Predicting the 2020 American Election using Post-Stratification on a Logistic Regression Model

Ruize Liu (1003320499), Yi Lei Feng (1003139356)

November 2nd, 2020

Code supporting this analysis is available at https://github.com/Raymastered/STA304_PS3.

Model

In this report, we are interested in predicting the popular vote outcome of the 2020 American federal election, similar to recent reports such as a Cambridge forecasting study which highlighted the many challenges of performing such analysis in a tumultuous year. To do this, we are employing a post-stratification technique. In the following sub-sections we will describe the model specifics and the post-stratification calculations.

Model Specifics

We will be using a logistic regression model to model the proportion of voters who will vote for either Donald Trump or Joe Biden. A logistic model is appropriate in this case as we are trying to predict a binary response value corresponding to predicting either a Trump or Biden win. Our model will consist of the input parameters: work status, education, and race. Our output will be a binary value that indicates if that particular candidate wins the upcoming election. We can represent our model as follows:

$$log(\frac{y}{1-y}) = \beta_0 + \beta_1 x_{work} + \beta_2 x_{edu} + \beta_3 x_{race} + \epsilon$$

Where y represents the likelihood of a vote for the candidate under a particular instance. Similarly, β_0 represents the intercept of the model, and is the likelihood of voting for that candidate when all paramaters are a non-factor (i.e. not working, no education, unspecified race). Additionally, $\beta_1 \ldots \beta_3$ represent the respective slopes of the model based on each parameter, where an increase in said parameter is expected to have a correlated increase of some β_i amount towards the output likelihood.

Post-Stratification

To estimate the proportion of voters who will vote for Donald Trump or Joe Biden, we will perform a post-stratification analysis. This technique is effectively mapping the likelihoods we will be calculating for our logistic regression model to a larger population dataset.

For instance, assume we calculate the likelihood of voting Trump to be 55% for a white bachelors' holder who is in the workforce. Then for the "white bachelors' holder" group in the census dataset, say 30000/100000 people, we assume all 30000 of them have a 55% chance to vote Trump. Doing so, we can scale the % by all groups, and average out the total expected value based on population/group distribution.

The inputs we chose are therefore quite useful because they distinctly partition different aspects of a general population. Whether someone is employed or not is likely to change their goals and motivations. The same can be said for education level, and while race is a little generalized, there are often some visible differentiation between racial groups. Overall, we believe these inputs are flexible as they can distinctly create strata that could lead to interesting output predictions.

Thus, using the model described in the previous sub-section, we will estimate the proportion of voters in each bin comprising of a sub-grouping of the 3 inputs (i.e. white bachelors' holder that's employed currently). We will then weigh each proportion estimate (within each bin) by the respective population size of that bin and sum those values. Finally, we'll divide that by the entire population size to result in an overall population likelihood value.

Results

Taking our post-stratification data, we will now analyze the likelihood of the candidates being elected across each input demographic, and then extrapolate across all groups to obtain a population percentage estimation. Our model will be passed through the predict function with our census data groups along with the flag "type = response" which is required for logistic models. This will process our log estimates from the model and output a numerical estimate between 0 to 1 for each census group, representing the average likelihood of someone in that group to vote for one of the presidential candidates.

First, let's look at the results for Donald Trump, beginning with grouping by education levels:

Education Group	Likelihood of Voting Trump	Data Points
3rd Grade or less	0.4667029	59183
Associate Degree	0.3804567	262054
College Degree (such as B.A., B.S.)	0.4082705	593308
Completed some college, but no degree	0.3820363	495326
Completed some high school	0.3778908	906078
Doctorate degree	0.5312042	43373
High school graduate	0.3795134	521513
Masters degree	0.4511598	326973
Middle School - Grades 4 - 8	0.2921701	88155

[Figure 1]

The education groups fluctuate quite a bit from 29.2% to 53.1%, but the groups with a higher number of data points are typically within the 35-45% range for Trump. Next, we'll look at the group predictions by race:

Racial Group	Likelihood of Voting Trump	Data Points
american indian or alaska native	0.2916638	34216
black/african american/negro	0.0876740	338361
chinese	0.1663099	46070
japanese	0.1947627	9715
other asian or pacific islander	0.2916479	126788
other race, nec	0.2369947	115291
white	0.4501645	2625522

[Figure 2]

The data fluctuates quite a bit again, with the range going from a low 8.8% to 45.0%, and incidentally, these

extremes are the two strata with the most data points. Finally, we'll look at the likelihoods split by whether an individual is employed:

Is Employed/In Laborforce	Likelihood of Voting Trump	Data Points
no, not in the labor force	0.3667534	1305469
yes, in the labor force	0.4102657	1990494

[Figure 3]

It seems like there is a slight difference between the two groups, with unemployed people on average having a 36.7% chance to vote Trump, compared to 41.0% for employed people.

We can calculate the overall likelihood of Trump being voted by performing an aggregation over the entire post-stratification dataset, including all our input factors of education, employment, and race.

Overall Likelihood of Voting Trump
0.3930313

[Figure 4]

Based on this post-stratification analysis of our data using a logistic regression model, we can estimate that the overall predicated chance for a person over 18 (on average) voting for Trump is about 39.3%.

Additionally, we ran the post stratification process on the model and inputs for Biden's chances of winning, and the results are as follows:

Biden vote chances by Education groups:

Education Group	Likelihood of Voting Biden	Data Points
3rd Grade or less	0.4635017	59183
Associate Degree	0.4464457	262054
College Degree (such as B.A., B.S.)	0.4545564	593308
Completed some college, but no degree	0.4142040	495326
Completed some high school	0.3345221	906078
Doctorate degree	0.3667853	43373
High school graduate	0.3399636	521513
Masters degree	0.4583668	326973
Middle School - Grades 4 - 8	0.3991067	88155

[Figure 5]

Biden vote chances by Racial groups:

Racial Group	Likelihood of Voting Biden	Data Points
american indian or alaska native	0.3679440	34216
black/african american/negro	0.6509535	338361
chinese	0.5614802	46070
japanese	0.7090053	9715
other asian or pacific islander	0.4863509	126788
other race, nec	0.4415001	115291
white	0.3513509	2625522

[Figure 6]

Biden vote chances by Employment groups:

Is Employed/In Laborforce	Likelihood to Voting Biden	Data Points
no, not in the labor force	0.4032764	1305469
yes, in the labor force	0.3889393	1990494

[Figure 7]

Overall chances for Biden to be voted:

Overall Likelihood of Voting Biden
0.394618

[Figure 8]

The overall chances of an individual voting for Biden is thus around 39.5%, slightly higher than Trump's estimated odds.

Discussion

Summary

Our goal was to make a prediction on the likely winner of the highly-anticipated American election that has sparked heated debates and stirred controversy for several months. We chose to model our sample survey data (obtained from Democracy Fund + UCLA Nationscape) with a logistic regression model since we are attempting to find an estimation on the likelihood of a campaign winner with probabilities ranging between 0 and 1. The next step was to apply our model to a post-stratification dataset obtained from American Community Surveys by dividing our model into bin splits classified by our chosen predictor variables: work status, education level, and race. Within each bin, we estimate the percentage chances of each campaign winning the election and apply a weight to this estimate relative to the entire population. Aggregating these estimations with their respective weights provided us with population-level odds for each party winning the 2020 election. The results of our post-stratification analysis showed an overall prediction of 39.3% of the voting-eligible population who are in favor of voting for the Republican Party and a 39.5% of the voting-eligible population in favor of voting for the Democratic Party.

Conclusion

Before we formally state our final prediction, we can make some observations from the results of each bin split and attempt to provide some explanations for support. The most interesting subgroup that stands out is the massive discrepancy for the racial demographic, as we observe that any non-white racial group is estimated to have a sub-30% likelihood of voting for Donald Trump whereas individuals of the white ethnicity group is estimated at a 45.02% likelihood of voting for Trump. To contrast, the estimated proportion of the white ethnicity group to vote for Biden is also the lowest among any racial group at 35.14%. To expand upon this, we draw upon a paper on the Harvard Institute of Politics on the role ethnicity plays in political attitudes, which studied the statistics behind former President Barack Obama's approval ratings in regards to race. It is shown that in this study that race was the number one factor, particularly among young Americans, when it comes to influencing politics with a 31% Obama approval rating among white Americans versus a 78% Obama approval rating for African-Americans. This study complements the results of our current analysis

calculations supporting just how big of a role race plays in politics in regards to this upcoming election. We can conclude that although it is not the sole factor, a racial divide is still evident in America when it comes to their projected odds of voting for either political party.

Although there are also small differences in voting predictions for the other stratification bins regarding education and work status, the results are nowhere near as egregious as the clear evidence displayed by the wide range of estimations for the racial sub-groups, which is the most important takeaway. Overall, based on our analysis summary and prediction of 39.3% of the population voting for Trump and 39.5% of the population voting for Biden, we will make the prediction that the Democratic Party will win the popular vote for the 2020 American election by a narrow margin.

Weaknesses

One of the limitations of this analysis comes from the fluid situation of the unorthodox year 2020 has been. The datasets we are currently using for our model and census data are not completely up-to-date and in a world where opinions are easily swayed by the unpredictable nature that has been the pandemic, this can lead to outdated data. For instance, there is a strong argument that can be had to show that the COVID-19 crisis has altered Americans' view on the election, as there have been sharp declines in Donald Trump approval rates based on his actions as the pandemic has raged on, based on a project conducted by Bycoffe, Groskopf and Mehta. This may be reasonable given how it is well-known that Trump has prioritized the US economy over taking extreme safety precautions in regards to COVID-19. The point here is that the rapidly evolving real-world situation we are in could result in a higher limitation than usual when it comes to the accuracy of any model analysis.

Furthermore, there is also the possibility of external circumstances that goes beyond our scope of analysis, which is potential controversy that Donald Trump has hinted towards throughout his presidential campaign along with his past history of the 2016 campaign, in which Trump made several unusual claims and even "questioned the integrity of the electoral process". This could of course combine to be a meaningful factor that once again ties into the results of the current election but nevertheless this cannot be a priority for this report as it is something that is out of our realm of control in statistics.

When it comes to areas of improvement, there is a case to be made to increase the flexibility of our current model by increasing the number of predictor variables. Although this could potentially be an improvement on the model and the number of predictors to use being a decision we discussed extensively, the trade-offs with a higher number of variables comes with an unreasonably high number of sub-groups. This could cause major issues when it comes to reliable analysis and displaying results. Ultimately, we settled on using our three input variables (work status, education level, race) that we believed would have significant influence on predicting the election results.

Next Steps

An interesting alternative that we considered was stratifying the sample data by states in order to predict the electoral vote outcome. We did not end up going this route as it can be argued that this approach is better suited for a different model (i.e. multi-level modeling) and obtaining an overall prediction for the election winner would require summing up the number of state victories by both election parties rather than finding an overall percentage of the population voting for one particular party. Since we went with the latter approach, it could be reasonable to re-do the calculations using state analytics and a multi-level model to see if a different projected winner would be the result, given how close the popular vote was projected to be. In addition, the logical next step would be to await the results of the 2020 election which is set to conclude on November 3rd, 2020. Following this, we can contrast our analysis from this report with the actual results and attempt to identify high-influence factors that were left unaccounted for through additional research. After gaining such insights, we could conduct a follow-up survey for people who voted in this election asking for some of the same input variables we used, which will help us determine any changes in votes.

References

Bycoffe, A., Groskopf, C., & Mehta, D. (2020, November 1). How Americans View The Coronavirus Crisis And Trump's Response. Retrieved from https://projects.fivethirtyeight.com/coronavirus-polls/

Dassonneville R., Tien, C. (2020, October 15), Introduction to Forecasting the 202 US Elections, Retrieved from https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/introduction-to-forecasting-the-2020-us-elections/78235400F6BB7E2E370214D1A2307028

Le, J. (2018, April 10). Logistic Regression in R Tutorial. Retrieved from https://www.datacamp.com/community/tutorials/logistic-regression-R

Logistic Regression | R Data Analysis Examples. (n.d.). Retrieved from https://stats.idre.ucla.edu/r/dae/logit-regression/

McCammon, S. (2016, November 05). Donald Trump Has Brought On Countless Controversies In An Unlikely Campaign. Retrieved from https://www.npr.org/2016/11/05/500782887/donald-trumps-road-to-election-day

Nationscape Data Set. (2020, September). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set

Race and Ethnicity Still Play a Role In Political Attitudes. (n.d.). Retrieved from https://iop.harvard.edu/race-and-ethnicity-still-play-role-political-attitudes

U.S. Census Data for Social, Economic, and Health Research. (n.d.). Retrieved from https://usa.ipums.org/usa/index.shtml

Wang, W., et al. Forecasting Elections with Non-Representative Polls. International Journal of Forecasting (2014). Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf