# Donald Trump estimated to win 40% of the popular vote in the 2020 presidential election

David Wong, Iris Shao, Jingjing Zhan, Muhammad Tsany

November 2, 2020

## Model

### Model Specifics

Our analysis is performed in the statistical language R (R Core Team, 2020), using haven (Wickham and Miller, 2019), broom (Robinson and Hayes, 2020), and tidyverse (Wickham et al., 2019) packages. Code and data supporting this analysis is available at: https://github.com/waviddong/STA304-A3-Stats-Don-t-Lie-V3.

We based our weights off of the surveyed data from the survey results collected from Nationscape. From modelling the survey data, we will then load in the post-stratification data which is microdata collected by Integrated Public Use Microdata Series (IPUMS). We will be using our model to apply to the post-stratified data. We chose to model our survey data with logistic regression as our model since our outcome of interest, candidate vote, our dependent variable (Y), is assumed to be binary (Donald Trump = 1 and Joe Biden = 0). We can make this assumption despite the presence of other possible candidates because the majority of votes for a candidate in a given state gets all the electoral delegates, regardless of other popular votes. Thus, since the democratic and republican parties hold strong majorities, we are assuming that other parties do not have enough representation to out-vote the other major parties.

Our model is the following:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{ethnicity} x_{ethnicity\ i} + \beta_{education} x_{education\ i} + \beta_{employment} x_{employment\ i} + \beta_{age} x_{age\ i}$$

where $i = 1, 2, ...n$ *and* $p$ $=$ *probability of a Trump vote*

Our model from the survey data reclassified and grouped some categories of our independent variables in order to achieve a statistically significant p-value. Some sub-categories of the given independent variables had very small sample sizes and thus, had a high p-values meaning no statistically significant correlation; when grouped with other similar variables, we were able to get statistically significant p-values. These same techniques for classification would be implemented in the Census Data accordingly.

Our number of Fisher Scoring iterations was 4, meaning that it took 4 iterations until our data fit the model. Thus, it took 4 steps to reach convergence.

### Post-Stratification

Multilevel regression with post-stratification (MRP) is a common method used to adjust dummy variables to analyze metrics related to opinion and other survey responses. It leverages and mathematically weighs

individual level survey responses (microdata) to various characteristics which is then used to adjust the sample to better represent the population. For our analysis, we followed these steps to post-stratify our data to find our $\hat{y}$ values:

1. Grab survey data
2. Clean survey data (quantify predictor variables)
3. Fit a model (glm() in our case)
4. Grab census data
5. Clean census data

5.1 Group data by our customized 'cells'. count the number of individuals in each cell

5.2 Quantify the same predictor variables in the same way as in 2. For example, if race is divided into 5 categories in survey data, race in census data should also be divided into the SAME 5 categories and the two new columns 'race_code' should have SAME name.i.e either both are 'race_code' or both are 'race_ethnicity_code'. )

6. Fit our census data into our model fitted in 3, get $log(\frac{p}{1-p})$ (logodds estimate) where P is the probability of voting for Trump
7. Calculate P based on the $log(\frac{p}{1-p})$ (logodds estimate) calculated in 6
8. Calculate $\hat{y}$ using the MRP function

We defined the following as our post-stratification cells: race, education, age and employment status

We split race into 5 categories that were generalized by region of ethnicity, if they were more than one, it would be placed into a mixed category. These categories were divided as such: White (1), Black/African American (2), Native American (3), Asian (including Pacific Islanders) (4), mixed races (5).

For education, we categorized the cell into 4 categories based off some assumptions; 1 as 8th grade or less completed, 2 as 9th to 12th grade (high school) completed, 3 as 1 to 4 years of college/university completed (degree or not), and 4 as 5+ years in college completed (degree or not). We are assuming that those completing 4 years most likely will have their degrees and 5+ years would mostly represent those pursuing higher education (MAs, PhDs etc. . . ).

For employment status, we categorized the cell into 3 categories; 1 as employed, 2 as unemployed, 3 as a combination of n/a and not in the labor force. We made this distinction between unemployed (2) and the combo split (3) because the combo split is defined to be people not actively looking. Circumstances for uncertain or non-labor force participants may be housewives, retirement, disabled, in prison etc. . . ; these people aren't unemployed but rather are classified as n/a or not a part of the labor force via economic conventions.

For age, we categorized the cell into groups of 20 starting at 18 (minimum voting age) to 100 (maximum age in the dataset is 97). These ranges are arbitrary and those aged closer to the fringes of the ranges may be misrepresented as they are closer to one division than the other.

# Results

Of 1256116 observations partitioned into 240 cells in our post-stratified census data, we estimate the proportion of voters in favour of voting for Donald Trump to be the following:

Table 1: American voter intention estimates by race, employment status, education level, and age brackets

| | Group | Mean | Lower | Upper |
|---|---|---|---|---|
| **Race** | | | | |
| White | 1 | 0.578 | 0.448 | 0.700 |
| African American | 2 | 0.134 | 0.081 | 0.203 |
| American Indian | 3 | 0.585 | 0.462 | 0.709 |
| Asian | 4 | 0.373 | 0.259 | 0.502 |
| Other races | 5 | 0.363 | 0.252 | 0.493 |
| **Employment Status** | | | | |
| Employed | 1 | 0.447 | 0.117 | 0.710 |
| Unemployed | 2 | 0.385 | 0.094 | 0.654 |
| Not in labor force | 3 | 0.378 | 0.086 | 0.639 |
| **Education Status** | | | | |
| Middle School | 1 | 0.453 | 0.128 | 0.709 |
| High School | 2 | 0.419 | 0.109 | 0.687 |
| University | 3 | 0.354 | 0.082 | 0.616 |
| Post-graduate | 4 | 0.392 | 0.097 | 0.660 |
| **Age Brackets** | | | | |
| Ages 18 - 40 | 1 | 0.357 | 0.083 | 0.612 |
| Ages 41 - 60 | 2 | 0.392 | 0.097 | 0.651 |
| Ages 61 - 80 | 3 | 0.425 | 0.112 | 0.689 |
| Ages 81 - 100 | 4 | 0.451 | 0.128 | 0.707 |

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j} = 0.404$$

This is based on our post-stratification analysis of the proportion of voters in favour of the Republican Party modelled by a logistic regression model, which accounted for four demographic variables: race, employment status, education level and age.

Table 1 lists the voter intent estimates by race, employment status, education level and age. Overall, Whites and American Indians are estimated to have the highest proportions of voters in favour of Trump ($\hat{y} = 0.578$ and $\hat{y} = 0.585$, respectively), while African Americans are estimated to have significantly lower voter intention for Trump ($\hat{y} = 0.134$). Respondents across all other race categories, employment statuses and education levels have similar intent estimates in favour of the Republican Party ($\hat{y}$ ranges from 0.354 to 0.453).

# Discussion

American election forecasts have long been a global interest. In this paper, we predicted the overall popular vote of the 2020 American federal election using a logistic regression model with post-stratification and using datasets from Democracy Fund + UCLA Nationscape and IPUMS, accounting for race, employment status, and education level.

Based on the estimated proportion of voters in favour of voting for the Republican Party being 0.404, we predict that the Democratic Party will win the election. Overall results indicate a deep polarization of voting intention, with most noticeable differences in the area of race. African Americans overwhelmingly intend to support Democratic candidates, which is in line with historical trends (Wallace et al., 2008), while the Whites tend to vote for Republican candidates.

## Weaknesses

There are a number of drawbacks in conducting election forecasts, as well as in our application of the poststratification technique on a logistic model.

Election forecasts may increase interests in elections and voting, but it could also be radically unreliable, especially for the 2020 presidential election due to polarization and the COVID-19 pandemic, as argued by Jennifer Nicoll Victor, Professor of Political Science at George Mason University (2020). Other downsides include affecting turnout—as evident in the 2016 election when Hiliary Clinton won the popular vote but lost the election, outsizing focus on the Election Horse Race, and giving a false impression of science and uncertainty (Victor, 2020).

Continue with Weakness in poststrat and in our model (can be short) In our model, we only chose four predictor variables to predict the value response variable, but there are more factors that can affect the election result in real life. Also, when we grabbed the census data, we reduced the census data size to 1.25 million, which is smaller than the population size.

Despite MRP being a strong statistical methodology to understand and analyze microdata, in this case, our Census data; it has some glaring weaknesses that impacts our analysis. Some of these errors in MRP include misclassification of our cells, missing data, or missing crucial demographic predictors (Kennedy et al., 2020). For the misclassification of cells, we had to make assumptions when choosing our cell splits, otherwise there would be sample size errors; niche ethnicities like "Samoan", while technically a Pacific Islander, would be culturally very different than a Malaysian who is also grouped with Pacific Islander. Despite this, we had to maintain consistency between the Survey Data and Census Data so we made the assumption that they are equivocally the same when they are not. Missing data was also an issue during our modelling since for example, employment for the Census Data had the field, "n/a" as a choice. We classified it the same as not in the labour force and assumed that if the respondents were employed/unemployed, they would respond accordingly; obviously this may not be the case since non-sampling error could have been a factor. Lastly, we could also be missing a crucial part of our demographic predictors, with so much microdata, it is important to be careful about the framework we choose for our analysis; some of our cell splits may not have been optimally or correctly classified and thus, lose explanatory power of our claims. Although our model assumptions are intuitively consistent and reasonable, it would be a stretch to say that this is anywhere near perfect.

## Next Steps

While our prediction from our model happens to be quite reasonable and within range. With many polling predictions claiming that Biden has the edge in the election, we can reaffirm their predictions with our $\hat{Y}^{PS} = 0.404$ (40.4%) chance of Trump winning the 2020 election. However, we must reassess our model to understand how accurate it is compared to the ground truth. We can do this by implementing our model to past elections' Survey Data and Census Data, with the election outcomes already set in stone, we can assess the predictive power of our model. From assessing the external validity, we can then discuss further implementations and modifications to our model to improve its accuracy, relevancy, and effectiveness in explaining election outcomes.

# References


Alexander, R and Caetano, S. (2020). *01-data_cleaning-post-strat1.* University of Toronto

Alexander, R and Caetano, S. (2020). *01-data_cleaning-survey1.* University of Toronto

Caetano, S. (2020). *ProblemSet3 - template-logistic.Rmd.* University of Toronto

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

Robinson, D. and Hayes, A. (2020). *broom: Convert Statistical Analysis Objects into Tidy Tibbles.* R package version 0.7.0.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. and Sobek M. (2020). IPUMS USA: Version 10.0 [20200625]. Minneapolis, MN: IPUMS, 2020. doi: 10.18128/D010.V10.0

Tausanovitch, Chris and Vavreck, Lynn. (2020). Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

Victor, J. N. (2020). Let's Be Honest about Election Forecasting. *PS: Political Science & Politics*, 1-3. doi:10.1017/s1049096520001432

Wallace, D. S., Abduk-Khaliq, A., Czuchry, M., & Sia, T. L. (2008). African American's Political Attitudes, Party Affiliation, and Voting Behavior. *Journal of African American Studies, 13*(2), 139-146. doi: 10.1007/s12111-008-9040-y

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting, 31*(3), 980-991. doi:10.1016/j.ijforecast.2014.06.001

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R.,... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. doi:10.21105/joss.01686

Wickham, H. and Miller, E. (2019). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* R package version 2.1.1.

Zhu, H. (2020). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* R package version 1.3.1.