Predicting Donald Trump's win*

using Post-Stratification by Age Groups and Education Level

Kar Hian Ong and Fadlan Arif

02 November 2020

Abstract

With the 2020 presidential election coming to an end soon, this paper looks into predicting who will win the presidency this November. By gathering data from the 2018 US census and survey data from nation scape, we were able to model and post-stratify our findings by age groups and education. We expect Donald Trump to win 59 percent of the national popular vote with 3 percent margin of error. This has implication on how we look at our future.

We use R Core Team (2020), Wickham et al. (2019), Larmarange (2020), Ushey et al. (2020), Allaire et al. (2020), Zhu (2020)

1 Introduction

With the US federal election 2020 swinging by, there is intense debate on who will be the upcoming US president of the United States. The election uses the electoral college system where each state is given a certain number of votes. Candidates with the 270 votes or more will win the election. Us is govern by 2 major parties, Democrats and Republicans. The candidate representing Democrats is Joe Biden while the candidate representing Republicans is Donald Trump. Both candidates have held positions prior to the 2020 election making them a strong contender for this election. Donald Trump is the current sitting president of the United States while Joe Biden was the vice president of the previous administration. Many polls have indicated that Joe Biden will win the popular vote as well as the overall election. Will the 2020 polls make the same mistake as they did by predicting that Hillary Clinton will win the election in 2016?

This research is to see if we can predict who will win the popular vote for the US election using age groups and education as predictors. The election result has massive impact for the global population. The sitting president will set the tone for international and domestic policies. This information could also be used for future references for looking into predicting later presidencies base on its accuracy.

Within this paper we look at the probability of and individual voting for Joe Biden, stratified by their age group and education. We were able to stratify by cross examining the survey data about election opinions with the US census for 2018. With this information, we were able to more accurately represent the survey data to help it be proportionate to the population of America. Through our research we were able to predict that Donald Trump will win the popular vote for the presidential election 2020. Read on to find out how we came to that conclusion.

The remaining section is structured as follows: Section 2 will talk about retrieval and cleaning of the survey data and American Community Survey (ACS) data, section 3 will introduce the model, section 4 we will get to see the results of our modeling, section 5 will be discussion on the findings and weaknesses.

*Code and data are available at: https://github.com/karhian/2020-election

2 Data

The survey data is provided by Nationscape (Tausanovitch and Vavreck (2020)). Nationscape conducts interview to 50,000 Americans covering the campaign and election. All respondents take the survey online and must complete an attention check before taking the survey. Nationscape handles non-response by using a set of weights. The weights are derived from the adult population of the 2017 American Community Survey of the U.S. Census Bureau. This survey generally asks political questions, policies, and views.

The post-stratification dataset that we will use is the ACS provided by IPUMS USA (Steven Ruggles and Sobek (2020)). Within the IPUMS USA site we chose the 2018 US census. From there we were able to choose some variables of interest. We chose based on what we thought were major indicators of one's political stance. Some of the important variables we picked were age, education level, school type and veteran status. We also picked a wider variety than we needed to be able to have flexibility in shaping our models and graphs. This acted as our population data. The ACS data is conducted once a year and is conducted by the United States Census Bureau. The ACS derived its sampling frame from which they draw the samples from Census Bureau's Master Address File (MAF). The ACS conducts a two-phase sampling technique and a 2-stage sample selection. The first stage of the first phase is by systematically assign new addresses to five existing sub-frames and the second stage is by systematically select sample from first-stage sample (sub-frame). The ASC handles non-responses in the second phase they select sample of unmailable addresses and non-responding addresses and send to Computer Assisted Personal Interviewing (CAPI). The sub-frames are assigned to specific year and are rotated annually. This is an example of stratified sampling where you divide the population into subgroups and do further sampling by using simple random sample or systemic sampling but for the ACS they did systemic sampling. Systemic sampling is when you start at a certain house and then you sample every kth sample after that.

The cleaning of survey data and stratification data uses `haven` (Wickham and Miller (2020)) package to read the data

After collecting both the needed datasets, we cleaned each of them respectively, picking the variables we needed and adding labels. For both datasets, we decided to generalise the ages and education levels. We made sure the categories on each data set match perfectly. This was done by adding a new column named 'age_groups' and 'educ_levels' respectively. The education levels represent the highest academic achievment so far. The reason we grouped the ages is due to their similar voting behavior.

With the new columns created we categorised the responses into it's category as shown in Table 1 and Table 2

Table 1: Age group categories

age groups
Below 18 years old
18-30 years old
31-45 years old
46-65 years old
Above 65 years old

Table 2: Education level categories

level of education
Grade school
College
Master's or above bachelor's degree
Doctorate Degree

For the Nationscape dataset, we only took in the into consideration the ones were explicitly said they were voting for either Donald Trump or Joe Biden. We made this decision as to not choose for those who were undecided. This is so that we can model a binary logistical problem where it's either 0 or 1.If the person planned on voting Joe Biden, it equalled 1, if they planned on Donald Trump, it equalled to 0. WE also remove those that are not eligible to vote as this create noise to the data.

For the ACS data, we recategorize "no schooling completed" as "grade school" due to similar voting patterns. We also removed responses from those below the age of 18 as the voting age is 18.

In order to get a bigger picture of the variables involved, first we plot the data for ages as shown in Figure 1 and Figure 2.

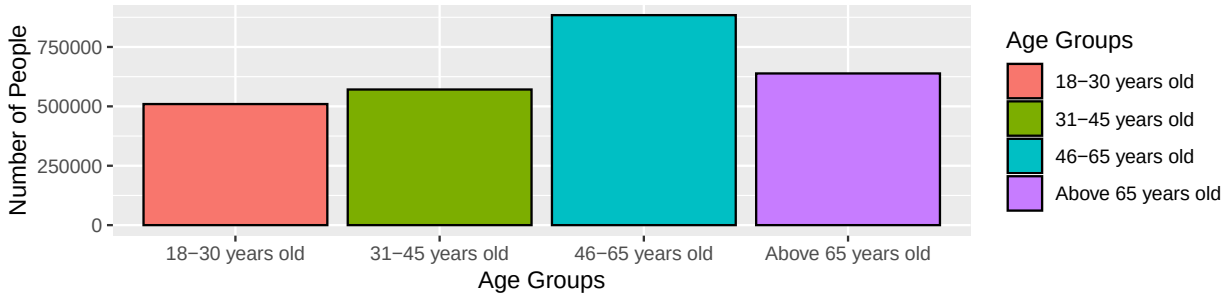

Figure 1: Age Groups for Population Data

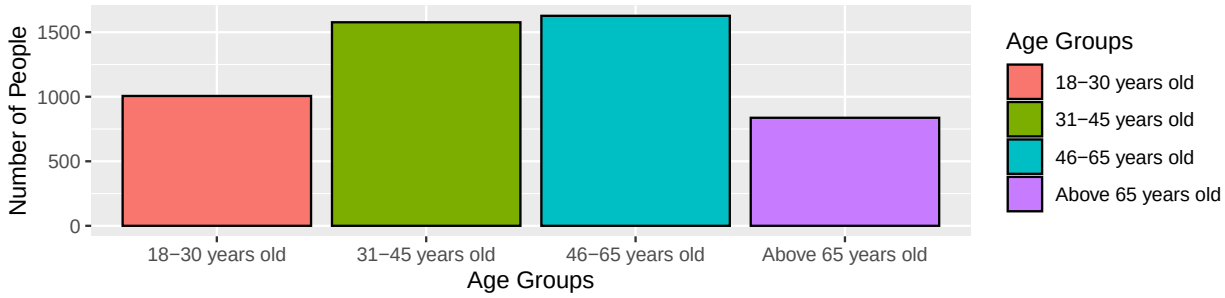

Figure 2: Age Groups for Survery Data

By comparing Figure 1 and Figure 2, we get to see that age groups for 46-65 years old and those above 65 is under represented while those between the ages 31-45 is over represented. This issue can be solved by using multilevel regression and post-stratification.

On the other hand, when we plot the other variable as shown in 3 and Figure 4,

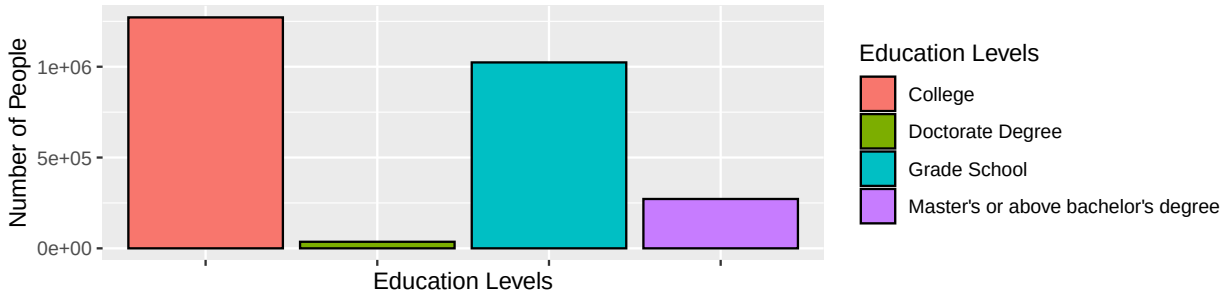

Figure 3: Education level for Population Data

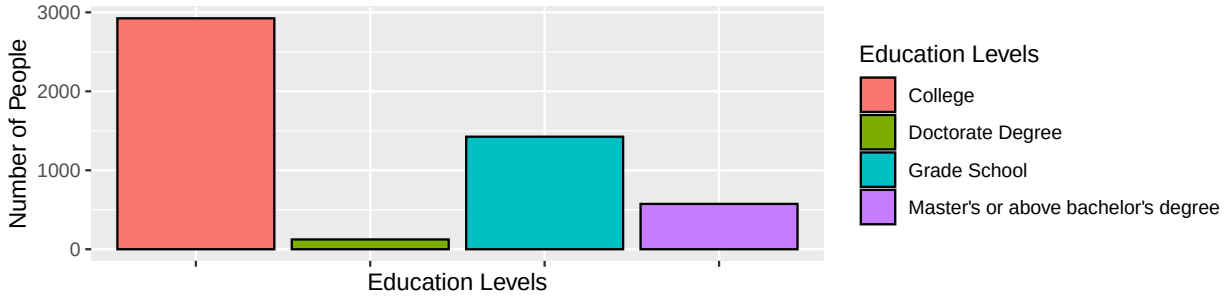

Figure 4: Education level for Survey Data

The plots in Figure 3 and Figure 4 shows a good propotion between the survey data and population data.

In Table 3 we can clearly see that Joe Biden has more votes. But this is just the survey data and does not represent the general US population. This is why we need a model that could generalise to the US population.

Table 3: voting intention 2020

Candidates	Frequency
Donald Trump	2417
Joe Biden	2630

After all the facts and figures, you might be wondering what multilevel regression with post-stratification is. Multilevel regression is a type of model where the parameters vary by more then one level. In our case we want to model the performance of Joe Biden which allow the grouping of age groups and education level. Post-stratification is an act of reweighing the survey data to the population data. When we conduct a survey, it is not possible to sample the entire US population. There are several factors to it. One of the factors is cost. It will cost the surveying company millions of dollars just to conduct a survey each time. So, we use an available census data to weight the prediction.

3 Model

Our approach towards handling our stratified data was through the scope of logistic regression. Logistic regression allows us to look at a binary variable, to see the probability of that event happening or not. We chose logistic regression as most of the votes land on the democratic or republican candidates thus making that our variable of interest. We mainly focus on those planning to vote for these candidates, as these will most likely be the deciding vote in the election. After narrowing down the election to two options, have turned the variable of interest into a binary one, thus allowing us to use logistic regression. The formula we use is:

$$log(\frac{p}{1-p}) = B_0 + B_1x_1 + B_2x_2 + B_2x_3 + B_4y_1 + B_5y_2 + B_6y_3 \tag{1}$$

Where p is the probability that the event of interest occurring (voting for Joe Biden), B_0 is the intercept and $B_i, 1 \leq i \leq 6$, coefficient represents change in log odds for one unit increase in x_j or $y_j, 1 \leq j \leq 3$.

As the age groups and education are categorical, they all have different weights on each category. The equation for this model is

$$log(\frac{p}{1-p}) = 0.828 - 0.711x_1 - 0.831x_2 - 0.800x_3 - 0.512y_1 - 0.290y_2 - 0.185y_3 \tag{2}$$

where the x values represent age groups and y values represent education levels. x_i is 1 when that sample contain a particular age group category while y_i is 1 when there is a particular level of education category. the value is 0 otherwise.

Table 4: switch for equation 2

	Categories
x1	31-45 years old
x2	46-65 years old
x3	Above 65 years old
y1	Doctorate Degree
y2	Grade School
y3	Master's or above bachelor's degree

If the age group is 18-30 years old, x1,x2,x3 is all 0. If the level of education is College, y1,y2,y3 is all 0. those are the baseline for the model.

The software we use to model is by using r's glm function.

4 Results

After applying the model, we visualise the distribution in Figure 5 and Figure 6

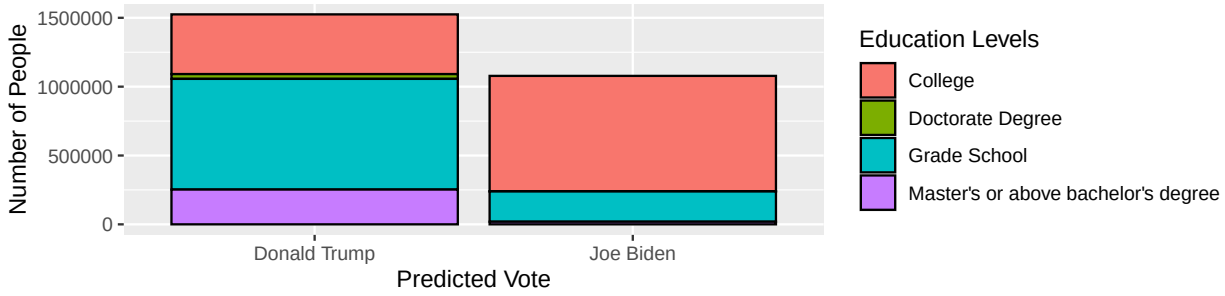

Figure 5: Voting Patterns according to education levels

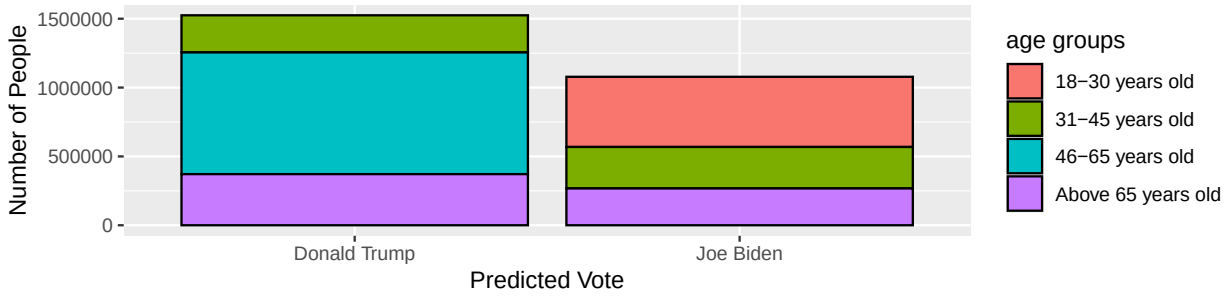

Figure 6: Voting Patterns according to age groups

The model predicted that those between the age of 18-30 will vote for Joe Biden.

Table 5 shows the propotion vote based on the model.

Table 5: voting prediction 2020

Candidates	Propotion
Donald Trump	0.59
Joe Biden	0.41

5 Discussion

5.1 First discussion point

The model uses an 80-20 split where 80% of the survey data is used to train the model and the remaining 20% is used to test the model. The 20% is use as cross validation to test whether our model fits the data well. Our model yield 56% accuracy for the training data and 58% accuracy on testing data. This might be signs of overfitting in the data. This model is not an ideal model as the accuracy is low. From the accuracy, we learnt that education and age groups is not a good deciding factor to predict the election. The margin or error is calculated by

$$me = 1.96 * \sigma/\sqrt{n} = 1.96 * 1.166/\sqrt{5047} = 0.032 \quad (3)$$

5.2 Second discussion point

Some of the other models that we can consider is by weighting the other polls and create a model among other polls from different polling company. The weights will be based on their reliability. 538 does its election prediction using the different weight for different polls. They do not use multilevel regression with post-stratification as it will smooth noisy data which may not accurately model the polls.

5.3 Weaknesses and next steps

Some weaknesses in the paper lie within the accuracy of the population data. It is from 2018, thus not fully representative of the 2020 population. This could lead to miscalculations within the younger age groups because they probably face the most change within the time difference. This problem could be fixed by taking into consideration new surveys done within 2020 and cross referencing the proportions to get a more accurate dataset.

Another possible weakness could be our key indicator variables which were age groups and education level. While these are strong indicators for someone's political stance, for others it might have nothing to do with how they vote. A possible solution for this is to survey the population about what are deciding life factors are for them when they are voting. What about the characteristics effect their vote more than others. With this information we could create new datasets that choose the key factors in what decides a person's vote.

Looking at the data in a binary setting also reveals many flaws within our research. Since we only took into consideration those who gave a definite answer on either Donald Trump or Joe Biden. We chose not to look into the unsure answers as to not impose our own opinion into the dataset. But this information could also be helpful in looking for how many votes were still up for grabs. We could have looked at similarities in the undecided votes with those who did vote a certain way to find a correlation that indicates their possible voting direction.

As for next steps, we may look into modelling the electoral colleges instead of the popular vote. This due the electoral college being the actual deciding factor of the election instead of the popular vote. As we could see in the 2016 election (Burns (2019)), Hillary Clinton received more votes overall but Donald Trump won because he won more electoral colleges. This would greatly increase the strength and accuracy of our research and help our future model.

References

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.

Burns, Sarah. 2019. *Curious Kids: How Come Donald Trump Won If Hillary Clinton Got More Votes?* https://theconversation.com/curious-kids-how-come-donald-trump-won-if-hillary-clinton-got-more-votes-126658.

Larmarange, Joseph. 2020. *Labelled: Manipulating Labelled Data.* https://CRAN.R-project.org/package=labelled.

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

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

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

Ushey, Kevin, JJ Allaire, Hadley Wickham, and Gary Ritchie. 2020. *Rstudioapi: Safely Access the Rstudio Api*. https://CRAN.R-project.org/package=rstudioapi.

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

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