Biden Predicted to Win 55% of Popular Vote

Forecasting the US 2020 Presidential Election Using a MRP Approach

Rebecca Yang

November 2, 2020

Abstract

In 2016, Donald Trump won the presidential election against all predictions, largely bolstered by voters who were undersampled in the polls. In this paper, I seek to predict the winner of the 2020 election, but to avoid this mistake through the use of multilevel regression with poststratification (MRP). I fitted a model using non-representative survey data from the Democracy Fund and UCLA Nationscape survey and post-stratified its results using data from the 2018 American Community Survey (ACS) to obtain representative estimates. The model predicts Biden to win 55% of the popular vote with a margin of error of 4% and 370 (out of 538) electoral votes.

Keywords: Forecasting; US 2020 Election; Trump; Biden; Multilevel Regression with Post-stratification

Introduction

The upcoming 2020 US presidential election is arguably the most high-stakes election in recent American political history. As such, there is extraordinary public interest in election forecasts, which means the onus on pollsters is also extraordinary, especially on the heels of the 2016 election. Donald Trump's 2016 come-from-behind victory to win the presidential election was a shock to many, but especially to pollsters who were practically certain of a Clinton victory, due to her superior performance in the polls. This "polling failure" has retrospectively been attributed to the under-sampling of uneducated white voters in battleground states who turned out for Trump in full force on election day (Jackman 2020). This issue is compounded as pollsters eschew traditional random sampling methods for online panel surveys (Wang et al. 2014). While they are quicker and cheaper, they are all but guaranteed not to be representative samples. In this paper, I attempt to predict the winner of the 2020 presidential election from non-representative survey data, but to avoid these pitfalls through the method of multilevel regression with post-stratification (MRP). Specifically, I am interested in three things: the proportion of votes received by each candidate (popular vote), the electoral college vote breakdown (the actual winner of the election)[1], and the probability of a Trump or Biden win.

I first fitted a Bayesian multilevel regression model to predict an individual's vote choice based on their age, sex, race, income, education, and stat. The data used to fit this model was from the Democracy Fund and UCLA Nationscape survey and used a non-random sample. I then post-stratified the results using representative data from the American Community Survey (ACS). The post-stratification data was divided into cells that

[1] For context, the winner of the US election is determined by the electoral college Each state is allocated a number of electoral votes, and the winner of the popular vote each state get those votes. There are two exceptions to this: Maine and Nebraska split their electoral votes based on the winners in each congressional district. The candidate who wins the majority of the electoral votes (at least 270 out of 538) is the winner of the election. This means that winning the right states is more important than simply winning the most votes, as was the case in 2016 ("Distribution of Electoral Votes," n.d.).

represent each combination of age, sex, race, income, education, and state. The predicted vote choice (Biden or Trump) was estimated for each cell. Finally, I reconstructed the overall vote share of the population based on each cell's relative weight in the population.

The post-stratified results predict Biden to win 55% of the popular vote, with a margin of error of four percentage points. I also found the predicted vote share of each state which allowed me to make predictions for the electoral college outcome. My prediction is that Biden will win 370 out of 538 electoral votes, thereby winning the presidency. Finally, I looked at all the different paths to victory for each candidate. In 4000 different potential scenarios in which the election could play out, Biden won in 99% of them. Interestingly, there was one scenario that ended in an electoral college tie.

These results bode well for Biden supporters. However, these results should be taken with a grain of salt given the unpredictability of elections, especially in this turbulent political climate (once again, 2016 is an example). While MRP allows us to account for some of the mistakes made in the 2016 election, it is by no means a panacea, and I can only resolve to correct the errors of which I am aware. The main limitation of these results is that it does not account for voter turnout,so the predictions are based on the assumption that every eligible voter is indeed a voter, but of course, this is not the case.

This report is divided into four sections. In the first, I describe the data used in fitting the model and post-stratification step. It is followed by the details of the model fitting approach. Then, I present the results from the model and my final predictions. Finally, I conclude with a discussion of these results, as well as weaknesses and directions for future work. The Appendix contains model checks and additional results.

Data

The method I use in my forecast, MRP, allows representative estimates to be obtained from non-representative data by post-stratifying the results with representative data. As such, two data sources are required. The first (non-representative) source is used to fit the model to predict individual-level outcomes. The second source is representative population-level data, which allows the population to be deconstructed into cells for different combinations of demographic traits. The model is then used to predict the outcome for each cell, and the population level outcome is reconstructed according to each cell's relative weight in the population.

The requirement of two data sources means that the variables used in the model are limited to those that are common between the two sources. For example, if my first survey included information on respondents' religion, but I do not have corresponding information about the religious composition of the population, I would not be able to use that as a term in my model. This means that the applications for MRP are mostly limited to outcomes that can be (or at least, believed to be) predicted by demographic characteristics, as this data is available through the US Census and American Community Survey (ACS). Luckily, vote choice is purported to be one of those outcomes (Wang et al. 2014; Park, Gelman, and Bafumi 2004).

Model Training Data

In my case, the first source is from the Democracy Fund + UCLA Nationscape survey (Tausanovitch and Vavreck 2020), which has been surveying people weekly since July 2019. The data used is from Wave 50 of Phase 2 of the survey and was collected between June 25 and July 1, 2020. The survey is a non-probability online survey, as respondents are selected from Lucid, a market research platform that purveys survey respondents for its clients. Respondents were selected from to meet a set of demographic quotas for age, gender, ethnicity, region, income, and education ("Democracy Fund + Ucla Nationscape User Guide" 2020; Chris Tausanovitch and Rudkin 2019). The average response rate across all their surveys was 75%, and the sample size of this particular survey was n=6479. Respondents are asked about their candidate preferences, policy views, and voting attitudes, but the main response of interest was the candidate they would vote for if the election were held now. Respondents were given the following options: Trump, Biden, someone else, "don't know", or "I would not vote".

Since I wanted to fit a model with a binary outcome (vote choice), I filtered out voters who did not support one of the two major party candidates, Trump or Biden. For undecided voters, I took the candidate that they leaned towards as the candidate they would vote for. I then removed respondents based on vote intention—those that said they were not going to vote, ineligible to vote, or uncertain about voting—which left 4,714 respondents in the sample. The idea was that, since the goal is to predict the outcome of the election, the model should be based on people who will actually vote in the election. I used this set of "voters" to train the model to predict an individual's vote choice. However, I remain conscious of the fact that vote intention is notoriously over-reported. I also created age groups from the ages of respondents, because I wanted to capture some generational effects, as well as the effect of the stage of life of the respondent, as this may have bearings on the issues that affect them, and consequently their vote choice.

From the map, it can be seen that there are more respondents in the survey from more certain states and regions, like California and the East Coast, and fewer respondents from the Midwest and Northwest areas. Most notably, Wyoming, North Dakota, and Alaska have 2, 3, and 5 respondents, respectively. The raw survey data suggests higher national support for Biden.

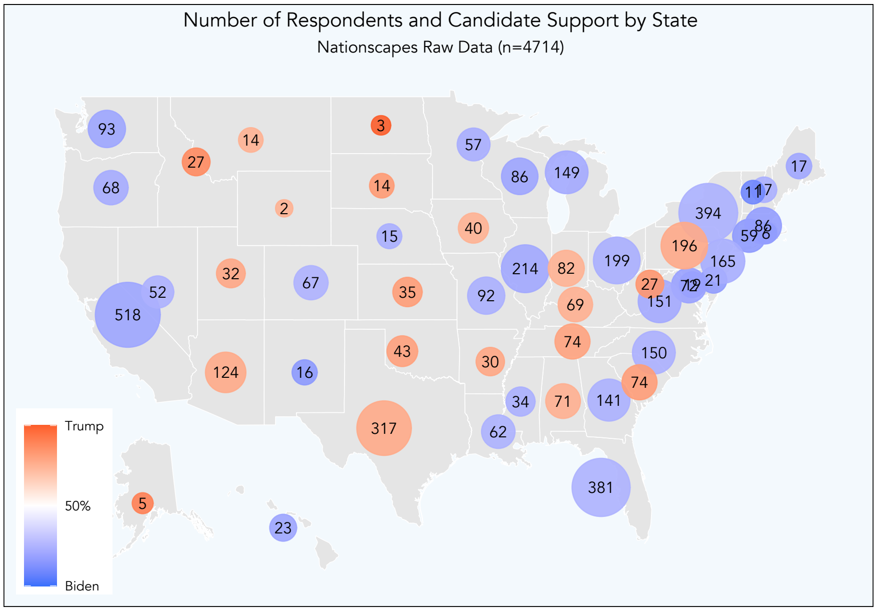

Figure 1: Geographic distribution of Nationscape survey respondents and candidate support by state

Initial comparison of Biden support suggest higher support among female, young, low-income, and black voters. It is similar between education levels.

Post-stratification Data

The data used to post-stratify the results from the model was from the 2018 American Community Survey (Steven Ruggles and Sobek 2020), a yearly survey carried out by the US Census Bureau. The goal of the ACS is to provide accurate estimates of demographic, social, and economic characteristics of the population. The frame, or the source from which respondents are selected, is the Master Address File from the US Census Bureau ("American Community Survey Operations Plan" 2003). It uses cluster sampling in that households are

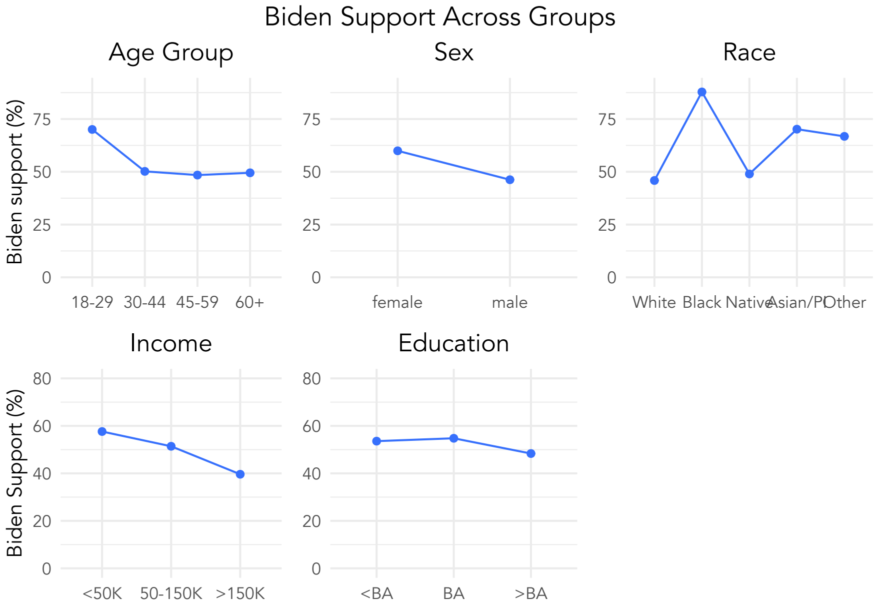

Figure 2: Comparison of Biden support across groups

selected and all members of that household (the cluster) are surveyed, as well as systematic sampling to select households from each county to survey every month. This means that there is a pattern to how households are selected (i.e. every other household).

Although it is not a census, it is a much larger random sample, which gives us more confidence in its ability to approximate the population (n=3,214,538). Nonetheless, it is still subject to non-response and issues like under-representation, which is why response weights have been provided by the ACS. I used these weights when I calculated the weight of each cell in order to better approximate the American population. The weight for each respondents describes the number of people that respondent represents in the population. Each cell represents a different combination of age group, sex, race, education, income, and state, for a total of 15,327 cells, or sub-groups, that we use for post-stratification.[2]

That being said, the American population is not actually the population of interest. The population of interest is voters in the 2020 election. In lieu of access to voter registration databases and voter history records, the best option is to only consider legally eligible voters—that is, citizens aged 18 and older. This left 2,446,456 respondents used to approximate the voting population.

One of the challenges of MRP is synchronizing the variables between the data used to fit the model and the post-stratification data. While variables like age are essentially standardized between surveys, there is a lot more variability in the way questions like income are posed. For example, the Nationscape survey asked respondents to identify their income bracket (i.e. $25-49,999). However, the ACS asked for a self-reported number (i.e. $54,000), so the ACS incomes had to be reduced into brackets to match.

These challenges are further highlighted by variables like race, for which different survey organizations have different treatments of this concept. In some cases, Nationscape offered more specific options for races, but they were similar to the ACS and able to be grouped together to match. The ACS had options for “Chinese”, “Japanese”, and “Other Asian or Pacific Islander”, while the Nationscape survey had seven Asian subgroups and four Pacific Islander subgroups, so these were combined to approximate an “Asian/Pacific Islander” category.

However, in the ACS, respondents were permitted to identify as belonging to two or more major races, but

[2] Evidently, not all combinations exist in the population.

there was not an equivalent multiracial option in the Nationscape survey. While there were indicator variables in the ACS that provided information about which races multiracial individuals belonged to, I did not want to use these as it would cause these individuals to be counted twice in the population. The ACS has indeed recognized the incompatibility of this option with other surveys and offers a "single-race" variable, but the was not available for the 2018 ACS[3]. I consequently grouped these responses together as "other" and considered them analogous to the "some other race" option in the Nationscape survey. Although this new "other" category only counts for 9% of the population, the ambiguity in the multiracial options and this category is far from ideal and leaves open the possibility for information to be muddled or lost. Looked at another way, 9% of a population with 300 million people is still a sizeable amount.

I also used state electoral vote counts (Augustyn et al. 2016) to supplement my predictions for the electoral college outcome.

From the table below, it is apparent that the Nationscape sample was indeed non-representative. The Nationscape sample was more educated and had higher incomes compared to the American population.

Table 1: Comparison of Nationscape and ACS Demographics

Variable	n	Nationscape	ACS
Age Group			
18-29	785	17%	21%
30-44	1482	31%	24%
45-59	1098	23%	24%
60+	1349	29%	31%
Sex			
female	2292	49%	51%
male	2422	51%	49%
Race			
White	3651	77%	76%
Black	526	11%	13%
American Indian/Alaska Native	49	1%	1%
Asian/Pacific Islander	208	4%	1%
Other	280	6%	9%
Education			
<BA	2630	56%	70%
BA	1203	26%	19%
>BA	881	19%	11%
Income			
<50K	2150	46%	71%
50-150K	2029	43%	23%
>150K	535	11%	4%

Model

The model was fitted using the `brms` (Bürkner 2017, 2018) and `rstan` (Stan Development Team 2020) packages in `R` (R Core Team 2020). I used the default settings of 4 parallel chains with 2000 iterations for each, of which

[3] Probabilities that an individual belongs to each race are calculated, and the variable takes the value of the "most-likely" race. It was last available in 2014. I would also hesitate to use it due to the amount of processing behind the scenes that takes place for this variable.

half were warmup chains.

I fit a Bayesian multilevel regression model to predict an individual's vote choice (Trump or Biden) based on their age group, sex, race, education, income, and state. Since the individual outcome was binary, the model was of the Bernoulli family.

The general form of the model is as follows:

$$P(\text{vote}_i = 1) = \text{logit}^{-1}\left(\beta_0 + \beta_1(\text{sex}_i) + \alpha^{state}_{s[i]} + \alpha^{age}_{a[i]} + \alpha^{race}_{r[i]} + \alpha^{income}_{n[i]} + \alpha^{educ}_{e[i]}\right)$$

$$\text{vote}_i = \begin{cases} 0 & \text{if Trump} \\ 1 & \text{if Biden} \end{cases}$$

$$\text{sex}_i = \begin{cases} 0 & \text{if Male} \\ 1 & \text{if Female} \end{cases}$$

$$\begin{aligned} \alpha^{state}_s &\sim \text{N}(0, \sigma_{state}) \\ \alpha^{age}_a &\sim N(0, \sigma_{age}) \\ \alpha^{race}_r &\sim N(0, \sigma_{race}) \\ \alpha^{income}_n &\sim N(0, \sigma_{income}) \\ \alpha^{educ}_e &\sim N(0, \sigma_{educ}) \\ \sigma_{state}, \sigma_{age}, \sigma_{race}, \sigma_{income}, \sigma_{educ} &\sim \text{t}^+(0, 3, 2.5) \end{aligned}$$

In other words, the probability that an individual voter i will support Biden depends on their state, their age group, race, income, and education.

As mentioned earlier, the variables included in the model are limited by the ACS, so they are all demographic variables. Nationscape selected their participants based on age, sex, income, race, and education. Given that respondents were selected based on their membership in these groups, a multilevel model was fitted according to these groups. These demographic variables, with the exception of sex, were all included in the model as group level effects, or varying intercepts, as each level has its own intercept that contributes to the predicted probability.

This multilevel structure allows for the data to be pooled between groups. This means that if there are is a small number of respondents for a group in the sample, the estimate for that group is not completely dependent on those few responses. Instead, those responses will have some effect, but the estimate will also be informed by the other groups. For example, even though North Dakota has three responses in the Nationscape sample, its state effect will be drawn towards the mean of the other states' effects, since three respondents is hardly a sufficient representation. The main underlying assumption here is that the groups are different, but not altogether independent from each other (Kennedy and Gelman 2020). While the rationale could extend the sex variable, grouping has been found to be inconsequential for groups with less than 3 levels, so it is included as a categorical indicator variable here (Park, Gelman, and Bafumi 2004). Since there are 15,327 post stratification cells, and 4714 observations in the original data, the ability to pool the data is indeed utilized here.

Income is included in the model as it has been found to affect voting patterns in different states (Gelman et al. 2008). In addition, the 2016 election, there were larger-than-normal gaps in vote choice between races, sexes, and education levels (Tyson and Maniam 2016). States are included in the model since the presidential election is essentially decided on a state-by-state basis, and the mainstream categorization of states as "red", "blue", and "swing" reflects the differences in political preferences between states. States also have their own governments, political systems, and cultures that affect their residents. Previous research supports the ability of these to predict election results (Park, Gelman, and Bafumi 2004, @yair_gelman, @xbox).

Table 2: Variable Descriptions

Effect	Levels	Description
Population		
Sex	2	Male or Female
Group		
State	51	50 states and District of Columbia
Age Group	4	18-29, 30-44, 45-59, 60+
Race	5	White, Black, Asian/Pacific Islander, American Indian/Alaska Native, Other
Education	3	<BA, BA, >BA
Income	3	<50K, 50-150K, >150K

Earlier versions of the model were fitted as generalized linear mixed effect regression models, but I opted for a Bayesian version instead. The Bayesian nature of the model leaves open the possibility for informative priors to be included in the model. Given the amount of data and research on elections and voting, there is definitely prior knowledge that could be used as a starting point for the model's predictions. Elections have happened in the past, and they indeed can inform future ones, so the model need not start from scratch. At this stage, I simply used the default prior distributions estimated from the data. The population level effects were given flat priors, and the standard deviations were given (half) t-distributions. In addition, this allowed for prediction intervals to be generated for the estimates and potential election outcomes to be generated from the data.

In the same vein, I considered using Trump's 2016 vote share in each state as a linear predictor in my model as was done by Ghitza and Gelman (2013) and Wang et al. (2014). The intention was to give the model an empirical foundation, so that it was not completely reliant on the polls. However, I found that the state effects conveyed similar information in the model and the predictions were nearly the same. While the Nationscape survey did ask about the respondent's vote in 2016, it was not a viable alternative to include in the model because the post-stratification data does not contain this information at the respondent level.

At this point, I would like to openly and matter-of-factly address the convergence of the model. Indeed, in the sampling for the model, there was one divergent transition. This is an indication that the model may not be a good fit for the data at this stage, but there are ways in which this could be fixed through the algorithm settings[4]. These, however, are computationally expensive, so for the purposes of this report, I proceed with the sub-par model. Thus, these results ought to be taken with the tiniest grain of salt. Additional model checks can be found in the Appendix.

After the model is fitted, the results can be post-stratified according to the following:

$$\text{Biden Popular Vote Share} = \frac{\sum_{j=1}^{J}(\text{vote}_j)(\text{total people in cell}_j)}{\sum_{j=1}^{J}(\text{total people in cell}_j)}$$

where J is the total number of cells for the population

[4] I used `adapt_delta` = 0.99, but it has also been suggested to change `step_size` and `max_treedepth` from the defaults

Results

Model Results

The model coefficient and intercept estimates are displayed below. The state estimates have been truncated for brevity so the most notable ones are shown below (additional results can be found in the Appendix). Bear in mind, these estimates are made on the logit scale.

Sex was found to have a significant effect, as its prediction intervals did not contain zero. According to the model, men are less likely to support Biden than women.

Younger voters, voters in lower income brackets, and less educated voters were predicted to be more likely to support Biden, while white and wealthier voters were more likely to support Trump. Black people were predicted to be more likely to support Biden; this was the most notable race effect as its prediction interval did not contain zero. For the states, California and Massachusetts were found to favor Biden the most heavily, while Texas and South Carolina strongly favored Trump.

Table 3: Model Estimates

Variable	Level	Estimate	2.5%	97.5%
Group Effects				
age_group	18-29	0.46	-0.10	1.14
	30-44	-0.11	-0.72	0.53
	45-59	-0.21	-0.83	0.44
	60+	-0.07	-0.68	0.58
education	<BA	-0.15	-0.85	0.61
	BA	0.10	-0.57	0.88
	>BA	0.08	-0.60	0.86
income	<50K	0.27	-0.70	1.43
	50-150K	0.11	-0.87	1.27
	>150K	-0.29	-1.31	0.84
race	White	-0.71	-1.74	0.36
	Black	1.32	0.30	2.43
	American.Indian/Alaska.Native	-0.70	-1.85	0.41
	Asian/Pacific.Islander	0.13	-0.91	1.25
	Other	-0.04	-1.06	1.05
state	CA	0.26	0.05	0.47
	MA	0.40	0.04	0.80
	SC	-0.45	-0.88	-0.07
	TX	-0.34	-0.59	-0.11
Population Effects				
(Intercept)	-	0.70	-1.08	2.27
sexmale	-	-0.41	-0.53	-0.28

The distributions for sex confirm the above findings. The distributions for the standard deviations for the states, income, education, and age groups are all very close to zero, with higher values for race. This suggests that the levels in the groups may not be drastically different from each other.

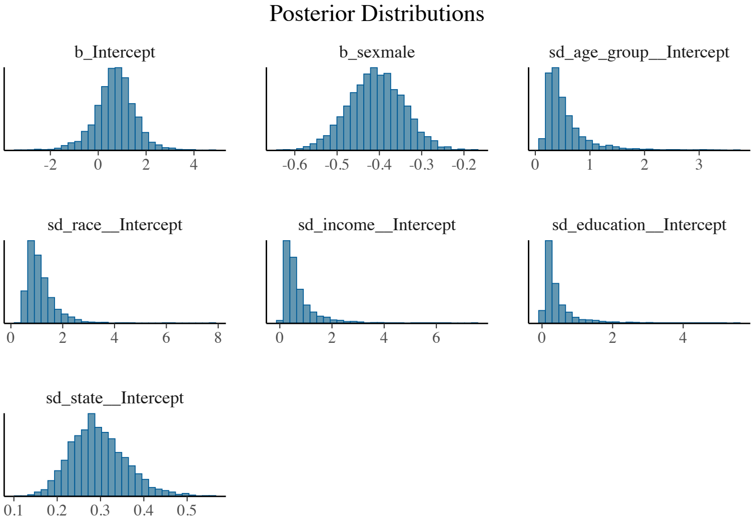

Figure 3: Posterior distribution of model terms

Post-stratification Results

In the figure below, the most notable differences between the raw data and post-stratified estimates for Biden support in each states are shown.

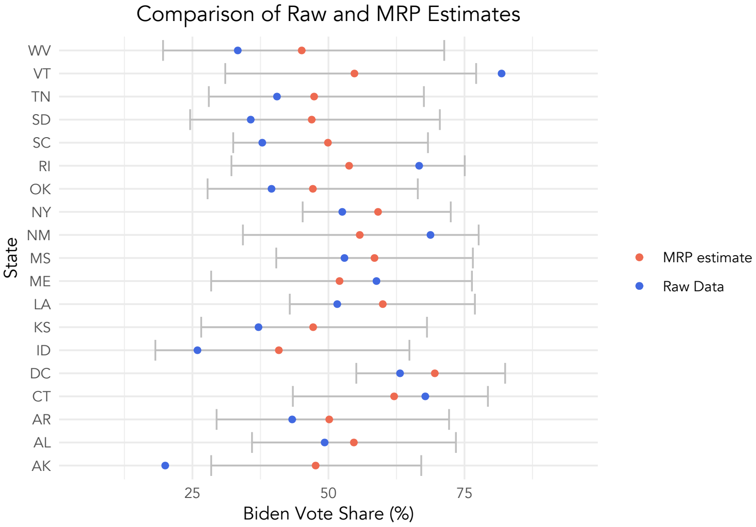

Figure 4: Comparison of MRP and Raw Estimates

The effect of post-stratification is evident. The most significant differences between the raw and MRP estimates are in Vermont and Alaska, as they are not even contained in the 95% interval for the MRP estimate.

Predictions

Here I present my final forecast for the outcome of the election.

I predict Biden to win the popular vote with 55% of the vote share, with a margin of error of 4 percentage points.

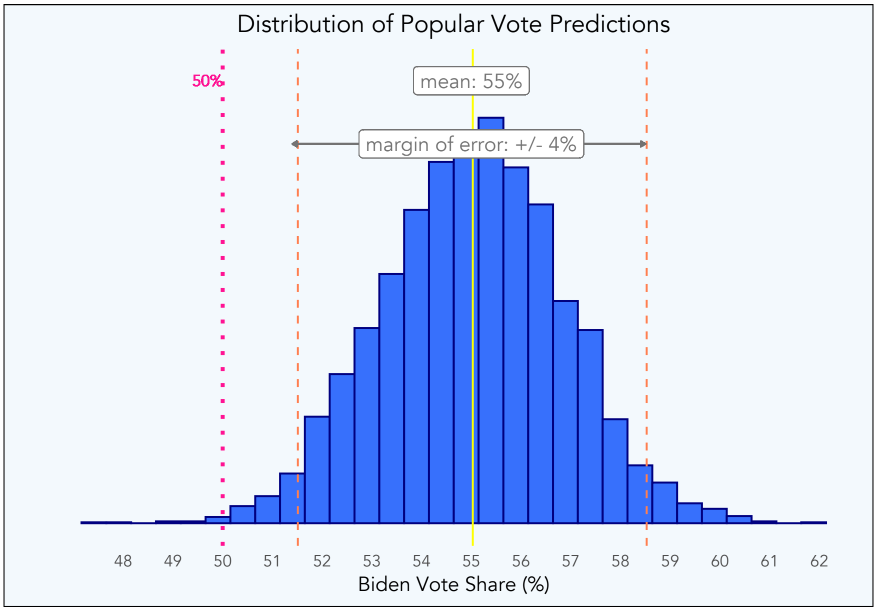

Figure 5: Distribution of popular vote predictions

Biden is predicted to win 370 electoral votes (Figure 6). The prediction interval for this estimate is between 284 and 444 votes. The map below shows the predicted electoral college outcome based on the predicted winner of the popular vote in each state.

I also looked at 4000 possible paths to victory, or ways the election could play out (Figure 7). This considered scenarios in which Trump won certain states or Biden lost certain states, and how the outcome of the election would be affected. Of these hypothetical elections, Biden won approximately 99% of them (3953 out of 4000), Trump won 1% (46 out of 4000), and interestingly, less than 0.1% of them ended in ties (1 out of 4000). Apparently, nothing is off the table in the election.

Discussion

According to the rather low standard deviations for the state intercept's distribution, the intercepts for the states do not appear to be drastically different. While intuitively the grouping should make sense, I hypothesize that this is because there are too many states (50, plus the District of Columbia), and they are not "different enough" from each other. When people think about state-level differences, they are usually thinking about the difference between, say, California and Alabama. However, I expect the differences between North and South Dakota are a little more nuanced. Broader groups, such as region (Northeast, Midwest, South, etc.) could potentially be more revealing groupings of the data. I could also consider grouping the states based on whether they are a state that historically favors one party (i.e. a red or blue state) or toggles between the two (i.e. a swing state).

Additionally, this would have implications on the way the data is pooled. For example, since North Dakota only had three respondents, its intercept would be pulled toward the mean of all the other states. However, it can be argued that it would make more sense for the North Dakota to pulled more towards South Dakota and its other neighbors. Gao et al. (2020) propose that this also could be controlled through the types of priors specified and improve MRP estimates as a result.

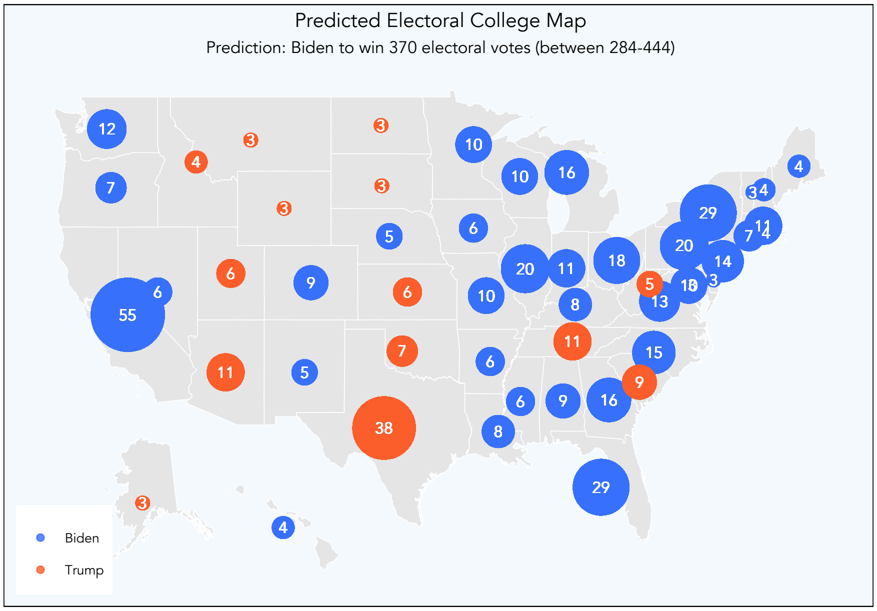
Predicted Electoral College Map
Prediction: Biden to win 370 electoral votes (between 284-444)
12
3
4
3
10
7
10
16
3
4
3
4
3
29
5
6
7
11
4
20
9
6
6
20
11
18
14
6
55
5
10
10
3
13
8
11
7
15
11
5
6
9
6
9
16
8
38
29
3
4
Biden
Trump

Figure 6: Predicted electoral college map

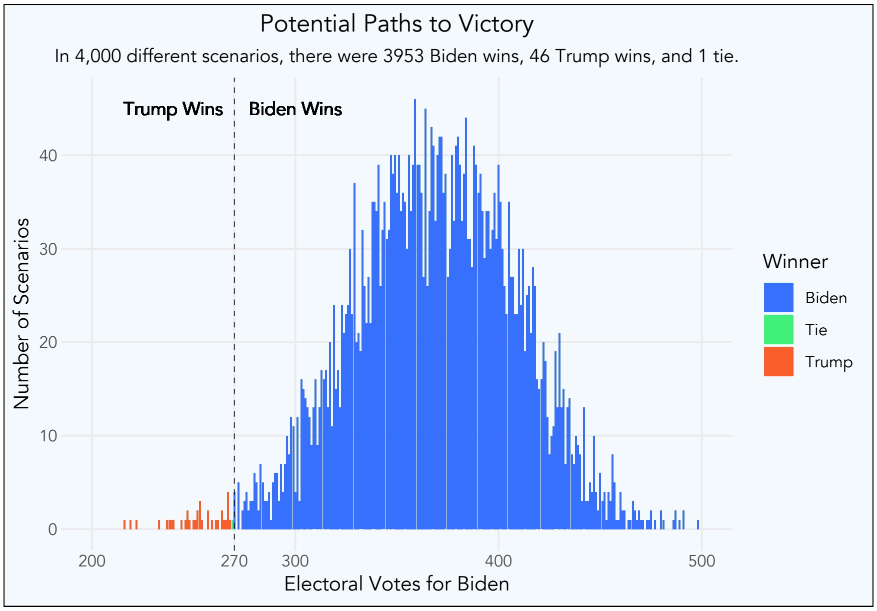

Figure 7: Potential Paths to Victory for Biden and Trump

On the opposite end, the most notable intercept is race, specifically with regard to black voters. Given the current political climate, Black Lives Matter, and the resurgence of white supremacists, this is not altogether surprising.

With regard to my predictions, they are very heavily in Biden's favor. Of course, this was the same mindset many had going into the 2016 election, which did not end well, so I err on the side of caution. If he wins the right states, and the swing states tip in his favor, a Trump win should not be written out entirely. According to Figure 7, there are, although few, worlds in which Trump wins, so a Biden win should not be taken as a given. These are only based on 4,000 different paths to victory as well; there are tens of thousands of possibilities. I also reiterate here that these results were based on a questionable model, so the appropriate weight should be placed on its results.

While the margin of error for the population level vote share is around 4%, the margins of error for the state level estimates are much larger, as illustrated by Figure 4. In some states, prediction interval spans 50%. Since the winner of the election hinges on state-level outcomes, these large margins are cause for concern. In the future, I could potentially improve this by setting stronger priors for the model. This model just used the default priors, so I would expect that informative priors would indeed provide more certainty to the estimates. Stronger priors could also help with the issue of divergent transitions.

Limitations

Models are only as good as the data used to fit them

In this case, the data is from the end of June, four months prior to the election. Since then, major events have occurred that may or may not have made an impression on voters — Trump's COVID-19 diagnosis, investigations into Hunter Biden's dealings in Ukraine and China, the debates, and the conventions, to name a few. In the same vein, surveys are subject to response bias: if Trump is having a bad day in the press, his supporters may be more likely to rush to his defense, or, alternatively, they may be reluctant to tell pollsters that they intend to vote for Trump. Even though the Nationscape survey sources its respondents through a research company and has relatively high response rates, it is not immune to the same issues that plague traditional surveys. Furthermore, Gelman and King (1993) found that elections are not very predictable from surveys conducted months in advance of the election; it is only in the lead up to the election that the polls stabilize and gain predictive power.

The Nationscape survey also supplied weights to help improve the representativeness of their sample. According to Ghitza and Gelman (2013), unweighted regression with MRP is acceptable when all the variables that were weighted on are included in the model, since they will be adjusted for in the post-stratification step. While they calculated their weights according to the 2017 ACS, they weighted on additional factors like language spoken at home, Hispanic ethnicity, birthplace, and language spoken at home, that I did not include in my model. Ghitza and Gelman (2013) suggest estimating a design effect in this case. This is something that could be explored further.

The model lives in a perfect world

The model's predictions are based on the assumption that every eligible voter performs their civic duty, and their vote indeed counts towards the final tally. However, the election process, of course, is not that seamless. There are a huge number of external factors that could influence the results of the election. For example, due to the COVID-19 pandemic, mail-in voting is supplanting in-person voting, bringing with it a slew of logistical challenges. There is also the possibility of voter suppression that disproportionately affects certain groups, the spreading of voter misinformation, or even measures to obstruct the outcome of the vote after it concludes

(Alba, Robertson, and Thrush, n.d.).

Furthermore, in the absence of compulsory voting, voter turnout plays a huge factor in the results of an election. As mentioned earlier, the basis for the post-stratification data set, the ACS, is meant to represent the American population at large, which means it includes voters and non-voters alike. While this can slightly be remedied by restricting the data to legally eligible voters, historical voter turnout rate has hovered around 62% (File, n.d.). Even though turnout is expected to be higher than average this election, it is still a great source of variability and uncertainty that influences the result of the election.

In the future, I could consider two options to account for this. First, an auxiliary model to predict voter turnout, since the Nationscape survey does ask about vote intention[5]. The voter turnout rate can be predicted for each post-stratification cell, and the support for each candidate calculated from this, as described by (Park, Gelman, and Bafumi 2004). However, these responses are subject to over-reporting, as mentioned earlier, and additional models always complicate things. A second, simpler, modern solution, is to utilize a better post-stratification data set. Voter-registration databases have been shown to be a new and highly effective resource for election forecasting with MRP (Ghitza and Gelman 2020). These databases contain records of registered voters and voting history, so they are much more well-calibrated to the the goals of the model. However, the ACS is freely available, which allows the model to be reproducible.

While the model produces state-level estimates which can be used to predict the electoral college results, it does not fully account for all features of the US electoral system. Namely, it fails to consider the two outlier states, Nebraska and Maine, which allocate their electoral college votes based on the winner in each congressional district, so such votes can be split between candidates. However, congressional-district-level data is not available in the ACS and county-level data is too involved, so this is a trade-off I am willing to make.

In addition, given that the model is configured to predict a binary outcome, it relies on there being two major party candidates. Of course, at this stage in the 2020 election, this is not an issue, but is something to consider for future models, should a third party challenger arise[6].

Future Work

In addition to the areas of improvement mentioned above, the model could also be expanded in several ways. There is a belief that the election can be predicted based on fundamental variables like the state of the economy and presidential popularity, and these are indeed incorporated in conjunction with survey data in the model used by The Economist ("How the Economist Presidential Forecast Works," n.d.). Were I to do this, it would give the model an empirical foundation so that it was not completely reliant on the polls. Complexity could also be added to the model through interaction terms. For example, Gelman et al. (2008) found that income had different effects on voting preferences in states depending on whether a state was relatively rich or poor.

Forecasting elections is a unique type of modeling exercise because there will actually be confirmation (or contradiction) of the model's predictions. When there are contradictions, there is something to be learned about the model and its shortcomings. In this paper, I have attempted to incorporate what was learned in the past election through the use of MRP to curb non-representativity. I conclude with a disclaimer of sorts. In the words of George Box, "All models are wrong, but some are useful."[7]

[5] Earlier iterations of the model fiddle around with this but were put on the back burner.

[6] In 1992, Ross Perot ran as an independent and won 19% of votes (Levy 2020).

[7] Although this is probably not one of those useful ones.

Appendix

Model Checks

The posterior predictive check is a way to compare the observed data to the model's predictions based on the same data. From the graph below, it appears that the model's predictions are centered around the observed data.

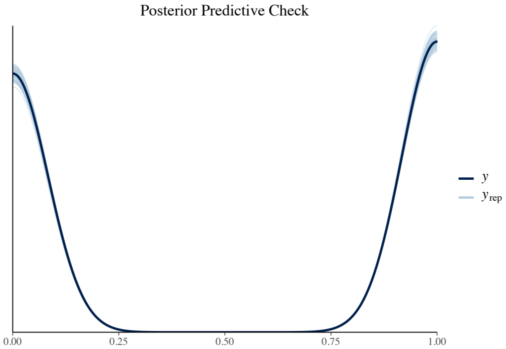

Figure 8: Posterior predictive check for the model

While the R-hat values are 1.00 and 1.01, it can be seen that the chains for age group, education, income, and race are not particularly well mixed.

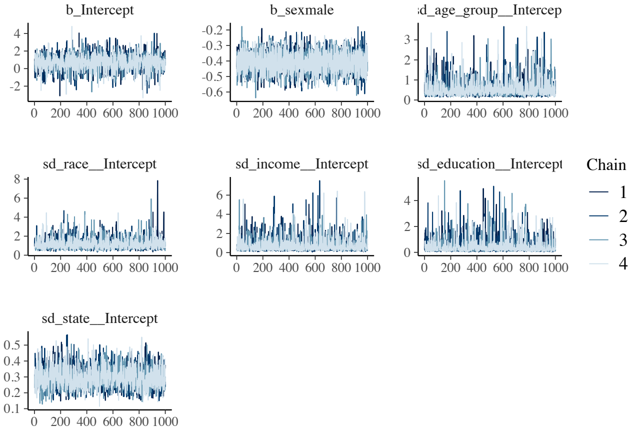

Figure 9: Trace plots to check model convergence

The loo output confirms that there are no influential observations, as measured by the Pareto k estimates.

```
loo(vote_model)
```

```
##
## Computed from 4000 by 4714 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -2980.3 22.4
## p_loo        37.9  0.5
## looic      5960.5 44.9
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

Model Results - More Details

Table 4: Model Results

group	Variable	Estimate	2.5%	97.5%
age_group	sd__(Intercept)	0.54	0.17	1.63
education	sd__(Intercept)	0.46	0.07	2.03
income	sd__(Intercept)	0.71	0.16	2.54
race	sd__(Intercept)	1.14	0.51	2.54
state	sd__(Intercept)	0.30	0.18	0.43

Codes

Code used in this analysis and instructions to reproduce can be found at: https://github.com/reb-yang/Forecasting-Election.git

Citations

Alba, Davey, Campbell Robertson, and Glenn Thrush. n.d. "2020 Election Live Updates: Early Votes Near 100 Million as Turnout Heads Toward a Record." *New York Times*. New York Times. https://www.nytimes.com/live/2020/11/02/us/trump-biden-election#battleground-states-are-seeing-the-most-voting-misinformation.

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown.

"American Community Survey Operations Plan." 2003. US Census Bureau. https://usa.ipums.org/usa/resources/codebooks/ACS_codebook.pdf.

Arnold, Jeffrey B. 2019. *Ggthemes: Extra Themes, Scales and Geoms for 'Ggplot2'*. https://CRAN.R-project.org/package=ggthemes.

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Augustyn, Adam, John M. Cunningham, Bruce Giles, Michael Levy, and Amy Tikkanen. 2016. "United States Electoral College Votes by State." Encycloaedia Britannica, inc. https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124.

Bolker, Ben, and David Robinson. 2020. *Broom.mixed: Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Bürkner, Paul-Christian. 2017. “Brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/doi:10.18637/jss.v080.i01.

———. 2018. “Advanced Bayesian Multilevel Modeling with the R Package Brms.” *The R Journal* 10 (1): 395–411. https://doi.org/doi:10.32614/RJ-2018-017.

Chris Tausanovitch, Tyler Reny, Lynn Vavreck, and Aaron Rudkin. 2019. “Democracy Fund + Ucla Nationscape Methodology and Representativeness Assessment.” https://www.voterstudygroup.org/uploads/reports/Data/Nationscape-User-Guide_2020sep10.pdf.

“Democracy Fund + Ucla Nationscape User Guide.” 2020. Democracy Fund + UCLA Nationscape. https://www.voterstudygroup.org/uploads/reports/Data/Nationscape-User-Guide_2020sep10.pdf.

“Distribution of Electoral Votes.” n.d. *National Archives and Records Administration*. National Archives; Records Administration. https://www.archives.gov/electoral-college/allocation.

File, Thom. n.d. “Voting in America: A Look at the 2016 Presidential Election.” *Random Samplings*. US Census Bureau. https://www.census.gov/newsroom/blogs/random-samplings/2017/05/voting_in_america.html.

Gao, Yuxiang, Lauren Kennedy, Daniel Simpson, and Andrew Gelman. 2020. “Improving Multilevel Regression and Poststratification with Structured Priors.” *Bayesian Analysis*, July. https://doi.org/10.1214/20-ba1223.

Gelman, Andrew, and Gary King. 1993. “Why Are American Presidential Election Campaign Polls so Variable When Votes Are so Predictable?” *British Journal of Political Science* 23: 409–51.

Gelman, Andrew, Boris Shor, Joseph Bafumi, and David Park. 2008. “Rich State, Poor State, Red State, Blue State: What’s the Matter with Connecticut?” *Quarterly Journal of Political Science* 2 (4): 345–67. https://doi.org/10.1561/100.00006026.

Ghitza, Yair, and Andrew Gelman. 2013. “Deep Interactions with Mrp: Election Turnout and Voting Patterns Among Small Electoral Subgroups.” *American Journal of Political Science* 57 (3): 762–76. https://doi.org/10.1111/ajps.12004.

———. 2020. “Voter Registration Databases and Mrp: Toward the Use of Large-Scale Databases in Public Opinion Research.” *Political Analysis* 28 (4): 507–31. https://doi.org/10.1017/pan.2020.3.

“How the Economist Presidential Forecast Works.” n.d. *The Economist*. The Economist. https://projects.economist.com/us-2020-forecast/president/how-this-works.

Jackman, Simon. 2020. “The Us Presidential Election Might Be Closer Than the Polls Suggest (If We Can Trust Them This Time).” *The Conversation*. https://theconversation.com/the-us-presidential-election-might-be-closer-than-the-polls-suggest-if-we-can-trust-them-this-time-141988.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Kennedy, Lauren, and Andrew Gelman. 2020. “Know Your Population and Know Your Model: Using Model-Based Regression and Poststratification to Generalize Findings Beyond the Observed Sample.” http://arxiv.org/abs/1906.11323.

Levy, Michael. 2020. “United States Presidential Election of 1992.” Encyclopaedia Britannica. https://www.britannica.com/event/United-States-presidential-election-of-1992.

Murphy, William. 2020. *Fiftystater: Map Data to Visualize the Fifty U.s. States with Alaska and Hawaii Insets*. https://github.com/wmurphyrd/fiftystater.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–85. https://doi.org/10.1093/pan/mph024.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2020. "RStan: The R Interface to Stan." http://mc-stan.org/.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS Usa: Version 10.0 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "June 25-July 1, 2020 (Version 20200814)." Democracy Fund + UCLA Nationscape. https://www.voterstudygroup.org/downloads?key=a19c1a98-554a-474a-b513-70b2afb78ed2.

Tyson, Alec, and Shiva Maniam. 2016. "Behind Trump's Victory: Divisions by Race, Gender, Education." Pew Research Center. https://www.pewresearch.org/fact-tank/2016/11/09/behind-trumps-victory-divisions-by-race-gender-education/.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. "Forecasting Election with Non-Representative Polls." *International Journal of Forecasting* 31 (September). https://doi.org/10.1016/j.ijforecast.2014.06.001.

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

———. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

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, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.