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%

*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 '2019 ACS' data (Ruggles et al. (2019)) 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 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.

In processing the raw post-stratification dataset, which initially contained approximately 3.2 million records, we refined it down to about 2.3 million records. This was achieved through a meticulous selection process, ensuring the data's integrity and relevance for our analysis. In our analysis, we've selected the variables 'sex', 'race', 'stateicp', 'age', and 'educd' from the dataset. To simplify our analysis, respondents who indicated 'other' or provided no data for their sex have been excluded. Consequently, 'sex' has been categorized strictly as 'Male' and 'Female'. We've refined 'race' into five categories: 'White', 'Black', 'Asian', 'American Indian', and 'Other', based on the composition of the U.S. population, with White, Black, and Asian categories accounting for approximately 93 percent of the total.

The 'stateicp' variable encompasses all U.S. states, using their standard abbreviations (e.g., 'CT' for Connecticut), and extends to 55 values to include 'Puerto Rico', 'State groupings (1980 Urban/rural sample)', 'Military/Military Reservations', 'District of Columbia', and an 'State not identified' category. Age has been grouped into four categories: '18-29', '30-49', '50-64', and '70+'. For educational attainment ('educd'), we've created four categories: 'High school or less', 'Some college', 'College degree', and 'Postgrad'.

We excluded any unknown responses to ensure clarity and accuracy in categorization and to enhance clarity and align with survey data, we've renamed 'sex', 'stateicp', 'age', and 'educd' to 'gender', 'state', 'age_group', and 'education', respectively. This restructuring aims to streamline our analysis by ensuring each respondent is accurately categorized.

Figure 1, Figure 2, Figure 3, and Figure 4 illustrate comparisons between survey data and post-stratification data across different variables. In these visuals, orange bars represent post-

Table 1: Voters Intention to Support Trump

[!h]

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

stratification data, while green bars signify survey data. The percentages displayed on each figure are rounded to the nearest tenth, introducing a potential margin of error of ±0.1% in the total values. Generally, the survey data aligns closely with the post-stratification data, maintaining a discrepancy of about 10% across most categories. Notable exceptions are observed in the '30-49' age group and the 'Some college' education level, where the differences exceed this margin.

```
# insert figure for state here
```

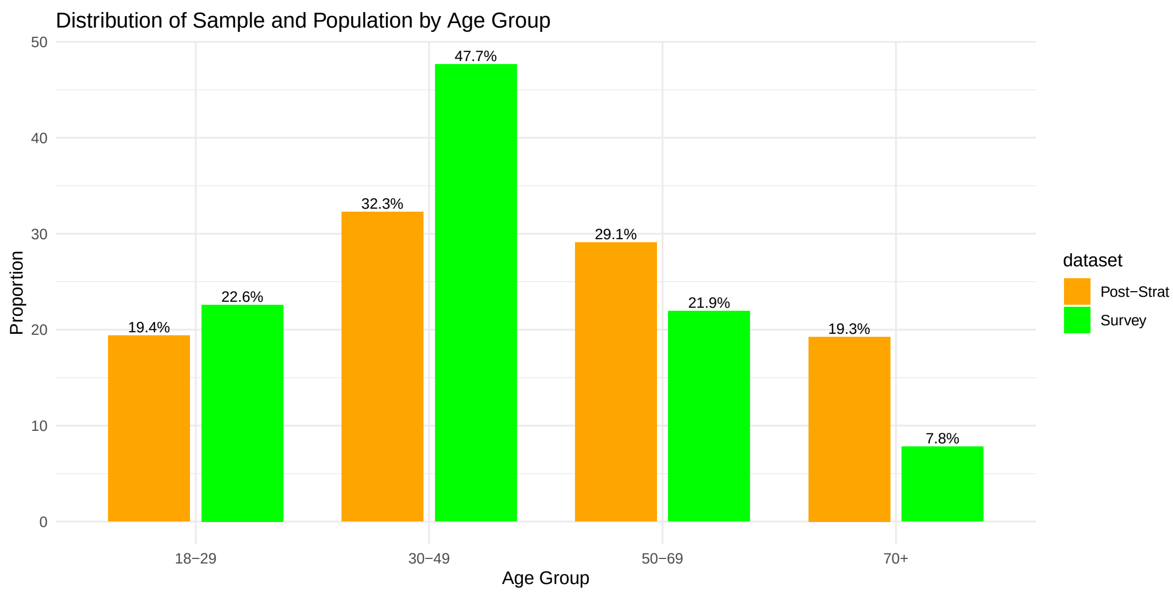

Figure 1: Distribution of Sample and Population by Age Group

Table 1 shows the proportion of voters who intend to vote for Donald Trump or not. Also, Table 2 shows the proportion of voters' supporting party. The data used to create these table is from the Tausanovitch and Vavreck (2020). We see that Donald Trump and his party Republican are not expected to win the popular vote before we implement the model.

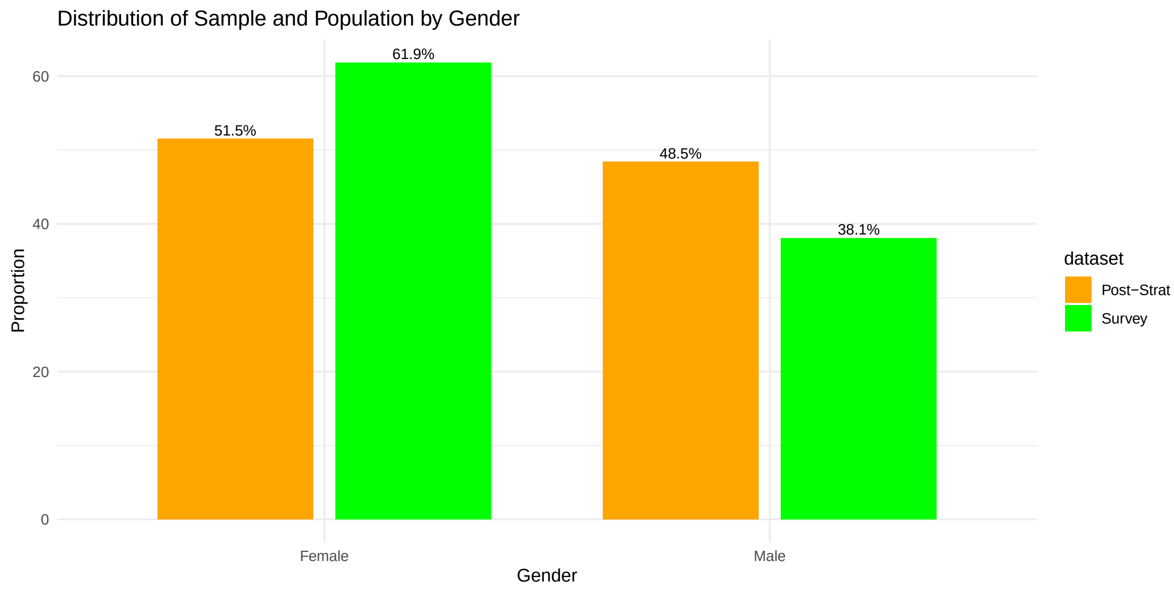

Figure 2: Distribution of Sample and Population by Gender

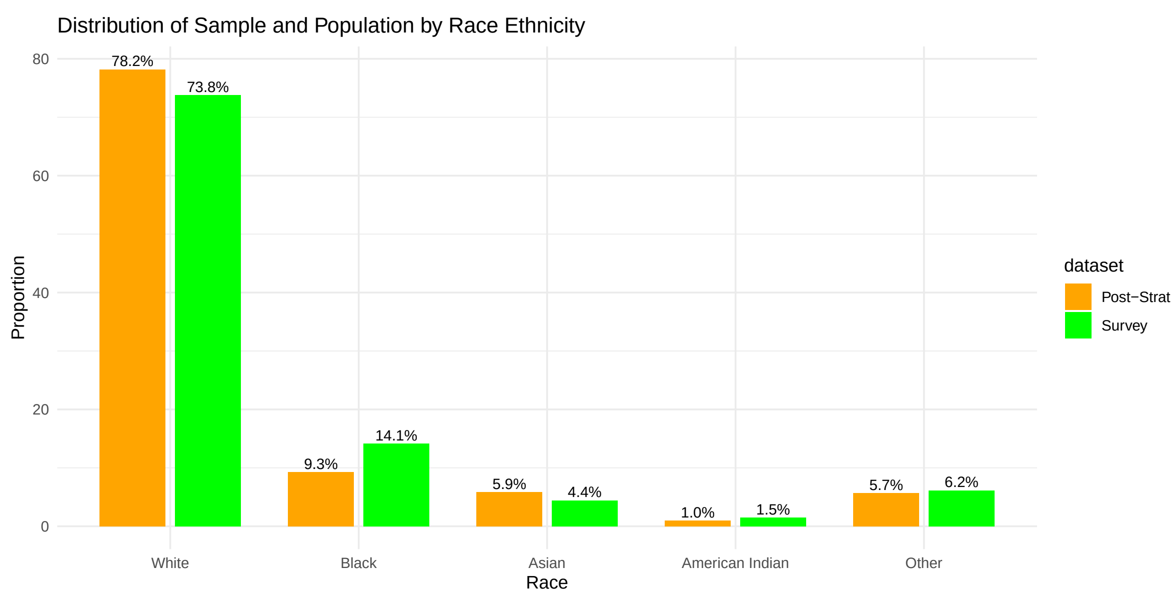

Figure 3: Distribution of Sample and Population by Race Ethnicity

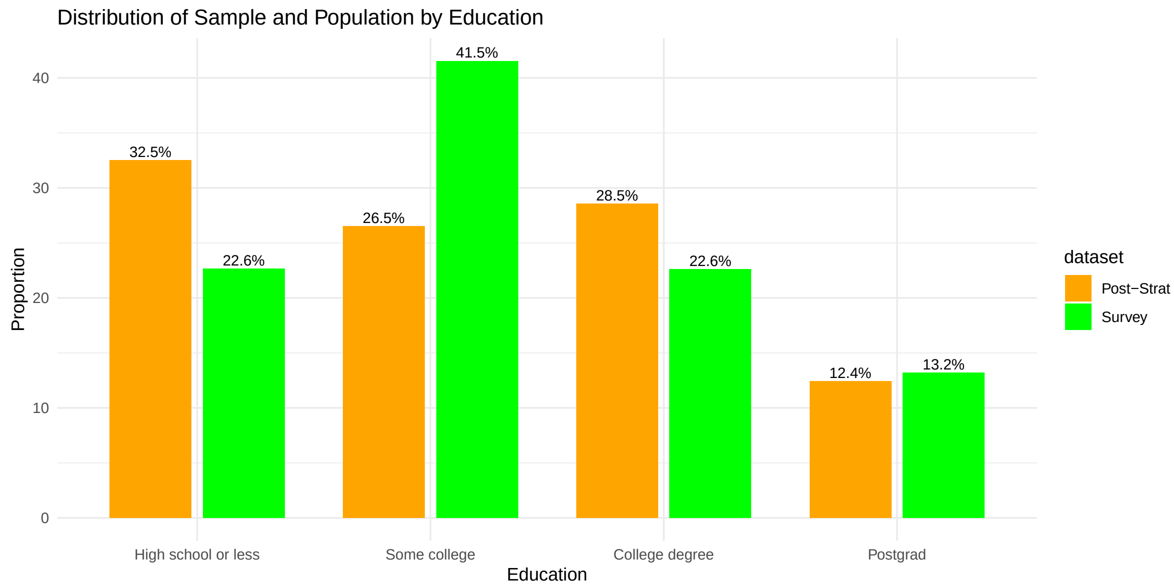

Figure 4: Distribution of Sample and Population by Education

Table 2: Voters Intention of Their Primary Party

[!h]

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.

The key steps in MRP involve initially selecting a dataset for model development. In this case, we utilized survey data from the Voter Study Group (Tausanovitch and Vavreck (2020)). The next step is to construct a model with this smaller dataset; here, we employed logistic regression based on the survey data, formulated as seen in equation 1. Following model creation, it is then applied to a broader dataset to estimate population characteristics. For our analysis, Census data from IPUMS (Ruggles et al. (2019)) served as this larger dataset.

To predict an individual's likelihood of voting for Donald Trump, we aim to construct a logistic regression model leveraging data from the Voter Study Group (Tausanovitch and Vavreck (2020)) and applying post-stratification with Census Data (Ruggles et al. (2019)). Given that logistic regression is suited for binary outcomes, we've introduced a variable, 'consider_trump', which assigns a 1 if the respondent indicates a plan to vote for Donald Trump, and a 0 for intentions to vote for other candidates, with 0 encompassing both "No" and "Other" responses.

The logistic regression model takes the form of:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{sex} + \beta_2 x_{agegroup} + \beta_3 x_{race} + \beta_4 x_{state} + \beta_5 x_{education} \tag{1}$$

In equation 1, each β represents a coefficient determined through regression analysis. The variables chosen for this project are sex, age, race, education, and state. These were selected because sex, age, and race are proven to be reliable indicators of voting preferences. This decision is based on patterns such as certain states consistently favoring the Republican party, while others alternate between Democratic and Republican. Education was chosen over income because it provides a clearer picture of an individual's background than income does.

Once the logistic regression model is developed, we'll use the `predict()` function in R (R Core Team (2023)) to apply our model to Census data (Ruggles et al. (2019)), breaking down the dataset into categories based on sex, race, age group, education level, and state. This will give us the likelihood of individuals within each category voting for Donald Trump. These predictions allow us to analyze potential outcomes like the popular vote winner or the electoral college vote distribution.

Table 3 shows the estimates for the coefficients that will fit into our logistic regression equation. These coefficients will fit into Equation 1, and were calculated using data from the Voter Study Group (Tausanovitch and Vavreck (2020)). The table is made using `kable` from `knitr` (Xie (2020)) and is formatted using `kableExtra` (Zhu (2020)).

Table 3: Coefficients from the Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	-1.1269207	0.8231922	-2.5930365	0.1878863
genderMale	0.6454282	0.0605598	0.5460791	0.7446367
educationHigh school or less	-0.0215196	0.0938906	-0.1785914	0.1291637
educationPostgrad	-0.0392885	0.1022382	-0.2097959	0.1332047
educationSome college	0.1320413	0.0781562	-0.0004280	0.2590242
age_group30-49	0.5119398	0.0850042	0.3699014	0.6477017
age_group50-69	0.6769174	0.0941263	0.5119061	0.8319957
age_group70+	0.8993021	0.1253618	0.6878129	1.1100246
raceAsian	-0.7916706	0.3008511	-1.2762432	-0.2728760
raceBlack	-1.7052759	0.2752683	-2.1485129	-1.2387060
raceOther	-0.9723473	0.2865172	-1.4448648	-0.4727599
raceWhite	0.1901461	0.2483476	-0.2079037	0.6111706
stateAL	0.1759303	0.8262695	-1.1488110	1.6604503
stateAR	-0.0657478	0.8344496	-1.4419087	1.4019364
stateAZ	-0.4574335	0.8034301	-1.7451418	0.9887553
stateCA	-0.5812949	0.7941797	-1.8467363	0.8398753
stateCO	-0.4103942	0.8165795	-1.7516658	1.0454097
stateCT	-0.8980631	0.8694824	-2.2666010	0.6056393
stateDC	0.1333770	0.9378865	-1.4644468	1.7361204
stateDE	-0.0329808	0.9211048	-1.5537404	1.5722435
stateFL	0.0218098	0.7992711	-1.2561402	1.4587686
stateGA	0.2273150	0.8093676	-1.0584034	1.6656086
stateHI	-0.4022495	1.0532898	-2.2097550	1.3718501
stateIA	-0.2792869	0.8523817	-1.6812761	1.2203228
stateID	-0.5715375	0.9117792	-2.1224329	1.0171239
stateIL	-0.4539897	0.7908061	-1.7369158	0.9760011
stateIN	-0.1124579	0.8034814	-1.3759652	1.3237964
stateKS	-0.9154592	0.8712311	-2.3169099	0.6127700
stateKY	0.1891792	0.8196904	-1.1163833	1.6735899
stateLA	0.5085951	0.8454838	-0.8286095	2.0124852
stateMA	-1.0144199	0.8087735	-2.3278399	0.4318518
stateMD	0.2540933	0.8153424	-1.0472043	1.7133075
stateME	0.3319872	0.8983963	-1.1254386	1.8934577
stateMI	-0.6201596	0.8147269	-1.9179094	0.8278040
stateMN	-0.1735514	0.8315900	-1.5049968	1.2975357
stateMO	0.2568405	0.8204754	-1.0284865	1.7210137
stateMS	1.0045240	0.8628145	-0.3393123	2.4995599
stateMT	0.9169121	1.0397878	-0.7536536	2.7395949
stateNC	-0.3230403	0.8028628	-1.6358424	1.1271320
stateND	-1.3821718	1.1464287	-3.4190661	0.5101015
stateNE	-0.7427467	0.8529747	-2.1674568	0.7907564
stateNH	-1.1627679	1.0026527	-2.8324897	0.5333387
stateNJ	-0.2755491	0.8195828	-1.5727529	1.1896796
stateNM	0.1284793	0.8789283	-1.2857051	1.6757967
stateNV	-0.4946525	0.8344399	-1.8263553	0.9859628
stateNY	-0.4028180	0.7889157	-1.6920330	1.0308732
stateOH	-0.2812188	0.8031339	-1.5499012	1.1728527
stateOK	-0.5510119	0.8230695	-1.8876618	0.9132757
stateOR	-0.6065868	0.8410369	-1.9839289	0.8792515
statePA	-0.3623052	0.8042335	-1.6496183	1.0852955
stateRI	-0.8718240	0.9623059	-2.4641792	0.7952202
stateSC	0.0476201	0.8259632	-1.2557306	1.5040781
stateSD	-0.1283511	0.9993491	-1.7289774	1.5773496
stateTN	0.0825080	0.8065814	-1.2181768	1.5502102
stateTX	-0.1073905	0.7918389	-1.3699153	1.3353886
stateUT	0.2037785	0.8368944	-1.1454047	1.6788839
stateVA	-0.0515024	0.7999330	-1.3606166	1.3835400
stateVT	-0.4688927	1.1001454	-2.2945844	1.3947062
stateWA	-0.7231677	0.8189467	-2.0323499	0.7530020
stateWI	-0.2468407	0.8044324	-1.5596314	1.1970349
stateWV	-0.1556233	0.8788915	-1.5628548	1.3558063
stateWY	0.8229283	1.0259232	-0.8515683	2.5749520

We use the `stan_glm()` function in R (R Core Team (2023)) for our regression analysis, specifically because we are dealing with a binary outcome: whether a voter supports Donald Trump or not. The nature of our data suggests an S-shaped distribution rather than a linear one, making logistic regression a better fit than linear regression. This approach is advantageous, especially when paired with post-stratification, as it allows us to better represent under-represented groups in our analysis. For instance, despite having only 7 responses from Alaska in our survey data (Tausanovitch and Vavreck (2020)), through multilevel regression and post-stratification, we can adjust this to effectively represent over 4500 individuals.

However, there are limitations to our model. The binary outcome does not allow for consideration of third-party candidates or non-voters, although this limitation is mitigated by our focus on the main candidates. More critically, our model's accuracy is heavily dependent on the quality of our survey data. Any inaccuracies or the need for adjustments in the survey can significantly impact our findings.

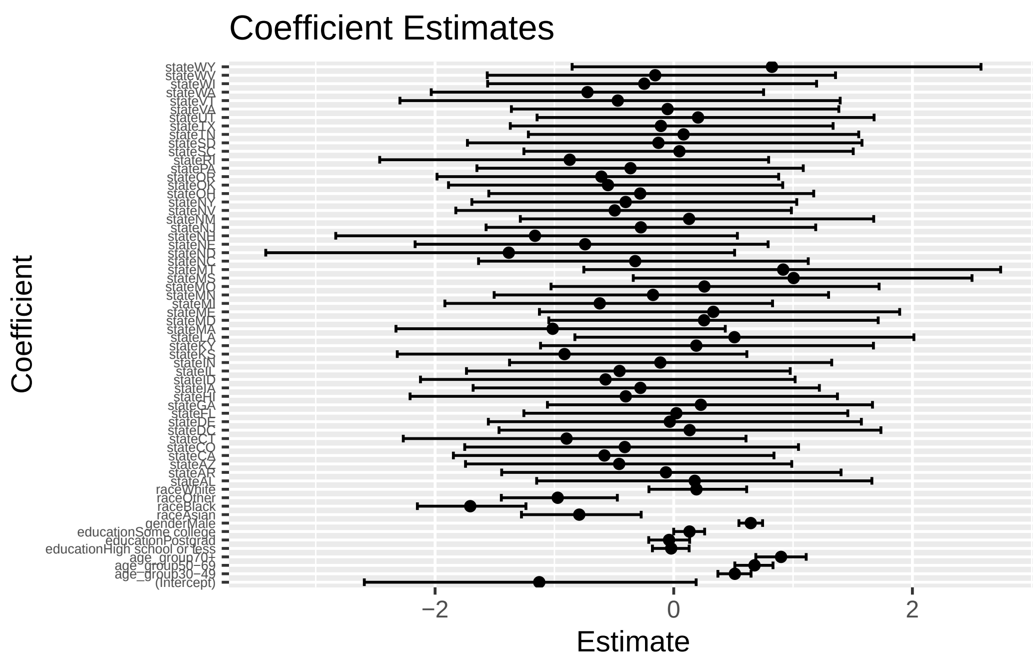

Figure 5: Coefficient Estimates

Figure 5 presents the coefficients derived from logistic regression on the survey data (Tausanovitch and Vavreck (2020)). It also includes error bars, indicating the confidence interval for each coefficient estimate. In interpreting these coefficients, it's essential to understand that positive values suggest a greater likelihood of voting for Donald Trump, whereas negative values indicate a tendency to vote for other candidates, such as Joe Biden.

By utilizing the results from our logistic regression model, we can formulate an equation that adheres to the structure outlined in equation 1, incorporating specific β coefficients for each

Table 4: Example of Prediction model

[!h]

gender	race	state	age_group	education	predicted_consider_trump_probability
Male	Black	MD	18-29	Some college	0.1463268
Female	White	TX	50-69	Some college	0.4450551
Male	Black	DC	70+	High school or less	0.2471943
Female	White	NY	30-49	Some college	0.3363712
Female	White	AL	30-49	College degree	0.4419403
Male	White	MI	50-69	High school or less	0.4404717
Female	White	MN	30-49	College degree	0.3594809
Male	White	CA	70+	High school or less	0.5060479
Female	Black	TN	70+	High school or less	0.1369608
Female	White	FL	18-29	College degree	0.2902058

variable. Given the number of variables, detailing the equation fully is challenging. Essentially, the equation integrates the β value of a variable if an individual's characteristic matches that variable. Table 4 offers examples of how the probability varies based on different variables.

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 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

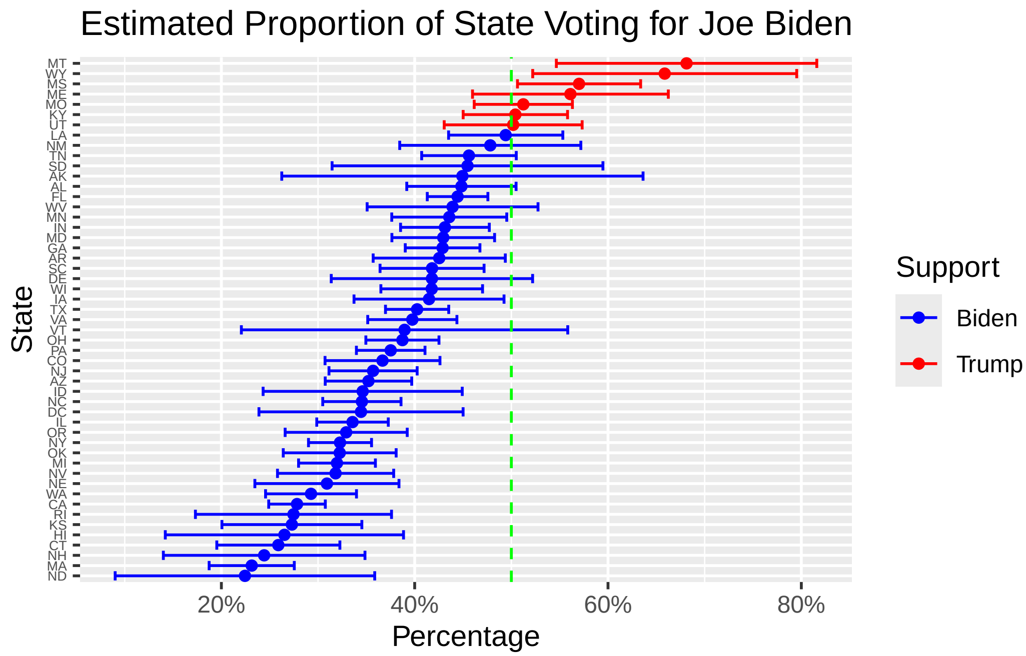

Figure 6: Distribution of Sample and Population by Education

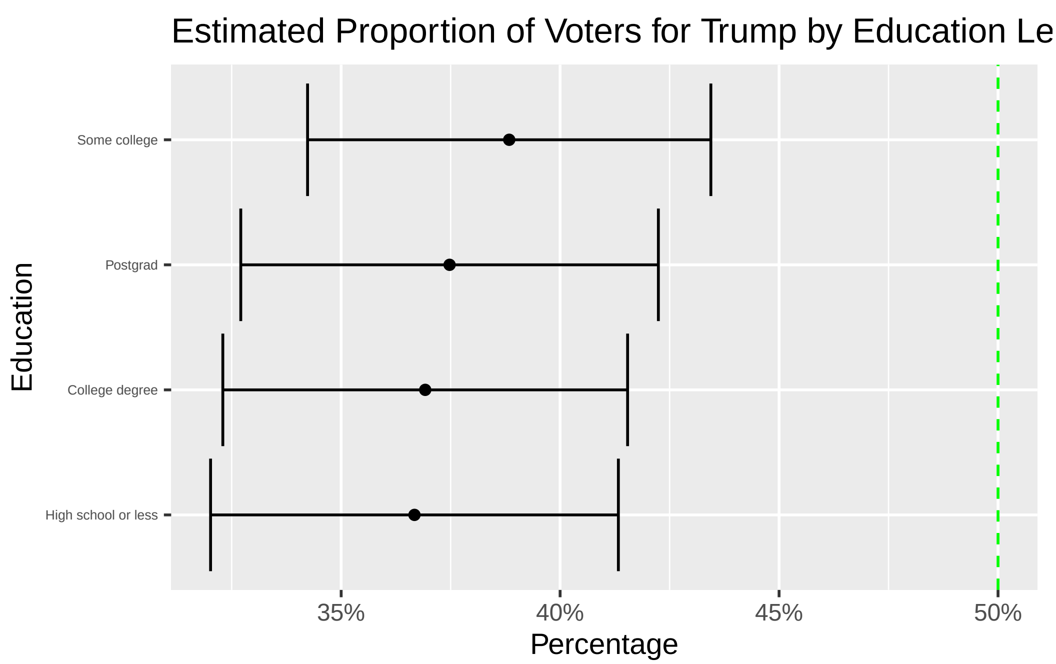

Figure 7: How different education levels of voters affect voting for Trump

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. Conversely, 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

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

References

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/.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2019. *IPUMS USA: VERSION 9.0*. Minneapolis: University of Minnesota. https://www.ipums.org/projects/ipums-usa/d010.v9.0.

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.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.