# PROBABILITY AND STATISTICS PROJECT

BY

DEVAKINANDAN PALLA

23MSP3128

UNDER THE FACULTY

Mrs. PITCHUMANI ANGAYARKANNI S

# Regression Analysis on Cement composition data using Minitab

**Student Name:** **Devakinandan Palla**

**Register number:** **23MSP3128**

**Submission Date:** **22-11-2023**

## TABLE OF CONTENTS:

## Introduction:

A materials scientist studies the heat that is generated in cement mixtures. The scientist varies the four ingredients in the mixtures to assess the impact on overall heat generation.

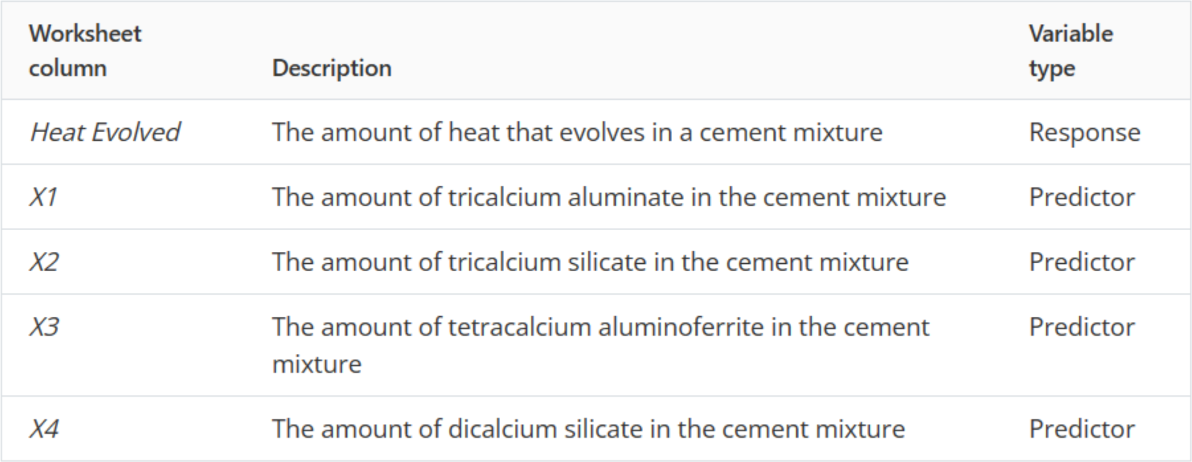
Because this data has 4 continuous predictor variables, I used it to demonstrate **Fit Regression Model** and **Best Subsets Regression**.

The dataset chosen is Cement Composition data

Significance-Using this data, the scientists can understand the contribution of the chemicals X1,X2,X3 and X4 in altering the heat generated by this mixture, and thus understand the relationship and therefore, predict the heat evolved based on the historical data about this mixture.

### Dataset Description:

Overview of this dataset:

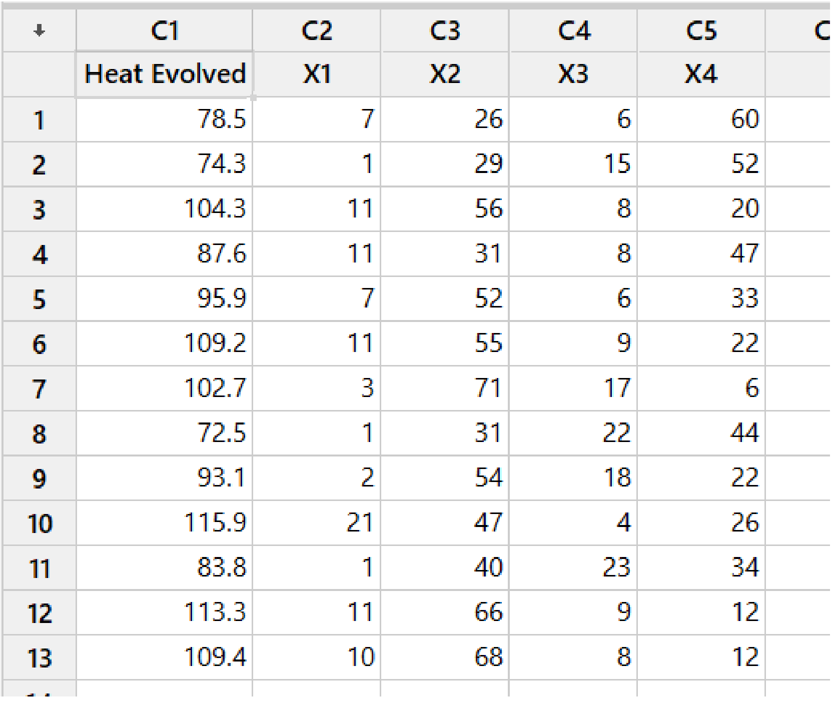


| Worksheet column | Description | Variable type |
|---|---|---|
| *Heat Evolved* | The amount of heat that evolves in a cement mixture | Response |
| *X1* | The amount of tricalcium aluminate in the cement mixture | Predictor |
| *X2* | The amount of tricalcium silicate in the cement mixture | Predictor |
| *X3* | The amount of tetracalcium aluminoferrite in the cement mixture | Predictor |
| *X4* | The amount of dicalcium silicate in the cement mixture | Predictor |

## Source:

Cement composition data - Data Set Library (minitab.com)

There are 13 instances in this data

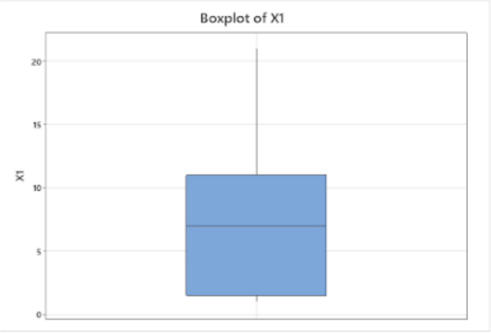


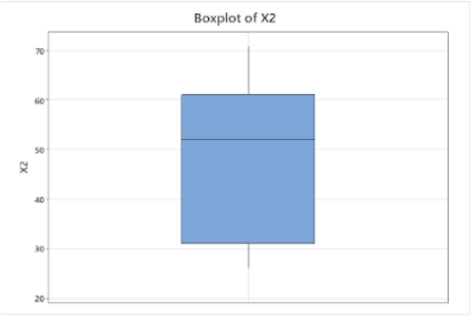
| ↓ | C1 | C2 | C3 | C4 | C5 | C |
|---|---|---|---|---|---|---|
| | Heat Evolved | X1 | X2 | X3 | X4 | |
| 1 | 78.5 | 7 | 26 | 6 | 60 | |
| 2 | 74.3 | 1 | 29 | 15 | 52 | |
| 3 | 104.3 | 11 | 56 | 8 | 20 | |
| 4 | 87.6 | 11 | 31 | 8 | 47 | |
| 5 | 95.9 | 7 | 52 | 6 | 33 | |
| 6 | 109.2 | 11 | 55 | 9 | 22 | |
| 7 | 102.7 | 3 | 71 | 17 | 6 | |
| 8 | 72.5 | 1 | 31 | 22 | 44 | |
| 9 | 93.1 | 2 | 54 | 18 | 22 | |
| 10 | 115.9 | 21 | 47 | 4 | 26 | |
| 11 | 83.8 | 1 | 40 | 23 | 34 | |
| 12 | 113.3 | 11 | 66 | 9 | 12 | |
| 13 | 109.4 | 10 | 68 | 8 | 12 | |

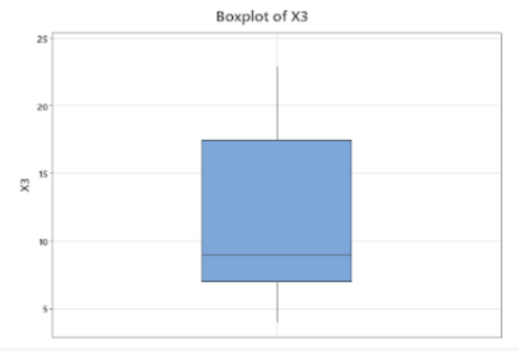
3.**Exploratory Data Analysis and Visualization:**

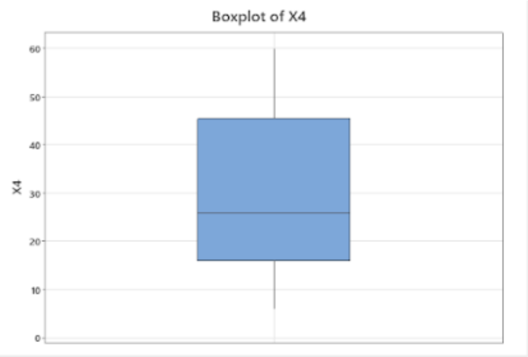
a)Data Analysis:

No anomalies or missing values detected, and box plot representation below concludes the lack of outliers.

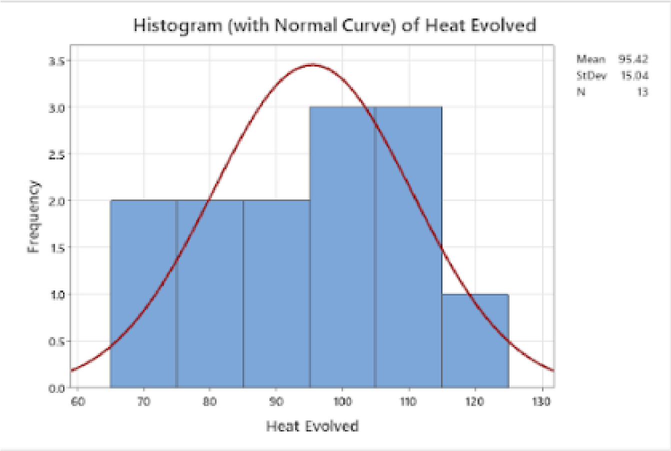




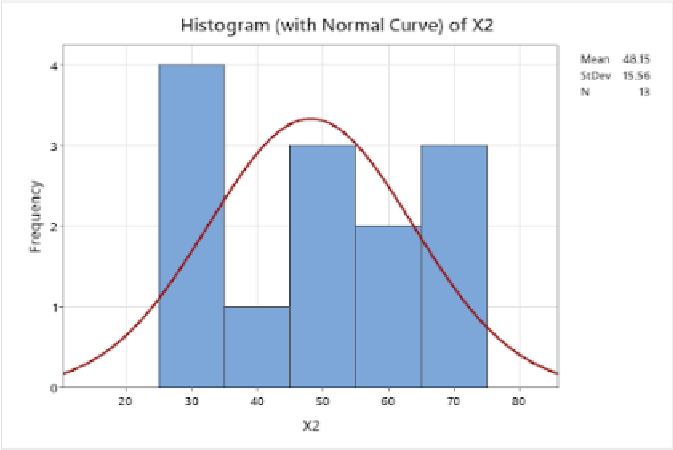




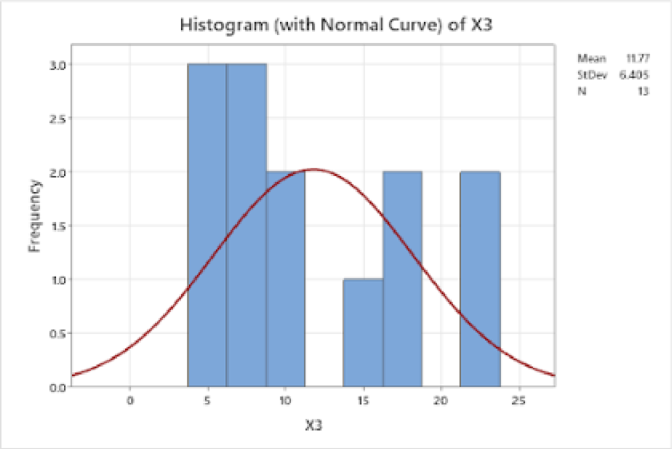
b)Visualization:



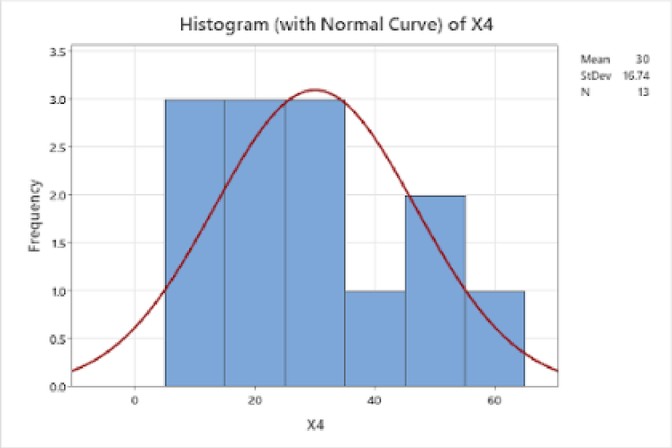
The <heat evolved> data points are normally distributed with a mean of 95.42 and a standard deviation of 15.04, and it is slightly skewed towards the left from the naked eye.



The predictor variable X2 data points are normally distributed with mean of 47.83 and standard deviation of 15.52, and skewed a bit to right.

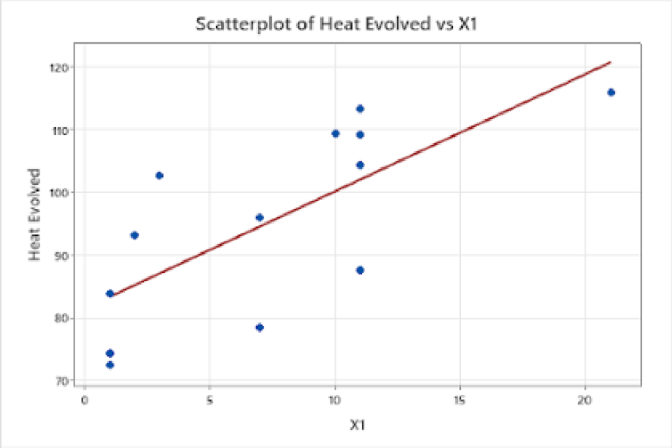


The predictor variable X3 data points are normally distributed with mean of 11.77 and standard deviation of 6.405 and positively skewed

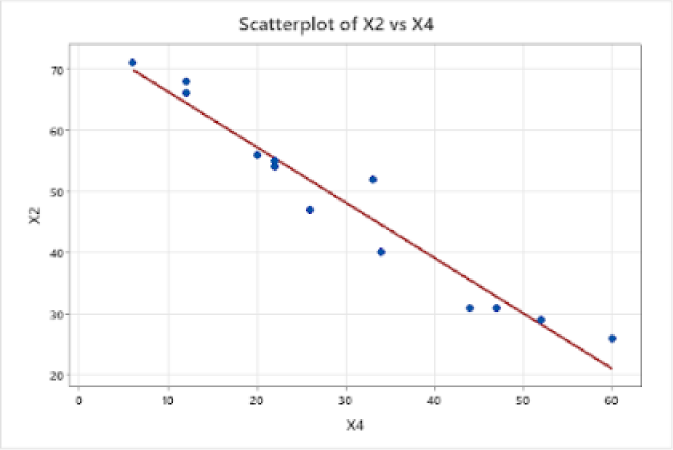


The predictor variable X4 are normally distributed with mean of 30 and standard deviation of 16.74 and slightly right skewed.

Scatterplots:

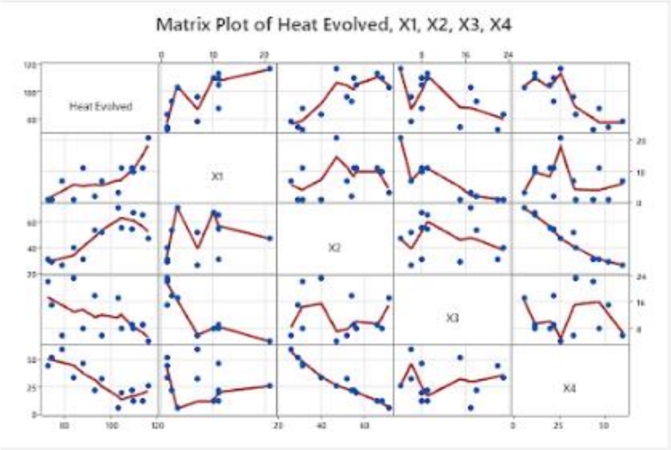


We can observe that there is a strong positive correlation between X1 and heat involved, which means that as X1 value increases, the heat evolve value also increases.



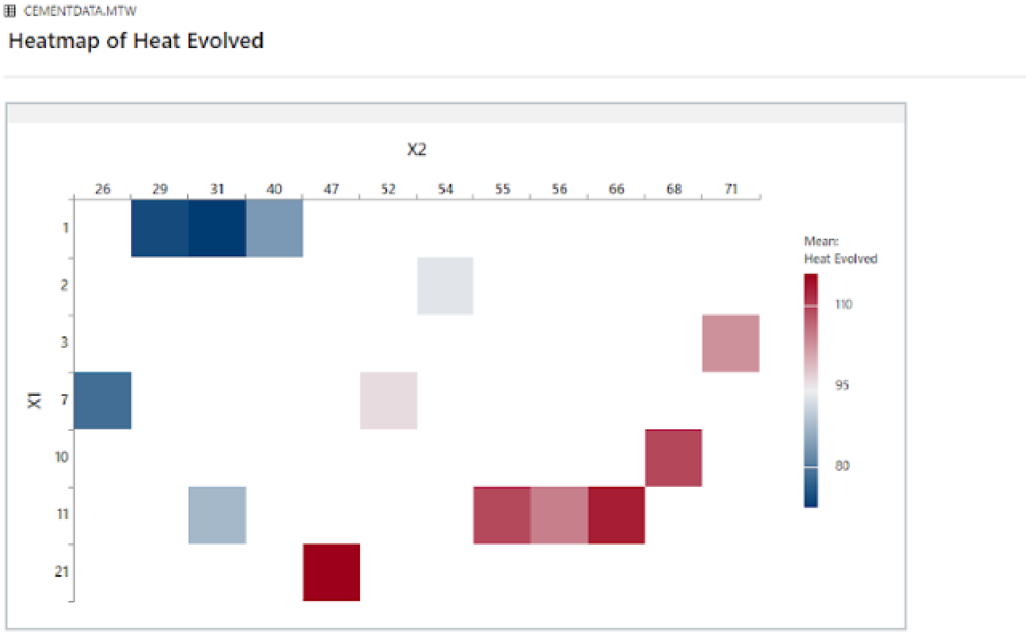
From this scatter plot, we can observe that there is a strong negative correlation between X4 and X2, which means that as X4 value increases, the X2 value decreases.

Matrix Plot:



Box Plot-provided earlier under Data Analysis

Heatmap:

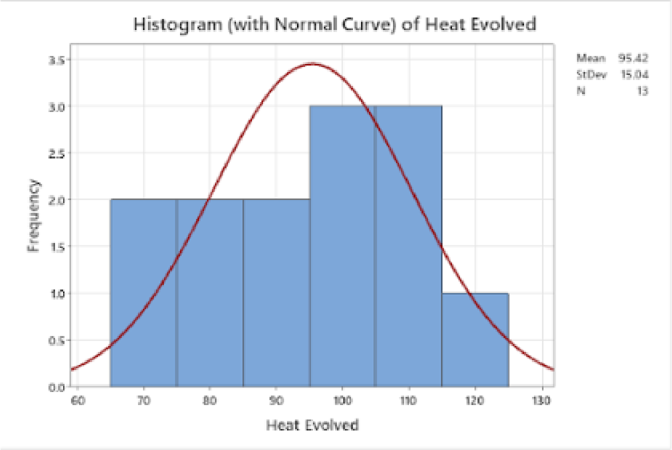


CEMENTDATA.MTW

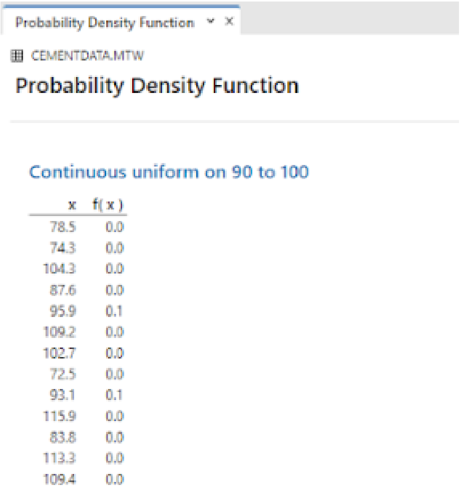
Heatmap of Heat Evolved

As X1 and X2 increase at the same time, this means that the mean of the heat evolved is also increasing alongside.

c)Probability Distribution Analysis:



The data is normally distributed for the target variable, but there could be a slight negative skewness to eye.



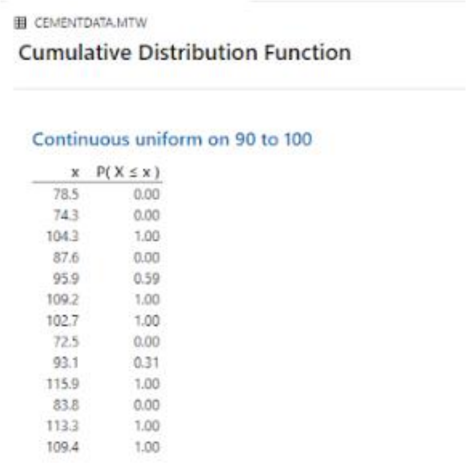
CEMENTDATA.MTW

Probability Density Function

Continuous uniform on 90 to 100

| x | f( x ) |
|---|---|
| 78.5 | 0.0 |
| 74.3 | 0.0 |
| 104.3 | 0.0 |
| 87.6 | 0.0 |
| 95.9 | 0.1 |
| 109.2 | 0.0 |
| 102.7 | 0.0 |
| 72.5 | 0.0 |
| 93.1 | 0.1 |
| 115.9 | 0.0 |
| 83.8 | 0.0 |
| 113.3 | 0.0 |
| 109.4 | 0.0 |

Pdf: Here we infer that f(x)=0.1, where x is between 90 and 100, and is 0 otherwise.


CEMENTDATA.MTW

Cumulative Distribution Function

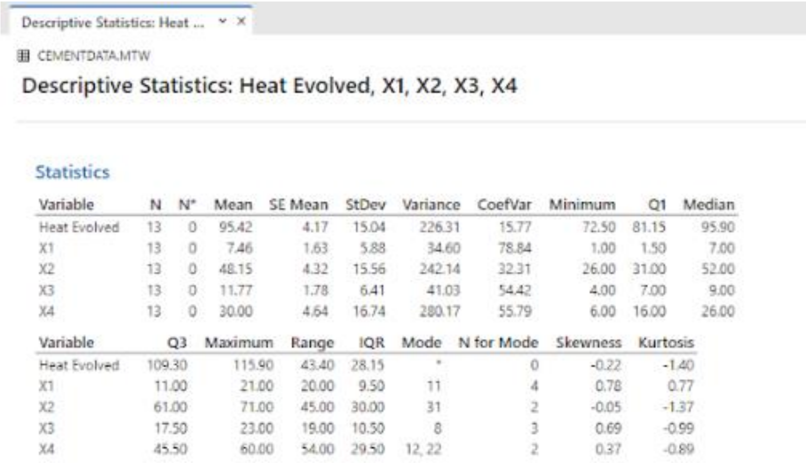
Continuous uniform on 90 to 100

| x | P( X ≤ x ) |
|---|---|
| 78.5 | 0.00 |
| 74.3 | 0.00 |
| 104.3 | 1.00 |
| 87.6 | 0.00 |
| 95.9 | 0.59 |
| 109.2 | 1.00 |
| 102.7 | 1.00 |
| 72.5 | 0.00 |
| 93.1 | 0.31 |
| 115.9 | 1.00 |
| 83.8 | 0.00 |
| 113.3 | 1.00 |
| 109.4 | 1.00 |

CDF:No skewness, and data is normally distributed

P(X=x)=0, where x<90,0 to 1, where 90 is between 90 and 100, and 1 above 100.

4.Descriptive Statistics:



Descriptive Statistics: Heat Evolved, X1, X2, X3, X4

Statistics

| Variable | N | N* | Mean | SE Mean | StDev | Variance | CoefVar | Minimum | Q1 | Median |
|---|---|---|---|---|---|---|---|---|---|---|
| Heat Evolved | 13 | 0 | 95.42 | 4.17 | 15.04 | 226.31 | 15.77 | 72.50 | 81.15 | 95.90 |
| X1 | 13 | 0 | 7.46 | 1.63 | 5.88 | 34.60 | 78.84 | 1.00 | 1.50 | 7.00 |
| X2 | 13 | 0 | 48.15 | 4.32 | 15.56 | 242.14 | 32.31 | 26.00 | 31.00 | 52.00 |
| X3 | 13 | 0 | 11.77 | 1.78 | 6.41 | 41.03 | 54.42 | 4.00 | 7.00 | 9.00 |
| X4 | 13 | 0 | 30.00 | 4.64 | 16.74 | 280.17 | 55.79 | 6.00 | 16.00 | 26.00 |

| Variable | Q3 | Maximum | Range | IQR | Mode | N for Mode | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| Heat Evolved | 109.30 | 115.90 | 43.40 | 28.15 | * | 0 | -0.22 | -1.40 |
| X1 | 11.00 | 21.00 | 20.00 | 9.50 | 11 | 4 | 0.78 | 0.77 |
| X2 | 61.00 | 71.00 | 45.00 | 30.00 | 31 | 2 | -0.05 | -1.37 |
| X3 | 17.50 | 23.00 | 19.00 | 10.50 | 8 | 3 | 0.69 | -0.99 |
| X4 | 45.50 | 60.00 | 54.00 | 29.50 | 12, 22 | 2 | 0.37 | -0.89 |

All the measure of central tendency including mean, median, mode is described iin the above table.

Measure of spread including range, variance and std deviation is also mentioned.

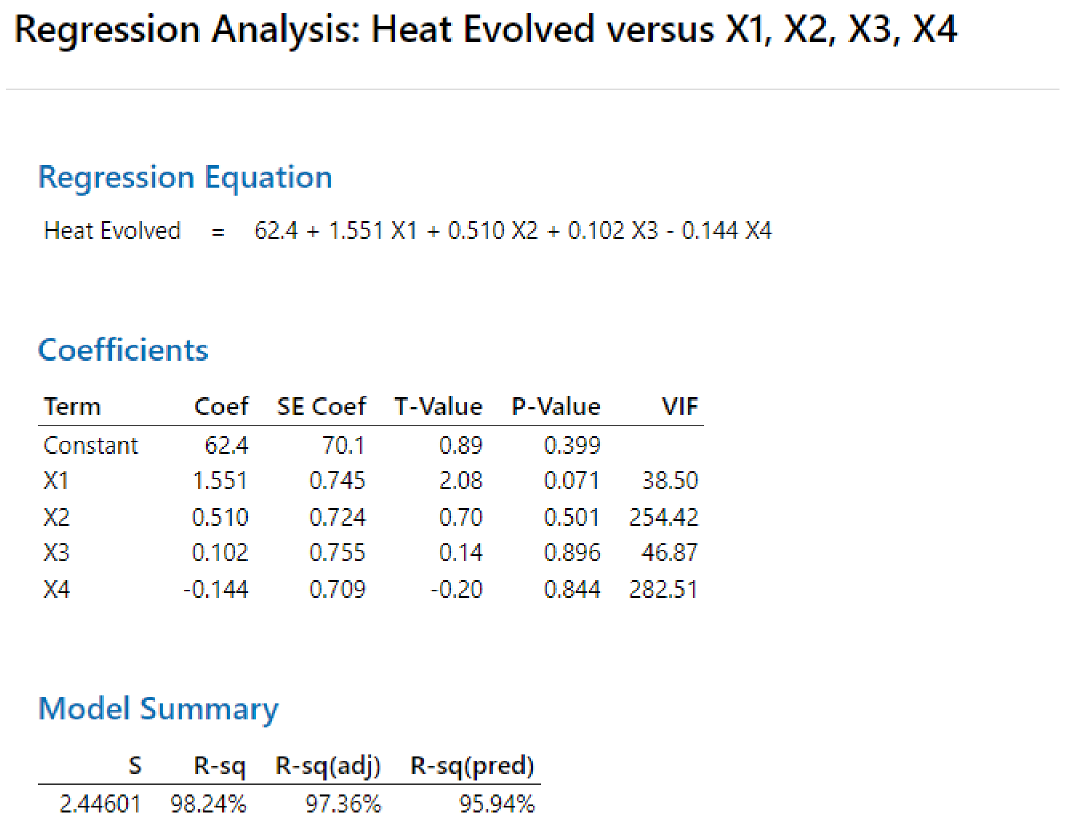
Heat evolved is left skewed and is mesokurtic

X1 is right skewed and is mesokurtic as well

X2 is positively skewed and is mesokurtic

X4 is slightly right skewed and is mesokurtic as well.

5.Regression Analysis:



## Regression Analysis: Heat Evolved versus X1, X2, X3, X4

### Regression Equation

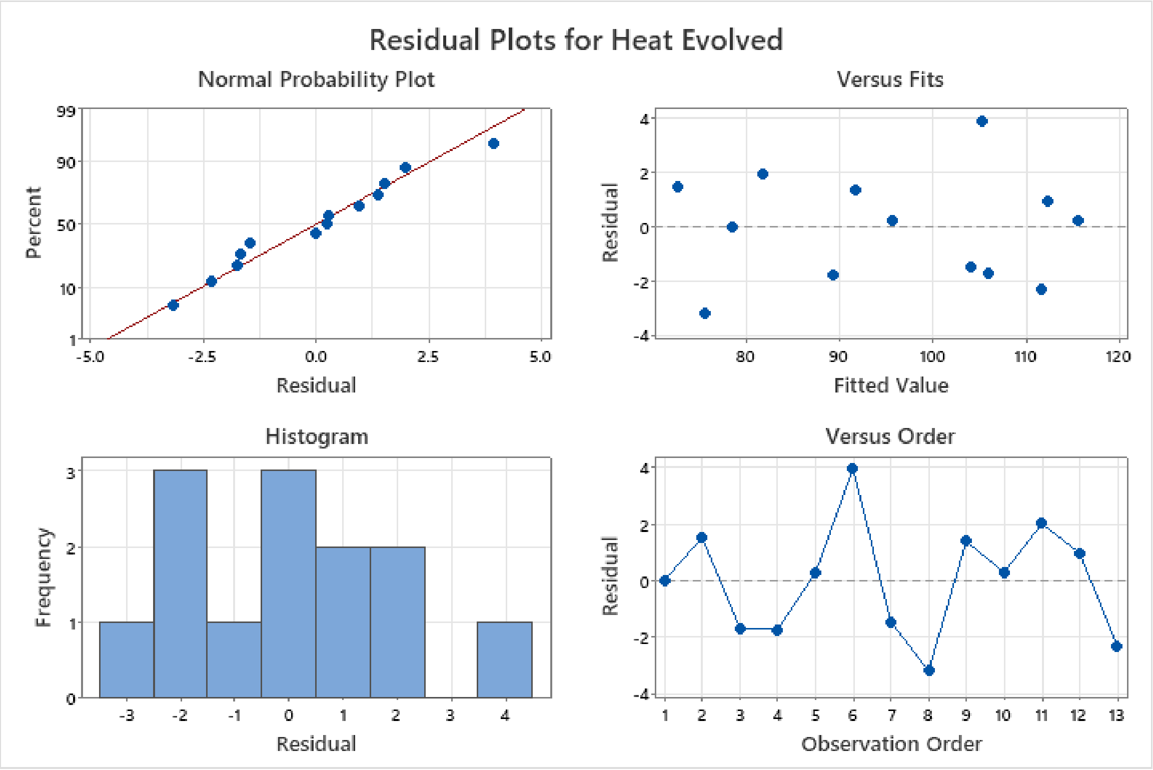
Heat Evolved = 62.4 + 1.551 X1 + 0.510 X2 + 0.102 X3 - 0.144 X4

### Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 62.4 | 70.1 | 0.89 | 0.399 | |
| X1 | 1.551 | 0.745 | 2.08 | 0.071 | 38.50 |
| X2 | 0.510 | 0.724 | 0.70 | 0.501 | 254.42 |
| X3 | 0.102 | 0.755 | 0.14 | 0.896 | 46.87 |
| X4 | -0.144 | 0.709 | -0.20 | 0.844 | 282.51 |

### Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---|---|---|---|
| 2.44601 | 98.24% | 97.36% | 95.94% |



When x1,x2,x3,x4 is equal to zero, then Heat evolved is constant with value of 62.4

X1 assists a heat increase of 1.551 rate of increase

X2 assists a heat inc of 0.510 rate of increase

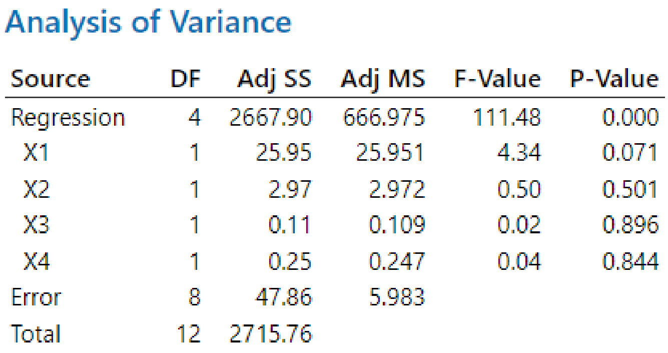
X3 :0.102 rate of increase

And x4 a rate of decrease of 0.144

All the above refer to a one unit increase of the respective variable.

6.Done in assignment 4

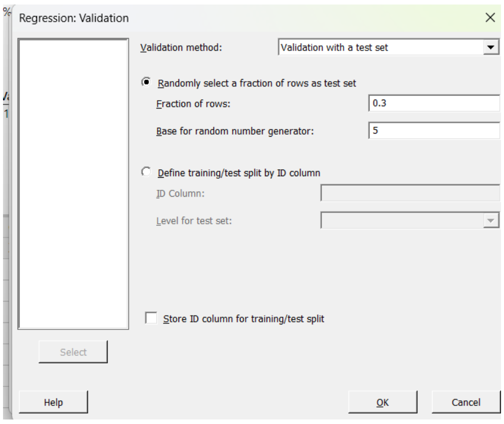
7.ANOVA:



Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 4 | 2667.90 | 666.975 | 111.48 | 0.000 |
| X1 | 1 | 25.95 | 25.951 | 4.34 | 0.071 |
| X2 | 1 | 2.97 | 2.972 | 0.50 | 0.501 |
| X3 | 1 | 0.11 | 0.109 | 0.02 | 0.896 |
| X4 | 1 | 0.25 | 0.247 | 0.04 | 0.844 |
| Error | 8 | 47.86 | 5.983 | | |
| Total | 12 | 2715.76 | | | |

8.Model Validation, diagnostic and Prediction:



Regression: Validation

Validation method: Validation with a test set

● Randomly select a fraction of rows as test set

Fraction of rows: 0.3

Base for random number generator: 5

○ Define training/test split by ID column

ID Column:

Level for test set:

□ Store ID column for training/test split

30%:70% is the testing:training proportion

CEMENTDATA.MTW

# Regression Analysis: Heat Evolved versus X1, X2, X3, X4

## Method

| | |
|---|---|
| Test set fraction | 30.8% |

## Regression Equation

Heat Evolved = 50 + 1.66 X1 + 0.66 X2 + 0.17 X3 + 0.01 X4

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 50 | 110 | 0.45 | 0.675 | |
| X1 | 1.66 | 1.11 | 1.50 | 0.207 | 50.21 |
| X2 | 0.66 | 1.16 | 0.56 | 0.603 | 208.97 |
| X3 | 0.17 | 1.11 | 0.15 | 0.887 | 53.77 |
| X4 | 0.01 | 1.12 | 0.01 | 0.996 | 209.84 |

## Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) | Test S | Test R-sq |
|---|---|---|---|---|---|
| 3.01818 | 98.03% | 96.06% | 91.33% | 2.05009 | 98.02% |

## Analysis of Variance

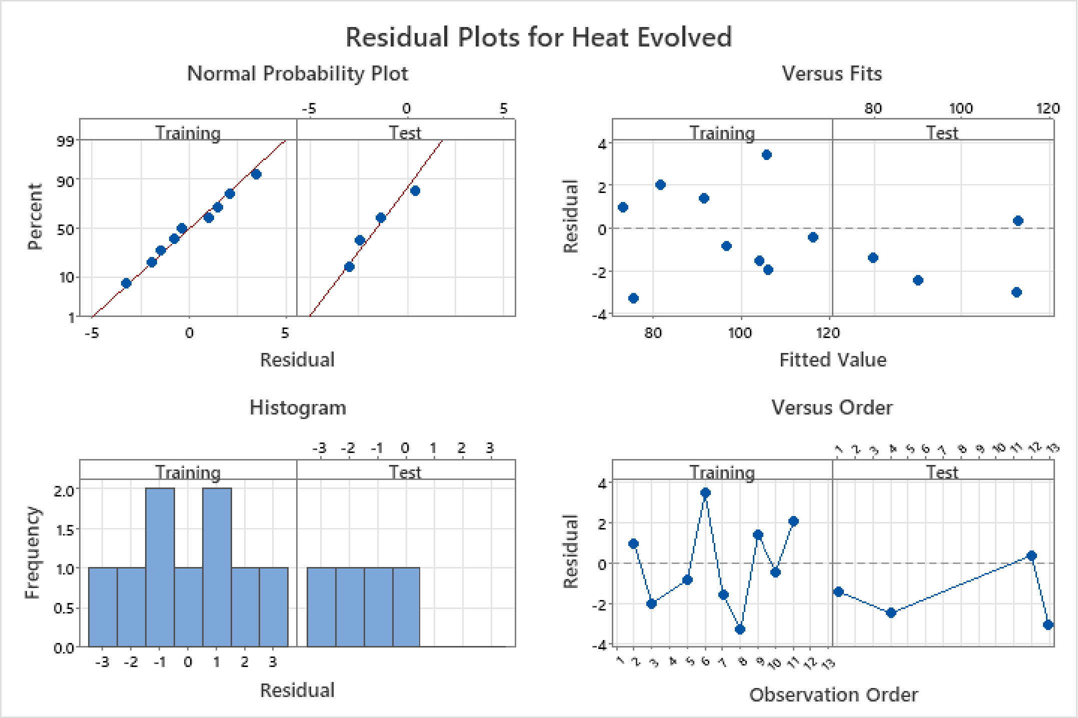
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 4 | 1811.18 | 452.796 | 49.71 | 0.001 |
| X1 | 1 | 20.60 | 20.596 | 2.26 | 0.207 |
| X2 | 1 | 2.89 | 2.892 | 0.32 | 0.603 |
| X3 | 1 | 0.21 | 0.208 | 0.02 | 0.887 |
| X4 | 1 | 0.00 | 0.000 | 0.00 | 0.996 |
| Error | 4 | 36.44 | 9.109 | | |
| Total | 8 | 1847.62 | | | |

## Fits and Diagnostics for Unusual Observations

Test Set

| Obs | Heat Evolved | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 1 | 78.50 | 79.85 | -1.35 | -0.29 | X |

*X Unusual X*



Inferences:

High VIF indicates more multi collinearity and thus x2 and x4 are major contributors.

Low T value means size of difference relative to variation in the sample data s low.

The MSE is 3.018 which is pretty low.

The R sq value ,r sq adj,pred having 90%+ suggest that this model is a best model with good accuracy.