Analysis of Predicted Votes for 2020 President Election

Siyi Lin, Hailan Huang, Zizhuo,Huang

2020-11-02

Model

Model Specifics

We will be using data collected from survey data to predict the overall popular vote of the 2020 American federal election. But this data is not representative to analyze our goal. Therefore we use MRP (multilevel regression with post-stratification) to model the proportion of voters who will vote for Joe Biden. Because of the policy of electoral college, the candidate who collects more votes in a state will gain all of the votes from the electors in that state, except for Maine state and Nebraska state. The candidate who owns the most elector votes will be the winner eventually.

To do this, we first partition the population into cells, two models will be created. For model 1, we set random intercept and coefficient model which means both intercepts and slopes are allowed to vary. In the random effects part, the standard deviation of cell is 0.8147 which is the standard deviation of normal distribution followed by intercept of baseline. In the fixed effects part, the list of estimates are the constant of each variables in cell. It is the same when applying model 2 where the only difference is that we substitute the state into age group in cells.

Post-Stratification

We transform the raw data into accurate estimates of voter intent in the general electorate, we make use of the rich demographic information that respondents provide. The core idea of post-stratification is to partition the population into cells based on combinations of various demographic and political attributes, use the sample to estimate the response variable within each cell, and finally aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population.

First model containing 18,400 cells will be using gender (2 categories), education (8 categories), state (50 categories), and household income (23 categories). Second model containing 1,840 cells will be using gender (2 categories), age group (5 categories), education (8 categories), and household income (23 categories). Then we compare these two models, select a better one and apply it to the census data of 2018. These variables are included because they are available in both the survey data and the census data, and they are likely to be significant and influential to our goal.

```
## Warning: Unknown or uninitialised column: 'electoral_votes'.
```

Results

AIC is a statistic that balances the goodness of fit of the model with a penalty term reflecting how complex the model is. And the penalty for BIC is more severe than the AIC so it will favor simpler models more. The preferred model is one with the lowest AIC. As with AIC, smaller values of BIC indicate the better model. After comparing, model 2 has much lower AIC and BIC so we choose model 2 to apply to the census data.

We use y to indicate the outcome of interest, the post-stratification estimate is defined by $\hat{y}_{ps} = \sum_{j=1}^{j} N_j \hat{y}_j / \sum_{j=1}^{j} N_j$ where y_j is the estimate of vote for which candidate in cell j, and N_j is the population size of the jth cell whether it votes for BIDEN in each demographic cell. To illustrate this approach, we

compute Xbox estimates of Bidden support for each level of our categorical variables, and compare those with the actual voting behaviors of those same groups, as estimated by the 2018 national exit poll.

We calculate that the proportion of voters intending to vote for BIDEN to be 0.5910781. It is according to the post-stratification analysis of the proportion of voters in favor of BIDEN modeled by multilevel regression model, which account for age group, gender, education, and household income.

```
predicted_states %>% group_by(winner) %>% summarise(total_votes=sum(electoral_votes))->election_result
election_result
```

```
## # A tibble: 2 x 2
##   winner       total_votes
##   <chr>              <dbl>
## 1 Donald Trump         220
## 2 Joe Biden            318
```

Discussion

Post-stratification is a common statistical technique to correct estimates when there are known differences between target population and study population. It is a method which aggregate cell-level value by weighting each cell by its relative proportion in population. It is useful because it increases the representativeness of the sample.

We figure out that Donald Trump will have 220 votes while Joe Biden will have 318 votes. According to the estimated proportion of voters intending to vote for Joe Biden being 0.5910781, we predict that Joe Biden will win the election.

Weaknesses

After setting the model and analysis, we find that there are some weakness in this model. First one is that we assume two states, ME state and NE state, are the same as other states but these two states have different voting policy. Everyone in these two states can vote for candidates they want while others have electoral college. Second one is that this census is done by 2018, which may not be effective in 2020 vote.

Next Steps

What we can do next is that adding more data in survey cause it may not be adequate to support our conclusion. We can also divide cell more specifically than we an get more detailed result to predict who will win the election. Besides, we can add a weight on the education. Cause education level is one important factor influencing the vote intention and it is supposed to account for more proportion when considering the vote for candidates.

References

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

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

Wei Wang, David Rothschild, Sharad Goel, Andrew Gelmana, chttps://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf, 2014