IDS 702 Team Project 2: StreetRx and Voting in NC

10/25/2021

Summary

This analysis is to investigate the potential factors, especially the demographic ones, that affect the turnout rates of the US 2020 General elections in North Carolina. During the process, data cleaning, aggregation, and merging have been performed to combine the registered voters information with the actual voter turnouts. Exploratory data analysis was then performed and the effects of demographic factors on the actual turnouts were examined. The finalized hierarchical model helps us dive into how demographics like age impact the turnout rates in the election, and serves to provide a possible explanation as to why the Republicans outperform the Democrats in certain counties.

Introduction

It is widely acknowledged that the voting behaviors in the election are related to several factors, e.g., demographic factors, location, or the party you are affiliated to. This analysis constructs a hierarchical model to answer the following questions:

- How did different demographic subgroups vote in the 2020 general elections?
- Did the odds of voting differ by county? Which counties differ the most?
- How did the turnout rate differ between males and females for different parties?
- How did the turnout rate differ among age groups for different parties?

Data

Data Pre-Processing

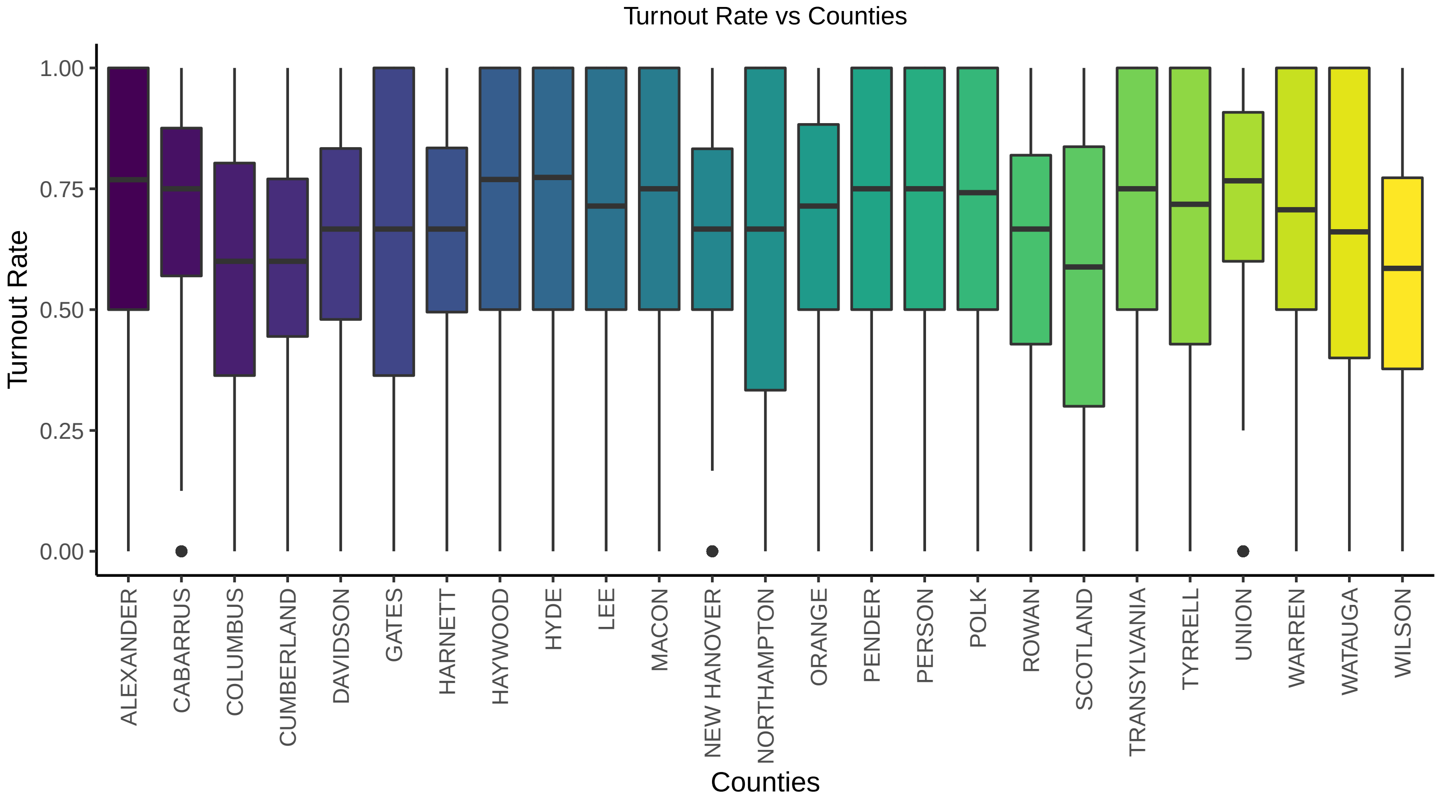
The data used for this analysis was extracted from two files available with The North Carolina State Board of Elections (NCSBE), which is the agency charged with the administration of the elections process and campaign finance disclosure and compliance. One file contains the voter registration records, while the other contains data on the actual turnout (https://www.ncsbe.gov/index.html, https://www.ncsbe.gov/results-data). The code book can be found in the appendix.

The unit of observation in the registered voters file is *county_desc*, *precinct_abbrv*, *vtd_abbrv*, *party_cd*, *race_code*, *ethnic_code*, *sex_code* and *age*. The turnouts file had data at a more granular level (*voting_method*), and it was aggregated to match the unit of observation in the registered voter file. The registered voters file had 592,265 observations, while the turnouts file had 928,532 observations. Post aggregation of the turnouts file, there were 492,567 observations remaining. The actual turnout numbers were then merged with the total voter file to create a model ready dataset. The dataset was further reduced to a sample of 25 counties, post which we were left with 13,162 observations, grouped at a *county_desc*, *party_cd*, *race_code*, *ethnic_code*, *sex_code* and *age* level, with the total number of registered voters and turnouts in demographic groups represented by these characteristics.

EDA

Our EDA is split into 4 different sections. First, we will look at a plot of our hierarchy, county, to determine if it has varying intercept or varying slope. Second, we will look at our main effects, which are all factor variables in our data, so we will only look at boxplots. Third, we will look at all of the interactions with our main effects, again by analyzing boxplots. Finally, we will look at the interactions between our main effects and our set hierarchy to see if we need to control for varying slope.

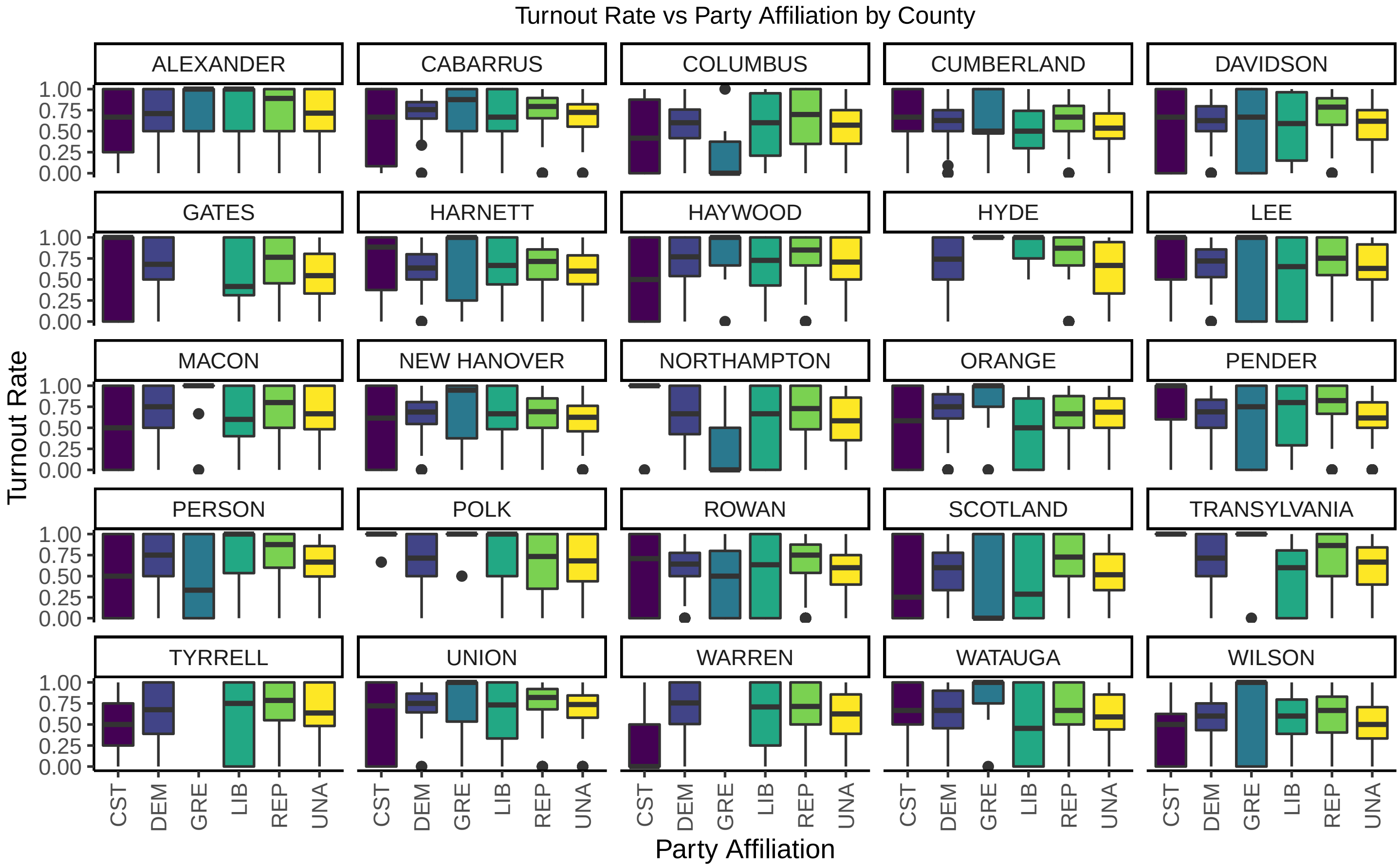
When we look at county, the hierarchy in our data, it clearly varies from county to county. Alexander, Haywood and Hyde have high voter turnout compared to Columbus, Cumberland, Scotland and Wilson. Therefore, we will move forward knowing that we need to include varying intercept in our model building. Next, when we look at our main effects, we can pull some basic information from the boxplots. For example, older age groups coming out to vote more than younger age groups, or that most people choose not disclose their sex when registering to vote.



When we go through all of the interactions of our main effects, we have to consider both the trends in the boxplots as well as the number of observations. This is because a changing trend from one category to the next could be due to a lack of observations rather than an interesting interaction taking place. One example of this is the race of P (Pacific Islander). We sometimes have zero observations of this race when we split it up into different categories, so we have to look at changing trends without consideration for it. This becomes complicated when we also want to look at the interaction between race and party because there are two parties with far fewer observations: CST and GRE. Now we have three different categories that we should *not* take into consideration when looking at trends. Due to these limitations, we decided not to include the interaction between race and party in our model building. On the other hand, there are two interactions of high importance to us: age versus party and sex versus party. These are questions of interest, so regardless of what kind of trend we see, we must include them when model building. We did see that there was a higher median for republicans of ages 26-40 as compared to democrats of the same age group, so we will keep that in mind moving forward.

	CST	DEM	GRE	LIB	REP	UNA
A	10	355	6	55	324	365
B	76	677	59	181	438	580
I	7	347	4	55	303	362
M	7	481	5	89	362	470
O	35	577	23	158	500	585
P	0	29	0	0	31	33
U	106	710	89	239	663	715
W	177	747	162	448	754	763

Finally, we looked at interactions of our main effects and county. We did not find the interaction between county and age or the interaction between county and ethnicity to be as interesting because the trend only changed in 8 or 9 counties out of all 25. When we look at interactions with race and party, we must still keep in mind the low count of observations for some categories. If we do not consider race P when analyzing our race interaction, the trend changes from one county to the next. We also saw the trend continually changing in the interaction between party and county, even when we did not take the CST or GRE parties into consideration. So now we have two interactions with county that we found interesting: race and county and party and county. Even though both of these interactions are interesting, we had to choose only one we wanted to focus on so that we could interpret our final model. We made the decision that we should control the random slope for county versus party, and we will further discuss this in our model building section.



Model

Model selection was performed by accounting for different interactions and effects, which included both random and fixed effects. Our main interactions of interest are analyzing how the turnouts differed by the sexes, and if party affiliations played a role. We are also interested in exploring by the turnouts differed for the age groups for different party affiliations.

We fit a hierarchical model to explore the random effects that different counties may contribute to the model. Counties are used as the only hierarchy. In addition, random slopes for party affiliations were also considered. The model with random intercepts by county was then compared to the model with random slopes for party affiliations with an ANOVA Chi-squared test, and we observe that incorporating the random slope significantly improves model fit.

The final model equation is as follows:

$$y_i | x_i \sim Bernoulli(\pi_i);\ i = 1, 2, \ldots, 13162;\ j = 1, 2, \ldots, 25$$

$$(\beta_0 + \gamma_{0j|i|}) + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4} + \beta_5 * x_{i5} + \beta_6 * x_{i6} + (\beta_7 + \gamma_{1j|i|}) x_{i7};$$

where, x_{i1} is age, x_{i2} is $race\_code$, x_{i3} is $ethnic\_code$, x_{i4} is sex_code, x_{i5} is the interaction effect of age and $party\_code$, x_{i6} is the interaction between sex_code and $party\_cd$, and x_{i7} is $party\_cd$.

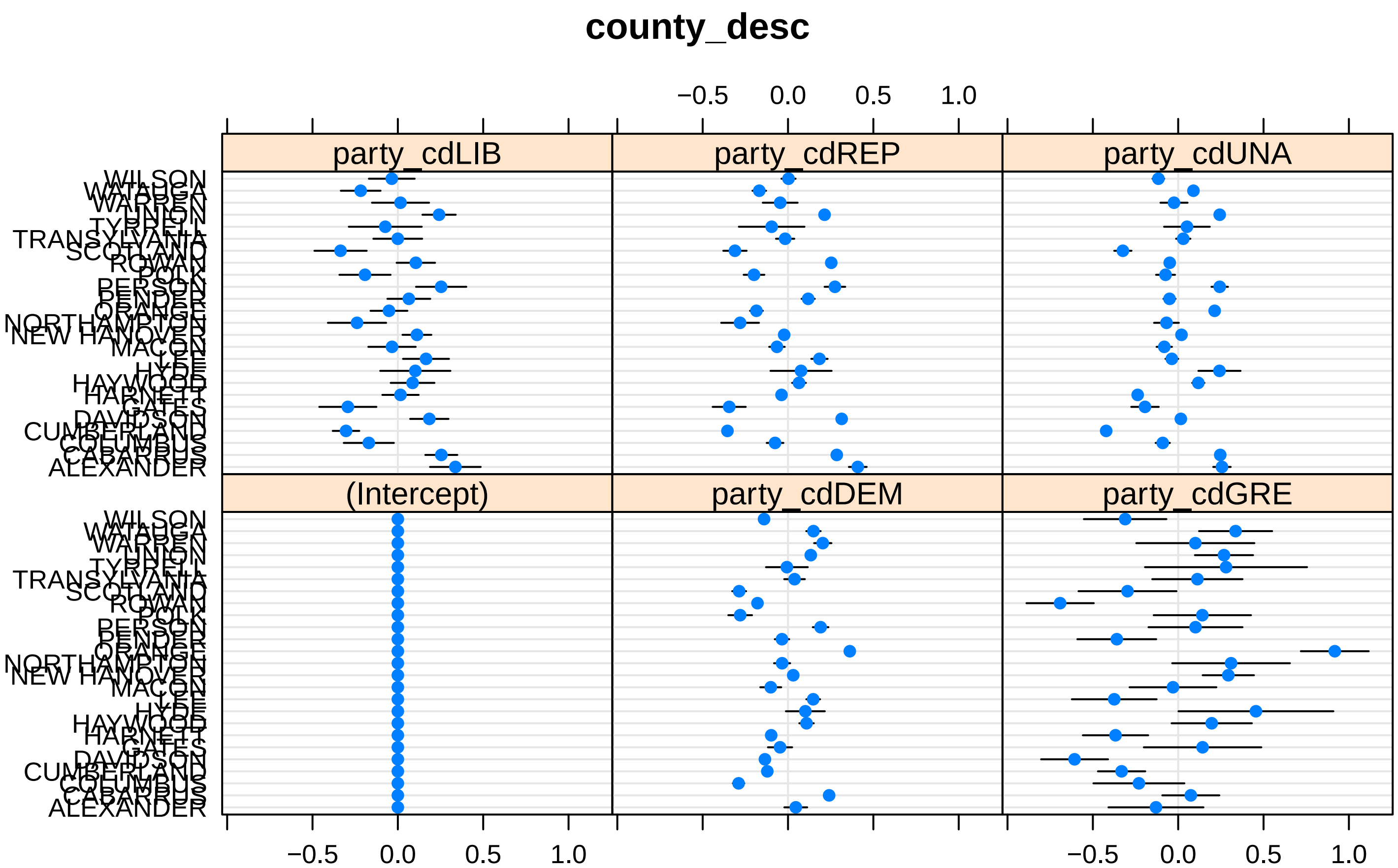
$$\gamma_{0j}, \gamma_{1j} \sim N_2(0, \Sigma)$$

The model results obtained are as shown below:

	Dependent variable:
	cbind(turnout, total_voters - turnout)
ageAge 26 - 40	0.559*** (0.175)
ageAge 41 - 65	1.000*** (0.182)
ageAge Over 66	0.804** (0.343)
party_cdDEM	0.677*** (0.184)
party_cdGRE	0.790*** (0.266)
party_cdLIB	0.114 (0.192)
party_cdREP	0.710*** (0.186)
party_cdUNA	0.288 (0.184)
race_codeB	−0.155*** (0.018)
race_codeI	−0.161*** (0.029)
race_codeM	−0.078*** (0.027)
race_codeO	−0.227*** (0.020)
race_codeP	1.934*** (0.338)
race_codeU	0.309*** (0.020)
race_codeW	0.197*** (0.017)
ethnic_codeNL	0.420*** (0.011)
ethnic_codeUN	0.334*** (0.012)
sex_codeM	0.058 (0.163)
sex_codeU	0.356* (0.197)
ageAge 26 - 40:party_cdDEM	−0.515*** (0.175)
ageAge 41 - 65:party_cdDEM	0.080 (0.183)
ageAge Over 66:party_cdDEM	0.334 (0.343)
ageAge 26 - 40:party_cdGRE	−0.490* (0.256)
ageAge 41 - 65:party_cdGRE	−0.217 (0.306)
ageAge Over 66:party_cdGRE	−0.347 (0.577)
ageAge 26 - 40:party_cdLIB	−0.429** (0.183)
ageAge 41 - 65:party_cdLIB	−0.307 (0.192)
ageAge Over 66:party_cdLIB	0.136 (0.365)
ageAge 26 - 40:party_cdREP	−0.389** (0.175)
ageAge 41 - 65:party_cdREP	0.051 (0.183)
ageAge Over 66:party_cdREP	0.316 (0.343)
ageAge 26 - 40:party_cdUNA	−0.362** (0.175)
ageAge 41 - 65:party_cdUNA	0.112 (0.183)
ageAge Over 66:party_cdUNA	0.624* (0.343)
party_cdDEM:sex_codeM	−0.303* (0.164)
party_cdGRE:sex_codeM	−0.047 (0.257)
party_cdLIB:sex_codeM	−0.052 (0.170)
party_cdREP:sex_codeM	−0.096 (0.164)
party_cdUNA:sex_codeM	−0.176 (0.164)
party_cdDEM:sex_codeU	−0.177 (0.197)
party_cdGRE:sex_codeU	−0.430 (0.281)
party_cdLIB:sex_codeU	0.391* (0.208)
party_cdREP:sex_codeU	−0.004 (0.198)
party_cdUNA:sex_codeU	−0.350* (0.197)
Constant	−0.631*** (0.182)
Observations	13,162
Log Likelihood	−32,112.000
Akaike Inf. Crit.	64,356.000
Bayesian Inf. Crit.	64,850.010
Note:	*p<0.1; **p<0.05; ***p<0.01

In this model, age group 18 - 25 year olds, the Constitution Party, the Asian race, the Hispanic/Latino ethnicity and the female sex are used as baseline factors that are absorbed into the intercept, for ease of interpretation of the impact of these predictors in voter turnouts. All the predictors and the interaction terms are categorical, and a total of 44 distinct factors can be seen in the final model. We observe that the fixed effects of the model are significant at the 5% level. The largest z-values can be observed for *race_code*, *ethnic_code* and *age*, signifying that these factors are the strongest predictors of whether a person if likely to turn up to vote in the elections.

$county_desc



From the dotplot of the random effects above, we observe that introducing the random slope effect of party shows that the log odds of a person voting across different counties differs greatly by the party they are affiliated to. Counties like Alexander, Cabarrus, Orange and Union have odds of voting which are significantly different than zero. People voting for the Green Party have highly varying odds, but this can also be attributed to the lower number of observations recorded against this party. Although the plot confirms that parties account for a lot of the variance in the voting odds we see across counties, it still does not explain all the variation in the model.

Conclusion

The answers to our questions of interest are:

- The odds of turnout for males is 1.06 times compared to females, which is six percent higher. However, it is not statistically significant. The odds of turnout for age group 41 to 65 is found to be the highest, which is 2.7 times compared to the age group 18 to 25. With respect to race, when the Asian race is set as the baseline, the odds of turnout for Pacific Islanders, Undesignated, and White have higher odds of turnout. However, there are very few observations for Pacific Islanders. With respect to ethnic groups, when the Hispanic/Latino category is set as the baseline, the non-hispanic/non-latino category have higher odds of turnout.
- The odds of voting differ by county in 2020 from our EDA, so we added varying intercepts by county in our model. Further, we also observed changes in trends when checking if there is an interaction between party and county. Thus, varying slopes by county were also added to the model. In this case, the varying intercept in our model is 0, but if we select a party, say Democrat, Wake county has the highest odds of turnout, while Randolph and Mcdowell have the lowest ones.
- Observing the interaction between age and party, there are many factors that are statistically significant. One of the interpretations for these combinations is that the odds of turnout for age group 26 to 40 who vote for Democrats is 2.03 times, increasing by 103%, compared to the age group 18 to 25 who vote for CST.

Besides the above questions of interest, we further dive into the voters' age groups and investigate which age group has a higher impact on the actual turnouts between Democrats and Republicans. Observing the interaction between age and party, the 26 to 40 age group is the only statistically significant factor compared with the two other age groups, with age group

18 to 25 as the baseline. Then, we further calculate the odds of turnout between the Democrats and Republicans for this age group, and we observe that the turnouts for Republicans is higher than for Democrats, with odds of turnout observed at 2.41 for the former and 2.05 for the latter when the baseline is the age group of 18 to 25 voting for CST. Another possible explanation behind this might also be the fact that Democrats and Republicans are the two biggest parties, and hence, many observations are available against these parties, compared to the Constitution and Green parties.

Limitations

One limitation of this analysis is that, the NCSBE does not provide the exact difference between the variables *party_cd* and *voted_party_cd* in the actual voter turnouts file. If voters changed their party affiliation at the time of casting their votes, this information is not captured in the dataset, and would not tally with the registered voter numbers. In addition, a random effect model with a varying slope and a varying intercept leads to very complicated interpretations of the model parameters. Since the data is grouped at a combination of different demographics, which makes it granular, we observe that not all groupings have enough observations recorded against them to fit a model on. Certain categories could be aggregated out so that there are sufficient observations against each grouping.