Depression Prediction and Classification using Bayesian Model

— Project Report —

Advanced Bayesian Data Analysis

Mohammad Istiak Mahbub
Mukta Ghosh
Naim Ahmad

March 17, 2024

TU Dortmund University

Contents

1 Introduction

Mental health is a critical aspect of overall well-being that impacts an individual's everyday life. Depression is one kind of mental health disorder characterized by feelings of sadness, hopelessness, and a lack of interest in everyday activities. The World Health Organization estimates that 1.3 million Kenyans suffer from depression. The Busara Center for Behavioral Economics organized a competition in collaboration with the data science community to explore the possibility of predicting the percentage of depression cases for Kenyans using routine survey data [Zindi, 2018].

The survey collects information about household composition, financial and economic activity, and health. The primary goal of this report is to examine if a regular lifestyle can affect to become depressed and if the Bayesian model predicts and classifies depressed and non-depressed people. For this project, a dataset is sourced from Kaggle, a platform for data science competitions and datasets Babativa [2019]. The dataset has 23 columns and a total of 1429 observations. As the pupil's basic personal information is the main focus, 8 columns within 23 columns are selected.

Descriptive statistics were used to describe the distribution of the covariates. A correlation analysis is used to provide insight into the relationship between the variables. Regression analysis is employed to assess the extent of the association between variables. Additionally, the analysis includes an evaluation of the nonlinearity of covariates, with the possibility of applying polynomial regression to mitigate nonlinearity. The second section, provides a more detailed overview of the dataset, data quality, and description of variables. The necessary statistical methods are presented and explained in the third section.

Firstly, data is preprocessed by checking missing values, downsampling, and factorization. Also, a descriptive analysis for selected columns is shown by histogram. For predicting and classifying depressed people, five Bayesian logistic regression models are applied with different covariates. As the outcome or target value is binary, Bernoulli and logit link functions are used. Two covariates are used for the first model. For the second model, $number\ of\ children$ is grouped by $village\_id$. For the third model, six covariates are fitted. In the fourth model with specific prior, a standardization process is applied as a preprocessing step to bring covariates in a similar scale. The last model is a multi-level model with different prior. Finally, a model comparison is shown using leave-one-out cross-validation. At last, in the fifth section, the main findings of this project are summarized.

2 Problem Statement

2.1 Description of the dataset

The selected dataset is sourced from Kaggle and survey data is collected by Busara Center for Behavioral Economics. The data consisted of a study about the life conditions of people in rural zones. The dataset has 1429 observations and 23 columns. As the people's basic personal information is the main focus, 8 columns within 23 columns are selected, where one is a dependent variable and the 7 others are independent variables. The $depressed$ is a binary dependent variable. And other independent variables are, sex, $Married$ with binary value and $Village_id$, Age, $Education\ level$, $Number\ of\ children$, $Total\ members$ are with numeric value.

There are no missing values in our selected dataset. So, data quality can be considered as good since there is no missing data. The data in its raw form can be found on the Zindi platform in Africa. [Zindi, 2018].

2.2 Project objectives

The objective of this project is to apply Bayesian modeling techniques to predict cases of major depressive disorder (MDD) using data. Specifically, the project aims to:

- Explore the use of Bayesian modeling as a tool for predicting depression.
- Preprocess and analyze the provided data to ensure its suitability for Bayesian modeling.
- Design and implement Bayesian models that can effectively identify at risk of depression.
- Consider the incorporation of prior knowledge and domain expertise into the Bayesian models, particularly emphasizing basic personal information factors known to influence mental health outcomes.
- Evaluate the performance of the Bayesian models using appropriate metrics and validation techniques.
- Interpret the results and insights generated by the Bayesian models to provide actionable recommendations for mental health

By achieving these objectives, the project seeks to demonstrate the utility of Bayesian modeling in predicting depression cases and provide insights into potential avenues for intervention and support for individuals experiencing depression, thereby contributing to the improvement of mental health outcomes in Kenya and beyond.

2.3 Data preprocessing

Histogram

Initially, a histogram was created to visualize the distribution of the selected columns from the dataset. This allowed for a better understanding of the data distribution and potential insights into the variables' characteristics.

Factorization

The categorical variables "sex" and "married" were factorized to represent them numerically. For example, "sex" was encoded as 0 for female and 1 for male, and "married" was encoded as 0 for no and 1 for yes. This transformation facilitated the inclusion of categorical data in regression models. Additionally, the target variable "depression" was encoded as 0 for no depression and 1 for depression, making it suitable for regression modeling.

Missing Values

A thorough check for missing values was conducted across the selected columns. Fortunately, no missing values were found in the chosen columns, ensuring the integrity of the dataset.

Downsampling

Prior to downsampling, the distribution of the target variable "depression" indicated a class imbalance, with a significantly higher number of "no" cases compared to "yes" cases.

Downsampling was performed to address the class imbalance by reducing the majority class ("no" cases) to match the minority class ("yes" cases). After downsampling, both "yes" and "no" classes comprised 238 instances each, ensuring a balanced dataset for model training and evaluation. This approach helped prevent bias towards the majority class and improved the predictive performance of the models on both classes.

3 Statistical Methods

3.1 Bayesian model

Bayesian models are a class of statistical models that use Bayes' theorem for inference. Unlike traditional "frequentist" statistics, Bayesian statistics incorporates "prior" knowledge or beliefs about the parameters in the model. This prior is then updated with observed data to provide a "posterior" distribution of the parameters.

The key idea behind Bayesian models is the concept of updating beliefs based on observed data. These beliefs are represented as probabilities, which are updated as new data is observed. This updating process is done using a fundamental theorem in probability theory, known as Bayes' theorem.

One of the unique aspects of Bayesian models is the incorporation of prior knowledge. This prior knowledge is represented as a probability distribution, which is then updated with the observed data to provide an updated probability distribution. This updated distribution is often referred to as the posterior distribution.

Bayesian models are particularly useful when we have limited data, as they allow us to incorporate prior knowledge into our inferences. This can lead to more robust and accurate estimates, especially in complex models with many parameters.

One of the challenges with Bayesian models is computational. The process of updating the prior often involves complex integrals that cannot be solved analytically. However, there are many numerical methods available for approximating these integrals, such as Markov chain Monte Carlo (MCMC) methods. Andrew Gelman [2013] McElreath [2019].

3.2 Prior

In Bayesian statistics, the term "prior" refers to the initial knowledge or belief about an unknown parameter before observing any data. This prior knowledge is represented as a probability distribution over the possible values of the parameter. Informative priors are chosen to incorporate substantial prior knowledge, while uninformative or weakly informative priors are used when little prior information is available. The prior distribution serves as a regularization mechanism, allowing for the incorporation of domain expertise and improving the stability

and interpretability of the Bayesian model. The choice of the prior can be subjective and is often based on expert knowledge or previous studies. The prior is an essential component of Bayesian models as it allows for the incorporation of external knowledge into the statistical analysis. Andrew Gelman [2013] McElreath [2019].

3.3 Bayesian multilevel

Multilevel models, also known as hierarchical or mixed effects models, have become essential in various scientific fields due to their ability to handle complex data structures and address issues like overfitting. These models utilize partial pooling, a statistical technique that pools information across units in the data to produce more accurate estimates for all units. Multilevel models are particularly useful in scenarios involving repeat sampling, imbalanced sampling, the study of variation among groups, and the preservation of uncertainty in pre-averaged data. They are applicable to a wide range of contexts, including missing data, measurement error, factor analysis, and spatial regression. Despite their benefits, fitting and interpreting multilevel models can be challenging, requiring careful consideration and expertise. However, as the field progresses, multilevel regression is increasingly advocated as the default approach, with papers needing to justify not using it. This shift mirrors the adoption of multivariate tools in applied statistics, suggesting that multilevel models will eventually become standard practice across disciplines. Andrew Gelman, 2013, 14-16.

3.4 Bayesian model comparison

Bayesian data analysis provides a way for different models to understand data. The two approaches cross-validation and information criteria aim to compare models based on expected predictive accuracy. First, these approaches provide useful expectations of predictive accuracy, rather than fitting the data precisely. In this way, they compare models where needed. Second, they give us an estimate of the overfitting tendency of a model, meaning giving information on how well data and model are interacting. Also helps us to find out the most influential observations [McElreath, 2019, p.13-14].

In this project, the leave-one-out cross-validation method is chosen for comparing models. Leave-one-out cross-validation (LOOCV) is a robust method for model comparison. In this method, one observation is used as the validation set, and the rest are used as the training set.

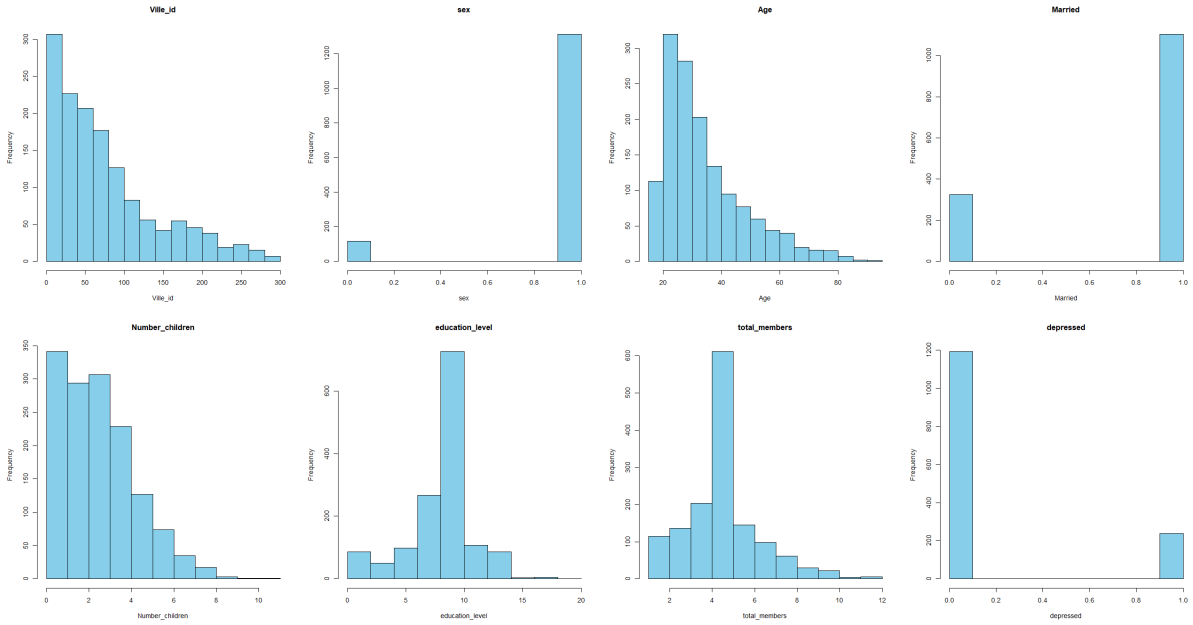

Figure 1: Histogram of all selected covariates including target variable

This process is repeated for each observation in the dataset, hence the term “leave-one-out”. The model’s performance is then averaged over all iterations to provide an overall measure of its predictive accuracy Hastie and Friedman [2016]

4 Statistical Analysis

4.1 Descriptive analysis

Histogram

To begin with, a histogram was generated in Figure 1 to illustrate the distribution of chosen columns from the dataset. This visualization facilitated a deeper comprehension of the data distribution and offered potential insights into the attributes of the variables.

Summary statistics

- For the variable Age, the mean age is approximately 36.03 years, with a median of 32 years. The youngest individual is 17 years old, while the oldest is 87 years old.

- The variable $Number\_children$ has a mean of approximately 2.975 children, with a median of 3 children. The range of the number of children is from 0 to 11.
- The $education\_level$ variable has a mean of approximately 8.223 years of education, with a median of 9 years. The range of education levels is from 1 to 18 years.
- The $total\_members$ variable has a mean of approximately 5.071 members, with a median of 5 members. The range of total members is from 1 to 12.

The covariates included in the analysis are presented in Table 1

Frequency distribution

Upon examining the dataset, found the following for sex and $Married$:

- Regarding gender distribution, the dataset now consists of 41 females and 435 males.
- In terms of marital status, the downsampling resulted in 127 individuals classified as not married and 349 individuals classified as married

Furthermore, after downsampling, the dataset contains an equal number of individuals diagnosed with depression (coded as 1) and individuals without depression (coded as 0), with 238 individuals in each category. This balanced representation ensures that the dataset is suitable for further analysis and modeling without bias towards any particular category.

Cross-tabulation

The cross-tabulation of sex and $Married$ shows the distribution of marital status among females and males in Table 2

4.2 Model with two covariates

In this model, select two covariates, namely sex and $Married$, to predict the target variable depressed. Logistic regression is employed, with the logit link function used to transform the linear combination of predictor variables into a probability scale. The logit link function is defined as:

$$logit(p) = log(p/(1p))$$

The linear predictor is then transformed using the logit link function to obtain the predicted probabilities of depression. This model aims to explore the relationship between the selected

covariates and the likelihood of depression, providing insights into how sex and marital status influence the probability of being depressed.

Model summary

```
> summary(M1)
 Family: bernoulli
  Links: mu = logit
Formula: depressed ~ sex + Married
   Data: df_updated (Number of observations: 476)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      0.04      0.33    -0.62     0.69 1.00     4632     3072
sexmale        0.13      0.35    -0.55     0.81 1.00     3719     3101
Marriedyes    -0.21      0.21    -0.64     0.21 1.00     3446     2785

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

The interpretation of the $intercept$, $sexmale$, and $Marriedyes$ parameters in the context of the model is as follows:

intercept: The $intercept$ parameter represents the average value of the outcome variable (depression) when both sex and marital status are held constant. In other words, it indicates the baseline level of depression when no other factors are considered.

sexmale: The $sexmale$ parameter represents the difference in the outcome variable (depression) between males and females while holding marital status constant. A positive value suggests that males are more likely to experience depression compared to females, while a negative value suggests the opposite. The magnitude of the parameter indicates the strength of this association.

Marriedyes: The $Marriedyes$ parameter represents the difference in the outcome variable (depression) between married and unmarried individuals while holding sex constant. A positive value suggests that married individuals are more likely to experience depression compared to

unmarried individuals, while a negative value suggests the opposite. Similar to sexmale, the magnitude of the parameter indicates the strength of this association.

These interpretations provide insights into how each covariate contributes to the likelihood of experiencing depression while accounting for the effects of other variables in the model.

Trace plot

The trace in Figure 8 provides valuable insights into the convergence and variability of the models. By examining trace plots, can assess whether the sampling process was reliable and stable.

Specifically, Figure 8 shows the trace plots and density plots for each parameter across the four chains. A well-behaved trace plot exhibits random fluctuations around a central value, indicating convergence to a stationary distribution. From the plots, it is evident that the models have converged to similar values across the four chains. This convergence indicates that the sampling process was consistent and reliable, enhancing confidence in the estimated parameter values.

Autocorrelation plot

Autocorrelation measures the correlation between a variable's values at different lags or time intervals. A flat autocorrelation plot indicates that there is little to no correlation between successive samples from the MCMC chain, which is indicative of convergence. In other words, the samples are effectively representing the posterior distribution of the parameters without being influenced by previous samples.

The autocorrelation plots in Figure 9 show that the $Intercept$, $sexmale$, and $Marriedyes$ variables display a relatively flat profile, with values close to zero. This observation suggests that the Markov Chain Monte Carlo (MCMC) chains for these variables have converged and exhibit randomness.

Parameter estimation

These odds ratios which are shown in Figure 10 offer a quantitative understanding of the impact of each covariate on the likelihood of depression, providing valuable information for

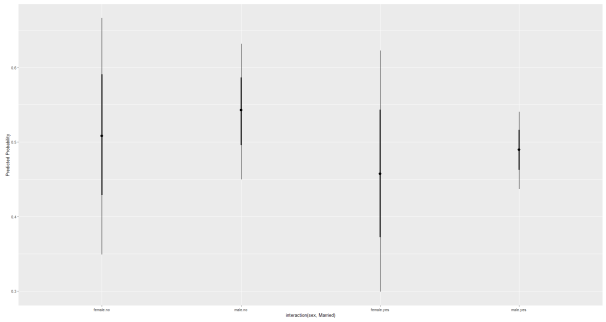

Figure 2: Parameter effects for Model 1

interpreting the model results and making informed decisions. The odds ratios provide additional insights into the relationship between the covariates and the likelihood of experiencing depression:

The odds ratio of the $intercept$ is 1.03, indicating that the odds of being depressed for the reference group (where sex and marital status are constant) are 1.03 times higher compared to not being depressed.

For $sexmale$, the odds ratio is 1.14. This suggests that the odds of being depressed for males are 1.14 times higher than for females, holding marital status constant. In other words, males have a 1.14 times higher likelihood of experiencing depression compared to females in the same marital status category.

The odds ratio of $Marriedyes$ is 0.81, implying that the odds of being depressed for married individuals are 0.81 times the odds of being depressed for unmarried individuals, holding sex constant. This indicates that being married is associated with lower odds of experiencing depression compared to being unmarried.

Parameter effects

The predicted probabilities for experiencing depression for different groups based on sex and marital status are as follows:

- For females who are not married:

$$p = \frac{1}{1 + e^{-(0.04 + 0.13 \times 0 - 0.21 \times 0)}} = \frac{1}{1 + e^{-0.041}} \approx 0.49$$

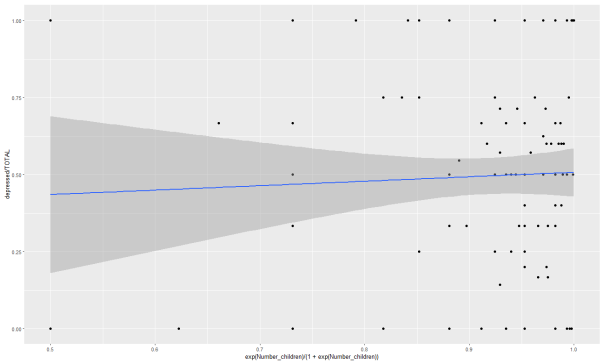

Figure 3: The proportion of people who are depressed related to the inverse-logit of number of children

These predicted probabilities provide estimates of the likelihood of experiencing depression for each group based on their sex and marital status, as determined by the logistic regression model.

4.3 Model with Group by covariate

Logistic regression can also be applied to model count or proportion data. Binary logistic regression assumes that the outcome variable follows a Bernoulli distribution, where the number of trials n is 1, allowing the outcome variable to take values of 1 or 0. On the other hand, binomial logistic regression extends this concept by assuming that the outcome variable follows a binomial distribution with n trials and probability q, allowing for handling count data where the outcome variable can take any non-negative integer value. Andrew Gelman [2013] McElreath [2019]

For Model 2, information about individual people is clustered within each village. By aggregating the data based on village id, a new dataset is created where each row represents a village. In this new dataset, the variable $depressed$ indicates the number of people who are depressed per village, $TOTAL$ represents the total number of people in each village, and $Number\_children$ denotes the average number of children for people per village.

The analysis reveals a slight positive relationship between the proportion of people who are depressed and the inverse-logit of the average number of children. It is important to note that for Model 2, variable $avg(number\_children)$ as its inverse-logit because in a binomial regression

model, also assumes a linear relationship between the inverse-logit of the linear predictor and the outcome (i.e., the proportion of events), rather than linearity between the predictor itself and the outcome.

```
> summary(M2)
 Family: binomial
  Links: mu = logit
Formula: depressed | trials(TOTAL) ~ Number_children
   Data: df_Prop (Number of observations: 180)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          -0.08      0.26    -0.60     0.43 1.00     3369     2078
Number_children     0.03      0.08    -0.13     0.19 1.00     3468     2168

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

The intercept parameter in Model 2 represents the average value of the outcome variable ($depression$) when the average number of children ($Number\_children$) is held constant. Additionally, the coefficient for $Number\_children$ indicates the change in the log odds of depression associated with a one-unit increase in the average number of children within each village. However, it's important to note that based on the provided summary and figures in Appendix A (Figure 12, Figure 13, Figure 14), the coefficient for $Number\_children$ is not statistically significant, as its 95% credible interval includes zero. Therefore, the relationship between the average number of children and the probability of depression may not be statistically meaningful in this model.

4.4 Model with six covariates

In the third model, a logistic regression is performed with six covariates: sex, Age, $Married$, $Number\_children$, $education\_level$, and $total\_members$. The target variable is depressed, which represents the likelihood of depression. The model aims to predict the likelihood of depression ($depressed$) based on these six covariates. The logistic regression estimates the

effect of each covariate on the log odds of depression while controlling for the effects of other covariates.

```
> summary(M3)
 Family: bernoulli
  Links: mu = logit
Formula: depressed ~ sex + Age + Married + Number_children + education_
    level + total_members
   Data: df_updated (Number of observations: 476)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept         -1.39      0.69    -2.73    -0.06 1.00     2996     2686
sexmale            0.25      0.36    -0.45     0.97 1.00     4135     3139
Age                0.02      0.01     0.01     0.04 1.00     3648     2882
Marriedyes         0.02      0.25    -0.45     0.51 1.00     3966     3335
Number_children   -0.13      0.08    -0.28     0.03 1.00     2960     2756
education_level    0.01      0.04    -0.06     0.08 1.00     3818     2822
total_members      0.14      0.08    -0.02     0.31 1.00     2863     2675

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Here's a brief interpretation of the output:

- **Intercept:** The estimated log odds of depression when all covariates are zero is -1.39. However, the interpretation of the intercept in logistic regression often doesn't have a practical meaning, especially when the covariates aren't centered or don't make sense at zero.
- **sexmale:** The log odds of depression increase by 0.25 for males compared to females, holding all other variables constant. However, the 95% CI includes zero (-0.45 to 0.97), suggesting that the effect of sex may not be statistically significant.
- **Age:** The log odds of depression increase by 0.02 for each additional year of age, holding all other variables constant. The 95% CI (0.01 to 0.04) does not include zero, suggesting that age has a statistically significant effect on depression.
- **Marriedyes:** Being married increases the log odds of depression by 0.02 compared to

not being married, holding all other variables constant. However, the 95% CI includes zero (-0.45 to 0.51), suggesting that marital status may not have a statistically significant effect on depression.

- **Number_children:** Each additional child decreases the log odds of depression by 0.13, holding all other variables constant. However, the 95% CI includes zero (-0.28 to 0.03), suggesting that the number of children may not have a statistically significant effect on depression.
- **education_level:** Each additional level of education increases the log odds of depression by 0.01, holding all other variables constant. However, the 95% CI includes zero (-0.06 to 0.08), suggesting that education level may not have a statistically significant effect on depression.
- **total_members:** Each additional member in the household increases the log odds of depression by 0.14, holding all other variables constant. However, the 95% CI includes zero (-0.02 to 0.31), suggesting that the total number of household members may not have a statistically significant effect on depression.

The **Rhat** values are all 1.00, indicating that the chains have likely converged. The $Bulk\_ESS$ and $Tail\_ESS$ are measures of the effective sample size, which suggest how well the posterior distribution is estimated. Higher values are better, and as a rule of thumb, values above 1000 are usually considered satisfactory.

Trace plot

In the trace plot Figure 16, there are seven different variables including $intercept$. Each plot shows the sampled values of a parameter at each iteration in the MCMC simulation. The four different lines in each plot represent four different chains.

Ideally, the lines in a trace plot should be centered around a constant value (the mean of the posterior distribution), and their fluctuations should resemble random noise around this value. If the chains appear to be stationary and well-mixed (i.e., the lines are overlapping with each other), this is usually a good indication that the MCMC simulation has converged.

Autocorrelation plot

Figure 17 shows the autocorrelation at different lags for each variable used in model 3. The autocorrelation should decrease toward zero as the lag increases, indicating that the samples from the MCMC simulation are not correlated with each other. If high autocorrelation is observed, it suggests that the MCMC chains are mixing slowly, which can affect the accuracy of the posterior estimates.

Parameter estimation

Figure 18 shows seven different parameters. Each parameter has an estimated effect (exp()) and a range of uncertainty represented by a horizontal line.

4.5 Covariate standardization for model with specific prior

Before proceeding with Model 4, the numerical columns in the dataset were standardized. Standardization is a preprocessing step in which the data is transformed to have a mean of 0 and a standard deviation of 1. This is done using the scale function in R, which centers and/or scales the columns of a numeric matrix.

In this case, a new dataframe was created, and each selected numerical covariate in the dataset was standardized. The standardized values are calculated as the difference between the original value and the mean, divided by the standard deviation. This process ensures that all the numerical variables are on the same scale, which can be beneficial for algorithms.

After standardizing the data, priors were specified for each covariate in the model. In Bayesian statistics, the prior represents our belief about the parameter before seeing the data. Here, set the prior for each covariate to be a normal distribution with a mean of 0 and a standard deviation of 1 since we have standardized covariates. Also, this is a common choice of prior when don't have strong beliefs about the parameter values before seeing the data.

```
> summary(M4)
 Family: bernoulli
  Links: mu = logit
Formula: depressed ~ sex + Age + Married + Number_children + education_
    level + total_members
   Data: data_standardized (Number of observations: 476)
```

```
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept         -0.21      0.33    -0.85     0.42 1.00     5072     3238
sexmale            0.22      0.33    -0.43     0.87 1.00     4365     2695
Age                0.30      0.11     0.08     0.52 1.00     3795     2763
Marriedyes         0.01      0.24    -0.46     0.48 1.00     4180     3214
Number_children   -0.24      0.15    -0.52     0.06 1.00     3047     3067
education_level    0.03      0.11    -0.18     0.25 1.00     3886     2732
total_members      0.24      0.14    -0.04     0.53 1.00     2917     3171

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Based on the summary of Model 4, Age appears to have a statistically significant effect on depression, as its 95% credible interval does not include zero. This suggests that as age increases, the log odds of depression also increase, holding all other variables constant.

The other covariates (sex, $Married$, $Number\_children$, $education\_level$, $total\_members$) may not have a statistically significant effect on depression, as their 95% credible intervals include zero. This means that we cannot rule out the possibility that these covariates have no effect on depression.

The trace plot and autocorrelation plot in Figure 20 and 21 respectively indicate the convergence and stability of the MCMC chains for the model parameters. The trace plot shows that the chains mix well, suggesting convergence to similar values across chains. The autocorrelation plot shows that the autocorrelation for each parameter is relatively low, indicating that the samples are independent and the chains have mixed effectively.

Finally, Figure 22 illustrates the parameter estimation of the M4 model, providing a visual representation of the estimated effects of each covariate on depression, along with their uncertainty intervals. This figure can help visualize the magnitude and direction of the effects, as well as identify which covariates have statistically significant effects on depression.

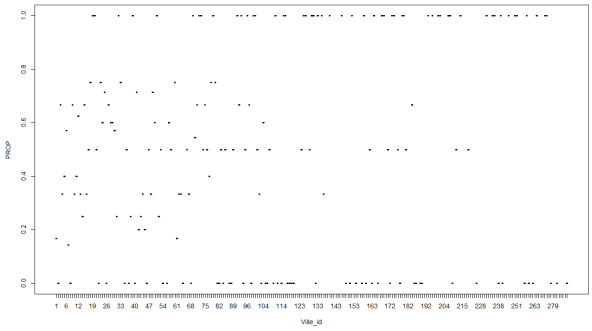

Figure 4: The proportions of people being depressed across Village ID

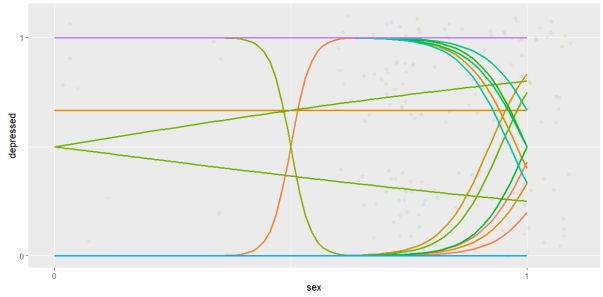

Figure 5: Relationship between SEX and $depressed$ by $Ville\_id$

4.6 Multilevel model with specific prior

The Bayesian binary logistic regression model introduced earlier is limited to modeling the effects of individual predictors, while the Bayesian binomial logistic regression is constrained to modeling the effects of village-level predictors. To incorporate both individual-level and village-level predictors, multilevel models, specifically Bayesian multilevel binary logistic regression, can be employed.

In addition to the motivation mentioned above, there are more reasons to employ multilevel models. For instance, as the data are clustered within villages, individuals from the same village are likely more similar to each other than those from other villages. Furthermore, even the relationship between the outcome and the predictor variables may vary across villages. Using multilevel models can appropriately address these issues.

Figure 4 illustrates the proportions of people being depressed across villages, providing insights into the variability of depression rates at the village level. Figure 5 shows the relation-

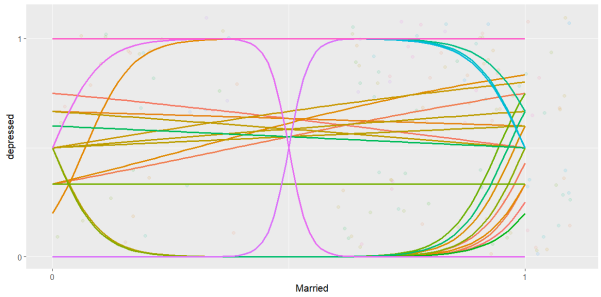

Figure 6: Relationship between$Married$ and $depressed$ by $Ville\_id$

ship between sex and $depression$ by $Ville\_id$, suggesting potential differences in this relationship across different villages. Similarly, Figure 6 depicts the relationship between marital status ($Married$) and $depression$ by $Ville\_id$, indicating potential variations in this relationship across villages as well. These plots highlight the importance of considering village-level effects when analyzing the relationship between predictors and depression.

In Model 5, used standardized data and set Gaussian priors for the covariates with a mean of 0 and a standard deviation of 1. This is a common choice when don't have strong prior beliefs about the parameter values.

For the village-level effects ($ville\_id$), set a Cauchy prior with a location parameter (average) of 0 and a scale parameter (standard deviation) of 50. The Cauchy distribution is a heavy-tailed distribution and is often used in multilevel modeling as it is more robust to outliers and can accommodate a wider range of values.

```
> summary(M5)
Family: bernoulli
  Links: mu = logit
Formula: depressed ~ 1 + sex + Age + Married + Number_children + education_
    level + total_members + (1 | Ville_id)
   Data: data_standardized (Number of observations: 476)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Group-Level Effects:
~Ville_id (Number of levels: 180)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.58      0.24     0.10     1.03 1.00      666      767
```

```
Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          -0.23      0.37    -0.95     0.48 1.00     6089     2898
sexmale             0.25      0.36    -0.45     0.94 1.00     6311     3060
Age                 0.33      0.12     0.09     0.59 1.00     3913     2925
Marriedyes          0.01      0.26    -0.48     0.53 1.00     4806     2991
Number_children    -0.25      0.15    -0.55     0.05 1.00     3507     3212
education_level     0.03      0.12    -0.20     0.27 1.00     4975     3243
total_members       0.26      0.15    -0.03     0.56 1.00     3488     3005

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Based on the summary of Model 5, the Bayesian multilevel binary logistic regression model was fitted to the standardized data. Here is a brief interpretation of the results:

- **Group-Level Effects:** The standard deviation of the random intercepts for $Ville\_id$ is estimated to be 0.58. This indicates that there is substantial variability in the baseline log odds of depression across different villages.
- **Population-Level Effects:** These are the fixed effects estimates for the predictors in the model. For example, the log odds of depression increase by 0.33 for each additional year of age, holding all other variables constant and averaging over the random effects of villages. However, the 95% credible intervals for most of the predictors include zero, suggesting that these effects may not be statistically significant at the population level.

From the trace plot in Figure 24, it appears that the MCMC chains for most variables have converged well, as indicated by the trace plots showing similar values across the four chains and the autocorrelation plot in Figure 25 is relatively flat and close to zero. This suggests that the sampling process was reliable and stable for these variables.

However, the $ville\_id$ variable shows higher levels of positive autocorrelation as the lag increases. This could be due to the nature of the data, which doesn't have a natural order or ranking, leading to a different distribution in the autocorrelation plot. High autocorrelation could also indicate slower mixing and less efficient sampling, which might require more iterations to accurately estimate the posterior distribution.

5 Convergence Diagnostics

5.1 Posterior predictive checks

To do the posterior predictive checking, the $posterior\_predict()$ function from the $brms$ package has been used to generate predicted probabilities for each observation in the test set, which is a way to evaluate the fit of the Bayesian model.

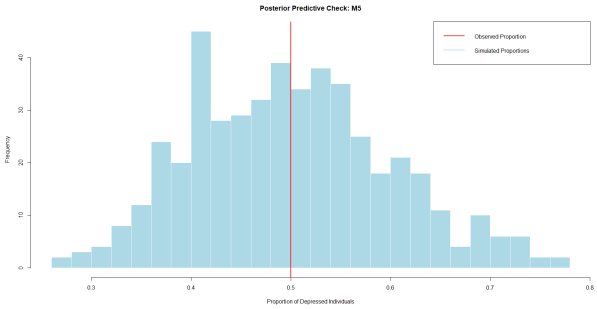

Figure 7: Posterior predictive check for M5

After generating posterior predictive samples, which are essentially simulated data generated from a fitted model, then calculate the proportion of 'depressed' individuals in both observed data and simulated data and compare these proportions visually using a histogram. The red line represents the observed proportion of 'depressed' individuals in the data, and the histogram represents the distribution of the proportion of 'depressed' individuals in simulated data.

If the model is a good fit for the data, the observed proportion should fall within the distribution of the simulated proportions. If it falls outside, it might indicate that the model is not capturing some aspects of the data well. Figure 7 shows the posterior predictive check for Model 5. The other model figures are included in the appendix as Figures 11, 15, 19, and 23.

5.2 Model comparison

To comprehensively compare all models based on their predictive performance, complexity, and fit to the data, analysis of various factors, employing methodologies like leave-one-out cross-validation (LOO-CV) offers a robust framework for this assessment, as demonstrated in Appendix D.

- **Predictive Performance:** This is measured by the expected log pointwise predictive density ($elpd\_loo$). A higher $elpd\_loo$ indicates better predictive performance. Based on this criterion, Model 5 has the highest $elpd\_loo$ and thus the best predictive performance among the four models.
- **Model Complexity:** This is measured by the effective number of parameters (p_loo). A higher p_loo indicates a more complex model. Based on this criterion, Model 5 is the most complex model among the four.
- **Model Fit:** This is measured by the LOO information criterion ($looic$), which is -2 times the $elpd\_loo$. A lower $looic$ indicates a better fit to the data. Based on this criterion, Model 5 has the lowest $looic$ and thus the best fit among the four models.

Based on the LOO-CV results, Model 5 appears to have the best predictive performance and fit to the data among the four models, despite being the most complex model. However, it's important to consider the context and the trade-off between model complexity and predictive performance when choosing the final model. For example, if Model 5 is much more complex than the other models but only slightly better in terms of predictive performance, it might be preferable to choose a simpler model to avoid overfitting.

5.3 Predictive performance

When assessing predictive performance using the Area Under the Curve (AUC), which is a performance measurement for classification problems at various threshold settings. In this context, the AUC results reveal distinct disparities among the models. Model 5 and Model 4 demonstrate the highest AUC of **0.84**, indicative of its superior discriminative power and overall effectiveness in classifying instances compared to the other models. Conversely, Models 3 and 1 exhibit slightly lower AUC values of 0.68 and 0.60 respectively, suggesting relatively inferior performance.

6 Limitations and Potential Improvements

While the models developed in this study have provided valuable insights into the factors associated with depression, several limitations should be noted:

- **Model Assumptions:** The models assume a linear relationship between the log odds of depression and the predictors. However, the true relationships may be non-linear or involve interactions between predictors.
- **Variable Selection:** The models only include a limited set of predictors. There may be other important predictors not included in the dataset, such as genetic factors, lifestyle habits, or environmental exposures.
- **Model Complexity:** The most predictive model (Model 5) is also the most complex model, which could lead to overfitting. While cross-validation methods were used to guard against overfitting, it's still a potential concern.

Potential improvements could include:

- **Incorporating Non-linear Relationships and Interactions:** More complex models could be developed to incorporate non-linear relationships and interactions between predictors.
- **Including Additional Predictors:** If additional data is available, other potentially important predictors could be included in the models.
- **Regularization:** Techniques such as ridge or lasso regression could be used to prevent overfitting in complex models.

7 Conclusion

The study has developed several Bayesian logistic regression models to predict the likelihood of depression based on individual-level and village-level predictors. Model 5, a multilevel model incorporating both individual-level and village-level predictors, was found to have the best predictive performance based on LOO-CV and AUC. However, it's also the most complex model, highlighting the trade-off between model complexity and predictive performance. Future work could explore more complex models and additional predictors to improve predictive performance. Despite the limitations, the models developed in this study provide valuable insights into the factors associated with depression and could be useful tools for identifying individuals at risk and informing interventions to reduce depression.

Bibliography

Hal S. Stern David B. Dunson Aki Vehtari Donald B. Rubin Andrew Gelman, John B. Carlin. *Bayesian Data Analysis*. Taylor Francis Ltd, 3rd edition, 2013. ISBN 978-1439840955.

Diego Babativa. depression. 2019. URL `https://www.kaggle.com/datasets/diegobabativa/depression`.

Robert Tibshirani Hastie, Trevor and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2016. ISBN 978-0387848570.

Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press/Taylor & Francis Group, 2nd edition, 2019. ISBN 9781482253443.

Zindi. Busara mental health prediction challenge. 2018. URL `https://zindi.africa/competitions/busara-mental-health-prediction-challenge`.

Appendix

A Additional figures

A.1 Additional figures for Model: M1

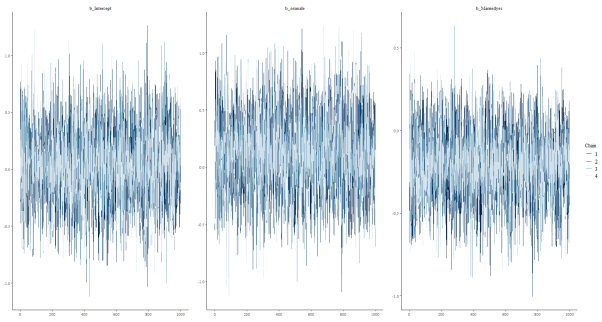

Figure 8: Trace plot for Model 1

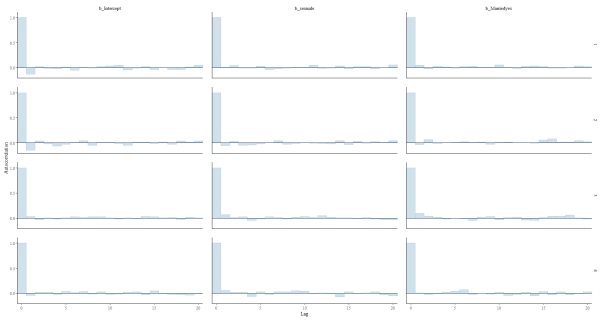

Figure 9: Autocorrelation plot for Model 1

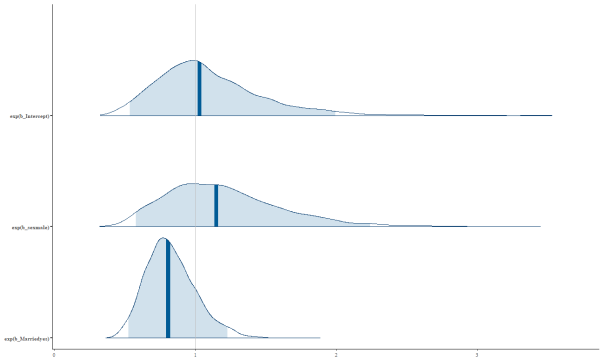

Figure 10: Parameter estimation plot for Model 1

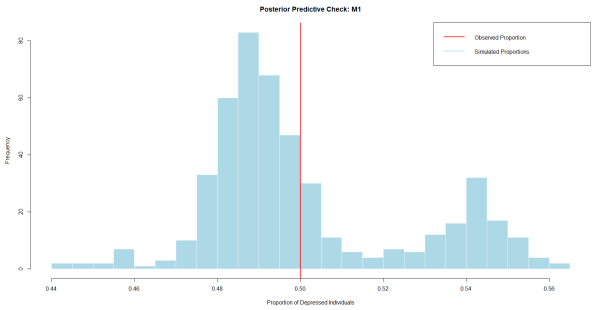

Figure 11: Posterior predictive check for M1

A.2 Additional figures for Model: M2

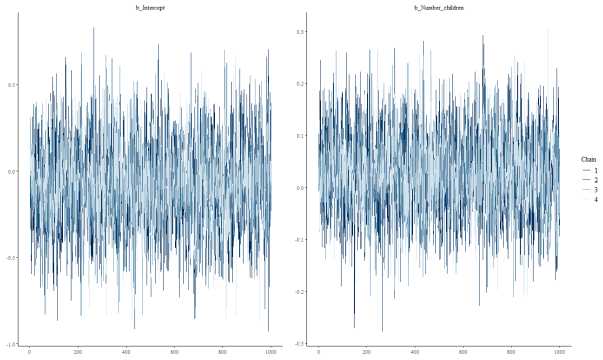

Figure 12: Trace plot for Model 2

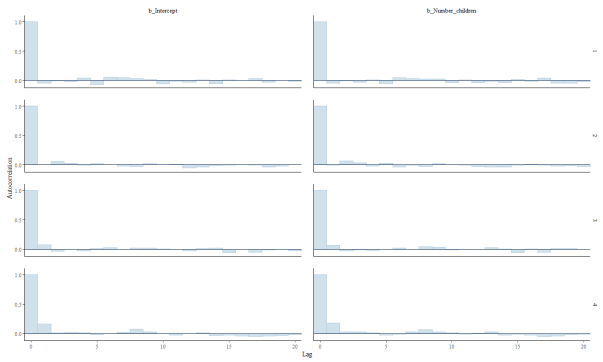

Figure 13: Autocorrelation plot for Model 2

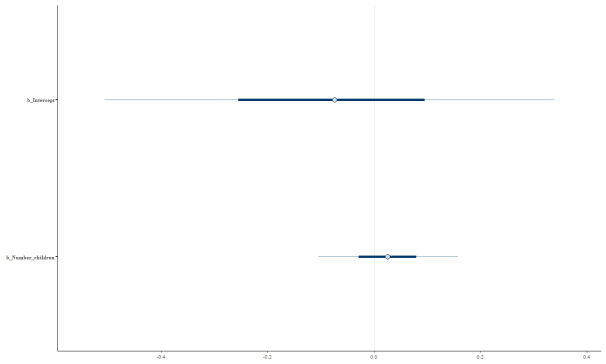

Figure 14: Parameter estimation for Model 2

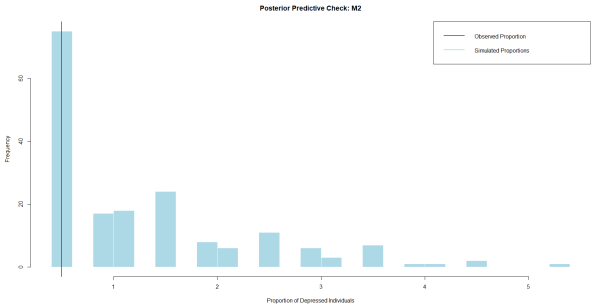

Figure 15: Posterior predictive check for M2

A.3 Additional figures for Model: M3

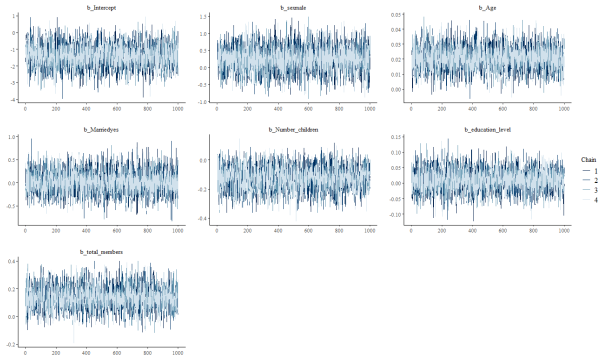

Figure 16: Trace plot for Model 3

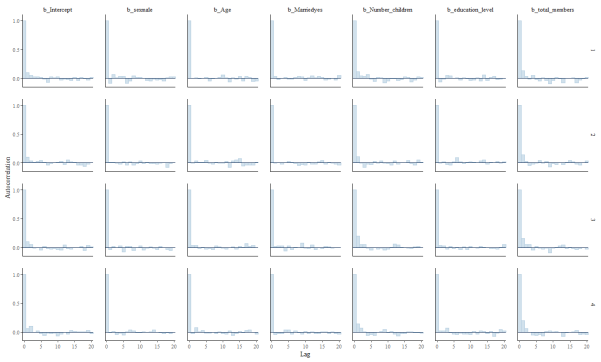

Figure 17: Autocorrelation plot for Model 3

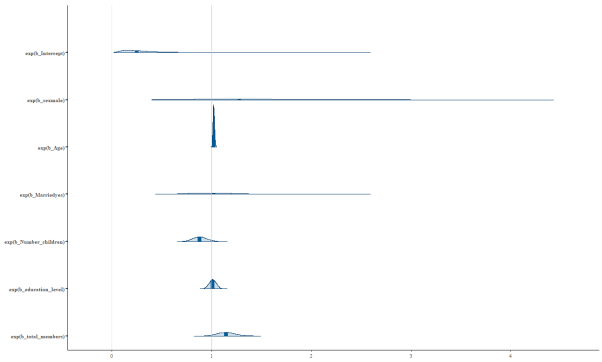

Figure 18: Parameter estimation for Model 3

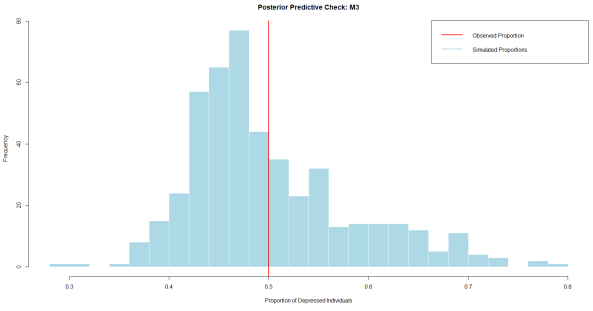

Figure 19: Posterior predictive check for M3

A.4 Additional figures for Model: M4

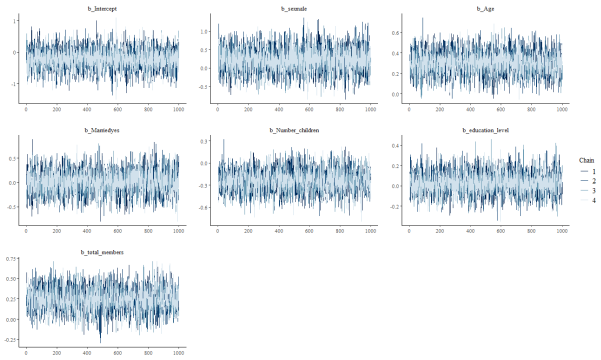

Figure 20: Trace plot for Model 4

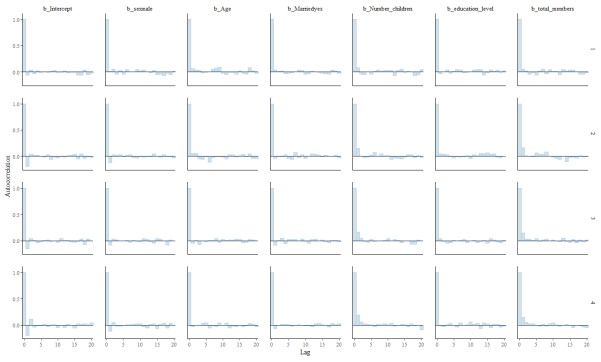

Figure 21: Autocorrelation plot for Model 4

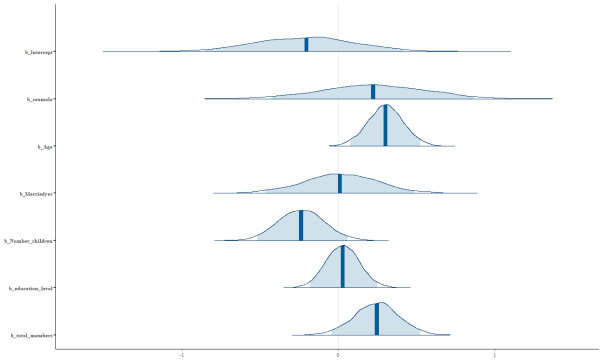

Figure 22: Parameter estimation for Model 4

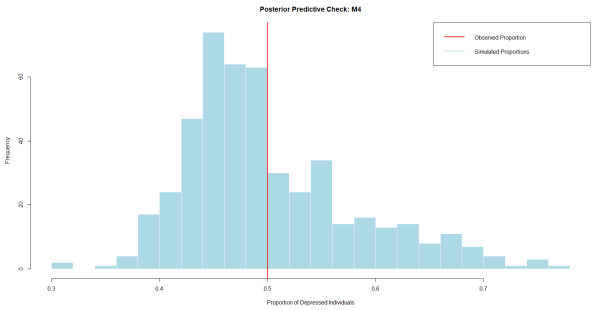

Figure 23: Posterior predictive check for M4

A.5 Additional figures for Model: M5

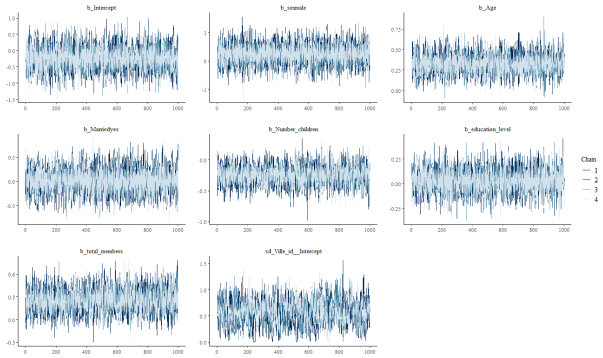

Figure 24: Trace plot for Model 5

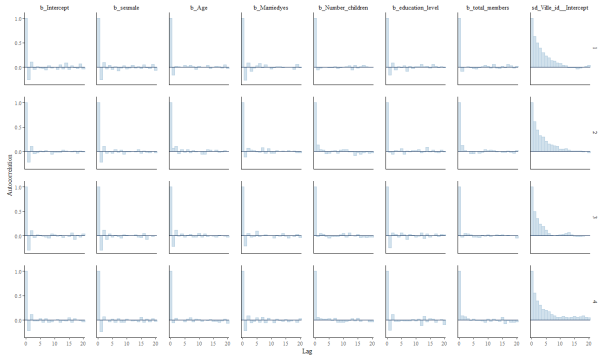

Figure 25: Autocorrelation plot for Model 5

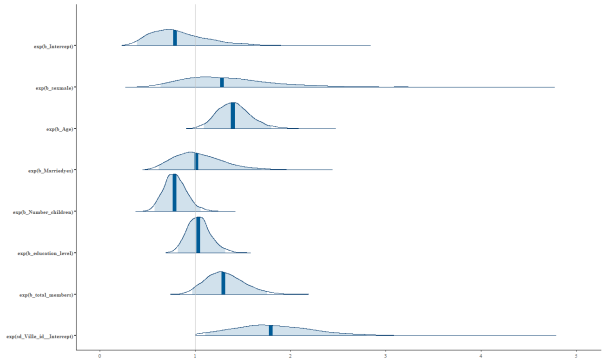

Figure 26: Parameter estimation for Model 5

B Additional calculation

The predicted probabilities for experiencing depression for different groups based on sex and marital status are as follows:

- For males who are not married:

$$p = \frac{1}{1 + e^{-(0.04+0.13\times1-0.21\times0)}} = \frac{1}{1 + e^{-0.071}} \approx 0.58$$

- For females who are married:

$$p = \frac{1}{1 + e^{-(0.04+0.13\times0-0.21\times1)}} = \frac{1}{1 + e^{0.071}} \approx 0.46$$

- For males who are married:

$$p = \frac{1}{1 + e^{-(0.04+0.13\times1-0.21\times1)}} = \frac{1}{1 + e^{-0.041}} \approx 0.49$$

C Additional tables

Table 1: Summary statistics: central tendencies and spread of the variables Age, $Number\_children$, $Education\_level$, and $Total\_members$.

Variable	**Mean**	**Median**	**Min**	**Max**	**Range**
Age	36.03	32	17	87	70
Number_children	2.975	3	0	11	11
Education_level	8.223	9	1	18	17
Total_members	5.071	5	1	12	11

Table 2: Cross-tabulation of sex and $Married$.

	No	**Yes**
Female	25	16
Male	102	333

D Additional output

```
> loo(M1, M3, M4, M5)
Output of model 'M1':

Computed from 4000 by 476 log-likelihood matrix

         Estimate  SE
elpd_loo   -332.5 1.0
p_loo         3.1 0.2
looic       665.0 2.1
------
Monte Carlo SE of elpd_loo is 0.0.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.

Output of model 'M3':

Computed from 4000 by 476 log-likelihood matrix

         Estimate  SE
elpd_loo   -331.0 3.7
p_loo         7.4 0.4
looic       661.9 7.3
------
Monte Carlo SE of elpd_loo is 0.0.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.

Output of model 'M4':

Computed from 4000 by 476 log-likelihood matrix

         Estimate  SE
elpd_loo   -330.6 3.6
p_loo         7.0 0.3
looic       661.2 7.2
------
Monte Carlo SE of elpd_loo is 0.0.
```

```
All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.

Output of model 'M5':

Computed from 4000 by 476 log-likelihood matrix

         Estimate  SE
elpd_loo   -329.8 4.4
p_loo        35.3 0.8
looic       659.7 8.8
------
Monte Carlo SE of elpd_loo is 0.1.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.

Model comparisons:
   elpd_diff se_diff
M5  0.0       0.0
M4 -0.8       2.4
M3 -1.1       2.4
M1 -2.7       4.3
```