Analyzing Voter Registration in the 2016 Presidential Election

Andrew Amore

2022-11-21

1 Introduction

Political campaigns analyze historical election data to understand how demographic factors influence voter registration rates. This information can inform optimal advertising strategies which can drum up more votes and win elections, especially during high profile presidential election years when voter turnout can be substantially higher.

In this report, 2016 election data from North Carolina counties was used to investigate how geography, gender, age and party affiliation influence registration behavior. Section 2 provides some background information on hierarchical modeling and discusses some favorable properties attractive under the voter registration context. Section 3 reviews the data used in the analysis, highlights a challenge arising from dated population estimates and motivates hierarchical modeling. Section 4 defines several multilevel model specifications, fitted to answer the main questions of interest using the voter and population data, and Section 5 reviews the results. Lastly, Section 6 introduces limitations of the analysis and summarizes relevant findings.

2 Background Information

Experimental data is often collected with an inherent nesting which can violate traditional modeling assumptions, like i.i.d. errors, and lead to poor uncertainty estimation if ignored. A hierarchical model is a statistical inference tool that addresses this concern by replicating the natural nesting structure of observed data. In multilevel or hierarchical modeling, model parameters are defined at different group membership levels and permitted to vary across groups. This can account for heterogeneity and improve uncertainty estimation attributed to group membership. Voter registration and population estimates are generally aggregated by geography (most naturally by county) during collection. In politics, it's generally known that individuals display different tendencies depending on geography (urban vs. rural). This structure lends itself well to hierarchical modeling, allowing parameters to vary by geography for more accurate effect estimates.

Hierarchical models can be estimated using Bayesian or Frequentist assumptions and I focus on Bayesian methods. In Bayesian estimation, parameters are treated as random quantities and derived from posterior distributions using Bayes rule. Posteriors provide more direct uncertainty quantification than Frequentist methods through credible regions (CR), which avoid caveats of traditional confidence intervals. However, this interpretability can come with increased computational costs, as Bayesian inference relies on expensive sampling methods compared to likelihood based optimization under Frequentist methods. This analysis utilizes conjugate priors to avoid excessive computational costs.

3 Data Overview

To investigate the main questions of interest data from the 2010 U.S. Census was enhanced with North Carolina voter registration records from the 2016 Presidential election. To combine datasets, demographic field values were standardized, as the coding structure varies slightly between State

and Federal agencies. In total, there were 6,858 county level observations with unknown gender specification (~15%), lacking a matching population estimate from the Census and these records were dropped. 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*

3.1 Population Migration

The 2010 Census is assumed to represent the voter population during the 2016 election, however, **16.3%** of all observations have more registered voters than the demographic population estimates from the Census. Are we observing potential voter fraud or is Census information from six years prior too dated to reflect accurate population metrics at the time of the election? To understand the scope of this issue, the difference between total registered voters and Census estimates were computed and aggregated by county. Figure 1 shows summaries for geographies with more than five observations for ease of viewing. Coincidentally, **16%** of counties have median population difference greater than **100 individuals**, however, **70%** of counties have median difference less than **50** and **35%** have medians less than **10**. The majority of moderately small differences assuages some concerns of major population shifts over the six year period. Population estimates for the "invalid" observations are set to the sum of registered voters, reflecting a 100% registration rate (a potential source of error), but could also be inflated with a correction factor learned from other observations.

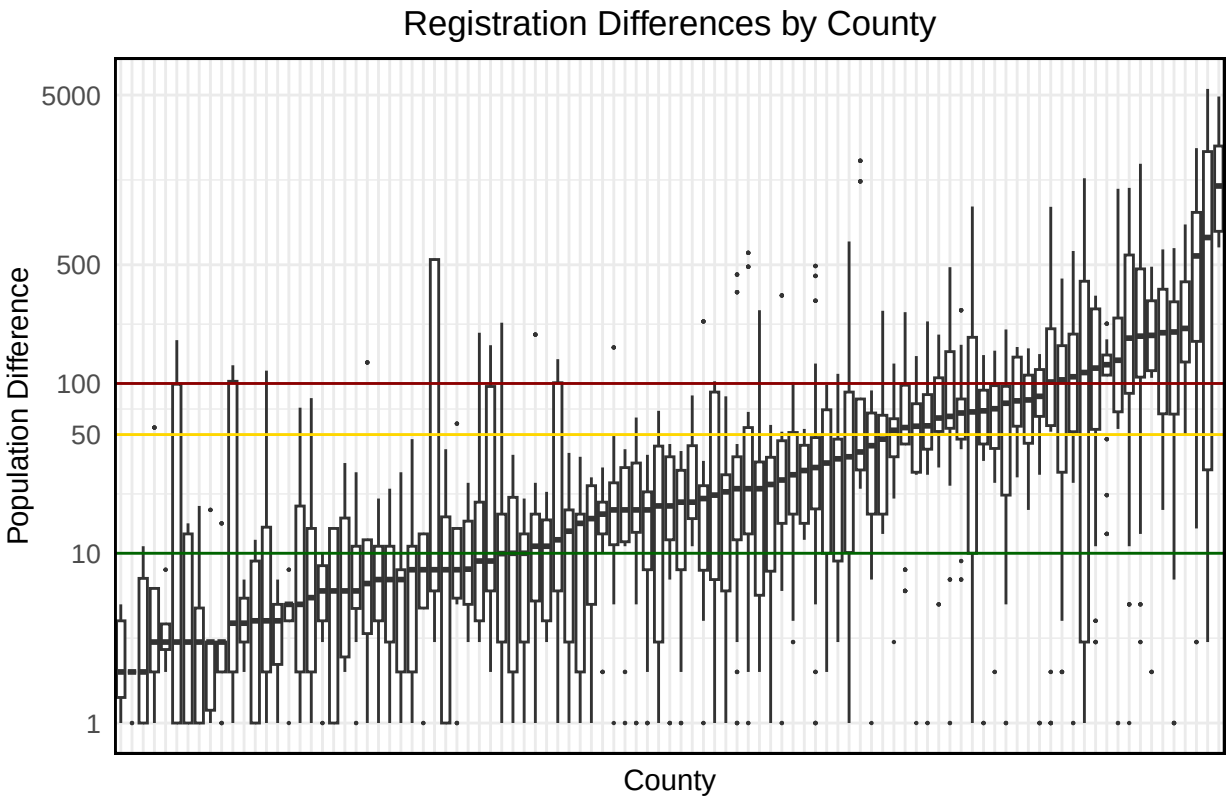


Figure 1: Registration Differences by County

To minimize error, 30 counties were randomly selected using sampling weights inversely proportional to the percentage of faulty observations (total registered voters > demographic population). This framework can be easily extended to a bootstrap resampling procedure for different county combinations to reduce variance. By using sampling weights I hope to reduce estimate variation by steering the model towards more accurate data at the expense of added bias. An additional benefit of sub-sampling is a reduction in computation time for MCMC inference.

3.2 Motivating a Multilevel Model

Voter data and population estimates are naturally grouped by geography, suggesting a multilevel model using county. We are interested in understanding how covariates, like age and political affiliation, influence registration rates. A model, without any hierarchical components, can be formatted as a binomial regression, where registered voters represent the number of "successes" from a set of "trials" reflected as population estimates. To evaluate modeling decisions, like where to apply random effects, an exploratory data analysis was conducted. Figure 2 displays facet plots for different binomial regressions for two demographic categories and color coded by county. (A) fits regressions by county/age and displays substantially different trends across/within facets, suggesting a random effect. While (B) fits regressions by county/race and displays much lower facet variability, suggesting a fixed effect. Additional combinations can by found in the Appendix. Continuous covariates, like total county population, hindered posterior mixing and were excluded.

Under the current specification it's impossible to determine the effect of political affiliation on voter registration, as the Census doesn't collect it with population estimates and it needs to be imputed. As an initial estimate, political affiliation can be assigned to unregistered voters from the observed party distribution of registered voters for each demographic group.

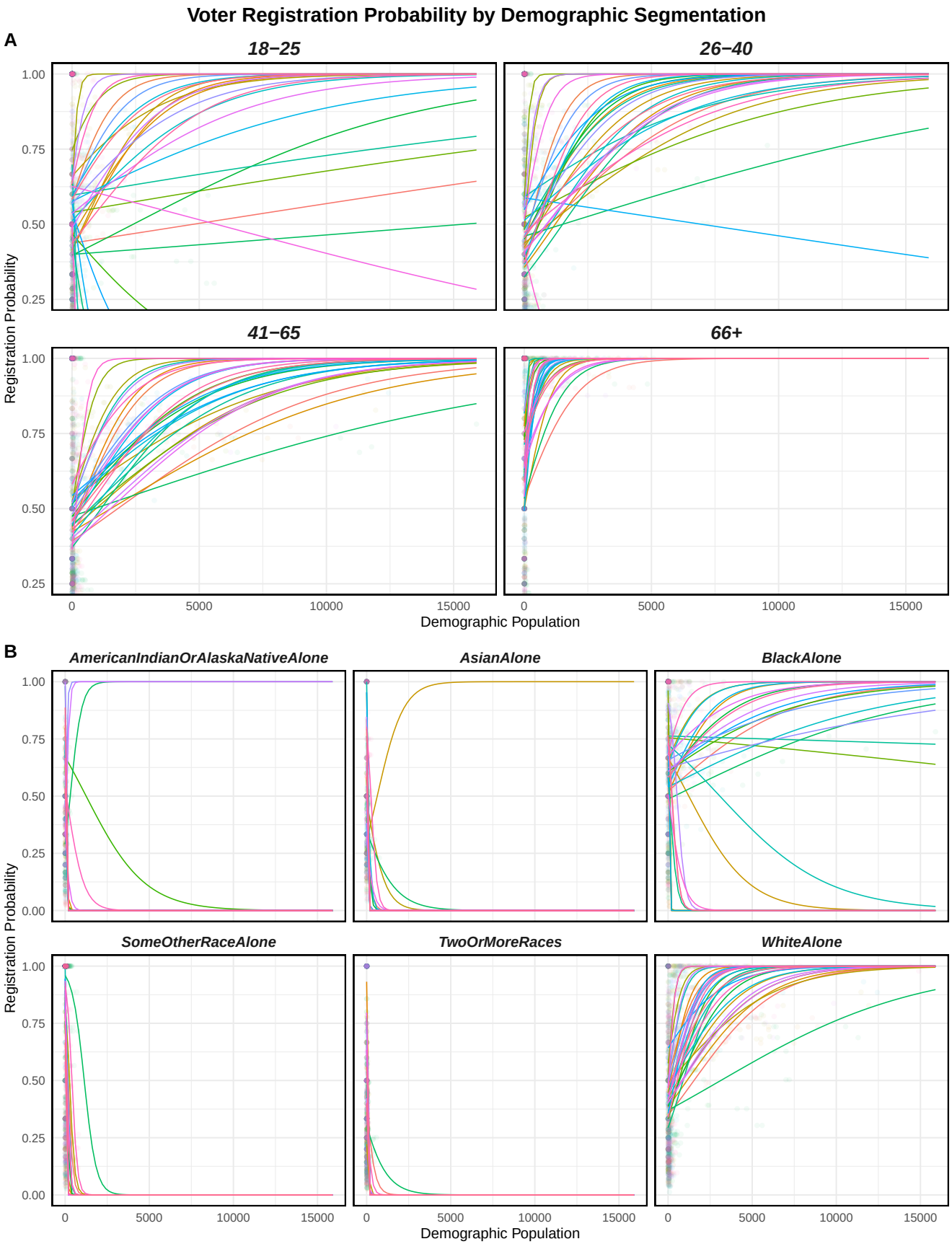


Figure 2: Voter Registration Behavior for (A) Age Category and (B) Race

4 Multilevel Models

4.1 Model Specifications

The modeling framework can now be specified as follows. Let (y_i, n_i, p_i) denote the number of voter registrations, demographic population and registration probability respectively for observation i. y_i is assumed to be distributed as a binomial random variable.

$$y_i \sim Binomial\left(n_i,\ p_i\right)$$

Individuals can only register to vote once, but binomial models assume independent sampling with replacement. This independence assumption seems reasonable, as the registration probability of any one individual is most likely independent of any other individual decision (more on this later). Additionally, $\sim 70\%$ of demographic categories have populations $n_i \geq 30$ which are sufficiently large. One might consider a hypergeometric model for more realism, but this would induce a dependence between y_i's which may be unwarranted. Under the binomial model, several multilevel structures, with varying complexity, were evaluated and Table 2 displays specifications for each.

Table 2: Random Effect Structures

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

4.1.1 Model I

Let μ denote a global intercept corresponding to a baseline demographic category with the following values: 'Female', 'Hispanic', 'AmericanIndianOrAlaskaNativeAlone', 'DEM' and '18-25'. Let θ_j denote a **county** random effect intercept for observations, i, in county j, X_i a vector of demographic indicator variables corresponding to Fixed Effect Fields in Table 2 and β the corresponding fixed effect estimates.

$$Logit(p_{ij})) = \mu + \theta_j + X_i\beta$$

Prior Specifications

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

4.1.2 Model II

In accordance with Table 2, let μ denote a global intercept corresponding to a baseline demographic category with the following values: 'Hispanic', 'AmericanIndianOrAlaskaNativeAlone' and '18-25'. Let Ω_p denote the **party affiliation** random effect for party p, let g_{ipk} denote the **gender** category, k, for observation i in political party p and Γ_{pk} the random effect. Finally, X_i and β have identical interpretations to what was described for Model I.

$$Logit(p_{ipk})) = \mu + \Omega_p + g_{ipk}\Gamma_{pk} + X_i\beta$$

Prior Specifications

$$\Omega_p \sim N\left(0, \sigma\right),\ \Gamma_{pk} \sim N\left(0, \gamma\right),\ \sigma, \gamma \sim HalfCauchy\left(0, \frac{1}{2}\right)$$
$$\mu \sim N\left(0, 1\right),\ \beta \sim N\left(0, 1\right)$$

4.1.3 Model III

In accordance with Table 2, let μ denote a global intercept corresponding to a baseline demographic category with the following values: 'Female', 'Hispanic' and 'AmericanIndianOrAlaskaNativeAlone'. Let Ω_p denote the **party affiliation** random effect for party p, let a_{ipk} denote the **age** category, k for observation i in political party p and α_{pk} the random effect slope. Finally, X_i and β have identical interpretations to what was described for Model I.

$$Logit(p_{ipk})) = \mu + \Omega_p + a_{ipk}\alpha_{pk} + X_i\beta$$

Prior Specifications

$$\Omega_p \sim N\left(0, \omega\right), \ \alpha_{pk} \sim N\left(0, \tau\right), \ \omega, \tau \sim HalfCauchy\left(0, \frac{1}{2}\right)$$

$$\mu \sim N\left(0, 1\right), \ \beta \sim N\left(0, 1\right)$$

5 Results

MCMC diagnostics, like trace plots, can be viewed in the Appendix for each model. The main questions of interest concern demographic factors influencing registration probabilities and are now discussed.

5.1 *How did different demographic subgroups register to vote?*

The fixed effect estimates from each model are plotted in Figure 3 with 95% CR. Note, not all models include identical fixed effects, but the overlapping estimates tend to display more shrinkage with increased model complexity. Significance is interpreted as CR not encompassing 0. Starting from the top of Figure 3, all Race indicators have significant impact on registration rate while party affiliation displays no significant effect. Hispanic & Gender estimates are significant and Older individuals are also more likely to register, but this can be attributed to having more opportunities/elections to do so.

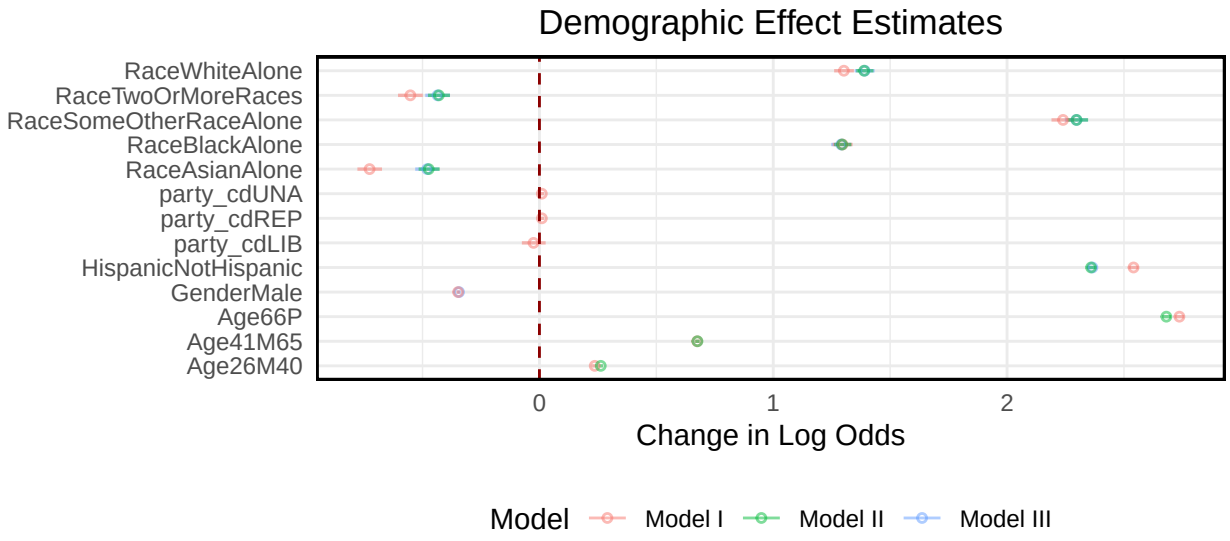


Figure 3: Fixed Effect Estimates From Each Model

5.2 *Did the overall odds of registering differ by county in 2016?*

Only 30 counties, viewable in the Appendix, are included in estimation. The random intercepts from Model I, with 95% CR are displayed in Figure 4. Tyrrell, the only county without any "invalid" observations, has the lowest registration effect, while **43%** of counties display no effect.

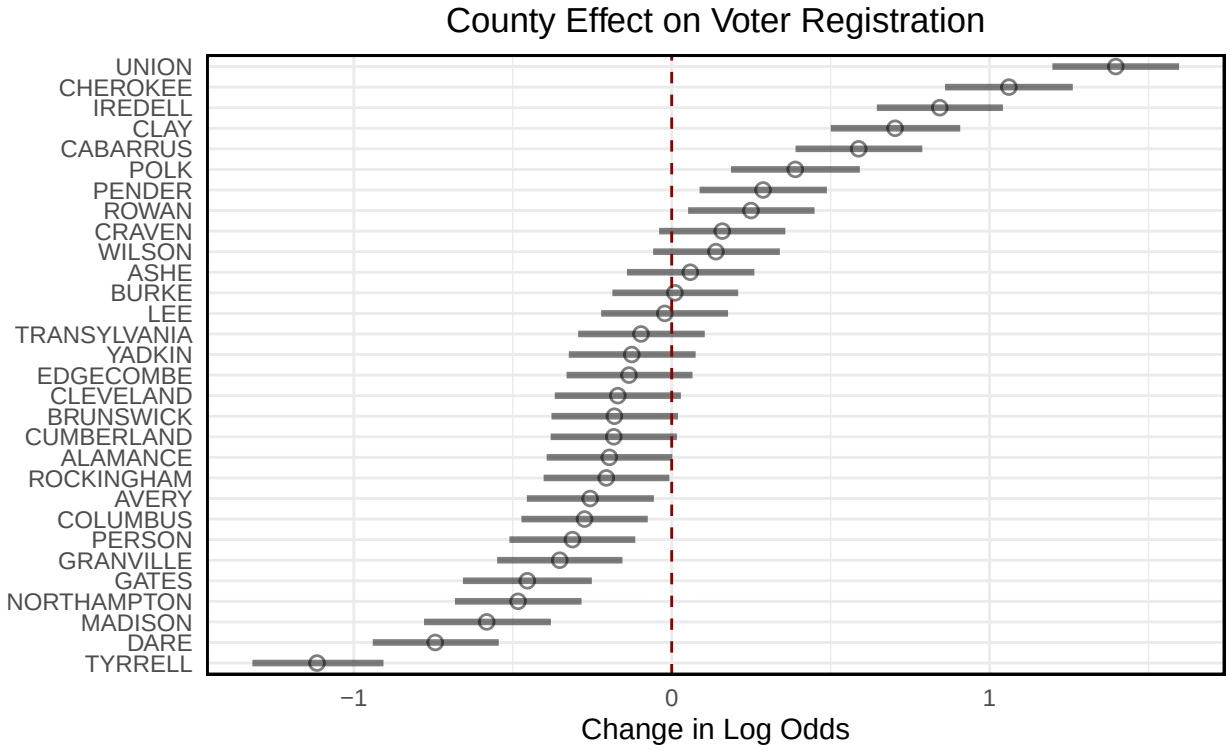


Figure 4: County Level Intercept Estimates from Model I

5.3 *How do registration rates differ by gender for each party?*

Using Model II we can look at the gender effect within each political party from random slope estimates and assess the relative change in log odds from the baseline (Female). Figure 5 shows 95% CRs and point estimates. Males across all political parties are less likely to register than corresponding females and republican males register at the highest rates.

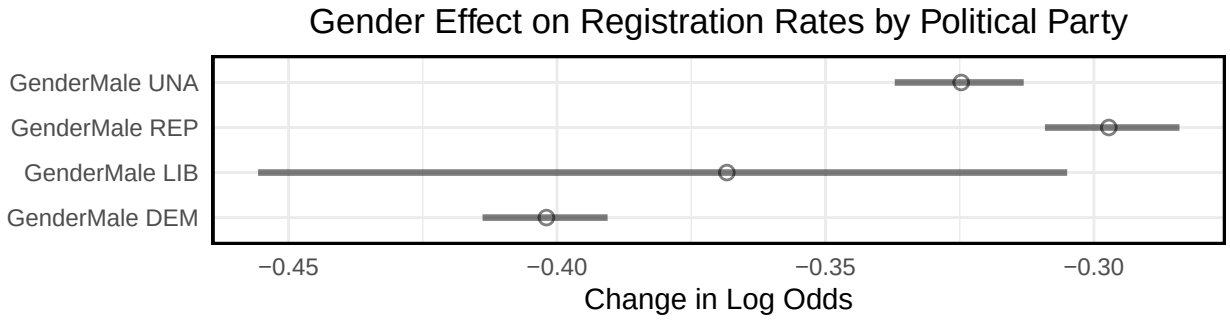


Figure 5: Random Slope Estimates (Model II) for Gender across Political Party

5.4 *Registration differences by age group for each party?*

Random slope estimates from Model III are the registration effects for different age groups relative to a baseline (18-25). Figure 6 shows 95% CRs and point estimates for different political parties

and age groups. Across age categories republicans are more likely to register than other political parties and registration rates tend to increase with age.

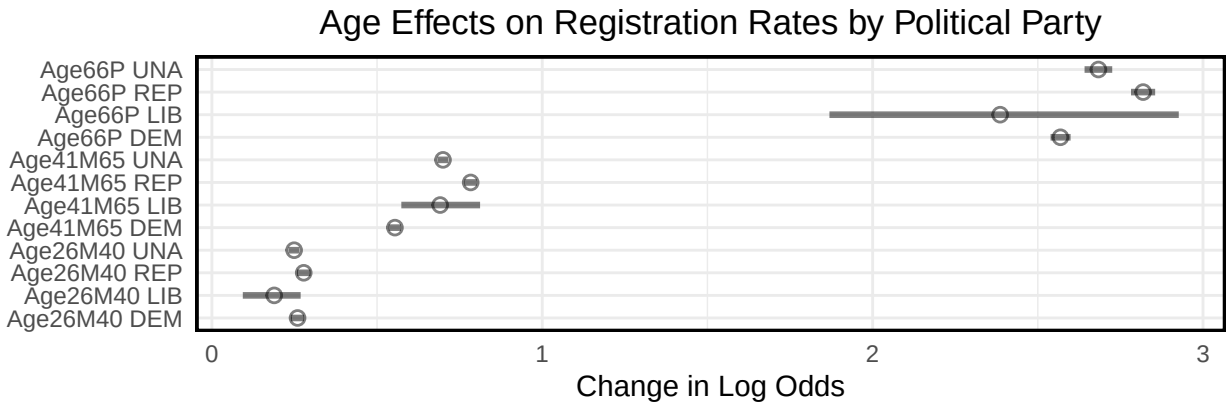


Figure 6: Random Slope Estimates for Age across Political Party

5.5 Predictive Accuracy

Comparing predicted voters to actual registrations using the out-of-sample data can be used to assess model validity. Figure 7 displays predictions for each demographic group within one county and dashed lines denote average error. Surprisingly, Model I has the best predictive performance, with average error of roughly 10 individuals, but median error less than one.

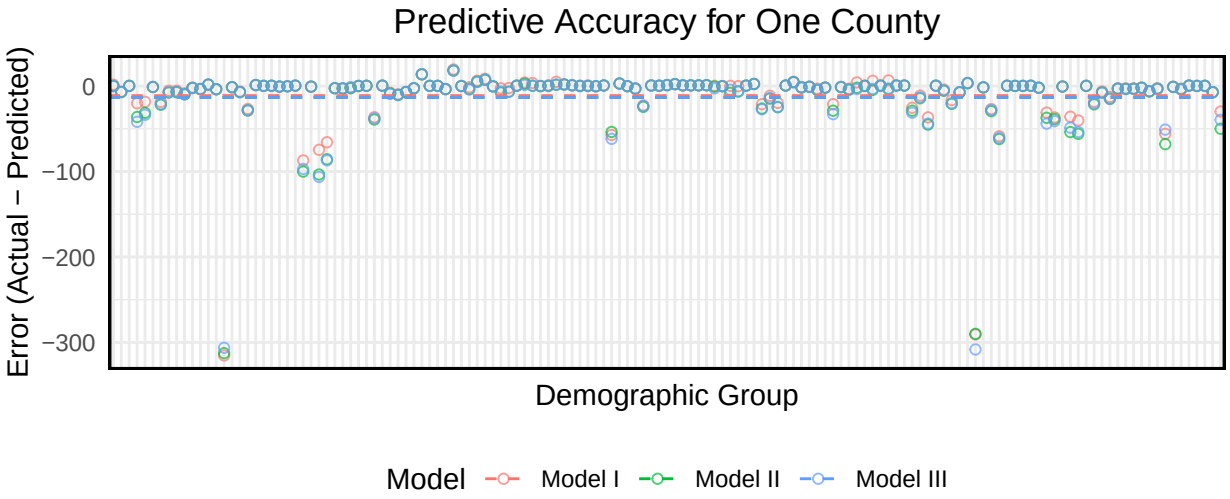


Figure 7: Predictive Accuracy across Demographic Groups

Table 3 displays the overall predictive performance for the full out-of-sample set. The average absolute error denotes the difference, in number of individuals, between model predictions and observed number of registered voters. The median error is substantial less than the average and indicates some predictions are significantly wrong.

Table 3: Prediction Summaries

	Avg. Absolute Error	Median Absolute Error
Model I	38.58	2.56
Model II	35.92	2.69
Model III	35.91	2.69

Table 4 shows the 95% credible region summaries for each model prediction. All models display substantially worse performance on the out-of-sample test set than the coverage level. Model I predictions have dramatically wider intervals than other models.

Table 4: Confidence Interval Summaries

	Confidence Interval Coverage	Avg. CI Width
Model I	0.8	150.88
Model II	0.6	16.4
Model III	0.61	16.46

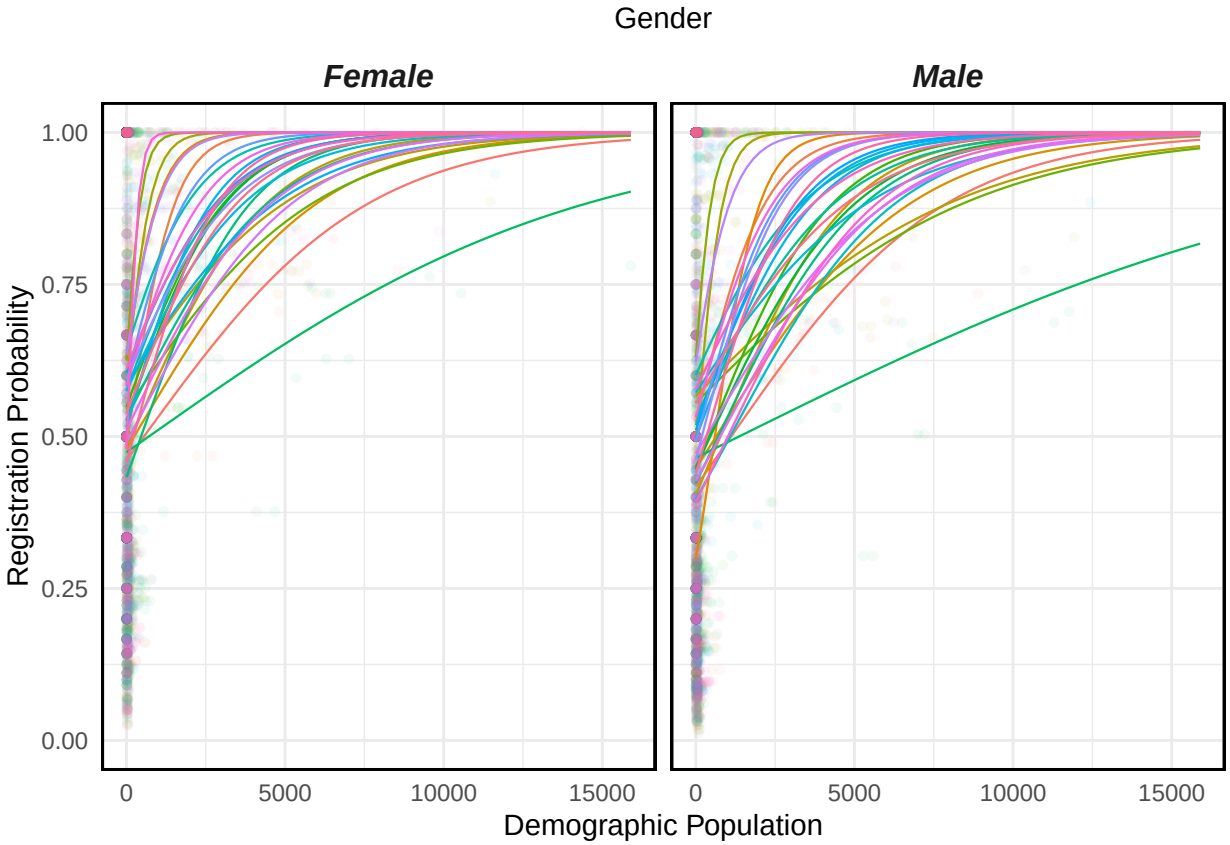
6 Limitations & Conclusion

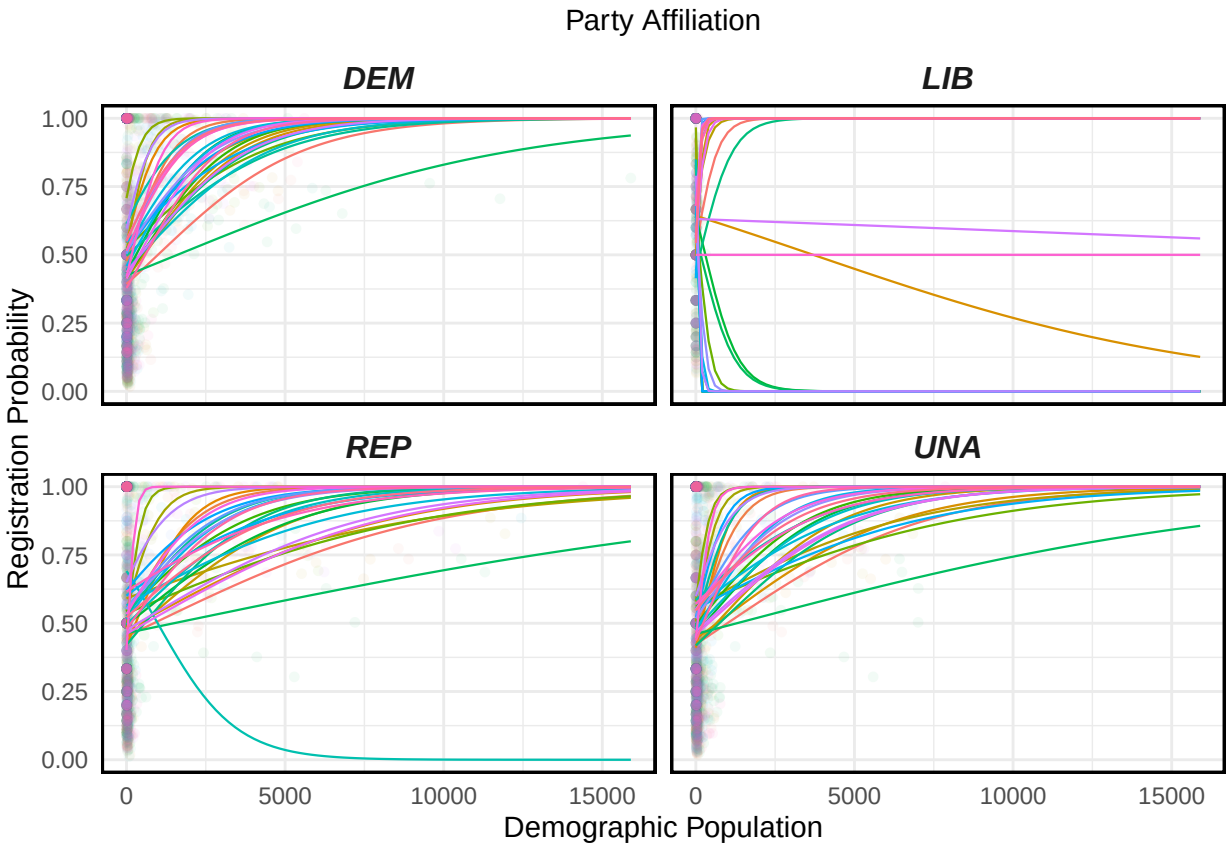
Historical voting records and population estimates were used to evaluate demographic differences in registration tendencies. Three Bayesian multilevel models were created to address specific questions of interest using different effect structures. The full dataset was sub-sampled to reduce computation time, but may lead to higher variance estimates. Preliminary results suggest significant differences in registration tendencies across demographic categories in North Carolina.

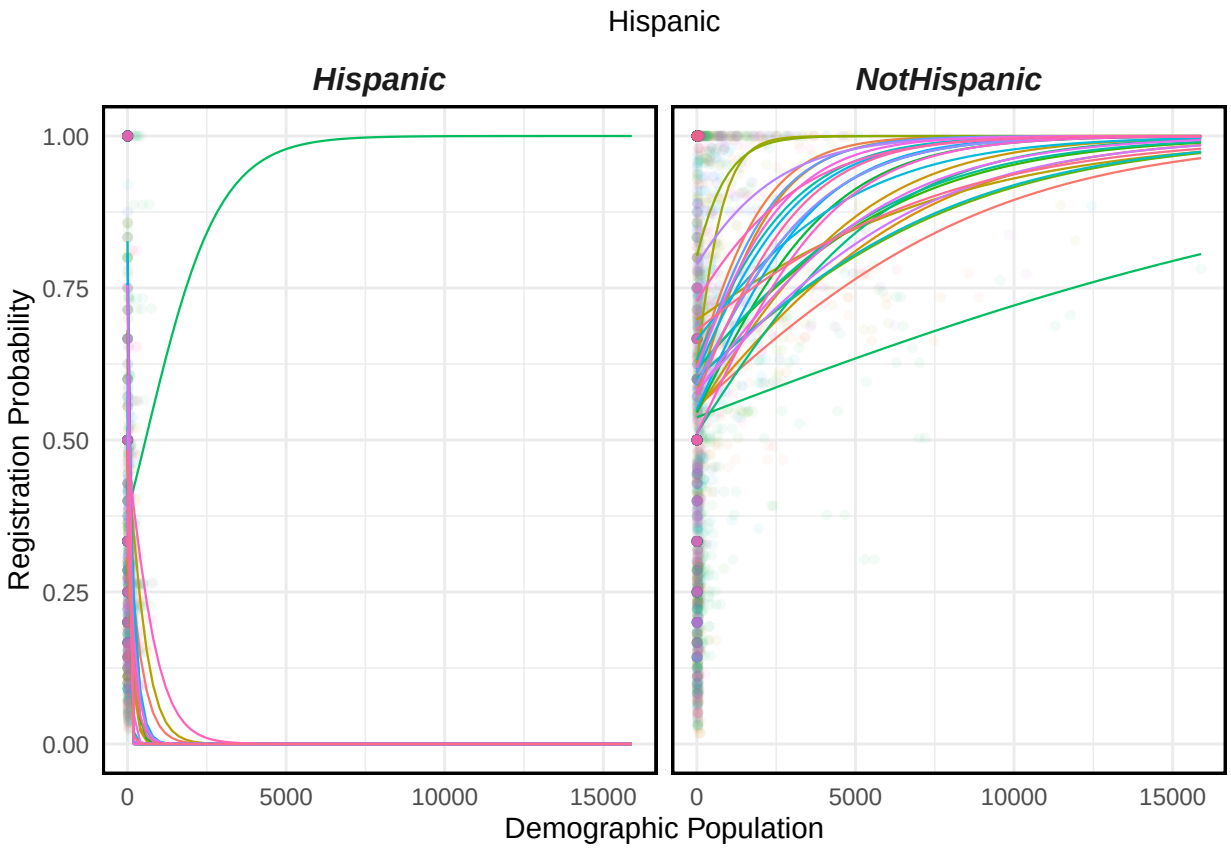
There are several limitations to this analysis. First, population estimates in 2016 are approximated from the 2010 Census. This may introduce unaccounted for variability if significant population changes occurred over the six year period. Multiple instances of observations with more registered voters than demographic population estimates, which should be impossible, may reflect this. To minimize potential variability a weighted sub-sample was drawn for inference. This addresses detectable population differences, but does not account for the possibility that all population estimates are different over the time period. In the future, one could/should look to include 2020 Census estimates, when they become available in 2023, to impute more realistic population measures that combine Census estimates from both years. The sub-sample also reduces the size of the dataset, adding variability, but required to reduce MCMC sampling time which can take over 12 hours on a laptop with the reduced dataset. One could leverage the entire dataset with a bootstrap resampling procedure and run multiple models in parallel to reduce variance concerns. An additional limitation is the latent political affiliation of unregistered voters, as the Census does not collect this information. These totals were imputed based on the party affiliations of registered voters as a rough approximation, but implies strong assumptions about how individuals register and may be better to adopt another modeling approach entirely. One alternative is to treat registration affiliation (democrats, republicans, libertarians, unaffiliated, not registered) as a multinomial draw from the total population and use registered individuals purely as response variables instead of imputing true affiliations. Lastly, overdisperion is almost always a problem in binomial regression models, as there is no independent variance term. If overdispersion is present, parameter estimates will be overconfident. Due to time constraints, overdispersion was neglected, but one could apply an uncertainty adjustment to estimates to account for this possibility.

7 Appendix

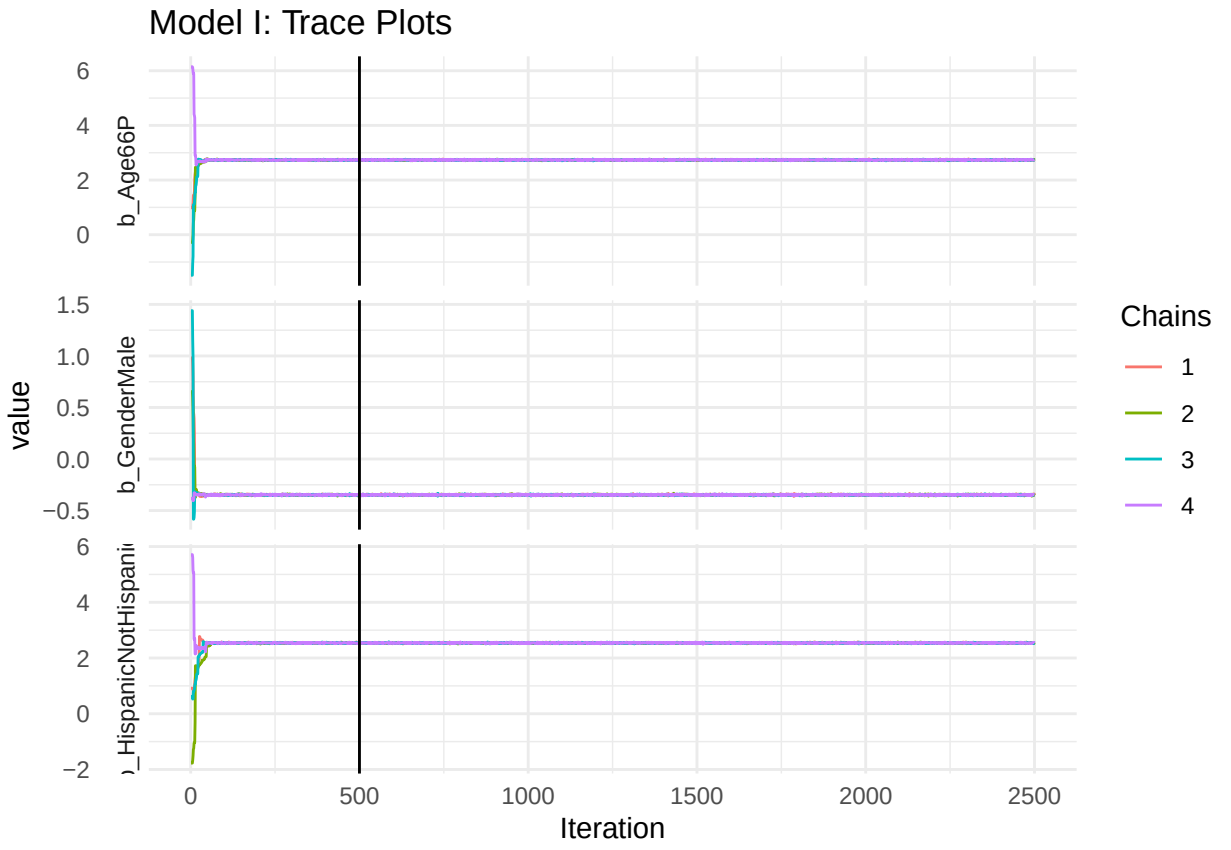
Additional Covariate Plots

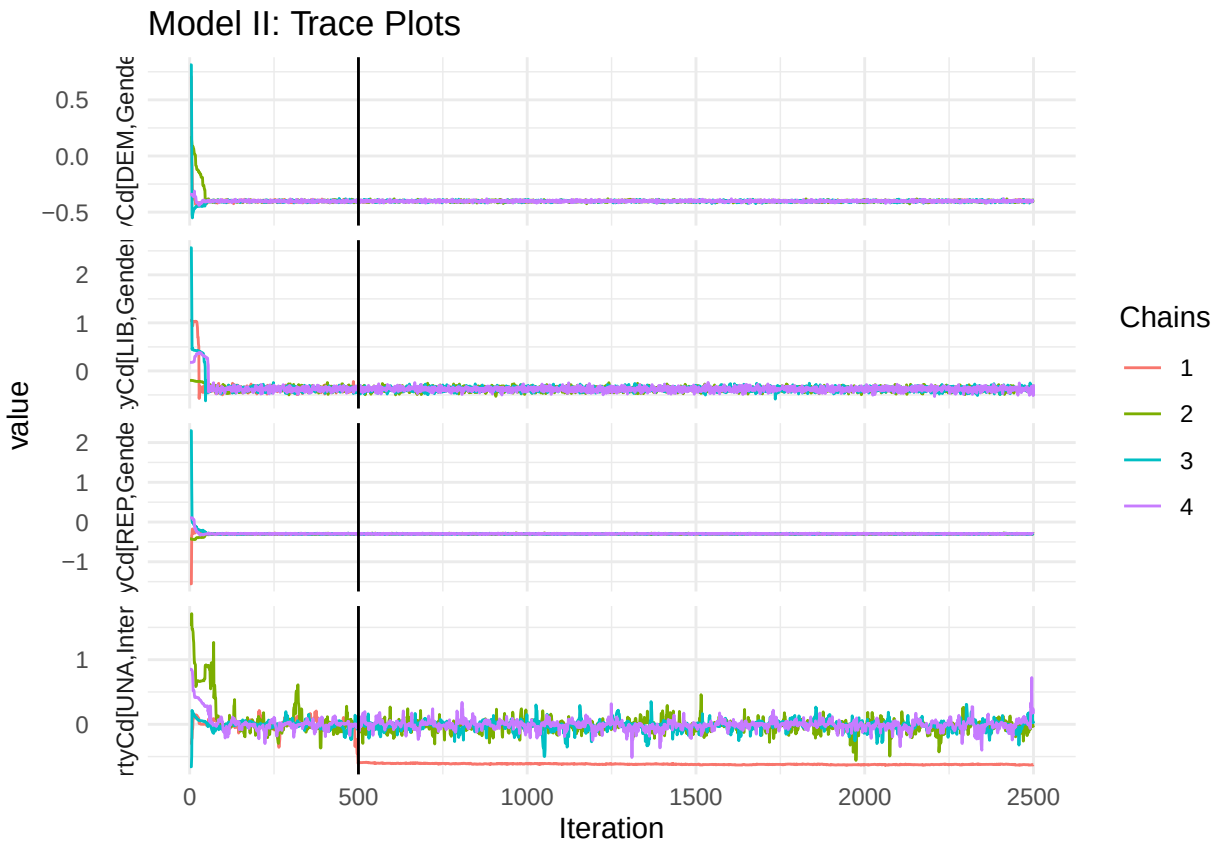


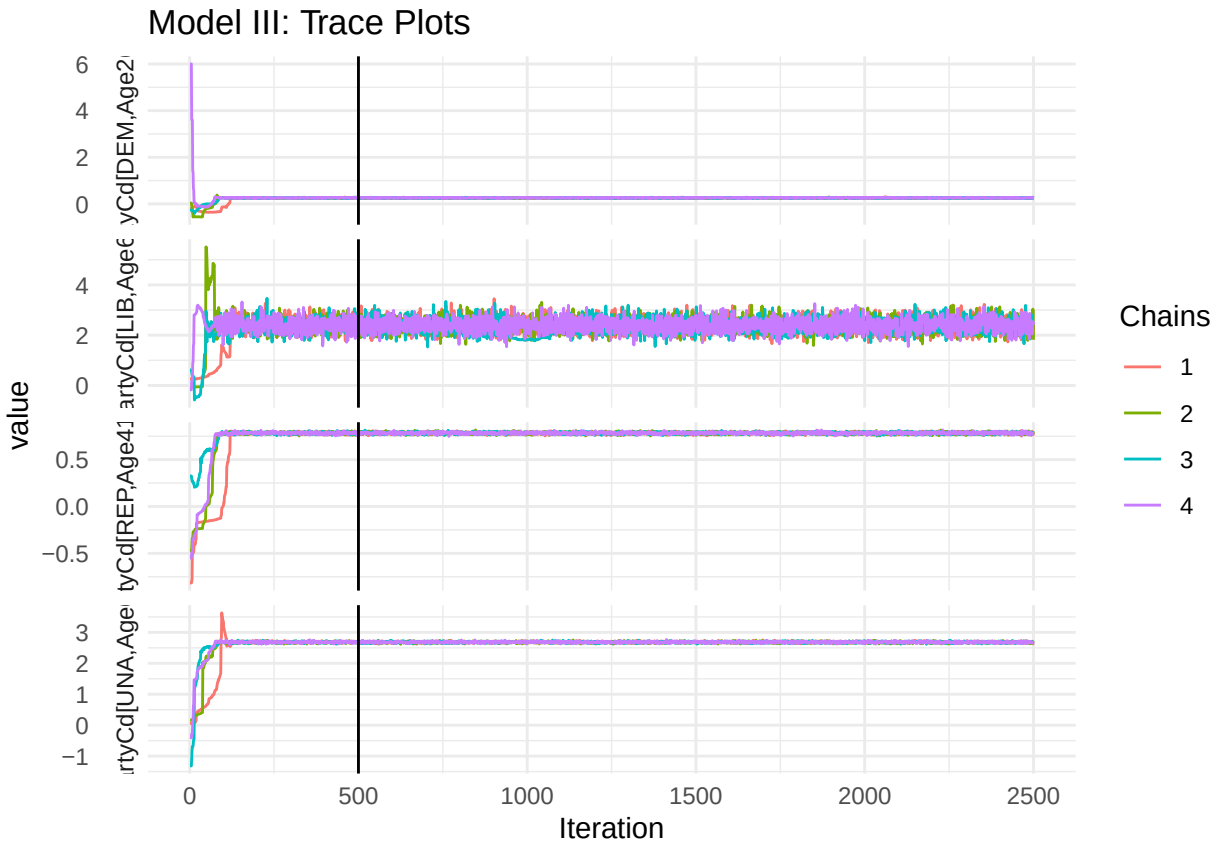



Hispanic
Hispanic
NotHispanic
Registration Probability
1.00
0.75
0.50
0.25
0.00
0
5000
10000
15000
Demographic Population

Trace Plots


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





County Sample List

Geography
ALAMANCE
ASHE
AVERY
BRUNSWICK
BURKE
CABARRUS
CHEROKEE
CLAY
CLEVELAND
COLUMBUS
CRAVEN
CUMBERLAND
DARE
EDGECOMBE
GATES
GRANVILLE
IREDELL
LEE
MADISON
NORTHAMPTON
PENDER
PERSON
POLK
ROCKINGHAM
ROWAN
TRANSYLVANIA
TYRRELL
UNION
WILSON
YADKIN