Introduction

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 (Wei Wang and Gelman 2014). 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 (Tausanovitch and Vavreck 2020; Steven Ruggles and Sobek 2020; Wei Wang and Gelman 2014).

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 (Wei Wang and Gelman 2014). 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 Canadian election polls

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 (Tausanovitch and Vavreck 2020).

The second dataset was Canadian national census data,

The target population of the census data is every resident of the Canada. The sample population is all residential areas

The characteristics of survey data and census data were summarized through tables (Table 1. and Table 2.).

Table. 1: Characteristics Summary of

Table. 2: Characteristics Summary of

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 (Wei Wang and Gelman 2014). 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.

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 (Wei Wang and Gelman 2014; Rohan Alexander and Leslie 2019). AIC was used to decide whether to include household income, employment and education.

The complete model was shown here:

$$Vote \sim Bernoulli(\frac{1}{1+exp(-(a+b_i x_i + b_{education}x_{education} + b_{householdincome}x_{householdincome} + b_{employment}x_{employment}))}) \quad (1)$$

$$Vote \sim Bernoulli(\frac{1}{1+exp(-(a+b_i x_i + b_{education}x_{education} + b_{householdincome}x_{householdincome}))}) \quad (2)$$

where the a is the intercept, b representing coefficients of different variables. Particularly, b_i and x_i represents the variables and corresponding coefficients set of age, gender, state and race, which are well-known variables and are closely associated with voting preference (Wei Wang and Gelman 2014; Rohan Alexander and Leslie 2019).

All multilevel logistic regression model were fitted with `glm` function in R (R Core Team 2020). Particularly, the AIC of the full model is Since the difference of AIC between the full model and model without employment is less than 2 (the empirical rule of AIC),

Equation (1) represents the complete model, and Equation (2) represents our final model, which did not include employment status.

All work were done in `R` (version 4.0.2) (R Core Team 2020) and `Rstudio` (version 1.3.1093). `Tidyverse` (version 1.3.0) was used for data wrangling and visualization (Wickham et al. 2019). R package `forcats` (version 0.5.0) was also used for data pre-processing (Wickham 2020). There are other packages used such as `captioner`, `gridExtra`, `broom`, `Haven`, `magrittr`, `knitr`, `labelled` and `arsenal` (Alathea 2015; Hlavac 2018; Heinzen et al. 2020; Xie 2020; Wickham and Miller 2020; Auguie 2017; Robinson, Hayes, and Couch 2020; Bache and Wickham 2014; Larmarange 2020). Code are available at: https://github.com/.

Results

Discussion

What have we learnt from the model

Our results show the similar pattern that age, gender, race and state are important variables in election prediction (Wei Wang and Gelman 2014; Rohan Alexander and Leslie 2019). Detailed information of our final model fitting can be found in the **Appendix**.

Why are we using MRP

Given random digit dialing (RDD) as an example, the response rates of RDD have declined to 9% in 2012 (Wei Wang and Gelman 2014). 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 (Wei Wang and Gelman 2014).

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 (Wei Wang and Gelman 2014). 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. One strength of the MRP is that state-specific estimates can be obtained by using a state-specific post-stratification dataset.

The sex and gender problem in modern survey

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. It is important to first understand that sex, by definition, refers to a set of biological attributes in humans and animals, while gender, by definition, refers to the socially constructed roles, behaviors, expressions and identities of girls, women, boys, men, and gender

diverse people (Kennedy et al. 2020). Another problem is that sex and gender are not binary nowadays, which makes it difficult in constructing features. Thus, the authors mentioned in their paper "Our challenge increases when we consider that we are not simply moving to a more diverse way of coding sex, but instead a recognition that the construct of gender, while the same as sex assigned at birth for many, is a different construct for others." (Kennedy et al. 2020). The similar issue also exists in other variables such as race vs. ethnicity.

Weaknesses and next steps

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.

One potential limitation of the MRP model is that only categorical variables such as socio-economic, demographic, and geographic features are suitable to build sub-cells. In future work, a separated model can be developed to include variables mentioned in the Keys model (Lichtman 2020). 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 un-correlated variables. All important "Keys" mentioned in the Keys model worth exploring in future study (Lichtman 2020; Gelman 2020).

Appendix

```
# the model broom::tidy(lg_employment) # there are 50 states
# plus 1 distric survey_data$state %>% levels() # barplots
# barplot(table(survey_data$gender ) )
# barplot(table(survey_data$education ) )
# barplot(table(survey_data$employment ) )
# barplot(table(survey_data$race ) )
# barplot(table(survey_data$household_income ) )
# barplot(table(survey_data$age ) ) # doule check the
# distribution of data, categorical
# barplot(table(census_data$gender ) )
# barplot(table(census_data$race ) )
# barplot(table(census_data$household_income ) )
# barplot(table(census_data$employment ) )
# barplot(table(census_data$age ) )
# barplot(table(census_data$education ) )
```

References

Alathea, Letaw. 2015. *Captioner: Numbers Figures and Creates Simple Captions*. https://CRAN.R-project.org/package=captioner.

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bache, Stefan Milton, and Hadley Wickham. 2014. *Magrittr: A Forward-Pipe Operator for R*. https://CRAN.R-project.org/package=magrittr.

Gelman, Andrew. 2020. "Concerns with Our Economist Election Forecast." https://statmodeling.stat.columbia.edu/2020/10/28/concerns-with-our-economist-election-forecast/.

Heinzen, Ethan, Jason Sinnwell, Elizabeth Atkinson, Tina Gunderson, and Gregory Dougherty. 2020. *Arsenal: An Arsenal of 'R' Functions for Large-Scale Statisticalsummaries*. https://CRAN.R-project.org/package=arsenal.

Hlavac, Marek. 2018. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Central European Labour Studies Institute (CELSI). https://CRAN.R-project.org/package=stargazer.

Kennedy, Lauren, Katharine Khanna, Daniel Simpson, and Andrew Gelman. 2020. "Using Sex and Gender in Survey Adjustment." https://arxiv.org/abs/2009.14401.

Larmarange, Joseph. 2020. *Labelled: Manipulating Labelled Data*. https://CRAN.R-project.org/package=labelled.

Lichtman, Allan. 2020. "The Keys to the White House: Forecast for 2020." https://doi.org/10.1162/99608f92.baaa8f68.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Rohan Alexander, Patrick Dumount, and Patrick Leslie. 2019. "Forecasting Multi-District Elections." https://github.com/RohanAlexander/ForecastingMultiDistrictElections/blob/master/outputs/paper/paper.pdf.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS Usa: American Community Survey 2018." https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "Nationscape Data Set." https://www.voterstudygroup.org/downloads?key=84a68f37-86ac-4871-b68a-a57b4d9e31d2.

Wei Wang, Sharad Goel, David Rothschild, and Andrew Gelman. 2014. "Forecasting Elections with Non-Representative Polls." https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf.

Wickham, Hadley. 2020. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Evan Miller. 2020. *Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*.