How Does Family Income Affect 2020 American Federal Election Result

Causal link between high-income and voter's choice

Xinyu Tan

21/12/2020

Code and data supporting this analysis available at:
https://github.com/Xinyu899/STA304-final-project

Abstract

This analysis aims to find out if there is a causal inference between American's household income and their choice in the 2020 election. A dataset from the Integrated Public Use Microdata Series(IPUMS) will be used in this analysis. To achieve our goal, we will use the propensity score matching method and logistic regression analysis method to see if the high income would affect the American people for choosing Joe Biden. The propensity score matching method is applied in the dataset to find out the causal inference on observational data.

Keywords: Propensity Score, Causal Inference, Observational Study, Household income, The U.S 2020 Persidential Election

Introduction

In the last part of our STA304 course, we learned how to use treatment to find causal inference, and this inspired me to analyze if American's household income would influence their choice on the 2020 presidential election. Causal inference helps us to understand what it means for some variables to affect others. Since the presidential election is a voluntary choice, observational study is more suitable in this situation. According to the information on the website usnews.com, a family of three would be considered as middle class if its household income is at least 53,431 dollars in 2019 (Snider, 2020), and there are 52% population in the U.S are in the middle class(Frankenfield, 2020). So, I define high income as household income is higher than 55,000 dollars adjusted by the number shows in the dataset.

One good way to make causal inference is using propensity score to match observations with the same treatment, and the treatment is high-income in this analysis. Propensity score matching method was first discovered by Paul Rosenbaum and Donald Rubin (Polsky & Baiocchi, 2014). The propensity scores can be used to reduce or eliminate selection bias in observational studies by balancing covariates between treated and control groups. (Glen, 2020). In this analysis, I will find out if there is a causal link between the level of income and people's choice on presidential election.

A survey dataset containing a series of basic information about election voters information will be used to find a causal link between high-income voters and voting for Biden. The following Methodology section will show you the details af the dataset and the model. The Results section will demonstrate e results of the analysis, and the Discussion section will discuss the conclusion drawn from the result. Also, the weaknesses of the analysis and next steps will be included.

Methodology

Data

The dataset is from the Integrated Public Use Microdata Series(IPUMS) website. The original dataset collected from a survey(June 25 to July 1, 2020), and it has 6479 observations and 265 variables. In this report, I selected 8 variables that can describe a person's basic information, such as choice of president(vote_2020), vote intention(vote_intention), gender(gender), age(age), race(race, ethnicity), education(education) and annual household income in USD(household_income). And I created a new binary variable called "high_income" which indicates high-income groups. A person whose annual household income is higher than 55,000 dollars is defined as a high-income individual. Then I filtered out people who said they are not going to vote or not eligible, and missing values are deleted from our data set. So, the dataset now has 4296 observations and 9 variables after cleaning. 8 of the Variables vote_2020, vote_intention, gender, race_ethinicity, education, household_income, high_income and registration are categorical variables, while only variable age is numerical variable.

Some of this dataset's strengths are that it has various variables, and the sample size is large. One issue this dataset might have is that the respondents were not honest in answering the income question, which means we might get inaccurate key information.

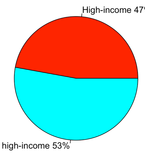
The observations in the dataset can be divided in to two groups by the varibale high_income(1 indicates high income, and 0 indicates non high-incom) : high-income group and non high-income group. Table 1 below shows the numbers of observations in two groups.

Table 1:

Group	Count
high-income	2025
non high-income	2271

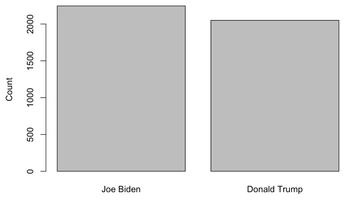
Below is a pie chart(Graph 1) which shows the ratio of high-income groups and non-high income group. There are 47% people in this sample are in high-income group, which is almost a half.

Pie Chart of high-income groups - Graph 1



Below is a barplot(graph 2) which demonstrates people's choices of president in this survey. And it shows that the vote for Joe Biden is slightly higher than Donald Trump.

Intentional Vote Chart - Graph 2



Model

Two logistic regression models were contructed in this analysis, one is calculating the propensity score and the other one is to find out the causal inference. Both models were using the glm function in Rstudio.

The logistic regression model calculating propensity score for each observation is modeled by

$$log(P_{high\ \ income}/1-P_{high\ \ income}) = \beta_0 + \beta_1 * Age + \beta_2 * Gender + \beta_3 * Education + \beta_4 * Race + \epsilon$$

where $P_{high\ \ income}$ represents the probability that an individual is categorized as high-income; β_1 represents the coefficient of the variable age; β_2 represents the coefficient of the variable gender; β_3 represents the coefficient of the variable education; β_4 represents the coefficient of the variable race and ϵ represent the error term. And varible age is a numerical, variables gender, education, race are categorical variables.

After matching the propensity scores, the remaining observations are used to construct a logistic model to predict the voting for Joe Biden, and the model is modelded by

$$log(P_{Biden}/1-P_{Biden}) = \beta_0 + \beta_1 * Age + \beta_2 * Gender + \beta_3 * Education + \beta_4 * Race + \beta_5 * Highincome + \epsilon$$

where P_{Biden} represents the probability that an individual is categorized as high-income; β_1 represents the coefficient of the variable age; β_2 represents the coefficient of the variable gender; β_3 represents the coefficient of the variable education; β_4 represents the coefficient of the variable race; β_5 represents the coefficient of the variable high_income and ϵ represent the error term. And varible age is a numerical; high_income is a dummy variables(household income > 55,000 dollars with the indicator "1" is the base level); variables gender, education, race are categorical variables.

Result

In the first model(logistic regression model calculating propensity sores), four estimated values of coefficients are significant at 5% significant level, and they are shown in the table below. If the individual is male, the log odds of being in the high-income group would decrease by 0.001137. If the individual has completed some high school, the log odds of being in the high-income group would decrease by 1.809186. If the individual is Black, or African American, the log odds of being in the high-income group would decrease by 0.652916. If the individual is other race, the log odds of being in the high-income group would decrease by 0.385337.

Coefficient	Estimate	P-value
Gender - Male	-0.001137	0.00121 **
Education - Completed some high school	-1.809186	0.02130 *
Race - Black, or African American	-0.652916	1.6e-08 ***
Race - Some other race	-0.385337	0.01169 *

In the second model(predict voting for Joe Biden), our interest is to find the causal link between high income and vote desicion, so value of β_5 and its p-value will help us to do this. And the values are shown below. The p-value is 0.003432 which is smaller than 5%, and this means the estimate is statistically significant. Moreover, the estimate is negative(-0.206602), which means if the individual is in the high-income group the log odds of voting for Joe Biden would decrease by 0.206602, so estimated probability of voting for Joe Biden also decreses.

Coefficient	Estimate	P-value
high_income(1)	-0.206602	0.003432 **

Discussion

Summary

The goal of this analysis is to find out the causal link between the high-income group(defined as people whose annual household income is higher than 55,000 USD) and the U.S people's choices on the 2020 presidential election. The survey dataset is from the Integrated Public Use Microdata Series(IPUMS), and only some variables were chosen. The technique of propensity score matching was used to match observation pairs(high-income and non-high-income), then a logistic regression model was built to verify the causal inference.

Conclusion

The propensity score analysis showed that there is no evidence for the causal link between the high-income group and the U.S people's choice on 2020 residential election based on the survey done by IPUMS. The p-value of β_5 in the second model is 0.003432, which is significant at a 5% significance level, but the estimated value is negative. Based on this result, it appears that if an individual is in a high-income group, the likelihood of voting for Joe Biden in the 2020 presidential election will decrease. In other words, the people with high annual household income do not have any obvious inclination towards Joe Biden in the 2020 election based on this analysis.

Weaknesses & Next steps

A problem in the methodology of observational studies is that the experimenters do not have control over the treatments given to participants, and it means that I can not group people into high-income and non-high-income groups randomly. So, the propensity score matching is used to find the confounding factor. One important weakness of this analysis comes from the method of propensity score matching. The true propensity score is never known in observational studies, so you can never be certain that the propensity score estimates are accurate. (Glen, 2020). Moreover, in observational studies, the propensity score analysis has the limitation that remaining unmeasured confounding variables may still be present, thus leading to biased results. (Nuttall,2008) In this report, the propensity score matching is failed to use all of the available information, and only 4 cases of predictors are statistically significant.

Another weakness is that the participant may provide inaccurate annual household income in the survey, meaning that the proportion of the two groups might be different.

The next steps could change or extract more variables from the dataset to do the propensity score matching and the regression model. Another possible way is to use another survey data that includes more samples and construct a random sample from it. And compare with the actual election result to verify the causal inference.

References

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.5. https://CRAN.R-project.org/package=broom

Frankenfield, J. (2020, September 14). Which Income Class Are You? Retrieved December 19, 2020, from https://www.investopedia.com/financial-edge/0912/which-income-class-are-you.aspx

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Integrated Public Use Microdata Series. (2020, October 29). American community surveys. https://usa.ipums.org/usa/index.shtml

Nuttall, G., & Houle, T. (2008, January 01). Liars, Damn Liars, and Propensity Scores. Retrieved December 22, 2020, from https://pubs.asahq.org/anesthesiology/article/108/1/3/7646/Liars-Damn-Liars-and-Propensity-Scores

Polsky, D., & Baiocchi, M. (2014). Observational Studies in Economic Evaluation. Encyclopedia of Health Economics, 399–408. doi:10.1016/b978-0-12-375678-7.01417-6

Snider, S. (2020, December 08). Where Do I Fall in the American Economic Class System? Retrieved December 19, 2020, from https://money.usnews.com/money/personal-finance/family-finance/articles/where-do-i-fall-in-the-american-economic-class-system

Stephanie Glen. "Propensity Score Matching: Definition & Overview" From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/propensity-score-matching/

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