Analyzing Voter Registration in the 2016 Presidential Election

Andrew Amore

2022-11-21

Abstract

To investigate demographic trends in voter registration rates across North Carolina, historical voter registration information from 2016 and U.S. Census estimates from 2010 were analyzed. To facilitate the analysis, datasets were standardized with matching field codes and combined into a holistic view for modeling. Several Bayesian hierarchical strategies were assessed with different multilevel structures and evaluated based on **metric XXX, YYY**.

The final model indicates...

Introduction

Political campaigns analyze historical election data to understand how different demographic groups register to vote, specifically during presidential election years, where voters are more likely to show up at the polls. This information can be used to determine optimal locations for advertising campaigns than can sway opinions, drum up more votes and even win elections. Campaigns are most interested in assessing how different demographic groups register to vote and how registration odds vary by county, gender, age and party affiliation.

Dataset Overview

To address the main concerns, two data sources were analyzed. Information from the 2010 U.S. Census was collected from the Federal Census Bureau website and combined with 2016 voter registration from North Carolina. To join datasets, field codes were standardized, as the Census Bureau uses slightly different coding than the North Carolina State Agency for demographic categories. Census data only quantifies male/female genders and entries in the voter registration file with unspecified gender were removed. Irrelevant fields denoting precinct locations were also removed. Metadata information for the combined dataset, with a sample observation, can be viewed in Table 1.

Table 1: Metadata Information

Field Name	Description	Sample
Geography	*County in North Carolina*	*MONTGOMERY*
Age	*Age Demographic Category*	*18−25*
Gender	*Gender Demographic Category*	*Female*
Hispanic	*Demographic Indicator of Hispanic Origin*	*Hispanic*
Race	*Race Demographic Category*	*BlackAlone*
VoterFreq	*Number of Registered Voters*	*1*
Freq	*Total Population Count for Specified Categories*	*2*
TotalCountyPopulation	*Total County Population*	*27798*
PartyCd	*Political Party Affiliation*	*DEM*
VoterTurnout	*Registration Percentage (VoterFreq/TotalCountyPopulation)*	*0.000036*

Data Issues & Subsetting for Analysis

An assumption of this analysis is that population numbers from 2010 are representative of 2016. Inspecting the combined dataset, there are numerous instances where the number of registered voters is greater than the total population for a demographic category (see Table 2). Voters may register outside of their residence county, however, a more plausible explanation is that the Census information is dated.

Table 2: Bad Data Sample

Geography	Age	Gender	Hispanic	Race	VoterFreq	Freq	PartyCd
ONSLOW	26−40	Female	NotHispanic	SomeOtherRaceAlone	114	82	REP
HOKE	41−65	Female	NotHispanic	SomeOtherRaceAlone	24	21	REP
DURHAM	66+	Male	NotHispanic	SomeOtherRaceAlone	67	6	UNA
CLEVELAND	18−25	Male	NotHispanic	SomeOtherRaceAlone	14	10	REP

In addition, the aggregate dataset contains too much information to analyze at once. To address computation concerns, 30 counties were randomly drawn using sampling weights inversely proportional to the percentage of observations where the number of registered voters (VoterFreq) is greater than the total demographic population (Freq).

Motivating the Model

An exploratory data analysis (EDA) was conducted to evaluate modeling decisions, like where to apply random effects. Figure 1 displays voter registration trends for three counties (color coded) and two demographic categories. The left plot fits regressions by county & age group, while the right fits identical data by county & race. Notice the differences in linear fits between the two plots suggesting voter behavior is similar across race, but varies significantly across age groups. Plots for remaining covariates can be viewed in the Appendix. In addition to age, party affiliation showed similar variation between categories, but other covariates displayed consistent trends.

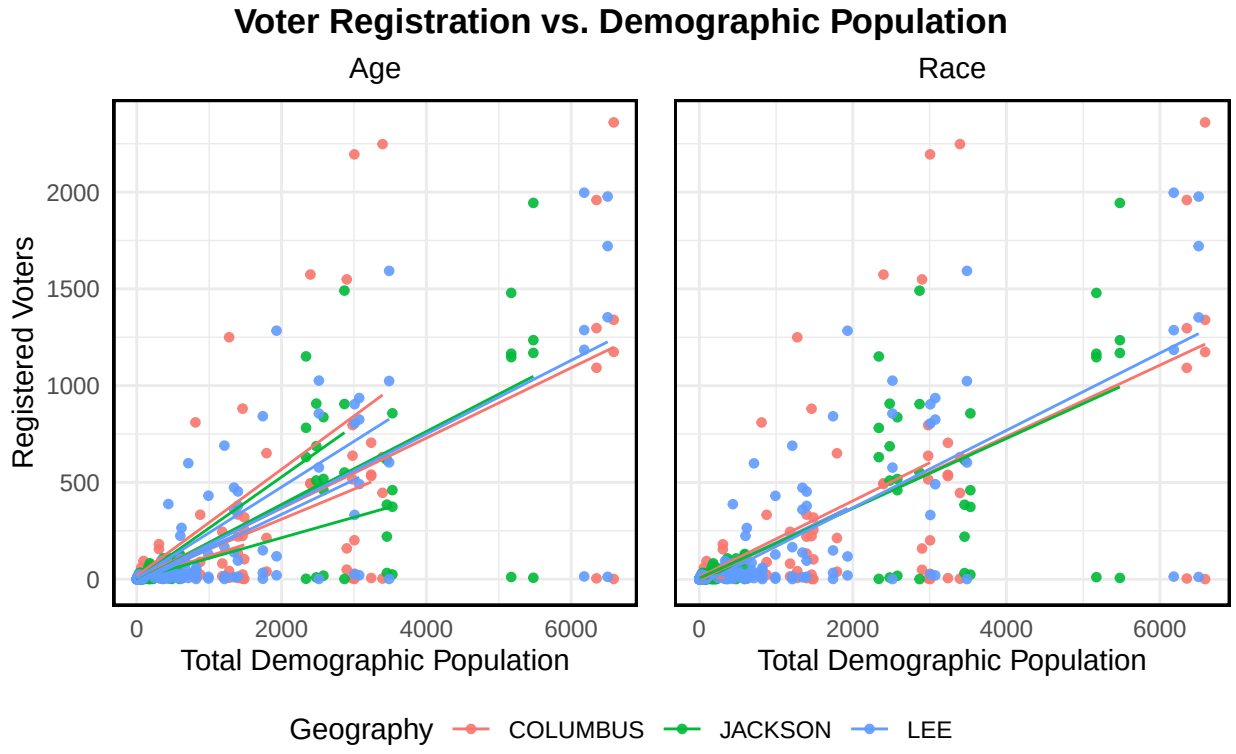


Figure 1: Voter Registration Behavior

Model Specification

The modeling framework can be specified as follows. Let n_{ijk}, p_{ijk} denote the total demographic population and probability of voter registration for voters in county i, age group j and remaining covariate indicator combination k respectively. We assume the number of voters registering to vote, $n_{voters,\{i,j,k\}}$, is distributed as a binomial random variable.

$$n_{voters,\{i,j,k\}} \sim Binomial(n_{ijk}, p_{ijk})$$

Binomial trials assume replacement, but voters can only register once. This simplifying assumption seems reasonable as most demographic categories have large populations ($n_{ijk} > 30$). A more realistic model could assume a

hypergeometric distribution to account for smaller population sizes, but this induces additional computation complexity and was avoided. Several multilevel models with binomial distributions were assessed (Table 3). The EDA suggests variation across counties, age groups and political affiliations, but how does this covariance structure affect coefficient estimates?

Table 3: Random Effect Structures

	Fixed Effect Fields	Random Intercept Fields	Random Slope Fields
Model I	Gender, Hispanic, Race, PartyCd, Age	Geography	
Model II	Gender, Hispanic, Race, PartyCd, Age	Geography	Age
Model III	Gender, Hispanic, Race, PartyCd, Age	Geography	Age, PartyCd

Model I
Let μ denote a global intercept, θ_i a **county** random effect intercept for observations in county i, $X_k, [ij]$ a vector of demographic indicator variables (gender, ethnicity, race, party affiliation and age) and β the corresponding fixed effect estimates.

$$Logit(p_{ijk})) = \mu + \theta_i + X_{k,[ij]}\beta$$
$$\mu \sim N(0,1)\,,\ \beta \sim N(0,1)\,,\ \theta_i \sim N(0,\sigma)$$
$$\sigma \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Model II
Model II is an extension of I with a new random effect for **Age**. Let $a_{ij,[k]}$ denote the age category, j, for observations in county i and Γ_{ij} the random effect slope. Other variables remain the same.

$$Logit(p_{ijk})) = \mu + \theta_i + a_{ij,[k]}\Gamma_{ij} + X_{k,[ij]}\beta$$
$$\Gamma_{ij} \sim N(0,\gamma)\,,\ \gamma \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Model III
An extension of II with an additional random effect for **party affiliation**. Let $q_{il,[jk]}$ denote the party, l, for observations in county i and Ω_{il} the random effect slope. Other variables remain identical.

$$Logit(p_{ijlk})) = \theta_i + a_{ij,[k]}\Gamma_{ij} + q_{il,[jk]}\Omega_{il} + X_{k,[i,j]}\beta$$
$$\Omega_{il} \sim N(0,\omega)\,,\ \omega \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

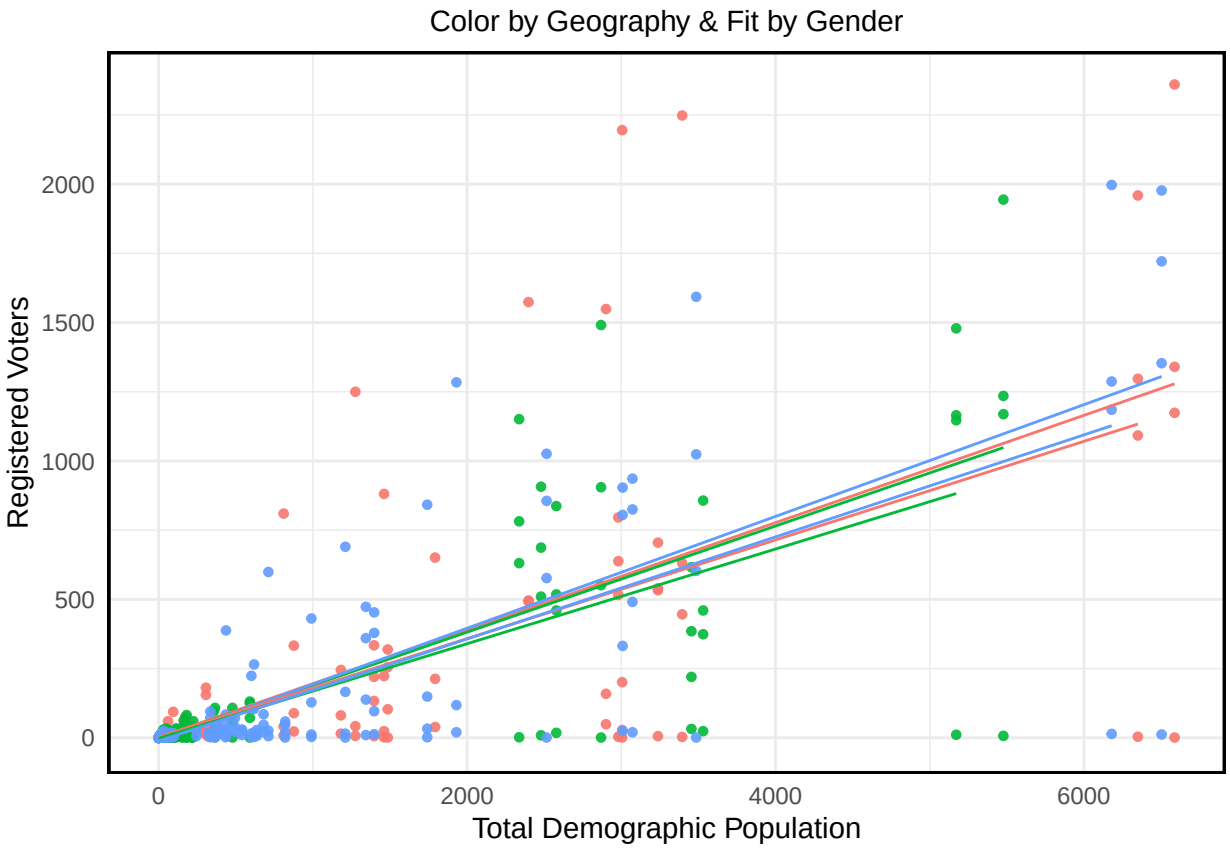
Assessing Model Fit & Comparing Models

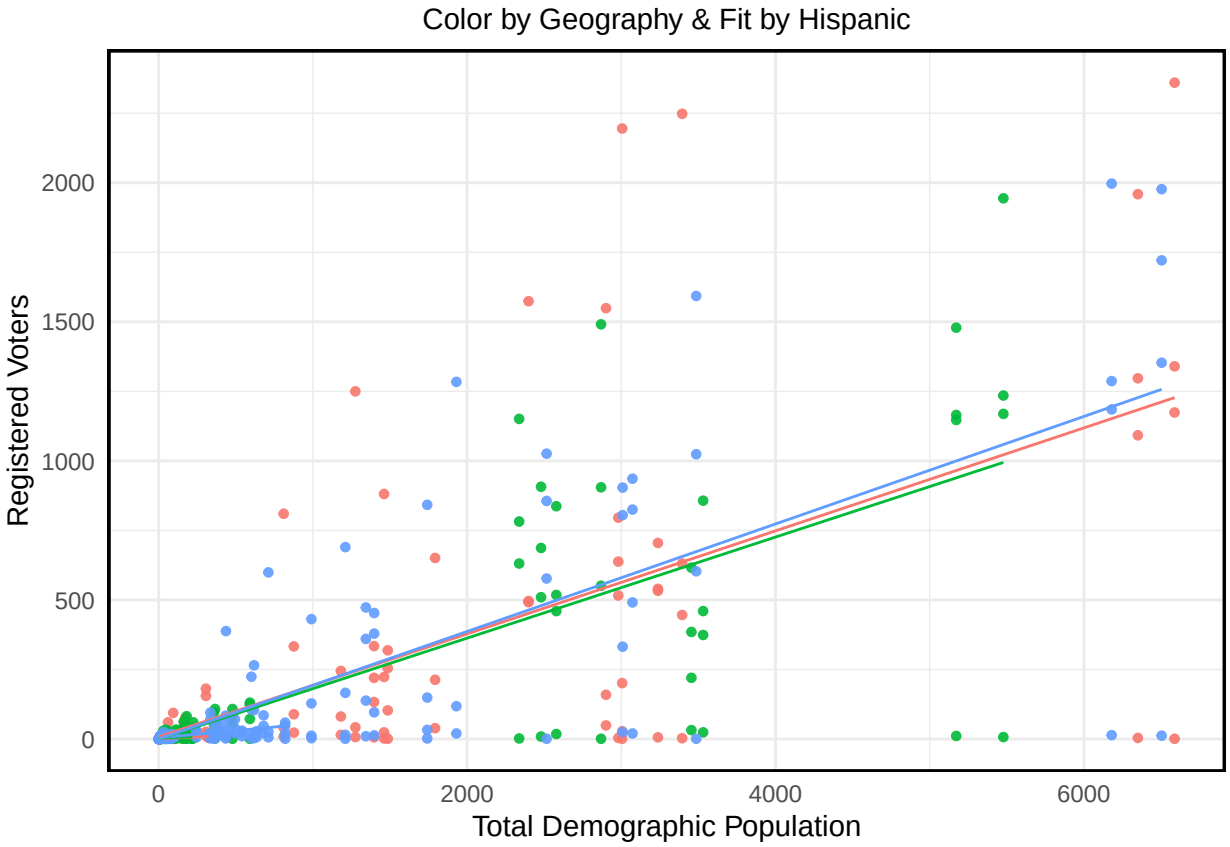
Analysis

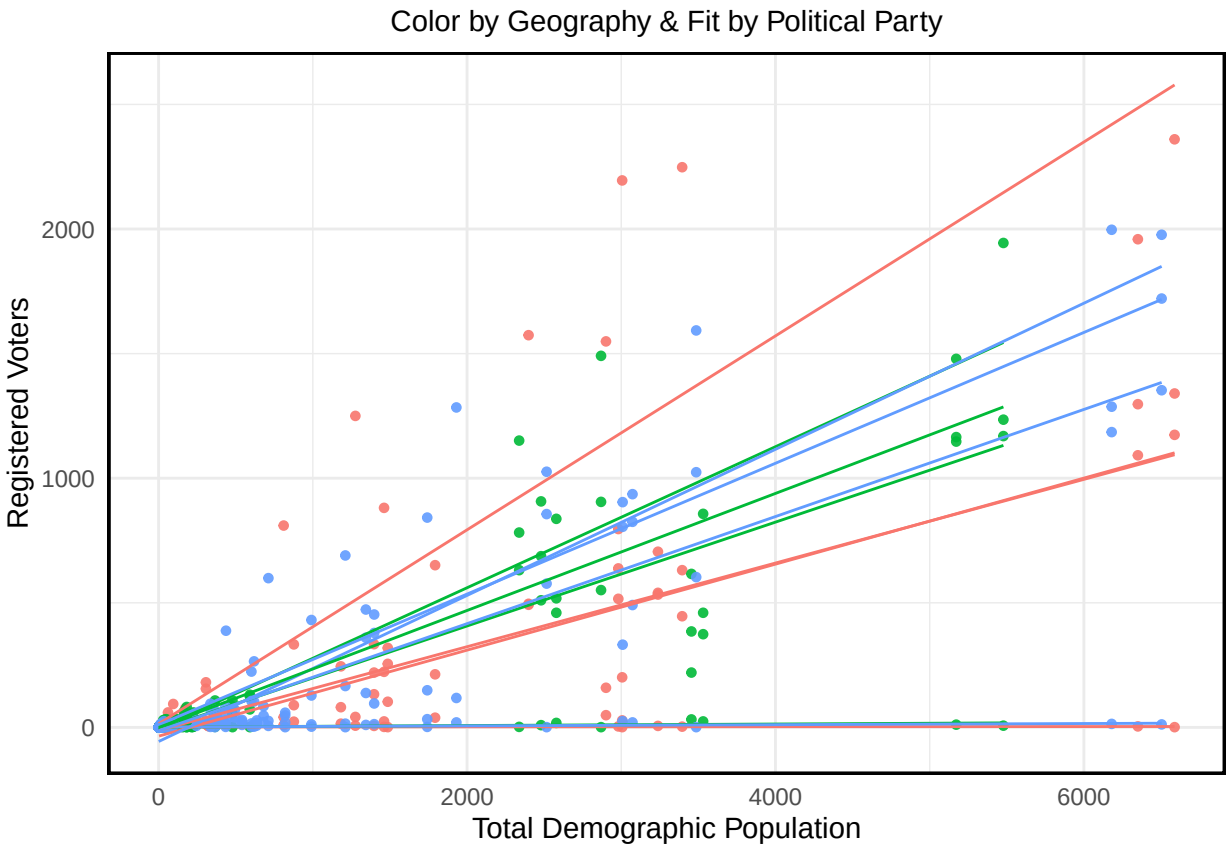
Predictive Performance

Conclusion

Appendix


Color by Geography & Fit by Gender
Registered Voters
2000
1500
1000
500
0
0
2000
4000
6000
Total Demographic Population


Color by Geography & Fit by Hispanic
Registered Voters
2000
1500
1000
500
0
0
2000
4000
6000
Total Demographic Population


Color by Geography & Fit by Political Party
Registered Voters
2000
1000
0
0
2000
4000
6000
Total Demographic Population