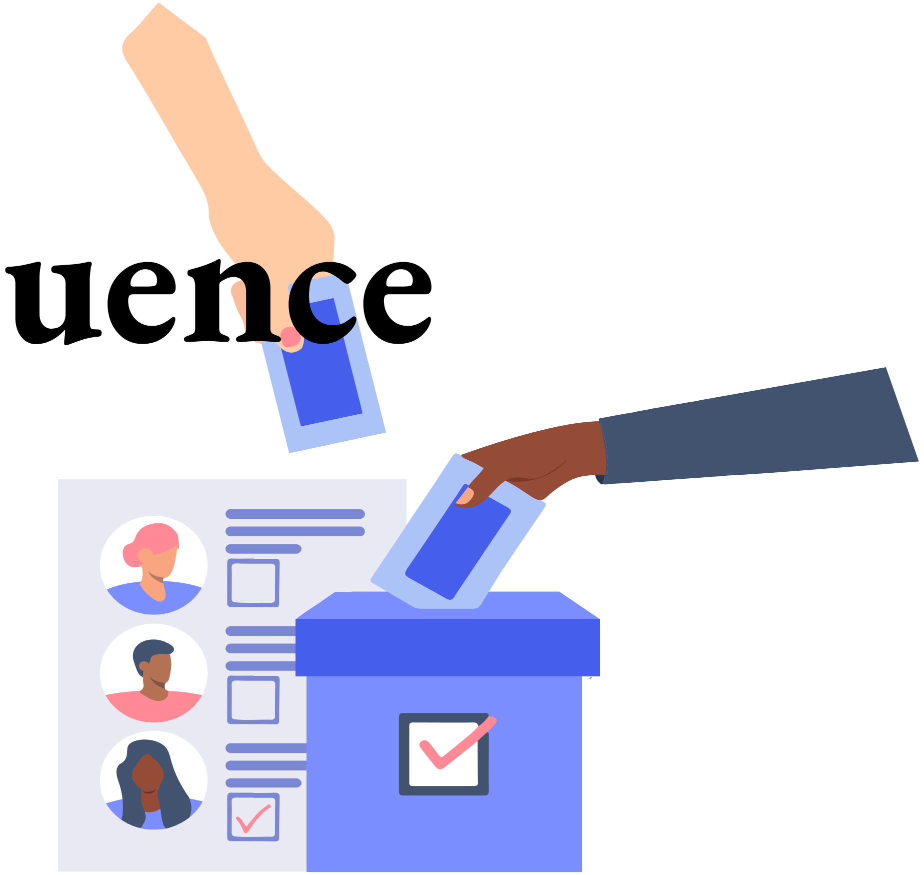
Applied Bayesian Data Analysis

# Factors that influence voting: A study

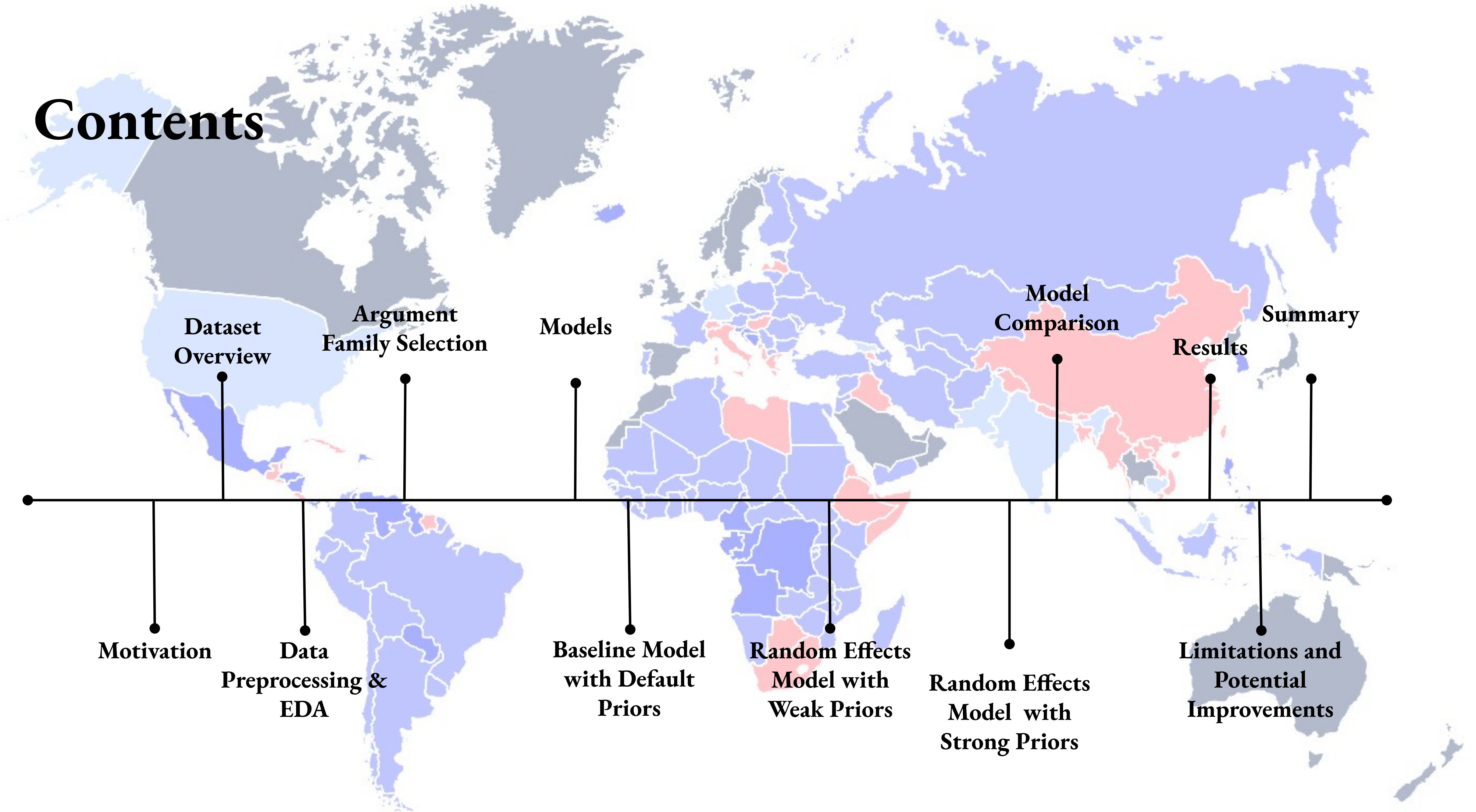
Based on World Values Survey



Anirudh Parameswaran Matr. No: 264787

Gautam D Hariharan Matr. No: 230237

# Contents



Motivation

Dataset Overview

Data Preprocessing & EDA

Argument Family Selection

Models

Baseline Model with Default Priors

Random Effects Model with Weak Priors

Random Effects Model with Strong Priors

Model Comparison

Results

Limitations and Potential Improvements

Summary

# Motivation

We want to understand key factors that influence voting behaviour in India and USA to investigate the dynamics of political participation.

It is especially vital in studying the impact of socioeconomic, cultural and demographic variables on civic engagement.

By examining variables such as age, sex, education, income etc. we aim to gain insights into universal and country-specific drivers of voter engagement.

# Dataset Overview: World Values Survey

The World Values Survey is a global research initiative that explores people's values, beliefs and cultural changes over time.

The dataset is derived from WVS, giving us a comparative analysis of factors that influence voting across different nations.

**Method**

Surveys over 100 countries done over multiple waves.

**Purpose**

Track values, beliefs, attitudes; analyze social, political, economic trends.

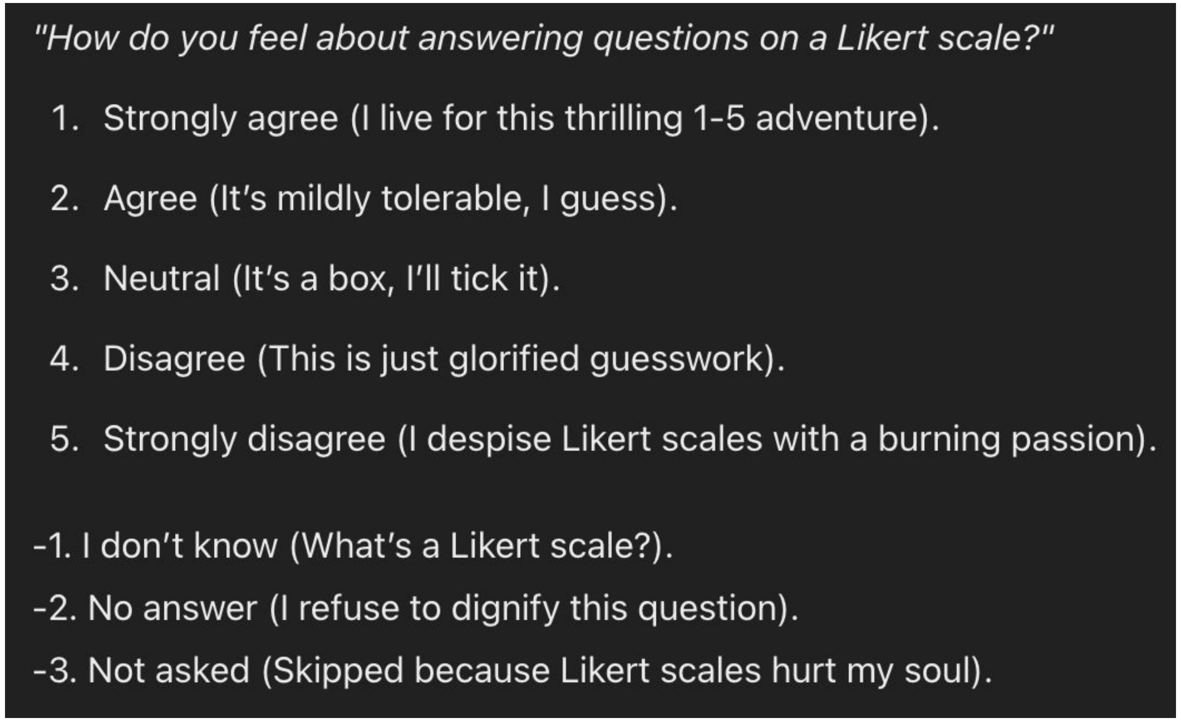
**Limitations**

Self-report bias

# Dataset Overview: World Values Survey

The dataset consists of survey responses to questions - so most variables are naturally categorical. Response type include:

- Numeric (age etc)
- Yes / No responses
- Likert scale responses
- Pure categorical (occupation types etc)
- Negative values are used to differentiate between different types of missing data



*"How do you feel about answering questions on a Likert scale?"*

1. Strongly agree (I live for this thrilling 1-5 adventure).
2. Agree (It’s mildly tolerable, I guess).
3. Neutral (It’s a box, I’ll tick it).
4. Disagree (This is just glorified guesswork).
5. Strongly disagree (I despise Likert scales with a burning passion).

-1. I don’t know (What’s a Likert scale?).

-2. No answer (I refuse to dignify this question).

-3. Not asked (Skipped because Likert scales hurt my soul).

# Data Preprocessing & EDA

**Dataset issues:**

- 3.3% rows contain missing values
- Target field - whether a person votes or not - is given as survey responses to two questions - local level and national level elections
- Data degeneracy - More than 80% respondents have mentioned that they vote!

**Solutions:**

- Variable Encoding:
    - Target Variable encoding
    - Other variables
- Resampling to mitigate data degeneracy issues and handle missing data

# Variable Encoding

For the voter column, we have 2 options:

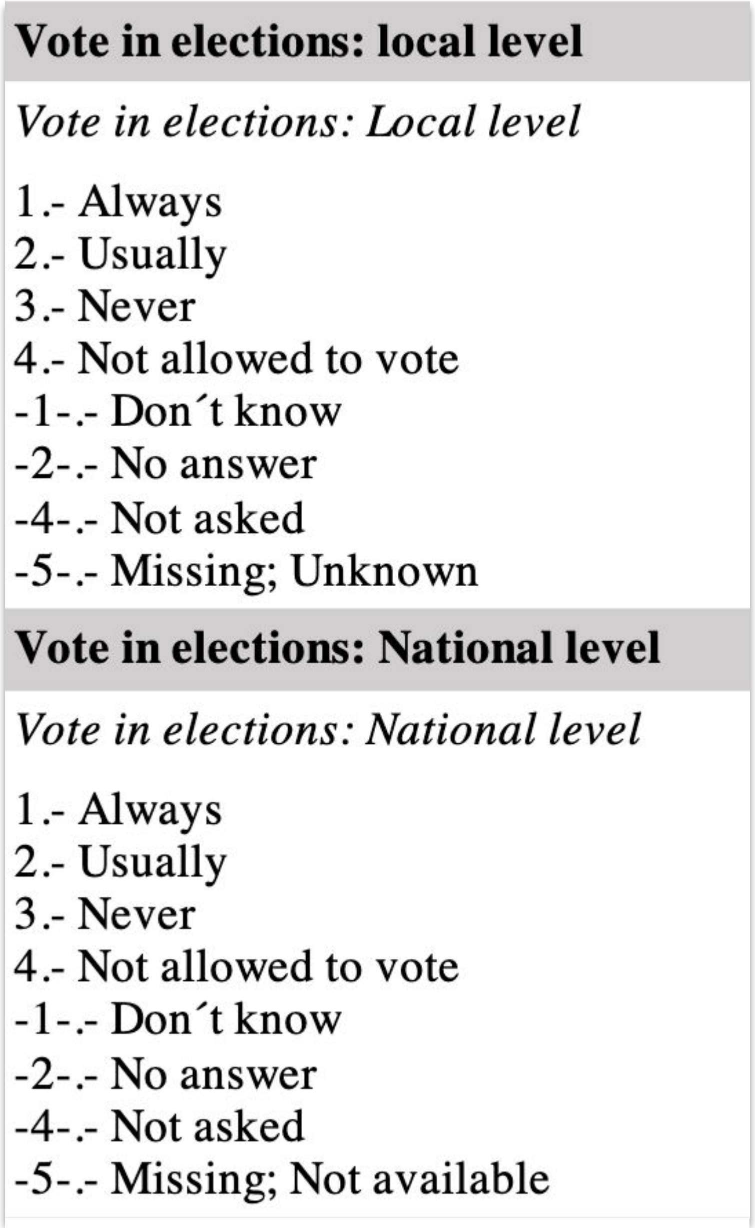
Two classes - Binary

- Always, Usually = 1
- Never, Not allowed to vote = 0
- Rest = NA

Three classes - Ordinal

- Always (either level) = 3
- Usually + Never = 2
- Never or Not allowed = 1
- Rest = NA

In this presentation we will only work with the binary encoded target.



**Vote in elections: local level**

*Vote in elections: Local level*

1.- Always
2.- Usually
3.- Never
4.- Not allowed to vote
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Unknown

**Vote in elections: National level**

*Vote in elections: National level*

1.- Always
2.- Usually
3.- Never
4.- Not allowed to vote
-1-.- Don´t know
-2-.- No answer
-4-.- Not asked
-5-.- Missing; Not available

Fig 1. Survey Questions about voting

# Resampling - Two birds with one stone!

Skip missing values

By undersampling majority class and oversampling minority class, helps remove bias in the target variable.

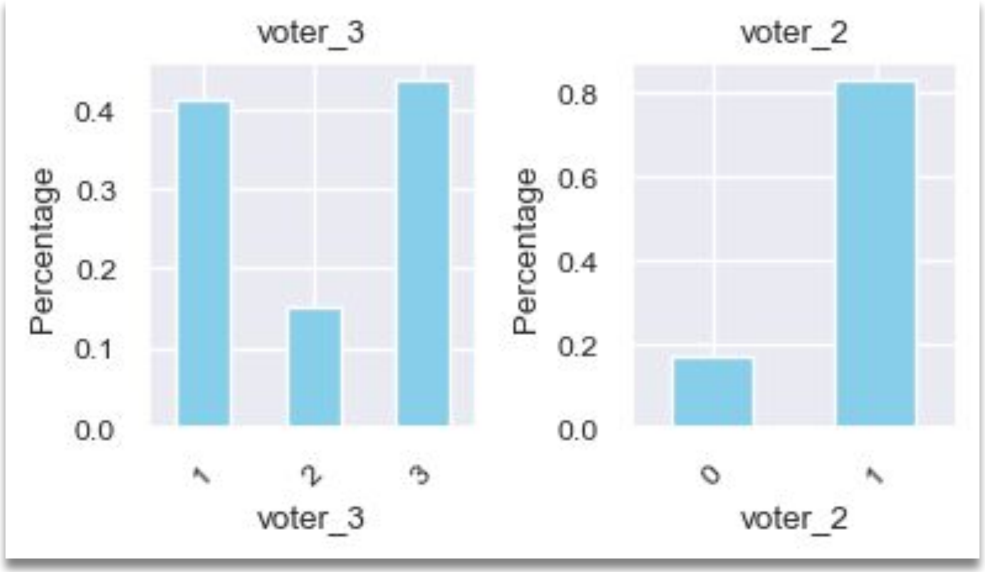


Fig 2a. Distribution before sampling

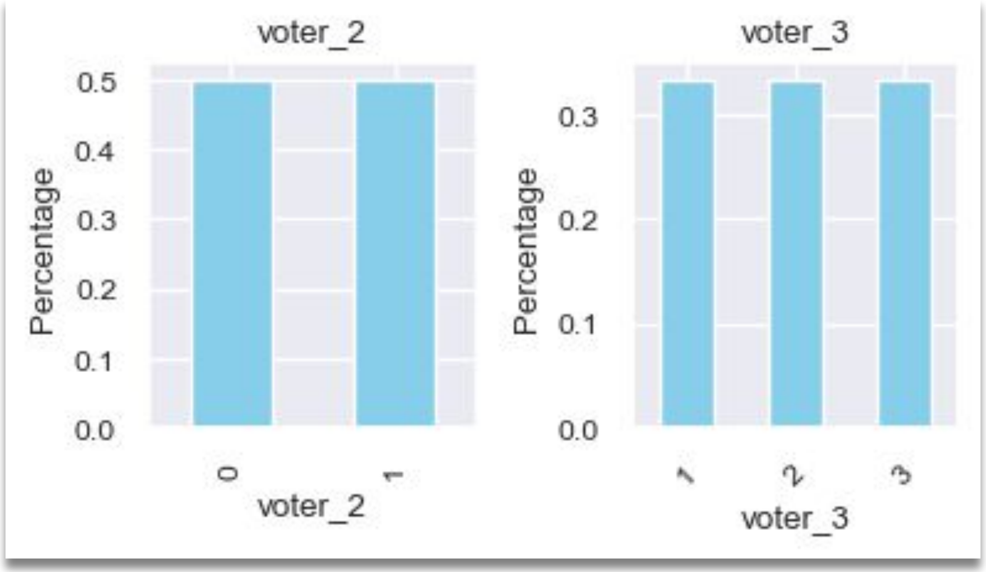


Fig 2b. Distribution after sampling

# Exploratory Data Analysis: Final Dataset Overview

| Field | Data type | Remarks |
|---|---|---|
| country | category | "IND", "USA" |
| satisfaction | numeric (1-10) | Likert scale |
| sex | category | "Male", "Female" |
| age | numeric | |
| immigrant | category | "Immigrant", "Not Immigrant" |
| children | numeric | |
| income_level | numeric (1-10) | Likert scale |
| education | numeric (0-8) | |
| god_importance | numeric (1-10) | Likert scale |
| praying_frequency | numeric (1-4) | |
| ethics_score | decimal (0-10) | Normalised response to ethical questions |
| voter_2 | category | 1 - yes, 0 - no |

Table 1. Overview of columns

# Exploratory Data Analysis: Final Dataset Overview

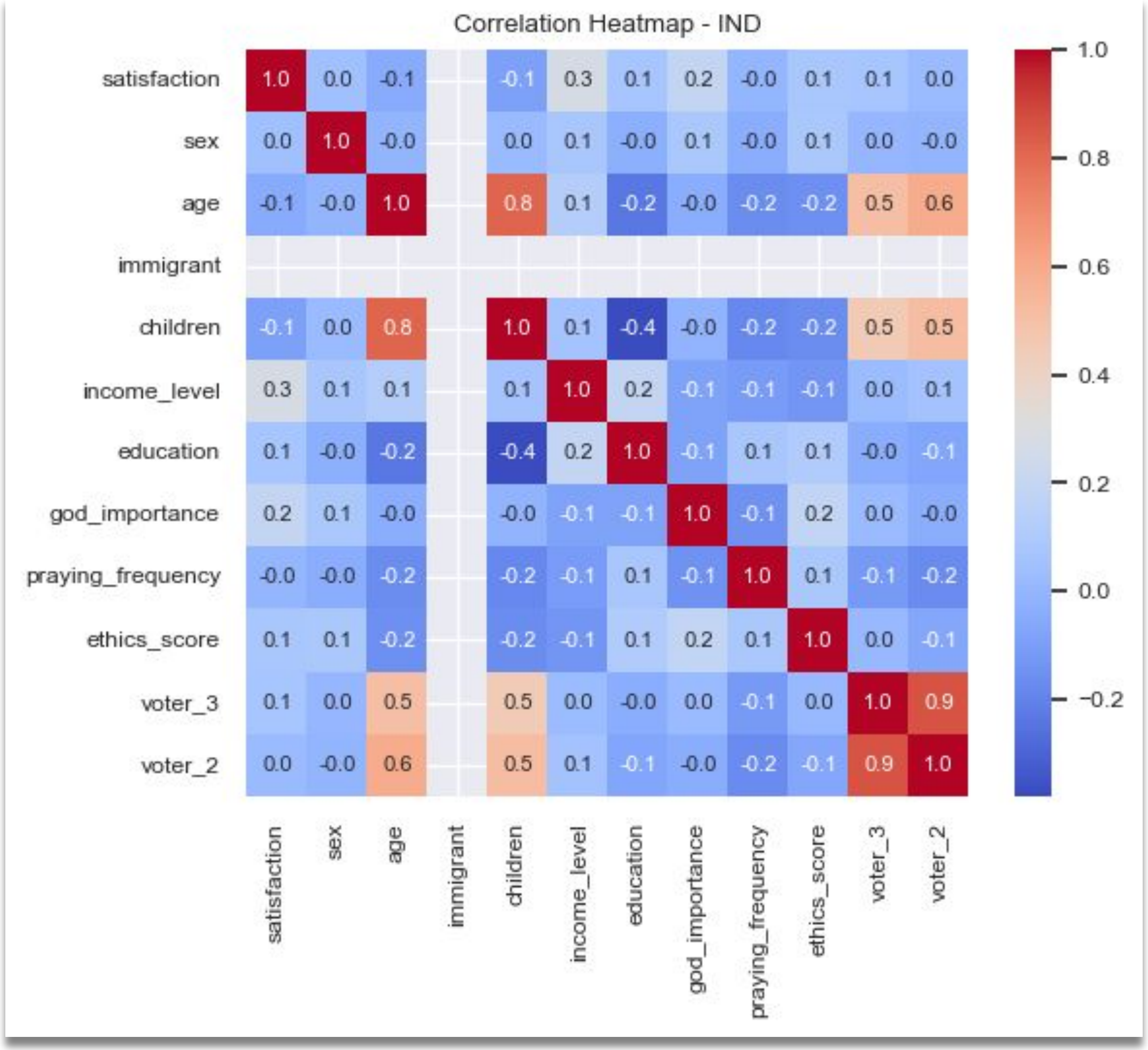


Fig 3a. Correlation Heatmap - India

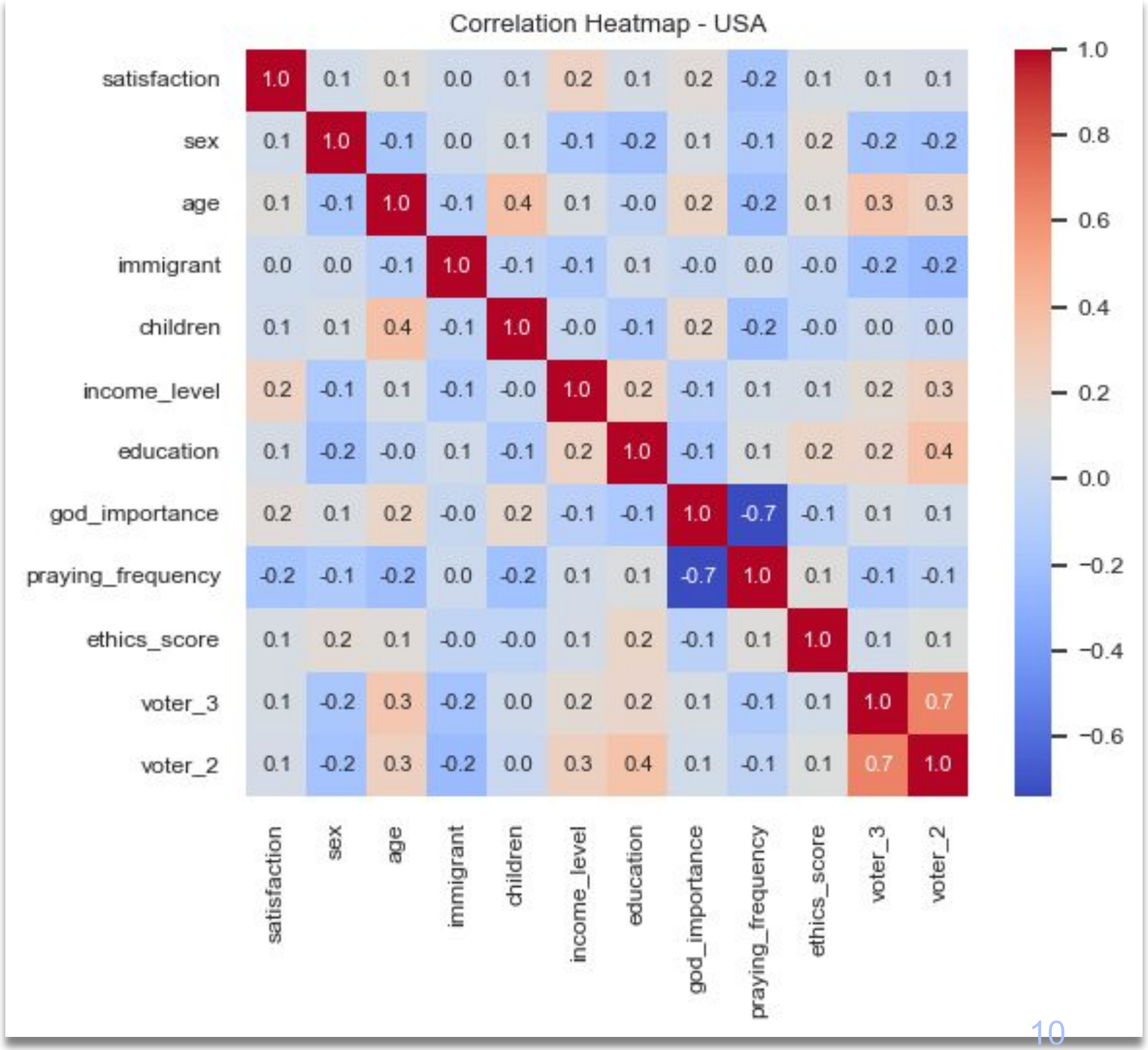


Fig 3b. Correlation Heatmap - USA

# Argument Family

**For voter_2:**

- **bernoulli - right choice**
- binomial - counts success so not appropriate

**For voter_3:**

- cumulative - equal effect of predictors on response variable across categories
- cratio - suitable for sequential transitions - moving from one category to next
- sratio - suitable for stopping at one category vs lower categories
- acat - adjacent categories have similar probabilities

In this project we will be focusing on the binary encoded target with the bernoulli distribution only.

# Model Selection

**Baseline**: Assuming equal effect of predictors across groups (IND vs USA); with default priors

```
voter_2 ~ age + sex + immigrant + income_level + education + god_importance +
praying_frequency + ethics_score + satisfaction + (1 | country)
```

**Random effects 1**: Assuming unequal effect of sex, education and income_level between countries; all predictors with weakly informative priors

```
voter_2 ~ age + sex + immigrant + income_level + education + god_importance +
praying_frequency + ethics_score + satisfaction + (1 + sex + age + education +
income_level | country)
```

**Random effects 2**: Assuming unequal effect of sex, education and income_level between countries; selected predictors with strongly informative priors

```
voter_2 ~ age + sex + immigrant + income_level + education + praying_frequency + (1 +
sex + age + education + income_level | country)
```

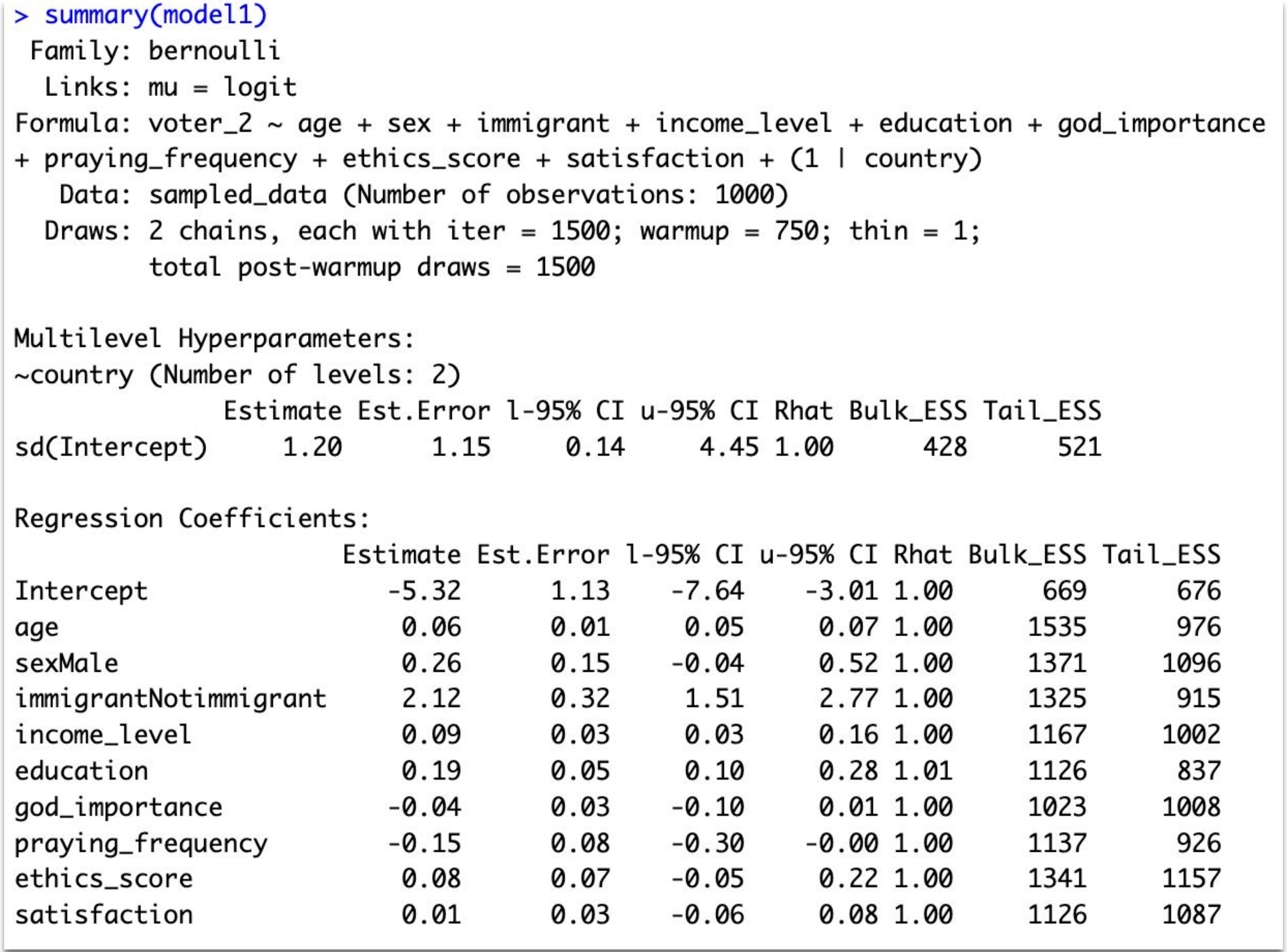
# Baseline

Significant covariates:

- Age
- Immigrant
- Income Level
- Education
- Praying Frequency

Values of Rhat close to 1.00 suggest that chains have converged to a common distribution

High error in sd(Intercept) indicates that we are not certain about the deviation in baseline voting pattern between India and USA.



```
> summary(model1)
 Family: bernoulli
  Links: mu = logit
Formula: voter_2 ~ age + sex + immigrant + income_level + education + god_importance
+ praying_frequency + ethics_score + satisfaction + (1 | country)
   Data: sampled_data (Number of observations: 1000)
  Draws: 2 chains, each with iter = 1500; warmup = 750; thin = 1;
         total post-warmup draws = 1500

Multilevel Hyperparameters:
~country (Number of levels: 2)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.20      1.15     0.14     4.45 1.00      428      521

Regression Coefficients:
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept               -5.32      1.13    -7.64    -3.01 1.00      669      676
age                      0.06      0.01     0.05     0.07 1.00     1535      976
sexMale                  0.26      0.15    -0.04     0.52 1.00     1371     1096
immigrantNotimmigrant    2.12      0.32     1.51     2.77 1.00     1325      915
income_level             0.09      0.03     0.03     0.16 1.00     1167     1002
education                0.19      0.05     0.10     0.28 1.01     1126      837
god_importance          -0.04      0.03    -0.10     0.01 1.00     1023     1008
praying_frequency       -0.15      0.08    -0.30    -0.00 1.00     1137      926
ethics_score             0.08      0.07    -0.05     0.22 1.00     1341     1157
satisfaction             0.01      0.03    -0.06     0.08 1.00     1126     1087
```

Fig 4 Baseline formula summary

# Baseline

**Age, Education, and Income Are Important Predictors**

- Older individuals, those with higher education, and those with higher income are more likely to vote.

**Non-immigrants are significantly more likely to vote than immigrants**

**Religious Predictors Have Mixed Effects**

- Praying frequency negatively impacts voting, while God's importance and ethics score have no significant effect.

**Country Effects**

- There's substantial variability in voting likelihood across countries, though the uncertainty in sd(Intercept) suggests limited data for precise estimates.

# Random Effects

What are Random Effects?

Fixed effects are constant across individuals, and random effects vary.

For example, in a growth study, a model with random intercepts a_i and fixed slope b corresponds to parallel lines for different individuals i, or the model y_it = a_i + b t. (Kreft and De Leeuw (1998))

We can use this to capture the difference in effects between India and USA. For example:

- Men and women vote differently in India vs USA
- Age, income level and education would have different effects in each country.

These ideas can be captured by the formula:

```
voter_2 ~ age + sex + immigrant + income_level + education + god_importance +
praying_frequency + ethics_score + satisfaction + (1 + sex + age + education +
income_level | country)
```

# Weakly Informative Priors

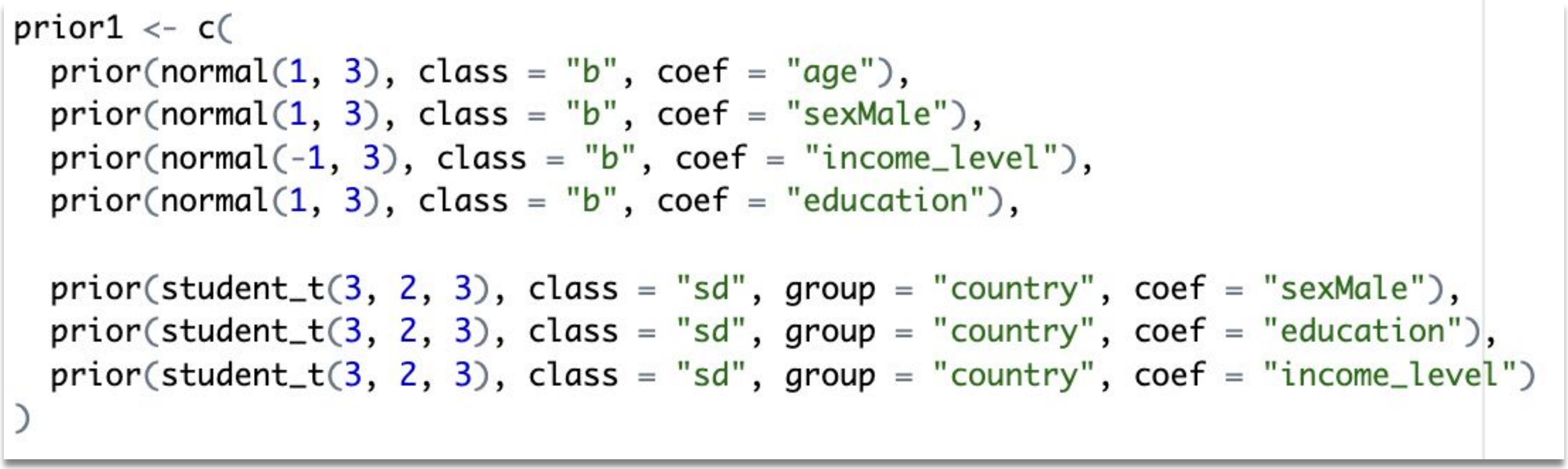
Let’s say we have some idea about how these factors should affect the probability of voting, but we are not very sure. We can encode these assumptions into weakly informative priors.

- Age should have a positive coefficient indicating that older people have a higher chance of voting.
- Men have a higher probability of voting than women.
- Income level should have a positive coefficient indicating that people who are doing well financially have a higher chance of voting.
- Education should have a positive coefficient indicating that highly educated people have a higher chance of voting.
- There would be some deviation in the slopes between India and USA - meaning that these factors would influence the voting chances different for each country.

# Weakly Informative Priors

For coefficients of age, sex, income_level and education, a normal distribution is considered since we do not expect extreme values to be very likely. However, the high SD represents our uncertainty about the true value.

For coefficients of SD between India and USA, we are using a student-t distribution with 3 degrees of freedom as there might be a higher chance of extreme values.



```
prior1 <- c(
  prior(normal(1, 3), class = "b", coef = "age"),
  prior(normal(1, 3), class = "b", coef = "sexMale"),
  prior(normal(-1, 3), class = "b", coef = "income_level"),
  prior(normal(1, 3), class = "b", coef = "education"),

  prior(student_t(3, 2, 3), class = "sd", group = "country", coef = "sexMale"),
  prior(student_t(3, 2, 3), class = "sd", group = "country", coef = "education"),
  prior(student_t(3, 2, 3), class = "sd", group = "country", coef = "income_level")
)
```

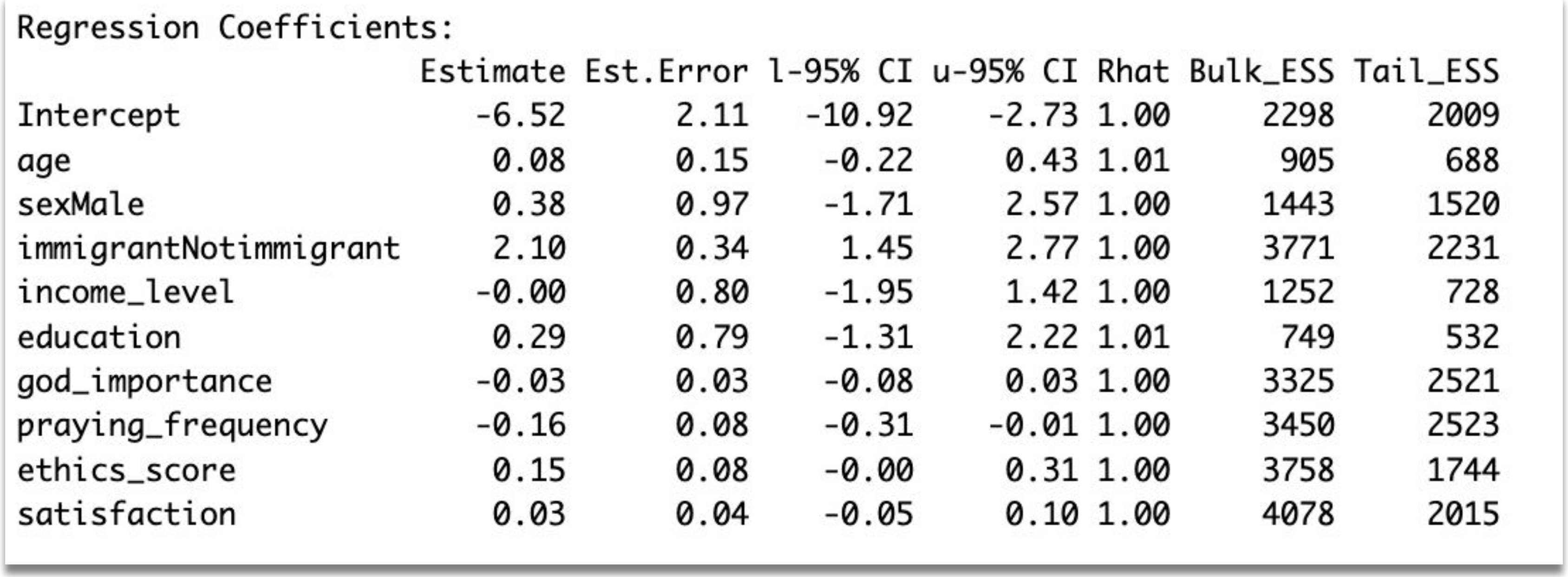
Fig 5 Weakly Informative priors specification

# Random Effects 1

Significant covariates:

- Immigrant
- Praying Frequency

Since the priors we specified had high uncertainty, our model outputs also reflect the uncertainty. Thus, we have very few significant covariates.

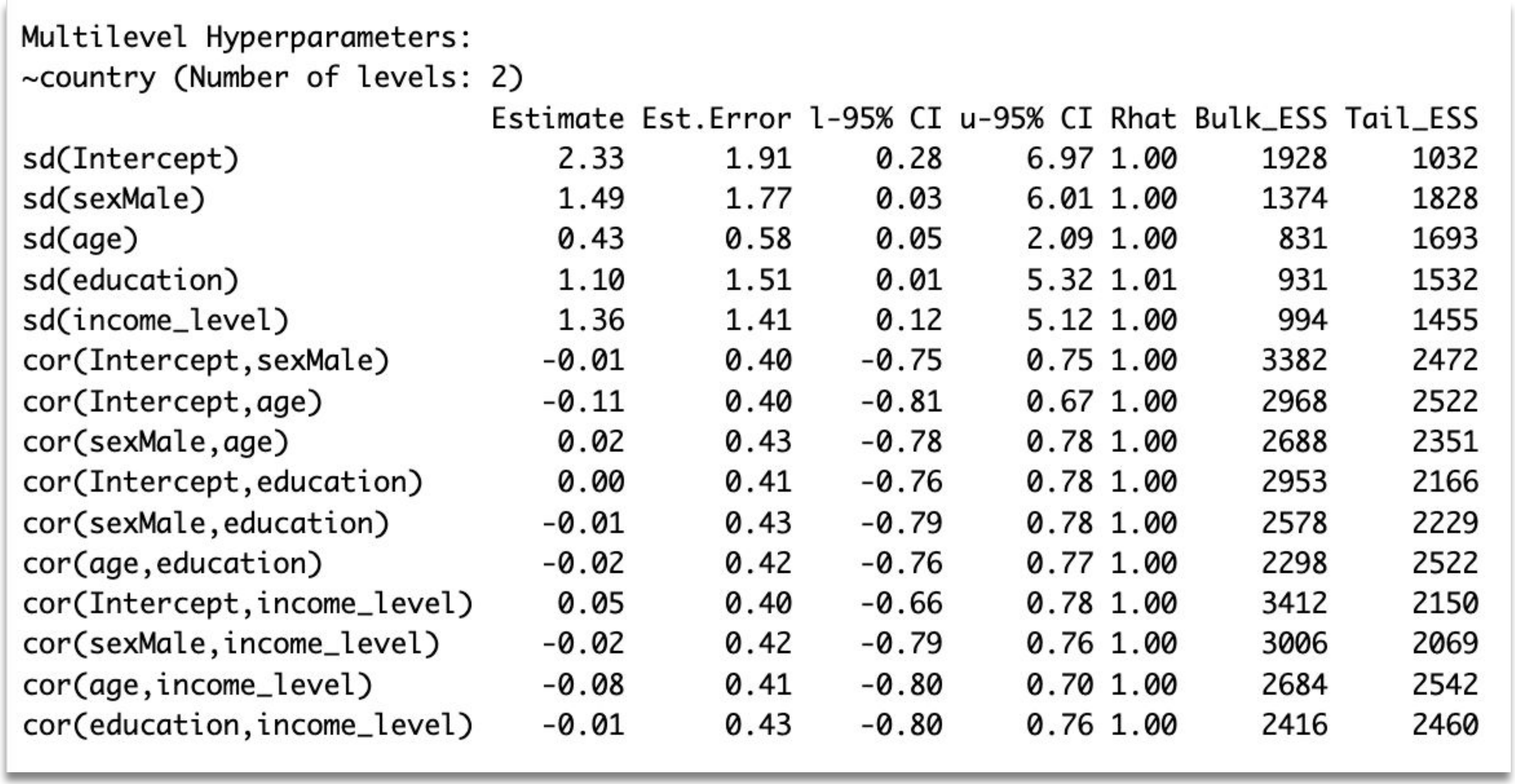


```
Regression Coefficients:
                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                 -6.52      2.11   -10.92    -2.73 1.00     2298     2009
age                        0.08      0.15    -0.22     0.43 1.01      905      688
sexMale                    0.38      0.97    -1.71     2.57 1.00     1443     1520
immigrantNotimmigrant      2.10      0.34     1.45     2.77 1.00     3771     2231
income_level              -0.00      0.80    -1.95     1.42 1.00     1252      728
education                  0.29      0.79    -1.31     2.22 1.01      749      532
god_importance            -0.03      0.03    -0.08     0.03 1.00     3325     2521
praying_frequency         -0.16      0.08    -0.31    -0.01 1.00     3450     2523
ethics_score               0.15      0.08    -0.00     0.31 1.00     3758     1744
satisfaction               0.03      0.04    -0.05     0.10 1.00     4078     2015
```

Fig 6a Fixed effects summary

# Random Effects 1

- None of the correlations are significant
- Each SD is significant but there is high error in the estimates. This is both because our priors have a lot of uncertainty and also because of the fact that there are only 2 groups (India vs USA).



```
Multilevel Hyperparameters:
~country (Number of levels: 2)
                            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                   2.33      1.91     0.28     6.97 1.00     1928     1032
sd(sexMale)                     1.49      1.77     0.03     6.01 1.00     1374     1828
sd(age)                         0.43      0.58     0.05     2.09 1.00      831     1693
sd(education)                   1.10      1.51     0.01     5.32 1.01      931     1532
sd(income_level)                1.36      1.41     0.12     5.12 1.00      994     1455
cor(Intercept,sexMale)         -0.01      0.40    -0.75     0.75 1.00     3382     2472
cor(Intercept,age)             -0.11      0.40    -0.81     0.67 1.00     2968     2522
cor(sexMale,age)                0.02      0.43    -0.78     0.78 1.00     2688     2351
cor(Intercept,education)        0.00      0.41    -0.76     0.78 1.00     2953     2166
cor(sexMale,education)         -0.01      0.43    -0.79     0.78 1.00     2578     2229
cor(age,education)             -0.02      0.42    -0.76     0.77 1.00     2298     2522
cor(Intercept,income_level)     0.05      0.40    -0.66     0.78 1.00     3412     2150
cor(sexMale,income_level)      -0.02      0.42    -0.79     0.76 1.00     3006     2069
cor(age,income_level)          -0.08      0.41    -0.80     0.70 1.00     2684     2542
cor(education,income_level)    -0.01      0.43    -0.80     0.76 1.00     2416     2460
```

Fig 6b Random effects summary

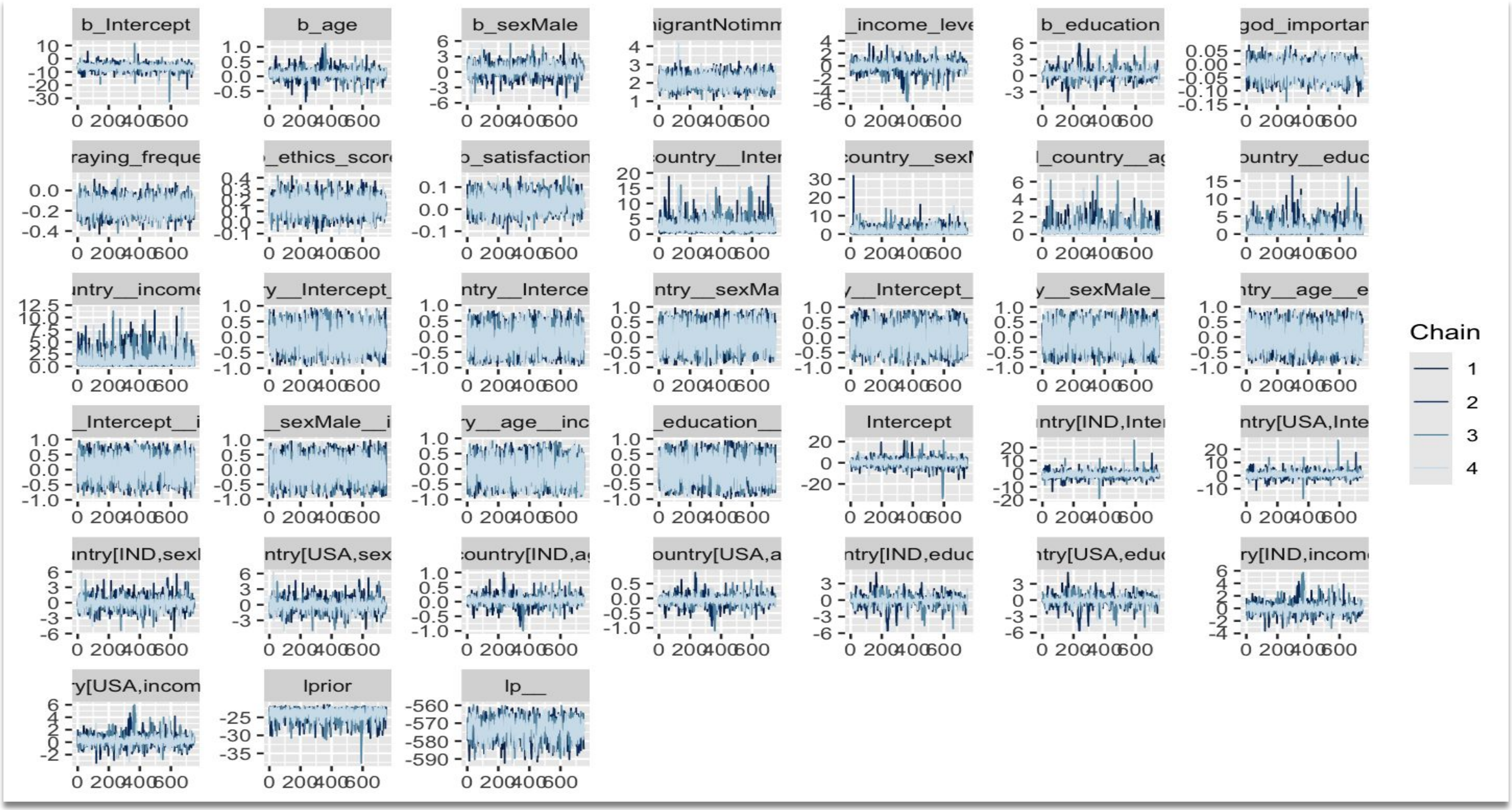


Fig 7 (Rather messy) caterpillar plots

# Posterior predictive check

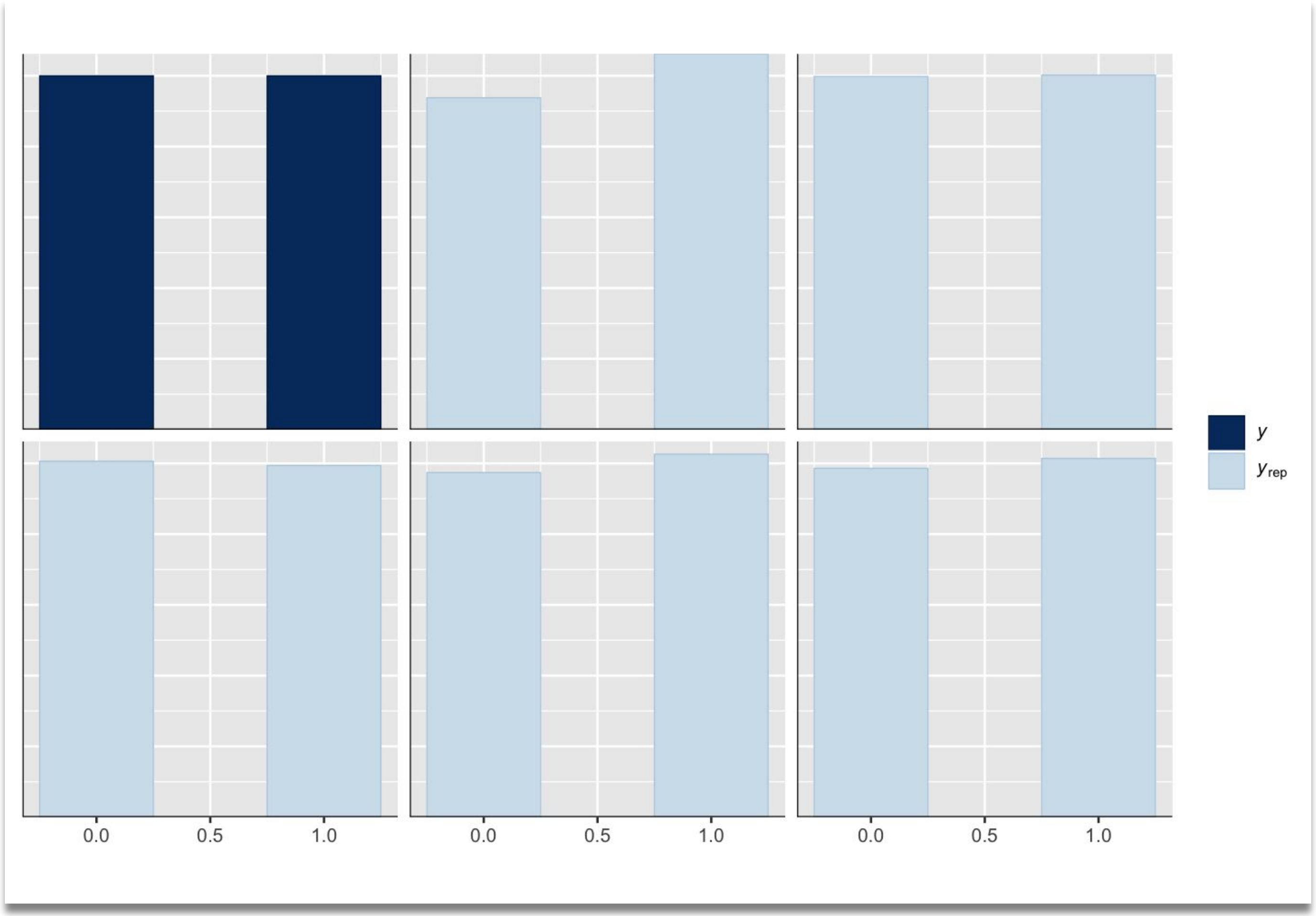


Fig 8a. Comparison between y and simulated draws of y

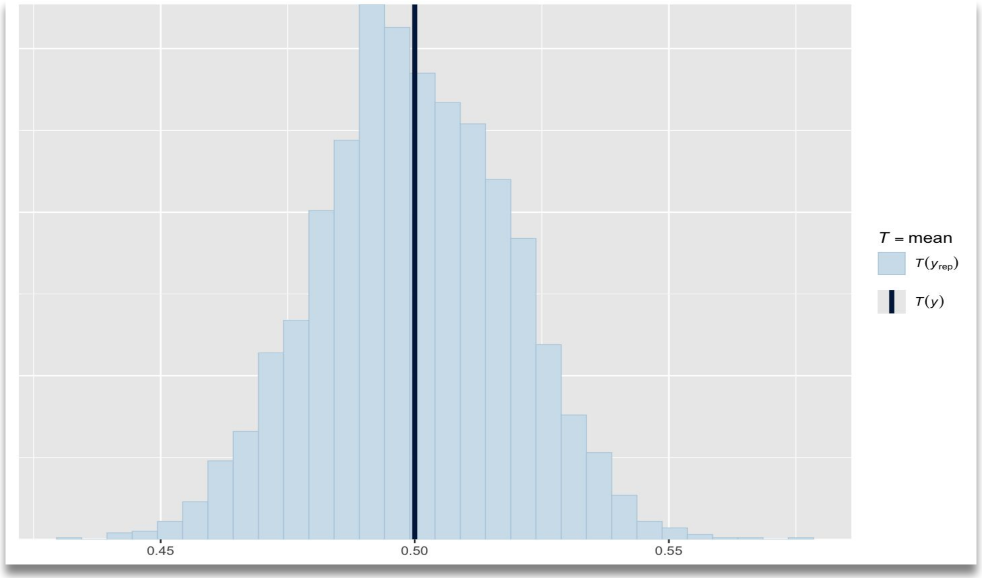


Fig 8b. Comparison of mean

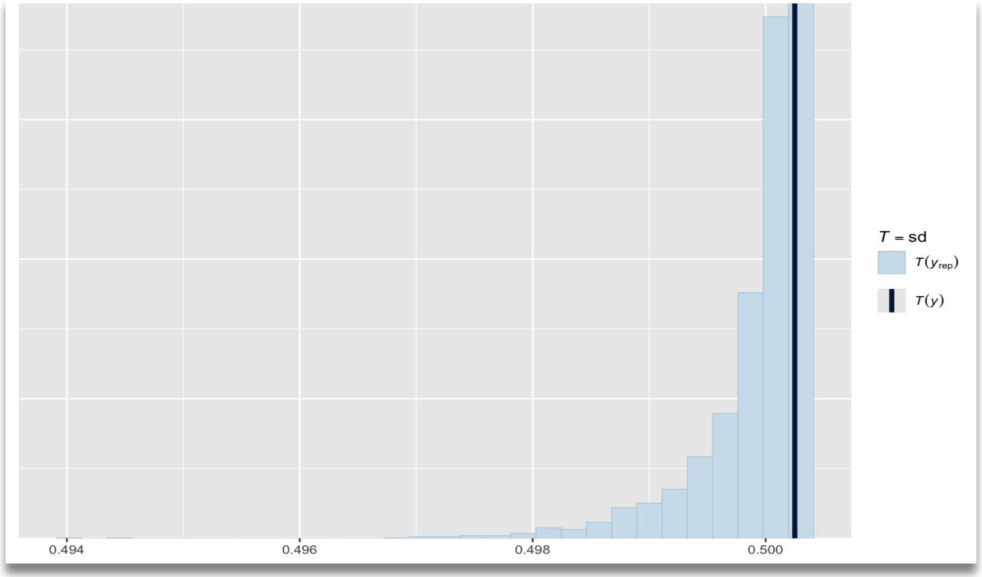


Fig 8c. Comparison of SD

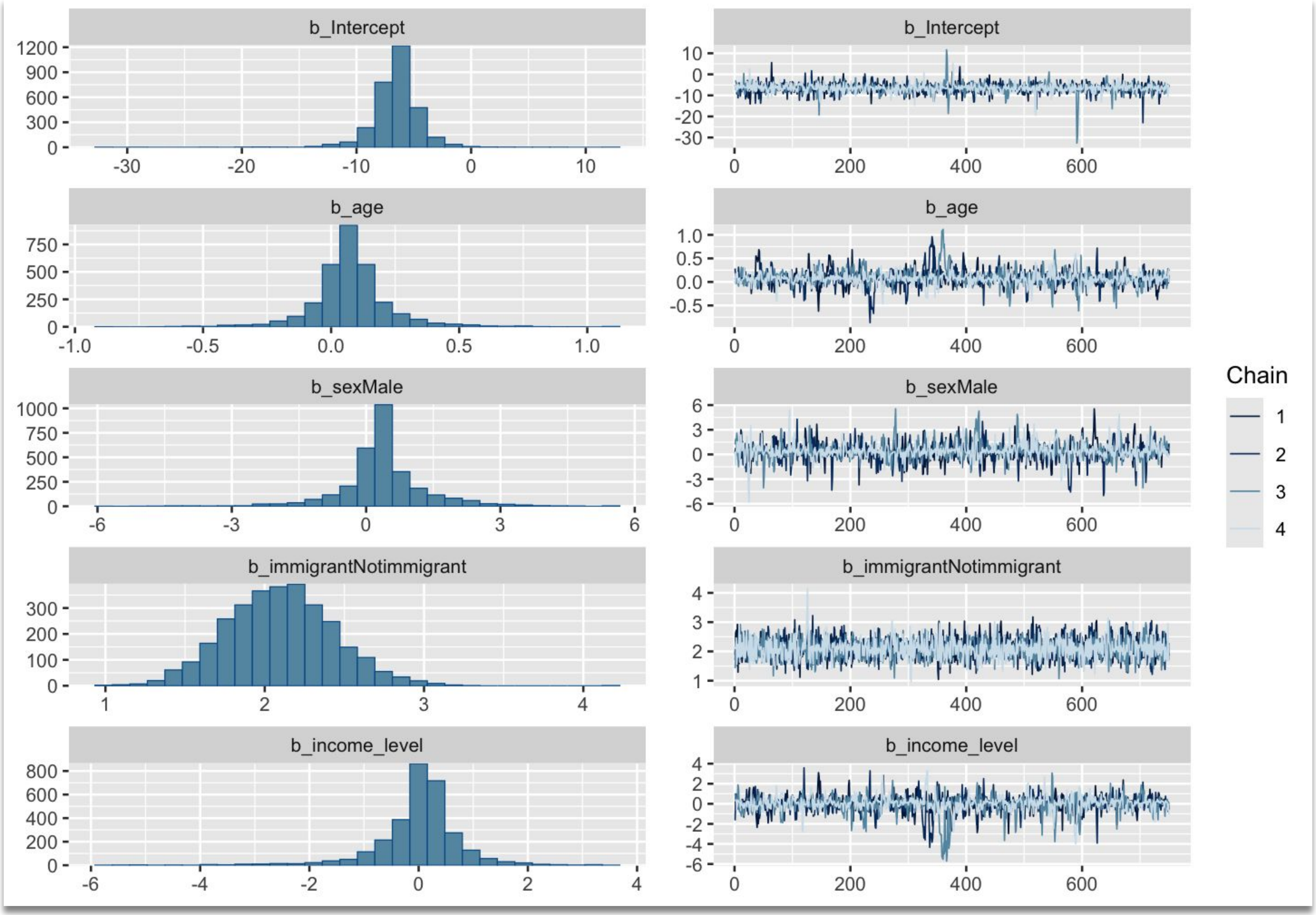


Fig 9a. Density and caterpillar plots - Coefficients

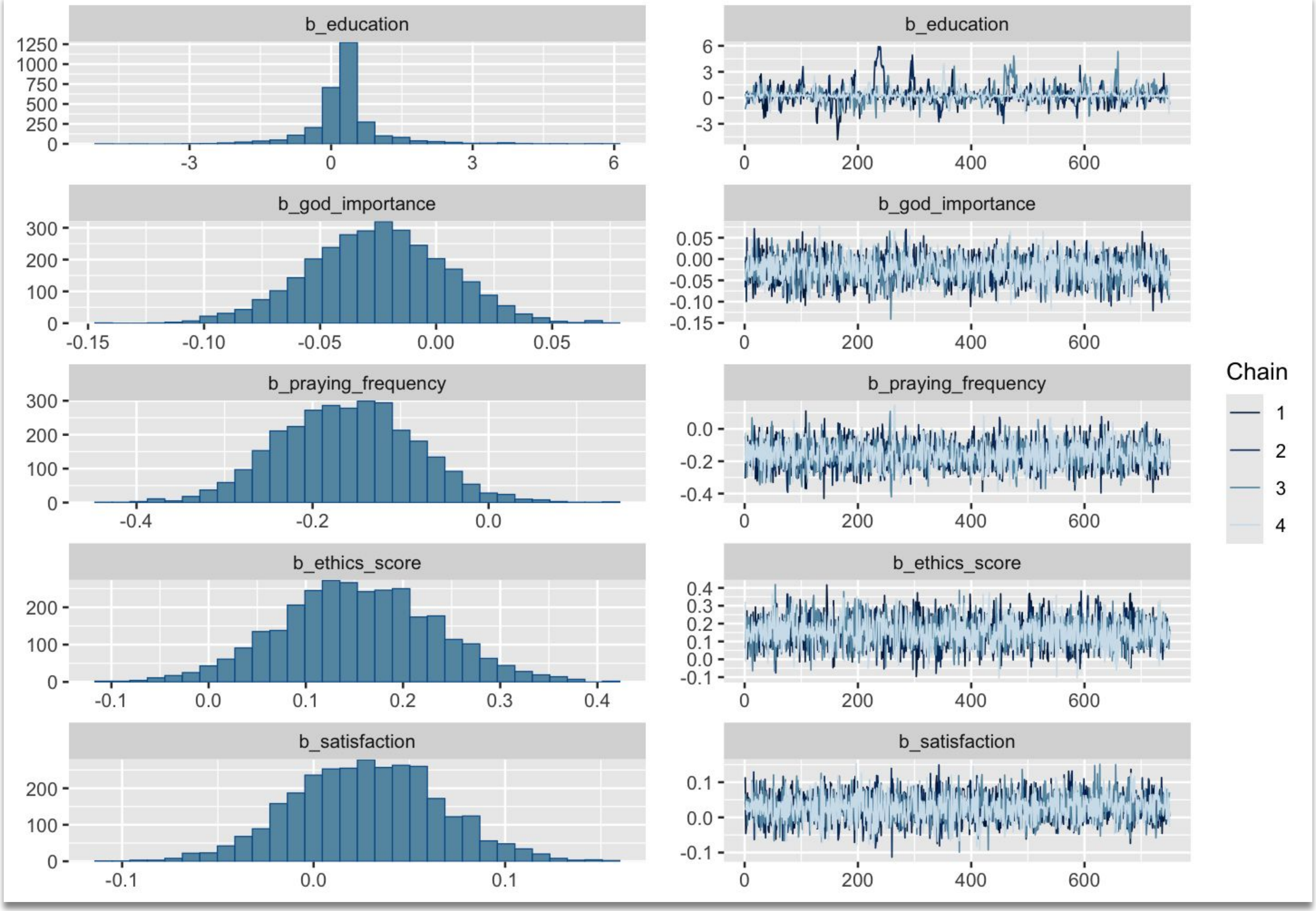


Fig 9b. Density and caterpillar plots - Coefficients

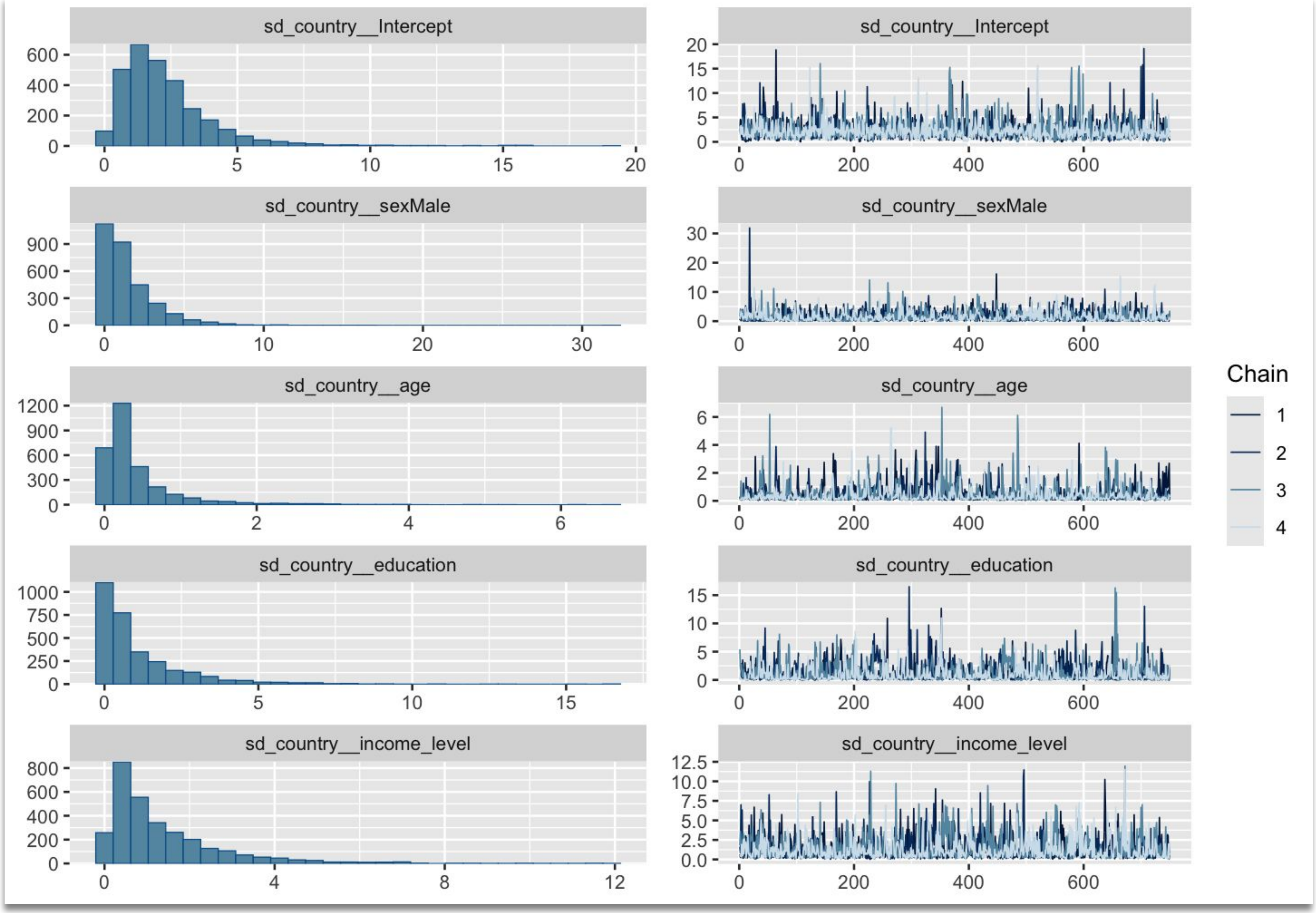


Fig 9c. Density and caterpillar plots - SD

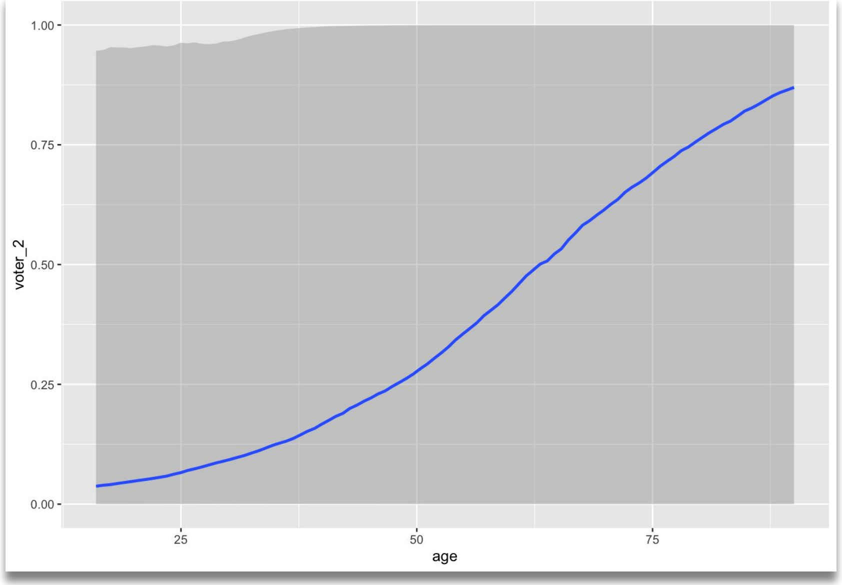


Fig 10a. Age vs Voter

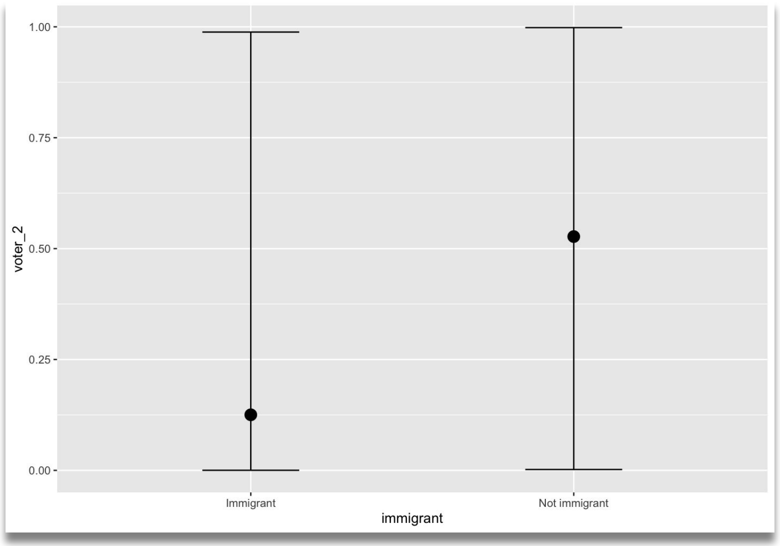


Fig 10b. Immigrant vs Voter

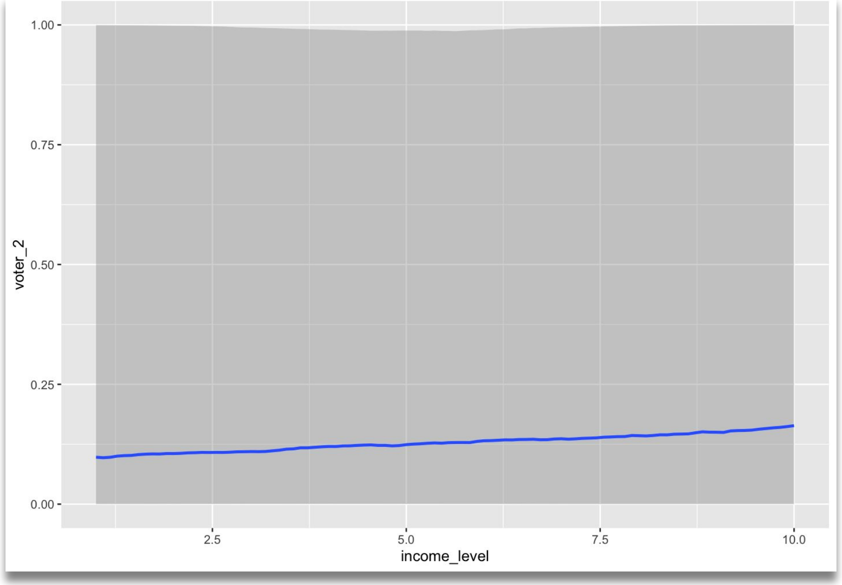


Fig 10c. Income level vs Voter

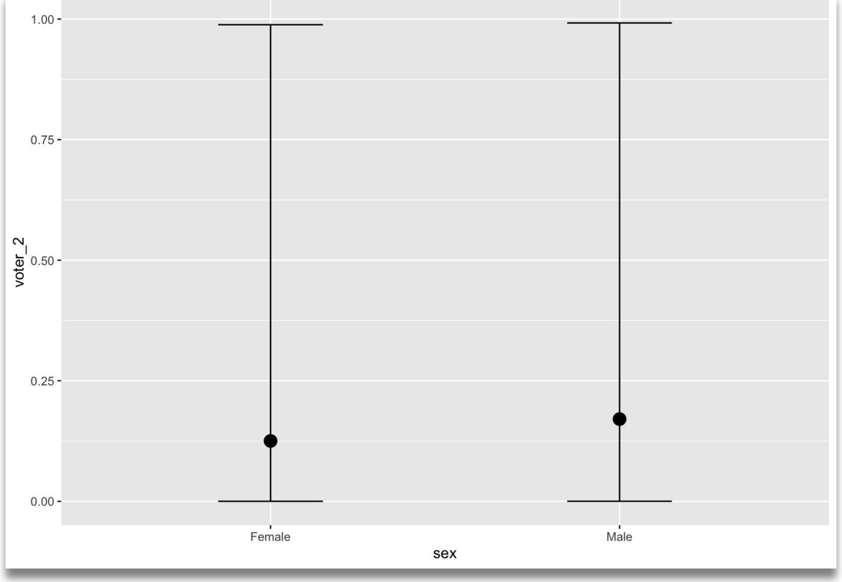


Fig 10d. Sex vs Voter

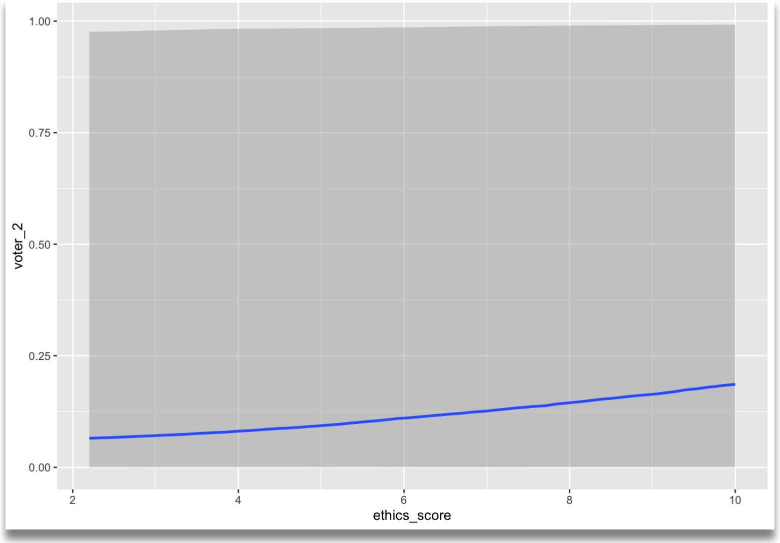


Fig 10e. Ethics score vs Voter

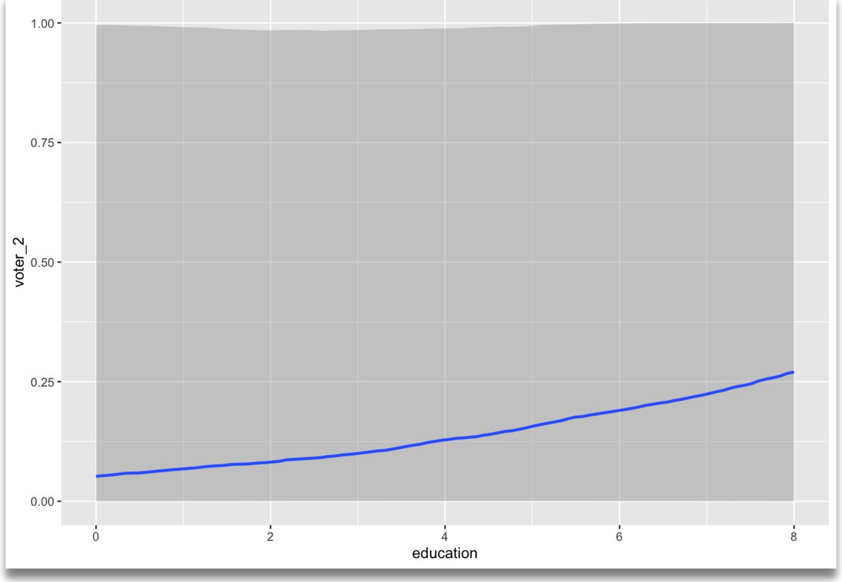


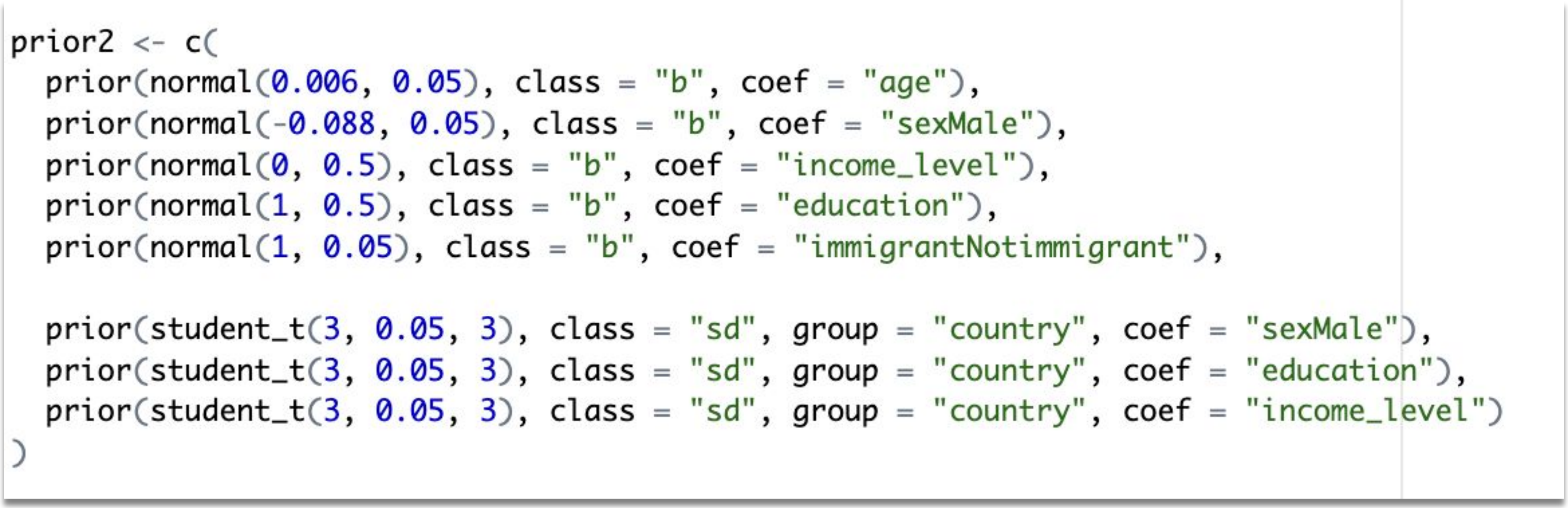
Fig 10f. Education vs Voter

# Strongly Informative Priors

Now we have a better idea about how these factors should affect the probability of voting, and we are quite sure about our intuition. We can encode these assumptions into strongly informative priors.

- Voting chances increase by ~0.6% with each 1 year increase in age. [1]
- Women (65%) have a slightly higher probability of voting than men (63%). [2]
- While the income has a positive coefficients, it is quite likely that this is the reverse in India. [3]
- Education has a positive effect on voting - highly educated people are more likely to vote.
- Both USA and India have quite similar overall voter turnout (66% vs 64%).

# Strongly Informative Priors



```
prior2 <- c(
  prior(normal(0.006, 0.05), class = "b", coef = "age"),
  prior(normal(-0.088, 0.05), class = "b", coef = "sexMale"),
  prior(normal(0, 0.5), class = "b", coef = "income_level"),
  prior(normal(1, 0.5), class = "b", coef = "education"),
  prior(normal(1, 0.05), class = "b", coef = "immigrantNotimmigrant"),

  prior(student_t(3, 0.05, 3), class = "sd", group = "country", coef = "sexMale"),
  prior(student_t(3, 0.05, 3), class = "sd", group = "country", coef = "education"),
  prior(student_t(3, 0.05, 3), class = "sd", group = "country", coef = "income_level")
)
```

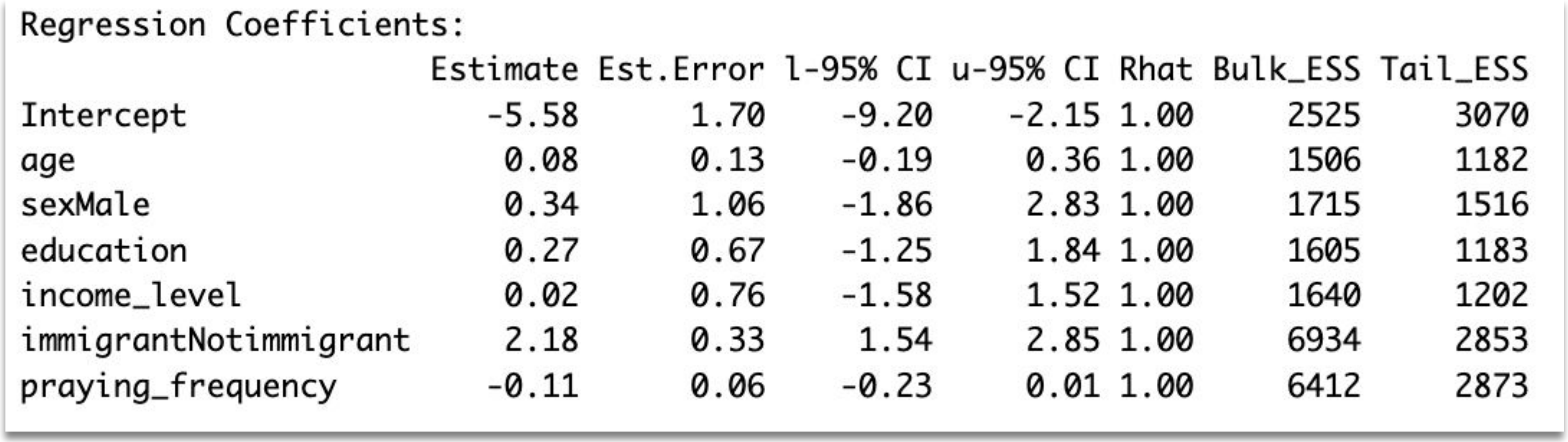
Fig 11 Strongly Informative prior specification

# Random Effects 2

```
voter_2 ~ age + sex + immigrant + income_level + education + praying_frequency + (1 +
sex + age + education + income_level | country)
```

Here we see that the significant coefficients have not changed - immigrant and praying_frequency!

Our choice of priors has not affected the output and the model is not very sensitive to priors.

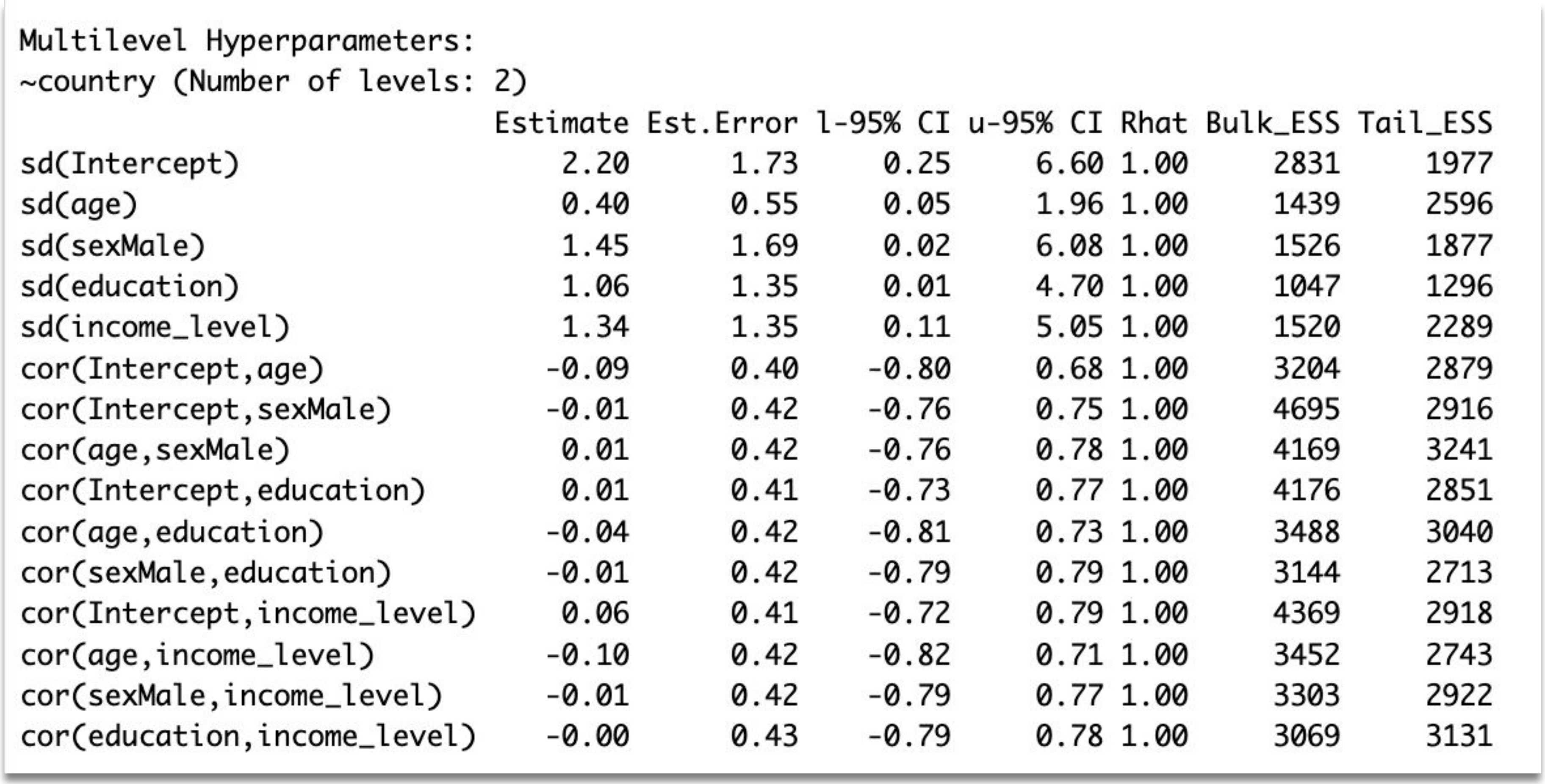


```
Regression Coefficients:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                -5.58      1.70    -9.20    -2.15 1.00     2525     3070
age                       0.08      0.13    -0.19     0.36 1.00     1506     1182
sexMale                   0.34      1.06    -1.86     2.83 1.00     1715     1516
education                 0.27      0.67    -1.25     1.84 1.00     1605     1183
income_level              0.02      0.76    -1.58     1.52 1.00     1640     1202
immigrantNotimmigrant     2.18      0.33     1.54     2.85 1.00     6934     2853
praying_frequency        -0.11      0.06    -0.23     0.01 1.00     6412     2873
```

Fig 12a. Model 3: Fixed effects summary

# Random Effects 2

It's the same story with the SD and correlation estimates!



```
Multilevel Hyperparameters:
~country (Number of levels: 2)
                           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                  2.20      1.73     0.25     6.60 1.00     2831     1977
sd(age)                        0.40      0.55     0.05     1.96 1.00     1439     2596
sd(sexMale)                    1.45      1.69     0.02     6.08 1.00     1526     1877
sd(education)                  1.06      1.35     0.01     4.70 1.00     1047     1296
sd(income_level)               1.34      1.35     0.11     5.05 1.00     1520     2289
cor(Intercept,age)            -0.09      0.40    -0.80     0.68 1.00     3204     2879
cor(Intercept,sexMale)        -0.01      0.42    -0.76     0.75 1.00     4695     2916
cor(age,sexMale)               0.01      0.42    -0.76     0.78 1.00     4169     3241
cor(Intercept,education)       0.01      0.41    -0.73     0.77 1.00     4176     2851
cor(age,education)            -0.04      0.42    -0.81     0.73 1.00     3488     3040
cor(sexMale,education)        -0.01      0.42    -0.79     0.79 1.00     3144     2713
cor(Intercept,income_level)    0.06      0.41    -0.72     0.79 1.00     4369     2918
cor(age,income_level)         -0.10      0.42    -0.82     0.71 1.00     3452     2743
cor(sexMale,income_level)     -0.01      0.42    -0.79     0.77 1.00     3303     2922
cor(education,income_level)   -0.00      0.43    -0.79     0.78 1.00     3069     3131
```

Fig 12b. Model 3: Random effects summary

# Model Comparison

Model3 performs the best as it has the highest Expected Log Predictive Density.

Additionally, since model2 and model3 are very similar, their elpd_diff is very small - indicating that they both have very similar predictive accuracy.

| | elpd_diff | se_diff |
|---|---|---|
| model3 | 0 | 0 |
| model2 | -0.8 | 2.6 |
| model1 | -45.7 | 10.8 |

Table 2: Model comparison using LOO

# Results

- Based on this analysis, we see that whether or not a person is an immigrant carries a lot of weight on whether they vote or not. Despite the fact that this data was only available for US, the results are significant - even if they cannot be extended directly to India.
- Age, income, education, sex also have some impact on voting chances, however these effects are quite small.
- Whether a person is satisfied, religious, ethical have little to no impact of deciding whether they vote or not.
- Despite having similar overall voter turnout, India and US have some variability across subgroups.

# Limitations and potential improvements

- Survey Data Limitations: Since the dataset is based on self-reported surveys, actual voter turnout might differ from responses. Access to official turnout records could enhance the accuracy and relevance of the findings.
- Modeling Voter Turnout as an Ordinal Variable: Treating voter turnout as an ordinal variable (e.g., "always votes," "sometimes votes," "never votes") could allow for the application of ordinal regression models, capturing nuanced patterns in voting behavior and improving interpretability.
- Temporal Analysis: Incorporating data from multiple years in the WVS dataset would enable the analysis of trends over time, offering insights into how voting patterns and their determinants evolve in different countries.

# Citations

[1] Voting chances by age:
https://usafacts.org/articles/how-many-americans-vote-and-how-do-voting-rates-vary-state/

[2] Election Commision of India - Voter turnout by gender:

https://elections24.eci.gov.in/docs/BnS4hhbvK9.pdf

[3] Good Authority - Is India unique in having higher voter turnout among the poor than the middle class and the rich?:
https://goodauthority.org/news/is-india-unique-in-having-higher-voter-turnout-among-the-poor-than-the-middle-class-and-rich/

# Summary | Take Home Message

Using Bayesian analysis of the WVS dataset, we modeled key factors influencing voting and explored tools like convergence diagnostics, caterpillar plots, and model comparison.

A critical takeaway: choosing priors carefully is essential, as they can shape results.

Lastly, remember that voting is vital for shaping our collective future—make your voice count!

Contact info:

Anirudh Parameswaran
anirudh.parameswaran@tu-dortmund.de

Gautam D Hariharan
gautam.hariharan@tu-dortmund.de