Analysis on the popular vote of the 2020 American federal election

Yena Joo, Woolim Kim, Guemin Kim

Nov 2, 2020

Predictions on the 2020 US Presidential Election based on the voter survey responses.

Code and data supporting this analysis is available at: https://github.com/Guemin/Problem_Set_3

Model

As the 2020 presidential election of the United States approaches, people across the world are interested in to which candidate the vote of the US citizens will be concentrated, either to Donald Trump or to Joe Biden. Since the election outcome will also affect our community in Canada, we are going to analyze and predict the winner of the popular vote in the 2020 American federal election.
Using the survey and census data obtained from Democracy Fund + UCLA Nationscape and IPUMS USA, we are going to predict the popular vote outcome of the election. To be more specific, we are going to use two logistic regression models, one for each candidate, and employ a post-stratification technique[1] with the models.
In the following sub-sections, we will describe the model specifics, the post-stratification calculation, and the result of the analysis.

Model specifics

As already mentioned, we will be using the logistic regression models and post-stratification technique with R software to predict the proportions of voters who will vote for either Donald Trump or Joe Biden. Specifically, we will create two models, each for proportions of voters for Trump or Biden, using 6 different variables(age_group, gender, race, education, household_income , and state)[2].

Since our response variables, vote_Trump and vote_Biden, are binary(either 'vote for' or 'not vote/not sure'), the logistic regression model is a suitable model to be used. Logistic regression is a mathematical model used to estimate the probability of an event occurring using binary data.
The logistic regression models we are using are:

[1]Post-stratification is a technique used in sample survey design to improve the quality of population estimates. In the post-stratification analysis, the population is partitioned into subgroups, and estimates are predicted within the subgroups. With the estimates, the sum of the estimate times the respective population size in each group is calculated, and finally, the sum is divided by the sum of the total population size. Detailed procedures on post-stratification for our analysis will be shown in the following sub-sections.

[2]* age_group is divided into 4 different groups: "18-29 year olds", "30-44 year olds", "45-64 year olds", "65 years and older".
* gender indicates either "Male" or "Female".
* race is divided into 5 different categories: "White", "Black", "Native", "Asian", "Other".
* education is divided into 4 different categories: "Didn't graduate from high school", "High school graduate", "Some college or associate degree", "Bachelor's degree or higher".
* household income consists of 9 categories range from "Less than $14,999" to "$150,000 and over".
* state indicates abbreviated names of 52 states in the United States.

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{age\ group} + \beta_2 x_{gender} + \beta_3 x_{race} + \beta_4 x_{education} + \beta_5 x_{household\ income} + \beta_6 x_{state}$$

where $log(\frac{p_i}{1-p_i})$ represents log odds in each model, and p_i is the proportions of voters who will vote for Donald trump or Joe Biden. Similarly, β_0 represents the intercept, and $\beta_1, \ldots, \beta_6$ indicate the slope parameters of the model. (Detailed descriptions on the x variables can be found in the footnote[3]).

Using the log odds estimates, we are going to find vote_Trump and vote_Biden (the proportions of voters each for Donald Trump and Joe Biden) in every possible combination of categories in our predictor variables, age_group, gender, race, education, household income, and state.

model diagnostics
Logistic regression could be well performed under several assumptions.

The assumptions on logistic regression model are:
1. linearity between the log odds and the predictor variables (independent variables should be linearly related to the log odds)
2. independent errors (logistic regression requires each observation to be independent.)
3. multicollinearity among predictors is not too high (predictor variables should be independent to each other)

Figure 1

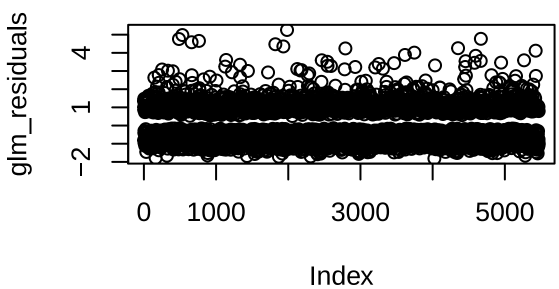


Figure 2

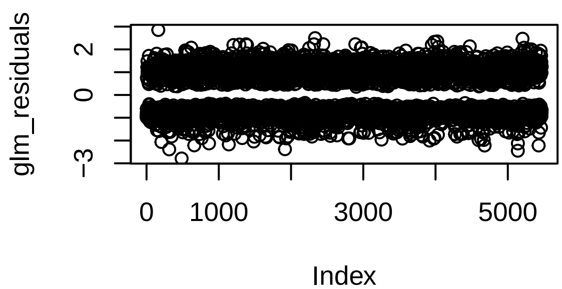


Figure 3
Cook's distance

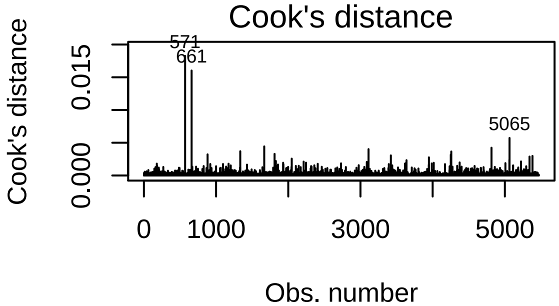
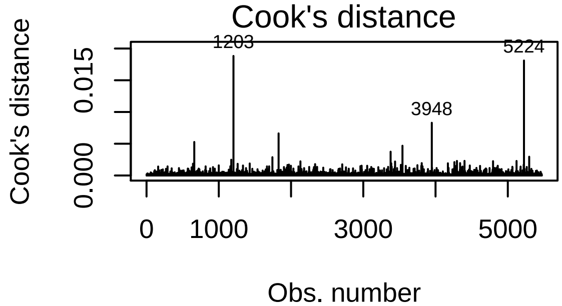


Figure 4
Cook's distance



###Yena 1. Model diagnostics:

[3]* $x_{age\ group}$ represents one of the four age groups that the respondent is in.
* x_{gender} indicates the gender of the respondent(either "Male" or "Female").
* x_{race} indicates the race ethnicity of the respondent.
* $x_{education}$ indicates the education attainment of the respondent.
* $x_{household\ income}$ indicates the total pre-tax income of the respondent's household.
* x_{state} indicates the state where the respondent is located in.

Table 1: VIF models

Predictor	VIF	Predictor	VIF
age_group	1.210003	age_group	1.246368
gender	1.068826	gender	1.072452
race	1.240050	race	1.353103
education	1.477839	education	1.452881
household_income	1.555977	household_income	1.564889
state	1.468859	state	1.461147

2. Cook's distance is used in regression analysis to find influential outliers in predictor variables, which identifies points that negatively affect your regression model. Cook's distance is higher when residuals and leverage are high. (Infulential/outlier points can be removed for a better analysis)
3. VIF is the Variance inflation factor, which measures the amount of multicollinearity in a set of multiple regression variables. Bigger the VIF, the bigger the multicollinearity is. As shown above, (Figure 3?) there is no sign of multicollinearity (not correlated) since VIF values do not exceed 2 for both models of Trump and Biden. Therefore, it is safe to say that the last assumption, multicollinearity among predictors is not too high is satisfied.

###Model content - Ages that are younger than 18 or older than 90(?) are removed from the data because… (also mention briefly what other samples are mutated/removed from the data)

Post-Stratification

- Any decisions that your group made should be explained and justified. (For example, are you looking at the proportion of people voting for Trump or Biden? why did you exclude sex in the cell split (practical explanations are acceptable)? etc.)

In order to estimate the probabilities of voting for both Donald Trump and Joe Biden, we are going to perform a post-stratification analysis. In order to use this technique, we need to subdivide the population having similar characteristics into cells. Hence, we are going to create a total of 55,325 cells based on different age groups, gender, race-ethnicity, education attainment, household income, and state.
Using the logistic regression models presented in the previous sub-section, we will estimate the probabilities of voting in each cell for each candidate. Then, we will weigh each estimate within each cell by the respective population size of the cell, sum those values, and divide that by the entire population size. This process can also be described by the expression:

$$\hat{y}^{ps} = \frac{\sum N_j * \hat{y_j}}{\sum N_j}$$

where $\hat{y_j}$ is the estimate of the probability of voting for either Trump or Biden in each cell, and N_j is the population size of the j^{th} cell based off demographics.

reason for Choice of the variables…

Results

Here you will include all results. This includes descriptive statistics, graphs, figures, tables, and model results. Please ensure that everything is well formatted and in a report style. You must also provide an explanation of the results in this section.

Table 2: Comparison of predicted estimate between Trump and Biden

total_predict_trump	total_predict_biden
0.4334444	0.3944298

Please ensure that everything is well labelled. So if you have multiple histograms and plots, calling them Figure 1, 2, 3, etc. and referencing them as Figure 1, Figure 2, etc. in your report will be expected. The reader should not get lost in a sea of information. Make sure to have the results be clean, well formatted and digestible.

```
## # A tibble: 70 x 5
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                               -0.707    0.741     -0.954 3.40e- 1
##  2 as.factor(age_group)30-44 year olds        0.575    0.0950     6.05  1.43e- 9
##  3 as.factor(age_group)45-64 year olds        0.743    0.0940     7.91  2.59e-15
##  4 as.factor(age_group)65 years and older     0.782    0.108      7.25  4.10e-13
##  5 as.factor(gender)Male                      0.422    0.0612     6.90  5.25e-12
##  6 as.factor(race)Black                      -1.42     0.209     -6.79  1.12e-11
##  7 as.factor(race)Native                      0.483    0.285      1.70  8.99e- 2
##  8 as.factor(race)Other                      -0.132    0.200     -0.661 5.08e- 1
##  9 as.factor(race)White                       0.589    0.160      3.68  2.37e- 4
## 10 as.factor(education)Didn't graduate fr~    0.357    0.119      3.01  2.61e- 3
## # ... with 60 more rows
```

```
## # A tibble: 70 x 5
##    term                                  estimate std.error statistic    p.value
##    <chr>                                    <dbl>     <dbl>     <dbl>      <dbl>
##  1 (Intercept)                           -0.418      0.839    -0.498      6.18e-1
##  2 as.factor(age_group)30-44 year olds   -0.200      0.0856   -2.34       1.93e-2
##  3 as.factor(age_group)45-64 year olds   -0.287      0.0855   -3.35       8.01e-4
##  4 as.factor(age_group)65 years and old~ -0.125      0.101    -1.24       2.14e-1
##  5 as.factor(gender)Male                 -0.302      0.0592   -5.11       3.27e-7
##  6 as.factor(race)Black                   0.999      0.166     6.03       1.67e-9
##  7 as.factor(race)Native                 -0.442      0.278    -1.59       1.12e-1
##  8 as.factor(race)Other                   0.00339    0.173     0.0196     9.84e-1
##  9 as.factor(race)White                  -0.449      0.143    -3.14       1.66e-3
## 10 as.factor(education)Didn't graduate ~ -0.667      0.114    -5.86       4.61e-9
## # ... with 60 more rows
```

We have created the logistic regression model on proportion of voters voting for Donald Trump and Joe Biden with 6 different following variables: age_group, gender, race, education, household_income, and state. Based off the post-stratification analysis we made, our estimation of the proportion of voters voting for Donald Trump is <0.433> and Joe Biden to be <0.394>.From the result of our estimation, We can predict that Donald Trump is more likely to win the 2020 president election.

- individuals with household_income "less than $14,999" are more likely to vote for Biden over Trump (due to Biden's election promises for lower income people?)
- state: **idk**
  Using the estimate proportion grouped by states, ...

```r
#compare the estimate for each state
trump_state <- predict_state_trump$predict_trump2
biden_state <- predict_state_biden$predict_biden2
trump_win <- 0
```

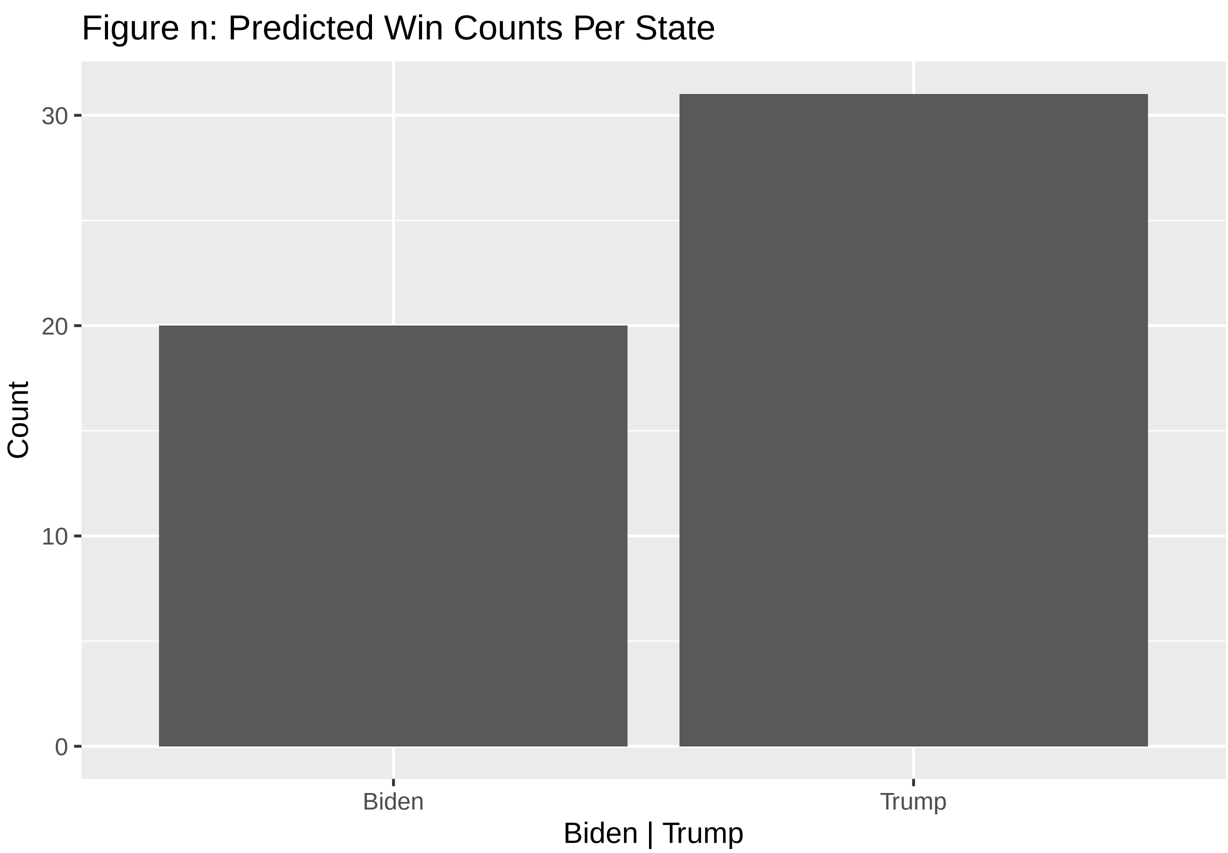
```r
biden_win <- 0
trump_biden <- c()
for (i in 1:51){
  if (trump_state[i] > biden_state[i]){
    trump_win <- trump_win + 1
    trump_biden <- c(trump_biden, "Trump")
  }
  else {biden_win <- biden_win + 1
  trump_biden <- c(trump_biden, "Biden")
}}
trump_win
```

```
## [1] 31
```

```r
biden_win
```

```
## [1] 20
```

```r
ggplot(data.frame(trump_biden), aes(x=trump_biden)) +
  geom_bar() + ggtitle("Figure n: Predicted Win Counts Per State") + xlab("Biden | Trump") + ylab("Coun
```



Under the assumption that whoever gets a higher expected proportion for each state wins in that state, Trump is expected to win in 31 states, and Biden is expected to win in 20 states. Both Figure x and Figure y show that Donald Trump has a higher possibility to win the election.

Discussion

Here you will summarize the previous sections and discuss conclusions drawn from the results. Make sure to elaborate and connect your analysis to the goal of the study.

Using the survey and census data obtained from Democracy Fund + UCLA Nationscape and IPUMS USA, we have predicted the popular vote outcome of the 2020 president election in USA. Logistic Regression is used to predict who is more likely to be elected for the 2020 presidential election. Explanatory variables used for the logistic regression model are age_group, gender, race, education, household_income, and state. Then, $\hat{y}^{ps}$ is measured using post-stratification technique to estimate the proportion of voters in favor of voting for each candidate.
- By using the post stratification, we created 55,325 cells based on the 6 variables that was used in the model, and found the probability of voting estimates for each cells. - Then we grouped each cell estimates into states and predicted who is expected to have more - The result shows that estimate value for proportion of voters voting for Joe Biden is 39.4% and Donald Trump 43.34%. Also Trump is ahead of Biden by 21 counts in estimate for each state. - *discuss about the result*

To conclude, based off the estimated proportion of voters in favor of voting for Donald Trump being 0.4334 (43.34%) and expected to win 31 states, we predict that Trump will win the 2020 president election. (. . . . . .)

Weaknesses

Here we discuss weaknesses of the study, data, analysis, etc. You can also discuss areas for improvement. 1. Weakness: Some variables could not be included in the generalized logistic model because either census data or survey data did not include the particular variables. If there is an important variable that could have affected the vote outcome, there might exist an omitted variable bias. (The omitted variables should be correlated with the dependent variable and with the explanatory variables included in the model).

- the Census data used in the analysis is 2018 data, so it might not reflect the most accurate vote outcome. 2020 data is more suitable to analyze more accurate results. Also, people who were underage in 2016, hence not included in the estimate would have the right to vote in 2020.

Next Steps

Here you discuss subsequent work to be done after this report. This can include next steps in terms of statistical analysis (perhaps there is a more efficient algorithm available, or perhaps there is a caveat in the data that would allow for some new technique). Future steps should also be specified in terms of the study setting (eg. including a follow-up survey on something, or a subsequent study that would complement the conclusions of your report).

- With the 2020 census data, we could estimate the proportion of voting for each candidate by state and estimate the winner of each state which would make a more reasonable and realistic prediction of the election.

- Create visualization of the results to view the groups of the voting estimates at once.

- In our future analysis, we can try to analyze the multilevel regression models using Bayes coding techniques.

- We can compare our prediction and the result of the actual 2020 president election.
  (something about comparing with the actual election results and do a post-hoc analysis (or at least a survey) of how to better improve estimation in future elections.)

References

1. Survey data: https://www.voterstudygroup.org/downloads?key=9337162e-e5ef-49d7-96fd-48a5c5dba31c
2. Census data: https://usa.ipums.org/usa-action/extract_requests/summary?
3. Post-Stratification technique: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf
4. Logit Regression Assumptions: https://rpubs.com/guptadeepak/logit-assumptions
5. Variance Inflation Factor(VIF): https://www.statisticshowto.com/variance-inflation-factor/
6. Tables side by side: https://bookdown.org/yihui/rmarkdown-cookbook/kable.html

#Appendix

state	predict_trump2
AK	0.6178175
AL	0.5294578
AR	0.5690489
AZ	0.4970620
CA	0.3500102
CO	0.4748248
CT	0.2840064
DC	0.2715509
DE	0.3901171
FL	0.4677497
GA	0.4716816
HI	0.3371692
IA	0.4501696
ID	0.6617140
IL	0.4152228
IN	0.4497127
KS	0.5724607
KY	0.4997506
LA	0.4574786
MA	0.2894075
MD	0.3519287
ME	0.4062306
MI	0.4074363
MN	0.4807777
MO	0.4489350
MS	0.4849136
MT	0.5407824
NC	0.4647314
ND	0.5234047
NE	0.4228730
NH	0.4164238
NJ	0.4045109
NM	0.2288712
NV	0.5159548
NY	0.3888962
OH	0.4457807
OK	0.4921180
OR	0.4084833
PA	0.4706141
RI	0.3574909
SC	0.5061444
SD	0.5185028
TN	0.5126872
TX	0.5087421
UT	0.4055765
VA	0.3846518
VT	0.1026603
WA	0.3766525
WI	0.3970139
WV	0.5386586
WY	0.1878584

state	predict_biden2
AK	0.2121776
AL	0.3429311
AR	0.2163157
AZ	0.3532261
CA	0.4605909
CO	0.3723088
CT	0.5343268
DC	0.7314580
DE	0.5308874
FL	0.3841190
GA	0.3827232
HI	0.5260513
IA	0.3894466
ID	0.2276048
IL	0.3984838
IN	0.3483643
KS	0.2903427
KY	0.4122667
LA	0.4197140
MA	0.5138926
MD	0.4938823
ME	0.4861133
MI	0.4562030
MN	0.4625594
MO	0.3787824
MS	0.3758363
MT	0.3513251
NC	0.4126773
ND	0.1747900
NE	0.3367801
NH	0.4753414
NJ	0.4240766
NM	0.5074545
NV	0.3375036
NY	0.4355101
OH	0.3743578
OK	0.2217403
OR	0.4259771
PA	0.3096477
RI	0.4515775
SC	0.2786692
SD	0.3387865
TN	0.2784533
TX	0.3048833
UT	0.2632457
VA	0.4476426
VT	0.7455384
WA	0.4615358
WI	0.4124047
WV	0.3194931
WY	0.2659662

Table 3: Figure n

household_income	predict_trump	household_income	predict_biden
$100,000 to $149,999	0.4788905	$100,000 to $149,999	0.3644576
$15,000 to $24,999	0.3819243	$15,000 to $24,999	0.4089684
$150,000 and over	0.5013108	$150,000 and over	0.3746106
$25,000 to $34,999	0.3917869	$25,000 to $34,999	0.3918452
$35,000 to $44,999	0.4036812	$35,000 to $44,999	0.4117412
$45,000 to $54,999	0.4363749	$45,000 to $54,999	0.3898337
$55,000 to $74,999	0.4219549	$55,000 to $74,999	0.4162607
$75,000 to $99,999	0.4191757	$75,000 to $99,999	0.4305907
Less than $14,999	0.3219549	Less than $14,999	0.3930797