Analysing the Tremendous Impact of Income on Voting Intention in 2020 American President Election using Propensity Score Matching

Shilun Dai

12/9/2020

Abstract

The 2020 America President Election was held on November 3, 2020, which is a milestone for the United States. As it has ended for more than a month now, the news that rich people prefer to vote for Joe Biden evokes the hypothesis that income has a tremendous impact on the voting preference. This report utilized propensity score matching process, along with logistic model to investigate the impact of income on voting preference based on the survey data of 6479 Americans about their personal information, including age, education, income and etc. The report successfully concluded that income is a key element in the 2020 American President Election.

Keywords

Propensity Score, Casual Inference, Observational Study, Matching, Logistic Regression Model, Income, Voting Preference, 2020 American President Election

Introduction

The questions that motivate most studies in the health, social and statistical sciences are not associational but causal in nature(Pearl, 2009). Casual inference method is ubiquitous to observational experiment. It's is often more feasible and reliable than other methods. Thus, having ability to make casual inference is key from both an economic and behavioral perspective.

One popular way to make causal inference is through propensity score matching. Propensity score analysis is a statistical technique developed for estimating treatment effects with nonexperimental or observational data(Guo & Fraser). It was initially introduced in 1983 and became popular recent years(Abadie, A., & Imbens, G., 2016).

As the exciting 2020 America President Election has come to an end for more than one month now, Joe Biden was elected to take charge of the government for the next four years. The competition is fierce, and the result is shown in Table 1 below:

Table 1: The Result of 2020 America President Election

Candidates	Results	Votes
Joe Biden	306	81,283,098
Donald Trump	232	74,222,958
Others	0	3,010,485

And the visualized results separated by states is shown in Figure 1 below:

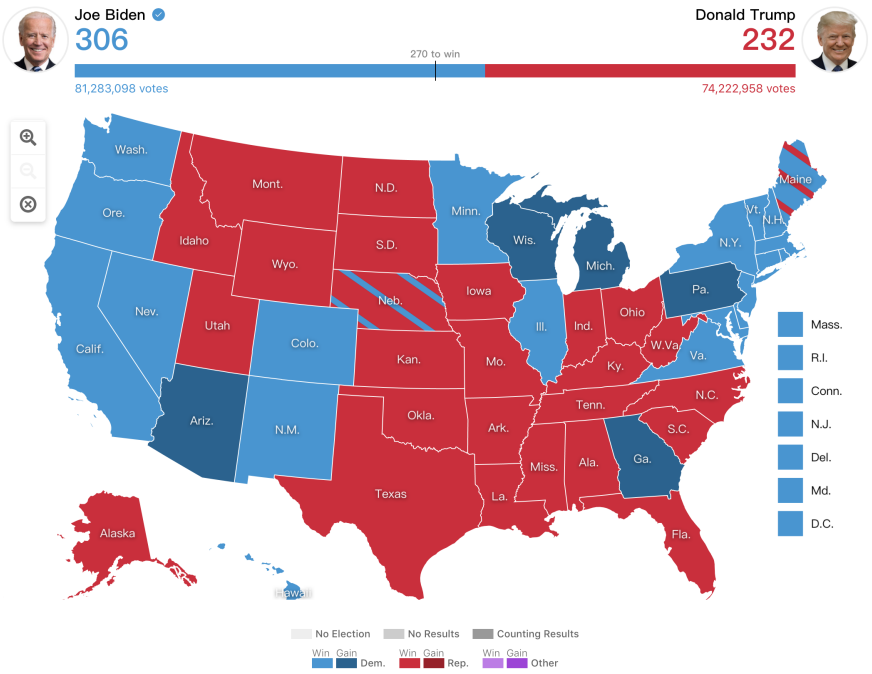

Figure 1: Election Results by States, abc NEWS

Supporters of different parties may have different features, especially income. Understanding the structure of voters of each party is a key point for candidates to win the election. To study this problem, this report will review the 2020 American President Election results using Nationscape data and use propensity score matching technique to conclude if there is a causal link between whether or not voters have high income and whether or not they are willing to voting for Joe Biden. The report successfully generates that income is a key element in 2020 American President Election.

Two data sets will be used to investigate how propensity score could be used to make inferences on the causal link between income and voting preference. In the Methodology section(Section 2), the way to simulate the data along with the logistic model that was used to perform the propensity score analysis was described. In the Results section(Section 3), it showed the results of the propensity score analysis. Finally, the inferences of this data along with conclusions were presented in the Conclusion section(Section 4).

Methodology

Data

We obtained the data set 'Nationscape' on the Voter Study Group website. Nationscape is one of the largest public opinion survey projects(Tausanovitch, Chris and Lynn Vavreck., 2020). It interviews people in nearly every county, congressional district, and city in the leadup to the 2020 election(Tausanovitch, Chris and Lynn Vavreck., 2020). 'Nationscape is fielded by Lucid, a market research platform that provides access to authentic, targeted audiences' (Tausanovitch, Chris and Lynn Vavreck., 2020).

In this study, the target population for this data is all persons who are eligible to vote, i.e., America citizens over the age of 18 or order(The America Voting Population, n.d.). The target frame is the respondents that is 'constructed to be representative of population in terms of a specified set of characteristics' using purposive sampling(Tausanovitch, Chris and Lynn Vavreck, 2020). The target frame is provided by Lucid, a market research platform that runs an online exchange for survey respondents(Tausanovitch, Chris and Lynn Vavreck, 2020). The sample data frame is drawn from this exchange to match a set of demographic quotas on age, gender, ethnicity, region, income, and education, which will be sent from Lucid directly to survey

software operated by the Nationscape team(Tausanovitch, Chris and Lynn Vavreck, 2020). All respondents take the survey online and must complete an attention check before taking the survey and the language used in the survey is English(Tausanovitch, Chris and Lynn Vavreck, 2020).

The data is observational data, meaning the data were collected based on random selection upon observations rather than a fully experimental data collection and introducing a systematic intervention to study any effects. The sampling method for this data set is weighting the survey data to represent the American population(Tausanovitch, Chris and Lynn Vavreck, 2020). The total observation of this data is 4296 and the estimated population size for this survey is 6479(Tausanovitch, Chris and Lynn Vavreck, 2020).

The survey uses appropriate weighted sampling strategy since the observations that are collected will represent the target population accurately(Tausanovitch, Chris and Lynn Vavreck, 2020). The multi-diversity and detailed information of each observation involved in this data make it a good source for propensity score matching. Another advantage of the survey is that the size of data set is large, with a large number of observations and diverse variables. Large sample size will provide more accurate estimators to estimate the population. However, the data we use were collected on 25th June, 2020 which was not the latest information. Thus, the model we built will only be accurate when it's applied to estimate the effect of income before June 2020.

The Nationscape data set has quite a lot of predictors, a few of which are considered important to this analysis are: state of residence, gender, age, race, education background, income and voting choice. We selected the variable *vote_2020* and eight other variables from the data set to do the analysis.

vote_2020 is a categorical variable representing the candidate a respondent wants to vote for. We only leave the observations voting for candidates Donald Trump and Joe Biden since other candidates only have minor supporters. *voting_intention* is a categorical variable demonstrating whether or not a respondent would like to vote in the election. We filter the observations who respond 'Yes, I will vote' and 'Not sure' in this part, ignoring those who respond 'No, I am not eligible to vote' and "No, I will not vote but I am eligible". *registration* is a categorical variable that shows whether or not voters will register to vote in 2020. Obviously, we only choose those who registered. *age* is a numerical variable, measured in years. *gender* is a categorical variable. It was treated as a dummy variable in the logistic regression model. *education* is a categorical variable and respondents will choose their education background. *state* is a categorical variable representing the geometric location that a respondent living in right now. *household_income* is a categorical variable. It shows the number of dollars the respondent earned per year. The last variable is *race_ethnicity* representing observations' race. Besides, we construct a new variable *income_high*. It is a categorical dummy variable representing whether or not the income of respondents is greater than $\$150,000$ in one year. The variable *income_high* equals to 1 (Yes) if they earn more than $\$150,000$ and equals to 0 (No) otherwise.

Those variables were selected because they are representative which have minor duplication between each respondent. And they contain more valid responses and are likely to have relationships with voting preference and income. In this section, We first filter the individuals over the age of 18 that are eligible to vote. Those who are registered and would have intention to vote in 2020 will be eventually obtained in our data set. Moreover, since most observations will vote for Donald Trump and Joe Biden, so we ignore voters who obtain other choices. What's more, we classify the income of these people into high income and low income. Those who earn more than $\$150,000$ in a year are labeled as having high income. Finally we removed NAs in the data set. The total variables in our data frame is 10.

Here is a glimpse of the data set:

Table 2: The Information of the Part of Respondents in data frame

vote_2020	vote_intention	registration	age	gender
Donald Trump	Yes, I will vote	Registered	49	Female
Donald Trump	Yes, I will vote	Registered	46	Female
Donald Trump	Yes, I will vote	Registered	75	Female
Donald Trump	Yes, I will vote	Registered	52	Female
Joe Biden	Yes, I will vote	Registered	21	Female

vote_2020	vote_intention	registration	age	gender
Joe Biden	Yes, I will vote	Registered	38	Female

education	state	household_income	race_ethnicity	income_high
Associate Degree	WI	$75,000 to $79,999	White	0
College Degree (such as B.A., B.S.)	VA	$175,000 to $199,999	White	1
High school graduate	TX	$65,000 to $69,999	White	0
High school graduate	WA	Less than $14,999	White	0
Completed some college, but no degree	MA	$80,000 to $84,999	White	0
Completed some college, but no degree	TX	Less than $14,999	Black, or African American	0

Besides, the data also presents some other interesting aspects. For example, the distribution of age. Based on 2020 Nationscape data, Figure 2 demonstrates that the mean age is approximately 48 years old. Most people in this data are order than 48 years old for the reason that the distribution of boxplot for age is right-skewed. Figure 3 demonstrates the barplot of gender and whether or not the respondent earns more than $\$150,000$ in one year. We notice that there are more males than females in this data frame and most of them does not have high income. However, the difference in number of males and females is not large, which means that the number of gender is balanced. It shows the distribution of respondents' income per year in Figure 4. Most voters earn less than $\$14,999$ for one year, those who earn $\$95,000$ to $\$99,999$ is chasing right behind. There are the least people who earn between $\$85,000$ to $\$89,999$ per year. Figure 5 states the education background of respondents. Most people graduated from college school, followed by those who completed some degree but get no degree. The number of people who graduate from 3rd grades or less is the least. From what we have seen in Figure 6 which demonstrates the distribution of race, white people lead in the distribution of race-ethnicity, and Black, or African American follow behind. Only very few portions of voters are other races.

Generally, from Figure 2 to Figure 6, we conclude that most people in the data frame are white males who are 48 years old or older with low income, obtaining college degree.

Model

In order to look into the potential impact of income on the result of 2020 American President Election, this report utilized Rstudio software to implement propensity score matching technique with logistic model to calculate and estimate the voting preference of different income.

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics(Austin, 2011). The propensity score allows one to design and analyze an observational study(Austin, 2011).

In this section, we first build a logistic regression model to calculate the probability of whether or not a voter has high income, which is propensity score. Then we match the voters with closest propensity score, with one is treated with high income and another is not. Finally, we reduced the data set with matched pairs to generate the impact of income on voting preference.

The logistic model will be focused on five large aspects of voters: age, gender, education background, state and race. Age will be a comparable larger aspect compared to other aspects, due to younger voters will tend to earn less money, while elder voters will tend to earn more. Sex could also be an interesting aspect, that voters of different sexes might have different income. Generally speaking, males earn more than females. The educational level is considered a high-weighted aspect. Although the income is not necessarily based on higher education backgrounds, the potential benefit from higher educations could still benefit to get a great

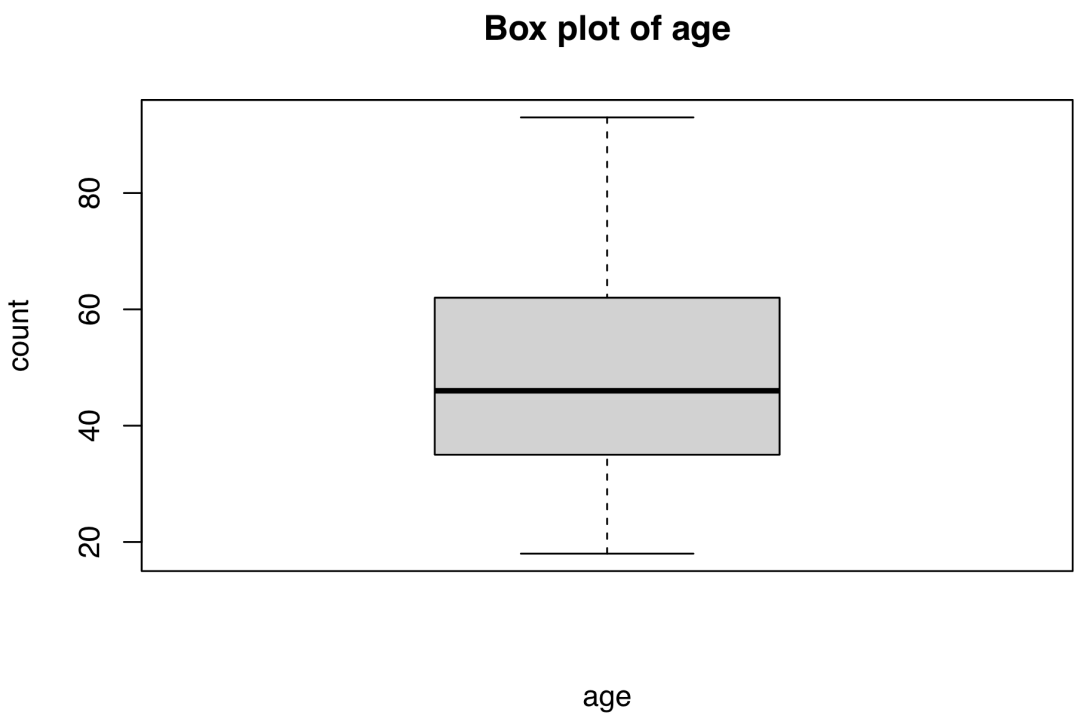

Figure 2: Box plot of age

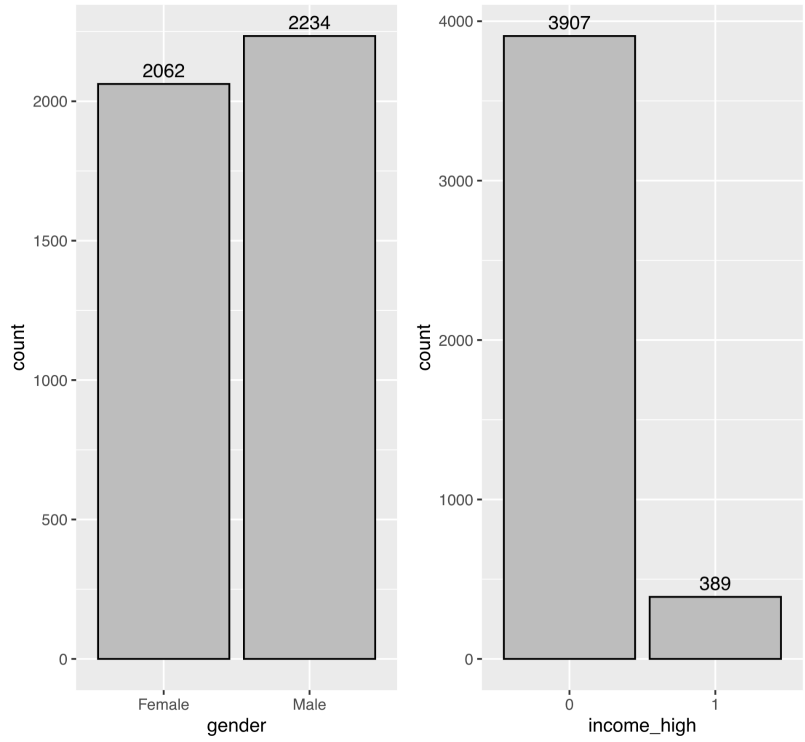

Figure 3: Bar plot of gender and income_high

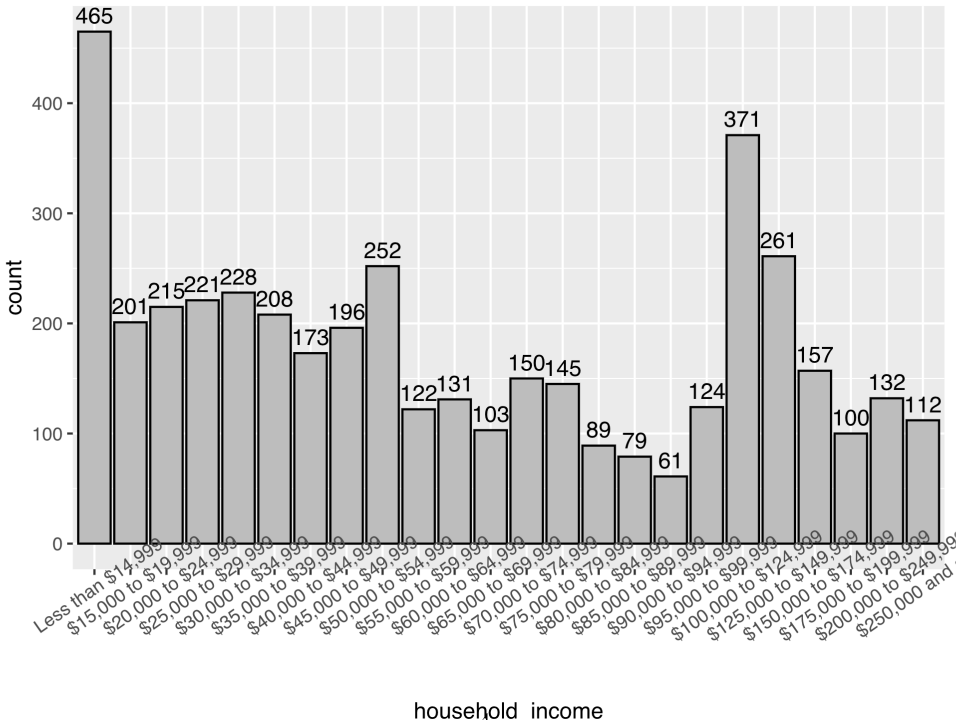

Figure 4: Bar plot of household income

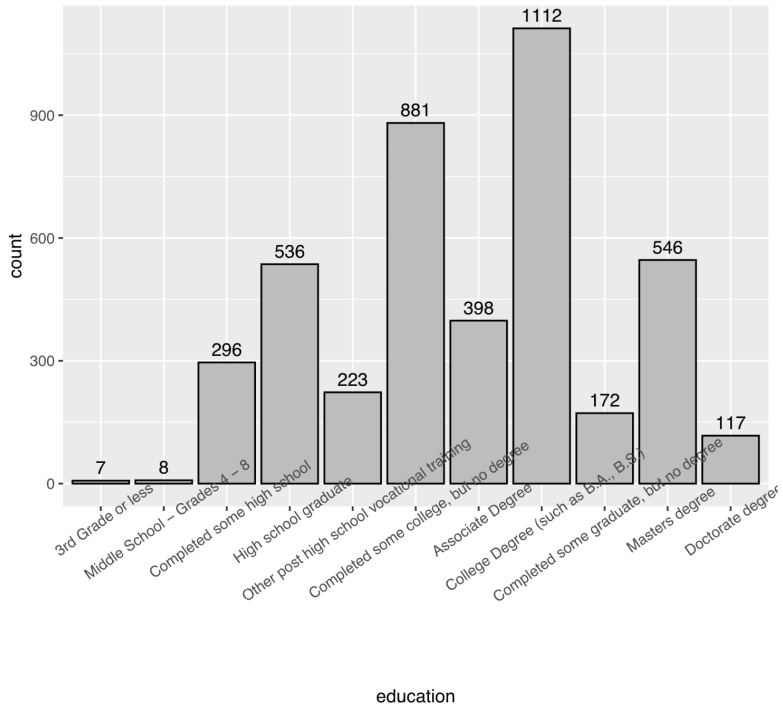

Figure 5: Bar plot of education

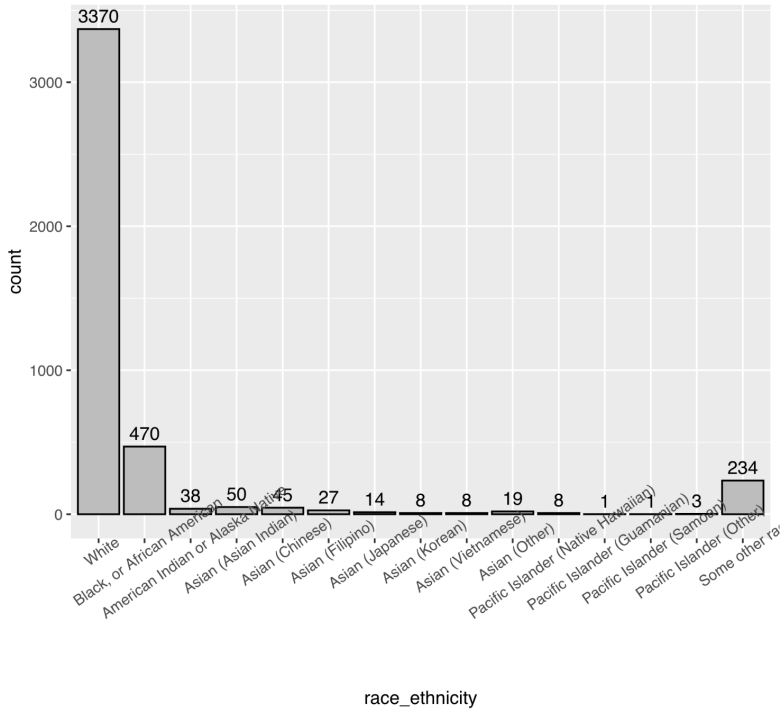

Figure 6: Bar plot of race

job. States may influence the voters' income level based on consumption level and culture. Finally, race is considered another important aspect. White males are considered to earn more salaries. Logistic model has the advantage of estimating binary outcomes, which is either 1 (yes) or 0 (no). Transforming income into whether or not they have high income can greatly take advantage of logistic model to calculate the propensity score.

Using the data selected, we build a logistic regression model to calculate the propensity score with the following formula:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{age,j} + \beta_2 x_{gender,j} + \beta_3 x_{education,j} + \beta_4 x_{state,j} + \beta_5 x_{race\_ethnicity,j}$$

$log(\frac{\hat{p}}{1-\hat{p}})$ represents the log odds of having high income. $\hat{p}$ is the probability of obtaining high income, which is calculated as the propensity score. β_0 represents the intercept parameter, which shows the log odds of having income greater than $\$150,000$ in a year when the age group is 20 or less; gender is female; education is 3rd Grade or less; state is AK(Alaska), and race ethnicity is white. For other βs, β_1, β_2, β_3, β_4, and $\beta_{5,j}$, they all represent the corresponding slope parameters. Each of them shows the change in log odds of obtaining high income when the x corresponding to β changes by 1 in dummy variable coding. Positive coefficients indicate the predictors have positive impacts on whether a voter has high income, for example, a positive $beta_1$ indicate the older the voter is, the higher the income, and vice versa. Also, the larger the absolute value is, the heavier the predictor can impact the income.

After using logistic model to fit the data, we get the following equation:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -33.5482 - 0.0115 x_{age,j} + 0.5908 x_{genderMale} + \beta_{3,j} x_{education,j} + \beta_{4,j} x_{state,j} + \beta_{5,j} x_{race\_ethnicity,j}$$

In which $\beta_{3,j}$, $\beta_{4,j}$ and $\beta_{5,j}$ are coefficients for corresponding jth variable.

In the next step, we start to match respondents with similar propensity score. Propensity score matching process entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score (Rosenbaum & Rubin, 1983a, 1985). Once a matched sample has been formed, the treatment effect can be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample.

To begin propensity score matching process in our data set, we would match the two voters that have similar propensity score while one is treated and others are not. In particular, this finds the closest respondents who have the similar gender, age, education, state and race that does not has high income, to each one that has high income. Then we reduce the data set to those that are matched so that there is a balance between treatment and comparison group on observable traits. One advantage of using propensity score matching is it balances the distribution of groups that are treated and untreated, which makes observed baseline covariates be similar(Austin, 2011). We had 389 treated, so we expect a dataset of 778 observations.

Part of reduced matched pairs are shown in Table 4 below:

Table 4: The Information of Part of Matched Pairs

vote_2020	vote_intention	registration	age	gender	education	state
Joe Biden	Yes, I will vote	Registered	19	Female	High school graduate	GA
Joe Biden	Not sure	Registered	19	Female	High school graduate	GA
Donald Trump	Yes, I will vote	Registered	65	Male	Other post high school vocational training	PA
Joe Biden	Yes, I will vote	Registered	60	Male	Other post high school vocational training	IN
Donald Trump	Yes, I will vote	Registered	58	Male	Other post high school vocational training	TN
Donald Trump	Yes, I will vote	Registered	63	Female	Other post high school vocational training	TX

education	state	household_income	race_ethnicity	income_high	.fitted
High school graduate	GA	$95,000 to $99,999	Black, or African American	0	0.0061748
High school graduate	GA	$150,000 to $174,999	Black, or African American	1	0.0061748
Other post high school vocational training	PA	$45,000 to $49,999	White	0	0.0084773
Other post high school vocational training	IN	$150,000 to $174,999	White	1	0.0084797
Other post high school vocational training	TN	$65,000 to $69,999	White	0	0.0090458
Other post high school vocational training	TX	$175,000 to $199,999	White	1	0.0090512

Finally, we perform a logistic regression to predict whether or not the indicidual will vote for Joe Biden using the following formula:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{age,j} + \beta_2 x_{gender} + \beta_3 x_{education,j} + \beta_4 x_{state,j} + \beta_5 x_{race\_ethnicity,j} + \beta_6 x_{income\_high,j}$$

$log(\frac{\hat{p}}{1-\hat{p}})$ represents the log odds of the voting for Joe Biden. $\hat{p}$ is the probability of voting for Joe Biden. β_0 represents the intercept parameter, which shows the log odds of voting for Joe Biden when the age group is 20 or less; gender is female; education is 3rd Grade or less; state is AK(Alaska), race-ethnicity is white and income is low(less than $\$150,000$). For other βs, β_1, β_2, β_3, β_4, β_5 and β_6 they all represent the corresponding slope parameters. Each of them shows the change in log odds of voting for Joe Biden when the x corresponding to β changes by 1 in dummy variable coding.

After using logistic model to fit the data, we get the following equation:

$$log(\frac{\hat{p}}{1-\hat{p}}) = 0.5377+0.0105x_{age,j}-0.9181507x_{genderMale}+\beta_{3,j}x_{education,j}+\beta_{4,j}x_{state,j}+\beta_{5,j}x_{race\_ethnicity,j}+0.4992x_{income\_h}$$

In which $\beta_{3,j}$, $\beta_{4,j}$ and $\beta_{5,j}$ are coefficients for corresponding jth variable.

Results

The following table represent the result of propensity score regression.

Table 6: The Result of Logistic Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.5377236	1.7070851	0.3149952	0.7527653
age	0.0105320	0.0060305	1.7464520	0.0807324
genderMale	-0.9181507	0.1911790	-4.8025696	0.0000016
educationHigh school graduate	-1.3908831	1.4507187	-0.9587546	0.3376824
educationOther post high school vocational training	-1.4834129	1.7631738	-0.8413311	0.4001625
educationCompleted some college, but no degree	-0.1487432	1.3684731	-0.1086928	0.9134462
educationAssociate Degree	-0.3981679	1.3815006	-0.2882140	0.7731829
educationCollege Degree (such as B.A., B.S.)	-0.4376015	1.3368780	-0.3273310	0.7434176
educationCompleted some graduate, but no degree	0.2085617	1.3819305	0.1509206	0.8800384
educationMasters degree	-0.3329751	1.3315876	-0.2500587	0.8025419
educationDoctorate degree	-1.0325814	1.3523447	-0.7635490	0.4451361
stateAZ	15.7282809	758.7219359	0.0207300	0.9834611
stateCA	0.1697587	1.0762275	0.1577349	0.8746657
stateCO	0.4205414	1.1793827	0.3565776	0.7214081
stateCT	0.9958414	1.2862338	0.7742305	0.4387945
stateDC	-0.0381212	1.3227328	-0.0288200	0.9770081
stateDE	1.0481354	1.4276098	0.7341890	0.4628336
stateFL	-0.0931323	1.0913347	-0.0853380	0.9319927
stateGA	-1.2040247	1.2260566	-0.9820303	0.3260849
stateHI	0.0246389	1.5766383	0.0156275	0.9875316
stateIA	15.5022810	905.9011838	0.0171126	0.9863468
stateID	-14.6596374	1455.3979364	-0.0100726	0.9919634
stateIL	0.1773754	1.1415153	0.1553860	0.8765171
stateIN	0.3615534	1.2903996	0.2801872	0.7793339
stateKS	-0.5606858	1.3863673	-0.4044281	0.6858980
stateLA	-1.0421187	1.5857375	-0.6571824	0.5110637
stateMA	0.3127182	1.2163460	0.2570964	0.7971044
stateMD	-0.5681020	1.1907778	-0.4770848	0.6333017
stateMI	-0.4994001	1.2526074	-0.3986884	0.6901228
stateMN	0.6185156	1.4821623	0.4173063	0.6764544
stateMO	-0.5206426	1.2114303	-0.4297751	0.6673592

term	estimate	std.error	statistic	p.value
stateMS	14.9160791	851.0920879	0.0175258	0.9860171
stateNC	-1.3501232	1.4507075	-0.9306654	0.3520267
stateND	-14.7853279	1455.3979262	-0.0101590	0.9918945
stateNH	0.8582036	1.3646123	0.6288992	0.5294150
stateNJ	-0.2941588	1.1126936	-0.2643664	0.7914976
stateNM	0.9579558	1.7456088	0.5487804	0.5831562
stateNV	-0.3995959	1.5840283	-0.2522657	0.8008357
stateNY	-0.2080342	1.0787028	-0.1928559	0.8470718
stateOH	-0.1090327	1.1510287	-0.0947263	0.9245322
stateOK	0.0074424	1.6484153	0.0045149	0.9963976
stateOR	0.8231532	1.2868289	0.6396758	0.5223834
statePA	-0.0007970	1.1465549	-0.0006952	0.9994453
stateSC	-1.4367223	1.5478464	-0.9282072	0.3533001
stateTN	-0.2014154	1.2833809	-0.1569413	0.8752911
stateTX	-0.4797285	1.1050483	-0.4341244	0.6641981
stateVA	0.1407039	1.1069546	0.1271090	0.8988541
stateVT	15.9144984	1026.7655355	0.0154996	0.9876336
stateWA	-0.2166528	1.3107603	-0.1652879	0.8687174
stateWI	0.3381863	1.2357214	0.2736752	0.7843342
stateWV	0.0757397	1.8417402	0.0411240	0.9671970
race_ethnicityBlack, or African American	3.8671477	1.1010527	3.5122276	0.0004444
race_ethnicityAsian (Asian Indian)	2.6889801	1.2171128	2.2093105	0.0271531
race_ethnicityAsian (Chinese)	2.3366472	1.0761674	2.1712673	0.0299110
race_ethnicityAsian (Filipino)	0.0058850	0.9719477	0.0060549	0.9951689
race_ethnicityAsian (Japanese)	0.2208460	1.1229571	0.1966647	0.8440900
race_ethnicityAsian (Korean)	0.1545801	1.0433728	0.1481542	0.8822211
race_ethnicityAsian (Vietnamese)	0.4847420	0.8583372	0.5647454	0.5722469
race_ethnicitySome other race	0.7033309	0.5171973	1.3598889	0.1738651
income_high1	-0.4991505	0.1609047	-3.1021493	0.0019212

The propensity score analysis showed that the people who have high income are more likely to vote for Joe Biden. It appears that if a voter earns more than $\$150,000$, there is 37.77% more possible to vote for Joe Biden. Using logistic regression model along with propensity score matching technique, this report is also able to calculate the p-value of income to indicate the impact of income on voting preference. Since the p-value of *income_high*1 is 0.0019 which less than 0.05, it is statistically significant. So we have strong evidence to reject the null hypothesis, which means that income significantly affects the result of election.

Disscussion

Summary

Observational study not only needs to be accurate but also relevant, time-efficient and cost-efficient. In this report, I have introduced the Nationscape data set containing information of respondents related to 2020 American president election. To evaluate the effect of whether or not a respondent earns high income in a year on the result of election, we perform logistic regression model to calculate propensity score and match closest respondents in pairs between treated and untreated. In this progress, I introduce the concept of the propensity score and described how to use propensity score matching to design and analyze the observational study.

First, the propensity score is a balancing score: the distribution of observed baseline covariates is similar between treated and untreated subjects based on the propensity score. We use whether or not a respondent has high income as response variable and all other variables, excluding *voting_intention*, as explanatory

variables to perform a logistic regression model using a binomial family. Whereas, the odds ratio of probability of containing high income is calculated when logistic regression model is used. The probability of containing high income is calculated as propensity score.

Second, propensity score matching method allows one to estimate treatment effects in metrics. Methods for assessing the specification of the propensity score model are based on finding the similar propensity score between treated and untreated subjects. After getting matched pairs, we reduce the data set to just those that are matched so that there is a balance between treatment and comparison group on observable traits.

Finally, we perform a logistic regression model again on reduced data set to predict respondents' voting intention, which is the response variable, using other variables as explanatory variables. With propensity score methods, one can more easily assess whether observed confounding has been adequately eliminated, whereas this is more difficult to assess when regression-based approaches are used.

Conclusions

After getting the p-value of obtaining high income from logistic regression model, we will conclude whether or not the high income will affect voting intention. If p-value is less than 0.05, then the high income is statistically significant. In this case, the p-value is 0.0019212, so the high income will significantly influence respondents' voting preference. The propensity score analysis showed that the people who have high income are 37.77% more likely to vote for Joe Biden.

Weakness & Next Steps

Generally, there are three weaknesses in our model. First of all, the data set we used to build model was generated on 25th June, 2020 which was not the latest data before the election and it is not difficult to predict that there could be a portion of people not actually voted for the exact same candidate as they mentioned when taking the survey. The nearer the election is, the fiercer the competition would be, and at the same time, situations could have changed drastically. Moreover, the size of Nationscape data is not big enough. Various NAs exist in the data set and we dropped some observations to avoid them, that's why our model is not close to perfect. Furthermore, our computing power is limited, so we cannot build more complex models.

Our model will be more accurate if we can get the latest data before election. And we can add more data to survey data set or create a follow-up survey to reduce NAs and generate more observations. As for modeling, a few more techniques could be implemented, for example, using Bayesian model could produce better results. Last but not least, there is a caveat in the model specific part that would allow for some new technique.

Appendix

Code and data supporting this analysis is available at: https://github.com/ShilunDai/FinalProject_304
Nationscape data was downloaded from: https://www.voterstudygroup.org/publication/nationscape-data-set

References

1. Austin, P. C. (2011, May). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/
2. Abadie, A., & Imbens, G. (2016). Matching on the Estimated Propensity Score. Retrived from https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11293
3. 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
4. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.2. https://CRAN.R-project.org/package=broom
5. Election 2020 Results and Live Updates. (n.d.). Retrieved from https://abcnews.go.com/Elections/2020-us-presidential-election-results-live-map/

6. 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
7. Guo, S., & Fraser, M. W. (n.d.). Propensity Score Analysis. Retrieved from https://books.google.com.hk/books?hl=zh-CN&lr=&id=5Y_MAwAAQBAJ&oi=fnd&pg =PP1&dq=the importance of propensity score&ots=WX55jLZA5x&sig=HS-OcK14mr_GPlIS_55kxedM9k0&redir_esc=y#v=onepage&q=the importance of propensity score&f=false
8. Pearl, J. (2009). Causal inference in statistics: An overview. Retrieved from https://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf
9. Tausanovitch, Chris and Lynn Vavreck., 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set.
10. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
11. The America Voting Population. (n.d.). Retrieved from https://www.arcgis.com/apps/Cascade/index.html?appid=4f72ccefe3444314830c821b7d42c723