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*	*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	*County Registration Percentage*	*0.000036*

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. Both examples in Table 2 have relatively small differences, but the large number of effected observations warrant additional exploration.

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

Figure 1 displays county summaries for the difference between registered voters and demographic population. Several counties have large differences, defined as 100 or more individuals, which may indicate significant population migration. It's plausible that the 2010 Census incorrectly reflects population densities in 2016, however, the majority of average county differences are less than 50, which is relatively small. To address this concern, 30 counties were randomly drawn from the 90 county population using sampling weights inversely proportional to the percentage of "incorrect" observations (registered voters > demographic population). This sub-sampling strategy also reduces computational overhead for an MCMC sampler in Bayesian models.

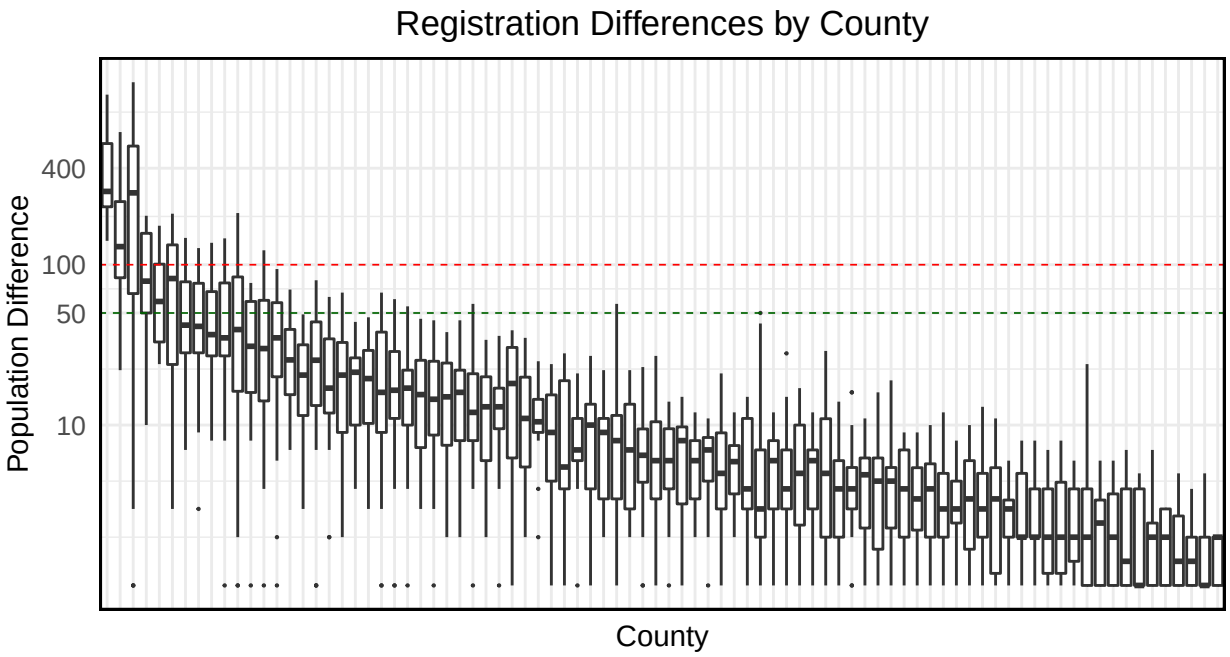


Figure 1: Registration Differences by County

Motivating a Multilevel 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.

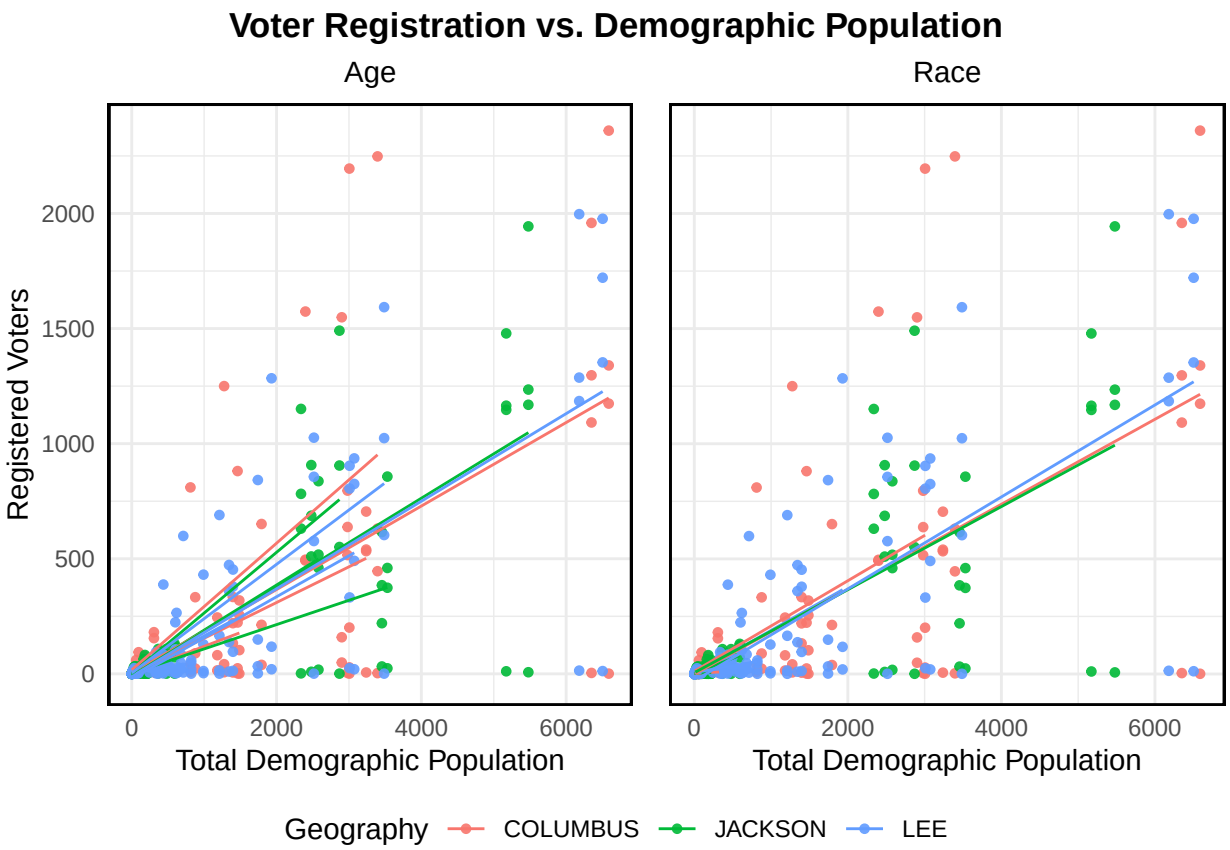


Figure 2: Voter Registration Behavior

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.

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

Model II
Model II is an extension 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.

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

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.

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

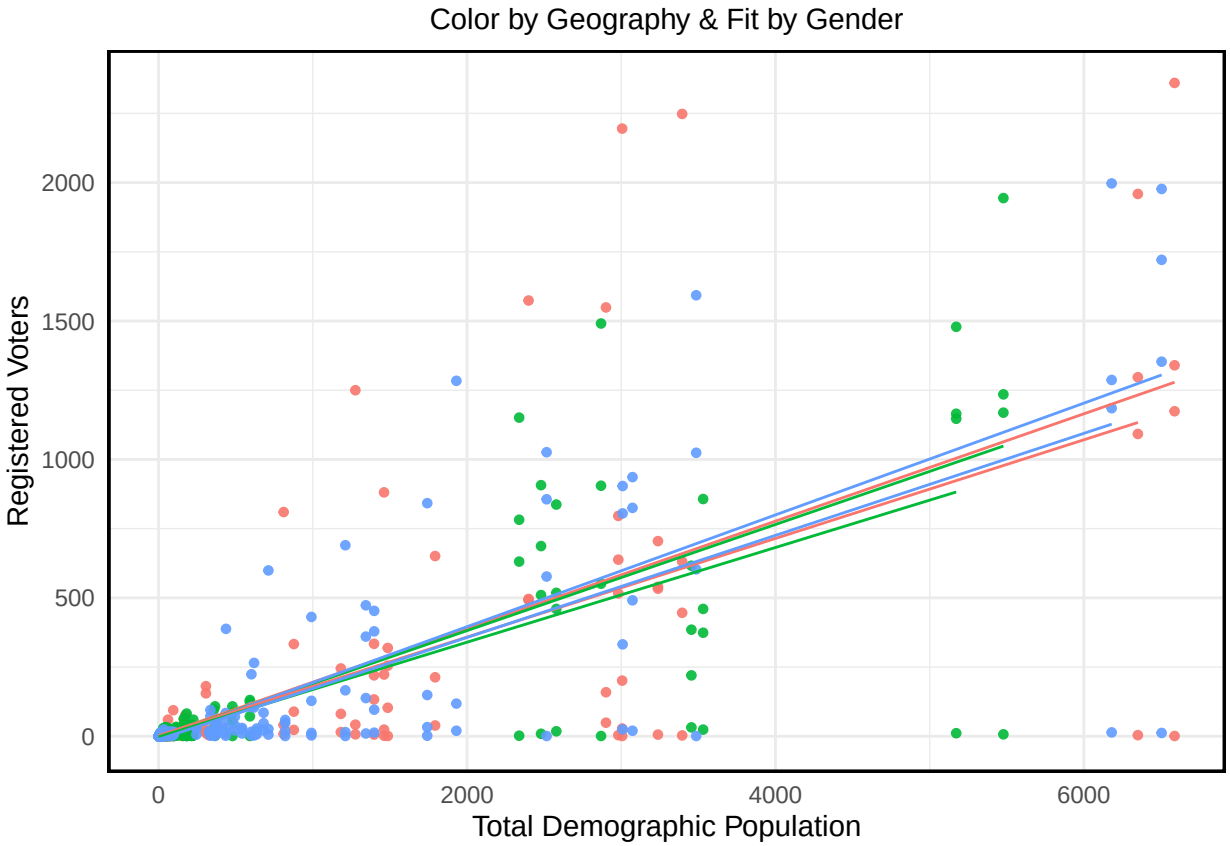
Assessing Model Fit & Comparing Models

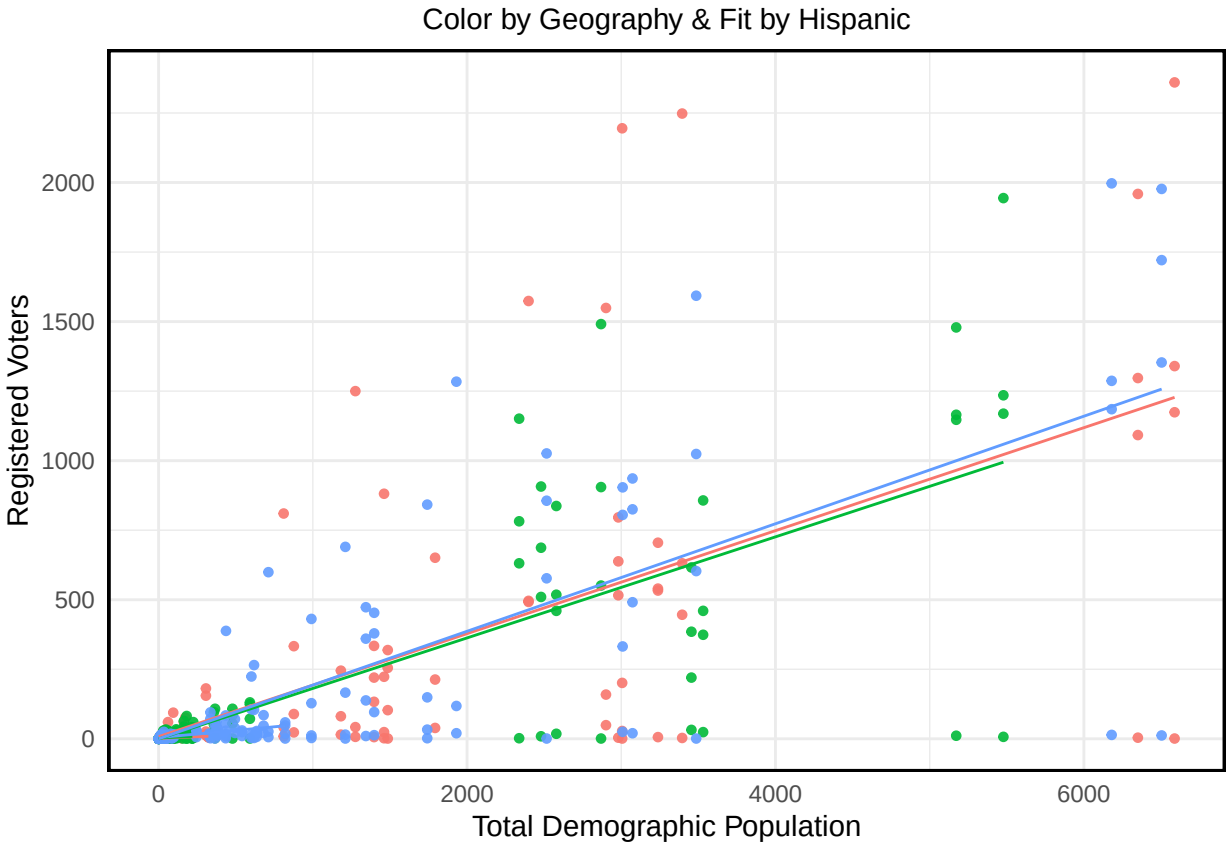
Analysis

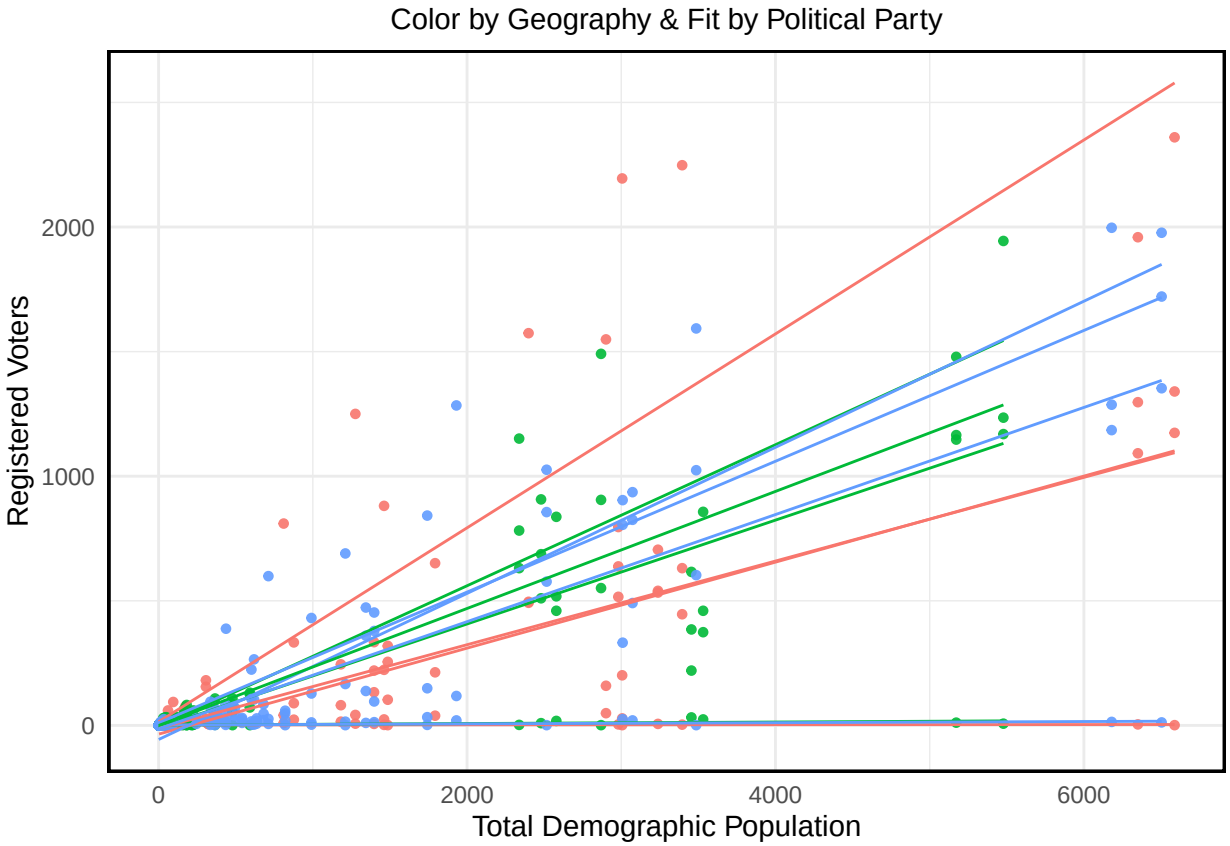
Predictive Performance

Conclusion

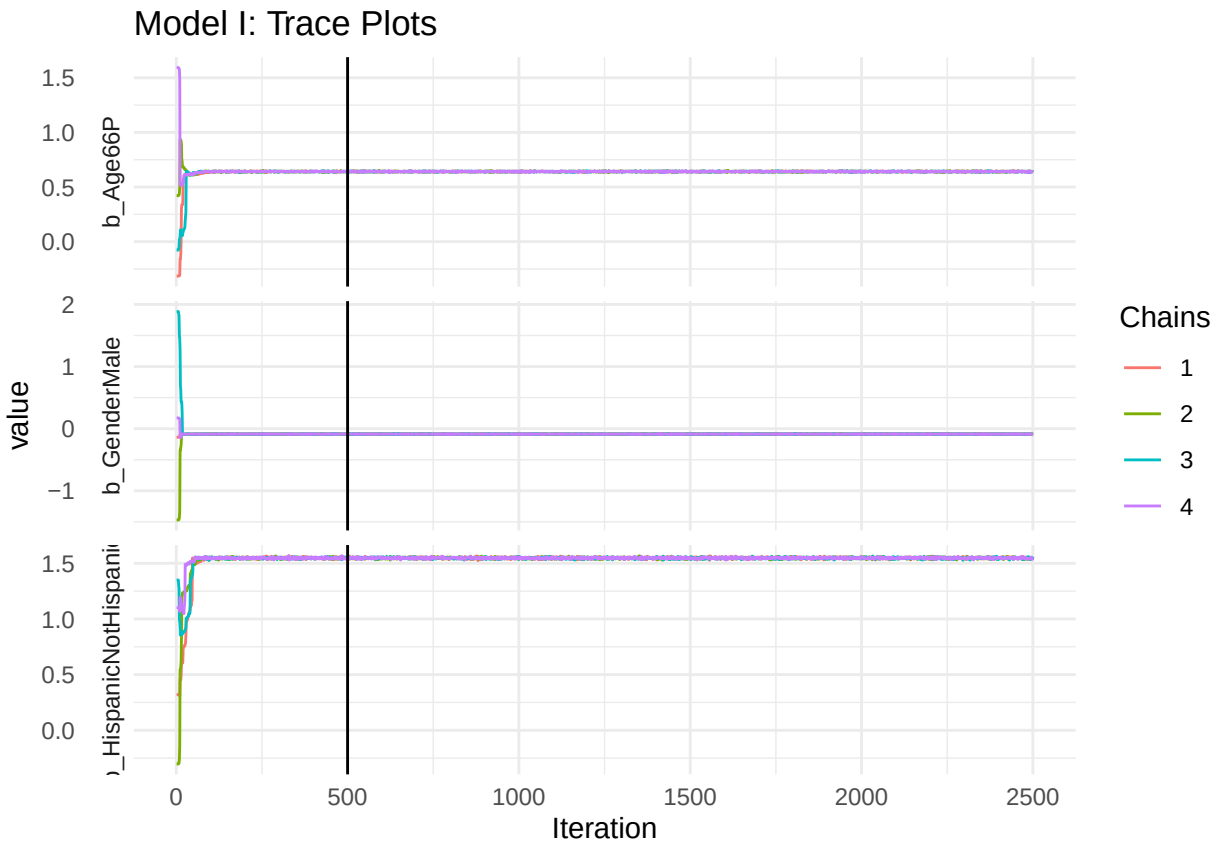
Appendix


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


Color by Geography & Fit by Hispanic
Registered Voters
0
500
1000
1500
2000
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

Trace Plots


Model I: Trace Plots
b_Age66P
b_GenderMale
b_HispanicNotHispani
value
1.5
1.0
0.5
0.0
2
1
0
−1
0
500
1000
1500
2000
2500
Iteration
Chains
1
2
3
4