Education Attainment Effect on Future Income

Causal Inference Project - 05.03.2023
Men Yevgeniy, 320647266, yevgenimen@campus.technion.ac.il
Naseem Yehya, 212001689, naseem.yehya@campus.technion.ac.il
Manar Awida, 207364332, manar.awida@campus.technion.ac.il

Abstract

Elevated levels of formal education are typically associated with elevated income, improved job prospects, and enhanced quality of life. This correlation is typically applied to salaried work, whereas its link to self-employment is less clear. The aim of this project is to explore this connection and determine whether attaining higher levels of education has a positive causal impact on annual earnings. We utilize the CPS 2022 dataset from the United States Census Bureau, which includes questionnaires on general characteristics as well as employment-related information, such as union membership and income. We construct two causal graph models of the problem setting and evaluate the identification criteria, which enables us to utilize causal inference estimators like IPW, S-learner, and Matching. We conclude that there is a rising positive effect of higher degrees on income for all tertiary degrees, except for college degrees, which provide inconclusive findings. Our code repository can be found at our github[1].

1. Introduction

Higher levels of education are usually associated with higher earnings, better employment opportunities and higher quality of life. In particular, the acquisition of a tertiary degree is correlated with increased income in all countries that belong to the OECD. Nevertheless, this benefit may fluctuate depending on factors such as age, gender, the level of tertiary

[1] github.com/yevgm/causal-inf-project-2023

education attained, and the field of study pursued. Those with advanced degrees and extensive work experience are more likely to receive higher salaries.

The approach of 'higher education means higher payoff' is widely acceptable in the general public, with typical families encouraging their young members to attain a tertiary degree in an esteemed field. However, recent reports such as [1] suggest that even though workers with more education tend to earn more, there is substantial variation in earnings at each level of education. A higher level of education does not guarantee higher earnings, while less education does not always result in lower earnings. In addition, there is a substantial number of self-made millionaires without a tertiary degree who don't rely on a monthly wage. This has motivated us to try to answer two questions: **whether higher education attainment has a positive effect on a self employed individual's annual income and whether higher education attainment has a positive effect on salaried employee's hourly rate**, whereas education is defined as major degrees such as bachelor's and master's while the control group is defined as under college education. By using Causal Inference tools we drew two causal graph models of our questions settings (Figure 1, Figure 2) which includes the treatment - education attainment and outcomes - yearly income and hourly rate with all of the confounders. In Causal Inference setting, the Back Door criterion is perhaps the most common approach to identifying causal effects in observational research. In essence it is to identify and condition on all possible confounders. In addition, for the same purpose we can use a less well known criterion which is called the Front Door. In our causal graphs we identified the necessary confounders and found out that both the Back door and the Front door conditions do hold (see section 3) by assuming a non-significant effect of an individual's ability. We focus on only one country - the United States as income relies heavily on the local market. By reducing our study to only one country we reduce the number of hidden market confounders and. In this project we will try to answer these questions using CPS 2022 dataset and causal inference tools.

2. Datasets

At the beginning, we wanted to use the IBM dataset from Kaggle[6] but we realized that it was artificially generated and the income values seemed random. Hence, we In this project we decided to use the CPS dataset of 2022[5]. The CPS is a monthly survey of about 60,000 U.S. households conducted by the United States Census Bureau for the Bureau of Labor Statistics (BLS). In our case, we took an aggregated dataset from the whole year of 2022. The dataset contains approximately 800 columns of information about household, family and individuals and all relative data about their labor such as demographics, educational

attainment, annual income, hourly income, unemployment and more. The relevant data for our study will be described below:

- “A_AGE” - the age of an individual
- “A_HGA” - the education of a certain individual. It is encoded as years of study in school and academic degrees. In our case we aggregated all those with under college education as “under college”, Bachelor’s, Master’s, Doctorate and Professional degree (such as medical doctors and law).
- “A_SEX” - Male\Female
- “A_MJIND” - Occupation industry such as agriculture, construction, manufacturing, etc.
- “PRDTRACE” - Race as defined by the CPS dataset, includes white\black\american native\asian and all mixes. We aggregated all the mixes as “other” for computational reasons and the common support assumption.
- “PEARNVAL” - Total annual earnings of an individual
- “ERN_SRCE” - source of income salaried work\self employment.
- “ERN_OTR” - wage and salary money earned from other work, y/n
- “A_CLSWKR” - Class of worker
- “A_USLHRS” - Usual number of working hours per week
- “A_GRSWK” - payment per week

We applied some preprocessing on those columns to get the desired data that will be described in the data analysis section.

The challenges in this dataset are understanding the definition of all the columns as there are more than 800 column data in it. In addition, aggregation techniques must be applied as the dataset is very informative, e.g. education attainment must be combined to a few major groups. Some outliers had to be removed as the survey contains children.

3. Causal Inference Elements

To assess the causal effect of an individual's education attainment on his yearly income or hourly rate we first build two causal graph models. We decide to split the models into two graphs as we wish to assess the causal effect among those who solely are self employed or solely working for a salary. An attempt to include both in one causal graph would create correlations between unrelated nodes, such as a backdoor correlation between annual income and education through self employment and job level which is non-existent and irrelevant for self employed individuals.

In the graphs we describe all possible causal relations between education (Treatment) and yearly income or hourly rate (Outcome).
Figure 1 describes the causal graph model of a salary paid individual. It describes the effect of Education through the Industry type and possible confounders such as age, race, sex, personal ability, experience and personal preferences. Figure 2 describes the causal graph model of a self employed individual which is similar but rather simpler. Nodes in purple are the observed confounders while yellow nodes are known hidden confounders. From the graphs we deduce that the important hidden confounder "Ability" which represents the personal ability of an individual, is not blocking the back door to the treatment (in a d-separation sense, for the backdoor criterion) and that the front door condition does not hold as not all backdoor paths from the outcome to the mediators are blocked. In addition, our data is from only one country (the United States) and to satisfy the Common support condition we must focus only on the US.

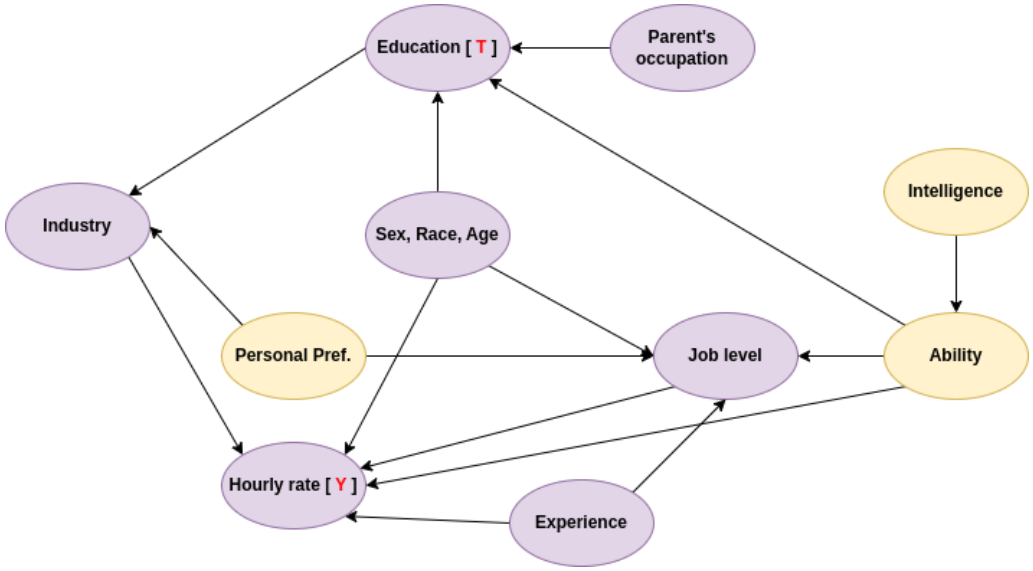

Figure 1. A causal graph model for salaried individuals. Yellow nodes are hidden confounders. The node "Sex, Race, Age" is a shorthand for 3 different nodes with the same edges.

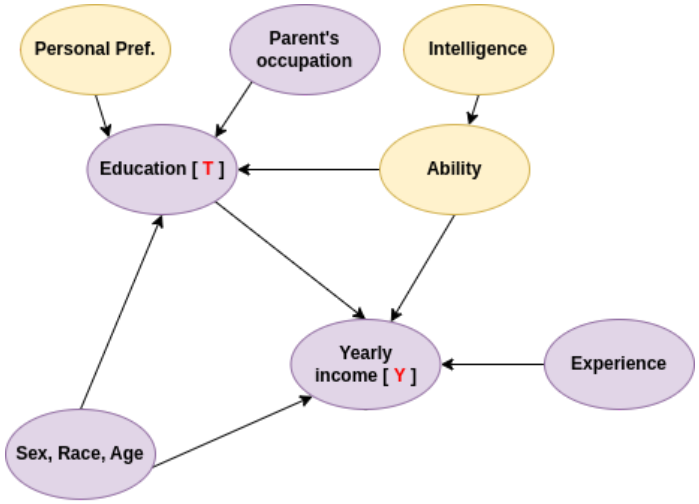

Figure 2. A causal graph model for self employed individuals. Yellow nodes are hidden confounders. The node "Sex, Race, Age" is a shorthand for 3 different nodes with the same edges.

3.1. Target trial

We attempt to describe our observational study as a randomized target trial in order to understand whether the observational study can emulate an hypothetical RCT. The target trial criteria are described below:

- Eligibility criteria: An individual over 18, with no college\university education. No self employment allowed.
- Treatment strategies: An individual must complete the assigned education program ('College', 'Bachelor', 'Master', 'Doctor').
- Assignment Strategies: The assignment of an education program will be random.
- Follow-up period: Starts at the completion of the program and ends 30 years after graduation.
- Outcomes: The hourly rate of the person, given the year after graduation.

Such a trial would have answered our question correctly. We conclude that if our analysis of the observational studies would be done right, we could emulate a RCT trial.

3.2. Identification Assumptions

When one is interested in establishing a causal effect, a set of four identification assumptions (SUTVA, Consistency, Ignorability, Common support) on the model must be met that allows for causal interpretation of the estimate. These conditions are widely used and a big part of causal inference is to understand when those conditions are plausible. In the following section we try to understand whether the assumptions plausibly hold.

3.2.1. Stable Unit Treatment Value Assumption (SUTVA)

To satisfy SUTVA assumption, the potential outcomes for any unit must not vary with the treatments assigned to other units, or in other words there is no interference between units. Furthermore, it requires that for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

In our case:

We assume that the hourly rate of other units has a small effect on the hourly rate of a particular unit. The small effect can be caused by market powers of supply and demand only. In addition, the education type and quality depends on the Institution. We conclude that SUTVA is **partially satisfied**.

3.2.2. Consistency

To satisfy the Consistency assumption, we must make sure that for a unit that receives treatment T, we observe the corresponding potential outcome Y_t. In our case, we rely on a self reported hourly rate, which can be noisy but we assume that most of the responders enter true values, hence the Consistency is **satisfied**.

3.2.3. Ignorability

Ignorability is satisfied when the potential outcomes are independent of treatment assignment, conditioned on observed covariates X, or in other words - there are no unmeasured confounders. In our case we tried to think of all possible confounders, but we have arrived at the conclusion that there are always some hidden confounders that we can't think of. In our opinion, most of the important confounders such as age, race, sex, job level, experience are measured. Hence, ignorability is **satisfied**.

3.2.4. Common Support

The common support condition is satisfied if:

$$P(T = t|X = x) > 0 \ \forall t, x$$

In our case, X consists of age, sex, job level, experience and the probability of attaining/not attaining an education is not 0. For example, both men and women have a higher than zero probability to get a high degree of any kind. Hence Common Support is **satisfied**.

3.3. Backdoor Criterion

Backdoor Criterion definition — Given an ordered pair of variables (T, Y) in a directed acyclic graph G, a set of variables X satisfies the backdoor criterion relative to (T, Y) if no node in X is a descendant of T, and X blocks every path between T and Y that contains an arrow into T. This definition is easy to understand intuitively: to understand the direct effect of T on Y we simply must make sure to keep all direct paths intact while blocking off any and all spurious paths. In addition, this criterion gives up a set of variables X that must be adjusted for, meaning that all confounders in X must be included in our analysis to assure correct causal effect. In figure 3 we redrew the two graphs from above but in a more technical form that allows us to review all the backdoor paths. In the left graph we can see that only four backdoor paths exist:

1. Y<-A->T
2. Y<-S.R.A->T (3 paths)

Thus to block both paths we must adjust for S.R.A and A.
In the right graph the backdoor paths are:

1. Y<-S.R.A->T (3 paths)
2. Y<-S.R.A->J.L<-A->T
3. Y<-J.L<-P.P->Ind<-T
4. Y<-J.L<-S.R.A->T
5. Y<-J.L<-A->T
6. Y<-A->T
7. Y<-Exp->J.L<-A->T
8. Y<-Exp->J.L<-P.P->Ind<-T
9. Y<-Exp->J.L<-S.R.A->T

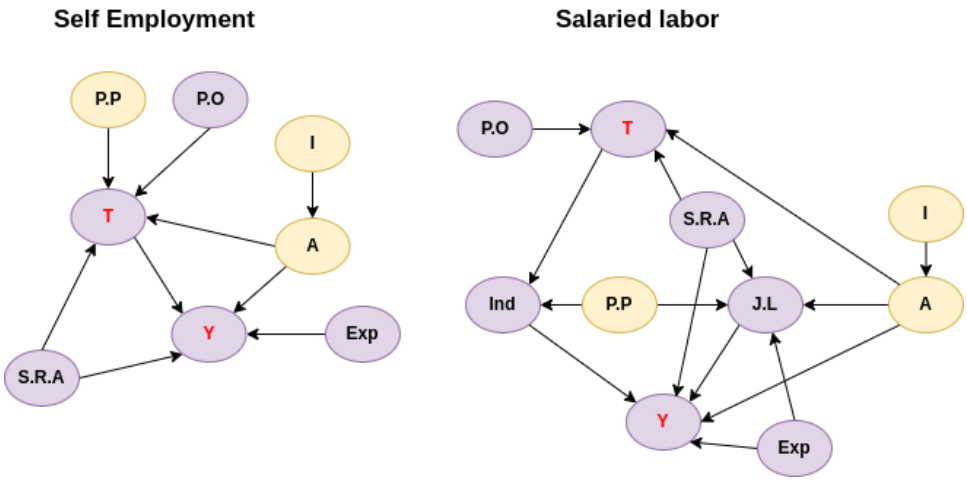

Figure 3. Both causal graphs from figure 1 and 2, drawn in a compact and easy to analyse manner.

In the paths we can see that the X set must contain A and S.R.A in order to block all backdoor paths in both graphs (paths' nodes colored in green). In our case we can't measure the Ability (other than some proxy as grades that we do not have) and we assume that the education level has a higher impact on the outcomes than the individual's ability. In conclusion, we adjust our calculations using three confounders - Sex, Age and Race in both cases, but keep in mind that the results may be limited by identifiability.

4. Data Analysis

First the dataset must be pre-processed in order to be able to train models effectively. The categorical columns were transformed into binary dummy data columns, and all the children and individuals without valid income values were removed from the analysis. At the end we are left with 9811 individuals. This final dataset was transformed into 5 subsets such that each includes the binary control "under college" and each treatment (e.g. Bachelor's), for easier further analysis. In the following plots we show the feature distribution of the data - Figure 4, Hourly rate given the Education attainment treatment - Figure 5, yearly income given the education attainment treatment - Figure 6 and the propensity scores of our 3 different models for example treatments figures 8.

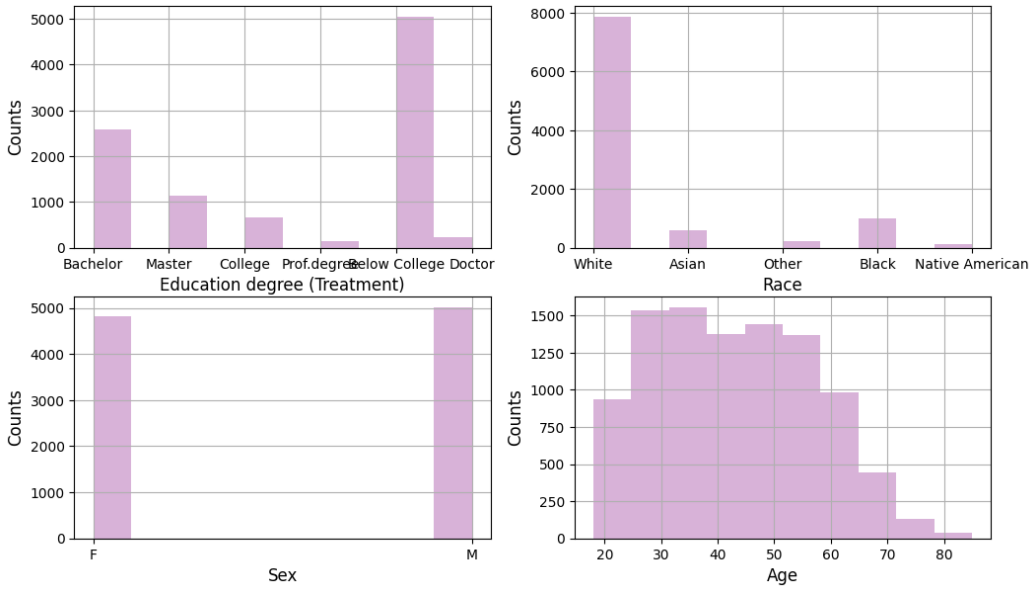

Figure 4. The distribution of the features of the CPS dataset. We can see that we have more control group individuals than the treated group, and imbalance between the races.

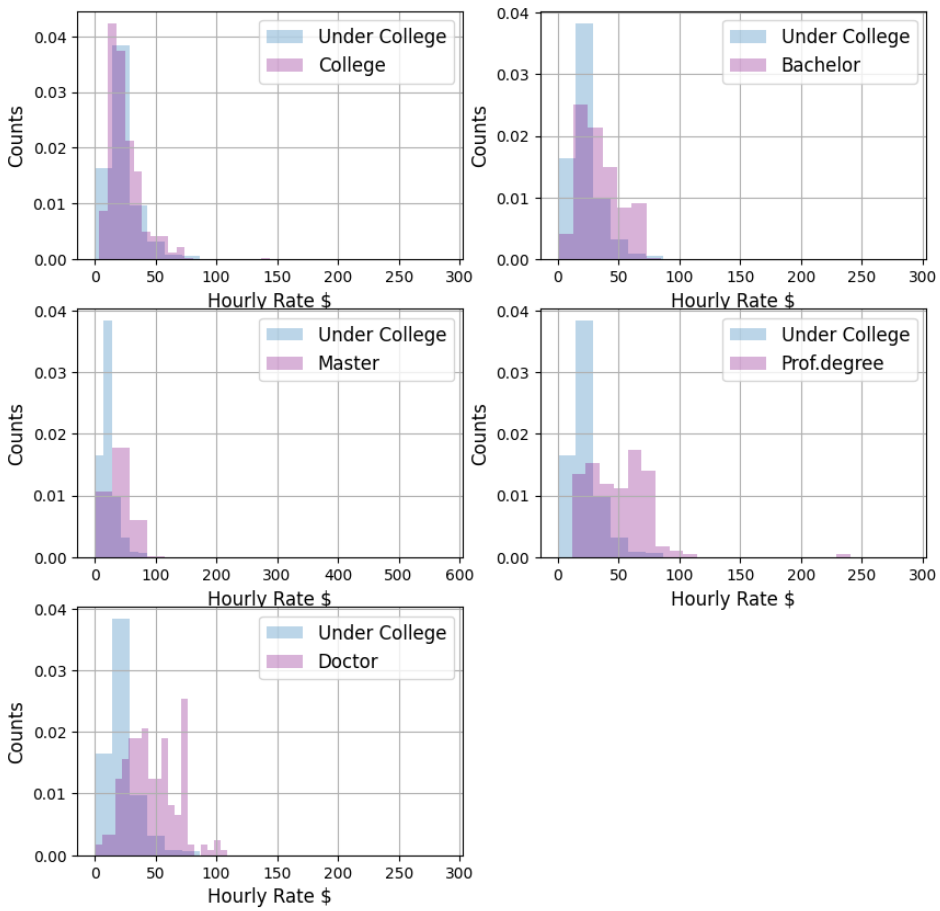

Figure 5. Bar plots of hourly rate of each treatment group against the control group. It seems that all the groups' distributions skewed towards the higher income except for the college degree.

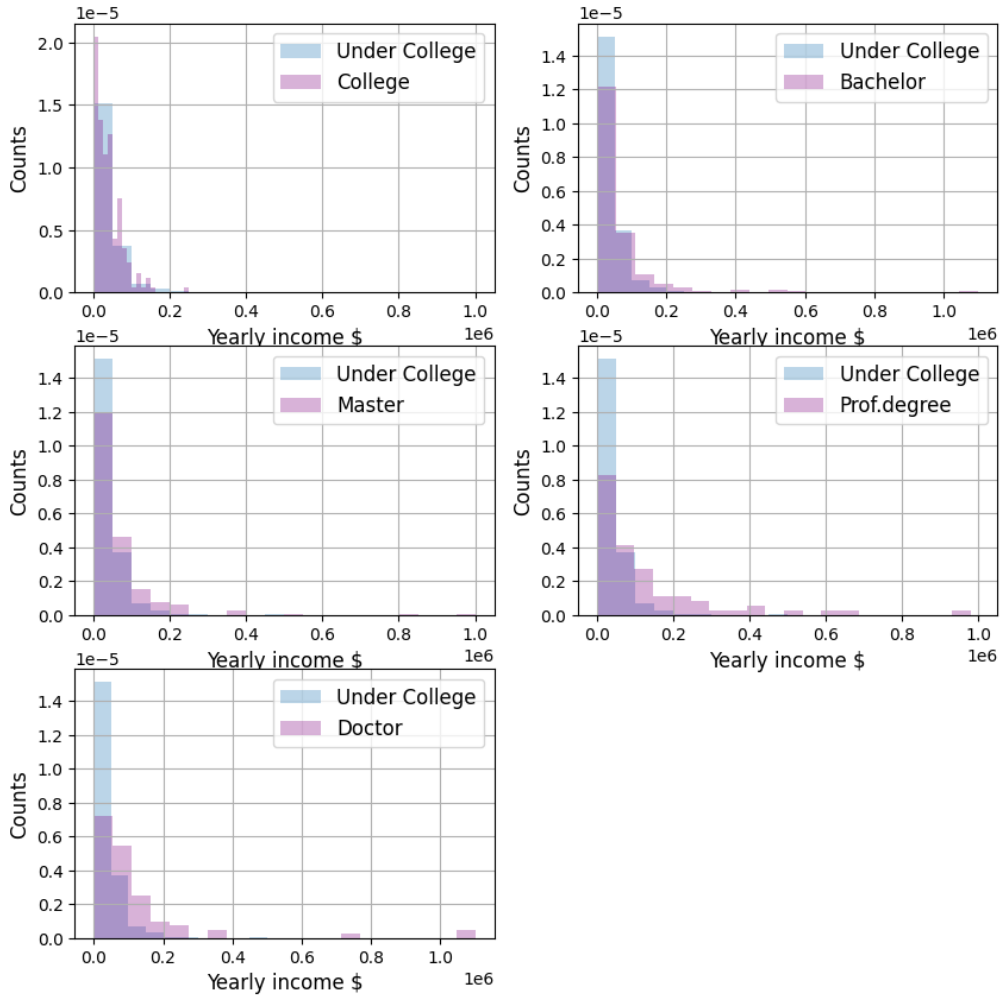

Figure 6. Bar plots of annual income of each treatment group against the control group. It seems that only doctorate and Professional degrees distributions' skewed towards higher income.

From the figures above we can learn that higher education is correlative both with higher hourly rate and yearly income of self employed individuals in the professional degree and doctorate group. Although it seems enough, we can't claim that there is a causal relation between each pair as there may be hidden confounders that can create this correlation.

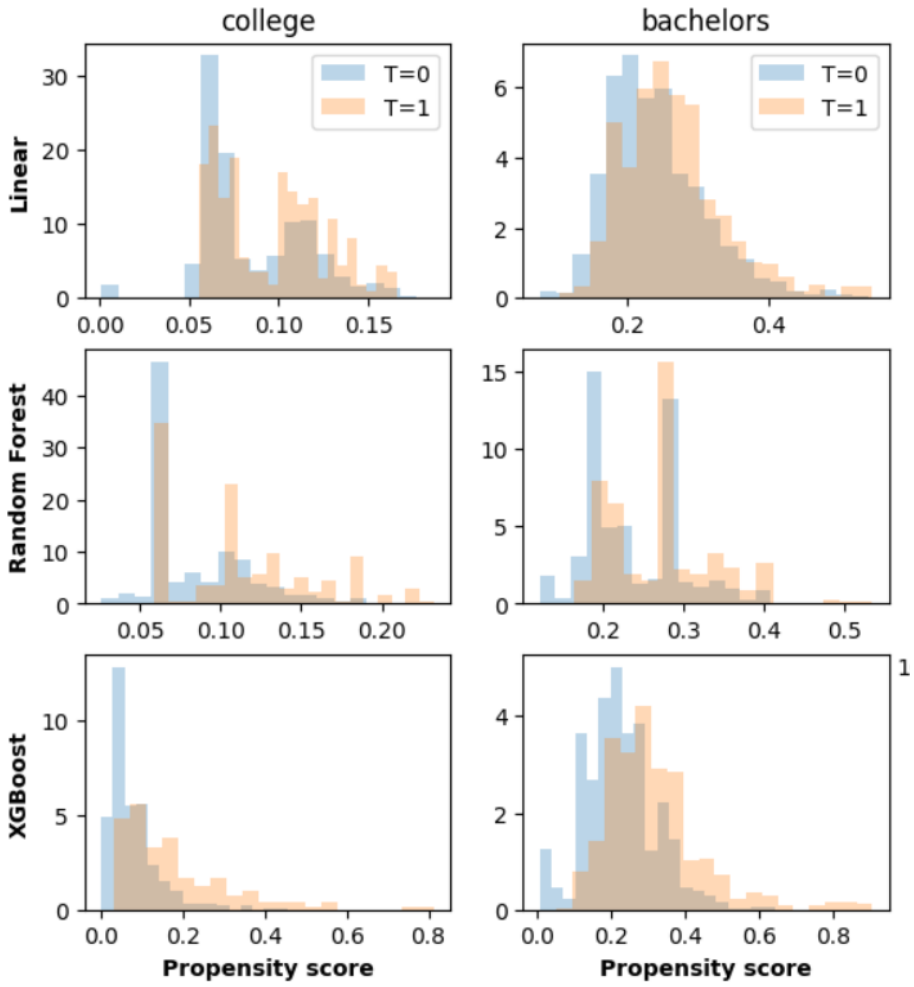

Figure 7. The propensity scores of our three different models colored by treatment. We can see that some scores show more overlap while others less overlap, Hence we assume that the IPW method should give us reliable results.

5. Estimation Methods

In this section we will present the estimation methods that we have used to estimate the causal effect. Our main goal is to estimate the ATE (Average treatment effect) and ATT (Average treatment effect among the treated). Theoretically speaking, to calculate the ATE we have to know the potential outcomes Y_0 and Y_1, then, $ATE = E[Y_1 - Y_0]$. As one of them can never be measured, the estimated ATE has the following formula:

$$ATE = E_{x \sim p(x)}[E[Y|X, T = 1] - E[Y|X, T = 0]]$$

To find both terms we must use estimation techniques such as Covariate adjustment and IPW. Another estimand is the ATT which is similar to the ATE but only for the treated group. In the case of RCT's, both estimates are the same, but empirically in observational studies they are almost always different as a result of hidden confounders.

$$ATT = E_{x\sim p(x)}[E[Y_1|X, T = 1] - E[Y_0|X, T = 0]]$$

The control group T_0 is defined as under college education, while T_1 is all other degrees separately, such that we compare all degrees to "under college" education and as a consequence we have multiple ATE's and ATT's for each comparison. Such formulation allows us to reduce the common support violation. Additionally, In all the following methods we report the median and the interquartile distance $[q_{25}, q_{75}]$ of a bootstrap of 100 subsets of 60% from the full dataset.

To get more reliable results that do not depend on model selection and training, the results of IPW, T-Learner and S-Learner were obtained by using 3 different models as estimators - Linear regression\classification, XGBoost and Random forest.

5.1. Matching

The idea behind matching is to create a comparison group that is similar to the treatment group in terms of observable characteristics, so that any differences in outcomes between the two groups can be attributed to the treatment effect. The group is constructed as 6-nearest neighbors and calculated as follows:

$$ATE = \frac{1}{N}\sum_{i=1}^{N} ITE(i)\ , ITE(i) = (y_{j(i)} - y_i) \cdot I_{[t_i=0]} + (y_i - y_{j(i)}) \cdot I_{[t_i=1]}$$

Where $j(i)$ is the set of 6 nearest neighbors defined by the ball tree algorithm of sklearn library.

The results are presented in the following tables:

		6 NearestNeighbors
ATE	College	(2.81, 2.46, 3.23)
	Bachelor	(12.48, 12.12, 12.91)
	Master	(17.85, 17.36, 18.33)
	Prof.degree	(25.27, 24.13, 26.28)
	Doctor	(22.86, 22.26, 23.57)
ATT	College	(2.89, 2.4, 3.46)
	Bachelor	(12.54, 12.15, 12.84)
	Master	(17.98, 17.56, 18.48)
	Prof.degree	(25.75, 24.21, 26.86)
	Doctor	(22.71, 21.87, 23.49)

		6 NearestNeighbors
ATE	College	(1443.96, -739.8, 3350.62)
	Bachelor	(26031.26, 23876.76, 28835.13)
	Master	(35847.49, 31642.15, 39615.94)
	Prof.degree	(83493.14, 71983.61, 95022.01)
	Doctor	(79101.06, 67589.11, 92017.36)
ATT	College	(1496.09, -192.47, 3342.46)
	Bachelor	(26034.98, 24353.43, 28410.61)
	Master	(34582.54, 30081.13, 38688.68)
	Prof.degree	(80132.47, 68655.76, 95484.66)
	Doctor	(86160.61, 73025.77, 100372.95)

Figure 9. ATE and ATT results for salaried work (left) and self employed (right) individuals for Matching method. How to read: [median, lower bound, upper bound].

From the tables above we can conclude that all models agree that each degree increases the hourly rate and the yearly income of self employed individuals, except the college degree level - which gave inconclusive results in terms of confidence interval in the self employed ATE.

5.2. IPW

IPW (Inverse propensity weighting) method relies on the estimation of the propensity score - $e(x) = P(T = 1|X = x)$, and weighting the ATE\ATT by the probability of the treatment given the confounders. Such a weighting allows for balancing between the imbalanced proportion of treated\controls in an observational study. In this project we will use the normalized IPW (IPWN[3]) which gave us better results in terms of stability and confidence intervals compared to the traditional IPW:

$$ATE = \frac{\sum_{i=1}^{N} t_i y_i / e(x_i)}{\sum_{i=1}^{N} t_i / e(x_i)} - \frac{\sum_{i=1}^{N} (1 - t_i) y_i / (1 - e(x_i))}{\sum_{i=1}^{N} (1 - t_i) / (1 - e(x_i))}$$

$$ATT = \frac{\sum_{i=1}^{N} t_i y_i}{\sum_{i=1}^{N} t_i} - \frac{\sum_{i=1}^{N} (1 - t_i) y_i \cdot e(x_i) / (1 - e(x_i))}{\sum_{i=1}^{N} (1 - t_i) \cdot e(x_i) / (1 - e(x_i))}$$

We have used 3 models for $e(X)$: Linear regression, Random forest regression and XGBoost regression. The results are presented in the following tables:

		Linear	Random Forest	XGBoost
ATE	0	(3.63, 3.22, 4.11)	(3.13, 2.83, 3.72)	(3.15, 2.68, 3.55)
	1	(12.98, 12.55, 13.4)	(12.5, 12.16, 12.93)	(12.24, 11.77, 12.7)
	2	(18.66, 18.06, 19.4)	(18.46, 17.8, 19.16)	(18.34, 17.72, 19.02)
	3	(27.05, 25.44, 28.58)	(26.58, 24.52, 28.31)	(25.37, 23.78, 27.26)
	4	(23.13, 22.02, 24.12)	(24.02, 22.23, 24.97)	(24.07, 22.7, 25.3)
ATT	0	(3.55, 3.06, 3.93)	(2.83, 2.5, 3.35)	(2.68, 2.32, 3.18)
	1	(13.09, 12.7, 13.46)	(12.34, 11.96, 12.68)	(12.48, 12.05, 12.91)
	2	(18.57, 17.94, 19.24)	(17.92, 17.36, 18.59)	(17.85, 17.43, 18.46)
	3	(26.03, 24.28, 28.1)	(25.86, 23.91, 27.55)	(25.24, 23.05, 26.82)
	4	(23.2, 21.89, 24.51)	(22.84, 21.7, 23.88)	(22.41, 20.84, 23.77)

Figure 10. ATE and ATT results for salaried work individuals for IPW method. How to read: [median, lower bound, upper bound].

		Linear	Random Forest	XGBoost
ATE	0	(2707.97, -585.2, 5314.02)	(3176.74, 175.83, 6291.57)	(5069.12, 1909.21, 8471.92)
	1	(27987.29, 25297.21, 30971.61)	(28245.38, 24771.41, 31631.01)	(29811.3, 26270.15, 34188.13)
	2	(42846.81, 31451.3, 52134.24)	(41159.12, 31621.6, 52574.61)	(48905.7, 37335.41, 65300.34)
	3	(93339.67, 70691.55, 117499.42)	(112062.12, 84170.47, 131457.63)	(114736.83, 99992.92, 151414.18)
	4	(149829.01, 108375.55, 197365.61)	(139191.1, 98362.11, 178211.77)	(167792.24, 88315.2, 236133.81)
ATT	0	(1292.19, -1619.03, 3801.55)	(1804.56, -426.77, 3760.87)	(1254.4, -1357.12, 3378.99)
	1	(26290.19, 23096.75, 29131.69)	(26568.27, 22281.53, 29221.91)	(25779.71, 22066.06, 28506.46)
	2	(33789.22, 28058.03, 39367.56)	(31741.23, 25663.33, 39004.04)	(32744.42, 28052.52, 39486.12)
	3	(86155.46, 65487.63, 101549.16)	(83999.41, 66862.62, 102296.23)	(78463.66, 64540.21, 91008.18)
	4	(80154.51, 65680.72, 100998.98)	(89865.67, 70895.25, 109729.46)	(85665.26, 61032.25, 101264.79)

Figure 11. ATE and ATT results for self employed individuals for IPW method. How to read: [median, lower bound, upper bound].

From the tables above we can conclude that all models agree that each degree increases the hourly rate and the yearly income of self employed individuals, except the college degree in self employed individuals.

5.3. T Learner

T-Learner is an estimation method for ATE/ATT that falls under the covariate adjustment family methods. T stands for Two, as we fit one model to each treatment $f(x) \approx E[Y_1|X, T = 1]$, $g(x) \approx E[Y_0|X, T = 0]$, and then:

$$ATE = \frac{1}{N}\sum_{i=0}^{N} f(x_i) - g(x_i)$$

$$ATT = \frac{1}{N}\sum_{i=0}^{N} f(x_i) - g(x_i)$$, Where X is actually X|T=1 and N is the number of treated.

We have used 3 models for $f(\cdot)$: Linear regression, Random forest regression and XGBoost regression. The results are presented in the following tables:

		Linear	Random Forest	XGBoost
ATE	**College**	(3.57, 3.2, 3.78)	(2.92, 2.52, 3.23)	(2.86, 2.46, 3.21)
	Bachelor	(12.98, 12.74, 13.17)	(11.78, 11.6, 12.05)	(11.65, 11.39, 11.94)
	Master	(18.6, 18.11, 18.9)	(17.59, 17.06, 17.99)	(17.55, 16.99, 18.08)
	Prof.degree	(25.92, 25.27, 26.81)	(26.07, 25.02, 27.45)	(25.49, 24.04, 27.3)
	Doctor	(23.18, 22.27, 23.78)	(23.77, 23.0, 25.01)	(24.44, 23.35, 25.43)
ATT	**College**	(3.18, 2.89, 3.53)	(2.99, 2.67, 3.23)	(2.83, 2.61, 3.1)
	Bachelor	(13.04, 12.79, 13.29)	(12.44, 12.29, 12.76)	(12.45, 12.23, 12.71)
	Master	(18.41, 17.99, 18.91)	(18.05, 17.47, 18.39)	(17.8, 17.44, 18.15)
	Prof.degree	(25.62, 24.6, 27.02)	(25.26, 23.75, 26.43)	(25.41, 24.07, 26.22)
	Doctor	(22.99, 22.19, 23.66)	(22.65, 22.0, 23.55)	(22.55, 21.9, 23.4)

Figure 12. ATE and ATT results for salaried work individuals for T-Learner method. How to read: [median, lower bound, upper bound].

		Linear	Random Forest	XGBoost
ATE	**College**	(3098.95, 1465.62, 4834.75)	(3570.77, 885.21, 5578.75)	(3824.82, 1271.11, 5701.73)
	Bachelor	(29873.77, 27214.47, 32047.8)	(30322.75, 27303.93, 32274.64)	(30845.92, 28326.89, 32933.8)
	Master	(44410.74, 38041.14, 51813.18)	(47254.24, 39127.67, 55351.98)	(47226.79, 34417.22, 60008.19)
	Prof.degree	(120433.15, 107013.93, 139373.98)	(112517.02, 96424.72, 130035.08)	(103362.41, 86109.43, 122822.11)
	Doctor	(148315.66, 127466.19, 180368.53)	(159517.38, 130607.11, 179765.93)	(139305.74, 109684.32, 177465.0)
ATT	**College**	(1643.87, 174.62, 2945.77)	(2268.86, -188.04, 3599.3)	(2087.11, 312.9, 3871.88)
	Bachelor	(25310.2, 23389.97, 27875.08)	(26223.38, 24407.04, 28150.9)	(26426.27, 23997.71, 28173.3)
	Master	(31590.4, 27895.02, 36765.44)	(36750.97, 31396.85, 41050.98)	(34858.15, 31002.54, 39084.23)
	Prof.degree	(82512.84, 70014.55, 91601.45)	(83290.26, 72581.0, 96143.89)	(84283.36, 73976.94, 93743.89)
	Doctor	(83705.79, 70794.99, 94763.61)	(89420.27, 77518.68, 102297.39)	(86377.87, 73770.14, 103549.67)

Figure 13. ATE and ATT results for self employed individuals for the T-Learner method. How to read: [median, lower bound, upper bound].

The tables above also agree with the previous estimation methods.

5.4. S Learner

S-Learner is an estimation method for ATE/ATT that falls under the covariate adjustment family methods. S stands for Single, as we fit one model to both terms $f(x,t) \approx E[Y_t|X, T = t]$, and then:

$$ATE = \frac{1}{N}\sum_{i=0}^{N} f(x_i, 1) - f(x_i, 0)$$

$$ATT = \frac{1}{N}\sum_{i=0}^{N} f(x_i, 1) - f(x_i, 0)$$, Where X is actually X|T=1 and N is the number of treated.

Similarly to T-learner, we have used 3 models for $f(\cdot)$. The results are presented in the following tables:

		Linear	Random Forest	XGBoost
ATE	College	(3.25, 2.99, 3.55)	(2.81, 2.46, 3.13)	(2.63, 2.38, 2.88)
	Bachelor	(13.07, 12.78, 13.25)	(11.96, 11.7, 12.16)	(11.82, 11.57, 12.02)
	Master	(18.49, 17.89, 18.77)	(17.52, 16.86, 18.15)	(17.61, 16.79, 17.99)
	Prof.degree	(26.19, 25.28, 27.32)	(25.88, 24.45, 27.22)	(25.18, 23.72, 26.63)
	Doctor	(22.76, 21.95, 23.44)	(23.71, 22.27, 25.07)	(23.64, 22.67, 24.43)
ATT	College	(3.43, 3.17, 3.73)	(3.01, 2.56, 3.25)	(2.76, 2.5, 3.18)
	Bachelor	(12.97, 12.77, 13.16)	(12.46, 12.26, 12.7)	(12.42, 12.14, 12.62)
	Master	(18.55, 18.11, 19.01)	(17.95, 17.57, 18.31)	(17.75, 17.2, 18.32)
	Prof.degree	(26.22, 25.12, 27.3)	(25.25, 24.03, 26.49)	(25.15, 24.19, 26.05)
	Doctor	(22.91, 22.17, 23.93)	(22.92, 22.21, 23.61)	(22.66, 21.97, 23.48)

Figure 14. ATE and ATT results for salaried work individuals for the S-Learner method. How to read: [median, lower bound, upper bound].

		Linear	Random Forest	XGBoost
ATE	College	(1458.42, 322.05, 3234.43)	(4508.94, 2965.33, 6156.64)	(3299.44, 1617.59, 5010.96)
	Bachelor	(26497.92, 24185.87, 28782.5)	(29996.01, 27389.21, 32641.25)	(29561.85, 26248.96, 31639.19)
	Master	(35195.59, 29872.51, 39162.46)	(46322.22, 38931.96, 57591.52)	(44960.9, 35736.31, 55200.04)
	Prof.degree	(78262.53, 66027.71, 91496.96)	(112771.19, 94701.01, 126651.87)	(108352.82, 89261.7, 125569.1)
	Doctor	(91397.98, 76253.68, 102246.25)	(156351.1, 119904.1, 194214.38)	(138217.64, 113481.12, 178767.95)
ATT	College	(1476.27, -12.49, 3062.79)	(2592.07, 855.67, 3944.96)	(1755.5, 254.15, 3650.26)
	Bachelor	(26965.86, 25141.81, 29341.34)	(26752.42, 24955.33, 29272.19)	(26461.48, 24395.92, 28843.18)
	Master	(34465.38, 30079.37, 39707.33)	(35325.05, 31445.02, 39568.44)	(35895.06, 30699.47, 39717.15)
	Prof.degree	(83511.27, 71422.03, 92951.46)	(87164.6, 75534.43, 95772.12)	(81488.7, 68435.67, 94669.24)
	Doctor	(81908.7, 68651.35, 98551.93)	(85391.31, 68027.09, 98663.78)	(83713.97, 71172.15, 97176.93)

Figure 15. ATE and ATT results for self employed individuals for the S-Learner method. How to read: [median, lower bound, upper bound].

The tables above also agree with the previous estimation methods.

6. Discussion

From our results we can conclude that all the estimation methods agree that among salaried work and self employment individuals - the education attainment of bachelor's degree and above will increase the future income in the form of hourly rate of yearly income. All the methods show inconclusive results regarding the college degree (positive and negative bounds), especially whether a college degree causes an increase in the income of the self-employed individuals, which makes sense in a way, since a college degree is helpful for being accepted to a specific job more than developing a skill to earn more as a self employed individual. Based on our findings, we provide empirical evidence supporting the widely held belief that education positively impacts future income in the context of salaried work. Additionally, we contribute to the existing literature by extending this relationship to self-employed individuals, which may not be immediately apparent given the absence of a direct correlation between education and payment rates.

7. Future Work and Improvements

1. Identification: It is not easy to build a correct model of a causal effect and it requires extensive prior knowledge. It is likely that our model is not perfect and that we skipped important confounders which may reduce the effect or even flip it's direcion.
2. Identification: As stated in the assumptions section, SUTVA condition holds only partially. Among each treatment value there are subgroups e.g. Bachelor's degree can be a BA (Arts) or BSc (Science). Both degrees are different and yet they share some similarities. Thus, the conclusion that education attainment increases future income may also be limited, as some degrees may not be so useful and some may be much more useful than others. One way to reduce its effect would be to consider a more specific degree such as scientific degrees, or only engineering. Such a formulation would require more complex calculations and thorough work.
3. Estimation: Our data was imbalanced - we had much more control of individuals that were treated. Such a setting imposes a challenge on machine learning estimators that we use in our methods, which can lead to incorrect models and estimations. To reduce this effect we must balance the data by downsampling or upsampling.

8. Bibliography

1. https://cew.georgetown.edu/cew-reports/collegepayoff2021/#resources
2. Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. Advances in Methods and Practices in Psychological Science. 2018;1(1):27-42. doi:10.1177/2515245917745629
3. Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. Biometrical Journal, 59(5), 967–985. doi:10.1002/bimj.201600094
4. Pearl Judea Madelyn Glymour and Nicholas P Jewell. 2016. Causal Inference in Statistics : A Primer. Chichester West Sussex: Wiley.
5. https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html
6. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset