Forecasting 2020 US Election Outcomes: The Role of Employment, Education, gender, and Race*

Sean Liu

March 16, 2024

This study employs logistic regression modeling on data from the Cooperative Election Study (CES) to forecast the 2020 US presidential election results. Focused on predicting voting choices between Trump and Biden, the analysis incorporates variables like employment status, education level, gender, and race. The research aims to gauge the probability of victory for each candidate, with a particular interest in assessing actual support levels. The dataset, sourced from CES, offers insight into voter behavior and demographic trends during this historic election.

[*] Code and data are available at: https://github.com/Sluuu/Politics.git

1 Introduction

Back in 2020, the US was having an important presidential election contest between the Democratic party and the Republican party. The Democratic ticket of former vice president Joe Biden and the Republican president, Donald Trump. The election saw the highest voter turnout by percentage since 1900, with each of the two main tickets receiving more than 74 million votes. In particular we are interested in whether we can forecast who a respondent is likely to vote for, based on knowing their employment status, highest level of education, and race. That means we are interested in a data-set with variables for who an individual voted for, and some of their characteristics, such as employment status, education level, gender, and their race. The data is from the 2020 election data set from Cooperative Election Study (CES) (Schaffner, Ansolabehere, and Luks 2021).

In this study, we used a logistic regression model to predict the 2020 election results, utilizing data from the Cooperative Election Study (CES). This model is well-suited for binary outcomes, such as predicting whether an individual voted for Trump or Biden. Our analysis aims to assess the probability of victory for each candidate, considering a variety of factors including employment status, education level, gender, and their race. The primary objective is to determine the actual support levels for Trump and Biden.

The paper is organized into distinct sections. In Section 2, we present the dataset utilized in our study, by assigning graphs to show various respondent groups. Section 3 outlines the development and rationale behind our model. Section 4 showcases prediction outcomes through tables and graphs. Lastly, Section 5 delves into discussions based on our findings, based on examining voting predictions relative to race, educational level, employment status, and gender.

2 Data

2.1 Source of Data

Our primary data is from the 2020 Cooperative Election Study, CES Schaffner, Ansolabehere, and Luks (2021). The data includes a nationally representative sample of 61,000 American adults. Schaffner, Ansolabehere, and Luks (2021) includes the data from the survey, a full guide to the data, and the questionnaires.

?@fig-education shows the relationship between respondents' educational level, gender and their voting preferences. In both gender, respondents with higher educational level tends to lean more towards Biden.

?@fig-race shows the relationship between respondents' race, gender and their voting preferences. For males, individuals identifying as Black, Asian, or Hispanic supported more towards Biden. For white males the numbers of supporters between Biden and Trump are close. For female respondents, almost all kind of race are more likely to support Biden.

?@fig-employment__status shows the relationship between respondents' employment status, gender and their voting preferences. For male full-time respondents, they are more likely to support Biden. For the rest of the male respondents, the number of supporters between Biden and Trump are close. For female respondents, regardless of the employment status, most of the them are supporting Biden.

3 Model

3.1 Model set-up

The model that we are interested in is:

[

$$
\begin{align}
y_i|\pi_i &\sim \text{Bern}(\pi_i) \tag{1}\\
\text{logit}(\pi_i) &= \alpha + \beta_1 \times \text{employment\_status}_i + \beta_2 \times \text{race}_i + \beta_3 \times \text{education}_i + \beta_4 \times \text{gender}_i \tag{2}\\
\alpha &\sim \text{Normal}(0, 2.5) \tag{3}\\
\beta_1 &\sim \text{Normal}(0, 2.5) \tag{4}\\
\beta_2 &\sim \text{Normal}(0, 2.5) \tag{5}\\
\beta_3 &\sim \text{Normal}(0, 2.5) \tag{6}\\
\beta_4 &\sim \text{Normal}(0, 2.5) \tag{7}
\end{align}
$$

]

Where y_i is the political preference of the respondent, it equals to 1 if the respondent voted for Biden and 0 for Trump, $\text{employment\_status}_i$ is the employment status of the respondent, race_i is the race of the respondent, education_i is the highest educational level of the respondent, and lastly gender_i is the gender of the respondent. The coefficients (β_1, β_2, β_3) are being assumed as a normal distribution with a mean of 0 and a standard deviation of 2.5 for each parameter. Lastly, pi_i is the probability of voting for Biden. We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022), `dplyr` package of Wickham et al. (2023), `Knitr` package of Xie (2014), `Tibble` package of Müller and Wickham (2023), `Tidyverse` package of Wickham et al. (2019), `ggplot2` package of Wickham (2016), and `modelsummary` package of Arel-Bundock (2022).

3.2 Model justification

We anticipate a positive correlation between individuals identifying as Black, Asian, or Hispanic and their support for Biden. As Biden's proposal to more advanced racial equity in the US (Sprunt 2020). Most traditional white voters are more likely to support Trump. As Trump's cultural and the atmosphere of Make America great again. Which tend to these group of people to support him (Gene Demby 2020).

For the education level, we expect respondents who has a higher educational level to be leaning towards to support Biden. As people with higher educational levels are more open to diverse perspectives and the new directions in education that Biden proposed (Wong 2020).

Lastly the voting preferences for different employment status voters and the difference in gender are harder to determine.

4 Results

Our results are summarized in Table 1.

Overall, our results were mostly matching with our predictions. Which individuals identifying as Black, Asian, or Hispanic and their support for Biden. Which White are more likely to support Trump. For the education level, the respondents who has a higher educational level to be leaning towards to support Biden. For gender, we observed that males are way more likely to vote for Trump than females. Lastly, the voting preferences for different employment status voters have huge differences between different employment status.

Table 1: Explanatory models of voting preferences

	Supported Biden
(Intercept)	0.718
	(0.228)
employment_statHomemaker	−0.141
	(0.096)
employment_statPart-time	0.090
	(0.078)
employment_statPermanently disabled	0.162
	(0.095)
employment_statRetired	−0.151
	(0.056)
employment_statStudent	1.061
	(0.178)
employment_statTemporarily laid off	0.322
	(0.155)
employment_statUnemployed	0.657
	(0.107)
educationHigh school graduate	−0.104
	(0.175)
educationSome college	0.456
	(0.173)
education2-year	0.223
	(0.178)
education4-year	0.803
	(0.170)
educationPost-grad	1.206
	(0.181)
raceBlack	1.291
	(0.194)
raceHispanic	−0.292
	(0.180)
raceMiddle Eastern	−0.595
	(0.222)
raceNative American	−0.864
	(0.306)
raceTwo or more races	−1.286
	(0.224)
raceWhite	−0.869
	(0.161)
genderMale	−0.405
	(0.046)
Num.Obs.	9780
R2	0.115
Log.Lik.	−5979.111
ELPD	−5999.6
ELPD s.e.	36.2
LOOIC	11 999.2
LOOIC s.e.	72.5
WAIC	11 999.1
RMSE	0.46

5 Discussion

5.1 Weaknesses and Next Step

The first weakness/limitation is that our regression only focuses on Trump and Biden. Which didn't go beyond other candidates.

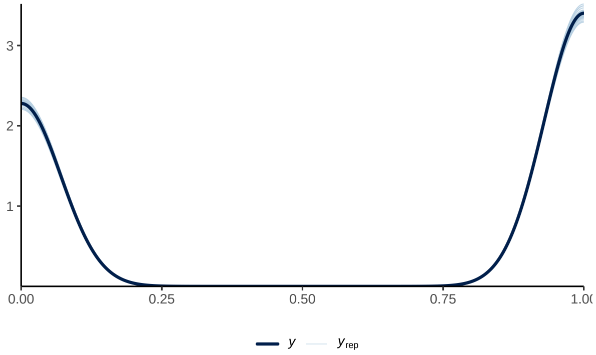
The second weakness we have was that our control variables were too less. Which might not fully capture the preferences.

Lastly, there might exist some sampling bias in our study. As our sample size is around 43554 respondents which does not even reach 1% of our total voter population, it is possible that the results from the survey respondents might not fully represent the whole population. In future studies, we should try to increase our sample size by raising the amount of target population. Besides, we hope to incorporate more demographic variables into our research. This helps us to have a more accurate prediction because the variation of prediction outcomes can be explained more comprehensively by the predictors we add.

A Appendix

A.1 Posterior predictive check

In Figure 1 we implement a posterior predictive check.



(a) Posterior prediction check

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

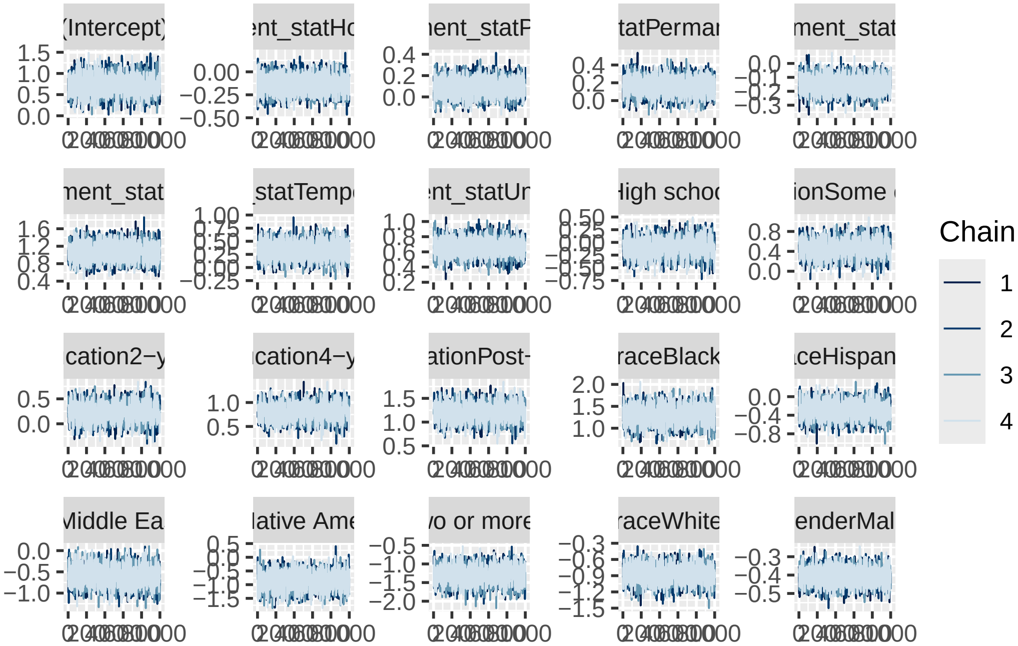
A.2 Diagnostics

Figure 2 is a trace plot. It shows a graphical analysis of Markov chain Monte Carlo (MCMC) simulations. In our graph the lines does not suggest anything out of the ordinary, meaning the chains are mixing well and the MCMC algorithm and is converging into the correct distribution.

Figure 3 is a Rhat plot. It is used to assess convergence in Markov chain Monte Carlo (MCMC) simulations. In our graph values are close to 1 and not over 1.1. This indicates that the chains are converging well to the target distribution.

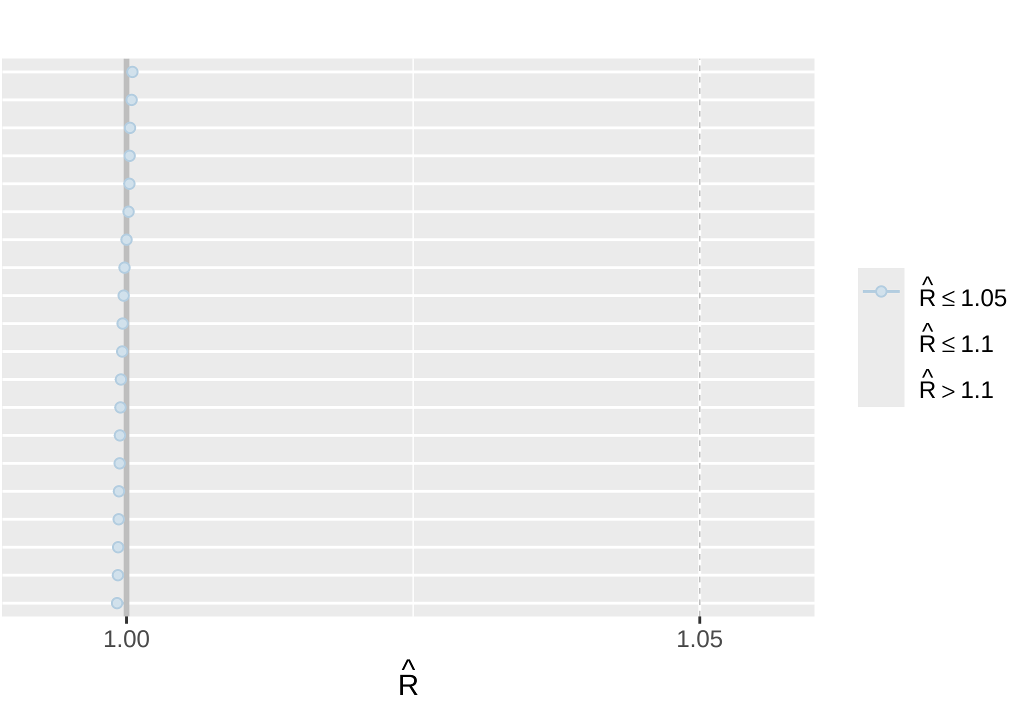
A.3 Creditibility Intervals

Figure 4 shows the 90% Credible intervals for predictors of support for Biden



(a) Trace plot

Figure 2: Checking the convergence of the MCMC algorithm



(a) Rhat plot

Figure 3: Checking the convergence of the MCMC algorithm

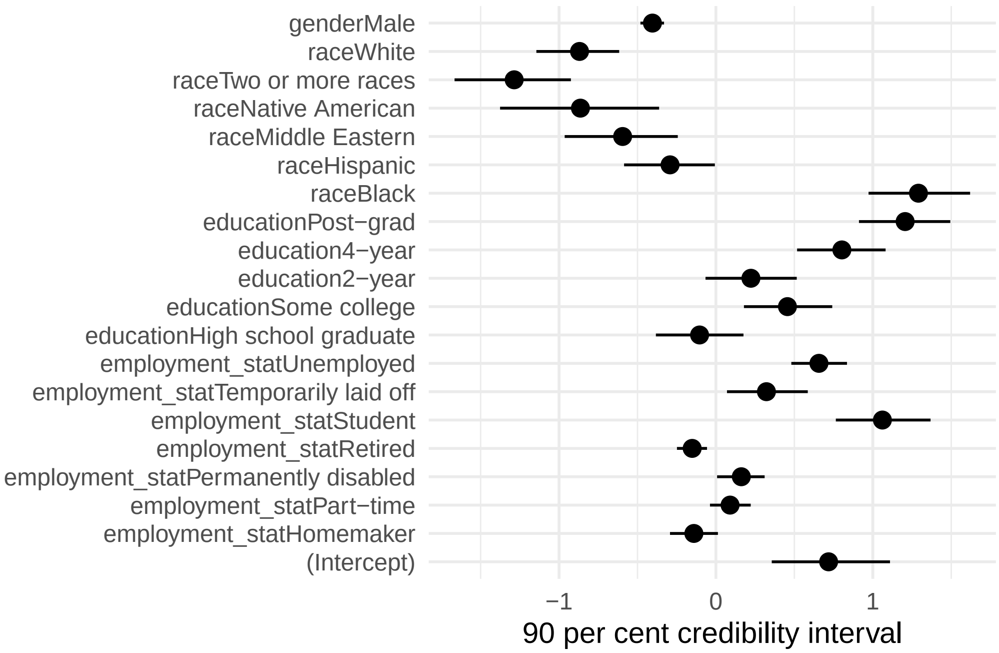


Figure 4: Credible intervals for predictors of support for Biden

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Gene Demby, Shereen Marisol Meraji. 2020. "The White Elephants in the Room." *Npr*. https://www.npr.org/2020/11/17/935910276/the-white-elephants-in-the-room.

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

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

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

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

Sprunt, Barbara. 2020. "Biden Details Proposal to Advance Racial Equity in America." *Npr*. https://www.npr.org/2020/07/28/896132054/biden-details-proposal-to-advance-racial-equity-in-america.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wong, Kenneth K. 2020. "The Biden Presidency and a New Direction in Education Policy." *Brookings*. https://www.brookings.edu/articles/the-biden-presidency-and-a-new-direction-in-education-policy/.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.