2020 American Federal Election Prediction with Post-Stratification

Hyunseok Rha

November 2, 2020

2020 American Federal Election Prediction with Post-Stratification

Hyunseok Rha

November 2, 2020

Source code of this report can be found at: https://github.com/hyunR/STA304/tree/main/PS3

Model

In this report, I will predict the 2020 American federal election with `Nationscape Data Set`[1] and `U.S. CENSUS DATA`[2] using multilevel logistic model with a post-stratification tenchnique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

I will be using a multilevel logistic regression model to model the proportion of voters who will vote for Donald Trump. This model has 2 levels that, the first level is an individual level with factors of age_group, gender, household_income, race_ethnicity and hispanic.The sescond level is a state level. There are many different factors that affects an individual's political stand, hence I will use the individual specific factors for my model. Also, I believe individuals from the same state would share similiar political stand, so I choose to use state as the second level of for my model. The multilevel logistic regression model I am using is:

$$y = \beta_0 + \beta_1 x_{age\_group} + \beta_2 x_{gender} + \beta_3 x_{household\_income} + \beta_4 x_{race_ethnicity} + \beta_5 x_{hispanic} + \epsilon$$

$$\beta_0 = W_{state}$$

Where y represents the probalility of an individual voter to vote for Donald Trump with certain factors.And β_0 represents the second level of the multilevel logistic regression which is `state`. And other βs represents the relationship between the probalility of an individual voter to vote for Donald Trump and each factor.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Using the model described in the previous sub-section I will estimate the chance of voters vote for Donal Trump. Post-Staratification works in a way that, we have a model that has a logic about how each factor would affect the chance of voting for Donal Trump. Then, we can apply the same logic to a larger data and predict what is the chance of each group to vote to Donal Trump would be. Post-Staratification is useful that we can use a large dataset such as `Census` that lacks of politicial information, to predict the missing information, in this case the chance of vote for Donal Trump, with a relativly small but well targeted survey such as `Nationscape Data Set`. Using the model described in the previous sub-section I will estimate the proportion of voters in each cell of age_group, gender, household_income, race_ehnicity, hispanic and state. For each voter, I will estimate the chance of voting for Donal Trump and save the values into `estimate` column. Then, since the `estimate` column is the chance of voting, having negative and greater than 1 as a value of `estimate` column does not make sense. Hence, I will squeeze them to be within 0

to 1 range and call it as `predict_vote_trump`. Then calculate the average of `predict_vote_trump` that represents the predicted portion of population who vote for Donal Trump.

Results

Based on my post-stratification calculation with my multilevel logistic regression model, the portion of population that vote for Donal Trump, which is $\hat{y}^{PS}$, about 10.1%.

Discussion

In this report, I predicted 2020 American Federal Election result with multilevel logistic regression and post-stratification using `Nationscape Data Set` for modeling and `U.S. CENSUS DATA` for post-stratification.

Based on the modeling and post-stratification calculation, the portion of voters who would like to vote for Donal Trump is about 10.1%, hence I predict taht the Democratic Party will win the 2020 American Federal Election and Joe Biden will be the president of United States of America.

Weaknesses

The weaknesses of my report is that the number of factors are not large enough to have more fine detailed model and result that can work with complecated politics.

Also my report does not capture the system of Electoral College that sometime makes unpredicted result.

Next Steps

After the election is done, do a post-hoc analysis with a follow up survey to see how well my model predict the result and check any factors that could be used for better prediction.

References

[1] Tausanovitch, Chris and Lynn Vavreck . 2020 . Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814) . Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

[2] Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0