Predict the overall popular vote of the 2020 American federal election using a regression model with post-stratification

Xinyi Chen, Zehui Yu, Yuxin Xie

2 November 2020

Predict the overall popular vote of the 2020 American federal election using a regression model with post-stratification

Xinyi Chen, Zehui Yu, Yuxin Xie

2 November 2020

Model

Model Specifics

We aimed at predicting the overall popular vote of the 2020 American federal election. It is a challenging question as there is a lot of people in the America and we cannot do a survey with each individual in the whole population in realistic and there are a lot of variables, ranging from background information such as age, gender, race, education, state to politics preference. However, multilevel regression has been shown to be an effective method of adjusting the sample to be more representative of the population for a set of key variables. As a result, we use a multilevel logistic regression model with random intercept to predict the probability of a voter who will vote for Donald Trump. We focused on 3 variables, education, gender and age. The reason why we choose this 3 variables is because people of different ages experience different politic backgrounds and their education level may have an effect on their politic preferences. One individual variables are education, which has 9 levels; and gender, which is either male or female. We set age-group to be the group level, which contains 6 groups: 20 or less, 21 to 35, 35 to 50, 50 to 65, 65 to 80, and above 80. We used these three variables from survey data to model the probability of voting for Donald Trump. The multilevel logistic regression model we are using is:

$$log(P/(1-P))_{ij} = \alpha + a_j + \beta_1 * eduassociatedegree + ... + \beta_8 * edumiddleschool + \beta_9 * male + \epsilon_{ij}$$

Where $log(P/(1-P))_{ij}$ represents the log odds of a voter voting for Donald Trump. α is the coefficient mean, it is the baseline of the model. a_j indicates random variable, which follow normal distribution. j indicates the age groups. β_1 to β_8 represent the slopes of dummy variables of education level. Which means for a voter's education level changed from "3rd Grade or less" to "Associate degree", we expect a β_1 increase in the probability of voting for Donald Trump. β_2 to β_8 has the same meaning as β_1. β_9 represents the slope of gender. which means when the gender changed from female to male, we expect a β_9 increase in the probability of voting for Donald Trump. ϵ_{ij} represents the residual of the model, it is a random variable and follows normal distribution.

Post-Stratification

Post-stratification is a common technique in survey analysis for incorporating population distributions of variables into survey estimates. The basic technique divides the sample into post-strata, and computes a post-stratification weight for each sample case in post-stratum. Inference about the population is one of the

main aims of statistical methodology. And post-stratification is an ideal method of adjusting the sample to be more representative of the population. In order to estimate the proportion of voters who will vote for Donald Trump, we perform a post-stratification analysis. We are interested in 3 variables ranged from age, education to gender. The gender gap in voting preference is quite obvious from the past two decades according to exit polls by the National Election Pool, such as 56% of women affiliate with the Democratic Party, compared with 44% of men in 2018's election as reported by CNN. And higher educational attainment is increasingly associated with Democratic Party affiliation and leaning. For instance, those without college experience that tilted more Democratic than Republican from the past 1o years according to BBC news. What's more, the generational gap which is reflected by age in partisanship is now more pronounced than in the past, and this echoes the widening generational gaps seen in many political values and preferences. As a result, we create cells based on different ages by separating it into 6 generation gaps. Using the model described in the previous sub-section, which is the multilevel regression model, we will estimate the proportion of voters in each generation. Then we proceed to weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

Results

```
## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
## # A tibble: 11 x 6
##    term                        estimate std.error statistic   p.value group
##    <chr>                          <dbl>     <dbl>     <dbl>     <dbl> <chr>
##  1 (Intercept)                   -0.904    0.924     -0.978  3.28e- 1 fixed
##  2 educationAssociate Degree      0.261    0.912      0.287  7.74e- 1 fixed
##  3 educationCollege Degree (~     0.317    0.908      0.349  7.27e- 1 fixed
##  4 educationCompleted some c~     0.386    0.909      0.425  6.71e- 1 fixed
##  5 educationCompleted some h~     0.576    0.914      0.630  5.29e- 1 fixed
##  6 educationDoctorate degree      0.780    0.926      0.842  4.00e- 1 fixed
##  7 educationHigh school grad~     0.638    0.909      0.701  4.83e- 1 fixed
##  8 educationMasters degree        0.486    0.910      0.534  5.93e- 1 fixed
##  9 educationMiddle School - ~    -0.243    1.15      -0.210  8.33e- 1 fixed
## 10 genderMale                     0.518    0.0626     8.27   1.34e-16 fixed
## 11 sd_(Intercept).agegroup        0.424   NA         NA     NA        agegr~
```

```
## # A tibble: 6 x 4
##   agegroup   trump_predict    num group_weight
##   <fct>              <dbl>  <int>        <dbl>
## 1 20 or less         0.572 129527       74146.
## 2 21 to 35           0.597 544711      324976.
## 3 35 to 50           0.621 532198      330629.
## 4 50 to 65           0.626 639259      400139.
## 5 65 to 80           0.622 456274      284007.
## 6 above 80           0.622 143106       89081.
```

```
## [1] 0.6146965
```

We estimate that the proportion of voters that lean towards of voting for Republican Party that votes for Donald Trump to be 0.6146965. This is based on our post-stratification analysis of the Republican Party modelled by a multilevel regression model, which accounted for age, gender and education. In our model, when people get more educated and further their study towards higher degree, such as from college degree to master degree, we could expect a 0.48647 increase in the probability of voting for Donald Trump. Democrats held a significant advantage among voters with a high school degree or less education for much of the time.

And when the gender changed from female to male, we expect a 0.51805 increase in the probability of voting for Donald Trump. This shows that the Republican gains among women have not come from increased affiliation with the party. What's more, it is interesting to note that people tend to vote for Republican as they getting older. For example, people who are younger as 20 or less only have a 0.5724373 preference for voting Donald Trump, while people from 50 to 65 would like to vote for Donald Trump with a probability of 0.6224489.

Discussion

Summary:

Our goal is predicting the overall popular vote of the 2020 American federal election. It is a challenging question as there is a lot of people in the America and we cannot do a survey with each individual in the whole population in realistic and there are many variables need to take account into. However, multilevel regression and post-stratification (MRP) has been shown to be an effective method of adjusting the sample to be more representative of the population for a set of key variables. As a result, this article provides an overview of multilevel regression and post-stratification. It reviews the stages in estimating opinion for some populations, and provides a worked example of the voting preferences for the whole population in the United States using publicly available data sources from VOTER STUDY ROUP and IPUMS USA. We select registration, vote_intention, vote_2020, gender, education, and age from the survey data for further analysis. The information we are interested in is only those are registered and have vote intention as well as aged over 18. Thus we have cleaned our data to meet the criterion. In order to match and compare the survey data with the census data, we mapped the data styles to make the same type of variable in the same format in both data. For instance, in both data files, we classified age to a generation group and unified the name of different education levels. Also, we convert the column name sex to gender in the census data. All of these cleaning steps are helpful and essential for later comparison and analysis. As a result, we made a multilevel regression model based on generation, gender and education level and proceed to post-stratification. Our model yields that Donald Trump who represents for Republican Party will win 2020's election with a probability of 0.6146965. And people who would like to vote for Republican Party are typically from elder generation.

Conclusion:

We think that Donald Trump will win the primary vote. Based off the estimated proportion of voters in favour of voting for Republican Party being 0.6146965, we predict that Republican Party will win the election. From the summary table, we noticed that as people getting elderly, they are more likely to vote for Donald Trump. The Silent Generation (that is 80 and above) and people aged from 65 to 80 have a relative similar preference for voting to Republican Party, more than half (62%) of Silent Generation voters identify with or lean toward the Republican Party, a larger share than a decade ago; while in contrast, Millennial voters (aged from 21 to 35 and those below 20) have had a Democratic tilt since they first entered adulthood; this advantage has only grown as they have aged. From immigration perspective and race to foreign policy and the scope of government, two younger generations, Millennials and Gen Xers, stand apart from the two older cohorts.And our mutilevel regression model shows that when gender changed from female to male, we would expect a 0.51805 increase in the probability of voting for Donald Trump.Since the violence of gender discrimination raised highly and causes a lot of attention this summer, some of the female may lose hope and distrust Donald Trump. Pew Research Center surveys also shows that women are significantly more likely than men to associate with the Democratic Party. Overall, the proportion of man voters who identify with the Republican Party has remained relatively constant and there is a slightly difference between that with women. Higher educational attainment is increasingly associated with Democratic Party affiliation and leaning. People who have achieved college degree and attained higher degree level tend to vote for Democratic Party with higher probability. At the same time, those without college experience are more likely to vote for Republican Party.

Weaknesses

Firstly, our dataset only includes 4392 observations which is a small sample size compared to the population. Hence the data value may not be that representative. Besides, in order to meet our assumption, we filtered out those people who do not registered and do not have an intention to vote and only focus on American citizen who has the voting right and vote for Donald Trump or Joe Biden. However, this may not be true in realistic. People do not registered does not actually mean they won't vote for their president. Since we have ignored these people's willingness to vote thus our sample may be a little bit bias. What's more, we have limited computing power and the data we used is several moths ago which may not catch up to date.

Next Steps

Since the election is an ongoing process that we could update our database with the newest voting result to make our prediction more time-sensitive. What's more, we could increase the sample size to get more data, getting more people involved and collect more information, thus make a more reliable and representative conclusion based a relatively larger sample size. And we can compare with the actual election results and do a survey of how to better improve estimation in future elections. Last but not least, we could also consider the influence of race and state that may make a change in our model prediction.

References

1.Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [URL].

2.American Community Surveys (ACS). (2018). Retrieved November, 2020, from https://usa.ipums.org/usa-action/data_requests/download

3.Post-Stratification: A Modeler's Perspective. (n.d.). Retrieved November 02, 2020, from https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1993.10476368

4.Tyson, A. (2020, July 28). The 2018 midterm vote: Divisions by race, gender, education. Retrieved November 02, 2020, from https://www.pewresearch.org/fact-tank/2018/11/08/the-2018-midterm-vote-divisions-by-race-gender-education

5.Trends in party affiliation among demographic groups. (2020, August 28). Retrieved November 02, 2020, from https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/

6.The Generation Gap in American Politics. (2020, August 14). Retrieved November 02, 2020, from https://www.pewresearch.org/politics/2018/03/01/the-generation-gap-in-american-politics/

7.Samantha-Jo Caetano(2020), data cleaning code (01-data_cleaning-post-strat1.R).https://q.utoronto.ca/courses/184060/files/

8.R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

9.Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

10.Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

11.Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

12.David Robinson and Alex Hayes (2020). broom: Convert Statistical Objects into Tidy Tibbles(Version 0.5.3) [R Package]. https://CRAN.R-project.org/package=broom

13.RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Github

https://github.com/xinyi-chen-16/Sta-304-PS3/upload/main