Forecasting the 2024 U.S. Presidential Election*

Biden Expected to Win Popular Vote Based on MRP Analysis

Talia Fabregas Fatimah Yunusa Aamishi Sundeep Avarsekar

March 19, 2024

The 2024 U.S. Presidential Election will take place on November 5th and it will be a rematch between President Joe Biden and Former President Donald Trump. In this report, we used a logistic regression model along with multi-level post-stratification (MRP) to predict the winner of the upcoming election. Using the results of our model, we predict that President Biden will win 56.59% of the popular vote and defeat former President Trump 467 to 71 in the electoral college, and win a second term in the White House. However, due to the limited nature of our survey data, the error range of our electoral college prediction is large.

1 Introduction

The U.S. Presidential election will take place on Tuesday, November 5th, 2024, as America faces unprecedented levels of political polarization and diminished levels of trust in democratic institutions. As of March 11, 2024, former President Trump has 1078 out of the 1215 delegates needed to win the Republican nomination and President Biden has 1872 out of the 1968 delegates needed to win the Democratic nomination, therefore this election will feature the same candidates as 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.

Our survey data set was provided by the Polarization Research Lab. The Polarization Research Lab was formed by a group of researchers to increase understanding of where partisan animosity comes from and what can be done to help (Iyengar, Lelkes, and Westwood 2024). Our post-stratification data set was provided by the Integrated Public Use Microdata Series (IPUMS). IPUMS is an organization which lays out survey and census data with the help of 105 statistical agencies (Ruggles et al. 2024). We used the data from the 2022 American Community Survey (ACS) for this report (Ruggles et al. 2024). We use multi-level regression

*Code and data are available at: https://github.com/taliafabs/US-Election-Forecast-2024.git with post-stratification (MRP) to predict the results of the 2024 U.S. Presidential Election. MRP uses a smaller survey data set to fit a model to predict vote preference. We built a logistic regression model, using sex, age, race, Hispanic, highest level of education, and whether the respondent lives in an urban or rural area as predictors to estimate support for President Biden. We fit this model using our survey data set, and apply it to our post-stratification data set to predict the popular vote and electoral college results of the 2024 election.

Our report has 4 sections. In Section 2, we examine the data sets that we used for our report and present how variables in our survey data set interact as well as summary statistics. In Section 3, we outline how we built a logistic regression model and used MRP to predict the results of the 2024 U.S. presidential election. In Section 4, we present the findings of our model, which include a prediction for the popular vote and electoral college results. In Section 5, we discuss what we can learn from our results, how the weaknesses of our survey data set translate into model weaknesses and limitations, and how this report can be extended and improved.

We used R (R Core Team 2023), and the `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023), `ggplot` (Wickham 2016), `rstanarm` (Brilleman et al. 2018), and `arrow` (Richardson et al. 2023) packages throughout our analysis to clean our survey and post-stratification data sets, fit our model, apply our model, and create our data visualizations.

2 Data

Our survey data is from the America's Political Pulse Survey conducted by the Polarization Research Lab. The Polarization Research Lab is a research group founded by top political science research scholars at Dartmouth College, Stanford University, and the University of Pennsylvania, dedicated to applying scientific research methods to the study of democracy and political polarization Iyengar, Lelkes, and Westwood (2024). The America's Political Pulse Survey interviews 1000 American adults each week and asks them questions that aim to track affective polarization, support for violations of democratic norms, and support for the use of political violence in America Iyengar, Lelkes, and Westwood (2024). High-level survey results are available on the Polarization Research Lab website.

Our post-stratification data is a subset of the 2022 American Communities Survey (ACS) provided by the Integrated Public Use Microdata Series (Ruggles et al. 2024).

2.1 Survey Data

We selected the America's Political Pulse Survey Data Week 3 2024 from the Polarization Research Lab as our survey data. It contains the responses of the 1000 American adults interviewed between January 12-19 2024. Samples were collected via interviews. The population of the survey, the respondents, are paid survey-takers from the YouGov survey platform Iyengar, Lelkes, and Westwood (2024). Demographic information about respondents, including sex, age, race, home state, employment status, marital status, and the size and type of city that they live in is included. Participants were also asked questions about who they voted for in the 2016 and 2020 U.S. presidential elections, stances on political violence, what their party affiliation is, the strength of their party affiliation, and what their political ideology is. While the America's Political Pulse survey aims to learn about affective polarization and respect for democratic norms by asking respondents questions regarding their feelings about the importance of voting, towards the Democratic Party and the Republican Party, faith in democratic institutions, and about the use of political violence, we are focusing on demographics and political preferences. As expected, responses vary by ideology and political affiliation.

We are focusing on demographics and other variables that might indicate who each respondent will vote for in the 2024 presidential election. The variables from the America's Political Pulse Survey that we selected include: `pid7`: the respondent's party affiliation and the strength of it; `presvote16post`: who the respondent voted for in the 2016 U.S. presidential election; `pres20votepost`: who the respondent voted for in the 2020 U.S. presidential election; `gender`: the gender of the respondent; `age`: the respondent's age as of January 2024; `race`: respondent's race or ethnicity; `educ`: highest level of education completed; `inputstate`: state of residence; `urbancity2`: whether or not the respondent lives in a city, and if so, the size of that city.

The survey does not contain a question about a preferred 2024 Presidential Candidate. In fact, it was conducted in January 2024, before Super Tuesday took place and Donald Trump had secured almost enough delegates to win the Republican (GOP) nomination. We created a variable `vote_biden`, which is equal to 1 if a respondent's preferred 2024 presidential candidate is Joe Biden, and 0 if it is Donald Trump. The data cleaning process used for the creation of the `vote_biden` variable is outlined in Section .1.

In Figure 1, we can see that support for Biden within our survey data varies by race and sex. Black women showed overwhelming support for President Biden's upcoming re-election campaign, while white men appear to be split between Biden and Trump. Among white survey respondents (69% of all respondents), women were more likely to support President Biden and men were more likely to support Trump. We used code from Chapter 13 of Telling Stories with Data and modified it to produce the bar graphs in this section (Alexander 2023).

We see differences in vote preference among survey respondents with different levels of education. Male voters with up to a high school education or some college education were more likely to prefer Trump than their 4-year college or post-graduate educated counterparts. Only male voters whose highest level of education is a 4-year college degree favored Biden over Trump, while males with a post-graduate education were split. Females with a post-graduate education heavily favored Biden, further highlighting the gender-gap in support for Biden and Trump. This makes sense, as Trump is strongly opposed to abortion rights and since the Supreme Court of the United States overturned Roe v. Wade in June 2022, there is nothing to stop him from signing an outright ban on abortion, contraceptives, or fertility treatments if elected. Females with a 4-year college or post-grad education were more likely to prefer Biden than those with only a high school education.

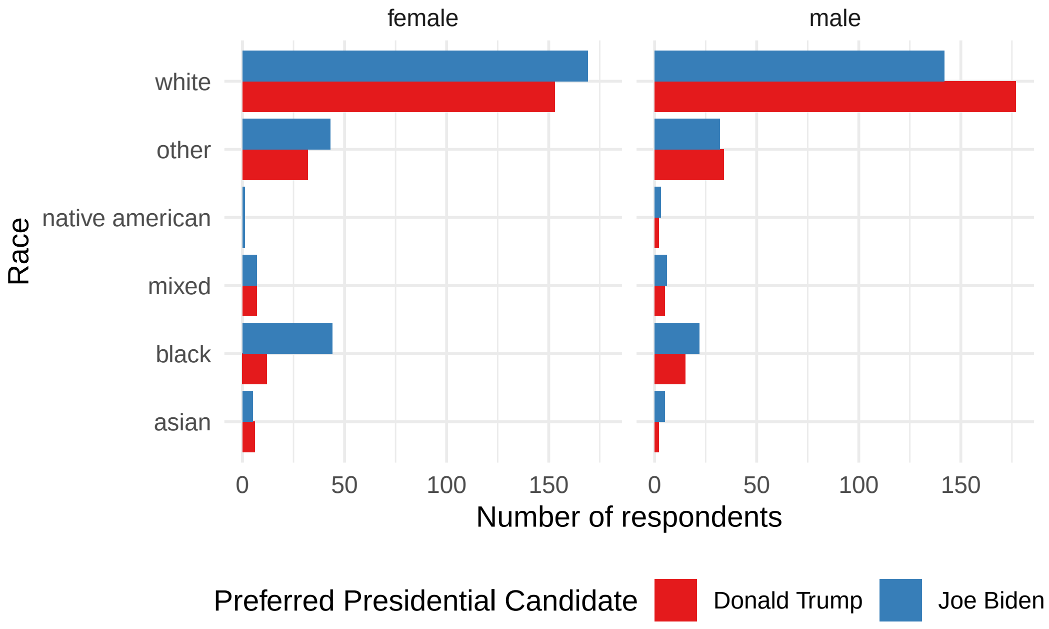


Figure 1: Preferred presidential candidates of survey respondents, by gender and race

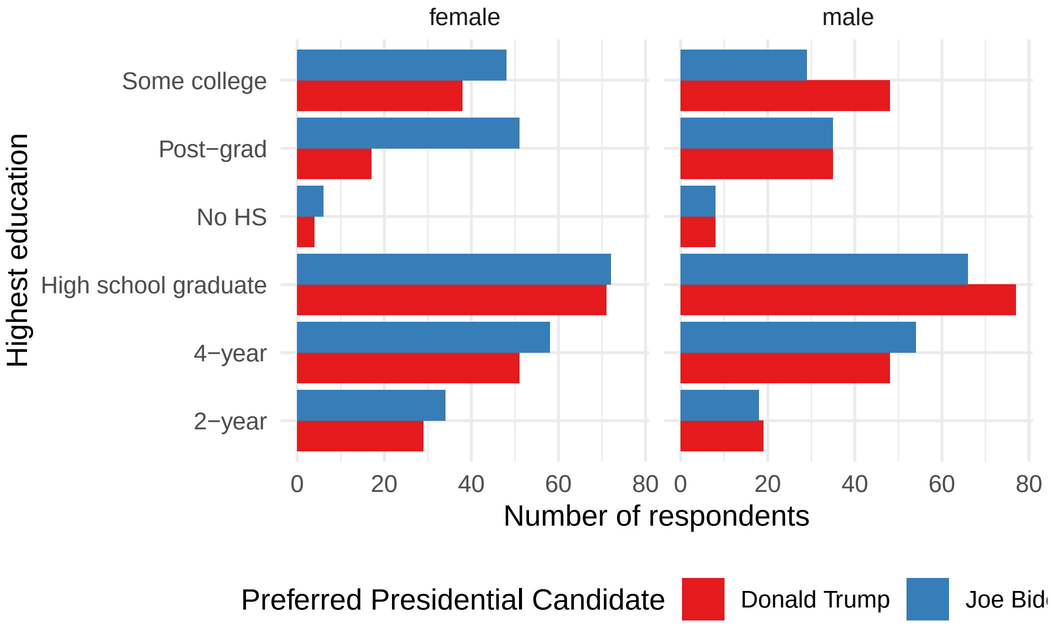


Figure 2: The Distribution of Presidential Preferences, by gender and highest level of education

Since the 1990s, America has developed an urban-rural ideological divide and it has kept growing (Cornellians Staff 2022). We created a new binary variable, `urban`, indicates whether the respondent lives in an urban or rural area. The data cleaning steps that we used to create this variable are outlined in Section .1. First, we added the `urban` variable to our survey data. It is equal to "urban" if `urbancity2` indicates that the respondent lives in a big city, a smaller city, or a suburb, and "rural" if the respondent lives in a small town or rural area. This effectively distinguishes city and suburban voters from small town and rural voters, making our analysis simpler. We found that female survey respondents living in urban areas heavily favored Biden, while male survey respondents living in rural areas heavily favored Trump. In Figure 3, we can see that while female survey respondents living in rural areas were split and male survey respondents living in urban areas favored Biden by a smaller margin, females in urban areas heavily favored Biden and males in rural areas heavily favored Trump.

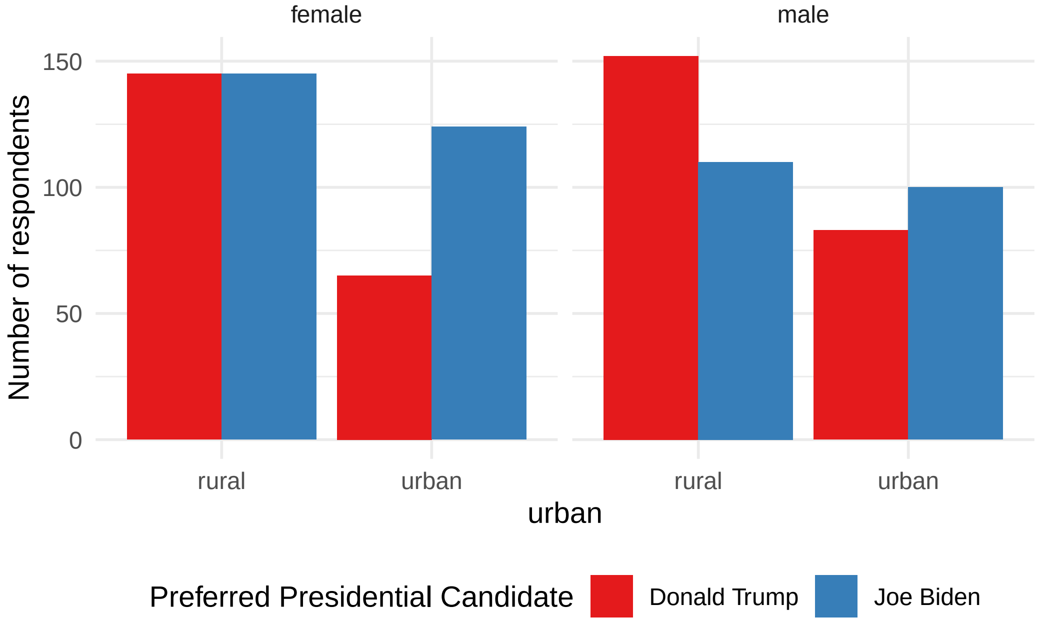


Figure 3: Preferred presidential candidate of respondents living in urban vs rural areas

Overall, Biden outperformed Trump in urban areas, with the support of 60.22% of survey respondents living in urban areas. This is unsurprising, and it is a continuation of the pattern observed in recent presidential elections. Scala and Johnson (2016) found that both former President Barack Obama and 2016 Democratic nominee Hillary Clinton performed well in densely populated urban areas, but faced increasingly challenging political climates and diminished voter support in rural areas. Hillary Clinton saw a particularly pronounced decline in support in rural areas in the 2016 election, where she was defeated by Donald Trump (Scala, 2016). Our survey data is somewhat consistent with this, as Biden trails Trump among survey respondents living in rural areas, with 46.20% support.

Using only the survey data, we can calculate the overall support for Biden and the support

Table 1: Presidential preferences of respondents living in urban and rural areas

Urban or Rural:	Biden %	Trump %
Urban	60.22	39.78
Rural	46.20	53.80

for Biden in each state to make a popular vote and electoral college prediction. Based on our earlier findings, we believe that being non-white, female, college or post-grad educated, and living in an urban area makes an individual more likely to support Biden. To calculate a popular vote prediction for President Biden based on the survey data, we simply calculated the mean of the `vote_biden` binary indicator variable, and to calculate an electoral college prediction, we added up the electoral college votes of the states where Biden received more than 50% support from survey respondents. We used the steps and code provided by Mitrovski, Yang, and Wankiewicz (2020) and the `statebins` package (Rudis 2020) to produce Figure 4.

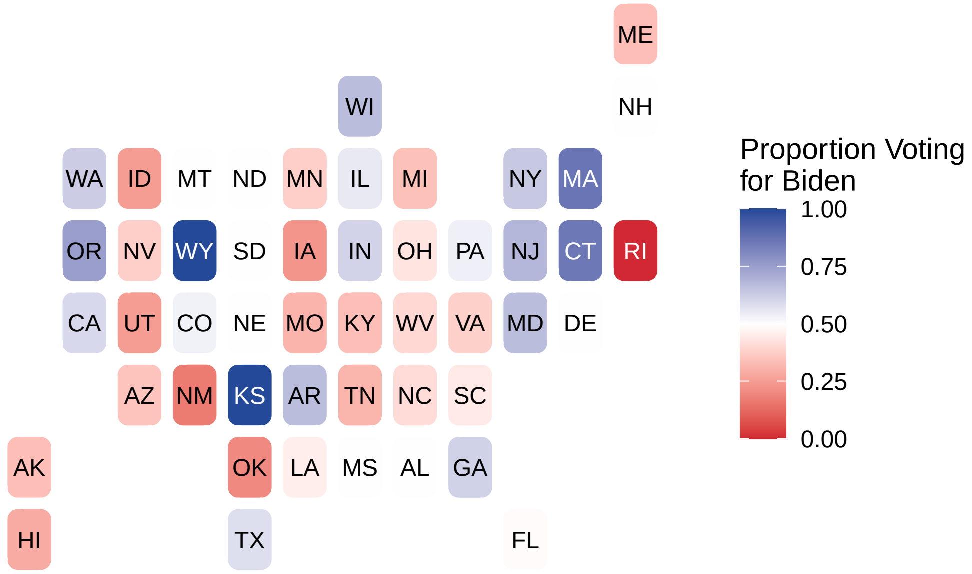


Figure 4: Electoral college map based on the survey data

Figure 4 illustrates the proportion of voters in each state who plan to support President Biden in the 2024 election. We can see that there is very strong support for President Biden in “deep blue” states such as Massachusetts, Connecticut, New Jersey, and Oregon. A majority of respondents in New York, New Jersey, Maryland, and California also support President Biden. On the contrary, Oklahoma, Idaho, Utah, and Alaska appear to be “deep red” states, which indicates that a high proportion of survey respondents from those states support former President Trump. Notably, Wyoming appears to be dark blue, indicating that 100% of survey respondents from that state support President Biden. However, Wyoming is a Republican stronghold, and the only Democratic presidential candidate who has won it since 1948 is Lyn-

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

Survey Estimate:	Biden	Trump
Num Votes	479.00	445.00
% Votes	51.84	48.16
Electoral College	324.00	211.00

don B. Johnson in 1964 (CNN Politics 2020). Wyoming was the most pro-Trump state in both the 2016 and 2020 presidential elections. Trump outperformed 2016 Democratic nominee Hillary Clinton and President Joe Biden in Wyoming by 46 and 43.3 percentage points, respectively (CNN Politics 2020). Our survey data contains one respondent from Wyoming, who leans Democrat and voted for Hillary Clinton in 2016 and Joe Biden in 2020. There are over half a million people in Wyoming, and Figure 4 reflects the political preferences of exactly one of them. Similarly, Hawaii (7 respondents) appears to be a "red" state, indicating that the majority of survey respondents from that state prefer former President Trump. However, Hawaii is a Democratic stronghold and it has only been carried by a Republican presidential candidate twice since it became a state in 1959 (CNN Politics 2020). President Biden and Hillary Clinton won 63.7% and 62.2% of the vote in Hawaii in 2020 and 2016, respectively (CNN Politics 2020).

Table 2 summarizes our predictions for the popular vote and electoral college based on our survey data. We can use this to predict that Joe Biden will win 51.84% of the popular vote, and win re-election by defeating Donald Trump 324 to 211 in the electoral college.

Our survey data set contains zero respondents from Vermont, so it is not shown Figure 4 and the number of electoral colleges in Table 2 adds up to 535 instead of 538 because Vermont's 3 electoral votes were not included. The absence of any respondent from Vermont and the fact that there are very few responses from Kansas, Hawaii, Rhode Island, and Wyoming may affect the accuracy of this prediction. This is a significant weakness, and we will discuss it in more depth, as well as how it contributes to the weaknesses and limitations of our model in Section 5.2.

2.2 Post Stratification Data

Our post-stratification data is a subset of the American Community Survey (ACS) 2022 provided by IPUMS (Ruggles et al. 2024). We visited the IPUMS data base, selected the ACS 2022, selected relevant variables, and downloaded the data set. Our raw post-stratification data set has 500,000 records. The variables that we selected mirror the ones that we have in our survey data set: `age`: the age of the census respondent; `sex`: only male and female are considered, for simplicity and to align with our survey data; `race`: the race of the census respondent; `hispan`: whether the respondent is of Hispanic origin or not; `educ`: highest level of education completed; `stateicp`: state of residence; `metro`: whether the respondent lives in a metropolitan area or not.

The sex, race, educ, and metro variables were cleaned and re-factored to mirror our survey data and to allow us to apply our model to our post-stratification data. Figure 5 shows the demographic statistics of our survey and post-stratification data sets. The steps and code for producing Figure 5 were obtained from Mitrovski, Yang, and Wankiewicz (2020). We also created a `race_hispanic`, which is equal to either "Hispanic" or "not Hispanic" to ensure that our survey and post-stratification data sets would align, because this was a separate variable from race in our post-stratification data. The data cleaning steps used to create the `urban` variable to make our post-stratification data set mirror our survey data set are outlined in Section .1.

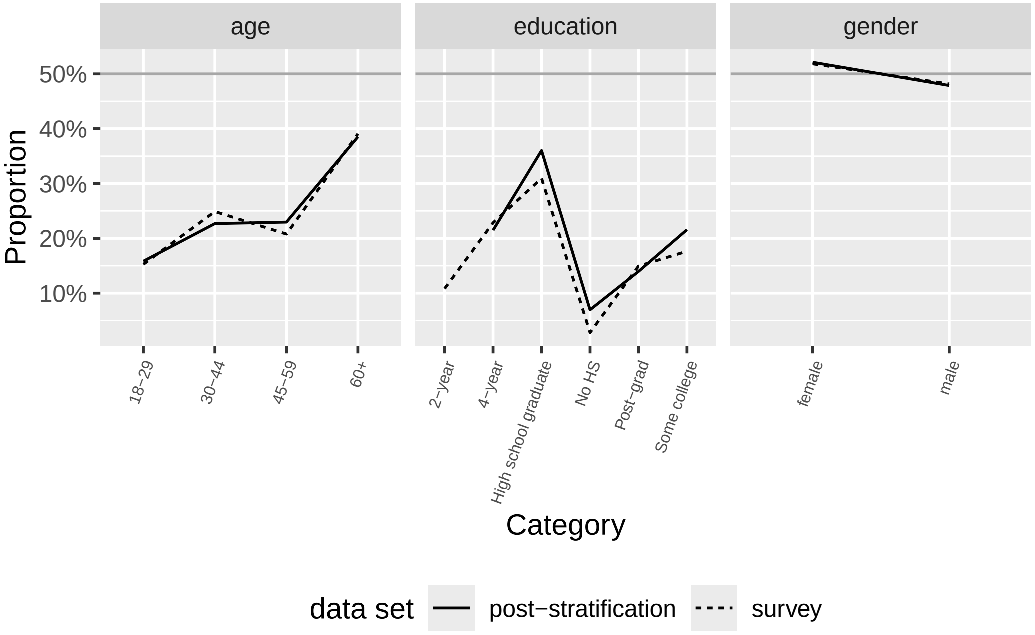


Figure 5: Survey vs post-stratification voter demographics

Figure 6 illustrates the demographic patterns in our post-stratification and survey data sets. The steps and code for producing Figure 7 were obtained from Mitrovski, Yang, and Wankiewicz (2020).

Overall, race demographics are consistent between our survey and post-stratification data sets. The proportion of Hispanic respondents in our survey and post-stratification data is comparable, at 12.7% and 13.5%, respectively. While there is a higher proportion of Hispanic respondents in our post-stratification data set, the difference is less than one percentage point. The proportions of white respondents in our survey (69.4%) and post-stratification (69.5%) data sets are nearly identical. Figure 7 illustrates the distribution of voters across the 50 states, including Washington D.C. in both our survey and post-stratification data sets. The steps and code for producing Figure 7 were obtained from Mitrovski, Yang, and Wankiewicz (2020).

We can see that there are some slight differences in the proportions of respondents in the

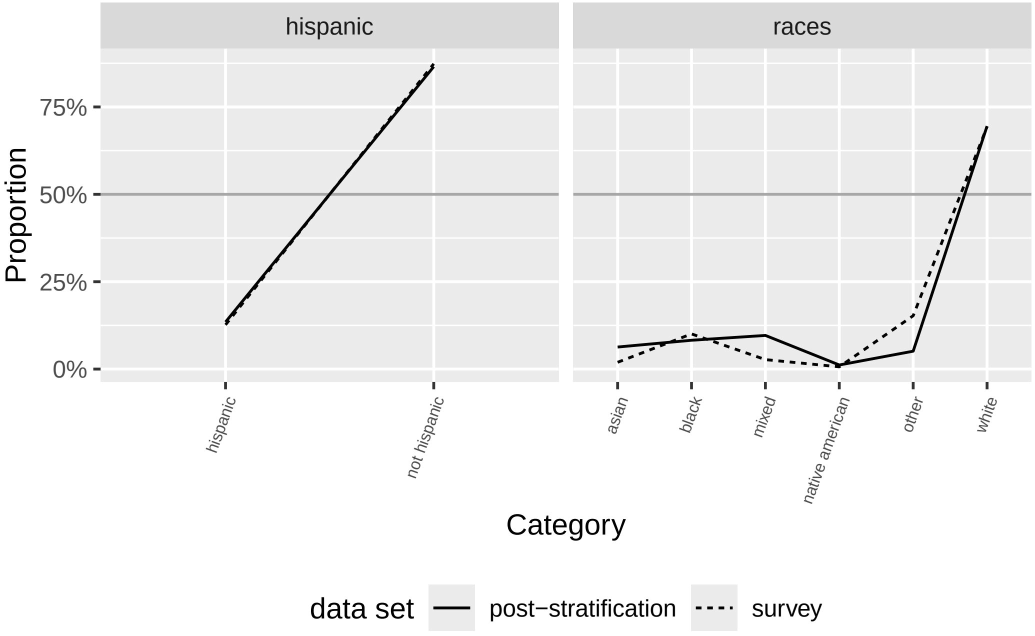


Figure 6: Survey vs post-stratification voter race demographics

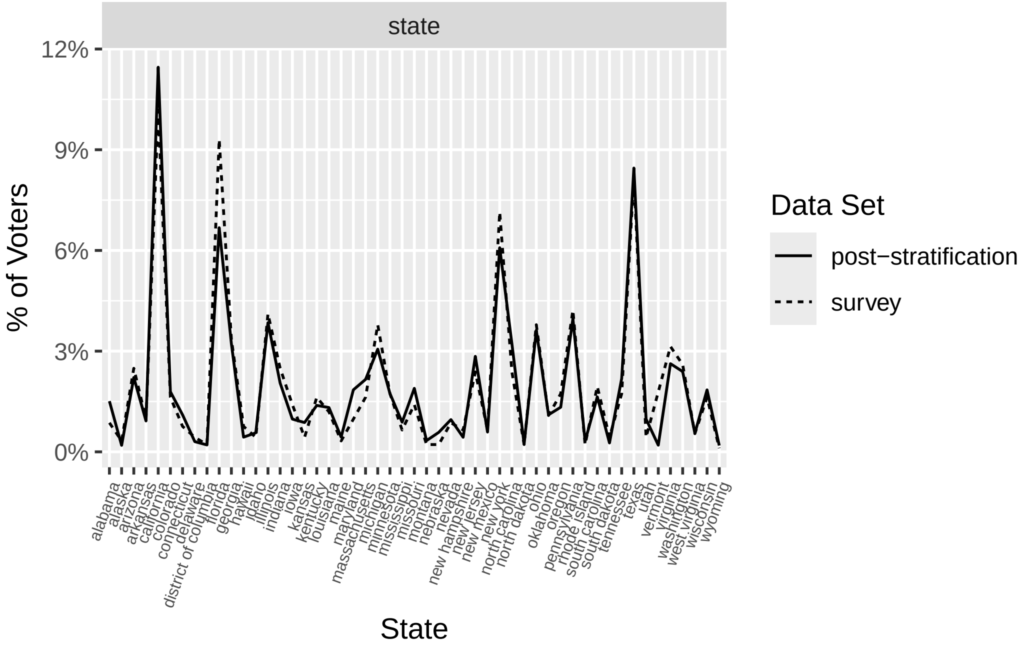


Figure 7: Survey and Post-Stratification Data Proportion of Voters by State

survey and post-stratification data from Alabama, California, Florida, Kansas, Maryland, Massachussetts, Michigan, New York, Nebraska, and Virginia. None of the voters in our survey data are from Vermont, but 0.198% of the voters in our post-stratification data are. Therefore, the results from applying our model to our post-stratification data will include a prediction for Vermont. Figure 7 shows that there is a comparable distribution of voters across U.S. states in the survey and post-stratification data. However, the survey data over represents Florida, Michigan, New York, and Virginia.

3 Model

We performed multi-level regression with post-stratification (MRP) to predict support for Joe Biden in the upcoming election. To perform MRP, we fit a logistic regression model to predict support for Biden using sex, age, race, race_hispanic, highest level of education, state, and urban as predictors on our survey data set (Iyengar, Lelkes, and Westwood 2024), and post-stratified it using ACS census data (Ruggles et al. 2024).

3.1 Model set-up

We built our Bayesian logistic regression model using the `stan_glm()` function of the `rstanarm` package (Goodrich et al. 2022). The model that we use is:

$$vote\_biden_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1\text{sex}_i + \beta_2\text{age\_bracket}_i + \beta_3\text{race}_i + \beta_4\text{race\_hispanic}_i + \beta_5\text{education\_level}_i + \beta_6\text{urban}_i \tag{2}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

$$\beta_4 \sim \text{Normal}(0, 2.5) \tag{7}$$

$$\beta_5 \sim \text{Normal}(0, 2.5) \tag{8}$$

$$\beta_6 \sim \text{Normal}(0, 2.5) \tag{9}$$

where binary variable $vote\_biden_i$ is equal to 1 if the respondent's preferred 2024 presidential candidate is Joe Biden, or 0 if the respondent's preferred 2024 presidential candidate is Donald Trump. We run our model in R (R Core Team 2023) using `stan_glm`, with the default priors from `rstanarm` (Brilleman et al. 2018). We then apply our model to our post-stratification data (Ruggles et al. 2024) to predict the popular vote and electoral college results of the 2024 U.S. Presidential election.

Our model uses logistic regression, therefore one of its weaknesses is the fact that it predicts a binary outcome and does not consider the possibility that some American adults might vote for a third-party candidate or abstain from voting. Tradeoffs and concerns about our model are discussed in more detail in Section 5.2.

3.2 Model Justification

We expect to see a positive relationship between support for President Biden and non-white race, college or post-graduate education, and living in an urban area. This expectation is based on the fact that Trump has spent much of his political career spewing anti-immigrant rhetoric and America's urban/rural ideological divide. In 2016, Trump campaigned on the promise of building a wall at the U.S.-Mexico border, and during his presidency, Trump labeled Mexicans as rapists and drug smugglers, referred to the coronavirus as the "Kung-flu" and the "China virus." Ghitza and Robinson (2020) found that 39% of Biden's supporters in the 2020 election were voters of color, compared to only 15% for Trump. Black and Asian-American voters turned out at an increased rate, heavily in favor of President Biden in 2020 (Ghitza and Robinson 2020). Scala (2016) found that even after social, demographic, and economic factors including education, age, and race were controlled for, the urban-rural divide was still statistically significant in estimating vote patterns in U.S. presidential elections. The 2020 election was no exception, as Trump had strong support in rural areas and Biden dominated in major cities (Ghitza and Robinson 2020). In Section 2.1, we found that respondents from rural areas were more likely to support Donald Trump, while respondents from urban areas were more likely to support Joe Biden.

4 Results

4.1 Popular Vote Prediction

We got our popular vote prediction by taking the weighted average of vote_biden across all U.S. states and converting it into a percent. The average is weighted according to the number of census respondents in our post-stratification data set from each state. Table 3 summarizes our mean, lower, and upper estimates for the popular vote percentage and number of electoral college votes that Joe Biden will receive on November 5th. Based on the results of our multilevel regression with post-stratification (MRP), which involved applying the logistic regression outlined in Section 3 to our post-stratification data set, we predict that Joe Biden and Donald Trump will receive 56.59% and 43.41%, respectively, of the popular vote. That means that our logistic regression model classified 56.59% of the people included in our subset of the ACS 2022 (Ruggles et al. 2024) as Biden voters based on sex, age bracket, race, highest level of

Table 3: 2024 U.S. election result estimates for Joe Biden based on post-stratification

(a) Popular Vote

Estimate:	Biden %	Trump %
Lower Estimate	51.43	48.57
Mean Estimate	56.59	43.41
Upper Estimate	61.42	38.58

(b) Electoral College

Electoral College Estimate:	Biden	Trump
Lower Estimate	363	175
Mean Estimate	471	67
Upper Estimate	538	0

education, and whether they live in an urban or rural area. Table 3 shows our lower, mean, and upper estimate for the percent of the popular vote that Joe Biden will receive. Conversely, the lower estimate for Joe Biden corresponds to the upper estimate for Donald Trump, and the upper estimate for Joe Biden corresponds to the lower estimate for Donald Trump.

4.2 Electoral College Prediction

The estimates for the support for Biden in each state are shown in Figure 8. We followed and modified the code provided in Chapter 16 of "Telling Stories with Data" (Alexander 2023) to produce Figure 8. The red or blue line for the predicted proportion of Biden votes in each state represents the 97.5% prediction interval. The mean estimate for the percentage of voters in each state supporting Biden is denoted by the blue or red dot at the center of each line. The line is blue if the mean prediction is over 50%, indicating that the state's electoral college votes would go to Joe Biden, and red if it is below 50%, indicating that the state's elecotral college votes would go to Donald Trump. This interval crosses the 50% line for numerous states, which means that based on our model, they could be toss-ups. We can see that the prediction interval is widest for Hawaii, which ranges from 44.5% to 66.9% with the mean prediction of 56.1%, and Alaska, which ranges from 41.4% to 63.6%, with a mean prediction of 53.2%

Figure 9 illustrates the same information as Figure 8, except it includes gray dots that represent the proportion of voters in each state that support Biden from the survey data. We followed and modified the code provided in Chapter 16 of "Telling Stories with Data" (Alexander 2023) to produce Figure 9.

We can see that there is a lot of variation between the post-stratification and survey estimates for the proportion of voters in each state that will vote for President Biden in the upcoming election. However, when we consider only the post-stratification estimates and add up the electoral college votes for the states in which the mean estimated support for Biden is over

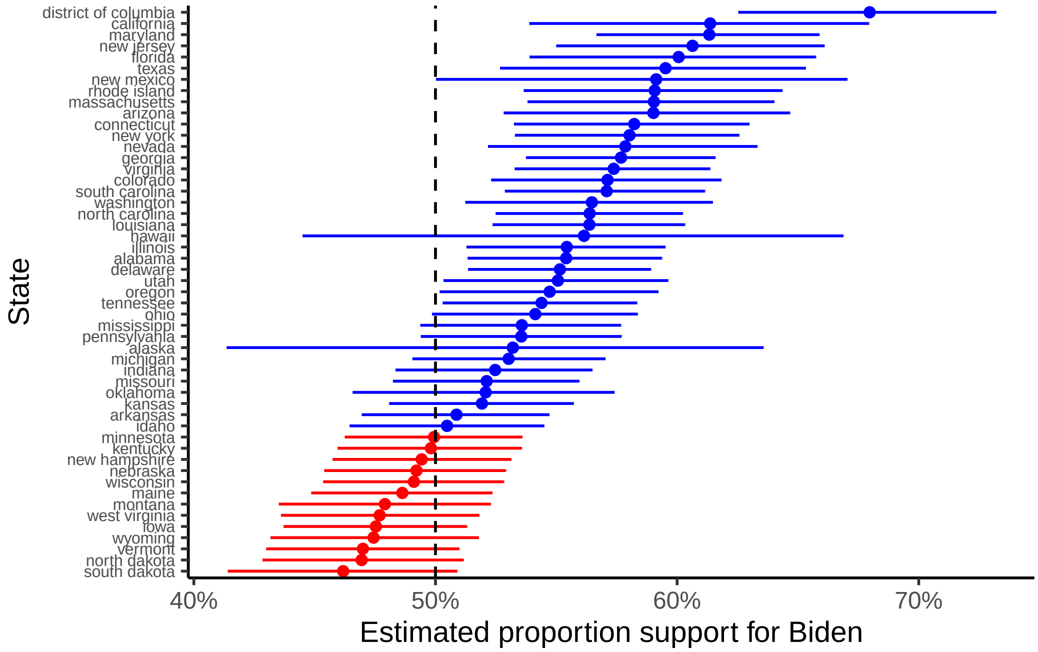


Figure 8: Estimated proportion of each state voting for Biden in 2024 (Post-Stratification)

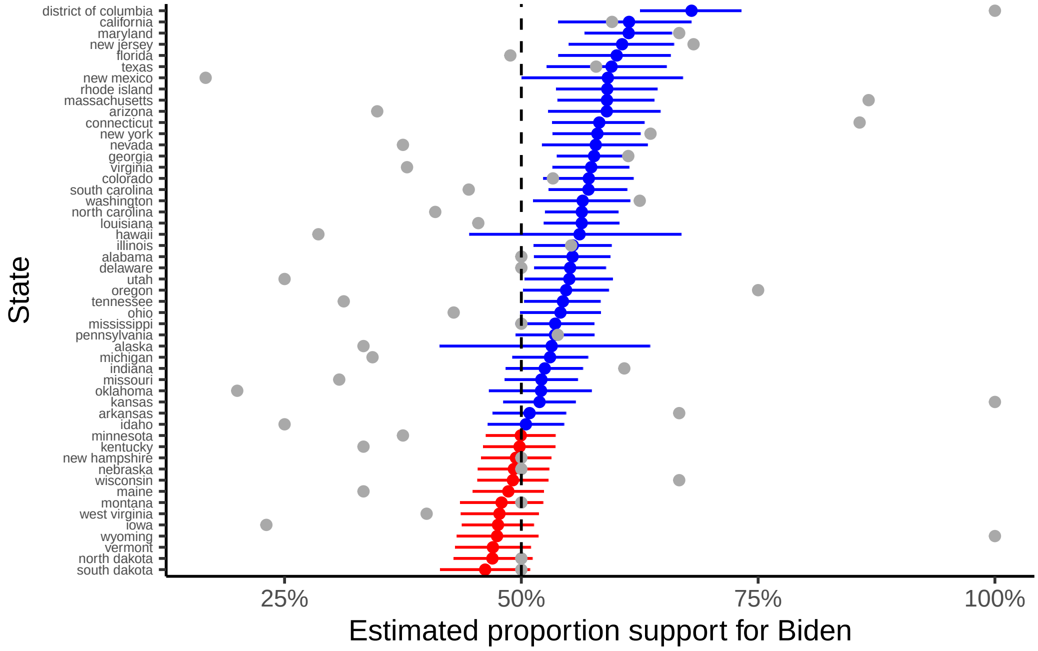


Figure 9: Estimated proportion of each state voting for Biden in 2024 Post-Stratification vs Survey Data

50% (shown in blue on Figure 8 and Figure 9), we can predict that Joe Biden will defeat Donald Trump in the electoral college 471 to 67. Table 3 shows the electoral college predictions based on the lower end and higher end of these intervals. At the lower end, we can predict that Joe Biden will defeat Donald Trump in the electoral college 363 to 175, and at the upper end, we can predict that Joe Biden will defeat Donald Trump in the electoral college 538 to 0.

Figure 10 illustrates the same information as Figure 8, but on an electoral map. We used the steps and code provided by Mitrovski, Yang, and Wankiewicz (2020) and the `statebins` package (Rudis 2020) to produce Figure 10. This map shows the mean prediction for the proportion of voters in each state prefer Joe Biden over Donald Trump. We see that California (CA), Maryland (MD), and New Jersey (NJ) are dark blue, indicating strong support for Joe Biden in those states. Wyoming (WY), South Dakota (SD), North Dakota (ND), and Montana (MT) are red, which indicates support for Donald Trump. In addition to this, Texas and Florida are blue in Figure 10. The Democrats have not won Florida since former President Barack Obama won when he was re-elected in 2012. Donald Trump carried Florida in both 2016 and 2020, so flipping it in 2024 would be huge for President Biden. The Republican presidential nominee has carried the state of Texas in every presidential election since 1976 (CNN Politics 2020), so flipping it in 2024 would give Biden an excellent chance of winning re-election.

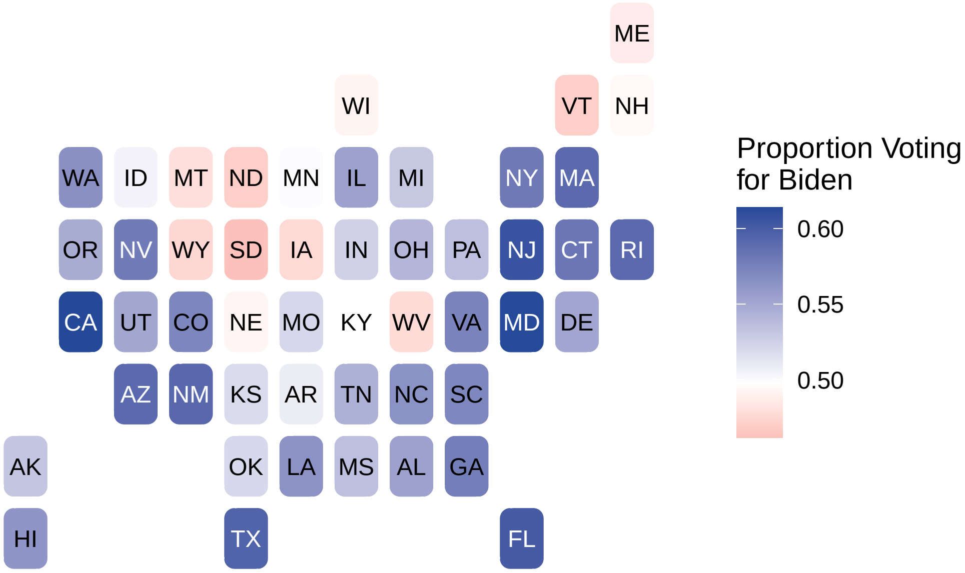


Figure 10: Electoral map based on post-stratification data

5 Discussion

5.1 The popular vote does not necessarily define the electoral college

When discussing the results of our model, it is crucial to consider the quality of the survey data that was used to fit it. Despite projecting Joe Biden to win an impressing, but not dominant, 56.59% of the popular vote in the 2024 presidential election when applied to the post-stratification data, our model produced a mean estimate of 471 electoral college votes for Joe Biden. We know that winning a certain percentage of the popular vote does not translate into a fixed numer of electoral college votes. Instead, the candidate with the most votes in each state receives all of its electoral college votes, and in America's two-party system, the candidate whose electoral votes add up to 270 or more wins the presidency. So, in theory, it is possible to win 538 electoral votes with exactly 50.1% of the vote in each state. However, this scenario is extremely unlikely. It is also possible to win the popular vote, but still lose the electoral college and the presidency. In fact, Hillary Clinton won the popular vote but lost the electoral college to Donald Trump, 232 to 306, in 2016.

The shortcomings of our survey data set are highlighted in Figure 9, where we can see that there is no visible pattern or relationship between predictions for the percent of voters that will support Biden in each state based on the survey and post-stratification data sets. The gray dots, which represent our prediction for the percent of voters who prefer Biden in each state based on our survey data set, are scattered all over the place. We used that same survey data set to fit the model that was applied to our post-stratification data set to produce the blue (where the mean prediction for Biden support was over 50%) and red (where the mean prediction for Biden support was below 50%) prediction intervals that show which candidate would win each state's electoral college votes. While state was not a predictor for our model, it is clear that in many cases, our survey and post-stratification data sets produced very different predictions for the percentage of votes that Biden will receive in many states.

The lower estimate for the number of electoral college votes that Biden will receive based on the results of our MRP in November is 363, while the upper estimate is 538, as shown in Table 3. That is a difference of 175 seats, which gives us little confidence in our current model's ability to accurately predict the electoral college. In comparison, our prediction for the number of electoral college votes that President Biden will win in the 2024 election based on our survey data is 324, as shown in Table 2. 471 electoral college votes is not even relatively comparable to the 306 that he received in 2020, or the number of electoral college votes that presidents have received in recent elections (CNN Politics 2020). Therefore, there is reason for us to believe that the seemingly random and far-off predictions for the proportion of votes that Biden will receive in numerous states based on our survey data limits our model's ability to accurately predict the electoral college.

5.2 Weaknesses and Limitations

Our model can only ever be as good as the data set that we used to build and fit it (Alexander 2023). The weaknesses of our model stem directly from the limited nature of the survey data set that we used, the America's Political Pulse Survey Week 3 2024. As seen in Section 2.1, the electoral college predictions from Kansas (4 respondents), Wyoming (1 respondent), and Hawaii (7 respondents) are drastically different from the results seen in recent elections. There are very few respondents from smaller states with fewer electoral college votes, such as Kansas, Wyoming, Hawaii, New Hampshire, and Rhode Island in our survey data set. In addition to this, there are no respondents from Vermont in our survey data set. As mentioned, there were no respondents from Vermont and very few respondents from several states with few electoral college votes in our survey data set. As a result, we were not able to include state as a predictor in our model. Including state as a predictor would not have been applicable to respondents in our post-stratification census data from the state of Vermont at all.

In practice, including state and considering third-party candidates would make sense, but it was not worth the trade-offs for the America's Political Pulse survey data set and our specific model. We initially did include state as a predictor, because we know that there are historical Democratic stronghold states, such as California, Illinois, Hawaii, Connecticut, Massachusetts, and Vermont, as well as historical Republican stronghold states such as Idaho, Utah, Kansas, Alabama, and Wyoming (CNN Politics 2020). However, this model resulted in warnings about the bulk effective sample size and the tail effective sample size being too low. This means that our model with state may have had unreliable posterior means, posterior medians, posterior variances, and tail quantiles (Stan Development Team 2022). This makes sense, as fitting a model that used state as a predictor on our specific survey data would have associated Kansas and Wyoming with support for President Biden and Hawaii with support for former President Trump. This would be reflective of the political preferences of fewer than five respondents from each of those states. It would fail to consider the historical Republican stronghold on Kansas and Wyoming and Democratic stronghold on Hawaii in presidential elections (CNN Politics 2020).

Logistic regression can only be used for binary classification. This limits the scope of this paper, as it only allows us to consider support for Joe Biden and Donald Trump in the upcoming 2024 presidential election. Our model does not consider support for third-party candidates, write-in votes, or the possibility of not voting, which could be a disadvantage when predicting results in swing states or for close elections. In the 2016 election, Donald Trump defeated Hillary Clinton by 11,612 votes (0.3 percentage points) in Michigan, which is considered a swing state. Third-party candidates Gary Johnson (Libertarian) and Jill Stein (Green) each received over 173,000 votes (3.6% of the popular vote) and 50,700 votes (1.1% of the popular vote) in Michigan (Politico 2016). This is just one example of at least one third-party candidate receiving more votes than the difference in votes between the Democratic and Republican nominees in a key battleground state. It highlights a major limitation of using logistic regression to predict vote preference.

5.3 Next Steps

We can address our model's inability to consider more than two voting outcomes by fitting a SoftMax regression model. SoftMax regression is a generalization of logistic regression that can predict the class probabilities for more than two classics and perform multi-class classification. However, a SoftMax regression model is more powerful than a logistic regression model. It would severely overfit our current survey data set because it only has 1000 observations, including less than five from some smaller states. While a SoftMax regression model might make sense for predicting vote preference when there are more than two candidates, this model would have been far too powerful for our survey data set and not worth the trade-offs. The bulk effective sample size and tail effective sample size would have been too low, which would mean that our posterior means, posterior medians, posterior variances, and tail quantiles would be unreliable (Stan Development Team 2022). We can justify using a logistic regression model because the United States has a two-party system, meaning that Joe Biden (Democrat) and Donald Trump (Republican) are the only candidates who can realistically win the presidency on November 5th. Although there will be third-party candidates, we do not yet know who they will be, nor are they expected to win any electoral college votes.

With a survey data set that is as limited as the one we currently have, SoftMax regression is not a viable option. However, the possibility of third-party candidates receiving vote shares larger than the margin of victory of the Democratic or Republican nominee in key battleground states highlights the need to consider the fact that a small percentage of voters will cast a ballot for someone not named Donald Trump or Joe Biden on November 5th. Therefore, the first step towards extending and improving this report in the future would be to find a survey data set that has significantly more respondents, is somewhat reflective of voting trends and demographics in each state, and specifically asks respondents who they plan to vote for in the 2024 U.S. presidential election. Once we address the issue of having very few, or even no respondents from smaller states and no question about 2024 vote intention, we will be able to build a stronger and more accurate model to predict vote choice based on state, urban or rural area, demographics such as sex, age, and race, and education. A robust survey data set with enough respondents (possibly upwards of 10,000) will allow us to train a more powerful SoftMax regression model to perform multi-class classification and classify voters as Trump, Biden, third-party, or abstaining voters using similar predictors to the ones outlined in Section 3 as well as state. We believe that we can extend and improve our study by using SoftMax regression, but it is crucial for us to find a better survey data set first.

Appendix

.1 Data Cleaning

We created a binary variable, vote_biden to indicate whether each survey respondent's preferred 2024 presidential candidate is Joe Biden or Donald Trump. vote_biden is equal to 1, which indicates that the respondent's preferred presidential candidate is Joe Biden, if their party affiliation is Democratic or if they voted for Joe Biden in 2020. vote_biden is equal to 0 otherwise, indicating that a respondent's preferred presidential candidate is Donald Trump. Donald Trump is arguably the most polarizing figure in American politics, therefore we are considering independents who voted for Joe Biden in 2020 to prefer Joe Biden over Donald Trump again in 2024. However, vote_biden will never be equal to 1 if the respondent voted for Donald Trump in 2016 or 2020. In the context of this report and the vote_biden variable, voting for a third-party candidate, a write-in vote, or not voting are not considered possible voting outcomes.

We also created a new variable, urban. First, we added the urban variable to our survey data. It is equal to "urban" if urbancity2 indicates that the respondent lives in a big city, a smaller city, or a suburb, and "rural" if the respondent lives in a small town or rural area. For our post-stratification data set, we also gave the urban variable two levels, "urban" and "rural" to align with our survey data set. If the metro variable indicated that a census respondent lives in a metropolitan area, whether inside or outside of the principal/central city, urban is equal to "urban," and we set it to rural for census respondents who do not live in a metropolitan area.

.2 Model Details

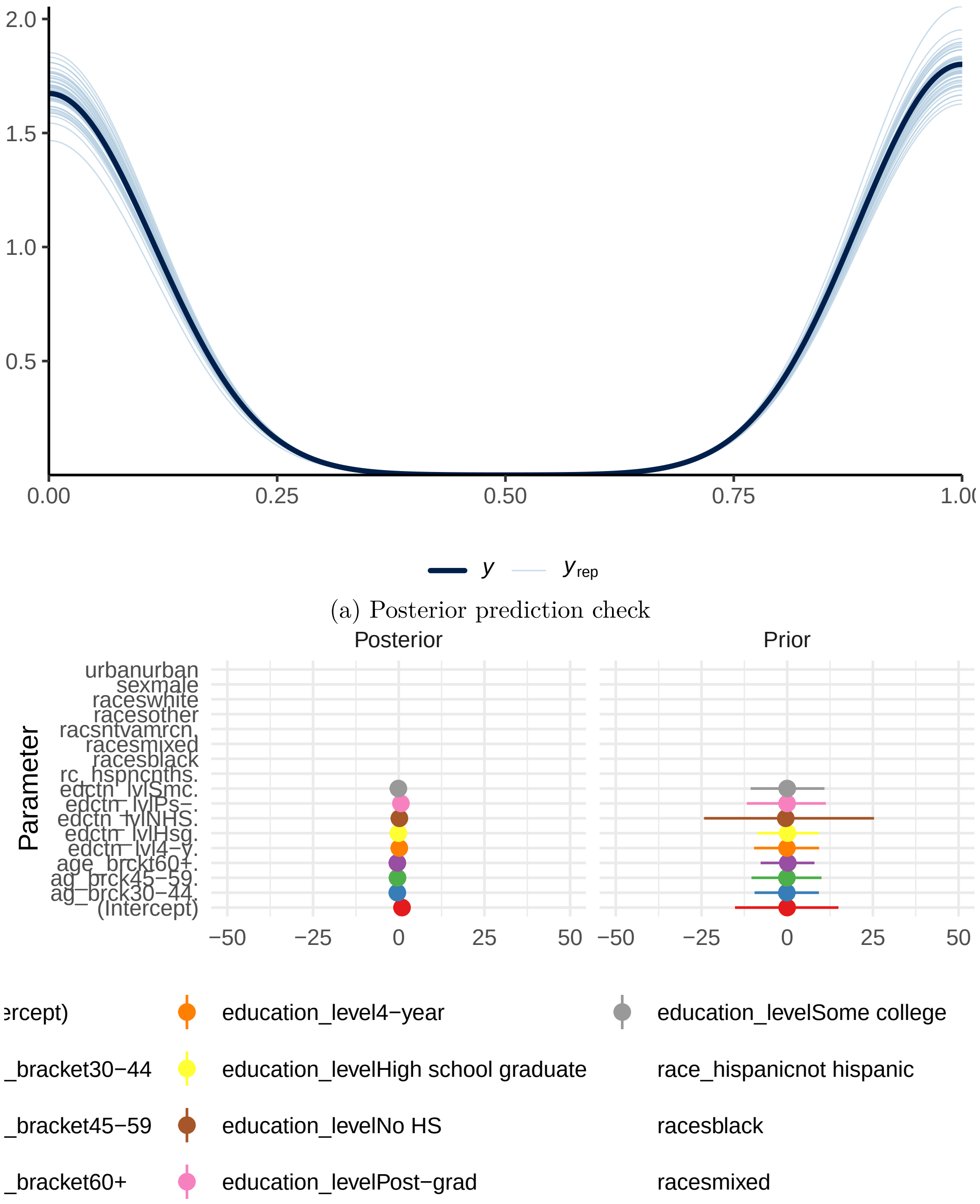
.2.1 Posterior Predictive Check

In Figure 11a we implement a posterior predictive check. In Figure 11b we compare the posterior with the prior.

.2.2 Markov Chain Monte Carlo Convergence

We check for signs that the Markov chain Monte Carlo (MCMC) algorithm ran into issues using an Rhat plot and a trace plot (Alexander 2023).

Figure 12a is an Rhat plot. It shows everything is close than 1 and less than 1.05. This means that the coefficients all converge to the same distribution, and the model is able to predict vote_biden in the survey data.



(a) Posterior prediction check

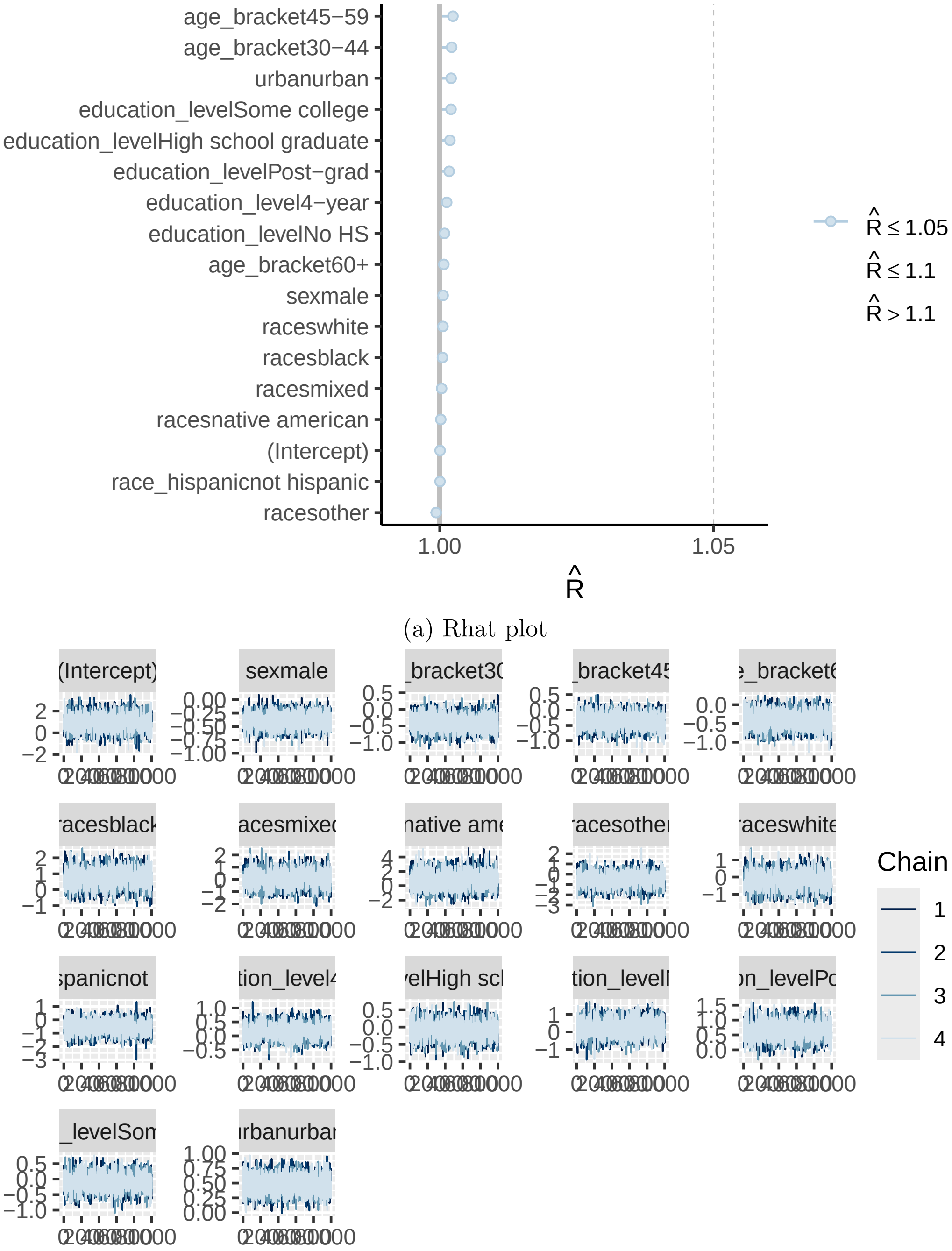
(b) Comparing the posterior with the prior

Figure 11: Examining how the model fits, and is affected by, the data

Figure 12b shows the trace plots for our predictors. We can see that the lines bounce, but remain horizontal and have a lot of overlap with each other (Alexander 2023). This suggests that the distribution behaves as expected, and we do not need to re-run our model with fewer predictors or different priors (Alexander 2023).

.2.3 Credibility Intervals

In Figure 13, we show the 90% credibility intervals for the predictors of `vote_biden`.



(a) Rhat plot

(b) Trace plot

Figure 12: Checking the convergence of the MCMC algorithm

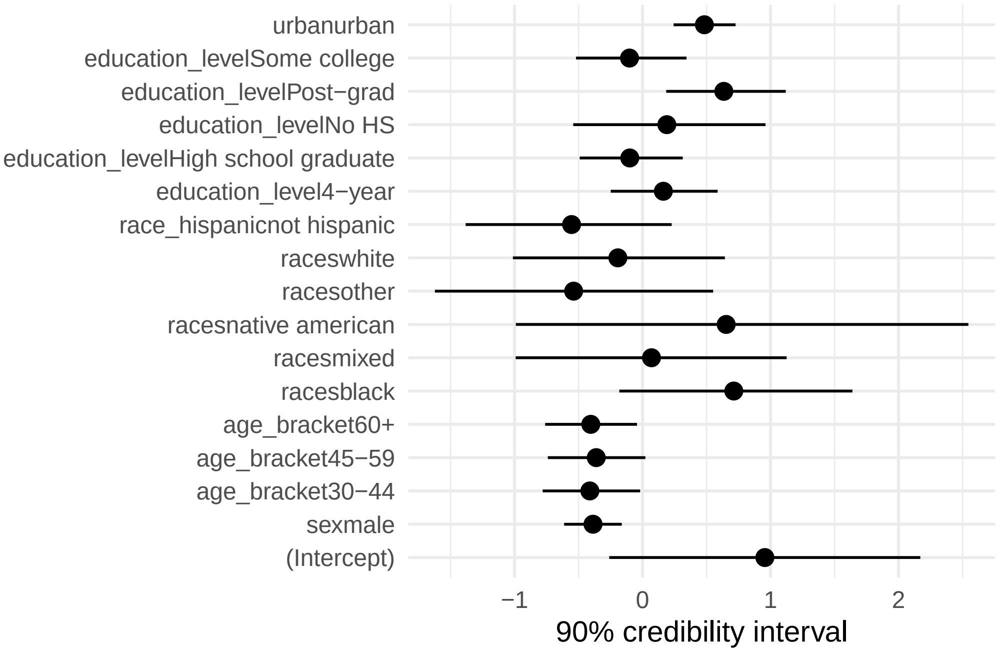


Figure 13: 90% Credibility intervals for predictors of vote_biden

References

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

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

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

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.

Ghitza, Yair, and Jonathan Robinson. 2020. *What Happened in 2020*. Catalist. https://catalist.us/wh-national/.

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/.

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.

Politico. 2016. *2016 Presidential Election Results*. https://www.politico.com/2016-election/results/map/president/.

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.

Rudis, Bob. 2020. *Statebins: Create United States Uniform Cartogram Heatmaps*. https://CRAN.R-project.org/package=statebins.

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/10.1177/00027162177126.

Stan Development Team. 2022. *Runtime Warnings and Convergence Problems*. R. https://mc-stan.org/misc/warnings.html#tail-ess.

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.