# EDA, Modeling

## 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 $county\_desc, precinct\_abbrv, vtd\_abbrv, party\_cd, race\_code, ethnic\_c$ 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 592265 observations, while the turnouts file had 928532 observations. Post aggregation of the turnouts file, there were 492567 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 13162 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 geoups represented by these characteristics.

| | *Dependent variable:* |
|---|---|
| | cbind(turnout, total_voters - turnout) |
| ageAge 26 - 40 | $0.559^{***}$ (0.175) |
| ageAge 41 - 65 | $1.000^{***}$ (0.183) |
| ageAge Over 66 | $0.804^{**}$ (0.346) |
| party_cdDEM | $0.677^{***}$ (0.184) |
| party_cdGRE | $0.790^{***}$ (0.265) |
| 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 | $1.934^{***}$ (0.337) |
| 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 | $-0.515^{***}$ (0.176) |
| ageAge 41 - 65:party_cdDEM | 0.080 (0.183) |
| ageAge Over 66:party_cdDEM | 0.334 (0.346) |
| 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 (0.579) |
| 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 (0.368) |
| ageAge 26 - 40:party_cdREP | $-0.389^{**}$ (0.176) |
| ageAge 41 - 65:party_cdREP | 0.050 (0.183) |
| ageAge Over 66:party_cdREP | 0.316 (0.346) |
| ageAge 26 - 40:party_cdUNA | $-0.362^{**}$ (0.176) |
| ageAge 41 - 65:party_cdUNA | 0.111 (0.183) |
| ageAge Over 66:party_cdUNA | $0.624^{*}$ (0.346) |
| party_cdDEM:sex_codeM | $-0.303^{*}$ (0.163) |
| party_cdGRE:sex_codeM | −0.047 (0.256) |
| party_cdLIB:sex_codeM | −0.052 (0.169) |
| party_cdREP:sex_codeM | −0.096 (0.163) |
| party_cdUNA:sex_codeM | −0.176 (0.163) |
| party_cdDEM:sex_codeU | −0.177 (0.197) |
| party_cdGRE:sex_codeU | −0.430 (0.280) |
| party_cdLIB:sex_codeU | $0.391^{*}$ (0.207) |
| party_cdREP:sex_codeU | −0.004 (0.197) |
| party_cdUNA:sex_codeU | $-0.351^{*}$ (0.197) |
| Constant | $-0.631^{***}$ (0.181) |
| 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 |

# 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 analysing 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 under:

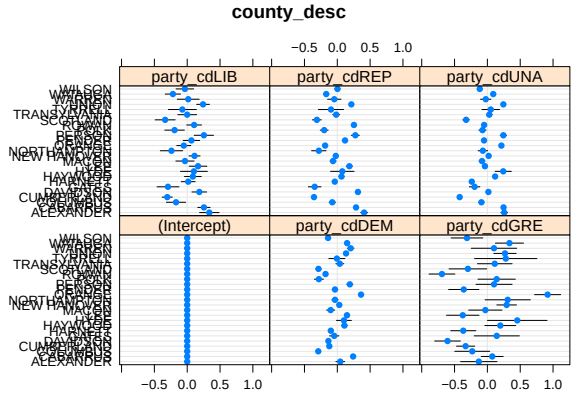
$$y_i|x_i \sim Bernoulli(\pi_i);\ i = 1, 2, \ldots, 13162;\ j = 1, 2, \ldots, 25$$

$$log(\frac{\pi_i}{1-\pi_i}) = (\beta_0+\gamma_{0j|i|})+\beta_1*age_{i1}+\beta_2*race\_code_{i2}+\beta_3*ethnic\_code_{i3}+\beta_4 sex\_code_{i4}+\beta_5 age : party\_cd_{i5}+\beta_6 sex\_code : p \quad (1)$$

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

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. The residuals obtained are shown below.

% Error: Unrecognized object type.