# Sage Plagarism Report

Powered by schoolworkspro.com

Hari Sharan Shrestha
150212

8 Results Found

12.93 % Match Percentage

| | |
|---|---|
| Submitted to: Softwarica | **5.66%** |
| Submitted to: Softwarica | **4.05%** |
| Submitted to: Softwarica | **0.89%** |
| Source: 3.5.3 - Time Series Plot \| STAT 200<br>Link: https://online.stat.psu.edu/stat200/lesson/3/3.5/3.5.3 | **0.85%** |
| Source: Static vs Dynamic<br>Link: https://en.wikipedia.org/wiki/Mathematical_model | **0.67%** |
| Source: An introduction to STRIDE<br>Link: https://en.wikipedia.org/wiki/Bivariate_data | **0.36%** |
| Submitted to: Softwarica | **0.36%** |
| Submitted to: Softwarica | **0.09%** |

REFLECTIVE REPORT
Modelling EEG signal using
polynomial regression
ST7089CEM - Introduction to statistical methods for data
Name: Hari sharan shrestha
College ID: 150212
Git Hub Repository Link:
https://github.com/softwarica-github/ 150212_hari_sharan_st7089cem
Contents

Introduction To detect abnormalities or electrical activities in our brains EEG is popular.
In our brain, the scalp is composed of soft tissue layers that cover cranium and during EEG procedure, electrodes consisting of small metal discs within wires are pasted in the scalp.
While doing any work, our cells activate and exchange electrical impulses with one another and that impulses are tracked as a wavy line during EEG.
In this assignment, the provided EEG signal was analyzed, and the optimal model was chosen from among five regression models.
Task 1 Time series plot By charting the essential data against time, a time series plot enables you to better understand the procedure that resulted in it.
In a time-series plot, the observed data is represented on the Y-axis and time is plotted on the X-axis.
The input and output EEG data were plotted as a time series using R, and the findings are displayed below.
When reading a time series display, it is essential to check for outliers or sharp changes in the plot that are brought on by inaccurate data.
Plots created using the provided data don't show any such sharp spikes, showing that neither external causes nor entry errors may have impacted the data.
In the first diagram, the four input values are plotted against a timestamp, and the output of the input values against the same date is given in the second diagram.
Examining the graphs closely reveals many tendencies in the data.
The time series plot of the input signal indicates significantly reduced noise and spikes.
We can see that the plot is settling down and exhibiting some recurring patterns after 120 milliseconds.
In X4 due to fewer spikes, patterns are clearly recognized.
After the 120 ms limit, the spikes are fewer in number.
Up until the 60 ms threshold, the output signal starts to emerge gradually.
We can see that the spikes are leveling off and adopting a consistent pattern after the 120ms barrier.
A noticeable rise can be seen at 60 ms. the spikes are now more obvious because of the noise.
Distribution plot:
The connection between observed values and sample data arranged from lowest to highest is known as distribution.
The graphic below is used to compare the input signal delivered to the brain.
When examining the distribution of the input signal from the histogram and density, the curve has a bell-shaped form and the mean of the input signal looks to be close to "0.".

In the diagram X signal’s peak is closer to 0.
It shows that minimum of -10 and maximum of between 8 and 10,
the brain can receive input signals.
The majority of the input increases as the form of the distribution begins to shrink from 5 and expand from -4.This distribution looks to be symmetric, making it simple to find the distribution's center because the lengths of both tails seem to be about identical.
Let's examine each signal that the brain receives individually from the four input signals one at a time.
According to the graph above, all input signals from X1 through X4 seem to follow a similar pattern.
The output signal, also known as the Y signal, is sometimes referred to as a left-skewed distribution because, on examining the figure, it seems that the left side's tail is significantly longer than the right..
The majority of output looks to be between -2.5 to 2.5 with maximum output of 4.5 and minimum of 7.
So Y signal’s mean can be lower than 0 or 1.
Correlation and scatter plots
A scatterplots diagram is used in our situation to highlight the link between the input and output datasets.
A single dot in a scatter plot diagram represents a single piece of data, and various dots throughout the diagram can each indicate a distinct pattern, depending on how closely the data is clustered.
When a data set tends to form a straight line beginning at the origin of the y axis, the relationship is called perfect positive correlation; when a straight line begins at the origin of the x axis, the relationship is called perfect negative correlation; otherwise, the relationship is called no correlation.
The connection is low positive or low negative correlation When there may be no perfect positive or negative correlation.
Examine the data's output signals and input signals right now..
The image below depicts the relationship between the X1, X2,X3 and X4 signals and the Y signal and exhibits a pattern in which the data appears to follow a low positive correlation since the dot plot has a higher Y- value on the X-axis and the data are not distributed.
As for the X2 and Y signals, there appears to be no association because the data does not follow any pattern and is spread unevenly.
Task 2 Task 2.1 Since the true value of a distribution is unknown, a random variable, also known as an estimator, is typically helpful for observing the various characteristics of a distribution.
An estimator is denoted as AB”.. Where AB = {B, B1, B2, …, Bbisa} T.".
From each individual candidate module of EEG data use least squares to calculate the estimator model parameters.
The regression equation's best fit line or the data point that illustrates the relationship between an unknown dependent value and the independent value are typically found using least
least square is indicated as where x and y are input and output data of EEG and
for calculating in R programming language here is the formula
First it needs to bind the value/column of input data
(x) from the provided data set to calculate least squares.
After performing the least square from the formula mentioned in the assignment.
the result are:

Task 2.2 To identify the error of the squared average estimate value, i.e.
variations between the average square actual and estimated value, utilize the model residual error (RSS), also known as the sum squared error of an estimator.
An estimator's quality is typically assessed using RSS, and a value of RSS that is nearer to zero is preferable because RSS can never be negative.
We use from task 2.1 to determine each candidate model's error in order to calculate the RSS first.
[ CITATION ada22 \l 1033 ]
R programming allows us to implement the aforementioned formula and calculate RSS as (RSS=sum((Y-Yhat) 2 where Yhat is a product of , n is the total length of Y, and ∑ is the sum of all the output Y.
All of the RSS values for the candidate model are
represented in the table below.
All Xi values are conceivable of X.
we determined from task 2.1, where each model's value varies.
The final RSS calculation
Where sum and Y are the total of all possible values and the
output signal respectively.
Task 2.3 To know how well the measured value fits the sample data of the provided model especially when parameters are unknown, then likelihood function is used.the main aim of likelihood is to find the best favorable way to fit a distribution to the data with a fixed observed value.
In order to operate conveniently for practical reasons by maximizing the estimation, log likelihood is a logarithmic transformation of the likelihood function that maximizes the likelihood, which is identical to maximum log-likelihood.
Result of likelihood are:
Task 2.4
The Akaike information criterion (ACI), which estimates prediction error, enables us to assess how well the model fits the given dataset without over fitting it.
AIC aids in evaluating the model's quality by contrasting various candidate models.
[ CITATION Aka22 \l 1033 ]
AIC calculations are given below:
One of the criteria for deciding which modules from a variety of candidate models are the best and ranking them is the Bayesian information criterion (BIC).
The candidate model with the lowest score value is chosen for choosing the BIC.
AIC and BIC are relatable and also based on likelihood function.
BIC calculations are given below:
Task 2.5 In task 2.2 RSS i.e Model Residual Error of each model has been calculated to depict the normal/Gaussian distribution with additive Gaussian noise of the output EEG signal.
Expected Y(Output), or Yhat value, was subtracted from output value(Y), which will be used in qqnorm and qqline, plotting functions of ggplot2 that are pre-defined in R programming, to calculate RSS.
A normal/gaussian distribution was plotted for each model in order to determine the most effective regression model and data trend.
[ CITATION res23 \l 1033 ]
Task 2.6

All necessary activities, from 2.1 to 2.5, were finished in order to choose the best candidate model.
For instance, computing RSS, determining likelihood, or drawing graphs of normal distribution.
Each step was meticulously carried out.
I will choose the most suitable candidate model for this regression based on the plot of 2.5 and the AIC and BIC calculations.
AIC and BIC are well known for being used frequently in the model selection process.
Since both models reduce the amount of mistakes when choosing a model.
The relative distance between the unknown likelihood of the data and the fit of the model is often used by AIC to choose a model.
Therefore, if the AIC is smaller and close to the truth value, a model may be chosen.
So, if the BIC is lower, a model can be chosen since it is most likely to be close to reality.
So, in order to choose the best model for this assignment, I'll use both AIC and BIC.
I determine that candidate model 2 is the best model among the other mentioned models by examining the produced output of each AIC and BIC from task 2.4 and the normal distribution plot from task 2.5.
Also, the produced value of model 2 seems to be lower for both AIC and BIC and the plot of model 2 appears to be similar to the normal distribution.
Task 2.7
The data that was provided as input(X) and output(Y) has been split into two parts at the beginning.
70% of each dataset was used for training and the remaining 30% for testing, as indicated by the information provided.
I have chosen model 2 for estimation of model parameters carried out on training data.
Following the estimation of model parameters by
computing the RSS, model output/prediction was performed on testing data.
After the RSS calculation and model prediction, 95% confidence interval was performed.
for calculating error bar od model 2, first calculated standard deviation by using formula
StandardDeviation = sqrt(Variance_model2)
then here are the screenshots of error bar
Task 3
Without determining the likelihood parameter, Approximation Bayesian computation is a sort of procedure that can assist in calculating and evaluating the posterior distribution of a model parameter.
With the aid of the top model chosen from task 2, computed the posterior distribution using the ABC technique.
Model 2 is fit as a regression module for task3 and will use it to finish this assignment.Two values were chosen from model 2 for the estimation of the model parameters, which was done on task 2.1 for determining the posterior distribution.
Two estimated parameter values from model 2 remain constant Theta bias and theta one chosen for the calculation of posterior value using model 2's least squares.
The runif function, which is built into R programming and provides two new values, should be used to calculate the range of those values, or the chosen parameter, after different values have been chosen.
By combining the two new values with the constant value left in the estimated parameter, the Y hat and error of the output signal can be calculated.I.
Since we produced the Y hat and error, RSS can be calculated in the same way as task 2.2.

ABC can be used for rejection after RSS calculation.
These values are rejected if the RSS value during rejection ABC exceeds the threshold value.
In any other case, if the RSS is below the threshold, the values are accepted and saved for plotting.
Plots of both parameters are displayed below.
In order to do rejection sampling and determine the range of estimated parameters for this work, the two parameters from model 2 with the highest values were chosen.
As in task 2.2, the range was calculated, and to find RSS Y hat and error were created.
Rejection ABC was used to
accept and plot RSS data that fell below the projected threshold and reject those that fell above the projected threshold.
The graphs for them are shown above.
Conclusion In this assignment I have to select the best regression model from given models.
After completing task 2 with findings of RSS, Variance, likelihood, AIC, BIC, I have chosen model 2 due to low AIC and BIC value as compared to others models.
R programming was used to find out results and that source codes are provided in the appendix section.
Anon., 2022.
Akaike Information Criterion.
[Online] Available at: https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/akaike-information-criterion [Accessed 2 feb 2023].
Anon., 2023. research data services.
[Online] Available at: https://data.library.virginia.edu/understanding-q-q-plots/ [Accessed 3 feb 2023].
borne, a., 2022.
Residual Sum of Squares (RSS).
[Online] Available at: https://www.investopedia.com/terms/r/residual-sum-of-squares.asp [Accessed 1 feb 2023].