My title*

My subtitle if needed

Jeongwoo Kim Jiwon Choi

March 14, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section 2....

2 Data

We have used two datasets for this study. One is the U.S election survey data of Democracy Fund + UCLA Nationscape dataset from the Voter Study Group, conducted on October 3, 2019. Second is the census data from IPUMS America Census Service, which is used as the post-stratification data for the survey data to adjust the weight.

2.1 Survey Data

This survey data is an 18-month election study conducted by UCLA researchers with roughly 6250 online interviews each from from July 2019 to February 2021 Tausanovitch and Vavreck (2020). The sample is weighted to represent the U.S. adult population Tausanovitch and Vavreck (2020). Nationscape groups weight on the following important factors: gender, the four major census regions, race, Hispanic ethnicity, household income, education, age, language spoken at home, nativity, 2016 presidential vote, and the urban-rural mix of the respondent's ZIP code Tausanovitch and Vavreck (2020). According to the data, Male make up 48.3% while

*Code and data are available at: https://github.com/Kjeongwoo99/STA302H_Paper3

female make up 51.3% Tausanovitch and Vavreck (2020). 74.2% of the respondents are White, 6.8% are Asian/Pacific, 12% are Black Tausanovitch and Vavreck (2020). 20.4% are those between 18-29, 33.4% are 30-49, 32.4% are 50-69, 3.3% are 70+ Tausanovitch and Vavreck (2020). On average, 5.1 percent declined immediately among those who are selected for the survey. 16.7 percent of the respondents did not complete the survey. Another 5.9 percent were categorized as speeding or straight-line which means they completed the survey in less than 6 minutes or selected the same response for every question in the three policy question batteries. Leaving these out leave 72.4 percent of the original sample for the analysis.

The Nationscape survey's strength lies in its methodological rigor - the effectiveness in collecting large samples from the U.S. citizen and its weighting strategy desinged to mirror the U.S. adult population by including weight factors such as age, gender, race and income and more. As they filter out inaccurate or missing data, it makes sure that the data collected are accurate and ensures data integrity. While other datasets such as the General Social Survey (GSS) and the American National Election Studies (ANES) are available, the Nationscape dataset's frequency (surveys collected every week) give it an advantage in analyzing electoral trends and shifts in real-time. Its' extensive sample size also justifies the choice of this dataset.

For our analysis, we decided to focus on five demographics: age, gender, education, race and state. Age is important because in general, voters tend to become more conservative as they get older. To account for the age difference, we divided the age group into four categories: 18-29, 30-49, 50-69 and 70+.

Gender is also an important category because in general, men tend be more conservative and women tend to be more liberal. Recently, gender issues are growing social issues and this may affect the election, hence we wanted to explore how this affects our model.

Education is also an interesting factor. In the past, non-college white voters used to support Democrats while college-educated white voters supported Republicans Harris (2018). However, there has been a switch in this trend as 61 percent of non-college white voters showed their support wheres just 45 percent of college-educated white voters did in the exit polls Harris (2018). Only 37 percent of those without a degree cast their votes for Democrats while 53 percent with a degree did so Harris (2018). We categorized education into four categories: 'High school or less', 'Some college', 'College degree', 'Postgrad'.

Race also needs some attention because normally non-white groups are highly in favour of Democrats regardless of candidates and white swing by depending on candidates. According to the statistics collected in 2016, 93% of black, 71% of Latino, 68% of Asian support democrats while only 41% of white support democrats Prokop (2021). As white voters make up 74% of the voting population, it is really important for both parties to attain this demographic group.

Lastly, states are very important as some states historically favor conservatives while some states vote for democrats. In general, the west and the east coasts are democrat supporters whereas south are conservative supporters.

2.2 Post-stratification Data

IPUMS ("Integrated Public Use Microdata Series") is a website that offers database of samples of the American population from the American Community Surveys of 2000-present. These samples provide rich qualitative information on the long-term changes in the population. We selected the ' ' data as the post-stratification dataset for our research. The ACS is an ongoing survey that collects data monthly, which is then combined into 1-year, 3-year, and 5-year aggregates. It then uses stratified sampling where the U.S population is broken down into sub-groups and initial weights are assigned to each respondent.

One strength of the IPUMS survey is the fact that it provides a comprehensive and extensive data with detailed demographic of the U.S. population with social, economic and housing characteristics, which is very useful in our analysis of the 2020 U.S presidential election forecast. The longitudinal data of this survey also allows researchers to analyze trends over time. The U.S. Census Bureau offers credibility of the data with high quality checks. The post-stratification process ensures correcting for sampling biases and non-response. On the other hand, since the survey relies on self-report, there lies a risk of response bias inherently. While it is an ongoing survey, there is still a time lag between the data collection and data availability. However, the large sample size, consistency and reliability of the data collection, the integrated data over time with post-stratification can justify the decision to utilize IPUMS data over other sources.

The variables we have decided to use in this analysis from the dataset are 'sex', 'race', 'stateicp', 'age_group', and 'educd'. We have filtered out the respondents who have answered 'other' or no available sex data from the raw dataset for simplicity of the analysis. Therefore, 'Sex' is categorized into 'Male' and 'Female'. 'Race' is divided into 'White', 'Black', 'Asian', 'American Indian' and 'Other'. As mentioned earlier, the fact that white, black and Asian together make up around 93 percent of the U.S population, this categorization is justified. 'stateicp' is a record of all the states in the U.S. We have used the abbreviations for each state, for example, 'CT' for Connecticut, 'ME' for Maine, 'MA' for Massachusetts. We have the data of 55 states in total. 'age_group' categorized the respondents' age into four groups: '18-29', '30-49', '50-64' and '65+'. 'education' is divided into four categories: 'High school or less', 'Some college', 'College degree', 'Postgrad'. We filtered out those unknowns, and cleaned the dataset by assigning each respondent to the right category accordingly.

3 Model

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.

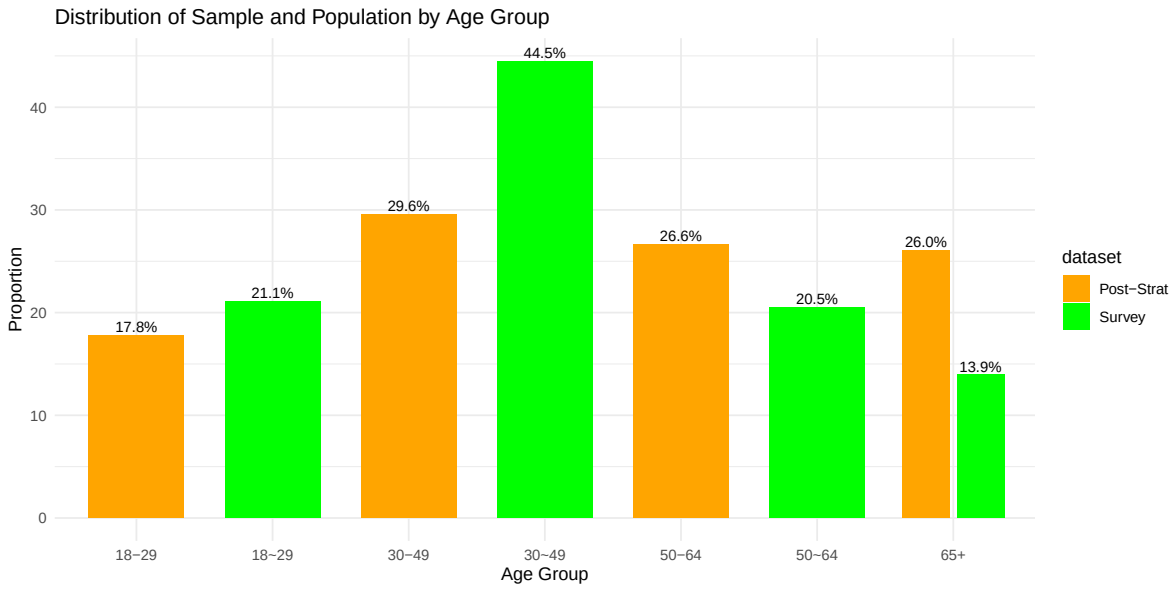

Figure 1: Distribution of Sample and Population by Age Group

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

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`.

4 Results

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

5 Discussion

5.1 Analyzing Voter Preferences Across Demographics

The analysis reveals critical insights into how education, age, race influence voting behavior. Contrary to conventional wisdom and past electoral analyses, our findings indicate a divergent pattern in voting behavior among different educational groups. The shift in party allegiance among non-college white voters from Democrats to Republicans and the nuanced preferences of voters with 'Some college' and 'Postgrad' education levels challenge the stereotypical narratives that educational attainment leads to an increase in support for the Democrats. This group, especially highly educated 'Postgrad' group being the second highest supporters suggest a reevaluation of Democrats' strategies to appeal to educated voters concerned with economic policies and national security. However, according to our data, the Republicans fail to attract voters as highest proportion of support among the four educational groups remain under 40 percent.

?@tbl-voters-intention-to-vote-for-trump suggests that

shows that majority of Biden's support

The increasing propensity of older voters to lean towards the Democratic nominee, potentially driven by concerns over COVID-19 and healthcare, marks a significant shift in voting patterns. Similarly, the racial dynamics observed, with non-white demographics showing strong support for the Democratic party, reflect broader national conversations on race, immigration, and identity.

5.2 Geographical Divides and Electoral Preferences

From **?@fig-distribution-by-state**, we can observe that Trump leads in mid-west of America, Montana (MT), Wyoming (WY), Mississippi (MS), Maine (ME), Missouri (MO), Kentucky (KY), and Utah (UT). Even the competing states like Louisiana (LA), New Mexico (NM), Tennessee (TN) are also mid-west part of the country, which are rural areas with low population density. This is correctly estimated since in the real world as Trump won in these states: MT, WY, MS, MA, MI, KY, UT. He also won in LA and TN, which we considered were competing states. We can go more in depth of this study with the benefit of hindsight. We estimated that the biggest victory for Trump would be Montana (MT) with just below 70%. In reality, he won by 59.6%, which is lower than we thought but still sit within our error values. The next state with the highest proportion of support for Trump was Wyoming (WY) from our model with 65%, which is close to the election poll result of 69%. We were very successful with Mississippi (MS) with an estimation of 57% and the poll result of 57.5%. Maine (ME) was another very close estimate with 52% and the poll result of 53.1%. 56.8% voted for Trump in Missouri (MO), which was slightly higher than our estimation of 51%. We estimated that only 26% of voters in Hawaii (HI) would vote for Trump with error margins of +-10%, which

matched 34.3% for Trump in the election. 41.9% for Trump in Colorado (CO) lie within our estimate of 37% with +-5% margins.

There were some cases where our estimation was incorrect in terms of who won the state but the results lie within our error margins. For instance, we estimated Utah (UT) in favor of the Democrats with a slight margin of 51% for Biden but Biden only received 37.6% in the election. This still lies within our error values that we estimated to be from 43% to 58%. However, there were some errors in our estimation where our model failed to predict the Trump winning in the number of states in general. Our model predicted Louisiana (LA) for Biden with 51% but he only received 39.9% of the votes. This value is lower than our error margins. 62% of the residents in Alabama, a Republican stronghold in the last eleven presidential elections had voted for Trump while our model estimated 45% of votes for Trump. The republicans won in Florida (FL) with a slight margin of 51.2% but we predicted an estimation of 45% with +-2% margins. What we estimated to be the strongest Democrat state was North Dakota (ND) with an estimation of only 23% with +-12% margins but in reality 65.1% of the residents voted for Trump.

We were able to estimate the winning of states with some success

5.3 Weaknesses and Implications for Future Research

Weaknesses and next steps should also be included.

5.4 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.

5.5 Second discussion point

5.6 Third discussion point

5.7 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…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected by, the data

Figure 2: **?(caption)**

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

Figure 3: **?(caption)**

References

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

Harris, Adam. 2018. "AMERICA IS DIVIDED BY EDUCATION." *The Atlantic*. https://www.theatlantic.com/education/archive/2018/11/education-gap-explains-american-politics/575113/.

Prokop, Andrew. 2021. "A New Report Complicates Simplistic Narratives about Race and the 2020 Election." *Vox*, May. https://www.vox.com/2021/5/10/22425178/catalist-report-2020-election-biden-trump-demographics.

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

Tausanovitch, Chris, and Lynn Vavreck. 2020. *Democracy Fund + UCLA Nationscape*. October 10-17, 2019 (version 20200814). https://www.voterstudygroup.org/publication/nationscape-data-set.

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.