Case Study 2

Emily Gentles (Writer) Weiyi Liu (Coordinator and Checker) Jack McCarthy (Presenter)
Qinzhe Wang (Programer)

02 December, 2021

Introduction

Voter turnout is a huge topic of interest for both pollsters and candidates, especially around presidential elections and with a particular focus on swing states, such as North Carolina. In this case study we will use North Carolina registration and turnout data, by county, for the 2020 election year. We wish to investigate what demographic factors such as age, race, and party affect turnout and whether turnout differs by county.

Data Cleaning

Note that we renamed the `history_stats_20201103` data as votes and the `voter_stats_20201103` data as registers. Within the votes and registers data we aggregated the total voters variable by taking the sum, ignoring NA values, and then left joined votes to registers on county and all the demographic features. We next replaced any NA values in total votes with 0. While these values are unlikely to be exactly 0, we chose to do this in order to preserve the information that was present as the alternative would be to throw the entire row away. After this we dropped all of the rows containing NA values in any of the other columns. Since total voters is one of the main variables we care about it made sense to preserve as many rows as possible but if we replaced every NA value in the data with 0 that may potentially introduce a lot of error into the data, as the values are unlikely to be exactly 0, so we instead dropped all of the rows containing NA values, leaving us with 51,883 rows. For some subgroups that have more actual voters than registered voters, we replaced the number of actual voters by the number of registers to ensure the turn out rate to be valid. We next randomly sampled 30 counties to utilize in our analysis, leaving our final data with 16,439 rows and, after de-aggregation, 1,908,907 individuals.

EDA

Below we see that most counties have a turnout rate over 70% with the average of all counties being 73%. The lowest turnout rate of 61% is from Robeson county. We also see that two counties, Forsyth and Cabarrus, have many more registered voters than the other sampled counties.

When looking at race, age, gender, ethnicity, and party individually (plots in appendix), we found that white people had the highest turnout and number of voters registered while multiracial people had the lowest turnout rate, along with people of race 'other', and the smallest number of voters registered. We also found the race 'P' has an extremely small sample size (115) compared with other race groups. Therefore, we decided to drop this group for computational efficiency.

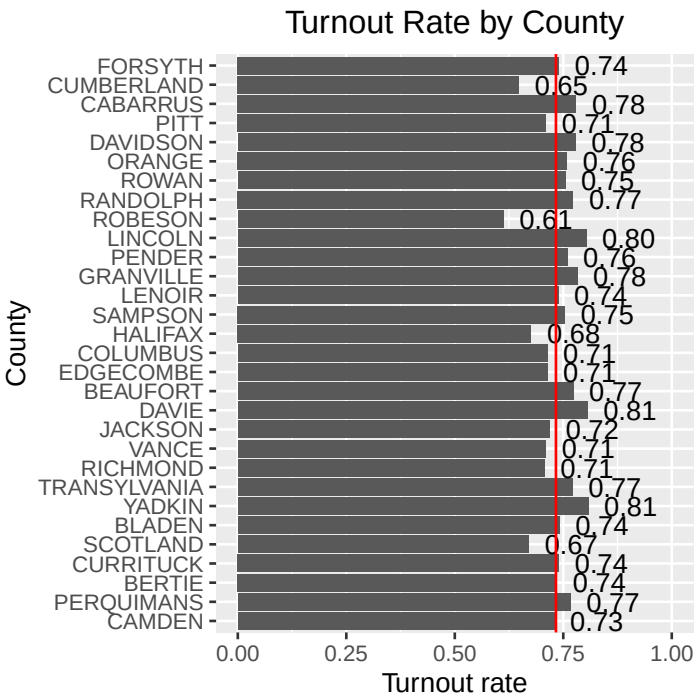
The Republican party had the highest turnout while the Libertarian party had the lowest turnout. Both the Green party and the Constitution party had the lowest number of voters registered, less than 1500 each, so we combined these categories to provide our model with computational stability.

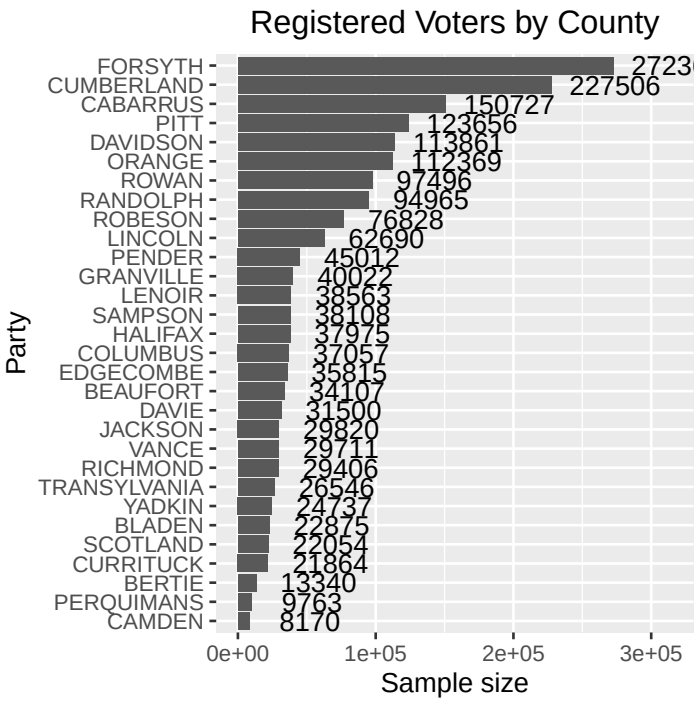
When looking at ethnicities, people of Hispanic/Latinx ethnicity had both the lowest number of voters registered and turnout, a 20 point difference compared to people who weren't of Hispanic/Latinx ethnicity.

Turnout rate varied very little by gender although people of unknown gender had the smallest number of voters registered.

Looking at age we see that the age group of people 18-25 had both the lowest number of voters registered and turnout rate, a difference of 27 points when compared to those 66 and older. However, the age group of people 41-65 had the largest number of voters registered, nearly triple that of the 18-25 age group.

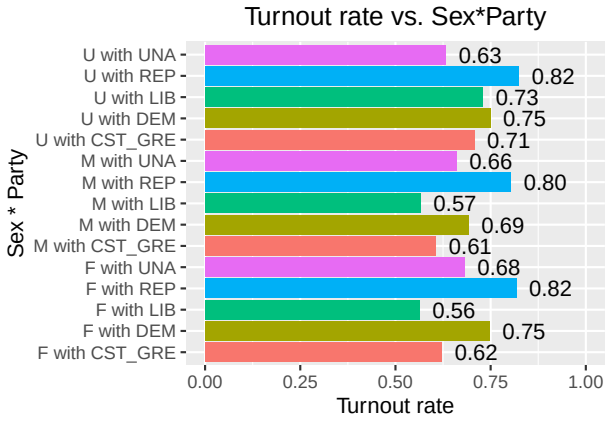
Turnout Rate by County

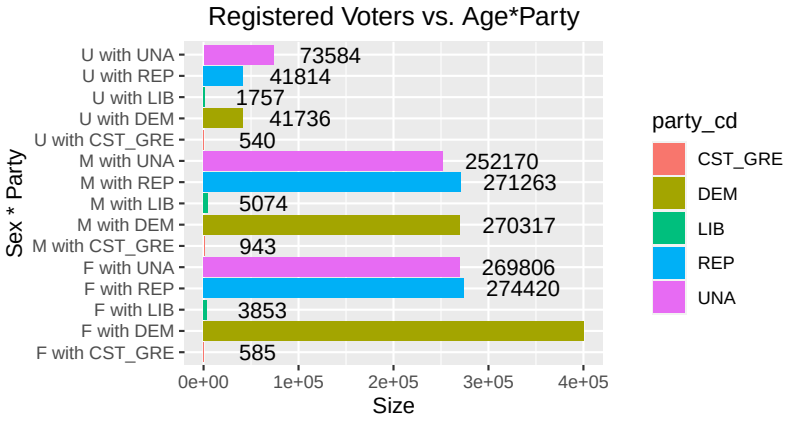




Gender & Party

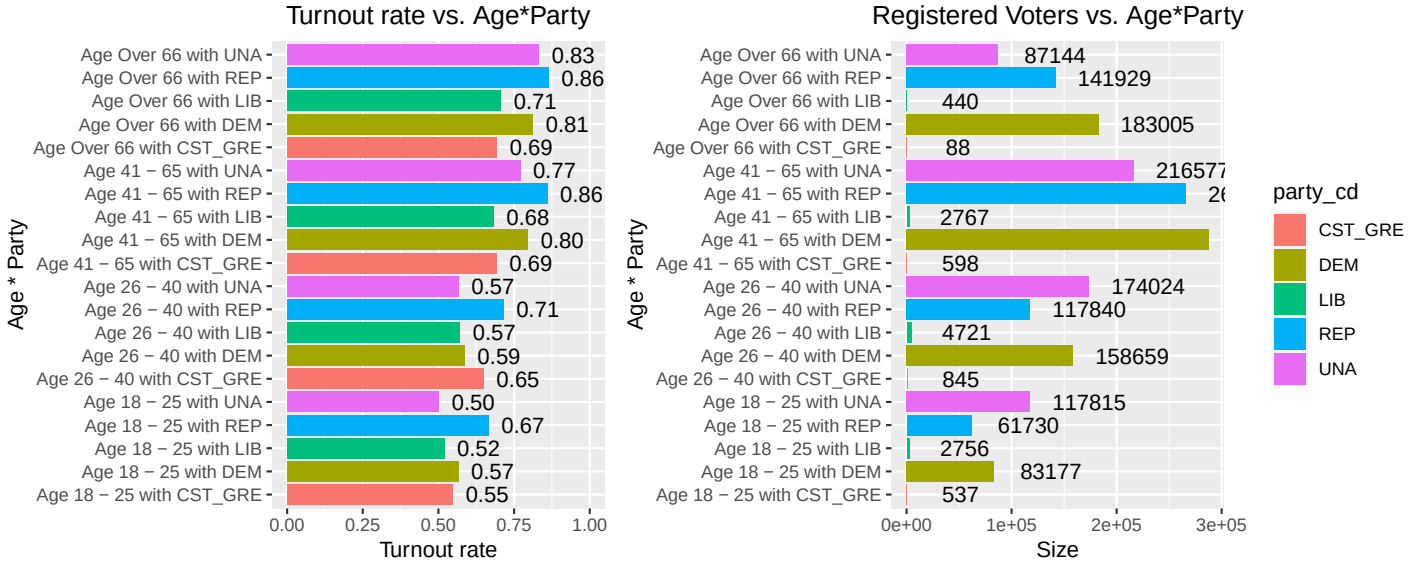
Next we move on to EDA related to interactions. We're interested in any potential differences in turnout rates for females and males with various party affiliations. In the graph below we see that females have higher turnout rates than males for each party. People with unknown genders, however, often have the highest turnout rate. For parties with substantial turnout rates, the Republican and Democrat parties, as well as those unaffiliated with a party, females had a larger number of voters registered than males and people with unknown genders had a much smaller number of voters registered. Note that, even after combining the Green and Constitutional parties, the number of registered voters for these parties is still quite small, relative to the other parties.





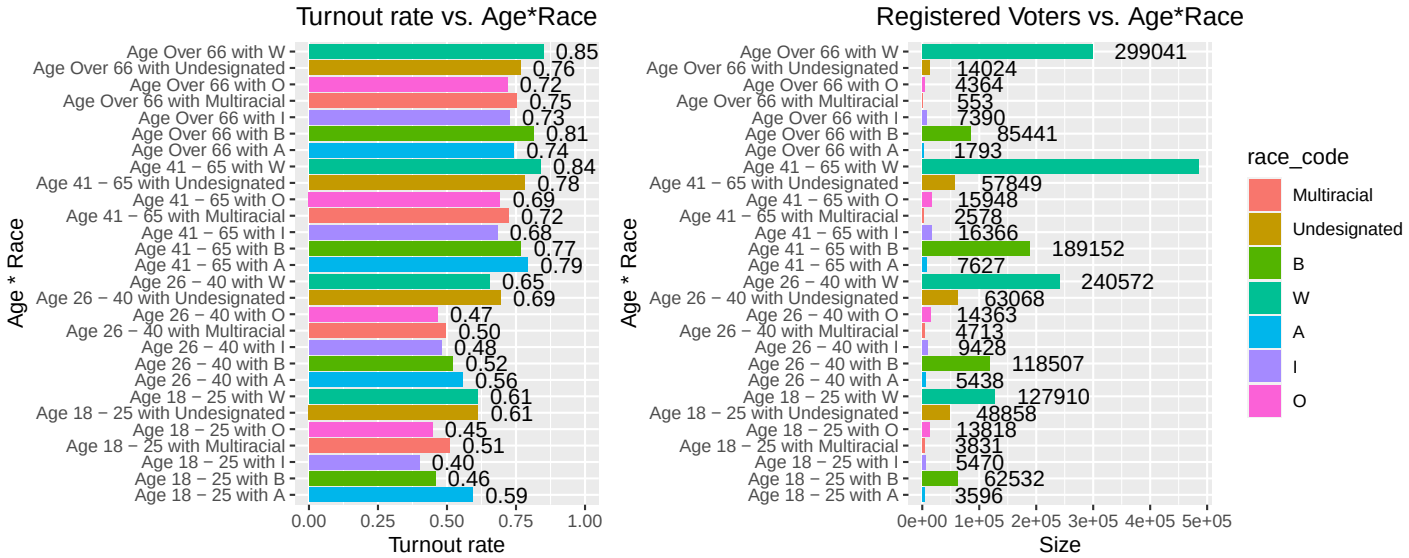
Age & Party

We are also interested in the potential difference between age groups in various political parties. From the graphs below we see that the age group of 66 and older have the highest turnout rate for every political party and the age group 18-25 has the lowest. In general, it seems that, across parties, as age increases, the turnout rate also increases. As noted previously, the age group 41-65 has the largest number of registered voters for both of the main parties, Republican and Democrat, and for unaffiliated voters.



Age & Race

We are also interested in a potential relationship between age and race. In the graphs below we see a general trend that, across races, as the age increase so too does the turnout rate. We also see that, within each age group, white people make up the majority of registered voters.



Model

With our exploration of the data complete, we now begin our model building. From the research questions we have an idea of what must be included in the model; this includes the main effect for party, sex, and age, as well as random intercepts by county. This will be our base model. From the research questions we also know

that we likely need an interaction term between age and party as well as sex and party. Additionally, since our dependent variable, turnout rate, is between 0 and 1, we know we should use a logistic mixed effects model. Below we see the output from the forward selection. We see that the last model, with interactions between sex and party, age and party, as well as age and race, has the lowest BIC and AIC so this will be our final model.

$$
\begin{aligned}
log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) &= \beta_0 + \sum_k \beta_{1k}\mathbb{1}[P_{ij}=k] + \sum_k \beta_{2k}\mathbb{1}[R_{ij}=k] + \sum_k \beta_{3k}\mathbb{1}[E_{ij}=k] + \sum_k \beta_{4k}\mathbb{1}[S_{ij}=k] \\
&+ \sum_k \beta_{5k}\mathbb{1}[A_{ij}=k] + \sum_{k,l} \beta_{6k}\mathbb{1}[S_{ij}=k]\mathbb{1}[P_{ij}=k] + \sum_{k,l} \beta_{7k}\mathbb{1}[A_{ij}=k]\mathbb{1}[P_{ij}=k] \\
&+ \sum_{k,l} \beta_{8k}\mathbb{1}[A_{ij}=k]\mathbb{1}[R_{ij}=k] + b_{0j}
\end{aligned}
$$

where $\pi_{ij} = Pr(y_{ij}=1)$ and $b_{0j} \overset{\text{iid}}{\sim} N(0,\sigma^2)$ and

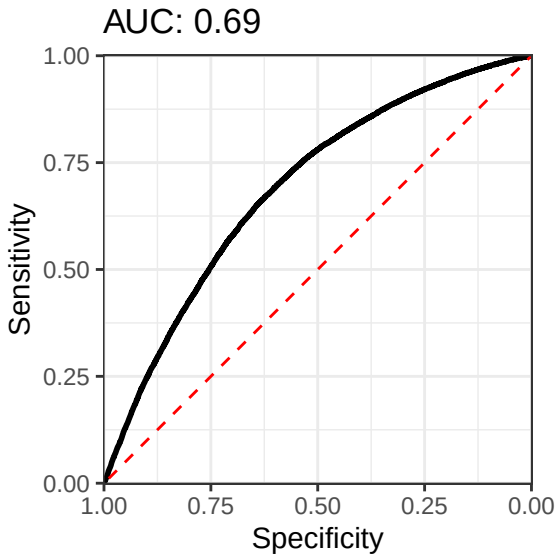
- j indexes county
- i indexes observation
- P_{ij} is the Party code
- R_{ij} is the Race code
- E_{ij} is the Ethnic code
- S_{ij} is the sex code
- A_{ij} is the Age group

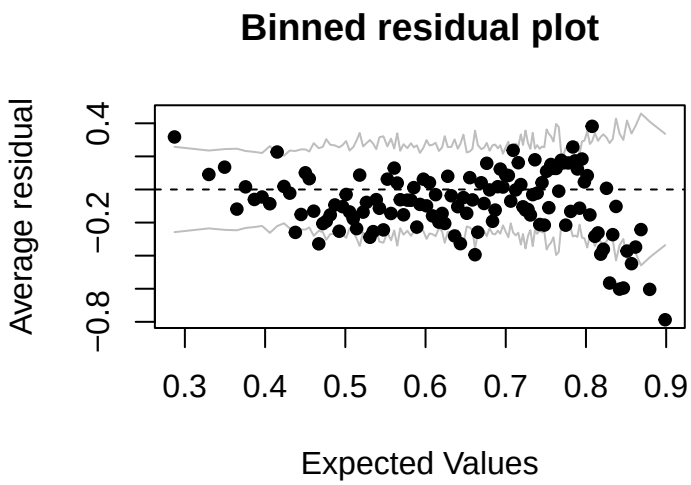
Table 1: Forward model selection

Model	LRT.p.value	AIC	BIC
Base model		96904.01	96988.73
Add race	0	88191.94	88322.86
Add race and ethnic	0	86526.87	86673.20
Without intercept	0	99839.91	99955.43
Add the interaction of sex:party, and age:party	0	83825.28	84125.65
Add the interaction of sex:party, age:party, and age:race	0	79688.45	80127.45

Model Assessment

From our EDA it can reasonably be assumed that the linearity condition is met for our model and that multicollinearity is not an issue. As we see in the binned residual plot below, there may cause for some concern as some points with expected values between 0.8 and 0.9 fall out of the range of the confidence bands. However, relative to the amount of data we have, these 8 points, in addition to 7 other scattered points outside the confidence bands, don't point to our model being incorrect. We see that the model has an AUC around 0.7 which is satisfactory and much better than a coin flip. Our model seems to more accurately identify positives than negatives as well.





Conclusion

Now, we can answer the research questions based on the model results.

Q1: How did different demographic subgroups vote in the 2020 general elections?

Party

Notice that our final model incorporated the interaction terms `sex:party` and `age:party`, we should interpret the effects of party at the baseline levels of the other two terms (female and age 18-25).

Table 2: Estiamtes and 95% CI of Fixed Effects for Party

	Estimate	Std. Error	95% lower	95% upper
party_cdDEM	0.3477	0.1128	0.1267	0.5688
party_cdLIB	-0.2886	0.1217	-0.5272	-0.0500
party_cdREP	0.3248	0.1129	0.1034	0.5461
party_cdUNA	-0.1791	0.1127	-0.4000	0.0417

Holding all other predictors unchanged:

an 18-25 year old female DEM voter has $e^{0.3477} = 1.4158$ times (a 41% increase) the odds of voting

an 18-25 year old female LIB voter has $e^{-0.2886} = 0.7493$ times (a 25% decrease) the odds of voting

an 18-25 year old female REP voter has $e^{0.3248} = 1.3838$ times (a 38% increase) the odds of voting

an 18-25 year old female UNA voter has $e^{-0.1791} = 0.8360$ times (a 17% decrease) the odds of voting (this effect is not significant since the 95% CI covers 0)

compared with an 18-25 year old female GRE/CST voter.

Race

Table 3: Estiamtes and 95% CI of Fixed Effects for Race

	Estimate	Std. Error	95% lower	95% upper
race_codeUndesignated	0.3584	0.0350	0.2899	0.4270
race_codeB	-0.3441	0.0341	-0.4110	-0.2772
race_codeW	0.2318	0.0337	0.1658	0.2979
race_codeA	0.1517	0.0477	0.0581	0.2453
race_codeI	-0.2409	0.0437	-0.3264	-0.1553
race_codeO	-0.1098	0.0374	-0.1832	-0.0364

Holding all other predictors unchanged:

an 18-25 year old voter with undesignated race has $e^{0.3584} = 1.4310$ times (a 43% increase) the odds of voting

an 18-25 year old Black voter has $e^{-0.3441} = 0.7089$ times (a 29% decrease) the odds of voting

an 18-25 year old White voter has $e^{0.2318} = 1.2609$ times (a 26% increase) the odds of voting

an 18-25 year old Asian voter has $e^{0.1517} = 1.1638$ times (a 16% increase) the odds of voting

an 18-25 year old Indian American voter has $e^{-0.2409} = 0.7859$ times (a 21% decrease) the odds of voting

an 18-25 year old voter in other race has $e^{-0.1098} = 0.8960$ times (a 10% decrease) the odds of voting

compared with an 18-25 year-old multiracial voter.

Ethnic

Table 4: Estiamtes and 95% CI of Fixed Effects for Ethnic

	Estimate	Std. Error	95% lower	95% upper
ethnic_codeNL	0.4438	0.0107	0.4228	0.4647
ethnic_codeUN	0.3618	0.0109	0.3404	0.3833

Holding all other predictors unchanged:

a non Hispanic/Latino voter has $e^{0.4438} = 1.5586$ times (a 56% increase) the odds of voting

a voter with unkown ethnic has $e^{0.3618} = 1.4359$ times (a 44% increase) the odds of voting

compared with a Hispanic/Latino voter.

Sex

Table 5: Estiamtes and 95% CI of Fixed Effects for Sex

	Estimate	Std. Error	95% lower	95% upper
sex_codeM	-0.0728	0.1111	-0.2907	0.1450
sex_codeU	0.1790	0.1303	-0.0763	0.4343

Holding all other predictors unchanged:

a male CST/GRE voter has $e^{-0.0728} = 0.9298$ times (a 7% decrease) the odds of voting

a CST/GRE voter with unknown gender has $e^{0.1790} = 1.1960$ times (a 20% increase) the odds of voting compared with a female CST/GRE voter.

However, neither of these coefficients are significant since none of the 95% CIs cover 0.

Age

Table 6: Estiamtes and 95% CI of Fixed Effects for Age

	Estimate	Std. Error	95% lower	95% upper
ageAge 26 - 40	0.0492	0.1237	-0.1933	0.2917
ageAge 41 - 65	0.3871	0.1379	0.1168	0.6575
ageAge Over 66	0.4609	0.2707	-0.0698	0.9915

Holding all other predictors unchanged:

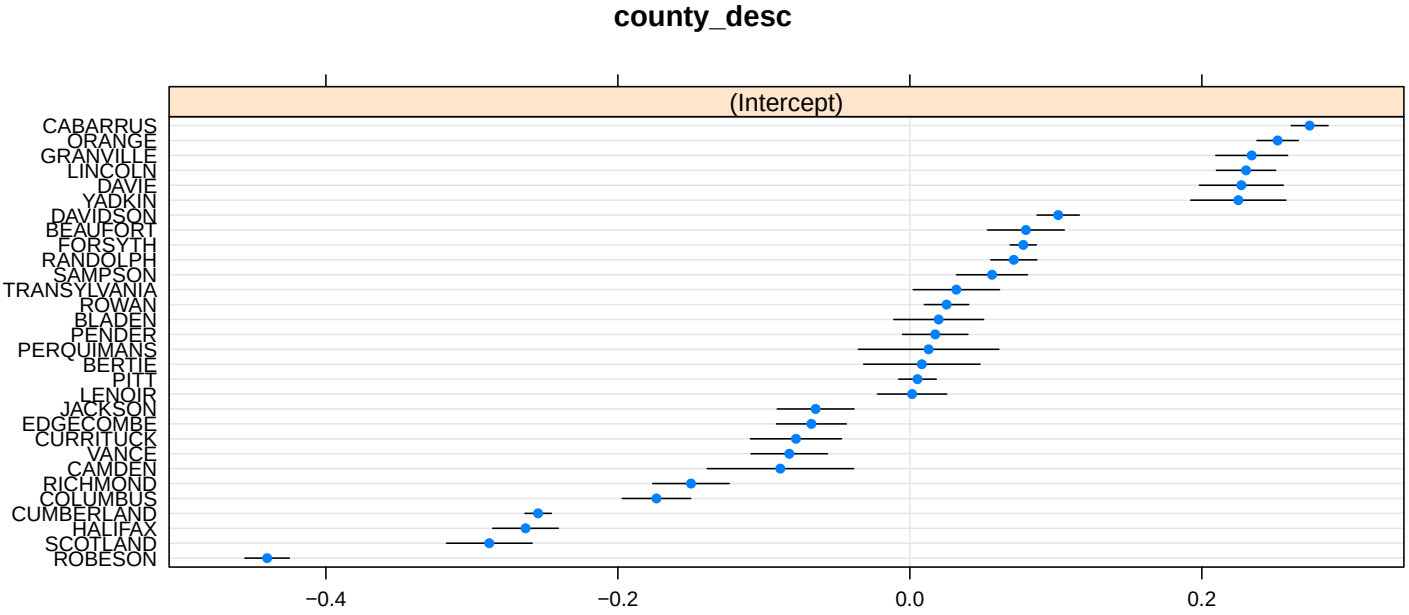
a 26-40 year old multiracial CST/GRE voter has $e^{0.0492} = 1.0504$ times (a 5% increase) the odds of voting (this effect is not significant since the 95% CI covers 0)

a 41-65 year old multiracial CST/GRE voter has $e^{0.3871} = 1.4727$ times (a 47% increase) the odds of voting

an over 66 year old multiracial CST/GRE voter has $e^{0.4609} = 1.5855$ times (a 59% increase) the odds of voting (this effect is not significant since the 95% CI covers 0)

compared with an 18-25 year old multiracial CST/GRE voter.

Q2: Did the overall probability or odds of voting differ by county in 2020? Which counties differ the most from other counties?



The above dotplot displays the random effects for each county. Since we observe clear heterogeneity across counties, we can conclude the overall probability of voting differed by county in 2020. Robeson County has the smallest estimate, while Cabarrus has the largest estimate. Furthermore, Robeson differs the most from all other counties. A table with the estimates and corresponding confidence intervals is included in the appendix.

Q3: How did the turnout rates differ between females and males for the different party affiliations?

Table 7: Estiamtes and 95% CI of Fixed Effects for Party:Sex

	Estimate	Std. Error	95% lower	95% upper
party_cdDEM:sex_codeM	-0.2085	0.1113	-0.4267	0.0096
party_cdLIB:sex_codeM	0.0302	0.1196	-0.2043	0.2647
party_cdREP:sex_codeM	0.0227	0.1114	-0.1956	0.2410
party_cdUNA:sex_codeM	-0.0401	0.1113	-0.2583	0.1780
party_cdDEM:sex_codeU	-0.0766	0.1306	-0.3324	0.1793
party_cdLIB:sex_codeU	0.4754	0.1448	0.1916	0.7592
party_cdREP:sex_codeU	0.1535	0.1307	-0.1028	0.4097
party_cdUNA:sex_codeU	-0.2499	0.1303	-0.5052	0.0055

From the table above, only the fixed effects for `party_cdLIB:sex_codeU` is significant (the 95% CI does not cover 0). For simplicity, we only interpret this fixed effect.

Q4: How did the turnout rates differ between age groups for the different party affiliations?

Table 8: Estiamtes and 95% CI of Fixed Effects for Party:Age

	Estimate	Std. Error	95% lower	95% upper
party_cdDEM:ageAge 26 - 40	-0.2391	0.1160	-0.4664	-0.0118
party_cdLIB:ageAge 26 - 40	-0.1267	0.1257	-0.3731	0.1196
party_cdREP:ageAge 26 - 40	-0.0635	0.1161	-0.2910	0.1641
party_cdUNA:ageAge 26 - 40	-0.0556	0.1158	-0.2825	0.1714
party_cdDEM:ageAge 41 - 65	0.4162	0.1267	0.1679	0.6646
party_cdLIB:ageAge 41 - 65	0.1036	0.1386	-0.1682	0.3753
party_cdREP:ageAge 41 - 65	0.5018	0.1268	0.2533	0.7504
party_cdUNA:ageAge 41 - 65	0.5563	0.1266	0.3082	0.8043
party_cdDEM:ageAge Over 66	0.2917	0.2504	-0.1991	0.7825
party_cdLIB:ageAge Over 66	0.2045	0.2745	-0.3334	0.7425
party_cdREP:ageAge Over 66	0.5102	0.2505	0.0192	1.0011
party_cdUNA:ageAge Over 66	0.8719	0.2504	0.3811	1.3627

Strengths and Limitations

Limitations of our model include the fact that during data cleaning we dropped rows that contained missing values. While we are working with almost 2 million observations, we do have quite a few small demographic groups, especially for our interaction terms. A possible alternative would have been to impute values for the missing data so that we could keep as many rows as possible. Obviously imputation comes with it's own set of assumptions and limitations.

Other limitations ?

A major strength of our model is the inclusion of random intercepts by county which allows us to understand the difference in turnout rate between each county.

Other strengths ?