North Carolina Election Analysis

Dapo Adegbile (Writer) Christopher Oblak (Presenter)
Sutianyi Wen (Coordinator,Checker) Jiaman Betty Wu (Programmer)

Oct 18, 2020

Summary

During years of elections, especially that of a presidential election year, understanding voter turnout is a powerful insight that could help either candidate in their desire to win an election. It also gives a snapshot of the type of individuals that participate in elections, giving us a sentiment of a population political leanings. The following analysis aims to gain more insight into doing just that. Looking at key factors of a voter and whether they voted, along with a voter's demographic information, to conduct analysis is the premise for the data collected. Through our research, we found that there was enough data to make most variables statistically significant, but that the overarching trends were: 1.Those of a minority tend to vote less frequently, 2.Individuals tend to vote more consistently as they age, 3.Women are more likely to vote than men, 4.Republicans are the most likely to vote.

Introduction

The data in this study is collected for county's specifically in North Carolina, in conjunction with the 2016 elections. We conduct our analysis on 20 random sample counties. We worked to provide analysis to answer critical questions about the voter population and what those factors can tell us about their propensity to vote. Understanding factors of voters turn out would allow campaigns to target groups that have lower turnout rates and could be swayed to vote where they may not otherwise. Specifically, we were interested in the trend by subgroups of demographics, both sex and race. We were also interested in if voter turn out differed by county, and if political affiliation by sex were strong voter indicators.

Exploratory Data Analysis

In order to make our data usable, for the voter and history datasets, we decided to collapse these datasets based on each demographic category and county. Our resulting datasets gave us the total number of registered voters from the voter dataset and the total number of actual voters from the history dataset. We then decided to merge these datasets based on county and demographic data. The final dataset we are using contains 9,864 observations of 9 different variables. We noticed in our new, merged dataset, there were some observations when the number of voters in the county exceeded the number of registered voters in the county. To remedy this issue we decided that for those specific observances, we set the number of voters in that county to the number of registered voters. After this we created a new variable, voter turnout rate, where we took the number of people that voted and divided by the number of registered voters.

To analyze our data more efficiently, we randomly sampled 20 counties in our data. The 20 counties we sampled are Alexander, Alleghany, Carteret, Catawba, Chatham, Chowan, Craven, Dare, Franklin, Gaston, Iredell, Macon, Nash, New Hanover, Onslow, Perquimans, Randolph, Scotland, Swain and Wake.

When plotting the distribution of our response variable, voter turnout rate, the distribution is fairly normal. But because there are some observances where our voter turnout rate is 0 or 100, there are a number of observations at both tail ends of the histogram.

When inspecting how voter turnout is impacted by our different variables, we plotted voter turnout vs age, and noticed that younger voters have a turnout rate of roughly 50%. But as the age group shifts from 26-40 to 41-65 we notice nearly a 50% increase in turnout rate. Additionally we sought to understand the relationship between voter turnout and county. While there is some variance, the median voter turnout by county typically falls within 60-70%.

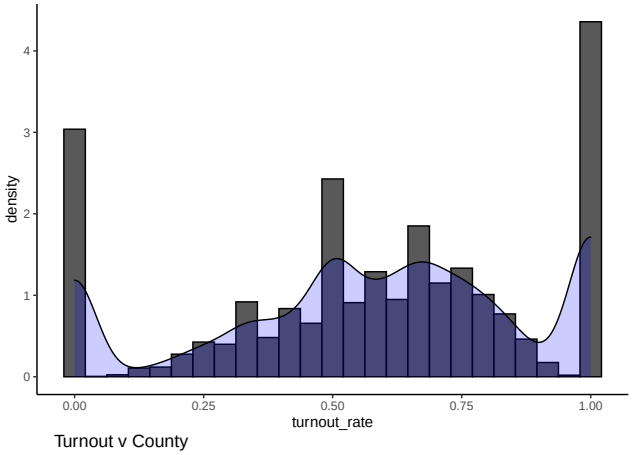
Additionally, we explored the relationship between the voter turnout and county and noticed some variance in by county. We need to conduct further statistical analysis to determine if there is statistically significant evidence to support the claim that voter turnout is impacted by county.

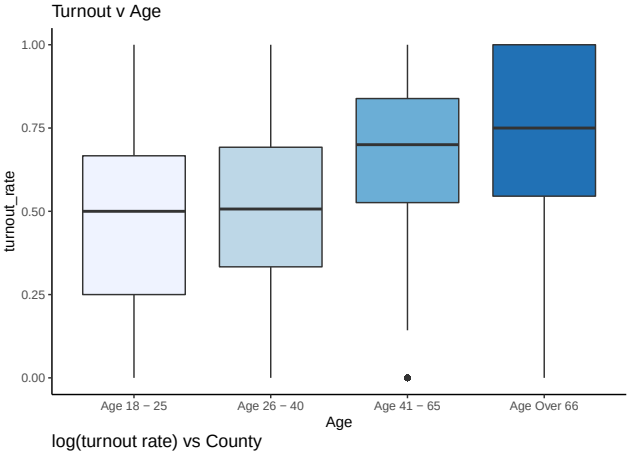
To get a better understanding of our data, we also plotted turnout rate vs county, by party. In order to explore this relationship, we randomly sampled voter data from 7 different counties. It's worth noting that for each party there is a decent amount of variance by county. But despite this, we can see that on average

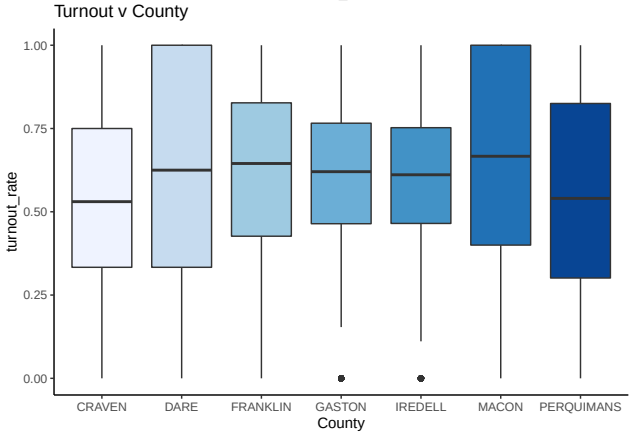
the median turnout rate for republican is higher than any other party, while on average, the median turnout rate for libertarians are lower than any party.

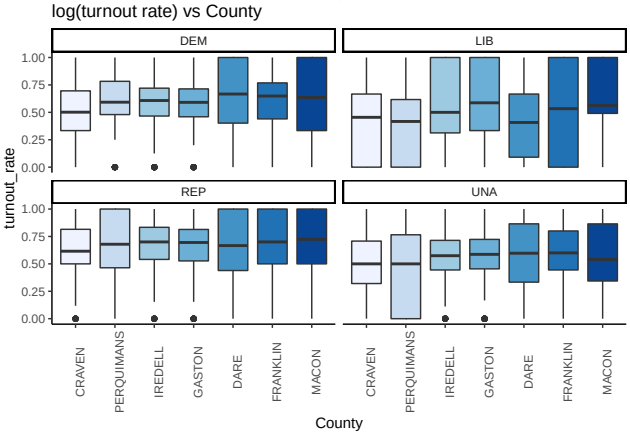
Moreover from continued exploration of our data we see that the median voter turnout rate for white and unidentified races are higher than that of any other race.

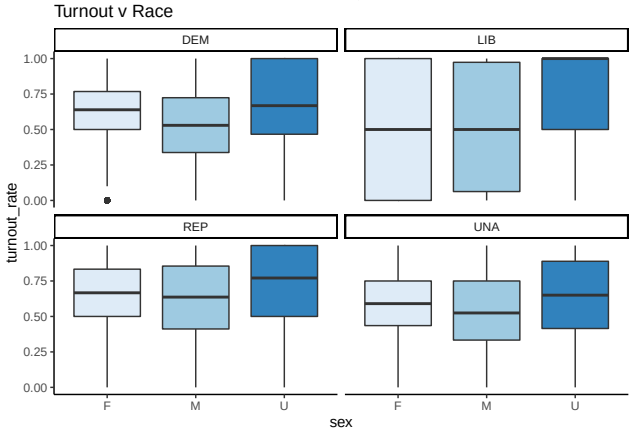
Another interaction we decided to explore was the relationship between sex and turnout rate, by party. Across all parties, people whose sex is unidentified are more likely to vote than any other race. While it appears that across all parties, males are the least likely to vote. Again, this data was taken from a random sample of 7 counties, so our results might be biased.











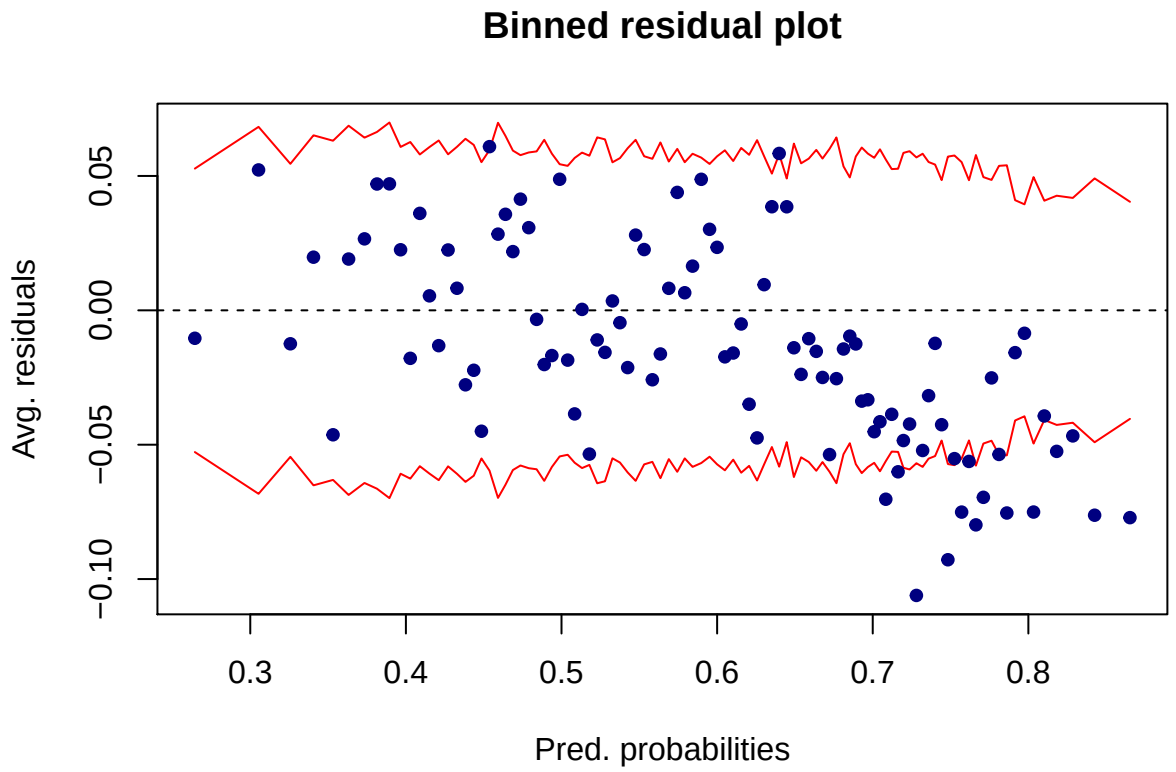
Model Building and Assessment

We decided to include all predictors from *merged* data set and one interaction terms between *sex* and *age* because we found there are different trends from EDA in our initial model. Besides these fixed predictors, we also included one grouping term to capture the varying-intercept effect of *County* since we are interested in whether the overall probability or odds of voting differ by county. After the initial model, we run ANOVA tests on each fixed predictor to see if we need to drop any fixed predictor from our model and we only drop *Party* because it's the only insignificant predictor. We then ran an AIC on the initial model and initial model without the grouping term. It turns out the AIC with grouping term is much smaller ($62180.15 < 166352.2$) than AIC with no grouping term so we have to keep the grouping term. And here comes our final model below. Because the original data set is really large and the number of predictors is pretty limited, it makes sense to include every predictor in our final model.

$$logit(turnout\_rate) = \beta_0 + \gamma_{0k[i]}^{County} + \beta_{race} race_i + \beta_{ethnic} ethnic_i + \beta_{sex} sex_i + \beta_{Age} Age_i + \beta_{sex*Party} sex_i * Party_i$$

$$\gamma_{0k} \sim N(0, \sigma^2_{County}).$$

In our binned residual plot, most of our points lie between the bounds of the 95% confidence interval, and they are roughly, randomly distributed. However there are some points that fall outside of the bounds, which signify some potential violations of our logistic regression model assumptions. Since there are potential violations of our model assumptions our model may experience inaccurate results.



Model Results

If the person's race is White, compared to Asians (baseline), they are 23% more likely to vote. While on the other hand compared to Asian voters, American Indian and "other" raced voters are 23% less likely to vote.

Holding all other variables the same, Men are 27% less likely to vote compared to Women.

We also noticed that the older you get the more likely you are to vote. Holding all other variables constant, across all counties, people who are over 66 are 228% more likely to vote compared to the 18-25 age group.

Additionally Republicans have the highest voter rate across all parties, as they are 14% more likely to vote than Democrats. While on the other hand Libertarians have the poorest sense of civic duty as they are 40% less likely to vote compared to Democrats.

We also noticed that an unidentified sex Libertarian is 3% more likely to vote than a Female Democrat.

Moreover the variance is relatively small so there isn't much change in voter turnout across these counties (after controlling for all the effects present in the model).

Predictors	Odds Ratios	CI	p
(Intercept)	0.88	0.79 – 0.98	0.025
race [B]	0.91	0.89 – 0.94	<0.001
race [I]	0.77	0.73 – 0.82	<0.001
race [M]	1.00	0.95 – 1.04	0.933
race [O]	0.77	0.75 – 0.80	<0.001
race [U]	1.19	1.15 – 1.22	<0.001
race [W]	1.23	1.20 – 1.26	<0.001
ethnic [NL]	1.15	1.13 – 1.18	<0.001
ethnic [UN]	1.07	1.04 – 1.09	<0.001
sex [M]	0.73	0.73 – 0.74	<0.001
sex [U]	0.86	0.83 – 0.90	<0.001
Age [Age 26 - 40]	1.32	1.31 – 1.34	<0.001
Age [Age 41 - 65]	2.87	2.84 – 2.90	<0.001
Age [Age Over 66]	3.28	3.24 – 3.32	<0.001
Party [LIB]	0.60	0.56 – 0.64	<0.001
Party [REP]	1.14	1.12 – 1.15	<0.001
Party [UNA]	0.78	0.77 – 0.79	<0.001
sex [M] * Party [LIB]	1.51	1.39 – 1.65	<0.001
sex [U] * Party [LIB]	1.99	1.54 – 2.57	<0.001
sex [M] * Party [REP]	1.28	1.26 – 1.30	<0.001
sex [U] * Party [REP]	1.27	1.19 – 1.34	<0.001
sex [M] * Party [UNA]	1.20	1.18 – 1.22	<0.001
sex [U] * Party [UNA]	1.04	0.99 – 1.08	0.130
Random Effects			
σ^2	3.29		
τ_{00} County	0.06		
ICC	0.02		
N County	20		
Observations	9863		
Marginal R2 / Conditional R2	0.091 / 0.107		

Conclusion

Compared to Females, Men are anywhere from 26 - 27% less likely to vote with 95% confidence level. Furthermore Blacks are anywhere from 6 - 11% less likely to vote when compared to Asians, with 95% confidence.

The Standard Deviation for the intercept for different Counties is relatively small (0.241) meaning that the overall probability of voter turnout did not differ much by county.

When evaluating the voter turnout of males and females across party lines, we see that a Male Libertarian is anywhere from 22-43% less likely to vote than a Female Democrat with 95% confidence. Continuing this analysis, we see that Male Republicans are anywhere from 3-10 percent more likely to vote than Female Democrats. Lastly, we can see that a Male unaffiliated with any party is anywhere from 29-34% less likely to vote when compared to Female Democrats with 95% confidence.

One of the limitations in are our analysis is that a decent amount of values fall outside of the 95% confidence bounds of our residual plot, leading us to believe that our results may not be entirely trustworthy. Additionally, we also had to alter our data for instances where there were more actual voters in a county compared to the registered voters in that county.

R Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(formatR)
library(car)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(knitr)
library(dplyr)
library(ggplot2)
library(pander)
library(gridExtra)
library(kableExtra)
library(stargazer)

opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

knitr::opts_chunk$set(echo = FALSE)



# LOAD DATA HERE!!!

addr = "~/Desktop/DukeFA20/IDS702/Team Assignment 2/merged.txt"
merged = read.csv(addr)[, -1]

# set registered as voted if voted > registered
merged$num_voted[is.na(merged$num_voted)] = 0
merged[merged$num_registered < merged$num_voted, ]$num_voted = merged$num_registered[merged$num_registe
    merged$num_voted]
merged$turnout_rate = merged$num_voted/merged$num_registered



# hist(merged$turnout_rate) hist(log(merged$turnout_rate))

merged = merged[!is.na(merged$race), ]
merged$race = factor(merged$race, levels = c("A", "B", "I", "M",
    "O", "U", "W"))  # Drop NA




ggplot(merged, aes(x = turnout_rate)) + geom_histogram(aes(y = ..density..),
    bins = 25, color = "black") + geom_density(alpha = 0.2, fill = "blue") +
    theme_classic()


ggplot(merged, aes(x = Age, y = turnout_rate, fill = Age)) +
    geom_boxplot() + scale_fill_brewer(palette = "Blues") + theme_classic() +
    theme(legend.position = "none") + labs(title = "Turnout v Age")
```

```r
# sample of 7 counties
library(forcats)
set.seed(10)
sample_counties = sample(levels(factor(merged$County)), 7)

ggplot(merged[merged$County %in% sample_counties, ], aes(x = County,
    y = turnout_rate, fill = County)) + geom_boxplot() + scale_fill_brewer(palette = "Blues") +
    theme_classic() + theme(legend.position = "none") + labs(title = "Turnout v County")

merged[merged$County %in% sample_counties, ] %>% dplyr::mutate(County = fct_reorder(County,
    turnout_rate, .fun = "median")) %>% ggplot(aes(y = turnout_rate,
    x = County, fill = County)) + geom_boxplot() + scale_fill_brewer(palette = "Blues") +
    theme_classic() + theme(legend.position = "none", axis.text.x = element_text(angle = 90)) +
    labs(title = "log(turnout rate) vs County") + facet_wrap(~Party)

ggplot(merged, aes(x = sex, y = turnout_rate, fill = sex)) +
    geom_boxplot() + scale_fill_brewer(palette = "Blues") + theme_classic() +
    theme(legend.position = "none") + labs(title = "Turnout v Race") +
    facet_wrap(~Party)




library(lme4)

form1 = cbind(num_voted, num_registered - num_voted) ~ (1 | County) +
    race + ethnic + sex + Age + sex * Party
model1 = glmer(form1, merged, family = binomial)
# summary(model1)


library(arm)
binnedplot(fitted(model1), residuals(model1, "resp"), xlab = "Pred. probabilities",
    col.int = "red", ylab = "Avg. residuals", main = "Binned residual plot",
    col.pts = "navy")




library(sjPlot)
# tab_model(model1)
```