Forecast of voting results in the 2020 US election*

Ruoyun Wang,1005472597

01/04/2022

Abstract

This paper uses the multilevel regression with post-stratification (MRP) method to predict the 2020 election results, survey data Democracy Fund + UCLA Nationscape 'Full Data Set' from July 2019 to December 26, 2019. The post-stratification dataset selected the American community survey (ACS) data in 2010. Both datasets contain age, gender, region, race, and income information. The above variables are used as key variables to model the survey data. The modelling results were then applied to ACS data to complete the MRP estimates of the 2020 election results. The results show that Biden will win 21 states and 50.25% of the votes.

Keywords: Forecasting, US 2020 Election, multilevel regression with post-stratification;

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

1 Introduction

The U.S. election has four steps (U.S. Presidential Election Process, 2020), Primaries and caucuses: party members vote for the best candidate National conventions: each party finalizes the presidential nominee General Election: Presidential candidates campaign throughout the country to win support from the populace. Electoral College: Each state gets a certain number of electors. There are 538 electoral votes, and the election will be won if more than half of the votes are 270.

Before each election, many agencies make predictions about the election results. The Economist predicts that Biden will win 356 electoral votes on November 3, 2020. However, the anticipated results change over time (President—Forecasting the U.S. 2020 elections, 2020). At the beginning of April 2020, Trump's approval rating was slightly higher than Biden's. The prediction of the election is not necessarily accurate. For example, in the 2016 election, Hillary's chance of winning was 71.4%, much higher than Trump's, according to the prediction on Nov. 8, 2016, by the FiveThirtyEight website (Silver, N, 2016). The same statistician, Silver, predicts that Biden will have an 89% chance to win in the 2020 election (Silver, N, 2020).

Considering that electoral decisions are related to presidential candidates' policies, the different candidates' policies may be biased towards specific groups, so people have different opinions on election results. When conducting electoral research, the questionnaire may be designed with obvious political inclinations and lead to some systematic errors. Selection bias on the survey population will make the electoral study biased. Thus, we should try to eliminate the apparent political questionnaire and adopt a good survey design to get the data. For the selection bias of the survey population, multilevel regression with post-stratification (MRP) is used to balance the bias.

The MRP method will use two data sets, a survey and post-stratification data. The survey data set has apparent purposes, such as the opinions on the 2020 U.S. election. There may be some selection bias in the survey data, which cannot represent the entire population. Another dataset is post-stratification data to represent the population. Post-stratification data are not a random sampling, and we often use census data as post-stratification data. After we get the census data, we can use the random sampling method to select parts of the census data, and the reduced size data can still represent the whole population. The MRP uses regression models to correlate individual-level survey responses with various characteristics, then uses the regression results to apply to the post-stratification data to get the new estimation.

Pew research center (PRC) analyzes the reasons for Biden's 2020 election victory. Due to the impact of COVID-19, the voting method changed in 2020. The voters sent the vote by mail, and about 7% more voters participated in the voting. In 2020, approximately 66% of U.S. adult citizens were voting.

There were many changes in the 2020 U.S. election. The PRC found that Biden got more support from suburban voters, with about 54 percent supporting Biden. Trump's approval rating has improved among Hispanic voters, especially those without a college degree. And Trump's approval rating is higher in Hispanic voters (41%) than that of those College-educated Hispanic voters (30%). In addition to changes in Hispanic voters' approval ratings, minority support for Biden is significantly more excellent than Trump's, with black voters supporting Biden by an overwhelming 92%. Among White non-college voters, Biden's approval rate improved, and the approval rate is about 36 percent, eight percentage points higher than Clinton's approval rate in 2016. In addition, people of different age groups have other preferences for elections. Gen Z and Millennial voters favoured Biden over Trump, while Gen Xers and Boomers had evenly preferred Biden and Trump. The Gen Z voter constituted 8% of the voters, while Millennial and Gen Xers were 47% in 2020. For the men voters, Trump and Biden also had preferences evenly. But, for women voters, Trump won a slightly larger share than the share in 2016.

From the above analysis, it can be seen that factors such as age, gender, education, race, Hispanic, and income will affect the election results. Therefore, this paper uses the MRP method to predict the outcome of the 2020 U.S. election. Use Democracy Fund + UCLA Nationscape data set to get individual-level survey data about election polls, and then use 2010 American community survey (ACS) data as a post-stratification dataset. Both datasets include age, gender, region, race, income and other information, so the above variables are used as key variables to model the survey data. Then, the modelling results are applied to ACS data to complete the MRP estimation of the 2020 general election results.

This report consists of 4 parts:

1. We will introduce two databases, survey data and post-stratification data, to study the distribution of the above key variables through the EDA method.
2. The working principle and steps of MRP are introduced in detail in the model.
3. We display the MRP results through tables, graphs, etc.
4. The advantages and disadvantages of MRP are discussed by comparing the estimated results of MRP with the actual election results.

2 Data

2.1 Survey Data

The individual-level survey data selects the National Data set. This dataset collaborates between the Democracy Fund Voter Study Group and UCLA political scientists and is one of the largest polling projects ever conducted. The project interviewed people in every country and mid-sized city, and it had three phases. The first phase was released in January 2020 and included 24 weeks of data from July 2019 to December 26, 2019, with 156,000 samples collected. The second phase ran from January 2020 to July 2020 and released data in September 2020. The third phase runs from August 2020 to February 2021. The collected data was sorted by week. The survey data in this article uses the data from June 25, 2020, to July 1, 2020. The data set contains 6046 US adults and 266 variables.

In data cleaning, first, only relevant variables such as age, gender, state, Hispanic, race, education, household income, and vote_2020 are retained. Then, deleting observations other than Donald Trump and Joe Biden in the election for president. In addition, observations under 18 do not meet the voting qualification and will also be deleted. After that, 4803 units remained.

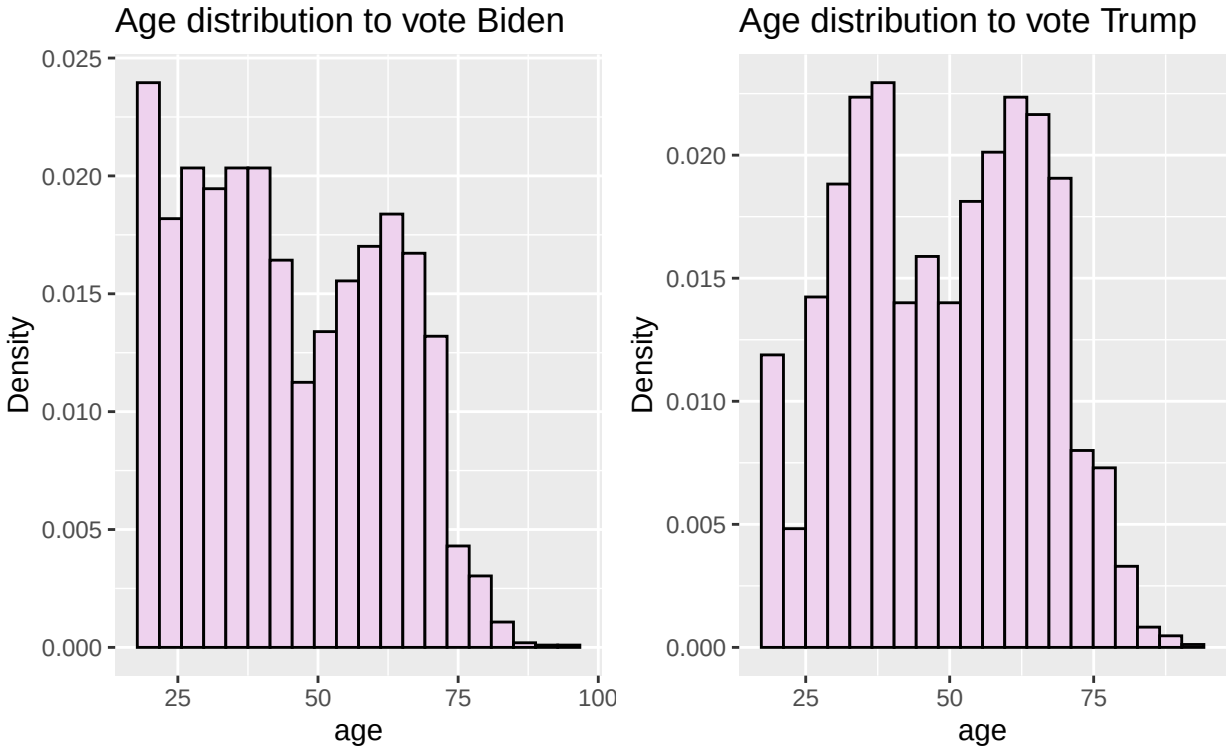

Figure 1: Distribution of age

Figure 1 shows the age distribution for the participants in the survey who select Biden and Trump. The

graph on the left indicates that young voters make up a high proportion of voters preferring Biden. The ages of voters under 44 are distributed evenly and reach a local minimum at age 50. Subsequently, the age reaches a local maximum around age 60. We found that the younger voter share was not as good as that of Biden for those who chose Trump. Among the 25-44 age group, voters choosing Trump increased with age. A lower proportion for the 45-50 population also appeared. Then around the age of 60, the ratio peaks again.

The age distributions show significant differences between the pro-Biden and pro-Trump populations. Therefore, age will be an essential factor affecting the approval rate. In Bannerbook, the age distribution is divided into four categories: 18-29, 30-44, 45-64 and 65+, so in the clean data process, we keep the above classification method and create an age group variable.

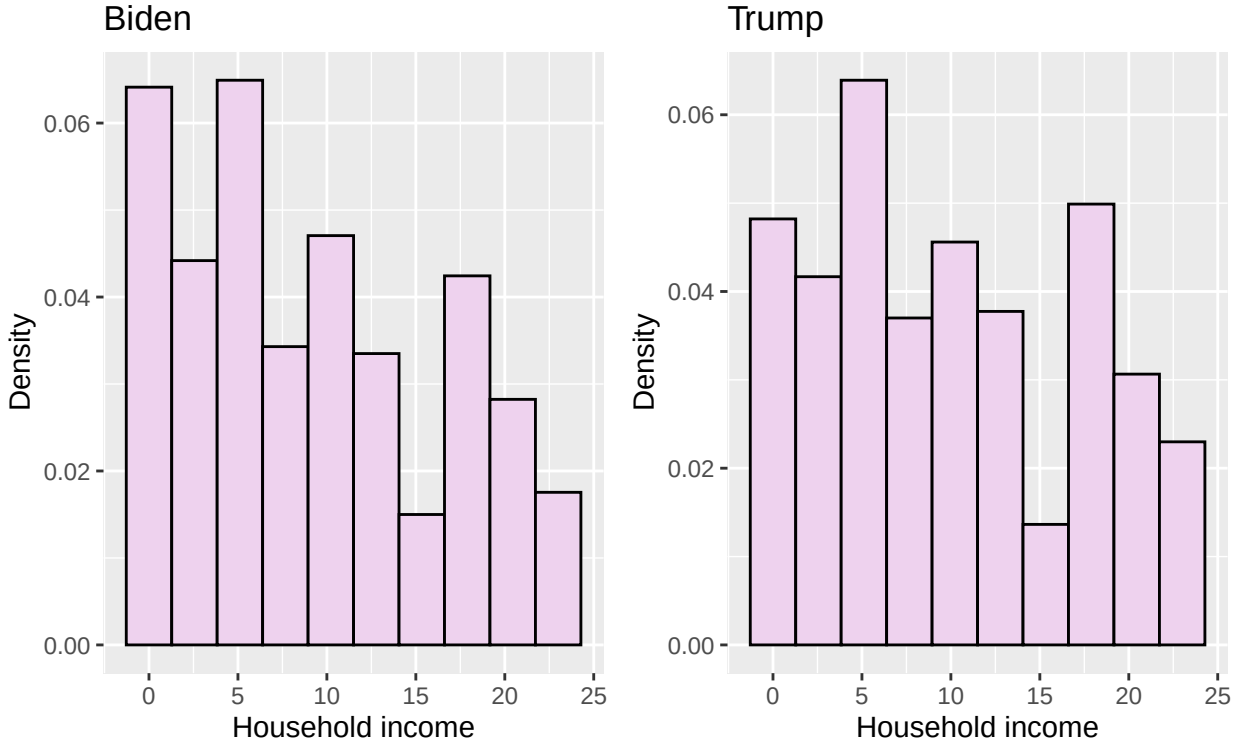

Figure 2: distribution of income

Figure 2 shows the income distribution for choosing Biden and choosing Trump in the survey data. The graph shows that the lowest income group of those who selected Biden accounted for more than 6%, while among those who chose Trump, the proportion was less than 5%. Overall, the income distributions are similar. Moreover, the survey data records household income, while the census data is individual income, which does not match. Therefore, the follow-up research will ignore the influence of income on election results.

The factor education has 11 levels, and I rearrange these levels into four new levels: post-secondary or higher, High school or less, some post-secondary, and graduate degree. For the Hispanic variable, there are 15 levels, and some levels only have several units. I also rearrange this Hispanic variable into four classes: Not Hispanic, Mexican, other Hispanic, Cuban and Puerto Rican. From the introduction, race is an essential factor in voting. There are 15 levels for race-ethnicity. I also create a new race variable to identify the levels: White, Black or African American, Chinese, Japanese, American Indian, and others.

Table 1 shows the summary of the selected survey data. It can be seen that the proportion of participants with age 18-29 is 19.5%, which is smaller than that of other groups, and the ratio of the age range of 30-44 and >60 is the largest, which are 28.9% and 28.5, respectively. The proportion of the 45-59 years old participation was 23.2%, slightly smaller than the proportion in groups 30-44 and >60.

There is a gender difference in the survey data, with 53.6% female participants and only 46.4% male

Table 1: survey data summary

		N	%
age_group	age_45-59	1113	23.2
	age_60_or_more	1386	28.9
	age_18-29	937	19.5
	age_30-44	1367	28.5
gender	female	2576	53.6
	male	2227	46.4
education_level	Post secondary or higher	1796	37.4
	High school or less	1235	25.7
	Some post secondary	1304	27.1
	Graduate degree	468	9.7
hispanic_level	Not Hispanic	4108	85.5
	Mexican	446	9.3
	other	219	4.6
	Cuban	27	0.6
	Puerto Rican	3	0.1
race_level	White	3608	75.1
	Black or African American	574	12.0
	other	164	3.4
	Chinese	63	1.3
	Japanese	19	0.4
	American Indian	59	1.2
vote_biden	0	2212	46.1
	1	2591	53.9

participants. The difference in the number of males and females will lead to bias estimation in the model estimated by the survey data.

Post-secondary or higher degrees have the largest proportion among the four education levels, about 37.4%. High school or less and some post-secondary degrees account for 25.7 and 27.1 percent, respectively. The graduate degree has the smallest proportion, with less than 10%. According to the survey results, there are differences in the distribution of the population participating in the survey.

The newly generated Hispanic level shows that about 4108 observation units are not Hispanic, accounting for 85.5%. And within these Hispanic participants, Mexican has the largest number, about 446, accounting for 10% of the entire survey dataset, and Cuban and Puerto Ricans make up less than 1%.

At the race level, we found that the proportion of White who participated in the survey accounted for 75%, much higher than that of other ethnic groups, followed by Black and African Americans, with 574 participants, accounting for 12% of the total number. The number of Chinese and American Indians is relatively small, accounting for about 1.2%, while the number of Japanese is only 0.4%.

2.2 Post-stratification Data

The census data is collected from the IPUMS. IPMUS collects and preserves U.S. Census data, including the decennial census data from 1790 to 2010 and the American Community Survey (ACS) from 2000 to the present. This paper uses the 2010 ACS data as the post-stratification dataset. Due to a large amount of full data, we only extract 0.1% data from the full data file, and 306194 observations are kept in the raw dataset.

After clean data procedure and keeping observations with age no less than 18, the census data's age distribution is shown in Figure 3. The left histogram shows the age distribution of females, and the rightside figure describes the age distribution for males. It can be seen that the male and female age distributions have

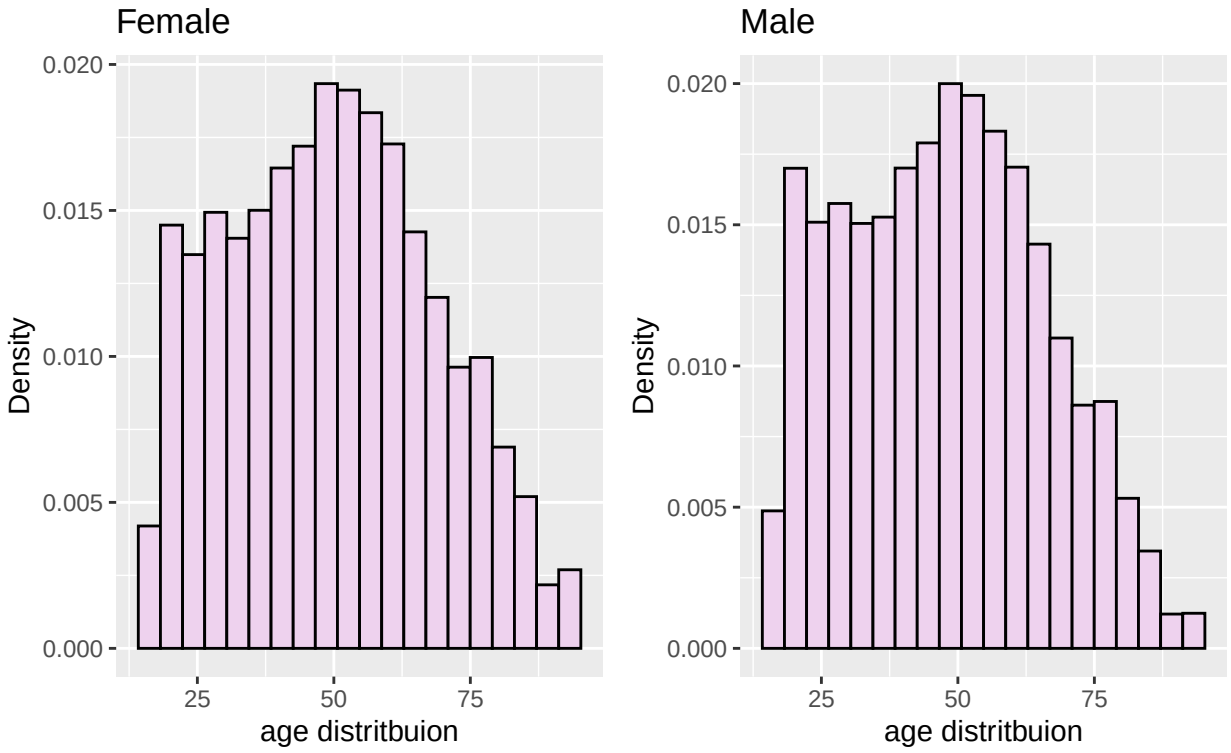

Figure 3: distribution of age in census

similar shapes. The distribution of people aged 20 to 40 years is evenly, and the proportion of people aged 50 to 60 is higher than other age groups. After 55 years old, the ratio of the population decreases with increasing age. Comparing Figure 3 and Figure 1, we can find a clear difference between the age distribution in the survey data and the census data. These differences reflect some selection bias in the survey process, and the results estimated by the survey data will lead to bias from the actual data.

Comparing the data in table 1 and table 2, we find that in the age group, the proportion of age groups in Table 2 is 28.8%, which is higher than that of the age group in the survey data, 23.2%. The proportion of the 30-44 age group was 23.8% in post-stratification data, which is lower than the proportion of participants in this age group in the survey data in Table 1.

There is also a big difference in the education level. In the survey data, the ratio of post-secondary or higher is 37.4, which is 13% higher than the ratio in the post-stratification data, which is 24.4. In the post-stratification data, the proportion of high school or less is the largest, about 42.7%, while in the survey data, the proportion is only 25.7%. In addition, the values of gender, Hispanic level and race level are similar in the two tables. From the data in Table 1 and Table 2, we can see a big difference between survey data and post-stratification data. Therefore, the MRP method can reduce the influence of the selection bias in the survey data on the result estimation.

Model

logit regression

Logistic regression is a statistical model that uses a logistic function to model a binary response variable. Since the output variable for predicting election results is binary, a logistic regression model is used to predict the voting results. The predictor variables are age, race, education, Hispanic and gender. The regression model is:

Table 2: post-stratification data summary

		N	%
age_group	age_30-44	56360	23.8
	age_18-29	43975	18.6
	age_60_or_more	68444	28.9
	age_45-59	68143	28.8
gender	male	113542	47.9
	female	123380	52.1
education_level	Post secondary or higher	57695	24.4
	Some post secondary	54523	23.0
	Graduate degree	23635	10.0
	High school or less	101069	42.7
hispanic_level	Not Hispanic	209241	88.3
	other	6445	2.7
	Mexican	17415	7.4
	Puerto Rican	2517	1.1
	Cuban	1304	0.6
race_level	White	188385	79.5
	Black or African American	23552	9.9
	Chinese	2918	1.2
	other	19296	8.1
	American Indian	2062	0.9
	Japanese	709	0.3

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{education} + \beta_4 x_{Hispanic} + \beta_5 x_{race}$$

Where:

$P = Pr(vote_biden == 1)$ is the propability to vote biden.

$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients to be estimated.

After we get the estimation of the regression model, the estimated probability is calculated by:

$$\hat{p} = \frac{exp(\beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{education} + \beta_4 x_{Hispanic} + \beta_5 x_{race})}{1 + exp(\beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{education} + \beta_4 x_{Hispanic} + \beta_5 x_{race}}$$

The assumption for logisitic regrssion model is (Z, 2020):

Assumption 1: the response variable is binary

Assumption 2: The observations are independent

Assumption 3: There is no multicollinearity among explanatory variables

Assumption 4: There are no extreme outliers

Assumption 5: There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable

Assumption 6: The Sample Size is Sufficiently Large

post-stratification method

MRP is a popular way to adjust a non-Representative sample to better analyze the survey. It uses a regression model to build an individual-level model and then rebuild the sample to better match the population. To rebuild the sample, we choose the non-probability data like census data as the post-stratification data set.

In the previous logistic model, age has 4 levels, gender has 2 levels, education has 4 levels, race has 6 levels, Hispanic has 5 levels, and the state has 51 levels. Thus, these predictor variables can build $4 * 2 * 4 * 6 * 5 * 51 = 48960$ cells. We count the number of observations N_i in each cell and then use the regression model to predict the probability Biden to win in each cell $\widehat{y}_j$. The MRP estimation of the probability Biden to win will be :

$$\hat{y}^{PS} = \frac{\sum N_j \widehat{y}_j}{\sum N_j}$$

Where:

$\hat{y}^{PS}$ is the estimated probability of Bidne win

- N_j represents the population size of the j^{th} cell
- $\widehat{y}_j$ represents the estimation of proportion cell j.

Results

The logistic regression results are shown in Table 3. I omit the 50 coefficients of state in the table. 45-60 is the baseline age group. The regression results show that the coefficients of age groups 60 or more, 18-29 and 30-44 are positive, so people in these groups will have a higher probability of voting for Biden than the baseline group. The age group 18-29 has the largest value 2.20, and the conclusion that young people prefer to select Biden is consistent with the Histogram in Figure 1. From the t value calculated by the ratio between an estimated value and its standard error, all these three age group coefficients have t values large than 2. Thus, these age group coefficients are all statistically significant. The scale of these coefficients is larger than 1, so these coefficients are also economically significant.

For the gender, the male coefficient is 0.69, which is positive and significant by the calculated t value. Therefore, we conclude that male voters will have more chances to vote for Biden than females.

Compared with the education level, the reference group is Postsecondary or higher. The results show that voters with graduate degrees would have the highest preference to vote for Biden. The second highest coefficient is some postsecondary education level. The postsecondary or higher level has the lowest chance to vote for Biden.

For the Hispanics, the non-Hispanics group is the baseline. Mexican voters have the highest chance to vote for Biden, and Cuban and Puerto Rican have no difference from other non-Hispanics. The other Hispanic voters also have a significantly higher chance to vote for Biden.

White is the baseline group in the race, and the Black or African Americans have the highest chance to vote for Biden for the race levels. The other minorities, such as Chinese, Japanese, and American Indians, also prefer Biden to Trump compared to the White.

These coefficients are consistent with the exploratory data analysis. Based on this regression model, I predict the probability in each post-stratification cell. The estimation probability in each state is shown in Figure 4. The figure's points are the mean probability to vote for Biden in each cell, and the line is the 95% CI for the probability in each cell. We can see that 40 states have a mean value larger than 0.5. According to this figure, Biden will have a higher chance to win the vote.

After that, I make another estimation by the MRP. The estimated probability of voting for Biden is shown in Figure 5. In this case, only 21 states will get a probability larger than 0.5, less than half of the states. The MRP estimated proportion is 0.5025, which makes it hard to say if Biden will win the results.

The contradictory conclusion is because the mean probability in Figure 4 considers that each cell has the same size and ignores the weight effect on the prediction. In this case, the sample can not represent the population, leading to bias. The MRP probability is more credible, and Biden will lose in the prediction.

```
sum(post_stratified_estimates$mean>0.5)
```

```
## [1] 41
```

```
sum(post_stratified_estimates$alp_predict>0.5)
```

```
## [1] 22
```

```
mean(post_stratified_estimates$alp_predict)
```

```
## [1] 0.5026003
```

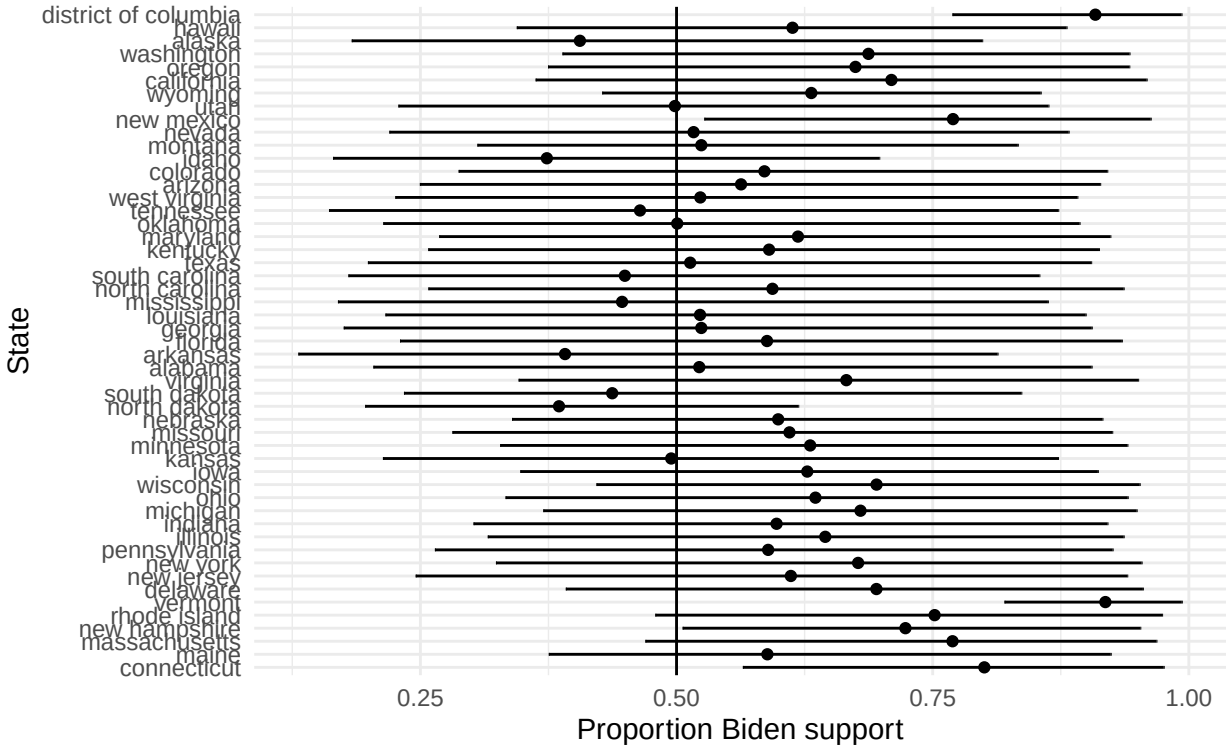

Figure 4: prediction in each state

Discussion

This paper uses the MRP method to predict the 2020 election results. The individual-level survey data in the MRP method uses the data of the election survey report completed by the Democracy Fund Voter Study Group in cooperation with UCLA political scientists from July 2019 to December 26, 2019. According to the survey results, Biden will win the election with a narrow advantage of 53.9%. However, the survey data population cannot reflect the entire population. To overcome the influence of selection bias, we selected and used the 2010 American community survey (ACS) data as the post-stratification dataset. Both datasets include age, gender, region, race, income and other information, so the above variables are used as key variables to model the survey data. Then, the modelling results are applied to ACS data to complete the MRP estimation of the 2020 general election results. The results show that Biden will win the votes of 21 states and 50.25%, which is far from the 53% voting rate of the statistical results.

Through the research on the prediction of the U.S. election results by the MRP method, we found that the

Table 3: Logistic regression results

	Model 1
(Intercept)	1.36
	(0.24)
age_groupage_60_or_more	1.22
	(0.09)
age_groupage_18-29	2.20
	(0.11)
age_groupage_30-44	1.24
	(0.09)
gendermale	0.69
	(0.07)
education_levelHigh school or less	0.64
	(0.09)
education_levelSome post secondary	0.84
	(0.08)
education_levelGraduate degree	1.24
	(0.12)
hispanic_levelMexican	1.86
	(0.15)
hispanic_levelother	1.05
	(0.18)
hispanic_levelCuban	0.46
	(0.44)
hispanic_levelPuerto Rican	1.41
	(1.34)
race_levelBlack or African American	9.57
	(0.14)
race_levelother	1.61
	(0.18)
race_levelChinese	2.90
	(0.31)
race_levelJapanese	3.69
	(0.61)
race_levelAmerican Indian	0.85
	(0.28)
Num.Obs.	4487
AIC	5611.2
BIC	6040.6
Log.Lik.	−2738.620
F	7.888
RMSE	1.11

Notes: The data in brackets is standard error

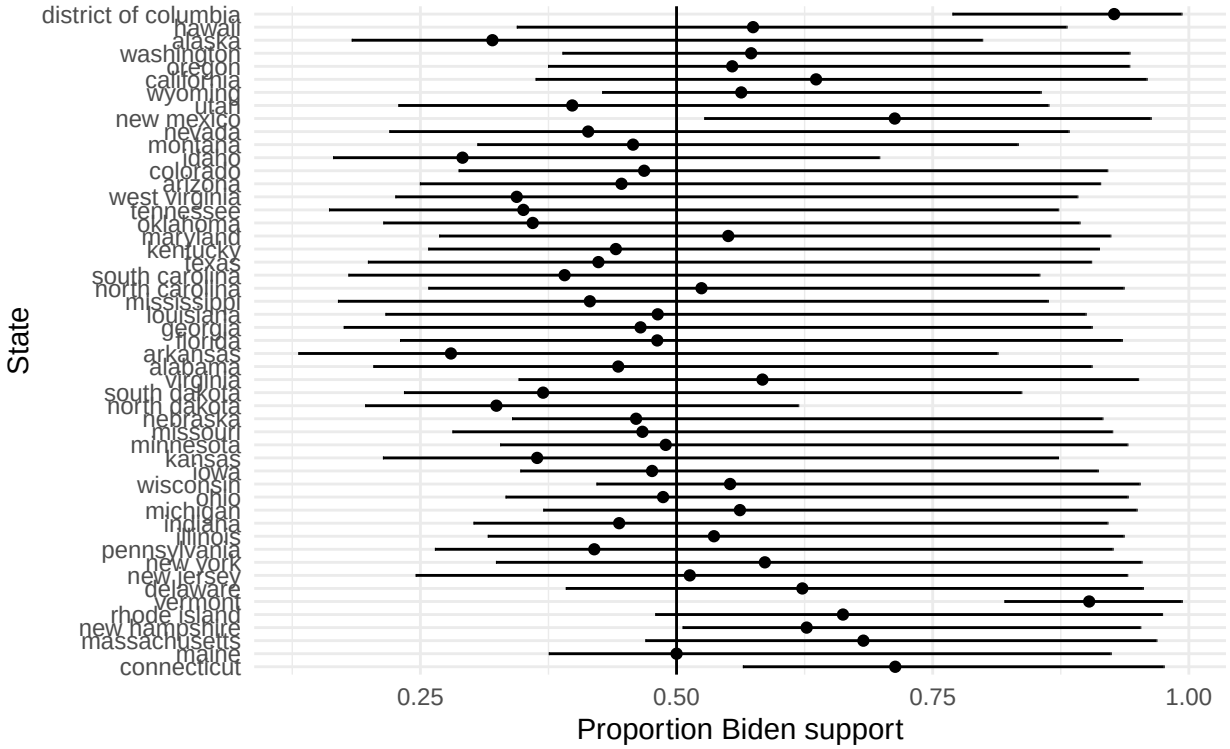

Figure 5: MRP prediction in each state

poll survey results often have a certain bias, and this bias may appear in many aspects, such as selection bias. Voter participation in election polls varies across age groups, with those in their mid-50s having the lowest participation rates. We found that Biden's policies favoured and gained more support from minorities through poll surveys. However, over 70% of the American population is white, and winning the favour of white voters will determine the general election outcome. This is also the most striking point of Trump's governing policy. As for this kind of voting system in the U.S., winning the majority of the people looks more important than minority ethnic. Thus, it may be a reason why racial discrimination is becoming more and more serious in the United States.

The weakness of this paper is reflected in the fact that when obtaining the election results, it does not consider the difference in the number of votes in each state, and only considering whether the election is won in a state cannot fully reflect the general election results. Therefore, in the follow-up research, we can add the information on the number of votes in each state to realize the prediction of national public opinion. The estimated results will be more accurate. Secondly, the survey data were collected at the end of June, and the election date is still relatively long, so the MRP results reflect the results of the June election prediction. As the time approaches the election, opinion polls will become more accurate. Therefore, we can use the survey data near the general election to estimate the results accurately predict. At the same time, we also realize that the survey results do not reflect the real situation. For example, in the 2016 public poll survey, although Clinton won the poll, however, the election results were the exact opposite.

Reference

Tausanovitch, Chris and Lynn Vavreck. 2021. Democracy Fund + UCLA Nationscape Project, October 10-17, 2019 (version 20211215). Retrieved from [URL].

Z. (2020, October 13). The 6 Assumptions of Logistic Regression (With Examples). Statology. https://www.statology.org/assumptions-of-logistic-regression/

U.S. Presidential Election Process. (2020, August 28). U.S. Embassy in the Philippines. https://ph.usembassy.gov/us-elections/us-presidential-election-process/#:%7E:text=In%20the%20Electoral%20College%20system,(270)%20wins%20the%20election.

President—Forecasting the US 2020 elections. (2020, November 3). The Economist. https://projects.economist.com/us-2020-forecast/president

Silver, N. (2016, November 8). 2016 Election Forecast. FiveThirtyEight. https://projects.fivethirtyeight.com/2016-election-forecast/

Silver, N. (2020, November 3). 2020 Election Forecast. FiveThirtyEight. https://projects.fivethirtyeight.com/2020-election-forecast/

Alexander, R. (2022). Chapter 16 Multilevel regression with post-stratification | Telling Stories With Data. Telling Stories With Data. https://www.tellingstorieswithdata.com/mrp.html

Igielnik, R., Keeter, S., & Hartig, H. (2021, September 30). Behind Biden's 2020 Victory. Pew Research Center - U.S. Politics & Policy. https://www.pewresearch.org/politics/2021/06/30/behind-bidens-2020-victory/

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Vincent Arel-Bundock (2022). modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready. R package version 0.9.6. https://CRAN.R-project.org/package=modelsummary

Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Hadley Wickham and Evan Miller (2021). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.4.3. https://CRAN.R-project.org/package=haven