Higher education, the key to greater job opportunities

Ziyu Hao

22 December 2020

Abstract

Education plays a substantial role on employment opportunities. In this paper I will analyze the effect of receiving post-secondary or higher education, as opposed to secondary or lower education, on employment status. Since students who have received higher than post-secondary level of education, as opposed to student who have not, on average have more exposure to all aspects of life, we will use propensity score matching to obtain more credible causal estimates.

To examine the effect of receiving post-secondary or higher education (with treatment) versus receiving secondary or lower education (controlled) on employment using matching, I will conduct the following steps: 1. Estimate propensity scores (The probability of being treated) 2. Apply propensity score matching to the filtered data 3. Create matches and reduce the dataset to include only matched observations 4. Estimate treatment effects with reduced dataset using glm model

Keywords

propensity score matching, 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. Thus, it is 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. However, confounding variable influences both the response and explanatory variables and causes a spurious correlation, thereby impeding the ability to draw causal inferences from observational data. Thus, we use methods to minimize the effects of confounding variables, such as propensity score matching, to examine the causal effect of education on employment.

Propensity score matching is an experimental method in which a control group is constructed consisting of pairs whose individuals have matching characteristics, but differ in treatment. The dataset can then be further reduced and the matched pairs can be used to estimate the impact of an intervention. I will implement this method to discern a causal link between level of education and employment status.

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 level of education and employment status. 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 in the previous section. The inference drawn from 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].

The random digit dialing method is riddled with ineffectual responses. The success of making generalized findings has reduced greatly due to the low response rate from the challenge of reaching more targeted survey participants. To deal with such problems Nationscape, a partnership of Democracy Fund Voter Study Group and UCLA Political Scientist, provides a dataset containing data of certain demographic criteria. 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. The 7 variables selected are employment, gender, race_ethnicity, education, state, household_income, and age. Employment and education are selected since we aim to infer causal relationship between them. All other variables are chosen for which they potentially affect employment rate and can be combined with achieved education levels to assess change in employment status.

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, the members of the unemployment variable. In categorizing our treatment into 2 categories, a variable named 'secondary_or_lower' is created to represent 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, members of the variable secondary_or_lower.

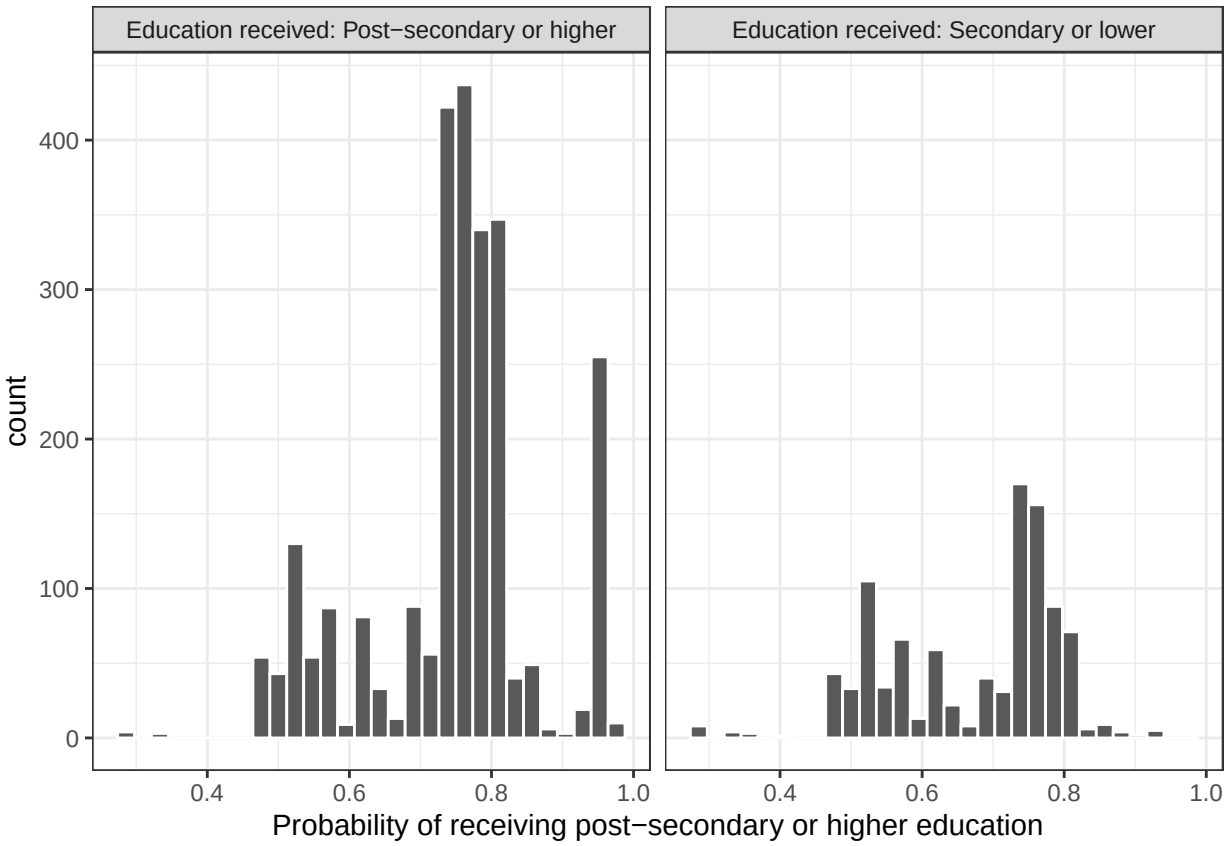
The predictors are edu, gender, race_ethnicity, state, home_income, and age. Race_ethnicity was categorized into 4 groups namely Asian, Black, White, and Others. state was categorized into 4 categories namely Northeast, Midwest, South, and West. Home_income was created based on the original variable household_income, with 1 representing high income, greater than or equal to $150,000 and 0 represents low income, which is less than $150,000.

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

```
##
## Call:
## glm(formula = edu ~ home_income + age + gender + state + race_ethnicity,
##     family = binomial, data = filtered_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3671  -1.1946   0.6773   0.7766   1.5873
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.50489    0.19055   2.650  0.00806 **
## home_income1          1.75385    0.25558   6.862 6.78e-12 ***
## ageAge_30-44          0.91412    0.09398   9.727  < 2e-16 ***
## ageAge_45-59          0.91957    0.10219   8.999  < 2e-16 ***
```

```
## genderMale            0.12035    0.07952   1.514  0.13015
## stateNortheast        0.33007    0.12210   2.703  0.00687 **
## stateSouth           -0.05915    0.10242  -0.577  0.56363
## stateWest             0.18538    0.11525   1.609  0.10771
## race_ethnicityBlack  -0.58479    0.19361  -3.020  0.00252 **
## race_ethnicityOthers -1.37185    0.29271  -4.687 2.78e-06 ***
## race_ethnicityWhite  -0.37476    0.17498  -2.142  0.03221 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4197.4  on 3565  degrees of freedom
## Residual deviance: 3904.7  on 3555  degrees of freedom
## AIC: 3926.7
##
## Number of Fisher Scoring iterations: 5
```

pr_score	education
0.741	1
0.728	1
0.94	1
0.74	1
0.613	1
0.685	1



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

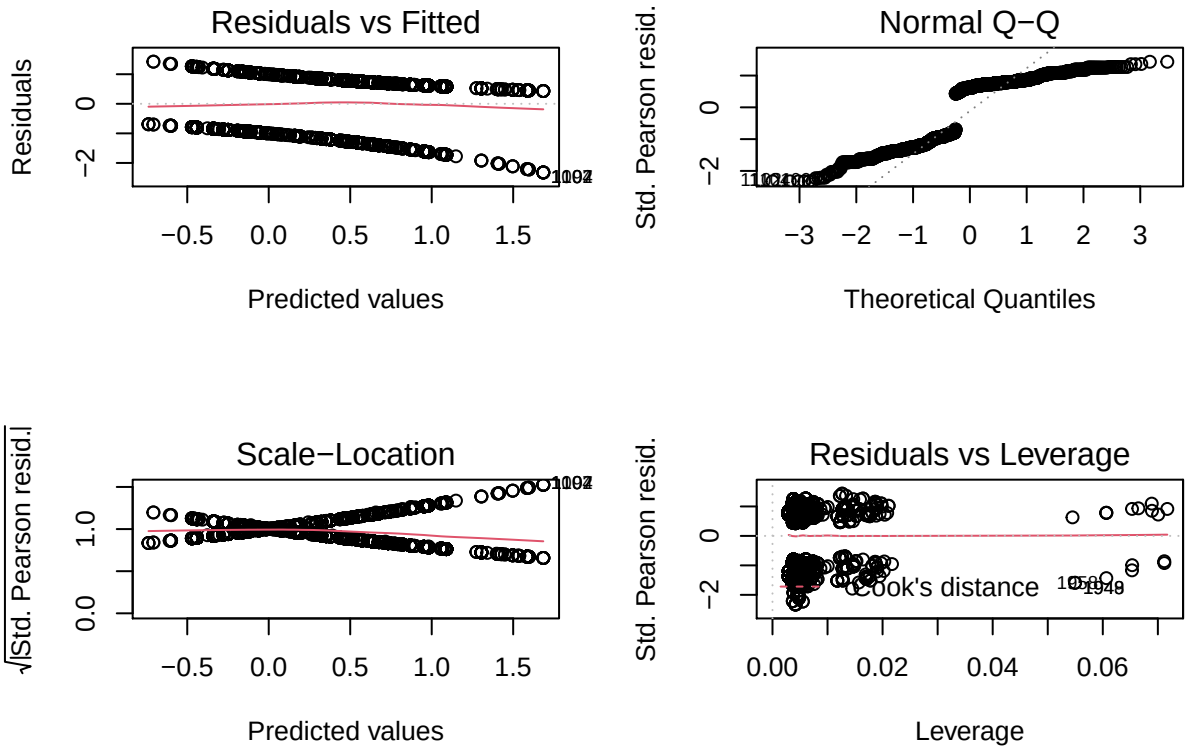
Predictors are categorical variables including home_income, age, gender, edu, state, and race_ethnicity.

Our model equation is:

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

To check if the assumptions for the generalized linear model are satisfied, we generated the following plots. Three key assumptions behind glm are: 1. Linearity 2. Response distribution, and 3. Independence

We can conclude that the linearity assumption is respected by looking at the Residuals vs Fitted plot, no trend is observed in the residuals. By looking at the QQ-plot, we can check if the standardized residuals follow a standardized normal distribution. Though we expect the points to align with the reference line, it is reasonable to have some departures. The plot for dataset respects the response distribution assumption for logistic regression. The Scale-Location plot is used for testing the assumption of equal variance(homoscedasticity). The assumptions are satisfied since the residuals shows equal distribution above and below the line. Three assumptions are checked.

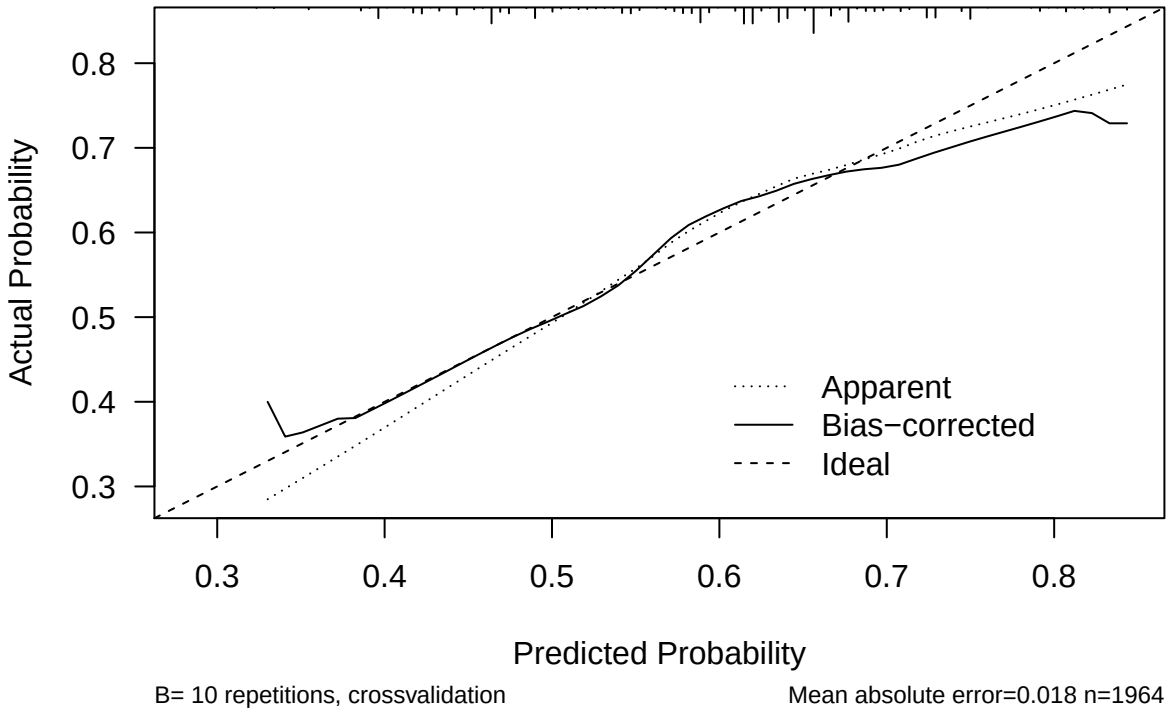


Overfitting is a condition where the fitted statistical model describes the random error in the data instead of relationship between variables. This is common when the fitted model is too complex. Generalizability can be reduced if the regression model is overfitted. Thus Cross-validation is employed in this paper. Cross-validation is used to prevent overfitting and to ensure well-fitting of the proposed model. Here we use a function called 'calibrate' from CRAN package. Calibrate is a resampling model calibration method and it allows us to use cross-validation to obtain overfitting-corrected estimates of predicted versus observed values. The function type for calibration in this case is lrm(binary/ordinal logistic model).

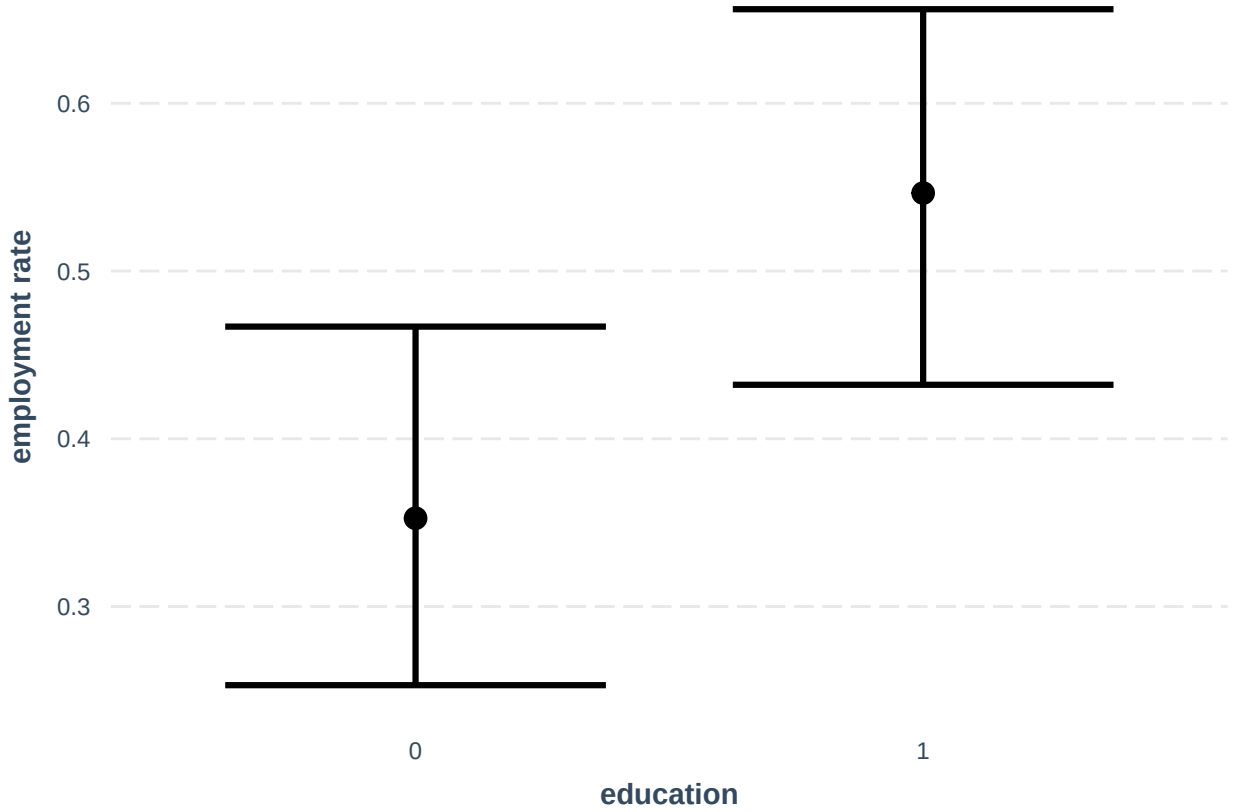
A non-parametric calibration curve is generated for logistic model. It is estimated over predicted values. The mean absolute error is 0.02, referring to the difference between the predicted values and the corresponding overfitting-corrected calibrated values. The curve shows that the data is almost consistent with the ideal line. Hence validating our proposed model.

```
##
## n=1964   Mean absolute error=0.018   Mean squared error=0.00071
## 0.9 Quantile of absolute error=0.039
```

Figure 1: Cross Validation



Results and Discussion



The model is log(p/(1-p)) = -0.607 + 0.361* home_income1 + 0.280* ageAge_30-44 + 0.192 * ageAge_45-59 + 0.629* genderMale + 0.794* edu1- 0.130* stateNortheast- 0.100stateSouth +0.008 stateWest+ 0.590* race_ethnicityBlack + 0.434 * race_ethnicityOthers +0.267*race_ethnicityWhite

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

We interpret the coefficients of dummy variables as the percentage change in y when changing from the reference variable to the corresponding dummy variable, while holding all other covariates fixed, when y is log transformed.

The coefficient of the dummy variable edu1 is 0.794, and it suggests that the log odds of getting employed increases by 79.4% when the education achieved changes from secondary or lower education(reference category-edu0) to post-secondary or higher education.

The p-value for edu1 is less than 0.001, which is smaller than the 0.05 significance level. We can conclude that there is a statistically significant association between receiving post-secondary or higher education and being employed. Change in the variable 'edu' from edu0 to edu1 (from secondary or lower education to post-secondary or higher education) is associated with changes in the mean of employment.

Other covariates also affect job opportunites and likelihood of finding a job. Older peple have more job opportunities and thus have higher chances of getting employed. People who fall into the age range of 30-44 has 28% higher chance of getting employed than those who are less than 30 years old. The employment rate increases by 19.2% for people who are 45 to 59 years old compared to people who are 30 years old or younger. This is reasonable since older people tend to have more experience in the job market and know what kind of job they want and can best make use of their skills. Gender is also believed to be one of the most important factor. Male have 62.9% higher chances of employed than female. As for location, we can see that people live in Northeast and South of the United States have less chances of getting employed compared

to those who live in Midwest. However, people who live in West have a really slight higher possibility of getting employed than those in Midwest. Asian have the lowest employment oppurtunity. The probability of employed increases by almost 60 percent for a black people compared to asian.

Given the two assumptions of propensity score matching holds, (1. conditional independence: there exists a set of x observational covariates such that when holding these covariates fixed, the outcomes are independent of treatment. and 2. Common support: for each value of x, the probability of being treated and untreated falls between 0 and 1.) we can infer that there is a causal relationship between level of education and employment status.

To visualize effect of categorical predictor, effect_plot is used . The predictor of interest is the treatment edu. We can tell that there is a clear superiority of receiving post-secondary or higher education over lower education.

In this paper, we successfully infered a causal realtionship between education and employment by implementing a method called propensity score matching. We constructed an artificial control group and matched treated and non-treated units with other covariates being similar to each other. Propensity scores are calculated and applied to the filtered dataset and reduced the data to include only observations that successfully matched. A generalized linear model was then used to estimate the effects of education on employment rate.

Conclusions

We found that receiving a high level of education, as opposed to secondary or lower education, has a positive effect on employment rate. There is a correlation between attained level of education and probability of securing employment. There are also other covariates that essentially effect our job opportunities such as age, gender, home location, and ethnicity. People with greater age tend to have higher chances of acquiring employment than those of lower ages. Males have higher likelihood of gaining employment than females, which is the second most influential factor affecting employment. People living in the west overall have higher chances of getting employmed and this is reasonable since it is where most world-renowned universities are located and college opens doors to more career opportunities. As for ethnicity, Black, White and people of other ethnicities all have higher chances of employment than those of Asian ethnicity. Holding all other covariates fixed, unemployment decreases as educational attainment increases.

As the educational level increases and more people are pursuing job positions, the workforce is supplied with more high skilled workers. This increase in supply of highly educated people in the workforce affects the requirements for jobs. Nowadays, it is impossible to get the position you want without a desirable degree. The emphasis on educational requirement has greatly increased, only with advanced studies in a related field can you become an competitive candidate.

Higher academia can give people time to improve writing, reading, communication and technical skills. These are essential in a competitive job market. Higher education offers people an opportunity to interact with a wider variety of people, improving social skills that are important in job search processes.

Weaknesses & Next Steps

Propensity score matching method has one disadvantage. PSM accounts only for observable covariates. There are hidden factors that affects assignment to treatment, and they are not accounted for in the matching process. There are certain biases due to latent variables and it can remain after matching.

The inappropriateness of the link function is also a problem in the generalized linear model. The lack of a fixed level of predictiveness of the models is another problem.

What we can do next is to conduct a survey with the leading businessess on the categories they recruit. By comparing with the categories I have used in this analysis, we can add more categories or eliminate certain confounding variables to make the analysis more precise since businesses are defined by themselves.

Appendices

You can find our codes in: (https://github.com/js25-lab/Education_employment_causal_inference_Analysis/tree/main/Causal_analysis)

References

Wickham et al. (2019) Wickham and Miller (2020) Robinson, Hayes, and Couch (2020) Gelman and Su (2020) Hugh-Jones (2020) Robin et al. (2011) Harrell Jr (2020) Long (2020)

Gelman, Andrew, and Yu-Sung Su. 2020. Arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. https://CRAN.R-project.org/package=arm.

Harrell Jr, Frank E. 2020. Rms: Regression Modeling Strategies.

Hugh-Jones, David. 2020. Huxtable: Easily Create and Style Tables for Latex, Html and Other Formats. https://hughjonesd.github.io/huxtable/.

Long, Jacob A. 2020. Jtools: Analysis and Presentation of Social Scientific Data. https://cran.r-project.org/package=jtools.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "PROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." BMC Bioinformatics 12: 77.

Robinson, David, Alex Hayes, and Simon Couch. 2020. Broom: Convert Statistical Objects into Tidy Tibbles. https://CRAN.R-project.org/package=broom.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Evan Miller. 2020. Haven: Import and Export 'Spss', 'Stata' and 'Sas' Files.

	(1)
(Intercept)	-0.607 *
	(0.242)
home_income1	0.361
	(0.511)
ageAge_30-44	0.280 *
	(0.112)
ageAge_45-59	0.192
	(0.130)
genderMale	0.629 ***
	(0.104)
edu1	0.794 ***
	(0.100)
stateNortheast	-0.130
	(0.166)
stateSouth	-0.100
	(0.120)
stateWest	0.008
	(0.149)
race_ethnicityBlack	0.590 *
	(0.235)
race_ethnicityOthers	0.434
	(0.325)
race_ethnicityWhite	0.267
	(0.220)
N	1964
logLik	-1269.416
AIC	2562.832

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