My title*

My subtitle if needed

Jeongwoo Kim Jiwon Choi

March 13, 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%

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

Table 1: Voters Intention to Support Trump

Response	Number of Respondents	Proportion (%)
Yes	1908	34.25
No	3080	55.30
Other	582	10.45

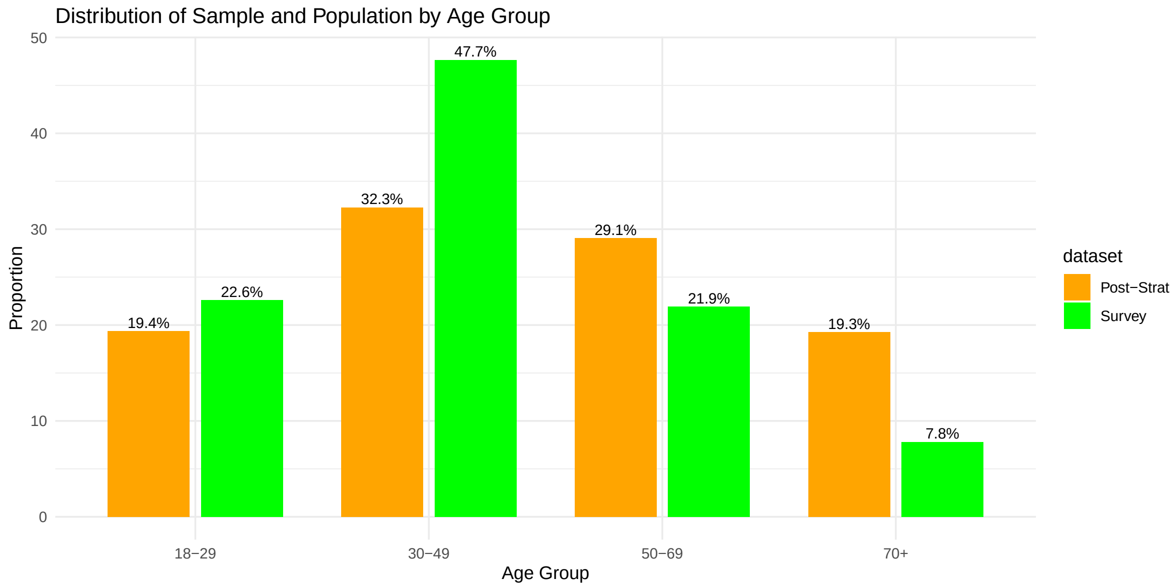

Figure 1: Distribution of Sample and Population by Age Group

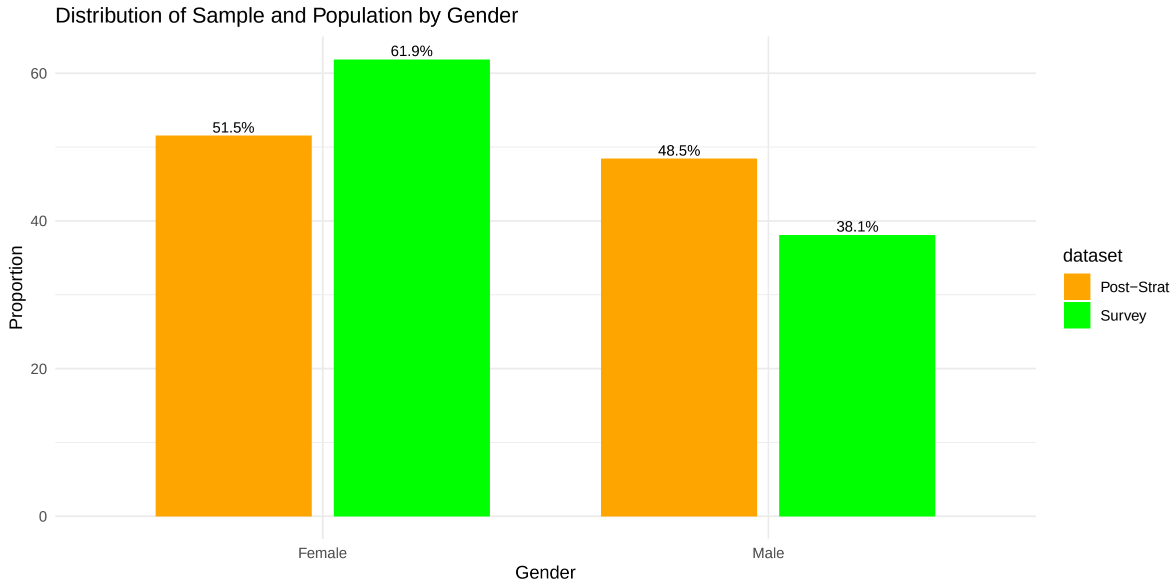

Figure 2: Distribution of Sample and Population by Gender

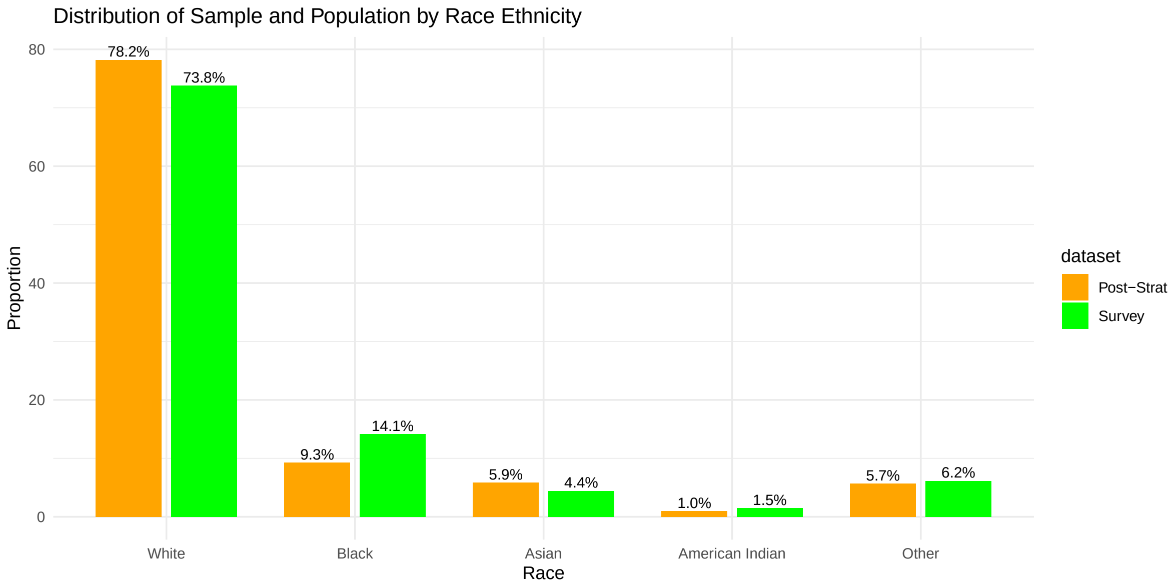

Figure 3: Distribution of Sample and Population by Race Ethnicity

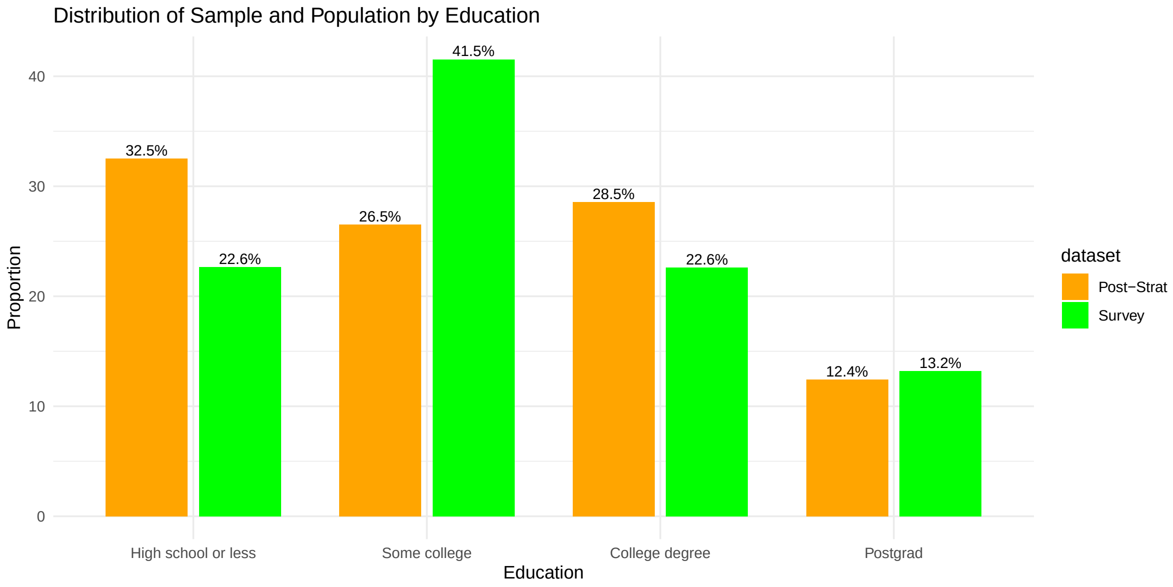

Figure 4: Distribution of Sample and Population by Education

Table 2: Voters Intention of Their Primary Party

Party Preference	Number of Respondents	Proportion (%)
Democratic	2180	39.14
Republican	1533	27.52
Other	1857	33.34

3 Model

For our study, we employ a technique called multilevel regression with post-stratification (MRP). This approach involves creating a model based on a smaller data set, such as our survey data, and then extending the model's findings to a larger population.

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.

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

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

Figure @ref(fig:fig-distribution-by-state) shows us the estimated of proportion of support for Trump and Biden by state using MRP with the inclusion of error terms. Each dot represents the point estimate of the proportion of support for Biden (blue) or Trump (red) in each state. Horizontal Lines extending from the dots represent confidence intervals for these estimates. The length of each line indicates the uncertainty associated with each estimate. For instance, we can see that this uncertainty lies between 50 percent to slightly higher than 80 percent for Trump in MT (Massachusetts). The dashed green line in the middle at the 50 percent

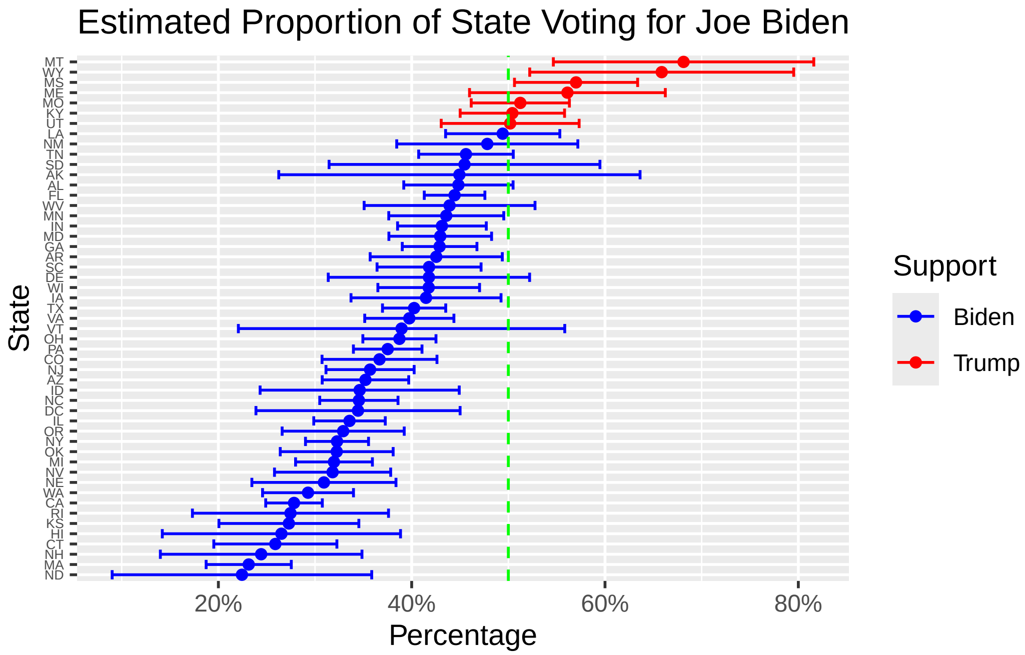

Figure 5: Distribution of Sample and Population by Education

mark represents the threshold for majority support. On the y-axis, each state is listed with its abbreviations and is ordered based on the proportion of support for Trump from the highest at the top to the lowest at the bottom.

From figure, it seems that majority of the states support Biden. Only 7 states out of 51 have its point estimate greater than 50 percent for Trump. However, the horizontal lines of confidence intervals of some states overlapping the green mark give some hope for the Republicans. However, excluding these contesting states, our model suggests that only 3 states are definitely in favor of Trump whereas 35 states are definitely supporting Biden.

Figure @ref(fig:fig-distribution-by-educationlevel) presents the estimated proportion of voters for Trump by education level, divided into four categories: 'High school or less', 'Some college', 'College degree', 'Postgrad'. Each black dot represents the point estimate of the proportion of voters within the corresponding education category who are predicted to vote for Trump. The horizontal lines extending to the left and right of each dot represent the confidence intervals around the estimate, which reflect the uncertainty.

It shows that regardless of education level, the level of support for Trump lies below 40 percent. The Republican party does not have majority, including the error bars across all education levels. Voters with "High school or less" education appear to have the lowest estimated support for Trump, which does not align with various exit polls and analyses from the 2020 election suggesting that Trump had substantial support among voters without a college degree. Con-

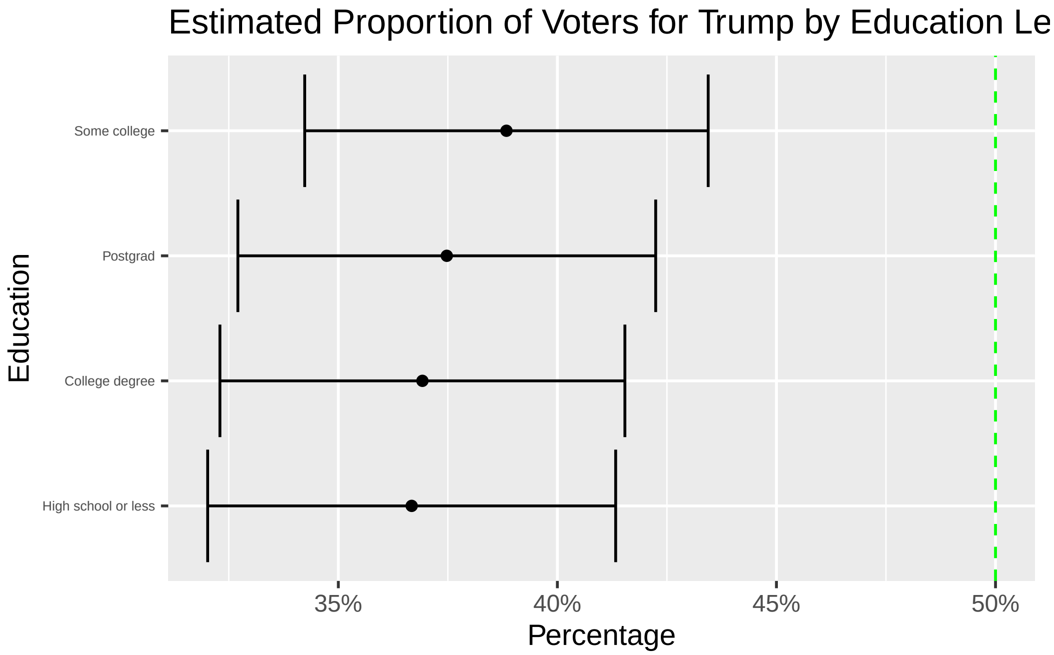

Figure 6: How different education levels of voters affect voting for Trump

versely, Voters with 'Some college' and 'Postgrad' education are the two groups that are more in support of Trump, which is exactly the opposite of what we have expected.

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.

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 7: **?(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 8: **?(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.