

TBD\*  
TBD

Arjun Dhatt, Benjamin Draskovic, Yiqu Ding, Gantavya Gupta

31 October 2020

### Abstract

First sentence. Second sentence. Third sentence. Fourth sentence.

## 1 Introduction

## 2 Data

## 3 Model

The model we use to predict the result of the election is MRP, multilevel regression with post-stratification. The advantage of this model is that it has better performance when estimating behaviors of a particular subgroup of the population. One can interpret MRP as a combination of multilevel logistic regression and post-stratification. Logistic regression allows us the simplicity of analyzing the categorical response variable while incorporating both numerical and categorical explanatory variables; while combined with post-stratification, it yields a much narrower confidence interval for the estimation.

We are interested in how variables such as age, education, race, family income, geographic region, and mask-wearing decision affect citizens' voting intentions, particularly between the major party candidates, Donald Trump for the Republicans and Joe Biden for the Democratic Party. The response variable that we are interested in is the chances of a citizen voting for Trump in the coming presidential election. It is a categorical variable with two levels: 1 represents intending to vote for Donald Trump, and 2 means planning to vote for Joe Biden.

Ideally, we ought to run a pre-analysis to determine the variables of interest. However, we will leave that open for further analysis and future elections due to the limited time and budget. We chose the explanatory variables based on our understanding of a respondent's significant characteristics that could potentially affect his/her voting intention. Except for respondents' age, education, income level, and state (common in statistical studies), we are very interested in the respondent's habits towards mask-wearing. Firstly because the US is the country with the most COVID cases globally, the influence of COVID on the electoral votes is inevitable. Secondly, since Trump and Biden hold quite opposite opinions towards handling the pandemic, this parameter will likely represent part of the respondent's voting intention in the same way as the more familiar political indicators.

First, we train a multilevel logistic regression model using raw data from the voter study group towards the variable of interest using the seven explanatory variables. We then apply this model to our post-stratification data set to get the estimates.

---

\*Code and data are available at: <https://github.com/STA304-PS4/Trump-vs-Biden>.

Logistic regression estimates  $\beta_0 \dots \beta_k$  in (1)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (1)$$

where  $p$  is the probability of event  $A$  that we are interested in,  $\beta_0$  is the intercept,  $x_1 \dots x_K$  are our variables of interest and  $\beta_1 \dots \beta_k$  are parameters for each of these variables. Based on the result, we are able to estimate  $p$  for a particular case given all the variables. We use `as.factors()` to incorporate dummy variables for all the categorical variable.

The main logic behind post-stratification is that we divide the sample data into strata based on a few attributes such as state, income, education level. Then we come up with a weight for each stratum to adjust its influence towards our prediction. Namely, if a stratum is over-represented, we want to reduce its impact on the prediction result; if a stratum is under-represented, we want to increase its influence. In this sense, the combination with logistic regression allows subgroups with a relatively small size to ‘reference’ from other subgroups with similar characteristics.

To mimic the population as closely as possible, we need a post-stratification dataset large enough to refer to. In this report, we use the ACS(reason).

The post-stratification estimate is defined by  $\hat{y}_{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$  where  $\hat{y}_j$  is the estimate of  $y$  in cell  $j$ , and  $N_j$  is the size of the  $j$ -th cell in the population. An illustration using the education variable is  $\hat{y}_{edu}^{PS} = \frac{\sum_{j \in J_{edu}} N_j \hat{y}_j}{\sum_{j \in J_{edu}} N_j}$ , we can get an estimation given any level of education of respondents,  $J_{edu}$  denotes all subsets of education levels,  $\cup_{i \in J} J_i^{edu} = \text{all possible education levels}$ .

(1) produces a proportion for the post-stratification based on regression, instead of merely averaging the sample in each stratum. Specifically, all possible combinations of cells are from combining:

- 
- 
- 
- 
- 
- 

We fit the logistic regressions for estimating candidate support in each cell. We used `function` from `package` to fit (2).

$$P(Trump) = \text{logit}^{-1}(\beta_0 + \beta_{age} + \beta_{j[i]}^{edu} + \beta_{j[i]}^{race} + \beta_{j[i]}^{income} + \beta_{j[i]}^{state} + \beta_{j[i]}^{mask}) \quad (2)$$

(2) is the model that we fit using the survey data, with  $\beta_{j[i]}^{var} \sim N(0, \sigma_{var}^2)$ .  $\beta_0$  is the intercept, each of  $\beta_{j[i]}^{var}$  represents the parameter for the  $i$ -th respondent in  $j$ -th cell. For example,  $\beta_{j[i]}^{state}$  can take values from  $\beta$  for all states in the US. Then the model sums up estimates for each cell to the population.

## **4 Results**

## **5 Discussion**

### **5.1 First discussion point**

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

### **5.2 Second discussion point**

### **5.3 Third discussion point**

### **5.4 Weaknesses and next steps**

#### **5.4.1 1: Nested Bayesian Model**

#### **5.4.2 2: Pre-analysis**

# Appendix

## 6 References