Deciphering Key Determinants of Biden Voter Support: Insights from Logistic Regression Analysis in the 2020 US Presidential Election*

Fares Alkorani Owais Zahid

March 16, 2024

In this paper, we discuss the correlation between various statistics of the respondents to the CES2020 election survey, and whether they voted for Trump or Biden. The statistics are: age, education, gender and race.

1 Introduction

On November 4, 2020, the 59th United States presidential election took place between Joe Biden and Donald Trump, where Joe Biden was elected as the 46th president of the United States. Historically, there tend to be common demographics among voters who support the Democratic or Republican party and by extension, the presidential candidates from these parties. In addition, voter demographics are strategically considered in presidential campaigns to appeal to certain citizens.

As a result, in this paper the estimand we hope to explore is the likelihood that an American citizen will vote for Joe Biden in the 2020 presidental election given their income, gender, race, and age. We will do this by using the 2020 Cooperative Election Study (CES) [CITATION] which is a national online survey that was administered to a representative sample of American citizens pre and post-election, gathering demographic information from 61,000 respondents. Using 8000 responses from this survey, we construct a logistic regression model using income, gender, race, and age as predictor variables to see which demographics aligned with support for Joe Biden. We find that (**I NEED YOUR MOST RELEVANT FINDINGS HERE**). This information is important as it can allow us to understand the campaign strategies in the 2024 United States presidential election toward citizens with specific demographics, especially

*Code and data are available at: https://github.com/FFFiend/linear_model_investigation.

as Donald Trump and Joe Biden will likely be representatives of their respective parties for the upcoming presidential election.

The remainder of this paper is structured as follows: Section 2 discusses the data from the 2020 CES, Section 3 discusses the logistic regression model that was constructed from the 2020 CES data, Section 4 presents the results, and finally Section 5 discusses our findings and some weaknesses. (**MOSTLY FROM https://tellingstorieswithdata.com/04-writing_research.html#introduction-1, DUNNO IF IT NEEDS REPHRASING OR GIVEN ADDITIONAL DETAILS LATER**).

2 Data

2.1 Data Source

In this paper, we will be using the 2020 Cooperative Election Study (CES) [CITATION], administered by YouGov, is a yearly online survey that gathers information regarding American adults and their political attitudes. This study was chosen because its dataset is publicly accessible, includes vote validation, and contains a diverse selection of attributes which make it possible to study how demographics of American adults correlates with preference for presidential candidate.

2.2 Data Measurement

Most, if not all variables in the 2020 CES dataset correspond to a question asked in the CCES online questionnaire. Most respondents to this questionnaire are YouGov panelists, but other respondents were recruited from online surveys or other survey providers. [CITATION]

However, in order to gain a representative sample of all American adults, not all respondents of the CCES questionnaire end up in the final dataset [CITATION]. The representative sample was created by using matching, "is a methodology for selection of "representative" samples from non-randomly selected pools of respondent" [CITATION] which is appropriate in this case as the pool of respondents is primarily made of YouGov panelists - which is nonrandom.

Lastly, "individual records were matched to the Catalist database of registered voters in the United States" to validate the votes. However, only respondents with a high confidence of being matched in the Catalist database were included. Thus, some respondents may have been omitted from the final dataset in this regard.

2.3 Data Features

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.5
v ggplot2   3.5.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

Table 1: A sample of CES 2020 survey data.

votereg	CC20_410	gender	educ	race	birthyr
1	2	1	4	1	1966
2	NA	2	6	1	1955
1	1	2	5	1	1946
1	1	2	5	1	1962
1	4	1	5	1	1967
1	2	1	3	1	1961

We use the R programming language (Team 2023) and dataverse [CITATION] to access the CES dataset which initially begins with 61,000 respondents. From this dataset, we begin with the following variables as shown in the [FIRST TABLE CROSS REFERENCE]. The values from these variables from the questionnaire corresponding to each variable can be found in the Appendix along with the value it was numerically coded to in the original dataset [CROSS REFERENCE: TABLE 1].

- votereg: This is a binary variable where the values 1 and 2 signify whether a person is registered to vote or not, respectively.
- cc20_410: This variable measures who the respondent voted for as President of the United States if they voted (with the values 1 and 2 corresponding to a vote for Biden and Trump, respectively) and values 4 through 7 that distinguish non-voters
- gender: This is a binary variable which measures if the respondent is male (1) or female (2)
- educ: This is a nominal variable which measures the level of education the respondent has ranging from 1 - "Did not graduate from high school" to 6 - "Postgraduate degree"
- race: This is a nominal variable which numerically codes races from the questionnaire to numbers in the dataset
- birthyr: This variable measures the year of birth of a respondent.

Since we are interested in registered voters who voted for either Biden or Trump, we filter the responses that have votereg equal to 1 and cc20_410 equal to 1 or 2. This leaves us with 43,554 responses. Also, because we are interested in the age of the respondents and not their year of birth, we create a new variable called "age" which subtracts the respondent's age from the year of the election, 2020, to get their age. While this may incorrectly represent the age of some people (since we do not have the month and day), for most, it will be an accurate representation of their age given that the election took place on November 4.

Lastly, we then process these responses and decode them to match with the questionnaire responses and obtain [CROSS REFERENCE: TABLE 2].

Table 2: Header for Processed 2020 CES survey data

voted_for	gender	education	race	age
Trump	Male	2-year	White	54
Biden	Female	4-year	White	74
Biden	Female	4-year	White	58
Trump	Male	Some college	White	59
Trump	Female	Some college	White	73
Trump	Female	High school graduate	White	50

[CROSS REFERENCE: FIGURE 1] shows that

3 Model

We resort to using a Logistic Regression model since we are interested in predicting the candidate a survey respondent voted for, given the former's age, gender, education and race. A logistic regression model is suitable as our response variables (1 for Biden and 0 for Trump) are clearly binary.

3.1 Model set-up

$$y_i|\pi_i, \sim \text{Bern}(\pi_i) \tag{1}$$

$$logit(\pi_i) = \beta_0 + \beta_1 \times gender_i + \beta_2 \times age_i + \beta_3 \times education_i + \beta_4 \times race_i \tag{2}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

$$\beta_4 \sim \text{Normal}(0, 2.5) \tag{7}$$

We define the response variable y_i to be the candidate that the ith respondent voted for. Note that the value of y_i is either 0 or 1 indicating whether the candidate in question voted for Trump or Biden respectively, and that y_i is conditioned on π_i and follows a Bernoulli distribution with parameter π_i

The explanatory variables $gender_i$, age_i, $race_i$ and $education_i$ are the selected attributes (with the same names, of) of the ith respondent.

We have defined normally distributed priors for the y-intercept β_0 as well as each of the coefficients of the explanatory variables β_1, β_2, β_3 & β_4 each with a mean of 0 and standard deviation of 2.5.

Note that the crux of our model, i.e $logit(\pi_i)$ is expressed as a linear combination of our explanatory variables along with the y-intercept β_0

Thus, the model enables us to estimate whether a candidate voted for Trump or Biden based on their education, gender, race and age. The logistic nature of the model ensures that any outputs or response variables obtained are either 0 or 1.

We run the model in R (Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.1.1 Model justification

Binary Outcome: logistic regression is specifically designed for binary outcomes, which is suitable for predicting whether a survey respondent voted for Biden (1) or Trump (0). Logistic

regression models the probability of the response variable falling into a particular category given the predictor variables.

Interpretability: Logistic regression provides easily interpretable results. The coefficients associated with each predictor variable represent the change in the log-odds of the outcome for a one-unit change in the predictor variable. This allows for clear interpretation of the effects of each predictor on the likelihood of voting for a particular candidate.

Efficiency: Logistic regression tends to perform well even with relatively small sample sizes compared to more complex models. Given that we have a limited set of predictor variables (age, gender, education, and race), logistic regression can efficiently model the relationship between these variables and the voting outcome.

Regularization Techniques: If necessary, logistic regression can be extended with regularization techniques such as L1 (Lasso) or L2 (Ridge) regularization to prevent overfitting and improve generalization performance, especially if there are many predictor variables or multicollinearity issues, which can be a useful strategy if overfitting has been observed upon analyzing the results of the model.

Statistical Inference: Logistic regression provides inferential statistics such as p-values and confidence intervals for the estimated coefficients, allowing you to assess the significance of each predictor variable in predicting the outcome.

Overall, logistic regression is a robust and interpretable method for modeling binary outcomes like voting preferences, making it a suitable choice for us to indicate what presidential candidate a survey respondent voted for.

4 Results

Our results are summarized in the table below:

As a note, the standard error for any value discussed below shall be abbreviated as sd-err.

Gender: The coefficient for Males is -0.45 (0.051 sd-err) which means Males are less likely to cast a vote for Biden compared to Females, or that Males would choose Trump in an election, given that Trump and Biden were the only candidates to choose from.

Education: We see that with an increase in literacy level, the likelihood that a candidate will vote for Biden increases. Observe that as we move upwards in the following list of categories: High school graduate, some college, 2-year, 4-year and post-grad, we see that the corresponding coefficients are -0.267 (sd-err 0.198), 0.054 (sd-err 0.195), 0.095 (sd-err 0.203), 0.445 (sd-err 0.196) and 0.906 (sd-err 0.199) respectively, meaning that the more educated a candidate is, the more likely they are to vote for Biden.

Race: Amongst the positive coefficients, observe that voters identifying on the survey as Black demonstrate the highest support for Biden amongst all other racial groups of respondents with

Table 3: Whether a respondent is likely to vote for Biden based on their gender, education, race and age (n=8000).

	Support Biden
(Intercept)	1.126
	(0.204)
genderMale	−0.454
	(0.051)
educationHigh school graduate	−0.267
	(0.198)
educationSome college	0.054
	(0.195)
education2-year	0.095
	(0.203)
education4-year	0.445
	(0.196)
educationPost-grad	0.906
	(0.199)
raceBlack	2.590
	(0.155)
raceHispanic	0.572
	(0.095)
raceAsian	0.635
	(0.162)
raceNative American	−0.241
	(0.259)
raceMiddle Eastern	0.356
	(0.181)
raceTwo or more races	−0.294
	(0.172)
raceOther	1.301
	(1.277)
age	−0.018
	(0.002)
Num.Obs.	8000
R2	0.126
Log.Lik.	−4829.465
ELPD	−4845.5
ELPD s.e.	33.6
LOOIC	9691.1
LOOIC s.e.	67.2
WAIC	9690.1
RMSE	0.46

a coefficient of 2.590 (sd-err 0.155), whereas Hispanic and Asian respondents bear coefficients 0.572 (sd-err 0.905) and 0.635 (sd-err 0.162) respectively. Coming in last place, we have Middle Eastern respondents with the smallest positive coefficient compared to the other racial groups, with a value of 0.356 (sd-err 0.181). Finally, we have respondents identifying with "Two or more races" that are actually LESS likely to support Biden, with a value of -0.294 (sd-err 0.172). Finally, we observe that respondents identifying as "Other" demonstrate a coefficient of 1.301, although with a comparatively higher standard error of 1.277.

Age: We see a coefficient value of -0.018 (sd-err 0.002) showing a weak negative correlation between the age of a respondent and whether they voted for Biden. All in all, it is bold to claim that there exists a definite negative correlation between a voter's age and whether they voted for Biden.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

B.2 Diagnostics

References

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