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 Simulation

We first start with simulating the dataset, and recall that we are concerned with each voter's education level, gender, race and age. Our chosen sample size is 8000, and the categories for education and race have been named akin to the categories in the dataset. Thus, the header of the table obtained for the simulation is shown below:

Table 1. Header for the political preferences dataset.

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

education	gender	race	age	supports_biden
Some college	Female	Black	14	yes
College	Male	Two or more races	36	yes
< High school	Female	Black	77	yes
High school	Female	White	80	yes
College	Female	Middle Eastern	20	yes
High school	Female	Asian	64	yes

3 Data

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

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

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 3. 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 4. 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.

4 Model

Since we wish to predict how likely it is for a respondent to cast a vote for Biden based on their gender, age, education and race, we shall use the rstanarm (citation needed) library to

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

4.1 Model set-up

We define y_i as the candidate the *ith* respondent likely voted for, mapping 1 and 0 to Biden and Trump respectively. The variables $gender_i$, age_i, $race_i$ and $education_i$ are the selected attributes (with the same names, of) of the *ith* respondent.

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

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

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

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

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

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

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

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

4.1.1 Model justification

I actually have no idea what to expect lol

5 Results

Our results are summarized in **?@tbl-modelresults**.

`results`

x
\begin{table}

	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

\end{table} |

6 Discussion

6.1 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.

6.2 Second discussion point

6.3 Third discussion point

6.4 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

C References