Forecasting the 2024 U.S. Presidential Election*

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

Talia Fabregas

April 19, 2024

The 2024 U.S. Presidential Election will take place on November 5. Voters will once again choose between President Joe Biden and former President Donald Trump. In this report, I used multi-level regression with post-stratification (MRP) to estimate the results of the upcoming presidential election. I fit a logistic regression model on the survey data and applied to the post-stratification data. Using the results of my MRP analysis, I can predict that President Joe Biden will win 55.47% of the popular vote and defeat former President Trump in the electoral college 413 to 125. However, the electoral college prediction raises questions because many states were predicted to see close a close margin between Biden and Trump.

1 Introduction

Every four years, on the first Tuesday of November, Americans head to the polls to elect their president. The 2024 United States presidential election will take place on Tuesday November 5, 2024. In an era of unprecedented 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 only U.S. president to return to office after losing his re-election bid is Grover Cleveland in 1893 (Waxman 2022).

This report builds on the lessons that I learned when I completed my first US Election Forecast last month. The survey data set that I used was provided by the Cooperative Election Study (CES) 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). The post-stratification data that I used was provided by the Integrated Public Use Microdata Series

*Code and data are available at: https://github.com/taliafabs/us-election-analysis.git

(IPUMS) USA online database. IPUMS provides American 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) to use as my post-stratification data. I performed MRP analysis to estimate the 2024 U.S. presidential election results. This involves using a smaller survey dataset (~10,000 respondents) to fit a logistic regression 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 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 estimate the popular vote and electoral college results of the 2024 U.S. presidential election.

This report features four sections. In Section 2, I discuss the context of the survey and post-stratification dataset, present the results of exploratory data analysis, and use tables and visualizations to show what the variables look like and explain how they interact. In Section 3 I explain, outline, and justify the Bayesian logistic regression model that I used to predict vote preference. In Section 4, I use tables and graphs to present the results of my MRP analysis, which include a popular vote and an electoral college prediction for the 2024 U.S. election. In Section 5, I discuss how my analysis was conducted in more detail, what we can learn from the popular vote and electoral college predictions, the limitations of my datasets and model, why logistic regression was used and the case for (and against) SoftMax regression in U.S. election analysis, and how I hope to extend and improve this report in the future.

I used R programming language (R Core Team 2023) and the `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot` (Wickham 2016), `knitr` (Xie 2014), `readr` (Wickham, Hester, and Bryan 2023), `arrow` (Richardson et al. 2023), and `rstanarm` (Goodrich et al. 2022) 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

My survey data comes from the 2022 Comprehensive Election Study (CES), provided by Harvard Dataverse. Sixty research teams participated in the 2022 CES. It is part of the ongoing Cooperative Election Study (CES), which has been conducted every year since 2006 to study elections in the United States using large-scale survey datasets (Schaffner, Ansolabehere, and Shih 2023). Schaffner, Ansolabehere, and Shih (2023) explain that the CES aims to build off of the work of the 2005 Massachusetts Institute of Technology Public Opinion Research and Training Lab (PORTL) study. The CES was known as the Cooperative Congressional Election Study (CCES) before 2020 (Schaffner, Ansolabehere, and Shih 2023).

Alternatively, I could have used the most recent weekly release of the America's Political Pulse Survey, provided by the Polarization Research Lab as my survey dataset. The Polarization Research Lab is a research group and resource hub led by top scholars at Dartmouth College,

Stanford University, and the University of Pennsylvania that collects and analyzes data to study polarization and democracy (Iyengar, Lelkes, and Westwood 2024). The America's Political Pulse Survey has been conducted weekly by the Polarization Research Lab since 2022 and its purpose is to study political polarization and respect for democratic norms in the United States (Iyengar, Lelkes, and Westwood 2024). There are America's Political Pulse Datasets that were put together as recently as March 2024, however, I opted not to use it because it is too sparse to fit an effective logistic regression model. Each America's Political Pulse dataset contains only 1000 observations, meaning that smaller states have few or no observations. For example, the January 12-19 2024 dataset contained noisy data including very few and overwhelmingly Democratic respondents from GOP stronghold states such as Wyoming and Kansas and no respondents from Vermont (Iyengar, Lelkes, and Westwood 2024).

My post-stratification data set was provided by IPUMS. It is a subset of the American Community Survey (ACS) 2022 (Ruggles et al. 2024) . I obtained it by visiting the IPUMS online database, selecting relevant variables, and downloading a subset containing 500,000 observations.

2.1 Survey data

I selected a random subset of the final release of the 2022 CES Common Content Dataset as my survey data. The 2022 CES Common Content Dataset is a nationally representative sample of 60,000 American adults and it includes sample identifiers such as state and congressional district, demographic profile questions, pre-election questions, post-election questions, and questions about candidate and party preferences (Schaffner, Ansolabehere, and Shih 2023). Most survey respondents are registered with the YouGov panel to receive notifications about surveys and are rewarded with points that can be exchanged for gift cards, however some come from other survey platforms and online ads (Schaffner, Ansolabehere, and Shih 2023). As outlined by Schaffner, Ansolabehere, and Shih (2023), not all responses are included in the 2022 CES Common Content Dataset to ensure that it is as representative as possible. I use a random subset of 10,000 of the 60,000 respondents. I explain my sub-setting process and why my random subset is representative of the 2022 CES Common Content Dataset in Section A.

I do not use the entire questionnaire in my study; I focus on survey questions about demographics, party identification, and vote choice. The variables from the 2022 CES Common Content Dataset that I selected include: `pid3`: 3-point party identification, `presvote16post`: who the respondent voted for in the 2016 U.S. presidential election, `presvote20post`: who the respondent voted for in the 2020 U.S. presidential election, `state`: the state that the respondent lives in, `gender4`: gender identity, `birthyr`: year of birth, `race`: race or ethnicity, `hispanic`: whether the respondent identifies as Hispanic, `educ`: highest level of education completed, and `urbancity`: the type of area that the respondent lives in. `gender4` initially included male, female, non-binary, and other. I re-named it to `sex` and only used male and female because

the `sex` variable in my post-stratification dataset provided by IPUMS only includes male and female. This is one example of how data science can ignore LGBTQ+ identities and reduce humanity to something that can easily be quantified (Keyes 2019).

The 2022 CES Common Content Dataset focuses on the periods leading up to (September 29 to November 8) and following (November 10 to December 15) the 2022 U.S. midterm elections (Schaffner, Ansolabehere, and Shih 2023). While the survey contains specific questions about party identification, strength of party identification, and votes congressional, senate, and gubernatorial elections, it does not ask about preferred 2024 presidential candidate. The survey was conducted in late 2022, when the Republican 2024 presidential nominee was not yet known. Party identification and recent vote choices, especially in the 2020 presidential election where the nominees were also Joe Biden and Donald Trump, are strong indicators of 2024 vote choice. I used `pid3`: party identification, `presvote16post`: 2016 presidential vote, and `presvote20post` to determine whether each respondent would vote for Joe Biden or Donald Trump in the 2024 presidential election and construct the `vote_biden` binary indicator variable. `vote_biden` is equal to 1 if a respondent's preferred 2024 presidential candidate is Joe Biden, and 0 if it is Donald Trump. I discuss the creation of the `vote_biden` variable in more detail in Section A.

The code for the following visualizations was obtained from Telling Stories with Data (Alexander 2023). As seen in Figure 1, support for Biden and Trump in the subset of the 2022 CES Common Content Dataset that I used varies by gender and race. White women were more likely to support President Biden than their white male counterparts. Overall, Black and Asian respondents showed strong support for Biden. Black women showed overwhelming support for Biden. This indicates that both gender and race are likely predictors of support for Biden.

The education gap in vote preference and party identification has widened since 2016 (Pew Research Center 2021a). The differences that I see in vote preference between male and female survey respondents is consistent with this. Figure 2 shows the differences in vote preference for male and female survey respondents by highest level of education. Overall, female respondents were more likely to support President Biden than their male counterparts with the same level of education. Male voters who did not graduate from high school or whose highest level of education is high school favored former President Trump. As highest level of education increased, so did male voters' support for President Biden. Among all male voters, those with a 4-year college degree or a post-graduate degree showed the most support for President Biden. On the contrary, female voters with a high school education or less slightly favored President Biden, while female voters with a 4-year college or post-graduate degree heavily favored President Biden.

Scala and Johnson (2016) found that even when social, demographic, and economic factors such as sex, age, race, and highest level of education were held constant, the urban-rural divide remained statistically significant for estimating vote patterns in American presidential elections. Furthermore, Parker et al. (2018) found that urban voters strongly favored President Biden, suburban voters were split, but slightly favored president Biden, and rural voters favored President Trump. Although the CES Common Content Dataset has strong overall support for

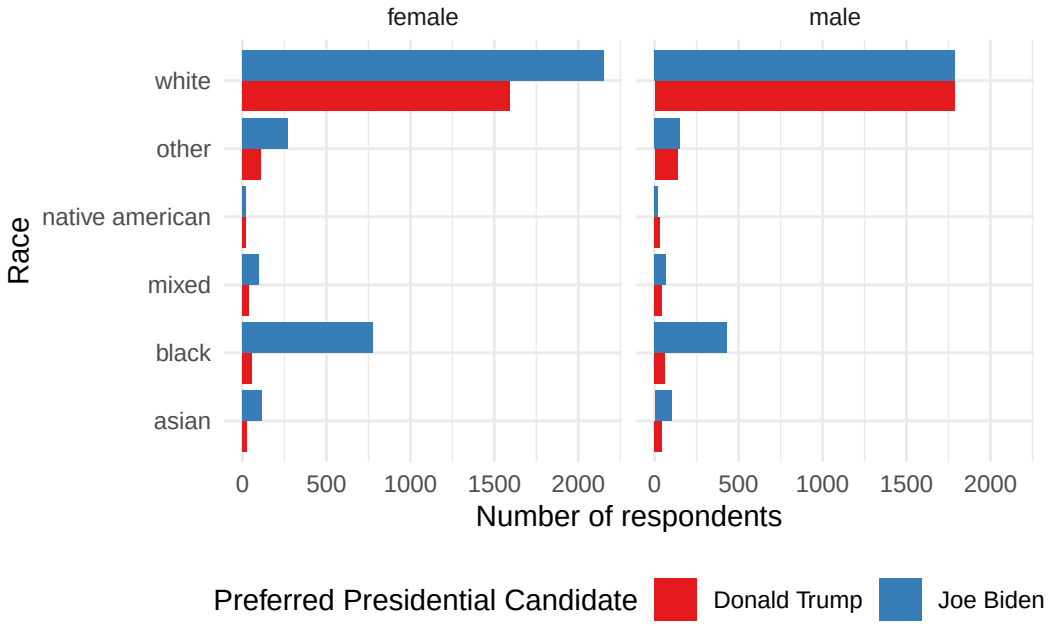

Figure 1: Preferred presidential candidates of survey subset respondents, by gender and race

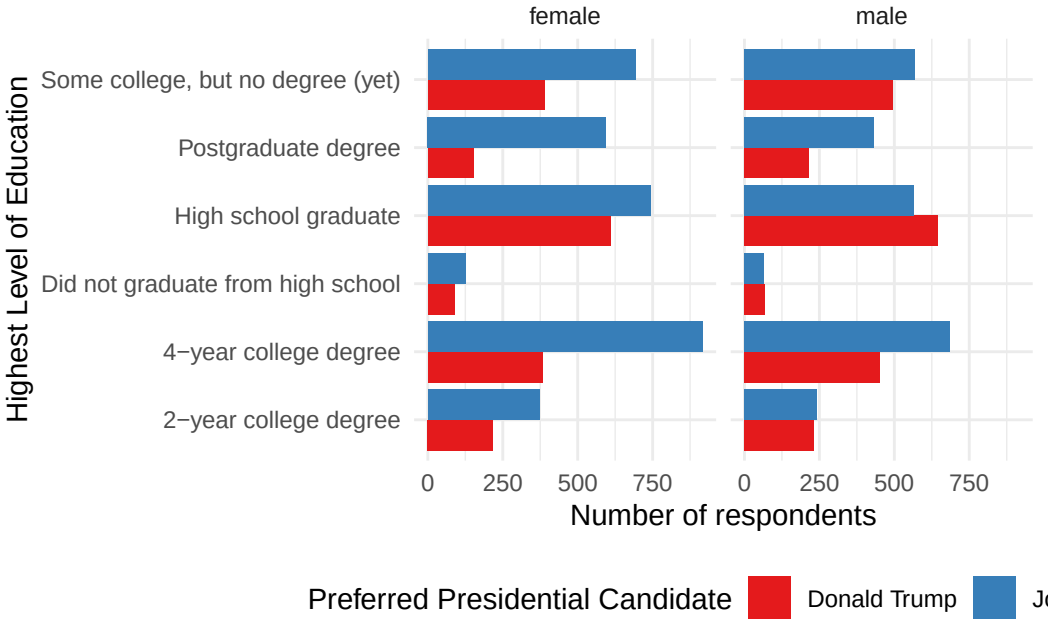

Figure 2: Preferred presidential candidates of survey subset respondents, by highest level of education

Table 1: Presidential preferences of survey subset respondents living in urban, rural, and suburban areas

	Biden %	Trump %
Urban	74.58	25.42
Suburban	59.19	40.81
Rural	44.31	55.69

Biden (60.34% of all respondents), the pattern that Parker et al. (2018) found still holds. As shown in Table 1, voters in urban areas heavily favored President Biden, the gap was narrower for suburban voters, and rural voters favored former President Trump. Figure 3 shows how female and male respondents' for Biden and Trump varied between urban, suburban, and rural voters. Female respondents living in rural areas favored former President Trump, while female respondents in suburban and urban areas favored President Biden. The pattern that Parker et al. (2018) found is stronger among male survey respondents. Male respondents in urban areas favored President Biden, the gap narrowed in suburban areas, and male voters in rural areas favored former President Trump.

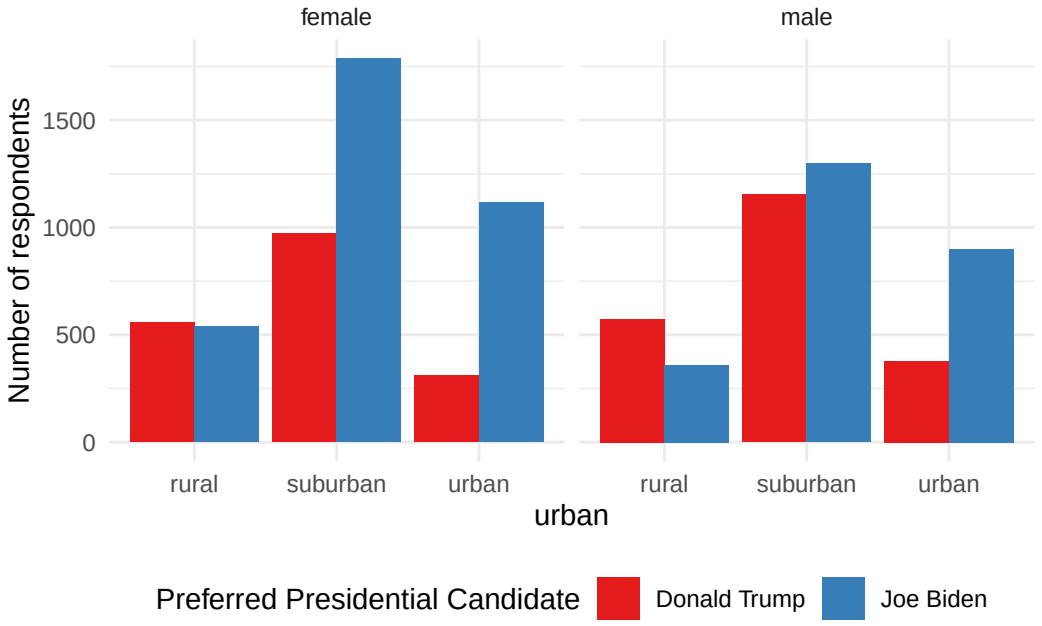

Figure 3: Preferred presidential candidate of subset respondents living in urban, suburban, and rural areas

Figure 5 illustrates the proportion of subset survey respondents in each state who plan to support President Biden in the 2024 election. My subset survey dataset appears to show stronger support for President Joe Biden than the general U.S. electorate both overall and at the state level. As shown in Section B, this is not unique to the subset data; the complete 2022

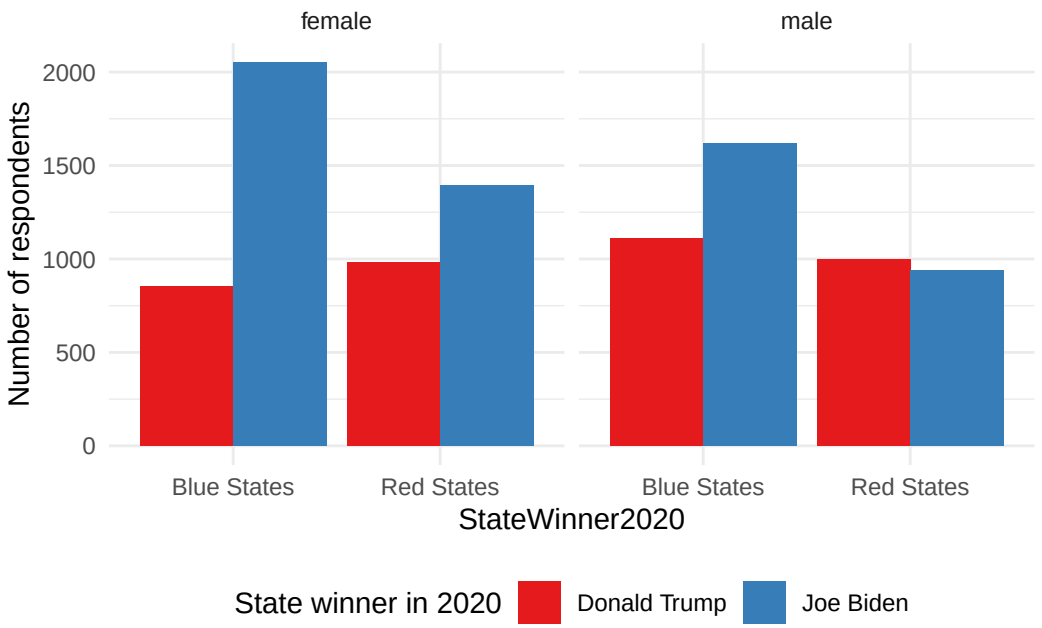

Figure 4: Preferred presidential candidate of subset respondents in states won by Trump vs Biden in 2020

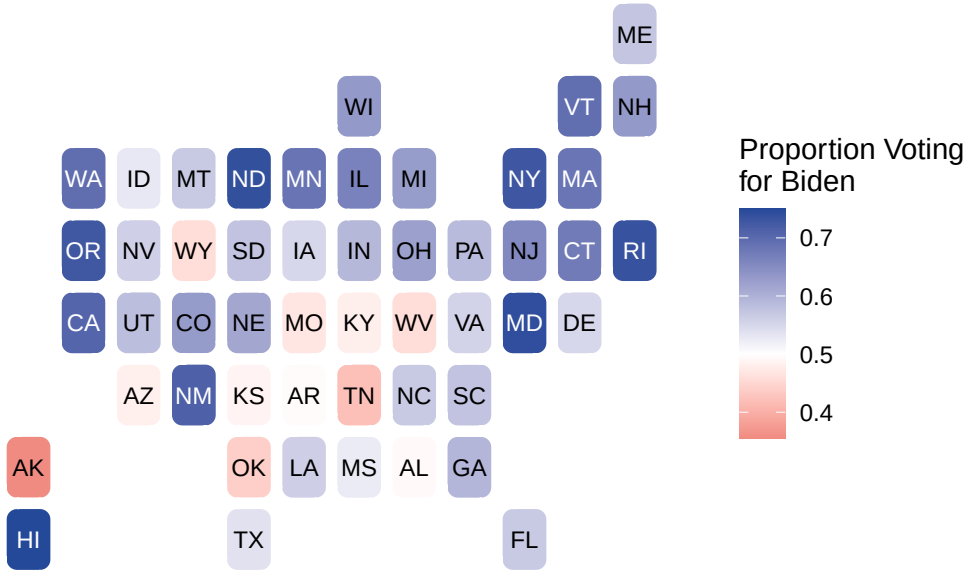

Figure 5: Electoral college map based on the subset survey data

Table 2: Popular vote and electoral college based on subset survey data

Survey Estimate:	Biden	Trump
Num Votes	6003.00	3946.00
% Votes	60.34	39.66
Electoral College	460.00	78.00

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`). 60.34% of survey respondents support BIden, while 39.66% support Trump. However, as seen in Figure 4, respondents who live in states won by Biden in 2020 (blue states) were more likely to support him over Trump 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. Table 2 summarizes the popular vote and electoral college predictions based solely on my survey data.

2.2 Post-stratification data

The post-stratification data that I used is a subset of the 2022 American Community Survey (ACS) from IPUMS (Ruggles et al. 2024). It includes 500,000 census respondents and was downloaded from the IPUMS online database. I selected variables that match the ones in my survey dataset so that my model could easily be applied to my post-stratification data. I carefully chose variables that would later be used in my logistic regression model that predicts 2024 preferred presidential candidate. The variables that I selected include: `stateicp`: the state that the respondent lives in, `age`, `sex`: the sex of the respondent (the ACS 2022 dataset only includes male and female), `race`: respondent's race or ethnicity, `hispan`: whether the respondent identifies as Hispanic or not, `educ`: highest level of education completed, and `metro`: the type of area that the respondent lives in. I re-named, cleaned, and re-factored variable levels to ensure that all variables in my post-stratification analysis dataset matched my survey analysis dataset.

The code for Figure 6, Figure 7, and Figure 8 was obtained from Mitrovski, Yang, and Wankiewicz (2020).

Demographic and geographic patterns in my survey and post-stratification datasets are comparable, but there are some discrepancies. Figure 6 compares the distribution of CES 2022 Common Content (survey) and ACS 2022 (post-stratification) respondents by age bracket, highest level of education completed, and gender. Approximately 10% of survey respondents are 18-29 years old, compared to 28.4% of post-stratification respondents. Consequently, the

survey dataset contains a higher percentage of respondents from the older age brackets than the post-stratification dataset. The percentages of respondents with a postgraduate degree are comparable in the survey and post-stratification datasets, at 13.8% and 10.8%, respectively. However we see a significant difference in the percentages of survey and post-stratification respondents who did not graduate from high school, at 34.1% and 20.7%, respectively. In addition to this, the post-stratification dataset is more gender balanced than the survey dataset; there are slightly more female respondents than male respondents in the survey dataset.

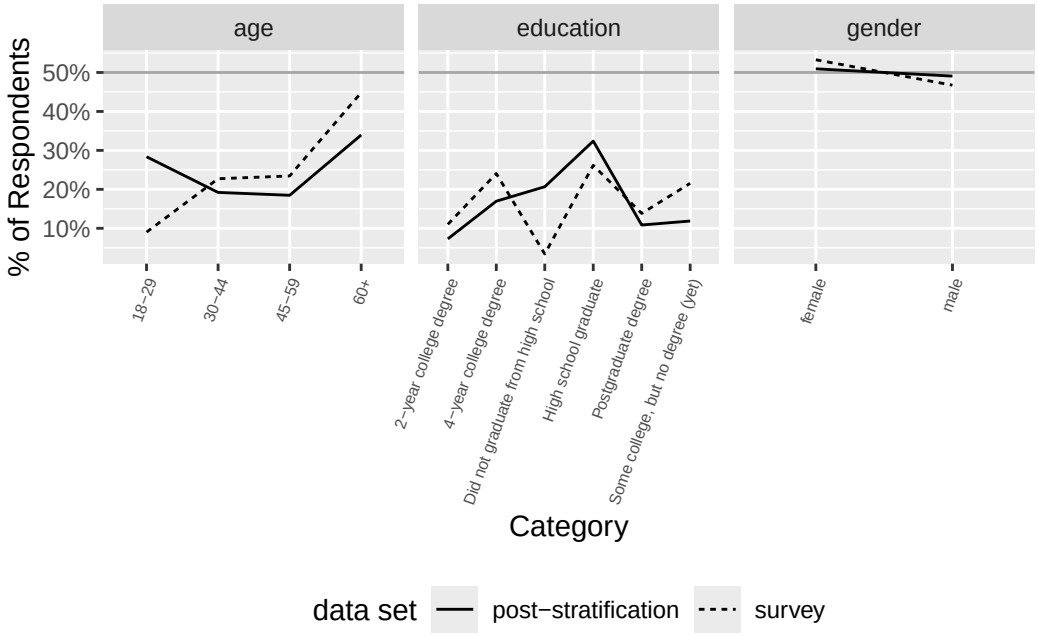

Figure 6: Survey vs post-stratification voter demographics

Figure 7 compares the distributions of survey and post-stratification respondents by race, which are comparable. However, Hispanic respondents have more representation in the post-stratification dataset (14.2%) compared to the survey dataset (8.55%).

Figure 8 shows the distribution of survey and post-stratification respondents across all 50 states. As expected, California and Texas have the highest percentages of voters in each of the survey and post-stratification datasets. However, the post-stratification dataset includes larger percentages of respondents from California and Texas, and, consequently, smaller percentages of respondents from smaller states. On the contrary, the survey dataset includes smaller percentages of respondents from California and Texas, and more respondents from Florida, Maryland, Michigan, New Jersey, New York, Ohio, and Pennsylvania.

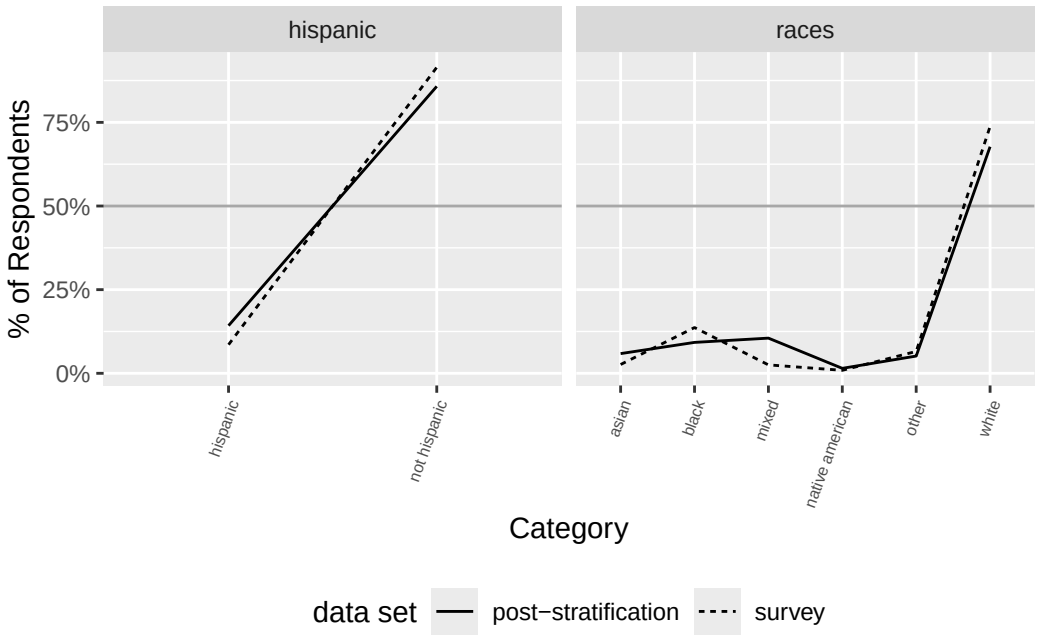

Figure 7: Survey vs post-stratification voter race demographics

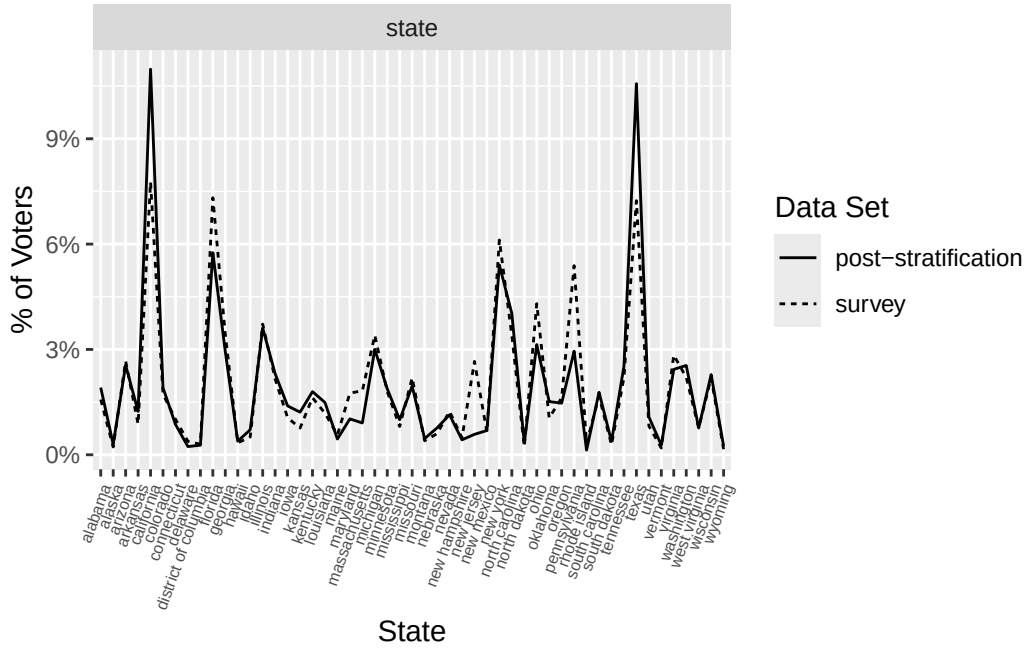

Figure 8: Survey and Post-Stratification Data Proportion of Voters by State

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. To perform MRP analysis, I fit a logistic regression model on the survey data and applied it to the post-stratification data. Logistic regression is a popular algorithm for binary classification; its goal is to learn how to classify examples (which in this case, are American voters) as Biden or Trump voters based on their features (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) (TensorFlow, n.d.).

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:

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

where the binary indicator variable `vote_biden` 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 possibility of voting for a third-party candidate

is 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 Section 5.3.

3.2 Model Justification

Logistic regression is a reasonable and appropriate choice for forecasting the U.S. presidential election because there are only two candidates who can win electoral college votes and the White House on November 5, 2024: Joe Biden and Donald Trump. In 2020, third-party candidates received only 1.9% of the national popular vote (Wasserman et al. 2020). I expect to see a positive relationship between support for President Biden and living in a state that he won in 2020, female gender, non-white race, college education or higher, and living in an urban or suburban area.

Across all race groups, women are more likely to support the Democratic Party than their male counterparts (Pew Research Center 2021b). Pew Research Center (2021b) also found sizable differences in party affiliation among different races; White voters are more likely to affiliate with or lean toward the Republican Party, while Black, Asian, and Hispanic voters overwhelmingly support the Democratic Party. Pew Research Center (2021a) found that Trump received a higher Hispanic vote share in 2021 than 2016, but Hispanic voters without a college education were substantially more likely to support Trump than Hispanic voters with a college education or higher. Similarly, Pew Research Center (2021b) found that higher levels of education are associated with increased support for and leaning towards the Democratic Party. America developed a rural-urban divide in the 1990s and it has kept widening since then (Cornellians Staff 2022). America's urban-suburban-rural divide applies to both party identification and political opinions (Parker et al. 2018). In a 2017 survey, Parker et al. (2018) found that 62% of registered voters who live in urban areas identify as Democrats or leaning Democratic, compared to only 31% who identify as Republican or lean Republican. The gap was narrower for registered voters who live in suburban areas, with 47% and 45%, respectively. Registered voters in rural areas leaned more Republican than Democratic, at 54% and 38%, respectively. I obtained similar results in Section 2.1, where I found that urban voters showed the most support for President Biden, followed by suburban voters.

4 Results

4.1 Popular Vote Prediction

Based on the results of my MRP analysis, which involved applying the logistic regression model outlined Section 3 to a subset of the ACS 2022 dataset, I can predict that Joe Biden will win the popular vote on November 5. President Biden is estimated to win 55.47% of the popular vote, while former President Trump is estimated to win 44.53% of the popular vote. This means that the logistic regression model that I fit on the CES 2022 Common Content Dataset

classified 55.47% of examples (ACS 2022 respondents) as Biden voters based on their features (`state`, `biden_won`, `age_bracket`, `sex`, `race`, `hispanic`, `educ`, and `urban`). The remaining 44.53% were classified as Trump voters.

The lower quantile, mean, and upper quantile estimates of the national support for President Biden, which correspond to the upper quantile, mean, and lower quantile estimates of national support for former President Trump are summarized in Table 3. The lower quantile estimate for Biden shows Trump winning the popular vote by a 2.9 percentage point margin; it predicts 51.45% of the popular vote for Trump. It is important to note that logistic regression performs binary classification, so the estimated popular vote percentages for Joe Biden and Donald Trump based on my MRP analysis will always add up to 100%.

Table 3: 2024 U.S. election popular vote estimates based on post-stratification analysis

	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

The lower quantile, mean, and upper quantile estimates for the percentage of votes that President Biden will receive in each state based on my MRP analysis are illustrated in Figure 9. The blue (states where the mean prediction for Biden is over 50%) or red line (states where the mean prediction for Trump is over 50%) shows the 95% prediction interval for the percentage of votes that Biden will receive. The mean estimates are denoted by the dot in the middle of each line. The 95% prediction interval falls on both sides of the 50% threshold for numerous states. This means that based on the results of my MRP analysis, we can expect to see close races in a number of states that saw close margins between Trump and Biden in 2020 including, but not limited to, Arizona, Iowa, New Hampshire, New Mexico, Georgia, Texas, Indiana, Ohio, Pennsylvania, and Wisconsin (CNN Politics 2020). There are a number of other states, including South Dakota, Maine, Minnesota, Delaware, Kentucky, and North Carolina that see the 95% prediction interval fall on both sides of the 50% support for Biden threshold.

Table 4: 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

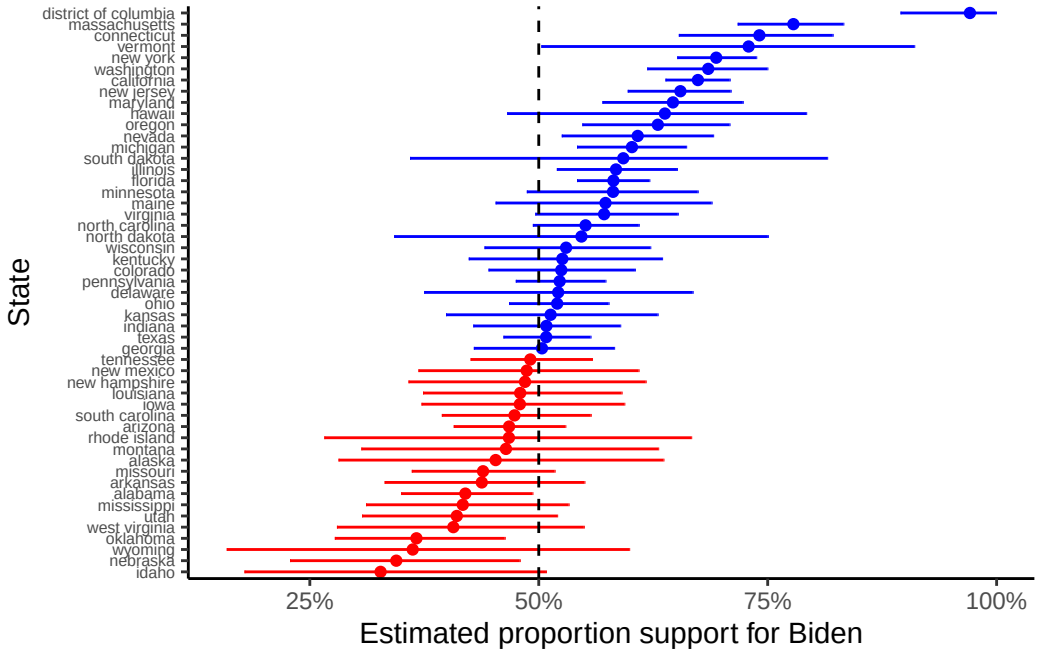

Figure 9: Estimated proportion of each state voting for Biden in 2024 based on MRP analysis

Figure 10 shows the same estimates as Figure 9, in comparison with the percentage of CES 2022 survey respondents who plan to vote for Joe Biden. The estimates based on the survey data are represented by the gray dots. There is some variation between survey and post-stratification estimates, but in most cases, the survey estimates fall within the 95% prediction interval. However, most states had higher support for President Biden in the survey dataset than in the MRP analysis results. This makes sense, because the CES 2022 survey dataset showed stronger support for President Biden (60.3%) than the general U.S. electorate and the exploratory data analysis conducted in Section 2.1 showed that respondents living in states carried by Biden in 2020 showed higher levels of support for President Biden than their counterparts in states carried by Trump in 2020 when gender was held constant.

Figure 11 uses an electoral map to illustrate the same information as Figure 9. It shows the mean estimated proportion of votes that President Joe Biden will receive in each state. The electoral college votes of the states that appear blue in Figure 11 would go to President Biden, and the electoral college votes of the states that appear red would go to President Trump. California, Oregon, Washington, Massachusetts, Connecticut, New York, New Jersey, and Vermont appear very blue on this map; this indicates that my MRP analysis predicted a comfortable margin of victory for Biden in each of these historically Democratic stronghold states (CNN Politics 2020). On the contrary, Idaho, Nebraska, Idaho, Wyoming, and Oklahoma appear quite red on the map. This indicates that my MRP analysis predicted a comfortable margin of victory for former President Trump in these states.

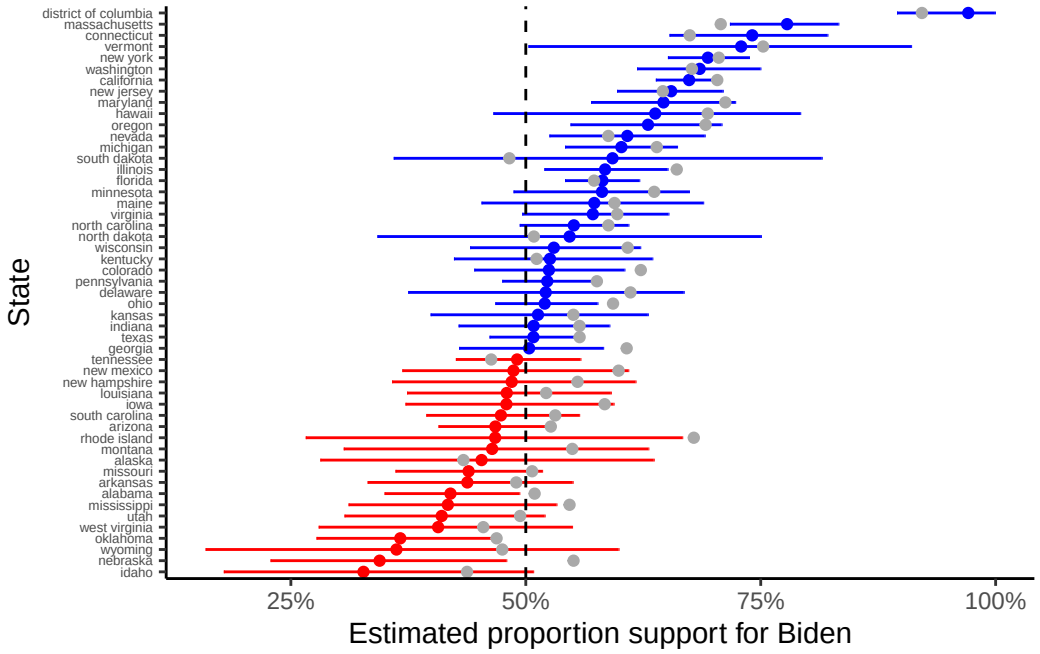

Figure 10: Estimated proportion of each state voting for Biden in 2024 Post-Stratification vs subset Survey Data

Table 4 shows the estimates for the electoral college based on the lower, mean, and upper estimates for the percentage of votes in each state that Biden will receive shown in Figure 9. The lower estimate for Joe Biden shows him losing the electoral college to Donald Trump, 220 to 318. However, the mean and upper estimates each show Joe Biden defeating Donald Trump in the electoral college by a wide margin, at 413 to 125 and 517 to 21, respectively. There is a huge difference between the lower and upper electoral college estimates for Joe Biden because estimates based on the lower and upper quantiles of the 95% prediction interval for support for his re-election campaign in each state often fall on opposite sides of the 50% threshold. One of those states is Texas, which is allotted the second-most electoral college votes, behind California (Wasserman et al. 2020). Wisconsin, Pennsylvania, Indiana, Georgia, Texas, and New Mexico appear very light-colored in Figure 11. These states also saw 95% prediction intervals for the percentage of votes that President Biden will receive fall on both sides of the 50% victory threshold. The winner of each of these states would get all of the electoral votes, regardless of the margin. Therefore, based on my MRP analysis, the results of 2024 U.S. presidential election may be defined close races in key battleground states, where a win for either Trump or Biden is within the 95% prediction interval.

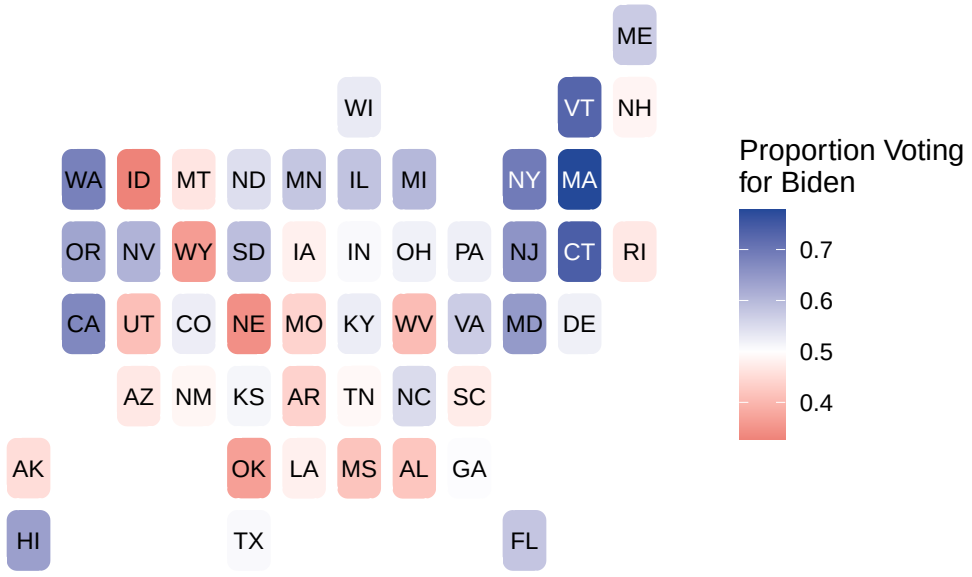

Figure 11: Electoral map based on MRP analysis

5 Discussion

5.1 Questions Raised by the use of Logistic Regression and the Resulting Electoral College Prediction

Based on my MRP analysis, I predicted that Joe Biden and Donald Trump would win 55.47% and 44.53% of the popular vote, respectively. As noted in Section 4, these percentages sum to 100% because the only candidates that my logistic regression model considers are Biden and Trump. In the 2020 U.S. election, Joe Biden and Donald Trump received 51.3% and 46.9% of the popular vote, respectively (CNN Politics 2020). Unlike the results of my MRP analysis, these predictions do not sum to 100%; third-party candidates received 1.8% of the national popular vote. While this percentage seems small, it is not evenly distributed across all states, and in some swing states, the third-party vote share was larger than Biden or Trump's margin of victory. Based on my MRP analysis, several battleground states including (but not limited to) Georgia, Pennsylvania, Wisconsin, and Arizona had 95% prediction intervals for support for President Biden fall on both sides of the 50% threshold. Some of these states saw a larger third-party vote share than margin of victory for President Biden in the 2020 election. In the 2020 presidential election, Georgia was decided by 0.3 percentage points, with President Biden receiving 49.5% of the vote, President Trump receiving 49.2%, and third-party candidates receiving 1.3% (CNN Politics 2020). Similarly, President Biden received 49.4%, former President Trump received 49.0%, and third-party candidates received 1.6% of the vote in Arizona (CNN Politics 2020). Logistic regression does not account for this. In general, that is a small tradeoff and using logistic regression is reasonable and appropriate because

the only two candidates who can win any electoral college votes and the White House are the Democratic and Republican nominees. The limitations of logistic regression can accumulate and seriously threaten the accuracy of an electoral college prediction. Not accounting for third-party candidates can be the difference between correctly and incorrectly predicting the results in swing states that have seen larger third-party vote shares than margins of victory in recent elections. Inaccurately predicting battleground states with tight margins will lead to an inaccurate electoral college and election forecast. In Section 5.3, I discuss the case for (and against) the use of SoftMax regression in U.S. election forecasting.

5.2 Weaknesses and Limitations of the Survey Dataset

The survey dataset that I selected, the CES 2022 Common Content Dataset, has an excellent sample of 60,000 respondents and it includes questions about race, sex, age, state, type of area, partisan affiliation, and recent presidential vote choices. However, it is not without weaknesses and limitations. Firstly, it does not contain a question about preferred 2024 presidential candidate. I had to construct the `vote_biden` variable myself, and that process is outlined in Section A. Secondly, it is almost two years old and its intended use is for studying the 2022 midterm elections, not forecasting the 2024 presidential election.

Another concern that I have about my survey dataset is bias in the `vote_biden` variable. 60.34% of survey respondents are labeled as Biden voters (`vote_biden = 1`), compared to 39.66% of survey respondents labeled as Trump voters (`vote_biden = 0`). The popular vote prediction that I have for President Biden based solely on my survey dataset, at 60.34%, is siginificantly higher than the popular vote shares that Presidents Clinton, Bush, Obama, and Trump ever received (Roper Center for Public Opinion Research 2024). The last U.S. president to win over 60% of the national popular vote was Richard Nixon in 1972 (Roper Center for Public Opinion Research 2024). This means that there is likely higher support for President Biden in my survey dataset than in the general U.S. electorate. I constructed the `biden_won` variable to try to account for this, but at the end of the day, biased survey data can lead to biased MRP results. My logistic regression model used the survey data to learn how to classify respondents based on state, whether Biden won that state, age, sex, race, education, and the type of area that they live in. If there was bias present in my survey dataset, my model would have learned it and produced biased results.

5.3 The Case for (and against) SoftMax Regression

As outlined in Section 5.1, the use of logistic regression in U.S. election forecasting is justified and reasonable, but not without limitations. SoftMax regression has the potential to solve the issues presented in Section 5.1; it is a generalization of logistic regression that can perform multi-class classification for K classes instead of binary classification for just two. At first, this sounds great. Considering third-party candidates sounds like a perfect solution to the issue that my current logistic regression model seems to have when predicting swing states such

as Georgia, Arizona, and Pennsylvania. However, SoftMax regression should be used with caution. It is far more complex than logistic regression and it includes more parameters. This means that it may severely overfit my survey dataset, which contains 10,000 observations. A model overfits when it is simply too complex and powerful to perform the intended task, so it picks up noise (outliers) in the data and uses those to make predictions. An overfit model will not perform well on the unseen post-stratification data. Given my current dataset, logistic regression remains the appropriate choice despite its inability to consider third-party voters and what this might mean for swing state predictions.

5.4 Next Steps

I intend to extend this study with a more recent and larger survey dataset in the future. My first step towards extending and improving this study would be finding a dataset that includes a question about preferred 2024 presidential candidate and is large enough to train, validate, and test a SoftMax regression model. I would use Python instead of R to conduct further analysis. I would then split my new dataset into training, validation, and test sets and teach a SoftMax regression model to classify voters as Biden, Trump, or third-party voters using the training data. I would then use the validation set to apply regularization to prevent overfitting and tune the learning rate to maximize validation accuracy. My goal would be to maximize validation accuracy because it is an indicator of how well the model will be able to classify unseen post-stratification dataset respondents. After it has been trained, tuned, and tested, I will apply the SoftMax regression model to the post-stratification dataset and use the results to once again predict the 2024 U.S. election electoral college and popular vote results, with third-party candidates taken into account.

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 Wasserman et al. (2020), 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 information 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 over 50,000 responses, but it was subset 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 sub-setting 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 more than 50,000 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. For the reasons outlined above, I am confident that my random subset is representative of the original 2022 CES Common Content Dataset (Schaffner, Ansolabehere, and Shih 2023).

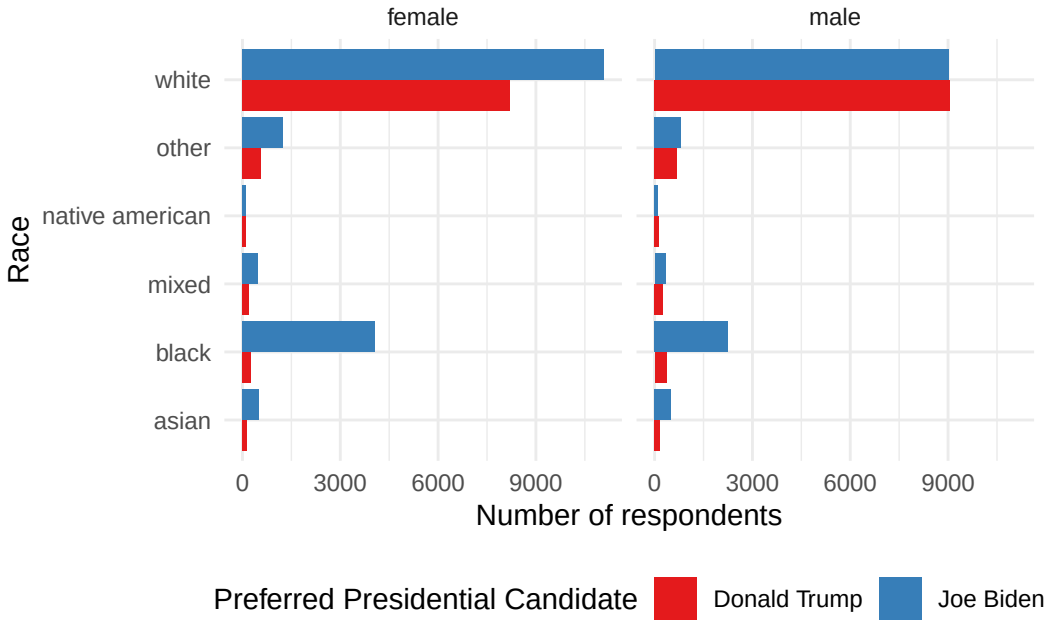

Figure 12: Preferred presidential candidates of survey respondents, by gender and race

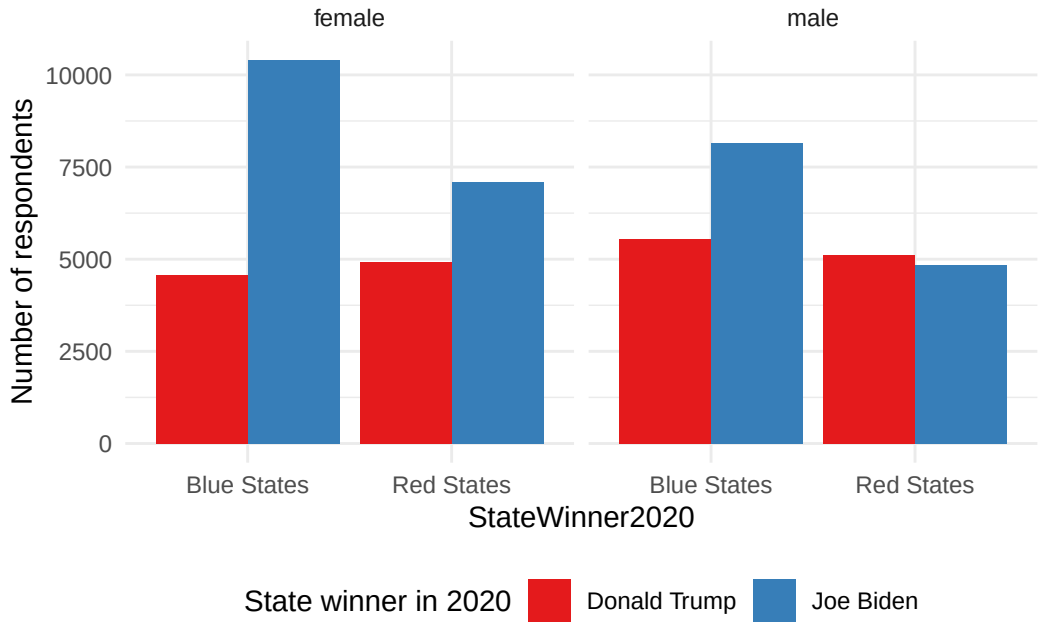

Figure 13: Preferred presidential candidate of survey respondents in states carried by Trump vs Biden in 2020

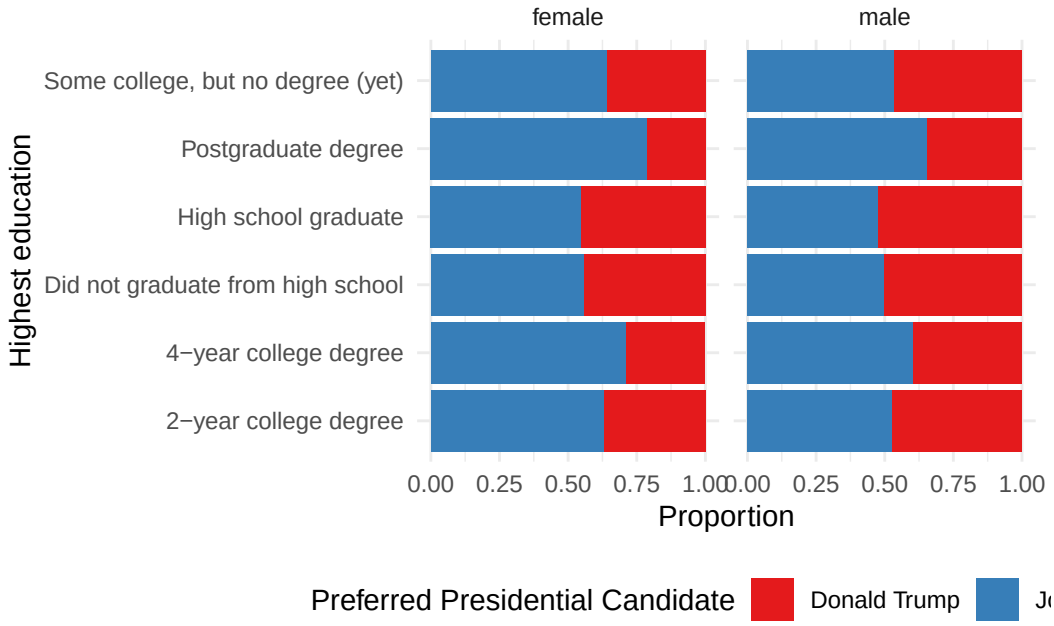

Figure 14: Preferred presidential candidates of survey respondents, by highest level of education

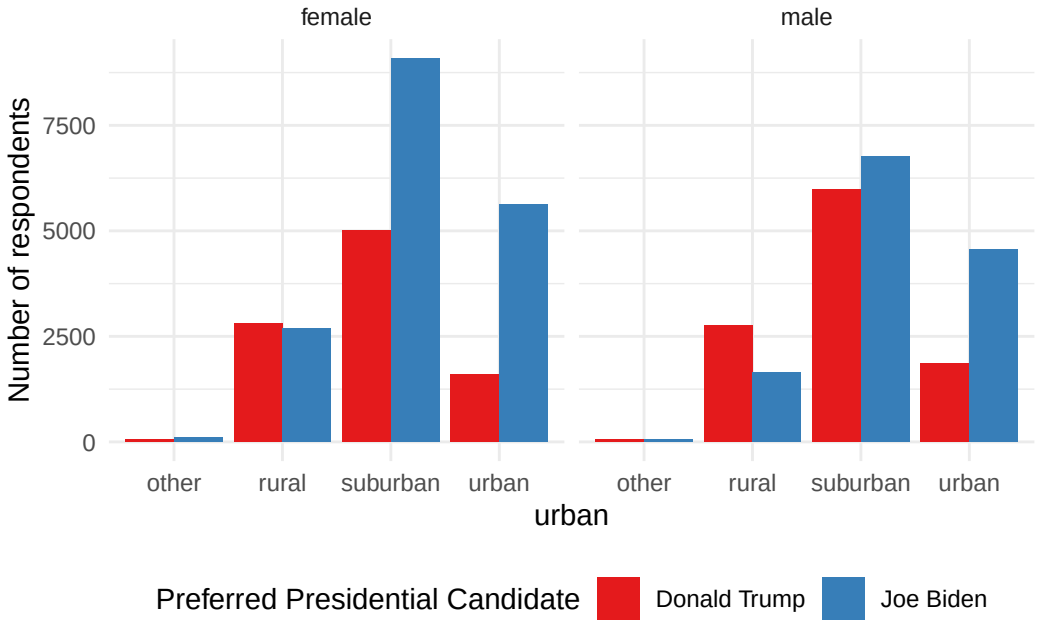

Figure 15: Preferred presidential candidate of survey respondents living in urban vs rural areas

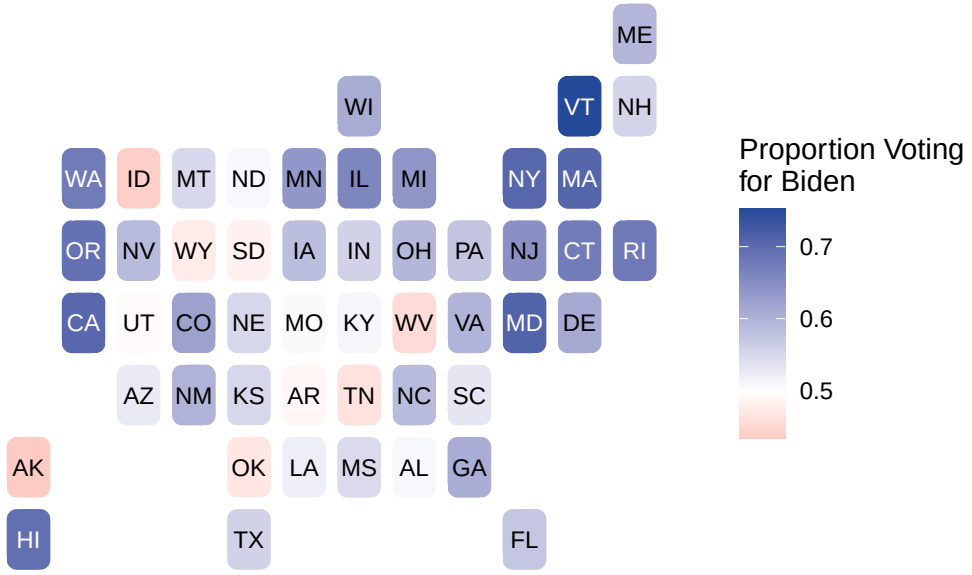

Figure 16: Electoral college map based on the survey dataset

Table 5: Presidential preferences of survey respondents living in urban, rural, and suburban areas

	Biden %	Trump %
Urban	74.67	25.33
Suburban	59.02	40.98
Rural	43.81	56.19

C Model details

C.1 Posterior predictive check

Figure 17 shows the posterior predictive check and Figure 18 shows the comparison of the posterior with the prior (Alexander 2023).

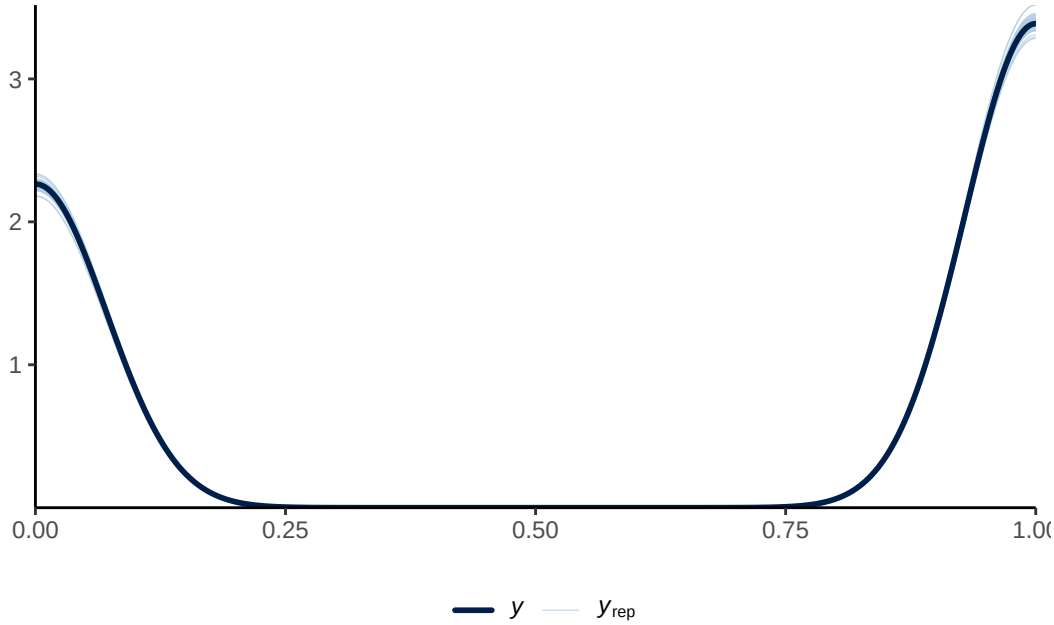

Figure 17: Posterior prediction check

C.2 Markov Chain Monte Carlo

I used an Rhat plot and a trace plot to check for signs that the Markov Chain Monte Carlo (MCMC) may have ran into any issues (Alexander 2023). As seen in Figure 19, everything is very close to 1 and below 1.05. This means that all the coefficients converge to the same

cept)	racewhite	stateillinois	s
bracket30−44	sexmale	stateindiana	s
bracket45−59	statealaska	stateiowa	s
bracket60+	statearizona	statekansas	s
4−year college degree	statearkansas	statekentucky	s
Did not graduate from high school	statecalifornia	statelouisiana	s
High school graduate	statecolorado	statemaine	s
Postgraduate degree	stateconnecticut	statemaryland	s
Some college, but no degree (yet)	statedelaware	statemassachusetts	s
inicnot hispanic	statedistrict of columbia	statemichigan	s
olack	stateflorida	stateminnesota	s
nixed	stategeorgia	statemississippi	s

Figure 18: Comparing the posterior with the prior

distribution and the logistic regression model defined in Section 3 can predict `vote_biden` (Alexander 2023).

Figure 20 shows the trace plots for each value for the predictors of `vote_biden`. If the lines are bouncy, horizontal, and overlapping, then the distribution behaves the way we expected it to and there is no need to re-run the model (Alexander 2023).

C.3 Credibility intervals

Figure 21 shows the 90% credibility intervals for the predictors of `vote_biden`. This visualization was produced using code provided by Alexander (2023).

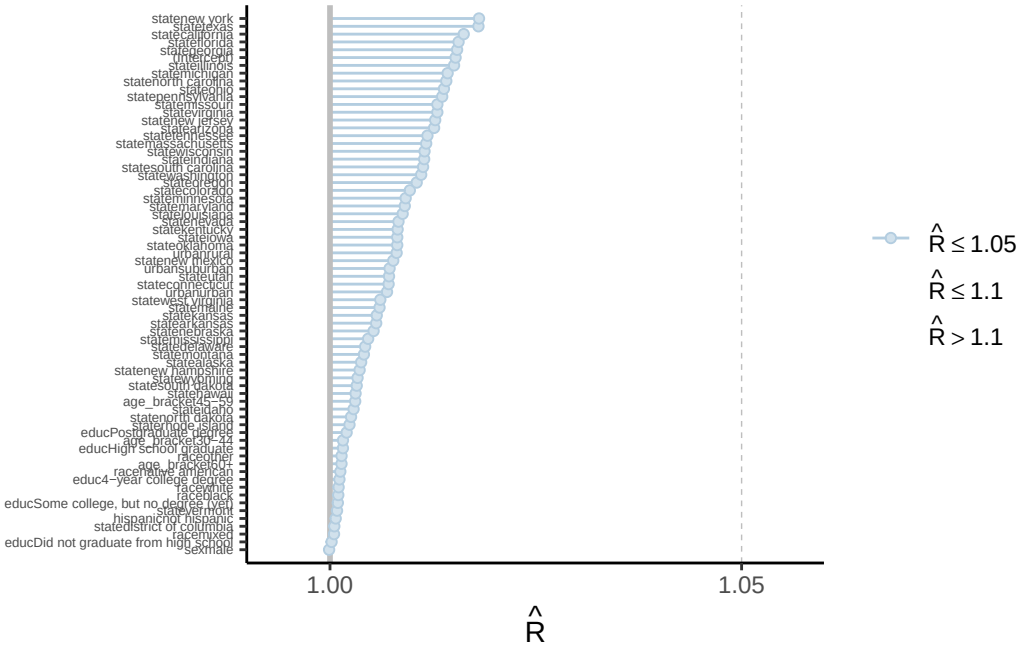

Figure 19: Checking the convergence of the Markov Chain Monte Carlo (MCMC) algorithm: Rhat

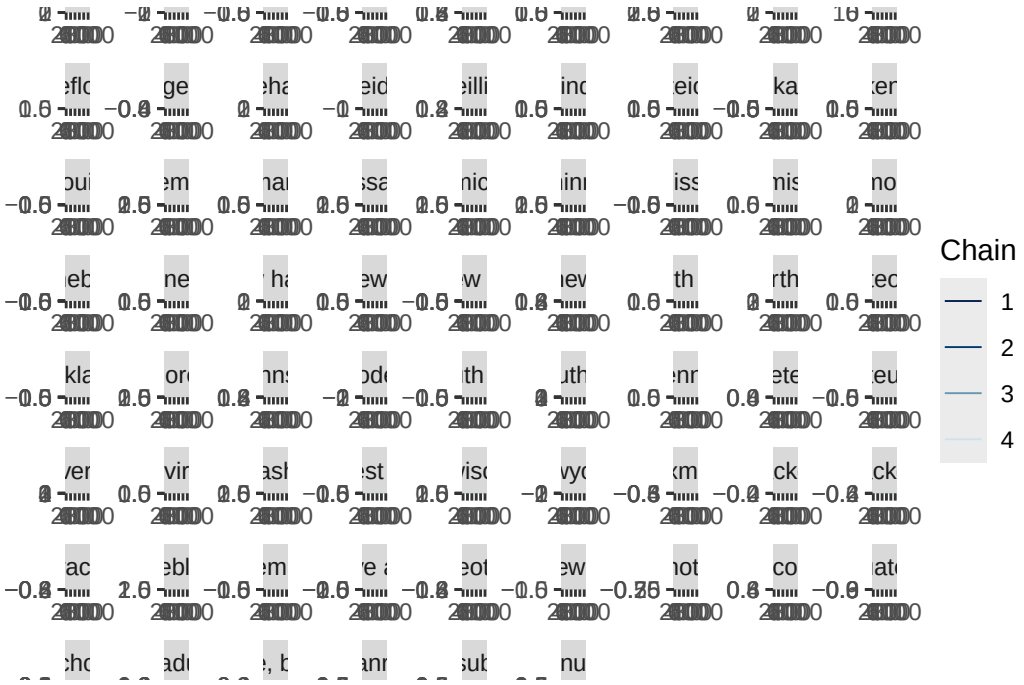

Figure 20: Checking the convergence of the Markov Chain Monte Carlo (MCMC) algorithm: Trace plot

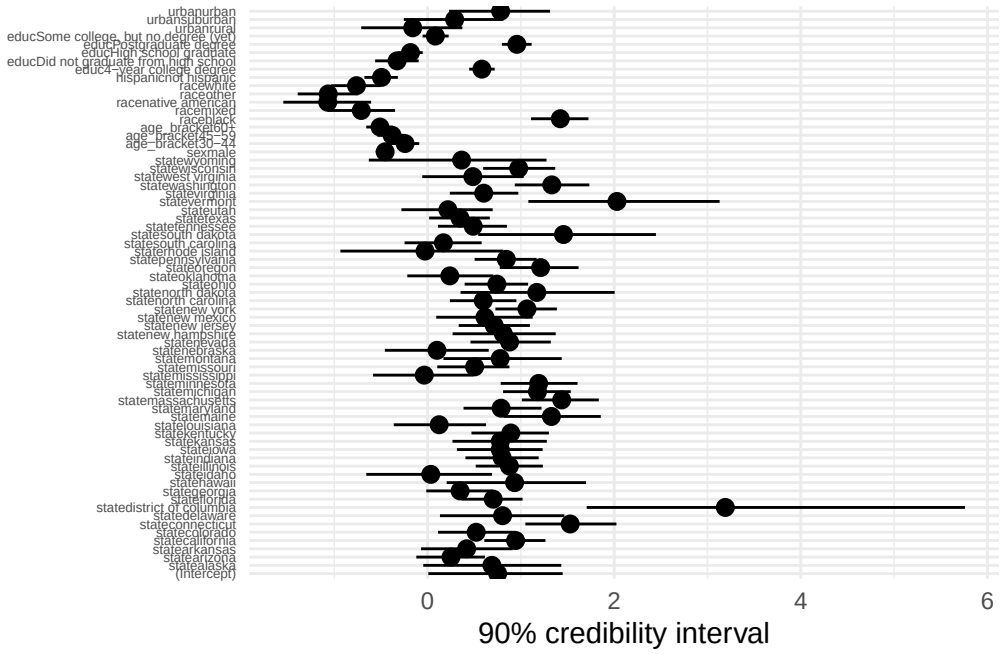

Figure 21: 90% Credibility intervals for the predictors of vote_biden

References

Alexander, Rohan. 2023. *Telling Stories with Data*. "University of Toronto". https://www.tellingstorieswithdata.com.

CNN Politics. 2020. *America's Choice 2020*. CNN. https://www.cnn.com/election/2020/results.

Cornellians Staff. 2022. *Exploring the Widening Chasm Between Urban and Rural Voters*. Cornell University Department of Government. https://government.cornell.edu/news/exploring-widening-chasm-between-urban-and-rural-voters.

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/.

Iyengar, Shanto, Yphtach Lelkes, and Sean Westwood. 2024. *America's Political Pulse*. https://polarizationresearchlab.org/americas-political-pulse/.

Keyes, Os. 2019. *Counting the Countless*. https://reallifemag.com/counting-the-countless/.

Mitrovski, Alen, Xiaoyan Yang, and Matthew Wankiewicz. 2020. *Joe Biden Projected to Win Popular Vote in 2020 US Election with 51*. Telling Stories with Data. https://github.com/matthewwankiewicz/US_election_forecast/tree/main.

Parker, Kim, Juliana Menasce Horowitz, Anna Brown, Richard Fry, Dvera Cohn, and Ruth Igielnik. 2018. *What Unites and Divides Urban, Suburban and Rural Communities*. Pew Research Center. https://www.pewresearch.org/social-trends/2018/05/22/what-unites-

and-divides-urban-suburban-and-rural-communities/.

Pew Research Center. 2021a. *Behind Biden's 2020 Victory*. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/.

———. 2021b. *Wide Gender Gap, Growing Educational Divide in Voters' Party Identification*. https://www.pewresearch.org/politics/2018/03/20/wide-gender-gap-growing-educational-divide-in-voters-party-identification/.

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.

Roper Center for Public Opinion Research. 2024. *Popular Votes 1940-2016*. Cornell University. https://ropercenter.cornell.edu/presidential-elections/popular-votes.

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.

Scala, Dante J., and Kenneth M. Johnson. 2016. *Political Polarization Along the Rural-Urban Continuum? The Geography of the Presidential Vote, 2000–2016*. Vol. 672. The American Academy of Political; Social Science. https://doi.org/https://doi.org/10.1177/0002716217 7126.

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

TensorFlow. n.d. *Logistic Regression for Binary Classification with Core APIs*. https://www.tensorflow.org/guide/core/logistic_regression_core#:~:text=Logistic%20regression%20is%20c belonging%20to%20a%20particular%20class.

Wasserman, David, Sophie Andrews, Leo Saenger, Lev Cohen, Ally Flinn, and and Griff Tatarsky. 2020. *2020 National Popular Vote Tracker*. https://www.cookpolitical.com/2020-national-popular-vote-tracker.

Waxman, Olivia. 2022. *Donald Trump, Grover Cleveland, and the History of Trying to Win Back the White House*. TIME. https://time.com/6234562/nonconsecutive-terms-president-grover-cleveland-donald-trump/.

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

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.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.