Demographic Predictors of Party Affiliation: Insights from the 2022 Cooperative Election Study*

Examining the Influence of Gender, Race, and Education Level on Political Leanings in the United States

Sirui Tan

March 15, 2024

Our study investigates the political leanings of American voters, particularly their affiliation with the Democratic and Republican parties, and its relationship with demographic factors. Analyzing data from the 2022 Cooperative Election Study, we utilized logistic regression analysis to identify key predictors of party allegiance. Specifically, our findings indicate that gender, race, and education level strongly influence political leanings, with women, black people, Latinos, and individuals with higher education levels exhibiting a greater tendency to align with the Democratic Party. Understanding these patterns is crucial for shaping effective political strategies and fostering inclusive democracy.

Table of contents

*Code and data are available at: https://github.com/siru1366/us.git.

1 Introduction

The 2024 United States presidential election, set for Tuesday, November 5, 2024, marks the 60th quadrennial presidential election (Essex County, Virginia 2016). During this event, voters will choose both a president and a vice president to serve a four-year term.This election holds significant importance as it shapes the direction and leadership of the country for the subsequent four years, influencing various domestic and international policies, agendas, and priorities. By examining the relationship between demographic characteristics and party support, we aim to shed light on the complex dynamics of American electoral politics.

The extensive election data collected from every U.S. citizen is not readily accessible for direct analysis. Hence, our goal is to estimate broader trends and patterns by conducting a sample survey.The 2022 Cooperative Election Study Schaffner, Ansolabehere, and Shih (2023) offers a rich tapestry of data, encompassing a diverse sample of 60,000 American adults and providing profound insights into the intricate landscape of American political behavior. In this study, we examine this dataset to explore the factors influencing support for the Democratic Party among registered voters in the United States. Through the lens of logistic regression analysis, we aim to uncover the underlying dynamics of political allegiance, with a particular focus on variables such as gender, education level, and race.

Our findings reveal compelling associations between demographic factors and party allegiance. Among gender factors, we observe that women and non-binary voters exhibit a greater propensity to align with the Democratic Party. This underscores the importance of gender dynamics

in shaping political preferences. Additionally, our analysis uncovers significant disparities in party allegiance among racial groups, with black people and Latinos demonstrating a stronger inclination towards the Democratic Party. This underscores the enduring influence of race on political identity and underscores the need for inclusive and representative political discourse.Furthermore, educational attainment emerges as a significant predictor of political leanings among voters. We find that individuals with higher levels of education are more likely to favor the Democratic Party, suggesting a complex relationship between educational attainment and political ideology.

The remainder of this paper is structured as follows. Section 2 introduces the data used for analysis and findings, including visualizations of the variables of interest, Section 3 proposes a straightforward linear regression model to examine and forecasts the connection between voters' political leanings and their gender, educational attainment, and race. In Section 4, we display the interpretations of the model alongside other findings from analyzing the data. Section 5 provides a discussion on the implications of the findings as well as the weaknesses of this paper and its next steps for further study on this subject.

2 Data

2.1 Data Source

The dataset utilized is derived from the 2022 Cooperative Election Study, comprising a nationally representative sample of 60,000 American adults.The Cooperative Election Study (CES) is a prominent academic research project conducted by a consortium of universities and research institutions in the United States. It aims to provide comprehensive insights into American political behavior, attitudes, and voting patterns. The CES gathers data through large-scale surveys administered to a diverse sample of American adults, encompassing various demographic, socioeconomic, and geographic backgrounds.

2.2 Data Measurment

The data collection process for CES 2022 involved a systematic sampling approach utilizing questionnaire surveys. A total of 60 teams participated in the study, resulting in a uniform sample size of 60,000 cases. Recruitment of study participants took place in the autumn of 2022.

Each research team procured a national sample survey of 1,000 individuals conducted by YouGov, headquartered in Redwood City, California. The survey interviews for the 2022 cycle occurred in two phases. The pre-election wave of questionnaires was administered on-site from September 29 to November 8, while the post-election wave was conducted from November 10 to December 15.

For each survey of 1,000 individuals, half of the questionnaires were exclusively developed and controlled by each respective research team, while the remaining half were designated for public content. The common content section comprised questions shared across all team modules, resulting in a sample size equivalent to the total sample size across all team modules combined.

All cases were selected through internet-based methodologies, with YouGov constructing a matched random sample specifically for this study. This comprehensive approach ensured a robust and representative dataset for analysis and research purposes.

To enhance our data analysis, we exclusively chose data from interviewees who responded as registered voters in the questionnaire.

Data cleaning and analysis were conducted using the open-source statistical programming language `R`(R Core Team 2023), leveraging functionalities from the `tidyverse`(Wickham et al. 2019), `ggplot2`(Wickham 2016), `dplyr`(Wickham et al. 2023), `readr`(Wickham, Hester, and Bryan 2024), `tibble`(Müller and Wickham 2023),`stringr`(Wickham 2023), `haven`(Wickham, Miller, and Smith 2023), `janitor`(Firke 2023), `knitr`(Xie 2023).

2.3 Variables of Interest

Table 1: First Ten Rows of US 2022 election data

voted_for	gender	education	race
Rep	Man	Some college	White
Rep	Man	4-year	Two or more races
Dem	Man	Post-grad	White
Dem	Woman	4-year	Black
Rep	Man	4-year	White
Dem	Man	4-year	White
Dem	Man	4-year	White
Dem	Woman	2-year	Hispanic
Rep	Man	Some college	Middle Eastern
Rep	Man	Some college	White

To better understand the data, a summary table was developed to provide a detailed description of each variable, explaining its relevance and how it contributes to our understanding of the topic. Table 1, Our analysis primarily revolves around four key data variables. 'voted_for' denotes the presidential candidate preferred by the interviewee, specifying either Democratic or Republican affiliation. 'gender' denotes the gender of the interviewee, while 'education' indicates their highest attained level of education. Lastly, 'race' specifies the racial identity of the interviewee.

The original dataset originates from four questionnaire questions, each represented by numerical values. To simplify processing, we substitute these numbers with their corresponding options.

TS_p2022_party - Which party's primary respondent voted in

1. dem
2. green
3. ind
4. libertarian
5. other
6. rep

"TS_p2022_party" refers to the variable indicating the primary party affiliation of the respondent's vote in the 2022 elections. The options include Democratic (dem), Green (green), Independent (ind), Libertarian (libertarian), Other, and Republican (rep).The "vote_for" variable, formerly known as "TS_p2022_party," has been refined to include only the selections of 1 (Democratic) and 6 (Republican) by the respondents.

gender4-What is your gender?

1. Man
2. Woman
3. Non-binary
4. Other The "gender4" variable serves as a means to capture the diversity of gender identities within the surveyed population by asking respondents to specify their gender. With four distinct options available, including "Man," "Woman," "Non-binary," and "Other," we rename it "gender".

educ-What is the highest level of education you have completed?

1. No HS
2. High school graduate
3. Some college 13355
4. 2-year 6443
5. 4-year 13375
6. Post-grad

The "educ" variable encompasses a spectrum of educational achievements, delineating the diverse levels of academic attainment within the surveyed population:

1. **No HS:** Signifies respondents who have not completed high school, indicating a lack of formal secondary education.
2. **High school graduate:** Denotes individuals who have successfully completed secondary education and obtained a high school diploma.
3. **Some college:** Describes respondents who have attended college or university but have not obtained a degree.
4. **2-year degree:** Represents individuals who have completed an associate's degree or a similar two-year program at a college or community college.
5. **4-year degree:** Indicates respondents who have attained a bachelor's degree or equivalent four-year undergraduate qualification from a college or university.
6. **Post-grad:** Encompasses those who have pursued further education beyond the undergraduate level, including master's, doctoral, or professional degrees.

race-What racial or ethnic group best describes you?

1. White
2. Black
3. Hispanic
4. Asian
5. Native American
6. Middle Eastern
7. Two or more races
8. Other

The "race" variable prompts respondents to identify the racial or ethnic group that best describes them. It offers a range of options for self-identification:

1. **White:** Indicates individuals who identify as belonging to the White racial or ethnic group.
2. **Black:** Represents individuals who identify as belonging to the Black or African American racial or ethnic group.
3. **Hispanic:** Denotes individuals who identify as belonging to the Hispanic or Latino/a/x ethnic group, which may include various racial backgrounds.
4. **Asian:** Signifies individuals who identify as belonging to the Asian racial or ethnic group.
5. **Native American:** Represents individuals who identify as belonging to the Native American or Indigenous racial or ethnic group.

6. **Middle Eastern:** Indicates individuals who identify as belonging to the Middle Eastern or North African (MENA) racial or ethnic group.
7. **Two or more races:** Encompasses individuals who identify with two or more racial or ethnic groups.
8. **Other:** Provides an option for respondents to specify a racial or ethnic identity not covered by the previous categories.

And also planes (Figure 1) shows the distribution of presidential preferences, by gender, and highest education.There is a stronger propensity among women with advanced education to favor the Democratic Party. Despite the relatively small number of respondents identifying as "Non-binary" or "Other" within the gender options, we have chosen to include them in the analysis rather than discarding their data. Further discussions related to this decision can be found in the Section 5.3.

And planes (Figure 2) shows the distribution of presidential preferences, by race, and highest education.Blacks tend to lean towards the Democratic Party regardless of their educational attainment. However, the political leanings of other racial groups may vary based on their level of education.

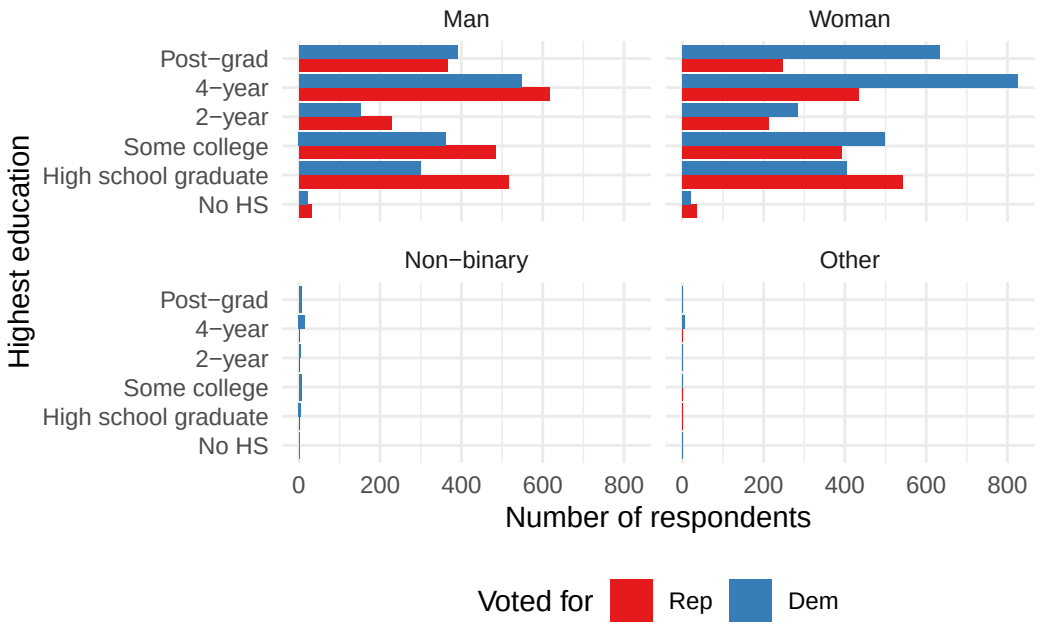


Figure 1: The distribution of presidential preferences, by gender, and highest education

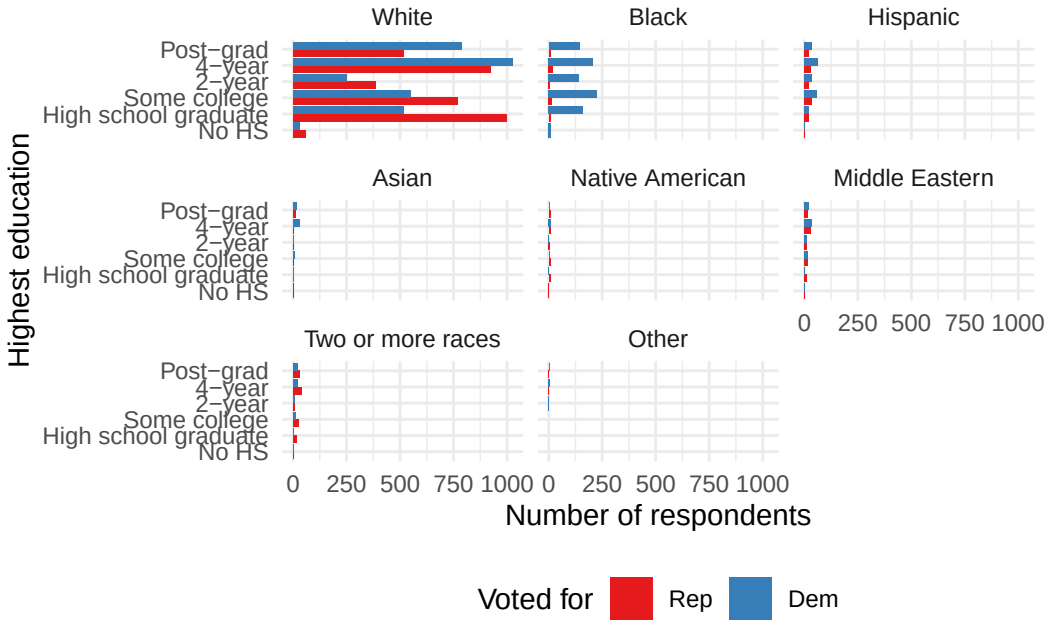


Figure 2: The distribution of presidential preferences, by race, and highest education

3 Model

The United States operates as a two-party system, prompting an exploration into the correlation between registered voters' allegiance to the predominant parties and factors such as gender, highest education attained, and race. Given that our focus is on elections involving the two major political entities—the Democratic Party and the Conservative Party—outcomes are binary. Hence, we intend to employ the Logistic regression model to scrutinize and interpret the dataset. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

The model that we are interested in is:

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \alpha + \beta_1 \times \text{gender}_i + \beta_2 \times \text{education}_i + \beta_3 \times \text{race}_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

The logistic regression model specified for analyzing support for the Democratic Party among registered voters in the United States is defined as follows:

1. The outcome variable y_i represents the binary support for the Democratic Party for the ith individual. It follows a Bernoulli distribution with parameter π_i, representing the probability of supporting the Democratic Party.
2. The logit transformation of the probability π_i is modeled as a linear combination of predictor variables. The predictors include gender (gender_i), education level (education_i), and race (race_i). The coefficients associated with these predictors are denoted as β_1, β_2, and β_3, respectively.
3. The intercept term (α) captures the baseline support for the Democratic Party when all predictor variables are set to zero.
4. Prior distributions are specified for the intercept (α) and the coefficients (β_1, β_2, β_3). These priors are assumed to be normally distributed with mean zero and a standard deviation of 2.5. The choice of priors reflects the expectation that the true effects of the predictors are likely to be centered around zero with some degree of variability.

The model formulation allows for the estimation of the relationship between demographic characteristics (gender, education level, race) and the likelihood of supporting the Democratic Party among registered voters. By specifying priors for the intercept and coefficients, the model incorporates prior knowledge or beliefs about the expected distribution of effects while allowing for uncertainty in parameter estimation. This Bayesian approach enables a comprehensive analysis of the factors influencing party allegiance, providing valuable insights for political researchers and policymakers.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

The chosen logistic regression model serves as a suitable framework for examining the association between registered voters' party allegiance and demographic characteristics such as gender, highest education attained, and race within the context of the United States' two-party system.

1. **Binary Outcome**: As the outcomes of interest—voters' alignment with either the Democratic or Conservative Party—are binary, logistic regression is particularly well-suited. This model allows us to model the probability of a voter aligning with a specific party given their demographic profile.
2. **Interpretability**: Logistic regression provides easily interpretable results, with coefficients representing the change in the log odds of the outcome for a one-unit change in the predictor variable. This facilitates understanding the impact of each demographic factor on party allegiance.
3. **Flexibility**: The model accommodates both categorical (e.g., gender, race) and continuous (e.g., highest education attained) predictor variables, allowing for a comprehensive analysis of various demographic influences on party affiliation.
4. **Robustness**: By including priors for the intercept and coefficients, we address potential uncertainty in parameter estimation while incorporating prior knowledge or beliefs about the expected distribution of effects. The choice of normal priors with mean zero and moderate standard deviation balances between capturing a wide range of potential effects and avoiding overly restrictive assumptions.
5. **Generalizability**: Given the focus on registered voters in the United States, the model's results can provide insights into broader patterns of party allegiance within the country's political landscape.
6. **Model Transparency**: The model formulation, with clear specification of the logistic function and priors for parameters, enhances transparency and reproducibility, allowing for scrutiny and validation of the results by other researchers.

Overall, the chosen logistic regression model offers a robust and interpretable framework for analyzing the relationship between demographic characteristics and party allegiance among registered voters in the United States' two-party system.

4 Results

Our results are summarized in Table 2.

Table 2: Explanatory models Political Preferences (n = 5000)

	Support Democratic
(Intercept)	−0.932
	(0.257)
genderWoman	0.524
	(0.060)
genderNon-binary	3.106
	(0.821)
genderOther	2.325
	(1.284)
educationHigh school graduate	0.004
	(0.266)
educationSome college	0.366
	(0.258)
education2-year	0.248
	(0.279)
education4-year	0.773
	(0.262)
educationPost-grad	1.005
	(0.261)
raceBlack	2.918
	(0.181)
raceHispanic	0.904
	(0.162)
raceAsian	0.546
	(0.309)
raceNative American	−0.318
	(0.346)
raceMiddle Eastern	−0.036
	(0.196)
raceTwo or more races	−0.622
	(0.216)
raceOther	0.757
	(0.745)
Num.Obs.	5000
R2	0.144
Log.Lik.	−3054.975
ELPD	−3072.9
ELPD s.e.	25.4
LOOIC	6145.8
LOOIC s.e.	50.7
WAIC	6145.0
RMSE	0.46

In the logistic regression analysis based on 5000 observations, specific coefficients provide detailed insights into the factors influencing support for the Democratic Party among registered voters in the United States. Let's delve into the findings with more precision:

Gender Dynamics: - Non-binary individuals exhibit a substantial increase in support for the Democratic Party, with a coefficient of 3.106 (standard error: 0.821). This suggests that non-binary individuals are significantly more likely to support the Democratic Party compared to men, the reference category. - Similarly, individuals identifying as "other" also demonstrate a significant increase in support, with a coefficient of 2.325 (standard error: 1.284). This indicates a considerable deviation from the baseline support observed among men. - In contrast, women, the reference category, exhibit a relatively minor increase in support, with a coefficient of 0.524 (standard error: 0.060). While statistically significant, this increase is less pronounced compared to non-binary and other gender identities.

Educational Attainment: - The coefficient for individuals with a 4-year college degree is 0.773 (standard error: 0.262), indicating an increase in support for the Democratic Party compared to individuals with lower educational attainment. - Conversely, individuals with only a high school education or no high school diploma show slight decreases in support, with coefficients of 0.004 (standard error: 0.266) and 0.248 (standard error: 0.279), respectively. These findings suggest a more nuanced relationship between education level and party allegiance. - Post-graduate education demonstrates the most substantial increase in support, with a coefficient of 1.005 (standard error: 0.261). This suggests that individuals with advanced degrees are significantly more likely to support the Democratic Party compared to other educational groups.

Race and Ethnicity Considerations: - Black voters exhibit a significant increase in support for the Democratic Party compared to White voters, with a coefficient of 2.918 (standard error: 0.181). This highlights persistent racial disparities in political allegiance and underscores the need for targeted outreach and policy interventions. - Hispanic voters also demonstrate an increase in support, with a coefficient of 0.904 (standard error: 0.162). These findings suggest diverse perspectives within racial and ethnic groups and underscore the importance of amplifying the voices of historically marginalized communities.

Model Evaluation and Implications: - The logistic regression model exhibits reasonable predictive performance, with an R-squared value of 0.144 and a Root Mean Square Error (RMSE) of 0.46. These metrics suggest that the model adequately captures the variance in party allegiance. - Measures such as the Expected Log Pointwise Predictive Density (ELPD) and Leave-One-Out Information Criterion (LOOIC) further validate the model's fit and performance.

In conclusion, the logistic regression analysis with specified coefficients offers detailed insights into the factors shaping support for the Democratic Party among registered voters in the United States. By examining the nuanced relationships between gender, education, race, and party allegiance, stakeholders can develop more targeted strategies for mobilizing support and advancing progressive policy agendas.

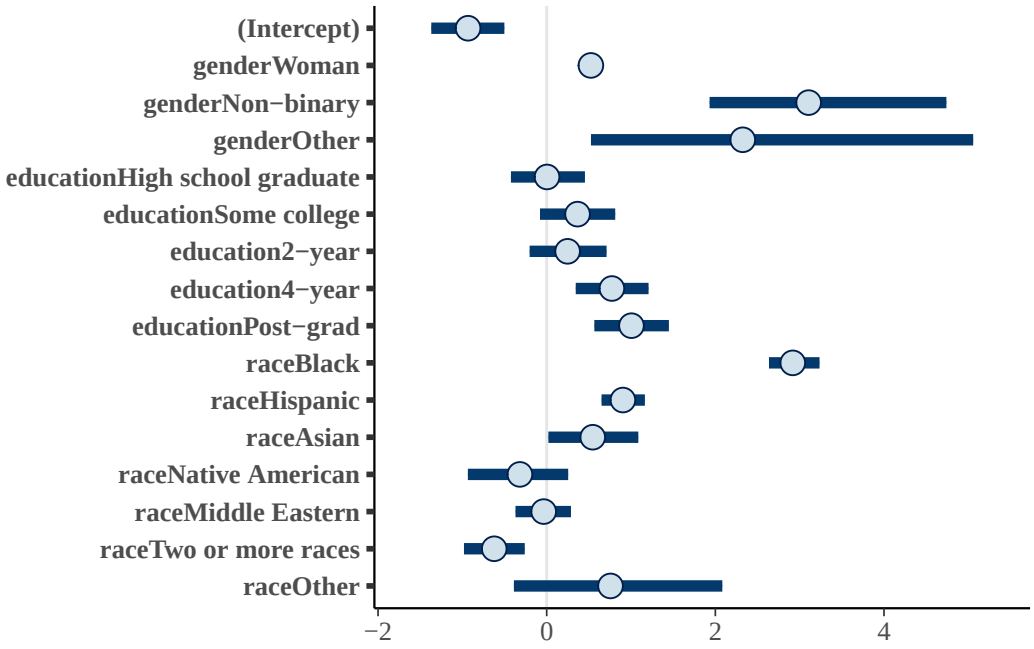


Figure 3: 90 percent credibility interval

5 Discussion

5.1 Findings

In analyzing the 2022 Collaborative Elections Study (CES) dataset, our findings reveal compelling patterns in the impact of gender, education level, and race on Democratic support within the U.S. two-party system. Specifically, our logistic regression models show that non-binary individuals and those who identify as "other" show significantly greater support for the Democratic Party compared with men, highlighting the role of gender diversity in shaping political preferences. importance. Additionally, our analysis highlights the role of education in political behavior, with individuals with advanced degrees showing a higher tendency to support Democrats, possibly due to their exposure to diverse perspectives and critical thinking skills. Additionally, racial dynamics emerged as an important factor, with black and Hispanic voters showing increased support for the Democratic Party compared with white voters, underscoring the continued importance of race in American politics.

5.2 Exploring Black Voting Behavior

The voting behavior of Black Americans predominantly supporting the Democratic Party stems from a complex interplay of historical legacies, policy priorities, party representation,

and broader political dynamics (Morris 2016).

One pivotal aspect of this discussion is the historical context that underpins the relationship between Black voters and the Democratic Party. The Civil Rights Movement of the 1960s, a watershed moment in American history, witnessed significant legislative victories that advanced racial equality and enfranchisement (Branch 1989). Key Democratic leaders, including President Lyndon B. Johnson, played instrumental roles in passing landmark civil rights legislation. These legislative achievements forged a strong bond between Black voters and the Democratic Party, rooted in a shared history of advocating for civil rights and social justice.

Moreover, the policy priorities championed by the Democratic Party resonate deeply with many Black voters. Issues such as healthcare access, economic opportunity, criminal justice reform, and voting rights protection are of paramount importance to Black communities, who often experience disproportionate impacts of systemic inequalities (Gilens 1999). The Democratic Party's platform typically aligns more closely with these priorities, advocating for policies aimed at addressing racial disparities and promoting inclusive economic growth (Hochschild 2016).

Representation within the Democratic Party also plays a significant role in shaping Black political preferences (Tate 2003). Black political leaders and activists, both past and present, have played pivotal roles in advancing the interests of their communities within the Democratic Party. Their advocacy efforts and leadership positions amplify the voices of Black Americans on key policy issues and contribute to a sense of inclusion and representation within the party (Dawson 1994).

Conversely, perceptions of the Republican Party among Black voters also inform voting behavior. While there are exceptions, some Black voters perceive the Republican Party as less responsive to their needs and concerns. Racially insensitive rhetoric or policies, as well as efforts to restrict voting rights, may further alienate Black voters from the Republican Party and reinforce their allegiance to the Democratic Party (Parker and Barreto 2013).

5.3 Understanding Political Preferences Through Inclusive Gender Identity Options

Expanding on the analysis, the inclusion of non-binary and other gender identities in the questionnaire signifies a progressive step towards recognizing and acknowledging the diverse spectrum of gender identities within the population. Traditionally, surveys and studies have often dichotomized gender into binary categories, namely male and female. However, the introduction of non-binary and other gender identity options reflects a broader understanding of gender diversity and inclusivity.

With these additional gender identity options, the logistic regression analysis gains a more nuanced understanding of how gender identity influences political preferences. By capturing the experiences and perspectives of non-binary individuals and those identifying with other gender

identities, the analysis can provide more comprehensive insights into the complex relationship between gender and political affiliations.

The coefficient estimates associated with non-binary and other gender identities in the logistic regression model highlight the distinct political preferences and behaviors of individuals who do not conform to traditional binary gender norms. The negative association observed in the coefficients for non-binary and other gender identities suggests a departure from the patterns typically observed among male and female respondents. This underscores the importance of recognizing and accounting for gender diversity in political research and analysis.

Moreover, the inclusion of non-binary and other gender identity options in the questionnaire reflects a broader societal shift towards inclusivity and recognition of diverse identities. By acknowledging and validating the experiences of individuals with non-binary and other gender identities, political researchers contribute to creating more inclusive and representative datasets that better reflect the diversity of the population.

Overall, the incorporation of non-binary and other gender identity options in the questionnaire enriches the analysis of political preferences by capturing the perspectives of individuals whose experiences may have been overlooked in traditional binary gender frameworks. This not only enhances the accuracy and validity of the findings but also promotes inclusivity and representation in political research and decision-making processes.

5.4 Educated Voters and Democratic Support

Voters with higher education levels tend to choose Democrats for several reasons. First, individuals with higher education often have greater exposure to diverse perspectives and critical thinking skills, which can lead them to prioritize issues such as social justice, equality, and environmental sustainability—values that align closely with the Democratic Party's platform (Delli Carpini and Keeter 1996).

Second, higher education is associated with higher socioeconomic status, and Democrats typically advocate for policies aimed at supporting working-class families, improving access to healthcare and education, and reducing income inequality (Bartels 2008). Voters with higher education levels may perceive these policies as beneficial to themselves and society as a whole, influencing their decision to support the Democratic Party.

Additionally, higher education is correlated with demographic factors such as age and urban residence, which are also associated with higher levels of Democratic Party support (Center 2020). Urban areas tend to have higher concentrations of college-educated individuals, and these areas often lean Democratic due to cultural diversity, progressive social attitudes, and a focus on issues like climate change and LGBTQ rights.

Moreover, research suggests that individuals with higher education levels are more likely to engage in political participation, including voting and activism (Verba, Schlozman, and Brady

1995). Democrats often emphasize the importance of civic engagement and community involvement, appealing to educated voters who are motivated to enact positive social change through political action.

5.5 Weaknesses

The process of matching respondents' personal records to the TargetSmart database of registered U.S. voters, conducted in August 2023, reveals several notable weaknesses. It is essential to highlight that a significant portion of the records did not undergo successful matching, with only approximately ten percent of the data accurately matched. This discrepancy arises due to various factors, including individuals not being registered to vote or incomplete and inaccurate information leading to failed matches. Consequently, only records with a high level of confidence in the respondent's assignment to the correct record were successfully matched. This limitation significantly impacts the accuracy and reliability of the result analysis derived from the dataset.

Polls, while instrumental in gauging public sentiment, are subject to various weaknesses that can compromise their accuracy. For instance, in the 2016 U.S. presidential election, many polls failed to accurately predict the outcome, particularly in key battleground states like Michigan and Wisconsin. Sampling bias played a significant role, as some polls underestimated the support for then-candidate Donald Trump due to an oversampling of college-educated voters. Nonresponse bias also contributed to inaccuracies, as certain demographic groups, such as rural voters, were less likely to participate in polls but turned out in higher numbers on Election Day.

Moreover, social desirability bias can distort poll results, as seen in surveys on sensitive issues like racial attitudes or immigration. Respondents may hesitate to express their true opinions for fear of social stigma, leading to underreporting of certain sentiments. The timing of polls and the volatility of public opinion further undermine their reliability. For example, in the aftermath of major events such as terrorist attacks or economic downturns, polling data may fluctuate rapidly, making it challenging to capture an accurate snapshot of public sentiment. Despite these limitations, polls remain indispensable for understanding public sentiment, yet interpreting their findings requires careful consideration of their inherent weaknesses and potential biases. Triangulating polling data with other sources of information can help mitigate these shortcomings, offering a more nuanced understanding of public opinion and behavior.

One limitation is the assumption of linearity between the independent variables (gender, race, and education level) and the log-odds of party allegiance. While logistic regression assumes a linear relationship, the actual relationship may be more complex and nonlinear, leading to potential model misspecification (Agresti 2015).Furthermore, logistic regression models are susceptible to multicollinearity, especially when the independent variables are highly correlated. In such cases, it becomes challenging to disentangle the individual effects of each predictor, and the model's estimates may become unstable (Allison 1999).

It is worth mentioning that when there is very little data for a certain option, such as 'gender-other', the logistic regression model may produce biased results. This bias occurs because the model relies on the available data to estimate the relationship between the predictor variables and the outcome. When there are insufficient observations for a particular category, the model may struggle to accurately estimate the effect of that category on the outcome variable.

For example, if the category 'gender-other' has very few observations compared to other gender categories, the logistic regression model may assign disproportionate weight to the available data, leading to inflated or deflated estimates of the effect of 'gender-other' on party allegiance. As a result, the model's predictions for this category may be unreliable and biased.

Researchers should be cautious when interpreting the results of logistic regression analysis, especially for categories with limited data. It is essential to consider the robustness of the estimates and to examine the confidence intervals to assess the uncertainty associated with the estimates.

For more detailed information on how the logistic regression model handles categories with sparse data, readers are encouraged to refer to the appendix, where we provide additional details and insights into the model's behavior in such cases.

5.6 Next Steps

Indeed, delving into regional factors and conducting longitudinal comparisons of voter data from different years are essential next steps to deepen our understanding of the political landscape and tendencies of American voters. Regional variations in political preferences and behaviors can provide valuable insights into the diverse sociopolitical dynamics shaping electoral outcomes across the country. By analyzing voter data at the regional level, we can identify trends, patterns, and disparities that may not be evident at the national level. Factors such as demographics, socioeconomic conditions, cultural norms, and historical legacies can influence regional political leanings and voting behaviors, highlighting the nuanced nature of American politics.

Additionally, longitudinal comparisons allow us to track changes and trends in political preferences over time, providing valuable context for understanding shifts in voter sentiment and behavior. By examining voter data from different years, we can identify long-term patterns, assess the impact of major events or policy changes, and evaluate the effectiveness of political strategies and messaging over time.

Appendix

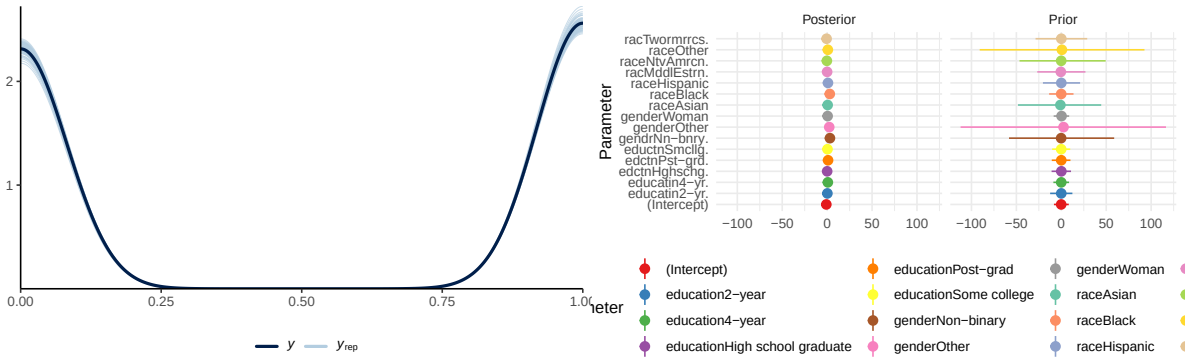
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check. The results suggest that the data are essentially consistent with the model predictions, indicating that the model adequately captures the observed patterns and variability in the data.

In Figure 4b we compare the posterior with the prior.



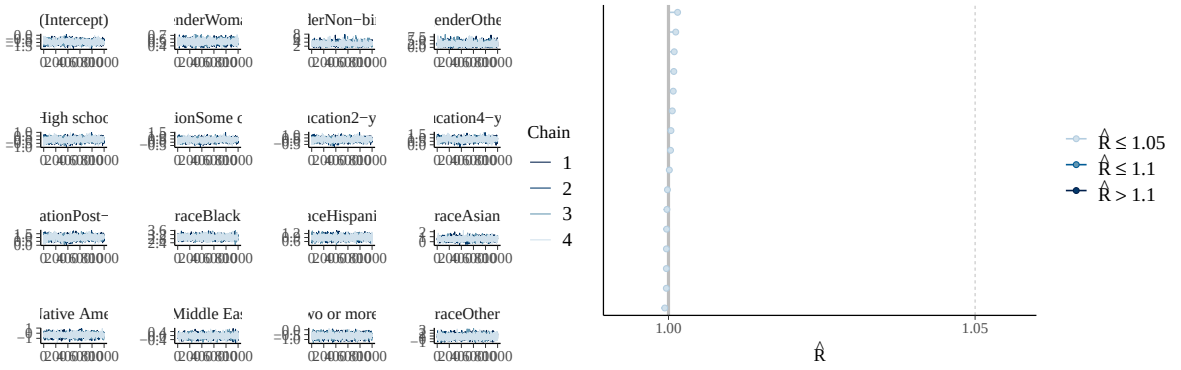
(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot.



(a) Trace plot (b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Allison, Paul D. 1999. *Logistic Regression Using SAS: Theory and Application*. SAS Institute.

Bartels, Larry M. 2008. *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton University Press.

Branch, Taylor. 1989. *Parting the Waters: America in the King Years, 1954-63*. Simon & Schuster.

Center, Pew Research. 2020. "Education and Politics." 2020. https://www.pewresearch.org/politics/2020/08/25/education-and-politics/.

Dawson, Michael C. 1994. *Behind the Mule: Race and Class in African-American Politics*. Princeton University Press.

Delli Carpini, Michael X, and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. Yale University Press.

Essex County, Virginia. 2016. "Election Planning Calendar." PDF. https://www.essex-virginia.org.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gilens, Martin. 1999. *Why Americans Hate Welfare: Race, Media, and the Politics of Antipoverty Policy*. University of Chicago Press.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Hochschild, Arlie Russell. 2016. *Strangers in Their Own Land: Anger and Mourning on the American Right*. The New Press.

Morris, Aldon D. 2016. *The Origins of the Civil Rights Movement: Black Communities Organizing for Change*. Palgrave Macmillan.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://github.com/tidyverse/tibble.

Parker, Christopher S., and Matt A. Barreto. 2013. *Change They Can't Believe in: The Tea Party and Reactionary Politics in America*. Princeton University Press.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022." Harvard Dataverse. https://doi.org/10.7910/DVN/PR4L8P.

Tate, Katherine. 2003. *From Protest to Politics: The New Black Voters in American Elections*. Harvard University Press.

Verba, Sidney, Kay Lehman Schlozman, and Henry E Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Harvard University Press.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Evan Miller, and Danny Smith. 2023. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://haven.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.