**Abstract**

There is no doubt that the most important political event taking place right now in the second half

year of 2020 is the United States Federal Election, either Donald Trump or Joe Biden, who can win

the throne in the White House remains unknown. As the COVID-19 pandemic continues, people all

over the world have never been this desperate to know the results of the 2020 United States Federal

Election, which might determine the global political and economic trends in the next several years. In

this study, the multilevel (logistic) regression with post-stratification (MRP) framework was trained

with the Nationscape survey data, then the prediction was made with American Community Survey

(ACS) national census data which is the dataset considered to be unbiased and representative. Our

work demonstrated that, Joe Biden might win the overall popular vote by 1.76%, but might lose the electoral vote, by including socio-economic, demographic, and geographic features such as gender, age groups, education and household income into the model.

The 2020 United States (US) election is coming to its final stage, and people still have no idea who will gain the key to the White House. The Guardian reported on Oct 24 2020, that the US 2020 election might have the highest voter turnout rate since 1908, which showed that this election has drawn more attentions than ever. Therefore, it is very important and interesting to forecast whether Donald Trump or Joe Biden will win this election, based on the current election polls.

There are several famous failures of federal election forecasts. In 1936, the popular magazine Literal Digest incorrectly predicted that Franklin Roosevelt would lose, based on a two-million responses mail-in survey [@Forecasting]. In the 2016 United States (US) federal election, most institutions failed to predict that Donald Trump won against Hillary Clinton. Those failed forecasts suggest that the pool of respondents can be highly biased, which would be detrimental to the results of the predictions. Hence, representative polls have been used extensively to make election predictions, where randomly sampled individuals are asked who they would like to vote for. Although this approach has been proved to be effective, it is faced with several challenges: 1) the cost of time and money is becoming enormous; 2) the response rates are keeping decreasing in the past several decades. Moreover, a trend which is worth noting is that large non-representative online surveys are much cheaper to collect and becoming more and more popular.

As a result, people are seeking alternatives, such as using non-representative polls, together with the benefits of proper statistical adjustment. With this approach, one could make accurate predictions based on faster and less expensive survey sampling methods. Hereby, the previously published method, multilevel regression with post-stratification (MRP) was adopted to forecast the federal election, using the model training dataset Nationscape survey data and the post-stratification representative census data ACS data [@surveyData; @Ipums; @Forecasting].

In this work, the multilevel logistic regression model was fitted using the `glm` function in `R`, the response variable, vote, was dichotomized as either Trump or Biden, where the Trump is the base level. At the first, the model was trained with the Nationscape survey data. The survey data was taken from the most recent week of survey, which was June 25 2020. There are 6479 person level observations in the survey data initially. Then the prediction was made with a slice of the full 2018 ACS census data, which contains 613777 person level observations. Both of the datasets were filtered by missing values and eligibility of voting. Particularly, only observations with age greater than 17 and have US citizenship eligible to vote were included. To fully take the advantages of the MRP framework, variables were all categorized and all levels of the categorical variables were matched between the two datasets. During the modeling stage, our model started with a set of variables which are known to be closely associated with voting preference, including age groups, gender, state of the voter and race [@Forecasting]. Then, three socio-economic variables including household income, employment status and education levels were incorporated into the models, and eventually the model with the lowest Akaike information criterion (AIC) was used to make the prediction. Because a lower AIC value indicates a better fit. With the ACS census data, the probability of voters voting for Biden was estimated. The popular vote is estimated as the population-level average of the probability. The vote preference was also estimated by probability cutoff 0.5, while probability greater than 0.5 means voting for Biden, otherwise Trump. With this approach, a rough estimate of electoral votes by state was generated.

Our prediction results demonstrated that Joe Biden might win the popular vote with 51.76% of all votes. However, if the voting probability was dichotomized as we did by the cutoff 0.5, Donald Trump might win the electoral votes of 29 states, including crucial swing states such as Arizona, Florida, Iowa, North Carolina, Ohio, Pennsylvania, while Joe Biden might only win swing states including Michigan and Wisconsin.

**Methods**

**Data**

The framework used in this work is multilevel regression with post-stratification (MRP). In general, a model was trained firstly using non-representative survey data, then the prediction was made with post-stratification dataset. Thus, two datasets were required by this framework and used in this work. Both of the datasets were used with permission, but sharing and uploading the datasets were not allowed.

The first dataset was survey data of US election polls, Nationscape Data, from the Voter Study Group [@surveyData]. The survey data of the June 25th 2020 wave from the Phase 2 of Nationscape Data was used in this work. Nationscape carries out surveys every week. July 18th, 2019 was the first wave that goes into the field [@surveyData]. For Phase 1 the last wave went into the field on December 26, 2019. For Phase 2 the last wave went into the field on June 25, 2020 [@surveyData].

Democracy Fund Voter Study Group is the source for used survey dataset. It carries out interviews of American voters every week throughout last year July to this year December [@surveyData]. The week of July 25 is the recent data in the dataset. The population is the electorate of the United States, and the sampling frame is a list of online respondents from a market research platform [@surveyData]. The sample is not a simple random sample [@surveyData]. On the contrary, they used purposive sampling to secure a sample, which is selecting respondents depend on their characteristics. The sample itself is constructed to be symbolic of the population regarding a certain set of characteristics [@surveyData]. The survey data is then weighted to represent the US population.

As for the mode of interview, it was run online anywhere as long as the respondent can have the access to a networked computer or mobile device [@surveyData]. The questionnaires every week are created for a 15-minute median administration time [@surveyData]. The Nationscape pilot survey is provided in alternative or in English, and the chooser can choose the language on a question-by-question basis [@surveyData]. Depending on the wave,the survey has an average yield of approximately 75% of the original invited samples [@surveyData].

The second dataset was US national census data, American Community Survey (ACS) 2018, which is person-level data as well [@Ipums]. In the U.S. census, respondents answered questionnaires sent by mail from the 1960 census onward. Particularly, the ACS 2018 is 1-in-100 national random weighted sample of the population, and the data include persons in group quarters [@Ipums].

One thing worth mentioning is that there is a data collection error in New Castle County, DE resulted in higher allocation rates for the several topics such as ancestry, commuting, disability and employment status. This did not affect our results since our final model did not contain those variables.

The target population of the census data is every resident of the United States. The sample population is all residential areas within the United States such as group homes, university dormitories and housing units, so that the sampling frame is a list of residential areas selected from the population [@Ipums]. There are two phases of sampling that has a list of all residential areas and specific non-residential buildings provided from “Master Address File” by US Census Bureau. Firstly, new samples are selected in September/October of the previous year, and then non-responding addresses are selected for personal interviewing (in January of the current year). This would decrease the percentage of non-respondents. The data collection is conducted by different approaches, including internet, mail, telephone and personal visit.

The characteristics of survey data and census data were summarized through tables (Table 1. and Table 2.). The by state vote preference of the survey data was shown in Figure 1. In general, this survey data suggests that Donald Trump is leading in crucial swing states such as Arizona, Florida and Pennsylvania, while Joe Biden is leading in swing states including Iowa, North Carolina, Ohio, Michigan and Wisconsin.

Moreover, the distribution of age groups and household income groups were compared in Figure 2 and Figure 3. With the assumption that the census data is unbiased representative data, one can tell that the age distribution and household income distribution of the survey data is different from the census data. In the census data, the largest age group is “60+”, and the household income is much less compared with the survey data, generally. In summary, it indicates that the post-stratification with census dataset is very important.

**Model**

Post-stratification is a well-known method to conduct statistical adjustment of biased samples, where the core technique is to partition the population into cells. Particularly, the cells are based on combinations of different socio-economic, demographic, and geographic features such as age groups and income levels, then the response variables are estimated within each cell, and eventually the population-level estimates are obtained through weighed aggregation of cell-level estimates where the weights are based on the relative proportion in the population [@Forecasting]. In order to generate stable and accurate cell-level estimates, a multilevel logistic regression model fitted, instead of simply taking the average within each cell. The combination of multilevel logistic regression and post-stratification is the MRP framework used in this work.

Initially, the survey data set had 6479 observations of 265 variables. The survey data was then filtered by age 18 years old and by eligibility of voting, and all null values were removed. The census data was also filtered by age 18 years old, only person older than 18 years old should be included in this study, and the citizenship was another filter criterier. All of the variables used in this dataset were categorized into groups. In this way, the requirement of the MRP framework was full-filled and all samples were grouped into small sub-cells. Several demographic, geographic, socio-economic variables which are known to be closely associated with voting preference were focused in this study [@Forecasting]. In details, variables such as state of the voter (state), education levels (education), gender, age groups (age), race/ethnicity (race), household income (household\_income) and employment status (employment) were chosen as variables of interest. The response variable was vote\_2020, which is the voting choice. Moreover, age, gender, state and race are key variables [@Forecasting; @multiDistrict]. AIC was used to decide whether to include household income, employment and education.

All multilevel logistic regression model were fitted with `glm` function in R [@R-base]. Particularly, the AIC of the full model is 5968.9, leaving out the variable household income, AIC is 5993.9, leaving out the employment status, AIC is 5970.1, and leaving out the education, AIC is 6007.5. Since the difference of AIC between the full model and model without employment is less than 2 (the empirical rule of AIC), the model without employment was chosen as the final model (Equation (2)). Specifically, we generated the cells by considering all possible combinations of gender (2 categories), race (4 categories), age (4 categories), education (4 categories), state (51 categories) and household\_income (5 categories), thus partitioning the data into 32640 sub-cells.

Results

The

**Discussions**

**What have we learned from the model?**

Some variables can be very important, discuss why

The limitation of the model, knowledge gained from the model, but be cautious of the scope of the knowledge can be applied

**Why are we doing MRP?**

Given random digit dialing (RDD) as an example, the response rates of RDD have declined to 9% in 2012 [@Forecasting].

Most people would agree that obtaining survey data from online community such as the generation of the Xbox data would be cheaper and faster compared with traditional methods such as RDD [@Forecasting].

Through post-stratification method, the final population-level estimates are actually estimated at cell-level. One could simply use the average of the response variable within each cell to obtain the cell-level estimates. However, taking average is an unbiased estimation only if the samples are randomly drawn from the population, which requires the partition of cell is fine and might result in sparsity in cells [@Forecasting]. Fortunately, the multilevel regression model can provide accurate and robust cell-level estimates, while avoiding those issues caused by taking empirical means.

As the number of variables and levels in each variable increases, the total number of cell can become enormous. Thus, it is very important to take robustness the small sub-groups estimation into consideration.

Fortunately, sparse cells can borrow information from other cells. The idea is that if some cell has very little information then it’s coefficients will be drawn from an average of those cells that are similar.

that state-specific estimates can be obtained by using a state-specific post-stratification dataset.

**The sex and gender problem in modern survey**

The Nationscape survey data gathered the information of gender, while the 2018 ACS data gathered the information of sex. It is worthwhile mentioning that sex and gender are two different things, however, in this study, we had to match sex to gender in order to make prediction with census data using survey data trained model.

While the measurements and the definition of sex and gender is continuously changing nowadays, challenges in response categories differ between sex and gender measurement.

**Limitations and future work**

As we mentioned above, the sex and gender adjustment, classification and construct might introduce biases into the framework. Another source of systematic error might be that the survey data used for modeling was collected in June. It is possible that public opinion of Trump and Biden have affected by the events bound to the election. It can therefore be assumed that this error will sooner or later influence the final result of US 2020 Presidential Election. Future work might extend the MRP framework to time series analysis area, which would be helpful in discovering trend of support rate over time.

Our final model partitioned the data into 32640 sub-cells and the census dataset to make the prediction has 466413 observations. On average, each sub-cell has about 15 samples. In terms of the number of sub-cells, the sample size about half a million might not be large enough. In our future work, a much larger census dataset could be used to make the prediction.

<https://statmodeling.stat.columbia.edu/2020/10/28/concerns-with-our-economist-election-forecast/>

In future work, a separated model can be developed to include variables mentioned in the Keys model [@Keys]. An popular idea is that the accuracy of the prediction can be improved through stacking different models together, where each model is in charge of different aspects of information, such as different sets of

<https://hdsr.mitpress.mit.edu/pub/xhgpcyoa/release/2>