

# Analyzing Voter Registration in the 2016 Presidential Election

Andrew Amore  
2022-11-21

## 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 show up at the polls. 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

### Voter Fraud?

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 historical Census information from six years ago too dated to reflect accurate estimates?

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, but could also be inflated with a correction factor learned from other observations.

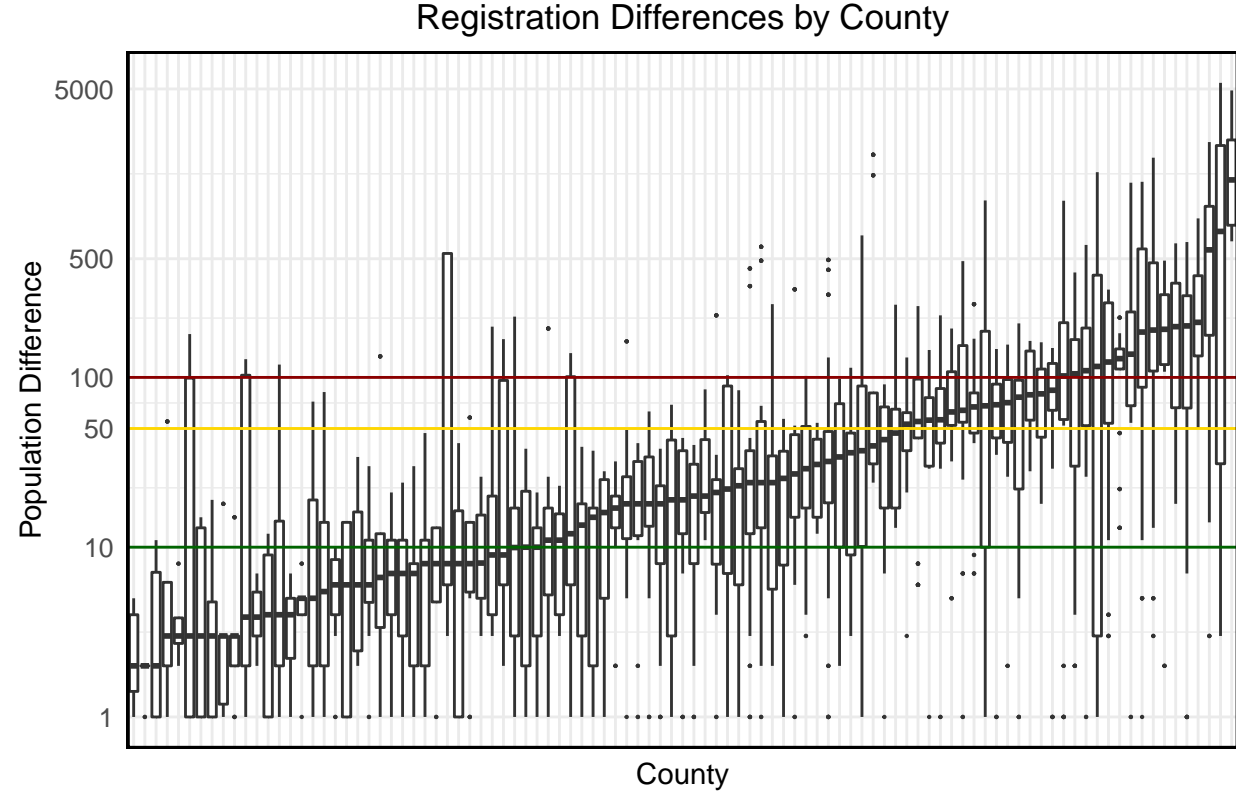


Figure 1: Registration Differences by County

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 extended to bootstrap resample different county combinations for parameter estimation which reduces variation by steering our model towards more accurate data. An additional benefit of sub-sampling is a reduction in computation time for MCMC sampling for parameter estimation in Bayesian models.

### Motivating a Multilevel Model

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. However, we do not know the political party affiliation for unregistered voters, as the Census does not ask about political affiliation and this membership is considered latent. This information is necessary to draw conclusions on party membership registration influence. As an initial estimate for this missing quantity, I impute it based on the original population Census estimates and the political party distribution for registered voters. This new quantity is used in place of the original estimates for inference.

To evaluate modeling decisions, like where to apply random effects, an exploratory data analysis (EDA) was conducted. Figure 2 is a facet plot of different binomial regressions on voter registration probability for two demographic categories and color coded by county. (A) fits regressions by county, age group and has substantially different trend lines across and within facets. (B) fits regressions by county, race and has much lower variability across facets. Plots for additional covariates can be found in the Appendix. Gender effects are consistent across geography, party affiliation displays consistent trends as does hispanic indicator. Continuous fields, like total county population, were also explored in the regression analysis, but were not included in the final model as including it lead to difficulty generating valid sampling distributions in brms (even with scaling and prior adjustments).

## Voter Registration Probability by Demographic Segmentation

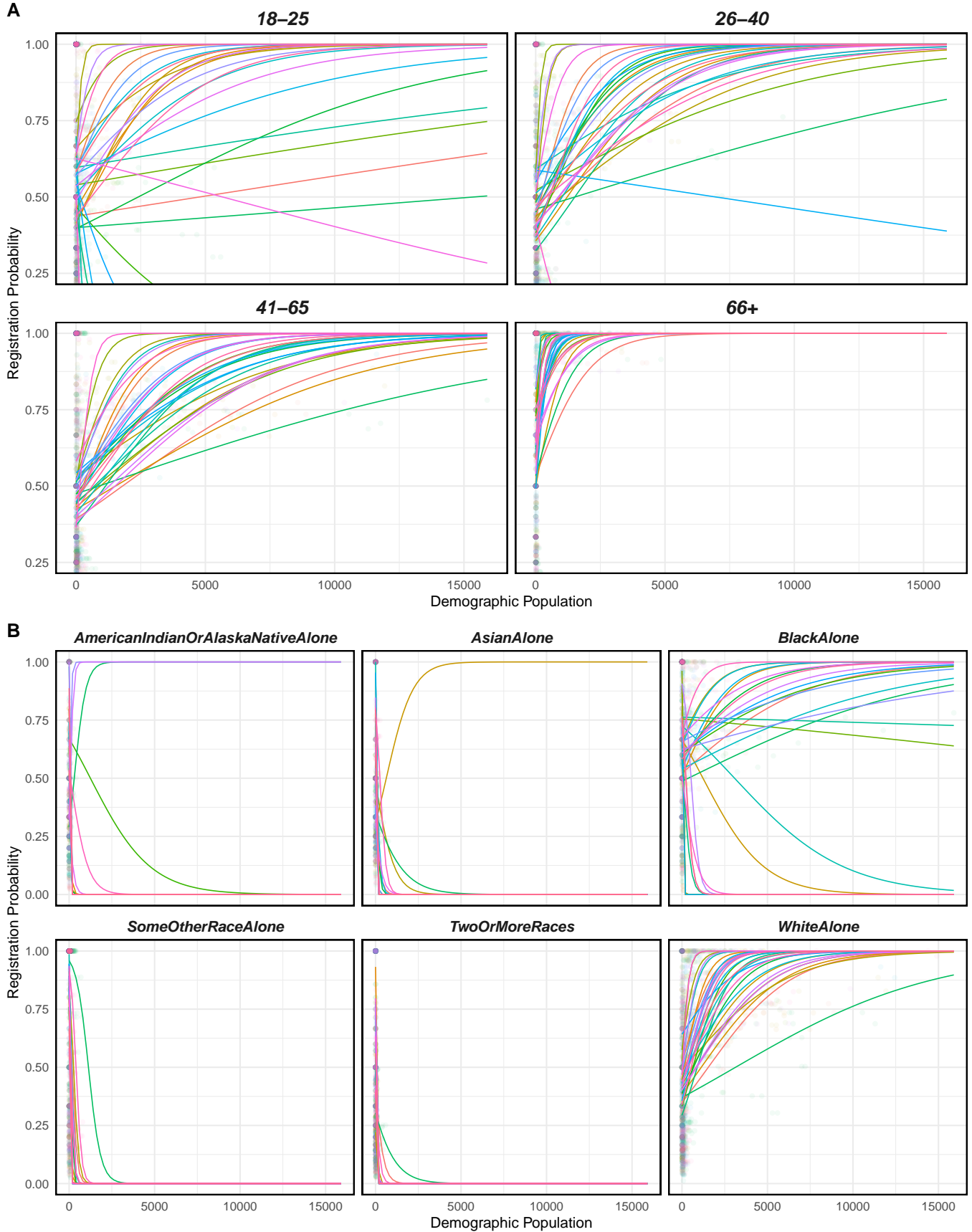


Figure 2: Voter Registration Behavior for (A) Age Category and (B) Race

## Model Specification

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.

$$y_i \sim \text{Binomial}(n_i, p_i)$$

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  $n_i \geq 30$ . However, the demographic samples can be considered population estimates and may warrant a hypergeometric distribution, but this induces a dependence between  $y_i$ 's which is unwarranted based on our independent registration assumption. Under the binomial model, several multilevel structures with varying complexity were evaluated. Table 3 displays the specifications.

Table 2: Random Effect Structures

	Fixed Effect Fields	Random Intercept Fields	Random Slope Fields
<i>Model I</i>	Gender, Hispanic, Race, PartyCd, Age	Geography	
<i>Model II</i>	Hispanic, Race, Age	PartyCD	Gender
<i>Model III</i>	Gender, Hispanic, Race	PartyCD	Age

### Model I

Let  $\mu$  denote a global intercept corresponding to a baseline demographic category with the following values: 'Female', 'Hispanic', 'AmericanIndianOrAlaskaNativeAlone', 'DEM' and '18-25'. Let  $\theta_j$  denote a **county** random effect intercept for observations,  $i$ , in county  $j$ ,  $X_i$  a vector of demographic indicator variables corresponding to column 1 in Table 2 and  $\beta$  the corresponding fixed effect estimates.

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

*Prior Specifications*

$$\mu \sim N(0, 1), \beta \sim N(0, 1), \theta_j \sim N(0, \sigma)$$

$$\sigma \sim \text{HalfCauchy}\left(0, \frac{1}{2}\right)$$

### Model II

In accordance with the model specification in Table 2, let  $\mu$  denote a global intercept corresponding to a baseline demographic category with the following values: 'Hispanic', 'AmericanIndianOrAlaskaNativeAlone' and '18-25'. Let  $\Omega_p$  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  $\Gamma_{pk}$  the random effect. Finally, let  $X_i$  be a vector of demographic indicator variables corresponding to column 1 in Table 2 and  $\beta$  the corresponding fixed effect estimates.

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

*Prior Specifications*

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

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

### Model III

In accordance with the model specification in Table 2, let  $\mu$  denote a global intercept corresponding to a baseline demographic category with the following values: ‘Female’, ‘Hispanic’ and ‘AmericanIndianOrAlaskaNativeAlone’. Let  $\Omega_p$  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  $\alpha_{pk}$  the random effect slope. Finally, let  $X_i$  be a vector of demographic indicator variables corresponding to column 1 in Table 2 and  $\beta$  the corresponding fixed effect estimates.

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

#### Prior Specifications

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

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

### 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. We are particularly interested in the following italicized questions.

#### 1. *How did different demographic subgroups register to vote in the 2016 election?*

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

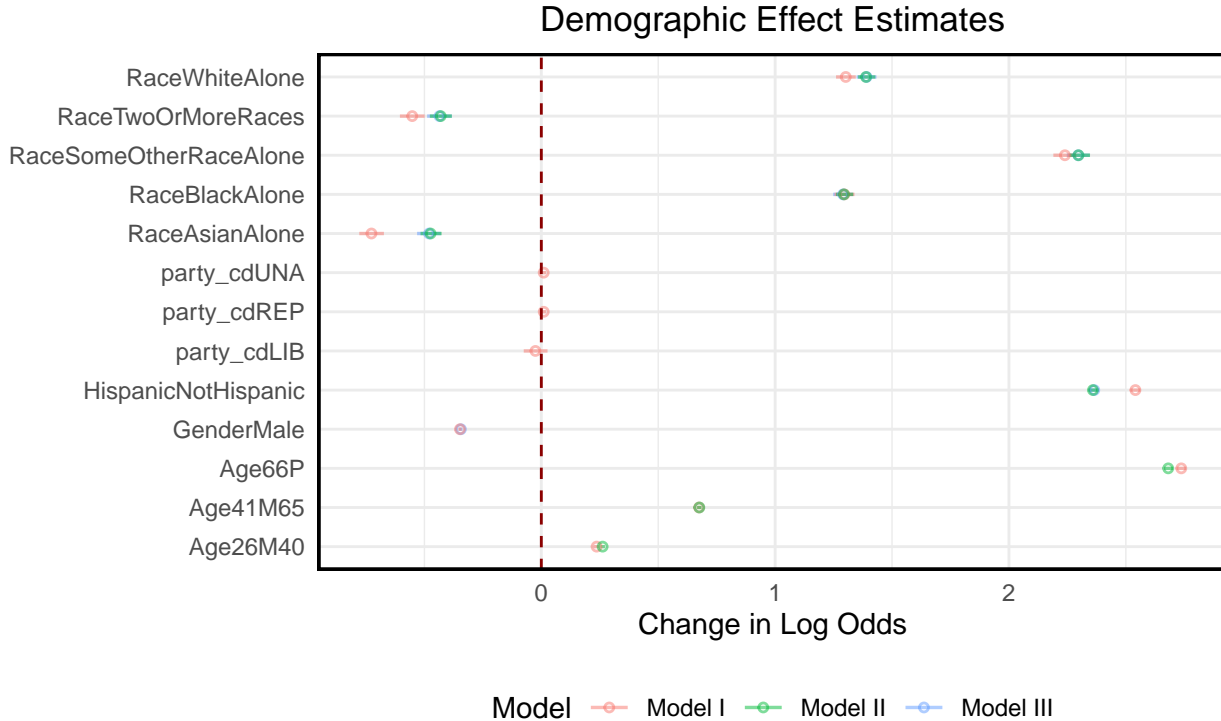


Figure 3: Fixed Effect Estimates From Each Model

**2. Did the overall odds of registering differ by county in 2016? Which counties differ the most from other counties?**

Only 30 counties are included in this analysis and the list can be viewed in the Appendix. The random intercepts from Model I, with 95% CR are displayed in Figure 4. Tyrrell, the only county in the dataset without any “invalid” observations has the lowest registration effect, while **43%** of counties have no detectable effect.

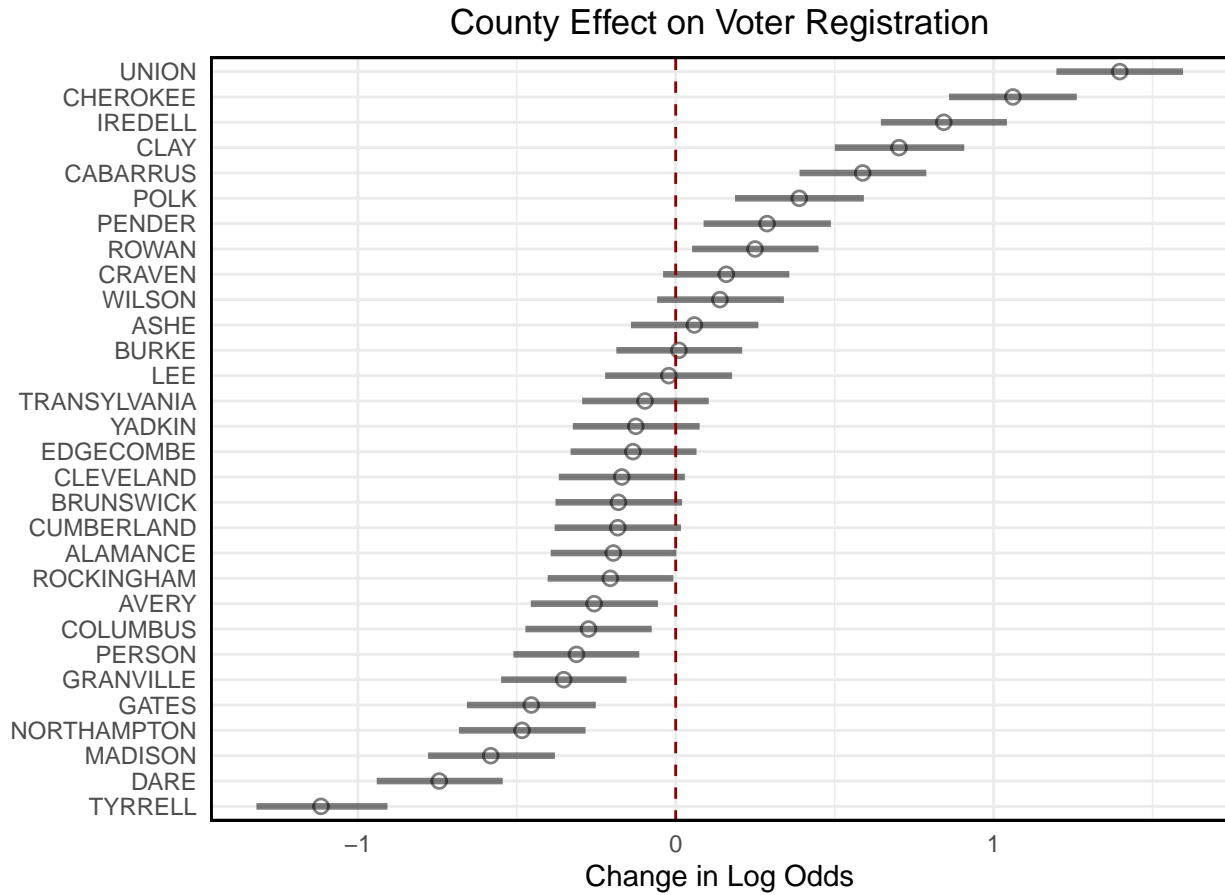


Figure 4: County Level Intercept Estimates from Model II

**3. How did the registration rates differ between genders for different party affiliations?**

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 gender (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.

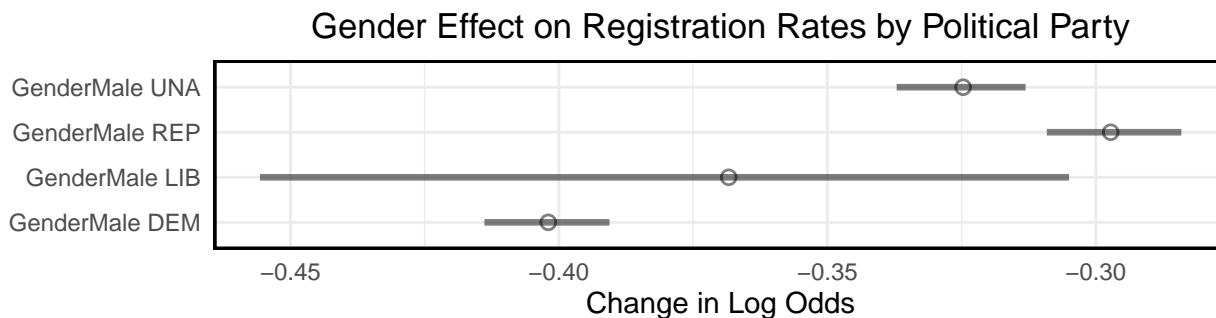


Figure 5: Random Slope Estimates (Model II) for Gender across Political Party

#### 4. How did registration rates differ between age groups for the different party affiliations?

Using Model III we can look at the age effect within each political party from random slope estimates assess the relative change in log odds from the baseline age (18-25). Figure 6 shows 95% CRs and point estimates. In almost all age categories republicans are more likely to register than other political parties. We also see registration rate across all parties as age increases.

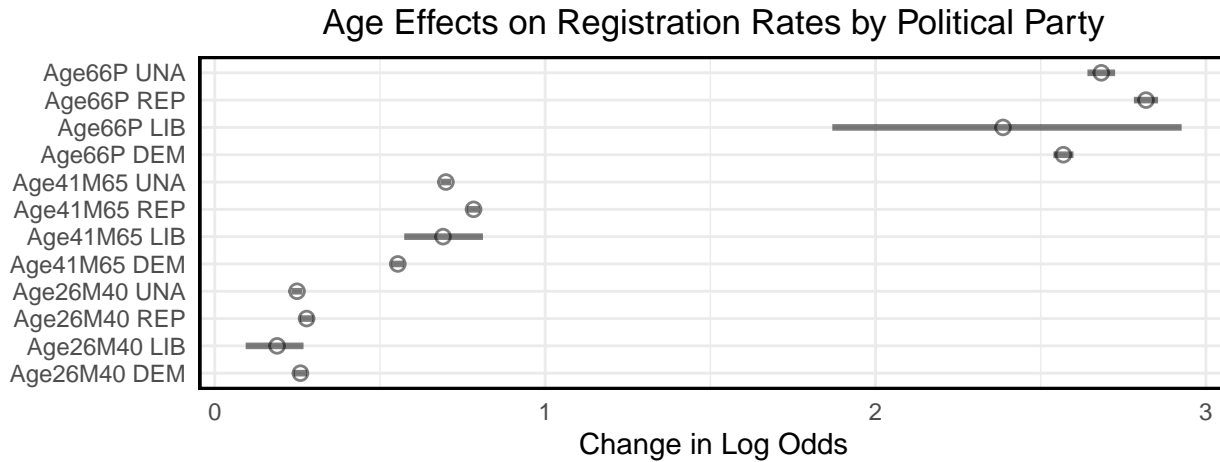


Figure 6: Random Slope Estimates (Model III) for Age across Political Party

#### Predictive Accuracy

In addition to inference, we are interested in predicting registered voters for each demographic combination. 70% of our data is unseen by the model and constitutes the test set. For ease of viewing, I've randomly selected one unseen county and display the predictive accuracy for every demographic combination in Figure 7. The error is computed as the difference between the actual number of registered voters and the model prediction. Dashed lines on Figure 7 denote average error. Surprisingly, Model I has the best predictive performance, with average error roughly 10 individuals, but median error of less than one.

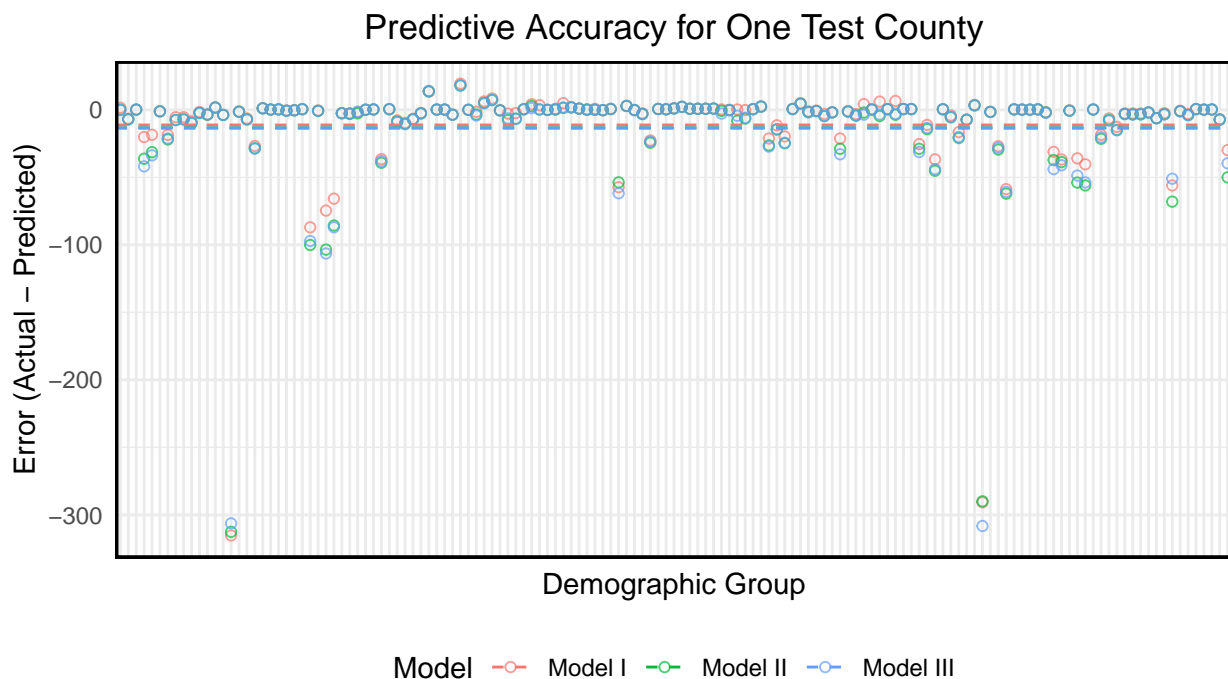


Figure 7: Predictive Accuracy for One Unseen County Across All Demographic Groups

## 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 which limits our data usage to only 30% of available counties and introduces additional variability. To alleviate this, one could leverage the entire dataset by either acquiring more computation power, letting models run for longer (most complex model takes  $\sim 12$  hours to run), adopt a bootstrap resampling procedure to run the model multiple times with different county combinations (also requires more computation time), and/or specify simpler models with less random effects that take less time to run.

An additional limitation is the unknown population estimates for political party affiliations. The Census data only quantifies innate categories and does not contain information regarding party affiliation. These totals were imputed from available fields based on the party distribution from registered voters within each demographic group. This introduces another source of error, but is necessary if we want to make inferences about party affiliation.

Lastly, the Gelman book indicates overdispersion is almost always a problem in binomial regression models as there is no independent variance term. If overdispersion is present, our estimates will be overconfident. Time constraints prevented testing for overdispersion and applying an uncertainty adjustment to estimates.

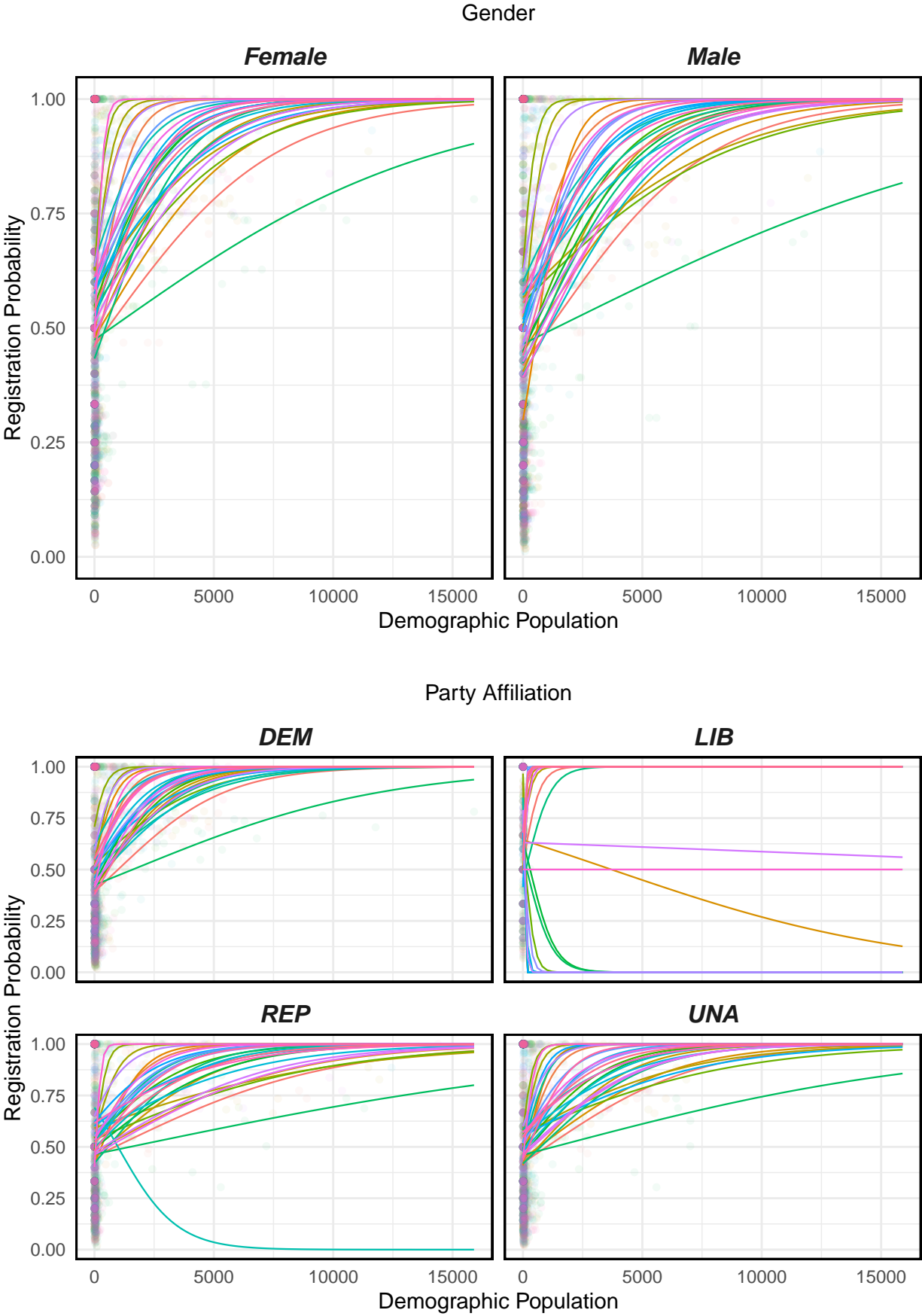
## Conclusion

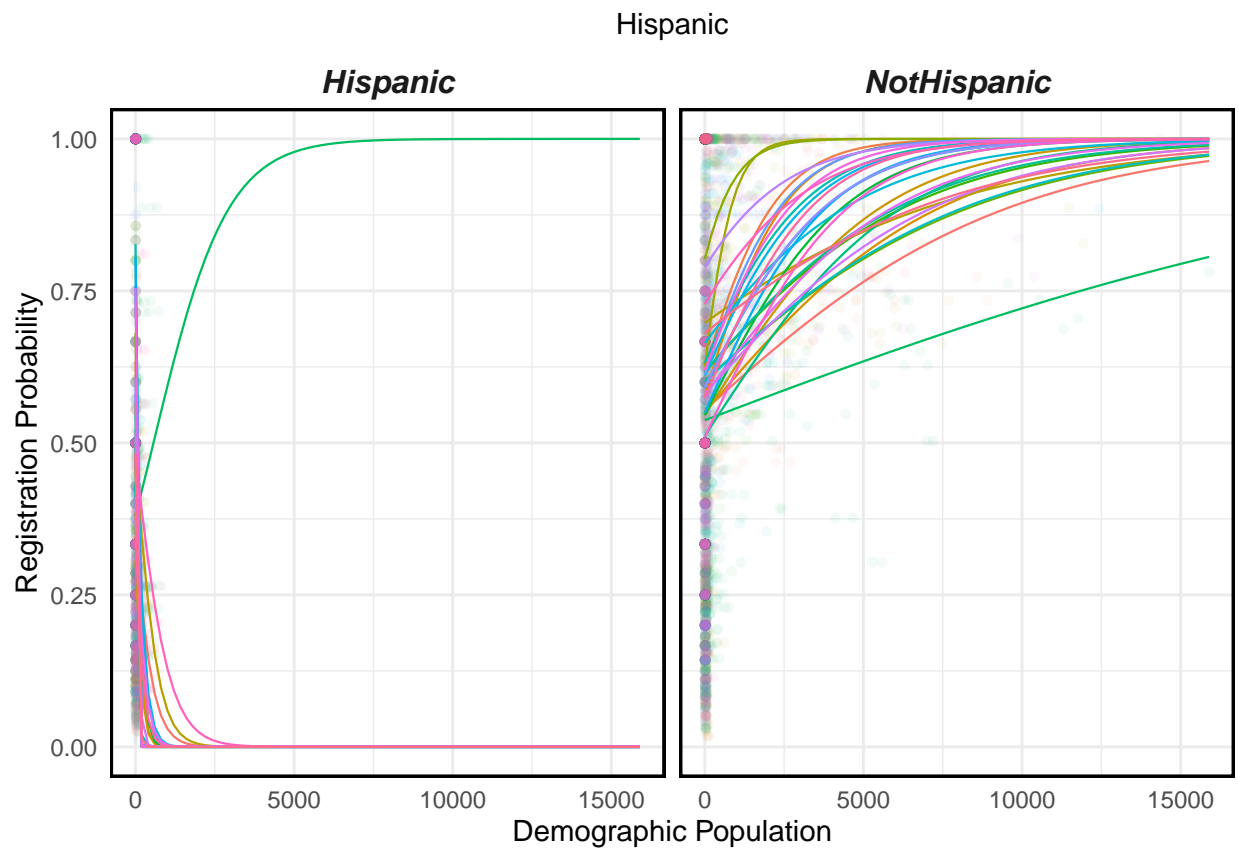
Historical voting records and Census population estimates were used to evaluate demographic differences in registration tendencies. Three Bayesian multilevel models were created to address specific questions of interest using different effect structures. The full dataset was sub-sample to improve computation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



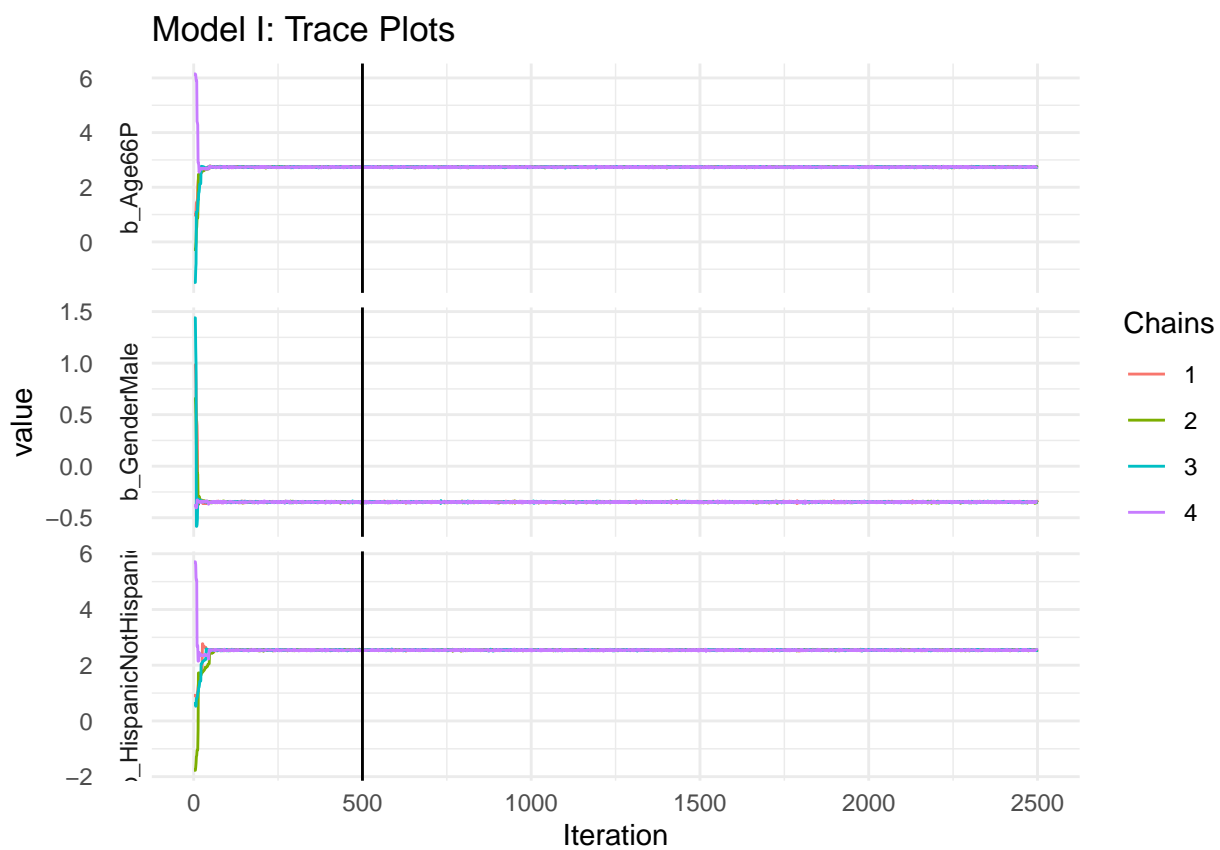
Appendix

Additional Covariate Plots

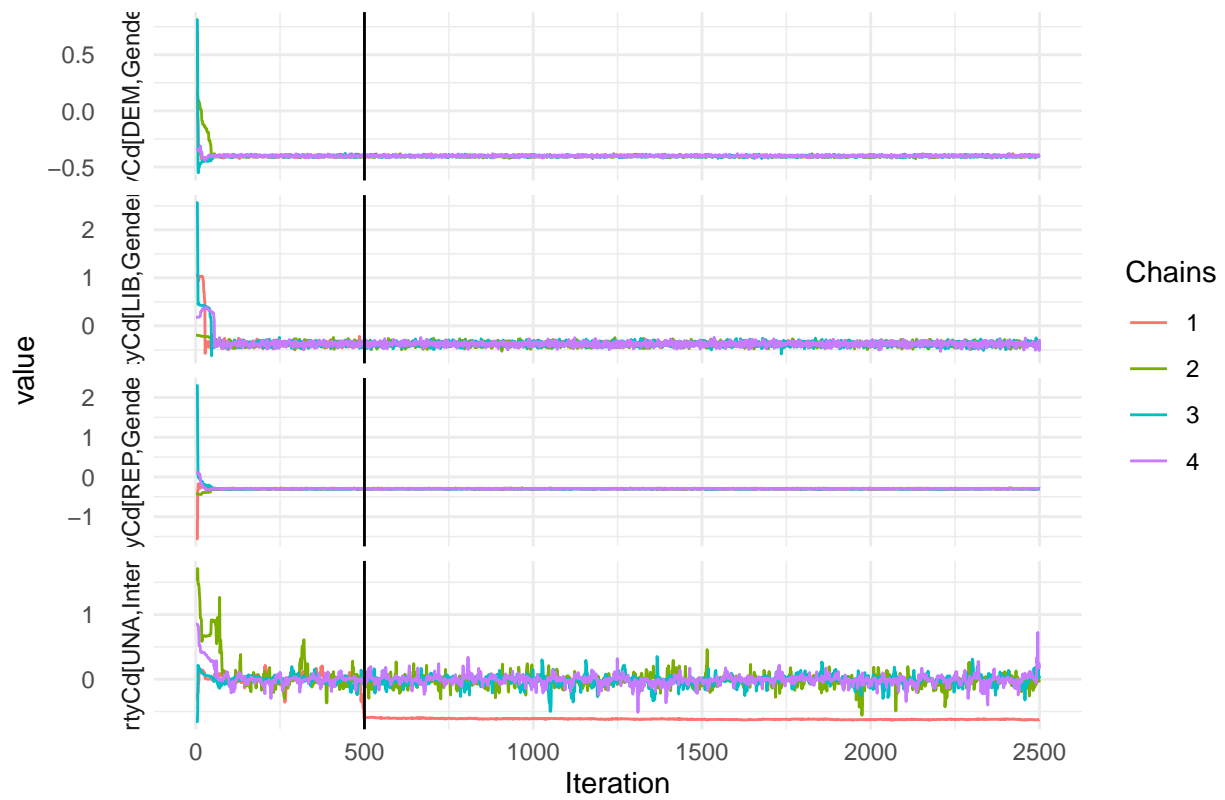




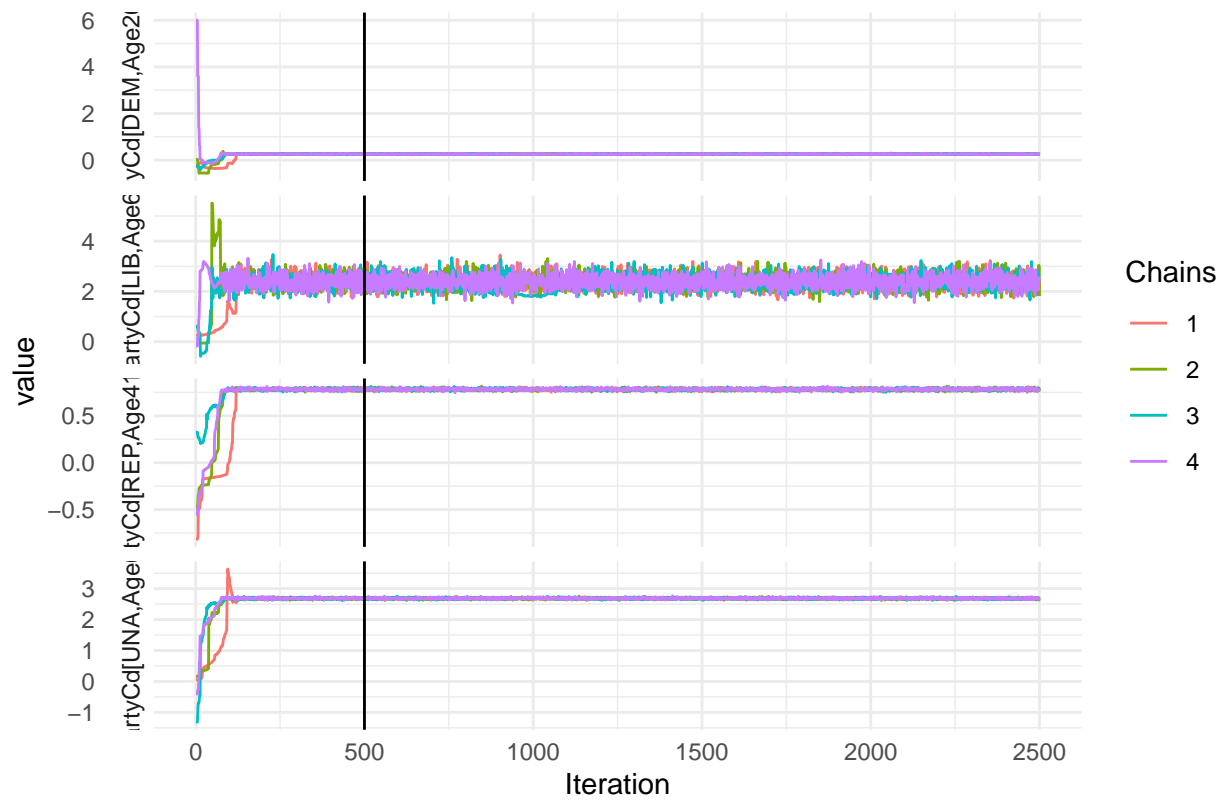
## Trace Plots



Model II: Trace Plots



Model III: Trace Plots



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

---