Investigating the correlation between a vote for Biden or Trump based on a persons education, gender, race and age.*

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

We are concerned with assessing whether we can predict what candidate a new survey respondent is going to vote for, given their age, education, gender and race. Accordingly, we shall be simulating a similar scenario for the 2020 US presidential election below.

2 Data

We will be using the 2020 Cooperative Election Study (CES) (CITATION TODO) as our dataset, which can be previewed below:

Table 1. Header for raw election 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

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

votereg	CC20_410	gender	educ	race	birthyr
1	1	2	5	1	1962
1	4	1	5	1	1967
1	2	1	3	1	1961

To elaborate further on the dataset starting with the `votereg` column, values 1 and 2 signifies whether a person has voted or not, respectively. Naturally, for entries with a value of 2 under this column, the corresponding `CC20_410` value is NA, and in the other case the values 1 and 2 signify a vote for Biden or Trump respectively. The naming of this column corresponds to the question in the survey from which the data for it was collected, which asks "For whom did you vote for President of the United States?" (citation needed), with a total of 6 options, where Biden and Trump correspond to the first two.

Next, there are two values for gender: 1 and 2 corresponding to Male and Female. There are also a total of 8 categories under race corresponding to numbers 1 through 8, as well as 6 education categories ranging from "No HS" to "Post-grad".

Since we are interested with ages and not respondent birth years, we may process the dataset as such to reflect this:

`Table 2. Header for 2020 US Election Survey data with the birthyr column converted to ages`

votereg	CC20_410	gender	educ	race	age
1	2	1	4	1	54
2	NA	2	6	1	65
1	1	2	5	1	74
1	1	2	5	1	58
1	4	1	5	1	53
1	2	1	3	1	59

We shall now filter for non NA values within the `CC20_410` column, as we are only concerned with respondents that cast a vote, which also means we filter for entries with a `votereg` value of 1. Additionally, we shall map each education level and race value to its corresponding value from the survey, and we finally obtain a table as follows:

`Table 3. Header for Processed 2020 US Election 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

TODO: complete bar chart section from the worked example.

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 (**citeR?**) 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:

```
Table 4. Whether a respondent is likely to vote for Biden based on their gender, education,
```

Gender: Test

Education: Test

Race: Test

Age: Test

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.

	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

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

C References