# STA 610 Case Study 2

Richard Fremgen

20 November, 2022

---

*Note: attached to my submittal is cs2_model.Rmd file that contains code for all of the different models fit, along with case_study_2_fremgen.Rmd which contains the code and text used to compile this pdf report.*

## 1. Introduction

One of the most prominent ways that individuals in the United States (U.S.) can have their voices heard are through elections, where people vote on local, state, and federal leadership. A primary requirement for one to be able to vote is that a person must register to vote, which can be done through a variety of ways, such as via the DMV, online, or through mail-in applications. Each state in the U.S. has some form of board of elections agency that oversees and administers this election process; elections in North Carolina for example are run by the NCSBE. Besides tracking election results, NCSBE also keeps record of every individual that is registered to vote, in addition to demographic (age, race, ethnicity, gender) data. While NCSBE provides excellent data about *who* registers to vote, if one wants to understand demographics on a population level, the U.S. census is an ideal place to look. In the U.S., a population census is taken every ten years to collect and record information about members of the U.S. population; such data is then used to better understand how funds and assistance are distributed, in addition to understanding the demographic groups that make up the United States. With that in mind, the objective of this case study is to investigate how different demographic groups registered to vote in North Carolina. In particular we are interested in using a Bayesian hierarchical model to address differences in voter registration for the 2016 election between sex, age groups, race, ethnicity, county, and political party affiliation in the counties across North Carolina.

## 2. Data Handling

In order to answer these questions concerning the relationship between demographic groups and voter registration, two data set were used: `voter_stats_20161108.txt` (referred to as the *voter* data set) and `Census2010_long.txt` (referred to as the *census* data set). The *voter* data set contains information from NCSBE about the aggregate count of registered votes by the demographic groups for each county in North Carolina (NC) for the 2016 election. The `total_voters` variable in this data set contains the total number of registered voters within a particular group, as the data are aggregated both by location (county, precinct, voting districts), in addition to by demographic groups such as: age, race, ethnicity, sex, and political party. The *census* data set contains the Census Bureau's demographic report of the state of North Carolina from the 2010 U.S. Census. The `Freq` variable in this data set represents how many people were recorded for a given location (NC county) and demographic group (age, gender, hispanic, race) combination.

Due to the complexity of the data, extensive data cleaning and data wrangling was conducted in order to make the data conducive for joining and analysis. One of the first steps taken was to reduce the size of the overall data set, 6.9 million registered voters, into a subset that was more manageable and convenient to run analysis on. As such, the decision was made to randomly select thirty counties in NC and then use this subset for the entirety of the report. Simple random sampling was applied via R's built-in `sample()` function to randomly select the thirty NC counties in `Table 1`. Both data sets were filtered to only these counties of interest before further data pre-processing occurred.

Table 1: 30 Counties Randomly Selected

| | | | | | |
|---|---|---|---|---|---|
| ALAMANCE | AVERY | BERTIE | BRUNSWICK | BURKE | CABARRUS |
| CHOWAN | CLAY | CURRITUCK | DARE | DAVIE | EDGECOMBE |
| GRANVILLE | GUILFORD | HARNETT | HENDERSON | HOKE | IREDELL |
| JACKSON | MCDOWELL | MITCHELL | NASH | POLK | RANDOLPH |
| ROCKINGHAM | SCOTLAND | SURRY | TRANSYLVANIA | WAYNE | WILSON |

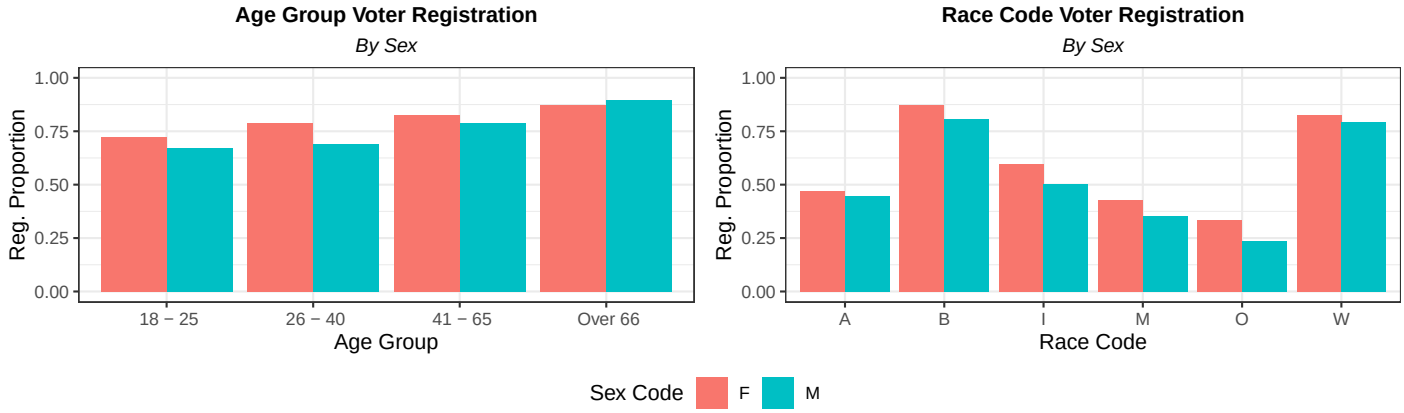
Next, steps were taken to further aggregate the *voter* data set to be on the same level as the *census* data. Unlike the *census* data that was recorded on the demographic-county level, the *voter* data set contains demographic registration values for each precinct, which comprise voting districts, which in turn roll-up to the one hundred counties in NC. As such, in order to make the data sets be on comparable geographic levels, the *voter* data set was aggregated to be on the demographic-county level, meaning that the precinct and voting districts were combined to be on the county level. Next, demographic groupings in the *census* data were modified in order to match the similar pairing of such group in the *voter* data set. For example, the *census* `age` variable grouping level `18-25` was changed to `Age 18 - 25` in order to match such level in the *voter* data set. Such a process was conducted on the age, gender, ethnicity, and race variables in the *census* data set as well, however, there were demographic group levels that existed in one data set, but did not exist in the other. For example the `sex_code` variable in the *voter* data set contains three unique genders: `F`, `M`, and `U`, where as the *census* column `gender` only contained two unique genders: `Female` and `Male`. In such a context, handling the gender of `U` presented challenges since there was not a `U` or unknown gender in the *census* data set. A similar situation also presented itself with the `race_code` variable where `race_code = U` only existed in one data set. As such, the decision was made to filter out rows where `sex_code = U` or `race_code = U`, since these instances were very infrequent ($< 2\%$ of the data).

The one problematic variable in the *voter* data set was `ethnic_code` which took on the values of `UN`, `NL`, and `HL` and encoded the Hispanic origin of a group. While this variable mapped to the `Hispanic` variable the *census* data set, the *census* data set only took on values of `Hispanic` and `NotHispanic`, meaning that `ethnic_code = UN` would cause problems when joining. Such occurrences could not simply be removed from the data, since `ethnic_code = UN` represented 16% of the number of registered voters. In order to cope with this issue, `ethnic_code = UN` was treated as missing completely at random (MCAR) data; where we assumed that the MCAR data would have the same distribution of the non-missing data . In order to impute these values in the *voter* data set, first the proportion of `NL` and `HL` groups was found for each county-party-race-sex-age combination of voters. Next, these proportions were then used to properly divide `ethnic_code = UN` data into `NL` and `HL` groups for the given demographic. The decision was made to calculate `NL:HL` proportions on such a granular demographic subgroup level, because *a priori* is it reasonable to think that different counties and demographic groups in NC have varying proportions of Hispanic to Non-Hispanic voters.

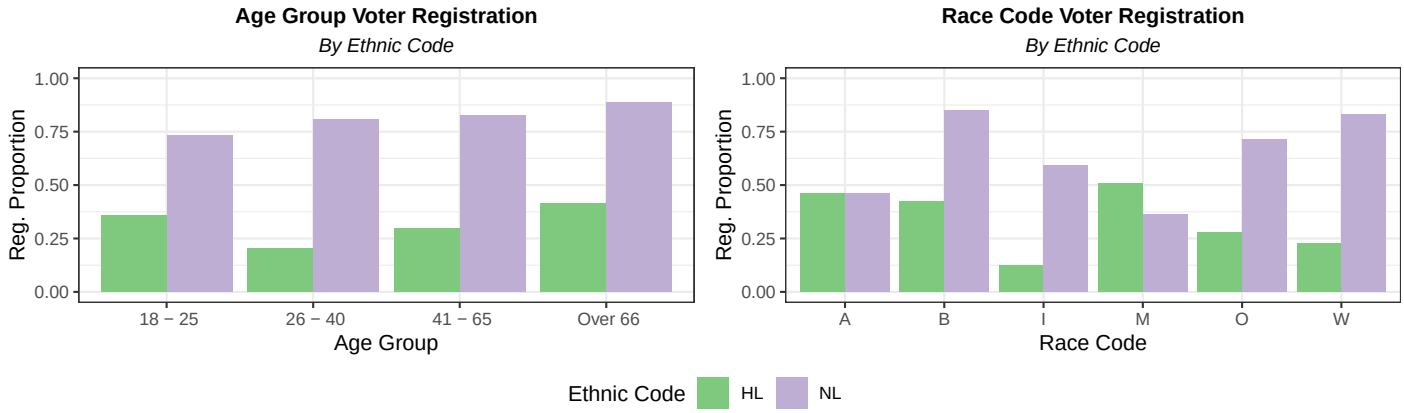
The final variable that was modified in the *voter* data set was `party_cd` which corresponds to the political party that a given individual registered to. Since `party_cd = LIB` made up only 0.4% of the registered voters, `LIB` party occurrences were grouped with the `UNA` party in order to represent the *Unaffiliated* party grouping. Next, these filtered data sets were then joined together based on the county, age, race, sex, and ethnicity variables. The resulting data set contained both the number of people that were registered to vote and the number of people accounted for during the census for each demographic group, as such values could be seen as binomally distributed and provided motivation for the type of models used. The one issue with this binomial assumption was that when the data was joined, there were certain demographic subgroups that had a greater number of registered voters than people recorded in the census. Such a finding is not entirely surprising, given that the voter registration data was taken six years after the census data, and generally speaking populations in the U.S. tend to grow each year. In order to cope with this issue, a naive approach was taken to uniformly increase each census value by 5.7%; this number was used since North Carolina's population grew by 9.5% between the 2010 and 2020 census and corresponds to a scaled value for the six year gap, if we naively assume that the population grew at a constant rate each year.

## 3. Exploratory Data Analysis (EDA)

One of the most important steps to take after the data has been processed, is to conduct EDA in order to investigate which variables could be used as potential covariates in the models fit. *A priori* it is feasible to think that the number of registered voters will differ between demographic sub-groups, but since each demographic makes up a different amount of the population, each group's voter registration totals were converted to a proportion out of how many total people there were recorded in the census for that group. To start, we consider how voter `sex_code` (male vs. female) registration rates varied by both `age`and `race_code`. Looking at the bar plots below, we can see that overall, registration rates tend to generally increase as age group increase for both `sex_code`, `M` and `F`, as females had slightly higher registration rates for three out of the four age groups. If one were to look just at `sex_code` registration differences (Appendix), it would be evident that females had a registration rate of 81% compared to 76% for males. Additionally, when just considering registration rates on just an `age` level, pooling both sexes together (Appendix), `age = Over 66`, had the highest registration rate of 88%, where as the youngest age group, `age = 18 - 25`, had the lowest at 70%. Aside from age group, we can also consider how `sex_code` registration rates vary by `race_code`, where we can observe much more variability between races. Looking at the output below, `race_code = B` (Black) and `race_code = W` (White) had much higher registration rates $> 75\%$ for both sexes when compared to the other four races.

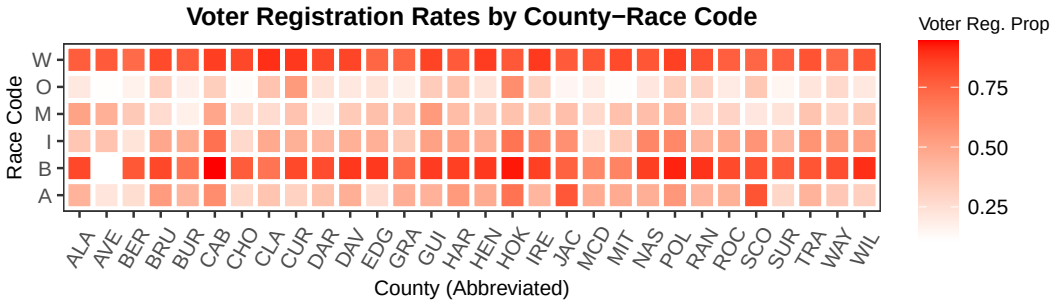



Another variable of interest is `ethnic_code` which can take on two possible values: `HL` (Hispanic) and `NL` (Non-Hispanic); as such we can reproduce the same plots above but segmenting by `ethnic code`. Clearly, Non-Hispanic groups had much higher voting registration rates compared to Hispanic individuals across age groups, as the differences between the two are very similar for every one of the four age groups. When segmenting `ethnic_code` by `race_code`, the difference between `NL` and `HL` ethnic groups becomes more variable, as for most races `NL` had higher registration rates than `HL`, however there were race codes such as `A` (Asian) and `M` (Multiple) where `HL` had higher or the same rates. It should be noted that while the differences between ethnicity looks to be extensive, the `HL` individuals only make up 1.5% of the data, even after imputing the MCAR data.




When we consider the different counties selected, *a priori* is it reasonable to think that out of all the demographic

groups selected, `race_code` would be the one to vary the most by geographic location (county), when compared to other variables such as `age` or `sex_code`. With that being said, the tiles in the heat map below correspond to the voter registration proportion for every `county` and `race_code` combination, as abbreviated county names (first three letters) were used to make the graphic more compact. We can see in the output below, that across all thirty counties, `W` and `B` race codes had the highest voter registration rate. However, we can discern that there do appear to be differences between counties, aside from `race_code = W`, as each race code does appear to differ to some degree for each for the thirty counties. For example `race_code = O` (Other races) has some counties, where the proportion is very low (white color tile), but many counties were the voter registration population appears to be moderate to high. A similar argument can be made for `race_code = A` (Asian), where the population in Scotland county (`SCO`) is $> 80\%$ compared to Avery county (`AVE`) where registration rate appears to be around 25%. Included in the Appendix is also a table calculating registrations for each county as a whole (ignoring any additional demographic effects), as each of the thirty counties range from an overall registration rate between 70% and 88%.




## 4. Model I

### 4.1 Model Motivation

In order to answer the questions of interest in this assignment, I have decided to fit two models. This section (`Model I`) will address the fist two questions of interest about differences in voter registration groups among demographic groups and between counties, and the next section (`Model II`) will address questions regarding political party affiliations and sex/age differences. Since the data we are working with is binomally distributed data, we can model such as $Y_g \sim Bin(n_g,\ p_g)$, where $Y_g$ is the number of registered voters for a given demographic subgroup $g$. Additionally, $n_g$ represents the number of people in the population for that given demographic group (based on the census data), and $p_g$ is the probability of someone in that demographic group registered to vote. We can further model $p_g$ by using the *generic* binomial regression equation as such: $\text{logit}(p_g) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. The generic form of this equation will be used in subsequent sections to fit a series of multilevel model, using a Bayesian approach; the final model in `Section 4.3` will contain the specific random effects and prior distributions that utilized in the model. A multilevel model is preferred in such a scenario, since they allow for information sharing between groups of interest (e.g. between counties or the different races), and provide the framework to account for both within-in and between-in group variances. Such would not be difficult to discern the difference between if one were to just use a normal logistic binomial regression model. Additionally, by using a Bayesian approach, we are able to quantify our uncertainty in the model and induce *a priori* knowledge about parameter estimates.

### 4.2 Model Evaluation

A logical first approach to address differences in registration rates is to fit fixed effects model with `brm()`, for baseline comparison. To start, binomial regression models that considered only fixed effects on `sex_code` and `county` were fit, and the results of such can be seen in in the Appendix. We can see that with the `sex_code` fixed effects model, the estimate for `sex_code = M` is $-0.30$, meaning that under this model, males are $e^{-0.3} = 0.74$ times the odds of registering to vote when compared to females; a finding that is supplemented by our EDA results that found that females had a higher probability of registering to vote across nearly every demographic. Additionally, under a fixed effect model that just uses `county` as a covariate, we can see a fair amount of

variability between counties, as counties such as Clay and Currituck had some of the highest registration odds, compared to Grainville that had the lowest in comparison to the baseline. While these models are interpretable and provide a simple solution to answer the questions of interest directly, they ignore the relevance of any other potential covariates and disregard group-level differences. Moving on with our analysis, our EDA revealed that other variables in the data set such as `sex_code`, `race_code` and `ethnic_code`, could help to explain the variability in the registration data. Additionally, our EDA revealed that there may be potential group level effects that a fixed effects model cannot account for, as placing a random effect on variables suspected of having group-level differences may help to explain additional variance in the data. To test such a theory, the following models were fit using `brm()` in R.

Table 2: Models Tested

| Model | Equation |
|---|---|
| model a | 1 + age + sex_code + (1\|county_desc) |
| model b | 1 + age + sex_code + ethnic_code + (1\|county_desc) |
| model c | 1 + age + sex_code + ethnic_code + race_code + (1 + sex_code\|county_desc) |
| model d | 1 + age + sex_code + race_code + ethnic_code + (1 + race_code\|county_desc) |

Aside from conducting posterior predictive checks, another way to compare Bayesian models is to use the Wantabe-Akaike Information Criterion (WAIC) , or `waic` in the table below, where lower `waic` scores indicate a better model fit. Looking at the results below, we can see that `model_d` that placed a random intercept on `county`, a random slope on `race_code`, and fixed effects on `age`, `sex_code`, `ethnic_code`, and `race_code` was the best performing model with the smallest `waic` value.

Table 3: Comparing Model Performance

| | elpd_diff | se_diff | elpd_waic | se_elpd_waic | p_waic | se_p_waic | waic | se_waic |
|---|---|---|---|---|---|---|---|---|
| model_d | 0.00 | 0.000 | -40706.08 | 2765.468 | 7547.379 | 777.0154 | 81412.16 | 5530.936 |
| model_c | -1024.89 | 1429.624 | -41730.97 | 2629.566 | 5926.631 | 662.1981 | 83461.94 | 5259.131 |
| model_b | -23249.38 | 3098.899 | -63955.46 | 4038.813 | 4571.356 | 535.1095 | 127910.92 | 8077.625 |
| model_a | -95064.42 | 6771.747 | -135770.50 | 7510.177 | 7482.078 | 746.5573 | 271541.01 | 15020.354 |

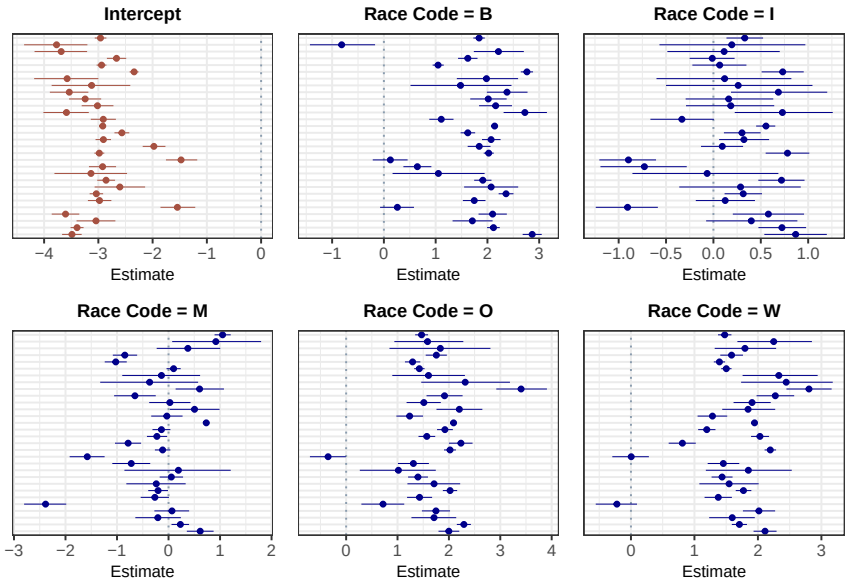
### 4.3 Final Model (`Model I`)

The final model (`model_d`) can be expressed mathematically as the following:

$$
\begin{aligned}
Y_{ijklm} &\sim Bin(n_{ijklm},\ p_{ijklm}) \\
\text{logit}(p_{ijklm}) &= (\beta_0 + b_{0,i}) + (\beta_1 + b_{1,j})\text{race}_j + \beta_2\text{sex}_k + \beta_3\text{age}_l + \beta_4\text{sex}_m \\
(b_{0,i} \quad b_{1,j})^\top &\sim N(\mathbf{0}, \mathbf{D}), \quad \mathbf{D} = \begin{pmatrix} \tau_0 & 0 \\ 0 & \tau_1 \end{pmatrix} \mathbf{W} \begin{pmatrix} \tau_0 & 0 \\ 0 & \tau_1 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & v \\ v & 1 \end{pmatrix}
\end{aligned}
$$

In this model $Y_{ijklm}$ is the the number of registered voters for county $i$, race $j$, sex $k$, age $l$, and ethnicity $m$, and $p_{ijklm}$ is the probability of someone in that demographic class being registered to vote. Additionally we can specify the following priors: $\tau_0 \sim Half\ Cauchy(0,1)$, $\tau_1 \sim Half\ Cauchy(0,1)$, $W = \text{LKJcorr}(2)$, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \sim N(0, 10)$ for all fixed effects. Half Cauchy priors were used on $\tau_0, \tau_1$ since these are variance values and by using this prior, we are forcing values to be $> 0$. Additionally, because *a priori* we are not that confident in the variance of the fixed effect estimators $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, a weakly informative prior of $N(0, 10)$ was used to prevent the posterior from becoming too concentrated around the prior. The LKJ prior was used for $\mathbf{W}$, since this prior is essentially an extension of the beta distribution and serves as a weakly information prior that is commonly used within Stan to model covariance matrices.

Even though this final model had the smallest `waic` there are several checks that must be done to verify that are findings and model selection is justified. First is to verify convergence in the MCMC chains used to fit the model; trace plots (see `Appendix`) validate that our model has converged, as no local trends can be observed. It is also important to verify that placing random effects on `county` and `race_code` was appropriate. One way to visualize these random effects, is to plot the coefficient estimates, along with their 95% credible interval as shown below. The `Intercept` plot illustrates the $b_{0,i}$ estimates for each of the thirty counties; each red dot (or row) corresponds to a different NC county as county names were removed to reduce clutter. Additionally, the blue dot plots correspond the the random slope estimates that our model generated for each `race_code` for all thirty counties.




Looking at the output above, we can see there is a fair amount of variability in the parameter estimates, as county-level estimates appear to be spread throughout the graphic. In particular there are clear differences in estimates of voter registration odds across the thirty counties of interest, as there are numerous instances in the `Intercept` plot where counties and their 95% credible interval do not overlap with one another. Such a finding validates our rationale for placing a random intercept on each county, meaning that there are significant differences between county registration rates, holding all else constant. On the high end, Jackson and Scotland county had the highest county estimates of $-1.48$ and $-1.54$ respectively, whereas Avery and Bertie county had the smallest county estimates at $-3.77$ and $-3.68$. An interesting difference between these two sets of counties that differ the most based on our model, is that Avery and Bertie are in the top 25% of counties by rural land percentage, and have considerable more rural land than Jackson and especially Scotland county. This leads one to infer that perhaps considering the *type* of county (rural vs. urban) can have an impact on registration rates, and could be explored in future analysis.

Aside from county-level differences, we can also see an extensive amount of variability in estimates for each county across the different race codes. Instances of non-overlapping 95% credible intervals in each `race_code` plot justify the decision to use a random effect on `race_code`, since there are clear group-level differences. We can see that `B` and `W` race codes generally had positive parameter estimates; while such becomes difficult to interpret because of the complexity of the model, we can say that generally speaking, holding all other variables constant, `B` and `W` race codes generally had a higher probability of registering to vote when compared to the baseline `A` race code. Additionally, we can see that according to our model the `M` race code had nearly all negative estimates across all counties, thus inferring that this race code is the least probable to register to vote holding all else constant.

Additionally, we can perform inference on the fixed effects components of the model, if we set all random effects equal to zero. Parameter estimates for the fixed effects variables are listed below, as one initial observation is that the estimates of `sex_code = M` is now $-0.22$ which translates to men being 0.80 times the odds of registering to vote when compared to females with a 95% credible interval of $(0.79, 0.81)$, holding all else constant. Similar to

our EDA findings, Non-Hispanic (`ethnic_code = NL`) are predicted to be 11.8 times the odds of registering to when compared to Hispanic individuals, with a 95% credible interval of $(11.59, 12.06)$, holding all else constant. Due to page constraints, included in the `Appendix` is a similar table but for the random effects in the final `Model I`.

Table 4: Final Model I Fixed Effects

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -2.9610044 | 0.1098707 | -3.1794071 | -2.7444319 | 1.0004163 | 2475.034 | 3406.667 |
| ageAge26M40 | 0.2949919 | 0.0058738 | 0.2834746 | 0.3064892 | 1.0000464 | 11220.177 | 6334.149 |
| ageAge41M65 | 0.4574619 | 0.0054943 | 0.4466245 | 0.4683658 | 0.9998574 | 11406.449 | 6329.303 |
| ageAgeOver66 | 1.0015527 | 0.0072435 | 0.9872546 | 1.0156228 | 1.0002586 | 11165.354 | 6391.593 |
| sex_codeM | -0.2240040 | 0.0036768 | -0.2312135 | -0.2167599 | 1.0002683 | 10251.039 | 5066.117 |
| race_codeB | 1.7097619 | 0.1637158 | 1.3834369 | 2.0305698 | 1.0005876 | 2499.786 | 3726.771 |
| race_codeI | 0.2260103 | 0.1021234 | 0.0208212 | 0.4224286 | 1.0001777 | 3613.396 | 4592.835 |
| race_codeM | -0.1461296 | 0.1415339 | -0.4228329 | 0.1394069 | 1.0002412 | 3459.768 | 4996.011 |
| race_codeO | 1.6806285 | 0.1258451 | 1.4360760 | 1.9352490 | 1.0005338 | 3859.409 | 5039.828 |
| race_codeW | 1.6595605 | 0.1166281 | 1.4236273 | 1.8827051 | 1.0003023 | 2963.463 | 4050.585 |
| ethnic_codeNL | 2.4685420 | 0.0101745 | 2.4486373 | 2.4887112 | 1.0006614 | 10038.526 | 4776.421 |

## 5. Model II

### 5.1 Data Preparation

As noted earlier, the objective of fitting a second model (`Model II`) was to evaluate differences in registration rates better age groups and sexes when considering different political party affiliations. The decision was made to fit separate models (`Model I` and `Model II`), since the data had to be aggregated slightly different to answer both sets of questions of interest for the modeling scheme devised. Since only the voters data set contained political party data, political party population values had to be imputed into the census data set in order to use the binomial regression approach described in Section 4.1. In order to impute these values for the Democratic (`DEM`), Republican (`REP`), and Unaffiliated (`UNA`) parties, the sum of the number of registered voters for the two opposite parties, was subtracted from the total census number for that particular subgroup. For example, $n_{DEM_g}$, the *theoretical* census number for demographic group $g$ that are Democrats was computed by $n_{DEM_g} = n_{total_g} - (Y_{REP_g} + Y_{UNA_g})$, where $n_{total_g}$ is the total census figure for group $g$, $Y_{REP_g}$ is the number of Republican registered voters for group $g$, and $Y_{UNA_g}$ is the number of Unaffiliated registered voters for group $g$. The same methodology was used to similarly to calculate $n_{REP_g}$ and $n_{UNA_g}$. Once we had these values, we could now carry out a similar analysis as we did for `Model I`, using the notations that $Y_{DEM_g} \sim Bin(n_{DEM_g},\ p_{DEM_g})$, $Y_{REP_g} \sim Bin(n_{REP_g},\ p_{REP_g})$, and $Y_{UNA_g} \sim Bin(n_{UNA_g},\ p_{UNA_g})$, where $g$ represented an abbreviated notation for a given demographic subgroup *ijklm*. Next, the same multilevel model described in was fit for each political party as outlined in Section 4.3, where a random intercept was placed on county, a random slope on race code, followed by fixed effects on the remaining age, race, sex code, and ethnicity covariates. Additionally the same prior scheme was used from earlier with an Half Cauchy prior on the variance values to force them to be positive, relatively weakly information Normal(0,10) priors on the fixed effects, followed by the LKJ(2) prior on the covariance matrix. Three separate, but identical models were fit, one for each political party, as the only difference was that the data source was filtered to only include entries for that particular political party of interest.

### 5.2 Final Models (`Model II`)

The table below represents the fixed effect estimates for the three different binomial regression models fit for each political party. Variable names were shortened to save space so that `Male` is the `sex_code` variable, and (`B`, `I`, `M`, `O`, `W`) are the different categories of `race_code`. Looking at the output below, we can that compared

to the baseline age group to `18-25`, voter registration probabilities tended to increase as one got older for both Democrats and Republicans, holding all else constant. Such can be seen below, where Republicans over the age of 65 were 3.06 times, with a 95% credible interval of (3, 3.12) to register to vote compared to the youngest age group of 18-25. Additionally, for Democrats, individuals over the age of 65 were 5.7 times, with a 95% credible interval of (5.58, 5.75) as likely to vote compared to 18-25 year old Democrats. Such a finding supported our EDA analysis, which revealed that voter participation increased as age increased. Interestingly enough, the Unaffiliated party group had mixed results as Unaffiliated individuals age 65 and over were only 1.19 times as likely to registered to vote when compared to the 18-25 age group population, as the 26-40 age group for the party actually had the highest odds of registering. Such a result is not entirely shocking, since the Unaffiliated political party is essentially made of up individuals who are undecided about whether they lean more conservative or liberal on certain issues, as it could be expected for this level indecisiveness to decrease as one grew in age.

Differences between males and females provided mixed results as well. The odds of a Democrat male registering to vote was only 0.66 times, with a 95% credible interval of (0.65, 0.67), the odds of a Democrat female. For Republicans, this sex code difference was almost non-existent as males were 0.95 times, with a 95% credible interval of (0.94, 0.96) of registering to vote when compared to females. Perhaps the most surprising results observed when separate models were fit for each political party was the vast differences in race groups. For example, the `B` race code estimate for the Democrat model was 2.79, which equates to a staggering 16.3 times as likely to register to vote when compared to the baseline race group for this political party. On the contrast, the model estimate for `B` race code that are Republicans is -0.27, which means that Black Republicans are only 0.75 as times as likely to vote compared to the baseline race, holding all else constant. Additionally, we can see that the estimates for the Democrat and Unaffiliated party models were very similar for the `W` race code, whereas white Republican voters had a significantly higher estimate and thus higher odds of registering the vote. Finally, when considering ethnicity, Non-Hispanic voters were more likely to register compared to Hispanic voters; The full fixed effects parameter estimate tables for these three models can be found in the Appendix of this report.

Table 5: Model II Fixed Effect Estimates

| Party | Int. | 26-40 | 41-65 | >65 | Male | B | I | M | O | W | NL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DEM | -4.10 | 0.52 | 0.83 | 1.74 | -0.41 | 2.79 | 0.36 | 0.13 | 1.72 | 1.46 | 2.04 |
| REP | -5.42 | 0.40 | 0.69 | 1.12 | -0.05 | -0.27 | 0.33 | -0.80 | 2.07 | 2.26 | 3.21 |
| UNA | -3.36 | 0.19 | -0.04 | 0.17 | -0.11 | 0.66 | 0.00 | -0.42 | 1.36 | 1.34 | 2.39 |

## 6. Conclusion and Model Limitations:

In this study, we explored differences in demographic subgroup voter registration rates for thirty randomly selected NC counties. In `Model I`, we observed clear across-county differences that females were more likely to vote than males, along with differences in rates between the counties selected. `Model II` was constructed to address questions regarding differences when considering different political party affiliation. As such, our results in `Model II` illustrated that both `sex_code` and `race_code` registration odds differed depending on whether one was a Democrat or Republican, while `age` group differences were generally the same for these parties. There are several limitations of the model that must be addressed; one of which has to do with how the data was collected, since there were numerous instances where different subgroups existed in one data set, but not the other. This resulted in some data (albeit very small amounts) having to be dropped from the final data set or imputed by assuming *MCAR*. Moving forward a better approach would to either improve the collection methods of both data sets so the same demographic exist in both surveys, or to assume *MNAR* and to use a more improved data imputation method to reduce the uncertainty in the data. Additionally, using data sets that differed by six years also limits our inferences, as perhaps future analysis should try to use data collected closer in time. Another limitation of the model was that simple random sampling was used to select the counties of interest. Perhaps further analysis could place more importance of using a sampling scheme such as stratified sampling, so that the counties selected are representative of the results from the state of North Carolina.

## 7 Appendix

**Note: this section of the case study is not to be included in the 8 page requirement and is used to portray supplemental material. Attached to this Sakai submittal is also the .Rmd that was used to generate this document along with all supporting code**

### 7.1 EDA Supplemental Material

Table 6: Sex Code Voter Registration Proportions

| Sex Code | Registration Number | Census Number | Proportion |
|---|---|---|---|
| F | 894413 | 1100468 | 0.8128 |
| M | 745664 | 976481 | 0.7636 |

Table 7: Race Code Voter Registration Proportions

| Race Code | Registration Number | Census Number | Proportion |
|---|---|---|---|
| A | 14057 | 30615 | 0.4592 |
| B | 349930 | 414180 | 0.8449 |
| I | 9426 | 17061 | 0.5525 |
| M | 9166 | 23211 | 0.3949 |
| O | 17769 | 63877 | 0.2782 |
| W | 1239729 | 1528005 | 0.8113 |

Table 8: Age Group Voter Registration Proportions

| Age Group | Registration Number | Census Number | Proportion |
|---|---|---|---|
| Age 18 - 25 | 192320 | 275862 | 0.6972 |
| Age 26 - 40 | 385930 | 519847 | 0.7424 |
| Age 41 - 65 | 743921 | 921372 | 0.8074 |
| Age Over 66 | 317906 | 359868 | 0.8834 |

Table 9: Ethnic Code Voter Registration Proportions

| Ethnic Code | Registration Number | Census Number | Proportion |
|---|---|---|---|
| HL | 32305 | 119555 | 0.2702 |
| NL | 1607772 | 1957394 | 0.8214 |

### 7.2 Model I - Fixed Effects Model Supplemental Material

Fixed Effect Model on `sex_code`:

Table 10: Fixed Effects Model on Sex

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 1.4680 | 0.0025 | 1.4632 | 1.4727 | 1.0030 | 1998.058 | 2323.125 |
| sex_codeM | -0.2954 | 0.0035 | -0.3022 | -0.2885 | 1.0055 | 1100.949 | 1730.458 |

Fixed Effect Model on `county`:

Table 11: Fixed Effects Model on County

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | 1.0945 | 0.0062 | 1.0826 | 1.1068 | 1.0007 | 993.2214 | 1558.321 |
| county_descAVERY | -0.1001 | 0.0189 | -0.1365 | -0.0630 | 1.0012 | 3772.2906 | 2751.922 |
| county_descBERTIE | 0.0360 | 0.0181 | 0.0006 | 0.0710 | 1.0004 | 3473.3583 | 2499.148 |
| county_descBRUNSWICK | 0.3109 | 0.0111 | 0.2895 | 0.3321 | 1.0002 | 2232.5551 | 2594.456 |
| county_descBURKE | -0.0386 | 0.0103 | -0.0588 | -0.0178 | 0.9999 | 1889.4406 | 2562.780 |
| county_descCABARRUS | 0.5230 | 0.0096 | 0.5045 | 0.5422 | 1.0005 | 1694.3836 | 2823.185 |
| county_descCHOWAN | 0.2865 | 0.0230 | 0.2422 | 0.3311 | 1.0012 | 4298.8050 | 2572.112 |
| county_descCLAY | 0.9083 | 0.0337 | 0.8435 | 0.9745 | 0.9998 | 5600.2565 | 2793.855 |
| county_descCURRITUCK | 0.6798 | 0.0221 | 0.6374 | 0.7249 | 1.0007 | 3917.3060 | 2966.718 |
| county_descDARE | 0.3445 | 0.0169 | 0.3111 | 0.3785 | 1.0003 | 3665.5427 | 2654.838 |
| county_descDAVIE | 0.4204 | 0.0153 | 0.3899 | 0.4505 | 1.0000 | 3156.4817 | 2897.661 |
| county_descEDGECOMBE | 0.3017 | 0.0129 | 0.2765 | 0.3261 | 1.0009 | 2426.8191 | 2688.029 |
| county_descGRANVILLE | -0.1938 | 0.0114 | -0.2174 | -0.1713 | 1.0003 | 2288.3461 | 2317.757 |
| county_descGUILFORD | 0.3925 | 0.0074 | 0.3781 | 0.4069 | 1.0003 | 1259.0808 | 2125.195 |
| county_descHARNETT | 0.0912 | 0.0100 | 0.0710 | 0.1101 | 1.0000 | 1917.1631 | 2664.120 |
| county_descHENDERSON | 0.5050 | 0.0107 | 0.4843 | 0.5256 | 1.0001 | 2138.7473 | 2569.232 |
| county_descHOKE | 0.3312 | 0.0146 | 0.3029 | 0.3595 | 0.9997 | 3038.5320 | 2428.705 |
| county_descIREDELL | 0.5730 | 0.0098 | 0.5536 | 0.5919 | 1.0005 | 1820.0496 | 2405.532 |
| county_descJACKSON | -0.0399 | 0.0132 | -0.0666 | -0.0142 | 1.0004 | 2684.9450 | 3090.420 |
| county_descMCDOWELL | 0.0733 | 0.0133 | 0.0463 | 0.0996 | 1.0001 | 2930.5629 | 2692.470 |
| county_descMITCHELL | 0.3430 | 0.0234 | 0.2977 | 0.3878 | 1.0005 | 4135.4579 | 2534.016 |
| county_descNASH | 0.2353 | 0.0108 | 0.2141 | 0.2563 | 0.9999 | 2230.4849 | 2836.410 |
| county_descPOLK | 0.5896 | 0.0208 | 0.5474 | 0.6291 | 1.0006 | 4050.2143 | 2924.388 |
| county_descRANDOLPH | 0.1649 | 0.0093 | 0.1465 | 0.1833 | 1.0003 | 1657.7399 | 2936.098 |
| county_descROCKINGHAM | 0.0335 | 0.0104 | 0.0126 | 0.0540 | 1.0001 | 2322.4896 | 2902.847 |
| county_descSCOTLAND | -0.0854 | 0.0152 | -0.1151 | -0.0560 | 0.9999 | 3239.7245 | 2522.769 |
| county_descSURRY | -0.1029 | 0.0109 | -0.1240 | -0.0812 | 1.0003 | 2168.2544 | 2791.345 |
| county_descTRANSYLVANIA | 0.2346 | 0.0148 | 0.2053 | 0.2636 | 1.0011 | 3217.3677 | 3064.044 |
| county_descWAYNE | -0.1273 | 0.0094 | -0.1461 | -0.1093 | 1.0002 | 1621.1270 | 2543.619 |
| county_descWILSON | 0.2446 | 0.0112 | 0.2228 | 0.2662 | 1.0013 | 2059.4211 | 2812.906 |

### 7.2 Model I - Final Mixed Effects Model Supplemental Material

Final Model County Estimates:

Table 12: Final Model I County Estimates

| | Estimate | Est.Error | Q2.5 | Q97.5 |
|---|---|---|---|---|
| ALAMANCE | -2.9601 | 0.0546 | -3.0667 | -2.8522 |
| AVERY | -3.7702 | 0.2990 | -4.3690 | -3.2022 |
| BERTIE | -3.6853 | 0.2478 | -4.1765 | -3.2075 |
| BRUNSWICK | -2.6651 | 0.0920 | -2.8427 | -2.4867 |
| BURKE | -2.9404 | 0.0459 | -3.0318 | -2.8518 |
| CABARRUS | -2.3408 | 0.0422 | -2.4243 | -2.2603 |
| CHOWAN | -3.5757 | 0.2987 | -4.1834 | -2.9996 |
| CLAY | -3.1235 | 0.3692 | -3.8613 | -2.4113 |

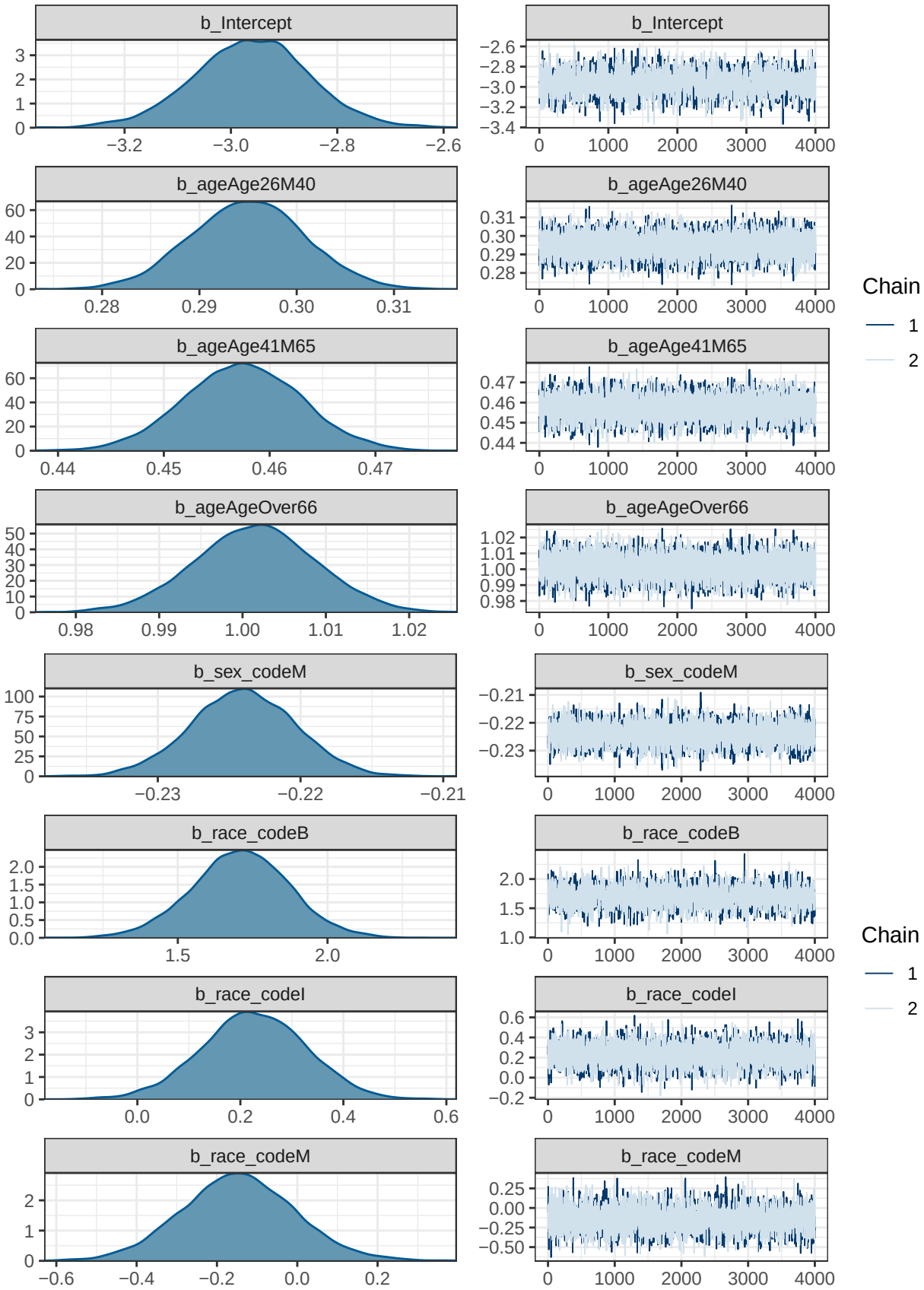
| | Estimate | Est.Error | Q2.5 | Q97.5 |
|---|---|---|---|---|
| CURRITUCK | -3.5364 | 0.1822 | -3.8940 | -3.1842 |
| DARE | -3.2429 | 0.1517 | -3.5396 | -2.9484 |
| DAVIE | -3.0159 | 0.1501 | -3.3074 | -2.7223 |
| EDGECOMBE | -3.5876 | 0.2129 | -4.0131 | -3.1763 |
| GRANVILLE | -2.9092 | 0.1185 | -3.1397 | -2.6767 |
| GUILFORD | -2.9189 | 0.0202 | -2.9579 | -2.8792 |
| HARNETT | -2.5663 | 0.0704 | -2.7034 | -2.4304 |
| HENDERSON | -2.9066 | 0.0719 | -3.0521 | -2.7671 |
| HOKE | -1.9762 | 0.1072 | -2.1833 | -1.7638 |
| IREDELL | -2.9825 | 0.0462 | -3.0728 | -2.8908 |
| JACKSON | -1.4750 | 0.1457 | -1.7539 | -1.1774 |
| MCDOWELL | -2.9216 | 0.1260 | -3.1735 | -2.6739 |
| MITCHELL | -3.1358 | 0.3453 | -3.8097 | -2.4715 |
| NASH | -2.8581 | 0.0845 | -3.0215 | -2.6923 |
| POLK | -2.6035 | 0.2357 | -3.0694 | -2.1381 |
| RANDOLPH | -3.0375 | 0.0634 | -3.1608 | -2.9124 |
| ROCKINGHAM | -2.9782 | 0.1099 | -3.1930 | -2.7602 |
| SCOTLAND | -1.5433 | 0.1633 | -1.8553 | -1.2139 |
| SURRY | -3.6048 | 0.1307 | -3.8599 | -3.3514 |
| TRANSYLVANIA | -3.0433 | 0.1825 | -3.4023 | -2.6864 |
| WAYNE | -3.3926 | 0.0615 | -3.5128 | -3.2729 |
| WILSON | -3.4864 | 0.0932 | -3.6687 | -3.3078 |

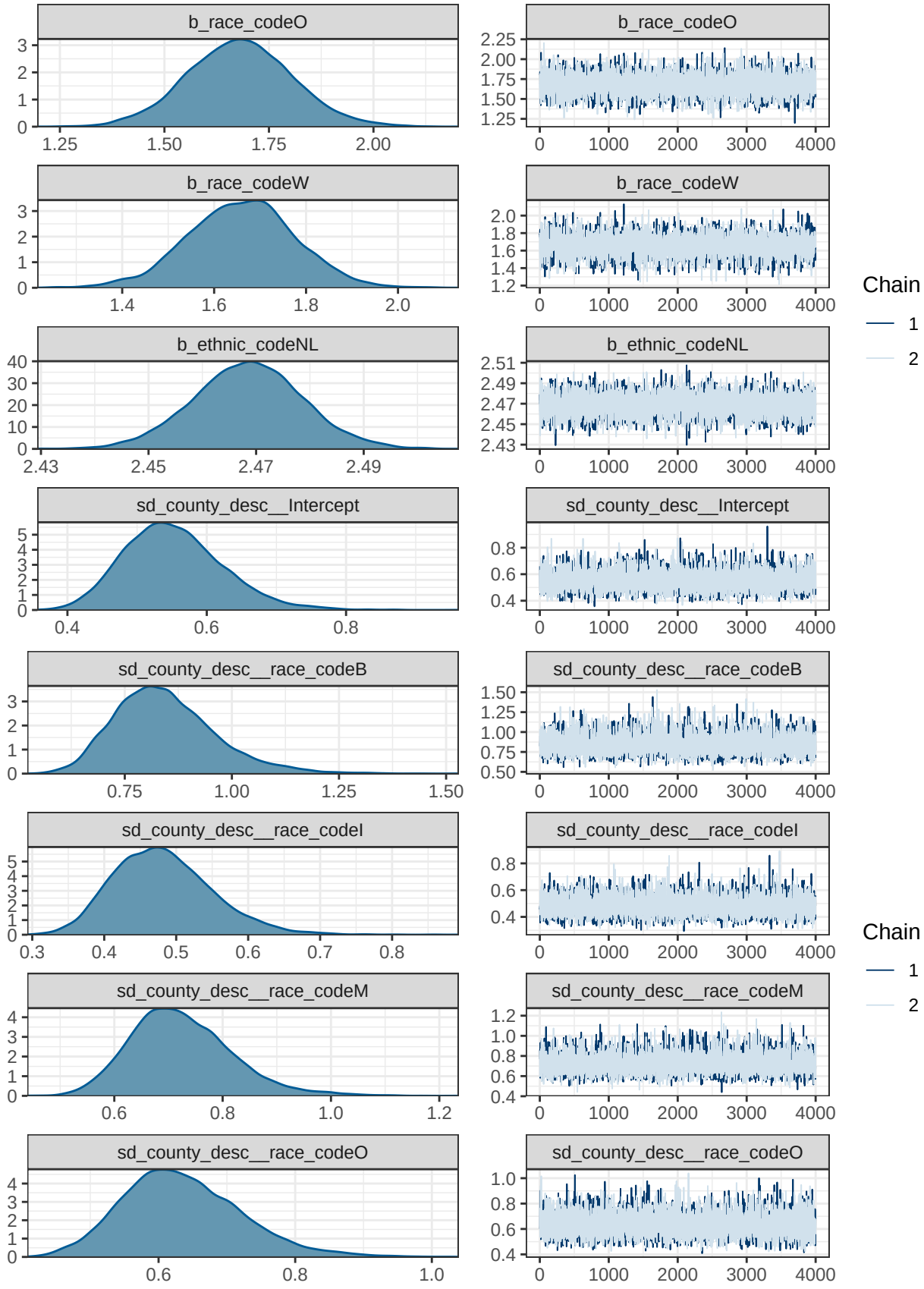
Final Model Random Effects:

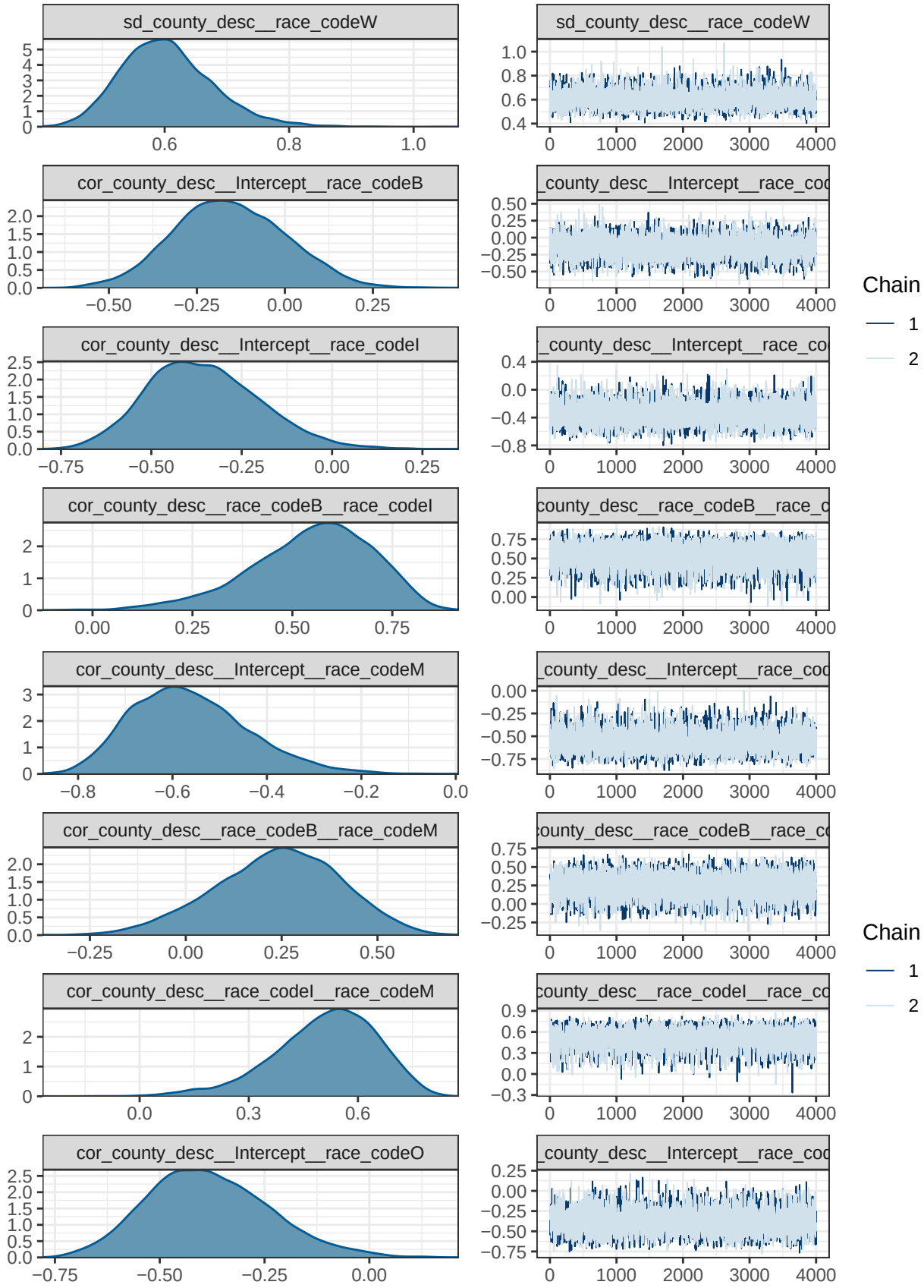
Table 13: Final Model I Random Effects

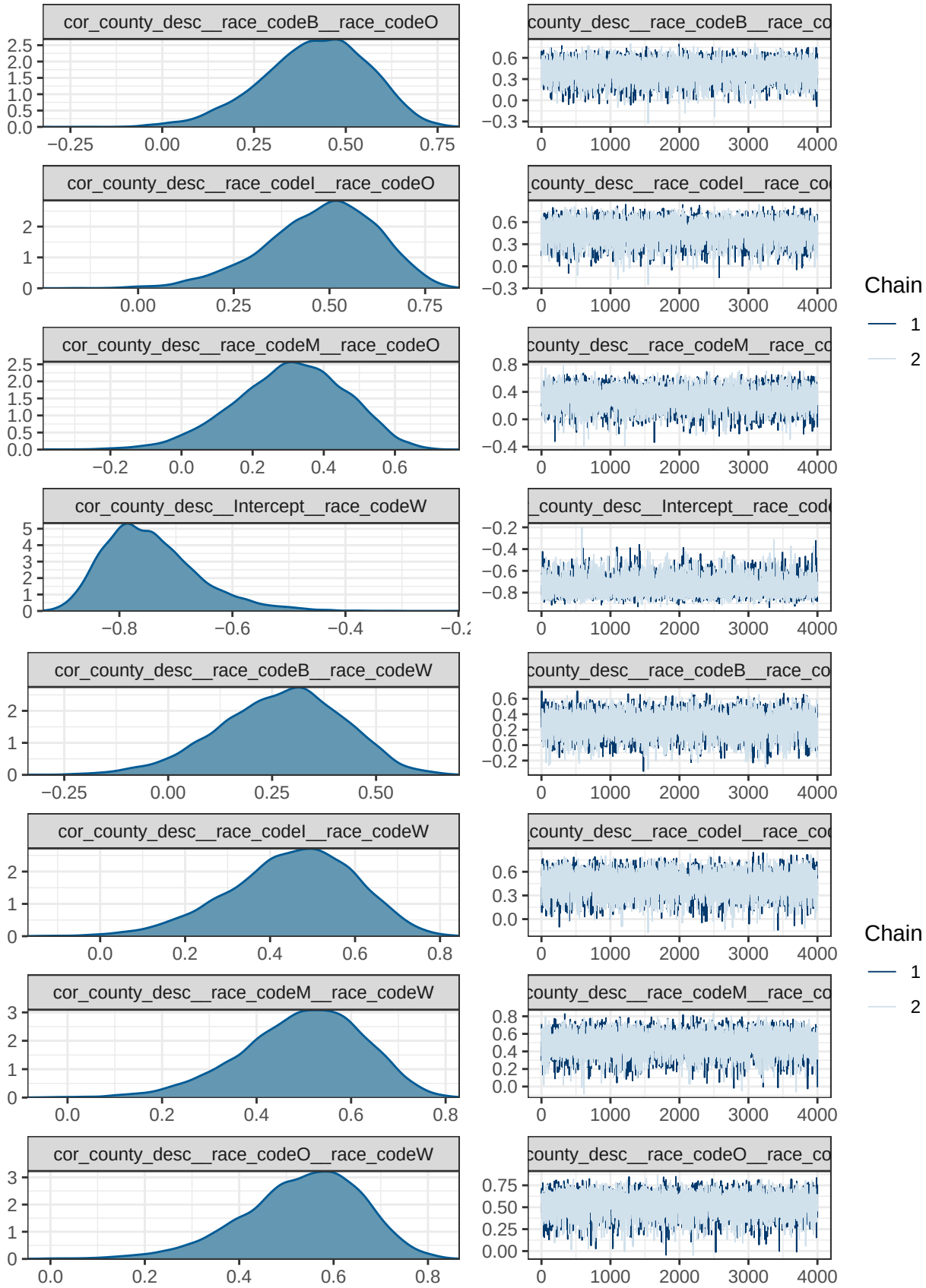
| | Est | Error | 95 CI L | 95% CI U | Rhat | ESS | Tail ESS |
|---|---|---|---|---|---|---|---|
| sd(Intercept) | 0.5526 | 0.0707 | 0.4297 | 0.7055 | 1.0003 | 4543.113 | 5256.864 |
| sd(race_codeB) | 0.8446 | 0.1178 | 0.6531 | 1.1168 | 1.0004 | 4373.895 | 5567.921 |
| sd(race_codeI) | 0.4820 | 0.0686 | 0.3662 | 0.6315 | 1.0008 | 4321.060 | 6049.189 |
| sd(race_codeM) | 0.7243 | 0.0937 | 0.5652 | 0.9363 | 1.0007 | 4688.549 | 5927.301 |
| sd(race_codeO) | 0.6350 | 0.0865 | 0.4860 | 0.8260 | 1.0009 | 4472.989 | 5179.972 |
| sd(race_codeW) | 0.6035 | 0.0737 | 0.4774 | 0.7679 | 0.9999 | 4753.205 | 4975.397 |
| cor(Intercept,race_codeB) | -0.1559 | 0.1582 | -0.4555 | 0.1565 | 1.0008 | 2883.254 | 4316.973 |
| cor(Intercept,race_codeI) | -0.3542 | 0.1552 | -0.6306 | -0.0261 | 1.0003 | 6398.399 | 5814.852 |
| cor(race_codeB,race_codeI) | 0.5479 | 0.1525 | 0.2016 | 0.7970 | 1.0002 | 6733.053 | 5795.040 |
| cor(Intercept,race_codeM) | -0.5624 | 0.1225 | -0.7658 | -0.2945 | 0.9999 | 5557.441 | 5223.449 |
| cor(race_codeB,race_codeM) | 0.2395 | 0.1631 | -0.0953 | 0.5377 | 1.0000 | 5941.495 | 5589.574 |
| cor(race_codeI,race_codeM) | 0.5105 | 0.1417 | 0.1800 | 0.7445 | 1.0017 | 4307.671 | 5085.338 |
| cor(Intercept,race_codeO) | -0.3823 | 0.1452 | -0.6399 | -0.0717 | 1.0005 | 5496.467 | 5508.323 |
| cor(race_codeB,race_codeO) | 0.4157 | 0.1452 | 0.1064 | 0.6669 | 1.0004 | 6486.193 | 5836.145 |
| cor(race_codeI,race_codeO) | 0.4777 | 0.1469 | 0.1511 | 0.7264 | 1.0004 | 4919.667 | 5760.931 |
| cor(race_codeM,race_codeO) | 0.3057 | 0.1567 | -0.0188 | 0.5797 | 1.0000 | 5709.568 | 5573.825 |
| cor(Intercept,race_codeW) | -0.7438 | 0.0834 | -0.8721 | -0.5457 | 1.0006 | 4565.411 | 5288.621 |
| cor(race_codeB,race_codeW) | 0.2711 | 0.1493 | -0.0460 | 0.5350 | 1.0000 | 4629.625 | 5175.819 |
| cor(race_codeI,race_codeW) | 0.4511 | 0.1481 | 0.1336 | 0.7060 | 1.0001 | 5316.228 | 5996.126 |
| cor(race_codeM,race_codeW) | 0.4961 | 0.1295 | 0.2153 | 0.7164 | 1.0002 | 5583.822 | 4942.291 |
| cor(race_codeO,race_codeW) | 0.5328 | 0.1274 | 0.2469 | 0.7447 | 1.0002 | 5821.429 | 5554.223 |

Final Model Trace Plots:




b_race_codeO
b_race_codeW
b_ethnic_codeNL
sd_county_desc__Intercept
sd_county_desc__race_codeB
sd_county_desc__race_codeI
sd_county_desc__race_codeM
sd_county_desc__race_codeO
Chain
1
2


sd_county_desc__race_codeW
cor_county_desc__Intercept__race_codeB
cor_county_desc__Intercept__race_codeI
cor_county_desc__race_codeB__race_codeI
cor_county_desc__Intercept__race_codeM
cor_county_desc__race_codeB__race_codeM
cor_county_desc__race_codeI__race_codeM
cor_county_desc__Intercept__race_codeO
Chain
1
2




### 7.3 Model II - Final Mixed Effects Model Supplemental Material

Final Political Party Models Fixed Effects Full Tables:

Table 14: Model II DEM Fixed Effects

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -4.0957856 | 0.1354995 | -4.3697120 | -3.8365138 | 1.0005467 | 1700.041 | 3220.409 |
| ageAge26M40 | 0.5207766 | 0.0077387 | 0.5056609 | 0.5358204 | 0.9998971 | 10691.654 | 6543.799 |
| ageAge41M65 | 0.8284283 | 0.0071908 | 0.8143319 | 0.8427780 | 1.0003848 | 10641.455 | 6138.019 |
| ageAgeOver66 | 1.7354033 | 0.0088722 | 1.7179583 | 1.7530126 | 0.9999913 | 10762.677 | 6284.438 |
| sex_codeM | -0.4054367 | 0.0046769 | -0.4145111 | -0.3965152 | 1.0000836 | 11823.269 | 3889.019 |
| race_codeB | 2.7881285 | 0.1617755 | 2.4678198 | 3.1095204 | 1.0002785 | 1960.836 | 3003.155 |
| race_codeI | 0.3594558 | 0.1349876 | 0.0808057 | 0.6097285 | 1.0007984 | 3020.971 | 4180.279 |
| race_codeM | 0.1310033 | 0.1698357 | -0.1933238 | 0.4711879 | 1.0007975 | 2940.081 | 4170.101 |
| race_codeO | 1.7191178 | 0.1250818 | 1.4781316 | 1.9644742 | 1.0002614 | 2717.641 | 3970.167 |
| race_codeW | 1.4610475 | 0.1301134 | 1.2131353 | 1.7257034 | 1.0004806 | 2048.656 | 3685.666 |
| ethnic_codeNL | 2.0421358 | 0.0151285 | 2.0130088 | 2.0719041 | 0.9999375 | 13879.988 | 5285.208 |

Table 15: Model II REP Fixed Effects

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -5.4247519 | 0.1165498 | -5.6527173 | -5.1982503 | 1.0007082 | 1543.358 | 2645.002 |
| ageAge26M40 | 0.3998820 | 0.0079606 | 0.3842160 | 0.4158438 | 0.9999669 | 9427.769 | 6861.797 |
| ageAge41M65 | 0.6875583 | 0.0072286 | 0.6734482 | 0.7016260 | 1.0000816 | 8224.657 | 6562.069 |
| ageAgeOver66 | 1.1204892 | 0.0088462 | 1.1034409 | 1.1375838 | 1.0002350 | 9176.537 | 6727.369 |
| sex_codeM | -0.0498676 | 0.0047048 | -0.0591724 | -0.0405259 | 1.0002913 | 16447.122 | 5132.104 |
| race_codeB | -0.2677326 | 0.1543950 | -0.5702894 | 0.0308390 | 1.0009310 | 1417.582 | 2509.172 |
| race_codeI | 0.3336726 | 0.1188879 | 0.0997151 | 0.5695929 | 1.0007781 | 2092.389 | 3592.013 |
| race_codeM | -0.7967584 | 0.1334035 | -1.0595819 | -0.5405484 | 1.0001701 | 2525.822 | 4543.233 |
| race_codeO | 2.0688865 | 0.1450838 | 1.7810548 | 2.3591524 | 1.0004491 | 2493.613 | 3398.717 |
| race_codeW | 2.2574488 | 0.1193615 | 2.0228220 | 2.4932463 | 1.0007373 | 1738.872 | 3033.629 |
| ethnic_codeNL | 3.2138868 | 0.0195475 | 3.1754432 | 3.2524148 | 0.9998555 | 12143.763 | 5761.104 |

Table 16: Model II UNA Fixed Effects

| | Estimate | Est.Error | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| Intercept | -3.3571416 | 0.1131037 | -3.5787155 | -3.1351793 | 1.0000528 | 1713.823 | 3311.249 |
| ageAge26M40 | 0.1852042 | 0.0069933 | 0.1713781 | 0.1987333 | 0.9998875 | 11999.213 | 7313.332 |
| ageAge41M65 | -0.0358419 | 0.0066082 | -0.0489299 | -0.0232552 | 1.0000532 | 10587.498 | 6209.007 |
| ageAgeOver66 | 0.1729840 | 0.0088033 | 0.1556737 | 0.1903056 | 1.0001802 | 11347.481 | 6587.102 |
| sex_codeM | -0.1061764 | 0.0045867 | -0.1150366 | -0.0970746 | 1.0004326 | 12041.836 | 4874.841 |
| race_codeB | 0.6595631 | 0.1440487 | 0.3723111 | 0.9403077 | 1.0009433 | 1941.139 | 3097.610 |
| race_codeI | -0.0036264 | 0.1083663 | -0.2195939 | 0.2085949 | 1.0005811 | 2582.233 | 3888.500 |
| race_codeM | -0.4233337 | 0.1260652 | -0.6762435 | -0.1736848 | 0.9998744 | 2605.035 | 4156.626 |
| race_codeO | 1.3567374 | 0.1339184 | 1.0958505 | 1.6248583 | 1.0004145 | 2823.034 | 3700.534 |
| race_codeW | 1.3402809 | 0.1185945 | 1.1095300 | 1.5731566 | 1.0012052 | 2411.974 | 3840.895 |
| ethnic_codeNL | 2.3880775 | 0.0148454 | 2.3594631 | 2.4170890 | 1.0002241 | 13886.421 | 5586.626 |