Interim Report: Who Votes in NC?

Alexander Bendeck, Lynn Fan, Cathy Lee, Alice Liao, Justina Zou

20 October 2020

Introduction

The United States is getting closer to the 2020 Congressional Elections on November 3, 2020. With the current polarizing political landscape, the congressional election outcomes are significant to determine the next stage of this country. As it becomes critical for statisticians to help build forecasting models to predict the election outcomes, we need to first understand the patterns of voter turnout. Voting has special importance in democratic systems, but only about half of the eligible U.S. citizenry votes, and there are real political consequences when voters differ systematically from nonvoters (Uhlaner et al.). There has been abundant literature proving that variation in voter turnout will have electoral consequences (Hansford, et al.), in a number of ways. First of all, the media conventional wisdom is that "higher turnout benefits Democrats," although there has been mixed evidence about this theory (Weinschenk, 2019). Second, literature has proven certain demographic factors to statistically significantly benefit one party over the other, in both congressional elections and presidential races. For example, gender, race and party registration could help profile voting patterns for congressional elections (Uhlaner et al.). Election prediction models need the baseline population of voters to predict the potential outcomes, and the demographic composition of voters will directly determine the forecasting results.

Among all states, North Carolina has been as a swing state in presidential and congressional elections for decades. In 2008, Obama won the state narrowly, but lost it narrowly after 4 years in 2012. Since 1996, the Republican statewide vote share in congressional elections has varied "from a low of 45% in 2008 to a high of 55% in 2014 (Perrin et al.)." It makes North Carolina an interesting battleground in which voter demographic changes could potentially lead to significant implications of election outcomes and "an excellent site for those interested in partisan voting trends (Perrin et al.)." This report seeks to understand the voter turnout of North Carolina for 2020 NC Congressional Elections, predicting who will vote in 2020.

Data Description

We are using public data provided by the NC State Board of Elections, which can all be accessed directly at the link https://dl.ncsbe.gov/list.html. The database contains voter history information for elections within the past 10 years in the ncvhis files, and all legally available voter specific information ion the ncvoter files. The ncvoter files contain point-in-time snapshot voter registration data. For privacy concerns, names, birth dates and drivers license are not included, but the two types of files could be matched by North Carolina identification (NCID) number. The database was last updated on September 9, 2020. While we understand that voters might register later than that as the voter registration deadline for North Carolina is October 9, 2020, we believe it is sufficient to represent the majority of NC potential voters.

From the ncvhis files, we only kept the voters that voted for the 2016 general election for our analysis. Studies have shown that presidential elections help mobilize voters, so voter turnout in presidential election years are significantly higher. In recent elections, voter turnout during presidential election years is around 60%, and only about 40% during midterm elections (FairVote.org). For North Carolina, voter turnout data in 2018 is also inappropriate to use because neither of North Carolina's U.S. senators nor the governor was up for reelection, further demotivated voters (Perrin et al.).

From the ncvoter files, we filtered demographic factors that are supported by existing literature to be significant in understanding voting patterns, including gender, race, party registration, and age (Kim et al.). We also have their county and congressional district information available.

Additionally, we found relevant literature proving the relationship between voter turnout and wage (Charles et al.), so we found county-level median household income data from Economic Research Service under United States Department of Agriculture (https://www.ers.usda.gov/data-products/county-level-data-sets)

Data Munging

Methods

After binding ncvhis files and ncvoter files by NCID and binding NC median household income by county, we started to process data for analysis. First of all, we identified those data points older than 116 years old and removed them as the oldest person in NC is 116 years old and anyone older should be wrong data points. Many data points are also missing congressional district information. We imputed the missing districts by matching the voter's registered county with congressional district. We removed the 4% of voters who reside in counties that span across more than one county. In the combined data set, there are party registrations for all parties, including The Libertarian Party and The Green Party. Because we are interested primarily only in the Republican Party and the Democratic Party and there are concerning class imbalance issues as the two parties take up the majority of registered voter population, we binded other parties as the third category `Other` for `Party`. Similarly, because of class imbalance, we binded the races other than White and African Americans as `Other` for `Race` as well. For those missing `Gender` information, we binded them with `Unspecified`.

Because we have eight million data points available, running models in a one-line-per-voter data set will be very computationally expensive. We instead decided to group data points by gender, race, party registration, county median income, and age, so that we can run models for the data set in a collapsed format. We divided (1) median county household income into four levels by the 25th, 50th, and 75th quantiles; (2) age into four levels for 18-29, 30-44, 45-59, and older than 60 years old, as it is a common way to analyze voter ages (McDonald, 2020); (3) gender into three categories, Female, Male and Other, and (4) race into three categories, Black, White, and Other.

Method

We will take a Bayesian approach to not only predict if a voter with a certain profile would vote, but also understand quantitatively how the geographic and demographic information of a registered voter is associated with his or her likelihood of actually casting a ballout. To model the binary outcome (vote vs not vote), we will first fit a simple logistic regression model with selected variables as a baseline for comparison. Then motivated by Y. Ghitza and A. Gelman's idea of poststratification (2013), we divide the population into mutually exclusive categories according to their demographic and geographic characteristics and fit a Bayesian model with group-level predictors as well as their interactions. With poststratification we can get average estimates for each of the subgroups.

The model takes the following form: *add latex?*

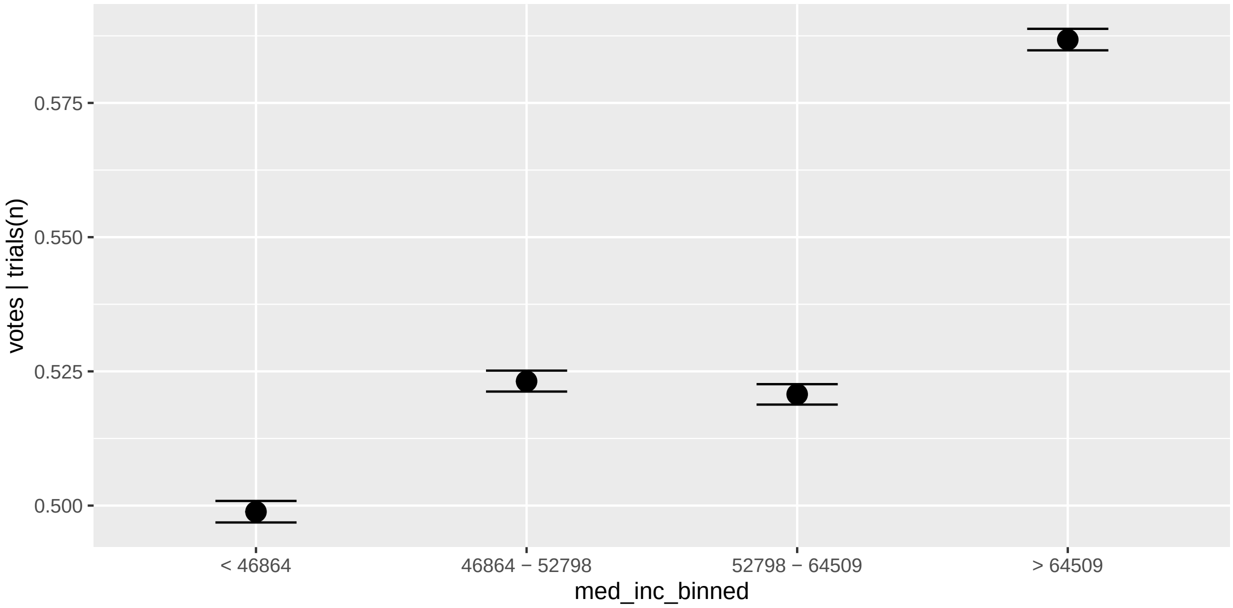
$$logit(Vote) = \beta_0 + \beta_1 I(Median\ Income\ > 64,509) + ...$$

In a later section, we will compare this Bayesian model with two additional models: one is a frequentist logistic regression model with the same predictors, interactions and random effect and the other a similar Bayesian model with additional random effect at the congressional district level. In this way we hope to assess if the Bayesian framework is superior than a frequentist approach when predicting voter turnout and if there is any salient unexplained variation within each congressional district. Before fitting any model, we preserved 20% of the full dataset as test data (*or are we doing 5-fold CV*) to quantitatively evaluate the models.

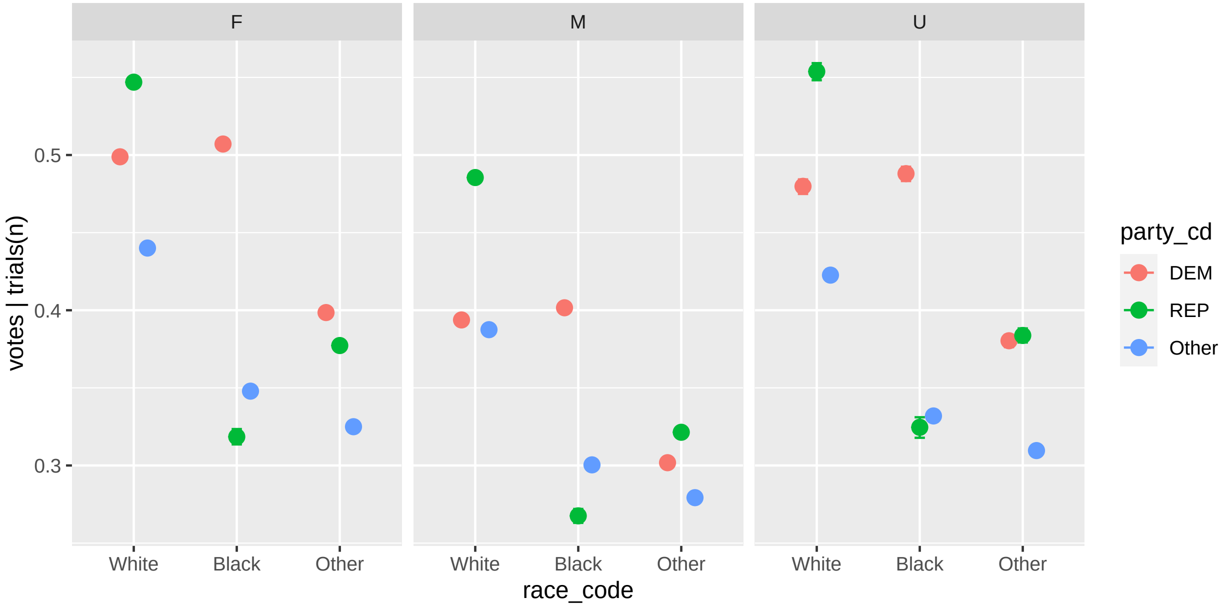
Results and Interpretations

	Estimate	Std. Error	2.5% Quantile	97.5% Quantile
Intercept	0.00	0.00	-0.01	0.00
Median Income > 64,509	0.10	0.00	0.09	0.10
Median Income 46,864-52,798	0.09	0.00	0.08	0.09
Median Income 52,798-64,509	0.36	0.00	0.35	0.36
Gender Male	-0.43	0.00	-0.44	-0.42
Gender Unspecified	-0.08	0.01	-0.09	-0.06
Race Black	0.03	0.00	0.03	0.04
Race Other	-0.41	0.00	-0.42	-0.40
Age 30-44	0.62	0.00	0.61	0.63
Age 45-59	1.08	0.00	1.07	1.09
Age 60+	0.94	0.00	0.93	0.95
Party Republican	0.19	0.01	0.18	0.20
Party Other	-0.24	0.00	-0.25	-0.23
Gender Male:Party Republican	0.18	0.00	0.17	0.19
Gender Unspecified:Party Republican	0.10	0.01	0.08	0.13
Gender Male:Party Other	0.21	0.00	0.20	0.22
Gender Unspecified:Party Other	0.01	0.01	-0.01	0.02
Race Black:Party Republican	-0.98	0.01	-1.00	-0.96
Race Other:Party Republican	-0.28	0.01	-0.30	-0.27
Race Black:Party Other	-0.42	0.01	-0.43	-0.41
Race Other:Party Other	-0.08	0.01	-0.10	-0.07
Gender Male:Age 30-44	0.02	0.00	0.01	0.03
Gender Unspecified:Age 30-44	-0.41	0.01	-0.43	-0.40
Gender Male:Age 45-59	0.11	0.00	0.10	0.12

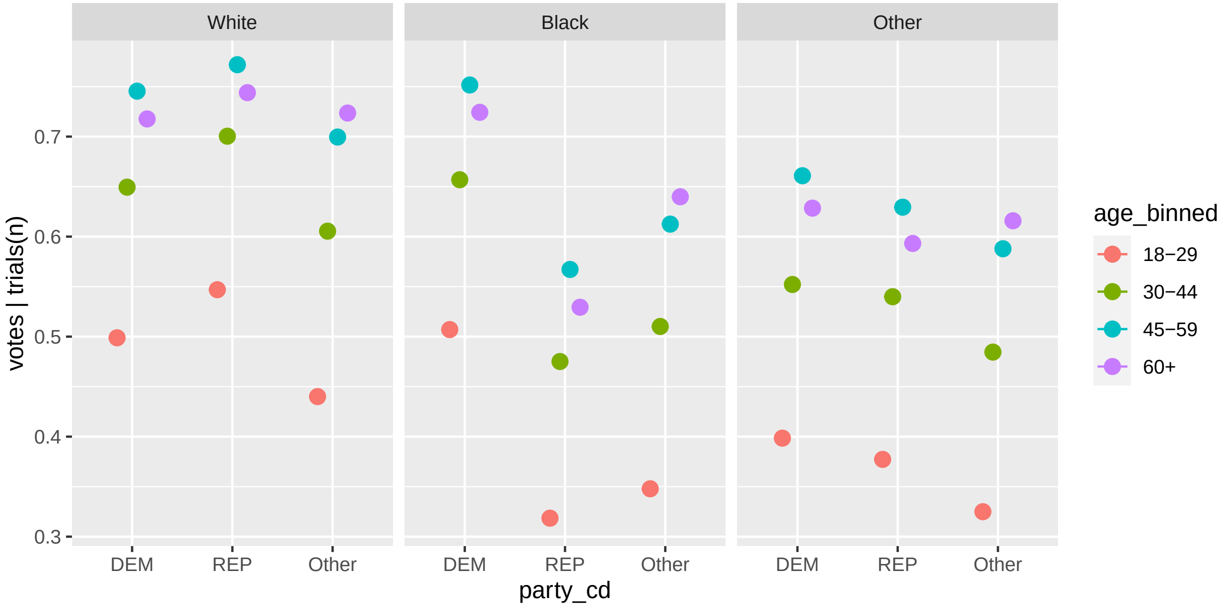
	Estimate	Std. Error	2.5% Quantile	97.5% Quantile
Gender Unspecified:Age 45-59	-0.60	0.01	-0.62	-0.58
Gender Male:Age 60+	0.25	0.00	0.24	0.26
Gender Unspecified:Age 60+	-0.47	0.01	-0.50	-0.45
Age 30-44:Party Republican	0.04	0.01	0.03	0.05
Age 45-59:Party Republican	-0.05	0.01	-0.06	-0.04
Age 60+:Party Republican	-0.06	0.01	-0.07	-0.05
Age 30-44:Party Other	0.05	0.01	0.04	0.06
Age 45-59:Party Other	0.01	0.01	0.00	0.02
Age 60+:Party Other	0.27	0.01	0.25	0.28



From the plot above, we see that the expected probability of voting is generaly greater that 50% for all median household income levels, but tends to increase as median household income increases, holding all other attributes constant (age, gender, race, party).

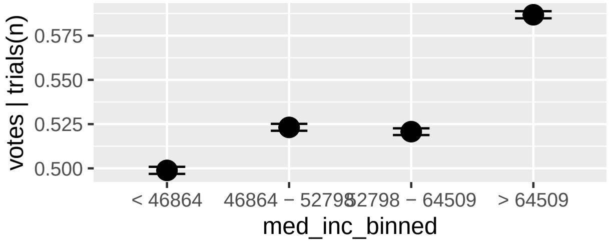


We can interpret each point in the plot above as follows: holding median household income level at baseline (less than $46,864) and age at baseline (ages 18-29), the y-axis value is the expected probability that a person of a particular race (x-axis), party (color), and gender (facet) votes. For example, the expected probability that a black, male, Democrat votes is 0.5, whereas the expected probability that a black, male, Republican votes is greater than 0.5. We can also see that women, regardless of race and party, are expected to be more likely to vote than men.

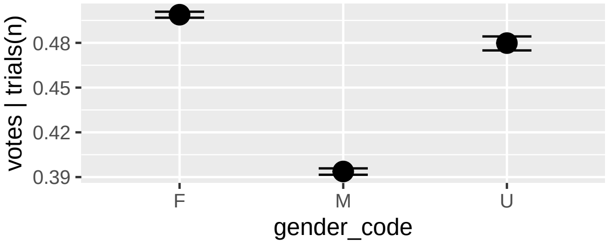


In the plot above, holding median household income level at baseline (less than $46,864) and gender at baseline (female), the expected probability of voting for white Democrats across all age groups is less than that for white Republicans. However, the expected probability of voting for black Democrats across all age groups is higher than that for black Republicans.

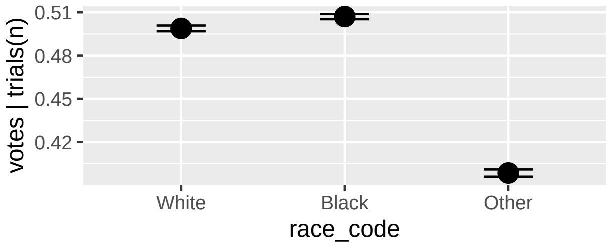
Appendix



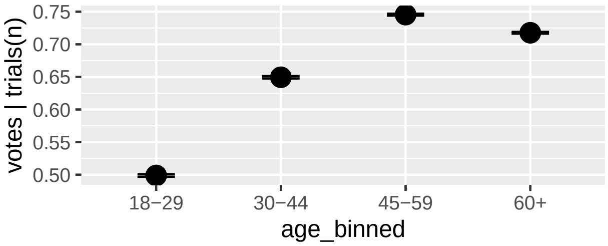
votes | trials(n)
0.575
0.550
0.525
0.500
< 46864
46864 − 52798
52798 − 64509
> 64509
med_inc_binned



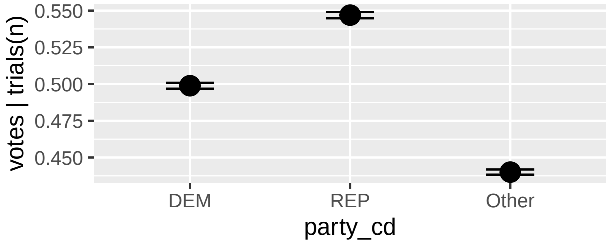
votes | trials(n)
0.48
0.45
0.42
0.39
F
M
U
gender_code



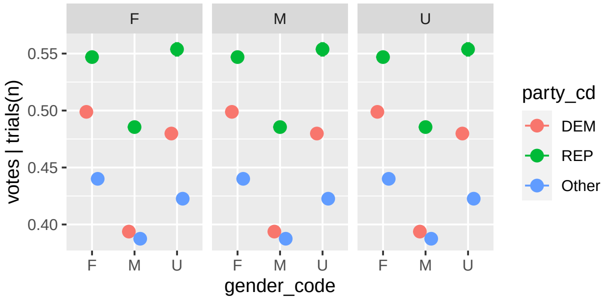
votes | trials(n)
0.51
0.48
0.45
0.42
White
Black
Other
race_code



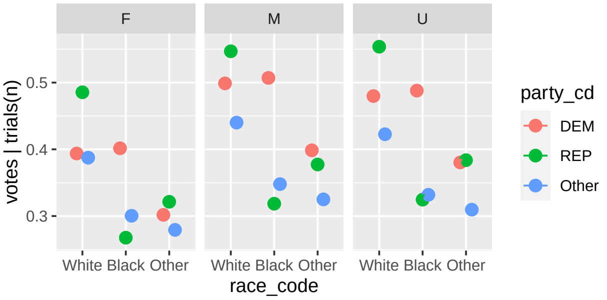
votes | trials(n)
0.75
0.70
0.65
0.60
0.55
0.50
18−29
30−44
45−59
60+
age_binned



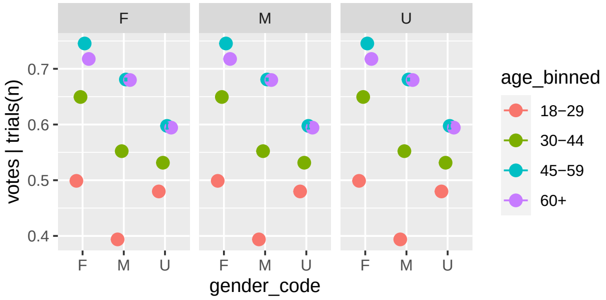
votes | trials(n)
0.550
0.525
0.500
0.475
0.450
DEM
REP
Other
party_cd



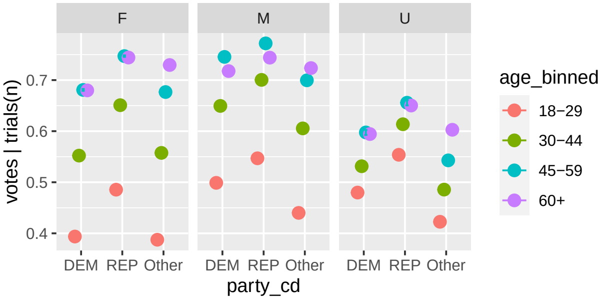
F
M
U
votes | trials(n)
0.55
0.50
0.45
0.40
F M U
F M U
F M U
gender_code
party_cd
DEM
REP
Other



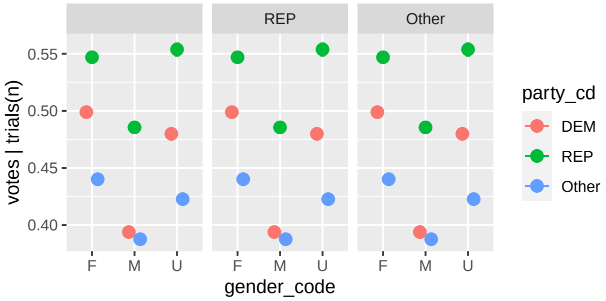
F
M
U
votes | trials(n)
0.5
0.4
0.3
White Black Other
White Black Other
White Black Other
race_code
party_cd
DEM
REP
Other

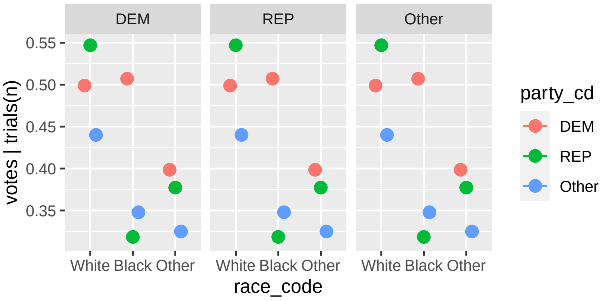


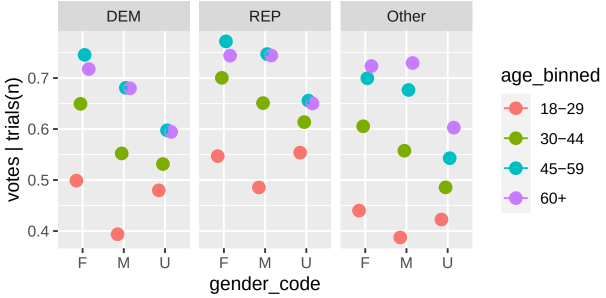
F
M
U
votes | trials(n)
0.7
0.6
0.5
0.4
F M U
F M U
F M U
gender_code
age_binned
18−29
30−44
45−59
60+

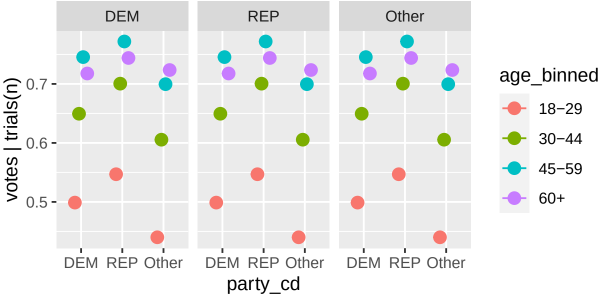


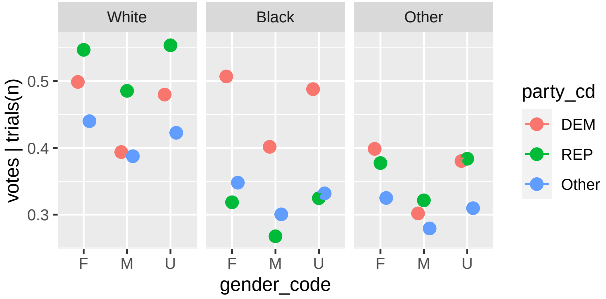
F
M
U
votes | trials(n)
0.7
0.6
0.5
0.4
DEM REP Other
DEM REP Other
DEM REP Other
party_cd
age_binned
18−29
30−44
45−59
60+

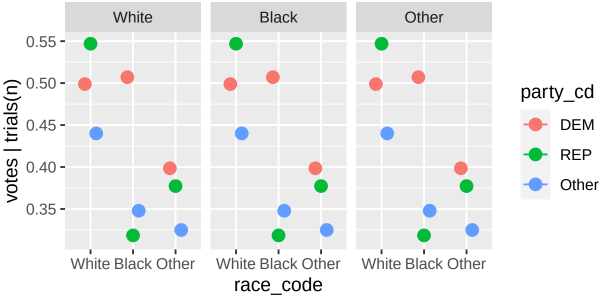


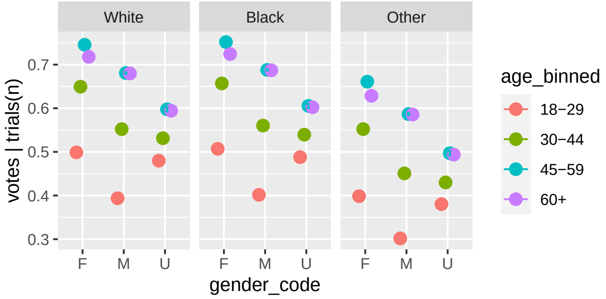


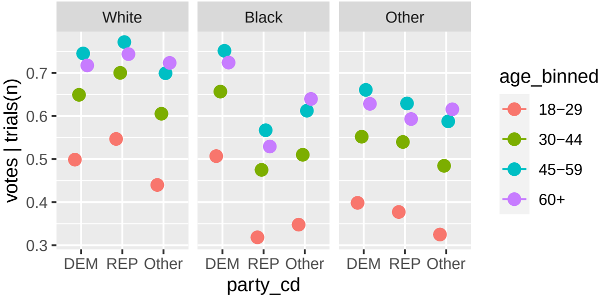












References

1. Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. Journal of the American Statistical Association, 108(501), 124-134. doi:10.1080/01621459.2012.737735

2. Hansford, T. G., & Gomez, B. T. (2010). Estimating the Electoral Effects of Voter Turnout. American Political Science Review, 104(2), 268-288. doi:10.1017/s0003055410000109

3. 2020 Election. (2020, October 20). Retrieved October 20, 2020, from https://fivethirtyeight.com/politics/elections/

4. Park, D. K., Gelman, A., & Bafumi, J. (2006). State-Level Opinions from National Surveys:. Public Opinion in State Politics, 209-228. doi:10.2307/j.ctvr33bdg.17

5. Mahler, V. A., Jesuit, D. K., & Paradowski, P. R. (2013). Electoral Turnout and State Redistribution. Political Research Quarterly, 67(2), 361-373. doi:10.1177/1065912913509306

6. Uhlaner, C. J., & Scola, B. (2015). Collective Representation as a Mobilizer. State Politics & Policy Quarterly, 16(2), 227-263. doi:10.1177/1532440015603576

7. Godbout, J. (2012). Turnout and presidential coattails in congressional elections. Public Choice, 157(1-2), 333-356. doi:10.1007/s11127-012-9947-7

8. Kim, S. S., Alvarez, R. M., & Ramirez, C. M. (2020). Who Voted in 2016? Using Fuzzy Forests to Understand Voter Turnout. doi:10.33774/apsa-2020-xzx29

9. Weinschenk, A. C. (2019) That's Why the Lady Lost to the Trump: Demographics and the 2016 Presidential Election, Journal of Political Marketing, 18:1-2, 69-91, DOI: 10.1080/15377857.2018.1478657

10. Charles, K. K., & Stephens, M. (2011). Employment, Wages and Voter Turnout. doi:10.3386/w17270

11. Hills, M. (2020, September 25). US election 2020: A really simple guide. Retrieved October 20, 2020, from https://www.bbc.com/news/election-us-2020-53785985

12. Railey, K. (2016). Federal Judges Let Stand North Carolina,'s New Congressional Map. The Hotline. https://link.gale.com/apps/doc/A498010836/ITOF?u=duke_perkins&sid=ITOF&xid=119d6ad9

13. Perrin, A. J., & Ifatunji, M. A. (2020). Race, Immigration, and Support for Donald Trump: Evidence From the 2018 North Carolina Election. Sociological Forum, 35(S1), 941-953. doi:10.1111/socf.12600

14. Redistricting in North Carolina. (2020). Retrieved October 21, 2020, from https://ballotpedia.org/Redistricting_in_North_Carolina

15. Â§ 132-1. Public Records. https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_132/GS_132-1.pdf

16. Â§ 163-82.10. Official record of voter registration. https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_163/GS_163-82.10.pdf

17. FairVote.org. (n.d.). Voter Turnout. Retrieved October 22, 2020, from https://www.fairvote.org/voter_turnout

18. McDonald, M. P. (2020). Voter Turnout Demographics. Retrieved October 23, 2020, from http://www.electproject.org/home/voter-turnout/demographics