

Estimating Treatment Effects of Higher Education Level: An Evaluation Using Propensity Score Matching

Ji Won Kim - 1002685212

18/06/2021

Abstract

While randomization is at the heart of observational study, self-selection bias and ethical issues plague in many randomized controlled trials. Using the monthly Labour Force Survey (LFS) data from Statistics Canada, we use propensity score matching (PSM) and appropriate regression techniques that allows us to post-hoc analyze the causal effects of attaining education level higher than Bachelor's degree having on the hourly wage earnings. We use logistic regression and nearest neighborhood approach to construct paired observations consisting of one treatment and one non-treatment respondent, followed by multiple regression to make predictions on hourly wage with several explanatory variables. Our result shows that there is indeed a statistically significant evidence that the earnings of treatment group is higher than non-treatment groups. However, we conclude that a more sophisticated approach on variable selection is required in order to increase the precision of estimated treatment effect and minimize the bias.

Introduction

It is generally accepted that graduates with higher educational attainment have higher earnings. According to the US Census Bureau (2019, as cited in Kerr, 2021), the median earnings of full-time workers holding a bachelor's degree was \$66,536 in 2019. With a master's degree, the median earnings jumped to \$81,636, which is over 22% increase. Nevertheless, it is not clear whether to accept that this result is actually due to differences in education level itself. For example, graduates from wealthy household could possibly be a factor that affects the overall income.

Researchers are often interested in performing a randomized controlled trial (RCT) to test hypothesis on effects of a treatment. RCT is a statistical methodology used to determine whether a cause and effect relation exists between treatment and the outcome of interest. The core idea behind RCT is that individuals are randomly assigned to either the treatment group or the control group, and conclusion is drawn that any observed differences in the outcome of interest is due to the treatment itself (Kendall, 2003). However, it is often a challenging task to design a truly randomized control group study because it is inefficient and faces ethical issues in many cases. For example, it is unrealistic to assign people to each group and force the treatment group attend graduate school while non-treatment group is not permitted to attend graduate school. So how must we account with this issue?

In this paper, we use the Labour Force Survey data to examine whether there exists a causal relationship between education level and hourly earnings among Millennials (born in between 1981-1997) can be shown without the use of randomized controlled trial. We hypothesize that, on average, graduates with higher education have higher earnings. In particular, we claim that there is a statistically significant evidence that graduates with academic attainment above bachelor's degree have higher hourly wage than those who

completed in degree equal or less than bachelor’s degree. We attempt to show this by using *propensity score matching (PSM)*, a technique for constructing an “artificial” control group by matching a treated unit with non-treated unit depending on observations’ propensity score- the likelihood of each unit experiencing treatment (World Bank, n.d.). The main benefit of using propensity score matching is that it reduces confounding (variable that influences both dependent & independent variable) between treated and non-treated group. In our study, respondents with education levels higher than a bachelor’s degree are assigned to the treatment group, while those with education levels lower than a bachelor’s degree are assigned to the non-treated group.

We define education level as the highest level of certificate, diploma, or degree completed. The term “higher than bachelor’s degree” refers to those who have completed certificate, diploma, or degree above bachelor’s level; this includes anyone with completion in masters/doctoral program or equivalent level. On the other hand, educational level “equal or lower than bachelor’s degree” includes those with bachelor’s degree, certificate/diploma in -secondary, some post-secondary with no certificate/diploma, high school graduate/some high school, or 0-8 years of education. Note that we will use the term “**higher education level**” to denote **education level above bachelor’s degree** throughout this paper.

Data

Data Collection Process

Our analysis is based on mandatory Labour Force Survey (LFS) conducted by the Statistics Canada. The LFS collects monthly information on the labour market activities of Canada’s working age population to measure the current state of the Canadian labour market and is used to calculate employment and unemployment rates at national, provincial, territorial and regional level (Statistics Canada, 2018).

The LFS involves probability sampling based on a stratified two-stage design, where each province is sub-divided into geographic stratum. The first stage of sampling consists of selecting smaller geographic areas, called clusters, from within each stratum. The second stage of sampling consists of selecting dwellings from within each selected cluster. Every member of each household represents approximately 290 Canadians of same age and sex (Statistics Canada, 2018). The LFS sends out an email or letter notifying the mandatory participation of the survey. Then, telephone interviews are conducted for all designated respondents, where respondents are asked to provide their demographic information following details into their current labour status (Statistics Canada, n.d.a). The full version of the questionnaire as well as recorded variables are available in the Appendix.

Bias

The survey data we are using for our analysis was collected in May 2021. When we originally planned this study, we attempted to obtain data collected before the Covid-19 pandemic sent shock waves through Canadian job market; however, any data prior to January 2021 was unavailable for access. Selecting an appropriate time frame of the survey data is especially important in this time of pandemic for the reason of avoiding potential time period bias. For example, employment since the onset of crisis affected the most in accommodation and food services, which usually tend to employ younger, lower-educated workers (Hirsch, 2004). Since the purpose of our study does not account for impact of Covid-19 on the job market, we selected data collected in March 2021 as the difference in the unemployment rate during pre-pandemic and post-pandemic was the lowest among 5 monthly datasets available for access (7.5% in March 2021 compared to 5.4% in May 2019) (Statistics Canada, 2021). Nevertheless, our data may not be the best representative of the “normal” job market.

Moreover, students who are currently pursuing a degree are not considered as graduates and therefore not counted towards the data, which can possibly lead to a bias. For example, the highest education

level attained by a master's degree student who graduated with a bachelor's degree will be counted as a bachelor's degree rather than a master's degree, since the questionnaire asks for the "highest level of schooling **completed**". This may lead to a bias since working labour currently pursuing a degree in master's program would likely be paid higher than those of undergrad students, yet be counted as a undergraduate.

Data Summary

Our original raw data includes 86,552 responses with 60 variables, having detailed breakdowns by demographic characteristics, industry and occupation, education level, job tenure, wage, etc. Among the entire dataset, we used 15,725 responses after data cleansing, of which 2,964 responses were actually matched after performing propensity score matching.

Raw data often does not come in the form where we can directly use it for analysis. There may be missing values, inconsistent responses, structural errors, and many other ways that can potentially impact the modeling process or result in an unexpected way. Therefore, it is important to appropriately clean the data that meets the requirements in the methodology.

In our LFS data, we filtered out anyone who is not currently in the labour force, meaning that anyone who is not currently employed was excluded from our sample. Seven variables were included in the propensity score matching: age group, sex, education level, hourly wage, province, student status, type of employment, and the respondent id number. The values of hourly wage was divided by 100 since the original data didn't contain the two decimal points. Next, we changed the every categorical variables' response values from numbers to factors and appropriately renamed the responses that matches the description of each categories. Every respondents who were in the age group between 15 to 19 and those over 39 were omitted (and deleted all factor levels in this age group), leaving out age group between 20 to 39. Then, we replaced every missing values in student status to non-student. Lastly, we excluded any responses that contains at least one missing response value.

The variables we selected are used are to appropriately estimate the propensity (or likelihood) of a respondent being in the treatment group or not, i.e. propensity of respondent having education degree above bachelor's degree or not. The following is a description of important variables used for our propensity score matching analysis:

- **Age group**

Age of the respondent is sub-divided to five-year age groups. In the original data, age is collected for every household member in the survey, and the information on labour market activity is collected for all persons aged 15 and over. However, we excluded anyone who is in the age group of 15-19 and those over 40, since our goal is to show if there is any statistical significance of attaining education level higher than bachelor's degree among the Millennials (born in between 1981-1996).

- **Education level**

Education level is a binary outcome categorical variable determining the highest level of schooling completed. Anyone who have completed degree above bachelor's degree are considered "treated", and therefore assigned to value 1. Note that treated respondents are those who have attained a degree or certificate equivalent or higher than master's degree. Those who have not obtained a degree above bachelor's degree are considered "untreated", and are assigned to value 0. Thus, education level is the core factor that identifies whether the person was treated or not.

- **Hourly wage**

Hourly wage is a numeric variable that indicates the usual hourly wage of the individual, including tips, commission and bonuses (before taxes and other deductions). Wage will be our outcome variable that determines whether respondents who have higher educational degree (treated group) are more likely to have higher earnings.

- **Type of employment**

Full-time or part-time work schedule. Full-time employment consists of persons who usually work 30 hours or more per week at their main or only job. Part-time employment consists of persons who usually work less than 3 hours per week at their main or only job. The type of employment is usually highly correlated to the overall wage as well as the educational attainment. For example, the wages of part-time workers are considerably lower than those of full-time workers (Hirsch, 2005).

- **Sex**

Sex is a binary outcome categorical variable referring to the biological sex assigned at birth. Sex plays an important role in determining the wage, since it is common that different sex (and gender) experience wage gaps.

- **Province**

Province is the geographical area in which the respondent resides is sub-divided into 10 provinces in Canada. Geographical is likely to impact the respondent's educational level as well as the wage, because it is observed that regional students were under-represented in higher education when compared to peers from urban areas (Baglin et. al., 2017).

- **Student status**

Student status is a categorical variable that refers to the differentiation of student, divided into full-time student, part-time student, and non-student. Note that anyone who have graduated or currently in leave of absence are categorized as non-student. Student status affect the overall wages since student workers are highly likely to be working part-time, and therefore paid substantially lower wages and benefits than do full-time jobs.

Characteristic	0, N = 14,243	1, N = 1,482
HOURLY WAGE	24.00 (17.75, 33.65)	35.10 (24.68, 46.23)
SEX		
Male	7,304 (51%)	595 (40%)
Female	6,939 (49%)	887 (60%)
AGE GROUP		
20 to 24	2,987 (21%)	29 (2.0%)
25 to 29	3,325 (23%)	280 (19%)
30 to 34	3,822 (27%)	539 (36%)
35 to 39	4,109 (29%)	634 (43%)
STUDENT STATUS		
Non-student	12,649 (89%)	1,364 (92%)
Full-time student	1,087 (7.6%)	71 (4.8%)
Part-time student	507 (3.6%)	47 (3.2%)
EMPLOYMENT TYPE		
Full-time	11,823 (83%)	1,315 (89%)
Part-time	2,420 (17%)	167 (11%)
PROVINCE		
NL	377 (2.6%)	36 (2.4%)
PE	409 (2.9%)	43 (2.9%)
NS	670 (4.7%)	81 (5.5%)
NB	666 (4.7%)	45 (3.0%)
QC	2,548 (18%)	277 (19%)
ON	3,602 (25%)	481 (32%)
MB	1,681 (12%)	142 (9.6%)
SK	1,084 (7.6%)	82 (5.5%)
AB	1,658 (12%)	127 (8.6%)
BC	1,548 (11%)	168 (11%)

Table 1. Summary statistics of variables grouped by education level (0: treated, 1: non-treated). The brackets within hourly wage shows the median IQR; for every categorical variables, the number of observations and its relative proportion within the group is shown.

The summary statistics for our LFS dataset are shown in Table 1, grouped by education level. Notice that every explanatory variables are categorical variables. The mean hourly wage of the entire sample is 27.75 with standard deviation of 12.6. The median (mean) hourly wage of treatment group is \$35.1 (\$36.44), and \$24.0 (\$26.85) for non-treatment group. This implies that there is some evidence that there exists some difference of hourly wage in the treatment group and non-treatment group.

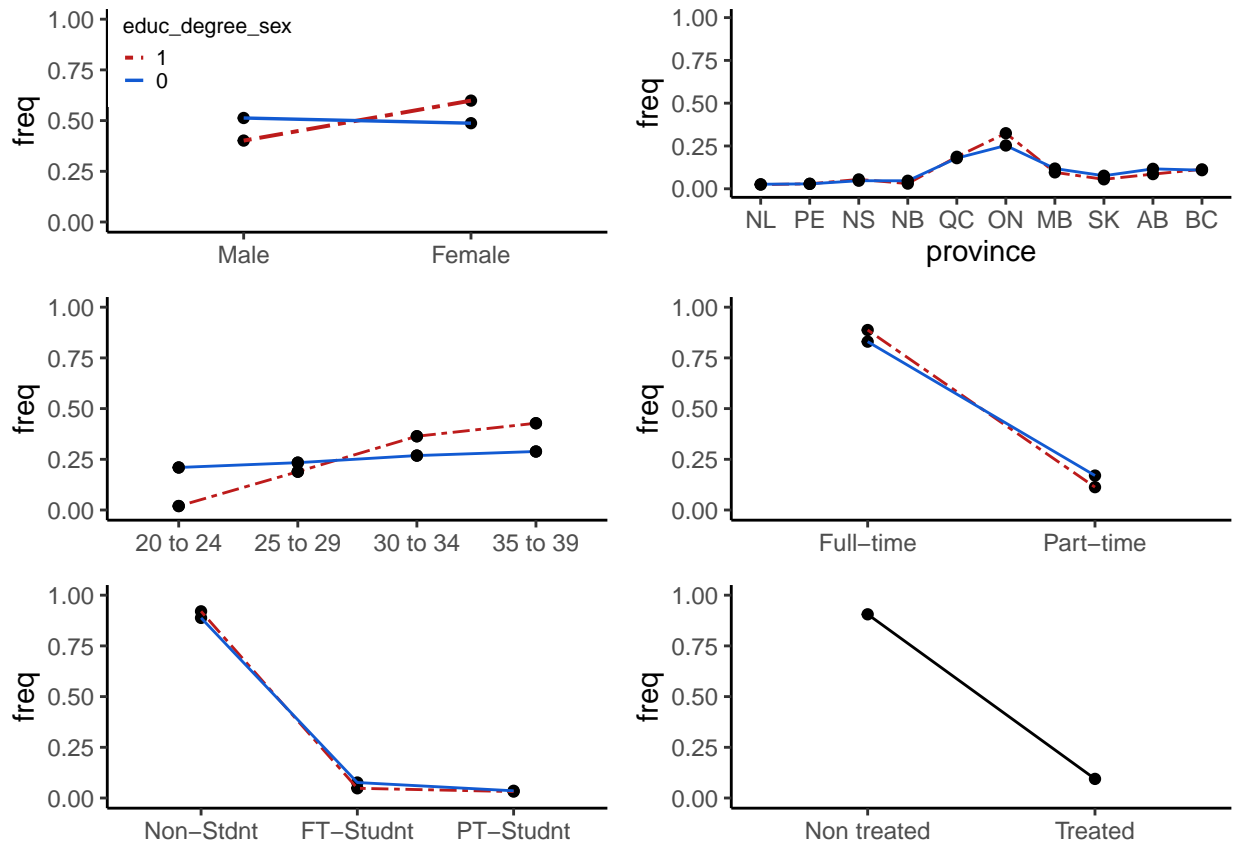


Figure 1. Comparison of ratios among various demographic subgroups between treated group (red dotted) and non-treated group (blue).

The figure above shows the proportion of each categories in the treatment group and non-treatment group, as well as a plot showing the proportion of treated and non-treated group in the whole sample. The proportion of treated and non-treated group is 9% and 9%, respectively. We can see that the distribution of each categories among the two groups are pretty similar. For sex, treatment group is more female dominant, consisting of 59.8%. In terms of geographical area, Ontario has the highest proportion of respondents currently residing in both groups, with Newfoundland and Labrador being the least concentrated area (2.4%~2.6%). As one might expect, the proportion of non-student and full-time workers are highly dominant.

All analysis for this report was programmed using R version 4.1.0.

Methods

I. Model

In order to estimate the effect of attaining higher education level on the wage, we make use of the **propensity score matching (PSM)** with use of linear regression and nearest neighborhood matching. Propensity score matching is a method for addressing selection bias and moving towards more causal estimates (Rosenbaum & Rubin, 1983), and has lately has been resurfaced due to rising popularity of artificial intelligence (Caetano, 2020). The reason for choosing PSM as our model is because it reduces greater portion of bias when precisely estimating the effect of a treatment (Carolyn et. al, 2010); it is realistically impossible (and unethical) to obtain a randomized control group by randomly assigning observations to each groups and determine whether or not they should receive higher education levels. Rather, the core idea behind PSM is motivated by forming a matched set of treated and untreated subjects sharing similar propensity score, analogous to forming a post-hoc randomized control group of study. Propensity score refers to the probability or likelihood of an observation being in the control group. This allows to examine whether there is any statistically significant effect of a treatment for example, if individual A and B share perfectly identical demographic characteristics except the treatment (education level) itself, then we can say that any difference between the outcome variable of interest (difference in hourly wage) can be explained by the treatment.

II. Methodology

1. Estimating Propensity Score

Based on given demographic information of the respondents, we use logistic linear regression model to estimate the propensity score, analogous to the probability of each respondent being treated. We use logistic model in our case since the response variable is education level, which has binary outcome- treated or non-treated. On the other hand, sex, age, province, employment type, and student status are the independent variables that will explain the relationship with the response variable. Using p to indicate the propensity score estimate, p is described by

$$\text{logit}(p) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_{i[j]}^{AGE}x + \hat{\beta}_{i[j]}^{SEX}x + \hat{\beta}_{i[j]}^{PROV}x + \hat{\beta}_{i[j]}^{STDNT}x + \hat{\beta}_{i[j]}^{EMPLOY}x$$

Where $\hat{\beta}_0$ is the fixed intercept and the terms $\hat{\beta}_{i[j]}^{AGE}$, $\hat{\beta}_{i[j]}^{SEX}$, $\hat{\beta}_{i[j]}^{PROV}$, $\hat{\beta}_{i[j]}^{STDNT}$, $\hat{\beta}_{i[j]}^{EMPLOY}$ corresponds to the coefficient associated with each categorical variable. Since each categories have multiple response values, the subscript $i[j]$ is added in the coefficients. For example, $\hat{\beta}_{i[j]}^{SEX}$ takes in values of $\hat{\beta}_{male}^{SEX}$ and $\hat{\beta}_{female}^{SEX}$. Note that the logit function $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$ in the equation above is also called the log odds. appropriate estimate of propensity score is calculated by solving for p in the left hand side of the equation. We then augment the results into our original dataset so that every observation is assigned to its relative propensity score.

2. Nearest Neighborhood Matching

Once the propensity score is estimated, we use our forecast to generate paired matches based on the propensity score of each observation by using single nearest neighborhood matching. This means that we match each respondent who has education level above bachelor's degree with another individual who do not have education levels above bachelor's degree but have a similar probability (i.e. a similar propensity score) of obtaining a bachelor's degree. In our case, the respondent with "similar probability/propensity score" is identified by locating another respondent who has the closest propensity score, which is the core function of nearest neighborhood matching. We reduce our dataset so that it only contains matched pairs of observations. The number of total matched pairs is equal to the total number of respondents in the treatment group. In case of our study, there were total of 1482 treated observations, thus we have 1482 pairs of matched dataset (total of $2 * 1482 = 2964$ observations).

3. Estimating Outcome Variable

The final step of our PSM method is to use multiple linear regression to estimate the response variable (hourly wage) based on all explanatory variables including the education level (age, sex, province, student status,

type of employment AND education level). This gives interpretation on the linear relationship between each explanatory variable and the hourly wage.

III. Underlying Assumptions

One of the key assumption of PSM is that untreated units can be compared to treated units based on some observable traits, as if the treatment had been fully randomized (World Bank, n.d.). In addition, it requires a conditional independence between the potential outcome (hourly wage) and the treatment status (Baum, 2013). Lastly, every observations have a positive probability of being both treated and untreated. In this way, PSM aims to mimic randomization in order to counter the selection bias issues that plague non-experimental methods (World Bank, n.d. a).

When estimating the hourly wage in step 3 of PSM, using multiple linear regression model entails several underlying assumptions (Statistics Solutions, n.d. a). First, the model assumes linearity: there is a linear relationship between the outcome variable (hourly wage) and all the explanatory variables. That is, the expected value of hourly wage is a straight-line function of every independent variable. Second, residuals are normally distributed (multivariate normality). Third, it assumes independence of individual observations. Fourth, variance of error terms are the same across every independent variables (homoscedasticity). Finally, the model assumes independent variables are not highly correlated to each other (no multicollinearity).

Note that logistic regression model do not make some of the key assumptions of linear regression (Statistics Solutions, n.d. b). In particular, logistic regression does not require assumption of normality of residuals and homoscedasticity. Rather, it assumes linearity of independent variables and the logit which is $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$, and that outcome of dependent variable to be binary.

Results

We first discuss the result of our logistic regression. **Figure 2** shows the regression output for prediction of education level with respect to age, sex, province, type of employment, and student status. Considering the coefficients with highest significance level, the estimation tells us that that females who are full-time students and living in Ontario are the most likely to be in the treatment group, i.e. receive higher level of education. Moreover, it is evident that there is a very strong, positive linear relationship with the treatment as the age group gets older- this is as predicted since many young students currently in the masters/doctoral program are counted as non-treatment group (since the survey question asks for the highest educational level completed).

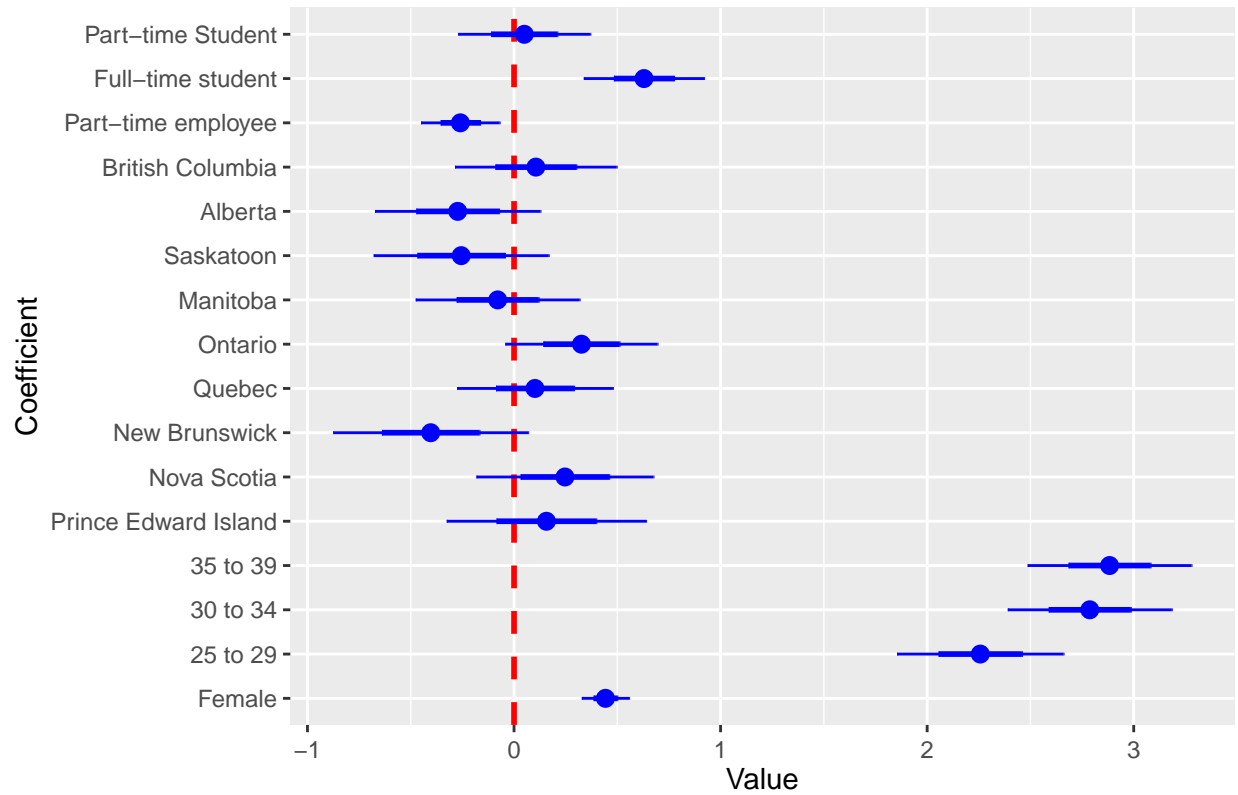


Figure 2. Coefficient plot showing logistic regression results with its respective 95% (thick), 99% (thin) confidence interval

Using the fitted values from the logistic regression output, we assigned each respondent by its respective propensity score, that is the likelihood of them being in the treatment group. We then used nearest neighborhood matching to create a new dataset that has pairs of observations that consists of treated and non-treated respondents. The dataset contains 2964 observations, of which consists of 1482 matched pairs, which is quite a serious data reduction compared to 15725 observations in our cleaned dataset.

Table 2: Fitting linear model: $\text{hrly_wage} \sim \text{sex} + \text{age_group} + \text{province} + \text{type_employ} + \text{student_status} + \text{educ_degree}$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20.52	2.355	8.715	4.779e-18
sexFemale	-1.611	0.4896	-3.29	0.001012
age_group25 to 29	5.383	1.789	3.008	0.00265
age_group30 to 34	9.783	1.754	5.579	2.645e-08
age_group35 to 39	13.33	1.749	7.623	3.328e-14
provincePE	-3.573	2.08	-1.718	0.08599
provinceNS	-2.047	1.837	-1.114	0.2652
provinceNB	-5.635	2.054	-2.743	0.006122
provinceQC	-0.5645	1.629	-0.3464	0.7291
provinceON	1.614	1.592	1.014	0.3107
provinceMB	-2.907	1.713	-1.697	0.08981
provinceSK	-1.964	1.845	-1.064	0.2872
provinceAB	3.322	1.738	1.911	0.05605
provinceBC	-0.05427	1.69	-0.03211	0.9744
type_employPart-time	-4.83	0.8027	-6.016	2.003e-09
student_statusFull-time	-3.353	1.206	-2.78	0.005476
student				
student_statusPart-time	0.4119	1.346	0.3061	0.7595
student				
educ_degreee1	7.372	0.474	15.55	1.78e-52

Table 2: Multiple regression analysis of age, sex, province, student status, type of employment, and education level as predictors of current hourly wage

With our matched dataset, we performed multiple linear regression to estimate the response variable (hourly wage) based on all explanatory variables including the education level (**Table 3**). This gives interpretation on the linear relationship between each explanatory variable and the hourly wage, as well as its respective standard error, test statistics, and p-value for two-sided hypothesis test.

As expected, the result clearly shows that education level has coefficient of 7.372 with p-value of $1.78e^{-52}$ making it the “strongest” predictor that has positive relationship with the hourly wage. The test result is highly significant since it implies that the chance of getting such result when the null hypothesis is true ($H_0 : \beta_{treated}^{EDUC} = 0$) is $1.78e^{-52}$, which is significantly less than 0.05. In terms of sex, females are less likely to have higher earnings. Similar to the results of the logistic regression, all age groups had positive linear relationship with hourly earnings, where older age groups are the most likely to be paid the highest. This is quite trivial as people in the older age group are likely to have more job experience, and therefore have higher wages.

Along 10 provinces across Canadian Territory, Ontario and Alberta had were the only provinces with positive linear relationship with wage- with New Brunswick being the lowest and considerably significant. Looking at the overall result, we believe that our model predictions are quite reasonable; however, it is difficult to assess their accuracy as our model relies on many underlying assumptions, such as linearity and low correlation of independent variables (World Bank, n.d.).

Table 3: One Sample t-test: `difference2`

Test statistic	df	P value	Alternative hypothesis	mean of x
15.63	1481	2.095e-51 * * *	greater	7.396

Finally, we performed a one-tailed t-test on to test our hypothesis by taking the hourly wage differences between the treatment group and non-treatment group in our matched data. Our null hypothesis H_0 was that there is no difference in the mean hourly wage, while the alternative hypothesis H_a claimed the treatment group have higher hourly earnings than non-treatment group. The test result shows that the p-value of the t-test is $2.0952932 \times 10^{-51}$, which means that given the null hypothesis is true, there is a $(2.0952932 \times 10^{-49})\%$ chance we get a result such that the mean hourly wage of treatment group is higher than the treatment. Therefore, we can reject the null hypothesis and accept the alternative hypothesis to conclude that hourly earnings of respondents with higher education is greater than those of who didn't receive higher education. In addition, the point estimate (mean) is 7.4, which we can interpret that the mean difference of hourly wage between the treatment group and non-treatment group is \$7.4. More generally, we are 95% confident that the average difference of hourly wage between the treatment group and non-treatment group is greater than \$6.61.

Conclusion

In this study, we conducted propensity score matching and linear regression to construct an artificial control group to prove our hypothesis on the existence of positive causal relationship between education level and hourly wage among Millennials. Propensity score matching is a method for addressing self-selection bias and moving towards more causal estimate. We used Statistics Canada’s Labour Force Survey data to calculate the respective propensity score for each observation using logistic linear regression, followed by matching treated individuals with non-treated individuals using the nearest neighborhood matching. Our findings showed that there is indeed a statistically significant evidence that individuals with academic attainment above bachelor’s degree have higher hourly earnings, with mean difference of \$7.4.

In a bigger picture, we can draw several conclusions based on the results of our study. We demonstrated how propensity score matching allows us to perform a post-hoc analysis on causal inferences without having to rely on a randomized control study. Using observed data, this study showed that we can effectively add a randomization component on the obtained sample without having to confront ethical issues that is inherent in many randomized trials. Moreover, this study showed how multiple stages of linear regression can be employed to determine the average treatment effect on hourly earnings. Finally, we conclude that variable selection could be an important factor that can increase the overall robustness of the propensity score model.

Weaknesses

The main weakness inherent in this study is that our particular model lies on many number of assumptions (mentioned in the methods section) that could make the overall result less “accurate”. Because of time constraints and limited knowledge in sophisticated methodologies that could possibly make our research more robust, our model does not go into depth, but rather gives an idea of how propensity score matching can be performed. We haven’t checked for the quality of matching nor a independence test for selected variables. Moreover, categorical variables dominate our data, which makes our study less quantitative. Lastly, our selection of data is not the best at this time of the moment since global job markets are still suffering to fully recover due to Covid-pandemic, which could possibly lead to a time period bias.

Next Steps

This paper could be further expanded to the following areas: first, it is highly suggested to perform Chi-square test or equivalent prior to selection of explanatory variables. According to Avorn et. al (2006), choices of variables included in a propensity score model can affect the bias, variance and mean-squared error of the overall estimated treatment effect. In particular, they suggest that including variables that are unrelated to the treatment but related to the outcome variable increases the precision of estimated treated effect without increasing the bias. Second, using difference in differences model could have been a better approach in our case, as the main interest of our study is to seek causal relationship between education level and hourly wages. Third, evaluating the quality of matching (percent bias reduction (PBR), histogram comparison, etc.) is highly desirable in order to check the distribution/spread is roughly normal than before matching. Perhaps performing ANOVA test would be helpful to analyze the differences among mean of treatment and non-treatment group.

Discussion

Propensity score matching offers a solution for reducing bias inherent in many observations studies and provide an effective way to make causal inferences by adding randomization component in post-hoc analysis. While propensity score seems to have potentials to be used in various study fields, it would be desirable to check and compare different models within PSM to evaluate which would be the most optimal choice in order for the result to be robust.

Bibliography

1. Rosenbaum, P., & Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>
2. KDnuggets. (n.d.). *Propensity Score Matching in R*. KDnuggets. Retrieved from <https://www.kdnuggets.com/2018/01/propensity-score-matching-r.html/2>
3. Avorn, J., Brookhart, M., Glynn, R., Rothman, K., Schneeweiss, S., & Stürmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163(12), 1149–1156. <https://doi.org/10.1093/aje/kwj149>
4. Bialik, K., & Fry, R. (2019 Jan 30). *Millennial life: How young adulthood today compares with prior generations*. Paw Research Center. Retrieved from <<https://www.pewresearch.org/social-trends/2019/02/14/millennial-life-how-young-adulthood-today-compares-with-prior-generations-2/>>
5. Baum, C. (2013). *Propensity Score Matching Regression Discontinuity Limited Dependent Variables*. [PowerPoint slides]. Retrieved from <<http://fmwww.bc.edu/EC-C/S2013/823/EC823.S2013.nn12.slides.pdf>>
6. Caetano, S. (2020 Nov 22). Propensity Score Matching - Theory. [Video file]. Microsoft Stream. Retrieved from <<https://web.microsoftstream.com/video/53989ba9-82ec-4fcb-96bf-56d2e3cc3f9c>>
7. Carolyn, H., Maffioli, A., & Vázquez, G. (2010). A Primer for Applying Propensity-Score Matching. *ResearchGate*. Retrieved from <<https://doi.org/10.1016/j.crvasa.2013.04.001>>
8. Cooper, G., Baglin, J., & Strathdee, R. (2017). *Access to higher education: Does distance impact students' intentions to attend university?* National Centre for Student Equity in Higher Education, Curtin University: Perth. Retrieved from <https://www.ncsehe.edu.au/wp-content/uploads/2017/02/RMIT_Cooper-Baglin-Strathdee.pdf>
9. Hirsch, B. (2004). Why do Part-Time Workers Earn Less? The Role of Worker and Job Skills. *IZA DP No.1261*. Retrieved from <<https://www.iza.org/publications/dp/1261/why-do-part-time-workers-earn-less-the-role-of-worker-and-job-skills>>
10. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. <<https://CRAN.R-project.org/package=stargazer>>
11. Kendall J. M. (2003). Designing a research project: randomised controlled trials and their principles. *Emergency medicine journal : EMJ*, 20(2), 164–168. <https://doi.org/10.1136/emj.20.2.164>
12. Kerr, E. (2021 March 30). *Is Graduate School Worth the Cost?*. US News. Retrieved from <<https://www.usnews.com/education/best-graduate-schools/paying/articles/is-graduate-school-worth-the-cost>>
13. Statistics Canada. (2021 May 07). *Labour Force Survey, April 2020*. <<https://www150.statcan.gc.ca/n1/daily-quotidien/210507/dq210507a-eng.htm>>
14. Statistics Canada. (2021). Labour Force Survey: Public Use Microdata File (pub0321) [CSV File]. Retrieved from <<https://doi.org/10.25318/71m0001x-eng>>
15. Statistics Canada. (2018). Tell us how Canada works: The Labour Force Survey. Statistics Canada. Retrieved from <<https://www.statcan.gc.ca/sites/default/files/lfs-brochure-epa-eng.pdf>>
16. Statistics Canada. (n.d.c). *Labour Force Survey (LSF)*. Statistics Canada. Retrieved from <https://www.statcan.gc.ca/eng/survey/household/3701>

17. Statistics Solutions. (n.d. a). *Assumptions of Multiple Linear Regression*. Complete Dissertation by Statistics Solutions. Retrieved from <<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/>>
18. Statistics Solutions. (n.d. b). *Assumptions of Logistic Regression*. Complete Dissertation by Statistics Solutions. Retrieved from <<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/>>
19. World Bank. (n.d.). *Propensity Score Matching*. World Bank. Retrieved from <https://dimewiki.worldbank.org/Propensity_Score_Matching>

Packages

1. Bates, D., Bolder, B., Maechler, M., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01./
2. Curry, M., Daniel, S., Hannum, M., Larmarange, J., Whiting K., & Zabor, E. (2021). gtsummary: Presentation-Ready Data Summary and Analytic Result Tables. R package version 1.4.1. <https://CRAN.R-project.org/package=gtsummary>
3. Venables, W. N., & Ripley, D. (2002). Modern Applied Statistics with S. Fourth Ed. *Springer*, New York. ISBN 0-387-95457-0
4. Gelman, A., & Su, Y. (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. <https://CRAN.R-project.org/package=arm>
5. Lander, J. (2021). coefplot: Plots Coefficients from Fitted Models. R package version 1.2.7. <https://CRAN.R-project.org/package=coefplot>
6. Daróczy, G., & Tsegelskyi, R. (2021). pander: An R ‘Pandoc’ Writer. R package version 0.6.4. <https://CRAN.R-project.org/package=pander>
7. Wickham et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>

Appendix

Section 1: Supplementary Data

1. Full version of Labour Force Survey questionnaire. (Note that there are over 100 questions in the original questionnaire)
[https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=1277826&UL=1V&]
2. Explanation of variables included in the original LFS data
[<https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvVariableList&Id=1313631>]
3. Glimpse of data

id	sex	age_group	province	type_employ	student_status	educ_degree	hourly_wage	fitted	treated	match.ind	ints	pairs
717	Male	20 to 24	NB	Part-time	Non-student	0	12.10	0.0032821	0	0	0	NA
3487	Male	20 to 24	NB	Part-time	Non-student	0	33.33	0.0032821	0	0	0	NA
17590	Male	20 to 24	NB	Part-time	Non-student	0	33.33	0.0032821	0	0	0	NA
31751	Male	20 to 24	NB	Part-time	Non-student	0	11.80	0.0032821	0	0	0	NA
40086	Male	20 to 24	NB	Part-time	Non-student	0	11.75	0.0032821	0	0	0	NA
48884	Male	20 to 24	NB	Part-time	Non-student	0	14.40	0.0032821	0	0	0	NA

Section 2: Supplementary Plots

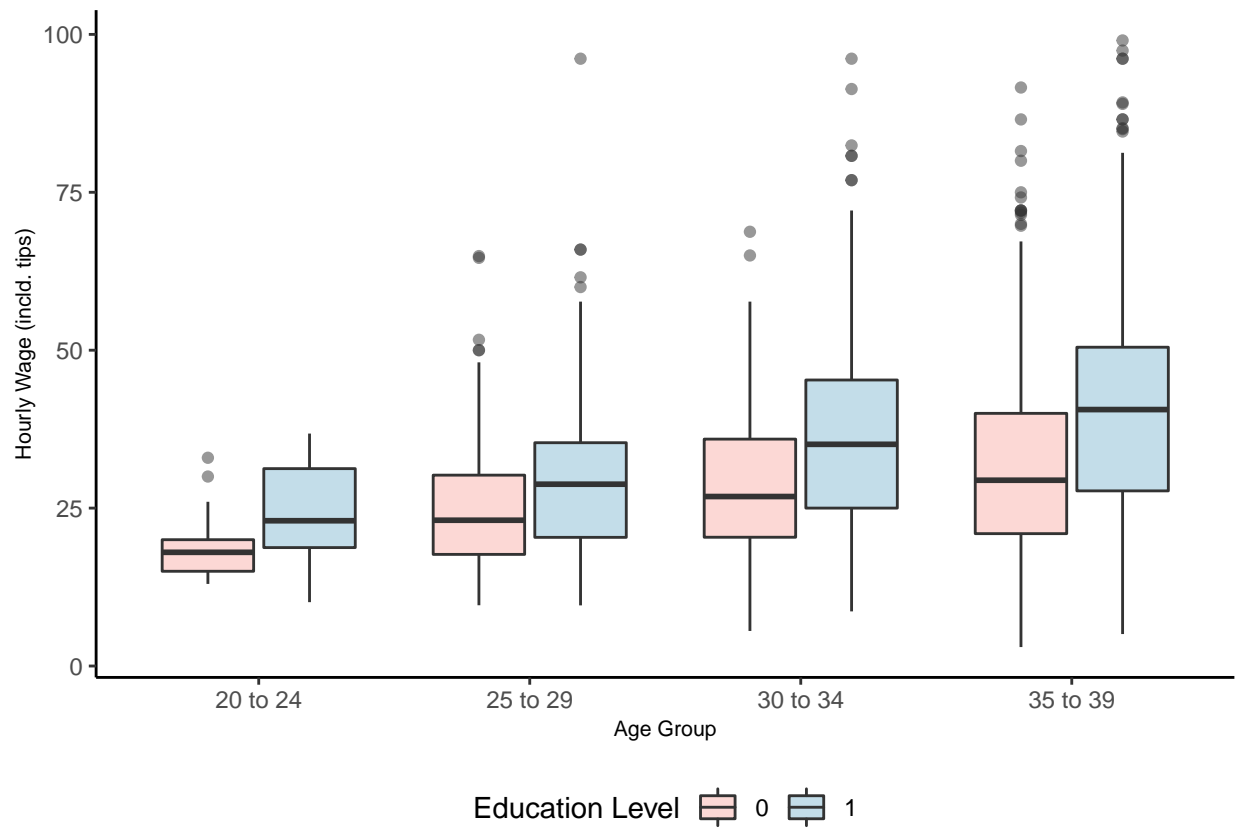


Figure 3. Comparison of hourly wages by education level and age group using matched dataset.

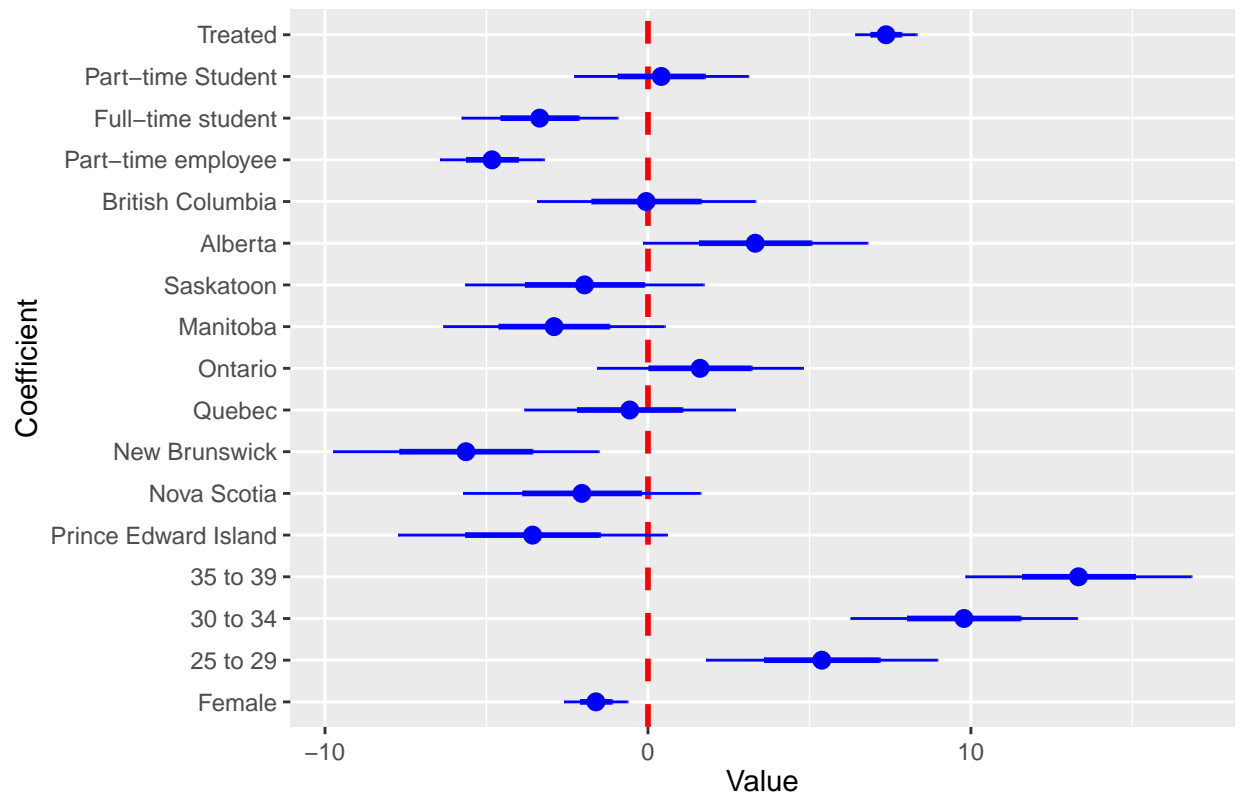


Figure 4. Plot showing the result coefficients of our regression output for matched set of data with 95% confidence interval. It predicts the overall hourly wage based on all explanatory variables including the education level of respondents.