Predicting the November Surprise Leaves Questions Unanswered, a statistical analysis on the 2020 US Presidential election popular vote

Adel Boufama

November 2, 2020

Predicting the November Surprise Leaves Questions Unanswered

A Statistical analysis on the 2020 US Presidential Election popular vote

Adel Boufama

November 2, 2020

Model

I will be predicting the popular vote of the 2020 American federal election (Dassonneville, R., & Tien,2020)[9]. I will use survey data from the Democracy Fund + UCLA Nationscape and census data from the American Community Surveys (ACS)[4]. To do this I am going to employ a post-stratification technique using Logistical Regression models. I will make two models, one to predict the portion of US citizens that will vote for Donald Trump and another to capture the predicted portion of citizens that will vote for Joe Biden. The reason why I decided to use two models was because there are undecided voters and US citizens that are deciding not to vote which renders a model for one candidate by itself to be rather uninformative. In the following sub-sections I will further describe the model specifics, the post-stratification calculation and the justification for their usage.

Model Specifics

The models used for this analysis are two Logistical Regression models based on the survey data. This type of statistical model uses a logarithmic based function to model the relationship between a binary response variable and one or more explanatory variables. A Logistical Regression is great for scenarios where one might need to predict a binary outcome given several factors that can be continuous or categorical. In my case, the binary outcome for my first model is whether or not a US citizen will vote for Donald Trump in the 2020 election, and for my second model, whether or not a US citizen will vote for Joe Biden in the 2020 election. Not every citizen on the survey data will participate in voting due to various reasons which makes a 3rd category of voters that cannot be considered as supporting either candidate. As a result, the binary outcome cannot be interpreted as: Not voting for Biden means voting for Trump or vice versa. This is the reasoning behind using two models, one for each candidate, where each response variable could have a binary outcome of: 1) the US citizen voting for the indicated candidate. OR 2) the US citizen voting for the opposing candidate or not participating in the vote for what ever reason.

By using these two models in a Post-Stratification, I am hoping to find the predicted portion of US citizens who will vote for Donald Trump, the predicted portion of US citizens who will vote for Joe Biden and the predicted portion of US citizens who will not be participating in the vote. This will also give us an estimated voter turnout percentage as well by summing up all the predicted portion of US citizens participating in voting in the 2020 Federal Election.

The explanatory variables for my 2 models are: age groups, sex (male, female), race, household income level, and USA region groups. All these variables for the data will be put into the Logistic Regression models to create a formula that predicts probability of voting for the specified candidate given the characteristics of the US citizen by these 5 explanatory variables.

I chose age because it is one of the characteristics of a person that has an effect on their politics. An example is the common stereotype of older people being more conservative than younger folk. I chose to separate age into 5 groups: 18 to 24 years (just old enough to vote/mostly students), 25 to 34 years (early career/starting family), 35 to 49 years (mid career/settled down), 50 to 64 years (late career/empty nest), and 65 years and older (senior citizens/retirees). I believe these are good categories for ages because they most accurately represent the different stages in a person's life which can influence voting habits due to life experiences and circumstances. This categorization is in order to ensure that there is enough people for every single combination of categories. It is required for the Post-Stratification technique used which will be discussed below where I will give a justification for creating age groups.

I chose sex as an explanatory variable because it is also a characteristic of a person that influences politics as a result of different life experiences people have whether they are male or female. Because this is dealing with biological sex, and not "gender", there is no data on gender non-conforming individuals.

Race has been in center stage during political discourse as of recently, this is especially evident with the rising popularity of movements like Black Lives Matter. That is why I also decided to put "Race" as another explanatory variable as this characteristic can have a profound impact on a person's politics, similarly to sex because people of every specific race have their own unique experiences compared to people of another race. I decided to separate race into the categories: white, black, east Asian, other Asian, native American, Hispanic and other. I classified anyone who considered themselves as Hispanic as 'hispanic', even if they listed another race, this is because I believe that it is the self classification that more strongly determines political views compared to the actual ethnicity. I believe these categories are a good representation of the different racial groups of America.

Another explanatory variable I chose was household income level which is separated into 5 sections: less than 25k, 25k to 50k, 50k to 100k, 100k to 200k and above 200k. This one was an easy decision because wealth greatly affects a person's lifestyle and life experiences. It is generally well known that richer people are generally more conservative in their politics compared to their less affluent counterparts.

The final explanatory variable I decided to choose was the general US region which is separated into: Northeast, South, Midwest and West. There are differences in the politics of people in different regions in the US because of different lifestyles. I decided to go with regions instead of States because there are 50 states which is a lot of categories, this may cause issues with the Post Stratification technique as discussed below. Another important thing to note is that I wanted to have a categorical population density variable instead of regions but the surveys did not include that, only the census had that variable. The significance of a population density variable will be discussed in greater depth in the Discussion section.

Mathematically, the Logistic Regression model I am using is represented as:

Figure 1: Logistic Regression Formula

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_{age\_group} + \beta_{sex} + \beta_{race\_groups} + \beta_{region\_groups} + \beta_{hhincome\_bracket}$$

Notation Summary:

- p is the probability of the event of interest occurring. In my case, this is the probability of a US citizen voting for Trump for the 2020 US elections, given specific characteristics a person has which were the explanatory variables (age group, sex, racial group, US region groups, and household income bracket).
- Beta 0 (B0) is the intercept of the y axis which represents the log odds of a US citizen voting for Trump, given they fit in all the categories of the respective dummy (or reference) variables. The dummy values for the explanatory variables are. age_group = 18-24, sex = female, race_groups = black, region_groups = Midwest, and hhincome_bracket = 100k-200k. The intercept of this model is -2.964

so that means the log odds of voting for Trump for an 18 to 24 year old black woman in the Midwest with a household income of between 100k and 200k is -2.964 according to the data.

- The rest of the beta coefficients represent the change in log odds for each category of each variable one wants to represent in the model.
- The predictor variables can be numerical or categorical but for my model, they are all categorical.

As we can see based on the mathematical formula, logistical regression does not fit a line on the data plot like linear regression does, instead it fits an 'S' shaped natural logarithmic based function to the data. This model fits the curve using maximum likelihood estimation.

Logistic regressions are great for probabilistic interpretation, is less prone to over-fitting when dealing with fewer parameters and can be updated easily. A weaknesses however, is that it tends to perform poorly when there dealing with a higher number of parameters, which is why I decided for 5 as a good number.

Post-Stratification

I chose a Logistic Regression in order to categorize people to a certain degree into 'boxes' or 'cells', which are groups of people for a Post-Stratification. The following is the formula for a Post-Stratification.

Figure 2: Post-Stratification formula

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$$

A Post-Stratification works by first constructing strata/boxes or cells to put every single person in. These cells are a combination of all categories in the explanatory variables discussed above, for example, one cell would be, every single US citizen who is an race = east Asian, sex = Male, age_group = 25 to 34, living in the region = South and having a household income = between 100k and 200k. Every 'box' will contain the number of individuals in the census that fit in each unique combination of explanatory variable categories that I have detailed above. A cell is represented as an Nj in the formula. Given my explanatory variables, there are 5 age groups, 2 sexes, 5 income brackets, 6 racial groups and 4 regions so there will be a total of 5 x 2 x 5 x 6 x 4 = 1200 cells that will fit the data of around 1 million people in the census, my cells are mathematically represented as: N1, N2,…N1200.

The Yj estimate in Figure 2 is the probability of voting for Trump or Biden (depending on which model is put in the Post-Stratification), given the person fits all the categories of the j'th cell. As we discussed above, every combination of categories of a US citizen will give us an estimated probability that said person will vote for the candidate specified in the model. The denominator which is just the sum of Nj's is simply the entire population because we are summing up the number of people in each cell and each condition of every cell is mutually exclusive so summing Nj's will give us the total population. Finally the estimator y^PS gives us the estimated proportion of the US population that will vote for Trump or Biden (which ever model we put in here).

For a Post-Stratification to work well, generally the more unique cells you can produce, the better. There is a limit though, you should only produce as many cells as there will be a decent sample size of people in each unique cell. It is not optimal to have most cells containing only one person that satisfies those demographic categories but neither is it optimal to have too many. My decision to have 1200 cells would mean that on average, each cell for my data will contain 1000000/1200 = 800 people which seems like a lot. I found that even then, the variability of the number of people in a cell is very large to the point that there are several cells with only 1 person and a few that have over 5000. If I decided to make age into a continuous variable and/or split people up by the 50 states, there would be too many cells containing 1 person and perhaps some cells might not contain any people at all.

I used the census data for the Post-Stratification. How I did the Post-Stratification for each of my two models was to first organize the census data into cells (every row is a cell) and subsequently add another column 'n' containing the number of people in each cell (represented as Nj in Figure 2). Then I used the Logistical

Regression models to create respective probabilities (represented as the estimator Yj in Figure 2) for every single demographic characteristic of every cell. Lastly I run the model to let it compute the result using the Post-Stratification formula as shown in Figure 2 and give me the estimated proportion of US citizens that will vote for the specified candidate in the given model (represented by the estimator Yˆ PS in the model).

Results

After modelling the data, we find some interesting demographic trends for each different type of voter. All these explanatory variable demographics seem to be statistically significant with determining whether or not a US citizen votes for Trump but do not do so well to explain Biden voting trends. Only Sex and Racial groups are statistically significant in predicting whether or not a US citizen will vote for Biden. For Age groups, there seems to be a statistically significant positive relationship to the probability of voting Trump, which means with every increasing age group, there is an increase in likelihood for voting for Trump. On the other hand, age does not seem to be a statistically significant predictor for a Biden voter. For Sex, there seems to be a clear trend with more males favouring Trump and not Biden and more Females favouring Biden and not Trump. Household income level trends also show something interesting. Just like age, the higher the household income of a US citizen, the more likely for that US citizen to vote for Trump. Biden voters on the other hand, are not easily modelled using household income. Surprisingly, is the middle income individuals who seem to like Biden the most, the most wealthy and least wealthy individuals both tend to not want to vote for Biden, only the highest income bracket is highly statistically significant though. With US region groups, the only statistically significant data was with the US citizens living in the Southern states, they were more likely to vote for Trump and less likely to vote for Biden. Lastly, when it comes to Racial groups all of them were highly statistically significant for both models. whites are the most likely group to vote for Trump, followed by native Americans, other, Hispanics, other Asians, east Asians and the least likely racial group to vote for Trump are black Americans. For Biden, that trend is mostly reversed with black Americans being most likely to vote Biden, followed by east Asians, other Asians, Hispanics, other, white and the least likely racial group to vote for Biden being native Americans which is an interesting finding.

Post stratifying using these two models, we get the estimated portion of US citizens who will vote for Trump given by:

$$y^{\hat{PS}}{}_{Trump} = 0.421$$

This means an estimated 42.1% of US citizens will vote for Trump in the 2020 Federal Elections.

And the estimated portion of US citizens who will vote for Biden is given by:

$$y^{\hat{PS}}{}_{Biden} = 0.420$$

This means an estimated 42% of US citizens will vote for Biden in the 2020 Federal Elections.

Discussion

The goal of this report is to predict the overall popular vote of the 2020 US Federal Elections. The data used for this analysis were the Democracy Fund + UCLA Nationscape (for the survey data) from 2020 and American Community Surveys (ACS) from 2018 (for the census data). Two Logistic Regression models were created using the survey data to find demographic voter trends, one model for the Trump vote, another for the Biden vote. The census data was then organized into cells, each of which contained the number of people in the census that fit every unique demographic group created by the categories with the survey data. A Post-Stratification was then conducted, combining the organized survey and census data as discussed above.

Based on this analysis, we can see that the popular vote is estimated to go to Donald Trump by a very narrow margin of 0.1% (Trump's 42.1% to Biden's 42%). With such narrow margins, this analysis shows that it is quite a close race between both candidates as far as the popular vote goes. Because of the fact that voting for Trump and voting for Biden are mutually exclusive events, this information shows us that 100% - (42.1% + 42%) = 15.9% of people are not participating in the voting according to the survey. This could mean that the popular vote could swing in any direction if a percentage of the non-voters do decide to participate right before the election. We need to keep in mind though, that this estimate hints at a record voter turnout at 84.1% of eligible voters going to vote, for the last 40 years, the voter turnout ranged from 51.7% in 1996 to 61.6% in 2008 [10] which are both significantly lower turnouts than the 84.1% estimated for the 2020 election. As a result of a record voter turnout, I would predict that the likelihood of a significant portion of the 15.9% of non-voters going out to vote before the election is unlikely. That being said, a narrow 0.1% lead for Trump can easily be narrowed down or even swing in favour of Biden. As described by the title, this 'November surprise', leaves a lot of questions unanswered and this analysis does not give us any strong predictions for the winner of the popular vote.

The popular vote is also not very reliable in predicting the winning presidential candidate. As we all know, the popular vote of the 2016 election went to Hillary Clinton but Donald Trump still won the presidency because he won the Electoral College which is what ultimately counts. Predicting the Electoral College results for the 2020 election should be the next step to help give us a clearer picture for what is to come, come election day.

Weaknesses

Most of the weaknesses of this analysis come from statistically insignificant voter demographics, especially in capturing the demographic of Biden voters. As we have seen in the Results section, most demographic categories other than race and sex were statistically insignificant in predicting whether or not a US citizen will cast a vote for Biden or not. This means that the estimated 42% of eligible voters that will vote Biden is not very accurate which is a big weakness to the analysis.

While looking for statistically significant voter demographics, I found population density levels in the census data. It would have been another good characteristic to help pinpoint the type of people that will tend to vote for Trump or Biden. The reason why I did not us it is because the survey data (Democracy Fund + UCLA Nationscape) did not contain any population density parameter for the people surveyed which is considered a weakness in the data. People who live in places with a higher population density tend to lean more liberal in their politics compared to people who live in places with a lower population density (Goldin, 2017 & Wilkinson, 2019). This trend is seen even in people of a similar racial group or income levels and it explains why big cities and states with major metropolitan areas such as New York and California almost always vote for a Democrat candidate in a Federal election while less populated states tend to vote Republican. There are various reasons for this density caused voting bias which is detailed in some of my sources below. I believe that factoring density into my models if it was given in the survey data, would have greatly improved the accuracy of this analysis.

Lastly, some more weaknesses come from the census data. There might have been a slight demographic change from the 2018 census data to current year 2020. The effect of this is probably not too great but there is still some inaccuracies that can arise from the older data when extrapolated to the current year caused by new voters coming of age, the passing of some older people and immigration.

Next Steps

As we hear so often, more research is needed. This phrase also resonates after analyzing the results of my study on popular vote trends in the 2020 US Federal election. The next step would be to analyze the demographics of the non-voters in the survey dataset to find parallels and maybe predict which candidate the non-voters might vote for if they ultimately decide to vote before election day. This can help predict who will win the popular vote better than my analysis alone which has both candidates neck and neck in this race.

Another analysis should also be performed with parameters that can predict both Biden and Trump voters. As discussed earlier, my models were able to capture the Trump voter very well but perform rather poorly when trying to capture the Biden voter. The next step for this would be to find survey data or conduct a follow-up survey that contain more parameters (such as population density as discussed above) which can be used to capture the demographics characteristics and differences of both the Trump and Biden voter.

The Electoral College results are what ultimately decides the future president so another analysis should also be conducted on the predicted Electoral College results for the 2020 Federal Election. This will give us a better prediction to which candidate will win. As mentioned above, Donald Trump won the presidency in 2016 but did not win the popular vote. I am not very clear on how the Electoral College works but I believe there would need to be data on which state district each person in the survey and census resided in.

Typically, election predictions are done using poll data to predict upcoming results. Using poll data would also be considered a next step for further analysis to predict the future president. One thing to note though about polling data is that sometimes they could be inaccurate like most were in the 2016 election. The problem with polling typically arises when there exist many "shy" voters or voters who are afraid of revealing their preferred candidate. This is because polling works by asking a sample of eligible voters which candidate they will vote for. There are new and creative techniques though, to help avoid these mistakes and according to pollsters Arie Kapteyn and Robert Cahaly whom predicted the 2016 election more accurately, the key is to ask a "social-circle" type of question which is: "Who do you think your friends, family and neighbours will vote for?"(Stanton, 2020). This seems to solve the problem of the "shy" voters because most people have a cognitive bias that causes them to overestimate the degree to which their opinions and political views are shared by others and will assume other's are most likely to vote for the candidate that they themselves prefer [2]. This all happens without the person asked being afraid of expressing their own opinions because to them, they are not expressing what they themselves think but instead are expressing what they think others' around them are thinking. Ultimately, predictions are just that and what remains to be seen will remain to be seen. Let's look forward to our November surprise!

References

1. Dassonneville, R., & Tien, C. (2020). Introduction to Forecasting the 2020 US Elections. PS: Political Science & Politics, 1-5. doi:10.1017/S104909652000147X
2. Effectiviology. (n.d.). Retrieved November 03, 2020, from https://effectiviology.com/false-consensus/
3. Goldin, S. (2017, January 08). Why does density predict political preference? Retrieved November 2, 2020, from https://latentparadigm.wordpress.com/2017/01/08/why-does-density-predict-political-preference/
4. Second Nationscape Data Set Release. (2020, October 30). Retrieved November 2, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set
5. Stanton, Z. (2020, November 1). 'People Are Going To Be Shocked': Return of the 'Shy' Trump Voter? Retrieved November 2, 2020, from https://www.politico.com/news/magazine/2020/10/29/2020-polls-trump-biden-prediction-accurate-2016-433619
6. Stegmaier, M., & Norpoth, H. (2017, June 27). Election Forecasting. Retrieved November 2, 2020, from https://www.oxfordbibliographies.com/view/document/obo-9780199756223/obo-9780199756223-0023.xml
7. "Tidyverse Packages." Tidyverse, www.tidyverse.org/packages/.
8. The Government of the United States. Electoral College Fast Facts. Retrieved November 2, 2020, from https://history.house.gov/Institution/Electoral-College/Electoral-College/
9. U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. (n.d.). Retrieved November 2, 2020, from https://usa.ipums.org/usa/index.shtml
10. Voter Turnout in Presidential Elections. (n.d.). Retrieved November 2, 2020, from https://www.presidency.ucsb.edu/statistics/data/voter-turnout-in-presidential-elections

11. Wilkinson, W., VP. (2019, June). The Density Divide: Urbanization, Polarization, and Populist Backlash. Retrieved November 2, 2020, from https://www.niskanencenter.org/wp-content/uploads/2019/09/Wilkinson-Density-Divide-Final.pdf