Modeling Turnout in the North Carolina 2020 Election

Raza Lamb (Coordinator), Robert Wan (Writer)
Team: Erika Fox (Programmer), Minjung Lee (Presenter), Preet Khowaja (Checker),

10/24/2021

Summary

Through utilization of a heirarchical aggregated logistic regression model we analyzed data from the North Carolina State Board of Elections to determine how demographic factors and location affect odds of voting (i.e. voter turnout) in a subset of counties in North Carolina. The analysis demonstrated that the baseline probability of voting varies widely by county, from 47% to 68%. We also discovered that several demographic factors strongly influence turnout, including, ethnicity, race, age, and party affiliation.

Introduction

Despite the COVID-19 pandemic, the United States 2020 presidential election had the highest voter turnout of the 21st century, according to Census data. There are many potential reasons for this, including the often debated expansion of mail-in voting seen during the pandemic. However, the United States has an incredibly diverse landscape, and the turnout in the 2020 election could be driven by any number of factors. One specific insight that we hope to gain here is the effect of demographic variables and location on voter turnout.

In this analysis, we investigated the relationship between turnout in the 2020 U.S. presidential election and various demographic factors in North Carolina. The data used in this report derives from two separate precinct-level data sets from the North Carolina State Board of Elections—one consisting of votes and another consisting of registered voters. We utilized a hierarchical logistic regression model to examine how specific demographic factors, in addition to location (county), affect the odds of a registered voter voting in the election. Specifically, the questions of interest included:

- How did different demographic subgroups vote in the 2020 general elections in North Carolina?
- Did the overall probability or odds of voting differ by county in 2020? Which counties differ the most from other counties?
- How did the turnout rates differ between females and males for the different party affiliations?
- How did the turnout rates differ between age groups for the different party affiliations?

Data

As mentioned previously, the data employed in this analysis was initially in two separate data sets, both available from the North Carolina State Board of Elections (NCSBE). In both cases, the data is in aggregate form: each row represents an aggregated number of voters by demographic characteristics and location. Due to this format, the data required a significant amount of data preparation before beginning exploratory data analysis.

Data Preparation

In order to make this analysis more manageable, we used the data from only 25 of the 100 counties in North Carolina. In order to select which counties to use, we set the seed at 4747 and took a random sample of 25 counties out of North Carolina's 100 total counties, then subset both the registration and the turnout data sets—after this, there were 180,011 observations in the dataset containing turnout and 112,880 observations in the dataset containing registered voters. After collecting the subset, we noticed that the two datasets are structured slightly differently. The data set on voters from the 2020 election includes 3 more variables than the data on registered voters. These variables are: $voting_{m}ethod$, $voting_{m}ethod_{d}esc$, and $voted_{p}arty_{c}d$. In the data dictionary, $voted_{p}arty_{c}d$ is described as "voted party by ballot style," which while not immediately informative, further investigation revealed that this is the variable to match to the party variable in the registration data. Continually, further examination showed that $voting_{m}ethod$ and $voting_{m}ethod_{d}esc$ are identical in the subset. These variables describe the method by which the aggregated voter counts voted, i.e. in-person, curbside, etc. Intuitively, the registration data will not having a matching variables. Thus, before merging the data, we aggregated the voters data further, collapsing the $voting_{m}ethod$ column. Additionally, as discussed in the next section, we also decided to collapse both datasets by precinct, before merging. After this aggregation, the turnout data set had 11,296 observations, and the registration data set had 12,818 observations Following this, we performed a left join, with the voters data on the left and the registration data on the right. This is due to the fact that we observed several rows (~1500) in the registration data that do not have a matching row in the voter data, due to no voters in that aggregated data voting in the 2020 election.

After further aggregating the voting data and joining it with the registered voter data set, we further examined the data to ensure that it had been joined correctly. First of all, we examining the data for any missing values, we found 17 rows with missing values, and in each case the missing value was in the registered voter count. What this indicated was that there were demographic aggregates of voters who had a zero count in registration, and therefore not in the data set. There are a few reasons for this to occur, including the fact that voters can move and not update their registration, or that voters can register on the day of the election. However, because we do not have a registration figure for these voters, we simply drop these observations from the data. In a similar vein, we also examined the data for situations in which turnout exceeds registered count, for which there were 7 rows. When examining these rows, we see that in every case turnout exceeds registered voters by only one vote, and the total number of registered voter in each row is 7 or fewer. Because this is not a reasonable outcome to model (turnout cannot exceed 100%), we decided to reset these values so that turnout is equal to the number of registered voters in this row. While this may not be ideal, due to the very small magnitude of the change, we do not believe this affected the analysis. The final data set unsed for EDA and modeling contained 11,279 observations.

Exploratory Data Analaysis

After data cleaning and preparation, we proceeded to conduct exploratory data anlaysis. During EDA, we created a variable, $per_{t}urnout$, representing percent turnout for each row in the data. In addition to this response variable, there were 7 potential predictors in the data set, all categorical variables. The potential predictors were:

Code book:

- $county_{d}esc$: county in North Carolina
- $precinct_{a}bbrv$: precinct within the county
- age: age of voter, divided into four categories
- $party_{c}d$: political party of voter
- $race_{c}ode$: race of voter
- $ethnic_{c}ode$: ethnicity of voter (i.e. Hispanic/Latino)
- $sex_{c}ode$: sex of voter

First of all, in the exploratory analysis, we noticed that many precincts had very few voters, so we decided to exclude this variable from the analysis. Another important aspect of the data that we discovered was that many voters have unknown demographic variables. For example, 8% of all voters in the 2020 election have unknown sex, while 25% have unknown ethnicity. This was important to consider during both exploratory analysis and during the interpretation of our results. Finally, the last concerning aspect of the data noted here is that many of the demographic factors are unbalanced. While the age and sex are well balanced, the other variables have at least one category with relatively few observations. For example, within the party variable, there are more than 450,000 registered Republicans, Democrats, and unaffiliated voters, but there were less than 1,000 registered voters for the Green and Constitution parties, and less than 10,000 voters for the Libertarian party. This was especially important to consider when examining interactions between demographic variables.

For all of the other variables, we examined the potential relationship between percent turnout and the categorical variables through box-plots. The EDA revealed several relationships that directed the subsequent modeling process. First of all, there was significant variation in the turnout rate by county, with turnout rate ranging from 61% to 82%—overall turnout in the sample of 25 counties was 74%. We also noted that men have lower turnout than women and Hispanic voters have lower turnout than non-Hispanic voters. When examining age, there was a clear trend: as age increases, so does voter turnout. However, this trend does not look linear—the effect between the last two age groups, 41-65 and 66+, is much less pronounced. The variable encoding race also showed an interesting trend. White, Asian, Native American, and Unknown voters have similar turnout rates, while Black, Mixed, and Other voters have lower turnout rates. Native Hawaiian/Pacific Islanders have a very high turnout rate, but with a very small sample size. Finally, we examined how party affects registered voter turnout. As mentioned before, this variable is very unbalanced. However, keeping this in mind, we noticed that Republicans appear to have the highest voter turnout, with other parties (especially Libertarians and unaffiliated voters) lagging behind.

Following an analysis of potential predictors, we continued to examine potential interactions between these categorical variables. Interestingly, most relationships seemed fairly consistent across interactions between demographic variables. One interaction that showed a potential was the interaction between race and sex, included below. For the purpose of interpretation, we have condensed the graph to show only the relationship between White/Black voters and Male/Female, excluding several categories with few observations. Visibly, the trend for both women and men is similar: Black voters, on average, have lower turnout. However, the notable interaction here is that the difference is remarkably less pronounced for women voters. We also investigated whether or not the demographic voting trends vary by county. In many cases, this was difficult to interpret, due to the fact that we were visualizing 25 different counties. However, despite this, we only noted one interaction that was potentially interesting: race and county.

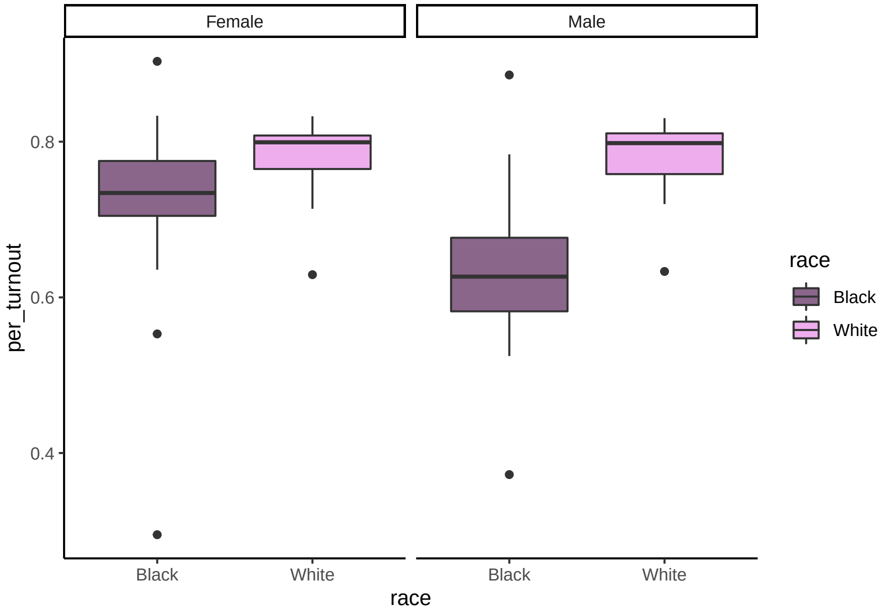


Figure 1: Interaction between Race and Sex

Finally, before moving on to modeling, we decided to condense some of the variables categorically, based on what we observed during EDA. First, as mentioned earlier, several parties have very few voters. To account for this, we created a new category: “other”, and placed the three minor parties into this cateogry (i.e. Constitution, Green, and Libertarian). For the variable $race_code$, we also noticed an unbalanced ratio.

There were more than 950,000 White voters in our data, more than 260,000 Black voters, and more than 141,000 voters with unknown race. However, in the remaining 5 more categories, each had less than 35,000 voters registered. While 35,000 is still a large number, because we are interested in the interaction between race and county, when broken down by county, the sample sizes become less manageable. Specifically, when examining these 5 races by county, nearly half have less than 50 registered voters. Due to this fact, we collapsed these 5 race categories (Asian, Mixed, Native American, Other, and Pacific Islander) into one "other" category.

Model

The final model we selected is a hierarchical aggregated logistic regression model with both varying-intercepts and varying-slopes, shown below. In this model, each county has it's distinct intercept, centered around the grand mean, and the slope for each race category also differs by county, again, centered around a grand mean.

$$y_i|x_i \sim \text{Bin}(n_i, \pi_i);$$

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \gamma_{0j} + \beta_1 party\_cd_{ij} + (\beta_2 + \gamma_{rj}) race\_code_{ij} + \beta_3 ethnic\_code_{ij} + \beta_4 age_{ij}$$

$$+\beta_5 sex\_code_{ij} + \beta_6 sex\_code_{ij} : party\_code_{ij} + \beta_7 age_{ij} : party\_code_{ij} + \beta_8 sex\_code_{ij} : race\_code_{ij};$$

$$\gamma_{0j} \sim N(0, \sigma^2_{county}), \gamma_{0r} \sim N(0, \tau^2_{race\_code}).$$

Model Selection

After conducting EDA, the model we initially started with was the one depicted above. In terms of fixed effects, we included all demographic variables and three interactions: sex and party, age and party, and sex and race. The first two enable us to answer our questions of interest, and the last interaction was observed during the EDA. For random effects, we included a varying-intercept by county and a varying-slope for race by county. In order to perform model selection, we first focused on fixed effects. To test this, we simply created a null model, which was identical, except for the exclusion of the interaction between sex and race. After performing a Chi-Squared test between these two models, we found that the original model was significantly better at the 95% confidence level. Based on this, we decided to utilize the extra interaction, and move on to setting the random effects. For this, we set a null model using **only** varying-intercept, as opposed to varying-intercept **and** slope. To compare these models, we compared the AIC value for each model, and discovered that once again, the full model had a noticeably decreased AIC value—XXX vs. XXX. Based on this result, we selected the final model depicted above.

Model Results

The output of the final model is included in Appendix 1. Using the final model, we were able to answer our inferential questions. First, we can interpret the random effects, using the plot in Appendix 2 and the table in Appendix 1. First of all, the overall intercept for this model is 0.51, which when converted to the probability scale, tells us that the probability of voting for a registered voter in the baseline categories (White, Female, Non-Hispanic, Democrat, age 18-25) is 63% overall. Each individual county has its own random effect that changes this baseline level. Based on the plot in Appendix 2, we can see that this intercept varies by county, and nearly all are significant at the 95% level, based on the confidence bands. For example, the probability of voting for an individual with baseline characteristics in Onslow county is only 47%, while the probability for the same individual in Carteret county is 68%. In addition to varying-intercepts, this model as has varying slopes for race by county. These plots, also included in the appendix, display how the coefficient for each individual race, i.e. Black, Mixed, Other, vary by county. One valuable insight from these slopes is that the distribution of slopes for Black voters is much tighter around the mean than for voters of either unknown race or other race. This observation is confirmed by the standard deviation of the random effects: the standard deviation for the slopes for Black voters represents less than 20% of variation of all slopes for race. One plausible interpretation of this is that this model is likely capturing the different races captured in the other and unknown group, and their distribution across counties. For example, an "Other" or "Unknown" population in a given county may be skewed towards one race, which has a higher or lower turnout rate. This again points to the fact that we did not have sufficient data to fully evaluate the various race categories.

Continually, we also interpreted the fixed effects to determine how effects range by demographics. Because of the inclusion of interactions in the model, interpretation of fixed effects must be interpreted in context of interactions. In order to do this, we visualized the effects using predicted probabilities. For each of the three interactions in our model, we created a visualization. The plots for the interactions between sex and party and age and party are included in Appendix 3. In these, plots, interestingly, we can see that while the effects may be statistically significant, the interactions do not have much scientific significance. In other words, the trend in voting by sex and by age does not differ by party. On the other hand, the graph below depicts the predicted probability of voting broken down by sex and race. Here, clearly the the odds of voting by race differs by sex. Black men are significantly less likely than White men vote, while Black women have turnout much closer to White women.

```
## [[1]]
```

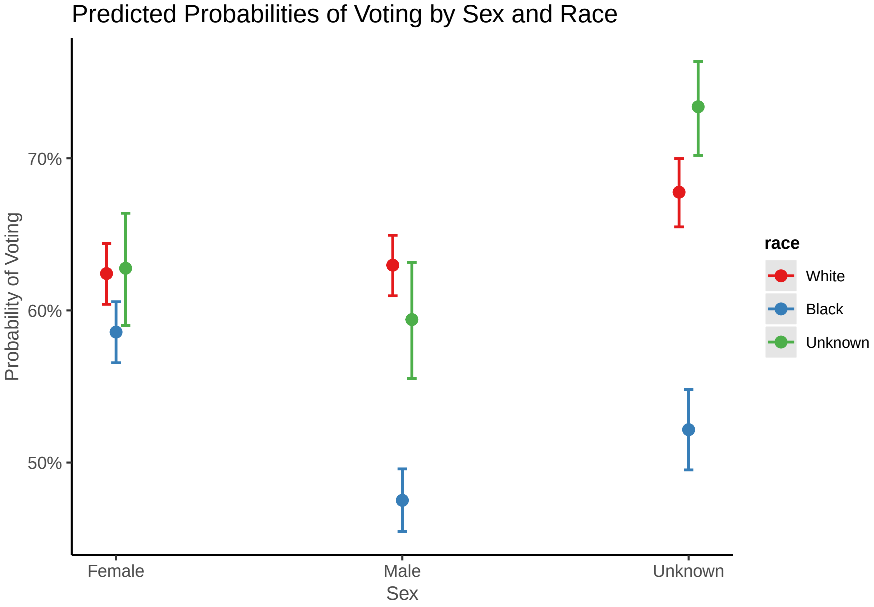


Figure 2: Interatcion of Sex and Race

Conclusion

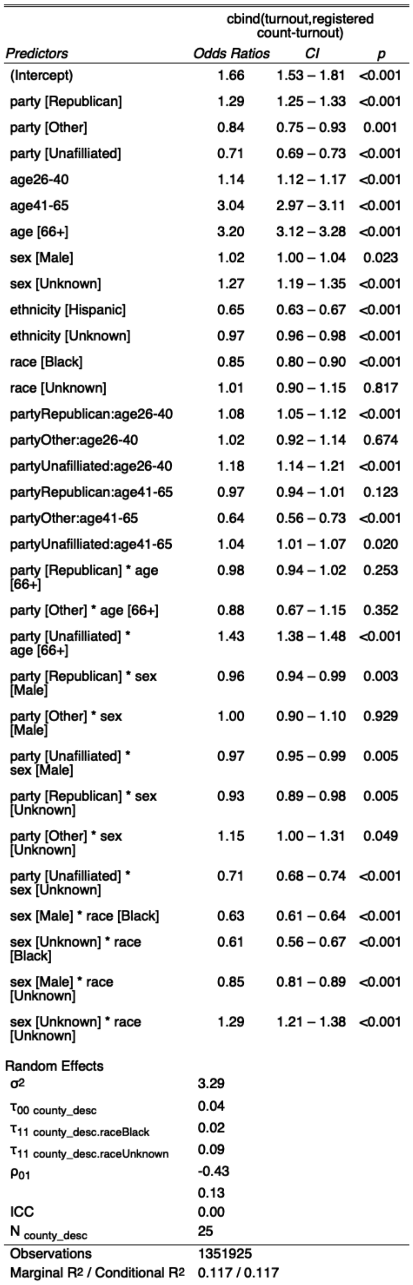
After fitting and interpreting our model, we summarized our findings and answered the inferential questions to the best of our ability. First, the probability of voting differs by demographic groups. The clearest trends is that age increases the predicted probability of voting, especially once someone is greater than 41 years old. Continually, Hispanic voters are less likely to vote than non-Hispanic voters. Republicans are the most likely political party to vote, compared to Democrats, other parties, and unaffiliated voters. We were also able to definitively state that the odds of voting differ by county, with a wide potential range of turnout.

The final two questions of interest deal with the plots included in Appendix 3. Interestingly, Male and Female voters have nearly identical turnout when split up by party. Continually, when investigating the interaction between political party and age, we found that the trend is again very similar by party—generally—turnout increases with age. However, there are some subtle differences. For example, for Republican and unaffiliated voters, the difference between the 18-25 and 26-40 age groups is significant, while in other groups the relationship is not (but trending in the same direction).

Limitations

While the results from this analysis are interesting and provide valuable insights into the effect of location and demographic groups on voter turnout, there are major limitations to consider. First of all, as mentioned during the data preparation section, there were some observations that had to be modified and some were dropped. While the overall number was small, this still has the potential to skew the results. Secondly, there are a significant number of observations where demographic features are unknown. If the true demographic characteristics for these voters were known, it's possible that this could shift our model, either erasing or exacerbating demonstrated effects. Finally, and perhaps most critically, this analysis only covers registered voters, not eligible voters. This casts a specific lens on the interpretation of this analysis. For example, while we demonstrate that registered Hispanic people are less likely to vote are more than registered non-Hispanic people, if eligible Hispanic voters are much more likely to register, than the turnout rate of eligible Hispanic voters could actually be higher than non-Hispanic voters. All in all, the results of this analysis are certainly valuable, but also deserve careful interpretation given these specific limitations.

Appendix 1: Final Model Results



	cbind(turnout,registered count-turnout)		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.66	1.53 – 1.81	<0.001
party [Republican]	1.29	1.25 – 1.33	<0.001
party [Other]	0.84	0.75 – 0.93	0.001
party [Unafilliated]	0.71	0.69 – 0.73	<0.001
age26-40	1.14	1.12 – 1.17	<0.001
age41-65	3.04	2.97 – 3.11	<0.001
age [66+]	3.20	3.12 – 3.28	<0.001
sex [Male]	1.02	1.00 – 1.04	0.023
sex [Unknown]	1.27	1.19 – 1.35	<0.001
ethnicity [Hispanic]	0.65	0.63 – 0.67	<0.001
ethnicity [Unknown]	0.97	0.96 – 0.98	<0.001
race [Black]	0.85	0.80 – 0.90	<0.001
race [Unknown]	1.01	0.90 – 1.15	0.817
partyRepublican:age26-40	1.08	1.05 – 1.12	<0.001
partyOther:age26-40	1.02	0.92 – 1.14	0.674
partyUnafilliated:age26-40	1.18	1.14 – 1.21	<0.001
partyRepublican:age41-65	0.97	0.94 – 1.01	0.123
partyOther:age41-65	0.64	0.56 – 0.73	<0.001
partyUnafilliated:age41-65	1.04	1.01 – 1.07	0.020
party [Republican] * age [66+]	0.98	0.94 – 1.02	0.253
party [Other] * age [66+]	0.88	0.67 – 1.15	0.352
party [Unafilliated] * age [66+]	1.43	1.38 – 1.48	<0.001
party [Republican] * sex [Male]	0.96	0.94 – 0.99	0.003
party [Other] * sex [Male]	1.00	0.90 – 1.10	0.929
party [Unafilliated] * sex [Male]	0.97	0.95 – 0.99	0.005
party [Republican] * sex [Unknown]	0.93	0.89 – 0.98	0.005
party [Other] * sex [Unknown]	1.15	1.00 – 1.31	0.049
party [Unafilliated] * sex [Unknown]	0.71	0.68 – 0.74	<0.001
sex [Male] * race [Black]	0.63	0.61 – 0.64	<0.001
sex [Unknown] * race [Black]	0.61	0.56 – 0.67	<0.001
sex [Male] * race [Unknown]	0.85	0.81 – 0.89	<0.001
sex [Unknown] * race [Unknown]	1.29	1.21 – 1.38	<0.001
Random Effects			
σ^2	3.29		
$\tau_{00\ county\_desc}$	0.04		
$\tau_{11\ county\_desc.raceBlack}$	0.02		
$\tau_{11\ county\_desc.raceUnknown}$	0.09		
ρ_{01}	-0.43		
	0.13		
ICC	0.00		
N_{county_desc}	25		
Observations	1351925		
Marginal R^2 / Conditional R^2	0.117 / 0.117		

Figure 3: Final Model Output

Appendix 2: Random Effect Plots

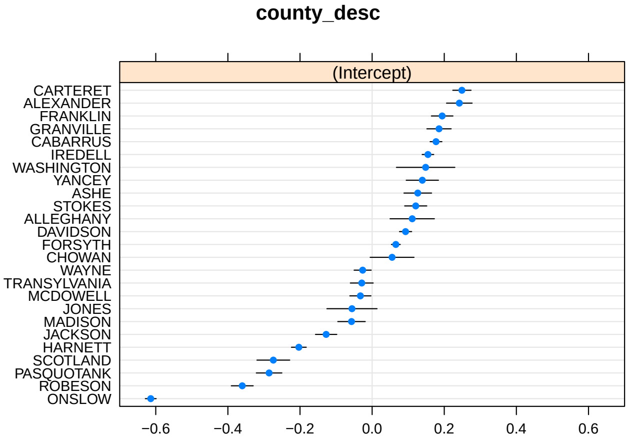


Figure 4: Dot Plot of Random Intercepts

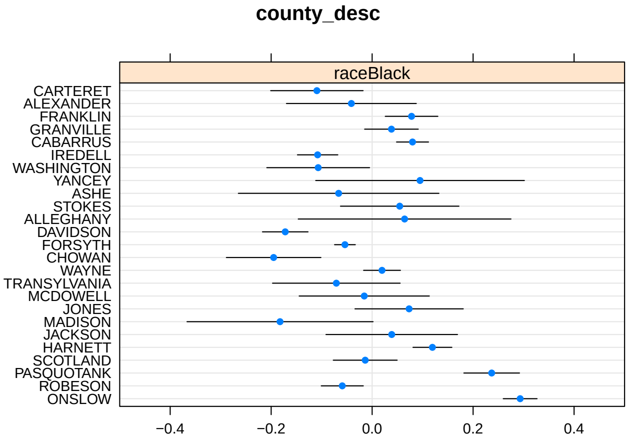


Figure 5: Dot Plot of Random Slopes: Black

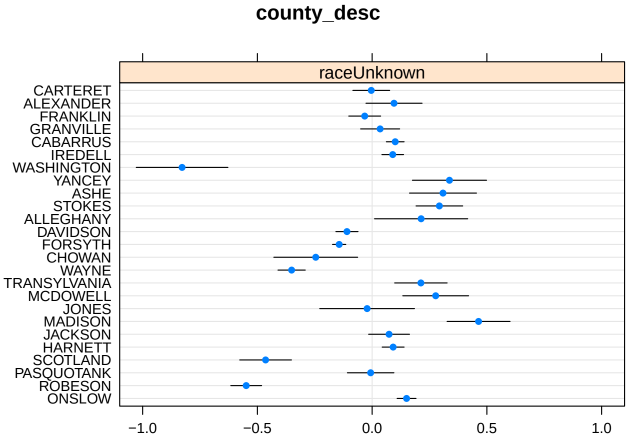


Figure 6: Dot Plot of Random Slopes: Unknown

Appendix 3: Interactions and Predicted Probability

```
## [[1]]
```

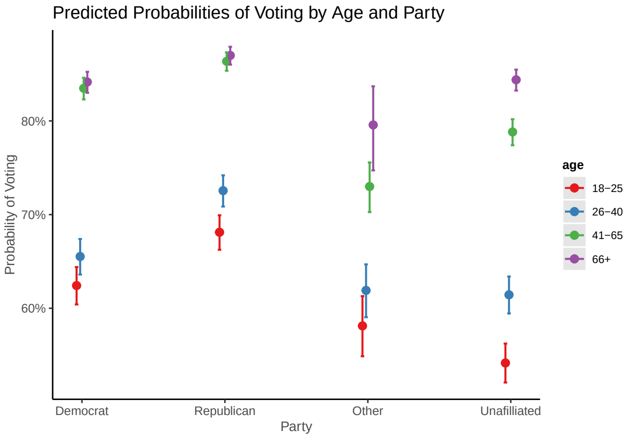


Figure 7: Interaction of Age and Party

```
## [[1]]
```

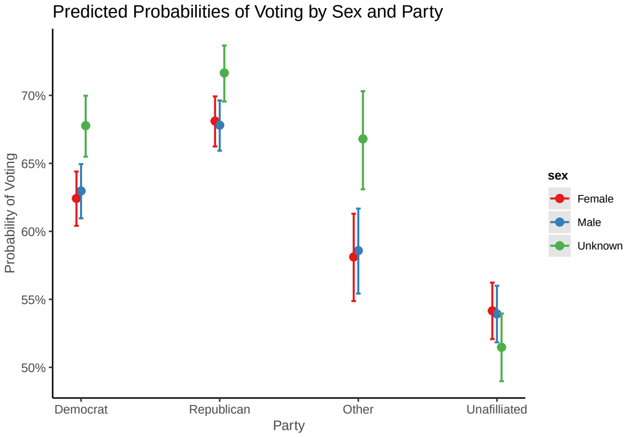


Figure 8: Interatcion of Sex and Party