Forecasting popular and electoral vote results for the 2020 US election using logistic regression with post stratification

The younger generation, favourability with women and stronger support amongst Blacks & African Americans could give Biden a slim margin

Ananya Jha, Aditi Khetwal, Sukhmani Khaira, Medha Srivastava

2nd November, 2020

Code and data supporting this analysis is available at: https://github.com/jhanan1/forecasting-US-election-using-MRP

Model

Data

Before constructing our model, we had to clean and simplify both our datasets. As we wanted to run two different models to predict the proportion of votes for Trump and Biden individually, we first formed two new binary variables: vote_biden and vote_trump. Being binary variables, they equaled 1 if the respondent indicated they would vote for the candidate and 0 otherwise. Since these variables only indicate whether the person will vote for Biden or Trump, we fail to account for the people who will not vote or those who haven't decided who they will vote for (a repercussion of this is seen in the post stratification result section where the proportion of votes the two candidates receive does not add up to 1). To match our survey variables to the census data variables, we converted the race and employment variables to include only 3 categories- "White", "Black or African American" or "Other" for race and "Unemployed", "Employed" or "Not in labour force" for employment. We also had to ensure the categories and column names in both our datasets (the survey and the census) matched in order to make predictions accurately.

Model Description and Features

The American society is currently tainted with a ravaging pandemic, revolutionary civil unrest, financial turmoil, and factionalism on various social issues. Given this backdrop that the U.S. 2020 Elections are happening in, the two candidates are starkly different. The New York Times Editorial Board publicly endorsed Biden's campaign in their take on the political scene in the USA. They called him both welcome and urgently needed after four years of the most divisive president in modern times[1](Kingsbury, 2020). Not only do Joe Biden and Donald Trump have disparate favourability ratings amongst different communities, they also have opposing positions on issues such as abortion and gun reform.

There were over 417 mass shootings in the US in 2019[2] (Silvestein, 2020) which ignited the conversation on gun control. While Biden has called for a ban on assault weapons and highlighted the importance of background checks[3] (Biden, 2020), Trump has mostly kept quiet on his stance on the issue. The Republicans believe that any form of gun control is an infringement of their 2nd Amendment rights, and thus those opposed to gun control associate with Trump.[4](Pearce, 2020). The death of Supreme Court Justice Ruth Bader Ginsburg sparked further conversations on abortion, and highlighted the divide between the two candidates' stances. Both personally and as a Democratic Nominee, Biden identifies as 'pro-choice' and supports a woman's right to make decisions about her body. Trump, on the other hand, has publicly attended 'pro-life' rallies and supports a near total ban on abortion. His controversial statements about women and orthodox notions about femininity have made him unpopular amongst women. Moreover, Trump's failed attempts at controlling protests the increased cases of police brutality against people of colour further reduced his support base in the BIPOC community. In contrast, Biden's running mate Kamala Harris, who is the first black and Asian-American nominee for the vice-presidential ticket, is a key factor in how his image plays out in the black and female community.

Generational differences are huge in 2020 as this will be the first time the election will be dominated by people younger than 40. The younger generations differ greatly from the Baby Boomers in terms of attitudes towards race, gender, and employment. Lastly, both the Republican and Democratic parties have different takes on how the economy should run. Because of the Democratic party's leftist ideologies, low and middle-income families favour Biden, while big businesses are in support of Trump's right-wing policies and tax cuts.

Given how the factors mentioned above influence the way Americans will vote in the US elections, we choose them as our predictor variables. We investigated income and education as variables too but decided against them as they could be correlated. As we are looking at popular vote over electoral vote, we choose not to include states in our variables (also opinions are changing and states are showing split verdicts over the popularity of the 2 candidates).

Model Specifics

We chose to use a logistic regression model in R to predict the probability of a vote for either Joe Biden or Donald Trump. We used explanatory variables such as abortion and gun control, as these are historically divisive issues in American politics, and many times present stark contrasts between Democratic and Republican nominees. Finally, we controlled for demographic characteristics such as age, gender and employment status.

Model for Joe Biden:
p : Probability of Vote cast for Joe Biden

$$log(\frac{p}{1-p}) = 0.839 - 0.0831S_m + 0.00081Age + 0.0488E_{LF} - 0.0199E_U - 0.199R_o - 0.275R_w - 0.148G_D - 0.116G_{ns}$$

$$+0.112NY_y - 0.188A_d - 0.112A_{na} - 0.0786A_{ns}$$

Model for Donald Trump:
k: Probability of Vote cast for Donald Trump

$$log(\frac{k}{1-k}) = -0.146 + 0.0945S_m + 0.00273Age - 0.0854E_{LF} - 0.0363E_U + 0.167R_o + 0.294R_w + 0.128G_D - 0.0673G_{ns}$$

$$-0.0428NY_y + 0.166A_d + 0.0847A_{na} - 0.000744A_{ns}$$

For our first model, 0.839 is the intercept and iy corresponds to a probability of 69.8% that a voter with a value of 0 for every characteristic controlled for in our model will vote for Joe Biden. Similarly, for our second model the intercept value of -0.146 suggests a probability of 46.4% that a voter who is female, age 0, employed, black, pro gun control, doesn't have the New York Times as their primary news source, and agrees with legal abortion at anytime, will vote for Donald Trump. Finally, for purposes of our post-stratification analysis, we denote the predicted percent of the popular vote that Joe Biden will receive as $\hat{y_B}^{PS}$ and the predicted percent of the popular vote Donald Trump will receive as $\hat{y_T}^{PS}$.

Post-Stratification

After completing the first part of our analysis, we moved to the second and most important part- Post Stratification. Post-Stratification is a common technique used in survey analysis to incorporate population distribution into the estimates. By grouping similar units during sampling, we can avoid nonsampling errors and reduce variance of estimates. To estimate the proportion of voters voting for Joe Biden and Donald Trump, we use this technique on our two models constructed above.

To calculate the proportion of voters, we used demographics and our survey data to estimate how the entire population would vote. Based on our model we predicted how individuals in each cell would vote, and then multiplied it by the number of voters in that cell. Then, we summed up this value for all our cells and finally, divided it by our total population to get a prediction of percentage of voters voting for our response variable.
We calculated

$$\hat{y_B}^{PS} = \frac{\sum N_i \hat{y_j}}{\sum N_j}$$

and

$$\hat{y_T}^{PS} = \frac{\sum N_i \hat{y_j}}{\sum N_j}$$

where $\hat{y_j}$ is the estimate for that candidate's proportion of votes and N_j is the population size of the j^{th} cell based off demographics. We used the variables that we could match to our census data. While three of our variables from our previous model couldn't be matched (due to unavailability of non demographic variables in census data), the remaining four variables- gender, age, employment and race_ethinicity (each with multiple levels) were used to create a post stratification dataset which contained the number of people in each possible combination of the four variables.

When looking at the proportion of voters, we modelled both candidates instead of just one candidate because we wanted to analyse the coefficients on variables and compare them too. This allowed us to compare how both candidates fare amongst the voters depending on various variables.

Additional Information

As seen 4 years ago, popular vote results are not very reliable and so, as an additional step, we decided to use post stratification to predict the electoral college votes for Joe Biden as well. In the U.S election system, candidates are elected directly by popular vote, but the president and vice president are not elected directly by citizens. They are chosen through the electoral college. To explain it in brief, each state gets a number of seats (which differs according to the state) and all seats from that state go to the candidate who won the popular vote (citizens vote for electors and the winning candidate's state political party selects the individuals who will be electors).

From our previous section, we had predictions of the popular vote division and to calculate the electoral vote, we used state wise prediction of how people in that state would vote. Using that proportion, we checked which candidate is expected to win in each state (by comparing their probabilities of winning) and finally summed up the number of electoral college seats in each state where Joe Biden was expected to win. For this section, we also added state as another predictor variable because five total predictor variables (4 from the previous section and state).

Results

Coefficient tables

names	model1	model2
(Intercept)	0.8376749	-0.1460695
abortion_any_timeDisagree	-0.1880305	0.1664765
abortion_any_timeNot Asked	-0.1124363	0.0847478
abortion_any_timeNot Sure	-0.0786284	-0.0007435
age	0.0008097	0.0027342
ban_gunsDisagree	-0.1483078	0.1283213
ban_gunsNot sure	-0.1158072	-0.0672644
employmentnot in labor force	0.0488497	-0.0854133
employmentunemployed	-0.0198691	-0.0362504
genderMale	-0.0831111	0.0944693
news_sources_new_york_timesYes	0.1117582	-0.0427714
race_ethnicityOther	-0.1987193	0.1665331
race_ethnicityWhite	-0.2757094	0.2941943

Model1 table

term	estimate	std.error	statistic	p.value
(Intercept)	0.8376749	0.0337930	24.788426	0.0000000
genderMale	-0.0831111	0.0121763	-6.825676	0.0000000
age	0.0008097	0.0003994	2.027372	0.0426657
employmentnot in labor force	0.0488497	0.0147360	3.314991	0.0009216
employmentunemployed	-0.0198691	0.0169540	-1.171938	0.2412656
race_ethnicityOther	-0.1987193	0.0234655	-8.468566	0.0000000
race_ethnicityWhite	-0.2757094	0.0189888	-14.519551	0.0000000

term	estimate	std.error	statistic	p.value
ban_gunsDisagree	-0.1483078	0.0153934	-9.634497	0.0000000
ban_gunsNot sure	-0.1158072	0.0213476	-5.424828	0.0000001
news_sources_new_york_timesYes	0.1117582	0.0130339	8.574397	0.0000000
abortion_any_timeDisagree	-0.1880305	0.0276388	-6.803128	0.0000000
abortion_any_timeNot Asked	-0.1124363	0.0239076	-4.702949	0.0000026
abortion_any_timeNot Sure	-0.0786284	0.0330671	-2.377839	0.0174436

Model2 table

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1460695	0.0324312	-4.5039756	0.0000068
genderMale	0.0944693	0.0116856	8.0842568	0.0000000
age	0.0027342	0.0003833	7.1332487	0.0000000
employmentnot in labor force	-0.0854133	0.0141422	-6.0396079	0.0000000
employmentunemployed	-0.0362504	0.0162709	-2.2279333	0.0259196
race_ethnicityOther	0.1665331	0.0225199	7.3949188	0.0000000
race_ethnicityWhite	0.2941943	0.0182237	16.1435387	0.0000000
ban_gunsDisagree	0.1283213	0.0147731	8.6861355	0.0000000
ban_gunsNot sure	-0.0672644	0.0204874	-3.2832105	0.0010318
news_sources_new_york_timesYes	-0.0427714	0.0125087	-3.4193316	0.0006317
abortion_any_timeDisagree	0.1664765	0.0265251	6.2761907	0.0000000
abortion_any_timeNot Asked	0.0847478	0.0229442	3.6936455	0.0002229
abortion_any_timeNot Sure	-0.0007435	0.0317347	-0.0234293	0.9813085

Our first model regressed whether voters would cast a vote for Joe Biden on factors such as their opinion of abortion, gun control, and their age. The model suggested that all factors we accounted for were statistically significant in our model, at a 5% significance level, except whether someone is unemployed. One interesting thing our model suggested is that people who said that their main news source was the New York Times were more likely to say they will cast their votes for Joe Biden, all else being equal. The coefficient on the New York Times variable is 0.112 and is statistically significant at the 5% significance level. We found this interesting because it may possibly suggest an effect on the Joe Biden vote from when the New York Times Editorial Board endorsed him for President. However, this would be a step we could possibly explore in the future in order to obtain a more sound and significant interpretation.

Further, our model suggests that those opposed to gun control are less likely to support Biden, with the coefficient on whether some disagrees with banning guns being equal to -0.148 and statistically significant. People who disagreed with legalized abortion at any time are less likely to support Biden as well as are men, with statistically significant coefficients of -0.188 and -0.0831, respectively.

We then ran a logistic regression with the same explanatory variables, but now looking at the probability someone would cast a vote for Donald Trump. All factors were statistically significant at a 5% level except whether someone was unemployed, whether their chief news source was the New York Times, and whether they were unsure or not asked about whether they support abortion anytime. Our model suggests, as we would expect, that older voters are more likely to support Donald Trump, holding all other factors constant, with a coefficient value of 0.00273. Also, those who oppose gun control or oppose abortion at any point, are more likely to support Donald Trump, with coefficients of 0.128 and 0.166, respectively.

biden_predict	trump_predict	biden_electoralVotes
0.4226211	0.4035821	273

We then conducted post-stratification to discern who our models predict will win the popular vote come Election Day. Our post-stratification analysis suggests that

$$\hat{y_B}^{PS} = 0.423$$

which suggests that our logistic regression model, based off factors such as race, gender, and opinion on hot button political issues, predicts that Joe Biden will receive 42.3% of the votes cast on Election Day. Moreover, our post-stratification suggests further that

$$\hat{y_T}^{PS} = 0.404$$

meaning that our model also predicts that Donald Trump will receive 40.4% of the popular vote.

From our additional analysis, we also found out that the states in which Biden is supposed to win the electoral college with maximum number of seats are the following:

	state	biden_seats
5	CA	55
34	NY	29
14	IL	20
22	MI	16
27	NC	15
31	NJ	14

Clearly, these are expected to be very important states as they have the maximum number of seats(except Florida which Biden is not expected to win in). Our Electoral college prediction results show that Joe Biden is expected to win 273 out of the total 538 seats in the electoral college. Since a majority of 270 seats is required to elect the President, according to our analysis, Joe Biden is expected to win this election with a small margin.

Discussion

Summary

To predict the overall popular vote of the 2020 American federal election, two multivariable logistic regression models were fit to predict the probability of voting for Biden and the probability of voting for Trump. Then, post-stratification techniques were used on both models to calculate the proportion of voters for Biden and the proportion of voters for Trump, based on demographics and survey data.

The models were fit using politically divisive and relevant variables such as New York Times as a news source, gun control opinions, abortion opinions, as well as basic demographic variables such as gender, age, and employment status. Gender, age, employment, and race were further used in the post-stratification step as politically relevant variables that were available in the census data. The choices for these variables were based on past elections and factors known to influence political beliefs based on current events. The use of a large, relatively representative dataset allows the predictions made from this analysis to be more accurate than with small-scale surveys and data.

Conclusion

The use of multilevel logistic regression models and post-stratification (MRP) methods allows for a strong prediction of the actual election results. This technique has been used in previous studies predicting election results, such as in studies predicting the 2016 election results— often more accurately than the polling data alone [5,6]. This report as well as the use of the MRP technique is useful not only in providing predictions that account for a large amount of influencing variables, but also for identifying and understanding the actual presidential election outcome once it is determined[5]. The results of this analysis showed 0.423 as the estimated proportion of voters for Biden and 0.404 as the estimated proportion of voters predicted for Trump. Based on these proportions, we predict that Biden will win the election.

Weaknesses

There are certain limitations in this analysis that should be considered. Firstly, the predictions made in this report are based on survey data and people's opinions. Although the data is recent from June, the months prior to election day are full of press statements and events that provide new information regarding election candidates[7]. This makes it highly plausible that the populations' opinions may have changed since the data was collected. Therefore, due to the versatile nature of opinions and their role as the basis of our predictions, the survey data may not be accurately representative of the population currently. Additionally, the census data is from 2018 which fails to account for new voters from the past 2 years and their information. This is a minor weakness since there has been no significant change in population since 2018, with immigration to the US even decreasing and no major change in government since the 2016 election[8](Tavernise, 2019). However this could potentially impact the predictions, as recent years may have had changes in the numbers of eligible voters and eligible immigrant voters[9] (Shoichet, 2020). A major limitation impacting the results of this analysis is sampling bias in the survey data. The survey data collected could have systemic errors, overestimating or underestimating a party's support across the country. Those errors are then funnelled through. Due to the large amount of data available, a solid model was made using specific significant variables and predictors. However, these choices despite their justifications may be a limitation of this analysis. The individual model in this analysis relies on a reasonable choice of demographic variables. Different models will produce different estimates — with different errors, based on these choices. Similarly, a substantive set of constituency-level predictors is needed. Picking poorly or even forgoing the use of a constituency-level predictor may increase error. The statistics for each constituency may be out-of-date: the census was last conducted eight years ago — in 2011. Recent estimates of each constituency should be available from the Office for National Statistics, but may be limited in terms of demographic dimensions[10](Masters, 2019). Another slight weakness of the analysis is that the model may not have been fully representative of the entire population, since the survey data only had three categories for gender (male, female, and N/A), which may have neglected to accommodate all genders. Since gender was an important variable in this analysis as a predictor, the prediction could be impacted by a lack of representation for other genders in the survey data.

Next Steps

Moving forward, we can enhance our analysis in various ways. Availability of some non demographic variables in census data such as opinions on abortion, gun laws, new green deal and other factors that the two candidates' ideologies greatly differ on would make our predictions significantly better. Having more recent data about the public's opinions on chief issues like the COVID-19 pandemic and the government's response to it will also be extremely helpful. It might also be interesting to investigate time trends and state level opinions on social topics. Furthermore, the model can be improved by using another approach such as a Bayesian model (as opposed to our current Frequentist logistic regression model) or a Multilevel Regression model (that partitions the data into multiple levels of demographic cells).

As the scale for this analysis increases, there is a lot that can be done for improvement and we hope to apply these in our future research.

References

1. Kingsbury, Kathleen. 2020. “Editor’s Note: Why the Times Editorial Board Endorsed Joe Biden for President.” New York Times, October 6, 2020. https://www.nytimes.com/2020/10/06/opinion/joe-biden-endorsement-editors-note.html

2. Silverstein, Jason. 2020. “There were more mass shootings than days in 2019” CBS News, January 2, 2020.
https://www.cbsnews.com/news/mass-shootings-2019-more-than-days-365/
3. Biden, Joe. 2020. “The Biden Plan to end our Gun Violence Epidemic” https://joebiden.com/gunsafety/

4. Pearce, Matt, 2020. “Trump and Biden on guns: Far apart on Policy and Perspective” Los Angeles Times, August 19, 2020. https://www.latimes.com/politics/story/2020-08-19/trump-biden-gun-policy

5. Langer, Gary. 2017. “The polls didn’t predict Trump’s win in 2016 but this technique did” Washington Post, June 13, 2017. https://www.washingtonpost.com/news/monkey-cage/wp/2017/06/13/this-new-polling-method-predicted-trumps-win-while-we-were-testing-it/

6. Jonge, Chad & Langer, Gary & Sinozich, Sofi. (2018). Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Poststratification (MRP). Public Opinion Quarterly. 82. 419-446. 10.1093/poq/nfy023. https://www.researchgate.net/publication/329759450_Predicting_State_Presidential_Election_Results_Using_National_Tracking_Polls_and_Multilevel_Regression_with_Poststratification_MRP

7. https://aceproject.org/ace-en/topics/me/onePage

8. Tavernise, Sabrina. 2019. “Immigrant Population Growth in the U.S. slows to a Trickle”
The New York Times, September 26, 2019. https://www.nytimes.com/2019/09/26/us/census-immigration.html

9. Schoichet, Catherine E. 2020. “1 in 10 eligible voters in 2020 are immigrants. That’s a record high” Cable News Network, February 26, 2020. https://www.cnn.com/2020/02/26/politics/2020-presidential-election-immigrant-voters/index.html

10. Masters, Anthony B. 2019. “MRP Estimates and the 2019 General Election” https://anthonybmasters.medium.com/mrp-estimates-and-the-2019-general-election-9ac1794120d6

11. Tausanovitch,Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814).

12. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

13. Park, David K., Andrew Gelman, and Joseph Bafumi. “Bayesian multilevel estimation with poststratification: State-level estimates from national polls.” Political Analysis (2004): 375-385.

14. “The Electoral College.” National Archives and Records Administration. National Archives and Records Administration. Accessed November 2, 2020. https://www.archives.gov/electoral-college.