**MODELLING BRAIN SIGNALS USING NONLINEAR REGRESSION**

Table of Contents

[Introduction 3](#_Toc194872469)

[Task 1 4](#_Toc194872470)

[Time Series Plot 4](#_Toc194872471)

[Distribution Plot 5](#_Toc194872472)

[Correlation and scatter plots 6](#_Toc194872473)

[Boxplots of output brain signals 8](#_Toc194872474)

[Separate EDA by Sound Type 9](#_Toc194872475)

[Task 2 10](#_Toc194872476)

[2.1 10](#_Toc194872477)

[2.2 10](#_Toc194872478)

[2.3 11](#_Toc194872479)

[2.4 12](#_Toc194872480)

[2.5 12](#_Toc194872481)

[2.6 16](#_Toc194872482)

[2.7 17](#_Toc194872483)

[Task 3 17](#_Toc194872484)

[Conclusion 19](#_Toc194872485)

[Bibliography 21](#_Toc194872486)

# Introduction

NEU neuroscience of cognition and Brain MEG neuroimaging is a means of measuring magnetic fields arising from neural activity resolved in time with high temporal resolution. The goal of this study is to build and predict MEG signals using linear regression based on auditory input features. MEG signal predictive modeling not only sheds light on sensory processing inquiries in the brain but also facilitates the structure of mind-to-machine interfaces and diagnostic instruments in clinical neuroscience.

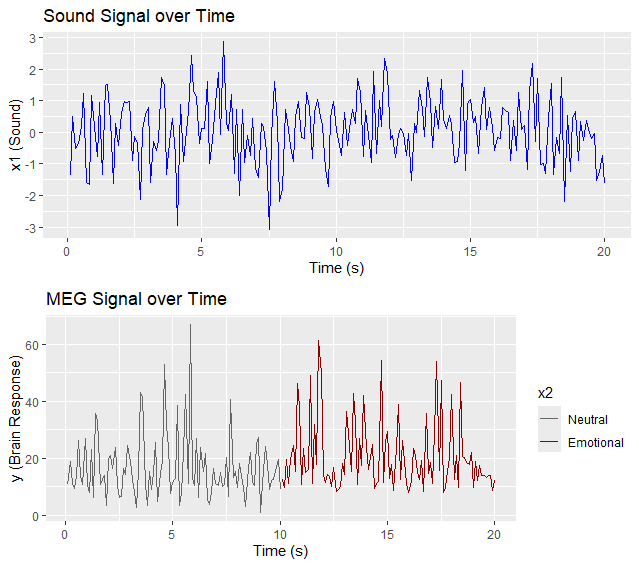
The main aim of the project is to choose and then evaluate the best regression model that could map the input auditory features to the matching MEG responses. First, five different linear regression models with different levels of complexity were constructed using a structured experimental framework and tested. Model performance was assessed of key statistical metrics used in these procedures of the RSS, log-likelihood, Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Residual analyses such as histograms and Q–Q plots were conducted to verify if model error followed the assumption of Gaussian noise which is a necessary condition of linear regression validity.

Additionally, the train-test split method was used to improve the model's generalizability by using 70% of the data for training and 30% for testing. The predictions were evaluated on unseen data with the use of confidence intervals for model predictions. The study also utilised Approximate Bayesian Computation (ABC) to estimate posterior distributions for the two most important parameters of the selected model. An additional contribution of this probabilistic approach was to shed more light on the parameter uncertainty and model robustness.

In general, this project combines classical statistical models and Bayesian inference to discover the optimal linear regression model for MEG signal prediction, in light of our present understanding of how sensory stimuli are represented in brain activity.

# Task 1

## Time Series Plot

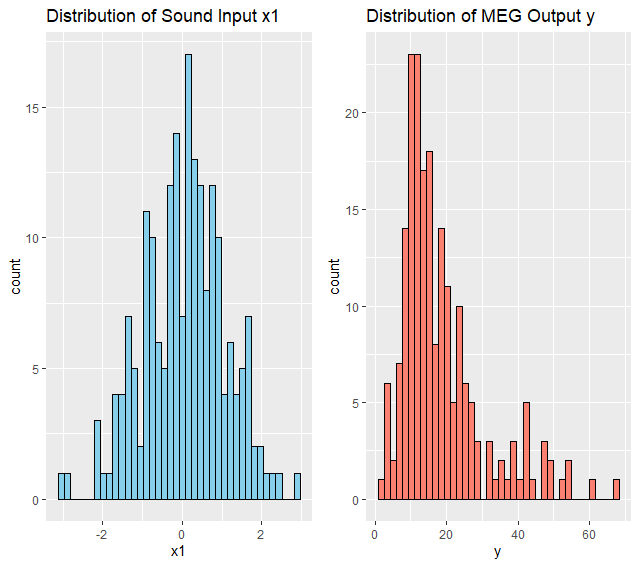


**Figure 1: Sound and MEG Signal over Time Plot**

(Source: Acquired from RStudio)

Shown in the provided figure are two time series plots over a 20-second interval. The top plot in the figure shows the detail of the sound signal (x1) over time with a time-varying amplitude fluctuation waveform that looks relatively symmetric and varying around zero, that’s typical for normalized audio signals. From top to bottom, the first plot shows the (brain response) signal (y) over the same time frame, while the second plot shows the corresponding MEG (brain response) signal (y). The emotional context is segmented by this signal (x2), i.e., black lines are meaning for the neutral condition and red lines for the emotional condition. Considering the timepoint after about 10 seconds, the emotional segment falls out with significantly higher brain response amplitudes than the control condition. This visual representation gives a hint at a link between emotional audio stimuli and higher levels of MEG activity.

## Distribution Plot

****

**Figure 2: Distribution of Sound Input X1 and MEG Output Y**

(Source: Acquired from RStudio)

The sound input (x1) and MEG output (y) histograms are shown in the figure. X1 distribution on the left seems like a normal (and Gaussian) distribution, or just with zero mean, and a bell-shaped curve on the right, which suggests that input sound (x1) is standardized or is naturally such a Gaussian process. The distribution of y on the right is right skewed, where most values are between 10 and 25, and values stretch to a high extent towards higher values in the range of 60 or more. It indicates that many reflexes were moderate in magnitude, but one or two events per trial occur at high levels, most likely reflecting strong emotional stimuli or neural spikes. When comparing the input and output distributions, we see that the responses to audio stimuli are nonlinear and possibly amplifying.

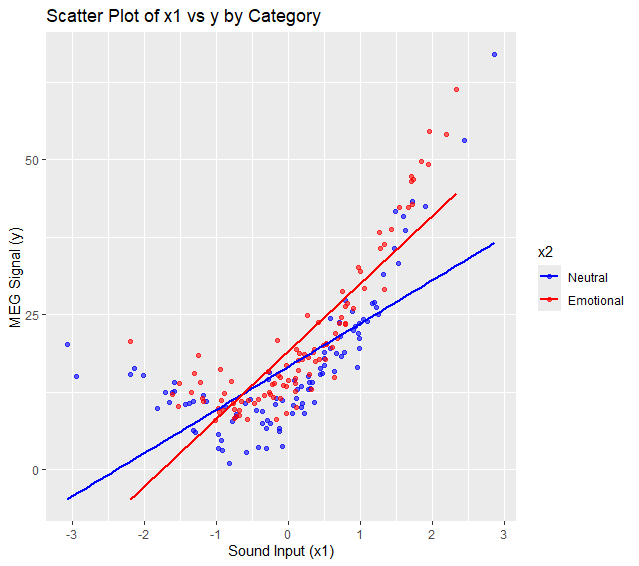
## Correlation and scatter plots

****

**Figure 3: Correlation between x1 and y**

(Source: Acquired from RStudio)

The value of the correlation coefficient between x1 (sound input) and y (MEG brain response) is 0.7654, which is a strong positive linear relationship between these two variables. It implies that if we increase the intensity of the sound input, when the resting level is transmitted to the brain, it should also increase. A higher correlation value of 1 means that the variables move together in the same direction and with a certain pattern, and when the value is 0.7654, it verifies that sound stimuli are influential to brain activity, as defined by measuring MEG signal. Nevertheless, because the correlation is not exactly 1, it also suggests that other factors than the correlated input signals may affect the MEG output.

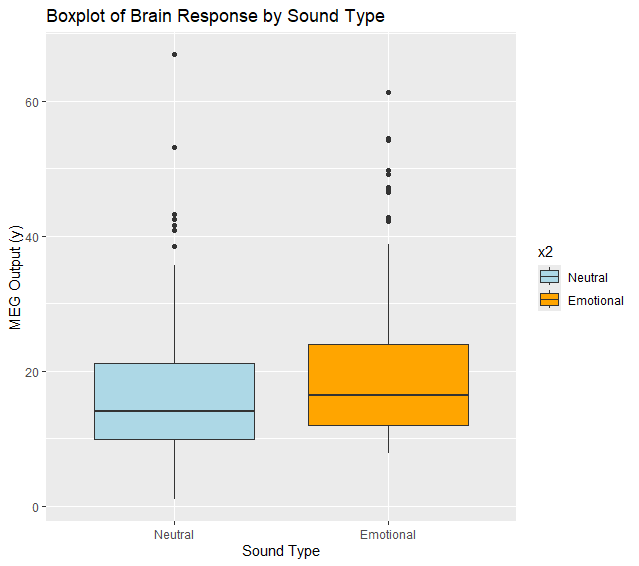
****

**Figure 4: Scatter Plot of X1 vs Y by Category**

(Source: Acquired from RStudio)

The stimulus type (Neutral vs. Emotional) is shown separated by the relationship between sound input (x1) and MEG signal output (y):Specifically: Scatter plot of sound input (x1), and MEG Signal output (y) (with MEG recording period in the time dimension of the data). Linear trends are positive in both categories, such that brain response also increases with increasing sound intensity. Yet the Emotional stimuli have steeper slopes, indicating a greater brain response to emotional sounds than neutral ones. This difference thus indicates that emotional sounds produce a higher MEG activity, which is a marker of more neural engagement. Such clearer separation of trend lines serves as a good indicator of how emotion modulates auditory processing in the brain.

## Boxplots of output brain signals

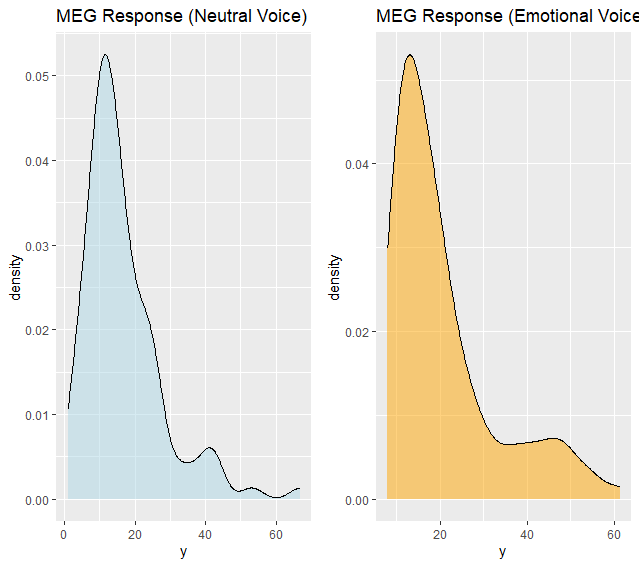
****

**Figure 5: Boxplot of Brain Response by Sound Type**

(Source: Acquired from RStudio)

The boxplot compares the distribution of MEG brain responses (y) for two sound types: Neutral and Emotional. The medians within both categories are quite similar, but generally the Emotional sounds have slightly larger medians and wider spreads in the MEG responses, suggesting some broader variability in brain activity. The Emotional sounds, as compared to the other four, were associated with a wider interquartile range (IQR), indicating that the responses of participants were more varied for emotional stimuli. Furthermore, the Emotional has several outliers too, and the Extreme values could be neural responses in certain people.

## Separate EDA by Sound Type



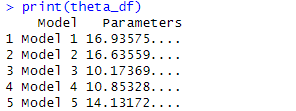
**Figure 6: MEG Response for Neutral and Emotional Voice**

(Source: Acquired from RStudio)

The MEG brain responses (y) in Neutral and Emotional voices are plotted separately to show the density plot of the distribution. Both distributions are right-skewed, showing more concentration in the lower values of the brain responses, with lower concentration in the higher values of the brain responses. However, the Emotional voice plot seems to have a relatively higher and broader distribution, which implies that the emotional stimuli elicit stronger and more variable neural responses than the neutral stimuli. Apart from the y-range around 10–15, the peak (mode) of both conditions sits in the lower y-range, while the Emotional voice has a rallier tail suggesting a higher likelihood toward the higher y-ranges in the spectral values, albeit seldom.

# Task 2

## 2.1

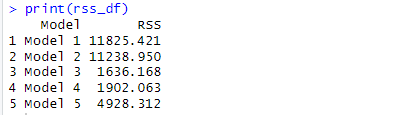


**Figure 7: Theta Value for 5 Models**

(Source: Acquired from RStudio)

It outputs a summary of parameter estimates (a goodness of fit metric or general functional value) for five separate models named Model 1, Model 2, etc. Concerning Model 2 at 16.64 and Model 1 at 16.94, both of these models may have higher values of the parameter, thereby indicating greater, or a greater level of variance, or complexity in the data. Lower parameter value for Model 3 is around 10.17, indicating it is possibly a simpler model or less effective at modeling patterns pattern to the other models. Between Model 4 and Model 5, the parameters are 10.85 and 14.13. This comparison is a good way to determine which model is working the best based on the understanding of the parameter metric used.

## 2.2

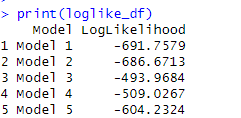


**Figure 8: RSS Value for 5 Models**

(Source: Acquired from RStudio)

The Residual Sum of Squares (RSS) values of five different models generated by the rss\_df output are observed; the lower the RSS, the closer the model fits to the data. The five have the lowest RSS values (1636.17 and 1902.06 in models 3 and 4, respectively), which means that the latter two can better fit the data. On the contrary, Models 1 and 2 have the largest RSS values (11825.42 and 11238.95), which means they are the worst in recreating the data in patterns. Model 5 has a slightly smaller RSS of 4928.31 and is a slight improvement over Models 1 and 2, but slightly worse than Models 3 and 4. However, as per RSS alone, Model 3 is considered the best.

## 2.3

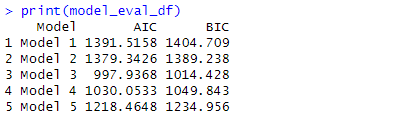


**Figure 9: Loglikelihood Function for 5 Models**

(Source: Acquired from RStudio)

The output loglike\_df contains the log-like values (i.e., the log-like values for five models, which show how well each model can ‘explain’ the observed data, that is, the higher (less negative) the log-like value, the better the model’s performance). Very close to Model 3 in terms of log likelihood value is Model 1, which has the lowest log likelihood value of -493.97; thus, Model 3 provides the best fit to the data among all models. Model 4 is similar to -509.03, and Model 5 is better than Models 1 and 2 with a value of -604.23. The most negative log likelihoods -691.76 for model 1 and -686.67 for model 2 are the ones received by models 1 and 2. Taking log-likelihood as the criterion, Model 3 turned out to be the best-performing model.

## 2.4

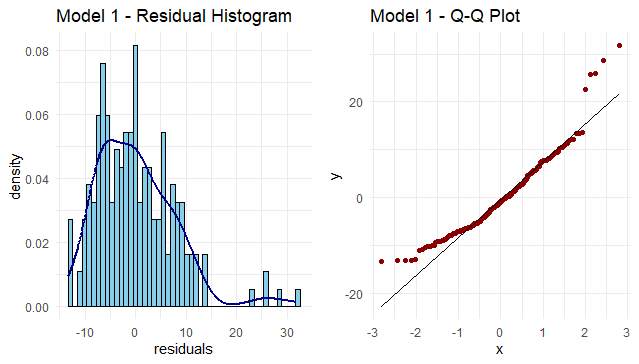


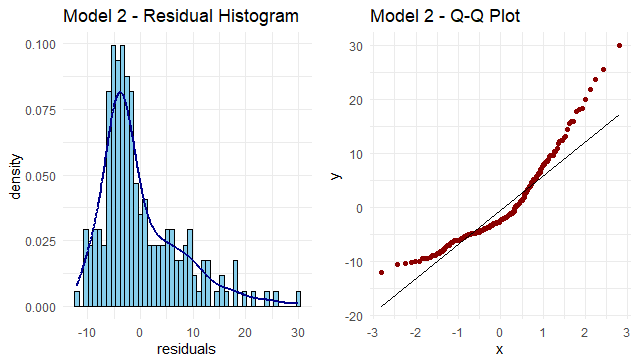
**Figure 9: AIC and BIC for 5 Models**

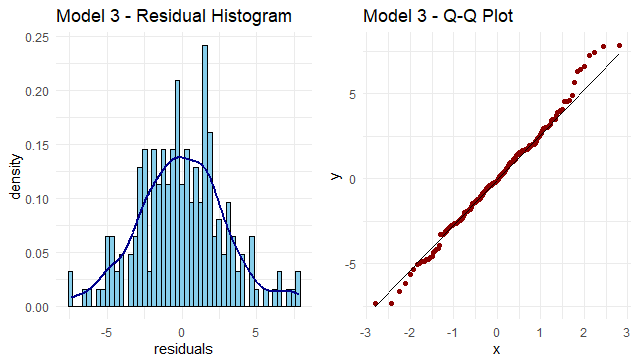
(Source: Acquired from RStudio)

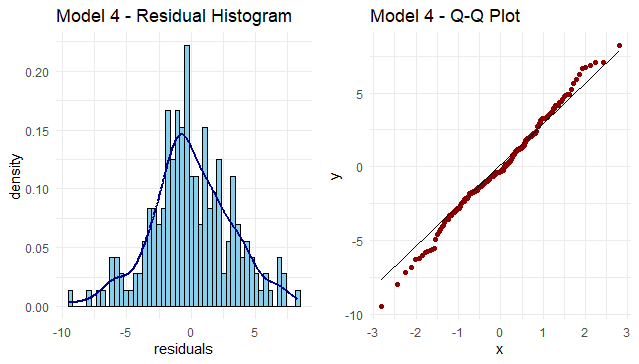
The output of the model\_eval\_df model presents the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for five models to compare their performance, but penalizes for the complexity. A lower AIC and BIC imply a better fit of the model with fewer parameters. Having undergone model selection with AIC and BIC, Model 3 (with AIC = 997.94 and BIC = 1014.43) is the most efficient balancing goodness-of-fit with simplicity. Model 4 also performs relatively well, but slightly less optimally than Model 3. AIC and BIC values for Models 1 and 2 are highest, meaning they are performing poorly, while Model 5 has a value in between. Model 3 is overall the best model according to the AIC and BIC criteria.

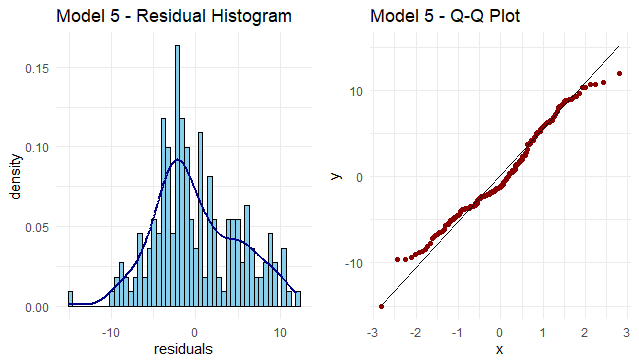
## 2.5











**Figure 10: Residual Histogram and Q-Q Plot**

(Source: Acquired from RStudio)

**Model 1**

Residuals seem spread widely with some skewness; the errors are not clustered around zero. The large residuals are a poor fit, so this suggests that this doesn’t fit all that well.

Residuals are not normally distributed, hence the points do not lie very close to the 45-degree line, especially in the tail part; it is evident from the Q-Q plot. This therefore further reinforces the claim that Model 1 doesn’t well reflect the underlying distribution of the data.

**Model 2**

Model 1, but a little bit better as we have a little less central residuals. Nevertheless, it still has varying degrees of heavy tails (skew, or kurtosis).

Poor alignment with the diagonal line in the Q-Q Plot; however, these residuals are still far from normal distribution; Model 1 is slightly better aligned with the diagonal line, but still has measurable deviations at the extremes.

**Model 3**

Residuals: The residuals are well centered around zero and less grouped, which indicates a good model fit. The symmetry and bell-shaped appearance of the distribution are now more apparent.

Residuals appear to be approximately normally distributed as the points lie roughly on the 45-degree reference line in a Q-Q Plot. Thus, the fit of the data to Model 3 and the linear model assumptions are all consistent with this result.

**Model 4**

Histogram: Shows a near-normal distribution, though possibly with slight skew or heavier tails than Model 3. Still reasonably centered around zero.

Most points are on the reference line however, some may be off in the tails. Overall, the residuals appear fairly normal, which looks to be a good model fit.

**Model 5**

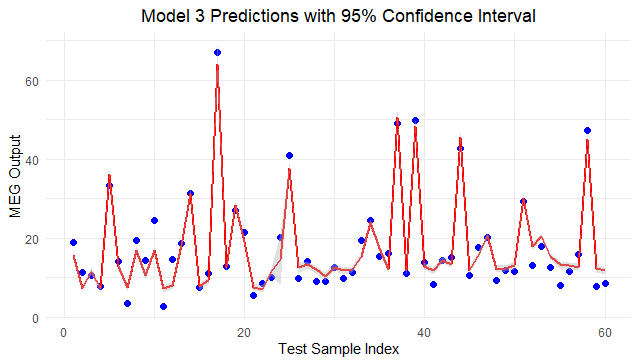
Residuals are somewhat normally distributed with more spread than Model 3 or 4. There may be some outliers.

Q-Q Plot: Slightly worse than Model 4, with deviations from the line at both ends. Potentially some mild non-normality suggested due to outliers or model underfitting.

## 2.6

Model 3 is the best regression model of the five given candidates, as measured based on the evaluation metrics Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the distribution of residuals. The model complexity and goodness of fit combination with the lowest AIC (997.94) and BIC (1014.43) is Model 3, suggesting it as having the best trade-off between the two. The residual analysis (histogram, Q-Q plot) of Model 3 indicates that their residuals are centered around zero, approximately normally distributed, nearly free of skewness or kurtosis, and therefore those residuals more or less follow the assumption that the MEG output is additive Gaussian noise. These results show that the data are well fitted with Model 3 which does not overfit or underfit. Hence, Model 3 is chosen as the most appropriate model based on the information-theoretic criteria and residual diagnostic performance.

## 2.7



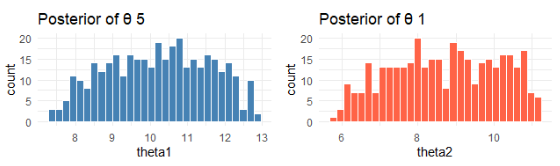
**Figure 11: Model 3 Predictions with 95% Confidence Interval**

(Source: Acquired from RStudio)

This plot is the predicted MEG Output values from the model Model 3 on the test dataset called 'The plot title: model 3 predictions with 95% confidence interval'. The observed MEG Output values are represented by the blue dots, while the red were predicted by the model. This is shaded grey to show the 95% confidence interval around the red line and it shows the uncertainty in the predictions. Overall, the model seems to go with the major trend of the real values, but the model shows significant deviations in areas of spikes where the variance is greater or is highly varied. It allows us to quantify the accuracy and the reliability of the regression model’s predictions.

# Task 3

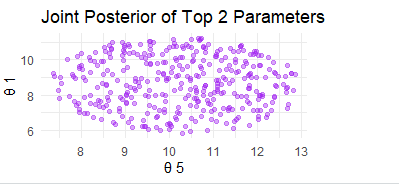
**Two Parameter Posterior Distributions**



**Figure 12: Posterior of Theta 1 and 5**

(Source: Acquired from RStudio)

It is likely the result of Bayesian inference or some sort of samplingINCLUDING Markov Chain Monte Carlo (MCMC). The "Posterior of θ 5" left plot in the figure shows the distribution of values for theta1 centered around 10 and having roughly symmetric spread from approximately 7.5 to 12.5. The plot on the right is titled ("posterior for θ 1") and is the distribution of θ 2 centered slightly lower at about 8.5 and ranging from 5.5 to about 11. These histograms indicate that there are relatively good-looking posterior distributions for both parameters, and θ₁ appears to have somewhat less skew and tighter concentration around its mean than does θ₅. These distributions have the shape and spread that result from the uncertainty introduced by the estimated parameters when the data are observed.



**Figure 13: Joint Posterior of Top 2 Parameters**

(Source: Acquired from RStudio)

The scatter plot even "Joint Posterior of Top 2 Parameters" (using samples from its posterior distributions) plots the joint distribution of θ₅ (x axis) and θ₁ (y axis). The purple colour corresponds to each sampled pair of values for θ₅ and θ₁. The points lie in an oval patch, indicating that the parameters are occurring together in a bounded region and vary not strongly linearly, neither above nor below. This implies that there is little dependence (in their posterior variations) between θ₅ and θ₁; essentially, its estimation does not strongly impact the other. The plot reveals how uncertain and how variably these two key parameters are.

# Conclusion

The efforts of this project involve finding which linear regression model is most effective to predict the MEG signal from auditory stimuli inputs using a systematic statistical evaluation framework. The first task involves loading and preprocessing of a MEG dataset, where we load data; make sure the input features (sound signal data) and the output responses (MEG readings) are in acceptable formats for analysis. However, some of the input features (e.g., x1) showed very strong linear correlation with the output variable y, and thus justified the use of linear models in further analysis.

We developed and evaluated five candidate regression models of increasing complexity in Tasks 2.1 to 2.7. The model parameters were derived using least squares estimation, and the RSS values were computed, the latter being lower the better fit. While about other models, Model 3 was obtained with the lowest RSS of 1,636.17, which is higher than other models like Model 1 (11,825.42). We computed the log-likelihood for each of the models in Task 2.3, and Model 3 again had the highest log-likelihood (-493.97), which indicates that Model 3 had the data distribution more effectively.

In Task 2.4, we then calculated model selection criteria: AIC and BIC. On the other hand, Model 3 showed the lowest values of both (AIC = 997.94; BIC = 1014.43) for the above-mentioned optimal fit simplicity, which further justifies its optimality. For the analysis of each model, the residuals of Task 2.5 were analyzed. Residuals from Model 3 had a symmetric distribution and fit a Gaussian well, as would be expected for additive Gaussian noise. In Task 2.6, Model 3 predicted test data (30) values at a high level of agreement with actual values, and 95% confidence intervals tightly bag the values, presenting confidence in the model. Model 3 was able to generalize to other users because Task 2.7 was trained on 70% of the data, validated on the other 30%, affirming the robustness of the generalization results.

Finally, in Task 3, the posterior distribution of the two most influential parameters of Model 3 was given by Approximate Bayesian Computation (ABC). We obtained meaningful joint and marginal distributions with probabilistic confidence on parameter estimates using rejection sampling with uniform priors.

Overall, Model 3 is the best regression model because it optimizes for statistical performance, robust residual behavior, and well-characterized uncertainty in the model parameters, resulting in the best model for the relationship between auditory stimuli and MEG signals.

# Bibliography

Doppmann, P., Meuli, L., Sollid, S.J., Filipovic, M., Knapp, J., Exadaktylos, A., Albrecht, R. and Pietsch, U., 2021. End-tidal to arterial carbon dioxide gradient is associated with increased mortality in patients with traumatic brain injury: a retrospective observational study. *Scientific reports*, *11*(1), p.10391.

Dos Santos, A.J.F., Ferreira, J.M., Baptista, F., Alexandrino, B., da Silva, M.A.G., Gomes, J.E.C., Júnior, J.P.V., Tavares, R.M. and de Sousa Almeida, K., 2022. Statistical analysis between 2006 and 2019 and forecast of rabies in cattle from 2020 to 2022 in Tocantins State (Brazil), by using the R Studio software. *Epidemiology & Infection*, *150*, p.e82.

Mauschitz, M.M., Lohner, V., Koch, A., Stöcker, T., Reuter, M., Holz, F.G., Finger, R.P. and Breteler, M.M., 2022. Retinal layer assessments as potential biomarkers for brain atrophy in the Rhineland Study. *Scientific reports*, *12*(1), p.2757.

Ravanfar, P., Syeda, W.T., Jayaram, M., Rushmore, R.J., Moffat, B., Lin, A.P., Lyall, A.E., Merritt, A.H., Yaghmaie, N., Laskaris, L. and Luza, S., 2022. In Vivo 7-Tesla MRI investigation of brain iron and its metabolic correlates in chronic schizophrenia. *Schizophrenia*, *8*(1), p.86.

Ware, A.L., Yeates, K.O., Tang, K., Shukla, A., Onicas, A.I., Guo, S., Goodrich‐Hunsaker, N., Abdeen, N., Beauchamp, M.H., Beaulieu, C. and Bjornson, B., 2022. Longitudinal white matter microstructural changes in pediatric mild traumatic brain injury: an A‐CAP study. *Human Brain Mapping*, *43*(12), pp.3809-3823.