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 presedential election coming to an end soon, this paper looks into predicting who will win the the presidency this November.First sentence. By gathering data from the 2018 US census and survey data from nation scape, we were able to stratify our findings by age groups. After stratifying, we were able to more accurately model the population of the United States, letting us predict who will win. We find in this paper that (insert winner's name) is most likely to win the 2020 US election.

1 Introduction

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

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.

This research is about predicting who will win the popular vote for the US election. 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. 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.

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 ot not. We chose logisitic regression as the majority 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 willl 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:

$$log(\frac{p}{1-p}) = B_0 + B_1 x_1 + ... + B_k x_k \tag{1}$$

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

Where p is the probability that the event of interest occuring (voting for Joe Biden), B_0 is the y-intercept and $B_i, 1 \leq i \leq k$, coefficient represents change in log odds for one unit increase in x_i.

There are several sections to this paper. Section 2 will talk about retrieval and cleaning of the survey data and ACS data, section 3 will detail the model that us use to predict the winner of the US election, section 4 we will get to see the results, section 5 will be discussion on the findings and weaknesses.

Through our research we were able to predict that (winner's name) will win the the presidential election of 2020. That is they have the highest probability of winning with the survey information we have used.

2 Data

The survey data is provided by Nationscape (Tausanovitch and Vavreck (2020)).Nationscapre conducts interview to 50,000 americans corvering the campaign and election. All responses 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 devived from the the adult population of the 2017 American Community Survey of the U.S. Census Bureau.

The stratification dataset is 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 pur 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 wee needed to be able to have flexibility in shaping our models and graphs. This acted as our population data.

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. This was done by adding anew colemn named 'age_groups' or 'age_range'. With this new column we categorised each person into one of five different age groups:

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

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.

Then we began the post-stratification by age groups. When looking at the population data from IPUMS US we only focused on those 18 and above, to only represent the population that could actually vote. When calculating the percentage we got:

- '18-30 years old' = 19.6%
- '31-45 years old' = 21.9%
- '46-65 years old' = 30.7%
- 'Above 65 years old' = 15.1%

With these new proportions, we turned to the sample dataset from Nationscape. We chose randomly from each age group proportionate to the percentage of the population data and cut it down the 2000 entries. This was our final dataset that had been stratified by age groups. We then created a new column named 'vote2020_bin'. This was a binary variable that had either 0 or 1. If the person planned on voting Joe Biden, it equalled 1, if they planned on Donald Trump, it eaqualled to 0. With this new column, we found our binary variable of interest that could be used with logistic regression to predict the probability of who is voting for who.

Figure 1 shows something

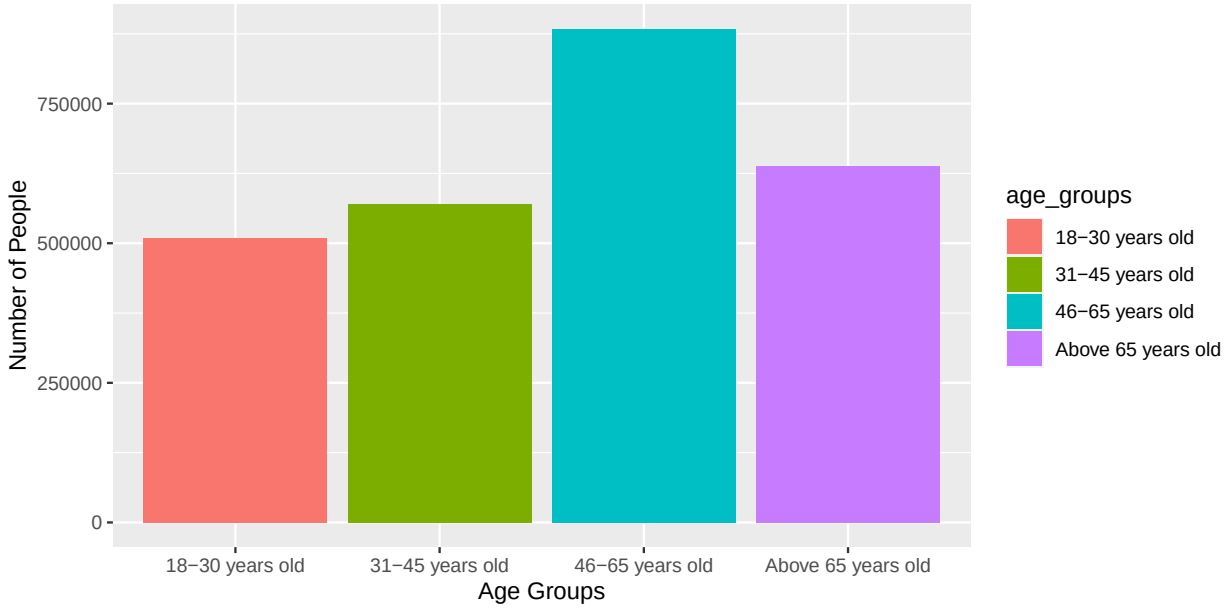

Figure 1: Frequency of Age Groups for Population Data

Figure 2 aslo shows something

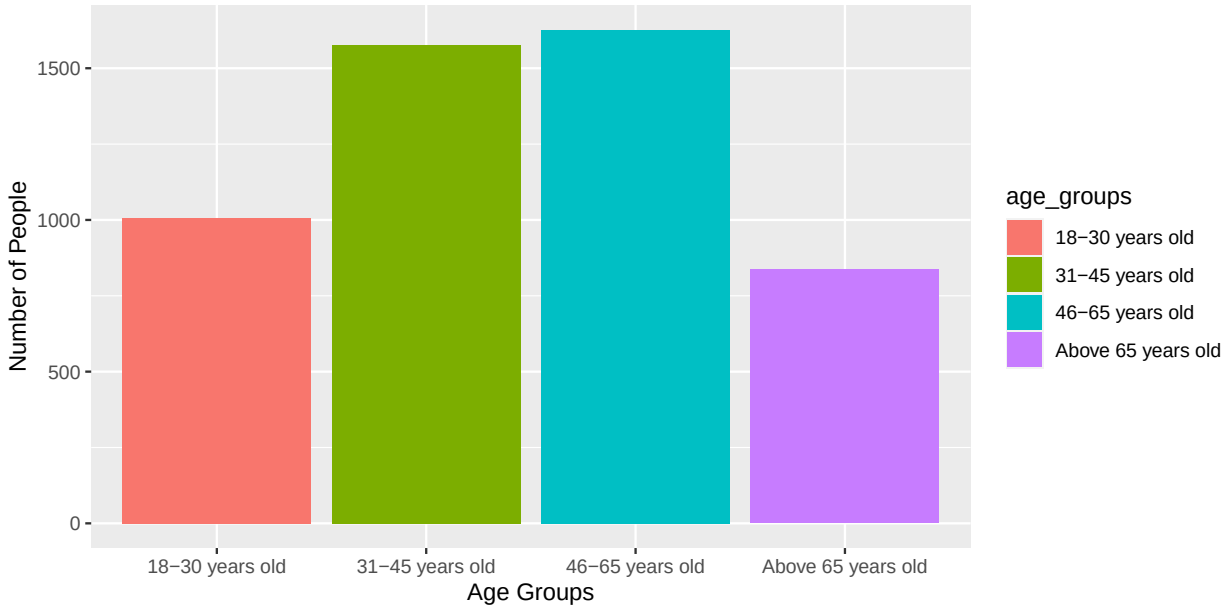

Figure 2: Frequency of Age Groups for Survery Data

```
##
## Donald Trump    Joe Biden
```

```
##        2417        2630
```

3 Model

After molding the data to our needs, we compute the generalised linear model with vote2020_bin as the dependent variable and age and education level being out independent variable.

We then called onto the summary(first_logit) function to retrieve all the needed coefficients and assigned simpler variable names to each value to form our regression formula:

4 Results

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

5.4 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 nad help our future model.

Appendix

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.