Case Study 2

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

10/28/2021

Introduction

Voter turnout is a huge topic of interest for both pollsters and candidates, especially around presidential elections and with 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.

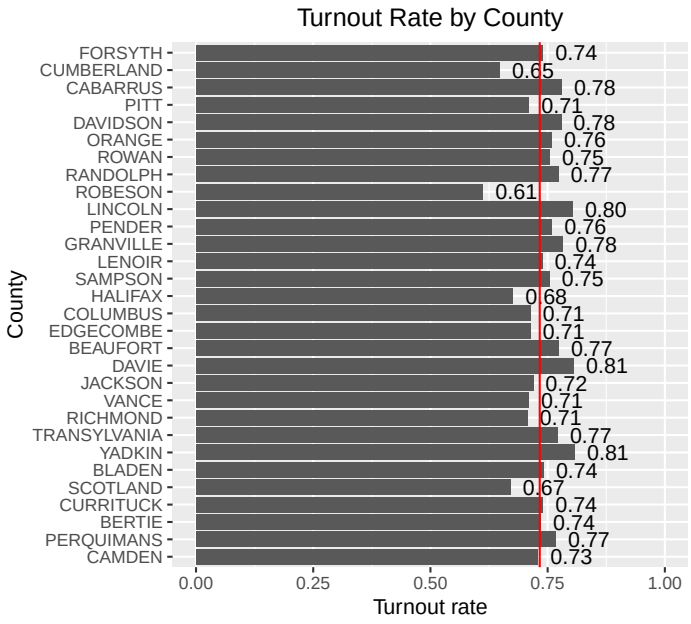
Data Cleaning

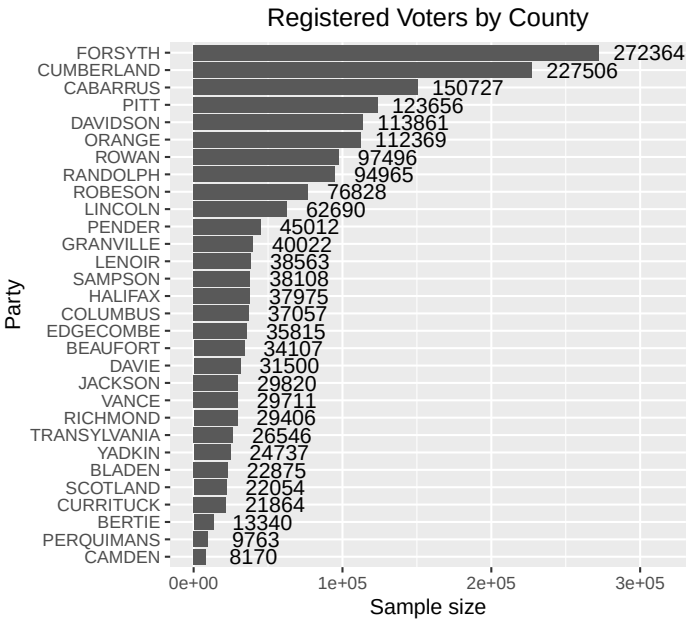
Note that we renamed the history stats 2020-11-03 data as votes and the voterstats 2020-11-03 data as registers. Within the votes and registers data we aggregated the total voters variable by taking the sum, ignoring NA values, and then left joined votes to registers. We next replaced any NA values in 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. We next randomly sampled 30 counties to utilize in our analysis, leaving our final data with 16,439 rows and, after deaggregation, 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. 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.

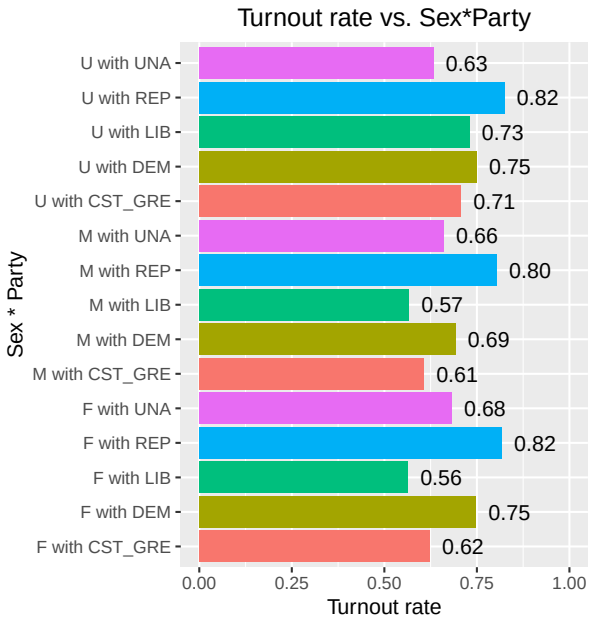
Turnout Rate

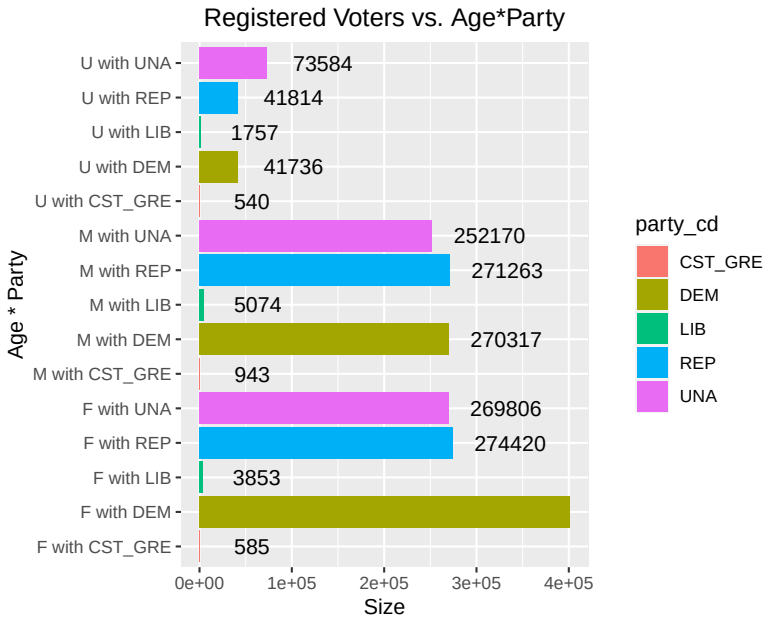




Gender & Party

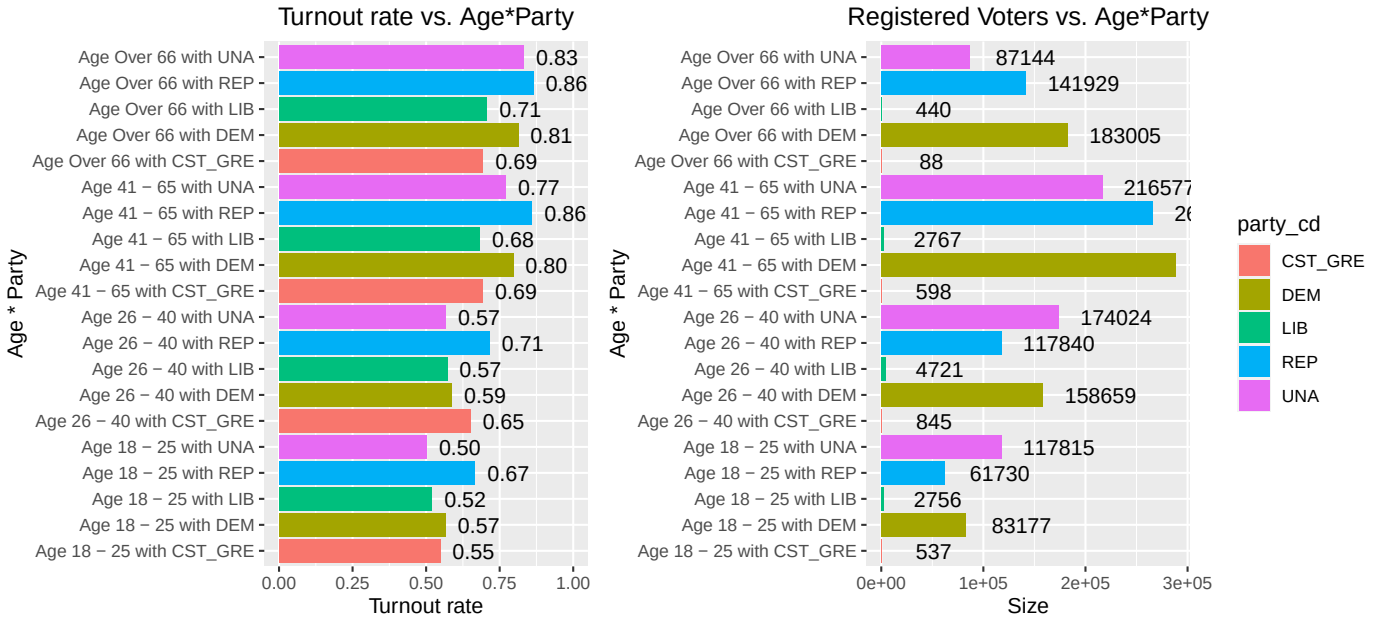
Next we move on to EDA related to interactions. We're interested in any potential differences in turnout rate for females and males with various party affiliations. In the graph below we see that females have higher turnout rates than males for each party. People with unknown genders, however, often 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.





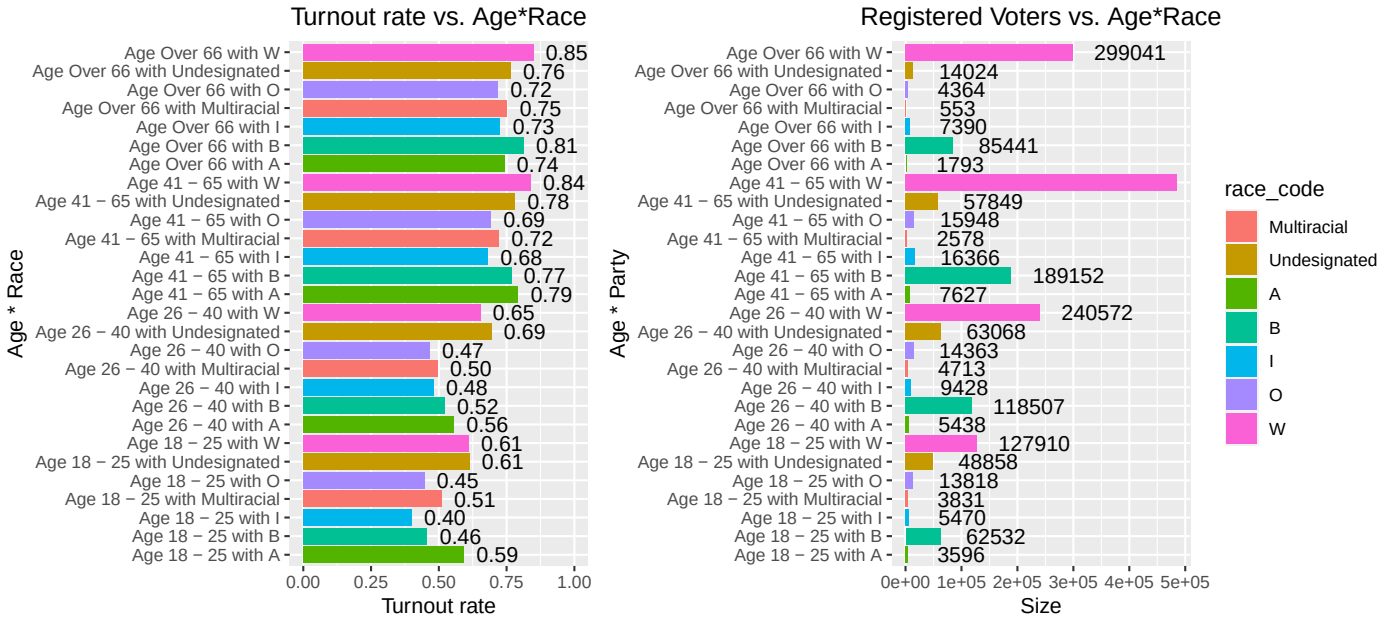
Age & Party

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 has 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, 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 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 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.

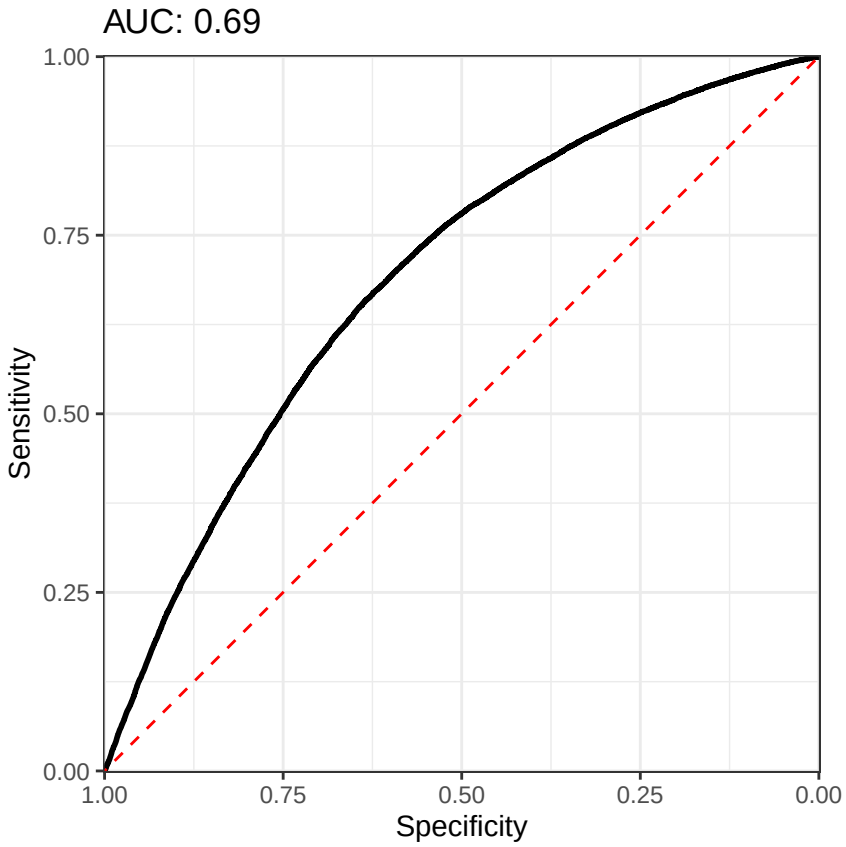
$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \sum_k \beta_{1k}\mathbb{1}[P_{ij}=k] + \sum_k \beta_{2k}\mathbb{1}[R_{ij}=k] + \sum_k \beta_{3k}\mathbb{1}[E_{ij}=k] + \sum_k \beta_{4k}\mathbb{1}[S_{ij}=k] + \sum_k \beta_{5k}\mathbb{1}[A_{ij}=k]$$
$$+ \sum_{k,l} \beta_{6k}\mathbb{1}[S_{ij}=k]\mathbb{1}[P_{ij}=k]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

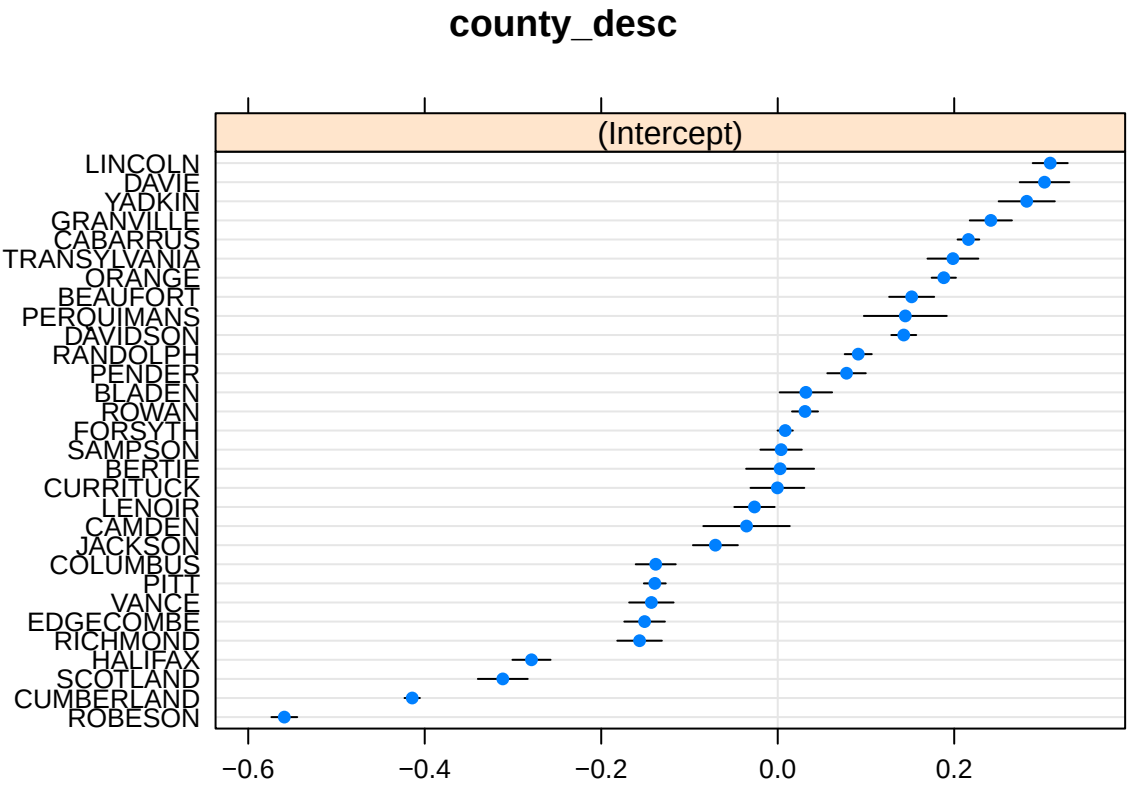
where $\pi_{ij} = Pr(y_{ij}=1)$ and $b_{0j} \overset{\text{iid}}{\sim} N(0,\sigma^2)$ and

- 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 and party_cd, and age and party	0	83825.28	84125.65
Add the interaction of sex and party_cd, age and party, and age and race	0	79688.45	80127.45


AUC: 0.69
1.00
0.75
0.50
0.25
0.00
Sensitivity
1.00
0.75
0.50
0.25
0.00
Specificity


county_desc
(Intercept)
LINCOLN
DAVIE
YADKIN
GRANVILLE
CABARRUS
TRANSYLVANIA
ORANGE
BEAUFORT
PERQUIMANS
DAVIDSON
RANDOLPH
PENDER
BLADEN
ROWAN
FORSYTH
SAMPSON
BERTIE
CURRITUCK
LENOIR
CAMDEN
JACKSON
COLUMBUS
PITT
VANCE
EDGECOMBE
RICHMOND
HALIFAX
SCOTLAND
CUMBERLAND
ROBESON
−0.6
−0.4
−0.2
0.0
0.2

	Estimate	95% CI
(Intercept)	3.232	[1.195, 5.268]
party_cd5	0.412	[-1.614, 2.437]
party_cd6	1.369	[-0.603, 3.342]
party_cd7	0.681	[-1.29, 2.652]
party_cdCST_GRE	0.533	[-1.603, 2.669]
party_cd3:sex_codeM	0.758	[-1.213, 2.729]
party_cd5:sex_codeM	1.006	[-1.041, 3.053]
party_cd6:sex_codeM	0.915	[-1.059, 2.888]
party_cd7:sex_codeM	0.903	[-1.069, 2.875]
party_cdCST_GRE:sex_codeM	0.953	[-1.233, 3.138]
party_cd3:sex_codeU	0.979	[-1.005, 2.963]
party_cd5:sex_codeU	2.071	[-0.017, 4.159]
party_cd6:sex_codeU	1.049	[-0.939, 3.036]
party_cd7:sex_codeU	0.804	[-1.174, 2.781]
party_cdCST_GRE:sex_codeU	1.403	[-0.825, 3.63]

County	E.Estimate	E.Lower	E.Upper
BEAUFORT	1.164	-0.862	3.190
BERTIE	1.003	-1.037	3.042
BLADEN	1.032	-0.998	3.063
CABARRUS	1.241	-0.771	3.254
CAMDEN	0.965	-1.085	3.016
COLUMBUS	0.871	-1.152	2.894
CUMBERLAND	0.661	-1.348	2.670
CURRITUCK	1.000	-1.032	3.031
DAVIDSON	1.154	-0.861	3.168
DAVIE	1.353	-0.676	3.382
EDGECOMBE	0.860	-1.164	2.884
FORSYTH	1.009	-1.000	3.017
GRANVILLE	1.273	-0.751	3.298
HALIFAX	0.757	-1.266	2.779
JACKSON	0.932	-1.094	2.958
LENOIR	0.974	-1.049	2.997
LINCOLN	1.362	-0.659	3.382
ORANGE	1.207	-0.807	3.221
PENDER	1.081	-0.941	3.103
PERQUIMANS	1.156	-0.893	3.204
PITT	0.870	-1.143	2.883
RANDOLPH	1.096	-0.920	3.111
RICHMOND	0.855	-1.171	2.881
ROBESON	0.572	-1.443	2.587
ROWAN	1.031	-0.984	3.046
SAMPSON	1.004	-1.020	3.028
SCOTLAND	0.732	-1.297	2.761
TRANSYLVANIA	1.220	-0.810	3.249
VANCE	0.867	-1.159	2.892
YADKIN	1.326	-0.707	3.359

Interpretation

Our baseline is that of a female aligned with party ?? of mixed race and hispanic/latina ethnicity, in the age category 18-25. The odds of this person voting is .

Research Questions

- How did different demographic subgroups vote in the 2020 general elections? For example, how did the turnout for males compare to the turnout for females after controlling for other potential predictors?
- 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?