Education_employment_causal_inference_analysis

StudentA

12/9/2020

Abstract

Education believes to have a substantial effect on employment prospects. This paper aims to investigate the causal links between education and employment by employing the propensity score method. Demonstration of the propensity score matching method is presented in examining the causal relationship between education and employment.

Keywords

propensity scores, causal inference, employment, education

Introduction

Statistical analysis has been widely implemented in carrying out research. It involves collecting and selecting data, drawing meaningful conclusions, and reporting findings. Causal inference is the crucial process in concluding a causal link based on varying conditions of an effect. Causal inference analyzes the response of the effect variables with the changes of the cause. It is thus of great importance to use causal inference in investigating a potential relationship between two variables.

Observational data is often considered more feasible and reliable than experimental data. One problem present in observational data is confounding. It impedes one's ability to draw causal inferences. Propensity score matching methods is thus introduced and widely implemented. The propensity score matching method is used in this context of examining a causal effect of education on employment.

Propensity score matching is a quasi-experimental method in which a statistical techniques is used to construct an artificial control group by matching units with similar characteristics with whether or not being treated being the only difference. The dataset can then be further reduced and the matchings can be used to estimate the impact of an intervention. I will implement this method to discern a causal link between whether or not a person receives high education and whether or not a person is employed.

Dataset obtained from the Voter study group will be used. The propensity score matching method is implemented in making inferences on the causal link between education and employment. In the Methodology section, data and the model used for propensity analysis are described. The results section includes the result derived from the propensity score analysis from the previous section. The inference drawn on the data and conclusions is included in the Conclusion section.

Methodology

Data

The data was generously provided by Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [https://www.voterstudygroup.org/downloads?key=181743d7-a018-45c1-b048-bdb2eceac7dd].

Nowadays, the random digit dialing method is characterized by inadequate responses. It is challenging to reach more targeted respondents, and thus such a low response rate can essentially jeopardize any attempt to make generalized findings. Hence, Nationscape, a partnership between Democracy Fund Voter Study Group and UCLA Political Scientist, provides a convenience dataset containing data selected on a set of demographic criteria. Purposive sampling is employed in intentionally determining information based on their representativeness of the population in terms of specific characteristics. The survey conducted 500,000 interviews of Americans from July 2019 through December 2020, covering the 2020 campaign and election. The survey includes interviews with roughly 6,250 people per week. The survey is available online, and an attention check is required before the study. The survey is conducted entirely in English.

There are 6479 observations in the original dataset containing 265 variables. For the purpose of this study, we first filter the data to 3635 observations and 7 variables.

A variable named 'unemployment' is created to include observations characterized by inability to be employed or not in the workforce. Our response variable 'emp' is then created to have 2 categories, with 1 representing employment and 0 representing unemployment. In categorizing our treatment into 2 categories, a variable named 'secondary_or_lower' is created to include observations characterized by secondary or lower levels of education. Our treatment variable 'edu' is then created to have 2 categories, with 1 representing post-secondary or higher levels of education and 0 representing secondary or lower levels of education.

The predictors are edu, gender, race_ethnicity, state, home_income, and age. race_ethnicity was categorized into 4 groups namely Asian, Black, White, and Others. state was categorized into 4 categories namely Northeast, Midwest, South, and West. home_income was created based on the original variable household_income, with 1 representing high income and 0 represents low income.

After propensity score matching, the filtered data was reduced to 2000 observations, thus 1000 pairs of matched observations.

Table 1-Baseline of characteristics of the data: home_income0: low income genderfemale: female edu0: secondary or lower education including race_ethnicity_asian: asian home_income0: low household income state_midwest: states belongs to midwest region

Model

Generalized Linear Model is used in this research. For the purpose of implementing propensity score matching, we have coined a new variable 'emp' to have 2 categories, employed and unemployed, where: - 1 represents the number of people that are currently full-time employed, part-time employed, or self-employed - 0 represents the number of people that are currently a homemaker, Unemployed or temporarily on layoff, Permanently disabled, or a Student

Predictors like home_income, gender, edu, state, and race_ethnicity are categorical with one expection, age, which is numerical.

Our model equation is:

$$Pr(emp = 1) = \text{logit}^{-1}\left(\alpha_{a[i]}^{home_income} + \alpha_{e[i]}^{age} + \alpha_{s[i]}^{gender} + \alpha_{d[i]}^{edu} + \alpha_{e[i]}^{state} + \alpha_{s[i]}^{race_ethnicity}\right)$$

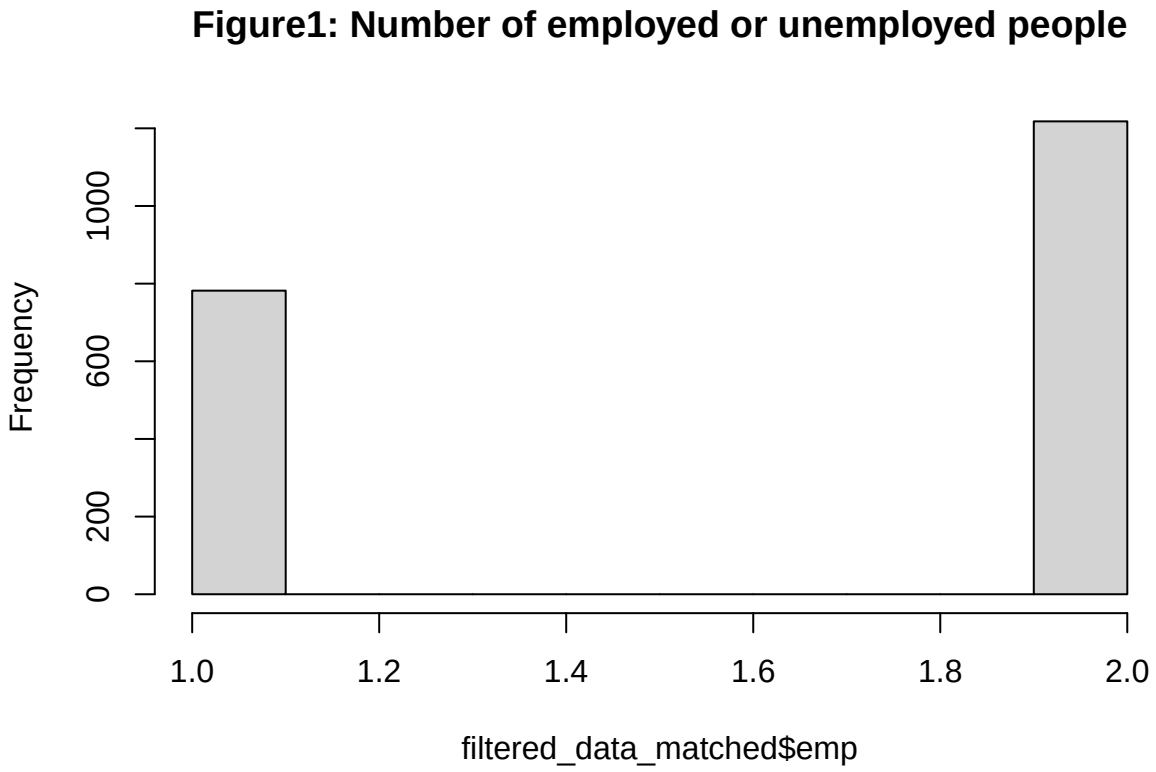


Figure 1 histogram shows that our data is significantly unbalanced. It does not follow a normal distribution. Therefore, a logistic model with a binomial family would be more appropriate for this data.

```
filtered_data_matched$emp<-as.factor(filtered_data_matched$emp)
propensity_score_regression <- glm(emp ~ home_income+age+gender+edu+state+race_ethnicity,
     family = binomial,
        data = filtered_data_matched)

huxtable::huxreg(propensity_score_regression)
```

```
## Warning in huxtable::huxreg(propensity_score_regression): Unrecognized statistics: r.squared
## Try setting ‘statistics‘ explicitly in the call to ‘huxreg()‘
```

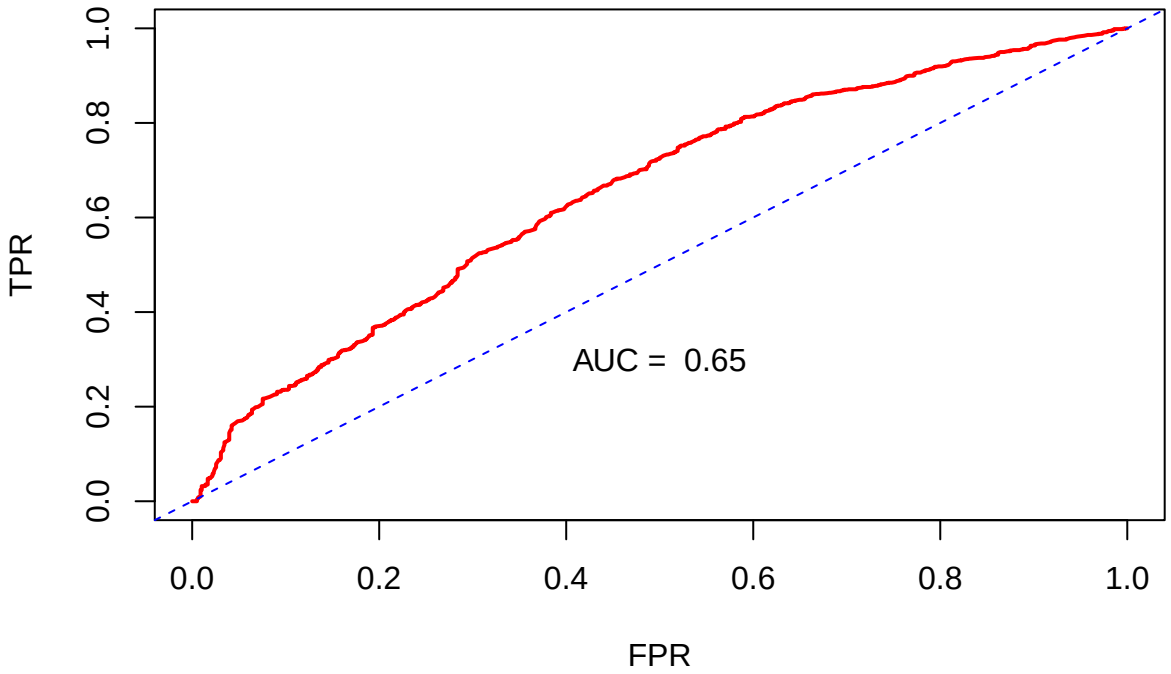
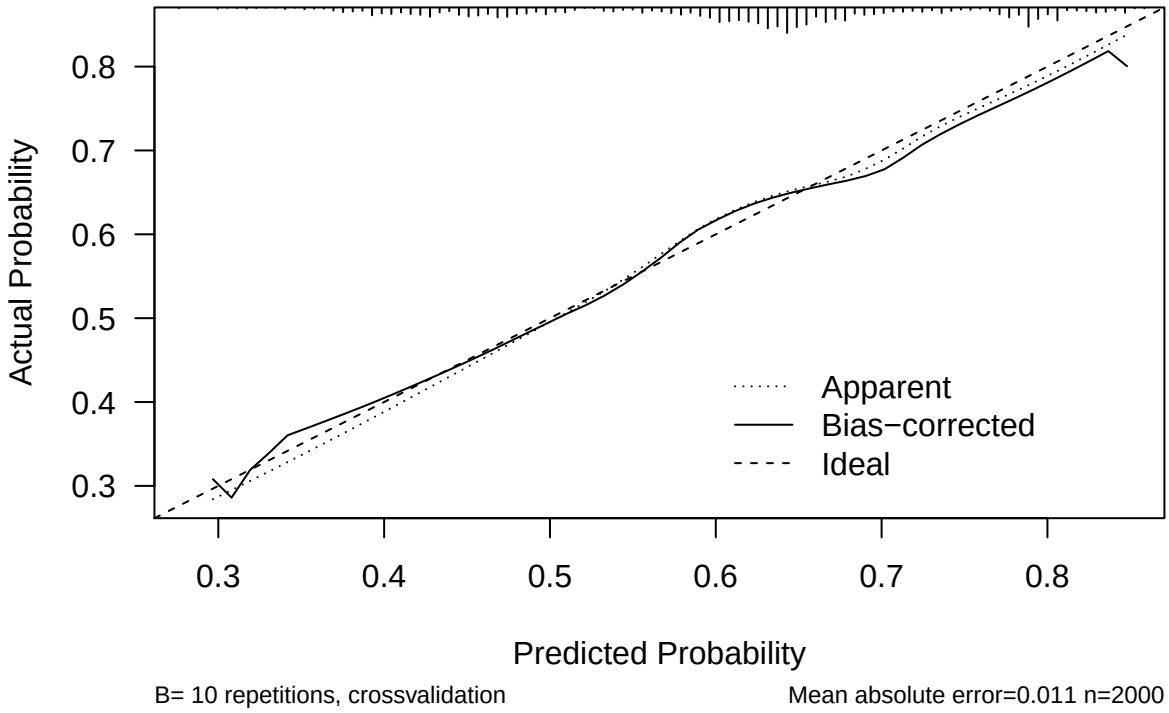


Figure 2: The area under the curve indicates that the model accurately predicts 65% of the time. Cross-validation is implemented to ensure the model fits well

```
##
## n=2000   Mean absolute error=0.011   Mean squared error=0.00017
## 0.9 Quantile of absolute error=0.019
```

Figure 3: Cross Validation



As we can see from figure 3, the data are mostly consistent on the ideal line.

Results

The Huxtable shows that the coefficient for treatment is 0.889, exhibiting a positive correlation between treatment(edu) and the response variable(employment-emp). The treatment is also one of the most significant variables among gender, age, and race_ethnicity. This means that the log odds of having employed increases when receiving a higher level of education, hence the probability of getting employed. It is confident to say that there is a causal relationship between education and employment. Receiving higher education increases one's probability of getting employed.

Besides higher education, gender and race_ethnicity also proves to be significant factors in ensuring employment. Black males seem to have high chances of getting employed. Age is less significant in determining employment compared to other factors. A negative correlation is observed between all three state areas and employment, indicating that people from all three areas have lower chances of getting employed.

	(1)
(Intercept)	-0.975 ***
	(0.293)
home_income1	0.417
	(0.517)
age	0.012 *
	(0.005)
genderMale	0.745 ***
	(0.099)
edu1	0.889 ***
	(0.100)
stateNortheast	-0.195
	(0.163)
stateSouth	-0.107
	(0.120)
stateWest	-0.035
	(0.142)
race_ethnicityBlack	0.648 *
	(0.260)
race_ethnicityOthers	0.518
	(0.345)
race_ethnicityWhite	0.338
	(0.245)
N	2000
logLik	-1265.414
AIC	2552.829

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.