BIOST 579 Project

Benjamin Stan
June 11, 2021

Introduction

Background and Rationale

The topic for this analysis is voter likelihood in the United States. The US has a low turnout relative to peer countries in the Organization for Economic Cooperation and Development, most of which consists of developed democratic nations. Out of the 35 countries in this group, the US ranks 30th in voter turnout.[1] In the 2020 election, 33% of eligible voters decided not to vote, and the 2016 election saw 40% fail to participate[2]. As a result, any election can be swung by changes in turnout among key demographics. By understanding the subpopulations least likely to vote, it would be possible to create targeted interventions to increase accessibility and participation. The data used in this analysis comes from a 2020 poll conducted by FiveThirtyEight and Ipsos on attitudes towards the government and participation in voting.

Objectives

The objective of this analysis is to identify associations between characteristics of survey respondents and their likelihood to vote. Below are scientific questions to be addressed:

- Controlling for other variables such as age, sex, education level, and income, is race associated with voting patterns for respondents of the FiveThirtyEight/Ipsos survey?
- In total, 32.9% reported at least one previous barrier to voting. Controlling for other variables such as race, age, sex, education level, and income, how are previous barriers to voting associated with voting patterns for survey respondents?
- Controlling for other variables such as age, sex, education level, race, and income, is attitude towards the government associated with voting patterns for respondents of the FiveThirtyEight/Ipsos survey?

Study Methods

Study Design

The study design of interest is a cross-sectional (observational) survey.

Sample Size and Population

The sample size is 5836 survey respondents, who took the survey between September 15 and September 25, 2020. The original number of respondents was 8,327, but this was filtered down following exclusion of those whose voting record could not be matched. The survey includes 28 questions, which can contain multiple parts, along with demographic information such as age, education level, income, sex, and race. In total, there are 117 features in the data set.

[1] https://www.pewresearch.org/fact-tank/2020/11/03/in-past-elections-u-s-trailed-most-developed-countries-in-voter-turnout/
[2] https://www.fairvote.org/voter_turnout#voter_turnout_101

Recruitment

The poll was based on a nationally representative probability sample of adults with an oversampling of young adults (18-29 years old), non-Hispanic Black and African Americans, and Hispanic and Latino Americans. Sampling was done based on address using a database with access to USPS delivery information.

Missing Data

Out of the 117 features in the dataset, 22 columns had null values. These were largely due to conditional questions, such as “If you voted ‘yes’ to question 5,...” As these columns are not directly relevant to the features of interest, they will be excluded from the analysis. There are no missing values in the covariates of interest.

Statistical Principles

Outcome Definitions

The outcome to be analyzed in this study will be the voter likelihood category. As part of the data collection, the conductors of the poll categorized all respondents into the following three categories based on actual voting history: Always (31%), Sporadic (44%), and Rarely/Never (25%). Note that any respondents who were eligible to vote in three national elections or fewer were dropped from the survey. Always voters voted in all or all-but-one elections in which they were eligible since 2000. Sporadic voters voted in at least two elections in which they were eligible over the same time period, and Rarely/Never voters comprised the remainder of the population. We combined the voter categories of Always and Sporadic into one, which turned the voter category into a binary classifier.

Data Encoding

In addition to reencoding the response category into a binary variable, the covariates will also require processing in order to perform the desired analysis. Indicator variables will be used to encode the variables of race (4 categories: Black, Hispanic, Other/Mixed and White), annual income (4 categories: Over $125k, $75-125k, $40-75k, and under $40k), sex (2 categories: male and female), and education (3 categories: high school diploma or less, some college, and college degree). Age is a continuous variable which does not require modified encoding. The variable for experiencing a barrier to voting will collapse 10 survey responses into a binary variable. In total, 32.9% of respondents reported at least one of these barriers:

- Waited in line to vote for more than one hour (18.9%)
- Couldn’t get off work to vote (9.6%)
- Missed voter registration deadline (5.6%)
- Didn’t receive absentee ballot in time to vote (5.6%)
- Couldn’t find their polling place (4.9%)
- Was told their name wasn’t on the registered voter list despite registering (4.6%)
- Couldn’t physically access their polling place (4.0%)
- Had to cast a provisional ballot (3.3%)
- Was told they didn’t have correct identification (2.8%)
- Couldn’t get necessary help to fill out a ballot (1.5%)

The following question will have a binary outcome:

- Thinking about the design and structure of American government, which would you say is more in line with your view?
    - 0: Changes are not really needed

- 1: A lot of changes are needed

Analysis Methods

Multiple logistic regression will be performed in this analysis with two response categories. The predictors are listed below. The coefficients will be interpreted to understand which variables have the greatest impact on likelihood to vote, and confidence intervals and p-values will be analyzed appropriately.

The model for the logistic regression to assess the impact of race on voting patterns is shown below.

$$log - odds(P(voter)) = \beta_0 + \beta_1 * I_{black} + \beta_2 * I_{hispanic} + \beta_3 * I_{other/mixed} + \beta_4 * I_{75-125k} + \beta_5 * I_{40-75k} + \beta_6 * I_{<40k} + \beta_7 * I_{some\ college} + \beta_8 * I_{HS\ or\ less} + \beta_9 * age + \beta_{10} * I_{female}$$

Where

- $voter$ is 1 if the respondent is categorized as Sporadic or Always voting and 0 otherwise
- I_{black} is 1 if the subject was identified as black and 0 otherwise
  - Variables $I_{hispanic}$ and $I_{other-mixed}$ are similarly defined
- $I_{75-125k}$ is 1 if the respondent had an income between $75-125k annually and 0 otherwise
  - Variables I_{40-75k} and $I_{<40k}$ are similarly defined
- $I_{some\ college}$ is 1 if the respondent had some college education, but not a full degree, and 0 otherwise
  - Variable $I_{HS\ or\ less}$ is similarly defined
- age is the age of the respondent in years
- I_{female} is 1 if the subject is female and 0 otherwise

Note that the intercept captures the baseline values of all variables; this corresponds to a white male of age 0 with a college degree and over $125k in annual income. In the case of the subsequent two scientific questions, the model would be expanded to include binary variables for having experienced a barrier to voting and believing that the structure of the US government requires change, respectively. For completeness, these are listed below.

- $I_{barrier}$ is 1 if the respondent had previously experienced a barrier to voting and 0 otherwise
- $I_{changes}$ is 1 if the respondent believed changes are needed to the structure of the American government and 0 otherwise

Confidence Intervals and P-values

The confidence intervals that will be used to provide estimates for effect size will be computed using the Wald method. The p-values associated with each of the variables of interest could be computed using a likelihood ratio test, though if each scientific question requires a separate test, then a correction for multiple testing needs to be implemented. As the number of tests is relatively small (3), the Bonferroni correction appears to be a reasonable solution to maintain an overall alpha level of 0.05.

Potential Confounding Covariates

As the survey collects a wide array of information from each respondent, choosing a subset that avoids confounders will be particularly important. The initial proposal is to include all key demographic data, so as to adjust for their effect in the regression. This includes age, education level, income, sex, and race. By including all of these factors, the goal is to account for any factors that may determine answers to select survey questions. For instance, people of a certain race or income level may be more likely to encounter a barrier to

voting, which is a covariate of interest in the analysis. Despite these efforts, there will likely be confounding due to variables not represented in the data.

Results

Exploratory Analysis

As an exploratory analysis, the percent of respondents classified as always/sporadic is plotted along with the prevalence of each covariate in the charts below.

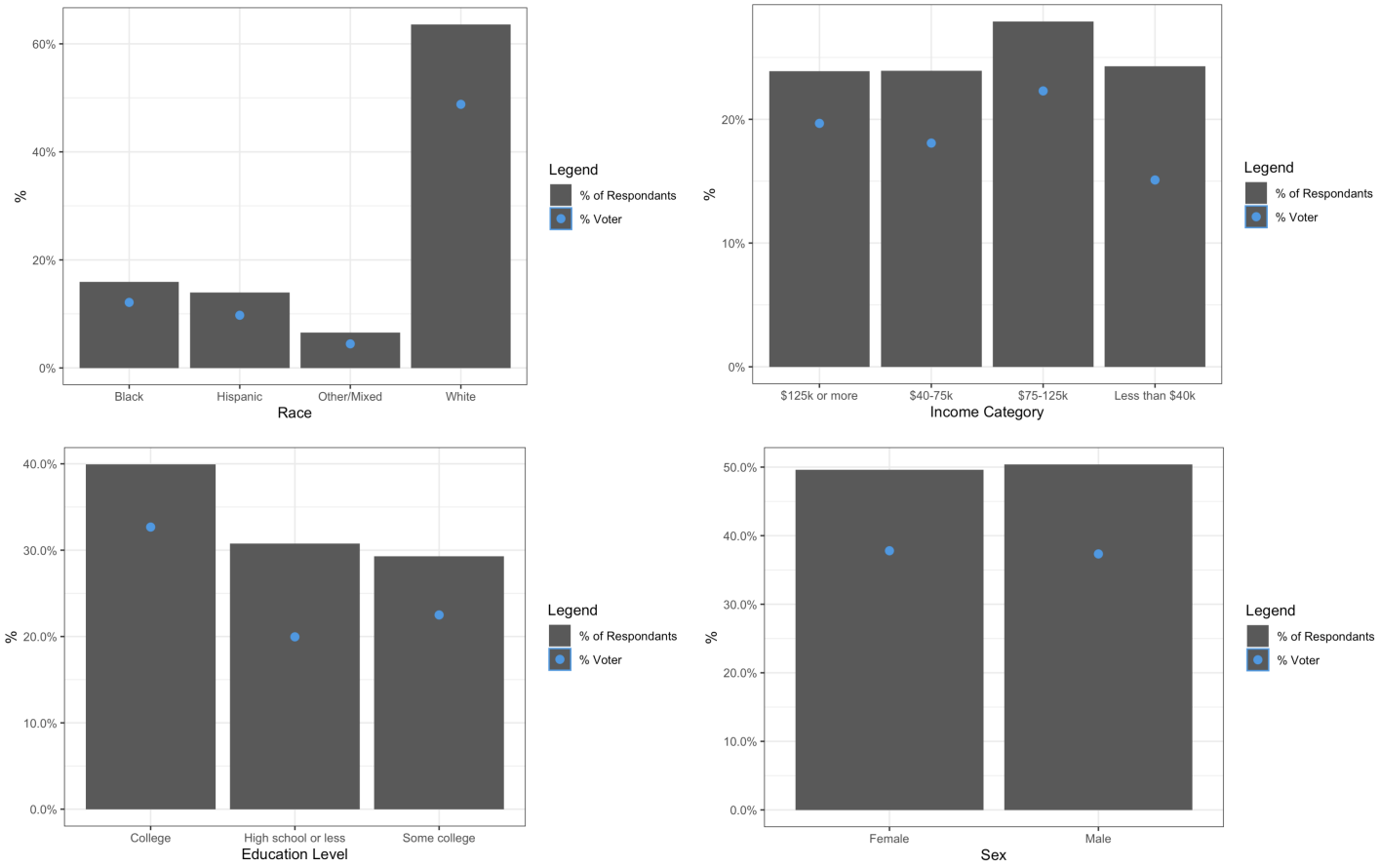

Figure 1: Percent of respondents and percent classified as always/sporadic voters by various attributes.

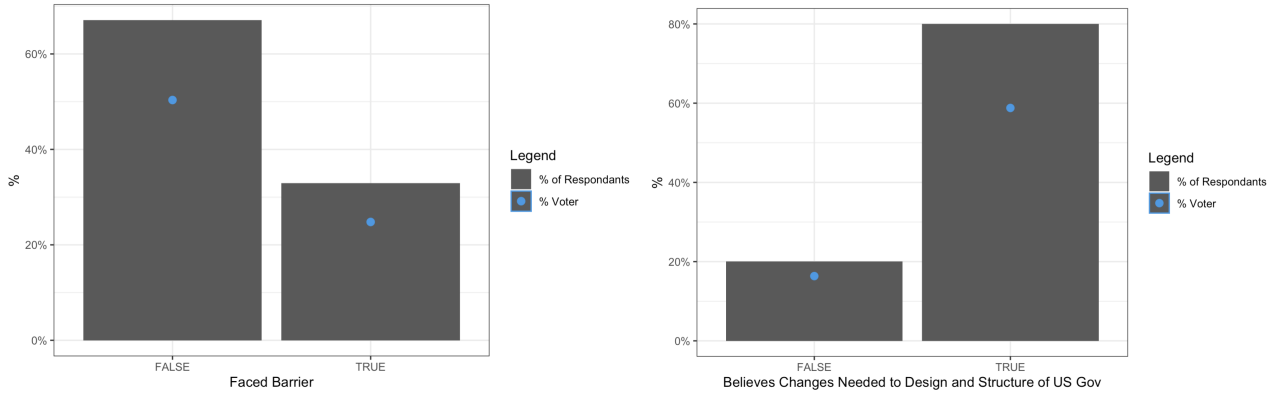

Figure 2: Percent of respondents and percent classified as always/sporadic voters by survey responses.

The correlation matrix of the variables was examined to identify any potential collinearity. This is shown below.

	I_{black}	$I_{hispanic}$	$I_{other/mixed}$	$I_{<40k}$	I_{40-75k}	$I_{75-125k}$	$I_{HS\ or\ less}$	$I_{some\ college}$	I_{female}	age	$I_{barrier}$	$I_{changes}$	I_{voter}
I_{black}	1.00												
$I_{hispanic}$	-0.175	1.00											
$I_{other/mixed}$	-0.115	-0.106	1.00										
$I_{<40k}$	0.117	0.027	-0.022	1.00									
I_{40-75k}	0.023	0.010	-0.025	-0.318	1.00								
$I_{75-125k}$	-0.034	0.005	-0.020	-0.352	-0.349	1.00							
$I_{HS\ or\ less}$	-0.009	0.039	-0.065	0.299	0.073	-0..125	1.00						
$I_{some\ college}$	0.052	0.038	-0.041	0.030	0.044	0.029	-0.429	1.00					
I_{female}	0.011	-0.036	0.000	0.084	-0.004	-0.011	0.079	0.000	1.00				
age	0.033	-0.089	-0.067	0.022	0.007	-0.033	0.102	0.015	0.053	1.00			
$I_{barrier}$	0.068	0.040	-0.002	-0.019	-0.007	0.019	-0.090	0.000	0.031	-0.080	1.00		
$I_{changes}$	0.148	0.033	0.032	0.081	0.027	-0.013	0.064	0.028	0.059	-0.103	0.076	1.00	
I_{voter}	0.008	-0.048	-0.042	-0.170	0.006	0.069	-0.158	0.025	0.024	0.316	0.002	-0.076	1.00

Table 1: Correlation matrix among features in the analysis.

The greatest correlation between any two variables is between the indicator variable for having a high school education or less and the indicator variable for earning less than $40,000 in annual salary. With a correlation value of 0.3, this will require consideration when interpreting the coefficients on these variables. As neither of these terms are included in the scientific questions, however, they should not meaningfully impact the conclusion of the analysis.

Modeling Results

Following this exploratory analysis, the initial logistic regression to determine the impact of race on voting patterns was performed. Below are the resulting coefficients from the regression.

Variable	Coefficient	95% CI Lower	95% CI Upper	P-value
Intercept	-0.784	-1.033	-0.535	$6.5*10^{-10}$
I_{black}	0.066	-0.122	0.256	0.50
$I_{hispanic}$	-0.068	-0.252	0.118	0.47
$I_{other/mixed}$	-0.414	-0.663	-0.161	0.0012

$I_{<40k}$	-0.693	-0.902	-0.486	$6.4*10^{-11}$
I_{40-75k}	-0.105	-0.311	0.100	0.32
$I_{75-125k}$	0.054	-0.143	0.252	0.59
$I_{HS\ or\ less}$	-1.08	-1.256	-0.901	$< 2.0*10^{-16}$
$I_{some\ college}$	-0.378	-0.552	-0.203	$2.1*10^{-5}$
I_{female}	0.138	0.007	0.270	0.038
age	0.052	0.048	0.057	$< 2.0*10^{-16}$

Table 2: Logistic regression coefficients for scientific question 1.

Based on the results of this regression, the estimated odds ratio of being an always or sporadic voter between two groups of otherwise equal demographic values, one white and one black, is 1.07 (95% CI: 0.89 - 1.29), with the group of black people having greater odds of voting. The corresponding odds ratios for a group of hispanic people and a group of other/mixed people relative to white people are 0.93 (95% CI: 0.77 - 1.13) and 0.66 (95% CI: 0.52 - 0.85), respectively, with the group of white people being more likely to be an always/sporadic voter. There is strong evidence to reject the null hypothesis that the odds of being an always/sporadic voter are equal across the races identified (p-value: 0.0085).

The second model was fitted in order to determine the impact of facing a barrier to vote. Below are the resulting coefficients from this regression.

Variable	Coefficient	95% CI Lower	95% CI Upper	P-value
Intercept	-0.798	-1.054	-0.543	$8.9*10^{-10}$
I_{black}	0.062	-0.126	0.253	0.52
$I_{hispanic}$	-0.071	-0.255	0.116	0.45
$I_{other/mixed}$	-0.413	-0.662	-0.161	0.0012
$I_{<40k}$	-0.694	-0.902	-0.486	$6.1*10^{-11}$
I_{40-75k}	-0.106	-0.312	0.099	0.31
$I_{75-125k}$	0.054	-0.144	0.251	0.59
$I_{HS\ or\ less}$	-1.074	-1.253	-0.896	$< 2.0*10^{-16}$
$I_{some\ college}$	-0.375	-0.550	-0.201	$2.4*10^{-5}$

I_{female}	0.137	0.006	0.269	0.040
age	0.052	0.048	0.057	$< 2.0*10^{-16}$
$I_{barrier}$	0.036	-0.104	0.176	0.62

Table 3: Logistic regression coefficients for scientific question 2.

The estimated odds ratio of being an always or sporadic voter between two groups of equal demographic values, one that faced a barrier to voting and one that did not, is 1.04 (95% CI: 0.90 - 1.19), with the group that faced a barrier having greater odds of voting. We do not have strong evidence to reject the null of equal odds of being an always/sporadic voter between those that experienced a barrier to voting and those that did not (p-value: 0.62).

Lastly, the third model was fitted to evaluate the impact of believing that the US government required change on voting patterns.

Variable	Coefficient	95% CI Lower	95% CI Upper	P-value
Intercept	-0.631	-0.919	-0.342	$1.8*10^{-5}$
I_{black}	0.092	-0.097	0.285	0.34
$I_{hispanic}$	-0.061	-0.245	0.125	0.52
$I_{other/mixed}$	-0.400	-0.649	-0.147	0.0018
$I_{<40k}$	-0.682	-0.891	-0.475	$1.3*10^{-10}$
I_{40-75k}	-0.095	-0.301	0.110	0.36
$I_{75-125k}$	0.060	-0.138	0.257	0.55
$I_{HS\ or\ less}$	-1.070	-1.249	-0.893	$< 2.0*10^{-16}$
$I_{some\ college}$	-0.372	-0.547	-0.198	$2.8*10^{-5}$
I_{female}	0.144	0.013	0.276	0.031
age	0.052	0.048	0.056	$< 2.0*10^{-16}$
$I_{changes}$	-0.189	-0.369	-0.011	0.039

Table 4: Logistic regression coefficients for scientific question 3.

The estimated odds ratio of being an always or sporadic voter between two groups of equal demographic values, one that believes change is needed to the structure of the US government and one that does not, is 0.83 (95% CI: 0.69 - 0.99), with the group that does not believe change is needed having a greater odds of voting. We do not have strong evidence to reject the null of equal odds of being an always/sporadic voter

between these two groups (p-value: 0.038). Note that this conclusion is made under the adjusted alpha value of 0.017 resulting from the Bonferroni correction.

Assumptions and Limitations

This analysis requires assumptions regarding the population being surveyed, the responses that they give, and the models that are utilized to analyze them. First, we will consider the assumptions necessary to draw broader conclusions from the respondents of this survey. As with many surveys, there may be biases in the subpopulations that are likely to respond. In order to ensure proper representation, the organizations conducting the survey deliberately oversampled certain minority groups (listed in the Methods section). This may be due to differences in response rates between demographic groups. Assuming the respondents are representative of the US population overall, there still may be selection bias among those that respond. For instance, respondents may be more likely to have a strong opinion on the issue of voting or may have had a specific experience that prompted them to participate. There is also a temporal assumption being made around sentiment at the time of the survey. The voter category for each respondent considers many elections of participation, dating back as far as decades. We are making conclusions assuming that the current sentiment of the respondent was consistent over the entire voting history period. Lastly, we are assuming that respondents answered truthfully to the questions on the survey. It may have been the case that some individuals answered questions with the socially desirable answer rather than their true sentiment. This would have affected the responses, and possibly the conclusions, regarding the attitude towards change in the US government. However, it is important to note that voting patterns were defined by actual voting history (not self-reported), so these values are not subject to response biases. Additionally, the predictors of race and experiencing a barrier to voting should not be subject to biases with the exception of the respondent outright not telling the truth.

Regarding modeling assumptions and limitations, we require all those pertaining to logistic regression. The assumption of independent responses seems reasonable, as the participants were randomly sampled from the population. One detail that may be violated is the assumption of little or no multicollinearity among covariates. The correlation matrix is shown above, and the greatest correlation between predictors is 0.29. Regardless, all features remain included due to the need to adjust for confounding variables. The final modeling assumption is a linear relationship between independent variables and the log odds of the outcome, which we believe to be valid for the continuous variable of age. The greatest limitation of the models is the context in which they are interpreted. In order to adjust for covariates, the interpretations of each coefficient in the regression are subject to the constraints imposed by other features. This is the reason why the results from the regression may seem counter to the trends in the exploratory analysis.

Discussion

The results shown by the exploratory analysis and the models may appear contradictory but require consideration of the questions being addressed. The first scientific question asks, controlling for other variables such as age, sex, education level, and income, is race associated with voting patterns for respondents of the FiveThirtyEight/Ipsos survey? To answer this, the variables associated with race were included in a multivariate logistic regression. As discussed previously, the estimated odds of being an always/sporadic voter are higher for a population of black respondents than for a white population of equal income level, education, sex and age. For populations of hispanic and other/mixed respondents, the estimated odds are lower. The results for the black population may seem counter to the exploratory analysis, which showed that white respondents were more likely to be an always/sporadic voter by a factor of three. However, this initial analysis did not consider the other variables being adjusted for in the analysis. The result of the likelihood ratio test showed that there is,

indeed, strong evidence to suggest that the difference in odds of being an always/sporadic voter is not zero when comparing races in this survey population.

The results of the second scientific question contained similar contrasting findings between initial analysis and modeling. The logistic regression found that the odds of voting increase for respondents who experienced a barrier to voting when adjusting for demographic variables. This was despite the finding that the prevalence of being an always/sporadic voter was substantially lower among those that experienced a barrier to voting. Lastly, the estimated odds of being a always/sporadic voter were lower for those who believed changes are needed to the structure of the US government. The hypothesis tests on the second two scientific questions did not find that the voting patterns differed significantly for respondents who experienced a barrier to voting or believed that changes are needed to the structure of the US government.

These findings could be used to inform voter turnout efforts. By providing summary information about voting patterns and the results of the model, political organizations could prioritize different groups for outreach in order to maximize voter turnout. For instance, knowing that Black Americans have an increased likelihood to vote when adjusting for other demographic values, investment in this population could be prioritized.

Resources

- FiveThirtyEight article: https://projects.fivethirtyeight.com/non-voters-poll-2020-election/
- Github folder with data and survey questions: https://github.com/fivethirtyeight/data/tree/master/non-voters