Forecasting the 2024 U.S. Presidential Election*

President Joe Biden Projected to win the Popular Vote Based on MRP Analysis

Talia Fabregas

April 18, 2024

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

1 Introduction

Every four years, American adults head to the polls to elect their president. The 2024 United States presidential election will take place on Tuesday November 5, 2024. Amidst unprecedented levels of political polarization and distrust in democratic institutions, America will see a rematch of the 2020 election. President Joe Biden will seek a second term and former president Donald Trump will try to become the second president to serve two non-consecutive terms.

The survey data set that I used was provided by the 2022 Cooperative Election Study (CES) and downloaded from Harvard Dataverse. The CES is a nationally representative survey of 60,000 American adults, conducted before and after U.S. presidential and midterm elections (Schaffner, Ansolabehere, and Shih 2023). It aims to study the voting behavior of American adults and how it is influenced by geographic and demographic factors (Schaffner, Ansolabehere, and Shih 2023). I selected a subset of the final release of the 2022 CES Common Content Dataset to use as my survey data. The post-stratification data that I used was downloaded from the Integrated Public Use Microdata Series (IPUMS) USA online database. IPUMS provides survey and census data, dating back to 1850, with the help of 105 statistical organizations (Ruggles et al. 2024). I selected a subset of the 2022 American Communities Survey (ACS). I use multi-level regression with post-stratification (MRP) to use as my post-stratification data. I performed MRP analysis to forecast the 2024 U.S. presidential election. This involves using a smaller survey dataset (~10,000 respondents) to fit a model to predict vote preference based on geographic and demographic characteristics and then applying it to a larger post-stratification dataset (~500,000 respondents). The model will learn how to classify respondents as Trump or Biden voters using the survey dataset. It will then use what

*Code and data are available at: https://github.com/taliafabs/us-election-analysis.git it has learned to classify ACS respondents as Trump or Biden voters when applied to the post-stratification dataset. I will use these results to predict the popular vote and electoral college results of the 2024 U.S. presidential election.

I used R programming language (R Core Team 2023) and the `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot` (**citeggplot2?**), `knitr` (**citeknitr?**), `readr` (Wickham, Hester, and Bryan 2023), `arrow` (Richardson et al. 2023), and `rstanarm` (**citerstanarm?**) packages to clean my survey and post-stratification datasets, create my data visualizations, fit my logistic regression model, and apply my logistic regression model.

2 Data

2.1 Survey Data

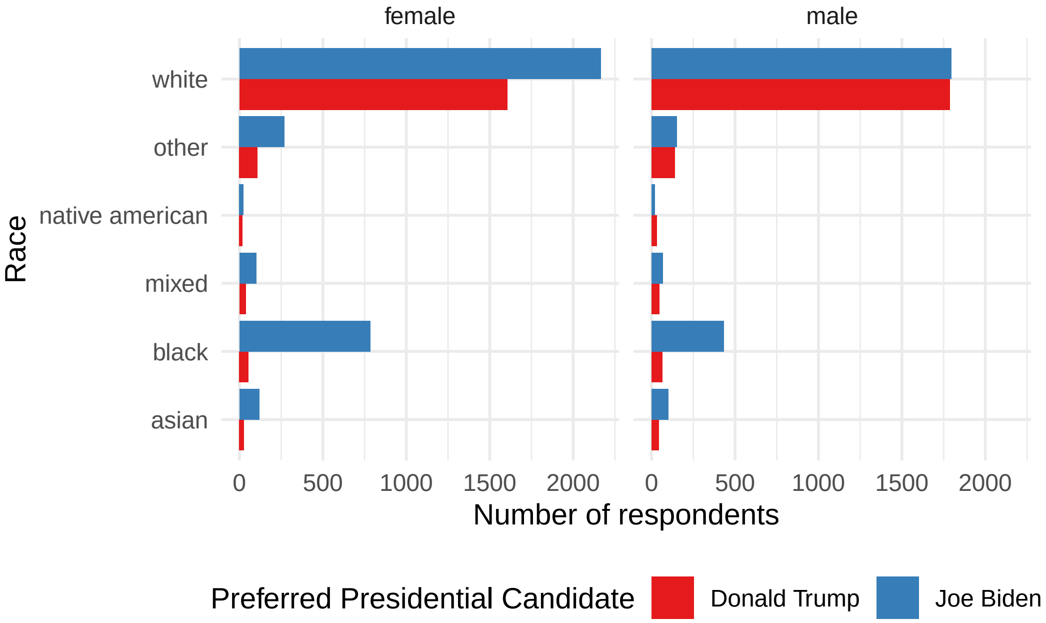


Figure 1: Preferred presidential candidates of survey subset respondents, by gender and race

Figure 16 illustrates the proportion of subsetted survey respondents in each state who plan to support President Biden in the 2024 election. My subsetted survey dataset appears to show stronger support for President Joe Biden than the genral U.S. electorate both overall and at the state level. As shown in Section B, this is not unique to the subsetted data; the complete 2022 CES Common Content Dataset shows strong support for President Biden and the Democratic Party based on 2016 vote choice (`presvote16post`), 2020 vote choice (`presvote20post`), and party identification (`pid3`). However, as seen in subset-state2020_subset, respondents who live in states won by Biden in 2020 (blue states) were more likely to support him over Trump

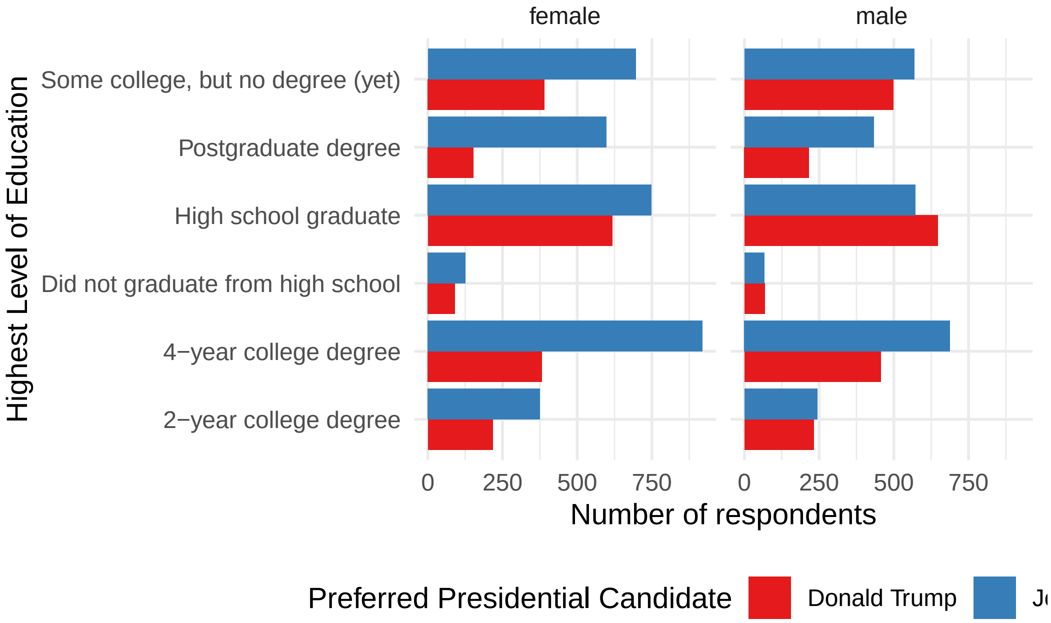


Figure 2: Preferred presidential candidates of survey subset respondents, by highest level of education

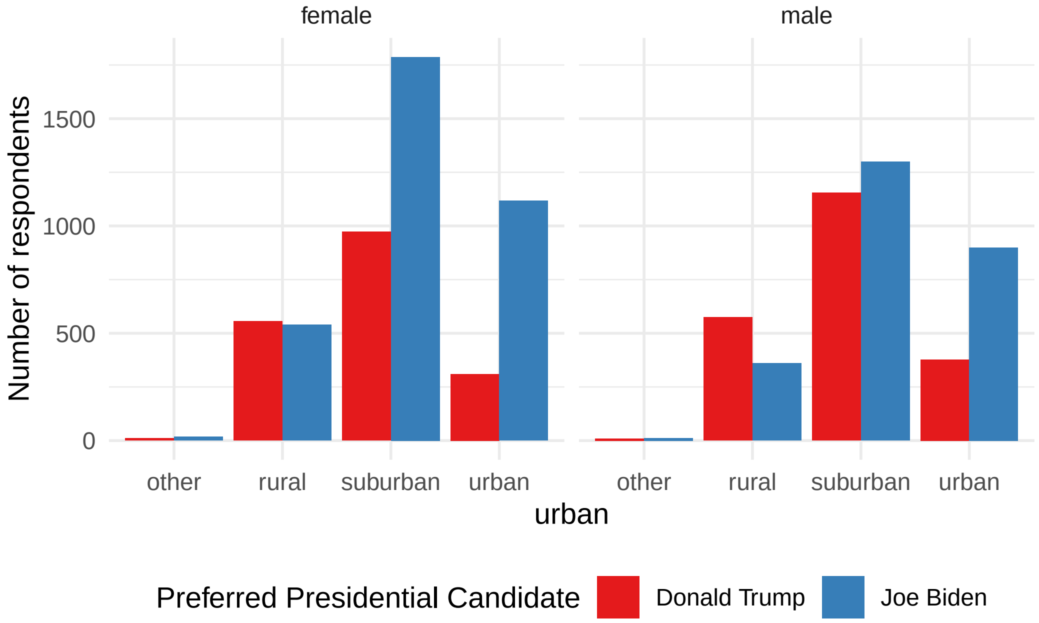


Figure 3: Preferred presidential candidate of subset respondents living in urban vs rural areas

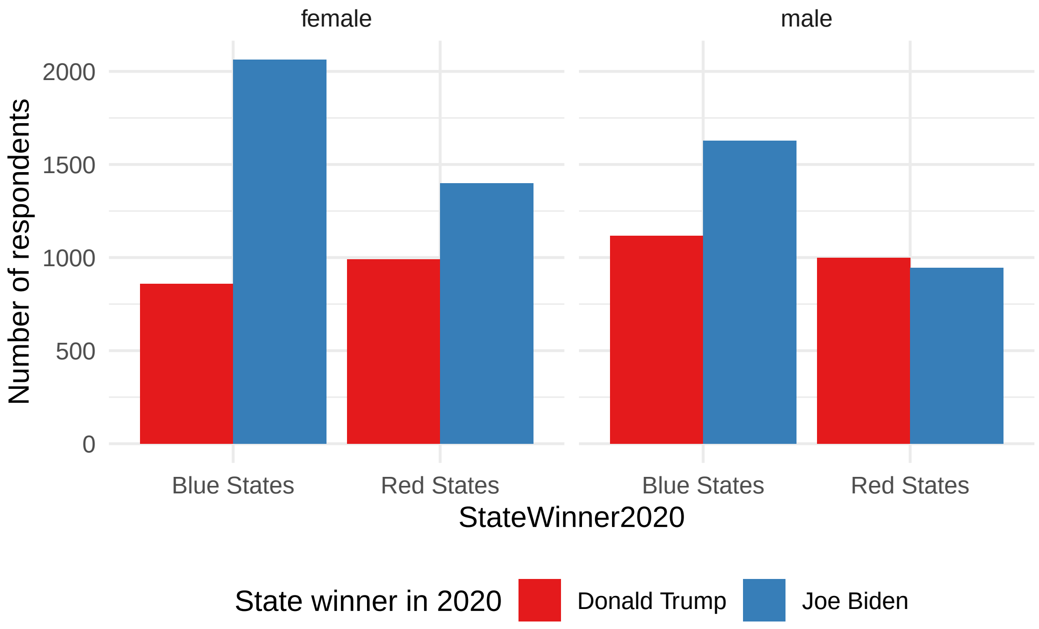


Figure 4: Preferred presidential candidate of subset respondents in states won by Trump vs Biden in 2020

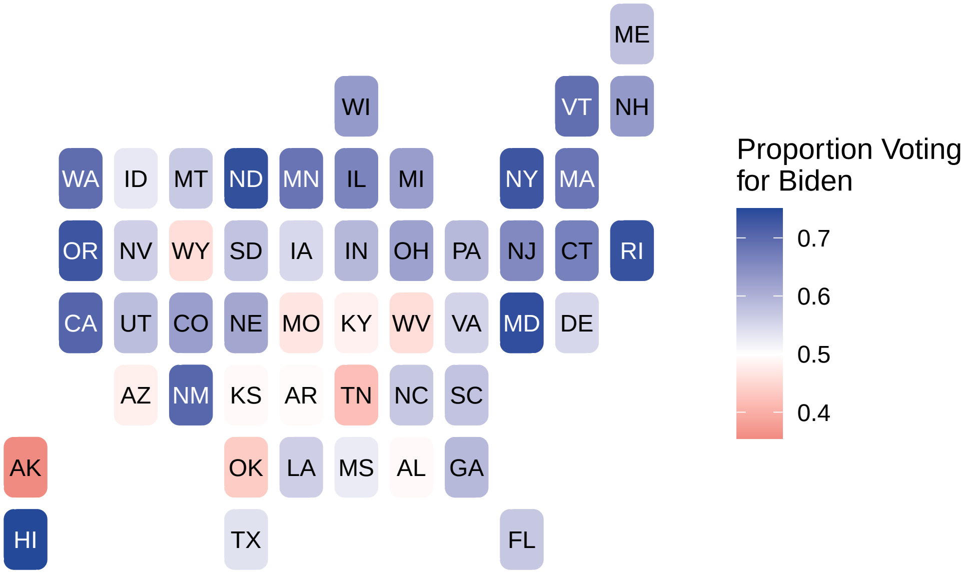


Figure 5: Electoral college map based on the subsetted survey data

Table 1: Popular vote and electoral college based on subset survey data

Survey Estimate:	Biden	Trump
Num Votes	6033.00	3967.00
% Votes	60.33	39.67
Electoral College	460.00	78.00

in 2024 than those who live in states won by former president Trump in 2020 (red states). To account for this difference, and the fact that the survey dataset I used has strong support for President Biden, I created the `biden_won` variable, which is equal to 1 if a state was carried by President Biden in 2020 and 0 if it was carried by former President Trump in 2020. The data cleaning steps that I used to create the `biden_won` variable are outlined in Section A.

2.2 Poststratification Data

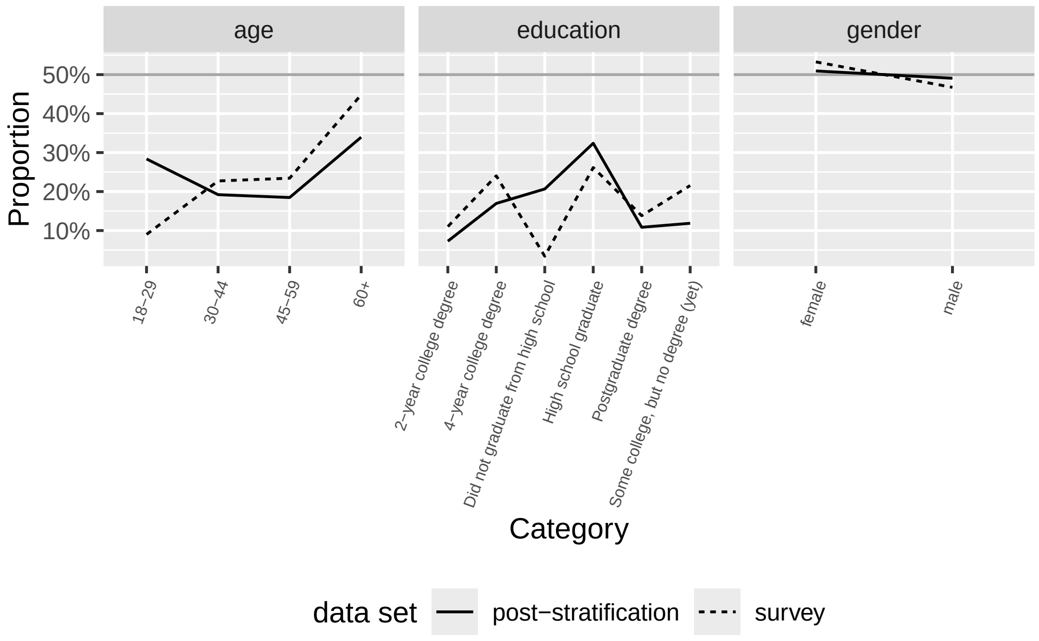


Figure 6: Survey vs post-stratification voter demographics

3 Model

I performed multi-level regression with post-stratification (MRP) to predict support for president Joe Biden and former president Donald Trump in the 2024 U.S. presidential election.

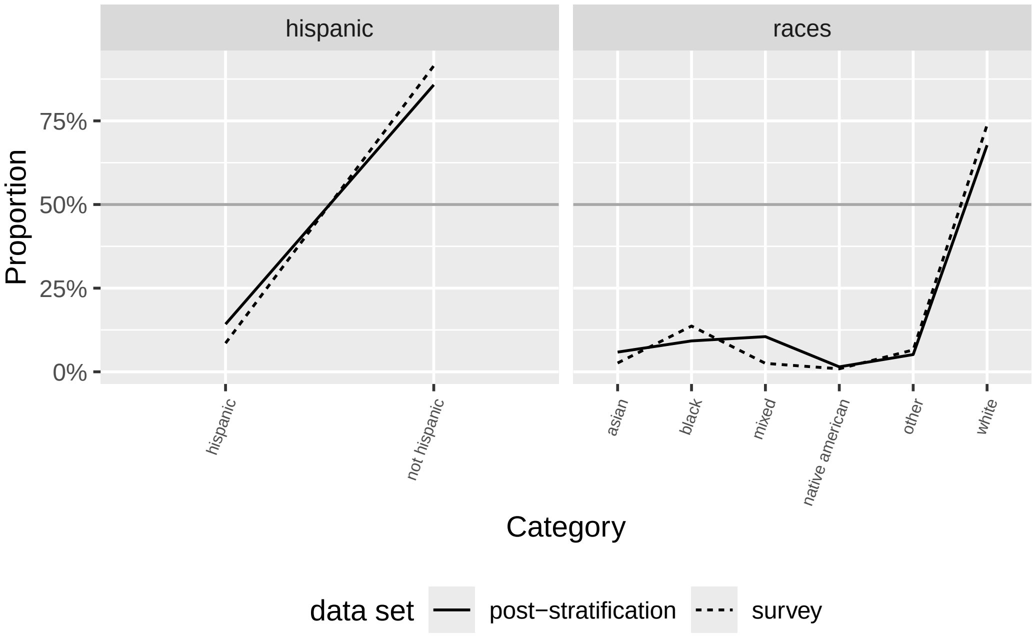


Figure 7: Survey vs post-stratification voter race demographics

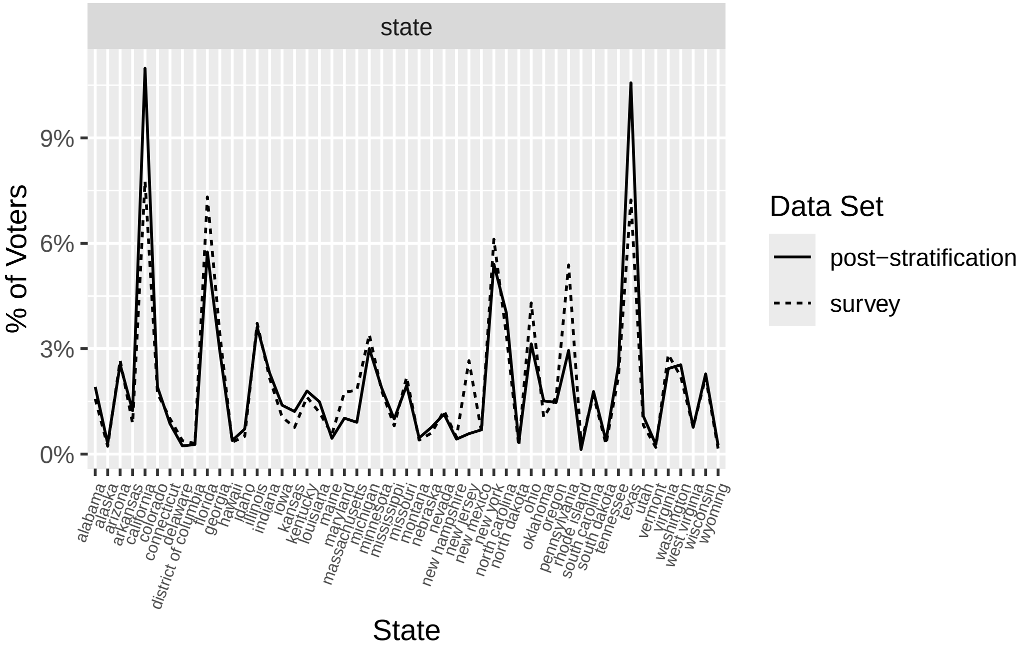


Figure 8: Survey and Post-Stratification Data Proportion of Voters by State

To perform MRP analysis, I fit the model on the survey data and applied it to the post-stratification data. Fitting the logistic regression model on the survey data teaches it to classify each respondent as a Biden or Trump voter based on the state that they live in, whether Biden or Trump won that state in 2020, their sex, age bracket, race, highest level of education, and whether they live in an urban area. I then apply the model to my post-stratification dataset (Ruggles et al. 2024) to forecast the popular vote and electoral college results for the 2024 U.S. presidential election. When applied to my post-stratification dataset, the logistic regression model uses the same variables (state, whether Biden or Trump won that state in 2020, sex, age bracket, race, highest level of education, and urban) and what it learned from being fit on the survey dataset to classify each census respondent as a Biden or Trump voter.

3.1 Model set-up

I built my Bayesian Logistic Regression model in R (R Core Team 2023) using the `stan_glm` function and the default priors of the `rstanarm` package (Goodrich et al. 2022). My model is as follows:

$$
\begin{aligned}
vote\_biden_i | \pi_i &\sim \mathrm{Bern}(\pi_i) \\
\mathrm{logit}(\pi_i) &= \beta_0 + \beta_1 \mathrm{state}_i + \beta_2 \mathrm{biden\_won}_i + \beta_3 \mathrm{sex}_i + \beta_4 \mathrm{age\_bracket}_i \\
&\quad + \beta_5 \mathrm{race}_i + \beta_6 \mathrm{hispanic}_i + \beta_7 \mathrm{educ}_i + \beta_8 \mathrm{urban}_i \\
\beta_0 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_1 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_2 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_3 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_4 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_5 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_6 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_7 &\sim \mathrm{Normal}(0, 2.5) \\
\beta_8 &\sim \mathrm{Normal}(0, 2.5)
\end{aligned}
$$

where the binary indicator variable vote_biden_{i} is equal to 1 if the respondent's preferred 2024 presidential candidate is President Joe Biden (D) , or 0 if their preferred candidate is former President Donald Trump (R). My model uses logistic regression, it is not without tradeoffs. Firstly, logistic regression can only be used to predict a binary outcome variable. As far as my model is concerned, the only two vote choices for U.S. adults in the upcoming presidential election are Joe Biden and Donald Trump. The possibilities of voting third-party or not voting at all are not considered. I discuss the tradeoffs, benefits, weaknesses, and limitations associated with the use of a logistic regression model to forecast the U.S. election in more depth in Section 5.4.

3.1.1 Model justification

4 Results

4.1 Popular Vote Prediction

I got my popular vote prediction by applying the model outlined in Section 3.1 to my post-stratification data set to predict the 2024 preferred presidential candidate of each ACS 2022 respondent (Ruggles et al. 2024). Table 2

Table 2: 2024 U.S. election popular vote estimates based on post-stratification analysis

Estimate:	Biden %	Trump %
Lower Estimate	48.55	51.45
Mean Estimate	55.47	44.53
Upper Estimate	62.56	37.44

4.2 Electoral College Prediction

Table 3: 2024 U.S. election electoral college estimates based on multilevel regression with post-stratification (MRP) analysis

Electoral College Estimate:	Biden	Trump
Lower Estimate	220	318
Mean Estimate	413	125
Upper Estimate	517	21

5 Discussion

5.1 Popular Vote Prediction

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this. The survey dataset appears to favor President Joe Biden more than the general U.S. electorate.

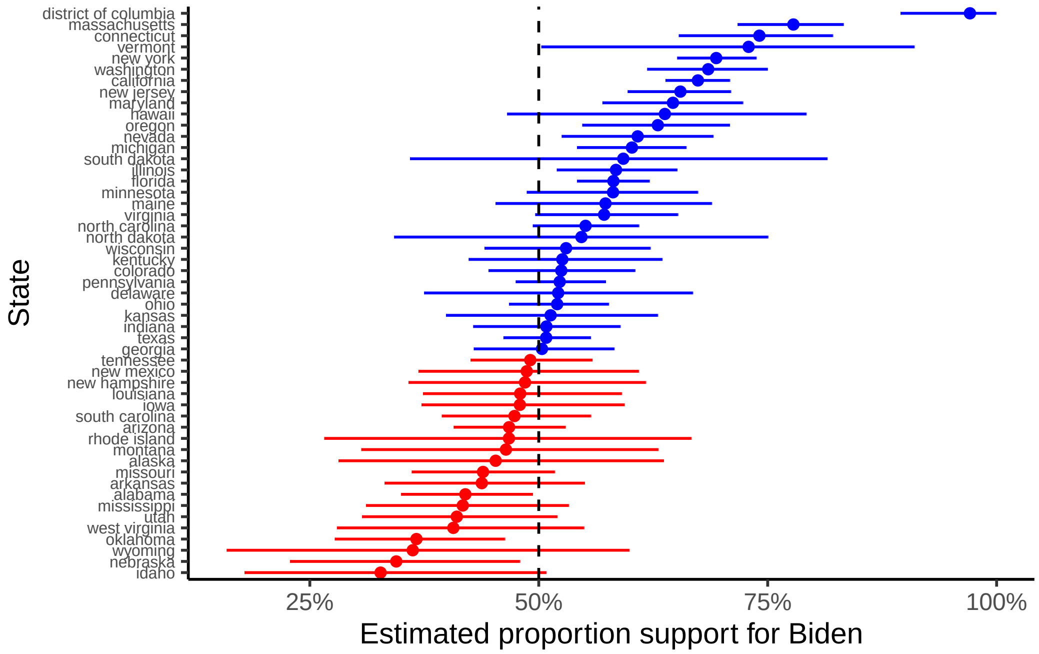


Figure 9: Estimated proportion of each state voting for Biden in 2024 based on MRP analysis

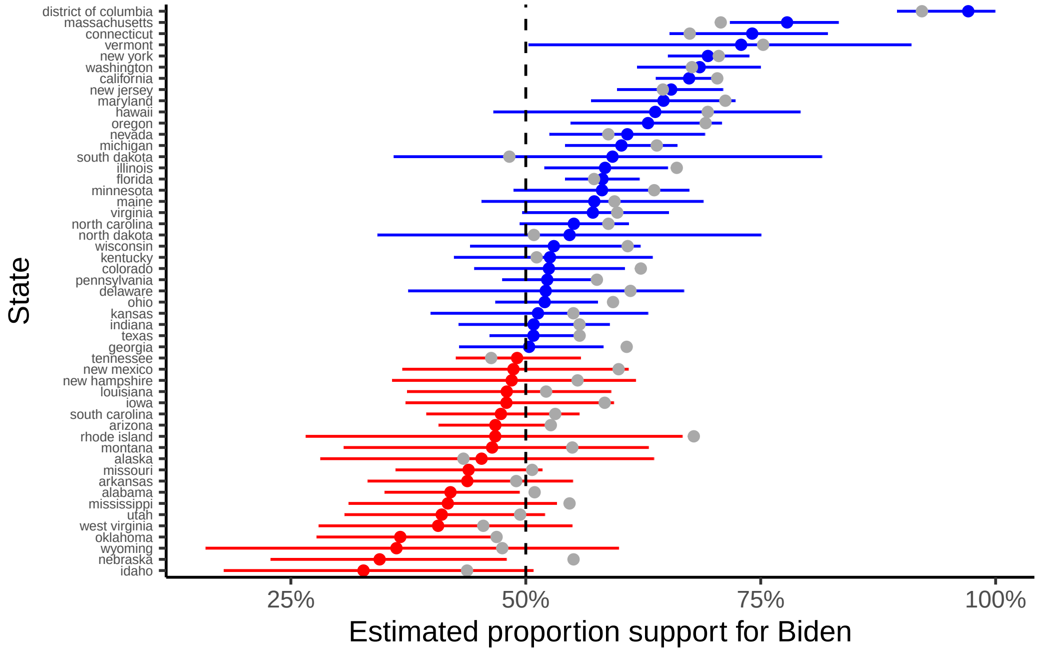


Figure 10: Estimated proportion of each state voting for Biden in 2024 Post-Stratification vs Subsetted Survey Data

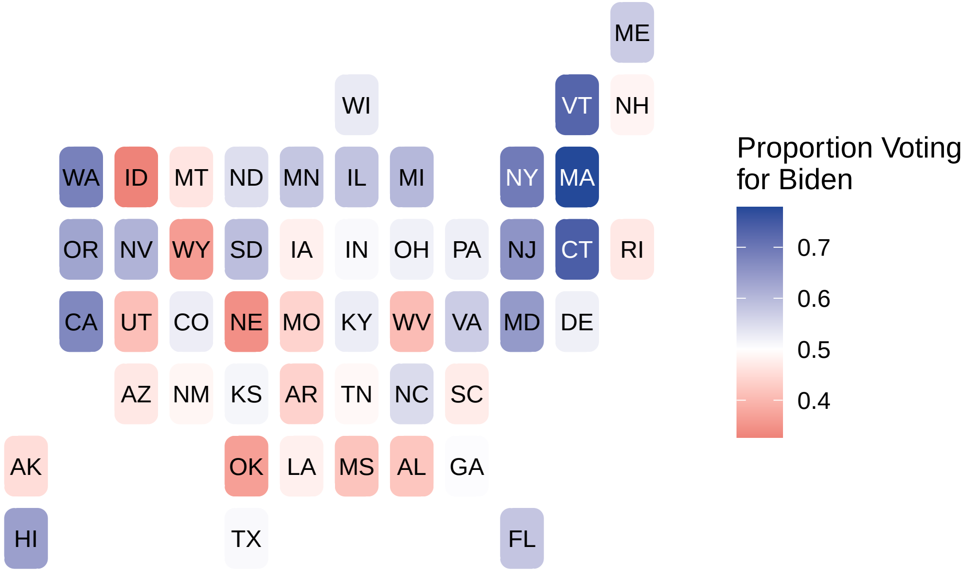


Figure 11: Electoral map based on MRP analysis

5.2 Swing States had Close Margins and Large Error Ranges

Swing states and error ranges in the MRP analysis electoral college prediction

5.3 Weaknesses and Limitations of the Datasets

5.4 The Limitations of Logistic Regression and the Case for (and against) SoftMax Regression

5.5 Next Steps

Python, more recent data set, softmax regression, gradient descent Split the survey data into training, validation, and test Use gradient descent to find the optimal weights to maximize validation accuracy Apply the model to the post-stratification data Softmax regression does risk overfitting

Appendix

A Additional data cleaning details

I created a binary variable, `vote_biden` that is equal to 1 if a respondent's preferred 2024 presidential candidate is Joe Biden, or 0 if it is Donald Trump. My Cooperative Election Study Common Content survey dataset was put together in 2022, so respondents were not asked about their preferred 2024 presidential candidate. However, they were asked about their party identification (`pid3`), who they voted for in the 2016 presidential election (`presvote16post`), and who they voted for in the 2020 presidential election (`presvote20post`). I started by filtering out respondents who have no party affiliation and did not vote for either major party nominee in the 2016 and 2020 presidential elections. This was done because I am using a logistic regression model, which is only capable of performing binary classification. I used this information to create the `vote_biden` indicator variable; if a respondent's `pid3` is Democratic, or they voted for the Democratic nominee in 2016 or 2020, I label them as a Biden voter (`vote_biden = 1`). Otherwise, I label them as a Trump voter (`vote_biden = 0`). As previously stated, respondents who both voted third-party and have no party affiliation were not considered because my model is not capable of performing multi-class classification and there is no indication of whether they prefer Joe Biden or Donald Trump.

I downloaded the 2020 National Popular Vote Tracker from (**cook?**), cleaned it to change the state column to match my survey and post-stratification data sets, and selected the `state`, `biden_won`, `dem_votes`, `rep_votes`, `other_votes`, and `dem_percent` columns. I then left-joined the 2020 National Popular Vote Tracker with both my survey and post-stratification datasets so that I could include Biden's performance in each state in the 2020 election to my model as a predictor for `vote_biden`. I added the binary variable, `biden_won`, to both my survey and post-stratification analysis datasets. `biden_won` is equal to 1 if Joe Biden won the electoral college votes of the state that the respondent lives in in the 2020 presidential election, and 0 if Donald Trump won the state in the 2020 presidential election. Maine and Nebraska have split electoral college votes; this means that the presidential nominee who wins each congressional district receives an electoral college vote, and an additional 2 electoral college votes are awarded for winning the statewide popular vote. In the 2020 presidential election, Joe Biden won the statewide popular vote in Maine, while Donald Trump won the statewide popular vote in Nebraska, although Biden won one congressional district in Nebraska, and Trump won one congressional district in Maine. `biden\_won` corresponds to the presidential nominee who won the statewide popular vote in the 2020 election, so`biden\_won = 1` for respondents who live in Maine, and `biden\_won = 0` for respondents who live in Nebraska. I use informatoin from the 2020 National Popular Vote tracker in various places throughout my analysis and discussion.

B Additional survey data details

The original survey dataset contained 60,000 responses, but it was subsetted to 10,000 so that R and `rstanarm` could handle it (R Core Team 2023). The `glm` function of the `rstanarm` package was used to fit the logistic regression model to predict 2024 presidential vote choice based on state, whether Biden or Trump won that state in 2020, sex, age bracket, race, and highest level of education completed. Although Schaffner, Ansolabehere, and Shih (2023) advised against subsetting the 2022 CES Common Content Dataset, this was a necessary step for me. When I tried to fit the model using the original survey data set, it took hours to run and I was not able to post-stratify due to the following error: `Error: vector memory exhausted (limit reached?)`. 10,000 of the 60,000 responses were randomly selected using the `sample` function of base R. The visualizations below show the results of the exploratory data analysis conducted on the original survey data set. As seen in Figure 12, Figure 13, Figure 14, and Figure 16, the results are similar to the ones shown in Section 2.1. Therefore, I am confident that my random subset is representative of the original 2022 CES Common Content Dataset (Schaffner, Ansolabehere, and Shih 2023).

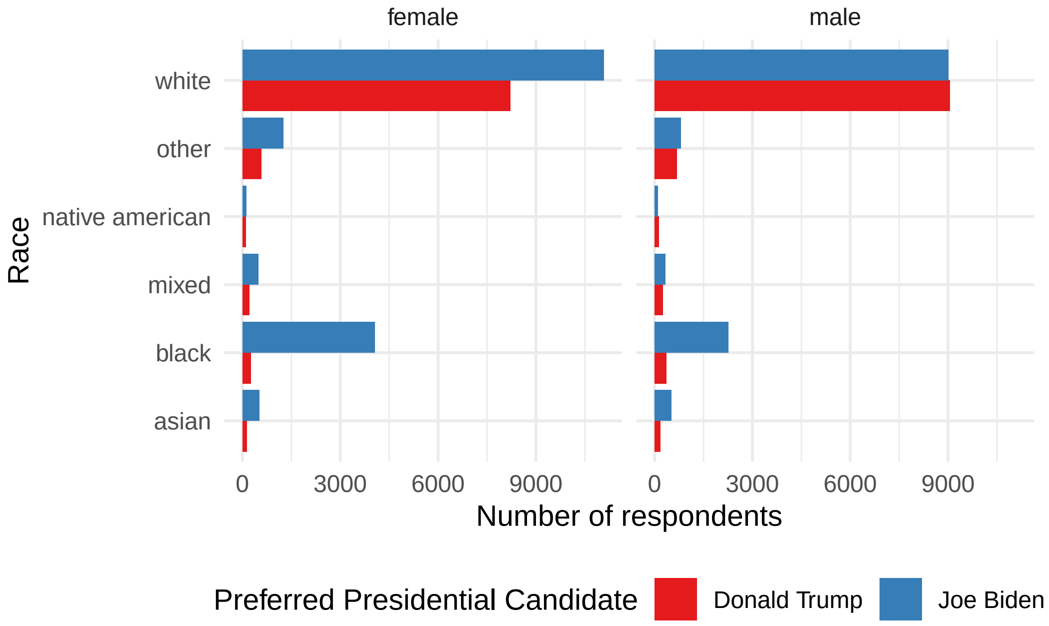


Figure 12: Preferred presidential candidates of survey respondents, by gender and race

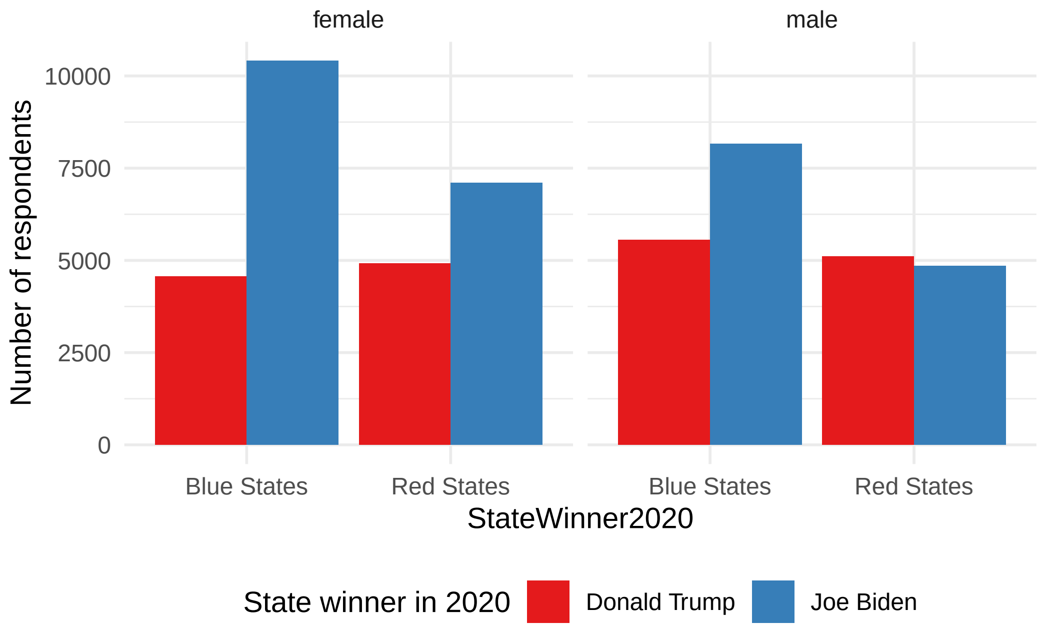


Figure 13: Preferred presidential candidate of survey respondents in states carried by Trump vs Biden in 2020

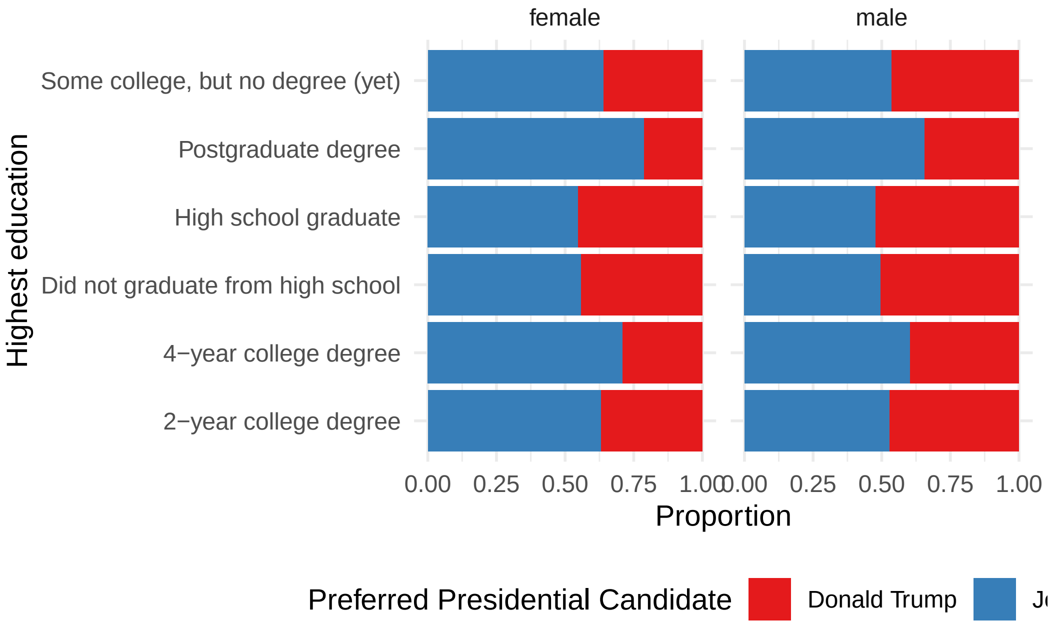


Figure 14: Preferred presidential candidates of survey respondents, by highest level of education

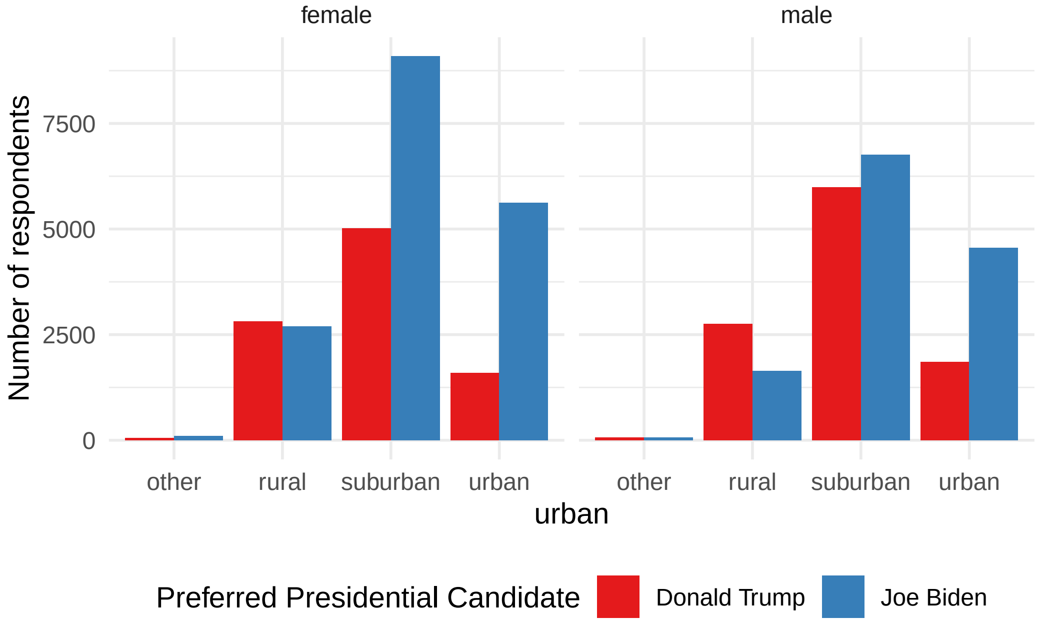


Figure 15: Preferred presidential candidate of survey respondents living in urban vs rural areas

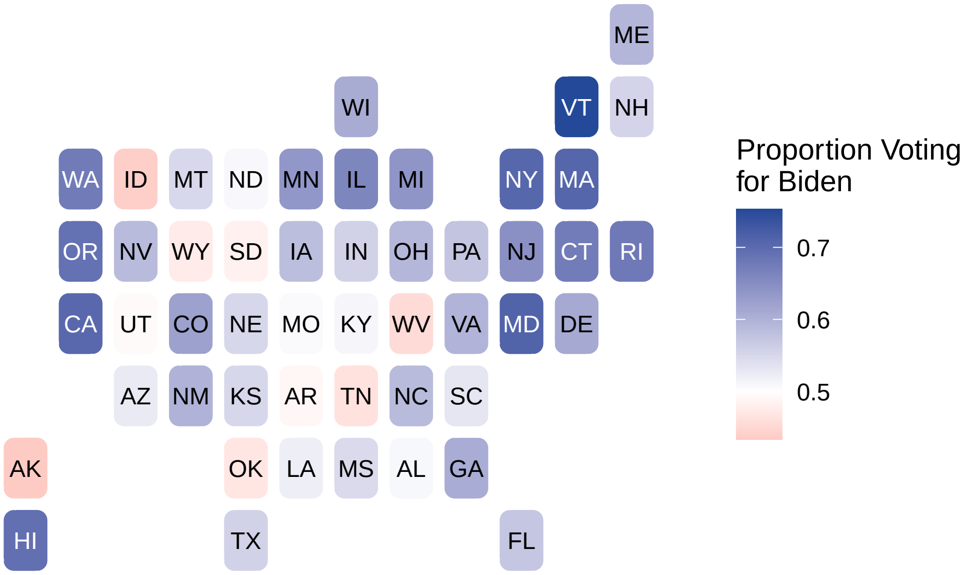
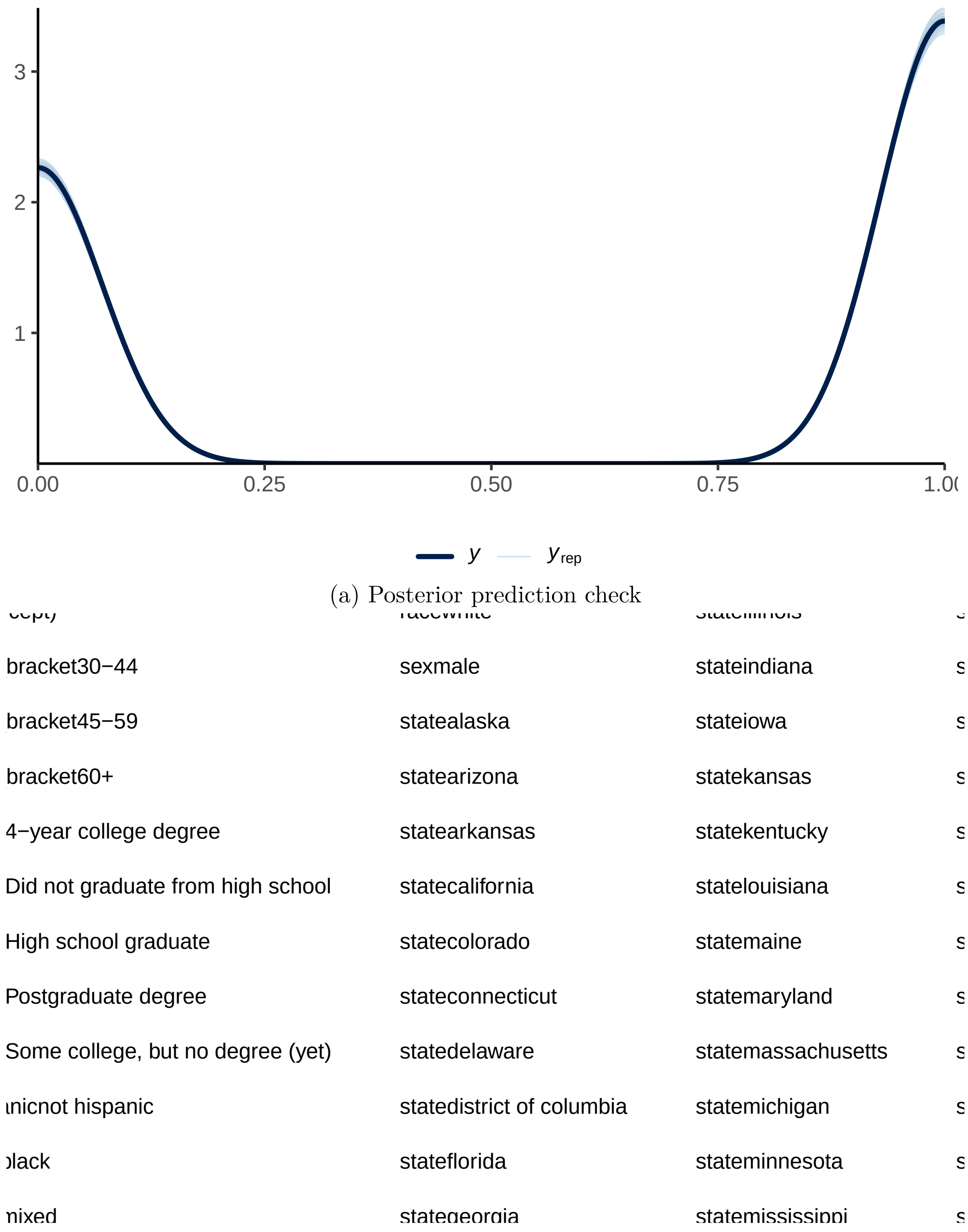


Figure 16: Electoral college map based on the survey dataset



(a) Posterior prediction check

(b) Comparing the posterior with the prior

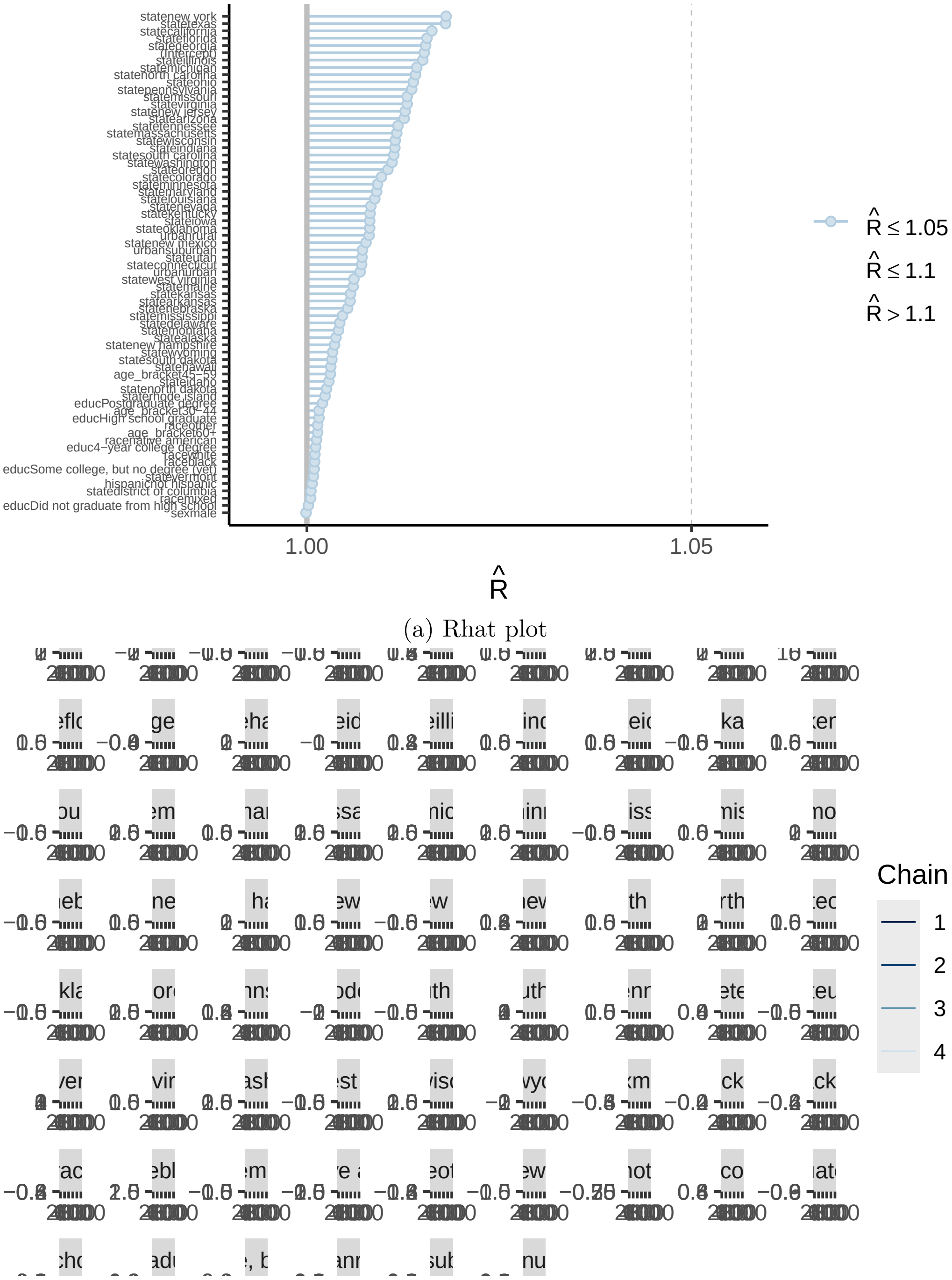
Figure 17: Examining how the model fits, and is affected by, the data

C Model details

C.1 Posterior predictive check

C.2 Markov Chain Monte Carlo

C.3 Credibility intervals



(a) Rhat plot

(b) Trace plot

Figure 18: Checking the convergence of the Markov Chain Monte Carlo (MCMC) algorithm

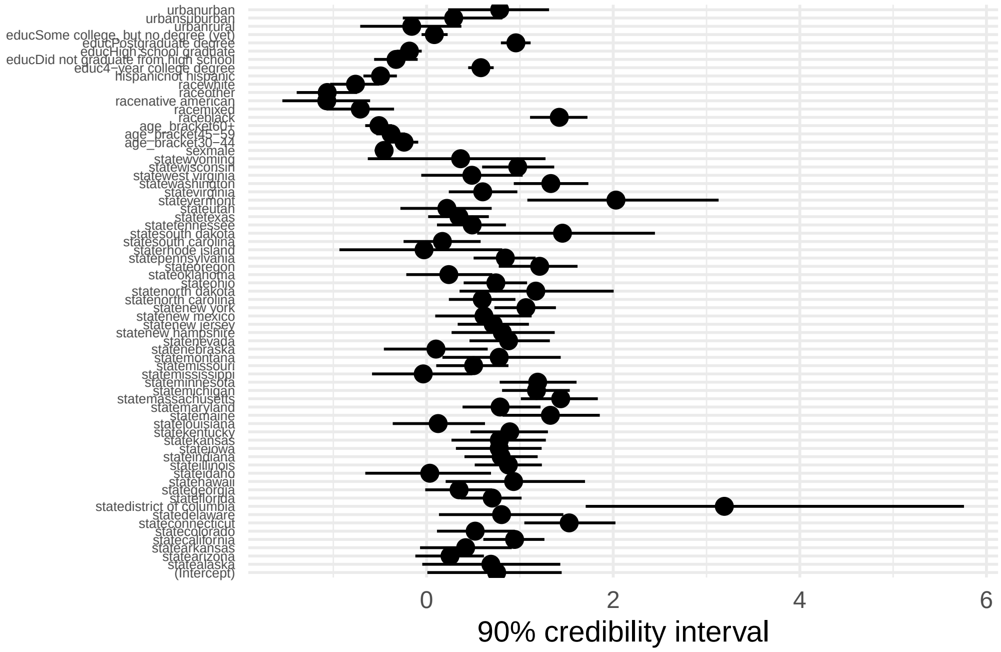


Figure 19: 90% Credibility intervals for the predictors of vote_biden

References

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Renae Rodgers Grace Cooper Stephanie Richards, and Megan Schouweiler. 2024. *IPUMS USA: Version 15.0 [ACS 2022]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022." Harvard Dataverse. https://doi.org/10.7910/DVN/PR4L8P.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.