

Biden To Win the 2020 U.S. Popular Vote and Electoral College

Jin Long Cao, Brian Diep, Johnathan Tillman, Tanya Woloshansky

2/11/2020

Keywords

BIDEN, TRUMP, 2020 US ELECTION, FORECASTING, MULTILEVEL REGRESSION WITH POST-STRATIFICATION

Abstract

The U.S. election is one of the most important political processes in the world, the outcome of which can completely change the direction of American domestic and foreign policy. Thus it is relevant to many observers whether the Republican Party under Donald Trump will remain a second-term or if the Democratic Party under Joe Biden will replace him in the 2020 election. We approach forecasting the election by combining survey and census data from the Nationscapes and ACS respectively, in order to perform a multilevel regression using a Bayesian hierarchical model with poststratification. Through this methodology we predict that Biden is likely to win the popular vote with approximately 50.4% of the votes as well as the electoral college by winning 284 electoral votes over Trump's 254 votes.

Introduction

The 2020 U.S. federal election is shaping up to be one of the most divisive and hotly contested elections in recent history. Important issues such as climate change, gun control, COVID-19 response, and foreign trade policy are being discussed during this election cycle. The two major parties: the Republicans and the Democrats, are at odds for many of these key issues. This makes forecasting the election an important task for Americans and non-Americans alike. Although a key part of the electoral tradition is placed on the significance of each individual's vote, we can still find that almost all voters fall into broad categories that constitute voting blocs. These blocs can be united from many disparate categories such as age, race, and state of residence and many voters will identify with multiple blocs simultaneously. Although this is not an easy task, an analysis of these blocs can allow pollsters to predict the outcome of the federal election. For our purposes we combine individual survey data from the Democratic Fund's Nationscape Study and census data from the 2018 American Community Survey (ACS) to use a multilevel regression with poststratification.

In order to create a Bayesian multilevel logistic model, variables associated with voter preference were extracted from the Nationscape study. In particular, age, state, income, race, gender, Hispanic identity and employment were decided upon as predictors since exploratory analysis of the data showed a significant relationship (p-value less than 0.05 via logistic regression models) between these variables and voting intentions. Poststratification was performed using the ACS data to incorporate the population distributions of the U.S. population. The outcome of our simulation, which paid attention to the Electoral college wins and the popular vote, predicted Joe Biden as the winner of the 2020 election.

This work will continue to further explore the data sets used in our modelling, summarizing features of the data and visualizing variables of interest to our model. Next the features of our model and post-stratification process are described prior to simulating the election on the Electoral College. Finally, the prediction of Joe

Biden's win will be investigated further and a discussion follows into strengths and weaknesses of the model, why this information is useful and further avenues of study.

Data

Nationscape Data

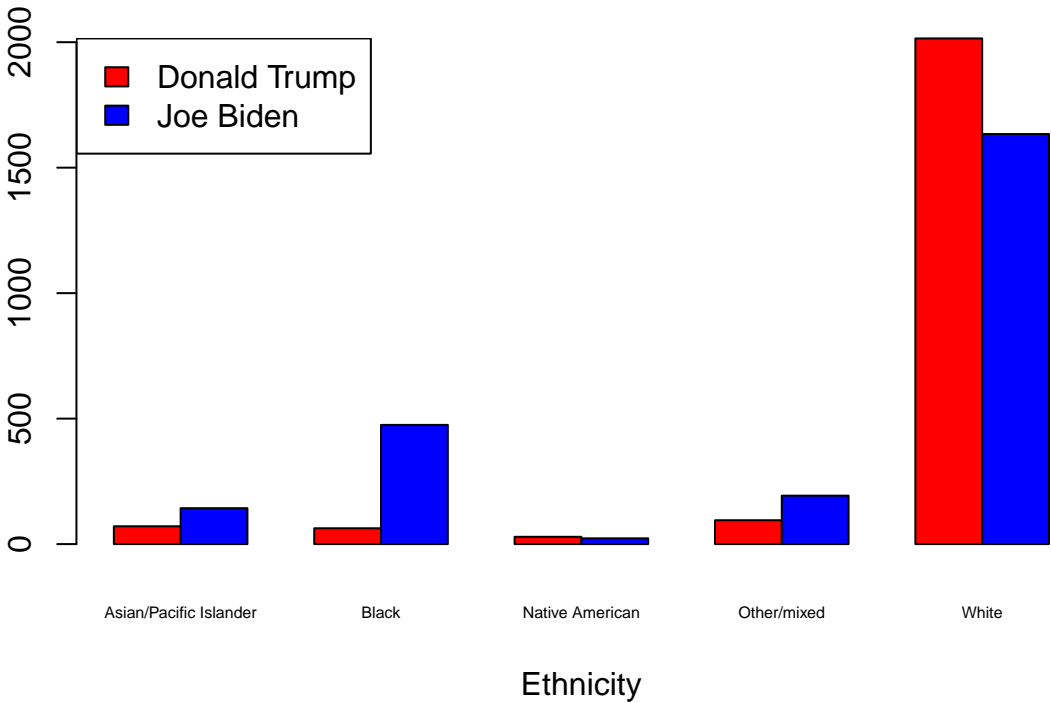
The dataset we used for our analysis came from responses to the Nationscape survey. The survey is the result of a partnership between the Democracy Fund Voter Study Group and two Political Scientists at UCLA (Chris Tausanovitch and Lynn Vavreck). The Nationscape survey runs from July 2019 to December 2020, interviewing 6250 people per week for a total of 500,000 respondents. The population of interest for the study is everyone eligible to vote in the US and the population frame are all the respondents sent by the market research platform Lucid to the Nationscape team. The team sampled the respondents sent until a quota on different demographic groups of age, gender, ethnicity, region, income, and education were reached.

The individual samples of 6250 respondents were sent directly from Lucid to the Nationscape survey software. Of the 6250 survey invitations sent, only about 75% of those surveys are included in the data as 12% decline immediately. Around 5% of the respondents are omitted because they stop participating during the survey and 8% are removed for speeding through the survey or straight line selecting answers. Lucid also requires an attention check before the survey is conducted and their supply partners that help form samples are evaluated by third party data specialists to ensure accurate representation of the target population..

The Nationscape team used a simple raking technique to transform their dataset to more accurately represent the US voting population. Raking involves adjusting weights of different demographic groups of a dataset to be proportional to the population of interest. The data is also checked for representativeness in a process used by Pew Research center. To do so, a number of questions are asked which are found on reliable, large scale, government surveys. Estimates from Nationscape surveys are then compared with government surveys, and the difference was in line with that produced by Pew Research.

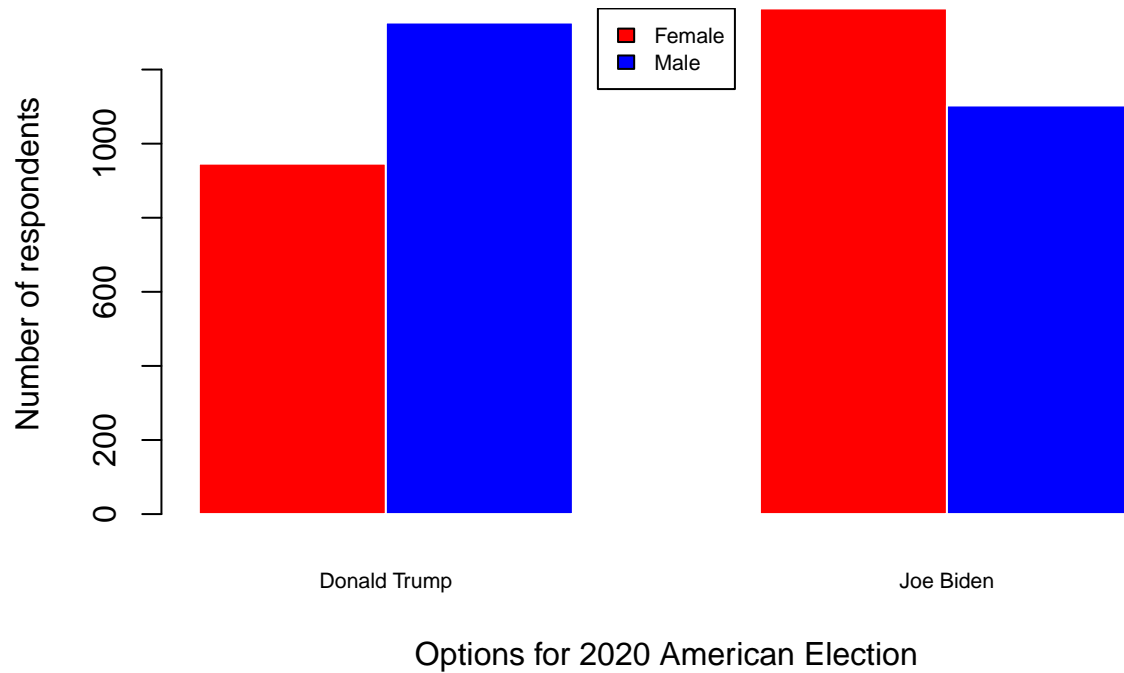
When mining for important variables, there were many categories where questions were not asked of all participants. While this is necessary to avoid too many people quitting the survey, some questions had a concrete answer for only 25% of respondents (such as the query regarding Medicare). Another weakness of the survey is that it is only available online and in English, which may deter those who are not comfortable with technology or English. Since the survey is done online, the frame is limited to all Americans that have access to the internet, are reachable online and own a device that is compatible with the software. The survey itself was also only available in English, thus limiting representation from non-native english speakers.

Figure 1. Respondents Voting Habits By Ethnicity



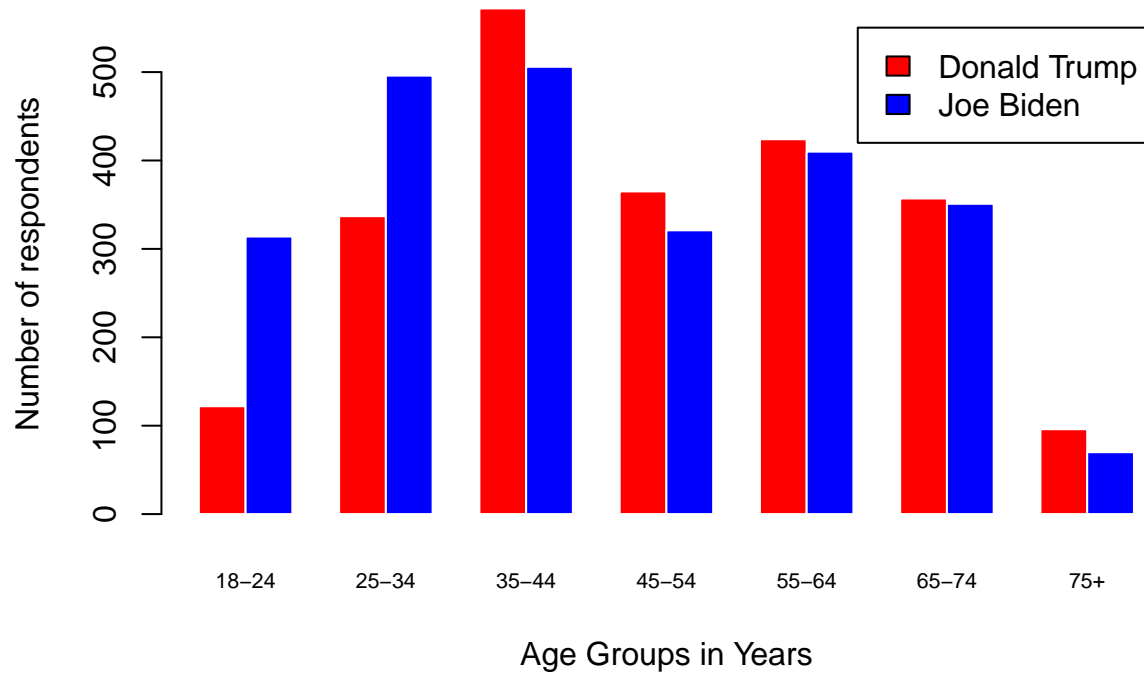
Ethnic minorities, as shown by the Nationscape data, tend to favour the democratic candidate, Joe Biden. White voters on the other hand, appear to be planning their vote for Trump

Figure 2. How Respondents Plan to Vote in 2020 By Gender



Another trend that the Nationscape data demonstrates is that female voters tend to prefer Joe Biden over Donald Trump. Male voters planned to vote the other way and support the incumbent president.

Figure 3. How Respondents Plan to Vote in 2020 By Age Group



Voters younger than 35 years of age strongly prefer Joe Biden. Respondents over this age seem to slightly prefer Donald Trump but some age groups are too close to call.

Figure 4. Political Party Association of Respondents Voting for Donald Trump

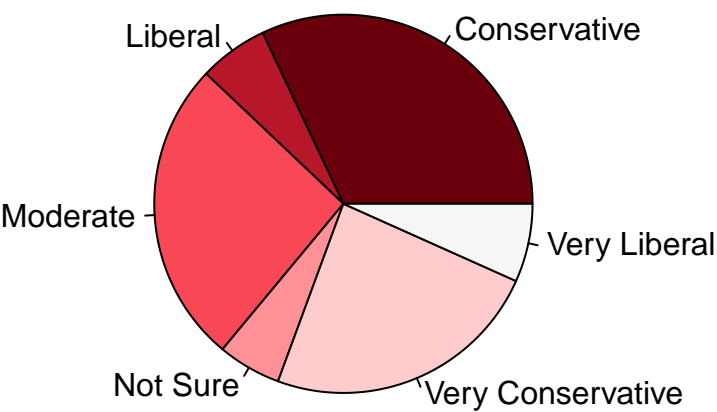
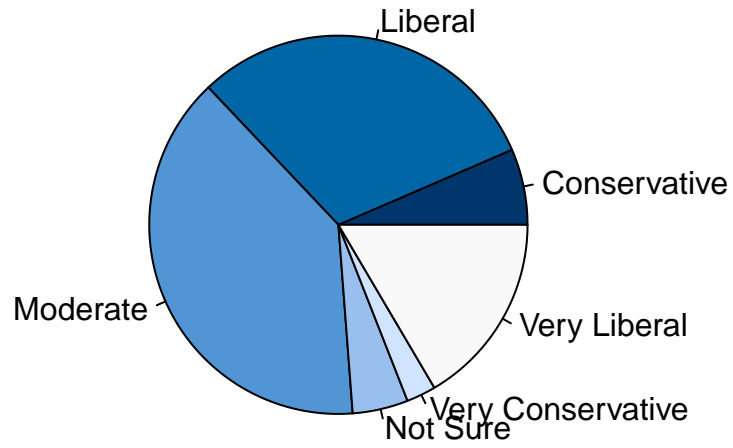
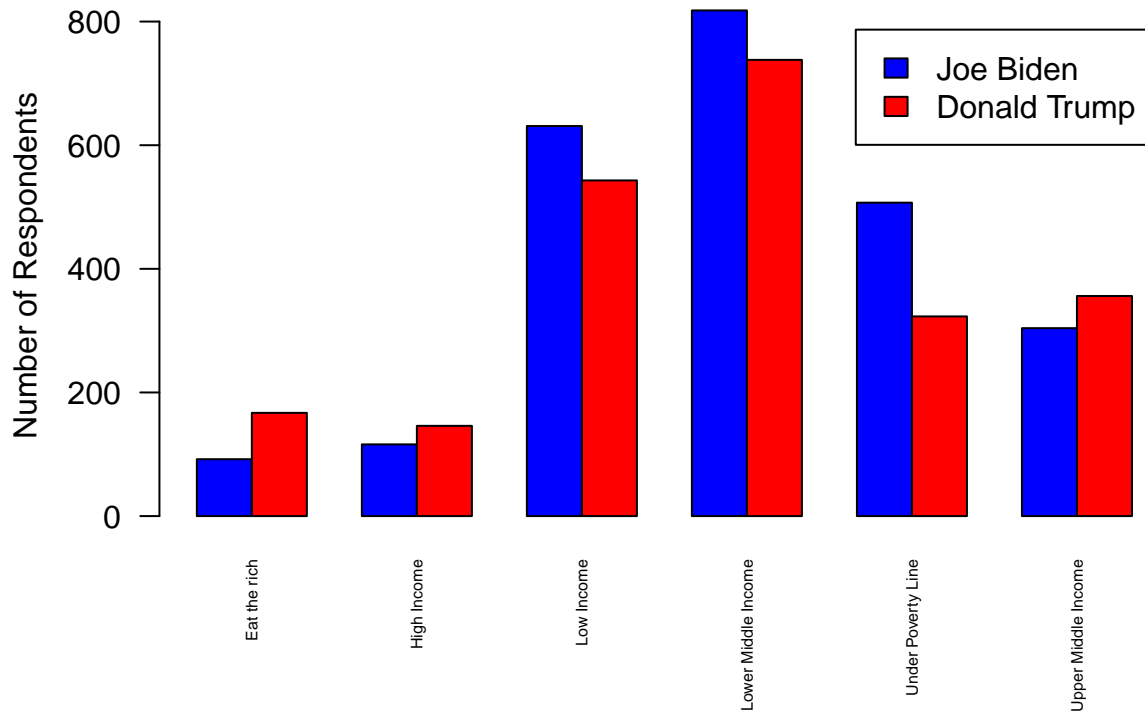


Figure 5. Political Party Association of Respondents Voting for Joe Biden



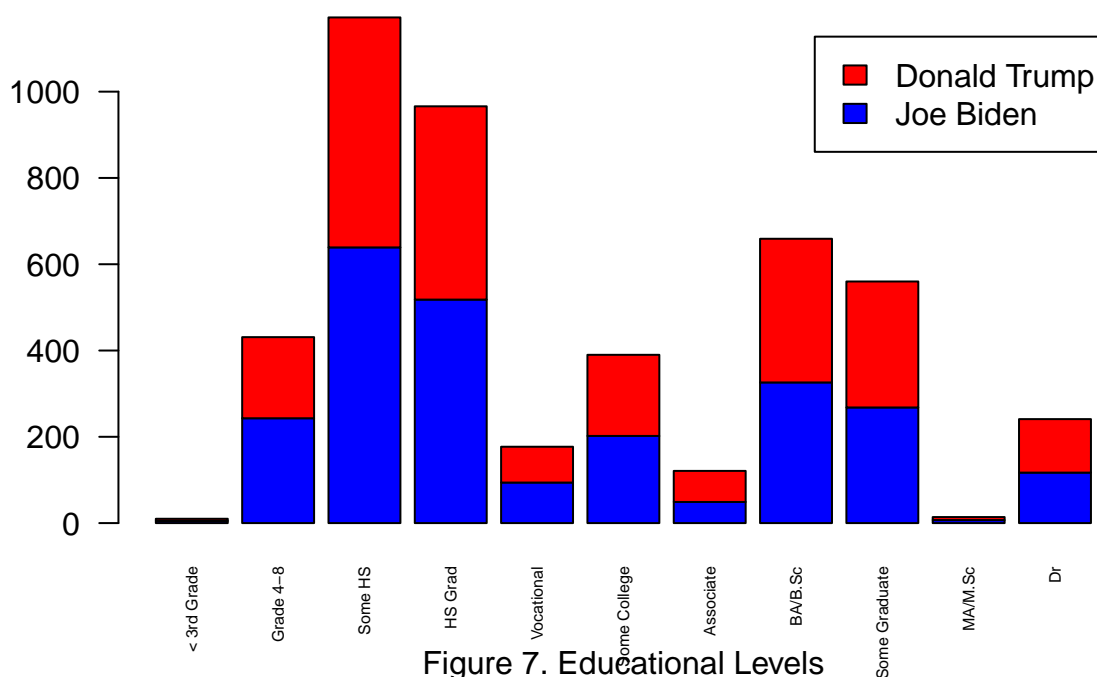
Not much crossing of party lines appears to be happening with the Nationscape data. Most of those who plan to vote for Biden range from very liberal to moderate. Likewise, respondents who consider themselves very conservative, conservative or moderate will be voting for Trump.

Figure 6. Voting Tendencies of Respondents by Income Brackets



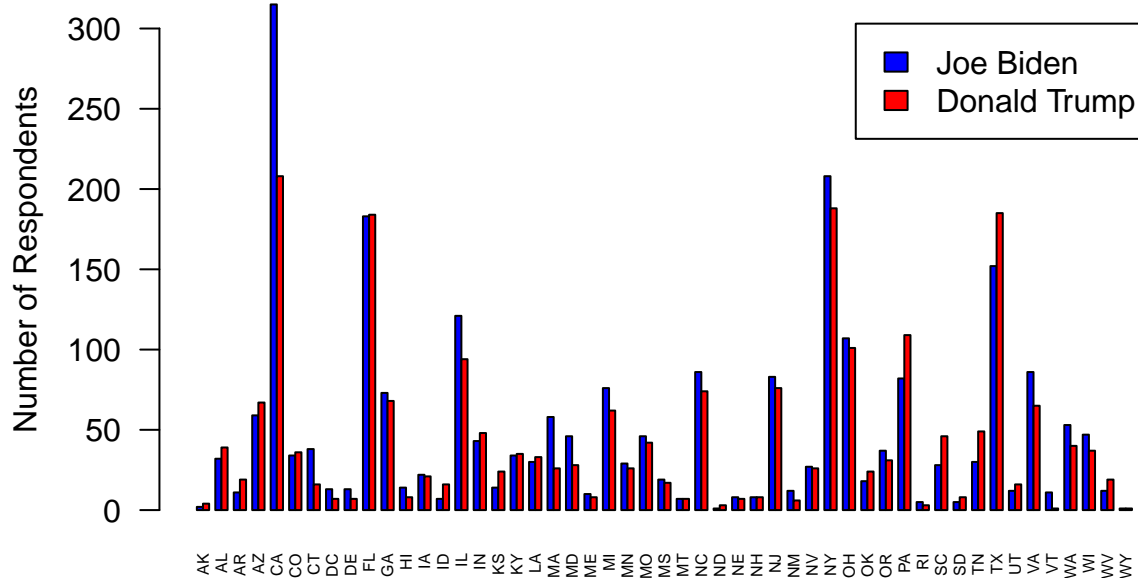
Income has also shown to be associated with voter preference with our Nationscape data. Lower incomes (those close to the poverty line or minimum wage earners) plan to strongly support Joe Biden. The Democratic Party continues to lead in all income levels (although more modestly) up until six figure earners where respondents flip and plan to vote for Trump.

How Respondents at Varying Levels of Education Will Vote in 2020



Interestingly, education level of respondent did not show a relationship with voting intentions for the 2020 election. Across all education levels queried in the survey, an association with either candidate did not exist.

Figure 8. Voting Tendencies of Respondents by State Brackets



Similar to education, it did not appear that states tended to favour any particular candidate in the Nation-scope survey data. The race for the Whitehouse appeared to be fair game in all states at the time of this survey.

ACS Data

Since we decided to poststratify via adjusting the weights of respondent groups using an additional dataset, we used the American Community Surveys (ACS) from 2018. The ACS is a reliable source of data on population parameters and is sent out to a sample of 3.5 million households and group quarters monthly. The population of interest for the ACS are all citizens of the US. The population frame for the ACS is the Master Address File (MAF) which consists of the Census Bureau's known living quarters and some nonresidential units. The MAF includes households and group quarters where groups of people live together. In 2013 the survey methodology was updated to stop sending surveys to College dorms during summer months to improve response rates and lower costs.

The Census Bureau first divides the US into five distinct sub-frames, each of which is utilized for a survey every five years to avoid oversampling specific areas. New addresses in each of these sub-frames are added each year systematically. After selecting the sub-frame for the year every block in the sub-frame is assigned a sampling rate based on sampling identities that affect their population like being in a school district or being in a Native reserve. The smallest sampling identity is used to determine the sampling rate for the block. After selection is finished the assigned blocks are grouped via stratum and geography to be systematically assigned to a month in the year to be surveyed.

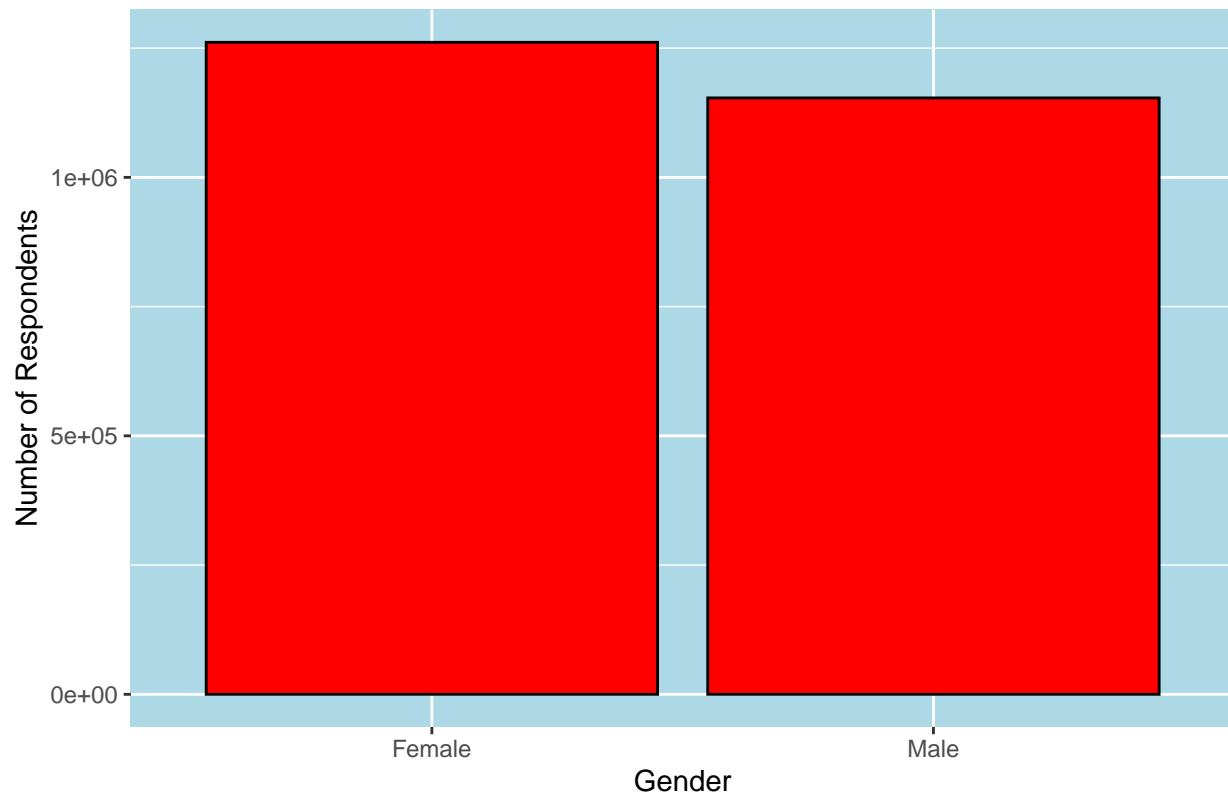
Surveying each household in assigned blocks involves four phases of attempting contact. First the Census Bureau sends each household a request form to fill the survey out online and if not completed they mail a questionnaire to be returned by mail. If they still haven't received a response they will follow up with a Computer Assisted Telephone Interview. Finally if a household has not answered any of the previous survey

attempts they will conduct a Computer Assisted Personal Interview with a representative of the Bureau travelling to the address.

After data selection occurs the Census Bureau weights the household respondents in three stages. First they weigh to account for the sampling rates they calculated previously via geography. Next they weigh the households to account for non-responding households in the same block to account for the reduced sampling rate. Finally they weigh the households via age, sex, race and Hispanic origin to match their population estimates program.

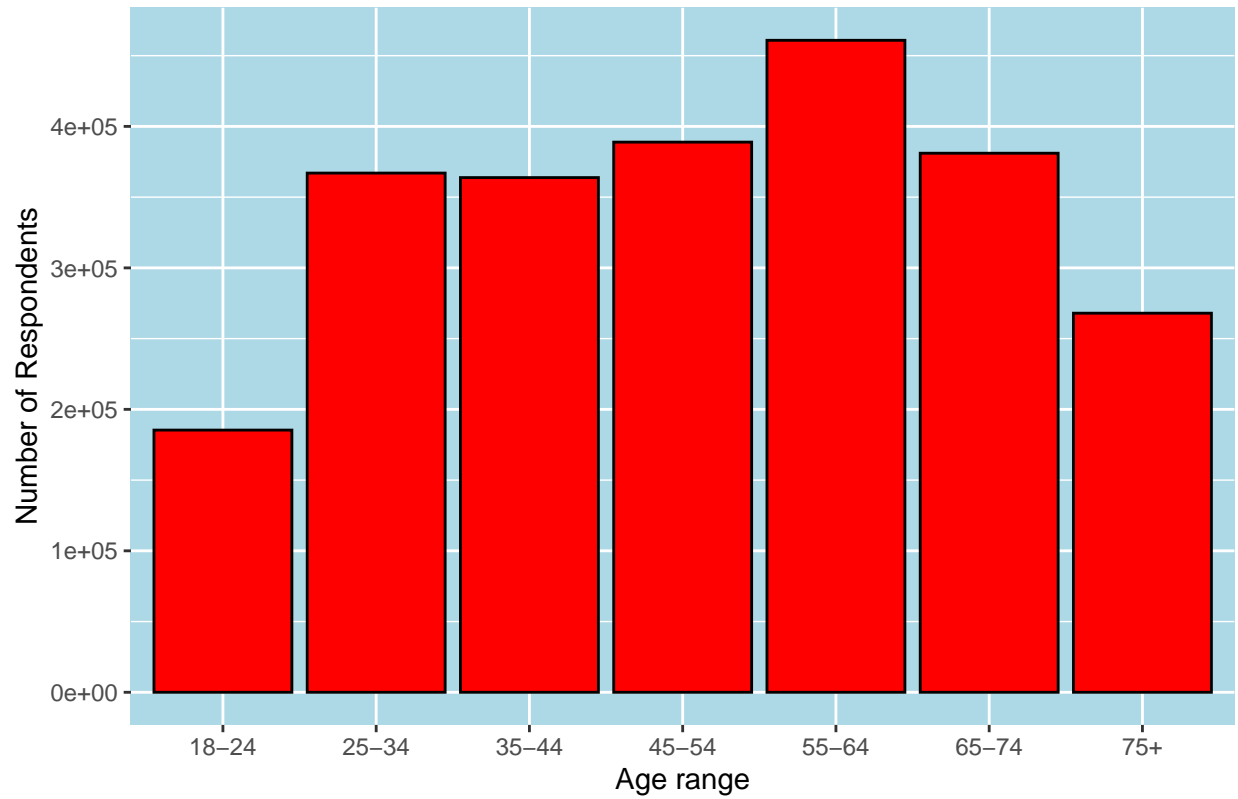
Although the ACS has a very comprehensive sampling design they only select a small sample size of 295 000 of more than 180 million households to be sampled each month preventing accurate representation through a robust sample size. Additionally since the Census Bureau uses multiple methods of completing the ACS survey differing measurement error are present and not consistent among households sampled.

Figure 9. Gender Proportion in the American Community Survey (2018) c

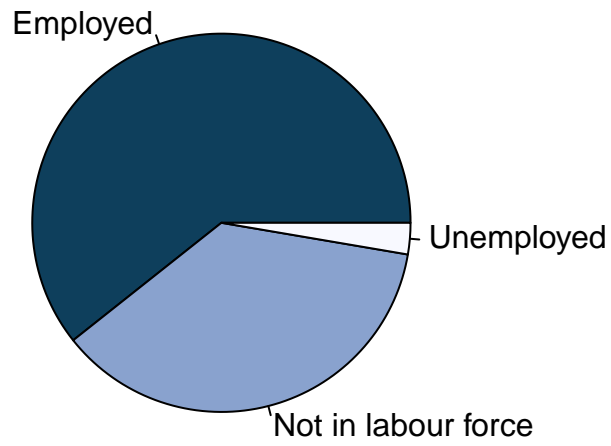


To show that it is appropriate to use the 2018 ACS data for post stratification, the variables used in the model must be in proportions representative of the American population. Since there are slightly more females in the general American population, it is suitable that the proportion of females is a bit greater in the ACS.

Figure 10. Age range in the US Estimated Through American Community

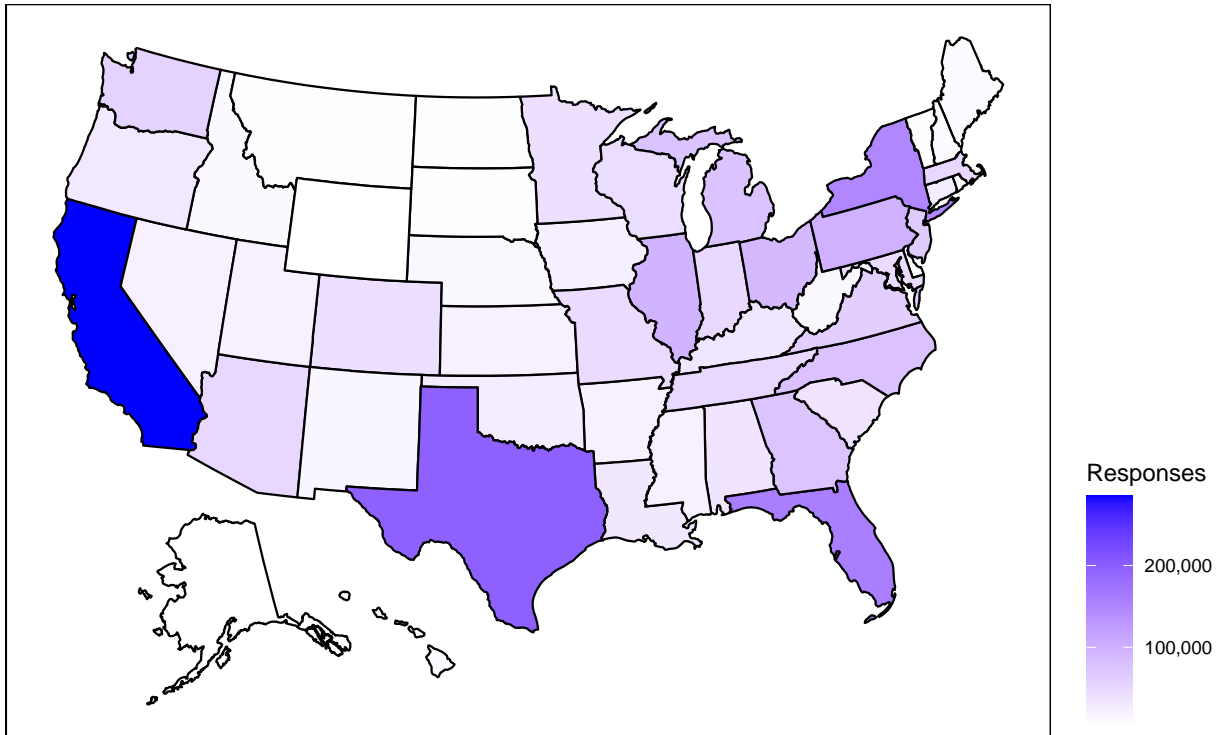


Aligned with population data, a slight left skew is demonstrated by the age group data in the 2018 ACS survey.



Since the majority of the American public was employed at the time of the 2018 ACS survey, these results appear fitting to simulate our presidential prediction model. However, this could be a source of error as the unemployment rate is currently higher due to the pandemic.

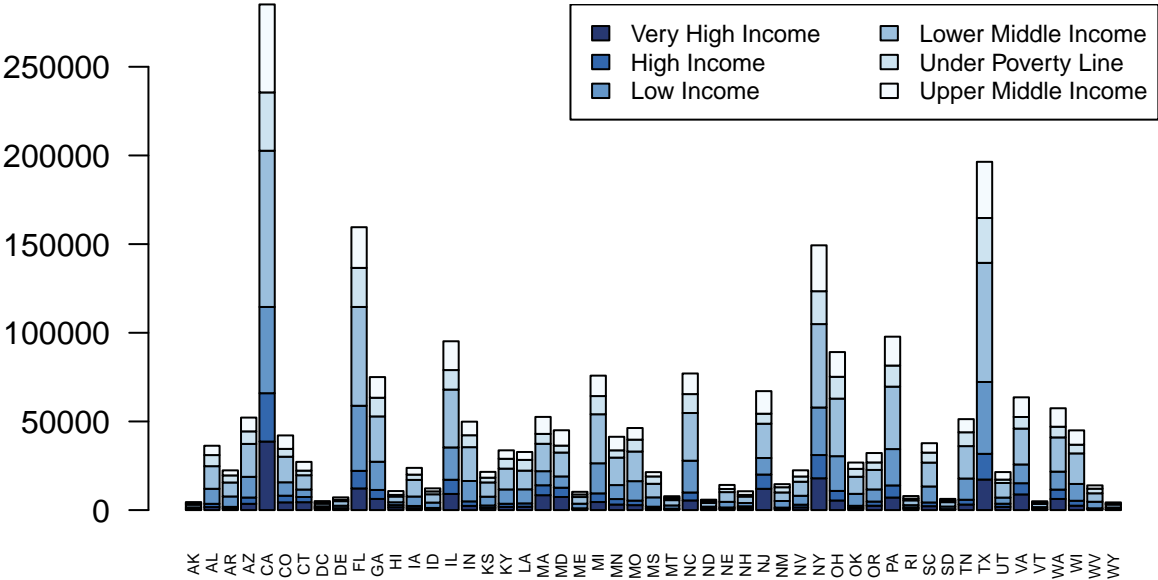
Figure 12: Responses by State



Source: ACS 2018

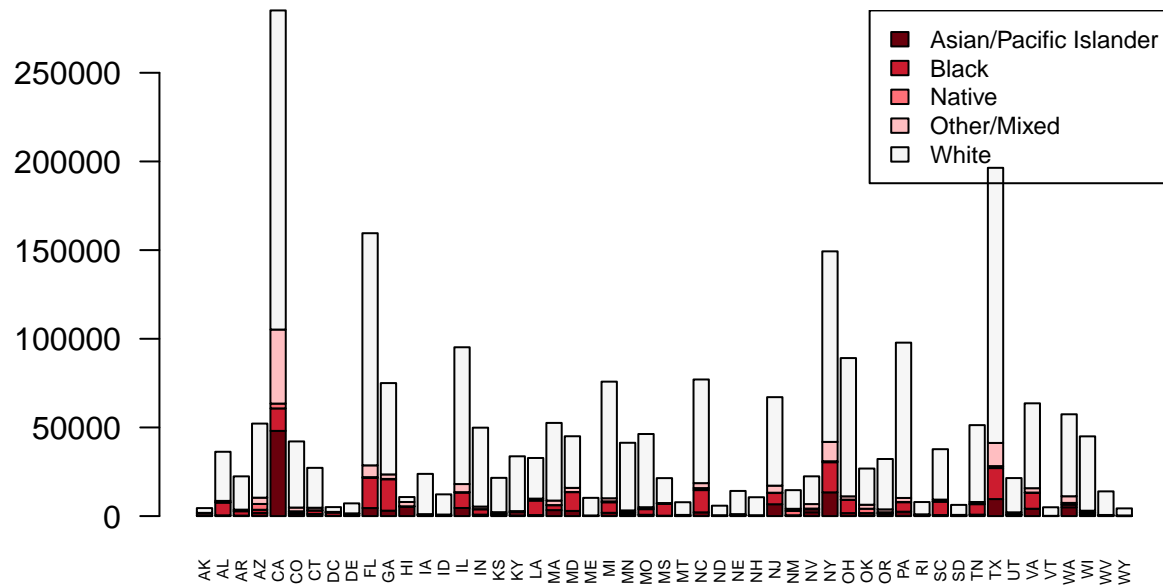
State-by-state responses are visualized to show where the highest number of respondents resides. California, Texas, Florida and states closer to the coasts tend to have the most respondents in the ACS survey. Since California, Texas and Florida are the most populated states (in that order); it is good news as a base for our simulation.

Figure 13: Estimates of Income Stratification By State



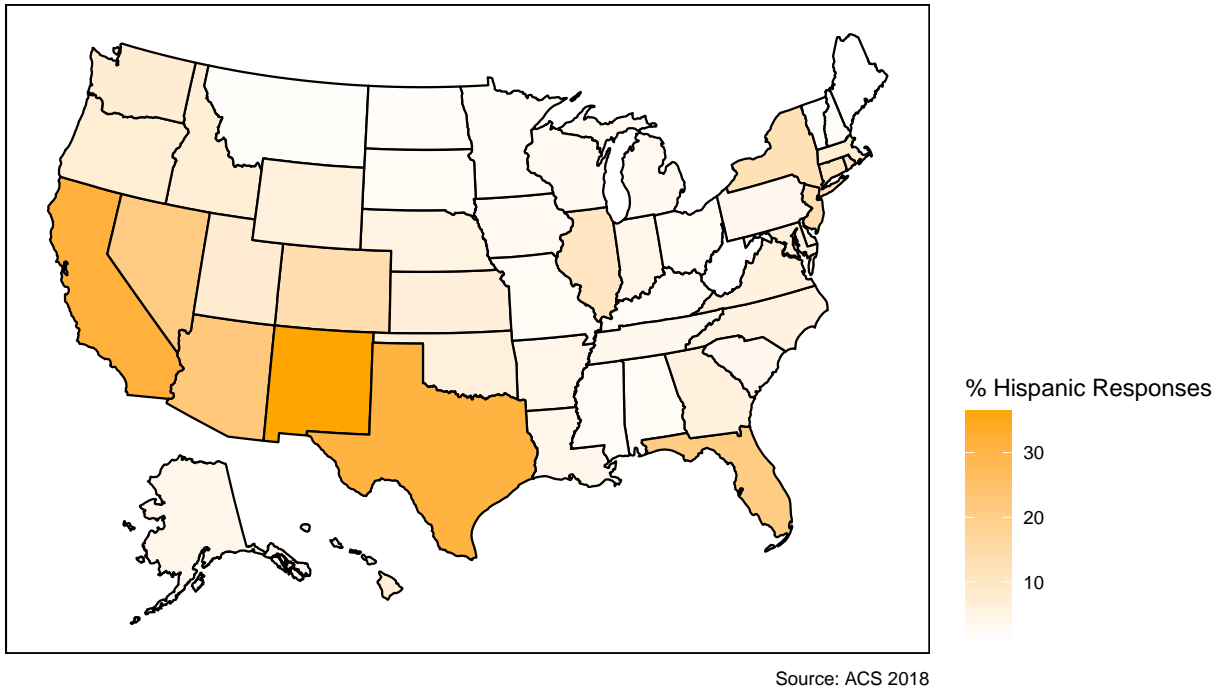
Income was another key variable in our simulation model and it can be shown here that states vary by proportion of respondents at various income levels.

Figure 14: Distribution of Race Ethnicities By State



Distribution of race ethnicities throughout America are shown here to be more common in the southern states as well as states near the Pacific Ocean (excluding Alaska)

Figure 15: Proportion of Hispanic Respondents



Proportion of Hispanic residents in the 2018 ACS is distributed throughout the southern United States and states along the coast, similar to race ethnicities.

Data Cleaning

From the ACS and Nationscape data, we removed invalid incomes/null values, null/other employment, third party/non-voters and education from the sample. As those values did not highly affect the outcome and in the case of non-voters, their vote is not relevant to an accurate analysis of who will actually cast votes during the election.

We took income and age which were continuous variables in the survey and census and grouped them into income and age groups that we could poststratify more easily. These groups are also done in such a way that they represent voting blocs in general.

A similar process was done with the race and ethnicity data to aggregate smaller racial groups into broader ones that could be matched easily between the survey and census. The variable Hispanic was also aggregated into a binary variable as opposed to a more specific country specific categorical variable.

In our analysis, gender and sex were made equivalent since there was an absence of other options of gender identity and we did not have any meaningful demographic information on non-binary respondents. Employment was also compressed from the survey data to match broad categories of employed, unemployed, and out of labour force. This captured the majority of the nuance between employment status that we were looking to examine in this analysis.

Model

A Bayesian multilevel logistic model with post-stratification was used to predict the plurality for each state in the US 2020 election. We used a logistic model as they can associate binary and continuous independent variables with a binary categorical response variable. Since the decision of which major party to vote for is condensed to a binary response variable, the logistic model is well suited for our goal of predicting the election winner. Linear models were unsuited for our goals as they require the response variable to be a continuous variable such as height. Our response variable was either 0 (representing a vote for Trump) or 1 (representing a vote for Biden), which is binary in nature.

Mathematically with a logistic model we can express the probability ‘p’ that an individual will vote for the Democratic Party as a logarithmic function of the various dependent variables.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$$

For a logistic model to be valid, we also need certain conditions to be true with our data. This includes the independence of each observation and that there is low covariance between any pair of explanatory variables.

We know that the first assumption holds due to the survey design which is sampled randomly. We also know that the second assumption holds due to our covariance matrix where all values are low which indicates little to no covariance between our explanatory variables (the code for covariance matrix is located in the model section of the rmd file).

We decided to make the model multilevel in order to group voters by state. Since each state varies significantly in voting preference and demographic groups we saw it fit to analyze plurality of each state separately. Most importantly, the electoral college also assigns electoral votes based on plurality of each state so to predict the winner of the US 2020 election we needed to separate the predictions via state. MRP also allows us to generalize our findings based on a possibly unrepresentative sample by correcting with external demographic data (Wu & Thompson 2020). However, multilevel modelling with poststratification does not allow us to firmly establish a causal relationship between our independent and dependent variables (Gelman & Kennedy 2020). MRP also requires a large dataset containing the demographics of the population in order to be useful. For our purposes, this condition is fulfilled since the ACS comprehensively details all the demographic information that is relevant to our analysis.

We used a Bayesian approach rather than a Frequentist to allow a credible interval to be produced. The Bayesian approach involves updating the probability of an event based on prior knowledge affecting the event. Even though we use noninformative prior information—known as a prior—the method allows us to create a credible interval in which we are 95% certain our results will occur. If we used a Frequentist method that uses proportions or frequencies of the data we can only make a confidence interval that shows if the elections were repeated 100 times our result would occur 95 times. Bayesian models are known to account for any cause of uncertainty in inferences specifically for predictions.

We decided to use a Bernoulli distribution as a noninformative prior to make our model since it has two outcomes: either a failure (0) or a success (1). Other distributions with binary outcomes like the Binomial Distribution involve sampling with replacement of individuals in each sample, which is not observed in the sampling methods of the Nationscape and ACS datasets.

The Bayesian approach to logistic regression involves predicting the parameters using the typical approach to Bayesian inference involving the priors. If we express our model in mathematical notation we are returned the form:

$$P(\beta_0, \beta_1, \dots, \beta_k | y, X_1, \dots, X_k) \propto P(y, X_1, \dots, X_k | \beta_0, \beta_1, \dots, \beta_k) * P(\beta_0, \beta_1, \dots, \beta_k)$$

where y represents the probability of an individual voter voting for Biden. Specifically our model predicts the outcome of the average voter that we can generalize to the entire US voting population. The beta values represent our independent variables, specifically age, income, employment status, sex, state, race,

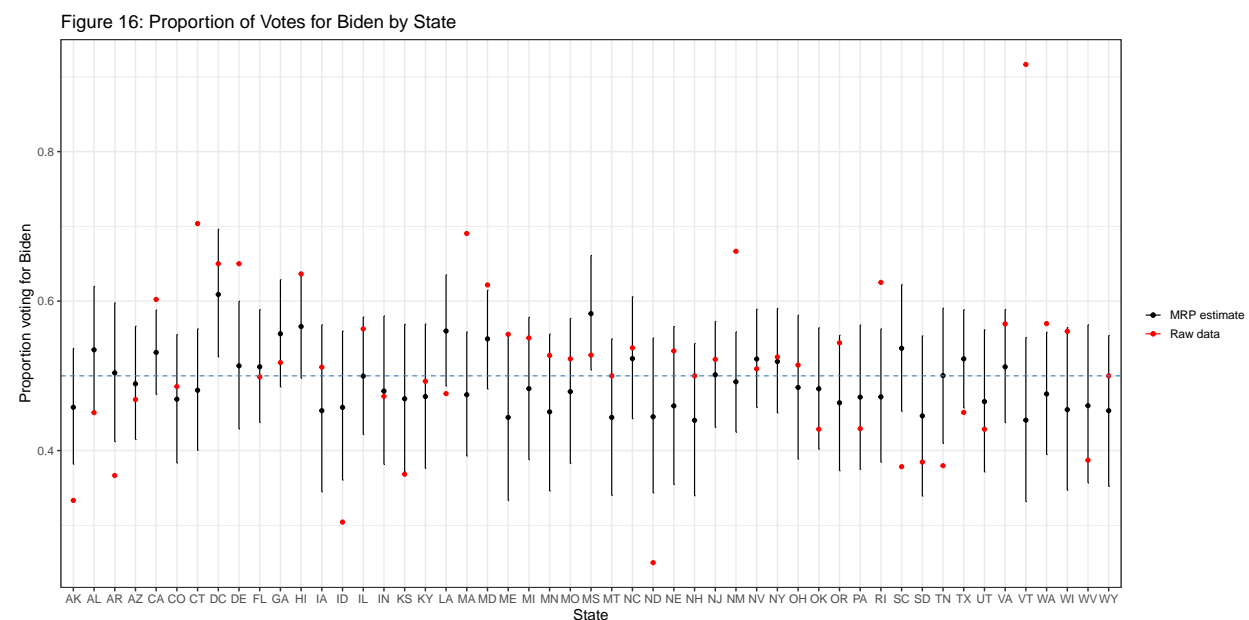
and whether or not an individual is Hispanic. We decided to use poststratification when making our model to accurately depict the US voting population. To reduce any non-response bias we weighted different demographic groups to represent the population of interest.

The dependent variables that we decided to fit our model on were some of the relevant variables found in the data exploration section such as gender, age, employment status, household income, and race/ethnicity including Hispanic heritage.

We decided to group ages of voters in brackets of 10 years and income based on varying brackets to distinguish different economic groups. This allowed us to easily interpret relationships found between voting preference and age and economic groups. Modelling these variables as continuous does not show us the relationships between differing groups and makes it harder to view the representation of these groups for poststratification. The other categories were group based on convenience as the samples came from our census and survey.

Our model was fit using brms and Stan in R.

Results



After fitting our Bayesian multilevel regression model with poststratification, we find that Biden is at an advantage for winning the election. We see in Figure 16 above, where we plot the predicted proportion of Biden votes in each state according to our estimate from MRP, as well as the raw estimate according to the survey. In particular we plot the mean of the predicted proportion of Biden votes among voters and the 95% credibility interval. We notice that the raw estimate of Biden proportions does not well represent the estimate after MRP in many cases, especially in states that have small sample sizes.

Figure 17: Proportion of Votes for Biden by Gender

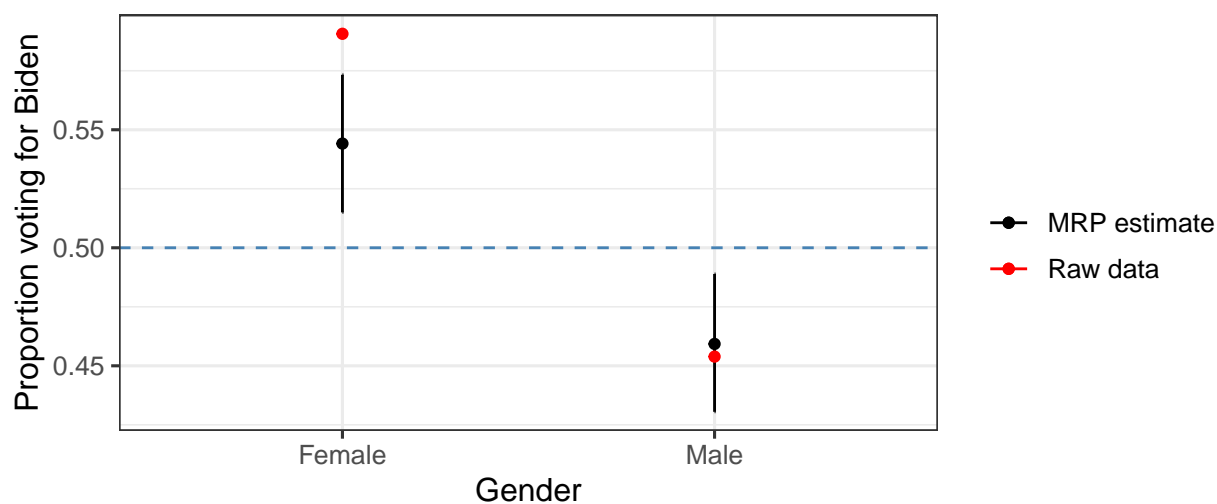
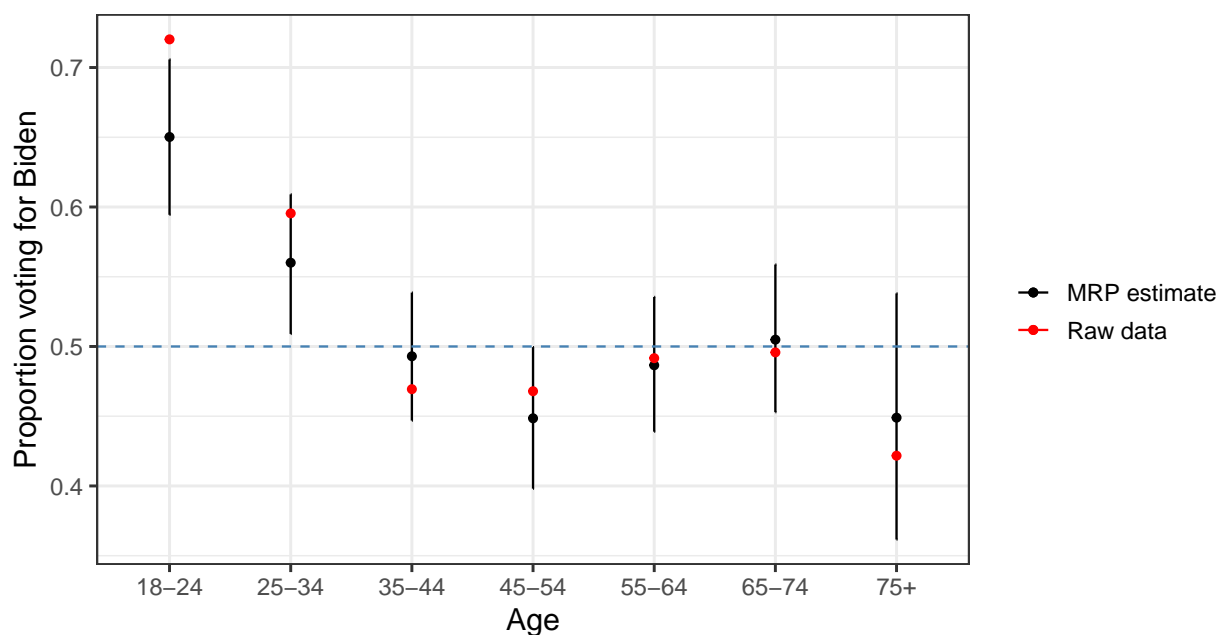


Figure 18: Proportion of Votes for Biden by Age



We also plot similar confidence intervals for voters after poststratification based on the other categories such as gender, income, age, race, and employment status. Based on gender in Figure 17, we find that women are more likely to vote for the Democrats as opposed to men. We also see based on age in Figure 18 that younger voters under the age of 34 tend to vote Democratic and those above that threshold are more likely to vote Republican.

Figure 19: Proportion of Votes for Biden by Race/Ethnicity

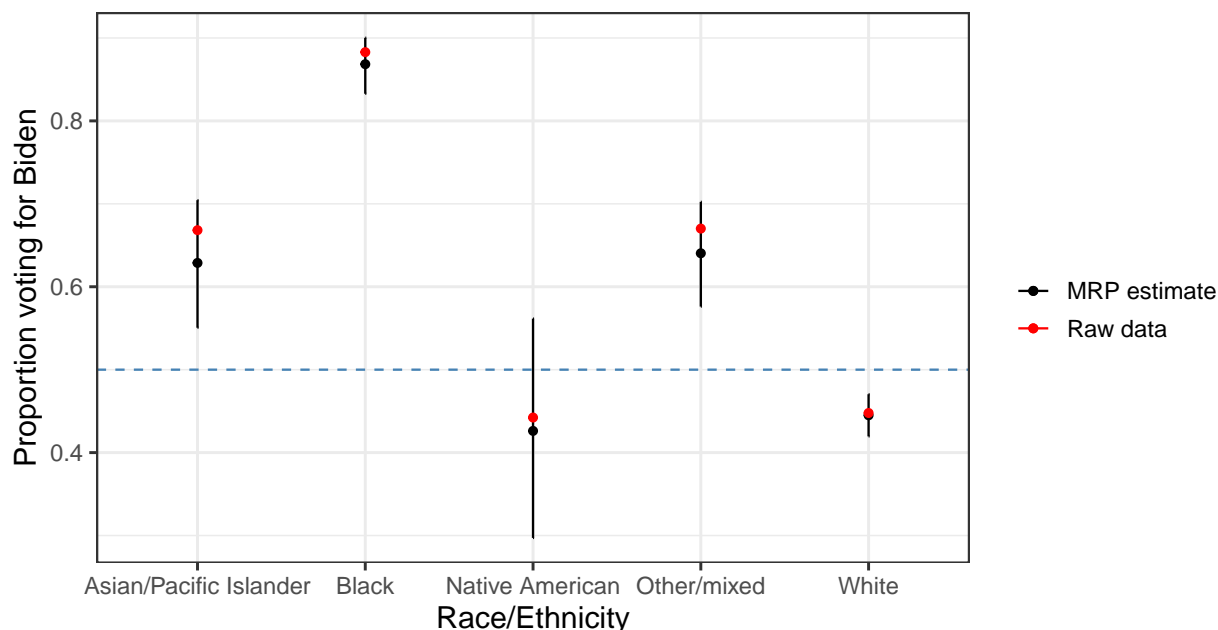
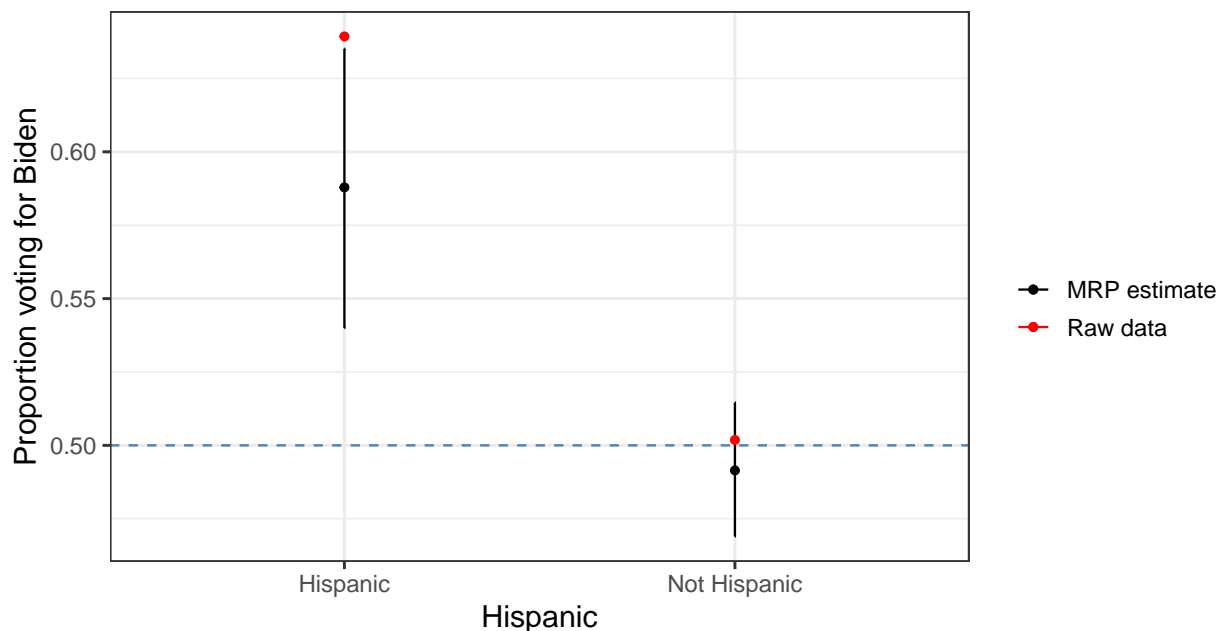


Figure 20: Proportion of Votes for Biden by Hispanic Heritage



We find in Figure 20 that in terms of race, self-identified Hispanic individuals are much less likely to vote for Trump as compared to their non-Hispanic counterparts. As for race in Figure 19, white Americans and native Americans are the only two racial groups that are more likely to vote for Trump than Biden. However one thing of note, the credibility interval for native Americans is much larger than the one for white Americans. We also see that black Americans vote overwhelmingly for Biden whereas voters of Asian/Pacific Islander descent as those self-identified as mixed or other racial groups skew towards the Democrats to a smaller degree.

Figure 21: Proportion of Votes for Biden by Income

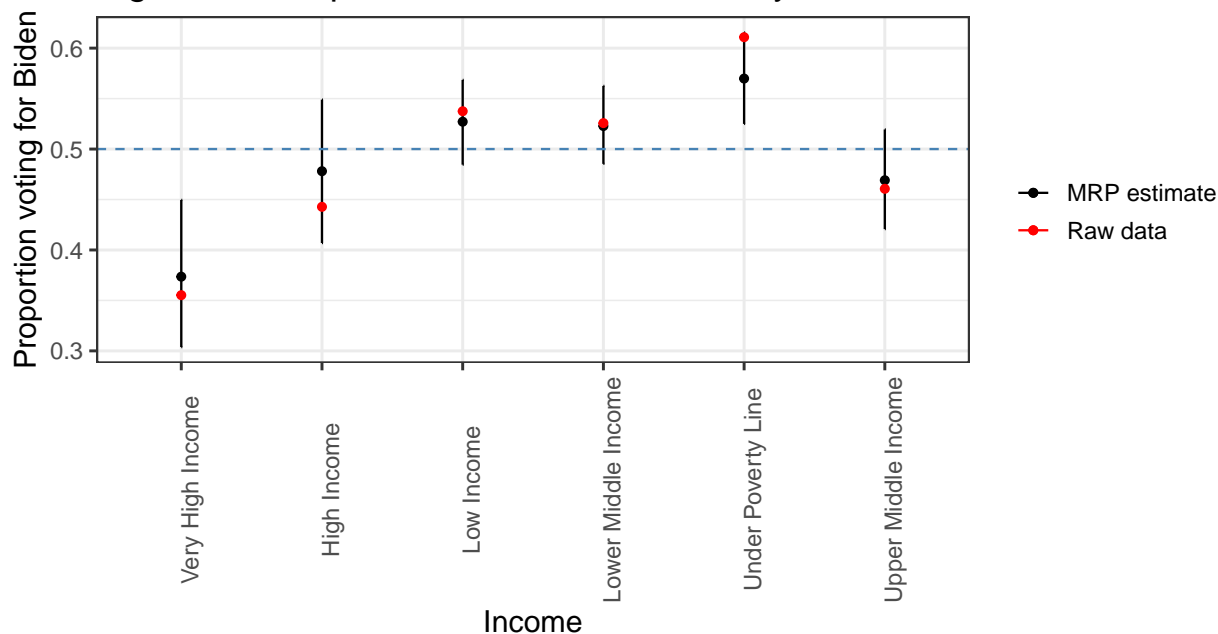


Figure 22: Proportion of Votes for Biden by Employment

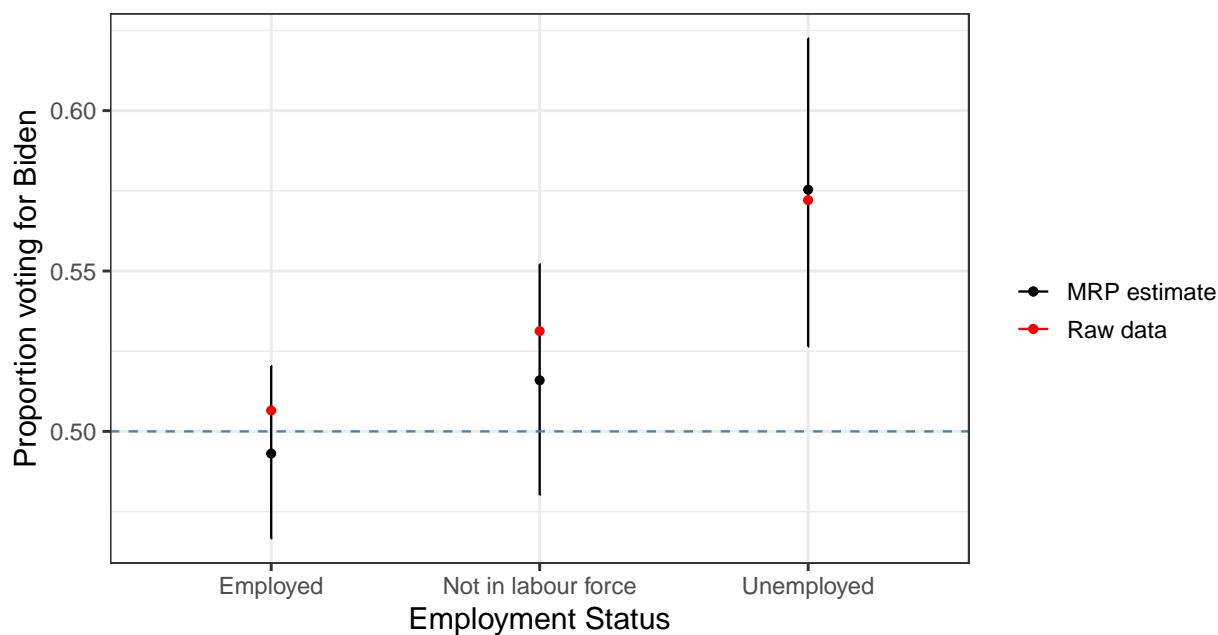
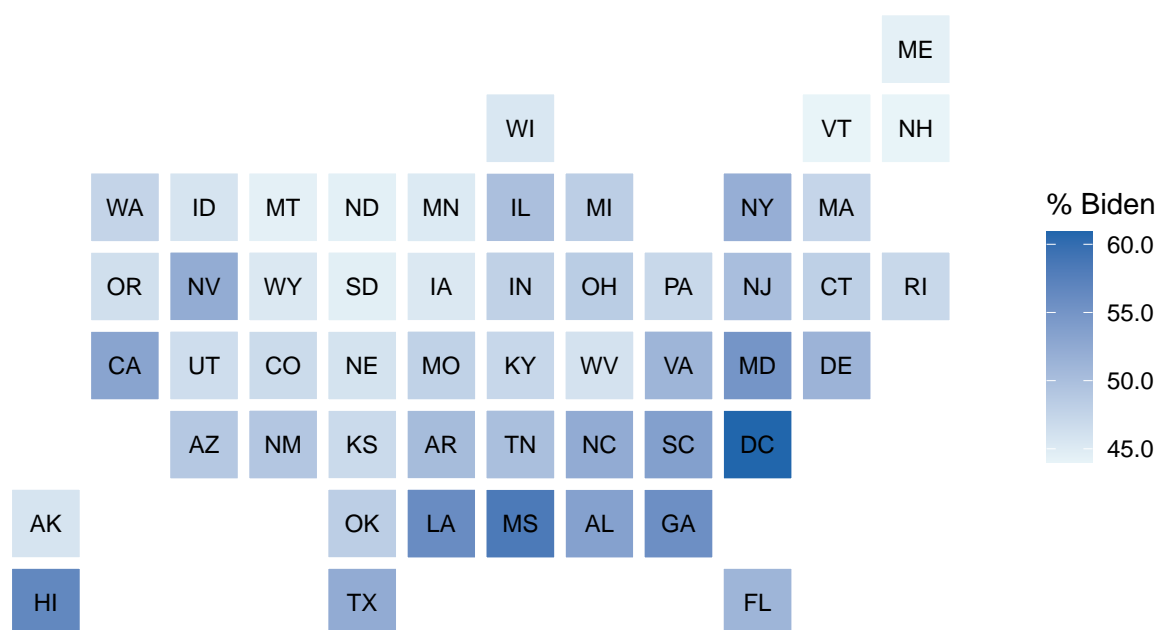


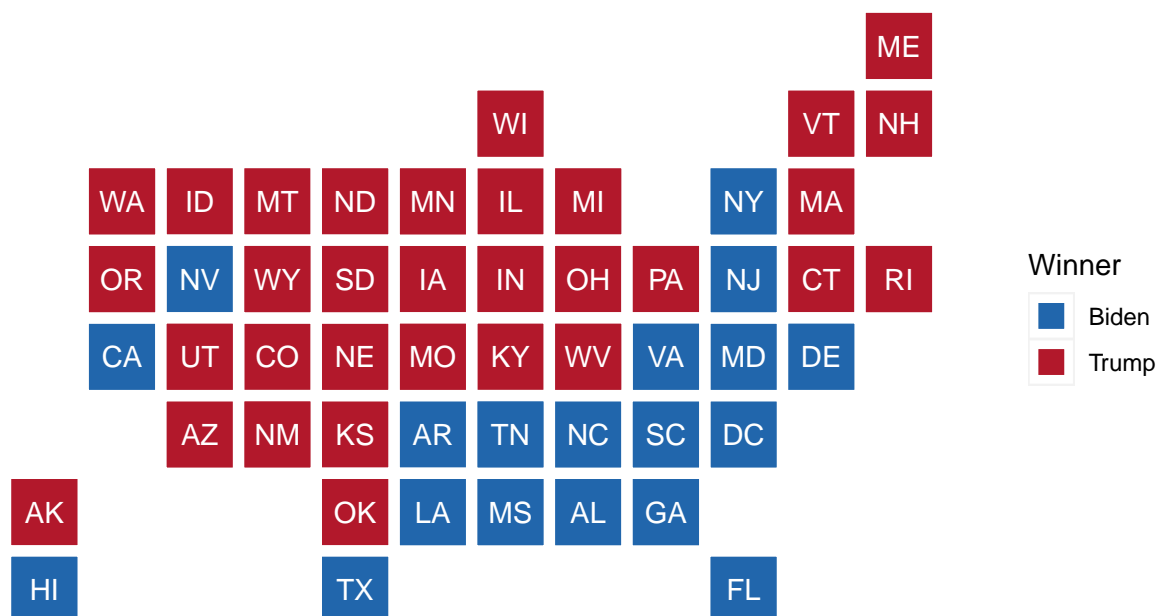
Figure 21 and 22 when we examine voter preferences in terms of economic status, those who are unemployed or not in the labour force tend to swing towards Biden and the Democrats whereas the Republicans are favoured by those who are actively employed. We also find a clear trend that as an individual's household income increases, they are more likely to vote for the Republican Party.

Figure 23: Distribution of Votes for Biden in 2020 US Election



Combining all of this data using poststratification to adjust for the relative proportion of each group. Our model predicts that we have Biden narrowly winning the popular vote with almost 50.4% of all ballots cast nationally.

Figure 24: 2020 US Election Electoral College Predictions



However, since the U.S. election is decided by the electoral college and not by a national popular vote, we must examine which states each candidate achieved a plurality. For the vast majority of states, the winner of the popular vote in each state is awarded all of the state's electoral college votes. For our analysis, we assume that Biden will win any states where the mean proportion of Biden voters is greater than 50% and that Trump will win the remaining states. As seen in Figure 24, we predict that Biden will win a lot of the American South, the East, Hawaii and California whereas Trump is predicted to win much of New England, Midwest, Alaska and the Western U.S. When we sum up the weight of these states and their electoral votes we predict Biden to win 284 electoral college votes to Trump's 254.

Discussion

Using a Bayesian multilevel logistic regression model composed of age, race, hispanic identity, home state, gender, employment status and income level variables; we found that Joe Biden would win the 2020 presidential election in terms of the electoral college as well as the popular vote. However, we should continue to analyze the reasons as to why Joe Biden is favoured to win the elections and whether these trends can be used to predict future elections.

Our model shows that the core supporters for the Democratic Party as well as the Republican Party can be deconstructed to consist of clear groups based on social and economic class. It is evident from our results that Republicans tend to skew towards being white, male, older, and have higher household incomes compared to Democrats who gather much of their support from women, racial minorities, and lower/middle class voters.

As seen from our census data, there are more people above the age of 45 than there are below as well as more white people than any other ethnicity. These make up a core part of Donald Trump's and the Republicans' voter base. Traditionally, higher income earners also tend to vote Republican due to their favourable policies regarding wealth, which our data also supports.

However, we also note that there are more female, middle class and lower economic class voters which are all groups that are predicted to favour Joe Biden. Not only were these variables significantly associated with voter outcomes, but current literature agrees that younger, female, non-white voters tend to be associated with the Democratic Party (Pew Research Center 2020).

Although our model is being fit using data from a year unique in several ways (i.e. global pandemic), the fact that the variables chosen are verified by literature based on previous elections is important. It helps ensure that the model has not been overfit and can possibly be used as a template for future elections.

Racial data seen in our results also shows that the vote is heavily divided on racial lines, this suggests that Biden is much more likely to win in states that have higher proportions of ethnic minorities, in particular black/African Americans and those of Hispanic heritage. This is clear from graphs above where Biden has significant wins in the American South, where black and Hispanic voters make up a large percentage of the state population.

Our analysis only took into account demographic information about voters, and completely omitted ideological factors. Thus, the voting trends that are seen in our model exposes a major rift within American society. This rift appears to be drawn particularly on racial and class lines and suggests that they still play a heavy role in the political process. We expect that these divisions will continue to be exacerbated as these themes become increasingly politicized. As American society continues to diversify, if we do not see a bridging of social and economic inequality, we are likely to see greater polarization between the parties.

The model described in this paper is based on the outcomes of voting within Electoral colleges rather than the popular vote. Recent elections in 2016 and 2000 have shown that the popular vote and Electoral college may not agree so it is important to simulate models using the latter. Studies have found a bias in the Electoral college towards Republicans historically but other simulations for the 2020 elections have determined similar outcomes with the Democrats leading.

We also find that our analysis is most confident in states where the credibility interval for the proportion of votes for Biden in states that have their mean further away from 0.5. These states are more likely to be “safe” and their races will be less competitive than others since there is likely to be a clear front runner. However, for states like Tennessee, New Jersey, and Arkansas, we see that the mean is incredibly close to the 0.5 threshold and we predict that these will be competitive swing states that could determine the outcome of the election.

Limitations and Weaknesses

One limitation of our model is that it was restricted to only one phase of the Nationscape Data. Although 5200 respondents is a good size survey, it may not properly model a population of 350 million. For example, current literature has shown that well educated citizens, those holding degrees higher than a Bachelor’s, identify as liberal more than republican. However, the data set used to model our prediction did not show a significant difference in voter ideologies based on education level.

A second issue with the source data for the model is that it is limited to those who are adept at technology. While many younger people would be confident taking the survey, the older generations are the most reliable voters. Older voters who are less comfortable with technology may be underrepresented possibly biasing the results. (this is fixed by post-stratification tho?)

A third limitation of our model is the lack of segregating Nebraska and Maine into congressional districts for predictions of plurality. The other forty-eight states of the US use plurality to decide the electors, however in Nebraska and Maine only two electors are decided this way and the rest are decided based on plurality in each congressional district. This is a minor issue however, as the congressional districts should each follow the plurality of the state and the two states only have 9 of 538 electoral votes.

The 2020 U.S. election is also particularly unique because it is the first federal election for which a majority of ballots will be cast via mail-in voting. The effects of this on voter turnout were not taken into account as a factor in our model.

Another limitation with our predictions is that the survey we use for our model is from June. Recent developments that could swing voter preference are not accounted for in our model and represent a possible bias that can affect the credibility of our predictions. For instance Trump's recent management of the Covid-19 pandemic and the effect it's taken on the economy has experts believing his approval rating is currently declining (The Economist 2020). After contracting Covid-19 himself, many believe Trump will not be able to fulfill his duties as president and cannot handle managing the pandemic response. Overall, current news regarding Trump can cause a bias in our data set of inflating the number of Trump voters. Particularly older age demographics that are concerned the most with Covid-19 will see the most inflation. This will affect not only the model's voting percentage for Trump, but also the relationship between age group and vote. This is the most concerning due to older age demographics typically having a higher proportion of Republican voters (Desilver 2014).

Future Work

Future studies for the next election should primarily study voting preferences of swing states in the US. Since the elections work via plurality of each state rather than the country as a whole, studies should first determine swing states via a quick survey. States with less than a ten percent perceived difference between the two leading parties should be considered for a further study to predict the election winner. This method will be more time and cost-effective than conducting surveys in states where there is a clear frontrunner.

Mail-in voting also provides an additional avenue for future work as mail-in ballots will be used in the 2020 election on a wider scale than any previous U.S. election. An analysis on the habits of voters who voted using a mail-in ballot warrants further investigation.

Finally our model offers the Republican and Democratic Parties future targets to focus their campaigns towards in the next election. They can view states that are close in votership but are favoured towards their opposition and collect more data on the demographics of the state and policies people in the state favour. The parties can also look at the average demographics that support their opponent using our model and survey these specific demographics for changes they'd like to see in their party to increase votership.

Appendix

The code, scripts, and RMarkdown file in this report's analysis can also be found at the following link: https://github.com/brdiep113/forecasting_us_election

References

- Brandon, E. (2020). "Why Older Citizens are More Likely to Vote." Retrieved from: <https://money.usnews.com/money/retirement/aging/articles/why-older-citizens-are-more-likely-to-vote>.
- Bürkner, P (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01
- Bürkner, P (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, 10(1), 395-411. doi:10.32614/RJ-2018-017
- Desilver, D. (2014). "The politics of American generations: How age affects attitudes and voting behavior." Retrieved from: <https://www.pewresearch.org/fact-tank/2014/07/09/the-politics-of-american-generations-how-age-affects-attitudes-and-voting-behavior/>.
- Di Lorenzo, P (2020). usmap: US Maps Including Alaska and Hawaii. R package version 0.5.1. Retrieved from: <https://CRAN.R-project.org/package=usmap>.
- Encyclopedia Britannica. (n.d.). "United States Electoral College Votes by State." Retrieved from: <https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124>.

- Erikson, R., Sigman, K., & Yao, L. (2020). "Electoral College bias and the 2020 presidential election". Retrieved from: <https://www.pnas.org/content/early/2020/10/22/2013581117>.
- Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., & Sobek, M (2020). IPUMS USA: Version 10.0 [American Community Survey 2018]. Minneapolis, MN: IPUMS. Retrieved from: <https://doi.org/10.18128/D010.V10.0>.
- Frey, W. (2020). "Older Voters May Secure A Biden Victory in 2020's Swing States". Retrieved November from: <https://www.brookings.edu/blog/the-avenue/2020/10/28/older-voters-may-secure-a-biden-victory-in-2020s-swing-states/>.
- Gelman, A., & Kennedy, L. (2020). "Know your population and know your model: Using model-based regression and post-stratification to generalize findings beyond the observed sample." Retrieved from: <https://arxiv.org/pdf/1906.11323.pdf>
- JJ Allaire et al. (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. Retrieved from: <https://rmarkdown.rstudio.com>.
- Pew Research Center. (2016). "Ideological Gap Widens Between More, Less Educated Adults". Retrieved from: <https://www.pewresearch.org/politics/2016/04/26/a-wider-ideological-gap-between-more-and-less-educated-adults/>.
- Pew Research Center. (2020). "Trends in party affiliation among demographic groups". Retrieved from: <https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/>.
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: <https://www.R-project.org/>.
- Robinson, D., Hayes, A., & Couch, S. (2020). broom: Convert Statistical Objects into Tidy Tibbles. R Package version 0.7.0 Retrieved from: <https://CRAN.R-project.org/package=broom>.
- Rosentiel, T. (2007). "Money Walks". Retrieved from: <https://www.pewresearch.org/2007/04/12/money-walks/>.
- Rudis, B. (2020). statebins: Create United States Uniform Cartogram Heatmaps. R package version 1.4.0. Retrieved from: <https://CRAN.R-project.org/package=statebins>.
- Tausanovitch, C., & Lynn V. (2019). Democracy Fund + UCLA Nationscape(version 20200814). Retrieved from <https://www.voterstudygroup.org/downloads?key=1b155ffb-aa7a-432a-826e-6b8876e493ac>.
- The Economist. (2020). "How will Donald Trump's covid-19 infection affect the election?" Retrieved from: https://www.economist.com/united-states/2020/10/02/how-will-donald-trumps-covid-19-infection-affect-the-election?fbclid=IwAR33_5og5cqlosoP36fplLcNn6w1qd0smgqQpZZrBXOUwReann7RtvR0bw0.
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag. New York.
- Wickham, H., & Miller, E. (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. Retrieved from: <https://CRAN.R-project.org/package=haven>.
- Wickham et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. Retrieved from: <https://doi.org/10.21105/joss.01686>.
- Wu, C., & Thompson, M. (2020). Sampling Theory and Practice, Springer.