IDS702 Team Yellow Project II

Anna Dai (Presenter), Athena Liu (Checker), Dauren Bizhanov (Writer), Himangshu Raj Bhantana (Co

October 17, 2021

Summary

- Recap focus
- Recap methods
- Summarize result

Introduction

The North Carolina State Board of Elections (NCSBE) oversees election administration as well as campaign finance disclosure and compliance. They give voter registration and turnout information online, among other things, via the ncsbe.gov website. Using the NC voter files for the general elections in November 2020, we will attempt to identify/estimate how various groups voted in the 2020 elections, at least among those who registered. We also employ the hierarchical model to narrow our attention on a few areas of particular interest. How did different demographic groups vote in the 2020 general election? If the overall chance of voting in 2020 varies by county? What was the difference in turnout rates between men and women for each party affiliation? Finally, how did turnout rates differ by age group for different 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 592 265 rows. The second one - `History` contains actual votes - `total_voters.actual` with other eleven variables and 928 532 rows. 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 in the analysis 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. After the filtration the resulting data sets have 130 709 and 205 303 observations for `Voters` and `History` respectively. Variables `stats_type`, `election_date`, `update_date` are deleted immediately as they are out of our interest. In order to merge the data from two tables, we have performed aggregations of `History` and `Voters` data sets to join the tables correctly. Grouping variables used for the aggregations are: `county_desc`, `age`, `party_cd` (`voted_party_cd`), `race_code`, `ethnic_code`, `sex_code` and `total_voters.actual` and `total_voters.registered` are summed across these groups. Post aggregation `History` and `Voters` data sets have 11 374 and 12 935 rows. Having done with the aggregation and renaming `voted_party_cd` variable from the `History` data set to `party_cd`, left join has been performed. The keys for the left join are the same as the grouping variables above and after the left join has been performed we still have 12 935 rows as expected. All rows with missing values in `total_voters.actual` are dropped which reduced the merged data set to 11 374 observations. There are ten observations 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`. 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 probability than younger ones. The youngest age group "18-25" has probability of voting 58.3% and the oldest group "Over 66" has 84.2% probability. Moreover, probability of voting is different across ethnic groups. Hispanic group has lower voting probability 58.3% than Non-Hispanic 76.5% or Undesignated 73.7%. Voting rates varies as well as across races. So, "Other" group has the lowest probability of voting 58.9% and "White" group has 78.6% probability. Noticeable variation of conditional probabilities can be observed across parties. Democrats has 74.7% and Republicans 82% probability of voting. The variable `county_desc` may be a hierarchical one as turnout rate varies across different counties and the number of observation substantially differs, so rather than using this variable as categorical in Logistic regression we should try fitting hierarchical model. The most interesting interactions are `party_cd`:`sex_code` and `party_cd`:`age`. Turnout rate is the opposite for Libertarian party and Unaffiliated group for different gender. It is higher for female in Unaffiliated and higher for male in Libertarian party. Green party has higher turnout rate for younger people in 18-25 years category than any other parties have. We will examine these interaction further during modeling part.

Model

In order to answer the report questions we used Hierarchical Logistic regression model. As a first step we started our modelling with regular logistic regression and used stepwise approach for the variable selection. We created our Null model as an intercept only and Full model with `party_cd`, `race_code`, `ethnic_code`, `sex_code`, `age` variables. The resulting stepwise model keeps all the variables. The only insignificant categories in the stepwise model are "Democrats", "Republicans" in `party_cd` variable in comparison to the baseline "Constitution party" which gives us confidence to proceed further. We experienced model convergence issues while fitting a hierarchical model. Changing the data pre-processing part such as aggregating not only `History` data set, but `Voters` data set too. In addition, switching to a different optimizer eventually solved the problem. Our final model contains `party_cd`, `race_code`, `ethnic_code`, `sex_code`, `age` variables and `sex_code`:`party_cd` ,`age`:`party_cd` interactions. The hierarchical level uses `county_desc` variable for creating random intercepts.

$$y_{ij}|x_{ij} \sim Bernoulli(turnout_{ij})$$

$$\log(\frac{turnout_i}{1-turnout_i}) = (\beta_0 + \gamma_{0j}) + \beta_1 \cdot \text{party\_cd}_{ij} + \beta_2 \cdot \text{race\_code}_{ij} + \beta_3 \cdot \text{ethnic\_code}_{ij}$$
$$+\beta_4 \cdot \text{sex\_code}_{ij} + \beta_4 \cdot \text{age}_{ij} + \beta 5 \cdot \text{sex\_code} : \text{party\_cd}_{ij} + \beta 6 \cdot \text{age} : \text{party\_cd}_{ij}$$

$$\gamma_{0j} \sim \mathcal{N}(0, \sigma_0^2), i = 1, ..., n; j = 1, ..., J$$

All categories in `race_code` variable are significant and highly dispersed. Race code "B" - Black decreases the log odds ratio by 0.1869 which is 0.83 (exp(-0.1869)) multiplicative effect or 17% decrease in odds ratio in comparison to base level "A" - Asian. On the other hand, race code "U" - Undesignated increases log odds ratio by 0.2467 or by 1.28 (exp(0.2467)) multiplicative effect or increasing odds ration by 28% keeping all the other variables fixed. The variable `ethnic_code` has all significant categories too. Being Non-Hispanic is associated with increasing log odds by 0.42 (1.52 in odds ratio) in comparison to Hispanic ethnic group and being part of Undesignated ethnic group is associated with increasing log odds ratio by 0.30 (1.35 in odds ratio). There is no significant difference by `sex_code` variable in comparison with the baseline level. So, we can not conclude that being male or female or undesignated is associated with higher or lower turnout rates. Being part of Liberal party is associated with decreasing log odds ratio by the factor of 0.6728 (0.51 odds ratio) and this category is the only significant one in comparison to Constitutional party.

From the dotplot we can conclude that only four counties have non-significant random intercept. Counties such as Carteret, Alexander, Orange, Dare, Alamance, Johnston, Nash, Davidson, Haywood, Forsyth are associated with increasing log odds ratios in decreasing order of the increasing effect. Counties such as Robeson, Cherokee, Anson, Pasquotank, Gates, Craven, Camden, Edgecomb, Northhampton are associated

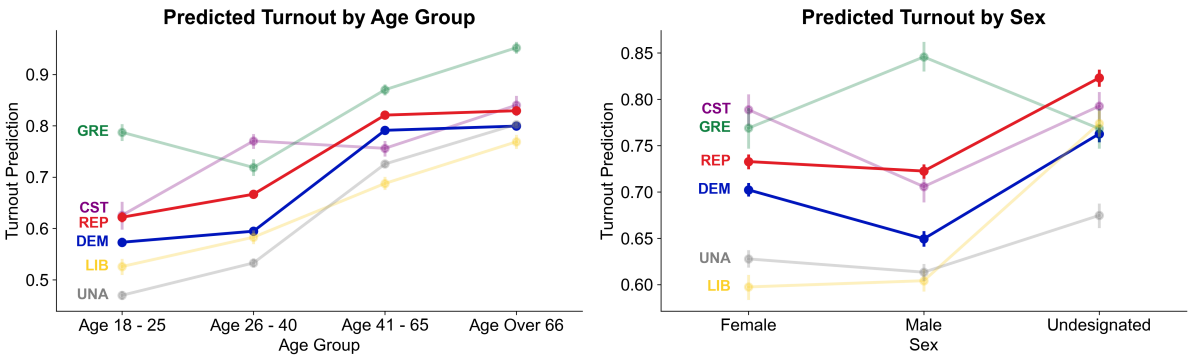

Figure 1: Turnout prediction plots

with increasing log odds ratios. For example, voters from Cartener county tend to have higher turnout and voters from Robeson lower turnout. We have two significant levels of party by sex interaction. Being a male and a member of Green party is associated with increasing log odds ratio by 0.786 (2.19 odds ratio) in comparison to a female member of Constitutional party. The second significant level is a Libertarian party member with undesignated sex category. This level is associated with increasing log odds ratio 0.74 (2.10 odds ratio). The coefficients for the party and age groups interaction have more variability. So, regardless of the party age group "26-40" is associated with decreasing log odds in comparison the baseline age group "18-25" and Constitutional party category. The opposite effect holds for the age group "Over 66" which is associated with increasing log odds ratio regardless of the party categories. The more interesting situation in the age group of "41-65" which has one level Green party with decreasing odds ratio -0.09 (0.91 odds ratio), however not significant.

Conclusion

- Recap focus
- Recap results
- Elaborate on interpretation
- limitation and future work