Towards a forecasing of American Presidential Election in 2020

Zhixing Hong, Lingyue Kong, Jinyu Luo

2020/11/02

Code and data suppor!ng this analysis is available at:

https://github.com/neverknowhen/Prediction_2020_Election

Introduction

The US presidential election is a process through the Electoral College. Although citizens are able to evolve in the voting process, the final result sometimes cannot be explained by the popular vote. For example, in 2016, Clinton won the majority popular votes, but her vote shares fell behind Trump's in several swing states, resulting the failure of winning the white house. This result showed the drawbacks of the Electoral College, so it is not sufficient to use the popular vote to predict the election result.

Due to our poor knowledge, it is necessary to refer the work done by experts in political science. Robert S. Erikson, Karl Sigman, and Linan Yao from Columbia University built model for predicting the Electoral College Bias for the 2020 presidential election. Their model suggested that the an important predictor is the state vote divisions in the last election. In other words, the data from 2016 will be very helpful for our forecasting.

The data set we used is the survey data collected by IPUMS USA, which including the latest investigation of American's intention to vote for the presidential election this year as well as the voting result in 2016. In the rest of paper, we compared the 2016 voting results with people's target candidate in this year by first visualizing the data. It is then followed by modeling the data and predicting through post-stratification.

Model

As we are now more and more closer to the American election day, the result of USA president election is also going to be revealed. Here, we are using the data from Nationalscape and the census data from Integrated Public Use Microdata Series (short for IPUMS) to predict the result of the 2020 American federal election with a post-stratification method. In the following sub-sections, model specifics and post-stratification method are presented.

Model Specifics

A logistic quasibinomial regression model is employed to predict the proportion of voters who will vote for Donald Trump in the 2020 election, with the designed survey methods.

`survey data` is obtained from the survey conducted by Nationscape and UCLA, and the sampling method they employed was purposive sampling. Therefore, to reduce the errors in the data which would influence the model result, the sampling method of the survey has been taken into consideration. The finite population is specified as the total population that has been sampled(the numbers of observations in the `census_data`

are used). The different states are viewed as different strata. To be noted here, the state are classified as `swing state` and `not swing state`. Moreover, the `weight` provided by the `survey data` are included in the survey design, as the weight enable the data to be used for a national sample.

After finalied the designed method, 5 different variables used in the model, `age group`, `gender`, `income`, `race` and `swing state`. All the five variables are representative for the voter's social and economics status and easy to get. As all the variables inclueded in the logstic model, `family = quasibinomial` is used instead of `family = binomial`. The `age group` categorical variable is used instead of `age`, since the people in the 15 years interval have similar characteristics. The `income` variable overwrites the original variable `household_income`: the `income` variable combined different levels in `household_income`. This aims to make the model more interpretative. All the five above mentioned variables are categorical variables with different levels. The logistic regression model used can be expressed as follows:

$$y = \beta_0 + \beta_1 gender_{male} + \sum_{i=2}^{5} \beta_i age\_group_i + \sum_{j=6}^{12} \beta_j income_j + \sum_{l=13}^{18} \beta_l race_l + \sum_{k=19}^{31} \beta_k swingstate_k + \epsilon$$

Where y represents the proportion of voters who will vote for Donald Trump. Similarly, β_0 represents the intercept of the logistic regression model. Additionally, $\beta_m (m \in [1, 31])$ represents the slope of each different variable. For example, β_3 is the coefficient for the age group (63, 78]. The coefficients in the logistic regression model are not able to show the relationship between the variable and probability vote for Trump directly, only if after the transformation. By applying the transformation on the coefficients, the general relationship is decreasing the log odds, increasing the real transformed coefficients. `R` software is used to build the logistic model and compute the model result.

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 the cells are based on different age groups. Using the model described in the previous sub-section to estimate the proportion of voters in each age group. Then we weight each proportion estimate (within each group) by the respective population size of that bin and sum those values and divide that by the entire population size. There are 5 different variables used in the model, `age group`, `gender`, `income`, `race` and `state`. We used `age group` instead of `age` since this variable divides people into age groups which every group has fifteen years difference which have similar characteristics and are more significant for the model. Therefore, there are 4 age groups from [18,93]. Also, there are 7 different groups separating the people which have income from $20,000 to $ 200,000. Note that there are several data that have negative income, and these are group into the income less than $19,999. Moreover, the `race` variable was cleaned by getting rid of the `two major races` and `three or more major races`, since these two groups contained into the census data but not the survey data, and this helps the samples to be one-to-one and divided into 6 different races. For the states, we first filtered all the swing states, since these states refer to any state that could reasonably be won by either the Democratic or Republican presidential candidate by a swing in votes, so the proportion may not change in a long century. After that, we get different unswaing states into different groups. For the total cells, we multiplied the number of groups of each variable and then we got 4\$7613($number of unswaing states$)\$2(two genders) = 4368 cells. We first calculate the predict proportion by multiply the estimate of each variable and the number of data in this variable, then the result could be the summation of the predict proportion divides the summation of the total number of data, the mathematical notation is

$$\overline{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_h$$

(where the population size be denoted N, the sample size n and the number of strata H. Index each stratum by h).

Additional Information

Descriptive statistics with continuous variables

Table 1: Proportion of Trump Supportion in survey data

	0	1
prop16	0.6561101	0.3438899
prop20	0.6118388	0.3881612

Table 2: Proportion of Trump Supportion in Census data

	mean	Std.error
voteTrump_2016	0.2734127	0.0081836
voteTrump_2020	0.3645291	0.0093893

We can see similar results from the original dataset and the one with finite population correction.

There are 38.8% of respondents are going to vote for Trump in 2020, about 4% higher than the proportion in 2016. We used `svymean()` function from the survey package to compute the design-based estimate of the distributions of factors "vote_trump" and "vote_trump_2016", representing respondents' intention of voting for Trump in 2020 and 2016 respectively. The standard errors were reported by the function as well, calculated by the formula $SE(p) = \sqrt[2]{\frac{p(1-p)}{n}}$.

Results

The logistic regression model used can be expressed as follows:

$$y = \beta_0 + \beta_1 gender_{male} + \sum_{i=2}^{5} \beta_i age\_group_i + \sum_{j=6}^{12} \beta_j income_j +$$
$$\sum_{l=13}^{18} \beta_l race_l + \sum_{k=19}^{31} \beta_k swingstate_k + \epsilon$$

Here, the β_0 to β_k represents the 'estimated' column of the table above, they are just coefficients of each variable, and according to the result of table3, it shows that most variables were negatively related to supporting Donald Trump in the 2020 election.

Table 3: Logistic Regression Model Result

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1087600	0.5694957	0.1909760	0.8485508
gendermale	0.4311423	0.0863279	4.9942395	0.0000006
age_group(48,63]	0.0756821	0.1137813	0.6651541	0.5059775

	Estimate	Std. Error	t value	Pr(>|t|)
age_group(63,78]	0.0988979	0.1222690	0.8088554	0.4186304
age_group(78,93]	0.2262642	0.3157025	0.7167005	0.4735869
age_group[18,33]	-0.4275258	0.1237089	-3.4559024	0.0005523
income$150,000 to $199,999	-0.0808477	0.1963073	-0.4118428	0.6804694
income$20,000 to $39,999	-0.1391672	0.1394355	-0.9980756	0.3182830
income$200,000 and above	0.3185109	0.1883268	1.6912677	0.0908378
income$40,000 to $59,999	-0.2067014	0.1406546	-1.4695669	0.1417317
income$60,000 to $79,999	-0.1196583	0.1533556	-0.7802664	0.4352649
income$80,000 to $99,999	-0.3220777	0.1675780	-1.9219576	0.0546585
incomeLess than $19,999	-0.4954037	0.1624662	-3.0492719	0.0023040
raceblack/african american/negro	-2.2586652	0.4879368	-4.6290117	0.0000038
racechinese	-1.3588383	0.6373274	-2.1320882	0.0330402
racejapanese	-0.4687864	0.9644007	-0.4860909	0.6269205
raceother asian or pacific islander	0.0287086	0.4894730	0.0586521	0.9532312
raceother race, nec	-0.6550380	0.4862807	-1.3470366	0.1780194
racewhite	0.1116692	0.4388378	0.2544658	0.7991445
swing_stateFL	-0.3268507	0.3755053	-0.8704290	0.3841009
swing_stateGA	0.2592875	0.4061090	0.6384677	0.5231937
swing_stateIA	-0.1283951	0.6335034	-0.2026747	0.8393962
swing_stateMI	-0.6385372	0.4267441	-1.4963002	0.1346282
swing_stateMN	-0.8853979	0.5594662	-1.5825762	0.1135709
swing_stateNC	-0.3477167	0.4166427	-0.8345681	0.4039942
swing_stateNH	-0.2789872	0.6670714	-0.4182269	0.6757962
swing_stateNV	-0.7882860	0.5741414	-1.3729824	0.1698093
swing_stateOH	-0.7735407	0.3840736	-2.0140430	0.0440497
swing_statePA	-0.6802627	0.3988181	-1.7056964	0.0881165
swing_stateTX	-0.1958693	0.3829564	-0.5114663	0.6090434
swing_stateunswing_state	-0.6708265	0.3484538	-1.9251521	0.0542577
swing_stateWI	-0.7732745	0.4562915	-1.6946940	0.0901855

The result, `prediction_result = 0.38672`, shows that people have a larger probability of not voting for Donald Trump. Here we assume 0 represents the people who are not vote to the Donald Trump and 1 represents to those who do vote to Donald Trump, since the prediction proportion is 0.38672 which is much closer to the 0, so people may have smaller probability to vote for Donald Trump, and he may fail in the election. This is based off our post-stratification analysis of the proportion of voters in favour of Donald Trump modelled by a logistic model, which accounted for 5 different variables used in the model, `age group`, `gender`, `income`, `race` and `state`. However, this result relied on all states which includes swing states and unswing states, it is surprising that after getting rid of the unswing states, the prediction proportion is much higher and he may win the election.

Discussion

Summary

The results reported by the logistic model are sufficient to explain the outcome variable. By first looking at the age groups, it is understandable that the young population, and minority race groups are less likely to support Trump. It is stated by James, G. (2013) that the estimated intercept is typically not of interest; its main purpose is to adjust the average fitted probabilities to the proportion of ones in the data. So in this model for the probability of supporting Donald Trump, the intercept with value -1.18 is not of interest. More meaningful estimates in the following are of interest. In terms of swing states, except Georgia, for every

additional number of people increase in the state, it is expected that the log adds of being satisfied with life to decrease by 0.1 to 0.9(i.e., probability of supporting Donald Trump is negatively related with age). "Dummy" variable is used for all five variables since all of them are categorical variables. Here, for estimates associated with income levels, if the person is in income $150,000 to $199,999, then it is expected that the log adds of supporting Donald Trump to decrease by 0.08. However, for income $200,000 and above, it is expected that the log adds of supporting Donald Trump to increase by 0.3185. It is shown that `income` is negatively related to supporting Donald Trump, it suggested that if a person has more income, then he or she is less likely to support Donald Trump, but after the threshold $200,000, the relationship will be opposite. Similarly for the `race`, for each person in Japanese increase, it is expected that the log adds of supporting Donald Trump to decrease by 0.4, but for each person in the race white increases, it is expected that the log adds of supporting Donald Trump to increase by 0.1.

According to the poll result provided by CNN, the proportion of polls Trump gained is only 3% behind Biden until 28th October 2020. As the model implies, our post-stratification prediction result indicated that the probability of Trump being re-elected is quite low, with only about 39%. However, if we focus on swing states, in other words, dividing stratas according to states, the chance of Trump being elected suddenly increased to 42%. In contrast, Biden was predicted to be leading the popular vote, but the percentage of votes Biden gained in swing states was about 5% less than Trump's share. The variations suggested that bias might also exist in popular vote (see detailed in Table 4).

Table 4: Supporter Number Summary

vote	count_swing_state	percent_swing_state	percent_total
Trump 2016	2074	0.4697622	0.3438899
Trump 2020	2341	0.5302378	0.3881612
Hillary Clinton	1899	0.4268375	0.3148732
Joe Biden	2550	0.5731625	0.4228155

Conclusion

As a conclusion, Trump will still have a big chance of winning the white house. Based on the raw data, we saw that there was a significant increase (+9%) in the number of Trump's supporters from 2016 (27.3%) to 2020 (36.4%) (see detailed information in Table2), although Biden was leading the population preference with approximately 42%. Having the 2016 election result as a reference, we are not able to draw the conclusion about the winner immediately. However, it is sufficient to say that the probability of Trump being elected is higher than the chance in 2016.

Weakness & Next Steps

There are some limitations and restrictions we found during the research. First, the dataset for stimulation is different from the dataset for prediction. As a result, the model fitted for the survey dataset was not able to be applied on the census dataset. In order to perform the prediction, we had to choose variables that contained in both dataset. However, the prediction accuracy was reduced immediately. Moreover, the survey dataset we used was the data collected in June, 2020. During these months, people's decisions might change due to some events happening during this time interval. Again, it would also influence our prediction results. It is hoped that in the future work, one can extend the model globally based on more global data, also we can do is waiting for the real election and training the model with the real results.

References

-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]

-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

-Wu, C., & Thompson, M. E. (2020). Sampling Theory and Practice (ICSA Book Series in Statistics) (1st ed. 2020 ed.). Springer.

-Balzer, Laura B. and van der Laan, Mark J., "Estimating Effects on Rare Outcomes: Knowledge is Power" (May 2013). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 310. https://biostats.bepress.com/ucbbiostat/paper310

-DataDhrumil. (2020, September 18). How Asian Americans Are Thinking About The 2020 Election. Retrieved November 02, 2020, from https://fivethirtyeight.com/features/how-asian-americans-are-thinking-about-the-2020-election/

-View latest 2020 presidential polling in Georgia. (n.d.). Retrieved November 02, 2020, from https://www.cnn.com/election/2020/presidential-polls/georgia

-Erikson, R. S., Sigman, K., & Yao, L. (n.d.). Proceedings of the National Academy of Sciences. Electoral College Bias and the 2020 Presidential Election. doi:10.1073/pnas