Report

2022-11-30

Contents

1 Introduction **2**

2 Data **2**
2.1 Dataset and Data Processing 2
2.2 Data analysis 3

3 Models **4**
3.1 Pooled Logistic Regression Model 4
3.2 Hierarchical Logistic Regression Model 6
3.3 Fitting data to the models 7

4 Convergence Diagnostics **8**
4.1 Rhat convergence diagnostics 9
4.2 HMC Convergence Diagnostics 9
4.3 Effective sample size 10

5 Posterior predictive checks **11**

6 Predictive performance assessment **11**

7 Model comparison **12**

8 Prior Sensitivity Analysis **13**
8.1 Pooled model 13
8.2 Hierarchical model 14

9 Discussion and Conclusion **16**

10 Self-reflection **17**

1 Introduction

This is part of Aalto University Bayesian Data Analysis 2022 course.

Recently, we have seen the number of patients diagnosed with depression increase substantially, especially during the COVID-19 Pandemic, causing a 25% increase in prevalence of anxiety and depression worldwide. Furthermore, depression is so common nowadays that around 3.8% of the worldwide population, around 298 millions people, is affected by this condition. This is why we chose depression as our topic as it is a very important and urgent matter that needs to be researched.

Thus, the topic of the analysis is "Predicting depression based on brain activity using Bayesian Data Analysis". The data was collected from a study on the prevalence of depression and its associated factors. The details of the dataset can be found here: https://www.kaggle.com/datasets/arashnic/the-depression-dataset. The dataset includes several variables such as brain activities, age, gender, education level, and family status that could be potentially associated with depression.

Our goal is to use Bayesian data analysis methods to predict if a person has depression or not as accurately as possible from data collected on patients with depression and a control group using R and RStan. We will try to apply the data to a non-hierarchical logistic regression model and a hierarchical logistic regression model. From these, we will then determine the best methods and models by using convergence diagnostics, comparisons, predictive performance as well as sensitivity analysis. We hope that our findings can contribute to the further research on depression and can even help with future diagnosis for those suffering from this condition. With that, we have great motivation to continue this project and make the best results possible.

This report contains the following sections: introduction, data analysis, model, implementations, convergence diagnostics, posterior predictive checks, sensitivity analysis, predictive performance assessment, prior sensitivity analysis, discussion and conclusion, and self-reflection.

2 Data

2.1 Dataset and Data Processing

The original dataset is "The Depression dataset" (2018) available on Kaggle: https://www.kaggle.com/datasets/arashnic/the-depression-dataset. This dataset is obtained from monitoring motor activity per minute in 55 participants, with 23 participants in the condition group (having depression), and 32 participants in the control group (no depression). The results are put into 2 folders, one for control group and one for condition group. In the folders, each participant has a separate file containing information about the measurements: timestamp, date, and activity (activity per minute). Finally, the dataset also contains a file scores storing an overview of the participants, including information such as number (indicator for the participants), gender, or age group.

"The Depression dataset" is used for various analysis and machine learning models, ranging from predicting depression to classifying depression types. In particular, an analysis by Samuel Oseh (available on Kaggle: https://www.kaggle.com/code/samuelkali/depression-dataset-analysis-and-machine-learning) inspired the choice of variable in the models by showing the zero proportion (proportion of zero motor activity measurements) of one participant from the control group (non-depressed) and one participant from the condition group (depressed). In this model, however, data will be further processed and summarized to investigate the potential trend in this property and depression to construct a complete model with Bayesian analysis.

This large dataset is processed and summarized to arrive at a final dataset containing the variables of interest. For each participant (file) in control and condition groups, the average motor activity per minute is calculated for each hour, as well as the zero-proportion. Then, data over different dates will be averaged to arrive at data of the participants for each hour throughout the day and combined with demographics information (gender) to create the final dataset.

The final dataset consists of 1320 rows with 7 columns for the variables of interest (futher elaborated on in the data analysis). The variables are as following:

Variable	Type	Description
entry	Textual	Indicator of the participants, e.g. control_1, condition_2
gender	Categorical	Gender of participants (1 = female, 2 = male)
depressed	Categorical	State of depression (0 = non-depressed, 1 = depressed)
hour	Numerical (Integer)	Time (in hours)
nof_measurements	Numerical (Integer)	Number of measurements taken
nof_zero	Numerical (Integer)	Number of measurements where motor activity per minute = 0
zero_proportion	Numerical (Real)	Percentage of zero-measurements
hourly_mean	Numerical (Real)	Hourly mean activity per minute

2.2 Data analysis

In the analysis by Samuel Oseh, the zero-proportions of one non-depressed and one depressed participants are compared and show significant difference. We aim to test if there are indeed such difference across non-depressed and depressed participants. Additionally, the hourly mean across non-depressed and depressed participants are also analysed to check for potential correlation. Such combined trend is calculated using the mean values of each group:

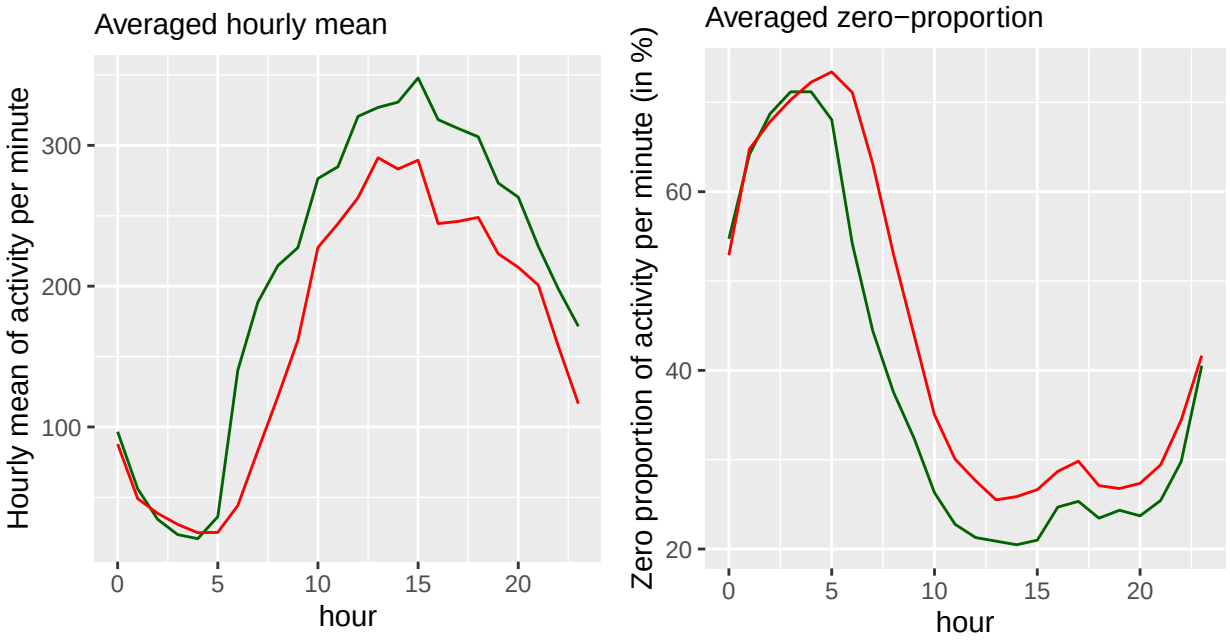
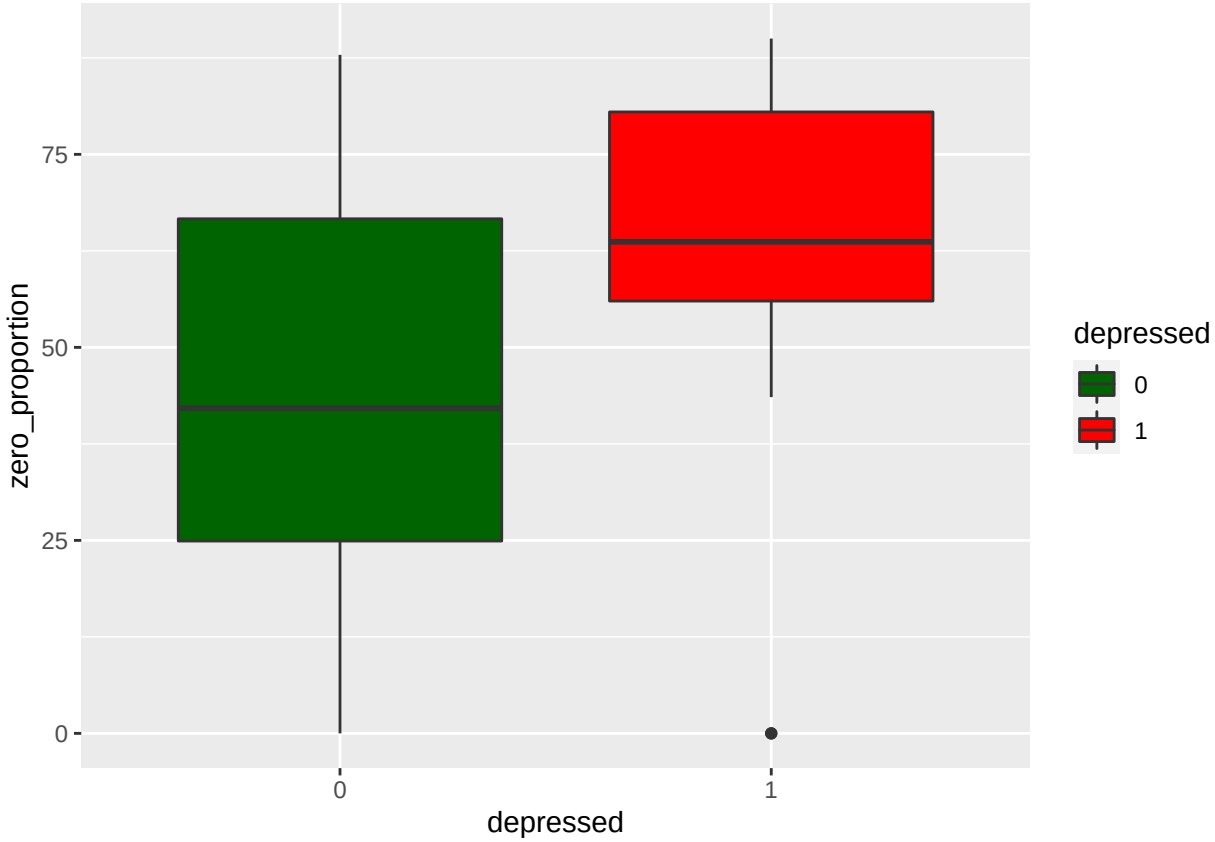


Figure 1: Trends analysis. Green = Non-depressed, Red = Depressed

There are no significant differences in hourly mean activity per minute between the control and condition groups throughout the day. However, there is a significant gap between the two groups with regards to the zero-proportion at the time 7:00 (up to 20% difference between approximately 40% and 60%). Hence, it is highly possible that the zero-proportion at the time 7:00 is correlated to depression state.

Further analysis of the zero-proportion at the time 7:00 shows that the majority of non-depressed people has lower zero-proportion than depressed people:



Therefore, the zero-proportion at the time 7:00 is chosen as the predictor variable in our model. Furthermore, as the demographics of the participants are available, another hierarchical model is developed assuming that there can be differences between these demographic groups. In this report, gender-based differences will be studied, as depression is known to exhibit different psychological patterns between males and females, and that the number of males and females are rather close in the participants.

3 Models

3.1 Pooled Logistic Regression Model

3.1.1 Model description

Logistic Regression would be a suitable choice for this problem, as the project is aimed to determine between depressed and non-depressed people (binary target label). In a logistic regression model, the target label y is the result of Bernoulli distribution, in which the outcomes are boolean value (0/1, yes/no, true/false) with a probability of true values denoted by p, or $P(y = 1) = p$.

In a logistic regression, this probability is expressed as a function of x (predictor variable):

$$p(x) = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

with α representing the log of the odds of the event, and β is the coefficient of the predictor variable (relationship).

Or that:

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

The $log\left(\frac{p}{1-p}\right)$ is denoted as the $logit^{-1}(p)$ (logit transformation of p). As such, the likelihood of this model is:

$$y \sim bernoulli(logit^{-1}(\alpha + \beta x))$$

y is the target label of the model, being either 1 (depressed) and 0 (non-depressed), x being the predictor variable, or the zero-proportion at 7:00.

The pooled model assumes that both gender groups follow the same model, with the same aforementioned distribution. In other words, both gender groups share the same α and β (same distribution for α and β).

3.1.2 Priors

A logistic regression model was fitted to the data, with the results being:

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-2.0027721	0.7571866	-2.645018	0.0081687
zero_proportion	0.0308387	0.0123964	2.487709	0.0128569

The intercept in the fitted model is α, while the coefficient for zero-proportion is β. From the summary of the fitted model, both these values have p-values $< 5\%$, meaning that they are statistically significant. Moreover, the estimated values for α and β are calculated, as well as the standard deviation.

Using these estimates, some information about the prior distribution of α and β. To be consistent with this analysis while also allowing room for variations, weakly informative priors are chosen, with α following a normal distribution with mean = -2 and standard deviation = 100, and β following a normal distribution with mean = 0 and standard deviation = 10.

$$\alpha \sim N(-2, 100^2)$$

$$\beta \sim N(0, 10^2)$$

3.1.3 Stan implementation

```
## data {
##   int<lower=0> N;  // number of observations
##   vector[N] x;     // zero-proportion of the observations
##   int<lower=0,upper=1> y[N];   // state of depression of the observation
##
##   int<lower=1> P;
##   vector[P] x_pred;
## }
## parameters {
##   real alpha;
##   real beta;
## }
## model {
##   // Priors
##   alpha ~ normal(-2, 100);
##   beta ~ normal(0, 10);
##
##   // Likelihood
##   y ~ bernoulli_logit(alpha + beta * x);
```

```
## }
##
## generated quantities {
##   int<lower=0,upper=1> y_pred[P];
##   vector[N] log_lik;
##   for (p in 1:P) {
##     y_pred[p] = bernoulli_logit_rng(alpha + beta*x_pred[p]);
##   }
##   for (n in 1:N) {
##     log_lik[n] = bernoulli_logit_lpmf(y[n] | alpha + beta*x[n]);
##   }
## }
```

3.2 Hierarchical Logistic Regression Model

3.2.1 Model descriptions

This model also uses Logistic Regression. However, this model will take into account the potential differences between female and male participants in the sample. This means that the parameter β (relationship between 2 variables) in the logistic regression are separate for each group, following a different distribution. Suppose that there are J groups, the likelihood y_i for each group is:

$$y_i \sim Bernoulli(logit^{-1}(\alpha + \beta_i x))$$

with $beta_i$ being the corresponding β for each group.

Nevertheless, in a hierarchical model, it is assumed that these different β parameters are generated from a hyper distribution, meaning that β_i are samples from another distribution. Assuming that this hyper distribution is a normal distribution with mean μ and standard deviation σ.

$$\beta \sim N(\mu, \sigma^2)$$

The distributions of μ and σ is unknown, which will be investigated by the Bayesian method with prior distributions and the given samples.

3.2.2 Priors

The prior distribution for α (log of the odds of the events) remains unchanged.

$$\alpha \sim N(-2, 100^2)$$

We previously assumed a weakly informative prior for μ similar to the previously used prior distribution for β in the pooled model. However, after conducting some runs and analysis, the model resulted in very low effective sample size. Hence, the prior distribution for μ is changed: the same mean but much higher variance to allow wider range. For the hyper parameter σ, a non-negative inverse chi squared prior distribution is chosen to ensure that the generated σ is non-negative. Some models were tested and the prior distribution that gives higher effective sample size is chosen for σ:

$$\mu \sim N(0, 100^2)$$

$$\sigma \sim Inv - \chi^2(10)$$

3.2.3 Stan implementation

```
## data {
##   int<lower=0> N;
##   int<lower=1> J;
##   vector[N] x;
##   int<lower=1,upper=J> g[N];
##   int<lower=0,upper=1> y[N];
##
##   int<lower=1> P;
##   vector[P] x_pred;
##   int<lower=1,upper=J> g_pred[P];
## }
## parameters {
##   real mu;
##   real<lower = 0> sigma;
##   real alpha;
##   vector[J] beta;
## }
## model {
##   mu ~ normal(0, 100);
##   sigma ~ inv_chi_square(10);
##   alpha ~ normal(-2, 100);
##   beta ~ normal(mu, sigma);
##   for (n in 1:N) {
##     y[n] ~ bernoulli_logit(alpha + beta[g[n]]*x[n]);
##   }
## }
## generated quantities {
##   int<lower=0,upper=1> y_pred[P];
##   vector[N] log_lik;
##   for (p in 1:P) {
##     y_pred[p] = bernoulli_logit_rng(alpha + beta[g_pred[p]]*x_pred[p]);
##   }
##   for (n in 1:N) {
##     log_lik[n] = bernoulli_logit_lpmf(y[n] | alpha + beta[g[n]]*x[n]);
##   }
## }
```

3.3 Fitting data to the models

```
pooled_data <- list(
  N = nrow(df),
  x = df$zero_proportion,
  y = df$depressed,
  P = nrow(df),
  x_pred = df$zero_proportion
)

hierarchical_data <- list(
  N = nrow(df),
  J = 2,
  x = df$zero_proportion,
  g = df$gender,
```

```
  y = df$depressed,
  P = nrow(df),
  x_pred = df$zero_proportion,
  g_pred = df$gender
)

SEED <- 1234
pooled_fit <- stan(file = "pooled_model.stan", data = pooled_data,
                   iter = 2000, chains = 4, seed = SEED,
                   control = list(adapt_delta = 0.99))
```

```
## Trying to compile a simple C file
```

```
hierarchical_fit <- stan(file = "hierarchical_model.stan", data = hierarchical_data,
                         iter = 2000, chains = 4, seed = SEED,
                   control = list(adapt_delta = 0.99))
```

```
## Trying to compile a simple C file
```

For each model, 4 chains are generated with length (iteration) = 2000. A seed is set for the results to be replicable. The parameter adapt_delta was increased as the hierarchical model encounters some convergences detected by the HMC diagnostic due to large step sizes in the sampler.

4 Convergence Diagnostics

The results from the fitted models are summarized as:

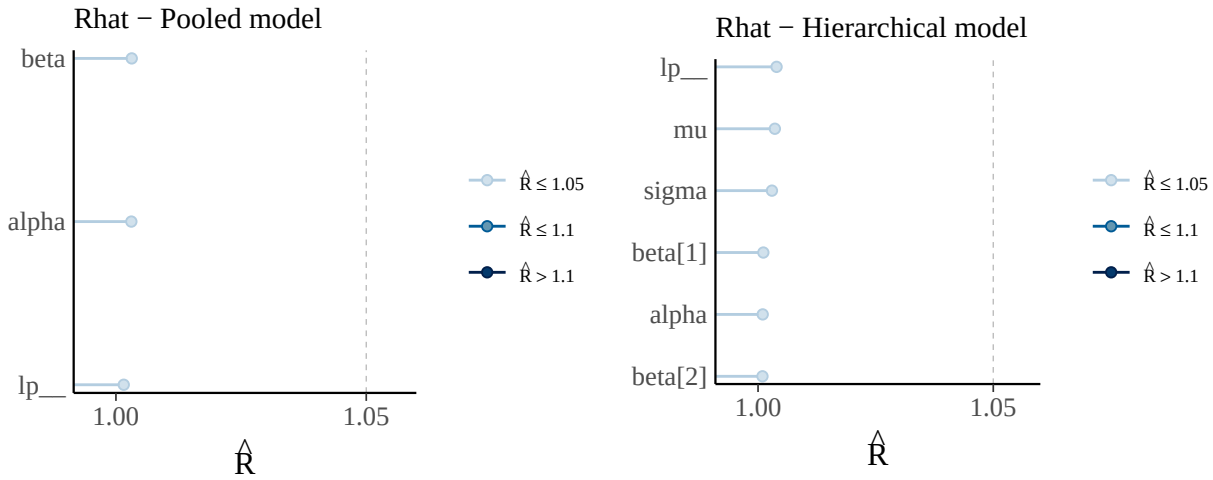
Pooled model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	-2.150963	0.0336123	0.8207466	-3.904662	-2.6837604	-2.0922651	-1.5843694	-0.6969685	596.2397	1.003077
beta	0.033271	0.0005435	0.0134735	0.008237	0.0239929	0.0326513	0.0418386	0.0619490	614.6494	1.003157
lp___	-34.831893	0.0377882	1.0909923	-37.662689	-35.2490260	-34.5124376	-34.0604897	-33.7729023	833.5505	1.001555

Hierarchical model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu	0.0298101	0.0025412	0.0798646	-0.1332911	-0.0122314	0.0319202	0.0755615	0.1796841	987.7057	1.003572
sigma	0.1015259	0.0016208	0.0491307	0.0428020	0.0684450	0.0901263	0.1201756	0.2307333	918.8984	1.002951
alpha	-2.1374294	0.0232885	0.7892894	-3.7778622	-2.6358149	-2.0962298	-1.6045717	-0.6909787	1148.6560	1.000992
beta[1]	0.0289447	0.0003895	0.0135686	0.0030220	0.0201142	0.0286764	0.0377072	0.0574669	1213.8206	1.001126
beta[2]	0.0376292	0.0003903	0.0138240	0.0123361	0.0279805	0.0371895	0.0465180	0.0665570	1254.2350	1.000942
lp___	-24.6553159	0.0574236	1.7530051	-29.1449093	-25.5517861	-24.2998860	-23.3597452	-22.3418950	931.9353	1.003940

4.1 Rhat convergence diagnostics



As seen from the summary tables and the plots, Rhat values for both models are very close to 1, much lower than the 1.05 threshold. Such low Rhat values indicate that the chains have converged and the estimates are relatively reliable.

4.2 HMC Convergence Diagnostics

HMC Convergence Diagnostics are checked using functions from rstan:

```
check_hmc_diagnostics(pooled_fit)
```

```
##
## Divergences:
## 0 of 4000 iterations ended with a divergence.
##
## Tree depth:
## 0 of 4000 iterations saturated the maximum tree depth of 10.
##
## Energy:
## E-BFMI indicated no pathological behavior.
```

```
check_hmc_diagnostics(hierarchical_fit)
```

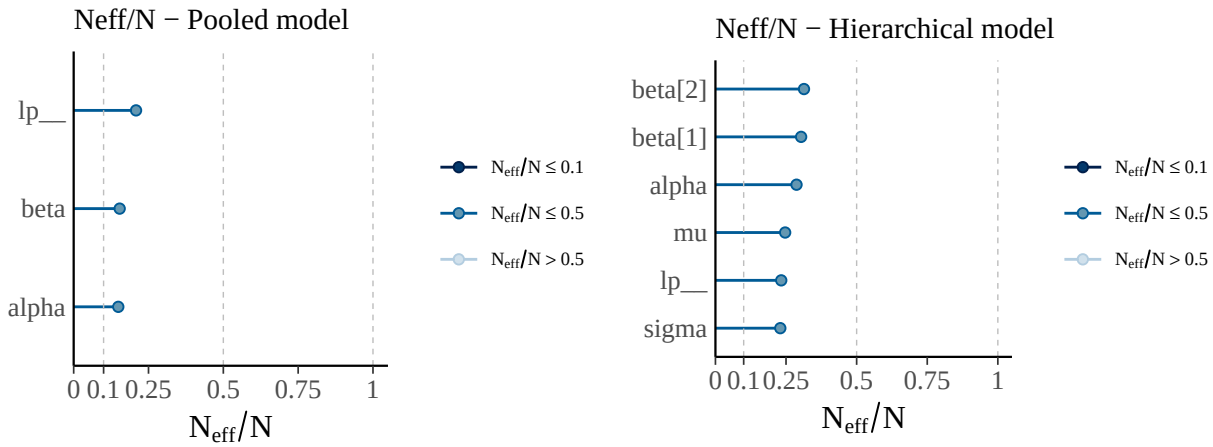
```
##
## Divergences:
```

```
## 0 of 4000 iterations ended with a divergence.
##
## Tree depth:
## 0 of 4000 iterations saturated the maximum tree depth of 10.
##
## Energy:
## E-BFMI indicated no pathological behavior.
```

As such, no divergences in the iterations are detected, and the maximum tree depth is not exceeded. Hence, the estimates are quite reliable.

4.3 Effective sample size

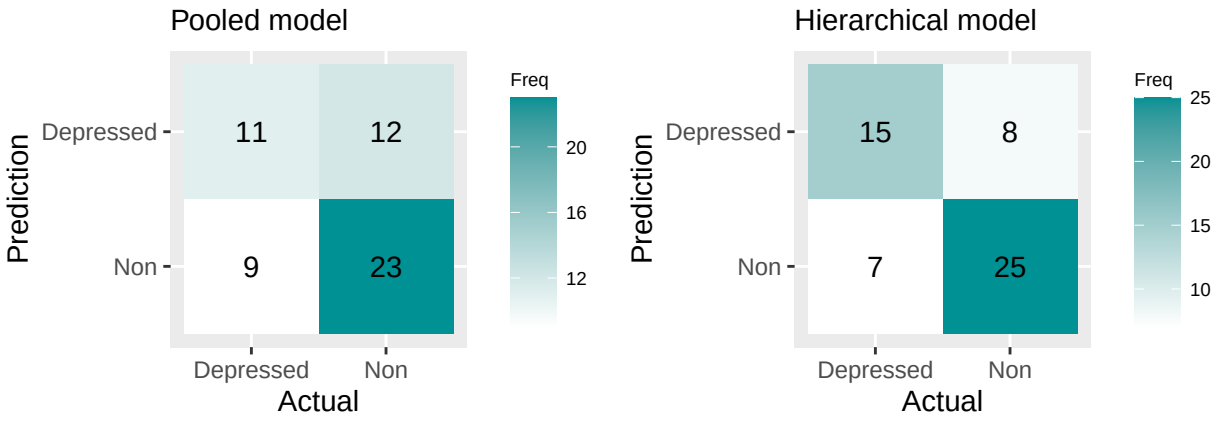
The effective sample size are seen from the summary tables (n_eff). Furthermore, the ratio of the effective sample size per the actual sample size can be visualized:



Effective sample size represents the number of independent samples from the drawn samples. Hence, the closer the effective sample size to the drawn sample size, the more reliable the samples. The threshold for the ratio of effective sample size to the actual sample size is 0.1, meaning that values below this threshold signifies potential problems. Nevertheless, the above plots show such ratios for both models. For both models, the ratios fall into the range between 0.1-0.5. While these are not a good effective sample sizes, they are still above the minimum threshold.

5 Posterior predictive checks

The same data (x-values) for training the models is used to calculate the accuracy of the models on training set. The mean (expected value) of the posterior predictive distribution y-values are calculated to determine the state of depression for each participant based on such distribution. Finally, the results are summarized for all participants to formulate a confusion matrix for both models:



Based on the confusion matrix, the accuracy of the pooled model on the training set is approximately 61.82%, while the accuracy of the hierarchical model on the training set is approximately 72.73%. Furthermore, the conditional probability of actual depression with positive prediction from the pooled model is rather low, only approximately 47.83%; but the conditional probability of actual depression with positive prediction from the hierarchical model is higher, up to 65.22%. As such, the hierarchical model performs better based on confusion matrix on the training set.

6 Predictive performance assessment

The predictive performance assessment will be conducted with a similar manner to the posterior predictive checks. Because of the small dataset, splitting into training and testing set will be extremely difficult and the result would be unreliable. Hence, the performance assessment will be done with KFold cross validation in which at each iteration (each fold), a different training and testing set will be generated. At the final step, the average accuracy score throughout the folds will be reported. The number of folds used in this report is 5, meaning that at each fold, the training set will be of size 44 and the test set is of size 11. Furthermore, the conditional probability of true positive (depression) with positive prediction is also calculated and averaged throughout the folds. The results are as following:

	Pooled model	Hierarchical model
Accuracy	0.6363636	0.6545455
Conditional probability	0.6188889	0.7157143

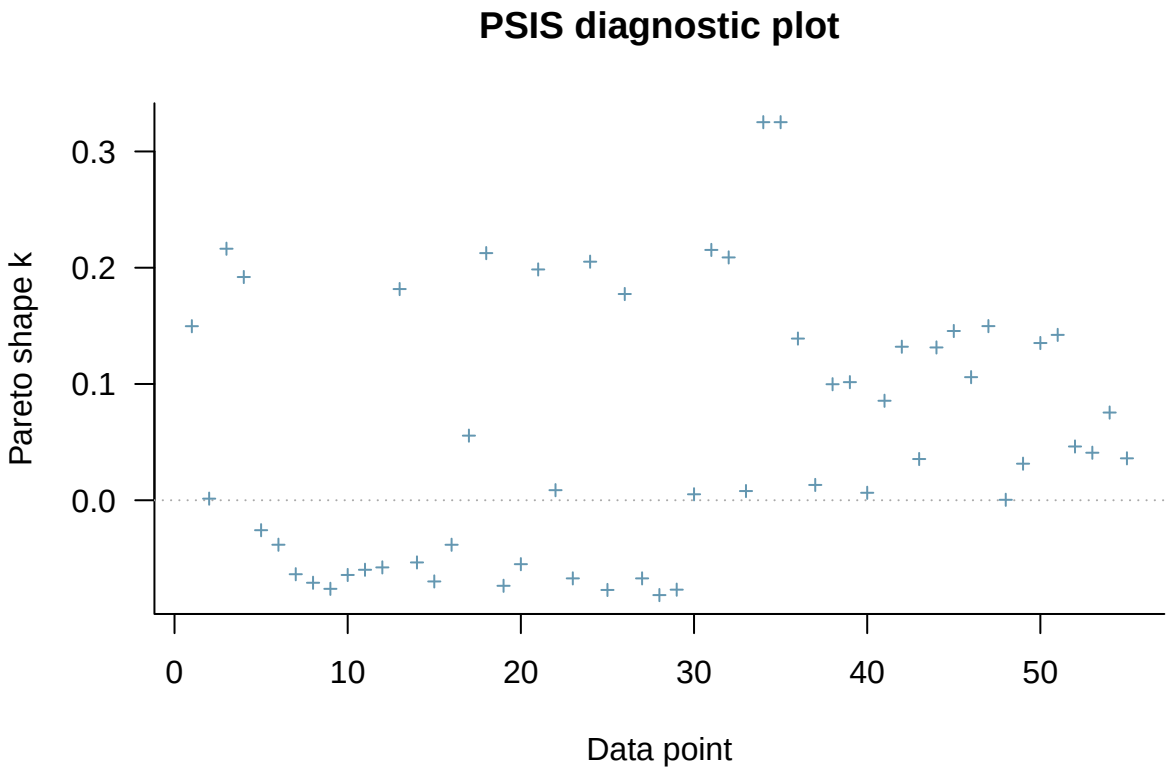
The performance of both models on the test sets are decent (providing from 60% to 70% accurate predictions). The conditional probability that a person is depressed with a positive (depressed = 1) prediction from the pooled model is decent, while such conditional probability by the hierarchical model falls into the good range (from 70% to 90%). Furthermore, the performance of the hierarchical model is better with both of these metrics.

7 Model comparison

The models are compared using PSIS-LOO diagnostics. The results for both models are as following:

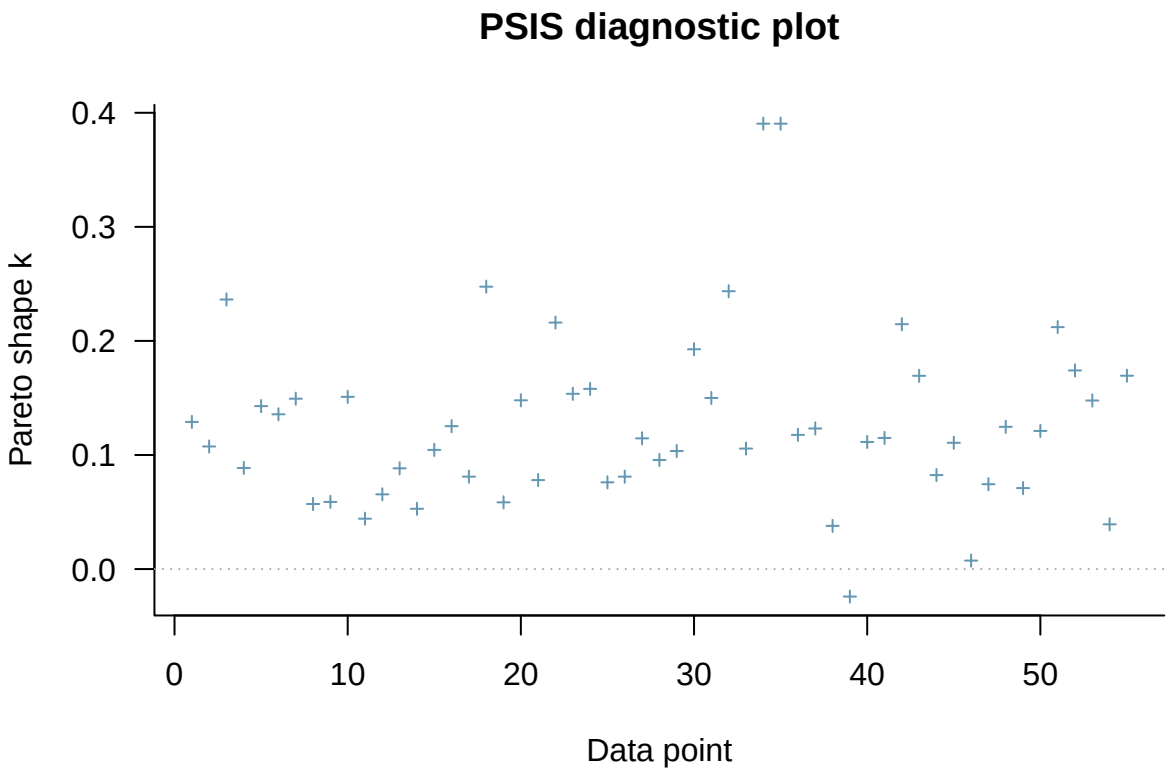
Pooled model

	Estimate	SE
elpd_loo	-36.298892	3.5997449
p_loo	2.697058	0.7420304
looic	72.597784	7.1994898



Hierarchical model

	Estimate	SE
elpd_loo	-36.954797	3.7704912
p_loo	3.739856	0.7399494
looic	73.909593	7.5409825



All Pareto-k values for both models are below the threshold 0.7, meaning that both models generate reliable estimates. However, the elpd_loo value of the pooled model is higher, meaning that the pooled model will be better according to the PSIS-LOO diagnostics.

8 Prior Sensitivity Analysis

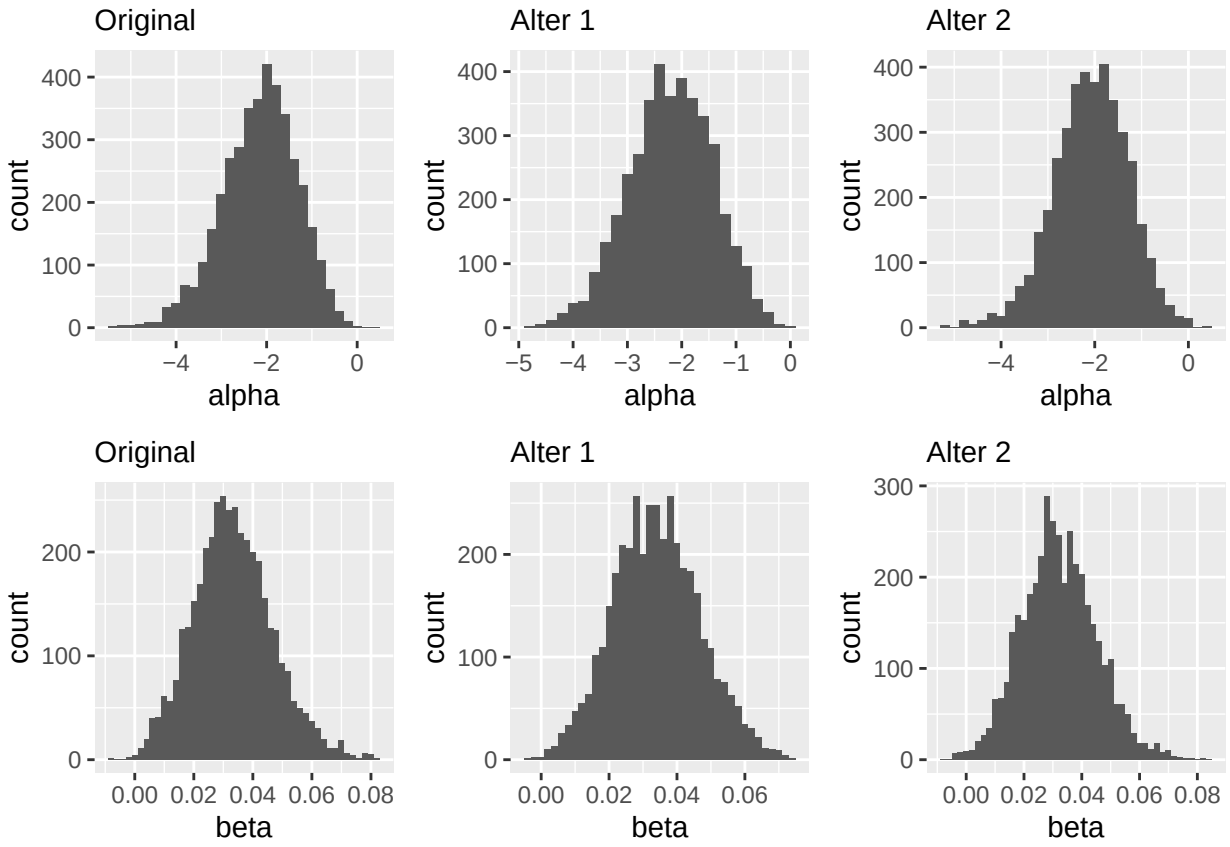
8.1 Pooled model

For the pooled model, the original model uses:

$$\alpha \sim N(-2, 100^2)$$

$$\beta \sim N(0, 10^2)$$

As such, for the alternative models, one model (alter1) will investigate the effect of changing prior distribution for α by using $\alpha \sim N(-10, 10)$ (changing the center point as well as giving smaller variances). Another model (alter2) will investigate the effect of changing prior distribution for β by using $\beta \sim N(0, 1)$ (much smaller variance). The posterior distributions of α and β between the 3 models are visualized as:



As such, these different prior distributions do not make a significant impact on the posterior distributions. As such, the pooled model is not sensitive to the choice of prior distribution.

8.2 Hierarchical model

For the hierarchical model, the prior distributions used are:

$$\alpha \sim N(-2, 100^2)$$

$$\mu \sim N(0, 100^2)$$
$$\sigma \sim Inv - \chi^2(10)$$

As such, several other models will be implemented using respective different prior distributions. The first alternative model (Alter 1) will use a different prior distribution for α:

$$\alpha \sim N(-10, 10^2)$$

The second alternative model (Alter 2) will use a different prior distribution for μ:

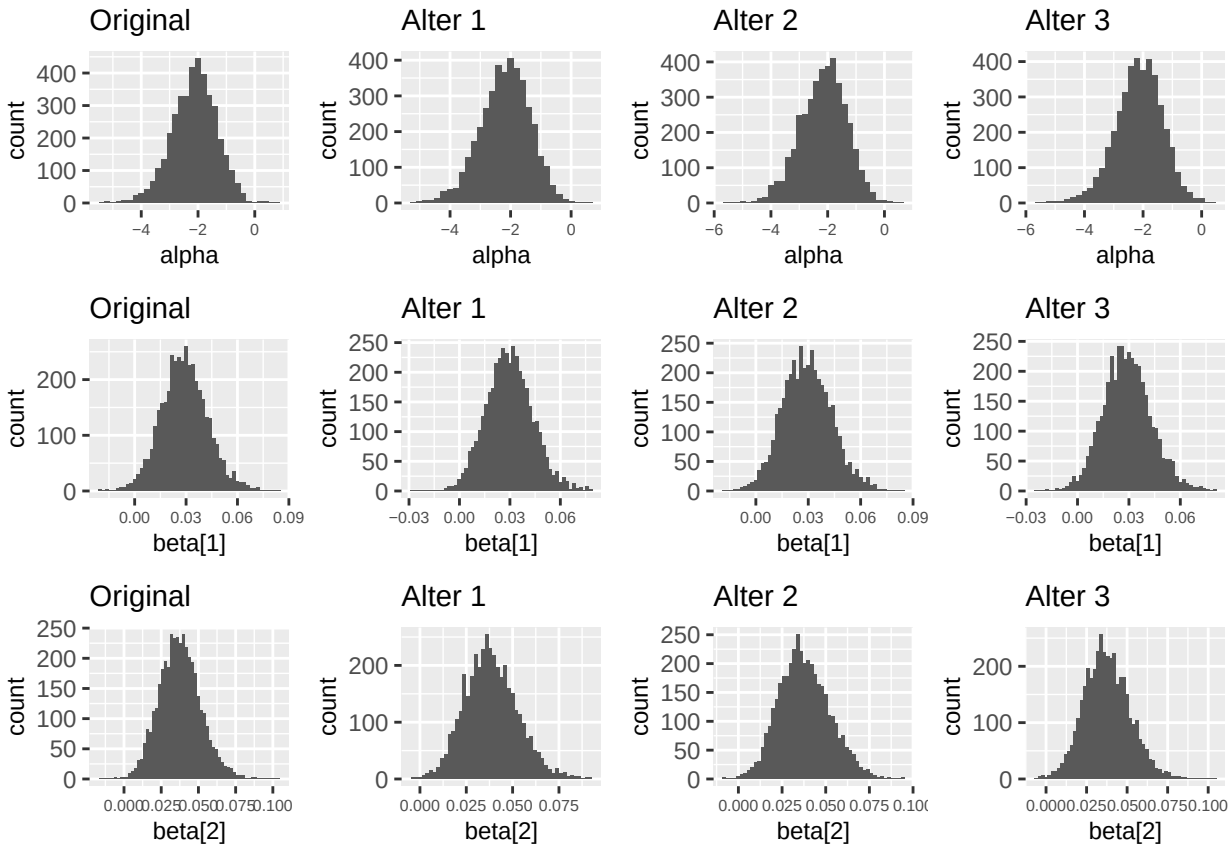
$$\mu \sim N(1, 10^2)$$

The third alternative model (Alter 3) will use a different prior distribution for σ:

$$\sigma \sim Inv - \chi^2(1)$$

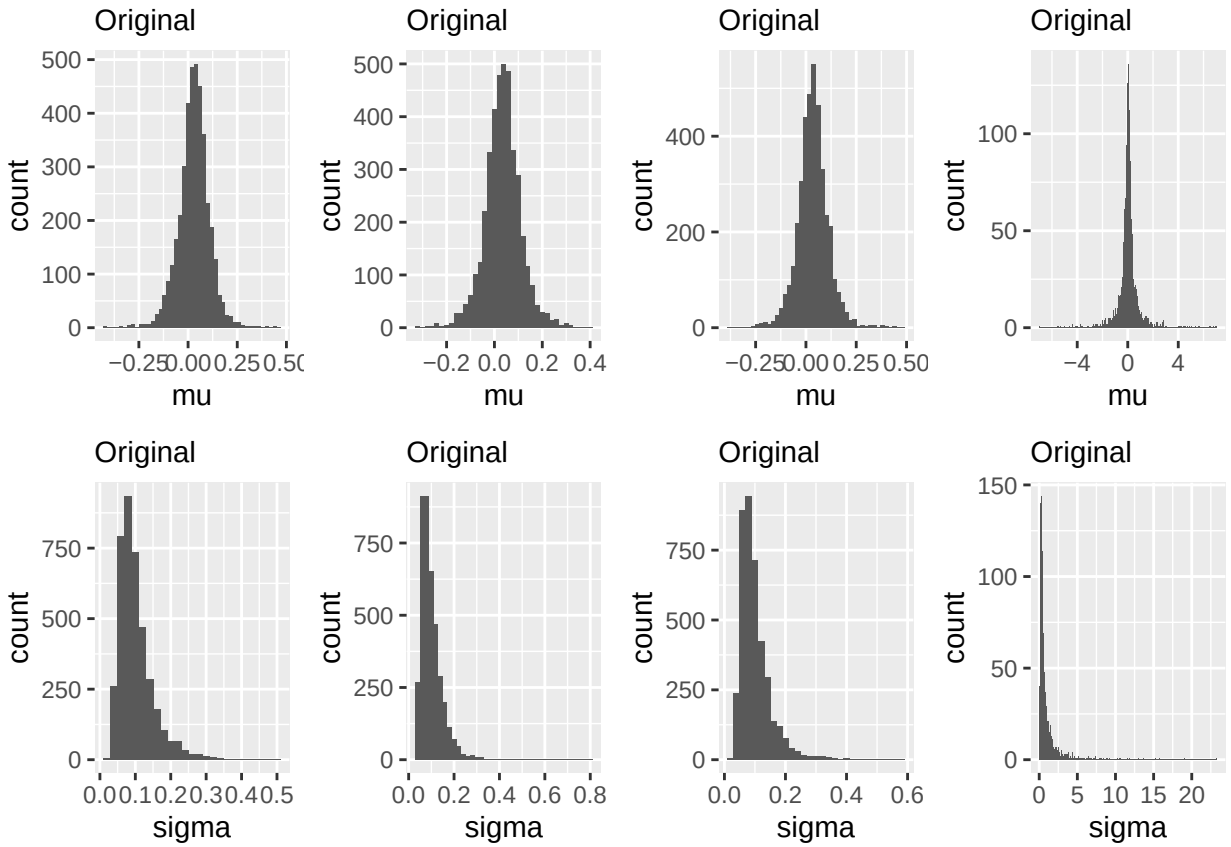
As previously noted, some choices of alternative prior distributions affect the convergence of the Markov chains in the hierarchical model. As such, the model is more sensitive to prior choice in this regards.

Comparing the posterior distribution of α, $\beta[1]$ and $\beta[2]$:



It is visible that different prior distributions do not lead to significant differences in the posterior distributions for α, $\beta[1]$ and $\beta[2]$, the direct parameters of the logistic regression model.

Comparing the posterior distributions for μ and σ:



There are no significant variations between the original, the first and second alternative models. However, there are some differences between the last alternative model and other models. As such, the posterior distribution of the hyper parameters can be influenced by the choice of prior distribution of the parameter σ. In conclusion, the hierarchical model is more sensitive to the choice of prior distribution than the pooled model, but the influence is not detrimental.

9 Discussion and Conclusion

The goal of this project is to create two different logistic regression models to predict whether a person has depression or not based on their brains' activity. The model's performance so far has been rather decent. The pooled model is better bases on PSIS-LOO analysis, but the predictive performance of the hierarchical model is better with the current dataset. Though the model might not be up to standard for medical usage, we hope to show that by applying Bayesian analysis methods, a reasonable model can be created to help professionals in their field to predict depression in patients.

Based on the model, zero proportion is a good predictor for depression, especially around 7:00 AM, where the difference between the zero proportion of the depressed and non-depressed is the highest, up to a 20% difference. This discovery is quite interesting, as this suggests that people who are less active in the morning have a higher chance to be diagnosed with depression than those who are. Another noteworthy point is that depression can be somewhat influenced by gender, with depression in male participants having a slightly higher coefficient for the correlation the zero_proportion at 7:00 AM based on the results of the hierarchical model.

The model performance is decent but can be improved further. This is due to the fact that our dataset is quite small, with only around 55 data points. Although there are a lot of parameters that can be chosen from the raw data, in the end, we decided on zero_proportion after some cleaning as it shows the clearest correlation. So, an obvious way to increase the model's accuracy is to have a larger dataset, this will also

help minimize outliers' influence on the result as the smaller the dataset, the more impressionable they are by the outliers.

Our group did not have much knowledge in the field of psychology, thus, our prior choices were very general and weakly informative. But this can be improved with the help of experts in the field, allowing us to have more informative priors, leading to a better model overall. Another way to improve the model with the help of experts is by changing the parameters chosen. In the report, we used zero proportion as the parameter but this can be changed to the MADRS score system to both improve the accuracy of the model and give us a more in-depth look at the severity of the depression.

In conclusion, the results we have gathered from the models are good but there are rooms for improvement. Furthermore, the models as of now are quite simple and can be a basis for further research, giving us more reasons and motivations to work on a more specialized and complete model of Bayesian analysis.

10 Self-reflection

The project so far has been a positive experience for us. This is because it allows us to internalize what we have learnt so far from the course and put them to use. Furthermore, we have also learnt various skills from planning the workload to implementing the actual Bayesian model. We have also learnt various skills that will be useful outside of the course like how to process some data, how to consolidate them, etc. And finally, we were able to improve our knowledge on Bayesian analysis methods and topics by using them, from diagnostics to convergence, comparing and seeing the difference between the model, choosing priors, etc. All in all, the course and the project itself has helped us a lot.