Analysis on the popular vote of the 2020 American federal election

Yena Joo, Woolim Kim, Guemin Kim

Nov 2, 2020

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 voterstudygroup.org 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 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 (woolim)

As already mentioned, we will be using the logistic regression models and post-stratification technique to predict the proportions of voters who will vote for Donald Trump and Joe Biden. Specifically, we will create two models, each for probabilites of voting for Trump or Biden, using 6 different variables(age_group, gender, race, education, household_income , and state)[1].

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.
The logistic regression models we are using are:

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

where $log(\frac{p_i}{1-p_i})$ represents log odds in each model, and p_i is the probability of voters who will vote for Donald trump or Joe Biden. Similarly, β_0 represents the intercept, and β_1,. . . , β_6 indicate the slope parameters of the model. (Detailed descriptions on the x variables can be found in the footnote[2]).

[1] * 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.

[2] * $x_{age\ group}$ 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.

Using the log odds estimates, we are going to find the probability of voting for Donald Trump and Joe Biden in every possible combinations of categories in our predictor variables, age_group, gender, race, education, household income, and state.

Post-Stratification

In order to estimate the proportion of voters who will vote for Donald Trump I need to perform a post-stratification analysis. Here I create cells based off different ages. Using the model described in the previous sub-section I will estimate the proportion of voters in each age bin. I will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

To find the proportion of voters who will vote for Trump and Biden estimates, we are going to use the post stratification analysis. Post-stratification is a technique used in sample survey design to improve the quality of population estimates.In order to use this technique, we subdivide the population having similar characteristics into cells. From previous logic regression model we created, we will estimate the probability of voters in each cells. The 6 categories that subdivides the population into a certain bin is age group, gender, race, education, household income, and state. Then we weight the each proportion of voting estimate by the respective population size of the cells and sum those values and divide it by the entire population size.

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.

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
```

```
## # A tibble: 1 x 1
##   total_predict_trump
##                 <dbl>
## 1               0.433
```

```
## # A tibble: 1 x 1
##   total_predict_biden
##                 <dbl>
## 1               0.394
```

```
## # A tibble: 9 x 2
##   household_income     predict_trump
##   <chr>                        <dbl>
## 1 $100,000 to $149,999         0.479
## 2 $15,000 to $24,999           0.382
## 3 $150,000 and over            0.501
## 4 $25,000 to $34,999           0.392
## 5 $35,000 to $44,999           0.404
## 6 $45,000 to $54,999           0.436
## 7 $55,000 to $74,999           0.422
## 8 $75,000 to $99,999           0.419
## 9 Less than $14,999            0.322
```

```
## # A tibble: 9 x 2
##   household_income     predict_biden
##   <chr>                        <dbl>
## 1 $100,000 to $149,999         0.364
## 2 $15,000 to $24,999           0.409
## 3 $150,000 and over            0.375
## 4 $25,000 to $34,999           0.392
## 5 $35,000 to $44,999           0.412
## 6 $45,000 to $54,999           0.390
## 7 $55,000 to $74,999           0.416
## 8 $75,000 to $99,999           0.431
## 9 Less than $14,999            0.393
```

```
## # A tibble: 51 x 2
##    state predict_trump2
##    <chr>          <dbl>
##  1 AK             0.618
##  2 AL             0.529
##  3 AR             0.569
##  4 AZ             0.497
##  5 CA             0.350
##  6 CO             0.475
##  7 CT             0.284
##  8 DC             0.272
##  9 DE             0.390
## 10 FL             0.468
## # ... with 41 more rows
```

```
## # A tibble: 51 x 2
##    state predict_biden2
##    <chr>          <dbl>
##  1 AK             0.212
##  2 AL             0.343
##  3 AR             0.216
##  4 AZ             0.353
##  5 CA             0.461
##  6 CO             0.372
##  7 CT             0.534
##  8 DC             0.731
##  9 DE             0.531
## 10 FL             0.384
## # ... with 41 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 <0.4>? and Joe Biden to be <0.3>?.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?) # 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.

Weaknesses

Here we discuss weaknesses of the study, data, analysis, etc. You can also discuss areas for improvement.

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).

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