Predicting the US Election Outcome Using Poststratification Techniques

Chen Hao Yu, Pritam Sinha, Chae Bae

November 2, 2020

Predicting the US Election Outcome Using Poststratification Techniques

Chen Hao Yu, Pritam Sinha, Chae Bae

October 30, 2020

Abstract

Given the upcoming election in the United States of America, this study aims to predict the final results of the election and the candidate who will become the forty-sixth president. In order to accomplish this, we have utilized U.S Census data from IPUMS, as well as the full Nationscape Data Set from the democracy fund Voter Study Group. By using the R language from the R Foundation for Statistical Computing, we have created a logistic regression model to show the outcomes of the election. As a result of the United States utilizing the Electoral College, a candidate will win presidency with at least 270 electoral votes. Furthermore, when looking at the popular vote, Donald Trump wins with about 58.6% of the votes. We have found that Donald Trump will win the election with 339 electoral votes, whereas Joe Biden will lose the election with 199 electoral votes. Meaning, Donald Trump will remain in the White House.

Introduction

As the 2020 US Presidential Election is just around the corner, it has drawn a lot of attention around the world. Many presses have published various forecasting models to predict the election result. Knowing that all other major election forcasters had turned out to be wrong in previous US presidential election, what indicators are appropriate to be used for the prediction model (Houseman, 2020)? It is prevalent that a certain group of voters can heavily influence the election result. Back in 2016, caucasian working-class female voters accounted for nearly 20% of the total electorate (Catanese, 2019). In this paper, we have performed a statistical prediction using logistic regression and post-stratification analysis with demographic factors of gender, race, education level, employment status, and the residing state.

The American election system is called the Electoral College, and works very differently from other democracies. Unlike most democracies, the president need not win with a popular national vote. In total there are 538 electors in the United States and in order to win, the presidential candidate must have the most electors to win, meaning that a majority of 270 votes are needed (Robertson, 2020). Each state is assigned a fixed number of electors based on the population of each state; this is equivalent to the number of congressional districts each state has plus two electors. Since each state has at least one congressional districts, each state is guaranteed to have a minimum of three electors (Marshall & Freelander, 2020). In addition, the Electoral College operates on a winner takes all system, that is, the presidential candidate that recieves the most votes in the state wins all of the electoral votes in that state, regardless of the percentage of voters who support either candidate (Murse 2020). This can cause presidents to lose the popular vote and still win the election;

in total this has occurred 5 times in US history, most notably in 2000 with Bush's victory over Al Gore, and the most recent Trump's victory over Clinton in 2016 (Marshall & Freelander, 2020).

Model

We created a logistic regression model with the percentage of voters voting for Trump as the binary response variable, and age, sex, race (ethnicity), education level, employment status, and state as the explanatory variable. Age, sex, race, employment status and education are our level 1 individual level variables, while state is a level 2 group level variable. Using this model, we will predict the probability that a voter will vote for Trump (or equivalently Biden) by using the above mentioned explanatory variables. Then using that probability for that given demographic, we will apply that probability to each different demographic for each US state. This will tell us which presidential candidate will likely win the popular vote (and all the electoral votes) of that state. After we tally up the final electoral vote for each state, we will arrive at our prediction for the president-elect.

Model Specifics

I will be using a logistic regression model to model the proportion of voters who will vote for Donald Trump. Our justification for using this model is because the election result is binary; our assumption is that either Biden or Trump will win this election. The following is the formula for the logistic model.

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{Chinese} + \beta_4 x_{Japanese} + \beta_5 x_{African-American}$$

$$+\beta_6 x_{OtherAsian_or Pacific Islander} + \beta_7 x_{OtherRace} + \beta_8 x_{White} + \beta_9 x_{Did Not Complete High School}$$

$$+\beta_{10} x_{High School Diploma} + \beta_{11} x_{Not in Labor Force} + \beta_{12} x_{Unemployed} + \beta_{13} x_{alaska} + ... + \beta_{62} x_{wyoming} + \epsilon$$

Where p represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the model, and is the logit probability of voting for Donald Trump if the person is age 0, a female, an Alaskan or American Indian, a college undergraduate or above, employed, and lives in the state of Alabama. β_1 represents the age of the person, which means for every increase in age by 1, the logit probability of voting for trump increases by β_1. β_2 represents the sex, and the logit probability increases by β_2 when the person is female. β_3 through β_8 represents the change in logit probability in race, β_9 represents the change in logit probability if the person has dropped out of high school and β_10 represents high school diploma. β_11 represents the change in logit probability when the person is not in labor force, and β_12 represents the change in logit probability if the person is unemployed. The rest of the parameters describe the change in logit probability for each US state and the District of Columbia.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Here we create cells based off different ages, sex, race, education, employment status and the state they reside in. Using the logistic model described in the previous sub-section we will estimate the proportion of voters in each demographic bin according to state. We will then weight each proportion estimate (within each state) by the respective population number of each demographic and sum those values and divide that by the entire state population number.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 51 x 2
##    state               alp_predict
##    <chr>                     <dbl>
##  1 alabama                   0.609
##  2 alaska                    0.505
##  3 arizona                   0.666
```

```
##  4 arkansas                0.573
##  5 california              0.00161
##  6 colorado                0.576
##  7 connecticut             0.398
##  8 delaware                0.494
##  9 district of columbia    0.497
## 10 florida                 0.989
## # ... with 41 more rows
```

This table shows the probability that each state will vote for Trump.

Results

```
## [1] 339
```

```
## [1] 199
```

By setting the vicotry threshold for Trump at the minimum of 50% of state popular vote, we see that Trump will win the election with 339 electoral votes, surpassing the 270 votes needed for him to stay in office. The below graph will show what the election outcome map should look like.

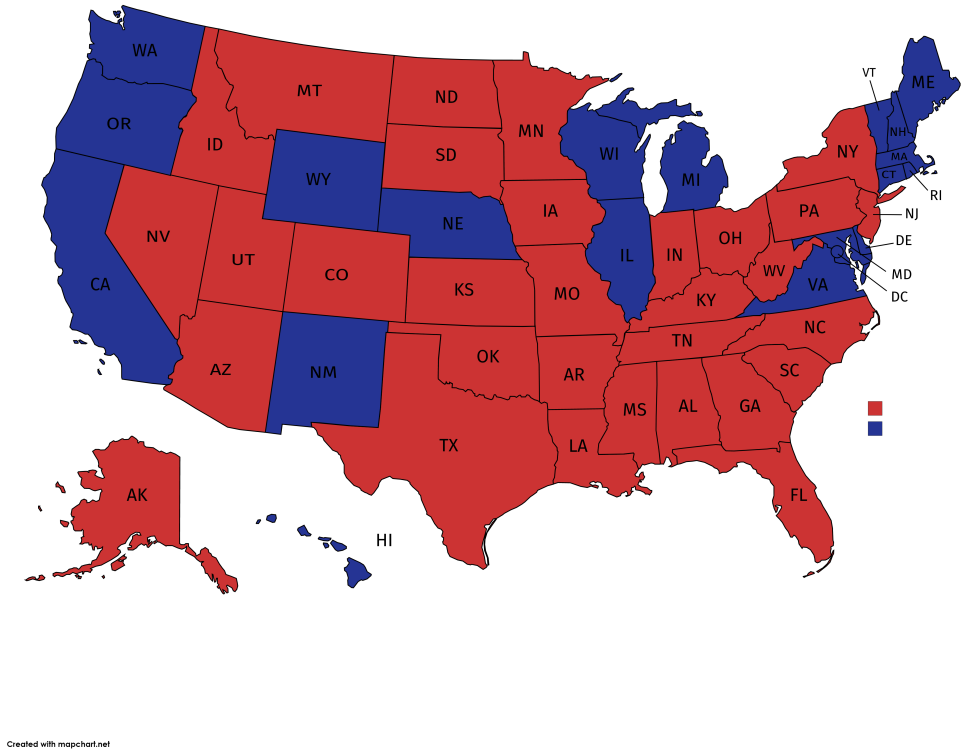

Figure 1: Election Results Based on the Model

The red color is the states Donald Trump is projected to win, while the blue color is the states Joe Biden is expected to win.

```
sum(popular_vote$trump_voters_per_state)/sum(popular_vote$pop_per_state)
```

```
## [1] 0.5855106
```

By looking at the popular vote using the poststratification formula for national trump voters, we find that $\hat{y}^{PS} = 0.586$. In other words, Trump will win the popular vote with 58.6% of voters.

Discussion

Summary

Earlier in our study we have used a logistic regression model to predict the final outcome of the election in the United States of America. In doing so, our findings show that Donald Trump will win the election with 339 electoral votes, while Joe Biden will loose with 199 electoral votes. We have found this by finding the probability an individual will vote for Trump based on the variables of race, age, state, employment, and education. As a result of the United States using the Electoral College, we have found candidate's electoral votes by seeing which candidates win in each state. To continue, when a candidate wins the state, they receive all of the electoral votes from it. We have summed up all of the electoral votes of Trump and Biden and have found they received 339 and 199 electoral votes respectively. As a result of all these steps, we were able to create a visual diagram showing which states voted for which candidate.

Conclusions

Based on the given model and results, we have reason to believe that Donald J. Trump will win both the popular vote for the country and the electoral college with a solid victory. We expect 339 electoral votes to go to Trump, with Biden far behind at 199 votes. The popular vote will likely swing in favor of Trump with 58.6% of voters supporting him.

Weaknesses

As we perform the analysis, we found that there some weaknesses that could possibly disturb the prediction result. One weakness could be the assumption that the demographic group of our data consists of every category in the group. For example, the census data under education level only tells you how many years of college, not if you completed a degree. Therefore, we could not determine the difference between someone in the middle of the college and those who dropped out. Other weakness could be an assumption that every stratum will represent the whole population of demographic group, meaning that belonging into a certain demographic does not mean everyone will vote in the same way as the majority of the people in that demographic. Finally, the race variable does not have Hispanics as a separate demographic group but included under the White demographic group. This can mislead the result with a bias.

Next Steps

Overall, our predictions followed an excellent analysis of 2020 US Presidential Election. However, like what happened in 2016, the forecasts can be wrong with any unexpected key variables (Houseman, 2020). The future steps could be done to improve our estimation after we get the actual result of the election. With use of the election result being released on November 3rd, we can further compare and improve our model to find which variables affected the most on the actual result and what other models could be better to reflect the actual election result.

References

Catanese, David. "The kind of voter Trump cant lose: Working-class white women drift toward Democrats" November 5-19. McClatchy DC. https://www.mcclatchydc.com/news/politics-government/election/article236982463.html

Helena Robertson, Ashley Kirk and Frank Hulley-Jones. "The Guardian Electoral college explained: how Biden faces an uphill battle in the US election" October 30-20. https://www.theguardian.com/us-news/ng-interactive/2020/oct/30/electoral-college-explained-how-biden-faces-an-uphill-battle-in-the-us-election

Housman, Patty. “Does Allan Lichtman Stand by His”13 Keys" Prediction of a Joe Biden Win?" October 28-20 https://www.american.edu/cas/news/13-keys-election-prediction.cfm

“List of State Electoral Votes For 2020.” List of State Electoral Votes For The 2020 Election, https://state.1keydata.com/state-electoral-votes.php.

Murse, Tom. “How Electoral Votes Are Awarded”. ThoughtCo. October 3-20 https://www.thoughtco.com/how-electoral-votes-are-distributed-3367484#citation-1

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200131). https://www.voterstudygroup.org/publication/nationscape-data-set

“The R Project for Statistical Computing”, https://www.r-project.org/

“United States”. MapChart, https://mapchart.net/usa.html

Vox. “The Electoral College, explained” Youtube, October 31-20. https://www.youtube.com/watch?v=ajavsMbCapY

Wickham, Hadley. “ggplot2: Elegant Graphics for Data Analysis”, Springer-Verlag New York, 2016. isbn 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wu, Changbao and Mary E. Thompson, “Sampling Theory and Practice”, Springer