Analyzing Voter Registration in the 2016 Presidential Election

Andrew Amore
2022-11-21

Abstract

Political candidates are often interested in understanding factors influencing voter registration rates across different demographic categories and how they effect overall registration probability. Understanding these aspects can better inform campaign staffs on election related decisions, like where to focus limited advertising budgets.

To conduct an analysis, official voter registration records and census population estimates for counties in North Carolina were collected and analyzed using Bayesian hierarchical modeling strategies to better quantify estimate uncertainty across different geographies. The final model indicates...

Analysis

Introduction

Political campaigns analyze historical election data to understand how different demographic groups register to vote, specifically during presidential election years when individuals are more likely to vote. This information can inform optimal advertising strategies, which can drum up more votes and win elections. Campaigns are generally interested in assessing registration differences amongst demographic groups and how registration tendencies vary by county, gender, age and party affiliation.

Dataset Overview

To address the main questions of interest, two data sources were analyzed. Information from the 2010 U.S. Census was collected from the Federal Census Bureau website and enhanced with official 2016 voter registration records from North Carolina. To combine the data, demographic field values were standardized, as the coding structure varies slightly between State and Federal agencies. Field values without a corresponding match were dropped from the analysis. For example, Census estimates quantify two genders (male/female) while registration records include a third unknown category. Irrelevant registration fields denoting precinct location were also removed. Metadata information for the combined dataset and a sample observation, can be viewed in Table 1.

Table 1: Metadata Information

Field Name	Description	Sample
Geography	*County in North Carolina*	*BURKE*
Age	*Age Demographic Category*	*26–40*
Gender	*Gender Demographic Category*	*Female*
Hispanic	*Demographic Indicator of Hispanic Origin*	*Hispanic*
Race	*Race Demographic Category*	*WhiteAlone*
VoterFreq	*Number of Registered Voters in Specified Demographics*	*14*
Freq	*Total Population Count for Specified Demographics*	*128*
TotalCountyPopulation	*Total County Population*	*90912*
PartyCd	*Political Party Affiliation*	*DEM*
VoterTurnout	*County Registration Percentage*	*0.000154*

Population Migration

The 2010 Census is assumed to represent the voter population during the 2016 election, however, almost 10% of observations have more registered voters than the demographic population estimates from the Census. Are we observing potential voter fraud or is the Census information too old to reflect accurate population densities six years later? Both examples in Table 2 have relatively small differences, but suggest additional exploration is warranted.

Table 2: Bad Data Sample

Geography	Age	Gender	Hispanic	Race	PartyCd	VoterFreq	Freq
ONSLOW	26−40	Female	NotHispanic	SomeOtherRaceAlone	REP	114	82
HOKE	41−65	Female	NotHispanic	SomeOtherRaceAlone	REP	24	21

To understand the magnitude of inaccurate estimates, the population difference (registered voters - demographic population) was computed for affected observations and aggregated by county. Figure 1 shows summaries for geographies with more than five observations, for ease of viewing, and displays interesting results. First, several counties have large differences, defined as more than 100 individuals (red line), indicating potentially significant population migration between 2010 and 2016. However, **80%** of counties have median difference less than **50** (green line) and **50%** of counties have median difference less than **10** (blue line). These relatively small differences, effecting only 10% of the total dataset, assuages any concern of major population shifts over the six year period. To minimize MSE, 30 counties (out of 100) were randomly selected using sampling weights inversely proportional to the percentage of faulty observations (registered voters > demographic population). By introducing bias, we hope to reduce estimate variation by steering our model to use counties with low numbers of faulty observations. The final dataset still includes a smaller percentage of invalid observations and voter registrations for these values were set to Demographic Population − 1. An additional benefit of the sub-sampling strategy is a reduction in computational overhead for an MCMC sampler in Bayesian models allowing our models to run in a reasonable amount of time on a personal computer.

Motivating a Multilevel Model

To evaluate modeling decisions, like where to apply random effects, an exploratory data analysis (EDA) was conducted. We are interested in understanding how demographic covariates affect the number of registered voters ("successes") relative to respective population estimates ("trials"), suggesting a binomial model. Figure 2 is a split plot of different binomial regressions on voter registration probability ($\frac{registrations}{demographic\ population}$) for different demographic categories and color coded by county.

The left plot fits regressions by county and age group. The substantially different trend lines by group (age and county) suggest a **random slope**.

The right plot fits regressions by county and race. The consistent slopes across counties for identical demographic groupings suggest a **fixed effect**.

Plots for additional covariates can by found in the Appendix, but findings are summarized here. Gender effects are consistent across geography and suggest a **fixed effect**. Party affiliation displays consistent trends, but the intercept varies by county suggesting a **random intercept**. Hispanic indicator shows consistent patterns and suggests a **fixed effect**.

Model Specification

The modeling framework can now be specified as follows. Let $(y_{ijk}, n_{ijk}, p_{ijk})$ denote the number of voter registrations, demographic population and registration probability respectively for county i, age group j and vector of fixed effect covariates k. y_{ijk} is assumed to be distributed as a binomial random variable.

$$y_{ijk} \sim Binomial\left(n_{ijk},\ p_{ijk}\right)$$

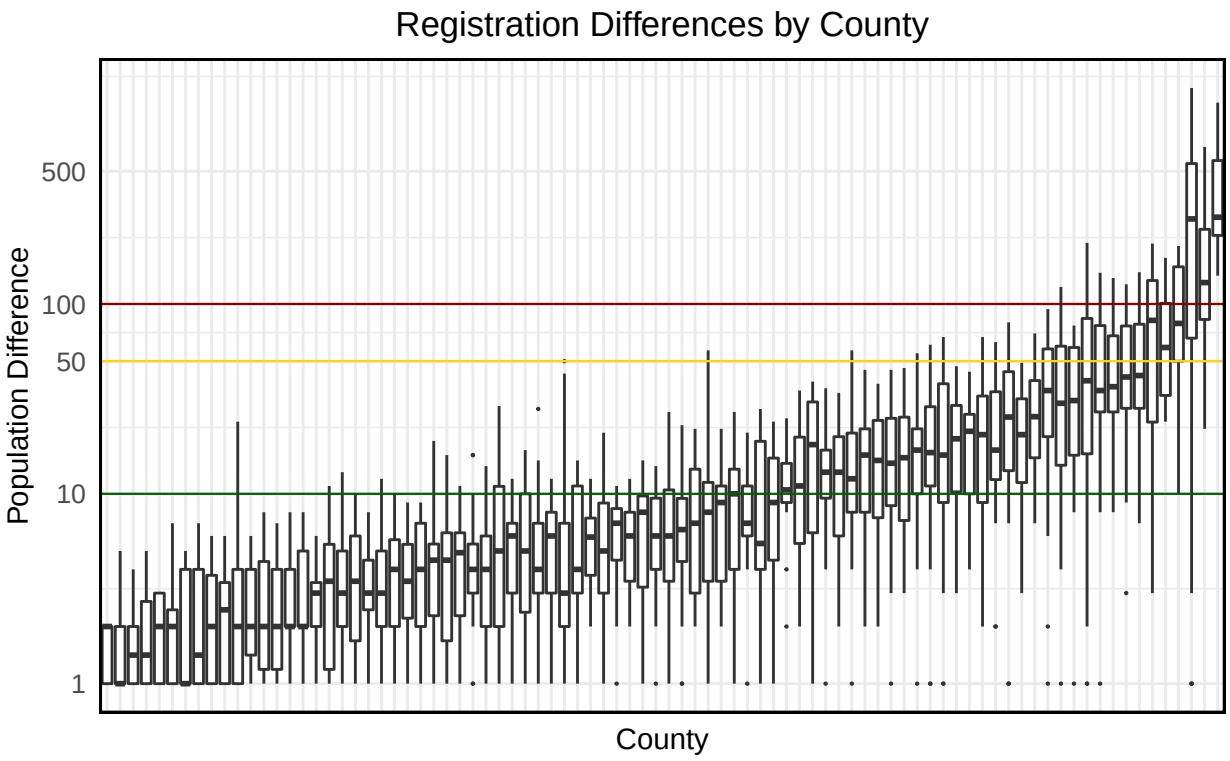


Figure 1: Registration Differences by County

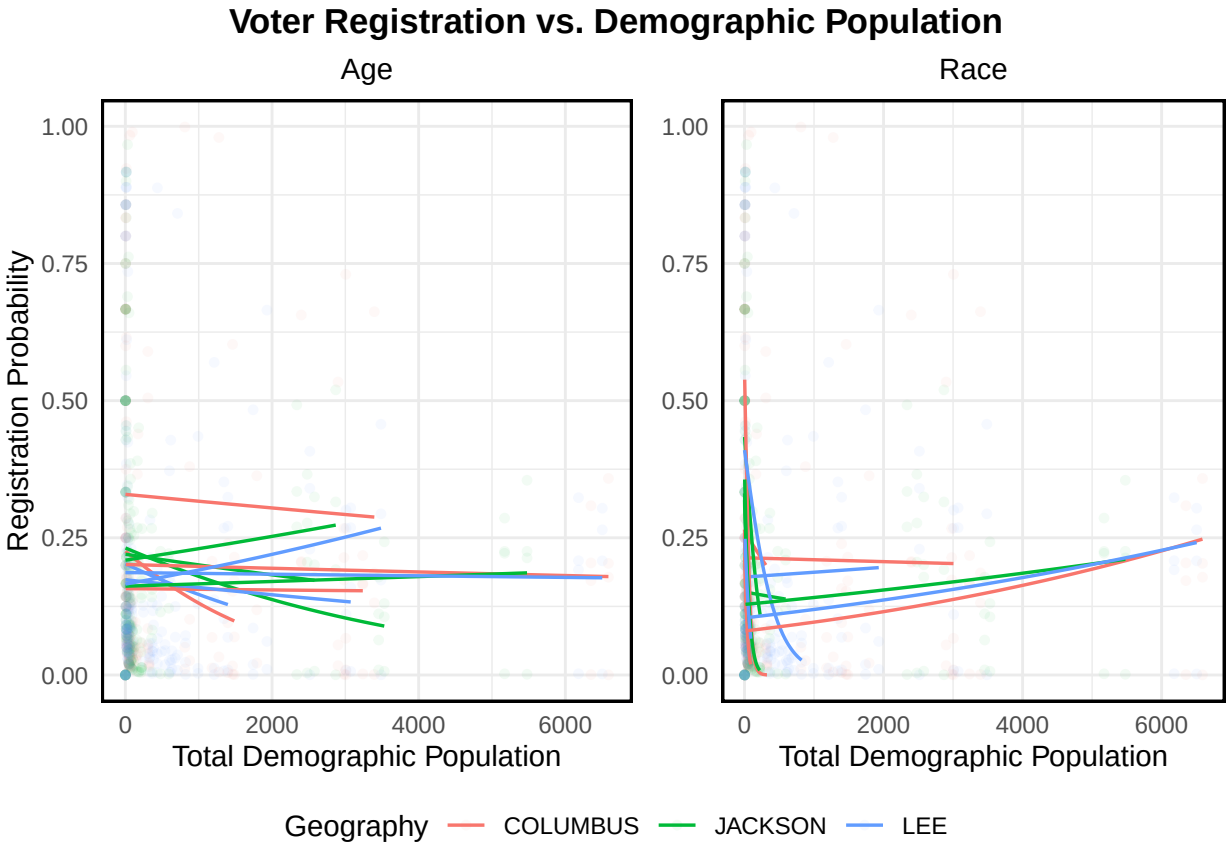


Figure 2: Voter Registration Behavior

Individuals can only register to vote once, but binomial trials sample independently with replacement. The independent draw assumption seems reasonable, as the probability of any one individual registering to vote is most likely independent of any other individual's decision (more on this later). Additionally, $\sim 70\%$ of demographic categories in our dataset have populations $n_{ijk} \geq 30$. However, the demographic samples can be considered population estimates and may warrant a hypergeometric distribution, but this induces a dependence between y_{ijk}'s which is unwarranted based on our independent registration assumption. Under the binomial model, several multilevel structures with varying complexity were evaluted. Table 3 displays the specifications.

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 IV	Gender, Hispanic, Race, PartyCd, Age	Geography, Age, PartyCd	Age

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.

$$Logit(p_{ijk})) = \mu + \theta_i + X_{k[ij]}\beta$$

Prior Specifications

$$\mu \sim N\left(0, 1\right), \ \beta \sim N\left(0, 1\right), \ \theta_i \sim N\left(0, \sigma\right)$$
$$\sigma \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Model II

Model II is an extension 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.

$$Logit(p_{ijk})) = \mu + \theta_i + a_{ij[k]}\Gamma_{ij} + X_{k[ij]}\beta$$

Additional Prior Specifications

$$\Gamma_{ij} \sim N\left(0, \gamma\right), \ \gamma \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Model III

Model III is an extension of II with an additional random effect for **Party Affiliation**. Let $q_{ip,[jk]}$ denote the party, p, for observations in county i and Ω_{ip} the random effect.

$$Logit(p_{ijkp})) = \mu + \theta_i + a_{ij[k]}\Gamma_{ij} + q_{ip[jk]}\Omega_{ip} + X_{k[i,j]}\beta$$

Additional Prior Specifications

$$\Omega_{il} \sim N\left(0, \omega\right), \ \omega \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Model IV

Model IV is an extension of II with an additional random effects for **Age**, α_j, and **Party Affiliation**, ζ_p.

$$Logit(p_{ijlk})) = \mu + \theta_i + \alpha_j + \zeta_p + a_{ij[k]}\Gamma_{ij} + X_{k[i,j]}\beta$$

Additional Prior Specifications

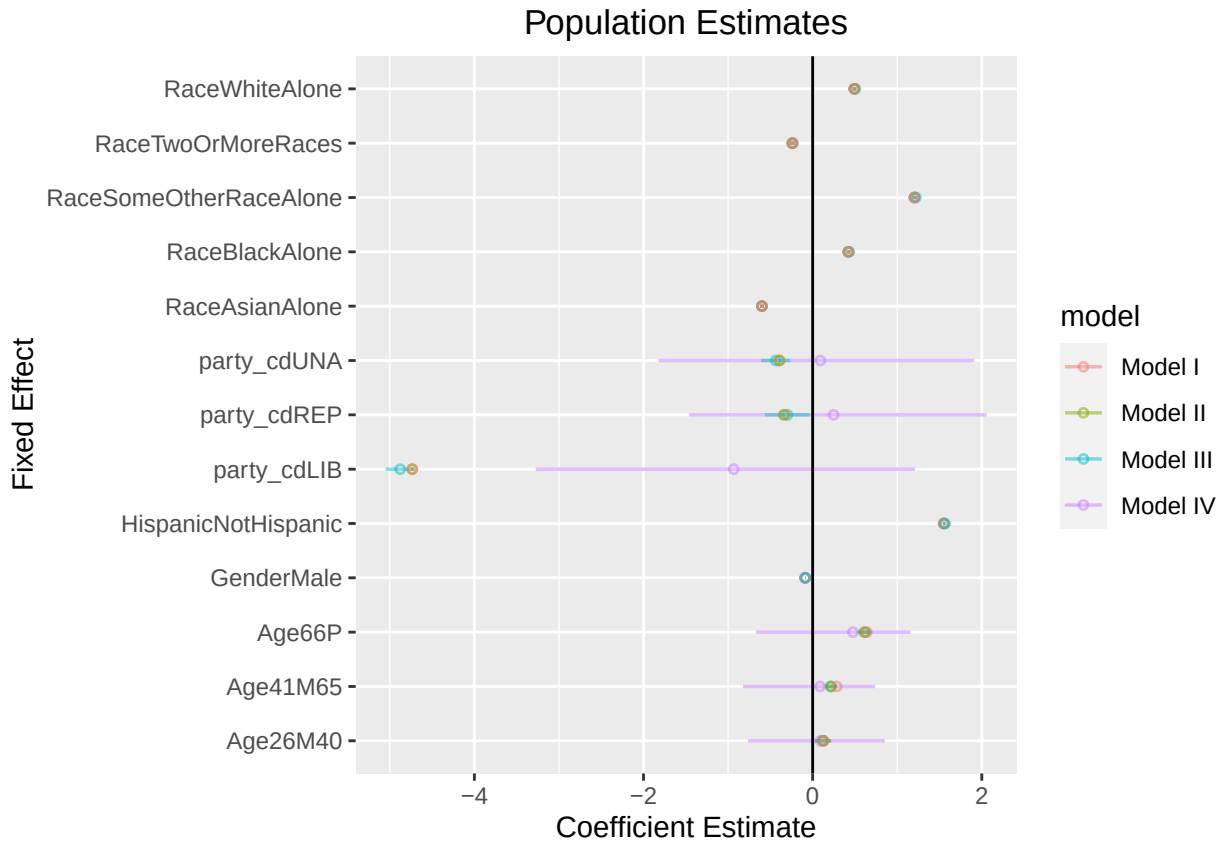
$$\alpha_j \sim N\left(0, \sigma_\alpha\right), \ \zeta_p \sim N\left(0, \sigma_\zeta\right)$$

$$\sigma_\alpha, \sigma_\zeta \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

Analysis

The main questions of interest concern...

Assessing Model Fit & Comparing Models



Limitations

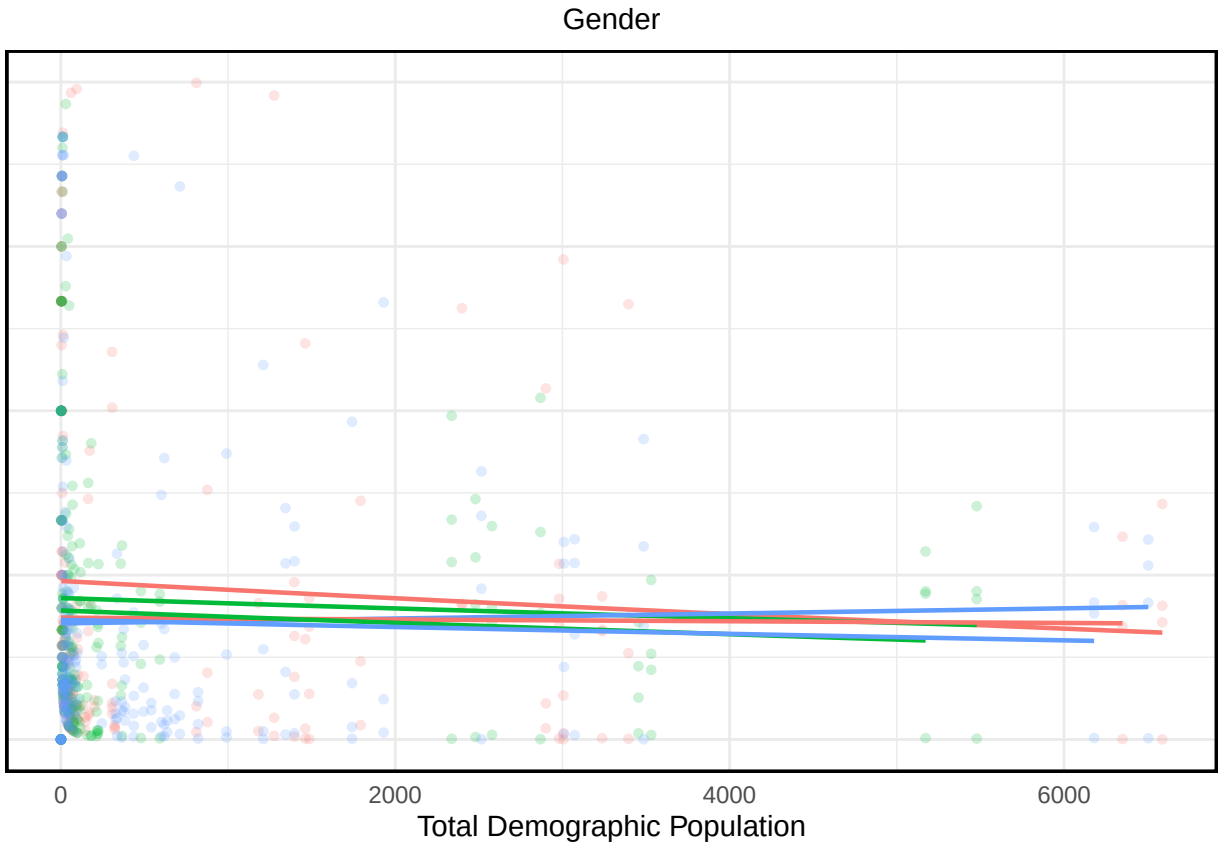
Several limitations have been mentioned throughout and are summarized. First, we have concerns about the viability of the 2010 Census for population estimates in 2016 as some observations have more voter registrations than population estimates. A weighted sub-sampling strategy was introduced to minimize the statistical impact, but only addresses detectable observations where $registration > population$. A valid concern is that all demographic populations undergo some change over the six year period. A potential solution is to impute or average the demographic populations in 2016 using estimates from both the 2010 and 2020 Censuses.

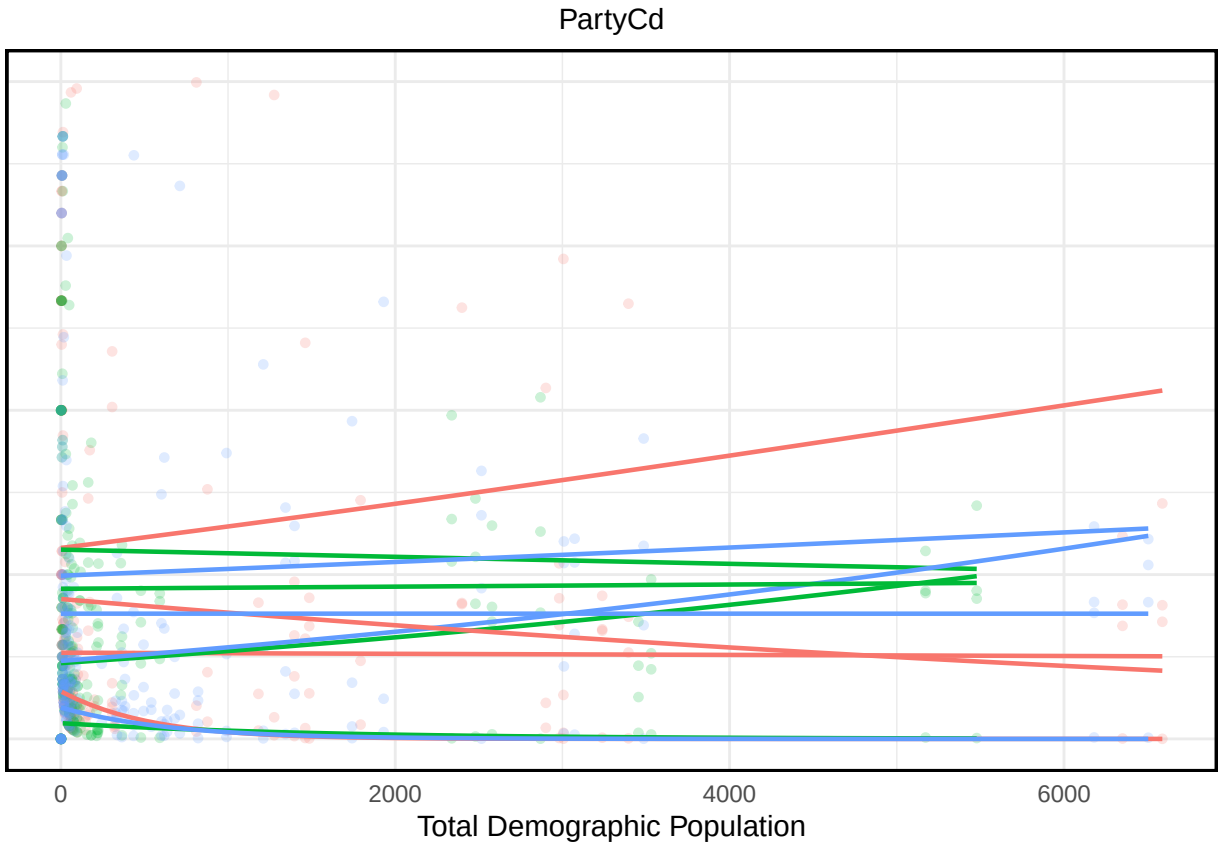
Another limitation, is from the sub-sampling strategy that limits the model to only 30% of counties which makes our statistical estimates more variable. To leverage the entire dataset one can either acquire more computation or specify simpler models that might not capture the true signal.

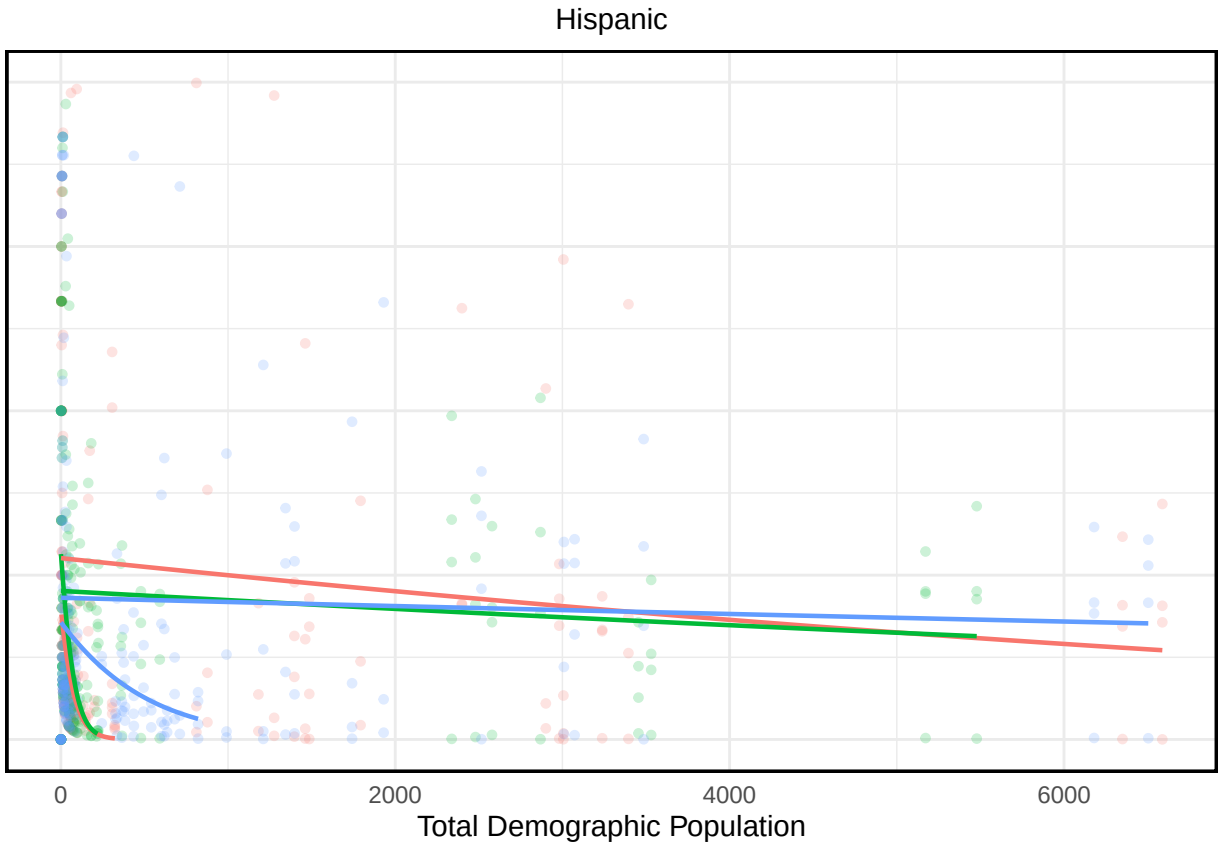
Conclusion

Historical voting records and Census population estimates were used to evaluate demographic differences in registration tendencies.

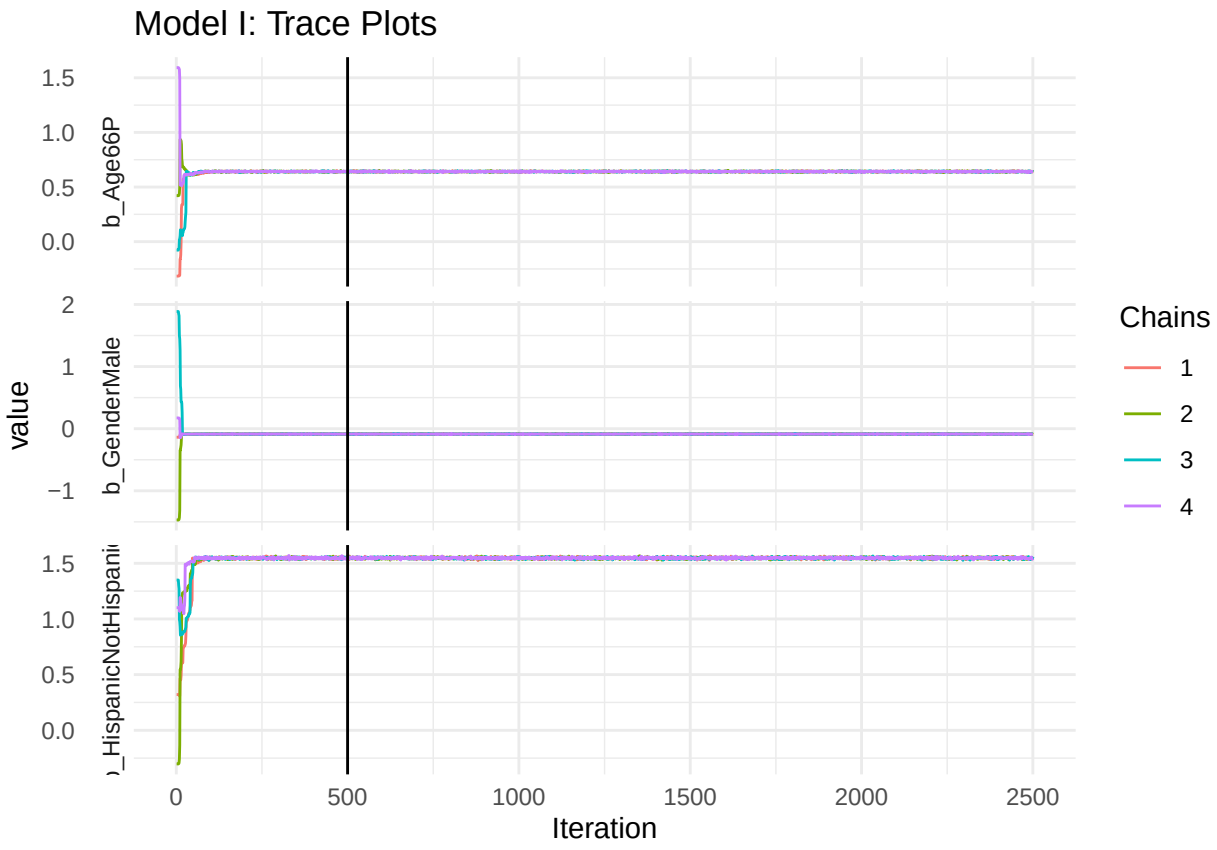
Appendix


Gender
0
2000
4000
6000
Total Demographic Population


PartyCd
0
2000
4000
6000
Total Demographic Population


Hispanic
0
2000
4000
6000
Total Demographic Population

Trace Plots


Model I: Trace Plots
b_Age66P
b_GenderMale
b_HispanicNotHispani
value
Iteration
Chains
1
2
3
4