Analysis of Voting in NC (2016 General Elections)

Akshay Punwatkar, Melody(Xinwen) Li, Derek Wales, Andrew Patterson, Tzu-Chun (Angela)

11/04/2019

SUMMARY

Analysis and modeling of the voter registration and participation data from the 2016 US General election for the state of North Carolina were performed. And the effects of the demographics on the voter turnout were analyzed and quantified. The analysis was preceded by several data processing steps and was performed on a subset of the data for 20 Counties within North Carolina. Voter turnout based on different genders, ethnicity, race, age groups, and party affiliations were analyzed. A Multilevel/Hierarchical logistic regression model used to quantify the effects of demographics on voter turnout.

INTRODUCTION

United States General Elections in 2016 was one of the most anticipated election. The election saw an average turnout of about 63%, which was reported as the lowest turnout in the past 20 years. Voter turnout is assumed to be affected by several demographic factors such as gender, age, race, ethnicity, county, and party affiliation. Other factors, such as the election campaign, accessibility to voting booths, and many more, also affects the turnout. The following analysis is aimed at analyzing and quantifying the effects of such demographic factors in determining the voter turnout. The scope of the analysis is limited to the state of North Carolina.

The North Carolina State Board of Elections (NCSBE) is the agency charged with the administration and the election process and campaign finance disclosure and compliance for the state. They are also required to keep extensive records to ensure electoral compliance, as part of their duties, they also keep information on likely voters and registered voters. Using the data obtained from NCSBE, the analysis was done primarily to gain insights about regarding a few questions :

- How did demographic subgroups vote in 2016?
- Did the overall probability or odds of voting differ by county in 2016?
- How did the turnout rates differ between females and males for the different party affiliations?

DATA

Dataset obtained from NCSBE contained demographic related information about the registered voters and voters who voted from 102 counties in North Carolina. Since the data was provided in two parts, extensive pre-processing of the data was performed. Overall, voter turnout as per the provided data was ~ 72.0 %, and it was made sure throughout the data processing process that the turnout occurs in a similar range.

Data Processing and Transformation :

The registered voter dataset, initially containing duplicates and null observations were processed to obtain the unique observations without any null values (0.2 % observations were null). Similarly, the voted voters dataset was processed to remove any null values (4.0 % observations were null). Subsequently, both the datasets were merged based on the demographics.

However, due to the variation in the methods used for voting or change in party affiliation, the voted dataset had multiple observations for the same demographics as in the registered voter's dataset. In order to eliminate redundancy of the total registered voter post merging, data were aggregated based on the demographics and total voters. In the process, few of the features, such as the voting method and voting method description, were dropped. Few more

- Since only a fraction of the population changed their party affiliations during voting, **party_cd** (original party affiliation) was used in the final dataset instead of **voted_party_cd**. Also, **voted_party_cd** was leading to duplication of total_registered voters.
- Few observations (0.84%) had **more voted voters than the registered voters** for a demographic. It could be explained by assuming that the voters might have changed their county/precinct and hence having the same demographic voted under different precincts. For such cases, the count of total registered voters was increased to match the voted voters, because eliminating such observations would have led to information loss for the entire demographic.
- **Precinct** and **voter district** was dropped from the final dataset. And the data was again aggregated based on the demographics.

Data Description :

The data dictionary used as part of the final analysis is as follows:

county_desc - *Name of the County of the voters belonging to a demographic group*
age - *Age group of the voters belonging to a demographic group*
race - *Race of the voters belonging to a demographic group*
ethnicity - *Enthnicity of the voters belonging to a demographic group*
sex_code - *Gender of the voters belonging to a demographic group*
party_cd - *Party affiliation of the voters belonging to a demographic group*
total_voters - *Number of registered voters belonging to a demographic group*
voted_voters - *Number of voters who voted belonging to a demographic group*

Features not used in the final dataset are not mentioned in the dictionary

Data Selection :

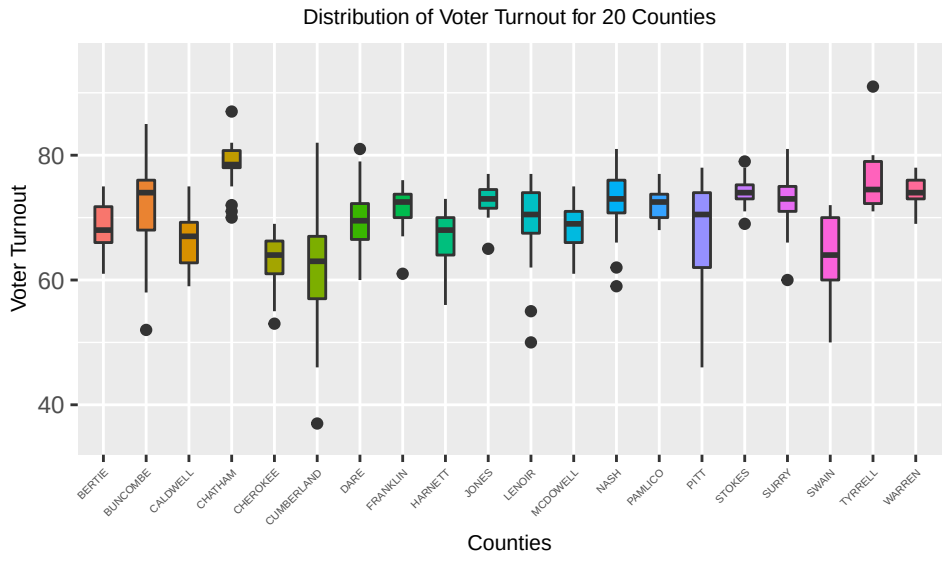
Voting data about 20 Counties was randomly selected from the primary dataset for the final analysis.

County (1-5)	County(6-10)	County (10-15)	County (15-20)
BERTIE	CALDWELL	FRANKLIN	DARE
HARNETT	JONES	CHATHAM	SWAIN
PAMLICO	BUNCOMBE	MCDOWELL	MCDOWELL
STOKES	PITT	NASH	SURRY
WARREN	TYRRELL	LENOIR	CUMBERLAND

Data Analysis :

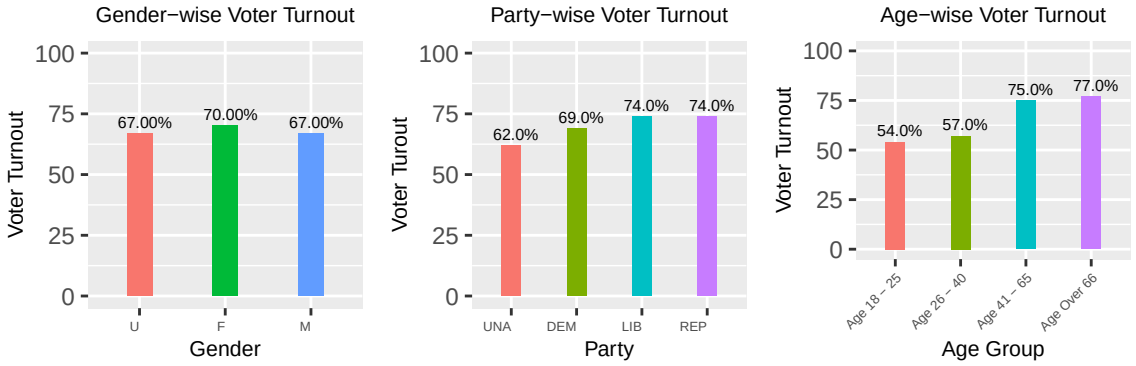
Initial analysis of data using visualization provides several keys insights about the variation of voter turnout among and within the counties. It also highlighted several relationships among gender, race, party, and counties. Following are the key observations from the analysis :

- Although **average voter turnout** for every **county** was in the **similar** range of ~70%, **distribution** of voter turnout within the counties didn't appear to be that similar.

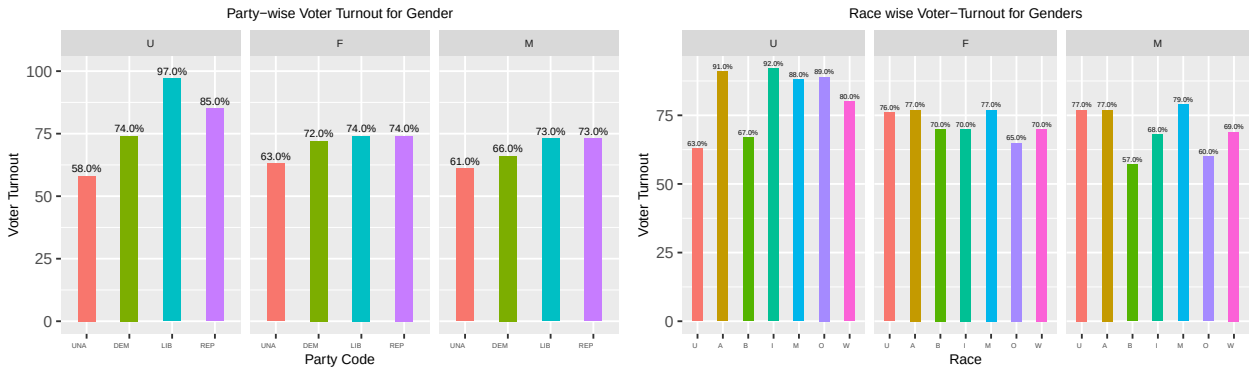


- **Female** voters had **3%** more turnout as compared to male counterparts.
- **Republican** and **Libertarian** voters had the highest turnout, 5% more as compared to Democrats.
- Only **half** of the registered voters **aged between 18-40** showed up for voting.

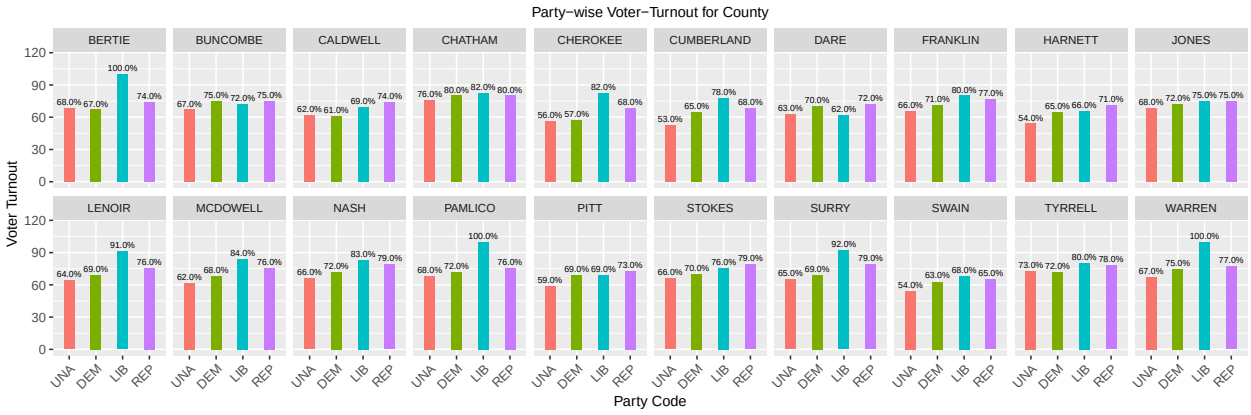
.



- **Female Democrats** had **8%** more turnout than the male democrats
- **Black-Female** had **13%** more turnout than their male counterparts.



- Interestingly several counties had **100%** voter turnout for **Libertarian** party which is much higher than the overall average turnout of ~ 70%.



MODEL

Post the initial analysis, to quantify the effects of the demographic variables, a logistic regression model was used. In addition, as discussed earlier, although all the counties had similar average voter turnout in the range of ~70 %, inter-county distributions were different. To capture this in-county variance in voter turnout, along with the information of overall voter turnout, a **multi-level** hierarchical model (random intercept for counties in this case) was used to quantify the demographic effects.

A series of modeling attempts, using *county* as a random intercept and rest of demographics as fixed effects along with few interactions between gender, age, party, and the race was made and tested using ANOVA. However, since voter turnout is primarily a function of the demographics, each model highlighted high significance towards all the fixed demographic variables. Moreover, with the increment of a number of interactions, effects were getting distributed and hence diminished among interactions.

Finally, a model (with lowest AIC and variation) using all the demographic features (age, gender, race, ethnicity, and party affiliations) along with an interaction between gender and party affiliations were selected.

$$log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = \hat{\beta}_0 + \gamma_{0j[i]}^{County} + \hat{\beta}_g G_i + \hat{\beta}_r R_i + \hat{\beta}_E E_i + \hat{\beta}_P P_i + \hat{\beta}_A A_i + \sum_{k=2}^{K} \hat{\beta}_6 G_{ik} : P_i$$

Where: **G** = *Gender*, **R** = *Race Code*, **E** = *Ethnicity code*, **P** = *Party Code*, **A** = *Age group*

RESULT

Quantifying the demographics effects provided several key insights :

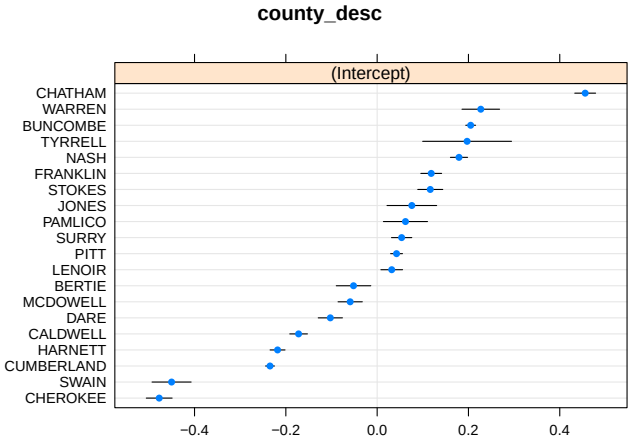
- The base odds of turnout is **1.09**. **Chatham** county had the highest baseline odds of turnout at **1.54**, while **Swain** and **Cherokee** had the lowest baseline odds of turnout at **0.5**

Given control of other potential predictors, **INCREASE** in Odds of Turnout being a :

- **Female** voters (1.14) is **13%** more the **Male** voters (1.01).
- **Libertarians** was surprisingly high, which could be explained by the very high voter turnout for the Libertarian party in several counties, as discussed during the analysis.
- **Republican** voters (3.82) is **117%** more than the **Democrats** voters (2.05).
- Voter in **age groups above 40 years** is nearly **150%** more than the lower age groups.
- **Hispanic or Latino** voter is nearly **30%** more than **Non-Hispanic or Latino** voters.

Controlling all other potential variables and varying only the gender and party, the prediction was made to quantify the effects of gender w.r.t to party. Below were the observations:

- Odds of Voter turnout being a **Female Democrats** exceeded **Male Democrats** by a factor of **0.65**.
- However, Odds of Voter turnout being a **Female Republic** exceeded **Male Republic** by only a factor of **0.05**.
- Moreover, Odds of Voter turnout being a **Male Republic** exceeded **Male Democrats** by a factor **0.93**.
- Moreover, Odds of Voter turnout being a **Female Republic** exceeded **Female Democrats** by a factor **0.35**.



Beta	Exp Val	Beta.	Exp Val.	Beta..	Exp Val..	Beta...	Exp Val...
Intercept	1.09	Age 26-40	1.16	>2 Races	1.91	F Dem	0.64
Female	1.14	Age 41-65	2.55	Other Race	0.69	M Dem	0.54
Male	1.01	Age over 66	2.85	White	0.76	F Lib	0.08
Democrats	2.05	Asian	1.44	Hispanic or Latino	1.36	M Lib	0.08
Libertarians	25.59	Black or AA	0.73	Non Hispanic or Latino	1.07	F Rep	0.39
Republican	3.82	AI or AN	0.99			M Rep	0.43

The above table contains the exponentiated estimates for the variables.

- The Model generated an **in-sample of accuracy** of **95%**.
- Based on the dot-plot highlighting the 95% confidence interval for the random effects of counties, a multi-level (random intercept) model seemed a good fit.

CONCLUSION

The analysis provided several important highlights about the demographics features such as County, Gender, age, race, and ethnicity, which seemed to affect voter turnout in the state of NC, among other variables. Quantifying the effect of these features using a multi-level logistic regression model provided an idea of the extent of effects of the demographics on the voter turnout. However, given more information about the polling stations and election campaigns along with the demographics, it could have assisted in better analysis. Moreover, since all the features were categorical, the model was highly susceptible to overfitting, and due to the absence of test data, testing model performance could not be done. Also, the interpretation of the model to quantify the effects of interactions could not be carried out directly. Prediction using controlled data was performed to quantify these effects. To summarize, an extensive analysis needs to done using more demographic features in the data, as mentioned earlier, and a better modeling algorithm to analyze and model the effects leading to voter turnout.