IDS702 Team Yellow Project II

Anna Dai, Athena Liu, Dauren Bizhanov, Himangshu Raj Bhantana, Moritz Wilksch

October 17, 2021

Summary

Introduction

The North Carolina State Board of Elections (NCSBE) is the body in charge of election administration as well as campaign money disclosure and compliance. They provide voter registration and turnout statistics online through the ncsbe.gov website, among other things. For this study, we will attempt to identify/estimate how various groups voted in the 2020 elections, at least among those who registered, using the NC voter files for the general elections in November 2020. We also use the hierarchical model to focus on a few areas of interest. In the 2020 general election, how did various demographic groupings vote? If the general likelihood or odds of voting in 2020 change by county? What was the difference in turnout rates between males and females for the various party affiliations? and finally How did turnout rates fluctuate by age group for the various party affiliations?

Data

Two data sets are available for this analysis. The first one - `Voters` contains registered voters - `total_voters.registered` with additional fourteen variables and the second one - `History` contains actual votes - `total_voters.actual` with other eleven variables. We use only a subset of these data sets for our purposes. We have randomly sampled 25 counties and used them as a filter. The counties used are: Cherokee, Dare, Alexander, Pasquotank, Camden, Mitchell, Anson, Vance, Carteret, Perquimans, Edgecombe, Brunswick, New Hanover, Robeson, Mcdowell, Nash, Davidson, Forsyth, Johnston, Northampton, Craven, Haywood, Gates, Alamance, Orange. In order to merge the data from two tables we have performed an aggregation of `History` data set to join the tables correctly. The group variables used for the aggregation are: `county_desc`, `precinct_abbrv`, `age`, `party_cd`, `race_code`, `ethnic_code`, `sex_code` and `total_voters.actual` is summed across these groups. Having done with the aggregation, left join has been performed. The keys for the left join are the same as the grouping variables above. There are around hundred rows with more actual votes than registered. It has been decided to lower the amount of actual votes in these cases, so that `total_voters.registered` are always greater than or equal to `total_voters.actual`. Missing values in `total_voters.actual` are replaced with zeroes. As a part of our EDA, we transformed aggregated actual voting variable to a binary variable with duplicated rows. Conditional probability tables have revealed interesting associations. The voting probability varies for different age groups. Mature people tend to vote with higher chance than younger ones. The youngest age group "18-25" has probability of voting 61.2% and the oldest group "Over 66" has 85.5% probability. Moreover, probability of voting is different across ethnic groups. Hispanic group has lower voting probability 58.6% than Non-Hispanic 78.6% or UN 75.5%.

- general description
  - random sample counties (name them!)
  - aggregated, describe used variables
  - mention diff between registered an actual

- joining
  - left, keys,
- cleaning
  - turnout <= 1, NA handling
- EDA
  - 0/1 dummy instead of turnout variable
- Interactions

Model

TBD

Conclusion