US 2020 Election prediction

Hung Dao and Hoang Huynh

2 November, 2020

US 2020 Election prediction

Hung Dao and Hoang Huynh

2 November, 2020

Model

Model Specifics

In preparation for the upcoming US 2020 election, we want to predict the result by building a logistic regression model based on a person's age and their self-identified race. We think this model is appropriate since we are interested in whether a person will vote for Donald Trump or not, hence it is a binary response. The model will look roughly like follow:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{age}x_{age} + \sum_{i=1}^{k} \beta_i x_i$$

Where p represents the proportion of voters that will vote for Trump. Furthermore, β_0 is the intercept which represents the *base log* proportion of voters under no other factor. Also, β_{age} represents the slope for the variable age in the model and β_i be the slope for the corresponding race x_i.

Post-Stratification

Post-stratification is a technique that examines a small group and extrapolates the result into a larger population. The group is split into a smaller subgroup with that upon closer inspection can provide more meaningful insight on the overall data. In our cleaned data, the **race** variable has 6 different responses. We decided to group some of the responses because our goal is to extrapolate our census data from the survey data, we need the response in both data set to be the same. We also removed observations where participants came from 2+ races because our survey data do not account for these categories. Another reason for that is to reduce ambiguity in our response, we did not outline a method to count people in those bins. Also, all responses with answered such as "unknown" or "don't know" are also removed from the data for the same reason.

Results

We estimate that the proportion of voters in favour of voting for Donald Trump to be 0.4556877. The result is based off on our post-stratification analysis of the proportion of voters in favour of Donald Trump modelled by a logistics regression model, which is built upon the variable *age* and *race*.

Discussion

Summary

This is our final logistic regression model, that is then used for post stratification based on the census and survey data:

$$log(\frac{p}{1-p}) = -0.399241 + 0.009244x_{age} - 2.064240x_1 - 1.343154x_2 - 1.393601x_3 + -0.619015x_4 + -0.752908x_5 + 0.068807x_6$$

x_1: represent race black/african american

x_2: represent race chinese

x_3: represent race japanese

x_4: represent race other asian or pacific islande

x_5: represent race other race

x_6: represent race white

Applying this model to each age and race category, we can calculate the proportion of people that will vote for Donald Trump in our survey data. Using the results and extrapolate into the census data, we can produce the proportion of voters who favour Donald Trump across our whole population by summing up the proportion for all of our bins.

Conclusion

Based off the estimated proportion of voters that will vote for Trump to be 0.4556877, we think that Trump got a very high chance of winning this year election because he has got more than 45% of the vote for himself. The remaining proportion of votes account for approximately 55% but this is for the rest of the candidates. Joe Biden is a very admirable competitor to Trump but we still believe, based on the result of our post-stratification, that Trump can win.

Weaknesses

A requirement for using post-stratification is that the variable names for both survey and census data must the same. So when we perform data cleaning, we have to modify some data points in these 2 data. In the survey data, we group different race so that they match with our census data variables. In the census data, we remove people who identify themselves with more than 1 race to reduce ambiguity. Also, we also filter out responses that stated 'don't know' or 'not going to vote' so our analysis might not represent the whole population accordingly.

Next Steps

We want to gain permission to access the upcoming election data to further check and improve our analysis/model. By having the current election data, we can improve our model significantly and it might be useful for future elections.

References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=ab8fe033-9152-414f-a49b-f556be29768f].

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