Joe Biden, Another Hillary Clinton? 2020 US Federal Election[*]

Multilevel Regression with Post-stratification (MRP) Framework Made the Prediction

Xi Cheng, Guangyu Du, Shichao Feng and Zhitong Liu

November 2, 2020

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 lose the electoral vote, by including socio-economic, demographic, and geographic features such as gender, age groups, education and household income into the model.

Keywords: forecasting; Trump; Biden; multilevel regression with post-stratification (MRP); US 2020 Election;

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

[*] Code are available at: https://github.com/Chelsea-Cheng99/STA304/tree/master/2020USElection

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 US election polls, Nationscape Data, from the Voter Study Group (Tausanovitch and Vavreck 2020). 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 (Tausanovitch and Vavreck 2020). 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 (Tausanovitch and Vavreck 2020).

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 (Tausanovitch and Vavreck 2020). 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 (Tausanovitch and Vavreck 2020). The sample is not a simple random sample (Tausanovitch and Vavreck 2020). 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 (Tausanovitch and Vavreck 2020). 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 (Tausanovitch and Vavreck 2020). The questionnaires every week are created for a 15-minute median administration time (Tausanovitch and Vavreck 2020). The Nationscape pilot survey is provided in alternative or in English, and the chooser can choose the language on a question-by-question basis (Tausanovitch and Vavreck 2020). Depending on the wave,the survey has an average yield of approximately 75% of the original invited samples (Tausanovitch and Vavreck 2020).

The second dataset was US national census data, American Community Survey (ACS) 2018, which is person-level data as well (Steven Ruggles and Sobek 2020). 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 (Steven Ruggles and Sobek 2020).

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 (Steven Ruggles and Sobek 2020). 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.

Table. 1: Characteristics Summary of the 2020-06-25 Nationscape Survey Data

	Overall (N=4741)
Age	
18-29	831 (17.5%)
30-44	1516 (32.0%)
45-55	1081 (22.8%)
60+	1313 (27.7%)
Gender	
female	2311 (48.7%)
male	2430 (51.3%)
Race/Ethnicity	
African American	538 (11.3%)
Asian	198 (4.2%)
Other	356 (7.5%)
White	3649 (77.0%)
Education	
High school and below	1073 (22.6%)
Some college	1207 (25.5%)
College	1603 (33.8%)
Graduate level	858 (18.1%)
Household Income	
Less than $29,999	1333 (28.1%)

	Overall (N=4741)
$30,000 to $69,999	1538 (32.4%)
$70,000 to $99,999	689 (14.5%)
$100,000 to $149,999	660 (13.9%)
$150,000 and above	521 (11.0%)
Employment Status	
employed	2760 (58.2%)
not in labor force	1551 (32.7%)
unemployed	430 (9.1%)

Table. 2: Characteristics Summary of the 2018 ACS Census Data

	Overall (N=467436)
Age	
18-29	84670 (18.1%)
30-44	97767 (20.9%)
45-55	113848 (24.4%)
60+	171151 (36.6%)
Gender	
female	241471 (51.7%)
male	225965 (48.3%)
Race/Ethnicity	
African American	45847 (9.8%)
Asian	20234 (4.3%)
Other	27546 (5.9%)
White	373809 (80.0%)
Education	
High school and below	179824 (38.5%)
Some college	105499 (22.6%)
College	127321 (27.2%)
Graduate level	54792 (11.7%)
Household Income	
Less than $29,999	244411 (52.3%)
$30,000 to $69,999	138729 (29.7%)
$70,000 to $99,999	39475 (8.4%)
$100,000 to $149,999	24607 (5.3%)
$150,000 and above	20214 (4.3%)
Employment Status	
employed	268326 (57.4%)
not in labor force	186547 (39.9%)
unemployed	12563 (2.7%)

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

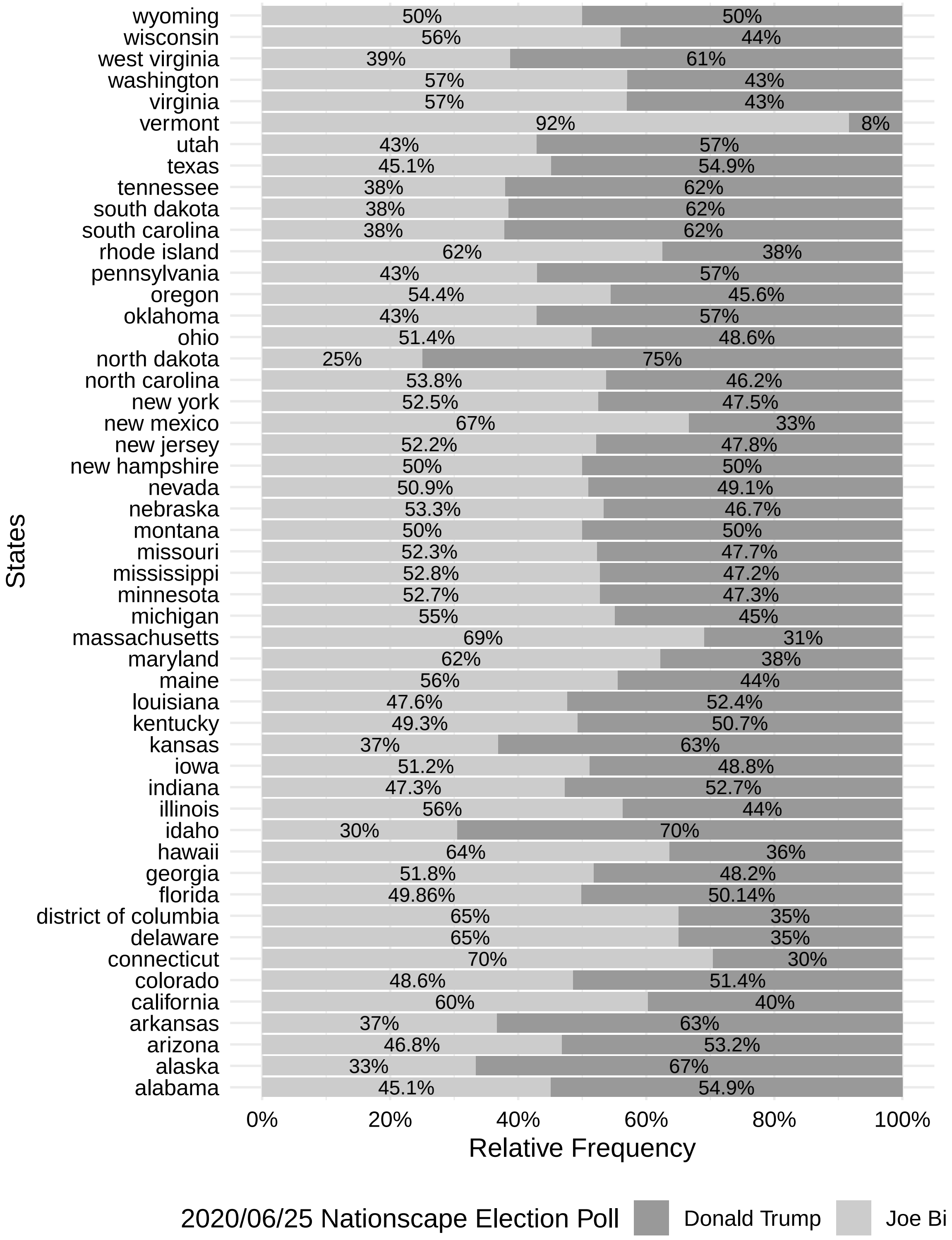
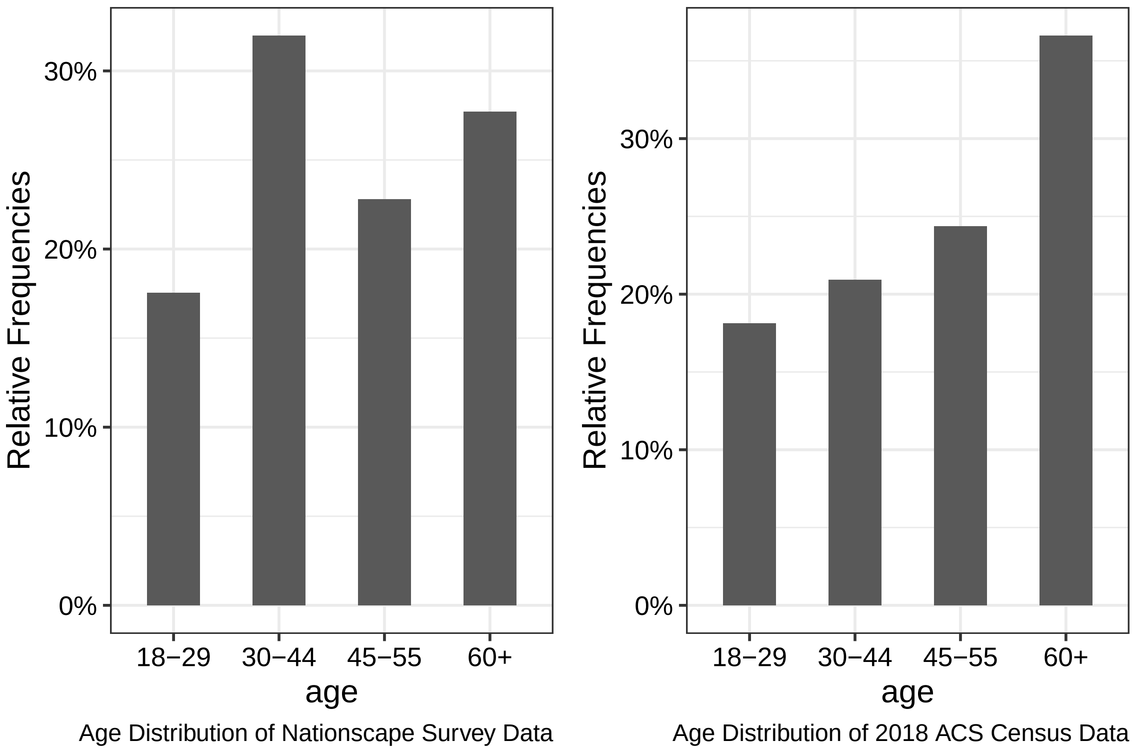


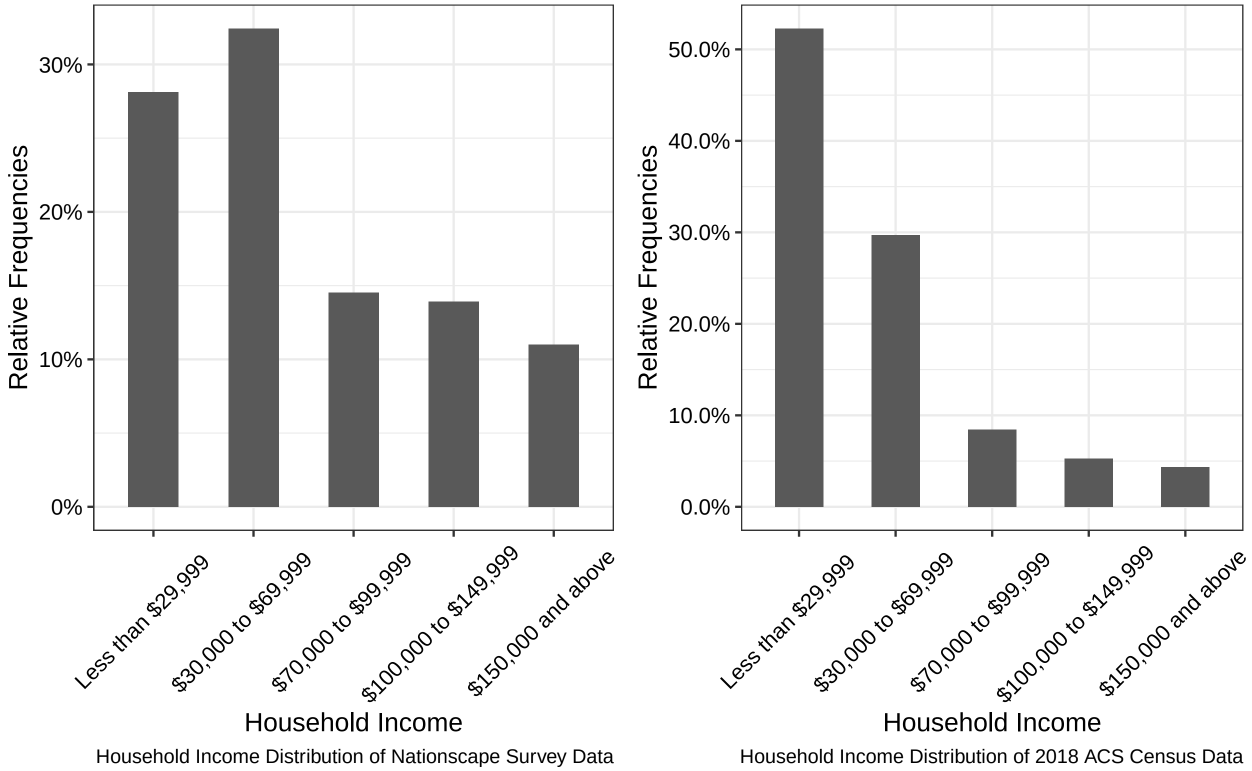
Figure 1: 2020-06-25 Nationscape Election Poll by State



Age Distribution of Nationscape Survey Data

Age Distribution of 2018 ACS Census Data

Figure 2: Comparison of Age Distributions Between Two Datasets



Household Income Distribution of Nationscape Survey Data

Household Income Distribution of 2018 ACS Census Data

Figure 3: Comparison of Household Income Distributions Between Two Datasets

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 (Wei Wang and Gelman 2014). 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:

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

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

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 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.

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/Chelsea-Cheng99/STA304/tree/master/2020USElection.

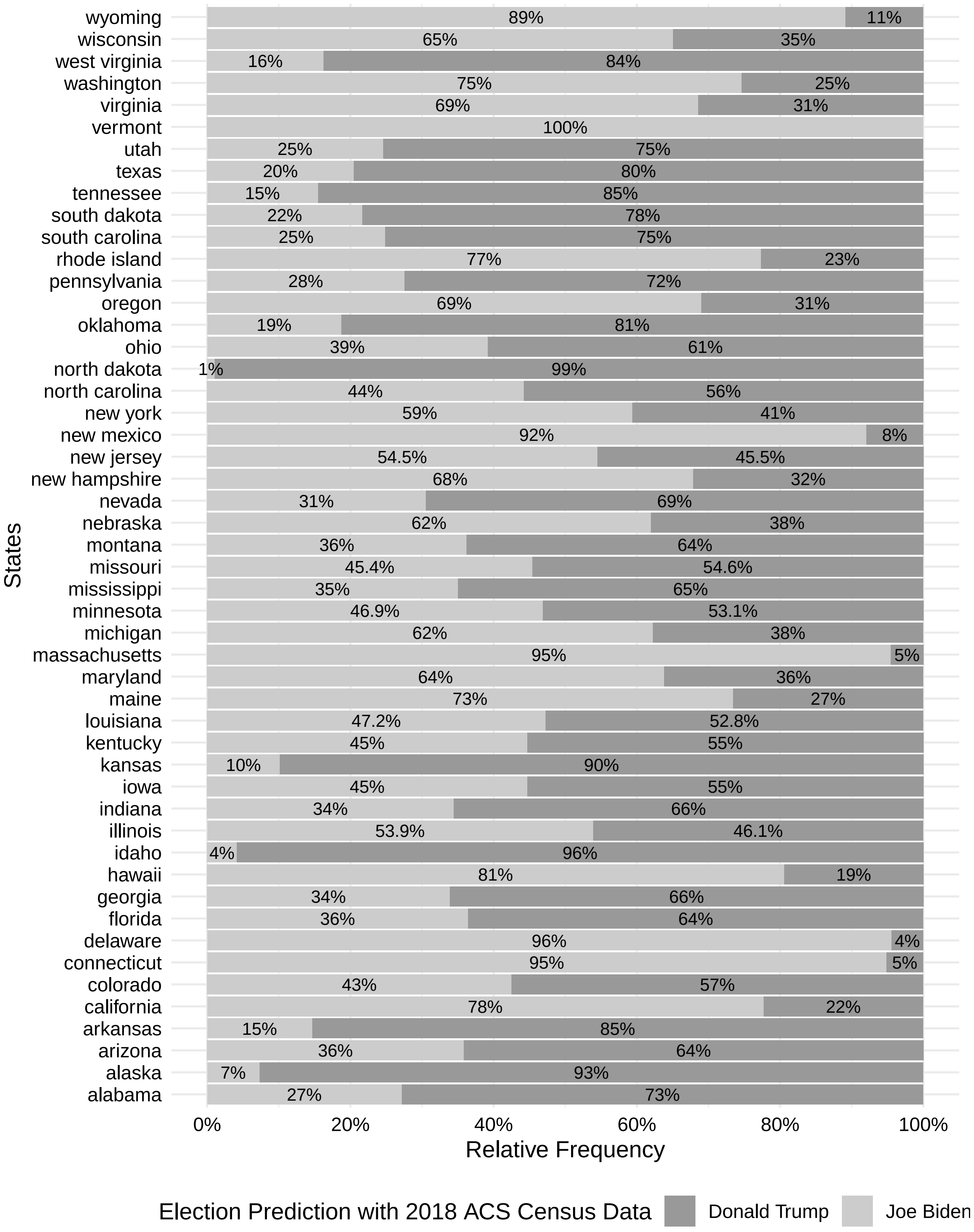


Figure 4: Election Forecast with 2018 ACS Census Data by State

Results

Discussion

First discussion point

Second discussion point

Third discussion point

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.

Appendix

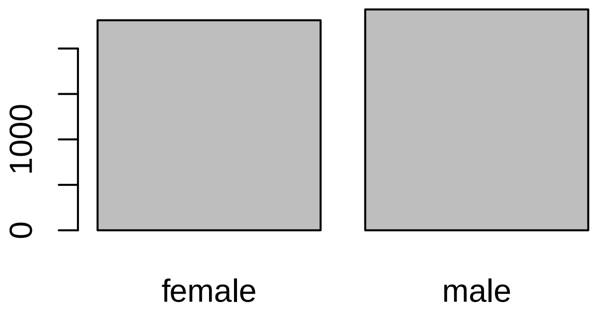
```r
# the model
broom::tidy(lg_employment)
```

```
## # A tibble: 65 x 5
##    term        estimate std.error statistic  p.value
##    <chr>          <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)   1.76      0.893     1.97    4.88e- 2
##  2 age30-44     -0.577     0.101    -5.73    1.02e- 8
##  3 age45-55     -0.751     0.106    -7.10    1.23e-12
##  4 age60+       -0.590     0.103    -5.70    1.21e- 8
##  5 gendermale   -0.445     0.0650   -6.85    7.18e-12
##  6 stateidAL     0.377     0.918     0.410   6.82e- 1
##  7 stateidAR     0.0473    0.970     0.0487  9.61e- 1
##  8 stateidAZ     0.686     0.900     0.763   4.46e- 1
##  9 stateidCA     1.21      0.886     1.37    1.71e- 1
## 10 stateidCO     0.871     0.916     0.951   3.42e- 1
## # ... with 55 more rows
```

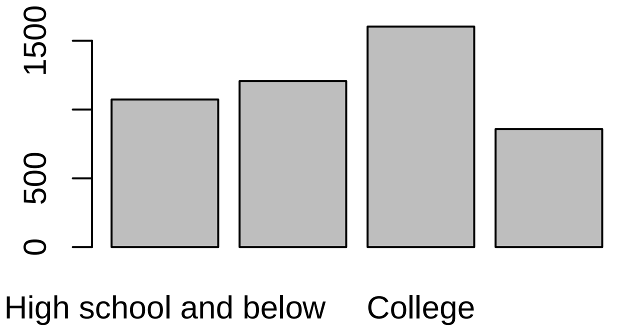
```r
# there are 50 states plus 1 distric
survey_data$state %>% levels()
```

```
##  [1] "alabama"              "alaska"               "arizona"             
##  [4] "arkansas"             "california"           "colorado"            
##  [7] "connecticut"          "delaware"             "district of columbia"
## [10] "florida"              "georgia"              "hawaii"              
## [13] "idaho"                "illinois"             "indiana"             
## [16] "iowa"                 "kansas"               "kentucky"            
## [19] "louisiana"            "maine"                "maryland"            
## [22] "massachusetts"        "michigan"             "minnesota"           
## [25] "mississippi"          "missouri"             "montana"             
## [28] "nebraska"             "nevada"               "new hampshire"       
## [31] "new jersey"           "new mexico"           "new york"            
## [34] "north carolina"       "north dakota"         "ohio"                
## [37] "oklahoma"             "oregon"               "pennsylvania"        
## [40] "rhode island"         "south carolina"       "south dakota"        
## [43] "tennessee"            "texas"                "utah"                
## [46] "vermont"              "virginia"             "washington"          
## [49] "west virginia"        "wisconsin"            "wyoming"
```

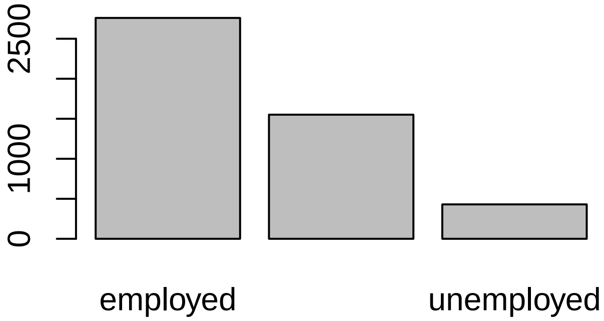
```r
# barplots
barplot(table(survey_data$gender))
```



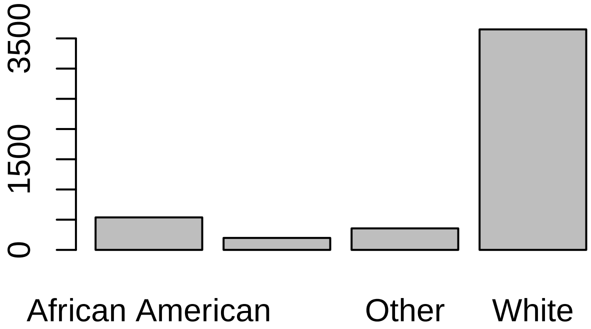
```r
barplot(table(survey_data$education))
```



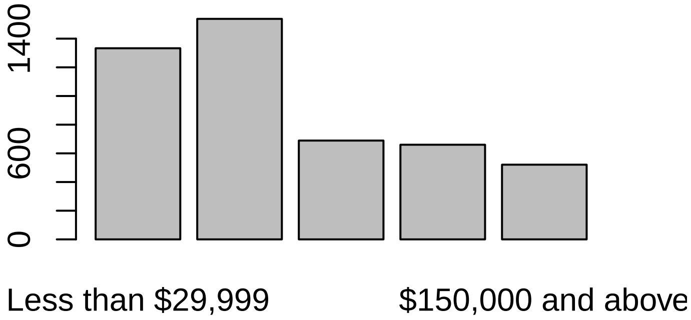
```r
barplot(table(survey_data$employment))
```



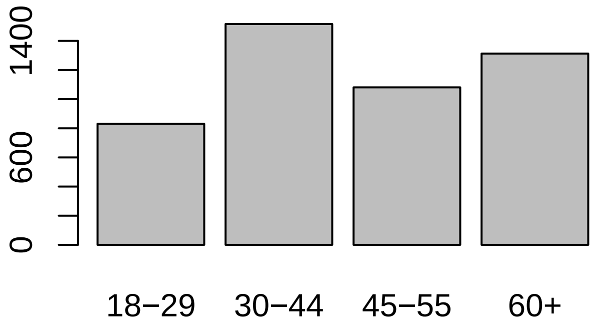
```r
barplot(table(survey_data$race))
```



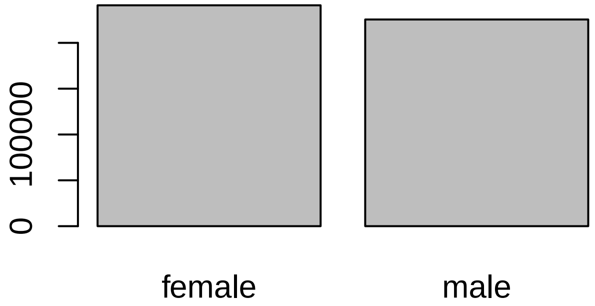
```r
barplot(table(survey_data$household_income))
```



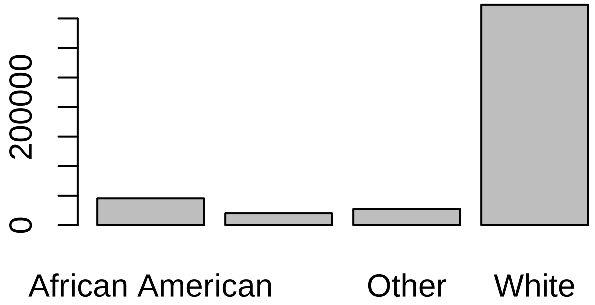
```
barplot(table(survey_data$age))
```



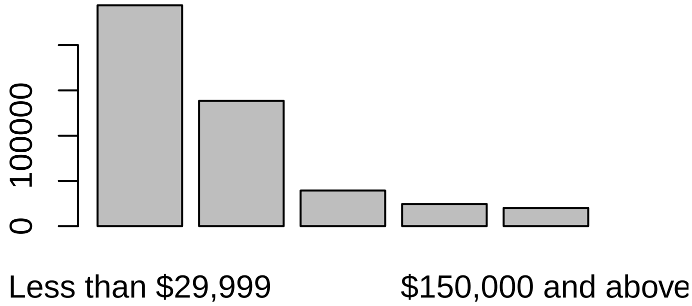
```
# doule check the distribution of data, categorical
barplot(table(census_data$gender))
```



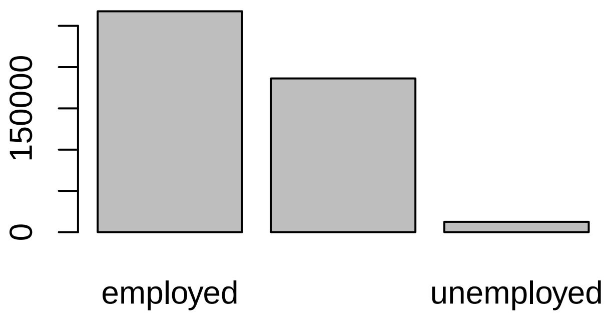
```
barplot(table(census_data$race))
```



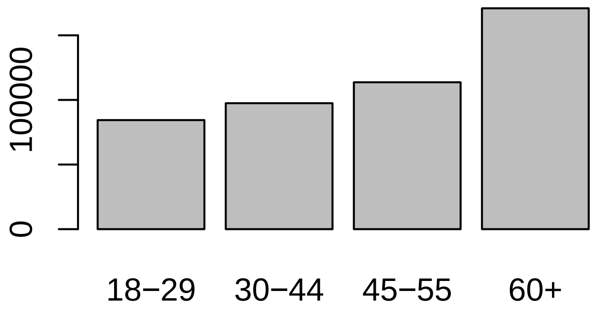
```
barplot(table(census_data$household_income))
```



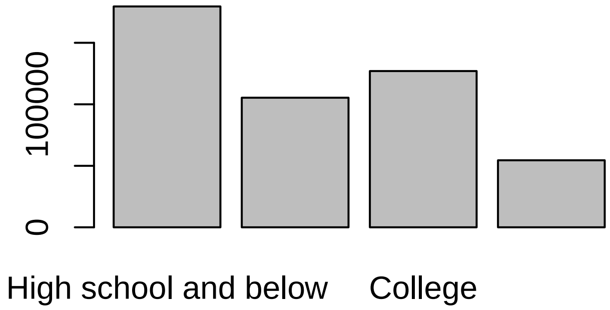
```
barplot(table(census_data$employment))
```



```r
barplot(table(census_data$age))
```



```r
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.

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.

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

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.*