Biden or Bye-den? Predicting the 2020 US Presidential Election

Karrie Chou (Student # 1005149153), Inaara Virani (Student # 1004596332)

November 2, 2020

Code and data supporting this analysis is available at: https://github.com/karriechou/sta304ps3

Data

The data to create this model came from the Democracy Fund UCLA Nationscape dataset, which was compiled based on a survey of public opinion regarding the current American presidency that was conducted in two separate phases throughout 2020. In order to make it usable for this analysis, we cleaned the data in the following ways.

From this dataset, we only selected variables which we would use in our final model construction. These variables include:

- vote_2020, a categorical variable indicating what a survey respondent indicated as their vote intent.
- employment, a categorical variable indicating a survey respondent's employment status at the time of taking the survey.
- gender, a categorical variable indicating the identified sex of the respondent.
- hispanic, a categorical variable that allowed respondents to indicate their specific Hispanic background, if they had one.
- education, a categorical variable indicating the highest educational attainment of a respondent.
- state, a respondent's state of residency.
- age, the age of a respondent.

From here, each variable, with the exception of vote_2020 and state, had its possible values adjusted so that their values matched corresponding ones from the IPUMS 2018 5-year American Community Survey (ACS) data, a set of census data used within this report's post-stratification analysis. These variables' values were transformed into numeric values to allow for a more streamlined summary output from the model.

For state, the values used in the model remained as factors. The state data in the IPUMS 2018 5-year ACS data was adjusted so that its values reflected state codes, which is how the Democracy Fund UCLA Nationscape dataset represented respondents' states of residency.

vote_2020 is used to create a binary dummy variable which is used as the response variable in this report's proposed model for the likelihood that an individual will vote for Joe Biden in the upcoming election. Further detail on how this was achieved is described in the Model section of this report.

Model

In the following sections, we describe a model which is used to predict the result of the 2020 United States presidential election. In the following subsections, we will describe the model's specifics as well as a post-stratification calculation to predict the proportion of voters in the United States who will vote for the Democrat candidate Joe Biden.

Model Specifics

The constructed model will be a logistic model based on Bayesian statistical inference.

Generally, a logistic model can be written as:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$

where the dependent variable of the model is a binary variable that has only 2 possible outcomes, a "success" and a "failure". The dependent variable has an associated probability p of being a "success". Each of the k predictor variables in the model $X_1, X_2, \ldots X_k$ can either be a numeric variable or a categorical variable with multiple factors.

For this model, the predictor variables are:

- X_1, employment status
- X_2, education attained
- X_3, sex
- X_4, if the respondent is Hispanic
- X_5, age

The response variable, vote_biden, is a binary variable which takes the value '1' for a vote for Biden and a '0' otherwise. This response variable was created by taking the vote_2020 variable from the cleaned Nationscape data set and using the tidyverse mutate() function to create the vote_biden variable, which translates a value of 'Biden' in the vote_2020 variable to a value of 1 for vote_biden. For any other values, vote_biden takes a value of 0.

Post-Stratification

The post-stratification calculation used for this model will be done by separating the sample into different groups by state and using the model to estimate the proportion of voters in each group who would vote for Biden. Then, a weighted average of each of these proportions is calculated to give us an estimate of the total population proportion of voters who will vote for Biden.

To create each group, or "bin", the age variable in the IPUMS Census data is used. From there, the values for the variables EMPSTAT, EDUC, SEX, HISPAN, and AGE of each recorded person in the IPUMS Census data are substituted into the appropriate predictor variables in the constructed model in order to obtain an estimated likelihood that a respondent would vote for Biden. These individual estimates within each bin are then added together and the sum is divided by the size of the bin, giving an average likelihood that a bin will vote for Biden. Then, each average likelihood will be multiplied by the proportion of the population that the bin comprises, and each of these weighted values will be added together to obtain a population estimate for the likelihood that Biden will be voted into the American presidency.

Using age for the group divisions is appropriate because age is a continuous variable and no individual can be two ages at once, making clustering very straightforward. The state variable, while initially included in the cleaned IPUMS 2018 5-year ACS census data, was ultimately removed from the post-stratification analysis and the final model due to technological constraints; including it would have resulted in a very large dataset which the predict() function used later to make predictions for each cluster's likelihood of voting for Joe Biden would not have been able to efficiently or effectively run.

Results

Model Summary

```
kable(summary)
```

predictor_variables	coef_estimate	standard_error	t_value	p_value
(Intercept)	0.4220393	0.0568489	7.423881	0.0000
employment_status	-0.0249758	0.0233882	-1.067881	0.2856
educ	0.0113639	0.0032860	3.458301	0.0005
sex	0.1315442	0.0137554	9.563126	0.0000
hispan	0.0147851	0.0078542	1.882444	0.0598
age	-0.0033021	0.0004221	-7.823065	0.0000

From the summary statistics, we obtain the estimated logistic regression model

$log(\frac{p}{1-p}) = 0.422 - 0.025X_1 + 0.011X_2 + 0.132X_3 + 0.015X_4 - 0.003X_5$

Assuming an $\alpha = 0.05$ significance level, the variables educ, sex, and age are the most significant ones in the proposed model. It is important to note that this is a very weak model; looking at the summary statistics, we see that the adjusted R-squared value is approximately 0.03. If the previously removed state variable were included, the adjusted R-squared would increase to 0.047. This is likely due to the large amount of degrees of freedom within the model allowing for a very large residual standard error. This model is however very practical based on the data used to complete the analysis; many of the variables in the Democracy Fund UCLA Nationscape dataset are categorical and include many levels which do not perfectly correspond to the levels for their corresponding IPUMS 2018 5-year ACS census data variables. The Discussion section will go over some of the larger weaknesses in this model as well as next steps to take if this experiment were to be repeated.

Post-Stratification Calculation Results

```
census_data$logodds_estimate <-
  model %>%
  predict(newdata = census_data)

census_data$estimate <-
  exp(census_data$logodds_estimate)/(1+exp(census_data$logodds_estimate))

census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict = sum(alp_predict_prop)/sum(n))
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.622
```

The post-stratification calculation yields an estimate of Joe Biden winning 62.2% of the vote. This value is based on a post-stratification analysis of the proportion of voters in favour of electing Joe Biden as the

next President of the United States, which was modelled by a logistic model which analyzed the effects of a person's employment status, highest level of education attained, sex, Hispanic background, and age.

The post-stratification calculation, as described in the Model section, is essentially a weighted average of the estimates of likelihood that each demographic group will vote for Biden based on the model's predictions of their behaviour. Therefore, we can assume that this is a statistically sound estimate for how the American population will vote in the upcoming 2020 election.

Discussion

Summary

To predict the result of the 2020 election for the United States presidency, a logistic model with predictor variables employment status, highest level of education attained, sex, Hispanic background, and age was used.

To account for potential non-response bias in some of the questions, as well as to ensure that our final prediction is representative of all those who are surveyed, a post-stratification analysis that clustered respondents according to age was conducted. For the post-stratification analysis, 2018 5-year ACS census data from IPUMS USA was used to estimate the proportions that different age groups in the United States made up of the entire population. From there, each unique group's values for other predictor variables were substituted into the constructed logistic model to calculate an estimate of the likelihood that that group would vote for Joe Biden. Finally, a weighted average of each of these likelihood estimates was calculated to give the final prediction that Joe Biden would win 62.2% of the primary vote in the upcoming election.

Conclusion

Based on the final prediction of Joe Biden winning 62.2% of the primary vote in the upcoming election, we predict that the Democratic Party will therefore win the election.

Weaknesses and Next Steps

The main weakness of the study is that using historical voter survey data and census data does not accurately capture the volatile and unpredictable nature of the 2020 US election. The use of historical data follows the philosophy of frequentist inference in statistics. With frequentist inference, there is an emphasis on using historical data to identify long-term trends. However, an election cycle is a non-repeatable event in which many factors come into play to determine its outcome. A Bayesian approach to statistical inference would be more appropriate in this situation, where focus is placed on achieving a plausible distribution for the value of the proportion of the population who would vote for Joe Biden in the upcoming election. By emphasizing that the true population proportion is not a fixed value, Bayesian inference better captures the uncertainty of an election cycle and may lead to a more well-founded prediction of election results.

Another weakness we found is that the assumptions for some of the variables might not be true. For example, the variable 'education' has multiple meanings since it's hard to determine what level of education and degree of literacy is necessary for someone to be considered educated. For many of the variables, translating their categorical values into numeric levels was a difficult task as there were not many times where the categorical values given by the Democracy Fund UCLA Nationscape perfectly matched up with the levels given by the accompanying codebook for the IPUMS 2018 5-year ACS census data.

The next steps would be to try and gather real time data about the individuals to help come up with a model to assist in the prediction of the presidential data rather than using historical data. This would mean collecting data about individuals at the time of voting. Another step would be to try and use the same model with different and more variables in order to see how the results vary from model to model and to determine

what are the most influential variables, or even to find a voter survey-based dataset whose variables more closely relate to the measurements used by the IPUMS 2018 5-year ACS census data.

References

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from voterstudygrou.org/downloads?key=6ecf3cee-e4e5-411c-b18d-92b39aff1da4

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0