Biden expected to win 40.6% of popular vote and Electoral College in 2020 U.S. Federal Election

A post-stratification analysis of UCLA Nationscape's Voter Study Group survey

Promiti Datta and Cathy Yang

November 2nd, 2020

Abstract

Poststratification is a method frequently used to aggregate survey-level estimates to the population and correct for non-probability sampling. Our approach uses two logistic regression models with survey data to model a voter's intention to vote for Trump and Biden based on sex, age, education, and race. We then employ post-stratification to estimate the proportion of voters for Trump and Biden at both the national and state level to determine the popular vote and distribution of Electoral College votes for the 2020 U.S. Presidential Election. Our findings forecast a Biden victory, who compared to Trump, is estimated to receive 40.6% vs. 39.7% the popular vote and 305 vs. 233 Electoral College votes.

Code supporting this analysis can be found at: https://github.com/promitid/GSS-2017-Analysis

Model

In this analysis, we are interested in predicting the popular vote and outcome of the 2020 U.S. federal election (Bacon, 2020). To do so, we are creating logistic regression models to model the probability of voting for candidates Trump and Biden, using survey data from the UCLA Nationscape Voter Study Group (Tausanovitch and Vavreck, 2020). We then employ a post-stratification technique to estimate the proportion of voters for each candidate at a national scale to determine the popular vote, then at the state scale to determine the distribution of Electoral College votes. All data cleaning, modelling, and calculations are carried out using R statistical language (R Core Team, 2020) and the tidyverse package (Wickham et al., 2019). The model specifics and post-stratification calculations are described in subsequent sections.

Variables and data cleaning

To aggregate the variables into meaningful groups and to ensure that the survey data corresponds to census data, each variable was reclassified into new categories. We chose the factors of sex, age, race, and education level. Age was aggregated into 4 groups: 18-29, 30-44, 45-60, and 60+ and subjects under 18 were filtered out. Race was categorized into 6 categories: White, Black/African American, Chinese or Japanese, Other Asian/Pacific Islander, and Native American/Alaskan Native, and Other. Finally, education level was categorized into 4 groups: less than high school, high school graduate, some college, and college graduate.

Our selection of variables was influenced by studies examining the links between voters' demographic characteristics and voting intention during the 2018 midterm election. One study by The Economist showed that the four most important demographic categories for predicting voter preference were religion, race, sexual orientation, and education (noa, 2018). We would have liked to consider all these factors in our model, but were limited by the census data which did not collect religion or sexual orientation. The other two factors we included in the model were age and sex, which were also ranked highly in the same study.

Model Specifics

We employ two logistic regression models to estimate the probabilities of voting for Trump and for Biden based on individual variables. Another model considered for this analysis was a multilevel logistic model with state as a group effect. However, this model assumes that individuals nested in the same groups display similar characteristics (Sommet and Morselli, 2017), which may be too large of an assumption to make for general demographic characteristics at a state scale. Rather, our aim was to examine voter intention based on individual characteristics, and aggregate to a state-level and national scale based on the electorate demographic composition, solidifying our choice to use a logistic regression model. The proposed logistic model equation is outlined in **Equation 1** below:

$$\begin{aligned} log(\frac{p}{1-p}) &= \beta_0 + \beta_1(age_{30-44}) + \beta_2(age_{45-59}) + \beta_3(age_{60+}) \\ &+ \beta_4(race_{CHN-JP}) + \beta_5(race_{native})\beta_6(race_{other}) + \beta_7(race_{Asian}) + \beta_8(race_{white}) \\ &+ \beta_9(gender_{male}) \\ &+ \beta_{10}(education_{HS\_Grad}) + \beta_{11}(education_{HS\_NoGrad}) + \beta_{12}(education_{College}) \end{aligned}$$

Where p represents the probability of whether or not a person will vote for Donald Trump in one version of the model, and Joe Biden in the other version. In both models, β_0 represents the intercept of the model. The coefficients $\beta_1, \beta_2, ..., \beta_{12}$ are the respective coefficients for each term, which represent the change in log odds of a person's intention to vote for Donald Trump. The binary variables corresponding to $beta_1$ to β_3 represent whether or not the person is within the age ranges 30-44, 45-59, or over 60. The binary variables corresponding to β_4 through β_8 represent race and classify whether the person is of the groups Chinese or Japanese, Native American, Non-Chinese or Japanese Asians, White, and Other races. The binary variables corresponding to β_{10} through β_{12} represent education level, in the 3 categories: Less than High School, High School Graduate, and Some College.

Post-Stratification

Poststratification is a method frequently used to aggregate survey-level estimates to the population and correct for non-probability sampling (Little, 1993; Wang et al., 2015). With poststratification, estimation of the response variable arises from cell-level estimates that are partitioned into different combinations of variables, which are then aggregated up to a population level by weighting each cell by its relative proportion in the population (Wang et al., 2015).

In our analysis, we use post-stratification to estimate the proportion of voters that will vote for Donald Trump and Joe Biden. We draw data from the 2018 American Community Survey (Ruggles et al., 2020) to partition the electorate population into all possible combinations of sex, age, race, and education level. With the response variable, $\hat{y}^{PS}$ representing the national proportion of voters for each candidate, the post-stratification estimate is defined by

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j}$$

where $\hat{y}_j$ is the estimated proportion of votes for either Trump or Biden within cell j, and N_j is the size of the j^{th} cell in the population.

Subpopulation estimation

While estimating the popular vote is useful, it may not be indicative of the outcome of the election as the allocation of Electoral College votes ultimately determines who becomes president (Robertson et al., 2020).

Under this system, the candidate who wins popular vote in each state receives the entirety of their Electoral College seats (with the exception of Maine and Nebraska) (Archives, 2019). Among a total of 538 electoral college votes, a majority of 270 votes are needed to win (cun, 2016). As such, we are interested in calculating the proportion of voters for Trump and Biden within each state, defined as

$$\hat{y}_s^{PS} = \frac{\sum_{j \epsilon J_s} N_j \hat{y}_j}{\sum_{j \epsilon J_s} N_j}$$

where $\hat{y}_s^{PS}$ is the estimate of the proportion of voters for a given candidate in state s, and J_s is the set of all cells that comprise state s. We then assign a Trump win or a Biden win to each state based on which candidate has a higher $\hat{y}_s^{PS}$. To determine the outcome of the election, we sum for each candidate the Electoral College votes from the states where they have a higher proportion of votes.

Results

Model results

Figure 1 compares the coefficient estimates for the logistic models for Trump and Biden voter intent. Values of these estimates can be found in Appendix A. As indicated, factors that positively impact on Trump voter intent include male sex, age groups 30-44, 45-59, 60+, and White, Asian and Pacific Islander, and Native American racial groups. Meanwhile, significant factors that negatively impact Biden voter intent include male sex, high school as the highest education level, and White, Asian and Pacific Islander, and Native American racial groups.

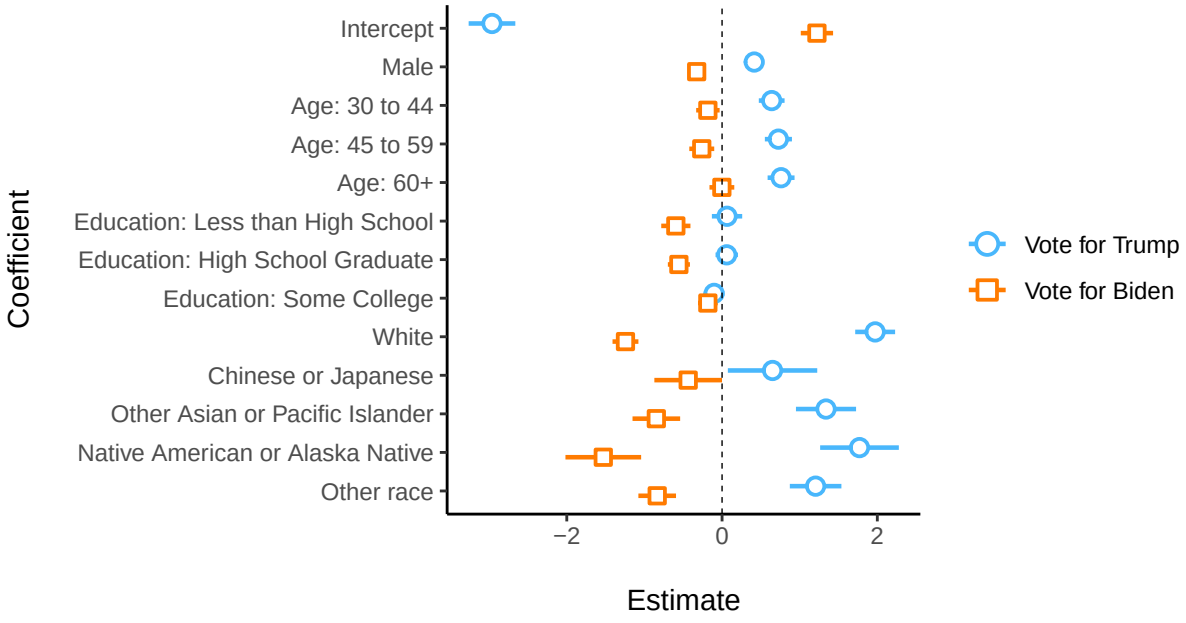

Figure 1: Coefficient estimates for logistic regression equations modelling the probability of voting for Trump and Biden, based on sex, age, race, and education level.

Post-stratification estimates

We estimate the national proportion of voters in favour of voting for Donald Trump and Joe Biden to be 0.397 and 0.406, respectively (Table 1). This is based on our post-stratification analysis of the proportion of voters in favour of each candidate modelled by logistic regression models accounting for sex, age, race, and education level.

Table 1: Estimates of national proportion of voters for 2020 U.S. Presidential Election candidates

Candidate	**Proportion**
Trump	0.397
Biden	0.406

Our post-stratification calculations by state estimates that Biden will receive 305 Electoral College votes, and Trump 233 (Table 2). The breakdown of electoral college seats allocated to each candidate are described in Table 3.

Table 2: Estimated total Electoral College votes per candidate in the 2020 U.S. Presidential Election

Candidate	**Electoral College votes**
Trump	233
Biden	305

Discussion

Our approach used two logistic regression models from voter survey data to model individual intention to vote for Trump and Biden based on sex, age, education, and race. We then employ post-stratification to the 2018 ACS census data to estimate the proportion of voters for Trump and Biden at both the national and state level. Our findings are able to forecast the popular vote and allocation of Electoral College votes for the 2020 U.S. Presidential Election. We found a narrow margin between Biden and Trump at the national scale, with Biden holding a 0.9% lead for the popular vote. However, when estimating the share of votes at a state level, we gain a clearer picture of the preferences for each candidate and are able to forecast the allocation of Electoral College seats, and thus the outcome of the election.

Weaknesses and next steps

While doing our analysis, we found various sources of weakness in the data, the type of model we chose, and other factors we did not consider. Within the data we excluded undecided voters, which may have created bias towards either candidate, and since the data is from June, voters' opinions may have changed since then. We discussed in the "Model Specifics" section why we chose a logistic model over a multilevel model, but a multilevel model by state would have helped to correct cells with smaller samples (Wang et al. 2015), since there were some states with fairly low sample sizes in the survey data. Finally, there were many factors not included in our analysis. We did not consider factors that may affect voter turnout, such as voting amid covid-19 and voter suppression. Voter turnout may skew the vote towards either candidate, and the effect will only be revealed after the election. There are also some states such as Maine and Nebraska who split their electoral college votes by congressional district, which we did not factor into our individual state analysis, but it would not have made a significant difference in our final predictions (Silver, N. 2018).

There are many possible next steps to this analysis, especially since the results of the 2020 election will soon be revealed. One way to get insight on possible inaccuracies on our model would be to do follow up surveys to see whether the respondents voted, and if they voted for the candidate they reported in June. We could use this data to see if any factors turned out to be poor predictors, or if any factors were more associated with not voting. We can also ask questions about the respondent's decision to vote, to get a better idea about how much current events impacted voter turnout, which can be useful to consider in future models.

Appendix

A: Coefficient estimates of logistic models for Trump and Biden voter intention

B: Distribution of sex, age, race groups, and education levels in the Voter Study Group Survey

Here is the distribution of demographic characteristics in our survey.

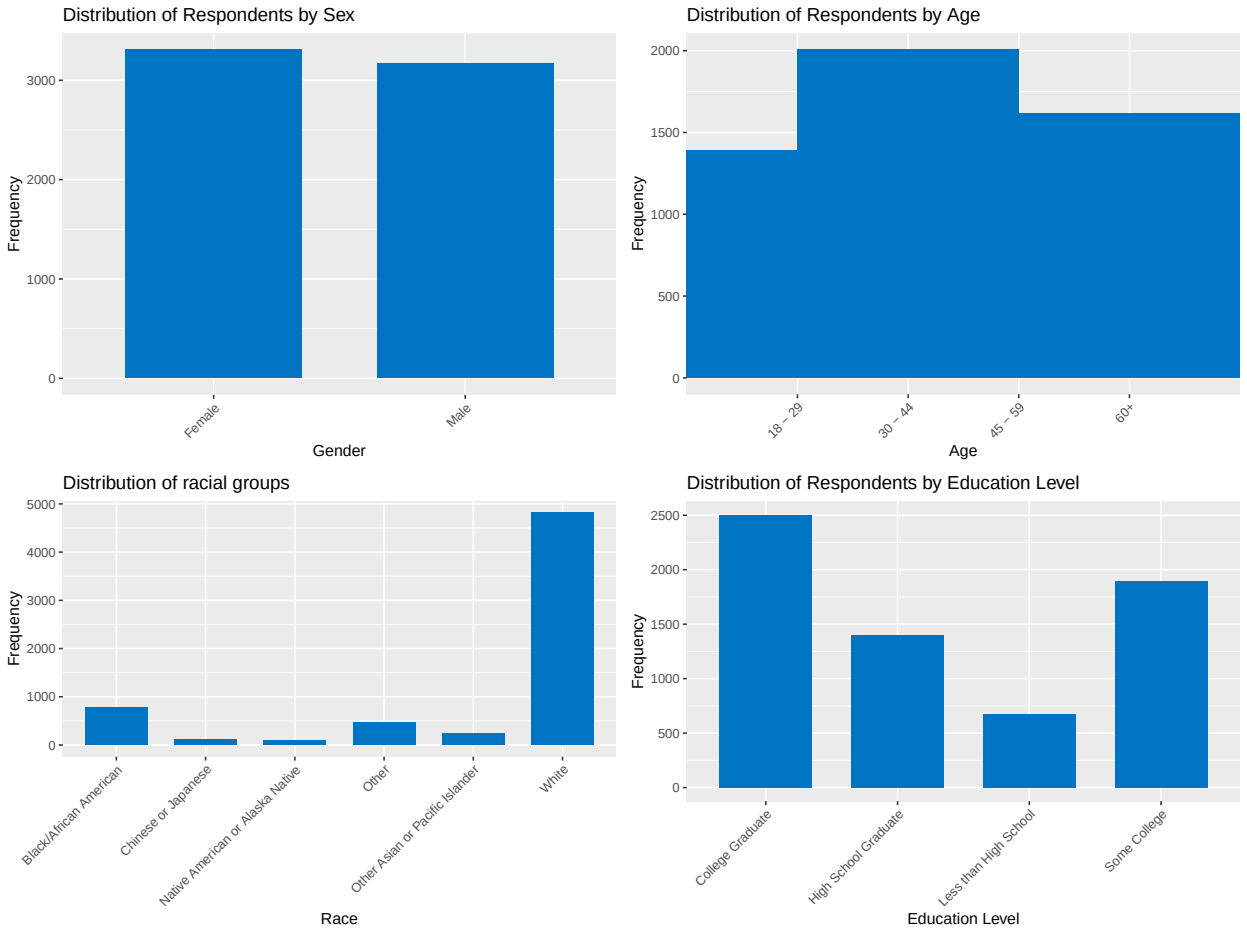

Figure 2: Distribution of sex, age, race groups, and education levels in the Voter Study Group Survey.

References

(2016). United States Electoral College Votes by State.

(2018). How to forecast an American's vote. The Economist.

Archives, N. (2019). What is the Electoral College?

Bacon, P. (2020). The Issues That Divide People Within Each Party.

Little, R. (1993). Post-stratification: A modeler's perspective. Journal of the American Statistical Association, 88(423):1001–1012.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robertson, H., Kirk, A., and Hulley-Jones, F. (2020). Electoral college explained: how Biden faces an uphill battle in the US election. The Guardian.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. (2020). IPUMS USA: Version 10.0 [2018 ACS].

Sommet, N. and Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. International Review of Social Psychology, 30(1):203–218. Number: 1 Publisher: Ubiquity Press.

Tausanovitch, C. and Vavreck, L. (2020). UCLA Democracy Fund Voter Study Group Nationscape Wave 50: June 25th - July 01, 2020.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3):980–991.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43):1686.

Table 3: Estimated proportion of voters and allocation of Electoral College seats by state for the 2020 U.S. Presidential Election

State	Seats	Trump	Biden	Winner	State	Seats	Trump	Biden	Winner
Alabama	9	0.374	0.424	Biden	Montana	3	0.448	0.368	Trump
Alaska	3	0.408	0.358	Trump	Nebraska	5	0.432	0.379	Trump
Arizona	11	0.415	0.386	Trump	Nevada	6	0.389	0.403	Biden
Arkansas	6	0.403	0.394	Trump	New Hampshire	4	0.44	0.381	Trump
California	55	0.373	0.418	Biden	New Jersey	14	0.389	0.421	Biden
Colorado	9	0.424	0.394	Trump	New Mexico	5	0.422	0.368	Trump
Connecticut	7	0.407	0.408	Biden	New York	29	0.377	0.424	Biden
Delaware	3	0.39	0.42	Biden	North Carolina	15	0.378	0.425	Biden
District of Columbia	3	0.284	0.515	Biden	North Dakota	3	0.438	0.368	Trump
Florida	29	0.404	0.409	Biden	Ohio	18	0.413	0.392	Trump
Georgia	16	0.351	0.445	Biden	Oklahoma	7	0.409	0.375	Trump
Hawaii	4	0.312	0.459	Biden	Oregon	7	0.427	0.387	Trump
Idaho	4	0.438	0.37	Trump	Pennsylvania	20	0.423	0.384	Trump
Illinois	20	0.399	0.407	Biden	Rhode Island	4	0.417	0.395	Trump
Indiana	11	0.418	0.382	Trump	South Carolina	9	0.369	0.432	Biden
Iowa	6	0.438	0.37	Trump	South Dakota	3	0.439	0.363	Trump
Kansas	6	0.425	0.384	Trump	Tennessee	11	0.402	0.4	Trump
Kentucky	8	0.425	0.378	Trump	Texas	38	0.395	0.403	Biden
Louisiana	8	0.354	0.438	Biden	Utah	6	0.418	0.383	Trump
Maine	4	0.45	0.37	Trump	Vermont	3	0.448	0.377	Trump
Maryland	10	0.346	0.46	Biden	Virginia	13	0.379	0.429	Biden
Massachusetts	11	0.405	0.409	Biden	Washington	12	0.411	0.398	Trump
Michigan	16	0.414	0.393	Trump	West Virginia	5	0.442	0.363	Trump
Minnesota	10	0.437	0.379	Trump	Wisconsin	10	0.439	0.372	Trump
Mississippi	6	0.334	0.458	Biden	Wyoming	3	0.441	0.369	Trump
Missouri	10	0.418	0.388	Trump					

Table 4: Vote intention coefficient estimates (Standard error), (t-stat, p-value)

	Vote for Trump	Vote for Biden
Intercept	-2.96 ***	1.22 ***
	(0.15), (-19.32, p = 0.00)	(0.11), (11.52, p = 0.00)
Male	0.41 ***	-0.33 ***
	(0.05), (7.54, p = 0.00)	(0.05), (-6.13, p = 0.00)
Age: 30 to 44	0.64 ***	-0.18 *
	(0.08), (7.55, p = 0.00)	(0.08), (-2.37, p = 0.02)
Age: 45 to 59	0.73 ***	-0.26 **
	(0.09), (8.15, p = 0.00)	(0.08), (-3.19, p = 0.00)
Age: 60+	0.76 ***	-0.00
	(0.09), (8.59, p = 0.00)	(0.08), (-0.03, p = 0.98)
Education: Less than High School	0.06	-0.60 ***
	(0.10), (0.64, p = 0.52)	(0.10), (-6.20, p = 0.00)
Education: High School Graduate	0.06	-0.56 ***
	(0.07), (0.82, p = 0.41)	(0.07), (-7.67, p = 0.00)
Education: Some College	-0.10	-0.18 **
	(0.07), (-1.52, p = 0.13)	(0.06), (-2.83, p = 0.00)
White	1.97 ***	-1.24 ***
	(0.13), (15.05, p = 0.00)	(0.09), (-14.63, p = 0.00)
Chinese or Japanese	0.65 *	-0.44 *
	(0.29), (2.21, p = 0.03)	(0.22), (-1.96, p = 0.05)
Other Asian or Pacific Islander	1.34 ***	-0.85 ***
	(0.20), (6.78, p = 0.00)	(0.16), (-5.41, p = 0.00)
Native American or Alaska Native	1.77 ***	-1.53 ***
	(0.26), (6.85, p = 0.00)	(0.25), (-6.15, p = 0.00)
Other race	1.21 ***	-0.84 ***
	(0.17), (7.11, p = 0.00)	(0.12), (-6.76, p = 0.00)
N	6475	6475
AIC	7941.69	8445.77

Standard errors are heteroskedasticity robust. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.