Analyzing Demographic and Socioeconomic Characteristics of Biden Voters in the 2020 US Presidential Election*

Chenhang Huang Zixuan Yang

March 12, 2024

This study examines the demographic and socioeconomic aspects that influence voter support for President Joe Biden in the 2020 US presidential election, using logistic regression models based on 2020 CES data. The findings show that Biden's triumph was driven by a varied coalition of supporters, including women, highly educated persons, and younger white voters. Understanding the demographics and inclinations of Biden voters provides vital insights into American politics and society, which can be used to shape future political campaigns and legislative goals. This work adds to our understanding of the processes affecting electoral outcomes and the implications for democratic government by shedding light on the elements influencing voter decisions.

Table of contents

*Code and data are available at:

1 Introduction

In the 2020 US presidential election, Joe Biden defeated Donald Trump to become the President of the United States. With the next election scheduled for November 2024, there is speculation about Trump's potential return to the political arena, leaving uncertainty about Biden's reelection. Therefore, understanding the demographic and socioeconomic characteristics of voters who supported President Joe Biden has become a focal point for researchers and political analysts. The estimand of this paper focuses on determining the average effect of demographic and socioeconomic factors on the likelihood of voting for Biden in the 2020 US Presidential Election.

This study, based on 2020 CES data, aims to delve into the characteristics of Biden voters to reveal the factors influencing their electoral decisions and the broader implications for American politics and society. Demographic factors analyzed include age, gender, and race, while socioeconomic factors encompass education. Utilizing logistic regression models, we examine how these four factors influence the probability of voting for Biden. The results indicate that President Biden's victory was driven by a diverse coalition of supporters, spanning different age groups, races, ethnic backgrounds, and socioeconomic statuses. Analyzing the demographic and socioeconomic characteristics of Biden voters reveals that women, highly educated individuals, and younger white voters are more inclined to choose Biden.

By elucidating the demographics and socioeconomic status of Biden voters, this study fills gaps in our understanding of the 2020 presidential election outcome and its implications for the future of American election. Understanding the composition and preferences of Biden voters will provide valuable insights into the Biden administration and aid in predicting the outcome of the 2024 US presidential election.

In the subsequent sections of this paper, we outline the methods used for data collection and analysis, discuss our findings.The sturcture of this paper is shown as follows. Section 2 will do the data analysis, and it includes the overview of the data, including the data source, the number of variables, and their definition. Section 3 will introduce the logistic model and the corresponding assumptions. Section 4 focuses on the analysis of the results of the logistic regression. In the end, discussion on these results are proposed in Section 5.

2 Data

2.1 Data Description

The 2020 Cooperative Election Study (CES) dataset by Schaffner, Ansolabehere, and Luks (2021) serves as the foundation for our analysis. This dataset represents the final release of the 2020 CES Common Content Dataset, encompassing responses from a nationally representative sample of 61,000 American adults. It includes survey data, a comprehensive data guide,

and questionnaires. Notably, the dataset incorporates vote validation conducted by Catalist, adding an additional layer of reliability to the electoral data.

The CES survey has a longstanding tradition of capturing US political opinion. The sampling methodology, as outlined by Schaffner, Ansolabehere, and Luks (2021), utilizes a matching approach, balancing sampling concerns with cost considerations effectively. While there may be similar datasets available, such as those from other reputable survey organizations or academic institutions, the CES dataset was selected for its extensive coverage, rigorous sampling methodology, and inclusion of vote validation data. Alternative datasets may not offer the same level of representativeness or comprehensive coverage of political opinion among American adults. Therefore, the CES dataset was deemed the most suitable and reliable choice for our research purposes.

This study aims to examine the impact of individual factors such as age, gender, race, and education on the choice of Biden voters in the election. Considering that the voting age for U.S. citizens is 18 years old, we first exclude sample units below this age from the survey results. Additionally, this study focuses solely on voters who chose either Biden or Trump, as the outcome of the 2020 election was determined between these two candidates. Given the diverse racial composition of the United States and the predominance of white Americans, accounting for approximately 60% of the total population, we categorize race into two groups: white and non-white. Furthermore, we incorporate gender and education information to assess the differences in the preferences for Biden and Trump among individuals of different ages and educational backgrounds.

2.2 Measurement

We can access the CES using the package dataverse by Kuriwaki, Beasley, and Leeper (2023) in the rstudio of R Core Team (2023). Upon examining the actual data, we noticed that there were no direct variables indicating the choice between Biden and Trump. However, upon consulting the data manual, we found that when the variable "CC20_410" is coded as 1, it signifies support for Biden, while a code of 2 indicates support for Trump. Therefore, we filtered the dataset to include only respondents with "CC20_410" values of 1 and 2. Furthermore, the CES dataset only provides respondents' birth years. To determine their ages, we subtracted their birth year from 2020. Gender information in the CES dataset is categorized as "female" and "male," with a variable value of 1 representing "male" and 2 representing "female." Finally, the variable "educ" ranges from 1 to 6 in the CES dataset, representing different levels of education attainment: "No Highschool," "High school graduate," "Some college," "2-year college," "4-year university," and "Post-grad."

After generating the necessary data and save it into csv file using readr package by Wickham, Hester, and Bryan (2023), we retained only the relevant five variables and removed observations with missing values using tidyverse package by Wickham et al. (2019). Ultimately, the dataset consists of 43,554 observations and 5 variables. Figure 1 illustrates presidential preferences

categorized by gender and highest education level by using the ggplot2 package by Wickham (2016). Among female voters, the number of individuals who voted for Biden significantly outweighs those who voted for Trump, especially among voters with a 4-year university degree. Trump only holds a slight edge over Biden among voters with a high school diploma. Overall, Figure 1 indicates that female and highly educated voters are more inclined to choose Biden, and Biden enjoys a notable advantage over Trump among all voters. Additionally, Figure 2 presents presidential preferences categorized by race and age. Among non-white voters, Biden maintains an overwhelming advantage, while among white voters, Biden's advantage is less pronounced. Particularly among white voters aged 50 and above, Biden does not hold a significant advantage, whereas among white voters under 50, Biden exhibits a slight edge.

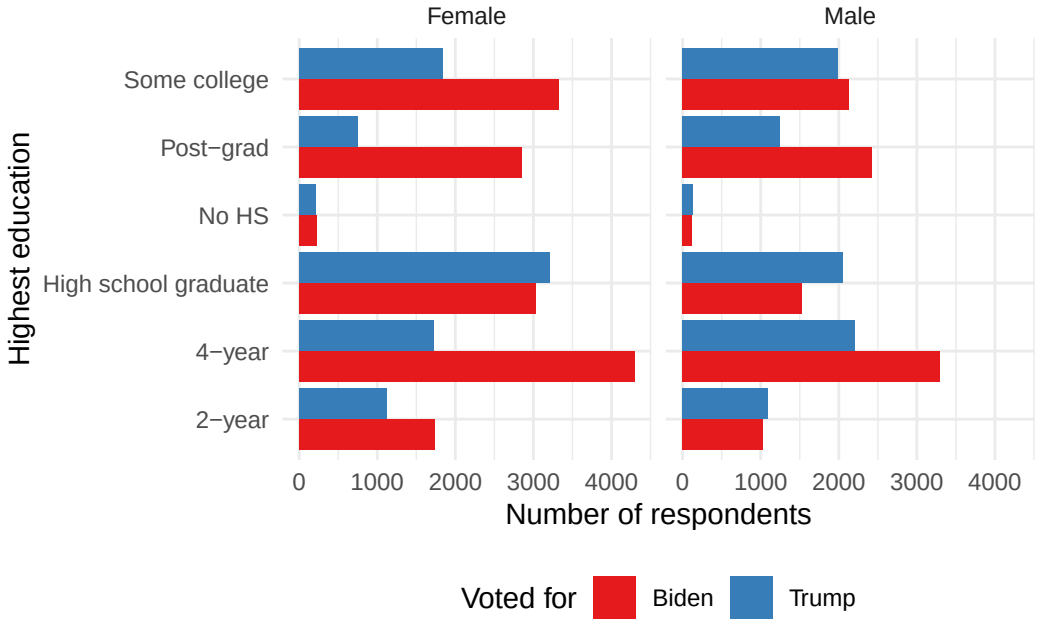

Figure 1: The distribution of presidential preferences, by gender, and highest education

3 Model

Considering that the output variable y_i only has two possible values (0 or 1), we assume that y_i follows a Bernoulli distribution with probability π_i. To establish the relationship between the predictors and the probability π_i., we employ logistic regression by Wang et al. (2015). The logistic regression model is defined as:

$$y_i|\pi_i \sim Bern(\pi_i)$$

$$logit(\pi_i) = \beta_0 + \beta_1 * gender_i + \beta_2 * education_i + \beta_3 * age_i + \beta_4 * white_i$$

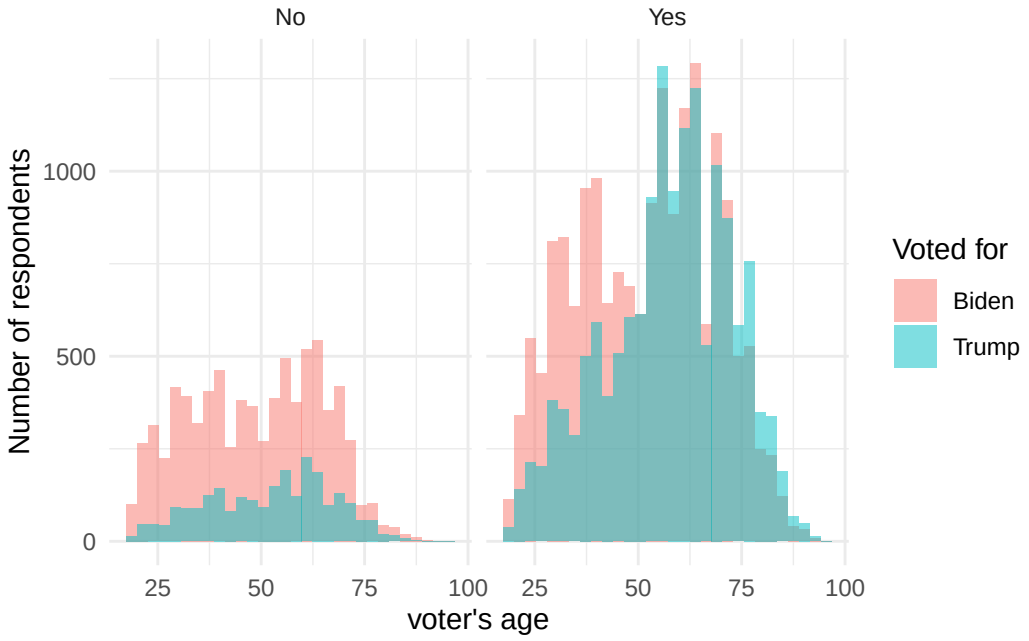

Figure 2: The distribution of presidential preferences, by race, and age

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$$

$$\beta_0 \sim Normal(0, 2.5)$$
$$\beta_1 \sim Normal(0, 2.5)$$
$$\beta_2 \sim Normal(0, 2.5)$$
$$\beta_3 \sim Normal(0, 2.5)$$
$$\beta_4 \sim Normal(0, 2.5)$$

where gender, education, age and white are the predictors in the model, and we assume their coefficients follow a normal distribution with standard derivation 2.5. In order to predict these coefficients, we use stan_glm() from the package rstanarm by Goodrich et al. (2022). This model is written in short-hand, in practice, these categorical variables will be converted into a series of indicator variables with different coefficients.

4 Result

After conducting regression analysis on all variables, the results are presented in Table 1 by using the modelsummary package by Arel-Bundock (2022). Table 1 displays the estimated

values of each parameter along with their corresponding standard deviations. For a normal distribution, the critical value at the 0.05 level is 1.96. Hence, we can use the ratio between beta and its standard deviation to obtain the test statistic and determine its significance. The absolute values of the test statistics for all variables in the table are greater than 1.96, indicating significance at the 0.05 level. The log odds ratio of voting for female voters is 0.399 higher than male voters when holding education, age, and race constant. This result is consistent with Figure 1.

Regarding education, there are five indicators, with 2-year college as the reference level. The coefficients for 4-year education and post-graduation are positive, with the coefficient for 4-year education being less than that for post-graduation education. This suggests that individuals with higher education levels are more likely to choose Biden. Conversely, the coefficients for high school graduate and no high school graduate are negative, indicating that individuals with lower levels of education are less likely to choose Biden. These results align with the findings from Figure 1. The coefficient for age is negative, indicating that as voters age, the probability of choosing Biden decreases, consistent with the distribution of white voters shown in Figure 2. Additionally, the coefficient for non-white voters is 0.89, indicating a stronger preference for Biden among non-white voters. The results are highly consistent with those from Figure 1 and Figure 2, demonstrating that logistic regression can effectively estimate the relationship between age, education, gender, race, and the choice of voting for Biden.

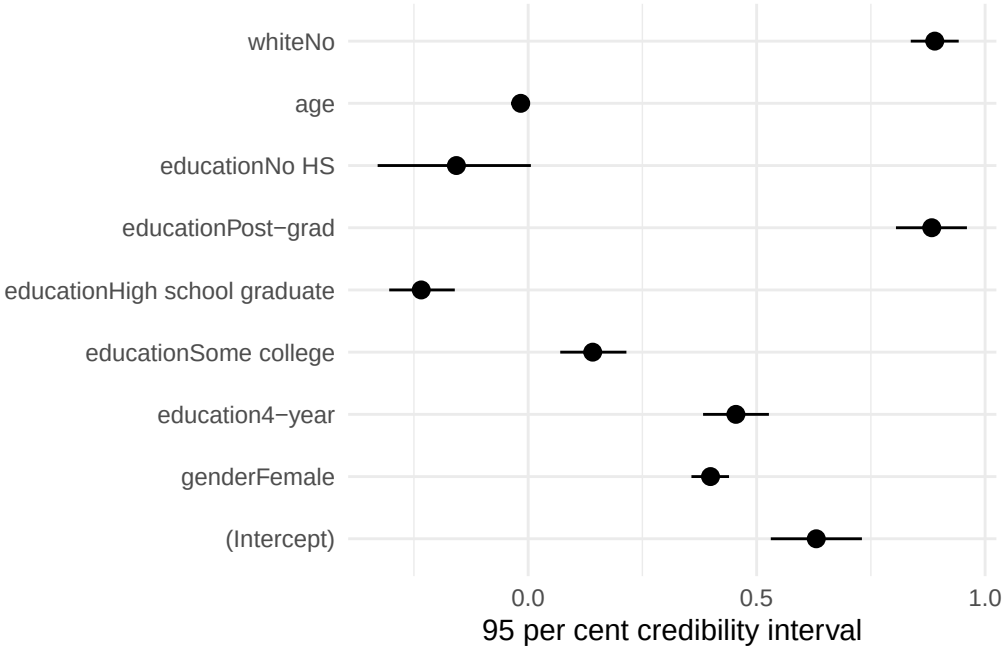

Figure 3: Credible intervals for predictors of support for Biden

Table 1: Results of logistic regression on the voted for Biden

	Support Biden
(Intercept)	0.630
	(0.047)
genderFemale	0.399
	(0.020)
education4-year	0.455
	(0.037)
educationSome college	0.141
	(0.036)
educationHigh school graduate	−0.234
	(0.036)
educationPost-grad	0.883
	(0.041)
educationNo HS	−0.157
	(0.082)
age	−0.016
	(0.001)
whiteNo	0.890
	(0.027)
Num.Obs.	43 554
R2	0.090
Log.Lik.	−27 322.328
ELPD	−27 331.4
ELPD s.e.	71.7
LOOIC	54 662.8
LOOIC s.e.	143.5
WAIC	54 662.8
RMSE	0.47

5 Discussion

5.1 The demographic and socioeconomic characteristics of Biden voters

During the 2020 election process, CES results showed that Biden's support was significantly higher than Trump's. There are distinct characteristics among Biden's supporters.

Firstly, Biden received support from non-white ethnicities and women. Biden's policy agenda includes addressing systemic racism, promoting social justice, and advancing policies aimed at improving the economic and social well-being of minority communities. These policy priorities may have attracted non-white voters who have historically faced inequality and discrimination. Biden emphasized diversity and inclusivity, with his running mate Kamala Harris being the first woman of color nominated for national office by a major political party. This fostered a sense of connection and trust among non-white and female voters.

Secondly, Biden's appeal to young people, particularly among white youth, stems from his policy proposals on climate change, student loan debt relief, healthcare, and social justice. His campaign embraced progressive ideas advocated by movements like Black Lives Matter and climate activism, resonating with many young voters passionate about social justice and progressive change. Disillusioned by the Trump administration, young people saw Biden as a more moderate and stable alternative. Biden effectively utilized social media platforms and grassroots organizing to engage and mobilize young voters. His alignment with progressive values, rejection of Trumpism, emphasis on diversity and inclusion, and strategic outreach efforts contributed to his popularity among young people, including white youth. Biden's policies and message appealed to the concerns and aspirations of younger generations, driving support for his candidacy.

Thirdly, Biden's appeal to voters with higher education can be attributed to several factors. Firstly, his emphasis on investing in education and providing relief for student loans directly benefits those with higher levels of education. Biden's policies resonate with educated voters by addressing their practical needs and concerns. Secondly, Biden's inclusive and moderate political stance garners admiration from voters with higher education, who tend to favor pragmatism and social stability. His approach aligns with the values of educated voters, who appreciate his emphasis on unity and compromise in governance. Furthermore, Biden's recognition of the importance of science and expertise aligns with the perspectives of those with higher levels of education. This emphasis on evidence-based decision-making resonates strongly with educated voters, who value rationality and informed policy choices.

5.2 The factors related to the next election

Through the above research, we have found that a candidate's policy stance will directly influence voters' decisions. To garner support from the general populace, candidates need to formulate policies that meet the needs of the majority, although these policies may potentially

harm the interests of a minority. Therefore, candidates must adapt their governing principles in response to opinion polls promptly, ensuring widespread support and eventual victory in the next election.

While Biden emerged victorious in the 2020 election, his tenure has been marked by numerous challenges. Firstly, there are domestic and foreign policy challenges that Biden faces. His call to dismantle the border wall has sparked dissatisfaction among Texans, as illegal immigration impacts the interests of Texas voters. Texas also garners support from many other states, so Biden's handling of immigration issues will influence his electoral outcomes. Secondly, in international relations, the ongoing conflict between Ukraine and Russia, with the U.S. supporting Ukraine, could impact Biden's reelection bid. The outcome of the conflict will affect Biden's administration, potentially garnering support if Ukraine emerges victorious, or providing ammunition for opponents if perceived as a misallocation of American resources. Economically, the U.S. remains in a high-interest rate environment, and decisions regarding interest rate cuts and economic development will be crucial factors determining the next election's outcome.

5.3 Weaknesses and next steps

The weaknesses of this study can be identified in several aspects. Firstly, the primary factor affecting the accuracy of predicting the next election lies in the outdated nature of the data. The analysis presented above is based on the 2020 election survey data, and after four years of Biden's presidency, the characteristics of the population supporting Biden may have evolved since 2020. The support for Biden was initially based on the policy agenda he proposed before taking office. However, during his tenure, Biden may not have fulfilled all of his promises, potentially causing disillusionment among some of his original supporters and leading them to lose trust in Biden, thereby opting for other candidates. Secondly, this study relies solely on CES data, which, while attempting to represent the population as accurately as possible, may suffer from biases inherent in survey data. To address potential biases in the survey results, future research could utilize census data to adjust the estimation results, thereby obtaining more accurate findings. Thirdly, as indicated by Table 1, the low R-square value of only 9% suggests that the explanatory power of the predictors for voting for Biden is limited. It is known that factors such as voter income, family situation, and the state in which they reside can influence voter decisions. In subsequent analyses, we will introduce additional predictors to enhance the model's explanatory power for voting for Biden.

References

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.

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

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

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 Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91.

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.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.