INTRODUCTION TO STATISTICAL METHODS FOR DATA SCIENCE

SOFTWARICA COLLEGE OF IT & E-COMMERCE

MODELING BRIAN RESPONSES TO MUSIC USING NONLINEAR REGRESSION

Table of Contents

[INTRODUCTION 2](#_Toc185227937)

[TASK 1: PRELIMINARY DATA ANALYSIS 3](#_Toc185227938)

[Time series plots 3](#_Toc185227939)

[Distribution for each signals (time- series) 7](#_Toc185227940)

[Correlation Plot 10](#_Toc185227941)

[Scatter Plot 11](#_Toc185227942)

[Task 2: Regression – modelling the relationship between signals 15](#_Toc185227943)

[Task 2.1: Estimation of Model Parameters 15](#_Toc185227944)

[Task 2.2: RSS 17](#_Toc185227945)

[Task 2.3: Log-likelihood Function 18](#_Toc185227946)

[Task 2.4: AIC and BIC 18](#_Toc185227947)

[Task 2.5: Distribution of model prediction errors 20](#_Toc185227948)

[Task 2.6: Selection of regression model 24](#_Toc185227949)

[Task 2.7: Model Training and Testing 25](#_Toc185227950)

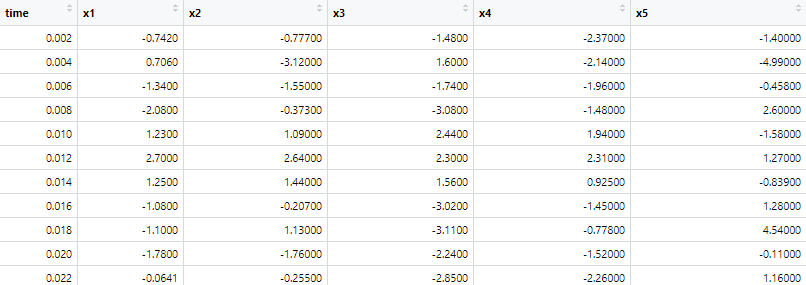
[Task 3: Approximate Bayesian Computation (ABC) 27](#_Toc185227951)

[References 28](#_Toc185227952)

# INTRODUCTION

Because music can effectively convey emotions via musical performances and evoke a range of feelings in listeners, listening to music is a generally cherished and well-liked pastime. At the most fundamental level, these feelings can be anything from pleasant to terrible, but music can also arouse more specialized negative and positive feelings in listeners, including fear, grief, joy, pleasure, and positive emotional goosebumps. In addition to evoking intense subjective emotions in listeners, music activates the human brain's extensive network for processing and identifying emotions. Since music is a nonverbal and apparently direct form of emotional communication, it can trigger neural processing for affect detection in many of these subsystems. Affective processing often involves a larger cortical and subcortical network of brain systems (Trosta, Trevora, Fernandeza, Steinera, & Frühholza, 2024).

Currently, for analysis we have been provided with 4 types of Input Signals (x1, x3, x4, x5) and (x2) as output signal in the given time period as below:



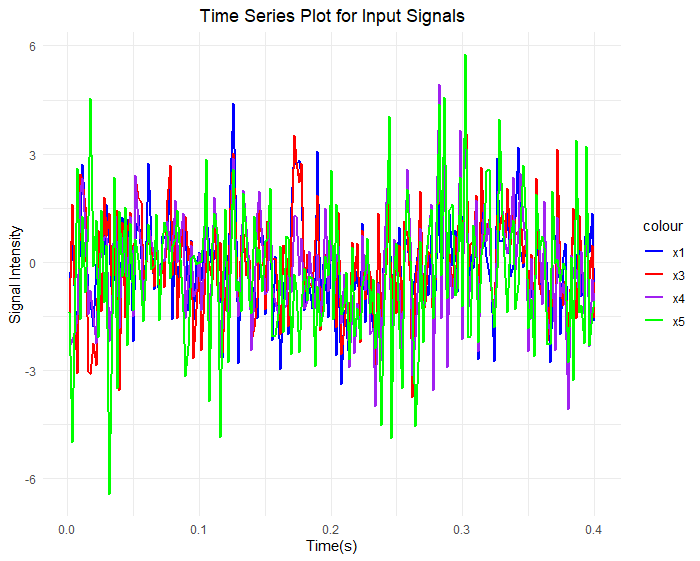
# TASK 1: PRELIMINARY DATA ANALYSIS

## Time series plots

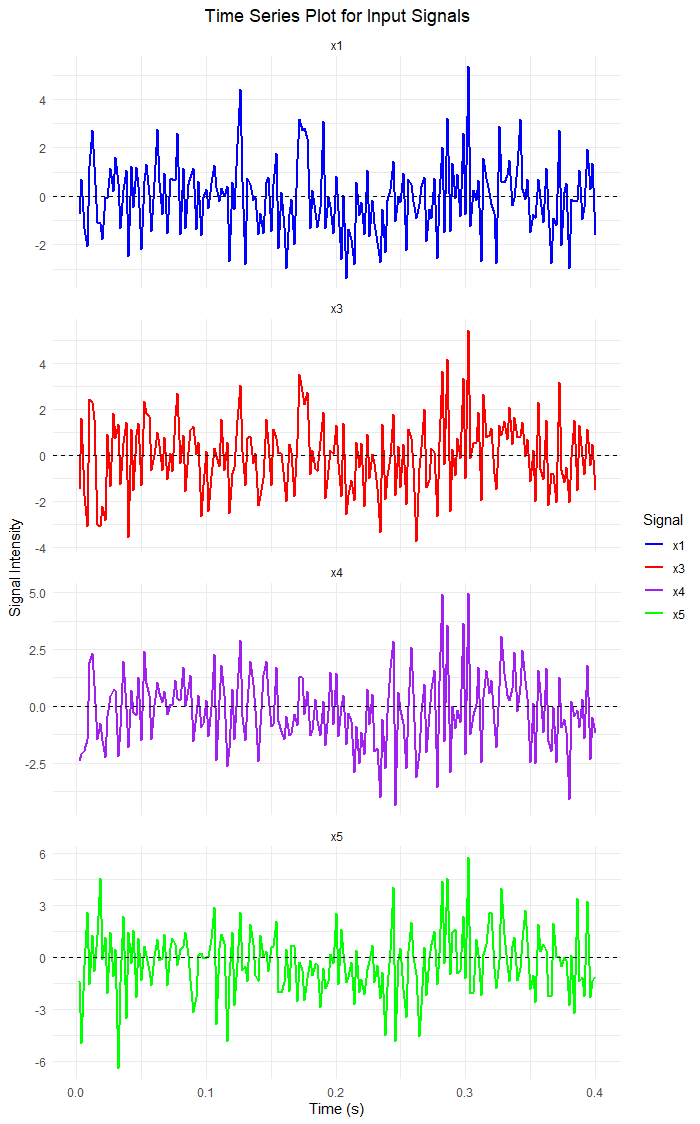
The most basic graphical representation of a time series {y(ti): i = 1,...,n} is as a time plot of the observed values y(ti) against the times of observation ti (Diggle & Giorgi, 2024). The given data consists x1, x3, x4, x5 as Input Signals, x2 as Output and time is in equal intervals for each trial.

Time series plot for input signals:

For the below visualization, each trial have their intensity fluctuated during the increment of time period for the trials.



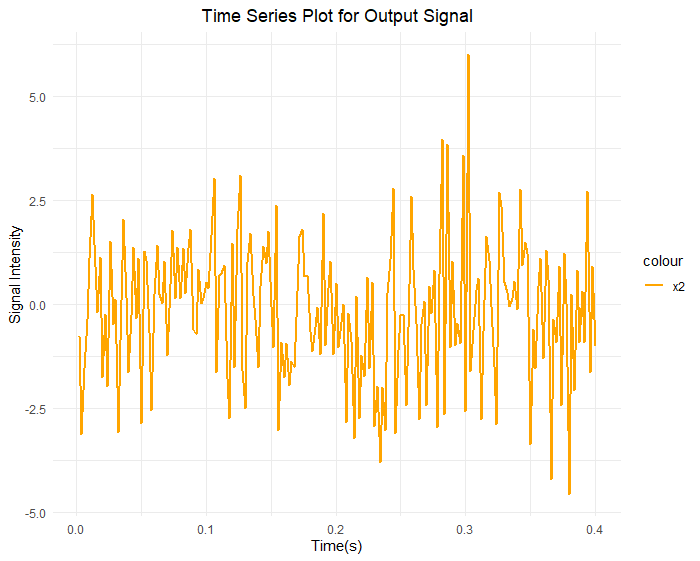
From the above time series plot, we can visualize that x5 trial signals are more fluctuated than others. For more clarification of each time series of each signals, below is the time series plot for each individual signals.



In this time series plot, we can clearly identify the variations for each signals are different. x1(blue) signal shows a mild fluctuations, x3(red) signal have larger spikes in intensity than other signals, likewise x4(purple) signal is locating to the baseline and x5(green) signal has similar pattern as x1 but have more fluctuations.

Time Series plot for output signal

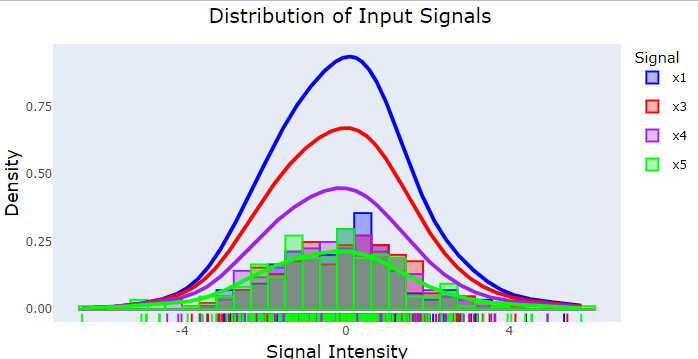
In the given data, x2 as an Output Signal while analyzing to support input signals it shows moderate fluctuations and is including combined patterns from all of the input signals.



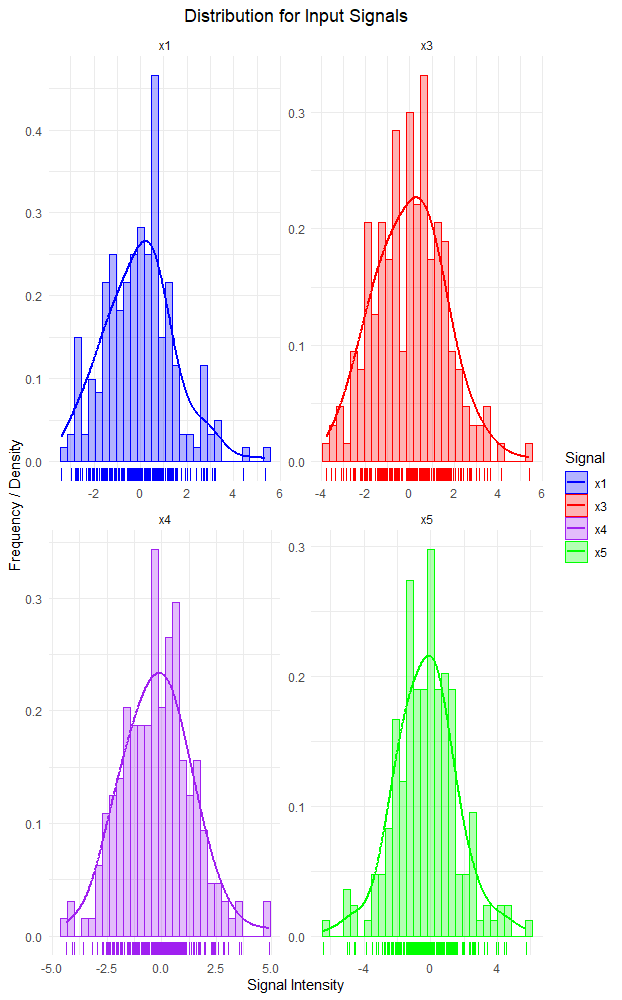
## Distribution for each signals (time- series)

To dig down the analysis for understanding its spread values, Distribution Analysis supports to identify overall shape of data and its central tendency.

To further analyze the dataset, below are the histograms including with their density curve.

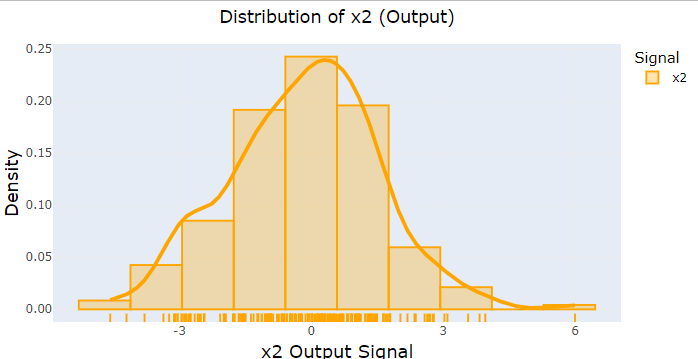


From the above Distribution table of the input signals, x1 signal has more density, then x3, x4 and x5 so on. All the distributions have somewhat bell shaped curve which means that the data are centered towards origin. The above distribution table is also plotted separately for each signals to summarize deeply as below:



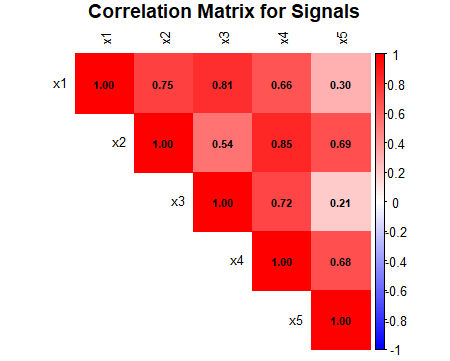
From the above diagram, it states that x1 and x3 are skiwed to the right side of the center whereas x4 and x5 are likely centered. Based upon the high point of curve, we can state the signals are mostly centered and scattered a bit to positive values.

Again, Distribution Analysis is performed for the output signal, which supports our findings above for the input signals as well. The output signal try’s to follow normal distribution table which means that the signals are centered towards origin.



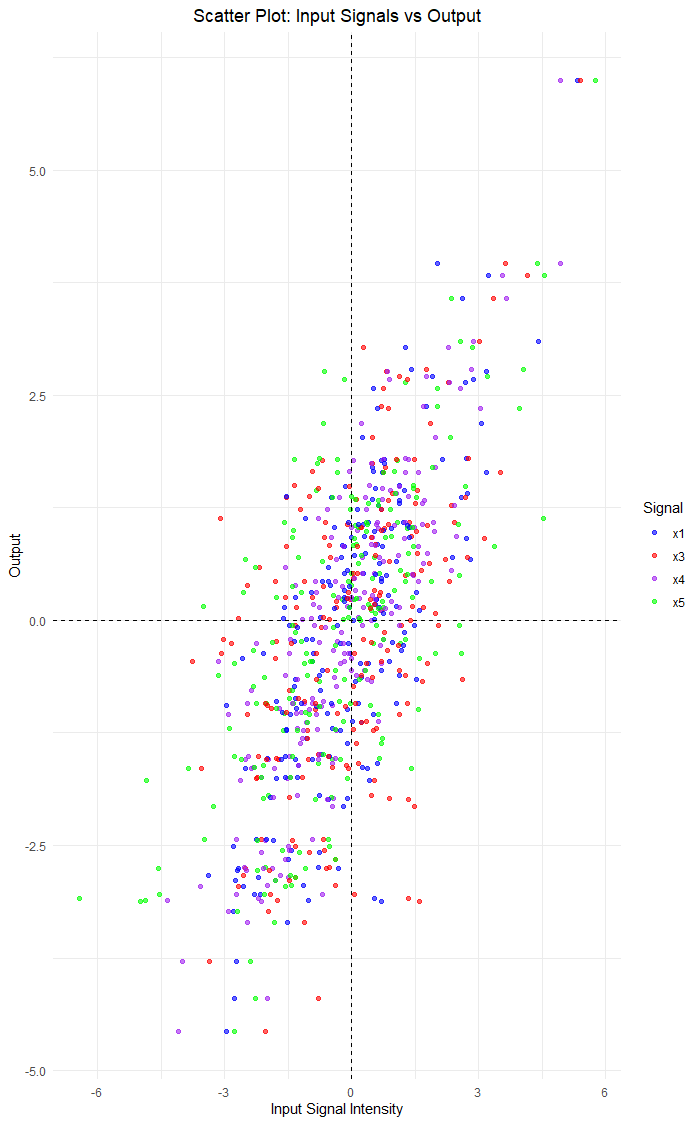
## Correlation Plot

When dealing with quantitative data, the Pearson correlation coefficient is employed. Both positive and negative numbers, as well as zeros, can be considered data. There are instances where all of the data values are negative, however they are uncommon. We may compute the correlation by simply dropping the signs of the data values, as corr(X,Y) = Corr(-X, -Y). To make the calculation easier, we may use Corr(-X,Y) = Corr(X,-Y) = -Corr(X,Y) if either variable has a majority of negative values (Chattamvelli, 2024). Below is the correlation matrix among between input signals and output signal.

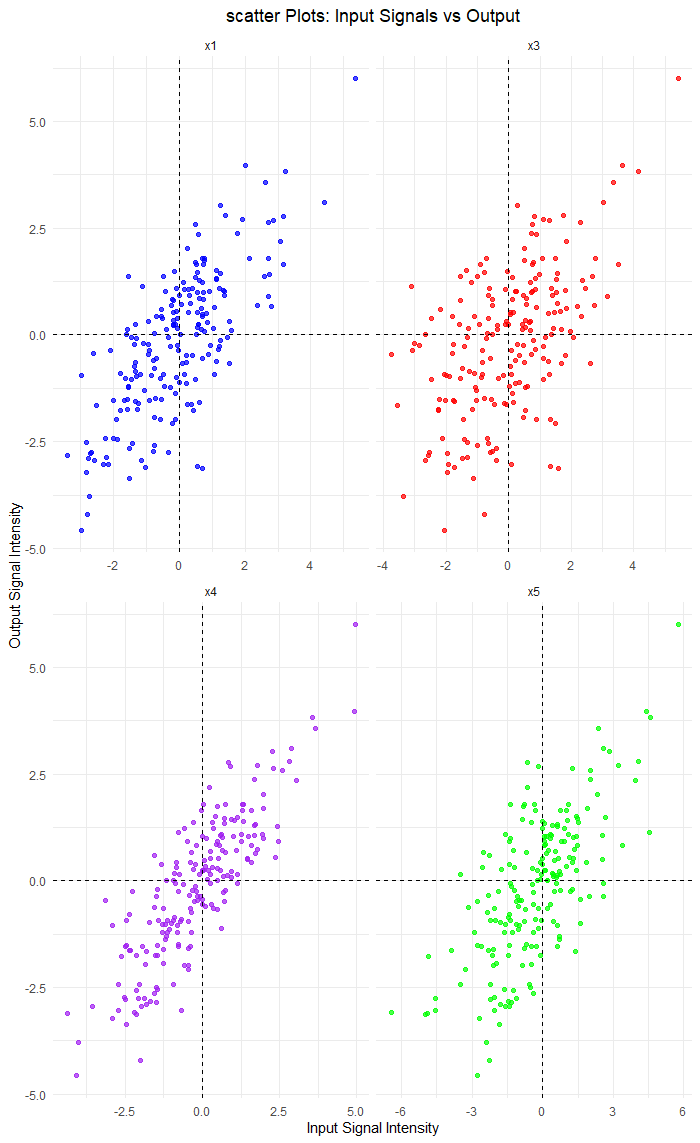


## Scatter Plot

As Scatter Plot makes more meaningful decisions on analyzing data distributions (Foo, Wong, & Goh, 2024). Below is the scatter plot between all the input signals in respect to output signals.



To make easier on examining the dependencies between each signal and output signal, below it is plotted separately for each (X) input signals in respect to (Y) output signal.



# Task 2: Regression – modelling the relationship between signals

A regression model to determine a suitable mathematical model in explaining the relationship between the input and output signals (i.e. x1, x3, x4, x5) that ‘regulate’ its expression. For this we have following candidate models:

Model 1: y = θ1x4 + θ2x32 + θbias

Model 2: y = θ1x4 + θ2x32 + θ3x5 + θbias

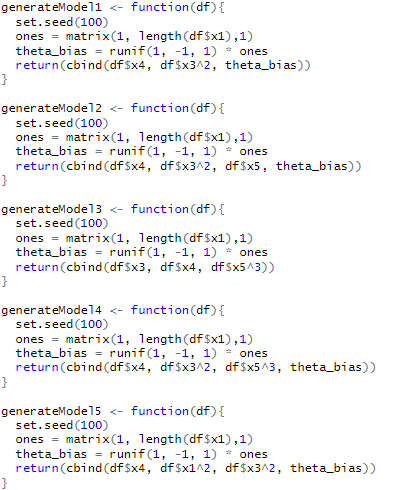
Model 3: y = θ1x3 + θ2x4 + θ3x53

Model 4: y = θ1x4 + θ2x32 + θ3x53 + θbias

Model 5: y = θ1x4 + θ2x12 + θ3x32 + θbias

## Task 2.1: Estimation of Model Parameters

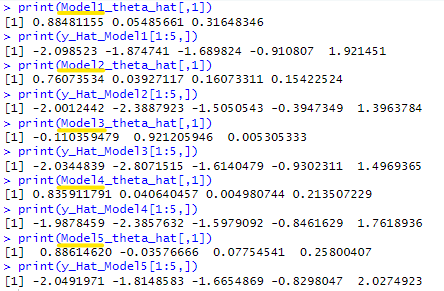
From the above candidate models, in Model 1 there are x4, x3 as input parameter selected from fMRI brain images signals and θbias is the constant value selected from between -1, 1.



The Least Squares equation is also denoted as **“**= **(XTX)-1XTy”** where X is the model estimator and Y is output signal from brain responses. In R, it can be written as:

****

The output of theta hat and y hat for each model is calculated and are given below:



## Task 2.2: RSS

RSS is a statistical measure that is used to evaluate the goodness of fit of a regression model to data. It calculates the total squared difference of the predicted values of the model from the actual values.

where yi is the observed value for the ­ith data point

is the predicted value for the ­ith data point from the model

n is the number of data points

In R, RSS can be written as below:



And, the value of each model is obtained as below:

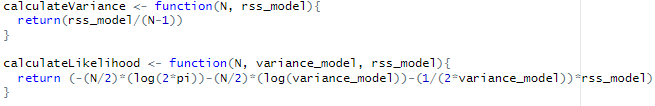


## Task 2.3: Log-likelihood Function

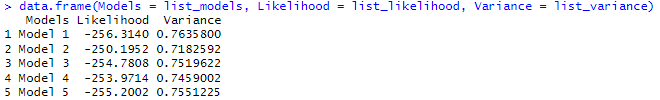
The log-likelihood function is a measure used in statistical modeling to calculate the likelihood of observing the given data under a specific model. Assuming that the residuals or errors are normally distributed, for regression models, a log-likelihood function can be derived based on the likelihood of the observed data.



This can be written as below in R programme:



The likelihood and variance for each models is obtained as below:



## Task 2.4: AIC and BIC

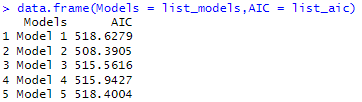
AIC is a statistical tool that allows one to make comparisons and assess the relative quality of regression models. It balances model fit and model complexity to find the best model without overfitting.



In R, AIC equation can be represented as below:



The AIC for each model is obtained as below:



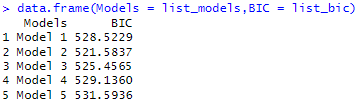
Another statistical measure applied for comparing models is BIC (Bayesian Information Criterion). It too involves a balance between the model fit and its complexity. However, BIC applies a heavier penalty on model complexity than AIC; hence, it tends to be more conservative when comparing models.



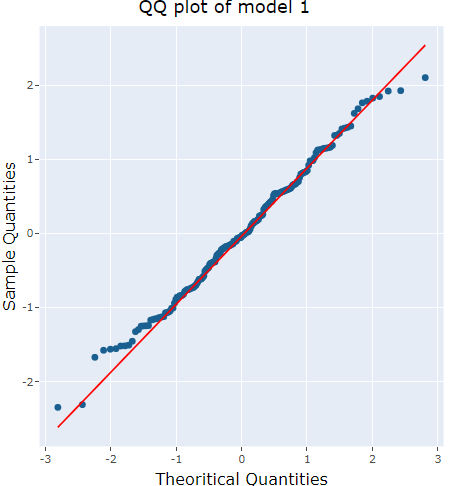
In R, BIC equation can be represented as below:

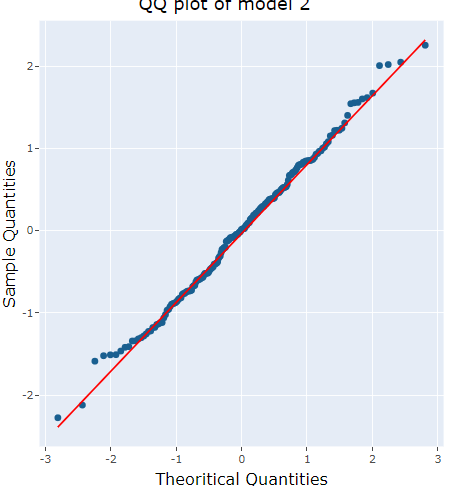


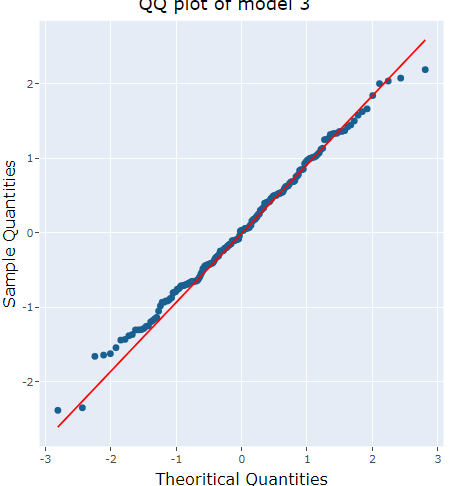
The BIC for each Models is obtained as below:

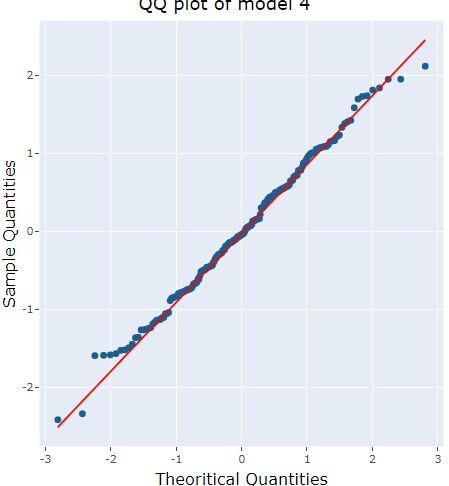


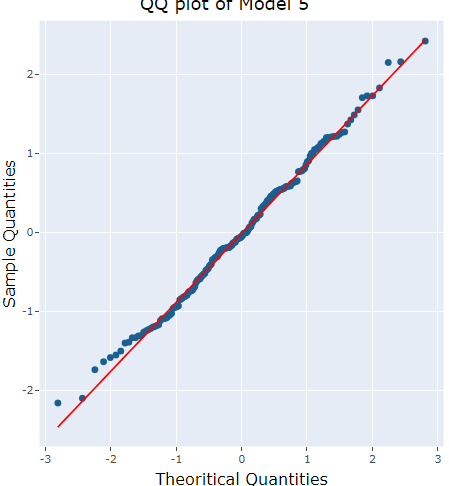
## Task 2.5: Distribution of model prediction errors





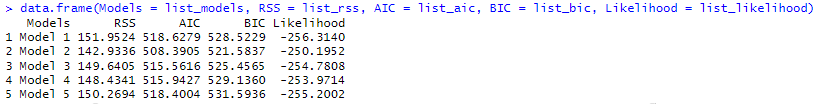






## Task 2.6: Selection of regression model

Analyzing the model based on AIC, lower AIC value denotes better model and Model 2 has the lowest AIC value i.e., 508.3905. Same as Lower BIC value denotes better model and Model 2 has the lowest BIC value i.e., 521.5837.

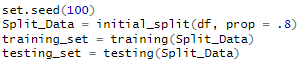


Because of lowest AIC and lowest BIC value, Model 2 is the best model as it strikes the good balance among model fit and complexity. Beyond this reason, Model 2 also includes x4, x32,x5 parameters which suggest it incorporates meaningful predictors without overfitting.

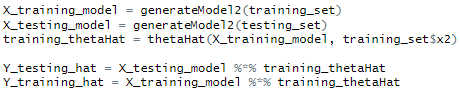
If the residuals of Model 2 are no significantly different in normality and independence, then the model fulfills the requirements of a proper regression model.

## Task 2.7: Model Training and Testing

After Model selection, X inputs and Y outputs have been split into 80 and 20 parts using initial\_split function in R.

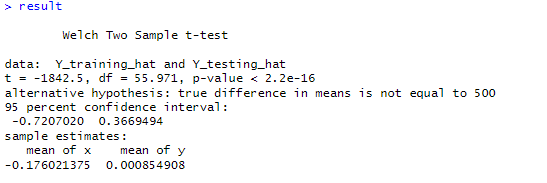


Then, the model parameter θ is calculated, to evaluate the training model into testing model.

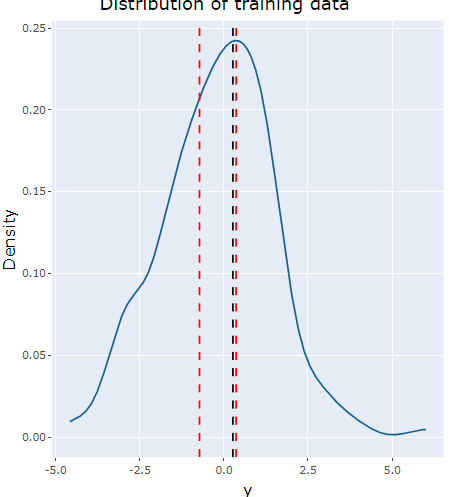


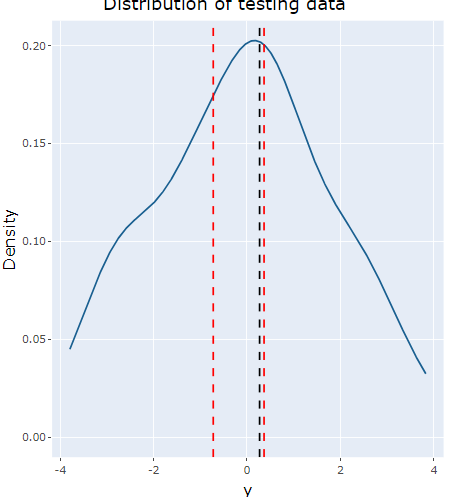
Again, the confidence interval is calculated as below:





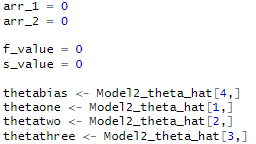
Once the confidence level is set up and evaluated at 95 percent for both training hat and testing hat.



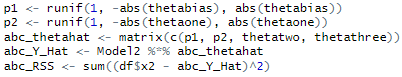


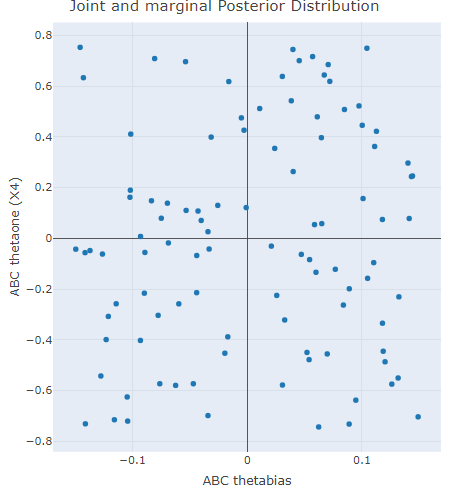
Two red vertical lines represent the confidence interval of 95 percent for both testing and training dataset which are -0.7207020 and 0.3669494 respectively, while black vertical line represents Standard error of prediction which is of 0.5311649.

# Task 3: Approximate Bayesian Computation (ABC)









# References

Chattamvelli, R. (2024). *Correlation in Engineering and the Applied Sciences.* doi:10.1007/978-3-031-51015-1

Diggle, P. J., & Giorgi, E. (2024). *Time Series* (2 ed.). Oxford University Press.

Foo, R., Wong, L., & Goh, W. W. (2024). What can scatterplots teach us about doing data science better? *International Journal of Data Science and Analytics, 2*, 111-125. doi:10.21203/rs.3.rs-1733113/v1

Trosta, W., Trevora, C., Fernandeza, N., Steinera, F., & Frühholza, S. (2024). Live music stimulates the affective brain and emotionally. *Proceedings of the National Academy of Sciences, 121*, 10. doi:e2316306121