Election Preferences in 2022 based on Gender and Education*

Rahul Gopeesingh

March 16, 2024

There are many factors that can influence someone's voting preference. Two of which that seem very relevant are a person's gender as well as their education level. This paper uses a logistic regression to investigate the relattionship between these factors and a person's political preference. We found that both females and highly educated people are significantly more likely to vote for Biden than for Trump

Table of contents

*Code and data are available at:https://github.com/Rahul-Uoft/election.git

1 Introduction

In this paper we investigate voter preferences of different demographics. One major factor that is believed to influence a persons political preferences is their gender. This is largely due to the gender gap and different governments taking different actions with respect to these differences. The next we look at is the highest education reveived in order to predict their political preference.

Since we want to investigate how these factors affect people presently, we look at the latest available dataset(2022) of the 2020 election between Biden and Trump. Here we are trying to estimate the log-odds that someone is more likely to vote for Trump than Biden.

We use R Core Team (2023) and Wickham et al. (2019) in order to conduct this analysis

The remainder of this paper is structured as follows. Section 2 discusses the data, how it was collected and the purpose in which we use this data. It also gives a brief analysis of the data. Section 3 walks through the logistic regression model used and the justification for using this logistic regression. Section 4 shows the results and an intepretation of what they mean. Finally, Section 5 discusses the potential reasons behind these results and their implications.

2 Data

2.1 Data Source

The data used in this paper is contained in the Cooperative Election Study 2022.@dataset This study examines American's views on representatives, electoral experiences and elections. This data is available for free at the Harvard University Dataverse.

Surverys were used to collect data from a sample large enough that it may be assumed is representative of the entire American population. Importantly, this covered a wide range of constituencies and the sample was large enough such that the data for each constituency is enough to assume that it is representative of its whole.@dataset There were 60 research teams, and each purchased 1000 surverys to hand out to their respective constituencies. Sample matching was used to pick the 1000 people in which the surverys were to be administered. This is a process which is not truly random but has the benefit of being practical. In a truly randomly selected group, the contact information for some may not be readily available and

therefore makes the data collectin process much more difficult. Sample matching involves selecting a sample at random, and anyone who does not fit the criteria to proceed(in this case having your contact information available) is then replaced by a person in the pool who is believed to have similar characteristics.

3 Model

We use the Bayesian Analysis model to conduct this by defining a linear relationship between our outcome variable and our predictor variables. These predictor variables are then assigned a distribution and we use these to predict the outcome variable.

3.1 Model set-up

Define y_i as the political preference of the individual and is 1 if the individual prefers Biden and 0 if the person prefers trump. Then gender_i is the individual's gender and education_i is the individual's education

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \alpha + \beta_1 \times \text{gender}_i + \beta_2 \times \text{education}_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
& && (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

A logistic regression model was used in this analysis as the outcome variable is Binary. This is because we are considering whether someone prefers Biden or Trump. Whenever the outcome variable is Binary, a logistic regression model makes sense to help us understand it more. We used the default priors for our input variables.

A logistic regression model works by instead of considering an error value, which is done in linear regression, it considers a distribution for each of the inputs. The variability of these distributions inherently create the variability of the outcome which is associated to the error value we get in linear regression. The main advantage to this comes from the assumption

made in linear regression models in which the error value is assumed to cancel out with each other(forming a normal distribution)

It is also important to note that we only consider votes for Biden and Trump as voting for any other candidate is generally seen as giving up one's vote as the other parties are extremely unlikely to win. Hence we can consider this Binary observation of Biden vs Trump.

4 Results

Our results are summarized in Table 1. This table includes values for the intercept under various conditions as well as their error value.(the value in parentheses below each intercept.) Each value of the intercept indicates the estimated log-odds of the outocme variable. What this means is that the log of the outcome is given as the intercept. Therefore by taking the exponential function we can predict the likelihood of a certain outcome given a certain variable. For instance, we can see the intercept for Post-grads were -0.94. taking the exponential function of this gives a result of 0.39. This means that a random post-grad is 0.39 times as likely to vote for trump as they are for Biden. We are not concerned with the second half of the table.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	−0.22
	(0.06)
genderMale	0.45
	(0.02)
educationHigh school graduate	0.07
	(0.06)
educationSome college	−0.31
	(0.06)
education2-year	−0.28
	(0.06)
education4-year	−0.61
	(0.06)
educationPost-grad	−0.94
	(0.06)
Num.Obs.	47 466
R2	0.037
Log.Lik.	−31 245.082
ELPD	−31 252.1
ELPD s.e.	55.6
LOOIC	62 504.2
LOOIC s.e.	111.1
WAIC	62 504.2
RMSE	0.48

5 Discussion

5.1 Female Prefereces

We observe that females are 0.8 times as likely to vote for Trump as they are for Biden. This follows a trend that has been observed since the beginning of the 21st century, women prefer the democratic party.@article1 This may be due to the history of the democratic party and its advocation for women-rights. It appears that the democratic party is the one that serves women best in terms of their economic interest. In fact the democratic party has paid a lot more attention to issues that directly pertain to women such as abortion and family leave than the republicans and there is evidence to suggest that this is a key part of its strategy. Kahn (2020)

5.2 Highly Educated People

One of the most apparent outcomes of this model would be the likelihood of highly educated people to prefer voting for Biden over Trump. One cause of this can be the fact that college graduates are far more likely to self-identify as liberal.@article3 It could also pertain to the way in which most colleges are structured. It is possible that the values taught there are aligned with those of the democrats and hence shape the thinking of these highly educated people in a way that increases their likelihood of voting for a specific candidate.

The proportion of graduates that make up the democratic voting pool has greatly increased over the last 20 years.@article3

5.3 Political Strategy

The most interesting part is how these observations apply to the political strategy of the two parties. Firstly, it was argued that the majority of females voting for the democrats were enough to sway the elections and therefore the main focus of the two parties, however this was disproven in 1980 and 1984 as the republican party won without fully closing the gender gap. They appealed to broad groups of women through issues that directly affected them and were not specific to women such as national security.@article2

Additionally, with the increased amount of liberal, educated individuals arguing for various policies that inconvenience some working class individuals, it creates an opportunity for the republican party to increase its share of this demographic in an attempt to balance the disproportionality. An exaple of this may be the advocation for an environmentally friendly policy that negatively impacts the working class people in this field. (Cohn (2021))

5.4 Weaknesses and next steps

A major weakness of this analysis is that it assumes that Biden and Trump are choices made because of preferences. However, as explained in the model section, these are the only rational choices as any other choice seems to be irrelevant. Therefore someone may vote for a candidate not because they resonate with their political ideals, but rather because they are the 'lesser of two evils.'

One way in which this can be analyzed is by considering someones thoughts on different administrations throughout the years or using a benchmark to determine how much they genuinely believe in the political ideals of a certain party.

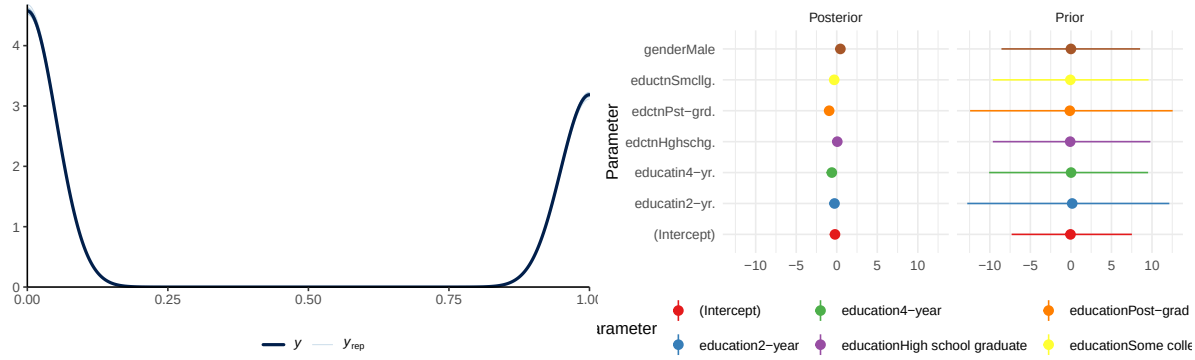
Appendix

A Model details

A.1 Posterior predictive check

In Figure 1a we implement a posterior predictive check. This shows the outcome variable with simulated variable from the posterior distribution.

In Figure 1b we compare the posterior with the prior. This shows how much the estimates change once the data is taken into account.



(a) Posterior prediction check

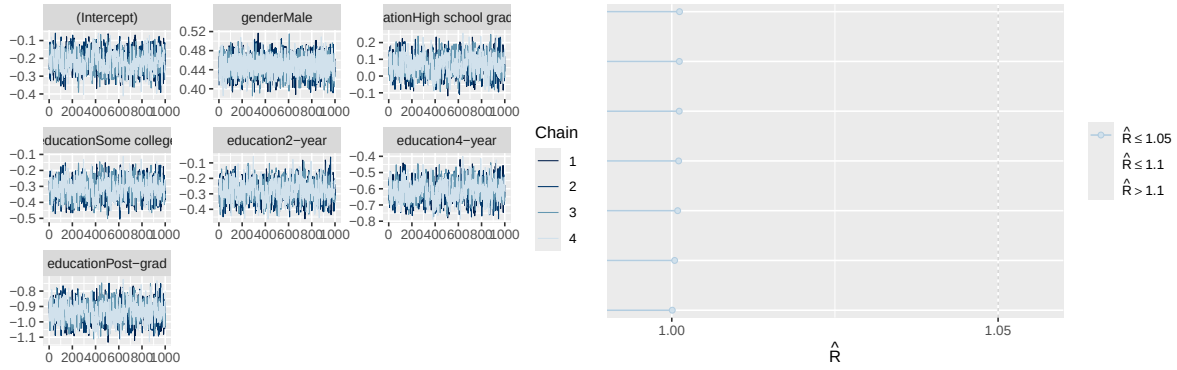
(b) Comparing the posterior with the prior

Figure 1: Examining how the model fits, and is affected by, the data

A.2 Diagnostics

Figure 2a is a trace plot. It shows lines that appear to bounce around and are horizontal. This suggests nothing wrong with the model we used.

Figure 2b is a Rhat plot. It shows every value close to 1 and nothing greater than 1.1 This suggests the model we used is fine.



(a) Trace plot

(b) Rhat plot

Figure 2: Checking the convergence of the MCMC algorithm

References

Cohn, Nate. 2021. "How Educational Differences Are Widening America's Political Rift." https://www.nytimes.com/2021/09/08/us/politics/how-college-graduates-vote.html.

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

Kahn, Suzanne. 2020. "Women Tend to Vote for Democratic Presidential Candidates More Than Men Do. Here's How That Gender Gap First Came to Be." https://time.com/5903399/gender-gap-politics.

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