United States Presidential Election - 2020 Forecast

Elric Lazaro

November 2, 2020

Code and data supporting this analysis is available at:
https://github.com/ElricL/United-States-Presidential-Election-2020-Forecast.git

Model

Model Specifics

Our model will be fitted using a survey sample obtained from Democracy Fund + UCLA Nationscape. From the collection of data given, we will be using the most recent survey available which is dated on June 25th, 2020. When processing the data I managed to create boolean (True/False) variables of wether to vote Trump or not and as well as Biden. Such variable translates well with logistic model regression. A logistic model will be used as it allows us to estimate and isolate probability. The following variables from our survey data will be used as I believe there is a correlation between them and the potential candidate vote: gender, race ethnicity, household income, and state. State in particular will be used as a group variable, making our model a multilevel model. This allows the model's intercept to vary from state to state. This is done as I believe that different states have different starting points in our model as there must be some correlation between the output of the other predictor variables and the state. Therefore, using a group variable should improve our model's performance in comparison to simply generalizing our probability through all of United States. The logistic regression model I am using can be written as a formula:

$$
\begin{aligned}
log(\frac{(p)}{1-(p)}) = \beta_0 + \beta_1 x_{gender:Male} + \beta_2 x_{gender:Female} + \\
\beta_2 x_{race\_ethnicity:Asian\ (Chinese)} + \\
\beta_3 x_{race\_ethnicity:Asian\ (Japanese)} + \\
\beta_4 x_{race\_ethnicity:Black\ or\ African\ American} + \\
\beta_5 x_{race\_ethnicity:Other\ Asian\ or\ Pacific\ Islander} + \\
\beta_6 x_{race\_ethnicity:Some\ other\ race} + \\
\beta_7 x_{race\_ethnicity:White} + \\
\beta_8 x_{household\_income:\$125000\ and\ above} + \\
\beta_9 x_{household\_income:\$100000\ to\ \$124999} + \\
\beta_1 0 x_{household\_income:\$75000\ to\ \$99999} + \\
\beta_1 1 x_{household\_income:\$50000\ to\ \$74999} + \\
\beta_1 2 x_{household\_income:\$15000\ to\ \$49999} + \\
\beta_1 3 x_{household\_income:\$75000\ to\ \$99999} + \\
\beta_1 4 x_{household\_income:Less\ than\ \$14999} + \epsilon
\end{aligned}
$$

Where p represents the probability that a voter or voters will vote the model's specific candidate. Similarly, β_0 represents the intercept of the model, which will vary depending on the state. Additionally, the following β represents the slopes of the features. The higher the slope the higher the probability is for given variable. For instance if the slope for a person with household income '$Less\,than\,\$14,999$' is -1 while one with income of $\$15,000\,to\,\$49,999$ is 1, the person with higher income is more likely to vote the model's candidate given their positive and large slope. Note that a negative slope also decreases likelihood while a positive increases.

Note that since we are analyzing votes for Trump and Biden we will perform this model twice with a fit for each candidate.

Post-Stratification

In order to estimate the proportion of voters who will vote for either candidates I need to perform a post-stratification analysis. The estimation for our logistic model will be done on American Community Surveys (ACS) data. This census data will provide us the Individual/Unit level features that can be entered into our model along with the individual's State, allowing us to make great amount of predictions. Predictions will come in form of cells based of diefferent gender, race ethnicity, household income, and State. Each will then be weighted based on their population size divided by the entire population size. Note that for post-stratification we can only use one model. We can simply choose the model that has a better correlation between our probability of vote and the variables. This can be achieved by observing the model's p-values and which one is usually lower. We can do this as a lower p-value means we can more likely reject the null hypothesis that there is no correlation.

Results

The following are the results of our model fit for each candidate:

factors	Trump_Estimate	Biden_Estimate
gender: Male	0.40	-0.37
race_ethnicity: Asian (Chinese)	-1.33	1.33
race_ethnicity: Asian (Japanese)	-1.14	1.50
race_ethnicity: Black or African American	-2.06	1.83
race_ethnicity: Other Asian or Pacific Islander	-0.71	0.79
race_ethnicity: Some other race	-0.76	0.76
race_ethnicity: White	0.06	0.21
household_income: $125,000 and above	0.12	0.00
household_income: $15,000 to $49,999	-0.27	0.28
household_income: $50,000 to $74,999	-0.21	0.25
household_income: $75,000 to $99,999	-0.25	0.39
household_income: Less than $14,999	-0.47	0.21

Figure 1: Summary of the logistic model for the likelihood of voting for Donald Trump and Jose Biden. Shows key details as to how each characteristic of an Individual affects their probability of voting the candidate.

The β coefficients or slopes of our models can be identified in the estimate column which we can then use to observe how each individual characteristic affect the probability of voting the candidate. The fit of our model has concluded that individuals with certain characteristics will more likely vote one than the other. For instance, we can observe that the estimate/[Beta] coefficient for male gender has a positive number for Trump's logistic model while Biden's is negative. We should be able to observe from this result that there is a higher chance that male-identified people will vote Trump over Biden. More on the results of these

estimates are further detailed in the Discussion section.

There were more lower p-values in Biden's model than Trump's and as a result I am more confident with rejecting the null hypothesis of no correlation between the logistic probability and the variables for Biden's logistic model. Therefore, Post-stratification was estimated using logistic model on probability of voting Biden.

Parameter	Trump_p_value	Biden_p_value
b_Intercept	0.7065	0.0470
b_as.factorgenderMale	0.0000	0.0000
b_as.factorrace_ethnicityAsianChinese	0.0010	0.0015
b_as.factorrace_ethnicityAsianJapanese	0.0730	0.0150
b_as.factorrace_ethnicityBlackorAfricanAmerican	0.0000	0.0000
b_as.factorrace_ethnicityOtherAsianorPacificIslander	0.0425	0.0230
b_as.factorrace_ethnicitySomeotherrace	0.0160	0.0155
b_as.factorrace_ethnicityWhite	0.8385	0.4690
b_as.factorhousehold_income.125000andabove	0.3390	0.9605
b_as.factorhousehold_income.15000to.49999	0.0165	0.0205
b_as.factorhousehold_income.50000to.74999	0.0945	0.0530
b_as.factorhousehold_income.75000to.99999	0.0715	0.0065
b_as.factorhousehold_incomeLessthan.14999	0.0020	0.1265

Figure 2: p-values for each coefficient in Trump's and Biden's logistic Model.

I estimated that the proportion of voters in favor of voting for Joe Biden to be 0.6059. This is based off our post-stratification analysis of the proportion of voters in favor of Joe Biden modeled by a multilevel logistic regression model, which accounted for gender, racial ethnicity, and household income. We can also then assume that Trump will likely have lesser proportion of 0.3941

When grouping the post-stratification values by states, we can also see that the prediction of majority of the votes going to Joe Biden is reflected for all states.

state	biden_mean	trump_mean
AK	0.5922632	0.4077368
AL	0.6076138	0.3923862
AR	0.5957750	0.4042250
AZ	0.5947300	0.4052700
CA	0.6225320	0.3774680
CO	0.6048940	0.3951060
CT	0.6203097	0.3796903
DE	0.6194105	0.3805895
FL	0.6046670	0.3953330
GA	0.6076003	0.3923997
HI	0.6299054	0.3700946
IA	0.5994707	0.4005293
ID	0.5853161	0.4146839
IL	0.6168898	0.3831102
IN	0.6002965	0.3997035
KS	0.5911766	0.4088234
KY	0.6065827	0.3934173
LA	0.6208646	0.3791354
MA	0.6239742	0.3760258
MD	0.6287050	0.3712950
ME	0.6039892	0.3960108
MI	0.6162567	0.3837433

state	biden_mean	trump_mean
MN	0.6080270	0.3919730
MO	0.6027434	0.3972566
MS	0.6204334	0.3795666
MT	0.5971926	0.4028074
NC	0.6122028	0.3877972
ND	0.5929837	0.4070163
NE	0.6057515	0.3942485
NH	0.5993031	0.4006969
NJ	0.6107848	0.3892152
NM	0.6066971	0.3933029
NV	0.6012416	0.3987584
NY	0.6213519	0.3786481
OH	0.6055521	0.3944479
OK	0.5961975	0.4038025
OR	0.6090936	0.3909064
PA	0.5945078	0.4054922
RI	0.6085158	0.3914842
SC	0.5954493	0.4045507
SD	0.5931790	0.4068210
TN	0.5926075	0.4073925
TX	0.5896227	0.4103773
UT	0.5875437	0.4124563
VA	0.6196940	0.3803060
VT	0.6112230	0.3887770
WA	0.6191030	0.3808970
WI	0.6107297	0.3892703
WV	0.5936119	0.4063881
WY	0.5967145	0.4032855

Figure 3: Post-stratification values of proportion voters for Joe Biden and Donald Trump per State.

Discussion

Using my multilevel logistic regression model that I fitted for both Trump vote and Biden vote intentions, I was able to observe which kind of individuals would gravitate towards Joe Biden or Trump based on their gender, ethnicity, and household income. With the two regression models, I chose Biden's model for our post-stratification process for having better correlation with my individual variables. Using the census data obtained from American Community Surveys, we then estimated on different cells based on the identified gender, race ethnicity, and household income.

Using the slopes that we have observed from our results we can make deductions on how each variable affect the likelihood of an individual voting for either candidate. Starting with trump I can observe the following:

- Race ethnicity other than 'White' have lesser likelihood of voting him.
- Males are more likely to vote Trump compared to Females.
- Individuals with high income ($125,000 and above) are more likely to vote him while a specified lesser income decreases the chances.

Biden on the other hand had the opposite situation with the following relationships:

- Non 'White' ethnicities have likelier chance of voting Biden. However, do note that individuals identified as 'White' still have decent chance of voting Biden given their positive slope.
- Males are less likely to vote Biden compared to females.
- Individuals with income less than $125,000 are more likely to vote Biden.

However, it appeared that overall there were more characteristics that leads to a higher chance of favoring Biden. This can be further strengthened by my post-stratification results which gave a higher estimate for Joe Biden. Based off the estimated proportion of voters in favor of voting for Joe Biden being 0.6059, I predict that Joe Biden will win the election.

Weaknesses

There are some limitations to my research. Given how expensive fitting a logistic model with Bayesian inference, variables such as household income had to be simplified. For same reason, age had to be omitted as a census dataset with it proved to be too large to calculate predictions on our model. The census data and survey data had some differences with their outputs resulting in needing to omit some information to generalize and match the two datasets. For instance, survey data contained more types of racial ethnicities but since we need to make predictions based off census data which had more limited output, many of the races had to be re-categorized. i.e. Asian (Asian Indian) outputs in the survey data had to be replaced with 'Other Asian or Pacific Islander'. Lastly, while the model with best p-values were chosen for post-stratification process, some were still above the standard alpha level of 0.05 and as a result it is possible correlation may be weak.

Next Steps

To further improve the model and post-stratification process, we can compare it to the final election results. If our prediction on Joe Biden winning the election ends up being wrong, it is critical to see where our analysis went wrong. There are many things we can look into that would result in our poor prediction in such scenario. Firstly, We could have missed an important variable in our model. Our dataset may not reflect well with general population. Further samples and surveys should have been used. Other ways we could have improve this model if given wrong prediction is by trying out different models if logistic regression ends up being inaccurate fit. For instance, using a model with Bayesian Inference may have allowed too much bias in the fit. Testing more efficient models may also be useful as it would allow me to use age as one of predictor variables as originally intended. Survey results also had "Don't know" or "Someone else" as the individual's vote intention. While their occurrence frequency should be too few to affect our final prediction result, it may still be interesting to find how these individuals have an affect on the outcome.

References

1. Join two tbls together. (n.d.). Retrieved from https://dplyr.tidyverse.org/reference/join.html
2. Extract or Replace Parts of a Data Frame. (n.d.). Retrieved November 02, 2020, from https://astrostatistics.psu.edu/su07/R/html/base/html/Extract.data.frame.html
3. Rename Data Frame Columns in R. (n.d.). Retrieved November 02, 2020, from https://www.datanovia.com/en/lessons/rename-data-frame-columns-in-r/
4. Convert case of a string - case. (n.d.). Retrieved November 02, 2020, from https://stringr.tidyverse.org/reference/case.html

5. Hadley Wickham [aut, C. (2020, August 18). Group_by_all: Group by a selection of variables in dplyr: A Grammar of Data Manipulation. Retrieved November 02, 2020, from https://rdrr.io/cran/dplyr/man/group_by_all.html

6. Daniel Lüdecke [aut, C. (2020, October 29). P_value: P-values in parameters: Processing of Model Parameters. Retrieved November 02, 2020, from https://rdrr.io/cran/parameters/man/p_value.html

7. Press, C., Finance, Y., & Newsweek. (2020, October 30). New: Second Nationscape Data Set Release. Retrieved November 02, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

8. Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 02, 2020, from https://usa.ipums.org/usa/index.shtml