


Regression Plots for Cement (component 1)(kg in a m^3 mixture)
Y and Fitted vs. X
Concrete compressive strength(MPa, megapascals)
fitted
Cement (component 1)(kg in a m^3 mixture)
Residuals versus Cement (component 1)(kg in a m^3 mixture)
resid
Partial regression plot
e(Concrete compressive strength(MPa, megapascals) | X)
e(Cement (component 1)(kg in a m^3 mixture) | X)
CCPR Plot
sidual + Cement (component 1)(kg in a m^3 mixture)*beta_1


Regression Plots for Blast Furnace Slag (component 2)(kg in a m^3 mixture)
Y and Fitted vs. X
Concrete compressive strength(MPa, megapascals)
fitted
Blast Furnace Slag (component 2)(kg in a m^3 mixture)
Residuals versus Blast Furnace Slag (component 2)(kg in a m^3 mixture)
resid
Partial regression plot
e(Concrete compressive strength(MPa, megapascals) | X)
e(Blast Furnace Slag (component 2)(kg in a m^3 mixture) | X)
CCPR Plot
+ Blast Furnace Slag (component 2)(kg in a m^3 mixture)*beta_2

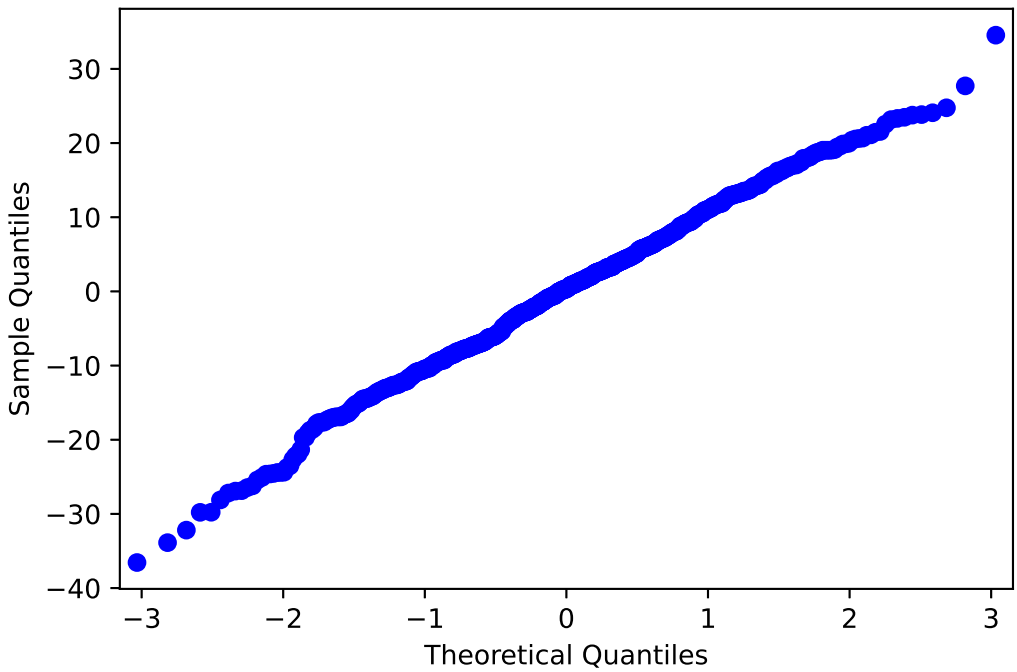

Regression Plots for Water (component 4)(kg in a m^3 mixture)
Y and Fitted vs. X
Concrete compressive strength(MPa, megapascals)
fitted
Concrete compressive strength(MPa, megapascals) | X
Water (component 4)(kg in a m^3 mixture)
Residuals versus Water (component 4)(kg in a m^3 mixture)
resid
Water (component 4)(kg in a m^3 mixture)
Partial regression plot
e(Concrete compressive strength(MPa, megapascals) | X)
e(Water (component 4)(kg in a m^3 mixture) | X)
CCPR Plot
residual + Water (component 4)(kg in a m^3 mixture)*beta_3
Water (component 4)(kg in a m^3 mixture)


Regression Plots for Superplasticizer (component 5)(kg in a m^3 mixture)
Y and Fitted vs. X
Concrete compressive strength(MPa, megapascals)
fitted
Concrete compressive strength(MPa, megapascals) | X
Superplasticizer (component 5)(kg in a m^3 mixture)
Residuals versus Superplasticizer (component 5)(kg in a m^3 mixture)
resid
Superplasticizer (component 5)(kg in a m^3 mixture)
Partial regression plot
e(Concrete compressive strength(MPa, megapascals) | X)
e(Superplasticizer (component 5)(kg in a m^3 mixture) | X)
CCPR Plot
residual + Superplasticizer (component 5)(kg in a m^3 mixture)*beta_4
Superplasticizer (component 5)(kg in a m^3 mixture)



9.3 Checking of normality assumption

Q-Q plots takes sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. In this case it is Normal distribution.



We can see that since we are plotting the data with the theoretical quantiles of a normal distribution, we are getting almost a straight line. Our assumption that error follows normal distribution is nor considered to be a fairly safe assumption. The points seem to fall about a straight line. Notice the x-axis plots the theoretical

quantiles. Those are the quantiles from the standard Normal distribution.

10 Predicted response for test dataset and checking the accuracy of our prediction

Let see how our model performs for our test data.

```
      Observed  Predicted
339  21.911547  31.468704
244  19.932751  22.426155
882  36.438807  22.808899
567  11.957582  23.504201
923  39.844818  28.604882
..         ...        ...
258  42.919881  37.519672
551   8.204075  17.800695
528  38.610656  43.191801
812  34.680643  32.419564
50   39.779318  46.990756

[206 rows x 2 columns]
R-squared value is :  0.49928116001908185
```

From this python we get our R^2value is 0.49928116001908185 which is approximately 50%. So we can conclude our model can explain 50% of actual data accurate. As the R^2 value is moderate so we can say that our model is moderate.