Federal Presidential Elections 2020

Langwen Guan, Yuhang Ju, Zike Peng

11/2/2020

Introduction

Elections in the United States of America are determined by a number of factors. Voters are primarily swayed by their identities, beliefs, and contexts. Since these issues are associated to their individual demographic classes, one can use a regression model to predict the outcome of the 2020 federal elections. We came up with a logistic regression model using the Nationscape dataset as a sample population. This dataset describes a number of demographic factors in the United States which affect the political affiliations of voters. The dataset enabled us to develop an accurate model to predict the outcome of this year's elections.

Model

The matters of race and ethnicity are bound to have a very pronounced impact on the outcome of the 2020 elections due to the cases of racial injustice and the protests against them which have been held in the recent past. In such a charged environment, political ideals are strengthened and their impact on the distribution of votes also increases. Age and educational levels of voters also have a significant impact on the outcome of the vote. As we would expect, household income determines a person's choice in voting since it reflects a person's satisfaction with the status quo of governance. These factors had greatly pronounced impacts on the 2016 election and in the feedback obtained from the survey. As such, we came up with a model which would predict the results of the presidential election in 2020. It used these factors as the variables which affect a voter's choice of presidential candidate as follows(Collier 2000).

$$vote\_trump\ income + age + edu + ideal + race + \varepsilon$$

Model Specifications

After downloading and unzipping the files, the dataset was cleaned to eliminate such rows as the voters who were not registered since they would not be allowed to vote. The cleaning process also assigned numeric values to the educational level, the political ideals, to facilitate the fitting of the model. The cleaning process was conducted in order to reduce the dataset to only include the essential information. This would facilitate the creation of a model from the data since the factors being used were converted into numbers(Group 2020).
The model can be summarised as shown below(Hlavac 2018). The logistic model was preferred since we aimed at getting a binary output. The output would simply state whether President Trump would win the election or lose, presumably to Joe Biden(Larsen et al. 2000).

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Nov 03, 2020 - 7:55:26 AM

Table 1:

	Dependent variable:
	vote_trump
age	0.010^{***}
	(0.002)
income	0.00000^{***}
	(0.00000)
race	1.184^{***}
	(0.095)
Constant	-2.076^{***}
	(0.136)
Observations	4,051
Log Likelihood	−2,625.030
Akaike Inf. Crit.	5,258.059
Note:	*p<0.1; **p<0.05; ***p<0.01

The model was then fitted to the census data using post stratification in order to predict the outcome of the elections. The result was the probability of Donald Trump winning the presidential elections(Pacas and Sobek, n.d.).

Results

The model predicted that the probability that a person of any age, income, or race would vote for President Trump was:

```
## [1] 0.605428
```

This means that President Trump is likely to win the 2020 election. This probability accounts for the three factors of age, household income, and race. However, there are many more variables which will affect the outcome. In addition, the model assumed that the determinants of the 2020 election results are the same as the ones for the 2016 elections.

Weaknesses

It is important to note one weakness of the model as the lack of the dataset about people's political ideals in the census data. Regardless, the use of political ideals as perceived by oneself is not an accurate way to determine the values which a person believes in. This is because people are susceptible to seeing a false image of themselves.

Discussion

We conducted model diagnostics with an ANOVA test which gave the following results:

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Nov 03, 2020 - 7:55:27 AM

The analysis of variance showed a large difference between the null deviance and the residual

Table 2:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Df	3	1.000	0.000	1.000	1.000	1.000	1.000
Deviance	3	94.642	71.933	52.029	53.117	115.949	177.694
Resid. Df	4	4,048.500	1.291	4,047	4,047.8	4,049.2	4,050
Resid. Dev	4	5,423.439	123.456	5,250.059	5,383.330	5,494.965	5,533.987
Pr(>Chi)	3	0.000	0.000	0.000	0.000	0.000	0.000

deviance. The variable of `race` showed a significant contribution to the deviance, followed by `income` and `age` respectively. This is because race had the greatest impact on the probability of any person voting for trump. This means that race is the most significant factor in swaying people's votes, followed by the other two factors in the respective order.

Next Steps

Besides compensating for the weakness of this model, studies and models created to determine the potential winner of the United States Presidential election should accommodate more factors. Although this model was prudent in its methodology and in the way it backed its arguments. As such, the models created in future for the same purpose should be, at least, of the same classification. That is, they should be logistic regression models.

This code can be found on github through this link: https://github.com/Juyuhang/Problemset-3

References

Collier, Paul. 2000. "Ethnicity, Politics and Economic Performance." *Economics & Politics* 12 (3): 225–45.

Group, Voter Study. 2020. "Nationscape Data Set." https://www.voterstudygroup.org/publication/nationscape-data-set.

Hlavac, Marek. 2018. "Stargazer: Well-Formatted Regression and Summary Statistics Tables. R Package Version 5.2.2."

Larsen, Klaus, Jørgen Holm Petersen, Esben Budtz-Jørgensen, and Lars Endahl. 2000. "Interpreting Parameters in the Logistic Regression Model with Random Effects." *Biometrics* 56 (3): 909–14.

Pacas, Steven Ruggles; Sarah Flood; Ronald Goeken; Josiah Grover; Erin Meyer; Jose, and Matthew Sobek. n.d. "IPUMS Usa: Version 10.0 [Dataset] Minneapolis, Mn: IPUMS, 2020." https://doi.org/https://doi.org/10.18128/D010.V10.0.