Georgia Voter Registration Analysis

Bilal Gilani

4/28/2021

Abstract

The analysis done within this project was to determine the factors or variables related to voter registration in Georgia. How more likely would it be for someone to register to vote if they were of a specific ethnicity, how they became U.S. citizens, the languages they speak, and finally education and economic level. The analysis showed interesting trends through the Census data, however the results of the analysis found that the variables selected were unable to create a proper model for voter registration in the state of Georgia. These variables were selected based off of qualitative intuition stemming from previous research on voter suppression strategies used in the United States. Alternate approaches are then offered on how to mitigate this issue so that further experimentation can be done via new variables.

Intro

The purpose of this project is to show how unique Georgia's 159 counties are from each other and to attempt to understand what factors/variables influence people to register to vote. According to the Pew Research Center, the United States trailed many "developed countries" in terms of voter turnout. For example, in the 2016 US Presidential Election, 245.5 million American were eligible to register to vote but just over half, 157.6 million of those were actually registered to vote [1]. The United States has an extensive history of voter registration suppression since its inception as a country. Jim Crow era laws aside, there are still voter registration suppression attempts being made today in more modern times. In 2011, for example, the state of Kansas passed a law that required citizens to show "proof of citzenship" documents in order to register to vote. Naturally, many people do not walk around with their passports or birth certificates, and over 30,000 potential registrants were unable to register to vote. In another instance, New York state required voters to registet to vote at least 25 days before the election. This law was suppresssive because in the ACLU's words own words:"By forcing voters to register before the election even becomes salient to the public, it discourages people from registering in the first place." [2]. More recently, Georgia Republicans passed a new law that created new voting registrations and voting restrictions such as adding voter ID requirements for absentee ballots, limiting access to ballot dropboxes, etc., moves that will undoubtedly effect black and brown voters [3]. Finally, voting/voter registrations centers are found too few in minority majority communities, such as in Union City, an Atlanta suburb, where of the city's 22,400 residents, 88% are black. One woman interviewed in an article for NPR waited on line for primary voting for five hours only to find the center closed and the voting machines off. Additionally, NPR research found that the nine counties of metro Atlanta "Fulton, Gwinnett, Forsyth, DeKalb, Cobb, Hall, Cherokee, Henry and Clayton — have nearly half of the state's active voters but only 38% of the polling places, according to the analysis." [4]. Voter suppression conversely affects minority communities and as such, variables such as **Race** and **Med_Income** were selected in order to determine correlations related to the number of registered voters per county.

Data

Georgia Voter Registration Data via TAP

The original data set used was the Georgia Voter Registration Data, the dataset consisted of 6.9 million observations with 25 variables. Examples of the variables in the data set were the names of registered voters and their full addresses. Of these variables, only two were needed, the number of registered voters by county (variable Reg_Voters) and the zip code (so that the two datasets could be joined correctly by county). Rather than needing the zip code, I required the name of the counties due to the later to be discussed Census Data being listed by county name. Because of the sheer amount of observations I could not go through all 6.9 million rows and place the correct county name next to each zip code. Thanks to the zipcodeR package, I successfully mapped each zip code to the associated zip codes. Finally, the Census data's "asc5" call gives 5 year estimates for variables, so since I would go on to select the year 2016, it would give me values from 2012-2016. Because of this, I subset my voter registration to only include voters who registered between the five year period of 2012 to 2016.

Tidycensus (Census Data)

Census data was acquired via the Tidycensus package, where variables are selected from a list of over 22,000 codes. Of those I selected Median Age, Gen. Population, and variables related to: Race, Citizenship, Language, and Education. As previously mentioned, the function get_acs's call gives me a five year estimate of each variable. Once the desired variables were acquired and correctly renamed, the two data sets were joined via the County variable.

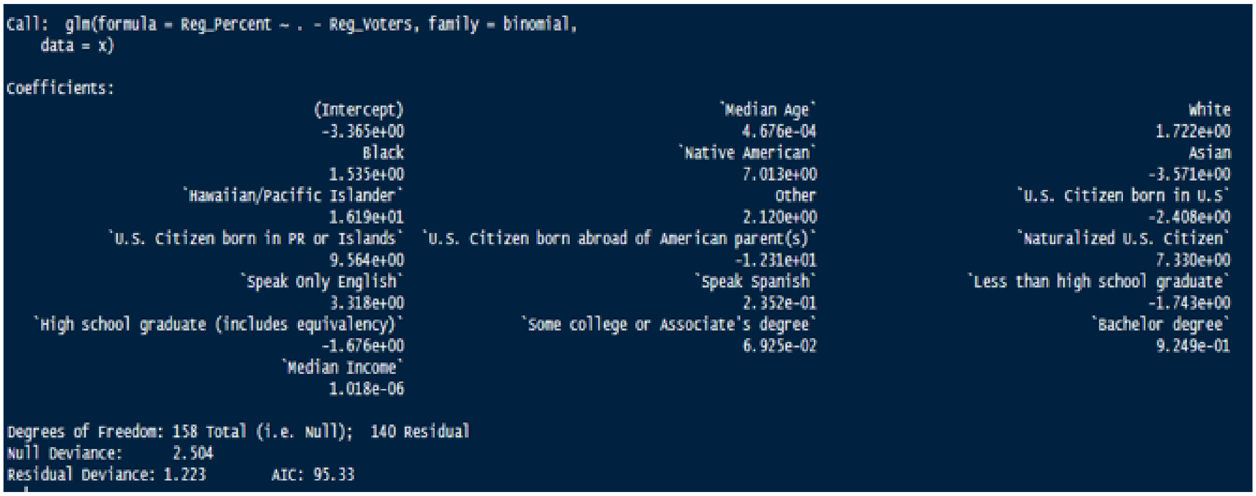
Analyses

Census Analysis

Data from the Georgia voter registration data revealed some interesting trends related to legislature passed in the United States in the 1960's and 1970's. Most notably, the drastic change in registered voters from 1975 to 1976. In 1975, just 8,471 people registered to vote in the state of Georgia. In 1976, however, 33,978 people registered to vote, this can likely be attributed to the Voting Rights Act of 1975. The Voting Rights Act of 1975 made the ban of literacy permanent, nationwide. Additionally, the Voting Rights act was amended to protect language minorities. Several civil rights organizations had argued that people of color were discriminated against when it came to voter registration. Congress prohibited laws that required ballots and voting information be exclusively in English for districts where a single-language minority group accounted for more than 5% of voting age people [5]. Additionally, between the years of 2012 and 2016, the highest number of registered voters came in years in which there was an election (2012 & 2016) as well as 2014 due to midterm elections. This trend is universal through the decades, more people register in election years.

Voter Registration Analysis

Initially, I needed to replace all the variable values in each row with ratio values, each variable has to be divided by the "Total" variable that was associated with the variable's group. For example, the race variables such as **White** or **Black** are to be divided by variables **Race_Total**. I chose to create these ratio values so that the difference in the size of counties, large or small, did not influence the results. By creating a ratio, the variables became scaled. Just as I created a ratio value for the variables that came from the Census data, I had to create a scaled value for registered voters which is **Reg_Percent**. In order to avoid perfect collinearity I drop four variables from the base model, which includes all the variables. The four variables that are dropped are: **Two or More Races**, **Not a U.S. citizen**, **Speak other languages**, **Graduate or professional degree**, or one from each group. The resulting model includes the estimate for the baseline demographic in the intercept term. Because all the dependent variable (Reg_Percent) value is bounded between 0 and 1, I ran a fractional logistic regression using binomial family (Figure A).



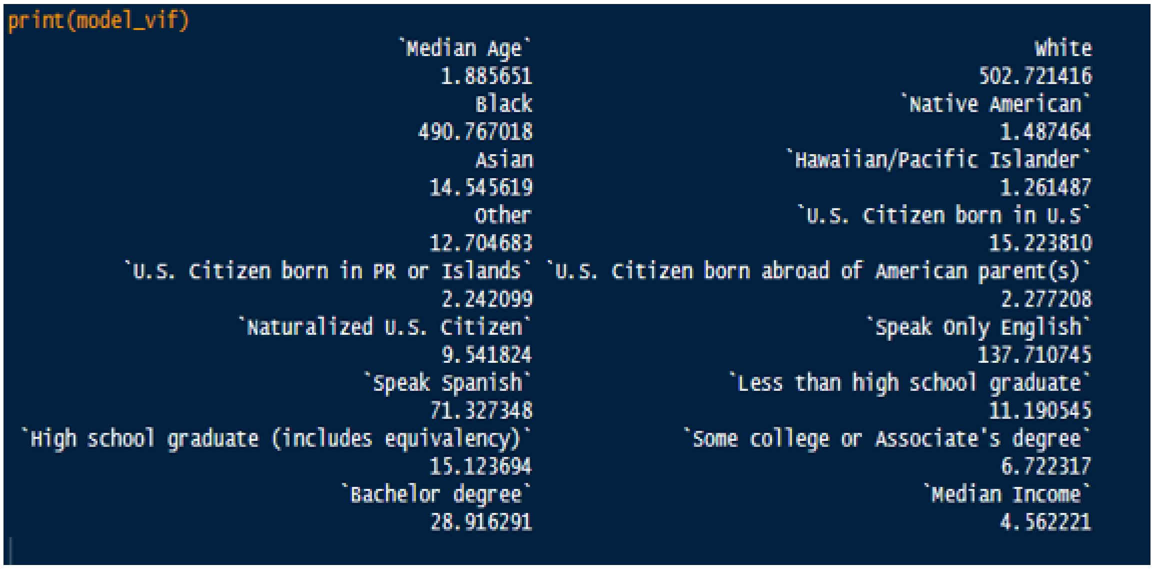
```
Call:  glm(formula = Reg_Percent ~ . - Reg_Voters, family = binomial, 
    data = x)

Coefficients:
                                   (Intercept)                                      `Median Age`                             White  
                                    -3.365e+00                                         4.676e-04                         1.722e+00  
                                         Black                                 `Native American`                             Asian  
                                     1.535e+00                                         7.013e+00                        -3.571e+00  
                   `Hawaiian/Pacific Islander`                                             Other        `U.S. Citizen born in U.S`  
                                     1.619e+01                                         2.120e+00                        -2.408e+00  
          `U.S. Citizen born in PR or Islands`  `U.S. Citizen born abroad of American parent(s)`        `Naturalized U.S. Citizen`  
                                     9.564e+00                                        -1.231e+01                         7.330e+00  
                          `Speak Only English`                                   `Speak Spanish`  `Less than high school graduate`  
                                     3.318e+00                                         2.352e-01                        -1.743e+00  
 `High school graduate (includes equivalency)`              `Some college or Associate's degree`                 `Bachelor degree`  
                                    -1.676e+00                                         6.925e-02                         9.249e-01  
                               `Median Income`  
                                     1.018e-06  

Degrees of Freedom: 158 Total (i.e. Null);  140 Residual
Null Deviance:	    2.504 
Residual Deviance: 1.223 	AIC: 95.33
```

Figure 1: Figure A

I then tested for multicollinearity (checking if the independent variables are not highly correlated with each other). Typically, a VIF above 5 indicates collinearity [6]. Many of the variables in the model show multicollinearity due to their high VIF value, this could be due to the fact that the variables are grouped together, such as **Speak English Only** and **Speak Spanish Only** which were shown to be highly correlated. After determining that the residuals mean was statistically different from zero via a t-test (-0.01158137), the model's Q-Q plot shows that the results are not normally distributed (Figure A).



```
print(model_vif)
                                 `Median Age`                                            White 
                                     1.885651                                       502.721416 
                                        Black                                `Native American` 
                                   490.767018                                         1.487464 
                                        Asian                      `Hawaiian/Pacific Islander` 
                                    14.545619                                         1.261487 
                                        Other                       `U.S. Citizen born in U.S` 
                                    12.704683                                        15.223810 
         `U.S. Citizen born in PR or Islands` `U.S. Citizen born abroad of American parent(s)` 
                                     2.242099                                         2.277208 
                   `Naturalized U.S. Citizen`                             `Speak Only English` 
                                     9.541824                                       137.710745 
                              `Speak Spanish`                 `Less than high school graduate` 
                                    71.327348                                        11.190545 
`High school graduate (includes equivalency)`             `Some college or Associate's degree` 
                                    15.123694                                         6.722317 
                            `Bachelor degree`                                  `Median Income` 
                                    28.916291                                         4.562221 
```

Figure 2: Figure B

Finally, I ran a Shapiro-Wilk test and a Kolmogorov-Smirnov test in order to check if the residuals are normally distributed. The Shapiro-Wilk test gave a p-value of less than 0.05, therefore we reject the null hypothesis that the residuals are normally distributed. Conversely, the Kolmogorov-Smirnov returned a conflicting p-value that was greater than 0.05, in which case we faily to reject the null hypothesis that

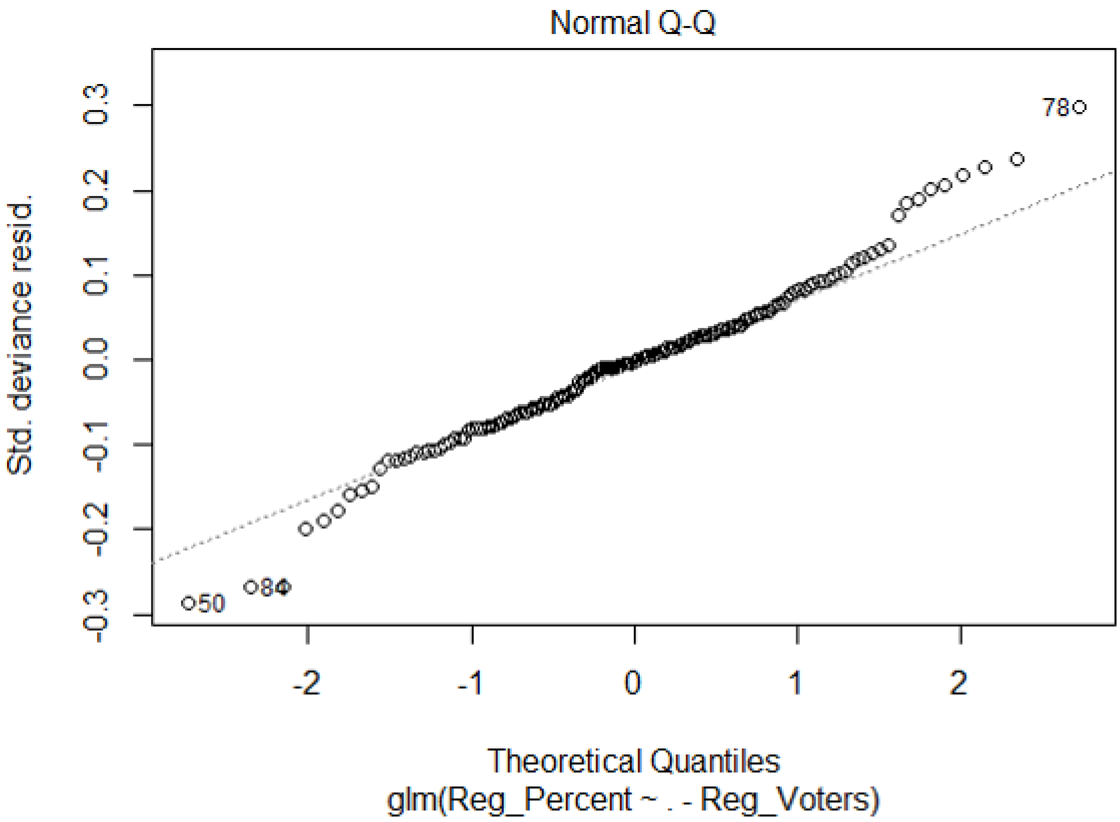
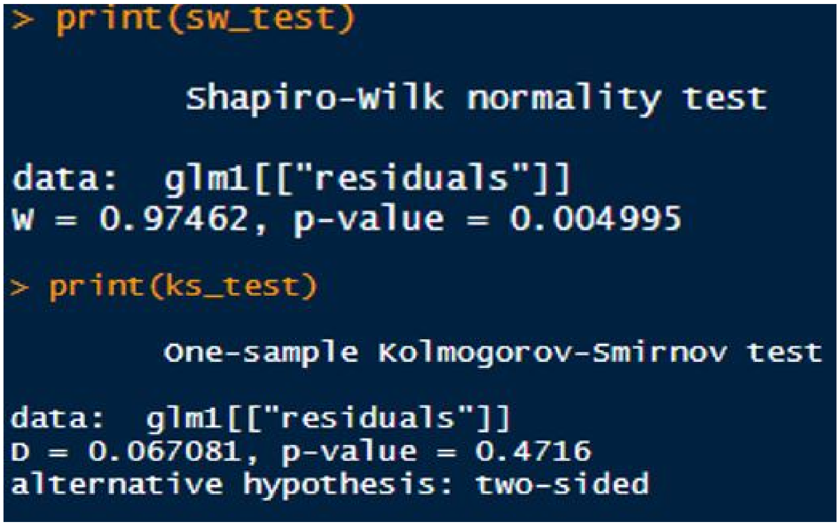


Figure 3: Figure C

the residuals are normally distributed. The Q-Q plot DID indicate that there might be a minor deviation from normality, the two tests mentioned provide different results, which do not confirm that the residuals are severely departed from a normal distribution. The residuals are statistically different from normal distribution, therefore, the p-values were correct (Figure B).



```
> print(sw_test)

        Shapiro-Wilk normality test

data:  glm1[["residuals"]]
W = 0.97462, p-value = 0.004995

> print(ks_test)

        One-sample Kolmogorov-Smirnov test

data:  glm1[["residuals"]]
D = 0.067081, p-value = 0.4716
alternative hypothesis: two-sided
```

Figure 4: Figure D

Conclusion

The VIF results show that there is multicollinearity between the independent variables, which can explain why some of the variables are statistically insignificant. Looking back this is not shocking, for example the variables for Median Income and Education level are likely correlated because logically, the higher education level a person possesses, the more they are likely to earn. Variable selection seems to have been the issue in this model, however I stand behind my initial selection of variables. As mentioned previously, there has been legislature in the past that attempted to both curb discrimination as well as stop it. Even today there are attempts to restrict voting rights, which is a strong qualitative justification for including Race, Language, and Education in the model despite not being statistically significant. The next step in this analysis would be selecting variables that have no clear correlation so that collinearity can be avoided. Examples of such variables could be “distance from registration drives”, “Wifi accessibility”, etc.

References

[1] Desilver, Drew. *In past elections, U.S. trailed most developed countries in voter turnout*, Pew Research Center, 3 Nov. 2020, www.pewresearch.org/fact-tank/2020/11/03/in-past-elections-u-s-trailed-most-developed-countries-in-voter-turnout/.

[2] *Block the Vote: Voter Suppression in 2020*, ACLU, 3 Feb. 2020, www.aclu.org/news/civil-liberties/block-the-vote-voter-suppression-in-2020/.

[3] Mansoor, Sanya. “Georgia Has Enacted Sweeping Changes to Its Voting Law. Here’s Why Voting Rights Advocates Are Worried.” *TIME*, 26 Mar. 2021, time.com/5950231/georgia-voting-rights-new-law/.

[4] Fowler, Stephen. *Why Do Nonwhite Georgia Voters Have To Wait In Line For Hours? Too Few Polling Places*, NPR, 17 Oct. 2020, www.npr.org/2020/10/17/924527679/why-do-nonwhite-georgia-voters-have-to-wait-in-line-for-hours-too-few-polling-pl.

[5] United States, Congress, House, Public Laq 94-73. *Voting Rights Act of 1975*, 6 Aug. 1975, *U.S. Government Publishing Office*, https://www.congress.gov/94/statute/STATUTE-89/STATUTE-89-Pg400.pdf

[6] Frost, Jim. “Multicollinearity in Regression Analysis: Problems, Detection, and Solutions.” Statistics by Jim, statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/#:~:text=Statistical%20software%20calculates
%20a%20VIF%20for%20each%20independent%20variable.&text=VIFs%20greater%20than%205%20represent,the%20strengt
20relationships.