****

**Modeling shopping mall sales data using Polynomial Regression**

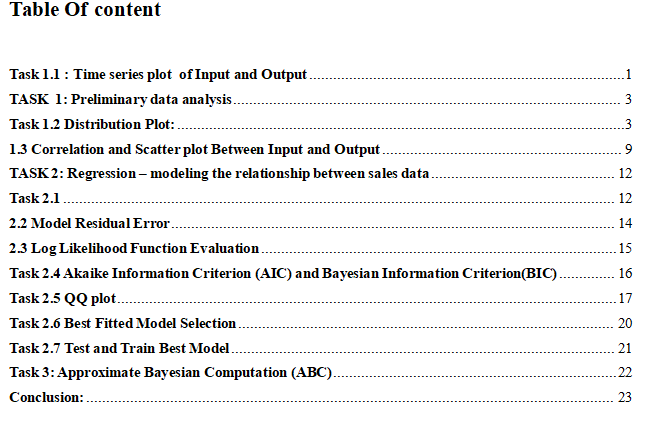
**Git Link:**

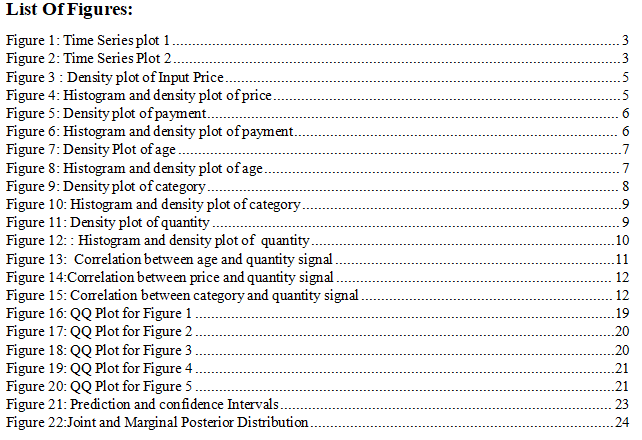
**Video Link:**

**Submitted By:**

**Anjana Jha**

**Coventry Id: 14825147**







# 

# **Task 1.1 : Time series plot of Input and Output**

A time series is a group of observations gathered over a specific amount of time. A graph with uniformly spaced time intervals between observations that displays data collected over an extended period of time is called a time series plot. In a timeline plot:  
The input signals are shown on the x-axis of the following figure with regard to the date on the y-axis.   
A time series plot, for instance, can display the total sales quantity over a period of days, months, or years in the context of client shopping data, assisting in the identification of patterns like weekly or seasonal swings in sales.

A line graph with text

Description automatically generated

Figure 1: Time Series plot 1

A graph of a line graph

Description automatically generated

Figure 2: Time Series Plot 2

To ensure that the time series plot is reliable, it is crucial to carefully examine it and search for any observed values that show unusual swings or depart from the expected range. The dataset has to be thoroughly inspected and verified because these anomalies could indicate mistakes or contaminated data.

**In figure 1:**

* It is a plot of four input signals represented according to the invoice date.
* Differences in the data plot are seen for each of the several input signals.
* The price input signal exhibits more fluctuations than the other signals.
* The payment method input shows a consistent trend over time, however the age and category inputs show distinct patterns over an extended period of time.

**In figure 2:**

* The output signals are represented by grouping them by year and month.
* There is some observable variation in the data.
* There is a sharp decline in signal around the 2023 limit, indicating that the overall quantity will decline following years of stability.
* Not much noise is apparent.

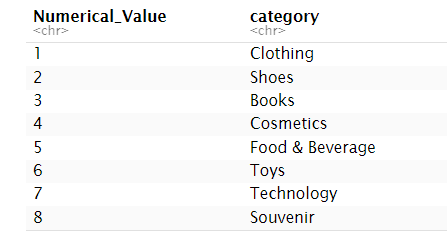
# **TASK 1: Preliminary data analysis**

A text on a white background

Description automatically generated

# **Task 1.2 Distribution Plot:**

A distribution plot is only a visual representation of a data set that compares observed and expected values for data types that share a particular characteristic.



For Input Signals:

1. **[Input = Price]**

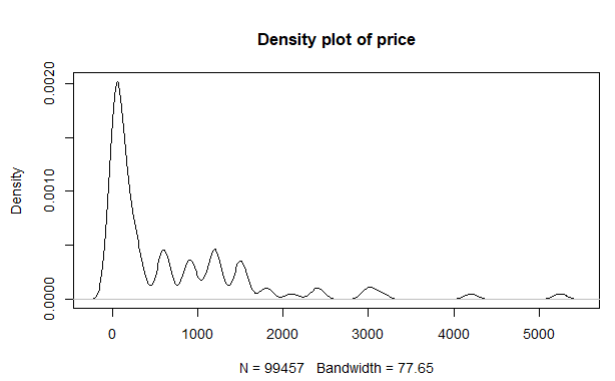


Figure 3 : Density plot of Input Price

* It appears that the price input signal follows a symmetrical curve.
* The observed bandwidth is 77.65.
* It is noted that the peak distribution is close to zero.

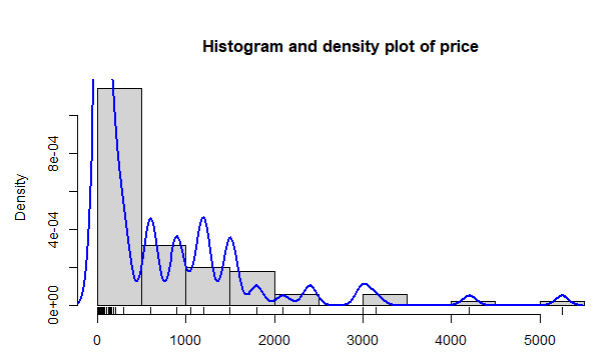


Figure 4: Histogram and density plot of price

Histogram and Density plot of Price:

* It has been noticed that the price input signal with the greatest value can range from 0 to 500, while the lowest value is represented by the remaining length.
* The majority of the input signals appear to be in the 500–3000 range, and the distribution starting at 500 is more dispersed than the other input signal ranges.

1. **[Input = Payment]**

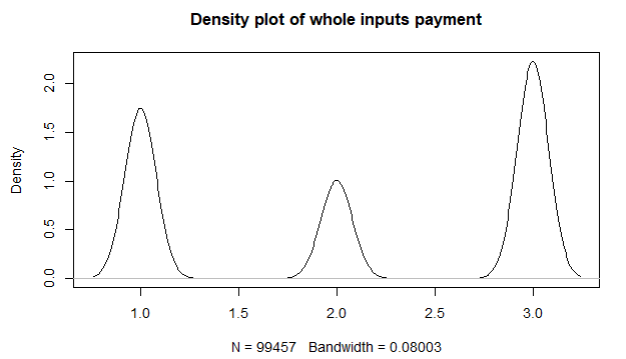


Figure 5: Density plot of payment

Density plot of Input Payment:

* The observation explains why there is nearly a symmetrical curve in the density plot of all the input signals for Payment.
* The observed bandwidth is 0.08003.
* It is noted that the peak distribution is close to 3.0.
* The skewness of the distribution is examined based on the fact that it is essentially screwed—that is, it is neither left nor right screwed.

A graph of a graph of payment method

Description automatically generated

Figure 6: Histogram and density plot of payment

**Histogram and Density plot of input signal**

* According to the diagram, the maximum input send would be at 3, with the others being lesser in comparison.
* Furthermore, a comparison between the plots and the density plot reveals that they are nearly symmetrical.
* Additionally, it is noted that the majority of input signals are spread along the numbers, with the distribution beginning at 0.5 and going up to 3.5 in accordance.

**[Input = age]**

A graph of a function

Description automatically generated

Figure 7: Density Plot of age

**Density plot of Input Age:**

* The observation explains why there is an unsymmetrical curve in the density plot of all the input signals for Payment.
* The observed bandwidth is 1.351. The input is dispersed approximately equally from number 15 to number 75, indicating the absence of a peak distribution.
* The skewness of the distribution is evaluated in order to evaluate its symmetry. A graph of a graph of age

  Description automatically generated

Figure 8: Histogram and density plot of age

**Histogram and Density plot of input signal**

* The density plot's age plots are shown in a continuous manner across the figure, and the plot's near-unsymmetrical nature is observed.
* It has been seen that nearly all of the inputs are collected between plots 20 and 70. Furthermore, a comparison between the plots and the density plot reveals that they are nearly symmetrical.

1. **[Input = category]**

A graph of a function

Description automatically generated

Figure 9: Density plot of category

**Density plot of Input Category:**

* The insight clarifies why the input signal density map for each category is nearly symmetrical.
* The observed bandwidth is 0.2022. The observed peak distribution is close to 1.0.
* The skewness of the distribution is examined based on the fact that it is essentially screwed—that is, it is neither left nor right screwed.

A graph with a blue line

Description automatically generated

Figure 10: Histogram and density plot of category

**Histogram and Density plot of input signal**

* The diagram depicts the highest input send would be at 1 where as others are smaller comparatively.
* It is also observed that the plots are almost symmetrical in comparison with the density plot.
* It is also observed that most input signals are distributed along the numbers and the distribution starts from 0.5 up to 8 gradually.

1. **[Input = Quantity]**

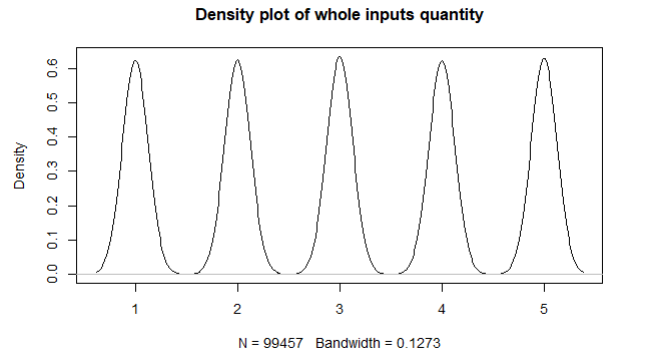


Figure 11: Density plot of quantity

**Density plot of Input Quantity:**

* The observation explains why the input signal density plot for each category is almost symmetrical.
* The observed width is 0.1273. The distribution of the curves created in the various plots is uniform.
* In order to ascertain the symmetry of the distribution, its skewness is examined by determining whether it is almost twisted or not, either to the right or left.

A graph of a function

Description automatically generated

Figure 12: : Histogram and density plot of quantity

**Histogram and Density plot of input signal**

* The diagram depicts the outputs appear equally distributed throughout the plot.
* It is also observed that the plots are almost symmetrical in comparison with the density plot.

**[\*\*Left Skewed => mean<median\*\*]**

**[\*\*Right Skewed => mean > Median\*\*]**

**[\*\*No Skew => mean= median\*\*]**

# **1.3 Correlation and Scatter plot Between Input and Output**

* A scatter plot is a type of visualization graph in which two numerical variables are plotted and their respective variables' relationships are displayed.
* There are three different kinds of correlation:  
  a) Positive: In a scatter plot, this occurs when one variable rises in proportion to the following variable.   
  b) Negative: In a scatter plot, this occurs when one variable falls as the following variable rises.   
   c)Non: In a scatter plot, this occurs when there isn't a discernible relationship between two variables.   
  The various scatter plots and their correlations are as follows:



Figure 13: Correlation between age and quantity signal

* This is weak positive correlation.



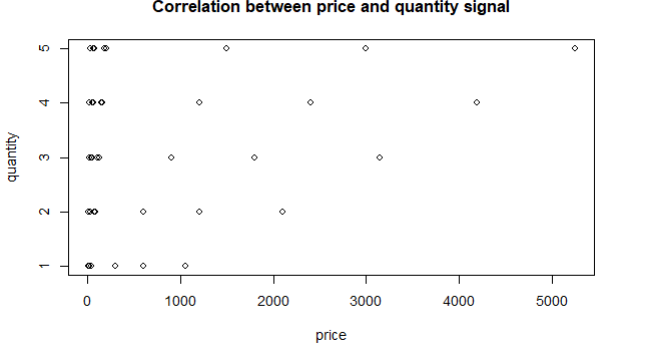


Figure 14:Correlation between price and quantity signal

* This is a strong positive correlation.



A graph with black dots and white text

Description automatically generated

Figure 15: Correlation between category and quantity signal

* This is a negative weak correlation.



A graph with numbers and lines

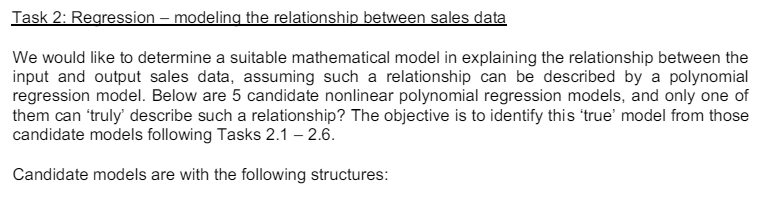
Description automatically generated with medium confidence

* This is negative correlation.



Plotting the individual quantitative values of age, price, category, and payment method in relation to the output value quantity yields a pattern with multiple circular dots, each of which identifies a unique piece of data on the graph that is further examined for correlation. Every pattern is carefully examined based on the three categories of correlation: non-correlation, positive correlation, and negative correlation.

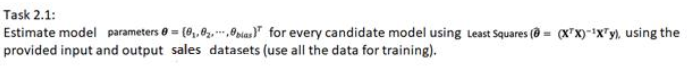
# **TASK 2: Regression – modeling the relationship between sales data**



A group of math symbols

Description automatically generated

# **Task 2.1**



In order to ascertain the various properties of the distribution, we will first designate 'θ' as the unknown variable, also referred to as the estimate, as the true value of the distribution is not given.

Given,

θ = {θ, θ1, θ2, …, θbias} ^T

We will compute the estimator model parameters for each candidate using Least Square.

By reducing the sum of the squares of the discrepancies between the observed and predicted values, the Least Squares approach, put simply, is a technique for determining the best-fitting line or curve for a set of data. The formula "θ = (X TX)^-1 X T Y" is used to calculate it, and it is represented by the symbol "θ". Minimizing the error between the actual values and the anticipated values (derived from the regression line) is our goal. In order to find the regression line that minimizes the error, the data points are plotted, and the Least Squares method is applied.

Since we are utilizing the R language, we are aware that the formula will be used to obtain "θ."

θ =solve(t(X) %\*% X) %\*% t(X) %\*% Y

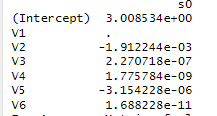
The value of the input data from the shopping mall dataset would first be bound.   
After that, we would bind ones using the input data.

X<-c cbind(ones,(x[,"X4"]),(x[,"X1"])^2,(x[,"X1"])^3,(x[,"X2"])^4,(x[,"X1"])^4)

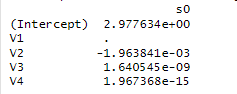
The least squares formula would be the last step in this process, and the possible outcome models would then be tabulated appropriately below.

For theta hat results, we have

1. For Model 1:



1. For Model 2



1. For Model 3

A number on a white background

Description automatically generated

1. For Model 4

A number and numbers on a white background

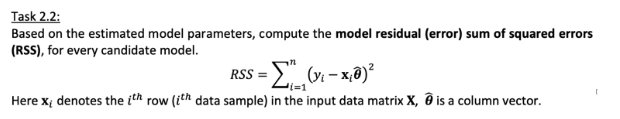
Description automatically generated

1. For Model 5

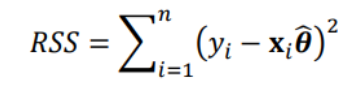
A close up of numbers

Description automatically generated

# **2.2 Model Residual Error**



**Model Residual Error (RSS):** Model Residual Error (RSS) also known as Sum Square error, RSS is considered one of the finest models for regression analysis because a large squared number indicates a large variation. This variation measure is computed using the sum of squares. This requires that we compute RSS, which is represented by:



Now that RSS is never negative, we use this formula to determine the inaccuracy of each candidate model, considering the fact that RSS is closer to 0.   
Since the R language is being used, we can use this formula to obtain RSS.

Given,

Yhat = product of Xiθ

N = total number of data points in Y

∑= sum of output Y

Y = output symbol

Sum = sum of every possible values.

Finally, each model with different θ value is put together and RSS value is obtained which is tabulated below.

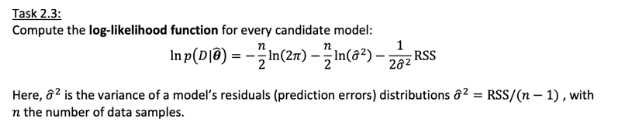
Ultimately, by combining all the models with varying θ values, the RSS value is calculated and presented in the table below.

A table with numbers and letters

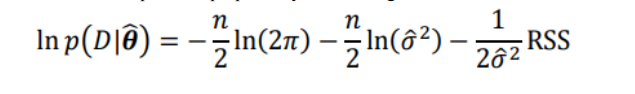
Description automatically generated

Table 1: RSS Values

# **2.3 Log Likelihood Function Evaluation**



* The log likelihood function, in its simplest form, is a technique for estimating how well the data fits the sample data and for determining the maximum likelihood estimator of a parameter.
* Formula is used to carry out this approach.



Its major objective is to determine the best technique to fit distribution to data with fixed observation value.

Given,

ln p(D|θ)= Log likelihood

n= total number of y signals

Ln= natural logarithm function(log)

σ 2=variance obtained from task 2.2

From RSS obtained it can be calculated as

σ 2 =RSS/(n-1)

Π=pi

RSS=Model Residual error



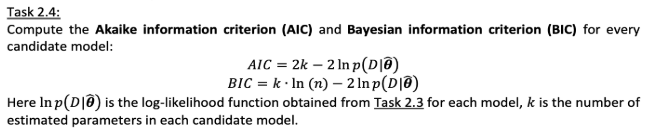
Finding the optimal method for fitting a distribution to data with a fixed observation value is its main goal.

A screenshot of a computer

Description automatically generated

Table 2: Likelihood

# **Task 2.4 Akaike Information Criterion (AIC) and Bayesian Information Criterion(BIC)**



**Akaike Information Criterion:** Akaike information criterion, to put it simply, is a statistical method for assessing how well a model mathematically matches the data.   
Generally, it is employed to evaluate various potential models and identify the most appropriate fit. To implement this strategy, a formula is used.



Given,

K= length of parameter from Task 2.1

L=likelihood function from Task 2.3

Given this formula we generate the Akaike Information Criterion

A table with numbers and symbols

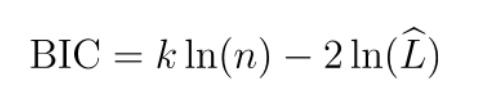
Description automatically generated

Table 3: AIC Score

**Bayesian Information Criterion:** To put it simply, the Bayesian Information Criterion is an additional statistical method for assessing similarity when choosing a model from a limited number of options.

It has already been applied to time series or line regression estimation.

* It is carried out by using the formula as below:



Given,

K=length of parameters to be identified from Task 2.1

L=likelihood function from Task 2.3

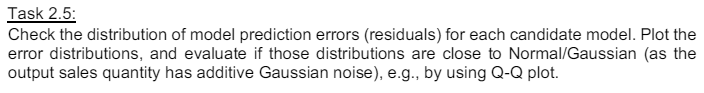
Given this formula we generate the Bayesian Information Criterion

A table with numbers and letters

Description automatically generated

Table 4: BIC Score

# **Task 2.5 QQ plot**



**Normal / Gaussian distribution** : The normal distribution, also known as the Gaussion distribution, is a probability distribution that is symmetric about the mean, meaning that data that are close to the mean are more common.   
Technical stock market analysis and other statistical dataset analyses are the main uses for it.  
According to the QQ plot, the data that most closely resembles the diagonal line with the fewest consistent deviations and the fewest troublesome patterns is probably the best fit.

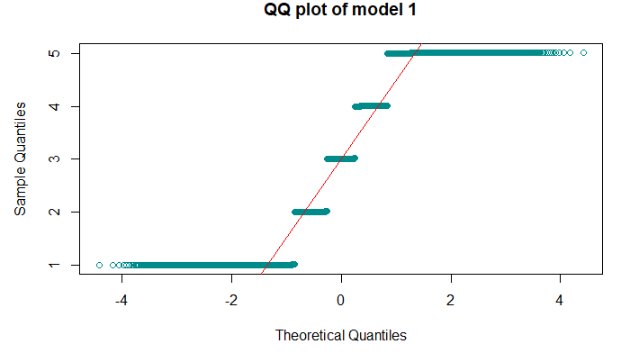


Figure 16: QQ Plot for Figure 1

A graph with a red line

Description automatically generated

Figure 17: QQ Plot for Figure 2

A graph with a red line

Description automatically generated

Figure 18: QQ Plot for Figure 3

A graph of a line

Description automatically generated

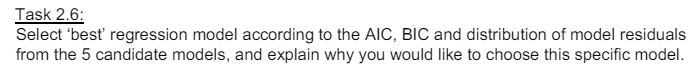
Figure 19: QQ Plot for Figure 4

A graph with a red line

Description automatically generated

Figure 20: QQ Plot for Figure 5

# **Task 2.6 Best Fitted Model Selection**



In order to select the best candidate model for this work, all tasks 2.1 through 2.5 were completed in order to acquire RSS, likelihood, normal distribution, and other relevant information. We would select the candidate based on the normal distribution attained as well as the AIC and BIC values obtained because both technique models contribute to a smaller error size. As is well known, AIC and BIC are frequently used in the model selection process. AIC tends to choose models based on the relative distance between unknown likelihood; therefore, an AIC with a lower value must be chosen for low error and is close to true value, whereas BIC compares a model's exterior posterior probability function in accordance with a specific Bayesian system; therefore, a BIC with a lower value is also near the true value.

In order to select the optimal model for a more accurate value, we would analyze the AIC and BIC values that were produced. Model 2, which is closest to the normal distribution and has the lowest AIC and BIC, is the one we will use.

# **Task 2.7 Test and Train Best Model**

A black text on a white background

Description automatically generated

The training dataset and the testing dataset were the first two sections of the data set that were made available. A priori, each dataset was split into two parts: the training dataset contained 70% of the data, and the testing dataset contained the remaining 10%. The best selected model, in this example, would be Model 2, from which the value of "q" was determined. Model prediction was then performed on testing data following the estimation of the model parameter with the use of RSS. The 95% confidence interval was used once the RSS and model selection were calculated.

A graph with red and green lines

Description automatically generated

Figure 21: Prediction and confidence Intervals

# **Task 3: Approximate Bayesian Computation (ABC)**

A screenshot of a computer program

Description automatically generated

Approximate Bayesian computation is the process of obtaining the posterior distribution of a model without computing the likelihood parameter. Utilizing ABC now to derive posterior distribution and execute ABC rejection from model 2, which is the best fit.

As a result of the assignment that we completed, I believe that model 2 is the best option for the regression module to finish the next task. Two values from model 3 were utilized to estimate the model parameters on task 2.1 in order to find the posterior distribution. You must choose the parameter with the maximum value feasible. Theta bias and Theta 1 were utilized in the least squares analysis of Model 2 to get the posterior value. Two estimated parameters are held constant in their values.

After selecting a new value, we must use the run if function—a feature of R programming—to ascertain the parameter's range. This function yields two new values. By multiplying the two new values by the constant value that is still present in the estimated parameters, we can use these values to find the output signal error and the Y hat. You can evaluate the RSS similarly to task 2 because a mistake has happened. 2. ABC rejection is possible following RSS computation. If the RSS value at discard ABC exceeds the threshold, these values are discarded. In other cases, the data will be accepted and kept for plotting if the RSS is less than the threshold.

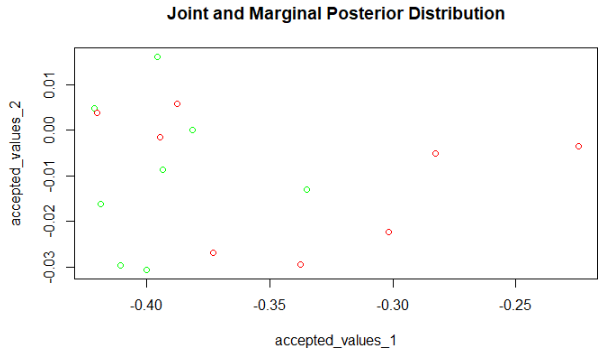
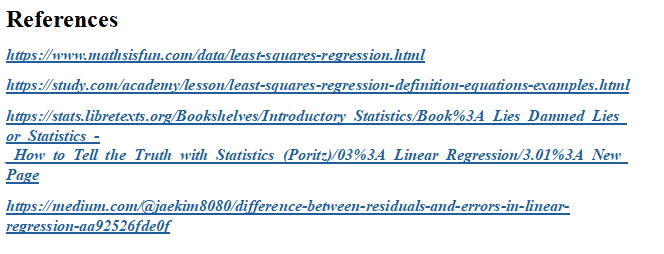


Figure 22:Joint and Marginal Posterior Distribution

For the range computation in this task, a total of two parameters from model 2 with the highest values were chosen.   
following the completion of reject sampling and range determination for the estimated parameters.

# **Conclusion:**

Among the potential models, we were to select the optimal regression model in accordance with the task. Model 2 was the most appropriate for this conduct, and the entire procedure was carried out after every task had been completed. Throughout the entire procedure, we utilized the R programming language to determine the value of the conducted results. The resulting code is shown below.



**Appendix**

