How Education Attainment Affects Average Annual Salary Income?

Ziqing Lyu(1005022093)

12/22/2020

Abstract

Salary income has always been a topic that people have never stopped caring about, whether education level will affect salary income is also worthy of our attention. In this report, we will study whether different levels of education will affect average annual wage income. We selected respondents aged 18 to 65 in the original database, and divided the education into two levels, one is a high school degree and the other is a college degree. Then we used the propensity score matching method to sort out the data, treatment is education level and outcome is salary income. This method of organizing data can better study the causal inference, and also better let us know the causal link between education and salary income. After sorting out the data, we created a multiple linear regression(MLR) and used the selected variables to predict salary income. Finally, we can know what factors affect income through this MLR.

Keywords

Education, Wage and Salary Income, Causal Inference, Propensity Score Matching, Observational Study, multiple linear regression(MLR)

Introduction

Nowadays, job competitions have become more intense, and the pressure for one to get a job is increasing. We believe that most people are often concerned about income wage related issues, and we have found that there could be multiple factors that influence personal income, such as age, gender, region of living, and education level etc. This report is a study of factors related to income wage level, which is an interesting topic. Main purpose of it is to study the relationship between an individual's education level and personal wage/salary income. For this research, we referenced data from USA IPUMS website (https://usa.ipums.org/usa/index.shtml), which contains survey results of social issues along with human being characteristics (like gender, age, family condition, etc). For this project, 2019 data was chosen for analysis.

The chosen dataset is of observational data type. In order to better study the relationship between education level and income, we make causal inference based on the dataset in our

study. There are multiple ways to do causal inference: one popular way is the propensity score matching method (Rohan Alexander, 2020). Propensity score matching is a quasi-experimental method in which the researcher uses statistical techniques to match treated and controlled observations on the estimated probability of being treated (Propensity Score Matching-DIME Wiki). It has been widely used for data organization and treatment analysis, and has been popular in recent years. In this report, this method is used to analyze if there is a causal relationship between education and income.

In the upcoming sections of the report, methodology of the entire analysis will be illustrated, including how raw dataset is being cleaned, how propensity score matching is applied to generate new dataset, and what the final model looks like and how it behaves, shown in the Methodology section. Furthermore, results and observations from the propensity score analysis based on the final model will be shown in the Result section. Summary, conclusion, weakness of analysis flow, and areas for improvements are shown in the Discussion section. Finally, references for the entire analysis flow are stated in the last section.

Methodology

Data

In the original data, I chose region, sex, age, current marital status, and the respondent's educational attainment as variables. The outcome variable is the total pre-tax wages and salary income - that is, money received as an employee for the previous year. Data contains a total of 50,660 observations. After the selection is completed, since the age is between 1 and 97 years old, we need to delete people under 18 and older 65. Since we assume people do not work under 18 and older 65, research on education and income in that range does not seem to be significant. So in this study, everyone is at least 18 years old and under 65, now the data becomes 30512 observations. In order to use propensity score matching, treatment must be binary, so now I will create a new column named "education", the value inside represents the overall level of education divided into two levels, these education attainment as measured by the highest year of school or degree completed. For "nursery school to grade 4", "grade 5, 6, 7, or 8", "grade 9", "grade 10", "grade 11", and "grade 12" considered as high school degrees, and "1 year of college", "2 years of college", "4 years of college", "5+ years of college" considered as college degrees. Among them, the high school degree is set as "0", and the college degree is set as "1", so our treatment, that is, the level of education, becomes a binary variable.

Propensity score

Before propensity score matching, a logistic regression model needs to be constructed. This model is used to predict the probability of the person getting a college degree. Predictor variables are region, age, sex, marital status. Therefore, probability of being treated, which is the probability of having a college degree, is calculated from this logistic regression. For every person who is actually a college degree, we want the untreated person who was

considered as similar to them as possible, based on propensity score. After the matching propensity score, we remove the observations who are not being matched. At this time, our process of propensity score matching is over, and we get a new set of data. The new data contains 27042 observations, which is 13521 treatment groups. In these 13521 pairs, we can think that all the features of each pair are "similar", the only difference is that the treatment is different, based on the propensity score of each pair is very close. Next, we can examine the 'effect' of being treated on salary income by using multiple linear regression.

Model

The model we need to build here is multiple linear regression (MLR). In order to get the best model, the AIC method is used here. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data, the lower AIC score provides a better model (Rebecca Bevans, 2020). So after performing the AIC method, we can get the best model. The final model contains region, sex, age, martial status, and education as predictors, outcome is respondent's wage and salary income. The formula of the MLR is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_{10} x_{10} + \beta_{11} x_{11} + \ldots + \beta_{15} x_{15} + \beta_{16} x_{16}$$

Where y represents the person's wage and salary income. In addition, β_0 represents the intercept of the model, 1means holding other variables constant, as x_1 increase 1 unit, y will increase on average by 1. For $x_3 \ldots x_{10}$, represent the dummy variables for the region, and $x_{11} \ldots x_{15}$ represent the dummy variables for marital status. For β_3, it means holding all other explanatory variables in the model fixed, if the respondent's region changes from baseline to dummy variable x_3, y will increase on average by β_3. Also, other coefficients have similar meanings. X16 represents the education, if education change from high school degree to college degree, the average of wage and salary income will increase by β_{16}.

Results

Model result

Here is the summary table of the model: Table1: summary table for multiple linear regression with all variables on predicting wage and salary income

term	estimate	std.error	statistic	p.value
(Intercept)	-6279.1123	2065.92919	-3.0393647	0.0023730
age	303.7213	30.04058	10.1103679	0.0000000
sexmale	19387.1253	701.18094	27.6492474	0.0000000
regioneast south central div	-4667.0563	1582.87648	-2.9484652	0.0031963
regionmiddle atlantic division	10222.0567	1311.86383	7.7920104	0.0000000
regionmountain division	-1936.8193	1495.46460	-1.2951288	0.1952869
regionnew england division	5008.7285	2030.56491	2.4666675	0.0136438

regionpacific division	7809.0514	1229.84663	6.3496140	0.0000000
regionsouth atlantic division	3238.1725	1181.63491	2.7404171	0.0061402
regionwest north central div	-1893.8714	1643.52822	-1.1523206	0.2491995
regionwest south central div	-1209.2077	1265.41803	-0.9555796	0.3392932
marstmarried, spouse absent	-5053.5010	2359.31174	-2.1419387	0.0322073
marstmarried, spouse present	16452.1448	1148.63051	14.3232698	0.0000000
marstnever married/single	-8543.0066	1301.84823	-6.5622139	0.0000000
marstseparated	-4594.0920	2647.51987	-1.7352436	0.0827091
marstwidowed	-5610.8357	2770.44243	-2.0252490	0.0428516
education	33043.7609	689.59623	47.9175486	0.0000000

As we can see from table1, the majority of these variable's p-values are less than 0.05, which are significant. For some extremely significant variables with P-value less than 0.01, such as age, sex, region at east south central division, region at middle atlantic division, region at pacific divisio. Also significant for some marital status is married, spouse present and never married/single. The last important extremely significant variable is education. So we use these significant variables that p-value less than 0.01 as predictors, our MLR will be:

$$incwage$$

$$= -6279.11 + (303.72)age + (19387.13)sexmale + (-4667.06)regioneastsouthcentraldiv + (10222.06)regionmiddleatlanticdivision + (7809.05)regionpacificdivision + (3238.17)regionsouthatlanticdivision + (16452.14)marstmarried, spousepresent + (-8543.01)marstnevermarried/single + (33043.76)education$$

The coefficient for age is 303.72, means holding other variables fixed, when respondent's age increases 1, the average annual salary income will increase by 303.72 dollars. As for sex, when the respondent's sex changes from female to male and keep everything else fixed, the average annual salary income will increase 19387.13 dollars. Others coefficients have similar meanings. As for education, the coefficient is 33043.76. This means holding other variables fixed, when the respondent's highest education level changes from high school degree to college degree, the average annual salary income will increase by $33043.76.

Baseline characteristics table

Since we have done the propensity score matching method, here is the table of the baseline characteristics of respondents.

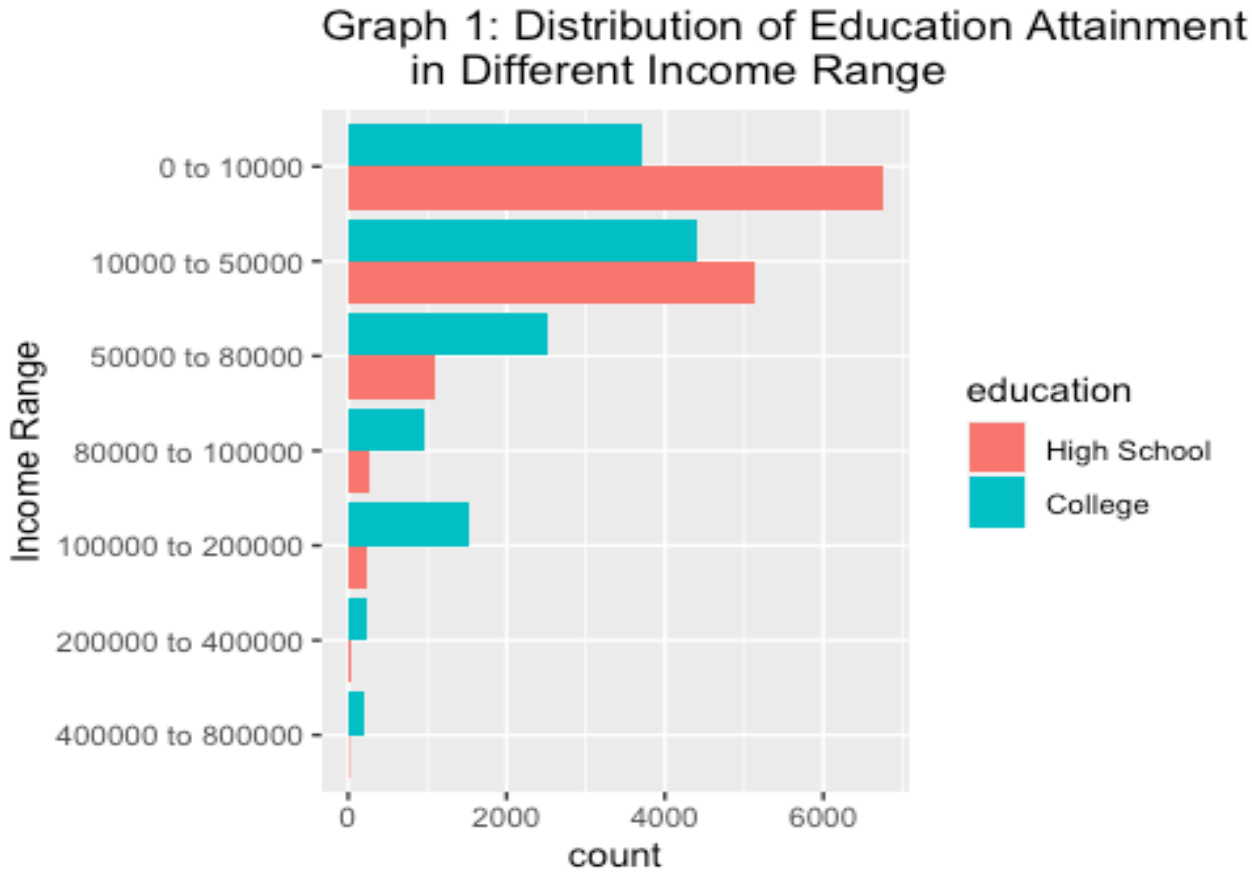
Table2: baseline characteristics of respondents, none of the differences between the groups was significant

	High School (N=13521)	College (N=13521)	Overall (N=27042)
sex			
female	6324 (46.8%)	5828 (43.1%)	12152 (44.9%)
male	7197 (53.2%)	7693 (56.9%)	14890 (55.1%)
age			
Mean (SD)	41.5 (15.3)	42.4 (13.5)	42.0 (14.4)
Median [Min, Max]	42.0 [18.0, 65.0]	42.0 [18.0, 65.0]	42.0 [18.0, 65.0]
region			
east north central div	2008 (14.9%)	2312 (17.1%)	4320 (16.0%)
east south central div	933 (6.9%)	876 (6.5%)	1809 (6.7%)
middle atlantic division	1677 (12.4%)	1623 (12.0%)	3300 (12.2%)
mountain division	967 (7.2%)	1165 (8.6%)	2132 (7.9%)
new england division	548 (4.1%)	409 (3.0%)	957 (3.5%)
pacific division	2093 (15.5%)	2119 (15.7%)	4212 (15.6%)
south atlantic division	2596 (19.2%)	2366 (17.5%)	4962 (18.3%)
west north central div	861 (6.4%)	778 (5.8%)	1639 (6.1%)
west south central div	1838 (13.6%)	1873 (13.9%)	3711 (13.7%)
marst			
divorced	1506 (11.1%)	1637 (12.1%)	3143 (11.6%)
married, spouse absent	387 (2.9%)	324 (2.4%)	711 (2.6%)
married, spouse present	5434 (40.2%)	6086 (45.0%)	11520 (42.6%)
never married/single	5608 (41.5%)	5043 (37.3%)	10651 (39.4%)
separated	323 (2.4%)	212 (1.6%)	535 (2.0%)
widowed	263 (1.9%)	219 (1.6%)	482 (1.8%)
incwage			
Mean (SD)	21900 (33000)	57000 (77100)	39500 (61800)
Median [Min, Max]	10500 [0, 665000]	40000 [0, 714000]	23300 [0, 714000]

The number of respondents with a high school degree and a college degree are the same, both have 13521 observations. It is trivial that the number difference across most characteristics shown in the table appears not so huge between the two groups (high school vs college). The only significant difference appears between the two groups on income wage. In the high school group, the average annual salary income is 21,900 dollars, while in the college group, the average annual salary income is $57,000. From this we can directly conclude that the average salary income of people with a degree in college is more than twice the average salary income of people with a high school degree.

Graph

In order to help us understand the relationship between education level and income more intuitively, the graph can show us clearly, the following graph shows the relationship between wage and salary income and education attainment.



From the above bar plots, several findings could be reasoned, shown as follows:

1. On the middle/low-end salary brackets (0-50k), we see that people with a high school degree earn a higher salary compared to ones with a college degree, percentage-wise. This scenario could potentially be due to middle/low-end job types being more often tied to technicians and small business units which do not require a college degree in order to operate.

2. As we move from the low-end bracket (0-10k) to higher-end brackets (100k-200k), we see that the number of people with a high school degree has decreased as the salary bracket increases. This is an interesting trend which shows an inverse linear relationship between salary range and number of people with only a high school degree. We have also observed that the ratio of the number of people with a college degree divided by the number of people with a high school degree increases from a negative to a positive value (10x and even 100x). This interesting trend shows us having a college degree contributes to a higher potential of high pay.

3. If we observe the high-end brackets (200k-400k and 400k-800k), we still see some outbeats from people with college versus with high school degrees, however, not so dramatically different compared to lower salary brackets. Reasons could be that even though obtaining a college degree may lead one to a higher salary, but it may become less effective because of a mixture of other influential factors that are also taken into account. These factors may be one's experience level, personal characteristics, industrial networking skills, and etc, and they then become more dominant to salary level in contrast to whether one has a college degree or not.

Discussion

Summary

In this report, we primarily studied the effect of education level on salary income. The first step was to select some variables of interest from the database, and then filter to obtain only people whose age is in the range of 18 to 65. Next step was to adapt the education level variable into a binary form to create a new database, and then use the method of propensity score matching for further analysis. Here, treatment is the education level is college, and outcome of interest is salary income. Hence, a logistic model was created to get a propensity score, and all data were then matched according to the propensity score. In this way, we would get similar features except treatment in each matched group. Following that, the next step was to do an MLR to predict salary income, get an MLR model (Table 1), and also get a baseline characteristic table (table 2), so that we know different education levels will lead to different salary income, reasoned from the result. Finally, based on the model, a graph (Graph 1) was established to show the relationship between education level and salary income, which can better help us analyze and affirm the relationship visually and more easily.

Conclusions

The propensity score analysis showed that people with a college degree will have a relatively higher salary than those with a high school degree. Under the circumstance that other conditions remain the same, the level of education changes from high school to college, and the average annual salary income will increase by more than $30,000 (p-value <0.001). This result tells us that the impact of academic qualifications on wages is indeed significant. In today's society, many students think that the impact of academic qualifications on future income is not important. However, in this research, if we want to have a relatively high salary in the future, getting a college degree is important. There are some other factors also impact salary income, like age, sex, martial statu. From age 18 to 65, as age increases, salary income increases as well, however the differences are not huge. The interesting result is, male are making more money than females, the average difference is more than $10000(p-value<0.001).

Weakness & Next Step

This report has some weakness in the following: First of all, the selection of variables in the original data is relatively subjective. If we choose different variables, we may get different results. Also, the data only includes people from 18 to 65 years old, and other age groups are not included in the analysis. When changing the level of education to binary, they are also divided according to their subjective consciousness, which may affect the results. The outcome (salary income) here is only income from wages and does not include income from other sources. People with low wages do not mean that their total income is low, so the results of the study may be different from the real situation. After the completion of the propensity score matching, some people were deleted, so our data became less, which may also lead to deviations in the results. There are also some flaws for the propensity score matching method. Propensity score matching cannot match on unobserved variables and we are using the data twice(Rohan Alexander, 2020).

For the future step of the analysis, we can try to choose different variables as the predictors to see if the results of the impact of education on salary income in MLR are different. When dividing the level of education, we can choose a different level of division, or choose another education-related variable as treatment to continue trying. We can also try different methods (for example, diff-in-diff, instrument variable) to study causal inference. Outcome can also be replaced by total personal income instead of just salary income. Of course, choosing one with more data than the current observation will also make the result more accurate.

References

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

(2) broom: David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1. https://CRAN.R-project.org/package=broom

(3) Code and flaw of propensity score matching: https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html#matching

(4) Code and flaw of propensity score matching: https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html#matching

(5) Summary table: https://cran.r-project.org/web/packages/sjPlot/vignettes/tab_model_estimates.html https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

(6) Propensity score matching : https://dimewiki.worldbank.org/wiki/Propensity_Score_Matching#:~:text=Propensity%20score%20matching%20(PSM)%20is,the%20impact%20of%20an%20intervention.

(7) Characteristic baseline table: https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html
(8) Change variable order in graph: https://www.r-graph-gallery.com/267-reorder-a-variable-in-ggplot2.html
(9) AIC: https://www.scribbr.com/statistics/akaike-information-criterion/#:~:text=The%20Akaike%20information%20criterion%20(AIC,best%20fit%20for%20the%20data.
(10) education and income artical: https://fcss-fesc.ca/news/2020/5/1/the-effects-of-income-inequality-in-education

Appendix

link for Github: https://github.com/ZiqingLyu/304-final-project