2020 elections

Langwen Guan, Yuhang Ju, Zike Peng

10/30/2020

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.

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

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.

The model can be summarised as shown below. 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.

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.

Results

The model predicted that the probability of a person voting for President Trump was: 0.61; regardless of his or her age, income, or race.

Discussion

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

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

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.

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.