# 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

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].

[CROSS REFERENCE: FIGURE 1] shows that

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