Understanding Voter Preferences: A Gender and Education Analysis of the 2022 United States Election*

Bernice(Yixuan) Bao, Zheng(Zeb) Yang, Dongli Sun

March 11, 2024

This paper investigates and studies the voter turnout of the 2022 the United Status election, and uses the CES research to sample the data and analyze the different groups. We used Bernoulli distribution to explore whether the gender and education of voters have an impact on candidates. We noticed that the analysis found that the majority of the group voted for Biden compared to Trump. Regardless of the level of education and gender differences, Biden's votes basically beat Trump's. Then, we're going to take a closer study in this phenomenon and we're going to explore why the majority of voters chose Biden. We need the outcome to determine whether they will affect future elections and adjust them.

1 Introduction

The purpose of CES is to investigate how Americans vote during elections in the United States and about voters' how they vote and their experiences during elections whether change according to political geography and social context. This study has an enormous sample to refer to, but we only need to start the calculation and evaluation on a state-by-state basis, so that we can do a better and more accurately understand on changes in the data. We will select some of these variables for in-depth analysis and optimization based on the articles by Schaffner, Ansolabehere, and Shih (2023), Brian and Ansolabehere, Stephen and Shih, Marissa.

During 2022, CES surveyed about 60 teams, producing a generic content sample of 60,000 cases. In the fall of 2022, the subjects for this study were recruited to do the test. There

*Code and data are available at: https://github.com/iloveyz12/US-Political-Support. Datasheet is inside the 'paper' folder.

were two interviews for this year's survey, and each research team bought a national sample of 1,000 people, which was conducted at YouGov in Redwood City, California. The paper will also analyze the sample of 1,000 people. CES used the data by taking samples and matching the data a second time. Finally, the survey results provide enough sample variables and the final data displayed, which can be re-entered into data cleaning and combined with other data to draw new conclusions. This paper is to select a few of the variables given and apply new filters and combinations to demonstrate whether the data generation is accurate.

We use R(R Core Team 2023) and the dataset from 2022 Cooperative Election Study Schaffner, Ansolabehere, and Shih (2023). To further enable the analysis I employed the use of the following packages: dataverse(Kuriwaki, Beasley, and Leeper 2023), ggplot(Wickham 2016), tidyverse(Wickham et al. 2019), arrow(Richardson et al. 2024), rstanarm(Goodrich et al. 2022), modelsummary(Arel-Bundock 2022) and here(Müller 2020).

2 Data

2.1 Data Source and variable

The data we used came from The 2022 Cooperative Election Study (CES) and because the database was too large and conducted on the web, with a total of 60,000 people participating in the campaign, YouGov used a matching random sample survey and sample to match these adult U.S. citizens and categorized these groups as gender, race and age. We mainly use two variables, gender and education, to analyze as Biden and Trump vote rating.

When we use dataset Kuriwaki, Beasley, and Leeper (2023) data, the data shows that several variables about "votereg", "presvote20post", "gender" and "edu" represent the voter, the voting status for 20 years, the gender of the voter, and the education level of the voter. First of all, we will use Wickham et al. (2019) for the simulate data part, and we get that there are 5 parts of the education level, from education level below high school to post-grad. At the same time, the data shows that there are four groups of genders, namely men, women, non-binary and other. In data cleaning, we're putting presidential candidates in the middle to show their respective approval ratings, and we're going to Wickham et al. (2019) and Richardson et al. (2024) Use "presvote0post" to filter, if "presvote0post" ==1, it will show Biden's approval rating, otherwise it will be Trump. This time, the data combines the voters' choice as well as their gender and education level, which are one of the factors.

We know from the data that 4,000 more women voted than men. Compare with man, women voters are more likely to vote for Biden, regardless of their educational stage. Men generally have higher approval ratings for Biden than Trump, except for men at the high school graduate level. Non-binary support for Biden almost crushed trump's votes, with the smallest number of people in the other group, with half and half turnout for Biden and Trump.

2.2 Data methodology

The 2022 United States Election Voter Survey is to select a representative sample within a sample group to conduct feedback and conduct in-depth surveys in order to ensure the generation of real data. In this way, when used for large-scale events such as elections in the United States, the fund of the web can greatly reduce the cost of expenses and the asymmetry of information generated by individuals. Digging deeper into the data, a matching method can be used for sample selectivity. The YouGov sample selection method is divided into two phases. First, they randomly selected a representative target sample from the target population. A matching sample is similar for the random sample and they will match it when the data can be matched. The matching process uses several variables to ensure that the matched sample is similar in measurement characteristics to the target sample. For example, gender will be divided into 4 different groups to vote within the group, in order to understand the choice by different genders people who chosen Biden. Further, we can combine the people of different genders with their educational backgrounds to analyze whether there are differences in the turnout rate of people in the same group due to different levels of education in different educational backgrounds.

2.3 Weighting

The target sample does not fit the demographics perfectly, and any remaining imbalances in the sample need to be weighted. Therefore, after the CES weighted the sample, the politically charged citizens were balanced in the distribution of multiple voter turnouts such as gender and education. Weighting is performed for samples in two parts. First, the 2020 presidential ballot was weighted by iterative proportional fitting ("skew") of joint distribution, and different states were selected as representative samples, so that the full cases could be weighted into the sampling framework using entropy balance. Then voters and other conditions were in a state of equilibrium, and the content of common content was cut out to reduce the proportion of data. In the second phase, CES used a matching method to conduct new weighting in 2022, which will ensure the diversity of the sample and understand the voting choices of different voters.

2.4 Vote Validation

In order to ensure that the data matches the TargetSmart, the lost of some data may not be matched. First, we download the file from dataverse Kuriwaki, Beasley, and Leeper (2023) to get the raw data, then we assign the sample, record the missing data, and test it. "TS_voterstatus" represents whether the voter is active or not. Then to "TS_g2022" about the status of whether voters vote or not, in 2022, CES recorded about 49% of people voting in the primary choice and 81% voting in the general election. He showed that if any non-missing value is less than 7, then the election has a verified voting record, which is possible

to be matched. We mainly use "CC22_401" to verify turnout, so that we can do the next step of matching and filter out data loss. If we want to verify the turnout of the CES, we can use three methods: variables and self-reported turnout - "TS_g2022" and whether it has been registered. To collect the primary election whether it is online or not, and then enter the system and survey and then determine whether they will participate in the general election for the second round, and how to participate in the selection method can also become a direction of analysis, because the data is too large, as mentioned above, it may be that the web will conduct the main voting method, but it is not excluded that other ways can also participate in the vote. Multivariate samples can lead to more data matching. Combined with the two main elements we used above about gender and education, we can end up with a complete and accurate output of data.

3 Model

3.1 Model set-up

Define y_i as the is the political preference of the respondent and equal to 1 if Biden and 0 if Trump. Then gender_i is the gender of the respondent and education_i is the highest education of the respondent. α represents the intercept term, which is the log-odds of the outcome variable when all predictor variables are equal to zero. β_1 and β_2 represent the coefficients associated with the predictor variables (gender and education level of respondent, respectively). We could estimate the parameters using `stan_glm()`. Note that the model is a generally accepted shorthand. In practice `rstanarm` converts categorical variables into a series of indicator variables and there are multiple coefficients estimated. In the interest of run-time we will randomly sample 1,000 observations and fit the model on that, rather than the full dataset.

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

In our model, these prior distributions are assumed to be normal distributions with mean 0 and standard deviation 2.5, where the prior distributions capture uncertainty about their values before observing the data. We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022), `modelsummary` package of Arel-Bundock (2022) and `here` package of Müller (2020). We use the default priors from `rstanarm`.

3.1.1 Model justification

Binomial logistic regression is a statistical method used to model the probability of a binary outcome variable. It's particularly suitable for situations where the dependent variable has two categories, such as in this study where we examine the likelihood of respondents voting for either Biden or Trump based on their gender and education level. The decision to employ binomial logistic regression for our study stems from its suitability for modeling binary outcome variables. As the outcome variable pertains to respondents' voting behavior, which involves choosing between Biden or Trump, binomial logistic regression is well-suited to capture this dichotomous outcome.

4 Results

Our results are summarized in Figure 1.

Based on the table outlining the likelihood of respondents supporting Biden based on gender and education levels, several key insights emerge. Firstly, the intercept value of 0.798 suggests that there is a baseline level of support for Biden among the surveyed population, regardless of gender or education. For men, their coefficient is not explicitly listed in the table, but it's implicitly represented by the intercept value. In this case, the intercept value of 0.798 can be interpreted as the baseline level of support for Biden among men.

Conversely, for women, there is a direct coefficient provided in the table, which is -0.608. This negative coefficient suggests that women exhibit a slightly lower level of support for Biden compared to the baseline represented by men. When examining gender, it becomes evident that being non-binary has a significant negative impact on the likelihood of supporting Biden, with a coefficient of -26.049. This indicates a substantial decrease in support compared to other genders. Conversely, the coefficient for individuals identifying as "Other" gender is negligible at 0.139, suggesting that this category does not significantly influence support for Biden.

The coefficients related to education levels illustrate a clear trend in support for Biden among respondents.High school graduates, individuals with some college education, and those with 2-year, 4-year, and post-graduate degrees all exhibit negative coefficients ranging from -0.504 to -1.649.Starting from high school graduates to post-graduates, there is a consistent decrease in support, with coefficients indicating diminishing likelihoods of supporting Biden as educational attainment increases. High school graduates exhibit a moderate decrease compared to the baseline, followed by individuals with some college education, 2-year, 4-year, and post-graduate degrees, showing progressively larger declines in support. Particularly striking is the substantial drop in support among those with post-graduate degrees, indicating that as education level increases, the likelihood of supporting Biden decreases.

Overall, the model's R-squared value of 0.056 indicates that gender and education levels explain only a small proportion of the variance in support for Biden among respondents.However, the

Table 1: Whether a respondent is likely to vote for Biden based on their gender and education

	Support Biden
(Intercept)	0.781
	(0.440)
genderNon-binary	−25.738
	(21.060)
genderOther	0.200
	(1.585)
genderWoman	−0.600
	(0.134)
educationHigh school graduate	−0.477
	(0.457)
educationSome college	−0.879
	(0.457)
education2-year	−1.087
	(0.489)
education4-year	−1.062
	(0.455)
educationPost-grad	−1.638
	(0.480)
Num.Obs.	1000
R2	0.056
Log.Lik.	−645.011
ELPD	−655.6
ELPD s.e.	9.8
LOOIC	1311.2
LOOIC s.e.	19.6
WAIC	1308.6
RMSE	0.48

coefficients provide valuable insights into how gender identity and educational attainment may influence political preferences, with non-binary gender and higher education levels being associated with decreased support for Biden.

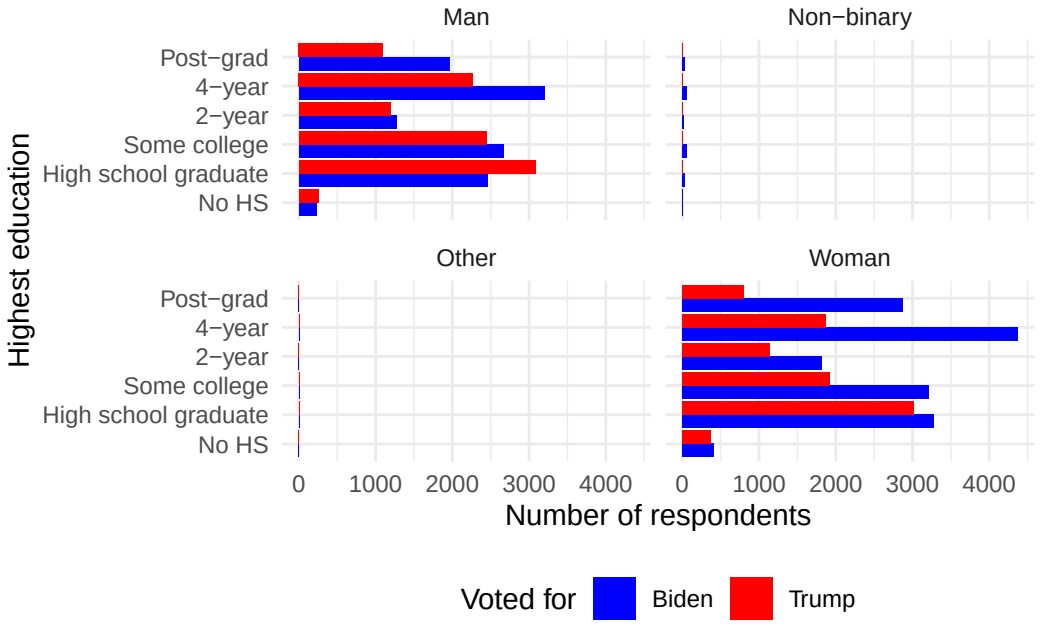

Figure 1: The distribution of presidential preferences, by gender, and highest education

5 Discussion

5.1 First discussion point: What does this paper do?

This paper delves into the voter turnout of the 2022 United States election, employing CES research to sample and analyze data across different demographic groups (Campbell, Gurin, and Miller (1954)). Utilizing the Bernoulli distribution, the study investigates whether the gender and education of voters impact their candidate preferences. The analysis reveals a consistent trend: regardless of gender or education level, the majority of voters cast their ballots for Biden over Trump (Bartels (2008)). Recognizing this phenomenon, the paper aims to conduct a deeper exploration into why Biden garnered widespread support and its potential implications for future elections (Ansolabehere and Iyengar (1995)).

5.2 Second discussion point: What have we learned from the world?

From existing literature and global observations, we've gathered insights into the dynamics of voter behavior and electoral outcomes. Studies have highlighted various factors influencing voter decisions, including socioeconomic status, political ideology, and candidate messaging (Bartels (2008)). Moreover, research into past elections has shown that demographic characteristics such as gender and education can play significant roles in shaping voting patterns (Campbell, Gurin, and Miller (1954)). Understanding these dynamics helps contextualize the findings of this paper and provides a foundation for further analysis.

5.3 Third discussion point: What's another thing we've learned from the world?

Another crucial insight gleaned from global observations is the importance of candidate appeal and campaign strategies in influencing voter preferences (Ansolabehere and Iyengar (1995)). Effective communication strategies, including resonant messaging and clear articulation of policy proposals, play a pivotal role in swaying undecided voters and mobilizing support from diverse demographic groups (Holbrook (1996)). Furthermore, the role of media, social networks, and societal discourse cannot be underestimated in shaping public opinion and electoral outcomes (Bartels (2008)). The intricate interplay between these factors, alongside demographic variables such as gender and education, offers a comprehensive understanding of the multifaceted dynamics that drive voter behavior. By delving into these nuanced interactions, researchers can glean invaluable insights into the complex mechanisms that underpin electoral decision-making processes, thus enriching our understanding of political phenomena and informing strategic approaches to future elections.

5.4 Weaknesses and next steps

Despite its contributions, this paper has several limitations that warrant acknowledgment (Ansolabehere and Iyengar (1995)). Firstly, while the analysis identifies correlations between gender, education, and candidate preference, it may overlook other influential factors such as race, age, and geographic location (Campbell, Gurin, and Miller (1954)). Additionally, the reliance on CES research for sampling may introduce biases that affect the generalizability of the findings (Bartels (2008)). Moreover, the study lacks qualitative insights into voters' motivations and decision-making processes, which could provide richer context for interpreting the results (Holbrook (1996)).

Moving forward, further research is needed to deepen our understanding of the factors driving voter behavior and electoral outcomes (Bartels (2008)). Future studies should employ more diverse sampling methods and incorporate qualitative methodologies to capture the nuances of voter preferences (Ansolabehere and Iyengar (1995)). Additionally, exploring the intersectionality of demographic variables and considering evolving societal trends will enhance the comprehensiveness of analyses (Campbell, Gurin, and Miller (1954)). Moreover, longitudinal

studies tracking voter preferences over time can provide valuable insights into the evolving political landscape and inform strategic adjustments for future campaigns (Holbrook (1996)). By addressing these areas, researchers can advance our understanding of electoral dynamics and contribute to more informed policymaking and political strategies.

References

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. Free Press.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

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

Campbell, Angus, Gerald Gurin, and Warren E Miller. 1954. *The Voter Decides*. Row, Peterson; Company.

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

Holbrook, Thomas M. 1996. *Do Campaigns Matter?* Sage Publications.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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

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