technische universität
dortmund

Bayesian Data Analysis on Suicides in India

— Project Report —
Advanced Bayesian Data Analysis

Author: Aakash Goyal (229975)
Jaykumar Savani (230443)

Supervisors: Prof. Dr. Paul Bürkner
Prof. Dr. Katja Ickstadt

March 17, 2024

TU Dortmund University

Contents

1 Introduction

Every year, more than $800,000$ people succumb to suicide, which is the fourth-leading cause of death globally. Suicide, or intentional self-killing, is a major public health concerns in all countries. Every suicide is a tragedy that affects families, communities and entire countries and has long-lasting effects on the people left behind. In 2021, a total of $164,033$ people took their own lives in India, and the number is increasing each year. A number of factors, such as health issues, drugs, financial and career problems, family conflicts, etc. can serve as potential triggers for suicidal thoughts and feelings.

The main objectives of the project are to investigate the relationship between different age groups, genders, causes, and the number of suicides in India. For this study, the dataset is sourced from an online data repository, data.world, a widely recognized platform for cataloging and governing enterprise data. The sourced dataset contains $1,048,576$ suicide records from 2002 to 2012, but a small sample is extracted to conduct the study. To achieve the aim of the project, initially, a small sample of the dataset is chosen by selecting the latest year, i.e., 2012 and only 8 states. Secondly, the pre-processing of the dataset is done, during which duplicates are eliminated, special characters in the *type* variable are removed, and *type* is mapped to the new cause category. Thirdly, an overview of the final dataset is presented. Fourthly, the dataset is modeled using three models, such as the Poisson model, the negative binomial model, and the Zero-inflated Negative Binomial model, and the best-performing model is chosen. Further, a brief discussion about the prior choices and prior sensitivity analysis of the chosen model is done, and the results are interpreted.

In Section 2, a detailed overview of the data is presented. In Section 3, statistical methods, including their mathematical formulation, are explained, which are used for the analysis. In Section 4, the introduced statistical methods are applied to the sourced dataset, and the results of the tasks are interpreted. In Section 5, all the findings of the analysis are summarized, and an outlook on potential improvements for further analysis is discussed. In the end, Section 6 offers insights into our learning journey and the skills we acquired during this project.

2 Data

In this section, the specifics of data preprocessing and description of dataset are presented. is provided.

2.1 Data Preprocessing

The dataset (Ilangovan, 2015) is sourced from data.world, and it contains suicide information for India from 2002 to 2012. The entire dataset consists of $1,048,576$ records, but a small fragment is considered for the study. India is a large country, and it has 28 states and 8 union territories. For the purpose of this study, randomly 8 states, i.e., Haryana, Uttar Pradesh, Andhra Pradesh, Tamil Nadu, Maharashtra, Gujarat, West Bengal, and

Assam, and the latest year 2012 are selected. Different age groups are available in the dataset, such as 0–14, $15-29, 30-44, 45-59, 60+$ and $0-100+$. All the records with the age group $0-100+$ are dropped as it contains the total sum of suicides for all the other age groups, which is redundant. Further, special characters in the *type* column are removed as there are 69 different types of methods used for committing suicide and these methods are remapped to 8 different categories, such as Drugs, Social, Family, Health, Educational, Financial & Career, Freewill and Others. Later, the dataset is aggregated based on $Cause$ to get the final processed dataset. The final dataset contains 640 records, and there are no missing values.

2.2 Description of Data

In Section 2.1, the final dataset contains 6 variables. Each record in the dataset represents the following information: $State$ (categorical) represents the state in India, $Year$ (discrete) represents the year, $Gender$ (categorical) represents the gender, and Age_group (categorical) represents different age groups of the people. There are 5 age groups present, i.e., $0-14, 15-29, 30-44, 45-60$, and $60+$. $Cause$ (categorical) represents the method used for committing suicide. There are 8 different categories, such as Drugs, Social, Family, Health, Educational, Financial & Career, Freewill and Others. Lastly, $Total$ (numeric) represents the total count of suicides and is the response variable.

3 Statistical Methods

In this section, several statistical methods are presented, which are later used for analyzing the data according to the project requirements.

3.1 Poisson Model

The Poisson model is used for the count data. If a single observation y is distributed according to the Poisson distribution with rate λ, then the probability distribution of y is represented as

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where $y = 0, 1, 2, \ldots$. Similarly, the vector form of y, which has independent and identically distributed observations, is represented as $y = (y_1, y_2, y_3, \ldots, y_n)$, the likelihood is represented as

$$L(y|\lambda) = \prod_{i=1}^{n} \frac{1}{y_i!} \lambda^{y_i} e^{-\lambda}$$

where n is the independent draw and λ is the Poisson parameter. The mean and variance are same and equal to λ (Gelman et al., 2022).

3.2 Negative Binomial Model

The negative binomial model is used to deal with the count data. This distribution model handles the issue of over-dispersion in the data. Over-dispersion means that the variability of the data is greater than its mean. The model is characterized by two parameters: μ (the mean) and k (the exponent), which control the degree of over-dispersion. The variance of the distribution is represented as $\mu + \frac{\mu^2}{k}$, and the probability function is represented as

$$f_{(\mu,k)}(y) = \frac{\Gamma(k+y)}{\Gamma(k)y!}\frac{k^k\mu^y}{(k+\mu)^{(y+k)}}, y = 0, 1, 2, 3, \ldots,$$

where $\Gamma(.)$ signifies the gamma function (Hwang et al., 2016).

3.3 Zero-inflated Negative Binomial Model (ZINB)

A Zero-inflated negative binomial model deals with two problems in count data, i.e., over-dispersion and excess zeroes. Let Y_i be the response variable, where $i = 1, 2, 3, ...$, the ZINB distribution is represented as:

$$P(Y_i = y_i) = \begin{cases} p_i + (1-p_i)(\frac{k}{\mu_i+k})^k, & y_i = 0, \\ (1-p_i)\frac{\Gamma(k+y_i)}{\Gamma(y_i+1)\Gamma(k)}(\frac{\mu_i}{\mu_i+k})^{y_i}(\frac{k}{\mu_i+k})^k, & y_i = 1, 2, 3, \ldots \end{cases}$$

where k represents the dispersion parameter and is greater than 0, $\Gamma(.)$ represents the gamma function. The mean and variance of the ZINB distribution is given as $E(Y_i) = (1-pi)\mu_i$, and $Var(Y_i) = (1-p_i)\mu_i(1+\mu_i k^{-1} + p_i\mu_i)$ (Garay et al., 2011).

3.4 Multilevel Modeling

The dependent variable y is estimated by estimating the parameters of θ_q of the response distribution D, also known as model family. It is represented as $y_i = D(\theta_{1i}, \theta_{2i}, \theta_{3i}, \ldots)$, dependent on i^{th} term. Each parameter θ_q can be modeled using its dedicated predictor η_q, by applying the inverse link transformation f_q, which results in $\theta_{qi} = f_q(\eta_{qi})$. These types of models are known as distributional models. The predictor η can generally be expressed as

$$\eta = X\beta + Zu + \textstyle\sum_{w=1}^{W} s_w(x_w),$$

where β represents the population-level effect, u represents the group-level effect, and X, Z are the design matrices respectively. The terms $s_w(x_w)$ represents the optional smooth functions of indeterminate form. They rely on covariates x_k, which are fitted via splines. The data is formed by the terms y (response variable), X, Z(design matrices) and x_w(covariates) and the terms β(population coefficient), u(group-level coefficient) and s_w(smooth functions) represents the model parameters that are to be estimated (Bürkner, 2018).

3.5 Leave-one Out Cross Validation (LOO)

To measure the predictive accuracy of the model or for model comparison, Leave-one out cross validation technique (LOO) is used. LOO uses Pareto smooth importance sampling (PSIS) for its computation as it provides credible estimates when a Pareto distribution is fitted to the upper tail of the distribution.

Let a dataset $y_1, y_2, y_3, \ldots, y_n$, where it is independent when conditioned on a set of parameters θ. It can be expressed as $p(y|\theta) = \Pi_{i=1}^{n} p)p(y_i|\theta)$. This approach also includes latent variable models which are represented by $p(y_i|f_i, \theta)$, with f_i being the latent variables. Additionally, let $p(\theta)$ be a prior distribution, which yields a posterior distribution $p(\theta|y)$ and a posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y_i}|\theta)p(\theta|y)d\theta$. A metric, called expected log pointwise predictive density, of predictive accuracy for n number of observations considered at a time is determined. This metric is designed to ensure comparability with the provided dataset and interpret the differences in the scale of vital parameters. The metric is represented as

$$elpd = \sum_{i=1}^{n} \int p_t(\tilde{y_i}) log\, p(\tilde{y_i}|y) d\tilde{y_i},$$

where $p_y(\tilde{y_i})$ represents the true data-generating process of y_i and it is approximated using cross validation. As an alternative, a log score for assessing the predictive density can be used. It is called as log pointwise predictive density (lpd) and can be represented as

$$lpd = \sum_{i=1}^{n} log \int p(y_i|\theta)p(\theta|y)d\theta.$$

Based on this, the Bayesian LOO estimate for out-of-sample predictive fit is represented as

$$elpd_{loo} = \sum_{i=1}^{n} logp(y_i|y_{-i}),$$

where leave-one-out predictive density $p(y_i|y_{-i})$ is represented as $intp(y_i|\theta)p(\theta|y_{-i})d\theta$.

3.6 Choice of Priors

In this section, different types of priors are discussed.

Non-Informative Priors: These priors are typically known as flat priors that lack prior knowledge about the parameters. These priors have no or minimal impact on the posterior distribution and are default priors of the Bayesian model. They can be used as a standard of comparison with the informative priors.

Weakly informative Priors: Weakly Informative priors express partial information about the parameters. These priors incorporates some degree of prior knowledge into the Bayesian model and do not strongly influence the posterior distribution.

Informative Prior: Informative priors are the priors that incorporates prior information about the parameters. These priors incorporates prior information into the Bayesian model and has a strong effect on posterior distribution.

3.7 Sensitivity Analysis

A sensitivity analysis enables the researcher to compare the outcomes of the final model with the reference priors against the outcomes of the model when different priors are applied. The process of the sensitivity analysis is given as

- The research selects a set of priors for the purpose of model estimation. The priors can be default priors or user-specified priors (based on previous knowledge).
- Ensuring convergence for all the parameters, the model estimation is carried out.
- Different sets of contending priors are selected by the researcher and applied to the model. The objective is to evaluate the stability of the model against different priors.
- Outcomes attained for the priors selected in Step 3 are compared with the reference priors outcomes using statistical methods and findings are interpreted.

4 Statistical Analysis

In this section, the above-mentioned statistical methods are applied to the dataset for analysis, and the results of the tasks are interpreted. For all calculations and visualizations, the software R (R Core Team, 2023a) and R packages rstan (Stan Development Team, 2024), brms (Bürkner, 2017), and bayesplot (Gabry and Mahr, 2022), loo (Vehtari et al., 2023), stats (R Core Team, 2023b), dplyr (Wickham et al., 2023a), posterior (Bürkner et al., 2023), ggplot (Wickham, 2016), magrittr (Bache and Wickham, 2022), car (Fox and Weisberg, 2019), scales Wickham et al. (2023b), and AER (Kleiber and Zeileis, 2008) are used.

4.1 Exploratory data analysis

In this subsection, the focus lies on understanding the variables present in the final dataset. As mentioned in Section 2, the final dataset consists of 6 variables, i.e. *State*, *Gender*, $Year$, *Age_group*, $Cause$ and the response variable $Total$. In Figure 1, a histogram of the response variable $Total$ is depicted. It represents the frequency of the total number of suicides; since it is count data, the variable cannot be negative. The mean value of $Total$ is 327.58, which signifies that the average number of people who committed suicide among 8 different states in India is around 327 in 2012. The histogram also shows that the frequency of the number of suicides is high in the range of $0 - 250$, and it tapers off as the number of suicides increases.

After a uni-variate analysis on the response variable, a bi-variate analysis of the independent variables (*Age_group*, *Gender*, and $Cause$) with the response variable is performed. Figure 2(a) depicts a bar plot between *Gender* and $Total$. It is observed that males have a higher number of suicides as compared to females.

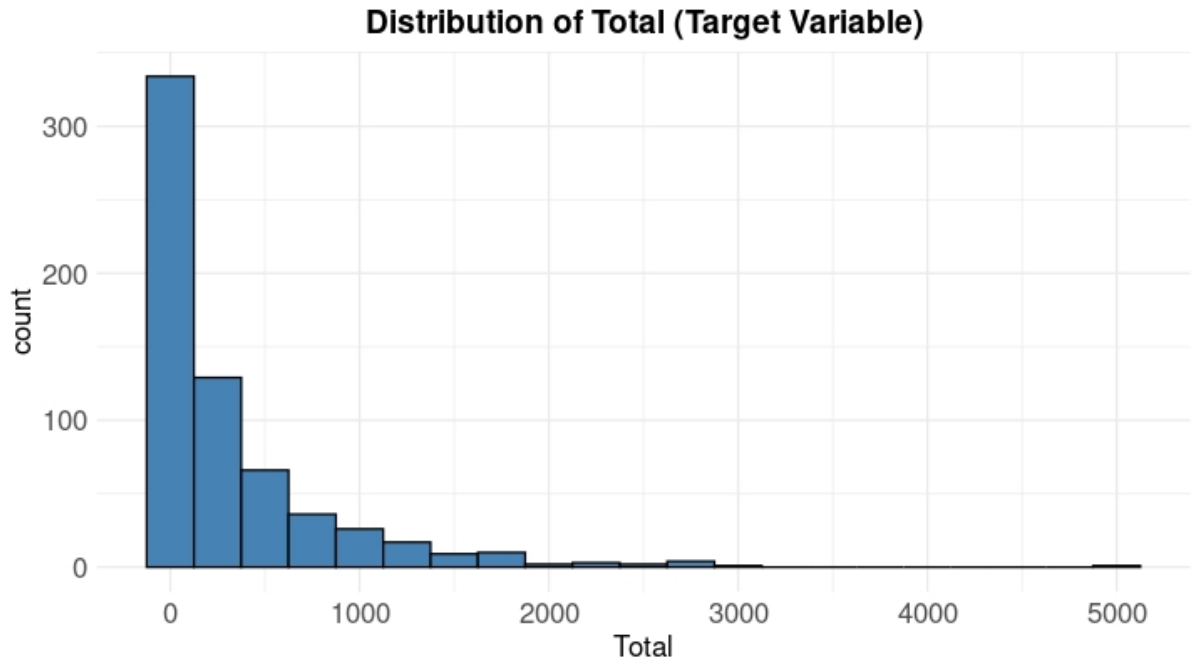
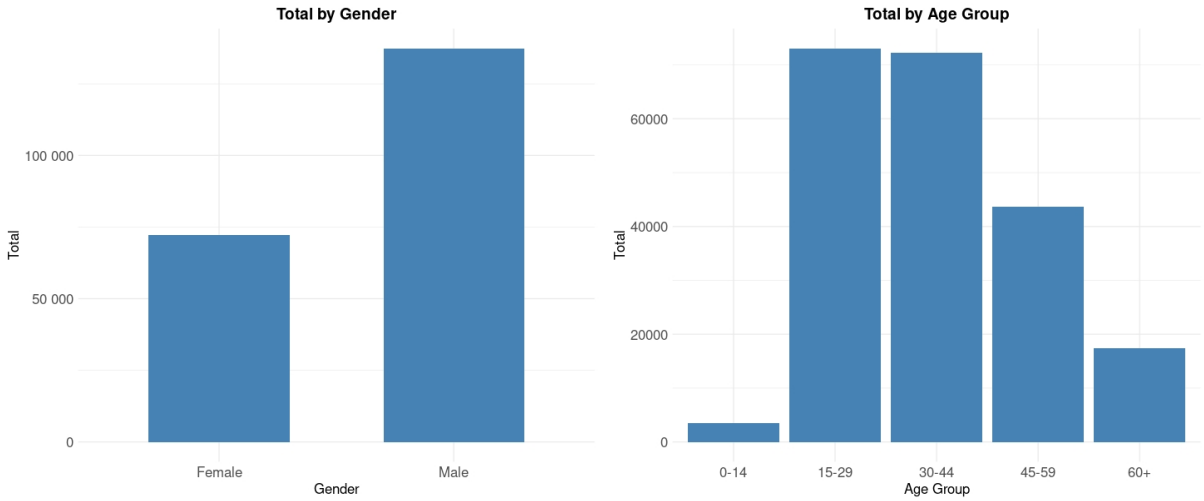


Figure 1: Distribution of Response variable

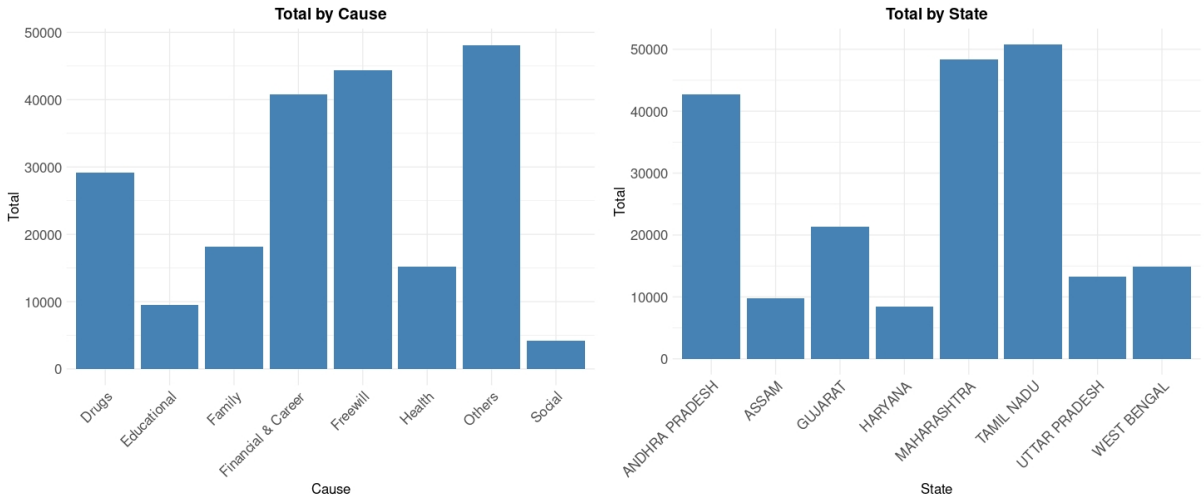


(a) (b)

Figure 2: Bar plot for (a) Total by Age Group (b) Total by Gender

Figure 2(b) depicts the bar plot between *Age_group* and *Total*. It indicates that the suicides are predominantly higher among people in the age groups of 15–29 and $30-44$. Meanwhile, the age group 0 to 14 has the least number of suicides. Figure 3(a) represents a bar plot and examines the relationship between *Cause* and *Total*. It is identified that the most common cause of suicide is Freewill, followed by Finance & Career, and Drugs consumption. In the last, Figure 3(b) represents a bar plot between *State* and *Total*, and

it is observed that Andhra Pradesh, Maharashtra, and Tamil Nadu are the states with the highest number of suicides. Conversely, Haryana has the lowest number of suicides in 2012.



(a) (b)

Figure 3: Bar plot for (a) Total by Cause (b) Total by State

4.2 Model interpretation

In this section, three different models are selected, i.e., Poisson, negative binomial and Zero-inflated negative binomial (ZINB). At the beginning of each model's output, details about the family, formula, total number of iterations, and chains are provided. Subsequently, group-level effects are presented individually for each grouping variable. They showcase the standard deviations and relationships between them. In the end, population-level effects, followed by family-specific parameters (if available), are displayed.

4.2.1 Poisson Model

The Poisson model is fitted using 4 chains, each running for 3000 iterations. The first 1500 iterations are warmup iterations that calibrate the sampler, leading to a total of 6000 posterior samples. Additionally, a control parameter is used, which helps in tackling the divergent transitions. This results in preventing the biasing of the posterior samples. The Poisson model is given as

```
Poisson_ ← brm(Total ~ Cause + Gender + (1 + Age_group|State),
               family=poisson(),
               data = subset_df1,
               control = list(max_treedepth = 15, adapt_delta = 0.99),
               core=26)

Family: poisson
```

```
Links: mu = log
Formula: Total ~ Cause + Gender + (1 + Age_group | State)
Data: subset_df1 (Number of observations: 640)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1; total
    post-warmup draws = 6000
```

Here, the Poisson model is conditioning the response variable $Total$ (number of suicides) on $Cause$, $Gender$ and the mixed effect modeling $(1 + Age\_group|State)$. Here, 1 represents the intercept, while Age_group represents a categorical variable indicating different age groups. The $|State$ term indicates that the effect of age group is allowed to vary by the levels of the $State$ variable. This hierarchical structure acknowledges that the impact of age group on the outcome may differ depending on the state in which the observation was made. Due to the inclusion of this random effect, the model output now presents standard deviation and correlation terms within the group-level effects. Table 1 represents the group-level effects and is given as

Table 1: Poisson model: group-level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	1.10	0.47	0.50	2.35	1.00	1768	2765
sd(Age_group15M29)	2.12	0.35	1.55	2.93	1.00	2086	3232
sd(Age_group30M44)	2.11	0.35	1.55	2.90	1.00	2146	3132
sd(Age_group45M59)	1.78	0.32	1.29	2.55	1.00	2101	3269
sd(Age_group60P)	1.29	0.33	0.85	2.10	1.00	2826	3543
cor(Intercept,Age_group15M29)	-0.33	0.39	-0.88	0.55	1.01	756	1159
cor(Intercept,Age_group30M44)	-0.28	0.40	-0.87	0.59	1.01	766	1186
cor(Age_group15M29,Age_group30M44)	0.95	0.06	0.78	1.00	1.00	1955	2731
cor(Intercept,Age_group45M59)	-0.23	0.39	-0.83	0.60	1.01	834	1258
cor(Age_group15M29,Age_group45M59)	0.90	0.11	0.59	0.99	1.00	2250	3348
cor(Age_group30M44,Age_group45M59)	0.95	0.07	0.75	1.00	1.00	2275	3873
cor(Intercept,Age_group60P)	0.15	0.36	-0.54	0.79	1.00	1072	1979
cor(Age_group15M29,Age_group60P)	0.66	0.16	0.28	0.89	1.00	4164	4459
cor(Age_group30M44,Age_group60P)	0.59	0.19	0.13	0.86	1.00	4184	4599
cor(Age_group45M59,Age_group60P)	0.58	0.21	0.08	0.87	1.00	4211	4556

$sd(Intercept)$ provides information about the standard deviation (sd) of the intercepts across the different states. It suggests the degree of variability in the baseline values of the outcome variable among the different states. In this case, the estimate of this standard deviation is 1.10, with a credible interval (CI) ranging from 0.50 to 2.35. $sd(Age\_group15M29)$ indicates the standard deviation of the random effect associated with the age group category $Age\_group15M29$ (representing individuals aged 15 to 29) across the different states. The estimate of this standard deviation is 2.12, with a credible interval ranging from 1.55 to 2.93. $cor(Intercept, Age\_group15M29)$ describes the correlation between the random intercept and the random effect associated with the age group category $Age\_group15M29$ (individuals aged 15 to 29). This correlation terms can be interpreted like Pearson's correlation coeffcient. The correlation coefficient is -0.33. However, the 95% credible interval of $cor(Intercept, Age\_group15M29)$ is quite wide $[-0.88, 0.55]$ and contains zero. This suggests that slope/intercept correlation has no convincing evidence of association with the data. Similarly, the 95% credible intervals of

the variables, such as $cor(Intercept, Age\_group30M44)$, $cor(Intercept, Age\_group45M59)$, and $cor(Intercept, Age\_group60P)$ contain zero and do not show any convincing evidence for slope/intercept correlation with this data.

Table 2: Poisson model: population-level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.45	0.88	1.94	5.48	1.01	715	1095
CauseEducational	-1.12	0.01	-1.14	-1.10	1.00	7165	4865
CauseFamily	-0.47	0.01	-0.49	-0.45	1.00	6599	5261
CauseFinancial&Career	0.33	0.01	0.32	0.35	1.00	5060	4897
CauseFreewill	0.42	0.01	0.40	0.43	1.00	5476	4870
CauseHealth	-0.65	0.01	-0.67	-0.63	1.00	6797	5096
CauseOthers	0.50	0.01	0.49	0.52	1.00	5649	4978
CauseSocial	-1.93	0.02	-1.96	-1.90	1.00	9023	5073
GenderMale	0.64	0.00	0.63	0.65	1.00	11753	3991

Following that, Table 2 describes the estimated population-level effects of the predictor variables on the outcome variable at the population level. $Intercept$ is the estimated intercept or baseline value of the outcome variable when all other predictors are zero. In this case, the estimated intercept is 3.45, with a 95% credible interval ranging from 1.94 to 5.48. This means that when all other predictor variables are zero, the expected outcome variable will range between 1.94 and 5.48. The other predictors, such as $CauseEducational$, $CauseFamily$, $CauseFinancial\&Career$, $CauseFreewill$, $CauseHealth$, $CauseOthers$, $CauseSocial$, and $GenderMale$ show the estimated effects on the outcome variable $Total$. These effects are interpreted as the change in the outcome variable associated with a one-unit increase in the respective predictor.

4.2.2 Negative binomial Model

The Poisson model does not accommodate over-dispersion which can be seen as a limitation in the real world scenario, hence an alternative, i.e., a negative binomial model, is considered. The negative binomial model handles the issue of over-dispersion in the dataset. The negative binomial model is fitted using 4 chains, each with 3000 iterations. The initial 1500 iterations are warmup iterations that calibrate the sampler, leading to a total of 6000 posterior samples. The model is given as

```
NegBino_ ← brm(Total ~ Cause + Gender + (1 + Age_group|State),
                    family = negbinomial(),
                    data = subset_df1,
                    control = list(max_treedepth = 15,
                                   adapt_delta = 0.99),
                    core = 26,
                    iter = 3000)

Family: negbinomial
Links: mu = log; shape = identity
```

```
Formula: Total ~ Cause + Gender + (1 + Age_group | State)
Data: subset_df1 (Number of observations: 640)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1; total
   post-warmup draws = 6000

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     0.89      0.05     0.79     0.99 1.00     8768     4314
```

Now, the family argument is changed to "negbinomial". The shape parameter in the model summary estimates the dispersion in the data. The higher value of the shape parameter indicates a low over-dispersion in the data. A shape parameter close to 1 suggests symmetry. Here, the shape parameter estimate is 0.89, which shows that the data is over-dispersed. Also, here the model coefficient can be interpreted in the same manner as the Poisson model, as the standard link function used in the negative binomial is the same as Poisson's log link function.

Table 3: Negative binomial model: group level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	1.29	0.51	0.62	2.56	1.00	2285	2994
sd(Age_group15M29)	2.31	0.44	1.64	3.37	1.00	1847	2793
sd(Age_group30M44)	2.08	0.40	1.48	2.99	1.00	1905	2846
sd(Age_group45M59)	1.62	0.33	1.12	2.42	1.00	2166	3111
sd(Age_group60P)	0.96	0.31	0.53	1.72	1.00	3441	3697
cor(Intercept,Age_group15M29)	-0.19	0.38	-0.82	0.61	1.00	691	1067
cor(Intercept,Age_group30M44)	-0.21	0.38	-0.84	0.57	1.00	720	1281
cor(Age_group15M29,Age_group30M44)	0.88	0.13	0.48	0.99	1.00	2325	3214
cor(Intercept,Age_group45M59)	-0.16	0.38	-0.81	0.61	1.00	724	1398
cor(Age_group15M29,Age_group45M59)	0.84	0.16	0.41	0.99	1.00	2915	3812
cor(Age_group30M44,Age_group45M59)	0.87	0.13	0.49	0.99	1.00	3485	4640
cor(Intercept,Age_group60P)	0.24	0.36	-0.50	0.85	1.00	938	1929
cor(Age_group15M29,Age_group60P)	0.54	0.23	0.00	0.88	1.00	4733	4789
cor(Age_group30M44,Age_group60P)	0.53	0.23	-0.00	0.87	1.00	4654	4830
cor(Age_group45M59,Age_group60P)	0.55	0.23	0.01	0.89	1.00	4769	4893

In Table 3, the group-level effects of the negative binomial model are presented. The 95% credible intervals of $cor(Intercept, Age\_group15M29)$, $cor(Intercept, Age\_group30M44)$, $cor(Intercept, Age\_group45M59)$, $cor(Intercept, Age\_group60P)$, and $cor(Age\_group15M29, Age\_grou$ are quite wide and contain zero, and hence do not have any convincing evidence of their correlation with the data.

In Table 3, the group-level effects of the negative binomial model are presented. $cor(Intercept, Age\_group15M29)$, $cor(Intercept, Age\_group30M44)$, $cor(Intercept, Age\_group45M59)$, $cor(Intercept, Age\_group60P)$, and $cor(Age\_group15M29, Age\_group60P)$ do not have any convincing evidence of their correlation with the data as their 95% credible intervals are quite wide and contain zero. Similary, in Table 4, the population-level effects of the negative binomial model are displayed and it is observed that all the predictors are significant except $CauseFinancial\&Career$.

Table 4: Negative binomial model: population level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.61	0.94	1.95	5.76	1.00	687	1185
CauseEducational	-1.13	0.19	-1.50	-0.77	1.00	3332	3529
CauseFamily	-0.96	0.18	-1.32	-0.61	1.00	3454	4343
CauseFinancial&Career	0.01	0.18	-0.34	0.35	1.00	3425	4245
CauseFreewill	0.51	0.17	0.17	0.86	1.00	3992	3929
CauseHealth	-0.87	0.18	-1.24	-0.52	1.00	3742	4354
CauseOthers	0.74	0.17	0.39	1.07	1.00	3322	4254
CauseSocial	-2.13	0.19	-2.51	-1.76	1.00	3647	3578
GenderMale	0.58	0.09	0.41	0.75	1.00	8996	4023

4.2.3 Zero-inflated negative binomial Model (ZINB)

As the dataset exhibits excess zeroes (10% of the dataset contains zeroes) and overdispersion, a Zero-inflated negative binomial model is considered to tackle the situation. In general, the model assumes that the additional zeroes observed originate from a different process that is distinct from the count process. The ZINB model is fitted using 4 chains, each with 3000 iterations. The initial 1500 iterations are warmup iterations that calibrate the sampler, leading to a total of 6000 posterior samples.

```
ZINB ← brm(Total ~ Cause + Gender + (1 + Age_group|State),
            family=zero_inflated_negbinomial(),
            data = subset_df1,
            control = list(max_treedepth = 20,
                           adapt_delta = 0.99),
            core=26,
            iter=3000)
```

```
Family: zero_inflated_negbinomial
Links: mu = log; shape = identity; zi = identity
Formula: Total ~ Cause + Gender + (1 + Age_group | State)
Data: subset_df1 (Number of observations: 640)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup draws = 6000

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     1.68      0.11     1.48     1.90 1.00     7915     4249
zi        0.09      0.01     0.06     0.11 1.00     9103     3983
```

The ZINB model estimates two family-specific parameters: shape and zero-inflation (zi). The shape parameter is estimated at 1.68 with an associated estimation error of 0.11 with a 95% credible interval ranging from 1.48 to 1.90. The zero-inflation parameter is estimated at 0.09 with an estimation error of 0.01 with a 95% credible interval between 0.06 and 0.11. It represents the probability of excess zeros in the data. All the parameters, including those from group-level effects and predictors from population-level effects, exhibit good convergence ($\hat{R}$ below 1.01) and sufficient effective sample sizes. Moreover,

convergence diagnostics for all the models are discussed later in the report in Section 4.4.

Table 5: Zero-inflated negative binomial model: group level effects

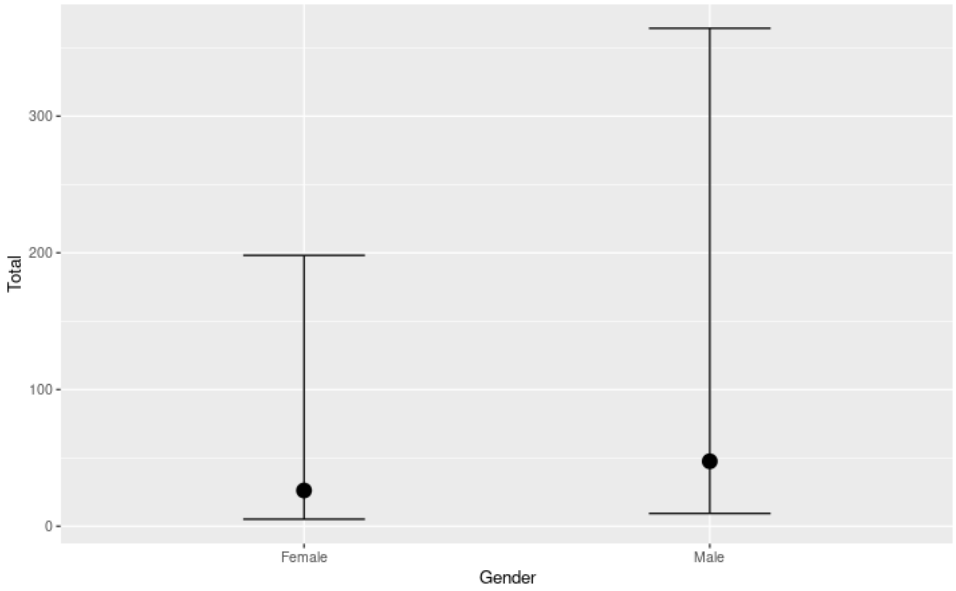
	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	1.29	0.50	0.64	2.53	1.00	1841	2797
sd(Age_group15M29)	2.24	0.41	1.59	3.19	1.00	1813	2893
sd(Age_group30M44)	2.03	0.37	1.46	2.90	1.00	2027	3154
sd(Age_group45M59)	1.60	0.31	1.12	2.32	1.00	2148	3448
sd(Age_group60P)	1.01	0.31	0.60	1.77	1.00	3193	3395
cor(Intercept,Age_group15M29)	-0.23	0.38	-0.83	0.55	1.00	500	1240
cor(Intercept,Age_group30M44)	-0.22	0.38	-0.84	0.56	1.00	480	1142
cor(Age_group15M29,Age_group30M44)	0.91	0.11	0.60	0.99	1.00	2146	3865
cor(Intercept,Age_group45M59)	-0.16	0.39	-0.80	0.62	1.00	524	1182
cor(Age_group15M29,Age_group45M59)	0.86	0.14	0.46	0.99	1.00	2495	3863
cor(Age_group30M44,Age_group45M59)	0.89	0.12	0.56	0.99	1.00	3580	4074
cor(Intercept,Age_group60P)	0.22	0.36	-0.48	0.84	1.00	691	2346
cor(Age_group15M29,Age_group60P)	0.56	0.21	0.06	0.87	1.00	3958	4317
cor(Age_group30M44,Age_group60P)	0.53	0.22	0.02	0.86	1.00	4170	4526
cor(Age_group45M59,Age_group60P)	0.54	0.22	0.00	0.87	1.00	4211	4177

Table 6: Zero-inflated negative binomial model: population level effects

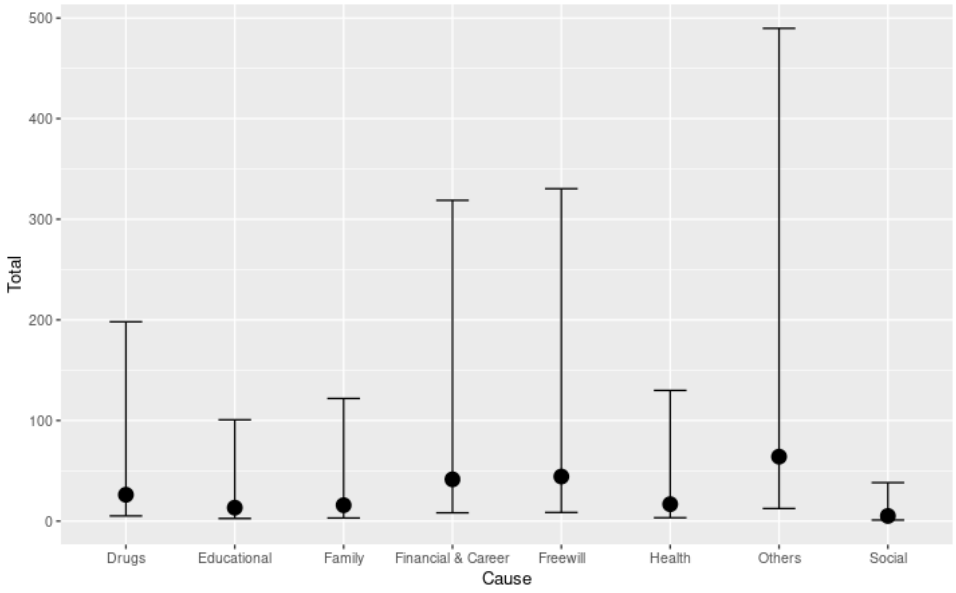
	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.42	0.92	1.73	5.39	1.01	470	1157
CauseEducational	-0.67	0.14	-0.95	-0.40	1.00	4127	4758
CauseFamily	-0.51	0.13	-0.77	-0.25	1.00	4039	4367
CauseFinancial&Career	0.46	0.13	0.21	0.72	1.00	3744	4423
CauseFreewill	0.53	0.13	0.27	0.78	1.00	3914	4385
CauseHealth	-0.44	0.13	-0.70	-0.18	1.00	4118	4397
CauseOthers	0.89	0.13	0.64	1.14	1.00	3963	4665
CauseSocial	-1.64	0.14	-1.91	-1.36	1.00	3827	3890
GenderMale	0.60	0.07	0.47	0.73	1.00	9252	4597

Further, the conditional effects of the ZINB model are observed and interpreted. In Figure 4(a) , the conditional effects plot for *Gender* displays two points along with their uncertainties. The x-axis represents the different categories available in *Gender*, and y-axis represents the $Total$ count. The point in the bars represents the median value of the predicted counts. It is observed from the plot that the median value of males is higher than that of females. Moreover, the uncertainty intervals for the male category are greater than those for the female category.

Similarly, in Figure 4(b), the conditional effects for $Cause$ represent eight points along with their uncertainty bands. Each point in the plot represents a different cause. Causes like 'Freewill' and 'Finance & Career' have a higher median value, while 'Educational' and 'Social' are on the lower end. The category 'Others' is not taken as the top cause, as it represents those records where no reason for suicide is mentioned.



(a)



(b)

Figure 4: Conditional effects plot of the ZINB Model

4.2.4 Posterior predictive checks

Posterior predictive checks provide a fundamental method for evaluating the quality of a model within a Bayesian framework. It is used to identify significant discrepancies between the real data and the data generated from the model. In cases of significant discrepancies, the model would not resonate with the true data-generating process. This provides a compelling reason to adjust the model. The function pp_check() is used to

perform posterior predictive checks for these models.

```
> pp_check(Poisson_, type = 'ecdf_overlay', ndraws = 100)

> pp_check(NegBino_, type = 'ecdf_overlay', ndraws = 100)

> pp_check(ZINB, type = 'ecdf_overlay', ndraws = 100)
```

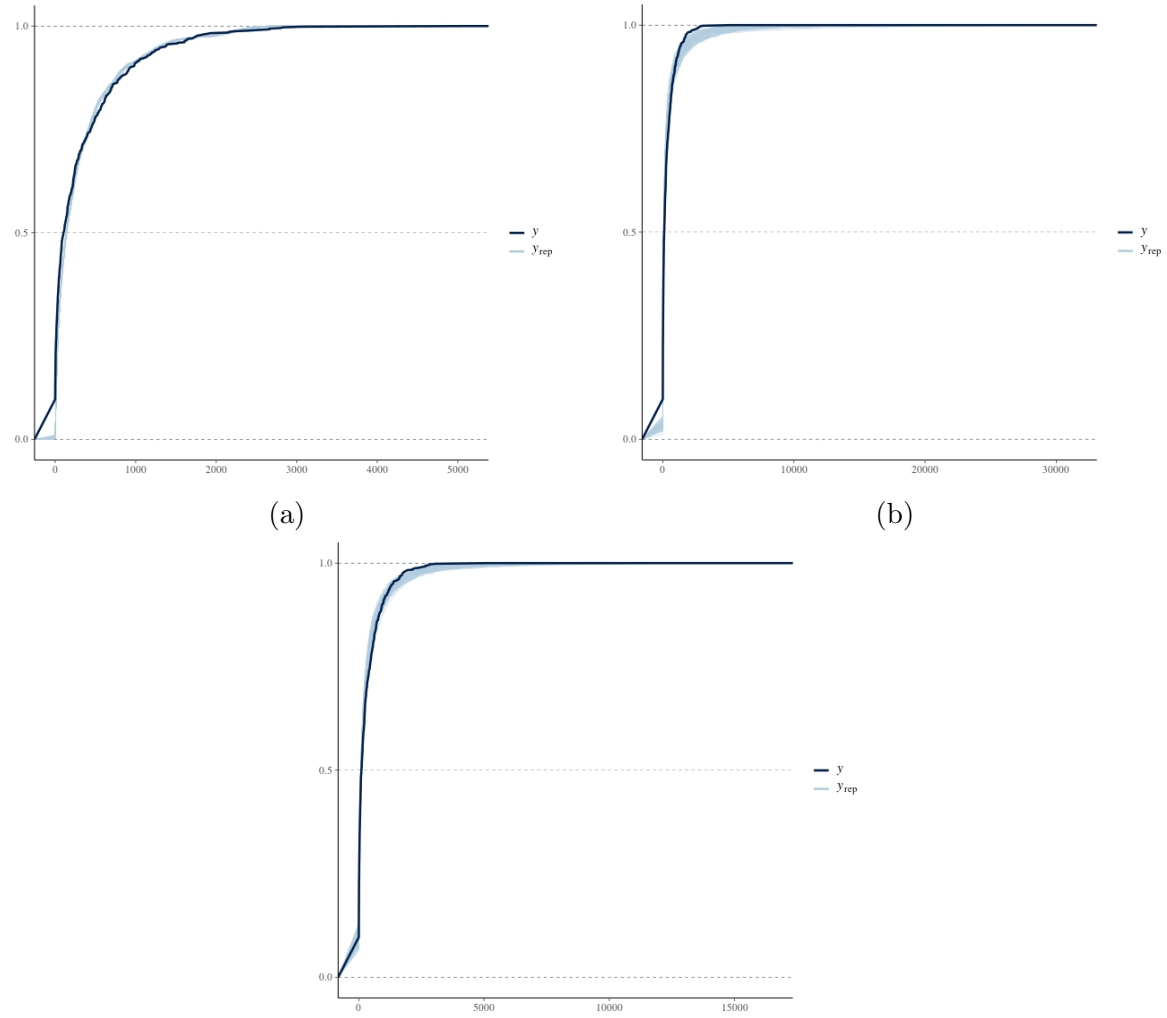


Figure 5: Posterior predictive checks (a) Poisson (b) Negative Binomial (c) ZINB

The output returned by the default *pp_check*() function may not be appropriate for the count or discrete data. Hence, the *ecdf_overlay* type is used to return an empirical cumulative distribution function (Winter and Bürkner, 2021). In Figure 4, the y-axis of the plot shows the proportion of values that are falling below the value on the x-axis. The original data is represented by the black line, and the simulated data is shown by the blue line. By visual inspection, the black line is almost within the simulated data in all three

models. It is observed that for all models, the black line falls in between the simulated data, but it also goes beyond it at some places. Hence, it is hard to compare the model just on visual inspection. In the next section, models are compared using leave-one out cross validation method.

4.2.5 Model Comparison of Poisson, Negative Binomials, and ZINB Models

The provided R code compares the models using the leave-one-out cross-validation (LOO) method, which evaluates the models based on their expected log predictive density (elpd).

```
#Model comparision
loo(Poisson_, NegBino_, ZINB, compare = TRUE)
```

Table 7: Model comparision: Poisson, Negative binomial, and ZINB

Model	elpd_diff	se_diff
ZINB	0.0	0.0
NegBino_	-76.8	15.6
Poisson_	-32643.4	2748.8

As shown in Table 7, the ZINB model is the reference model. The $elpd\_diff$ and se_diff for this model are both zero, as it serves as the baseline for comparison. Negative binomial has an $elpd\_diff$ of -76.8 compared to the reference model, which indicates that its expected log predictive density is lower by 76.8 units. The standard error associated with this difference is 15.6. Meanwhile, the Poisson model has $elpd\_diff$ of -32643.4 which is much larger compared to the reference model. This indicates a substantial decrease in the expected log predictive density compared to the reference model. The standard error associated with this difference is 2748.8. Hence, the ZINB model is the best-performing model among the ones compared, as it serves as the reference model with the highest expected log predictive density. On the other hand, the negative binomial model performs worse than the ZINB model, with a lower expected log predictive density. However, the difference is relatively small compared to the ZINB model. The Poisson model performs the worst among the three, with a significantly lower expected log predictive density compared to both the ZINB and negative binomial models. As the ZINB model is the best-performing model, a prior sensitivity analysis is performed in the next section with different prior settings and results are interpreted.

4.3 Prior sensitivity analysis

This section discusses the prior sensitivity checks between the two variations of ZINB models. Prior selections should exhibit moderate zero-inflation, over-dispersion, and uncertainty in the coefficient estimates, reflecting prior preconceptions about the parameters' values. By using the selected priors, the model integrates prior knowledge with observed data, resulting in more robust and interpretable inference.

Two different settings of the priors are selected, and the results are interpreted. In both the cases, the model uses a ZINB model framework to examine the association between the response variable (*Total*) and the predictors (*Cause*, *Gender*, and *Age_group*). A hierarchical structure is used to incorporate random effects, which account for fluctuations at different levels of the grouping variable (State). The *brm* function is used for model fitting, and it manages the model fitting process efficiently by performing 3000 iterations. The initial 1500 iterations are warmup iterations that calibrate the sampler, leading to a total of 6000 posterior samples. At first, a set of weakly informative priors are used, and the set is given as

4.3.1 Prior setup - I

```
#model summary:
priors_set1← c(
  set_prior("normal(0,50)", class = "b"),
  set_prior("normal(0,50)", class = "Intercept"),
  set_prior("cauchy(0,10)", class = "sd"),
  set_prior("beta(1,1)", class = "zi"),
  set_prior("gamma(1,0.01)", class = "shape")
)

# Fit the model with Weak priors
PSA1 ← brm(
  formula = Total ∼ Cause + Gender+(1 + Age_group|State),
  family = zero_inflated_negbinomial(),
  data = subset_df1,
  prior = priors_set1,
  control = list(max_treedepth = 20, adapt_delta = 0.99),
  core=26, iter=3000)
```

```
Family: zero_inflated_negbinomial
Links: mu = log; shape = identity; zi = identity
Formula: Total ∼ Cause + Gender + (1 + Age_group | State)
Data: subset_df1 (Number of observations: 640)
Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1; total
   post-warmup draws = 6000

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     1.69      0.11     1.48     1.91 1.00     7185     4619
zi        0.09      0.01     0.06     0.11 1.00     7867     3909
```

The result includes estimates and uncertainty intervals for the model's two family-specific parameters. The shape parameter, calculated at 1.69 with a 95% credible interval of 1.49 to 1.90 which describes the shape of the distribution, which indicates that the data is likely to over-dispersed. In contrast, the zero-inflation parameter is estimated at 0.09, with a narrower 95% credible interval of 0.06 to 0.11, indicating a moderate risk of extra zeros in the data. Both parameters show good convergence and sufficient effective sample sizes, which increases confidence in the estimations.

In Table 8, the estimates of the group-level effects are displayed. The terms, such as $(Intercept, Age\_group15M29)$, $(Intercept, Age\_group30M44)$, $(Intercept, Age\_group45M59)$, $(Intercept, Age\_group60P)$ and $(Age\_group15M29, Age\_group60P)$ do not have convincing evidence for correlations in the dataset as their 95% credible intervals contains zero, making them statistically insignificant. However, according to the population-level Table 9, all the predictors are statistically significant as they have convincing evidence at 95% credible intervals.

Table 8: Prior sensitivity analysis case-1: group level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	1.41	0.64	0.65	3.02	1.00	1557	2340
sd(Age_group15M29)	2.40	0.50	1.66	3.58	1.00	1067	1944
sd(Age_group30M44)	2.19	0.45	1.51	3.23	1.00	1204	2624
sd(Age_group45M59)	1.72	0.38	1.16	2.63	1.00	1265	2342
sd(Age_group60P)	1.07	0.34	0.62	1.91	1.00	2657	3319
cor(Intercept,Age_group15M29)	-0.20	0.41	-0.87	0.61	1.01	505	1209
cor(Intercept,Age_group30M44)	-0.19	0.41	-0.86	0.62	1.01	492	1297
cor(Age_group15M29,Age_group30M44)	0.91	0.11	0.57	0.99	1.00	2011	3039
cor(Intercept,Age_group45M59)	-0.13	0.41	-0.82	0.65	1.01	531	1300
cor(Age_group15M29,Age_group45M59)	0.85	0.15	0.42	0.99	1.00	2186	2944
cor(Age_group30M44,Age_group45M59)	0.89	0.13	0.53	0.99	1.00	2828	3526
cor(Intercept,Age_group60P)	0.23	0.38	-0.52	0.85	1.01	773	1982
cor(Age_group15M29,Age_group60P)	0.56	0.22	0.02	0.87	1.00	3624	3843
cor(Age_group30M44,Age_group60P)	0.52	0.23	-0.03	0.85	1.00	3658	4203
cor(Age_group45M59,Age_group60P)	0.54	0.24	-0.05	0.87	1.00	3522	3600

Table 9: Prior sensitivity analysis case-1: population level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.34	1.10	1.35	5.76	1.01	464	971
CauseEducational	-0.67	0.14	-0.94	-0.39	1.00	2977	4039
CauseFamily	-0.51	0.14	-0.77	-0.24	1.00	2639	3592
CauseFinancial&Career	0.47	0.13	0.20	0.73	1.00	2804	3531
CauseFreewill	0.53	0.13	0.28	0.78	1.00	3034	3674
CauseHealth	-0.44	0.13	-0.70	-0.18	1.00	3092	3590
CauseOthers	0.89	0.13	0.64	1.15	1.00	2898	3756
CauseSocial	-1.64	0.14	-1.92	-1.36	1.00	2994	3846
GenderMale	0.60	0.07	0.47	0.73	1.00	7867	5016

4.3.2 Prior setup - II

Here, the prior distribution parameters have been changed compared to the prior setup-I. This setup uses more informative priors for the intercept, slope and zero inflation probability, indicating a stronger shrinkage of the estimate towards the prior means.

```
#model summary:
priors_set2 ← c(
```

```
  set_prior("normal(0,5)", class = "b"),
  set_prior("normal(0,10)", class = "Intercept"),
  set_prior("cauchy(0,2)", class = "sd"),
  set_prior("beta(1,1)", class = "zi"),
  set_prior("gamma(2,0.5)", class = "shape")
)

# Fit the model with more informative priors
PSA2 ← brm(
  formula = Total ∼ Cause + Gender + (1 + Age_group|State),
  family = zero_inflated_negbinomial(),
  data = subset_df1,
  prior = priors_set2,
  control = list(max_treedepth = 20, adapt_delta = 0.99),
  core=26, iter=3000)

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     1.69      0.11     1.48     1.91 1.00     8273     4419
zi        0.09      0.01     0.06     0.11 1.00     10559     3868
```

The above summary result shows that the estimates are mostly identical with negligible difference in 95% credible intervals. Similarly, the group level effect Table 10 and population level effect Table 11 show negligible difference in estimates compared to the prior setup-I.

Table 10: Prior sensitivity analysis case-2: group level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(Intercept)	1.25	0.49	0.63	2.54	1.00	1992	2357
sd(Age_group15M29)	2.19	0.39	1.58	3.09	1.00	1927	2520
sd(Age_group30M44)	1.99	0.36	1.44	2.83	1.00	2151	3020
sd(Age_group45M59)	1.57	0.30	1.12	2.28	1.00	2469	3417
sd(Age_group60P)	0.98	0.28	0.59	1.72	1.00	3768	4471
cor(Intercept,Age_group15M29)	-0.19	0.40	-0.86	0.60	1.01	584	1270
cor(Intercept,Age_group30M44)	-0.17	0.40	-0.84	0.62	1.01	653	1463
cor(Age_group15M29,Age_group30M44)	0.91	0.11	0.59	0.99	1.00	2566	3603
cor(Intercept,Age_group45M59)	-0.11	0.40	-0.80	0.67	1.01	652	1653
cor(Age_group15M29,Age_group45M59)	0.86	0.14	0.49	0.99	1.00	3004	3326
cor(Age_group30M44,Age_group45M59)	0.90	0.12	0.56	0.99	1.00	3765	4049
cor(Intercept,Age_group60P)	0.26	0.36	-0.46	0.87	1.01	777	2051
cor(Age_group15M29,Age_group60P)	0.57	0.20	0.09	0.87	1.00	4563	4520
cor(Age_group30M44,Age_group60P)	0.53	0.22	0.03	0.86	1.00	4548	4007
cor(Age_group45M59,Age_group60P)	0.55	0.22	0.02	0.87	1.00	4613	4601

4.3.3 Model Comparision for different Prior Setting Models

The provided R code compares the models using the leave-one-out cross-validation (LOO) method, which evaluates the models based on their expected log predictive density (elpd).

```
#Model comparision
loo(ZINB, PSA1, PSA2, compare = TRUE)
```

Table 11: Prior sensitivity analysis case-2: population level effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	3.30	1.02	1.44	5.61	1.01	542	1041
CauseEducational	-0.66	0.14	-0.94	-0.39	1.00	4771	4921
CauseFamily	-0.50	0.13	-0.75	-0.25	1.00	4249	4838
CauseFinancial&Career	0.47	0.13	0.21	0.72	1.00	4343	4709
CauseFreewill	0.53	0.13	0.29	0.78	1.00	4635	4309
CauseHealth	-0.43	0.13	-0.69	-0.17	1.00	4445	4175
CauseOthers	0.89	0.12	0.66	1.14	1.00	4507	4659
CauseSocial	-1.63	0.14	-1.90	-1.37	1.00	4604	4758
GenderMale	0.60	0.07	0.47	0.73	1.00	11073	4106

Table 12: Model comparision: ZINB, ZINB Prior setting-1, ZINB Prior setting-2

Model	elpd_diff	se_diff
PSA2	0.0	0.0
PSA1	-0.5	0.3
ZINB	-0.5	0.2

In Section 4.2.5, it is determined that the ZINB model (with default prior) outperforms the other two models, i.e., the Poisson model and the negative binomial model. Now, the ZINB model (with default prior) is compared with the other ZINB models using different prior setups. Table 12 shows the model comparison between the ZINB model with default priors, prior setup-1, and prior setup-2. ZINB with prior setup-1 is the reference model here, with an elpd_diff and se_diff of 0.0. It indicates that it serves as the baseline for the comparison. Both ZINB with prior setup-2 model and ZINB (with default priors) have an $elpd\_diff$ of -0.5 compared to the reference model. It suggests that there is a slight decrease in the expected log predictive density compared to the ZINB model with more informative priors. The difference in standard error from the reference model associated with each model can be observed from Table 12. Based on the smaller elpd_diff value and the lower standard error, ZINB with prior setup-1 is the preferred model. It exhibits the highest expected log predictive density and the lowest uncertainty in comparison to the other two models.

4.4 Convergence Diagnostics

In Figures A3, the trace and density plots of the ZINB Model are displayed. A density plot depicts the posterior distribution, and a trace plot illustrates the sampled values for each chain and iteration. The trace plot provides a visual evaluation of the convergence and mixing of chains. All the trace plots of the ZINB model shows that chains have mixed well and converged. Moreover, the output of the trace and density plots is similar to the ZINB model, meaning the chains have mixed well for all the models used in this report.

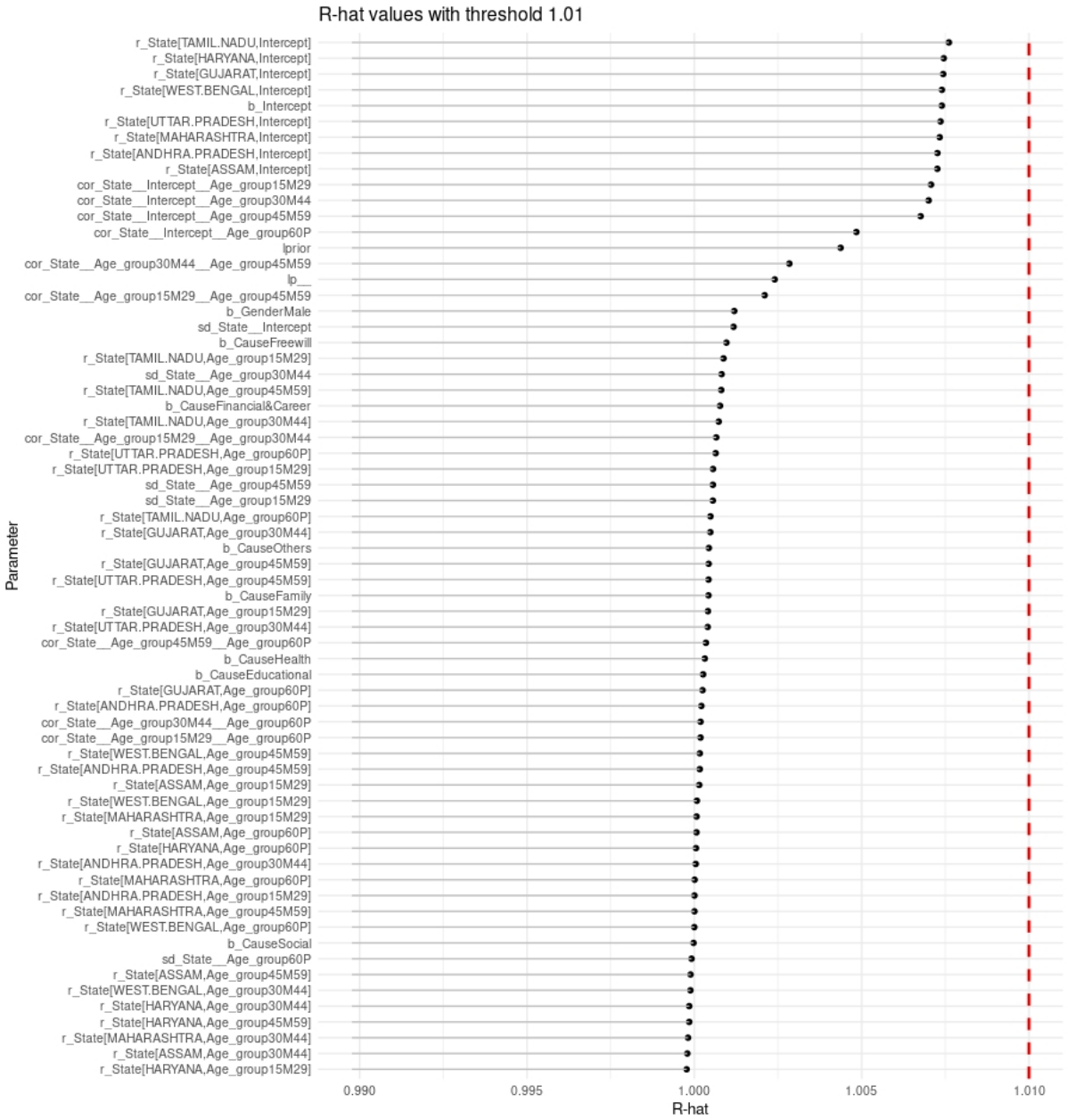


Figure 6: $\hat{R}$ value of Poisson Model

In Figures 6, A1, and A2, the $\hat{R}$ values of the Poisson model, negative binomial model, and ZINB model are depicted. The plot shows the variable names on the y-axis and the $\hat{R}$ values on the x-axis. Each horizontal line represents the $\hat{R}$ value for the particular parameter. The red dotted line represents the threshold line at 1.01. In this study, a tighter threshold of 1.01 is considered and it is suggested by (Vehtari et al., 2021). If the $\hat{R}$ value is greater than the threshold, it means that the convergence of chains has not taken place and there is a need to run more iterations or stronger priors should be

set. For all the parameters of all the models, $\hat{R}$ value is less than the threshold value of 1.01. This suggests that all the parameters of the all the models have converged to the target distribution.

Bulk_ESS (Bulk Effective Sample Size) measures the number of effective independent draws from the posterior distribution needed to produce the same standard error of the posterior mean as the dependent samples generated by MCMC (Markov Chain Monte Carlo) algorithm. At the time of fitting a model, four parallel chains are taken, so it is expected to have bulk_ESS greater than 400 (Vehtari et al., 2021). In all the Tables 1-10, it is observed that all the bulk_ESS values are greater than 400, indicating all the models have a good number of effective samples.

Tail_ESS (Tail Effective Sample Size) measures the effective sample size for the tails of the posterior distribution, i.e., 5% and 95% quantiles. Similar to bulk_ESS, higher values of the tail_ESS suggests a good number of effective samples for the tails. In all the Tables 1-10, the values of tail_ESS are decent, which shows that distributions' tails are well explored.

5 Summary

The main aim of the report was to perform Bayesian data analysis on the Suicides in India dataset, and to investigate the relationship between different age groups, genders and the number of suicides in India. The dataset, which was sourced from the data.world, consists of the suicides record from 2002 to 2012. For the purpose of analysis, the suicide data for 8 different states and the year 2012 were taken into account. To achieve the objectives of the project, initially, the exploratory data analysis of the dataset was done. It was observed that, firstly, the response variable represented the count data, which inherently cannot be negative. Secondly, males were more likely to commit suicide compared to females. Third, the prevalence of suicides was higher in the age groups of 15-29 and 30-44, while it was significantly lower in the age group of 0-14. Fourthly, the top three causes of suicides are Freewill, Finance & Career, and Drugs. In the last, Tamil Nadu, Maharashtra, and Andhra Pradesh are the leading states in terms of the number of suicides. Further, in order to model the count dataset, a Poisson model was selected and interpreted. The Poisson model does not accommodate over-dispersion, which can be limiting to real world data, hence an alternative, i.e., a negative binomial model, was applied and interpreted. It was observed that the dataset had a high over-dispersion as the value of the shape parameter was 0.89. Later, a Zero-inflated negative binomial model was also applied and interpreted as 10% of the dataset contains zero. This Zero-inflated model accommodates both over-dispersion and excess zeros in the dataset. In addition to this, the conditional effects plots for the ZINB model were depicted and explained. After fitting three models, posterior predictive checks were performed. But, it was hard to compare the models based on visual inspection, and then a leave-one-out cross-validation (LOO) comparison was performed. It was observed that the ZINB model was the best model, having the lowest ELPD score. Considering this, a prior sensitivity check was carried out on the ZINB model, and two different settings of priors were tested on the model. At first, a

set of weak priors were taken, and the model was fitted with these settings. Later, a set of more informative priors were taken for comparison. Three models, such as ZINB with default priors, the ZINB model with prior setup-1, and the ZINB model with prior setup-2, were compared. It was observed that the difference between the ELPD scores was too small, and the ZINB model with the prior setup-2 was the best-performing model. Later, the convergence diagnostics of all the models were discussed, and it was observed for all the models that all the chains mixed well, they all converged as their $\hat{R}$ was below the threshold value, and they had a good number of effective samples.

For further investigation, the focus of the study can be extended by expanding the geographical scope by including all the states and different years in the analysis. Moreover, hurdle models can be explored as an alternative to ZINB models, as they model zero observations and positive counts differently. Furthermore, more extensive prior sensitivity analysis can be conducted to understand the robustness of the models.

6 Self reflection

The project was a profound learning experience for us and enriched our knowledge of Bayesian approaches. During the project, we applied the techniques we learned in the course to a real-world dataset and tackled the problems that can occur in real-world settings, like those we encountered in this project, i.e., over-dispersion and excess zeros. Moreover, we learned how to apply various diagnostic methods and conduct prior sensitivity analysis to evaluate a good Bayesian model. This project not only provided us with an opportunity to delve into the world of Bayesian analysis but also solidified our understanding of the subject through practical implementation. The collaborative aspect of the project and the unwavering support from our supervisors were invaluable. Indeed, this course serves as a foundational framework for our future work in this field.

Bibliography

Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2022. URL https://CRAN.R-project.org/package=magrittr. R package version 2.0.3.

Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

Paul-Christian Bürkner. Advanced bayesian multilevel modeling with the r package brms. *R Journal*, 10:395–411, 07 2018. doi: 10.32614/RJ-2018-017.

Paul-Christian Bürkner, Jonah Gabry, Matthew Kay, and Aki Vehtari. posterior: Tools for working with posterior distributions, 2023. URL https://mc-stan.org/posterior/. R package version 1.5.0.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Jonah Gabry and Tristan Mahr. bayesplot: Plotting for bayesian models, 2022. URL https://mc-stan.org/bayesplot/. R package version 1.10.0.

Aldo M. Garay, Elizabeth M. Hashimoto, Edwin M.M. Ortega, and Víctor H. Lachos. On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318, 2011. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2010.09.019. URL https://www.sciencedirect.com/science/article/pii/S0167947310003622.

Andrew Gelman, John Carlin, Hal Stern, Donald Rubin, David Dunson, and Aki Vehtari. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2022. doi: http://www.stat.columbia.edu/~gelman/book/BDA3.pdf.

Wen-Han Hwang, Richard Huggins, and Jakub Stoklosa. Estimating negative binomial parameters from occurrence data with detection times. *Biometrical Journal*, 58 (6):1409–1427, 2016. doi: https://doi.org/10.1002/bimj.201500239. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201500239.

Rajanand Ilangovan. Suicides in india, 2015. URL https://data.world/rajanand/suicides-in-india.

Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. URL https://CRAN.R-project.org/package=AER. ISBN 978-0-387-77316-2.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023a. URL https://www.R-project.org/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023b. URL https://www.R-project.org/.

Stan Development Team. RStan: the R interface to Stan, 2024. URL https://mc-stan.org/. R package version 2.32.5.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667 – 718, 2021. doi: 10.1214/20-BA1221. URL https://doi.org/10.1214/20-BA1221.

Aki Vehtari, Jonah Gabry, Mans Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2023. URL https://mc-stan.org/loo/. R package version 2.6.0.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023a. URL https://CRAN.R-project.org/package=dplyr. R package version 1.1.4.

Hadley Wickham, Thomas Lin Pedersen, and Dana Seidel. *scales: Scale Functions for Visualization*, 2023b. URL https://CRAN.R-project.org/package=scales. R package version 1.3.0.

Bodo Winter and Paul-Christian Bürkner. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. *Language and Linguistics Compass*, 15 (11):e12439, 2021. doi: https://doi.org/10.1111/lnc3.12439. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12439.

Appendix

A Figures

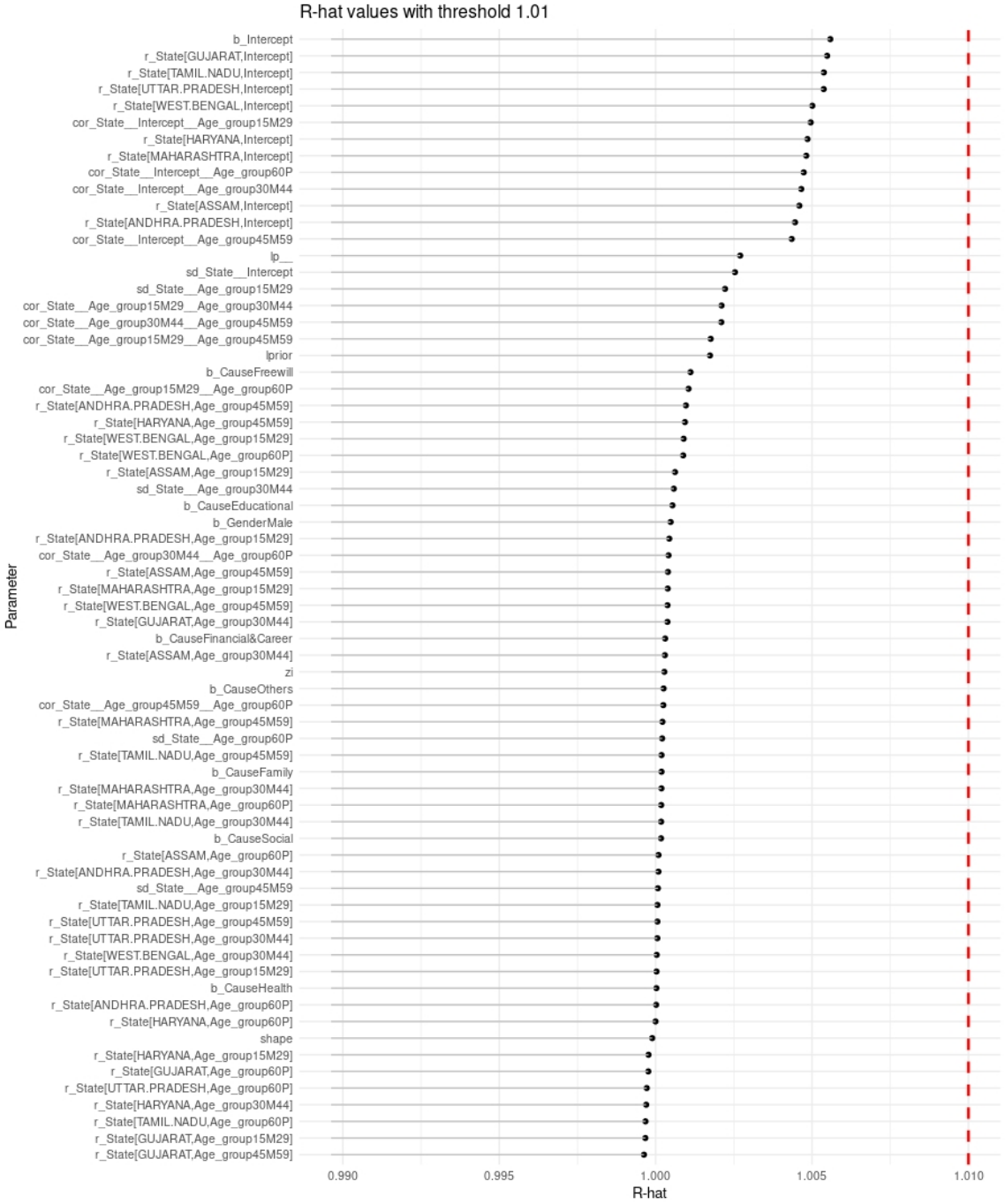
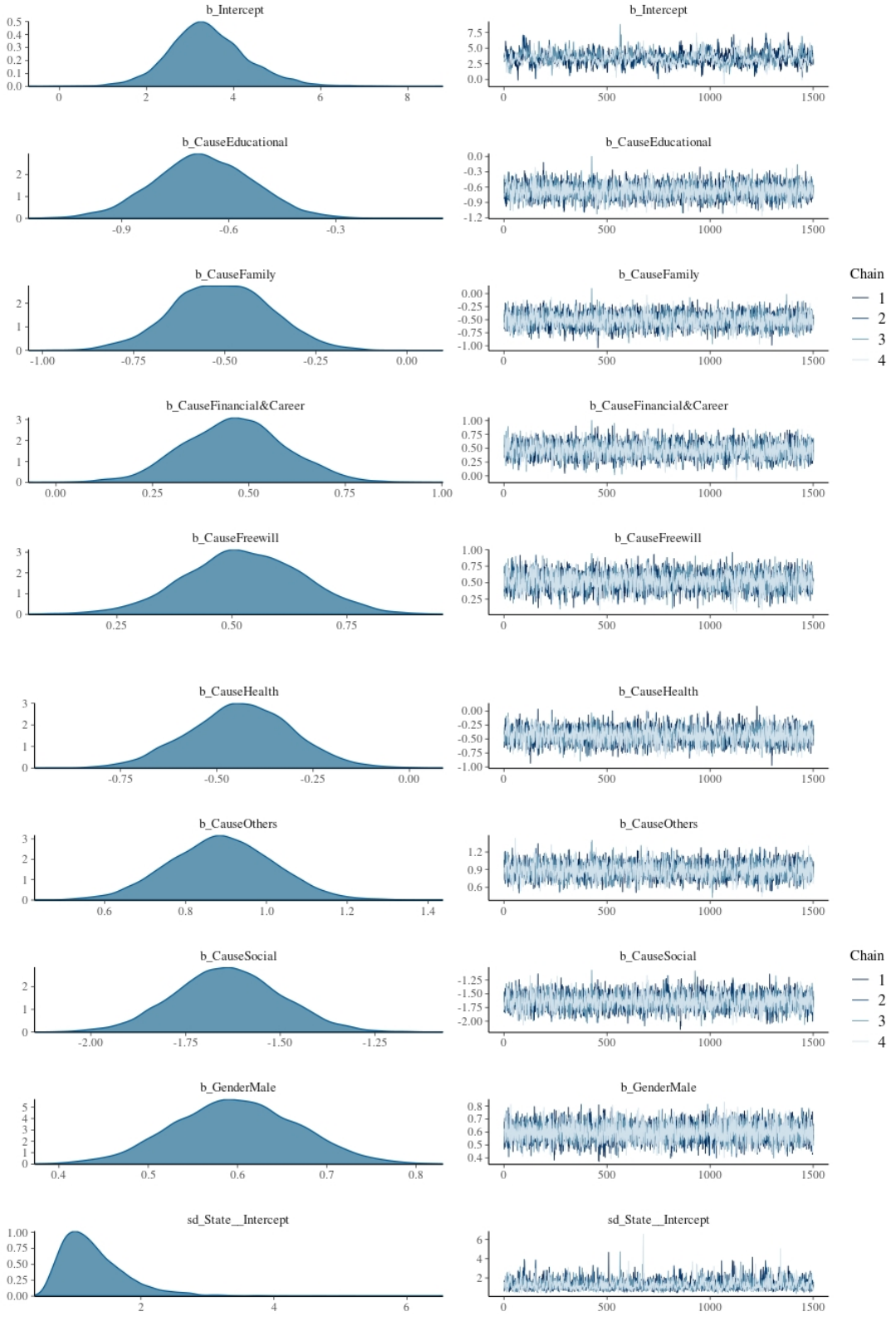
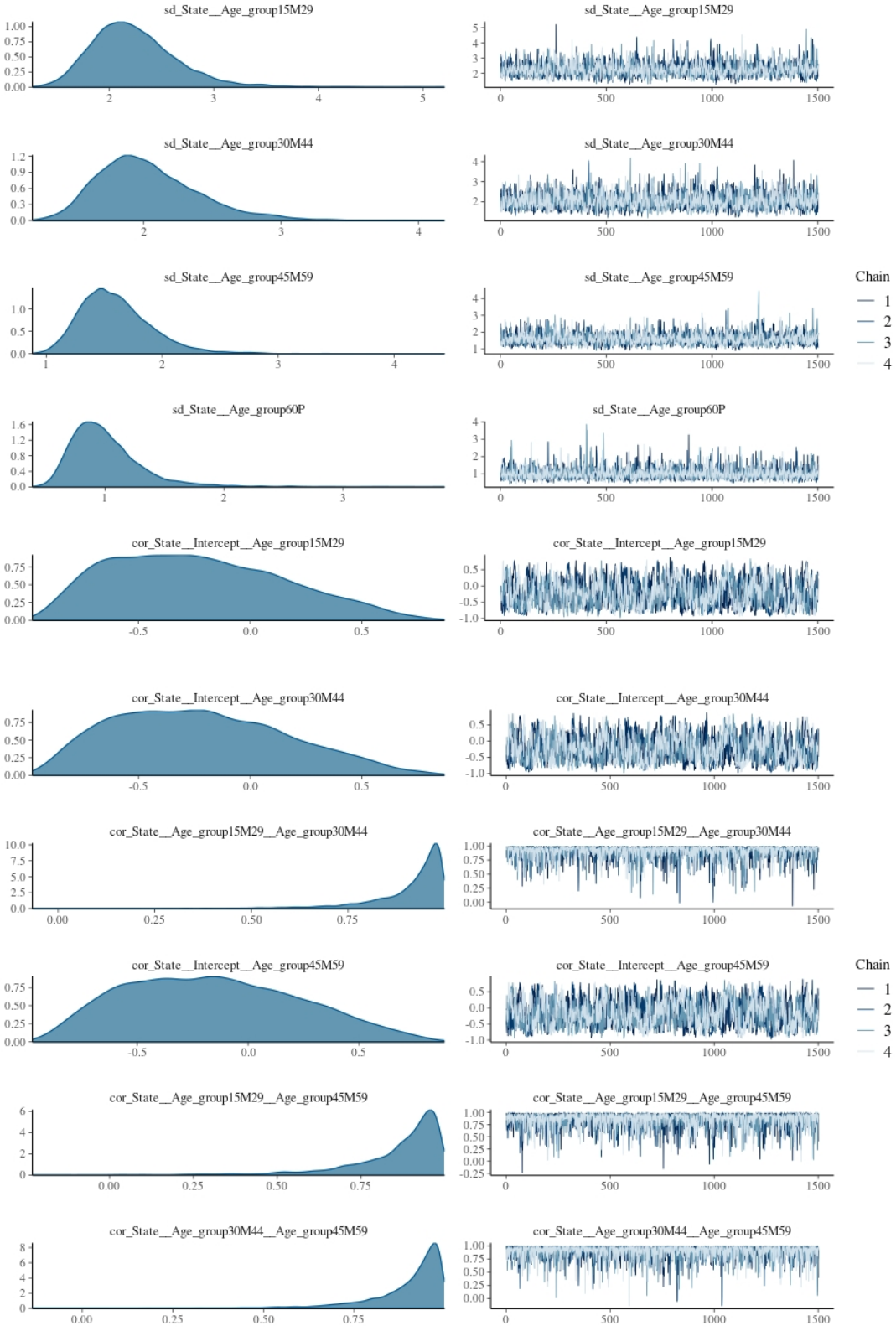


Figure A1: $\hat{R}$ value of Negative Binomial Model

Figure A2: $\hat{R}$ value of ZINB Model


b_Intercept
b_Intercept
b_CauseEducational
b_CauseEducational
b_CauseFamily
b_CauseFamily
b_CauseFinancial&Career
b_CauseFinancial&Career
b_CauseFreewill
b_CauseFreewill
b_CauseHealth
b_CauseHealth
b_CauseOthers
b_CauseOthers
b_CauseSocial
b_CauseSocial
b_GenderMale
b_GenderMale
sd_State__Intercept
sd_State__Intercept
Chain
1
2
3
4


sd_State__Age_group15M29
sd_State__Age_group30M44
sd_State__Age_group45M59
sd_State__Age_group60P
cor_State__Intercept__Age_group15M29
cor_State__Intercept__Age_group30M44
cor_State__Age_group15M29__Age_group30M44
cor_State__Intercept__Age_group45M59
cor_State__Age_group15M29__Age_group45M59
cor_State__Age_group30M44__Age_group45M59
Chain
1
2
3
4

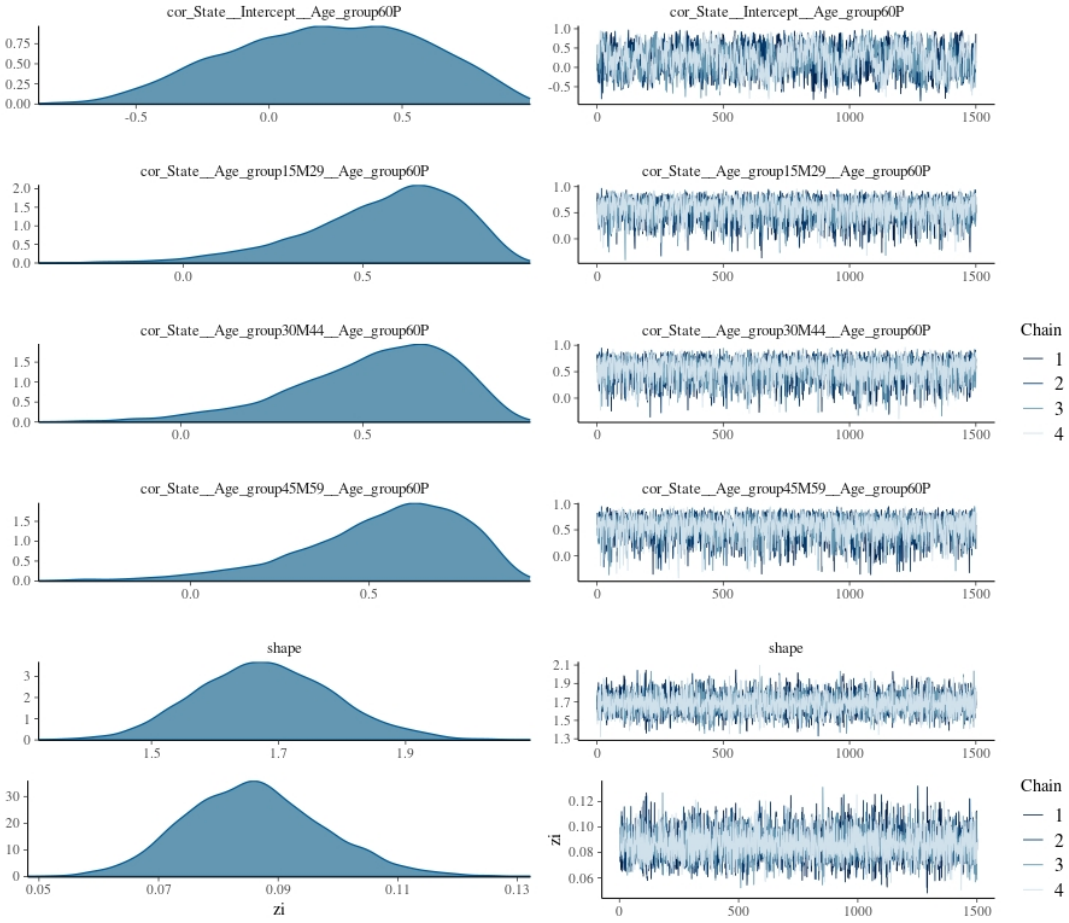


Figure A3: Trace and density plots for ZINB Model