Title: “Undecided Voters Hold Key To 2020 Election” Author: “Nick Callow, Jessica Glustein, Olivia Bi, Min Zhang” Date: “November 2, 2020” Abstract: “On November 3, 2020, Americans will go to the polls to elect their next President. Republican incumbent Donald Trump is seeking re-election against Democratic Nominee and former Vice-President Joe Biden. In this paper, we strive to answer one question, who will win one of the most contested presidential elections in history? In service of this goal, we employ three logistic regression models with post-stratification. Comparing these three results, we determine there is no clear winner of the popular vote, and undecided voters will play a large role in the outcome. Code and data supporting this analysis are available at https://github.com/jessglustien/sta304-p3” output: html_document: df_print: paged pdf_document: default

Model

Our election prediction rests on three separate logistic regression models, each with different assumptions about undecided voters. We selected logistic regression, given the binary nature of our outcome variables (Caetano, 2020a). Likewise, we chose a Frequentist over the Bayesian approach since we have no prior information about the distribution of the variables of interest (Caetano, 2020b). The first model examines the strength of each candidate’s base. As such, sampled individuals who indicated no known preference for either candidate, known as undecided voters, are excluded from the analysis. The second model assumes that all undecided voters in the sample data swing to Trump, while the third makes the opposite assumption. Please refer to Table 1 for a more thorough breakdown of the three models.

Table 1: Model Breakdown

Model	Outcome Variable	Reference Level	Description
Model I	Vote Biden	Vote Trump	In this model, we exclude undecided voters from the analysis.
Model II	Vote Biden	Vote *Not* Biden	In this model, we assume that Trump receives all of the undecided voters. Therefore, *Not* Biden includes sampled individuals who indicated they would vote for Trump and people who responded with “I don’t know.”
Model III	Vote Trump	Vote *Not* Trump	In this model, we assume that Biden receives all of the undecided voters. Therefore, *Not* Trump includes sampled individuals who indicated they would vote for Biden and people who responded with “I don’t know.”

Model Specifics

Our explanatory variables are static across all three models to make useful comparisons between these approaches. Therefore, each model takes the following mathematical form. Please refer to Table 2 for a breakdown of each variable and its associated categories, if applicable.

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \underbrace{\left(\hat{\beta}_3 x_{3,1} + \cdots + \hat{\beta}_8 x_{3,6}\right)}_{\text{Race Categorical Variables}} + \underbrace{\left(\hat{\beta}_9 x_{4,1} + \cdots + \hat{\beta}_{17} x_{4,10}\right)}_{\text{Education Categorical Variables}}$$

Results

Trump or Biden:

```
census_data$logodds_est <-
  model_1 %>%
  predict(newdata = census_data)

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

model_1_result <-
  census_data %>%
  mutate(biden_win = estimate*n) %>%
  summarise(biden_or_trump = sum(biden_win)/sum(n))
model_1_result
```

```
## # A tibble: 1 x 1
##   biden_or_trump
##            <dbl>
## 1          0.507
```

Trump or not Trump:

```
census_data$logodds_estimate <-
  model_2 %>%
  predict(newdata = census_data)

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

```
census_data %>%
  mutate(trump_predict_prop = estimate*n) %>%
  summarise(trump_predict = sum(trump_predict_prop)/sum(n))
```

```
## # A tibble: 1 x 1
##   trump_predict
##           <dbl>
## 1         0.423
```

Biden or not Biden:

```
census_data$logodds_est <-
  model_3 %>%
  predict(newdata = census_data)

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

model_3_result <-
  census_data %>%
  mutate(biden_predict_prop = estimate*n) %>%
  summarise(biden_predict = sum(biden_predict_prop)/sum(n))
model_3_result
```

```
## # A tibble: 1 x 1
##   biden_predict
##           <dbl>
## 1         0.430
```

Based on three models above, we estimate that the proportion of voters in favor of voting for Joe Biden to be 0.507. This is based off our post-stratification analysis of the proportion of voters in favor of Joe Biden modelled by a logistic regression model, which accounted for age, sex, race and education.

```r
library(sjPlot)
library(sjmisc)
```

```
##
## Attaching package: 'sjmisc'
```

```
## The following object is masked from 'package:purrr':
##
##     is_empty
```

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
## The following object is masked from 'package:tibble':
##
##     add_case
```

```r
library(sjlabelled)
```

```
##
## Attaching package: 'sjlabelled'
```

```
## The following objects are masked from 'package:sjmisc':
##
##     to_character, to_factor, to_label, to_numeric
```

```
## The following object is masked from 'package:forcats':
##
##     as_factor
```

```
## The following object is masked from 'package:dplyr':
##
##     as_label
```

```
tab_model(model_1)
```

	biden or trump		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.03	0.25 – 4.23	0.962
age	0.99	0.99 – 0.99	**<0.001**
sex [male]	0.64	0.57 – 0.72	**<0.001**
race [black/african american/negro]	9.90	5.57 – 17.81	**<0.001**
race [chinese]	4.39	2.02 – 9.98	**<0.001**
race [japanese]	4.87	1.53 – 18.90	**0.012**
race [other asian or pacific islander]	2.25	1.23 – 4.16	**0.009**
race [other race, nec]	2.52	1.43 – 4.49	**0.002**
race [white]	1.16	0.69 – 1.97	0.586
education [Associate Degree]	1.75	0.46 – 6.76	0.403
education [College Degree (such as B.A., B.S.)]	1.56	0.41 – 5.97	0.505
education [Completed some	1.39	0.37 – 5.34	0.618

college, but no degree]			
education [Completed some graduate, but no degree]	1.58	0.41 – 6.20	0.502
education [Completed some high school]	1.01	0.26 – 3.89	0.992
education [Doctorate degree]	0.95	0.24 – 3.81	0.941
education [High school graduate]	1.08	0.28 – 4.16	0.906
education [Masters degree]	1.46	0.38 – 5.62	0.573
education [Middle School - Grades 4 - 8]	1.42	0.27 – 7.74	0.677
education [Other post high school vocational training]	1.03	0.27 – 4.01	0.967
Observations	5200		
R^2 Tjur	0.107		

```
tab_model(model_2)
```

	vote trump		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.70	0.17 – 2.81	0.603
age	1.01	1.01 – 1.02	**<0.001**

sex [male]	1.56	1.40 – 1.74	**<0.001**
race [black/african american/negro]	0.15	0.09 – 0.25	**<0.001**
race [chinese]	0.29	0.14 – 0.60	**0.001**
race [japanese]	0.31	0.08 – 0.94	0.053
race [other asian or pacific islander]	0.58	0.34 – 1.01	0.053
race [other race, nec]	0.51	0.31 – 0.85	**0.009**
race [white]	1.10	0.70 – 1.76	0.670
education [Associate Degree]	0.43	0.11 – 1.62	0.202
education [College Degree (such as B.A., B.S.)]	0.49	0.13 – 1.83	0.274
education [Completed some college, but no degree]	0.49	0.13 – 1.86	0.285
education [Completed some graduate, but no degree]	0.49	0.13 – 1.87	0.282
education [Completed some high school]	0.64	0.17 – 2.42	0.499
education [Doctorate degree]	0.83	0.21 – 3.24	0.780
education [High school graduate]	0.57	0.15 – 2.15	0.396
education [Masters	0.57	0.15 – 2.15	0.396

degree]			
education [Middle School - Grades 4 - 8]	0.46	0.09 – 2.29	0.340
education [Other post high school vocational training]	0.69	0.18 – 2.64	0.582
Observations	6101		
R^2 Tjur	0.093		

```
tab_model(model_3)
```

	vote biden		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.78	0.19 – 3.07	0.713
age	1.00	0.99 – 1.00	0.126
sex [male]	0.71	0.64 – 0.79	**<0.001**
race [black/african american/negro]	5.63	3.44 – 9.43	**<0.001**
race [chinese]	3.29	1.70 – 6.53	**0.001**
race [japanese]	4.44	1.59 – 13.84	**0.006**
race [other asian or pacific islander]	2.10	1.23 – 3.67	**0.008**
race [other race, nec]	2.14	1.29 – 3.62	**0.004**
race [white]	1.30	0.81 – 2.12	0.287

education [Associate Degree]	1.02	0.27 – 3.84	0.970
education [College Degree (such as B.A., B.S.)]	1.02	0.27 – 3.80	0.976
education [Completed some college, but no degree]	0.83	0.22 – 3.10	0.778
education [Completed some graduate, but no degree]	0.94	0.25 – 3.57	0.924
education [Completed some high school]	0.62	0.17 – 2.34	0.472
education [Doctorate degree]	0.69	0.18 – 2.68	0.585
education [High school graduate]	0.62	0.17 – 2.33	0.468
education [Masters degree]	1.04	0.28 – 3.90	0.951
education [Middle School - Grades 4 - 8]	0.91	0.19 – 4.36	0.902
education [Other post high school vocational training]	0.72	0.19 – 2.74	0.625
Observations	6101		
R^2 Tjur	0.068		

Discussion

Our first model predicted that 50.65% of voters are going to vote for Biden. This is under the assumption that all voters are either voting for Biden or for Trump, and all undecided voters were not included. This implies that 49.45% of voters are going to vote for Trump. These values are so close, that it is hard to say with confidence which candidate will receive more votes. While this model did not produce a definitive winner, some interesting observations can still be found. The male coefficient is negative, meaning that being male decreases your chance of voting for Biden in comparison to being female. Age is also negatively correlated with voting for Biden, so the older a voter is, the less likely they are to vote for Biden, and in this model the more likely they are to vote for Trump.

Our second model assumed all undecided voters would not vote for Biden. This model predicted that Biden would receive 42.98% of the popular vote. A similar trend can be seen in the fact that male voters are less likely to vote for Biden than female voters, and age continues to be negatively correlated with voting for Biden.

Our third model assumed all undecided voters would not vote for Trump. This model predicted that Trump would receive 42.30% of the popular vote. Here we can see that the opposite relationship from the first two models is true. The male coefficient is positive, showing that men are more likely to vote for Trump than women, and age is positively correlated with voting for Trump.

One area that surprised us was that education did not appear to have a strong correlation with the voting result in our model. The p-values were very high for all of the education dummy variables, and there did not seem to be a clear trend among their coefficients. It has been shown that there is “a robust and positive relationship between education and political engagement” (Hillygus 2005), so it is a reasonable assumption that education would be correlated to voting habits in some fashion.

Overall our models show that the outcome of the election is going to be close. Biden is slightly ahead in our model excluding undecided voters, but the lead is so small that it could easily be shifted in either direction by undecided voters. It is clear that those undecided voters hold a large amount of power in this election, and the direction that they turn towards will shape the country for the next four years.

Weaknesses

There are limitations on the model we used. All the undecided voters have to pick one candidate since we can only have 2 outcome results in each logistic regression model. Ideally, it would be better if we can have three outcome results (Donald Trump, Biden, Other).

Another weakness is the survey was happened in June, but we used the survey data at the end of October. Since we only included people above 18, some people that were under 18 back then might have turned 18 after June and they are not included in our model.

In addition, in our model prediction, we predict the next president by popular vote. However, US president is not elected directly by citizens. Instead, they are chosen by “electors” through a process called the Electoral College. Each state has a number of electors and normally, the most popular candidate in each state will get all the votes from the electors. Thus, the real popularity nationwide might not be identical as the result of the election.

Next Steps

A possible future step would be to test for collinearity between the predictor variables used in our regression model. This would help to verify the accuracy of our model and the correlation exists between our predictor variables.

We could also Investigate on model with multiple levels outcomes since logistic model is limited because it has binary outcomes. It would reduce the inaccuracy if the model allows three or more outcomes (ex. two major candidates and other).

Going forward, we can compare the actual election results with our predicted results. Check how close the actual election result is to our result and which of the three models has the closest result to the actual result. We can also design a survey regarding the variables we chose and some additional variables to narrow down useful variables and help us pick more relevant predicted variables in the future.

References

Caetano, S. (2020a, October 5). *Introduction to Logistic Regression* [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/9309406?module_item_id=1855302

Caetano, S. (2020b, October 12). *Introduction to Bayesian Inference* [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/9490196?module_item_id=1872494

Caetano, S. (2020c, October 19). *Multilevel Regression and Post-Stratification* [PowerPoint Slides]. Quercus. https://q.utoronto.ca/courses/184060/files/9490196?module_item_id=1872494

Hillygus, D. S. (2005). The missing link: Exploring the relationship between higher education and political engagement. Political behavior, 27(1), 25-47.

Insights into the beliefs and behaviors of American voters. (2020, October 30). Retrieved November 02, 2020, from https://www.voterstudygroup.org/

Lüdecke D (2018). “sjmisc: Data and Variable Transformation Functions.” *Journal of Open Source Software*, *3*(26), 754. doi:10.21105/joss.00754 (URL: https://doi.org/10.21105/joss.00754).

Lüdecke D (2020). *sjlabelled: Labelled Data Utility Functions (Version 1.1.7)*. doi: 10.5281/zenodo.1249215 (URL: https://doi.org/10.5281/zenodo.1249215), <URL: https://CRAN.R-project.org/package=sjlabelled>.

Lüdecke D (2020). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.6, <URL: https://CRAN.R-project.org/package=sjPlot>.

Tausanovitch, C., & Vavreck, L. (2020, October 28). Nationscape Data Set. Retrieved October 30, 2020, from https://www.voterstudygroup.org/publication/nationscape-data-set

Pew Research. (2020, August 28). Trends in Party Affiliation Among Demographic Groups. Retrieved October 29, 2020, from https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/

Rohan Alexander and Sam Caetano. (2020) 01-data_cleaning-post-strat1

Rohan Alexander and Sam Caetano. (2020) 01-data_cleaning-survey1

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

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686