Forecasting 2020 US Presidential Election Result

Proportion of US Population that would vote for Donald Trump/Joe Biden

Ke Deng, Yongpeng Hua, Qihui Huang, Qing Wen

2 November 2020

Model

The interest of our study is to predict the proportion of people who would vote for Donald Trump along with the proportion of people who would vote for Joe Biden in the upcoming 2020 presidential election. To accomplish this, We would adopt two logistic regression models with the binary response variable indicating people's will whether to vote for Trump/Biden. The model is built using R(R Core Team 2020), and we use packages: tidyverse(Wickham et al. 2019), broom(Robinson, Hayes, and Couch 2020) and gridExtra(Auguie 2017).

In addition, we are using the multi-regression post-stratification(MRP) technique to arrive at the estimate of our interest. A detailed procedure would be described in the subsection: Post-Stratification. Data sets used are the survey data collected by the Voter Study Group(Tausanovitch and Vavreck 2020) and the census data(Steven Ruggles and Sobek 2020) collected by the IPUMS USA.

Model Specifics

As mentioned previously, we would adopt a logistic regression model to find the estimate for the proportion of voters who will vote for our two candidates. We are specifically choosing variables of *employment status*, *gender*, *race*, *household income*, *education level*, and *age group*.

The model is defined by:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 * x_{employment} + \hat{\beta}_2 * x_{gender} + \hat{\beta}_3 * x_{race} + \hat{\beta}_4 * x_{income} + \hat{\beta}_5 * x_{education} + \hat{\beta}_6 * x_{age}$$

with each $x_{variable}$ representing corresponding predictor variable, and each $\hat{\beta}_i$ representing the relative change in log odds as the value for the predictor variable increases by an additional unit. Furthermore, $log(\frac{\hat{p}}{1-\hat{p}})$ is the defined log odds rather than probability, so transformation on the log odds would be needed to find the corresponding $\hat{p}$.

Post-Stratification

Obviously, it would be hard to obtain a census data on how each US citizen is going to vote in the election, as the data collecting process would be extremely costly and time-consuming. With the sample data collected which consists of 6,479 observations, we want to form a representative sample to yield a meaningful analysis that could apply well to the general population. The sample size we obtained is really small compared to the total population of US citizens. To compensate for the potential non-representativeness, we want to adopt the post-stratification technique.

We intend to choose the number of variables so that we can obtain a sample as representative as possible meanwhile not over-complicating the model. So we limit the number of variables to 6 but they are all sound categories that could serve to split the population and produce meaningful results. For example, the *gender* and *age group* are both basic information pertaining to a certain individual. Variables like *employment status* and *education level* contain more specific information(i.e. more unique features to define a person's demographic than basic information does). Thus, we use the selected variables to split the cells.

What post-stratification allows us to do is that we can partition the population data into lots of different demographic cells using the selected variables and use the model built from the survey data to estimate the response variable in each cell(we would be using the logistic model as described in the previous section). Then, by weighing each cell by its relative proportion to the whole population, we could yield an estimate for the total proportion of US citizens voting for either Trump and Biden. The post-stratification estimate is defined as :

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$$

, where $\hat{y}_j$ is the estimate in each cell and N_j is the population size in the jth demographic cell. We will use $\hat{y}^{PS}_{Trump}$ to denote the proportion estimate for voting for Trump and $\hat{y}^{PS}_{Biden}$ to denote the proportion estimate for voting for Biden.

Notes on Data Cleaning Process

We use Lablled(Larmarange 2020) to help clean the data. We create two binary variables in the survey data, namely *vote_trump* and *vote_biden* in the survey data. Taking *vote_trump* for instance, we mutate all the responses for *vote_2020* that is "Donald Trump" to have the value 1, and 0 otherwise. The same procedure applied to *vote_biden*. One thing to notice is that in the *vote_2020*, there are responses of "I am not sure/don't know", which we also signed a value 0.

In order to let the model built on the survey data successfully yield desired estimates, we need to clean up both the survey data and census data so that each variable we choose contains the same levels of categorical values. For instance, we would like to have the *gender* variable in both the survey and census data to have two levels, namely *male* and *female*. So we would need to ensure that this is indeed the case in both data sets, so the model built from the survey data is being able to predict the estimates based on the input of census data.

However, when merging groups to match the groups in the other data set, there is an inevitable loss of information. We are trying to complete this cleaning process with the reservation of important details the variables contain. Below, we mention some of the cleaning processes when we decide to merge some groups together, with detailed reasoning outlined in the Appendix - Notes on Data Cleaning Process-detailed:

1. For *education*, we combine some of the small groups(population) into larger groups(sample) instead of keeping all the small groups. We end up with having 11 levels for this variable.
2. Considering variable *race_ethnicity*, voters with different races may affect their voting decisions. We combine some small groups(sample) of variables into larger groups(population), and we end up having 7 different levels.
3. Considering *age_group* as a possible factor, people at different ages with different experiences might affect their voting choices, and we end up with 8 groups.
4. For *employment*, we have combined the more specific divisions (sample) into a more general division (population). And we end up having 3 levels for this variable.

Results

Note: Here we are showing the first few rows of the regression output. For the complete tables of output, please refer to the Appendix for table 1 and table 2.

Table 1: Logistic regression output for predicting proportion of votes for Donald Trump

```
## # A tibble: 6 x 5
##   term                                      estimate std.error statistic  p.value
##   <chr>                                        <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                                -0.181    0.674     -0.268 7.88e- 1
## 2 as.factor(employment)not in labor force    -0.153    0.0732    -2.09  3.62e- 2
## 3 as.factor(employment)unemployed            -0.142    0.102     -1.39  1.63e- 1
## 4 as.factor(gender)Male                       0.362    0.0581     6.23  4.64e-10
## 5 as.factor(race_ethnicity)black/african ~   -1.81     0.265     -6.82  9.14e-12
## 6 as.factor(race_ethnicity)chinese           -1.30     0.397     -3.26  1.10e- 3
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.390
```

Table 1 shows the model output for voting Trump prediction. We get the coefficients for the regression equation, including the parameters of interest like intercepts and slopes for each factor). The estimated intercept for the regression equation here is -0.1809. Estimated values for slope parameters vary based on different fields and different cells.

Male increases the log odds while the female does not have contributions in this cell. Also, unemployed citizens lowered the result as well. In race and ethnicity fields, races except white have lowered the expected change in log odds of Trump winning the election; other factors like education and employment have lowered the log odds as well. People from different age groups contribute to the log odds of Trump winning the election greater or lesser.

One thing is worth mentioning is that not all classes contribute to the expected log odds of winning, the income of a family only between 200,000 to 249,999 U.S. dollars, 250,000 U.S. dollars above contribute, the others somewhat decrease the log odds. Finally, by extracting the probability from log odds and aggregating estimates from each cell. The estimated probability of Trump winning the election is 0.3897.

Table 2: Logistic regression output for predicting proportion of votes for Joe Biden

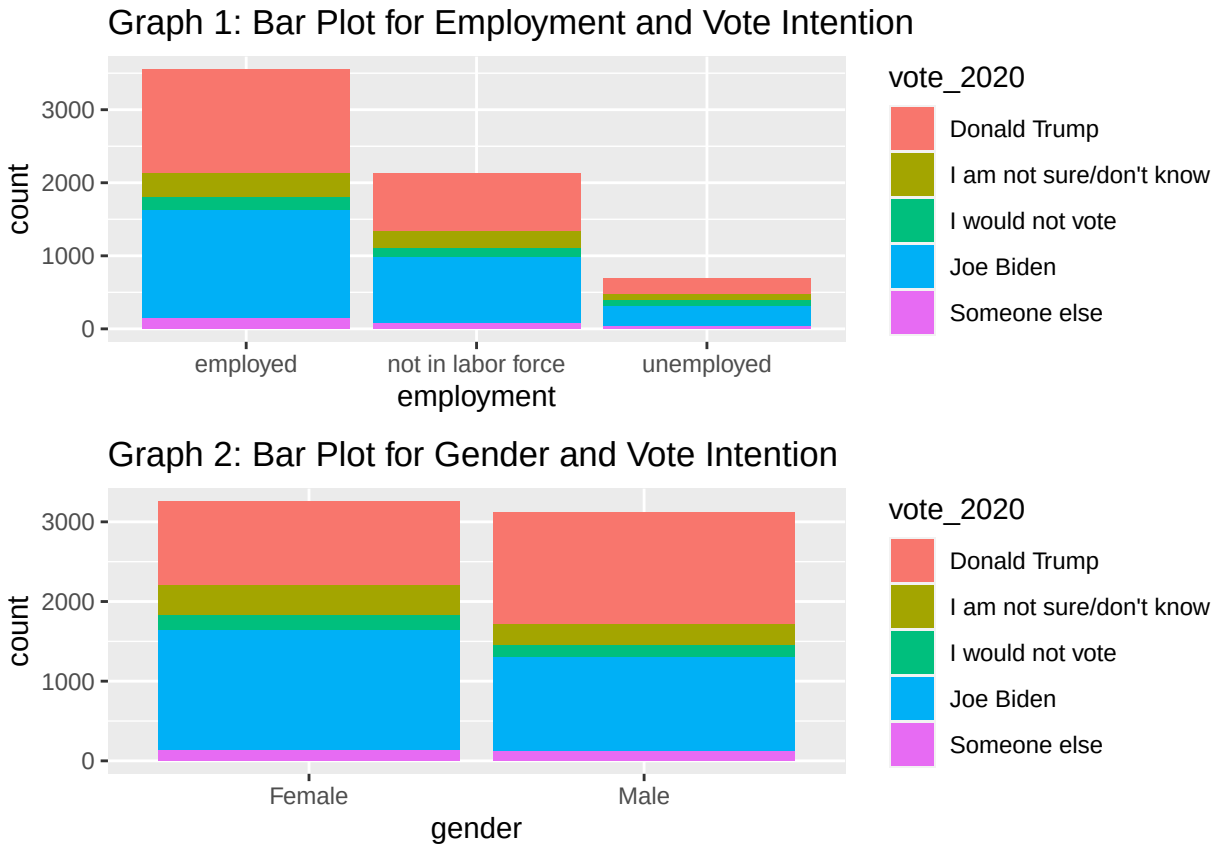
```
## # A tibble: 6 x 5
##   term                                      estimate std.error statistic  p.value
##   <chr>                                        <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                                -0.595    0.662     -0.899 3.69e- 1
## 2 as.factor(employment)not in labor force     0.110    0.0698     1.57  1.16e- 1
## 3 as.factor(employment)unemployed             0.0642   0.0957     0.670 5.03e- 1
## 4 as.factor(gender)Male                      -0.293    0.0562    -5.22  1.75e- 7
## 5 as.factor(race_ethnicity)black/african ~    1.57     0.256      6.14  8.14e-10
## 6 as.factor(race_ethnicity)chinese            0.946    0.344      2.75  6.01e- 3
```

```
## # A tibble: 1 x 1
##   alp_predict_b
##           <dbl>
## 1         0.406
```

Table 2 shows the logistic model output for voting Biden prediction. We use exactly the same factors to predict. The estimated intercept for the regression model is -0.5952. Contrary to what we got in predicting Trump's winning probability, unemployed/not in labor citizens rises log-odds output, and males decline it. Moreover, the race factor contributes but the age groups factor decreases the output. A College degree and

Master's degree is likely to increase the odds. For different income groups, only a few groups may lower the odds like income ranging from 175,000 to 199,999 U.S. dollars and above 250,000 U.S. dollars. The estimated probability of voting for Biden more than Trump a little, which is 0.4062.

The P-value for each factor in both logistic models is less than 0.005.



In Graph 1, the bar plot shows the proportion of people voting for Trump and Biden within each employment status. The precise proportion number is summarized in table3 below. Visually, within the three employment statues, voting for Trump and voting for Biden have similar proportions and small proportions for others.

Graph 2 above have shown the numbers as a visual bar plot. It has shown that in the female group, females voting for Biden have the highest proportion, then second-highest voting for Trump, with small proportions saying others. The male group has males voting for Trump being the highest proportion, then second-highest for voting for Biden, and small proportions for others.

Table 3: Proportion of people voting for Trump or Biden within each employment group

```
##           employment    vote_t    vote_b
## 1           employed 0.4015216 0.4170189
## 2 not in labor force 0.3770492 0.4290398
## 3         unemployed 0.3140376 0.4109986
```

Table 4: Proportion of people voting for Trump or Biden within each gender group

```
##   gender    vote_t   vote_b1
## 1 Female 0.3208359 0.4631223
## 2   Male 0.4495354 0.3758411
```

Table 3 shows that out of all employed people, 0.40 or 40% of them vote for Trump and 0.417 or 41.7% vote for Biden, and the other divisions are shown in Graph 1 (Bar plot for Employment and Vote Intention). Out of all people not in the labor force, 0.377 votes for Trump and 0.429 votes for Biden. Lastly out of all unemployed, 0.314 votes for Trump and 0.411 votes for Biden.

The precise proportion of people within each gender group voting for Trump and Biden are summarized in table 4. Out of all females, 0.32 (or 32%) vote for Trump and 0.46 vote for Biden. Within the males, 0.45 vote for Trump and 0.38 vote for Biden.

Discussion

Summary

In this study, we focus on the MRP technique to produce an estimate for the proportion of voters intending to vote for Donald Trump or for Joe Biden in the upcoming 2020 presidential election. We start off by reading in and cleaning the survey and census data to ensure that the variables in both data sets contain the same sets of levels for our model to yield prediction. After that, we fit in two logistic regression models to the survey data with the predictor and response variables identified previously. Lastly, we apply this model to the census data which already consists of split demographic cells to yield estimates for the proportion of voters in each cell and aggregate our calculation to the population level to obtain the final estimate. A quick glimpse of what the cells look like eventually is as follows:

```
## # A tibble: 6 x 11
##   gender household_income employment age_group race_ethnicity education     n
##   <chr>  <chr>            <chr>      <chr>     <chr>          <chr>     <dbl>
## 1 Female $100,000 to $12~ employed   18 to 27  american indi~ College ~     1
## 2 Female $100,000 to $12~ employed   18 to 27  black/african~ Associat~     1
## 3 Female $100,000 to $12~ employed   18 to 27  black/african~ College ~     4
## 4 Female $100,000 to $12~ employed   18 to 27  black/african~ Complete~     3
## 5 Female $100,000 to $12~ employed   18 to 27  black/african~ Doctorat~     1
## 6 Female $100,000 to $12~ employed   18 to 27  black/african~ High sch~     3
## # ... with 4 more variables: logodds_estimate <dbl>, estimate <dbl>,
## #   logodds_estimate_b <dbl>, estimate_b <dbl>
```

We finally arrived at two estimates for the proportion of voters: $\hat{y}^{PS}_{Trump} = 0.39$ and $\hat{y}^{PS}_{Biden} = 0.41$.

Conclusion

The building of this model gives society a timely and social effect. As in the political field, voting for the two candidates can be a popular subject and news. So we build models to roughly predict the result. Predicting people's choices can be a tricky process. However, with the given data and a properly built model, the prediction will help people see how their choices potentially affect the elections and notice the importance of personal choices. Even the data used in the model are not sufficiently representative and are not collected by ourselves, it can let society have an idea of what the predicted result is. The social effect of such a model is great and relevant.

From the model we generate, it gives us a clear view of the advantages and disadvantages in different fields of two U.S. presidential candidates by looking at the log-odds from the coefficient table. Some of the categories like *races* and *employment* do impose negative effects on the expected log odds in Trump's election model while increasing the log odds in Biden's model. Variable *gender* seems to increase log odds for Trump while decreasing log odds for Biden. Therefore, we make bar plots for two of the previous factors, namey *gender* and *employment*. As shown in table 3 and 4 in the result section, *males* are more likely to vote for Trump

while *females* are more likely to vote for Biden. On the other hand, people in either *employed*, *unemployed*, or *not in labor force* all have a higher tendency to vote for Biden than for Trump.

It is hard to come up with a pattern in other categories like household income and age groups, as these cells have brought the log odds up or down in a small amount, so we could not see a huge difference between these two categories. Also, the bar plots showing the precise proportion of citizens' voting intention help us confirm our previous result, and give us a detailed view of the difference of proportions in some categories.

The probability of winning the election does not have much difference between the two candidates, but Biden has higher possibilities of winning the election based on the census and survey data we have (39% compared to 41% for Trump). There are a lot of possible factors affecting people's choice this year, one guess is that ability to handle the COVID-19 period, whether a rapid pandemic response or is made, can either hold or change the public opinions. Also, disparities in races, like the anti-racism protests lasted two months from May, may be taken into consideration as well.

Weaknesses

One weakness that can be identified in our model is that due to the unmatch between the census data and the sample data, many variables that contain detailed information are being merged into smaller and more general divisions, which can cause a loss of information. For example, there are two variables in the census data that recorded the detailed employment status and more general employment status. We chose the more general variable due to the mismatch between the detailed employment status in the census and the sample data.

Next Steps

Going from the weakness section, the merging of groups within variables can make the model less informative. The next step to make a better model is to see if there are data sets (sample data and census data) that have variables that include more similar group divisions than the currently used data sets.

Also, we included 6 variables in our analysis to split the demographic cells. The result we get is sound and solid. However, if more potential variables can be included in this model, the cells could be more inclusive and yield a even more convincing result, since it helps further classify the population.

Concluding Remark

2020 has been a quite special year for a range of different reasons. The Coronavirus pandemic is certainly one of them, and it is still an ongoing issue. This pandemic has changed everything including how people live their lives and how to cope with any emerging challenges faced by them. With this situation in mind, the dynamic for the upcoming 2020 US presidential election might also alter, as Lichtman (2020) mentioned in his journal. As people are currently more focusing on the issue of public health, health care, and racial inequality, they may put more emphasis on the traits they want their president to have instead of focusing more on the economic perspective of the country. The vote result may well turn out to be a vote for supporting whoever the citizens think as the most trustworthy to lead them through the pandemic.

Appendix

Table 1: Complete result for table 1 in the result session

```
## # A tibble: 50 x 5
##    term                                  estimate std.error statistic  p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                            -0.181     0.674     -0.268 7.88e- 1
##  2 as.factor(employment)not in labor force -0.153    0.0732    -2.09  3.62e- 2
##  3 as.factor(employment)unemployed        -0.142     0.102     -1.39  1.63e- 1
##  4 as.factor(gender)Male                   0.362     0.0581     6.23  4.64e-10
##  5 as.factor(race_ethnicity)black/african~ -1.81     0.265     -6.82  9.14e-12
##  6 as.factor(race_ethnicity)chinese       -1.30      0.397     -3.26  1.10e- 3
##  7 as.factor(race_ethnicity)japanese      -0.874     0.627     -1.39  1.63e- 1
##  8 as.factor(race_ethnicity)other asian o~ -0.421    0.280     -1.51  1.32e- 1
##  9 as.factor(race_ethnicity)other race, n~ -0.623    0.258     -2.41  1.59e- 2
## 10 as.factor(race_ethnicity)white          0.104     0.234      0.444 6.57e- 1
## 11 as.factor(household_income)$125,000 to~ -0.0970   0.152     -0.639 5.23e- 1
## 12 as.factor(household_income)$15,000 to ~ -0.623    0.164     -3.79  1.51e- 4
## 13 as.factor(household_income)$150,000 to~ -0.176    0.186     -0.946 3.44e- 1
## 14 as.factor(household_income)$175,000 to~  0.267    0.219      1.22  2.24e- 1
## 15 as.factor(household_income)$20,000 to ~ -0.229    0.160     -1.43  1.52e- 1
## 16 as.factor(household_income)$200,000 to~  0.480    0.205      2.34  1.91e- 2
## 17 as.factor(household_income)$25,000 to ~ -0.387    0.159     -2.44  1.46e- 2
## 18 as.factor(household_income)$250,000 an~  0.218    0.211      1.03  3.01e- 1
## 19 as.factor(household_income)$30,000 to ~ -0.431    0.157     -2.75  6.02e- 3
## 20 as.factor(household_income)$35,000 to ~ -0.445    0.164     -2.71  6.68e- 3
## 21 as.factor(household_income)$40,000 to ~ -0.468    0.171     -2.73  6.30e- 3
## 22 as.factor(household_income)$45,000 to ~ -0.308    0.163     -1.89  5.87e- 2
## 23 as.factor(household_income)$50,000 to ~ -0.248    0.155     -1.60  1.09e- 1
## 24 as.factor(household_income)$55,000 to ~ -0.158    0.192     -0.823 4.10e- 1
## 25 as.factor(household_income)$60,000 to ~ -0.445    0.192     -2.32  2.03e- 2
## 26 as.factor(household_income)$65,000 to ~ -0.303    0.212     -1.43  1.53e- 1
## 27 as.factor(household_income)$70,000 to ~ -0.284    0.185     -1.53  1.25e- 1
## 28 as.factor(household_income)$75,000 to ~ -0.115    0.184     -0.625 5.32e- 1
## 29 as.factor(household_income)$80,000 to ~ -0.491    0.224     -2.19  2.86e- 2
## 30 as.factor(household_income)$85,000 to ~ -0.355    0.243     -1.46  1.44e- 1
## 31 as.factor(household_income)$90,000 to ~ -0.164    0.260     -0.631 5.28e- 1
## 32 as.factor(household_income)$95,000 to ~ -0.506    0.198     -2.55  1.06e- 2
## 33 as.factor(household_income)Less than $~ -0.615    0.134     -4.58  4.58e- 6
## 34 as.factor(education)Associate Degree    -0.717    0.647     -1.11  2.68e- 1
## 35 as.factor(education)College Degree (su~ -0.712    0.643     -1.11  2.68e- 1
## 36 as.factor(education)Completed some col~ -0.560    0.643     -0.871 3.84e- 1
## 37 as.factor(education)Completed some gra~ -0.709    0.658     -1.08  2.81e- 1
## 38 as.factor(education)Completed some hig~ -0.297    0.646     -0.460 6.45e- 1
## 39 as.factor(education)Doctorate degree    -0.487    0.670     -0.727 4.67e- 1
## 40 as.factor(education)High school gradua~ -0.418    0.644     -0.650 5.16e- 1
## 41 as.factor(education)Masters degree      -0.685    0.647     -1.06  2.90e- 1
## 42 as.factor(education)Middle School - Gr~ -0.598    0.797     -0.750 4.53e- 1
## 43 as.factor(education)Other post high sc~ -0.337    0.652     -0.517 6.05e- 1
## 44 as.factor(age_group)28 to 37            0.497     0.105      4.74  2.10e- 6
## 45 as.factor(age_group)38 to 47            0.672     0.106      6.34  2.31e-10
## 46 as.factor(age_group)48 to 57            0.738     0.110      6.73  1.74e-11
## 47 as.factor(age_group)58 to 67            0.830     0.110      7.53  4.91e-14
```

```
## 48 as.factor(age_group)68 to 77              0.812     0.129      6.29  3.14e-10
## 49 as.factor(age_group)78 to 87              1.28      0.240      5.32  1.04e- 7
## 50 as.factor(age_group)88+                   1.71      0.879      1.94  5.25e- 2
```

Table 2: Complete result for table 2 in the result session

```
## # A tibble: 50 x 5
##    term                                  estimate std.error statistic  p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                           -0.595     0.662    -0.899  3.69e- 1
##  2 as.factor(employment)not in labor force  0.110     0.0698    1.57   1.16e- 1
##  3 as.factor(employment)unemployed          0.0642    0.0957    0.670  5.03e- 1
##  4 as.factor(gender)Male                  -0.293     0.0562   -5.22   1.75e- 7
##  5 as.factor(race_ethnicity)black/african~  1.57      0.256     6.14   8.14e-10
##  6 as.factor(race_ethnicity)chinese         0.946     0.344     2.75   6.01e- 3
##  7 as.factor(race_ethnicity)japanese        1.65      0.596     2.77   5.65e- 3
##  8 as.factor(race_ethnicity)other asian o~  0.631     0.281     2.25   2.47e- 2
##  9 as.factor(race_ethnicity)other race, n~  0.727     0.262     2.77   5.57e- 3
## 10 as.factor(race_ethnicity)white           0.284     0.246     1.16   2.48e- 1
## 11 as.factor(household_income)$125,000 to~  0.182     0.152     1.20   2.31e- 1
## 12 as.factor(household_income)$15,000 to ~  0.279     0.157     1.77   7.59e- 2
## 13 as.factor(household_income)$150,000 to~  0.264     0.185     1.43   1.54e- 1
## 14 as.factor(household_income)$175,000 to~ -0.200     0.227    -0.881  3.78e- 1
## 15 as.factor(household_income)$20,000 to ~  0.294     0.157     1.87   6.11e- 2
## 16 as.factor(household_income)$200,000 to~ -0.450     0.215    -2.09   3.63e- 2
## 17 as.factor(household_income)$25,000 to ~  0.212     0.156     1.36   1.73e- 1
## 18 as.factor(household_income)$250,000 an~ -0.0424    0.210    -0.202  8.40e- 1
## 19 as.factor(household_income)$30,000 to ~  0.251     0.154     1.63   1.04e- 1
## 20 as.factor(household_income)$35,000 to ~  0.392     0.160     2.46   1.40e- 2
## 21 as.factor(household_income)$40,000 to ~  0.393     0.167     2.35   1.88e- 2
## 22 as.factor(household_income)$45,000 to ~  0.195     0.162     1.20   2.28e- 1
## 23 as.factor(household_income)$50,000 to ~  0.291     0.153     1.91   5.64e- 2
## 24 as.factor(household_income)$55,000 to ~  0.182     0.190     0.961  3.36e- 1
## 25 as.factor(household_income)$60,000 to ~  0.357     0.188     1.90   5.78e- 2
## 26 as.factor(household_income)$65,000 to ~  0.347     0.210     1.65   9.86e- 2
## 27 as.factor(household_income)$70,000 to ~  0.424     0.182     2.33   1.95e- 2
## 28 as.factor(household_income)$75,000 to ~  0.170     0.184     0.926  3.55e- 1
## 29 as.factor(household_income)$80,000 to ~  0.414     0.215     1.92   5.46e- 2
## 30 as.factor(household_income)$85,000 to ~  0.508     0.235     2.16   3.08e- 2
## 31 as.factor(household_income)$90,000 to ~  0.288     0.261     1.10   2.69e- 1
## 32 as.factor(household_income)$95,000 to ~  0.478     0.192     2.49   1.26e- 2
## 33 as.factor(household_income)Less than $~  0.181     0.130     1.39   1.63e- 1
## 34 as.factor(education)Associate Degree    -0.0232    0.630    -0.0368 9.71e- 1
## 35 as.factor(education)College Degree (su~  0.0814    0.626     0.130  8.97e- 1
## 36 as.factor(education)Completed some col~ -0.216     0.626    -0.346  7.30e- 1
## 37 as.factor(education)Completed some gra~  0.0110    0.640     0.0171 9.86e- 1
## 38 as.factor(education)Completed some hig~ -0.619     0.629    -0.984  3.25e- 1
## 39 as.factor(education)Doctorate degree    -0.0542    0.653    -0.0831 9.34e- 1
## 40 as.factor(education)High school gradua~ -0.577     0.627    -0.921  3.57e- 1
## 41 as.factor(education)Masters degree       0.213     0.630     0.337  7.36e- 1
## 42 as.factor(education)Middle School - Gr~ -0.597     0.769    -0.776  4.37e- 1
## 43 as.factor(education)Other post high sc~ -0.370     0.635    -0.582  5.60e- 1
## 44 as.factor(age_group)28 to 37            -0.123     0.0924   -1.33   1.84e- 1
## 45 as.factor(age_group)38 to 47            -0.258     0.0957   -2.70   7.03e- 3
```

```
## 46 as.factor(age_group)48 to 57            -0.301     0.100    -3.01   2.62e- 3
## 47 as.factor(age_group)58 to 67            -0.118     0.100    -1.18   2.39e- 1
## 48 as.factor(age_group)68 to 77            -0.0590    0.120    -0.493  6.22e- 1
## 49 as.factor(age_group)78 to 87            -0.492     0.242    -2.03   4.26e- 2
## 50 as.factor(age_group)88+                 -0.291     0.887    -0.327  7.43e- 1
```

Notes on Data Cleaning Process-Detailed:

1. Education: keeping the consistency between *census* data and *survey* data, so that they have the same number of groups in *education*; the groups in *census* are over-detailed, and some groups are repeated which means they represent the similar groups, we can combine them into a large group. In this case, we classify all 42 groups from *census* dataset into 11 large groups in *survey* dataset. For instance, **associate's degree, type not specified**, **associate's degree, occupational program** and **associate's degree, academic program** all goes into the **Associate Degree** group in *survey* dataset.

2. Race_ethnicity: the groups(*race_ethnicity*) in *survey* are over-detailed, and some groups of races can be included in a larger group. In this case, the original 15 groups are classified to 7 groups, which are *white*, *black/african american/negro*, *american indian or alaska native*, *chinese*, *japanese*, *other asian or pacific islander*, and *other race, nec.* In addition, the variable *race_ethnicity* is renamed to *race* in *census* dataset and also in the model.

3. Age group: *age_group* is a variable mutated from ages. In the sample data, it only records age groups that are over 18 years old. However, the population data consists of those that are under 18 and are not eligible to vote, which is being filtered out for the model.

4. Employment: in the sample data, the people with jobs are considered to be *employed*, including groups that are *Part-time employed*, *Full-time employed*, *Self-employed*. The other general division is people that are *not in labor force*, which in the sample are groups that are *Homemaker*, *Permanently disabled*, *Retired*, and *Student*. People who are *Unemployed or temporarily on layoff* are considered to be *unemployed*. One note is that those that have employment noted as *Other*, are grouped *n/a*. In the population data, when people under 18 years old are being filtered out, we have found out that those that have employment as *n/a* are those that are under 18 years old, so it does not cause any trouble in the model.

References

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

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*. Harvard Data Science Review. 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.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. *2018 1-Year Acs*. Minneapolis, USA: IPUMS USA. https://doi.org/10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. *Ns20200605*. Democracy Fund + UCLA Nationscape. https://www.voterstudygroup.org/publication/nationscape-data-set.

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.