Who Votes?

10/20/2020

Introduction

Election forecasting is of great interest to many people. Politicians might use it to adjust campaign strategies, economists might use it to predict how markets will react, and citizens might use it to decide whether to vote. Many polls are carried out in an attempt to predict the winner of an election, however these polls often don't capture the whole picture. Polls can be biased, they can change over time, and perhaps most importantly, they don't always reveal who will show up to vote on election day. In order to have a more accurate election prediction model, voter turnout is a critical factor to consider. For example, if one candidate is predicted to win by a large margin in the polls, that might cause some people not to show up to vote on election day. Additionally, some polls include people who might not actually be registered to vote.

In this report, we seek to build a comprehensive model predicting who votes, specifically in the state of North Carolina. First, we will cover a brief literature review of other voter turnout models that have been constructed in the past. Next, we will discuss the data used to build our models. After that, we will walk through the construction of our model, and we will discuss our results.

Literature Review

Previous studies have examined different ways of estimating who votes, but there does not appear to be any general consensus in the best way of modeling this. In one study (Wislek), the author attempts to predict voter turnout rates based on an area's violent crime rate, number of US House Representatives, and several other political factors. However, this data is examined at the state level, so there might be differences across counties that aren't being captured. Another paper by Challenor states the importance of likely voter models in election forecasting, and goes onto predict voter turnout based on US Census questions. They compared several models, the best of which had 87.82% accuracy, and determined that the features most correlated with voting included education level, marital status, and major job industry, while the variables with least correlated with voting were marital status based on armed forces participation, intermediate job industry, and age. In Grofman's paper, "Models of voter turnout: a brief idiosyncratic review," he examines many old studies about voting models, and some date back to as early as 1957. He says that the two main schools of thought in voter prediction modeling are predicting turnout in terms of demographic or attitudinal factors and the other is estimating voter turnout by thinking of voting as a "rational choice calculus" where citizens estimate the costs and benefits of voting. In the end, he determines that both methods produce similar results: it is very hard to explain which individuals will vote or not, but models have found more success in predicting which categories of people vote depending on group characteristics and the election.

Data

Several data sources were used to build the model. To start, we used information from the North Carolina voter registration database. The first dataset from this source was a file containing all registered voters in North Carolina. Each row represents an individual, and we also have their current voter status, the county they live in, the day they registered to vote, what political party they're a member of, as well as

other demographic information like race, ethnicity, gender, and age. The second dataset from this source contained actual voting records from the 2016 and 2018 elections. Each row represents a vote, and includes other information like county, voter registration number, election, voting method, which party was voted for, and the precinct.

To engineer an appropriate data set relevant to the research goals, we first merged both data sets together using the voting registration numbers to have a collective dataset with all the predictors relevant to the study. We then modified the ages of voters from their specific ages to age groups, so as to align with other census data, and allow for better interpretability of this predictor when perform model analysis. Finally, we grouped the data into different groups based on our factored predictors, resulting in a dataset with 30673 rows and 7 columns. Details of the predictor variables are described in Table 1.1.

Table 1: Table 1.1 Description Table of Data set variables

Metric	**Groups**
n	Number of voters
race	A = Asian or Pacific Islander, B = Black, not of Hispanic Origin, I =, M = O = , U = , W =
gender	M = Male, F = Female, U = Preferred not to respond
ethnicity	HL = , NL = , UN =
age group	Age groups of 18-24, 25-44, 45-64, 65+
county	North Carolina's 100 counties
election date	Election dates of 11/06/2018, 11/08/2016

Performing explanatory data analysis on the data set, we noted that there are unequal representation of certain groups within each predictor variables. Figures 1.1-1.4 show histograms of the counts of observations in terms of age group, race, ethnicity and gender. From these plots, we can observe that the majority of the observations that represent voters disproportionately identify their race as white, or identify their ethnicity as HL. We have slightly more female voters than male voters with little representation outside of these two specified genders, as well as having most of the voters between the ages of 45-64.

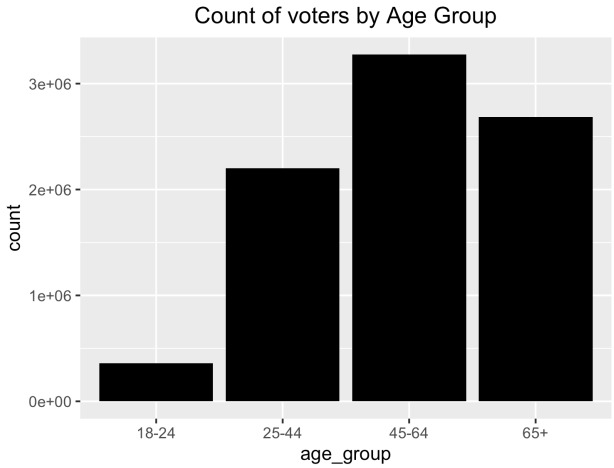

Figure 1.1 Histogram of voter counts by age group

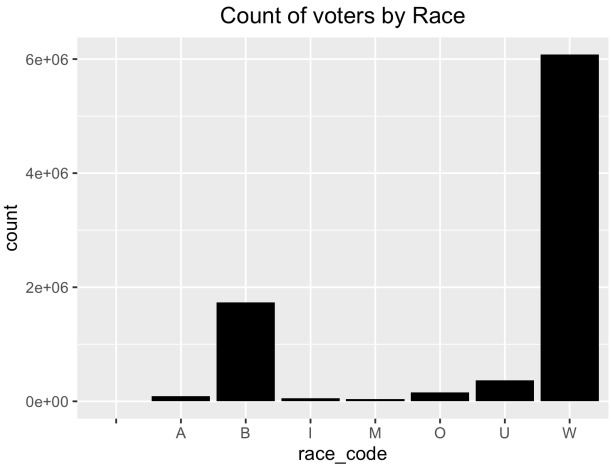

Figure 1.2 Histogram of voter counts by race

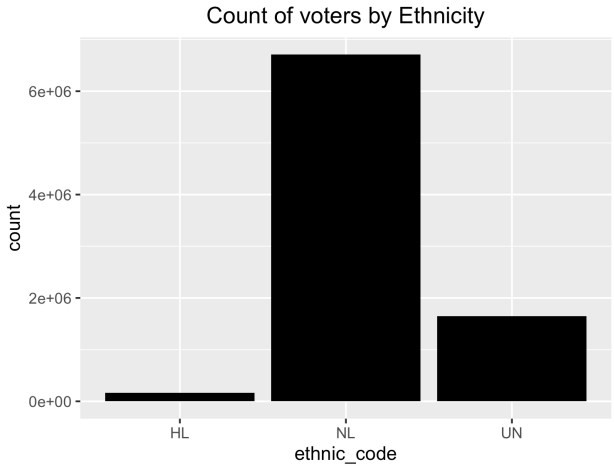

Figure 1.3 Histogram of voter counts by ethnicity

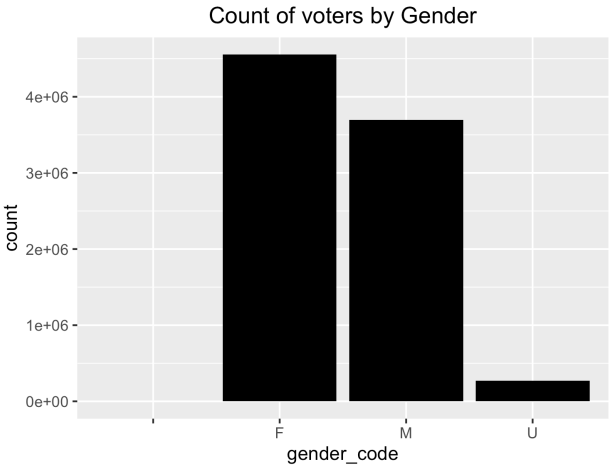

Figure 1.4 Histogram of voter counts by gender

Model

To build the model that predicts who votes specifically in the state of North Carolina, we fit a simple logistic regression model with the response variable being the number of people who voted, and the predictor variables being race, ethnicity, election date (2016 and 2018), gender and county, with all predictors being factor variables.

The model for race i, ethnicity j, election date k, gender l,county m and age group q to predict the number of people who voted $n_{voting,ijklm}$ can be written in statistical notation as:

$$n_{voting,ijklmq} \sim Binomial(n_{ijklm}, p_{ijklmq})$$

$$\begin{aligned} logit(p_{ijklmq}) = \alpha + \beta_1 * race_i + \beta_2 * ethnicity_j + \beta_3 electiondate_k \\ + \beta_4 * gender_l + \beta_5 * county_m + \beta_6 * agegroup_q \end{aligned}$$

$$\alpha \sim N(0,1), \beta_b \sim N(0,1)$$

As a start, we are using N(0,1) for all intercepts and coefficients with prior motivations being the intuitive expectation for various factors to have equal effects on whether someone votes or not. Furthermore, in order

to estimate the number of people who voted as a binomial probability, we give a number of "trials" using the number of registered voters in those groups as that number. This is so the predicted probabilities obtained from the model accurately reflect the proportion of registered voters who are actually voting. The baseline for our model predicts the number of people voting that are white, female, non-Hispanic and in the age group of 18-24.

Results

Citations

Can Likely U.S. Voter Models Be Improved? (2020, May 30). Retrieved October 22, 2020, from https://www.pewresearch.org/methods/2016/01/07/comparing-the-results-of-different-likely-voter-models/

Challenor, T. (2017, December 15). Predicting Votes From Census Data. Retrieved October 22, 2020, from http://cs229.stanford.edu/proj2017/final-reports/5232542.pdf

Grofman, B. (1983). Models of Voter Turnout: A Brief Idiosyncratic Review: A Comment. Public Choice, 41(1), 55-61. doi:https://www.jstor.org/stable/30024032?seq=1#metadata_info_tab_contents

Wislek, J. (n.d.). Predicting Voter Turnout. Retrieved October 22, 2020, from https://scholar.valpo.edu/cgi/viewcontent.cgi?article=1906&context=cus

Appendix