Voter Characteristics and Political Support in the US*

Tam Ly Renfrew Ao-Ieong Rahma Binth Mohammad

March 15, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

The United States (US) presidential election of 2020 resulted in President Joe Biden's victory, making him the 46th President of the US. The two top runners were Biden for the Democrats and former President Donald Trump for the Republicans. The presidential race between the two was of significant interest to statisticians and polling experts due to its potential impact on polling models and election forecasting. To provide some background, President Biden ran on his merit as a former Vice President of eight years and had faced lots of criticism during his time in office. Moreover, Former President Trump had controversial and career-damaging events. During the 2016 lections, Trump even faced Before the 2020 elections, Trump was charged with misconduct while in office(cite). However, the senate acquitted him of those charges. Thus, both Biden and Trump faced criticism from the electorate. An interesting question to consider is despite these, what characteristics did the voters have? Do they tell us a story on the type of people that voted for Biden and the type that voted for Trump? Answering these questions can help develop polling models and future election forecasting.

For our paper, we used a logistic regression model to estimate the likelihood of voting for Biden or Trump given certain characteristics. We use a binary outcome variable and three predictor variables. Our estimand is the probability that an individual voted for Biden or Trump based on three characteristics, their gender, race, and ownership of a gun. The reason for this choice will be further examined in Section 3. This data driven approach to elections can bring insight to voter characteristics and the importance of survey data and knowing the electorate.

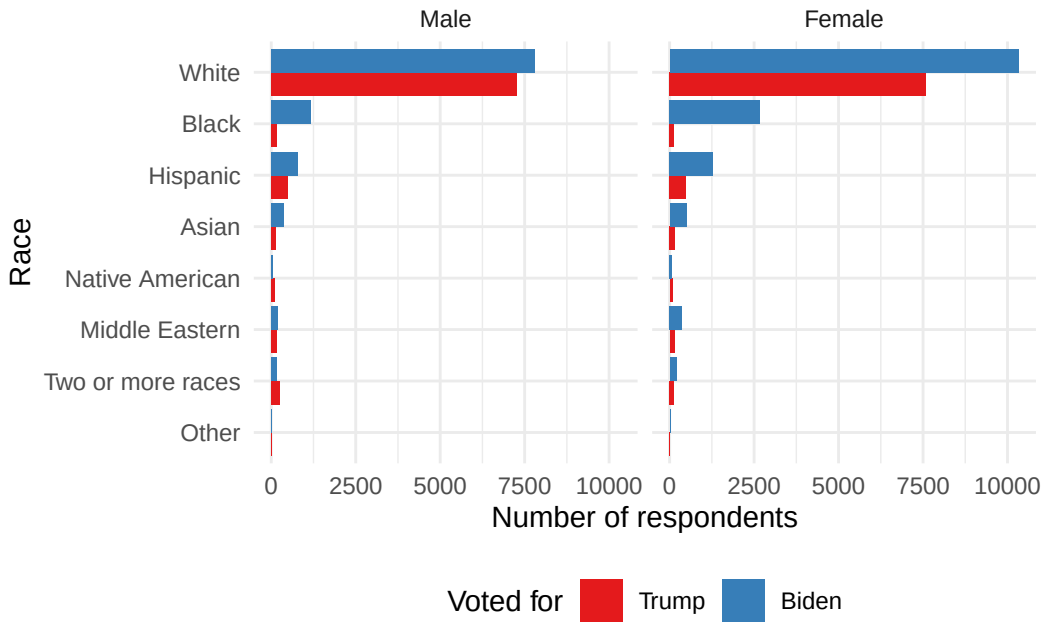
[insert more of what was done and what was found - overview of results section]

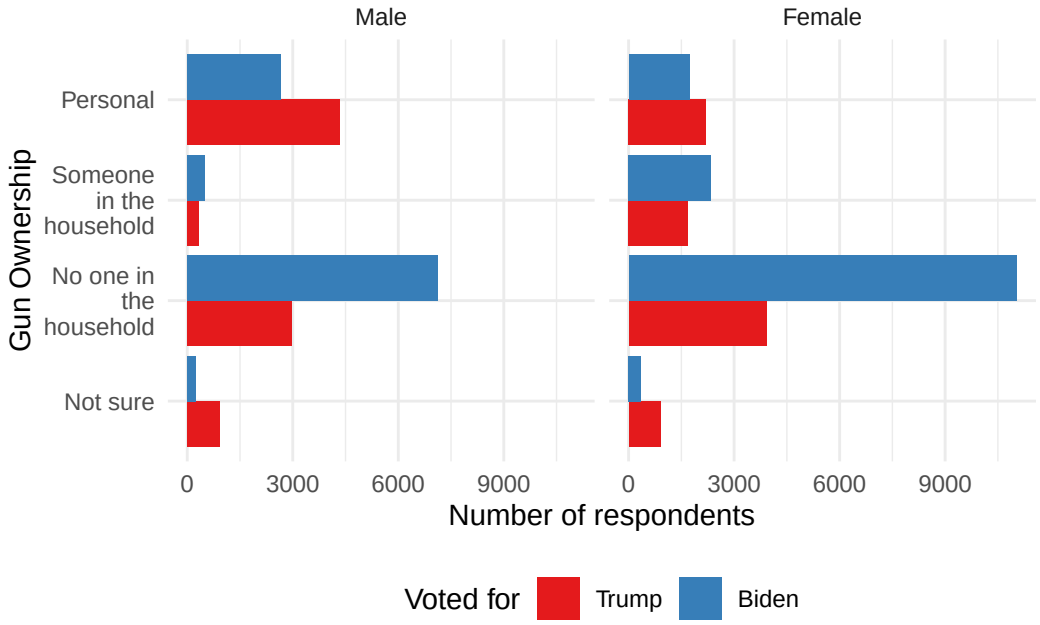
*Code and data are available at: https://github.com/atn-ly/political_support_in_the_us

The survey data for this paper is from the Harvard Dataverse Repository. The Cooperative Election Study is a sample survey that consists of pre-election and post-election questions. They provide a guide and a dataset for numerous years. It is also open to the public. [add more]

Data for this analysis and the different applications used will be further introduced in Section 2. Section 3 provides the model set-up and the justification for the use of that model. Section 4 will show results and Section 5 is a discussion of the results including the paper's weaknesses and biases.

2 Data





2.1 Sources and Collection

For this paper, we utilized the data from the 2020 Presidential Elections. We selected, gender, race, and whether an individual or a family member in the household owns a gun. We also selected the variable that recorded whether an individual voted for Biden or Trump. The broader context of the dataset is the political landscape of the US during the election cycle of 2020. The outcome of the 2020 election and every presidential election usually changes the direction of the country and demonstrates the perspectives of the majority of the electorate. There may have been similar dataset available, but we chose this one due to relevance to our research question. The questionnaire asked various questions on race, political party preference, voting preference, ownership of a gun and have numerous multiple choice answers available. Thus, this survey data provided us with the information on the demographic characteristics we wanted to consider.

Our paper aims to address the following two questions: (1) Does gender and race play a role in political preference? (2) Do those that own a gun favour one candidate over the other?

2.2 Methodology

2.3 Variables

2.4 Measurement

3 Model

3.1 Model set-up

The goal of our modeling strategy is to forecast if a person is likely to vote for Biden, based only on knowing their gender, race, and gun ownership status.

The model that we are interested in is:

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \alpha + \beta \times \text{gender}_i + \gamma \times \text{education}_i + \delta \times \text{gun}_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\delta \sim \text{Normal}(0, 2.5) \tag{6}$$

Where:

- y_i is the binary outcome variable, representing who respondent i voted for and equal to 1 if Biden and 0 if Trump,
- π_i is the probability that respondent i voted for Biden,
- $gender_i$ is a predictor variable, representing the gender of respondent i,
- $race_i$ is a predictor variable, representing the race of respondent i, and
- gun_i is a predictor variable, representing the gun ownership status of respondent i.

We used a logistic regression model in a Bayesian framework using the package `rstanarm` (Goodrich et al. 2022) in R (R Core Team 2023), which we will briefly describe here. Logistic regression is a type of generalized linear model. It is a tool for data exploration and used when we are interested in the relationship between a binary outcome variable and some predictor variables.

The foundation of logistic regression is the Bernoulli distribution and logit function. The Bernoulli distribution is a discrete probability distribution having two possible outcomes, "1" and "0", in which "1" occurs with probability p and "0" occurs with probability $1-p$. Logistic regression is still a linear model, because the predictor variables enter in a linear fashion

(**rohan?**). Hence, the logit function links the Bernoulli distribution to the machinery we use in linear models (**rohan?**).

In our model, we also have the parameters α, β, γ, and δ in addition to the variables. The parameter α is the intercept and the parameters, β, γ, and δ, are the slope coefficients. We specify prior probability distributions for each of the parameters in our model. However, these are just the default priors that `rstanarm` (Goodrich et al. 2022) uses, which are Normal distributions with a mean and standard deviation of 0 and 2.5, respectively.

3.2 Model justification

Given that Biden and Trump are far apart on women's issues, racial justice, and gun policy, we chose *gender*, *race*, and *gunown* to be the predictor variables in our model. First, since we are interested in forecasting who a respondent is likely to vote for, we used *gender* instead of *gender_post* because it is representative of all adult Americans rather than only adult Americans who are registered to vote. Second, we used *race* instead of variables by country or region to affirm that race is a social construct with no biological foundation. Third, we used *gunown* instead of variables about gun control because it captures broader opinions rather than specific policy proposals. After exploring the data in the previous section, we found differences in political preference based on these features and that it would be of interest to investigate further.

Logistic regression does not make the same assumptions as linear regression. First, linear regression assumes a continuous outcome variable that can take any number on the real line, whereas logistic regression assumes a binary outcome variable. Furthermore, linear regression requires that the outcome variable is a linear function of the predictor variables, while the outcome in logistic regression is part of the exponential family. However, logistic regression does assume that the relationship between the log-odds of the binary outcome and predictor variables is linear. Lastly, unlike in linear regression, logistic regression does not require the assumption of homoscedasticity of errors.

Alternative regression models were considered, but rejected because they were not appropriate for our outcome variable. Linear models assume a continuous outcome variable, and Poisson and Negative binomial models assume count outcome variables. Since our outcome variable is binary, we chose to use logistic regression.

To show that our model does a good job of fitting the data, we consider the posterior distribution and implement posterior predictive checks. Details and graphs can be found in Section B. Furthermore, we check if the Markov Chain Monte Carlo (MCMC) sampling algorithm that `rstanarm` uses to obtain samples from the posterior distributions of interest ran into any issues. Details and graphs can be found in Section B.

4 Results

We performed logistic regression analysis on 5000 observations; a subset of the total 43240 observations from our cleaned dataset. Since we are interested in which presidential candidate an individual voted for based on the categories explained earlier, our model made `genderMale`, `raceWhite`, `gun_ownershipPersonal` into reference groups for their respective categories.

Our results from the model given by Table 1 shows the coefficient of each indicator variable. This coefficient represents the log of the expected difference in support for Biden compared to Trump. When this value is negative, it shows a decrease in support for Biden, when this value is 0 it shows neither a decrease nor increase in support for Biden. When this value is positive, it shows an increase in support for Biden.

The intercept is -0.748 which means that the support for Biden decreases by about 0.784 units when all other variables are held constant. By holding all variables constant, we have an individual who is Male, White, and personally owns a gun.

4.1 Gender

We can see that for the variable `genderFemale`, we have a coefficient of 0.235 indicating an increase in support for Biden over the reference group being `genderMale`, while keeping the other variables constant.

4.2 Race

We can see that Black individuals show a much stronger support for Biden as the coefficient for `raceBlack` is 2.464. Asian, Hispanic, and Middle Eastern individuals all show an increase in support for Biden with coefficients of 0.790, 0.548, and 0.288 respectively. Native Americans show a decrease in support for Biden with a coefficient of -0.536 while individuals of two or more races show neither an increase nor decrease in support for Biden with a coefficient of -0.008.

4.3 Gun Ownership

For gun ownership, we can see that if an individual does not personally own a gun but someone in their household does, the coefficient is 0.870 which shows an increase in support for Biden. If nobody in an individual's household owns a gun, the coefficient is 1.332 showing a strong support for Biden. When an individual does not own a gun and is not sure if anyone in their household owns a gun, the coefficient is -0.815 showing a decrease in support for Biden.

Table 1: Explanatory models of political preferences based on gender, race, and gun ownership (n = 5000)

	Support Biden
(Intercept)	−0.748
	(0.067)
genderFemale	0.235
	(0.067)
raceBlack	2.464
	(0.184)
raceHispanic	0.548
	(0.122)
raceAsian	0.790
	(0.203)
raceNative American	−0.536
	(0.403)
raceMiddle Eastern	0.288
	(0.218)
raceTwo or more races	−0.008
	(0.252)
raceOther	−0.333
	(0.905)
gun_ownershipSomeone in the household	0.870
	(0.112)
gun_ownershipNo one in the household	1.332
	(0.072)
gun_ownershipNot sure	−0.815
	(0.157)
Num.Obs.	5000
R2	0.175
Log.Lik.	−2898.226
ELPD	−2910.6
ELPD s.e.	30.1
LOOIC	5821.1
LOOIC s.e.	60.2
WAIC	5821.0
RMSE	0.45

Table 2: Probability that an individual supports Biden given their demographic

rowid	estimate	conf.low	conf.high	voted_for	gender	race	gun_ownership
1	0.7511584	0.6631003	0.8273150	Biden	Female	Middle Eastern	No one in the household
2	0.7564311	0.7087319	0.7988593	Trump	Male	Hispanic	No one in the household
3	0.3741561	0.3412633	0.4101036	Biden	Female	White	Personal
4	0.6941046	0.6708337	0.7161842	Trump	Female	White	No one in the household
5	0.5875198	0.5431017	0.6310208	Trump	Female	White	Someone in the household
6	0.7971713	0.7557982	0.8334727	Biden	Female	Hispanic	No one in the household

From the model, we used `predictions()` from `marginaleffects` (Arel-Bundock 2023) to obtain the implied probability that an individual supports Biden. A sample of the result is given in Table 2 where each individual is given an `estimate` value which is the estimated probability that the individual supports Biden.

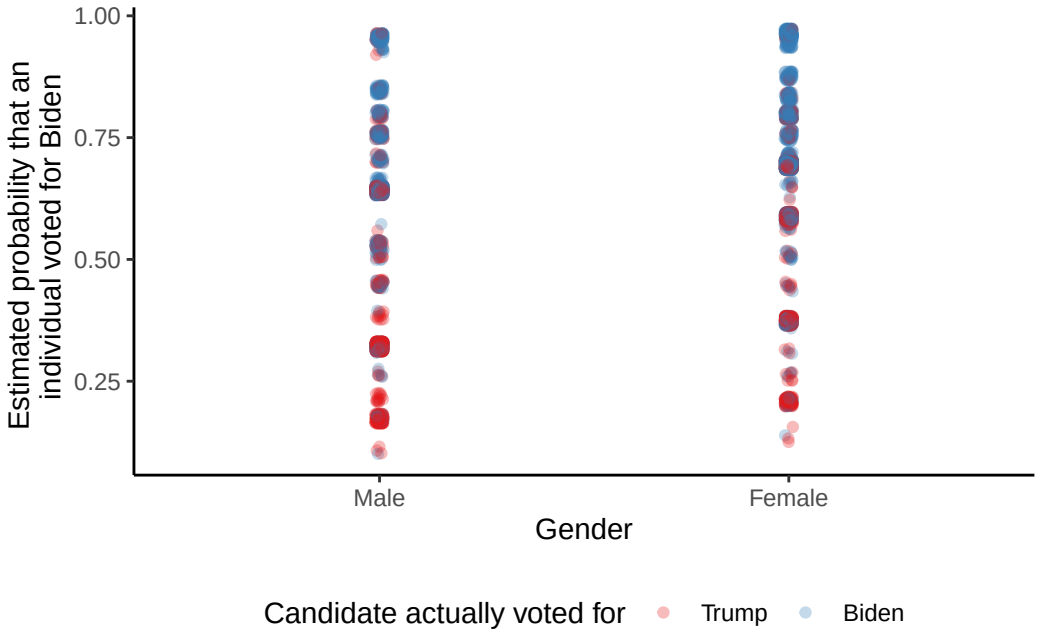


Figure 1: Predicted support of Biden given gender vs. actual support

Using the results in Table 2, we plotted three graphs using `ggplot2` (Wickham 2016) which

shows the estimated probability that an individual supports Biden given gender, race, and gun ownership with the colour of the point indicating their true support. Figure 1 shows the estimated probability that an individual supports Biden given gender. To support the accuracy of our model, blue dots should be higher up on the graph while red dots should be lower down indicating that our predicted probability matches with the actual result. We can see in Figure 1 that this trend holds true.

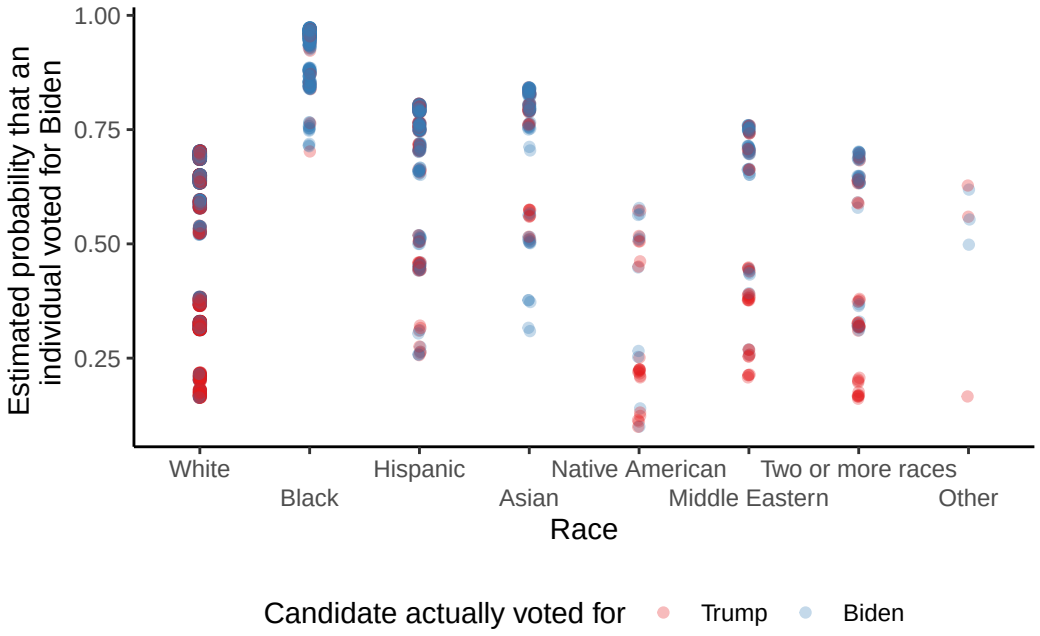


Figure 2: Predicted support of Biden given race vs. actual support

Figure 2 shows the estimated probability that an individual supports Biden given race. There appears to be red dots near the top for White, Native American, Middle Eastern, and Two or more races which indicates possible inaccuracy in our model. However, Black, Hispanic and Asian appears to follow the expected distribution of the dots.

Figure 3 shows the estimated probability that an individual supports Biden given their gun ownership status. We can see that Personal gun ownership and the Not Sure if anyone in the household owns a gun categories show a lower support of Biden, while the No One in the household and the Someone else in the household owns a gun categories show a higher trend of support for Biden. There appears to be an expected cluster of blue dots higher up and red dots lower down, which supports the accuracy of our model.

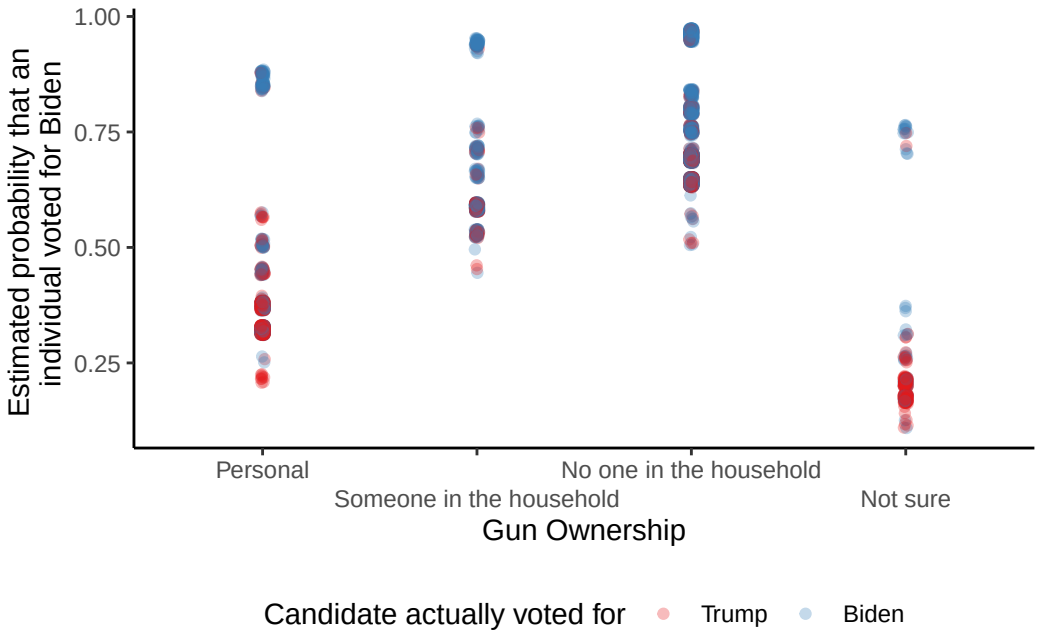


Figure 3: Predicted support of Biden given gun ownership vs. actual support

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Appendix

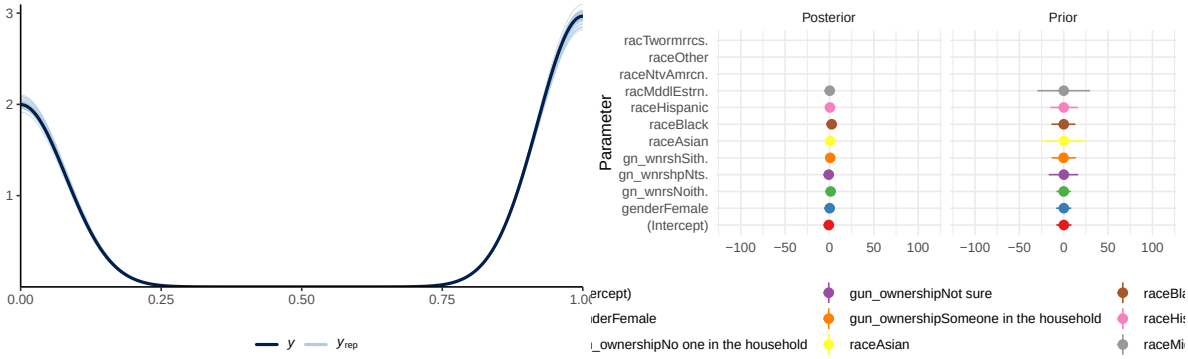
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check. It shows a very close match between the actual outcome variable with simulations from the posterior distribution. This suggests that our model does a good job of fitting the data.

In Figure 4b we compare the posterior with the prior. It shows that estimates changed minimally once the data was taken into account. This suggests that we specified good priors.



(a) Posterior prediction check

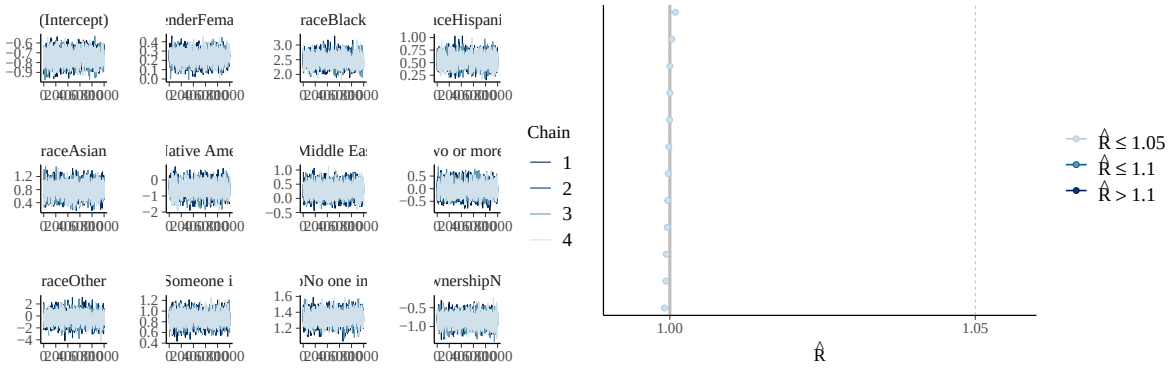
(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot. It shows that there are no horizontal lines that appear to bounce around and have a nice overlap between the chains. This does not suggest anything out of the ordinary.

Figure 5b is a Rhat plot. It shows that everything is close to 1 and no more than 1.1. This does not suggest any problems.



(a) Trace plot (b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2023. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.