# STW7089CEM: Introduction to Statistical Methods for data Science

**Submitted To**

Name: Hikmat Saud

**Submitted by**

Name: Prabesh Basnet

CUID:  16642701

SID: 250118

## Table of Contents

## Introduction to Non-Linear Regression

In Statistics, Non-linear regression refers to the form of regression analysis where observational data are modeled by a function which is a non-linear combination of the parameters in the model and depends on one or more independent variable.

## Objectives

The aim of this assignment is to select the best regression model (from a candidate set of nonlinear regression models) that can effectively describe the relationship between several continuous environmental variables and the net hourly electrical energy output (x5) of a Combined Cycle Power Plant (CCPP). Understanding this relationship is crucial for optimizing energy generation, improving efficiency, and managing operational constraints in power plants.

## Task 1 - Preliminary Data Analysis

The first task of the assignment follows preliminary data analysis of the given dataset. In this part the given dataset is loaded and the trends of the data is studied. There are four input variables (x1, x3, x4, and x5) and an output variable(x2). We will create a time series plot for these input and output variables by first separating these input variables in a separate csv file (x.csv) and output variable (y.csv) and for time we just use the number rows (t.csv).

Some of the key observations are:

- It's data collected from a power plant running at **full load** between **2006 and 2011.**

- It has **9,568 rows** — each row is one hour of recorded data.

- There are **5 columns (variables)**: 4 inputs and 1 output.

### Task 1.1 - Time Series Plots of input and output signals

A Time series plot is used to explain and describe the changes in a set of variables across a series of times. Our main objective is to create time series plot for the input and output signals based on the time variables for which we are using the number of rows since the dataset is recording data for each hour. We get the x, y and t data frames and we can now plot the time series plot of these variables by using plot () function.

## Time series plot of X Signal

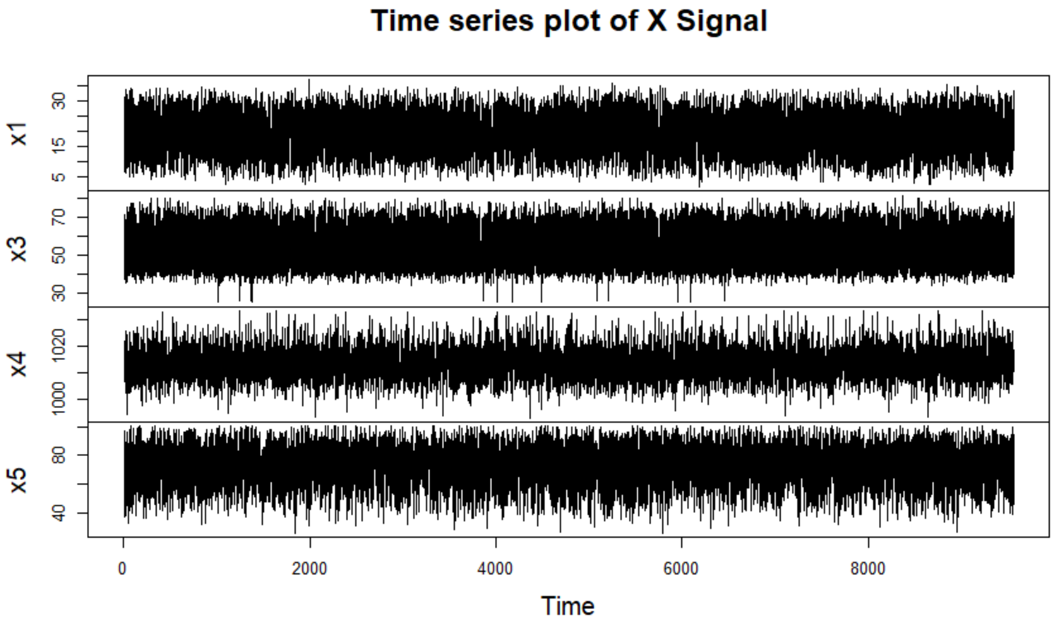


Figure 1 showing time series plot of input signals.

## Observations from the Input Time Plots

From the Time series plot we created for all the input variables (x) we get the above figure above (figure 1). This displays the variation of data based on time. Here we can observe that the variable x1 is varying in bet ween 5 and 30, x3 is varying in between 30 and 70, x4 is varying in between 1000 and 1020. Likewise, x5 varies in between 40 and 80.

### Time series plot of the output signal

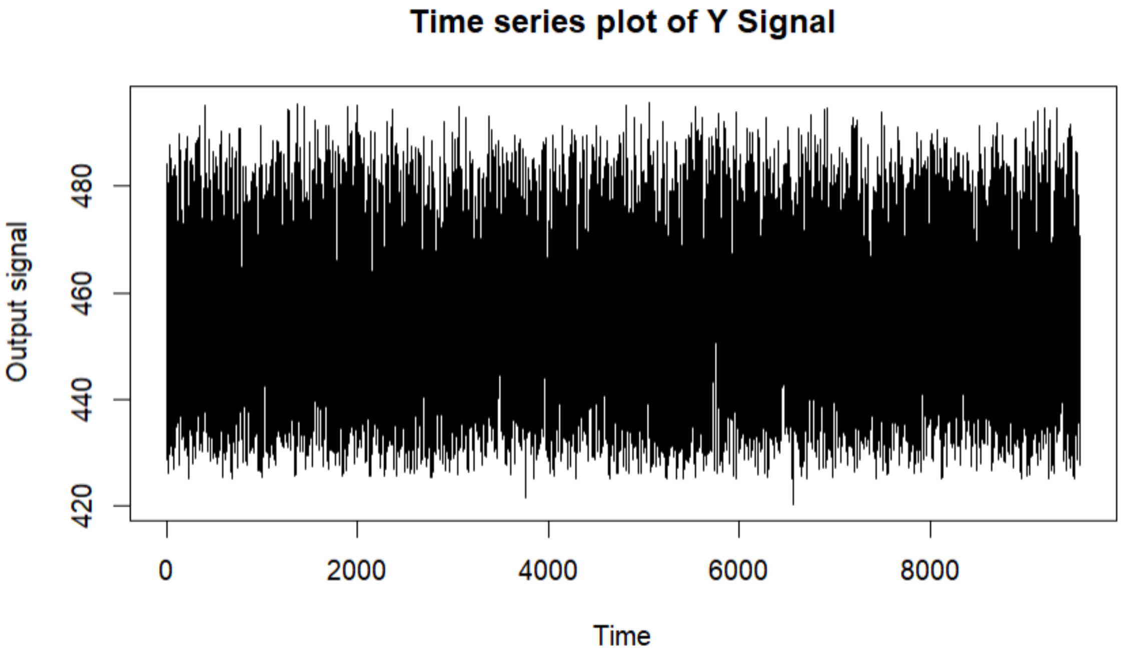


Figure 2 showing time series plot of output signal.

### Observations from the Output Time Plot

From the Time series plot we created for the output variables (y) we get the figure above (figure 2). This displays the variation of data based on time. Here we can observe that the variable y is varying in between 420 and 480.

## Task 1.2 Distribution for each signal

To display the distribution of data, we employ a histogram, density plot, box plot, or violin plot. However, the most common graph is the histogram, but in this case, we also made a density plot. Furthermore, histograms can provide us with a general notion of the distribution's overall form and the data's overall trend, and they work well with discrete data. But our data's nature was discovered must be continuous, thus a density plot is required.

## Density plot and Histogram of X

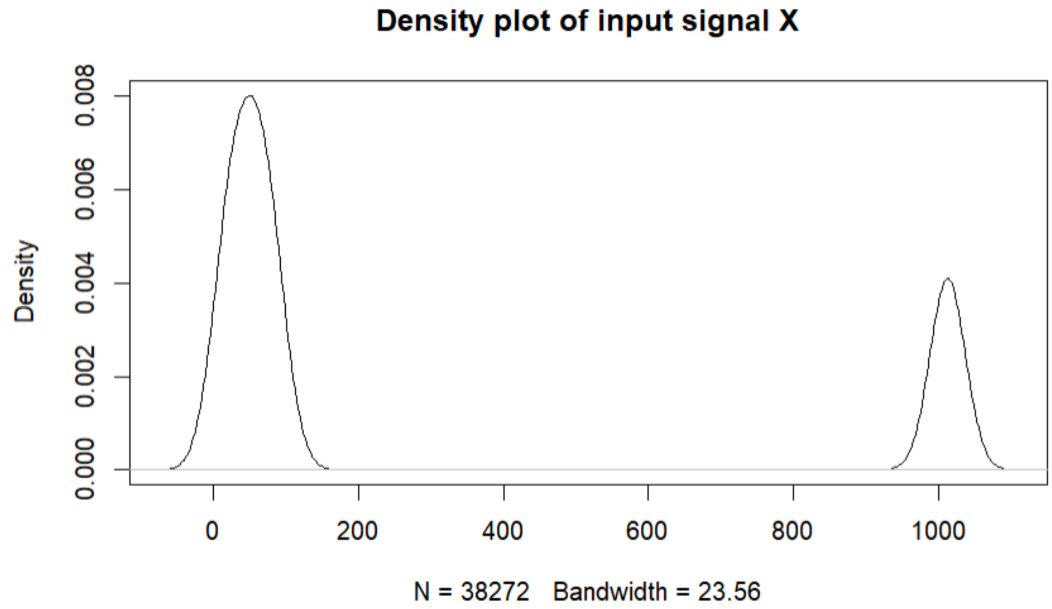


Figure 3 showing density plot of input signal.

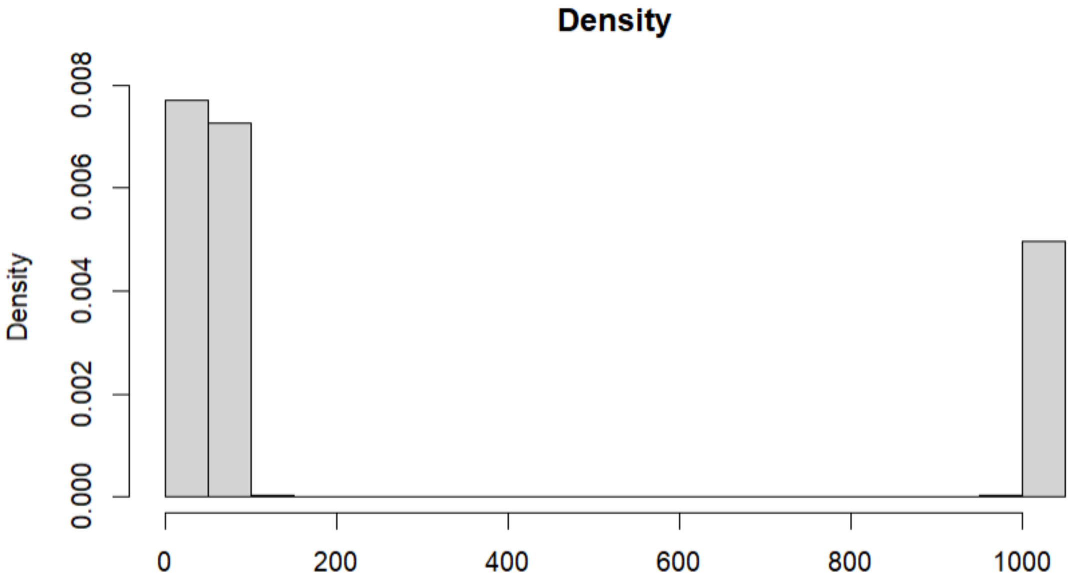


Figure 4 showing histogram of input signal.

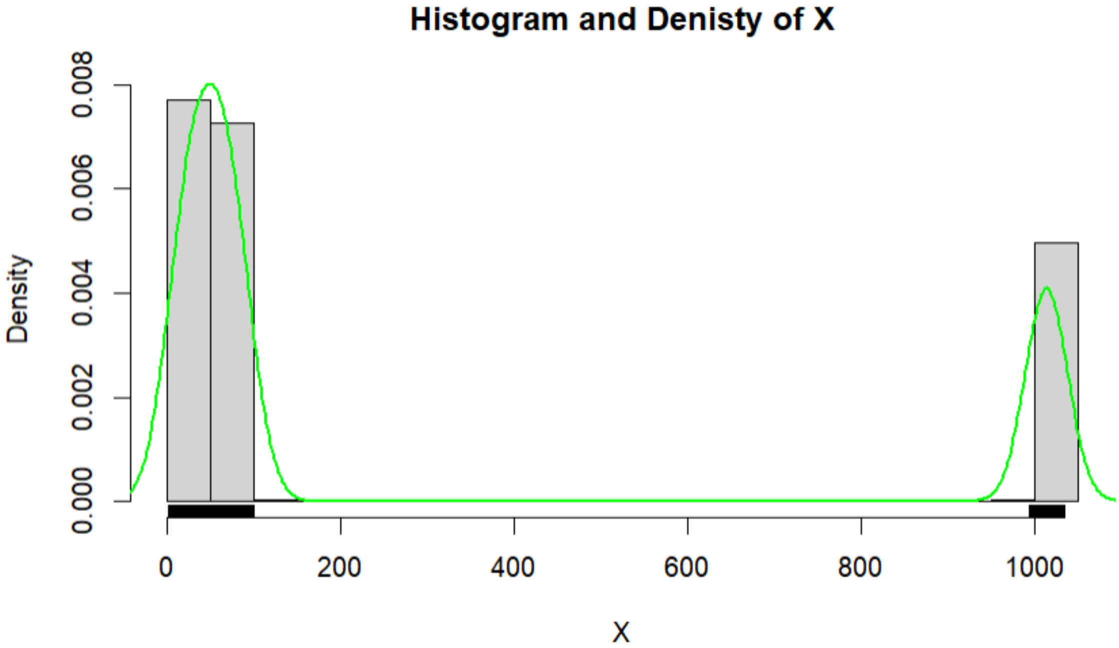


Figure 5 showing combined density plot and histogram of input signal.

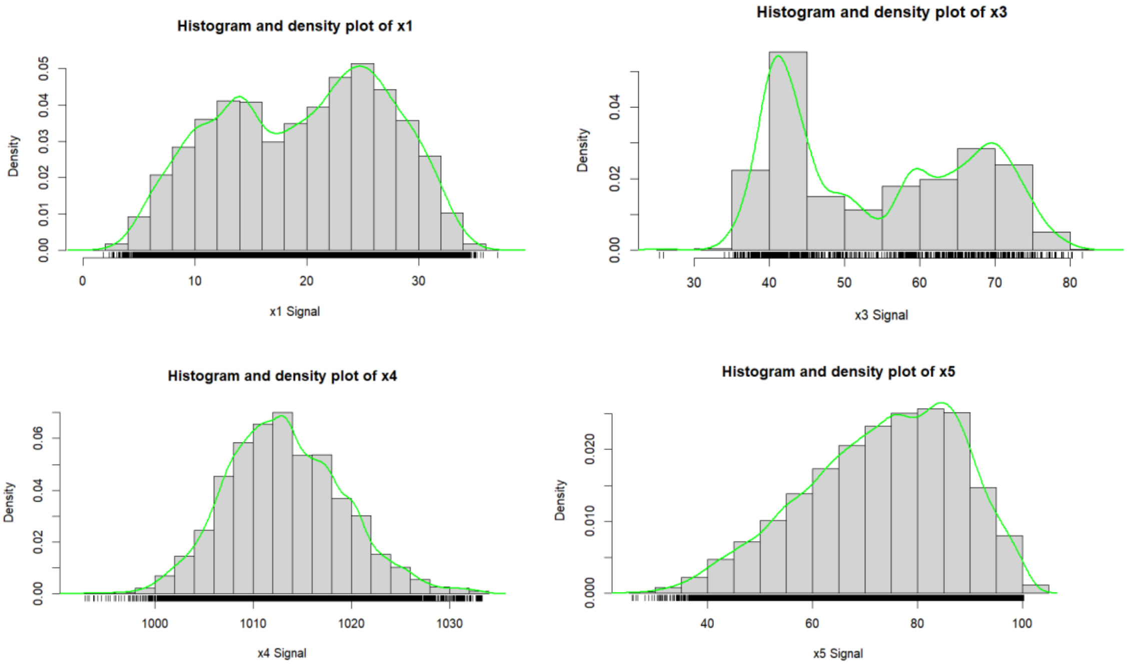


Figure 6 showing combined density plot and histogram of all the input signals.

## Observations from the input histogram and density plot

The integrated histograms and density plots presented in Figure 6 illustrate the distribution of four input signals: x1, x3, x4, and x5. The x1 signal appears to be approximately uniformly distributed, exhibiting several minor peaks, which implies it may not conform to a standard normal distribution. The x3 signal displays a multimodal and skewed distribution, characterized by a prominent peak around 40–45 and subsequent fluctuations, indicating possible non-normality and the likelihood of subgroups or mixed sources within the data. The x4 signal demonstrates a bell-shaped, symmetric distribution that aligns closely with a normal distribution, as evidenced by the smooth, centered density curve and the relatively even histogram. Finally, the x5 signal reveals a right-skewed distribution, featuring a longer tail on the right side, which suggests that higher values are less common yet still significantly present. The density curves superimposed on the histograms further validate these interpretations by visualizing the smooth estimated probability distribution of each signal.

## Density plot and Histogram of Y

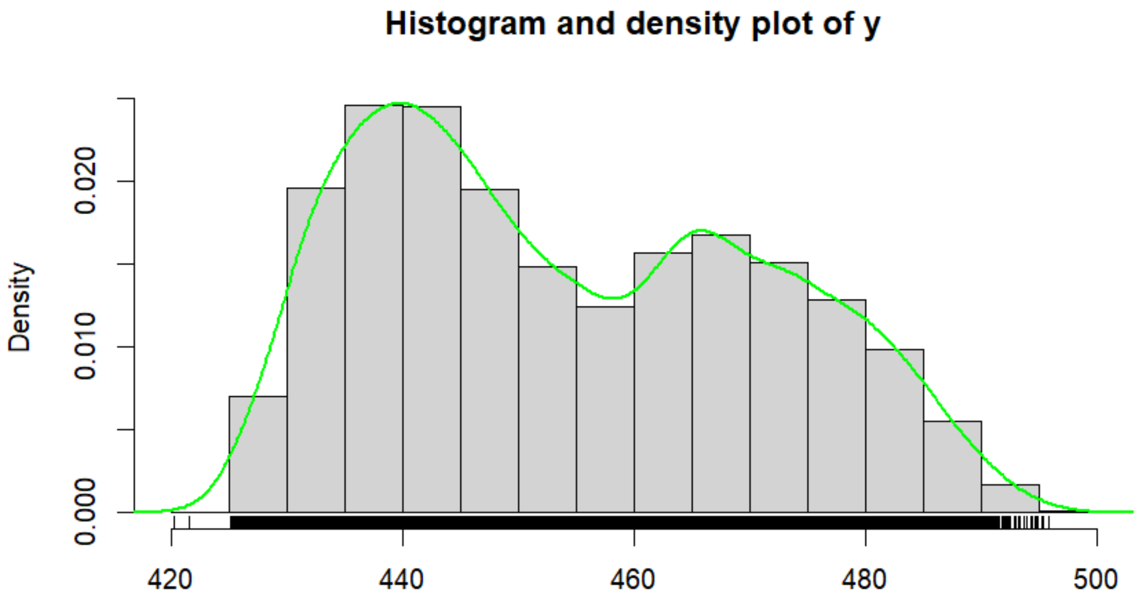


Figure 7 showing combined density plot and histogram of the output signal.

### Observations from the output histogram and density plot

The combined histogram and density plot of the variable *y* illustrates the distribution of the output signal. The histogram, represented by the gray bars, shows that the values of *y* are most frequently concentrated around 440–445, indicating a peak in this region. The density curve, shown in green, provides a smoothed estimate of the distribution and highlights a slightly skewed shape with a secondary mode or bump around 465. This suggests that while the data is primarily unimodal, there may be some multimodal tendencies or variability in the signal. Overall, the plot indicates that the output signal is not perfectly symmetric and may contain underlying structure or subgroups.

## Task 1.3 - Correlation Scatter plots

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect. Ref. (https://www.jmp.com/en/statistics-knowledge-portal/what-is-correlation#:~:text=Correlation%20is%20a%20statistical%20measure,together%20at%20a%20constant%20rate).) .

We continue by plotting the correlation between all the input signals and the output signal and we get the following correlation plot.

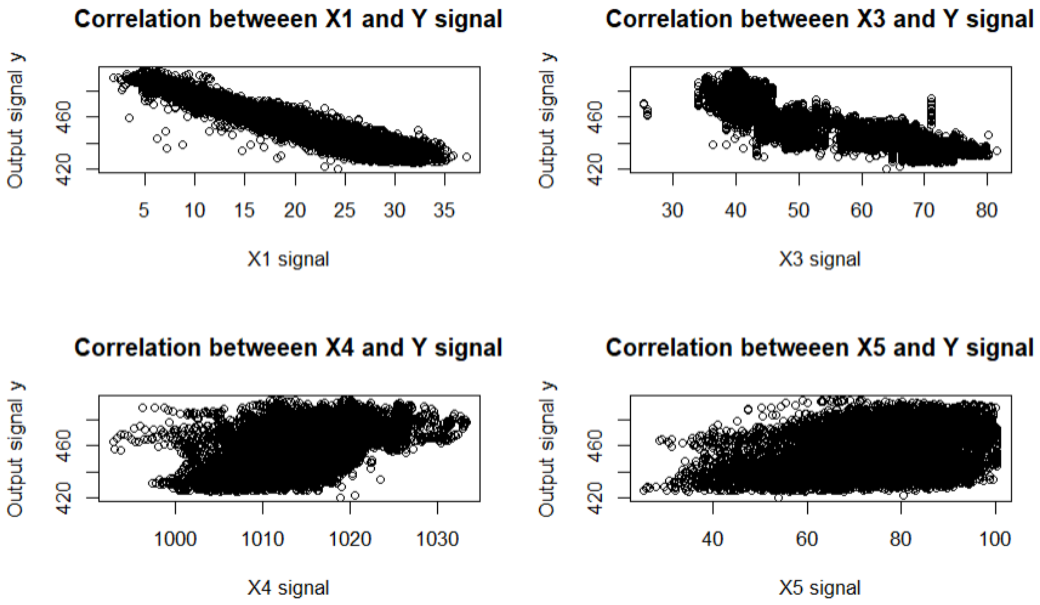


Figure 6 showing correlation.

### Observations from the correlation plot

Figure 6 gives us a visual sense of how different input signals relate to the output signal *y*. From the first plot, we can see that as the X1 signal increases, the output *y* consistently decreases, showing a strong and clear negative relationship. The plot for X3 tells a similar story—though not as clean, it still shows that higher X3 values tend to go along with lower values of *y*, indicating another negative correlation. On the other hand, the plots for X4 and X5 don't show such clear patterns. The relationship between X4 and *y* seems weak and scattered, with only a slight upward trend. Likewise, the correlation between X5 and *y* looks minimal, with the points spread widely and no obvious direction. Overall, X1 and X3 appear to be more strongly linked to the output signal than X4 or X5.

## Task 2 – Regression-Modelling the relationship between gene expressions

Assuming that the relationship can be represented as a polynomial regression model, we will now describe the relationship between the input and output EEG data. Five distinct nonlinear polynomial regression models were provided to us, and among them, we have to choose the most appropriate one.

We would like to determine a suitable mathematical model in explaining the relationship between the output Net hourly electrical energy (y) = x2 with other input x2: Net hourly electrical energy output (EP) in MW (dependent variable), x1: Temperature (T) – Ambient temperature (°C), x3: Ambient Pressure (AP) – Atmospheric pressure (millibar), x4: Relative Humidity (RH) – Humidity level (%), x5: Exhaust Vacuum (V) – Vacuum collected from the steam turbine (cm Hg) that 'regulate' its expression, which we assume can be described by a polynomial regression model.

Model 1: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta\_bias$

Model 2: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5 + \theta\_bias$

Model 3: $y = \theta_1 x_3 + \theta_2 x_4 + \theta_3 x_5^3$

Model 4: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5^3 + \theta\_bias$

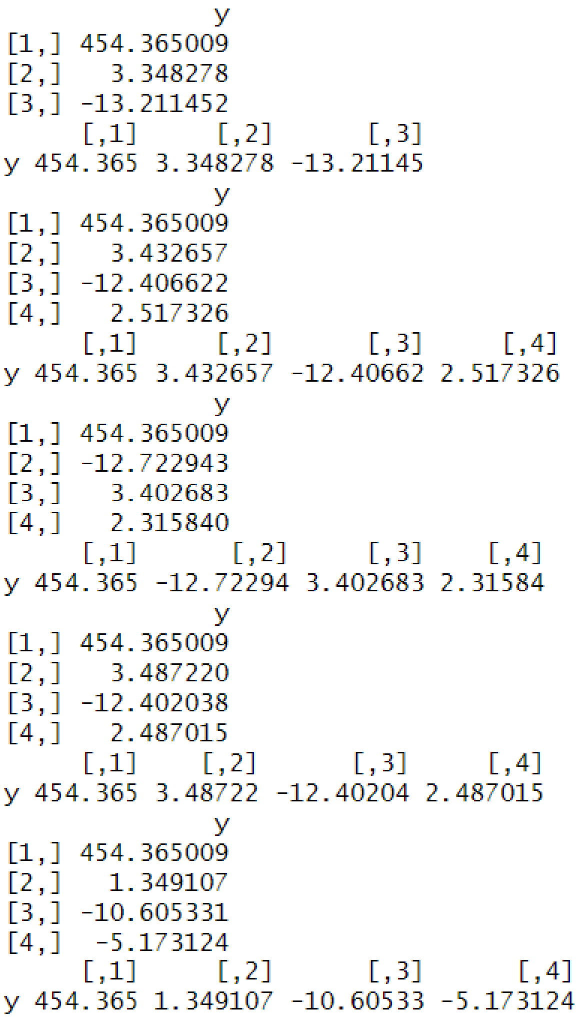
Model 5: $y = \theta_1 x_4 + \theta_2 x_1^2 + \theta_3 x_3^2 + \theta\_bias$

### Task 2.1 - Estimate model parameters $\theta = \{\theta_1, \theta_2, \cdots, \theta\_bias\}^T$ for every candidate model using Least Squares ($\hat{\theta} = (X^TX)^{-1}X^Ty$), using the provided input and output gene datasets (use all the data for training).

We start by calculating Theta-hat of all the given models. First we calculate the ones value for binding data.

ones = matrix(1 , length(X)/4,1)

We then proceed to calculate theta-hat value for all the models



```
              y
[1,] 454.365009
[2,]   3.348278
[3,] -13.211452
     [,1]     [,2]      [,3]
y 454.365 3.348278 -13.21145
              y
[1,] 454.365009
[2,]   3.432657
[3,] -12.406622
[4,]   2.517326
     [,1]     [,2]      [,3]     [,4]
y 454.365 3.432657 -12.40662 2.517326
              y
[1,] 454.365009
[2,] -12.722943
[3,]   3.402683
[4,]   2.315840
     [,1]      [,2]     [,3]    [,4]
y 454.365 -12.72294 3.402683 2.31584
              y
[1,] 454.365009
[2,]   3.487220
[3,] -12.402038
[4,]   2.487015
     [,1]    [,2]      [,3]     [,4]
y 454.365 3.48722 -12.40204 2.487015
              y
[1,] 454.365009
[2,]   1.349107
[3,] -10.605331
[4,]  -5.173124
     [,1]     [,2]      [,3]      [,4]
y 454.365 1.349107 -10.60533 -5.173124
```

Figure 7 showing theta-hat values of each model.

## Observations from the theta-hat values

Model 1: Utilizes pressure squared and humidity → Pressure squared decreases energy output, while humidity enhances it.

Model 2: Vacuum is added to Model 1 → the same as before, but there is also a slight benefit to vacuum.

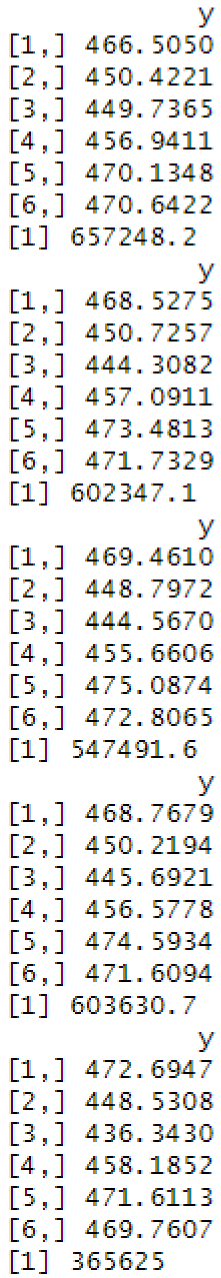
Model 3: Utilizes vacuum cubed, pressure, and humidity while pressure still reduces output, vacuum cubed helps increase it.

Model 4: Utilizes vacuum cubed, pressure squared, and humidity → whereas pressure squared reduces production, humidity and vacuum squared aid.

Model 5: Makes use of temperature squared, pressure squared, and humidity. Humidity helps

### Task 2.2 - Modal residual sum of squared errors (RSS Calculation)

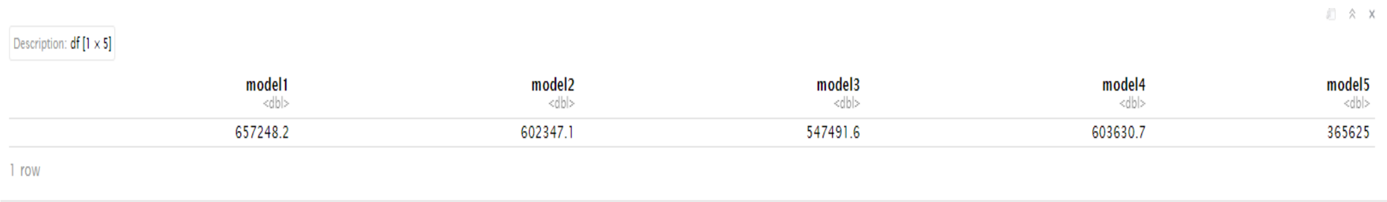
We start by calculating the Y-hat and RSS for each of the given five models.



```
              y
[1,] 466.5050
[2,] 450.4221
[3,] 449.7365
[4,] 456.9411
[5,] 470.1348
[6,] 470.6422
[1] 657248.2
              y
[1,] 468.5275
[2,] 450.7257
[3,] 444.3082
[4,] 457.0911
[5,] 473.4813
[6,] 471.7329
[1] 602347.1
              y
[1,] 469.4610
[2,] 448.7972
[3,] 444.5670
[4,] 455.6606
[5,] 475.0874
[6,] 472.8065
[1] 547491.6
              y
[1,] 468.7679
[2,] 450.2194
[3,] 445.6921
[4,] 456.5778
[5,] 474.5934
[6,] 471.6094
[1] 603630.7
              y
[1,] 472.6947
[2,] 448.5308
[3,] 436.3430
[4,] 458.1852
[5,] 471.6113
[6,] 469.7607
[1] 365625
```

Figure 8 showing Y-hat and RSS values of each model.

We then proceed with displaying all the RSS values



Description: df [1 x 5]

| model1<br><dbl> | model2<br><dbl> | model3<br><dbl> | model4<br><dbl> | model5<br><dbl> |
|---|---|---|---|---|
| 657248.2 | 602347.1 | 547491.6 | 603630.7 | 365625 |

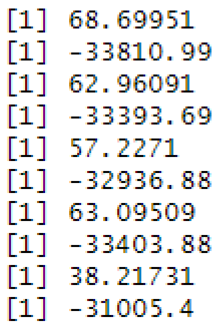
1 row

Figure 9 Displaying RSS values of each model.

We conclude that model 5 has the lowest RSS (36625) of all the models shown and is most likely the best fix from this procedure.

## Task 2.3 - Calculating the log likelihood and Variance of each model

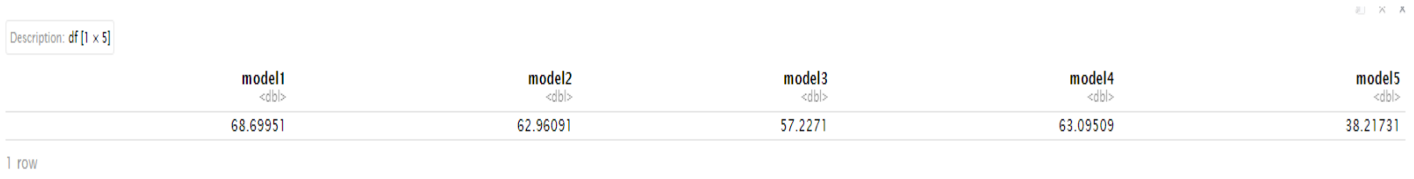
We start by calculating the variance to see how spread out our model prediction are and log likelihood is the measure of how likely our model would produce the data we observed. We fetch the number of variations for the variance model and start by calculating the variance and log-likelihood of each of the five given models.



```
[1] 68.69951
[1] -33810.99
[1] 62.96091
[1] -33393.69
[1] 57.2271
[1] -32936.88
[1] 63.09509
[1] -33403.88
[1] 38.21731
[1] -31005.4
```

Figure 10 Displaying Variance and log likelihood values of each model.

We then display the variance values.



Description: df [1 x 5]

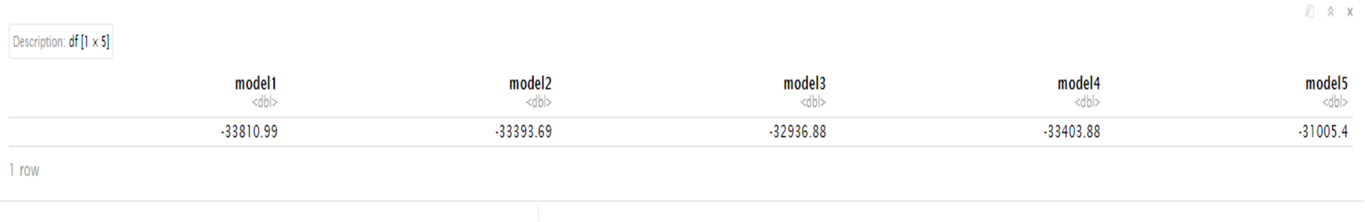
| model1 | model2 | model3 | model4 | model5 |
|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 68.69951 | 62.96091 | 57.2271 | 63.09509 | 38.21731 |

1 row

Figure 11 Displaying Variance values of each model.

We conclude that model 5 has the lowest variance value (38.21731) of all the models shown and is most likely the best fix from this procedure.

We then display the log-likelihood values.



Description: df [1 x 5]

| model1 | model2 | model3 | model4 | model5 |
|---|---|---|---|---|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| -33810.99 | -33393.69 | -32936.88 | -33403.88 | -31005.4 |

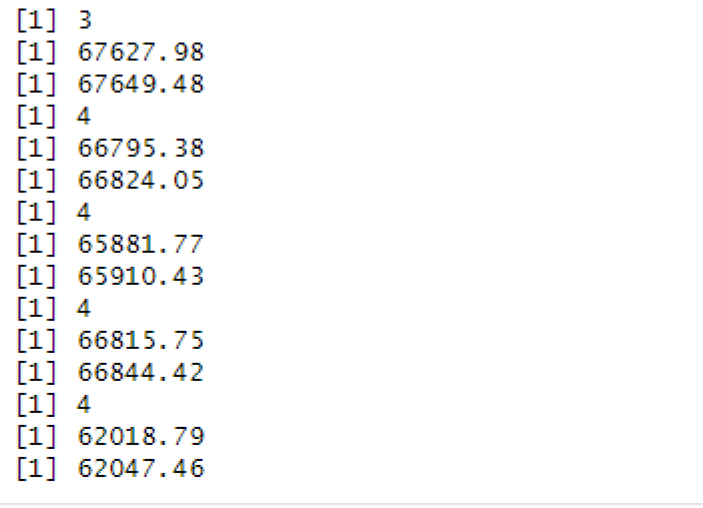
1 row

Figure 12 Displaying Log-likelihood values of each model.

We conclude that model 5 is the closest to 0(-31005.4) of all the models shown and is most likely the best fix from this procedure.

## Task 2.4 - Compute Akaike information criterion (AIC) and Bayesian information criterion (BIC) of all the models

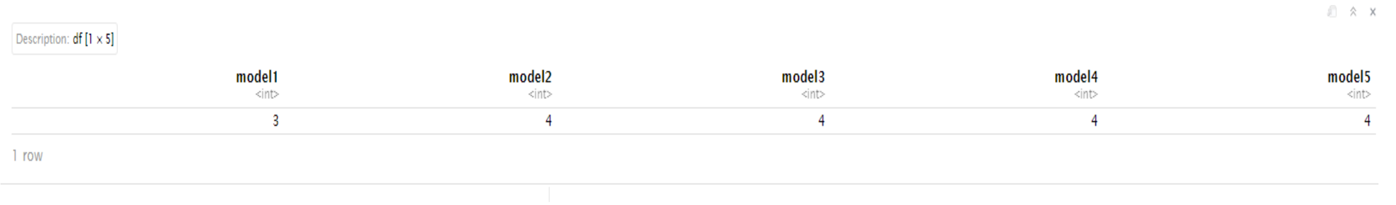
AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are both model selection criteria. These help us to figure out which model is suited better among all the models. Theta-hat of the models contain the estimated regression coefficients for Models — including: intercept (added by ones) and the coefficients for the predictors so, length (ModelX_thetahat) gives K, the total number of parameters.



```
[1] 3
[1] 67627.98
[1] 67649.48
[1] 4
[1] 66795.38
[1] 66824.05
[1] 4
[1] 65881.77
[1] 65910.43
[1] 4
[1] 66815.75
[1] 66844.42
[1] 4
[1] 62018.79
[1] 62047.46
```

Figure 13 Displaying k values, AIC values and BIC values.

Displaying K values from above



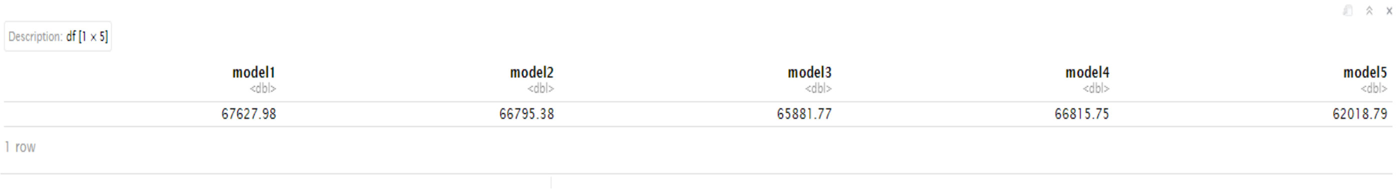
Description: df [1 × 5]

| model1<br><int> | model2<br><int> | model3<br><int> | model4<br><int> | model5<br><int> |
|---|---|---|---|---|
| 3 | 4 | 4 | 4 | 4 |

1 row

Figure 14 Displaying k values.

Displaying the AIC values



Description: df [1 × 5]

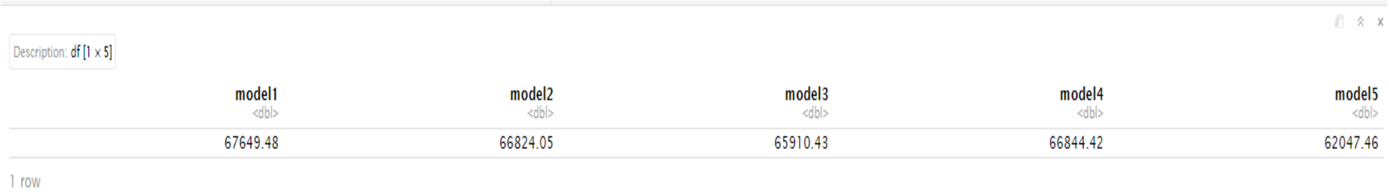
| model1<br><dbl> | model2<br><dbl> | model3<br><dbl> | model4<br><dbl> | model5<br><dbl> |
|---|---|---|---|---|
| 67627.98 | 66795.38 | 65881.77 | 66815.75 | 62018.79 |

1 row

Figure 15 Displaying AIC values.

We conclude that model 5 has the lowest AIC Value (62108.79) of all the models shown and is most likely the best fix from this procedure.

Displaying the BIC values



Description: df [1 x 5]

| model1<br><dbl> | model2<br><dbl> | model3<br><dbl> | model4<br><dbl> | model5<br><dbl> |
|---|---|---|---|---|
| 67649.48 | 66824.05 | 65910.43 | 66844.42 | 62047.46 |

1 row

Figure 15 Displaying BIC values.

We conclude that model 5 has the lowest AIC Value (62047.46) of all the models shown and is most likely the best fix from this procedure.

### Task 2.5 - Check error plotting normal/Gaussian distribution of each plot.

We continue by creating QQ plots for all the models. A Q-Q plot (Quantile-Quantile plot) is a graphical tool in statistics used to compare the quantiles of a dataset against the quantiles of a theoretical distribution (like the normal distribution). It's often used to check whether a dataset follows a specific distribution.

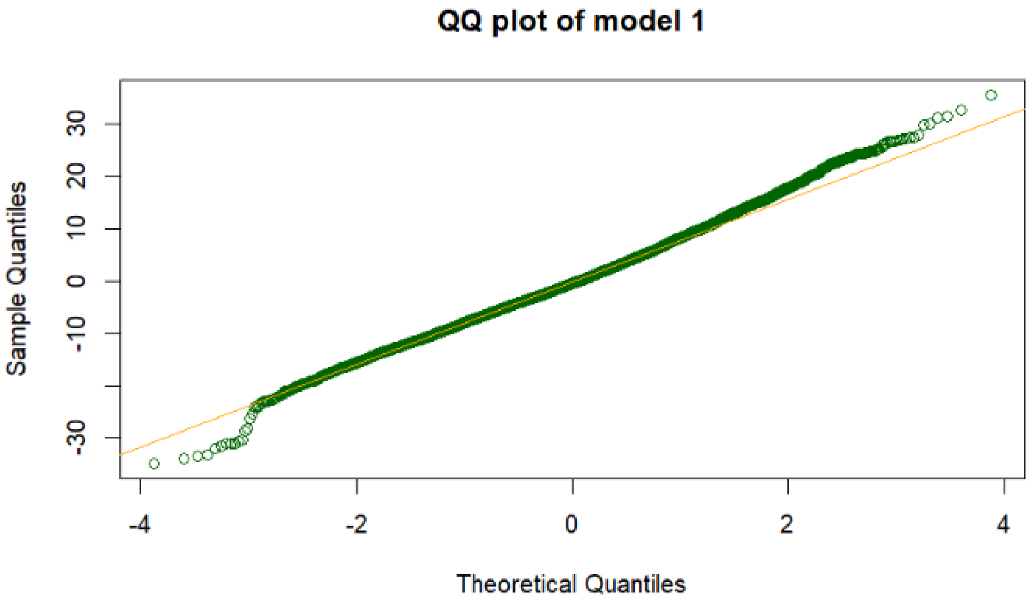


Figure 16 QQ plot of model 1

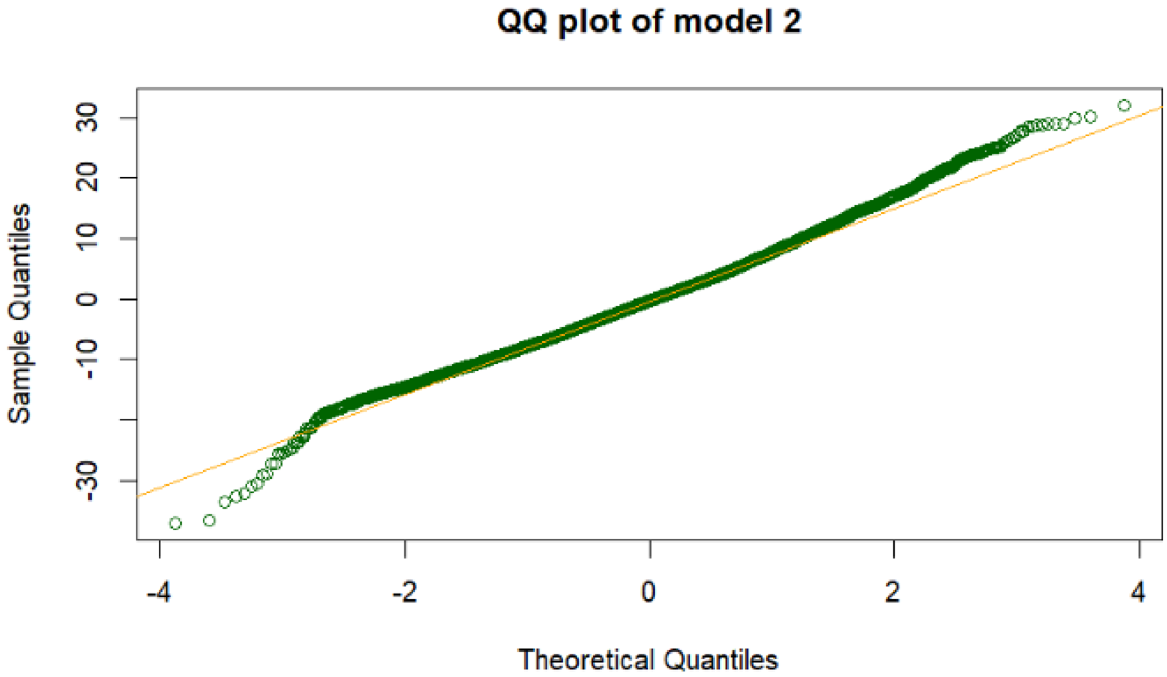


Figure 17 QQ plot of model 2

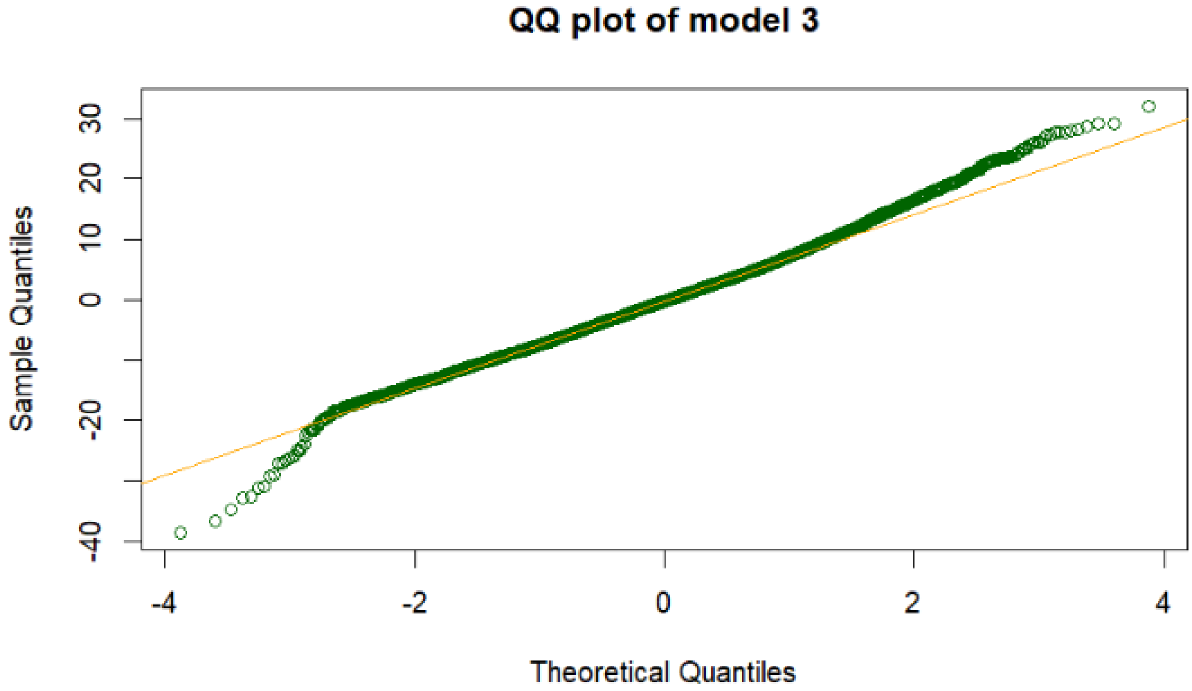


Figure 17 QQ plot of model 3

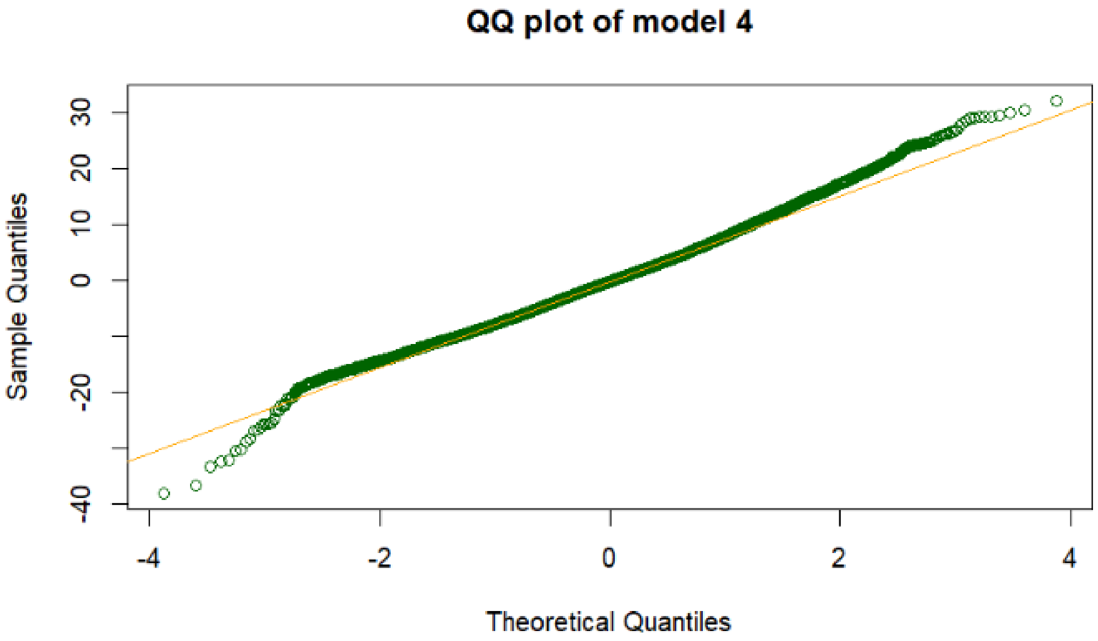


Figure 18 QQ plot of model 4

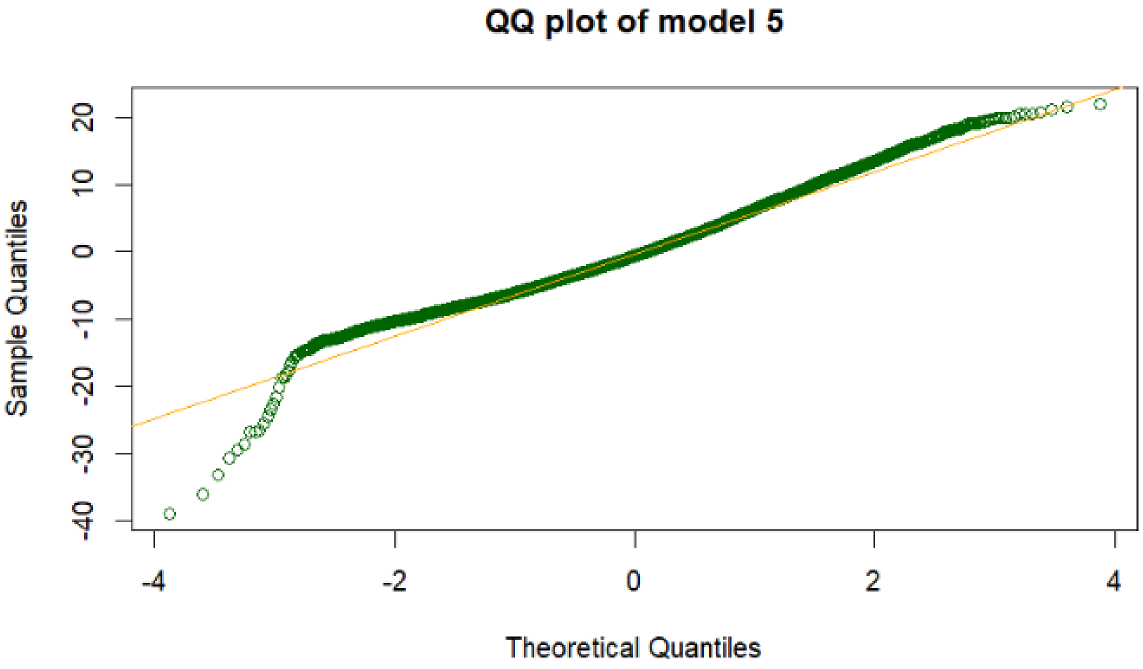


Figure 18 QQ plot of model 5

### Task 2.6 - Select ‘best’ regression model according to the AIC, BIC and distribution of model residuals

The best model up to now is model 5 as according to the AIC it had the lowest value according to BIC it is model 5 and according to distribution of model residuals (QQ Plot) it is still model 5 as it mostly follows the line and only has residuals in the left most part and is the most suitable one compared to the Q-Q plots of other models.

### Task 2.7 - Splitting the input and output dataset (X and y) into two parts:

We now proceed by splitting the input and output dataset into two parts. One part is used to train the model while the other is used for testing. This makes it so that the dataset is distributed in such a way that 70% of the data set is used for training whereas the remaining 30 % is used for testing purposes.

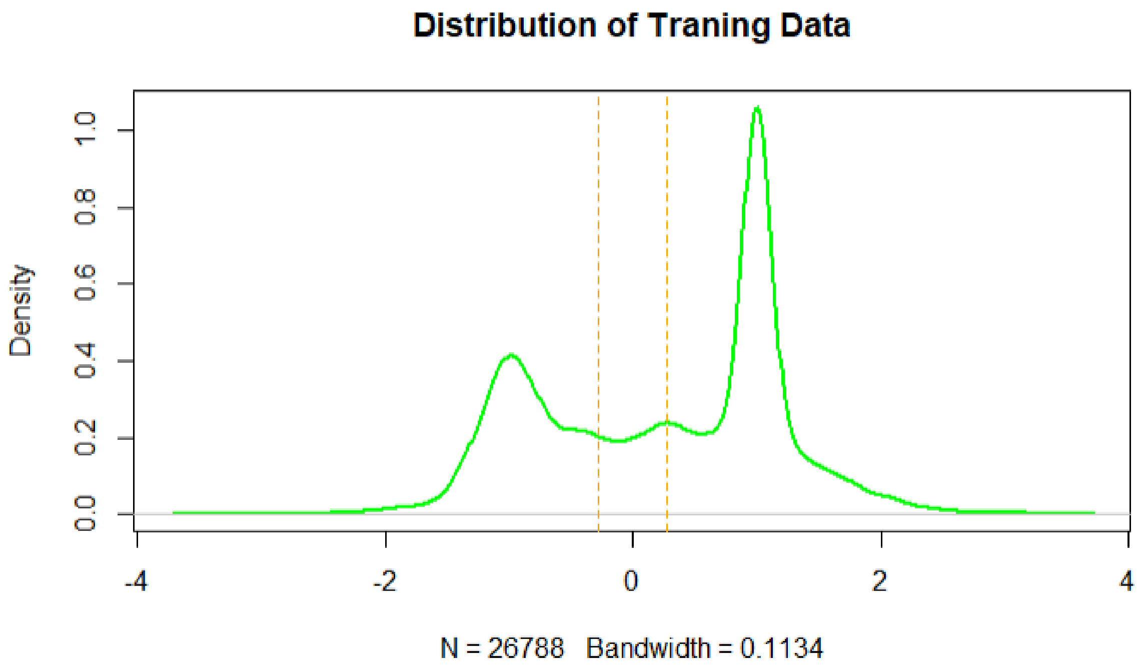


Figure 19 Distribution of training data

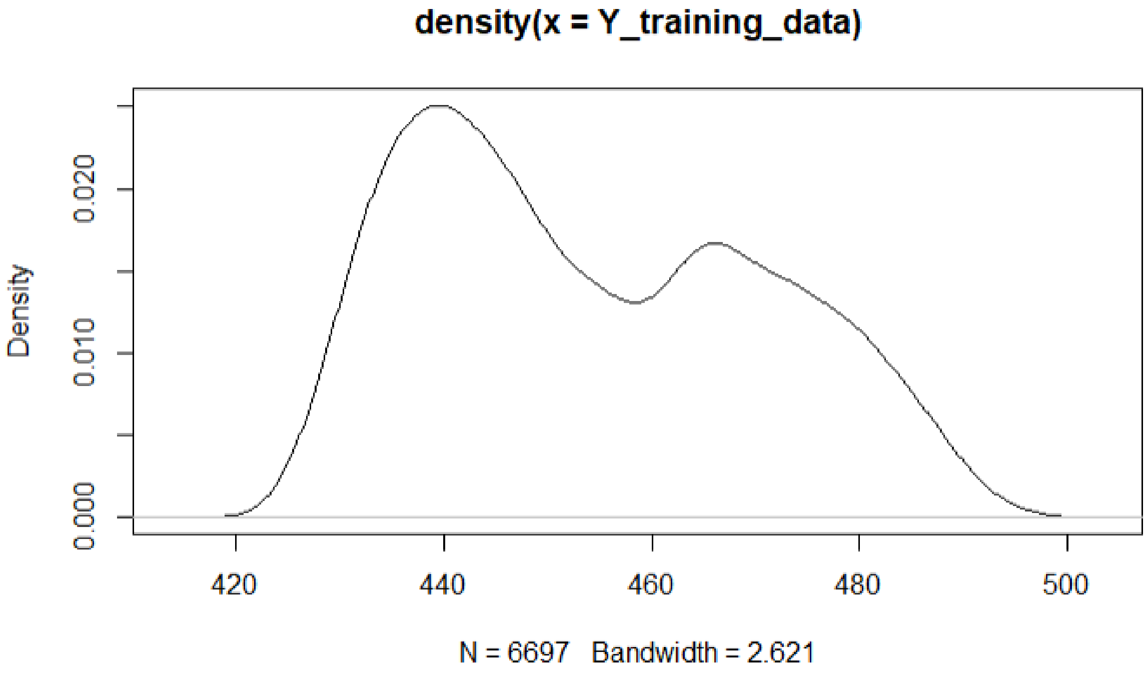


Figure 20 Density of training data

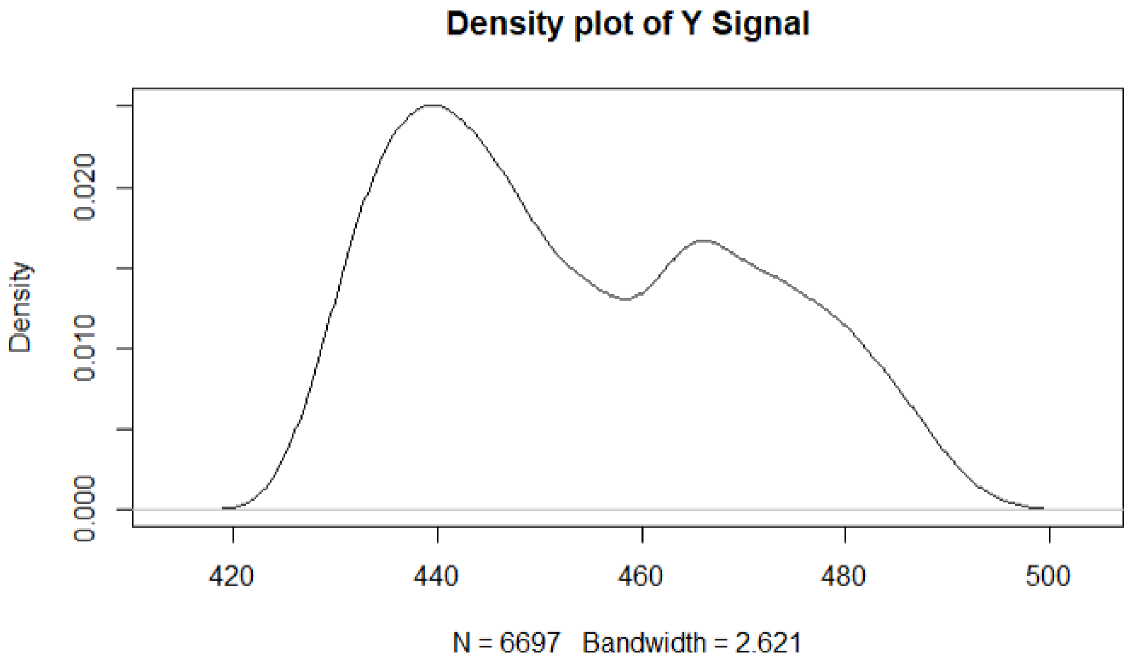


Figure 21 Density plot of Y signal

## Task 3: Approximate Bayesian Computation (ABC)

For the ABC, Model 5 will be used, parameter are selected and kept constant.

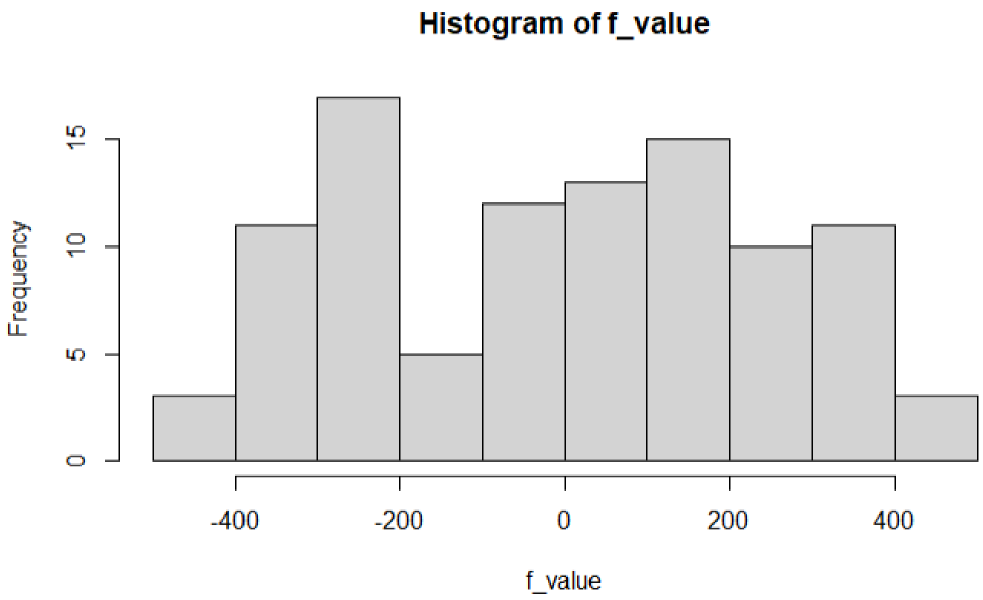


Figure 22 Histogram of f_value

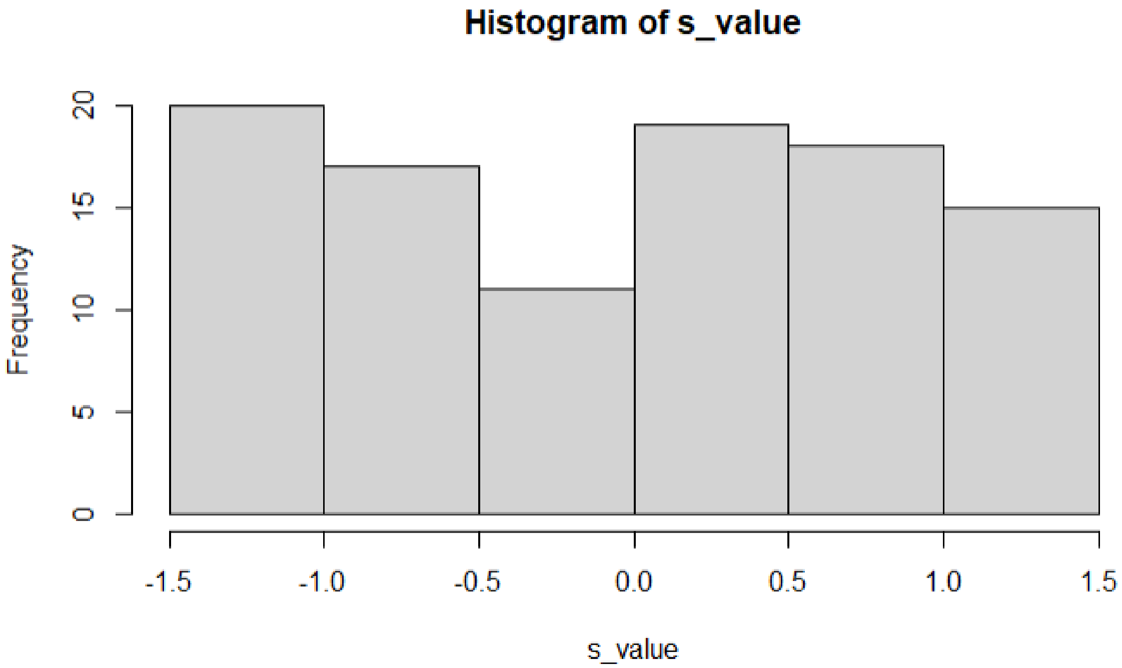


Figure 22 Histogram of s_value

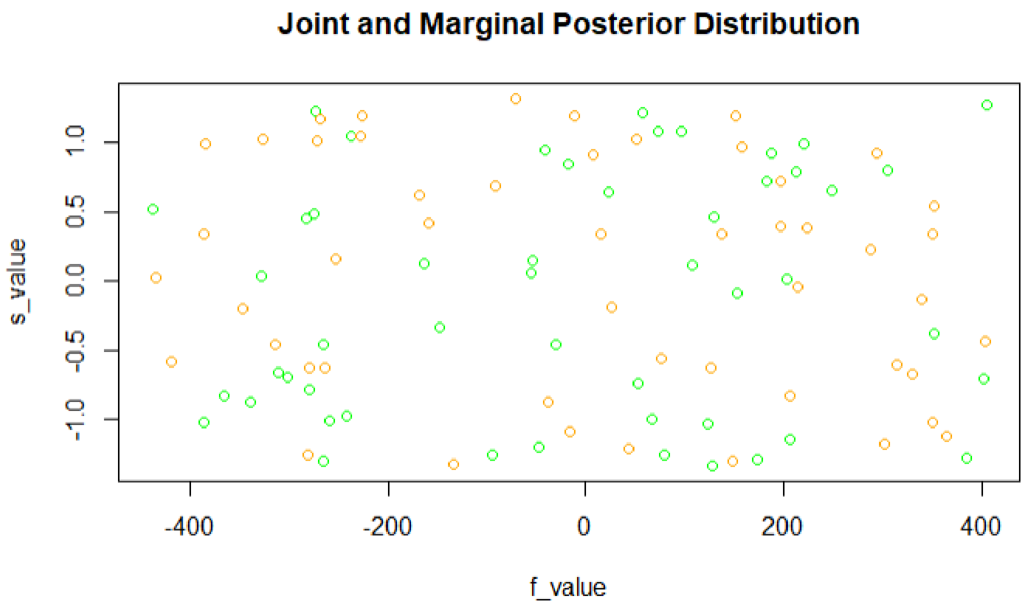


Figure 23 Joint and Marginal Posterior Distribution