Causal Inference in the 2020 US Federal Election: While COVID-19 Lowered Trump's Support, Loss of Confidence Among Wealthier Voters Seems More Decisive*

Yiliu Cao

December 7, 2023

This study used the election data from MIT EDSL and COVID-19 data from JHU CSSE, to investigate the causal inference between COVID-19 and Donald Trump's loss during the 2020 US Federal Election. The main methodology used in this paper is Propensity Score Matching with testing different treatments. The primary finding is that the counties with high death per case rate seems to vote less for Trump with approximate treatment effect -0.01. Besides, it seems that the interaction effect with income seems to be more critical: the medium-high income levels counties seems to vote at maximum 3% less for Trump compared to those low income counties. Furthermore, this paper also conduct counterfactual analysis and indicate the Trump would pre-elect if there was no COVID. Furture analysis should focus more on the choice of covariates to predict propensity scores.

Table of contents

*Code and data from this analysis are available at: https://github.com/yiliuc/covid_and_trump_loss.git

1 Introduction

2 Data

In this paper, the data used mainly consists of the voting patterns for the 2020 US Federal Election, COVID infection rate and socioeconomic variables in each county of the US. The voting data is taken from the MIT Election Data Science Lab, including voting for each party at the county level at each Federal Election from 2000 to 2020. The COVID data is from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The socioeconomic data is taken from the American Community Survey.

2.1 MIT Election Data Science Lab

The MIT Election Data Science Lab is a lab at MIT that collects and analyzes election data to support advances in election science. Their collected data is not limited to the presidential election results but also covers the midterm or US Senate results at the state or county level.

The data abstracted from MITEDSL in this paper is the “County Presidential Election Returns 2000-2020.” It contains the number of votes for each party and the total votes at each US county from 2000 to the 2020 US Federal Election. In this paper, I will only use the data for the 2020 US Election and focus on the counties on the mainland of the US. The **?@tbl-election** summarizes the essential variables.

The **?@tbl-election** summarizes the voting patterns for each party during the 2020 U.S. Presidential Election. Undoubtedly, the two most popular parties are the Democrats and the Republicans. However, even though Republican has almost double the mean percent vote than Democrat, the total votes for Republican is less than for Democrat. This pattern may indicate that more people living in high-population states, such as California, voted for Biden. This finding aligns with the fact that Biden defeated Trump in 2020. I will explain more later.

2.2 Center for Systems Science and Engineering at JHU

The Center for Systems Science and Engineering at JHU is at the Department of Civils and Engineering, which collects local, national, and global multidimensional data, including medicine, health care, disaster response, etc. During the pandemic, they collected the U.S. and international COVID cases and deaths and reported them on their GitHub. Their data is summarized by daily reports ranging from April 12, 2020, to March 9, 2023.

To illustrate the impact of COVID-19 on the 2020 Election to the greatest extent, the COVID data used in this paper will be the daily report on November 3, 2020, which is Election Day [citation]. This report contains the aggregated cases and deaths for each country and counties in the U.S.

2.3 American Community Survey (ACS)

The American Community Survey is the survey conducted by the U.S. Census Bureau. The survey contains a variety of socio-economic variables for each county. Although there are different data tables from ACS, I will use the 2020 five-year estimate of DP02, DP03, and DP05.

These three tables cover the social, economic and demographic characteristics of each county in the U.S. By referring to my previous research, where I found the variables with the highest correlations to the COVID mortality rate, I will use the same variables in this paper. That said, I will extract the data regarding the educational attainment from DP02, especially the proportion of people having at least a bachelor's degree. Regarding the economy, I will take the proportion of people with private health insurance and the mean household income. Lastly, regarding the demographic factors, I will use the total population, the ratios of children and individuals above 85. I will also take the White and Black people percentage as well.

2.4 Data Cleaning

Table 1: The summary of COVID cases and deaths as of Election Day.

Winning Party	Cases	Deaths	Infection Rate	dpc	Income
Democrat	10496.323	294.9605	3010.402	2540.958	83225.72
Republican	1427.794	28.2268	2971.093	1888.240	69794.49

Combining the three data sources, Table 1 shows the COVID cases and deaths and the infection and mortality rate for the county voting for the Democrat and Republican parties, respectively. Surprisingly, the average cases and fatalities of COVID-19 in the counties voting for the Democrats are significantly higher than that of the Republican party. However, the values of infection and mortality rate do not show the same dramatic difference. It also emphasizes that the counties (states) voting for the Democrat are population intensive, and counties with more residence generally have higher mean household income levels. We can observe this from the average income where the “Democratic” counties have about $13k mean income more elevated than the “Republic” ones.

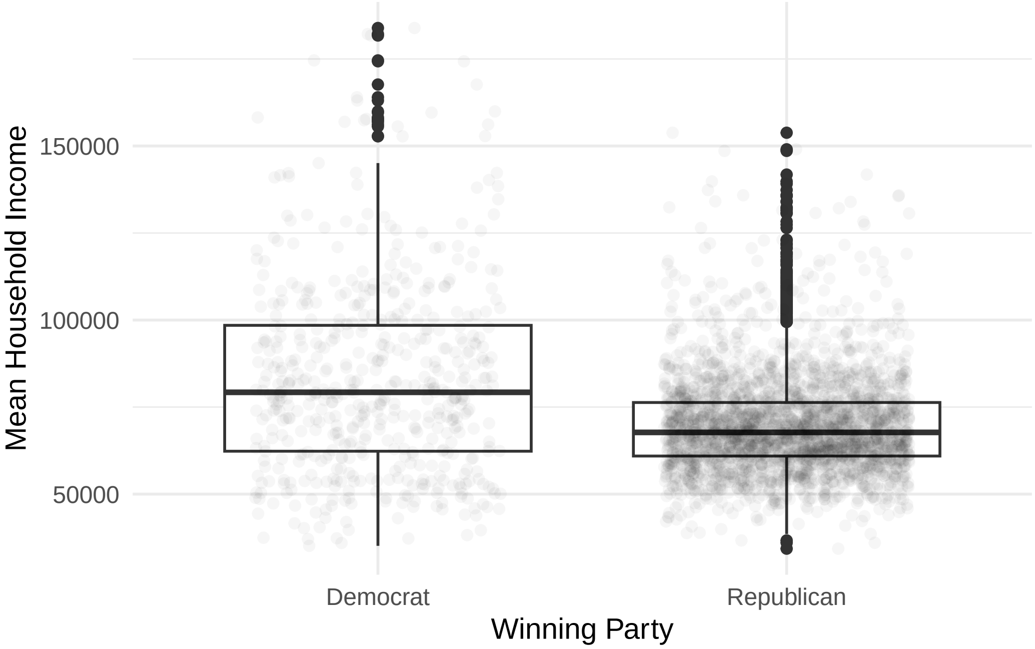
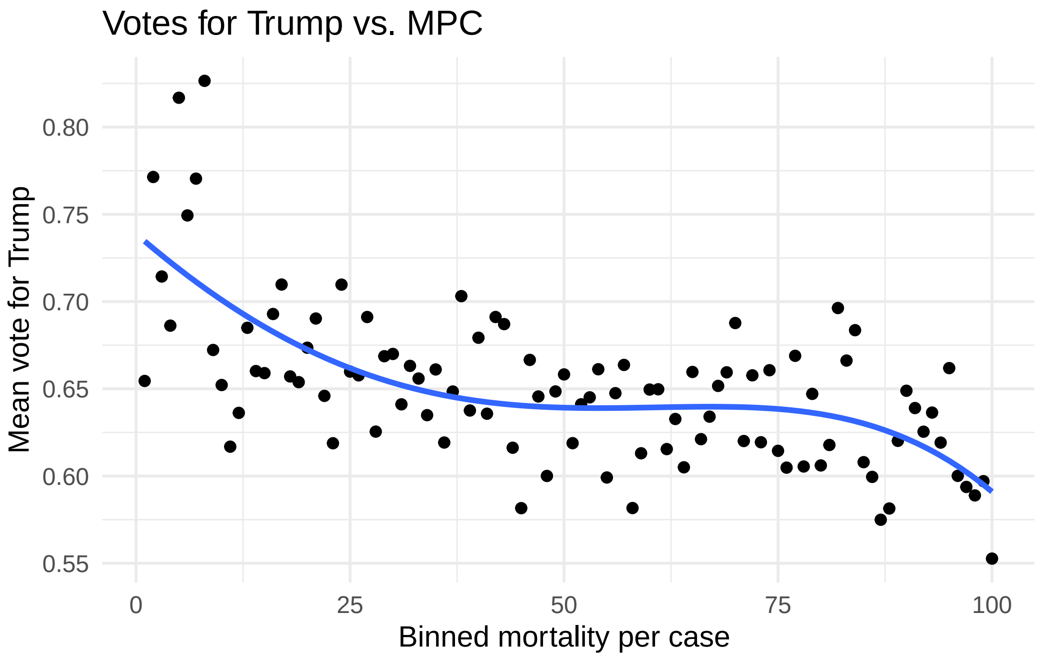


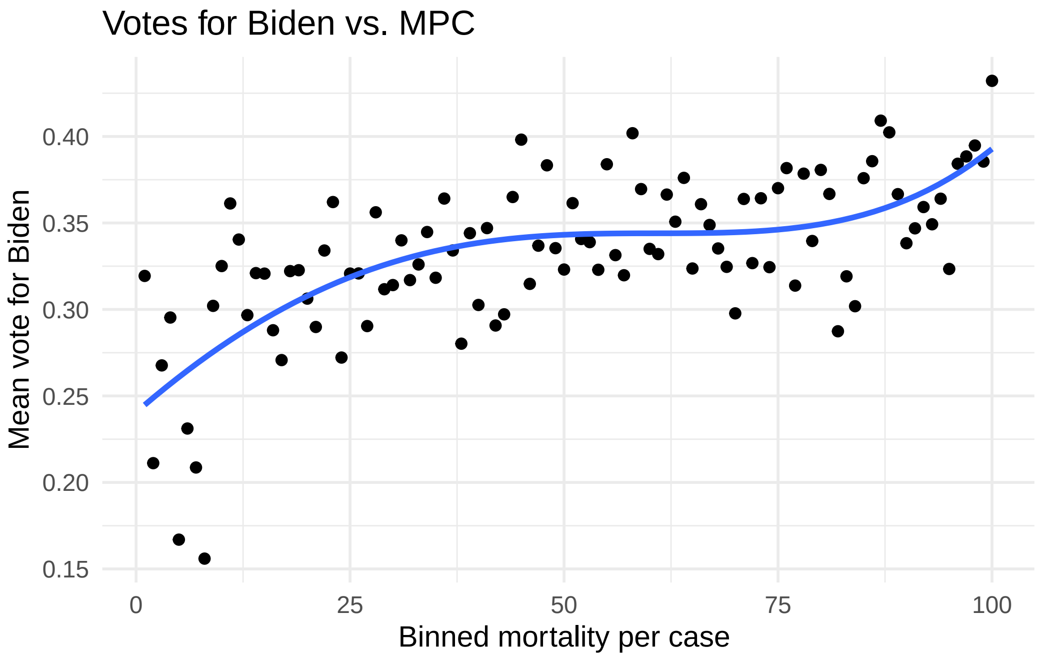
Figure 1: Summary of income levels for counties voting for Democratic and Republican

Figure 1 compares the distribution of county income levels for the two parties. In general, counties voting for the Democrats have higher income levels than the Republicans. In addition, we can observe that the counties are intensively concentrated around income levels $50k to $75k, compared to the Democrats, where the counties are approximately uniformly distributed at each income level. Furthermore, barely any county has at least 150k voting for the Republican party. Both Table 1 and Figure 1 indicates that the Democrat is in favour of wealthy counties but poorer counties for Republican. Consequently, the more affluent counties (states) usually have more electoral votes than the poorer ones; this may provide insights into why Trump lost.

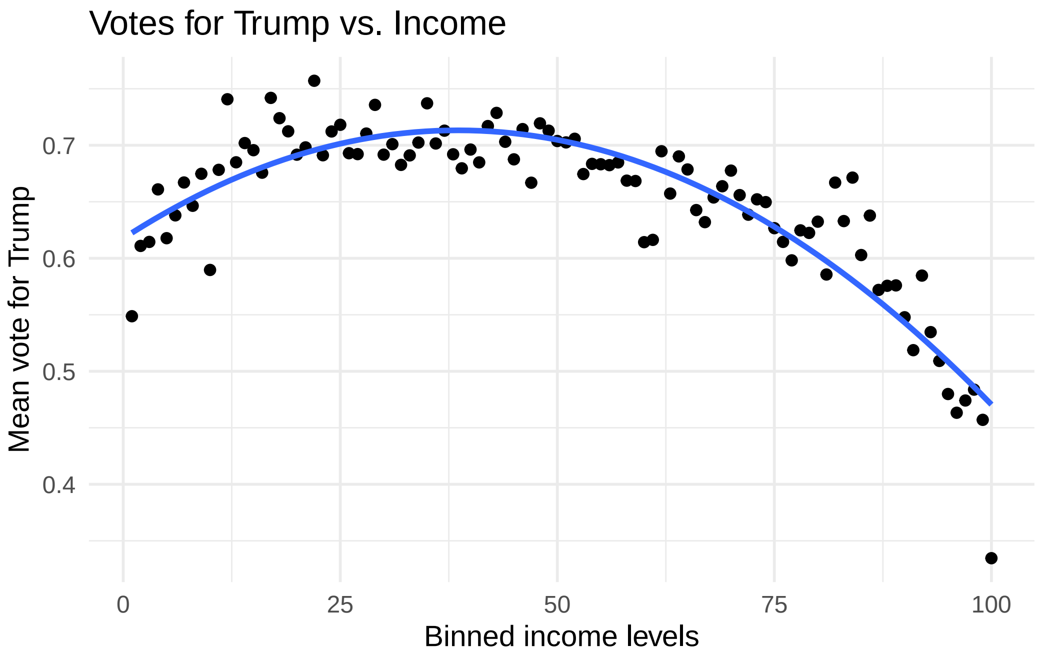
```
binned_data <- data %>%
  mutate(dpc_pctile = ntile(dpc, 100)) %>%  # Adjust 'breaks' to change number of bins
  group_by(dpc_pctile) %>%
  summarize(mean_rep = mean(pct_vote_rep, na.rm = TRUE),
            mean_demo = mean(pct_vote_demo, na.rm = TRUE)) %>%
  mutate(bin = 1:n())

binned_data2 <- data %>%
  mutate(dpc_pctile = ntile(mean_household_income, 100)) %>%  # Adjust 'breaks' to change
  group_by(dpc_pctile) %>%
  summarize(mean_rep = mean(pct_vote_rep, na.rm = TRUE),
            mean_demo = mean(pct_vote_demo, na.rm = TRUE)) %>%
  mutate(bin = 1:n(),
         sum = mean_rep + mean_demo)

# Plotting
ggplot(binned_data, aes(x = bin, y = mean_rep)) +
  geom_point() +  # or geom_line() depending on your preference
  geom_smooth(method = "lm",formula = y ~ poly(x, 4), se=FALSE) +
  theme_minimal() +
  labs(x = "Binned mortality per case ", y = "Mean vote for Trump",
       title = "Votes for Trump vs. MPC")
```



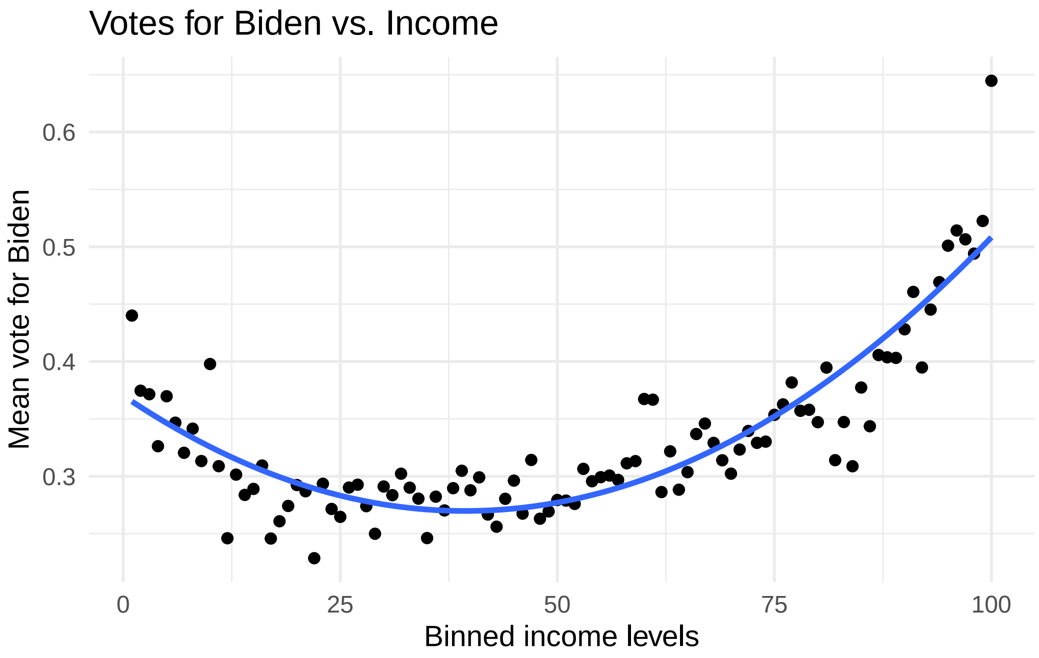
```
ggplot(binned_data, aes(x = bin, y = mean_demo)) +
  geom_point() +  # or geom_line() depending on your preference
  geom_smooth(method = "lm", formula = y ~ poly(x, 4),se=FALSE) +
  theme_minimal() +
  labs(x = "Binned mortality per case ", y = "Mean vote for Biden",
       title = "Votes for Biden vs. MPC")
```



```
ggplot(binned_data2, aes(x = bin, y = mean_rep)) +
  geom_point() +  # or geom_line() depending on your preference
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se=FALSE) +
  theme_minimal() +
  labs(x = "Binned income levels", y = "Mean vote for Trump",
       title = "Votes for Trump vs. Income")
```



```r
ggplot(binned_data2, aes(x = bin, y = mean_demo)) +
  geom_point() +  # or geom_line() depending on your preference
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), se=FALSE) +
  theme_minimal() +
  labs(x = "Binned income levels", y = "Mean vote for Biden",
       title = "Votes for Biden vs. Income")
```



```
library(sp)
library(geojsonio)
library(ggplot2)
library(dplyr)
library(broom)
hex_county <- geojson_read("us_county_hexgrid.geojson", what = "sp")

# Reformat the data as needed (depending on your data structure)
hex_county@data <- hex_county@data %>%
  mutate(adjusted_field = gsub("some_pattern", "replacement", original_field))

# Fortify the data for ggplot
hex_county_fortified <- tidy(hex_county, region = "desired_region_field")

# Plot the hexagon map
ggplot() +
  geom_polygon(data = hex_county_fortified, aes(x = long, y = lat, group = group), fill="#
  geom_text(data = hex_county_fortified, aes(x = long, y = lat, label = label_field), size
  theme_void() +
  coord_map()
```

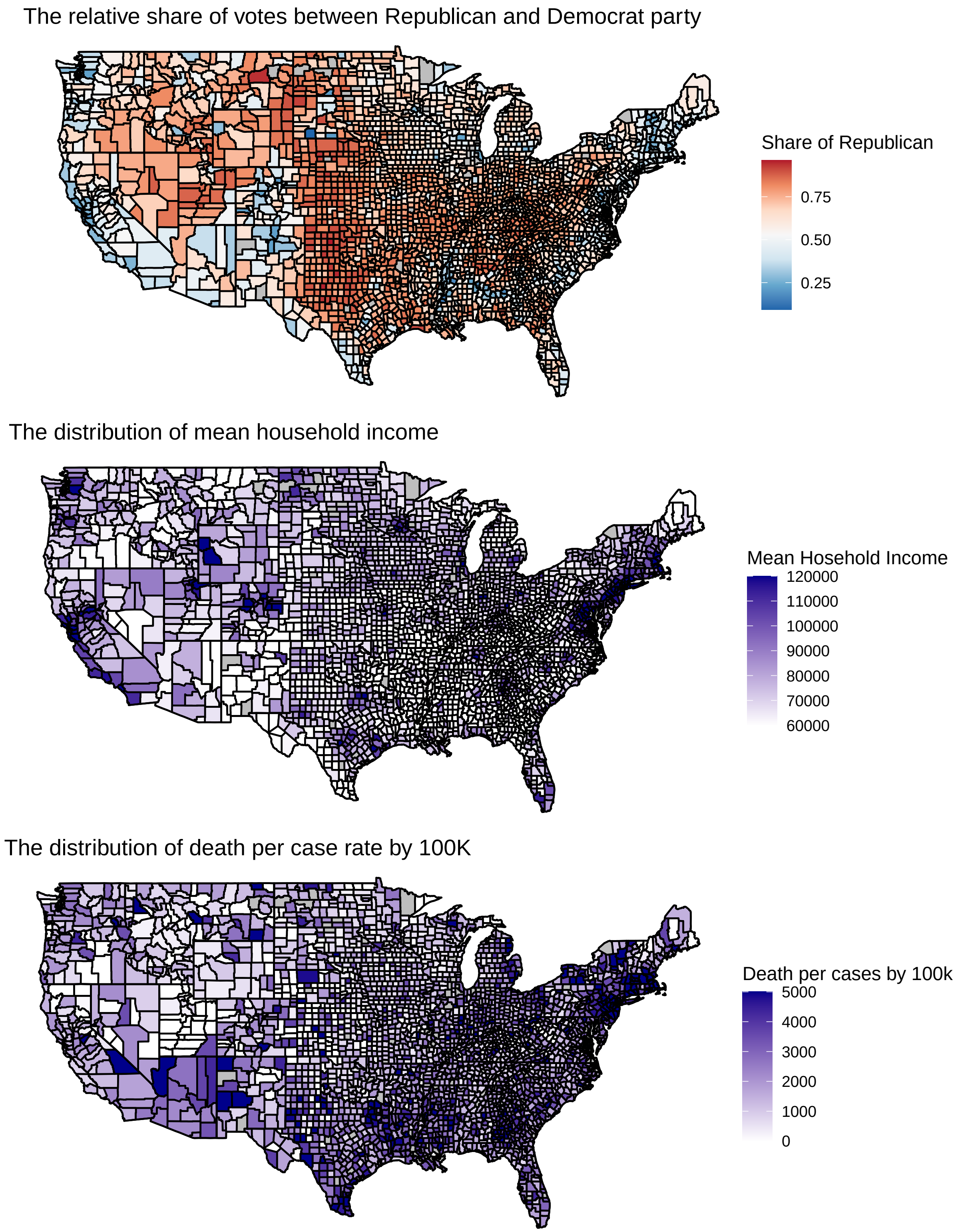

The relative share of votes between Republican and Democrat party
Share of Republican
0.75
0.50
0.25
The distribution of mean household income
Mean Hosehold Income
120000
110000
100000
90000
80000
70000
60000
The distribution of death per case rate by 100K
Death per cases by 100k
5000
4000
3000
2000
1000
0

Figure 2: The ratio of votes for the Republican and the infection rate per 100k in each county

2.5 Data Limitation

Although the merged data covers the voting patterns, socio-economic characteristics and COVID-19 infection rate by county, some counties are missing. From the above two maps, the counties in the middle west show grey, indicating that the data for these counties are absent. This absence of data may increase bias and influence the accuracy of data analyses.

3 Methods

This paper aims to make a causal inference between the 2020 voting patterns and COVID-19 to find whether Trump can contribute to his loss of COVID-19. To see the causal effect in the observational study, I will implement the methods of propensity scores to find the change in votes for Trump due to the pandemic. Based on the results, I will also conduct a counterfactual analysis to see what would have happened if there had been no COVID-19,

First and foremost, we need to define the treatment. The treatment in an experiment will be artificially imposed on specific groups of people to see the effect of interested variables with and without this treatment. Therefore, the people with the treatment are called the treatment group, and the rest are called the control group. The treatment and control group settings are to test whether the imposed factor impacts the interested variable. However, since this is an observational study, it is unrealistic to set a factor manually. Therefore, to find the causal effect of COVID-19 on Trump's voting, I will choose the treatment to be whether a county has a high infection rate.

3.1 Propensity Score

Propensity score is a technique that can help us to decide which county would be in the treatment group. That said, the propensity score in this study is the probability of a county having a high infection rate, given a set of socioeconomic variables. Using the propensity score can ensure that the two counties with similar propensity scores have the same distribution of the observed socioeconomic variables, such as income level, even though they are in different groups. This is fantastic, as we can have many pairs from the treatment of the control group with similar covariates by fixing propensity scores. The key drawback is that we can not estimate the causal effect from observational studies.

In addition, by finding the propensity scores, we can reconstruct our assignment mechanism to a strongly ignorable one. An assignment mechanism is "strongly ignorable" if and only if, given the observed covariates, the potential outcomes are independent of the treatment assignment. This indicates that the voting change for Trump is independent of our assignment that a county has a high or low infection rate.

As the propensity score represents the probability of a county having a high infection rate, we must find a logistic regression to predict this probability. In my previous paper, I already found the best model predicting the mortality rate of COVID in each US county. Even though that paper was to predict the mortality rate, the mortality is correlated with the infection rate. Therefore, I will take the dependent variables in that model to be the covariates in predicting the propensity score for each county, which is:

$$\begin{aligned}\frac{PS}{1-PS} = \beta_0 + & \\ & + \beta_1 \times \text{prop\_higher\_education} \\ & + \beta_2 \times \text{IncomePctile} \\ & + \beta_3 \times \text{no\_insurance} \\ & + \beta_4 \times \text{private\_insurance} \\ & + \beta_5 \times \text{males} \\ & + \beta_6 \times \text{old\_85} \\ & + \beta_7 \times \text{white\_pct} \\ & + \beta_8 \times \text{black\_pct}\end{aligned}$$

Therefore, we can find the PS for each county using the above model. However, there are three ways to use the propensity scores to find the treatment effect.

3.1.1 Propensity Score Matching

The package needed to implement the matching is the `Matching.` Using this package, we can match the observations in the treatment and control groups having the same or similar propensity scores. Then, take only the matched pairs to be the new data set and find the treatment effect between them. In addition, we can also check the balance between the treatment and control groups.

However, the most significant drawback is that the new data set can only contain the matched pairs. We may lose a considerable amount of observations during this process. Our predictions may also not be accurate.

3.1.2 Propensity Score Stratification

As introduced above, using PSM may cause the waste of data and results in bias. However, propensity score stratification will use the entire data set to avoid waste. The way to do this is to split the data into different quantiles and fit a model for each subclass, with the only dependent variable being the treatment variable. Then, calculate the average casual effect for the subclasses.

Even though the PSS can use the entire data set, however, the choice of quantiles will be closely related to the predictions. If we have a large sample and want to split the data into many subclasses, then this method may be computationally slow.

3.1.3 Propensity Score Regression Adjustment

The last method is the most straightforward, meaning we just fit a regression model with the treatment variable and propensity score as the two dependent variables.

Each of the above three methods has its unique advantages, but the results may vary between different data sets. I will perform both of the three methods and compare which one would be the best.

3.2 Counterfactual Analysis

Using the propensity score methodology, we can assess the causal effect of a high infection rate on the voting for Trump. However, it can not tell us the election results, and we can not draw conclusions about COVID and the election. Therefore, I will conduct a counterfactual analysis, meaning that I will re-calculate the votes for Trump and Biden to see if the final election results will change.

Eventually, I will re-calculate the votes for the two candidates to follow the official election rule. By finding the difference in voting percentage, I will re-allocate the voting for the two parties and calculate the electoral votes in each state using the winner-takes-all rule.

4 Results

$$\begin{aligned}
\text{Logit(PS)} = -2.008+ & \\
& -0.032 \times \text{prop\_higher\_education} \\
& -0.009 \times \text{IncomePctile} \\
& +0.088 \times \text{no\_insurance} \\
& +0.051 \times \text{private\_insurance} \\
& +0.019 \times \text{males} \\
& +0.055 \times \text{old\_85} \\
& -0.030 \times \text{white\_pct} \\
& +0.026 \times \text{black\_pct}
\end{aligned}$$

The above equation shows the logistic regression model to predict the propensity score for each county, given the covariates. It seems that insurance is a critical factor in the infection rate. Interestingly, the counties with a higher proportion of people with no insurance and private insurance will both increase the probability of this county having a high infection rate. Using this model, we can find the propensity score for each county and implement the three methods.

```r
model <- lm(pct_vote_rep ~ dpc + mean_household_income, data = data)
summary(model)
```

```

Call:
lm(formula = pct_vote_rep ~ dpc + mean_household_income, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65074 -0.08570  0.02651  0.10489  0.46201 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            9.005e-01  1.133e-02    79.5   <2e-16 ***
dpc                   -1.326e-05  1.524e-06    -8.7   <2e-16 ***
mean_household_income -3.101e-06  1.477e-07   -21.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1486 on 3104 degrees of freedom
Multiple R-squared:  0.1446,    Adjusted R-squared:  0.1441 
F-statistic: 262.4 on 2 and 3104 DF,  p-value: < 2.2e-16


Estimate...  -0.027732 
AI SE......  0.0092623 
T-stat.....  -2.994 
p.val......  0.0027531 

Original number of observations..............  3107 
Original number of treated obs...............  1485 
Matched number of observations...............  1485 
Matched number of observations  (unweighted).  3530 
```

```
Call:
glm(formula = change_vote_rep ~ treatment + prop_score, data = data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0298252  0.0009585  31.116  < 2e-16 ***
treatment   -0.0077685  0.0011398  -6.816 1.12e-11 ***
prop_score  -0.0186307  0.0020676  -9.011  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.0007112916)

    Null deviance: 2.4035  on 3106  degrees of freedom
Residual deviance: 2.2078  on 3104  degrees of freedom
AIC: -13699

Number of Fisher Scoring iterations: 2
```

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.

6 References