The Impact of College Education on Future Earnings

Meka Adegbola, Zaid Usmani and Alfredo Gago*

Carleton University

April 11, 2025

Abstract

This project investigates the causal relationship between college education on annual income earnings using data from the National Longitudinal Survey of Youth 1997 (NLSY97). Traditional methods of estimating returns to education often suffer from omitted variable bias and incorrect model specification. To address these limitations, we utilize a modern causal inference framework that combines Directed Acyclic Graphs (DAGs) with Double Debiased Machine Learning (DDML). We construct multiple DAGs to represent different theoretical assumptions about the underlying data-generating process and identify appropriate adjustment sets for estimating causal effects. Our treatment variable is the attainment of a college degree, while income serves as the primary outcome of interest. Control variables include GPA, PIAT scores, family income, parental education, age, and others. This project builds on recent studies by allowing for heterogeneous treatment effects, but in a data-driven way rather than through a structural model, as achieved in previous studies. DDML allows us to flexibly control for high-dimensional confounders without overfitting, producing more robust and unbiased estimates. The results provide insights into the magnitude of returns to education, consistent with our expectations, and highlight the importance of careful causal design when evaluating policy-relevant questions in labor economics.

Keywords: Income, College Degree, Double Debiased Machine Learning, Directed Acyclic Graphs, Random Forest, LASSO

We would like to thank Professor Thomas Russel for helpful comments. All errors are our own.

*Meka Adegbola, Zaid Usmani and Alfredo Gago, Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S5B6, Canada. Email: zaidnihaluddinusmani@cmail.carleton.ca, alfredogago@cmail.carleton.ca, mekaadegbola@cmail.carleton.ca.

1 Introduction

Education plays a vital role in modern labour market. It is a preconceived notion that a college degree greatly enhances your chances of increased earnings; studies in multiple countries have shown that better-educated individuals earn higher wages, face less unemployment and get access to work in much prestigious firms, as compared to their less-educated counterparts. However, in the absence of experimental evidence, social scientists are hesitant to conclude that better education is the main cause of higher earnings, even though there is a strong correlation between education and labor markets. The desire to pioneer the analysis of human capital efficiency was first initiated by Gary Becker Becker (1964), when he theoretically emphasized the importance of college education on earnings. At that time, modern empirical analysis was at its infancy and so, this topic was further taken up by Griliches (1977), when he performed empirical analysis using Instrumental Variables estimation. He also introduced the issue of ability bias- the possibility of earning higher income due to higher cognitive ability or higher parental income, instead of higher education. Card (1999) significantly improved on this by using compulsory schooling laws and geographic variation in college proximity as instruments for education, addressing the endogeneity concern. In recent studies such as that of Heckman et al. (2018), the authors use a structural model that accounts for selection bias, unobserved heterogeneity, and heterogeneous treatment effects. They model the decision to attend school and its impact in a more traditional econometric framework. However, we use a more recent machine learning–based method designed to flexibly estimate treatment effects in high-dimensional settings, which allows you to control for a large set of variables non-parametrically.

This study aims to quantify the causal effect of a college degree on earnings by employing Double Debiased Machine Learning (DDML), a robust econometric method that overcomes the challenges of endogeneity and high-dimensional covariates. Using data from the National Longitudinal Survey of Youth 1997(NLSY97), we apply DDML to estimate the returns to education while addressing potential sources of bias. Additionally, we explore heterogeneity in returns based on gender, race, and parental background, capturing dynamic treatment effects and unobserved heterogeneity in the education-earnings relationship. Almost every paper that investigates the relationship between college education and earnings uses the Mincer (1958) equation for their model, which is

$$\ln Y(S_i, X_i) = \gamma_i + \rho_i \underbrace{S_i}_{\text{Years of Schooling}} + \phi \underbrace{(X_i)}_{\text{Other Determinants}}$$

where Y(Si, Xi) is the earnings of individual i with Si years of education and a vector of other

determinants Xi.

Data and the empirical context

The dataset we chose is the NLSY97 (National Longitudinal Survey of Youth 1997), a nationally representative panel study conducted by the U.S. Bureau of Labor Statistics in collaboration with the Center for Human Resource Research at Ohio State University. The study began in 1997, surveying a cohort of approximately 9,000 youth who were between the ages of 12 and 16 as of December 31, 1996, corresponding to birth years 1980–1984. The sample was selected using a stratified multistage area probability design and included an oversample of Black and Hispanic youth to support subgroup analysis. While all respondents were living in the United States at the time of initial data collection, not all of those surveyed were U.S. citizens.

The NLSY97 collects data on a broad range of topics including education, employment, income, job training, family formation, criminal behavior, health, and demographics. Respondents were interviewed annually from 1997 to 2011 and biennially thereafter. As of March 2025, the study remains ongoing, with Round 22 scheduled to begin in September 2025. Because it follows the same individuals over time, the NLSY97 is structured as panel data, which is especially valuable for examining causal relationships and life-course dynamics, such as the long-term effect of education on wages.We initially chose the year 2006 for our analysis. However, after performing our regressions and analysis, we realised that that 2006 was too early for the respondents to realise the effect of education, which is why we were getting a negative relationship. We then proceeded to choose the year 2010 which provided us with a positive relationship between education and earnings.

Our Treatment Variables are Highest Degree attained, Gpa, Piat Score and the highest grade achieved by mother, father and resident father. We are trying to test if education exposure and attainment have a significant effect on income, our outcome variable. We are also accounting for covariates such as age, sex, race, urban/rural residence, household size, gross family income (1997), number of siblings, and household members below the age of 18; these have an effect on education and income, so we need to control for them to isolate the effect of education on income.

Variables and Explanation:

In this study, our **outcome variable** is **income**, which captures the economic returns to educational attainment and serves as the dependent variable in all regression models. The key **treatment variable** is **college degree attainment**, a binary indicator coded as 1 for individuals who have completed a college degree and 0 otherwise. This operationalization reflects the well-established role of post-secondary education as a signal of human capital accumulation and a determinant of labor market outcomes.

To address potential confounding and improve causal identification, we include a comprehensive set of **covariates**. **Age** is included to account for experience-related variation in both income and educational attainment, as older individuals are generally more likely to have completed higher levels of education and to earn higher wages. **Sex** is controlled for due to well-documented gender disparities in labor market returns and educational attainment, with women more likely to complete higher education but persistently earning less than men. **Race** is included as a categorical variable to adjust for racial disparities in educational access and income, particularly among Black and Hispanic populations, who have historically faced systemic disadvantages.

Household composition is captured through three variables: **total household size, number of siblings, and number of household members under the age of 18**. As Butcher and Case (1994) mention in their article, sibling sex composition has a significant effect on women's educational attainment. These serve as proxies for resource constraints during upbringing, under the assumption that larger family size may dilute parental attention and financial resources, potentially affecting educational outcomes. Urban or rural residence is included to account for geographic variation in access to both educational institutions and employment opportunities, with urban areas typically offering greater proximity to colleges and more robust labor markets.

We also control for **academic performance and ability** using **Grade Point Average (GPA)** and the **PIAT (Peabody Individual Achievement Test) Math Score**. GPA serves as a cumulative measure of academic achievement and college preparedness, while the PIAT math score provides an objective, standardized measure of cognitive ability in mathematics. Both are expected to positively correlate with educational attainment and future income.

Parental education is represented by the **highest grade completed by the respondent's mother, father, and resident father.** These variables capture the educational environment in the household and serve as indicators of inter-generational transmission of human capital. Prior

research suggests that exposure to more highly educated parents is associated with increased academic support, higher aspirations, and improved educational outcomes for children.

Weeks employed in the year is the employment variable included in our project. A respondent with more weeks employed is considered to have a higher income as compared to a respondent with lesser time employed. Since **employment** is one of the major sources of income, this variable is a very important element of our project.

Finally, we include PubID as a unique respondent identifier, as well as **month and year of birth** to ensure accurate age calculation and cohort alignment. Moreover, Angrist and Krueger (1991) showed that **season of birth** is related to educational attainment because of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at an older age, and can therefore drop out after completing less schooling than individuals born near the end of the year. Thus, birth month plays a crucial role in predicting the future earnings of respondents as well. These identifiers are essential for structuring the panel data and aligning responses across survey rounds, though they are not interpreted substantively in the outcome models.

Variable	Explanation	Mode	Min	Max	St Dev	Observations	Additional Information
PubID	Respondent unique identification #	-	-	-	-	1824	-
Birth Year	Year of Birth	1982	1980	1984	1.04	1824	
Birth Month	Month of Birth[1]	9	1	12	3.43	1824	
Sex	1 = male 2 = female		-	-	-	1824	
Age	Respondent Age at Interview Date	28	25	31	1.08	1824	(2010 age)
Race	Race/ Ethnicity of Respondent[2]	-	-	-	-	1824	
Marital	Marital status[3] (2010 status)	-	-	-	-	1824	
Gross Household Income	Yearly Gross Household Income	-3?	-48100	246474	41668	1824	(1997 income)
Gross Family Income	Yearly Gross Income	290810?	330	290810	54147	1824	(2010 income)
Youth Parent Guardian	Parent or Guardian in Household (1 = yes)	1	-	-	-	1824	(1997 status)
Siblings	Number of Siblings (Biological or Adoptive)	1	0	45	2.55	1824	
Income	Total income from wages and salary (Past year)	30000	0	130254	21511	1824	(2010 income)
Household Size	Household size (including respondent)	4	2	13	1.38	1824	(1997 status)
Household Members under 18	Household Members under 18 years of age	2	1	8	1.14	1824	(1997 status)
Sample Type	1 = cross-sectional 2 = oversample	-	-	-	-	1824	(1997 status)
Highest Grade Mother	Highest Grade Completed (Mother)[4]	12	2	20	2.91	1824	(1997 status)
Highest Grade Father	Highest Grade Completed (Father)[4]	12	2	20	3.21	1824	(1997 status)
Highest Grade Residential Father	Highest Grade Completed (Residential Father)[4]	12	2	20	3.30	1824	(1997 status)
Math Piat Percentile	Math Piat Grade Percentile	18	0	100	27.5	1824	(1997 score)
GPA	Grade Point Average (Transcript)	2.98	0.1	4.11	0.706	1824	Historical
Degree	Highest Degree Achieved[5]	0	0	1	0.480	1824	(2010 status)
Urban/Rural Residence	0 = Rural 1 = Urban 2 = Unknown	1	0	2	0.537	1824	(2010 status)

Figure 1: Variable Statistics

Identifiers:

PubID: Unique Identifier Number within the survey.

Year: Year of Birth

Month: Month of Birth

Notes:

Month of Birth

1 = January

2 = February

3 = March

4 = April

5 = May

6 = June

7 = July

8 = August

9 = September

10 = October

11 = November

12 = December

Race/Ethnicity

1 = Black

2 = Hispanic

3 = Mixed Race (Non-Hispanic)

4 = Non-Black Non-Hispanic

Marital Status

0 = Never Married, not Cohabitating

1 = Never Married, Cohabitating

2 = Married

3 = Legally Separated

4 = Divorced

5 = Widowed

Highest Grade Completed

0 = None

1 = 1st Grade

2 = 2nd Grade

3 = 3rd Grade

4 = 4th Grade

5 = 5th Grade

6 = 6th Grade

7 = 7th Grade

8 = 8th Grade

9 = 9th Grade

10 = 10th Grade

11 = 11th Grade

12 = 12th Grade

13 = 1st Year College

14 = 2nd Year College

15 = 3rd Year College

16 = 4th Year College

17 = 5th Year College

18 = 6th Year College

19 = 7th Year College

20 = 8th Year College

Highest Grade Achieved

0 = No Degree

1 = GED

2 = High-School Diploma

3 = Associate's Degree

4 = Bachelor's Degree

5 = Master's Degree

6 = PhD

7 = Professional Degree (MD, DDS, JD)

Additional Key Observations

-1 = Refusal to Answer

-2 = Do not Know

-3 = Invalid Skip

-4 = Valid Skip -5 = Non Interview

2 Methodology

The central focus of our study is to analyze the impact of college education on future earnings, while controlling for a vector of high-dimensional confounders. We plan to accomplish this using Double Debiased Machine Learning (DDML) to control for the confounders.

Our data contains quantitative variables from the National Longitudinal Survey of Youth 1997. The research mostly called for quantitative data for the purpose of building more accurate causal connections with variables that are more reliably measured. For the purpose of analyzing the causal connection between College Education and Future Earnings, cross-sectional data from the year 2010 was chosen, given that the survey started in 1997, choosing 2010 allowed the participants to settle down into their careers to allow for the full realization of the value of education.

Prior to running DDML, a naivé estimate was obtained with the equation: $Income_i = \beta_0 + Degree^*\beta_1 + \epsilon_i$, where Income represents our future earnings outcome variable, Degree is our binary variable for whether or not a participant attended college and epsilon is our error term, which represents all of the variables that we have not accounted for. More flexible OLS methods were also used to obtain estimates with the equations $Income_i = \beta_0 + Degree^*\beta_1 + X_i + \epsilon_i$, where X_i is a vector of confounders including family income, GPA (Grade Point Average), age, residential and biological fathers' highest grade, biological mothers' highest grade, number of siblings, sex, gross family income, the Math PIAT (Peabody Individual Achievement Test), month and year of birth, household size and occupants under 18. The final OLS method used before DDML followed the equation:

$$\begin{aligned} Income_i = \beta_0 &+ \beta_1 \cdot Degree_i + \beta_2 \cdot FamilyIncome_i + \beta_3 \cdot GPA_i + \beta_4 \cdot Age_i \\ &+ \beta_5 \cdot ResGrade_i + \beta_6 \cdot FathersGrade_i + \beta_7 \cdot MothersGrade_i + \beta_8 \cdot Siblings_i \\ &+ \beta_9 \cdot Sex_i + \beta_{10} \cdot GrossIncome_i + \beta_{11} \cdot PIAT_i + \beta_{12} \cdot Month_i \\ &+ \beta_{13} \cdot Year_i + \beta_{14} \cdot HouseholdSize_i + \beta_{15} \cdot Under18_i \\ &+ \beta_{16} \cdot WeeksWorked_i + \varepsilon_i \end{aligned}$$

Polynomials were placed on all the variables and an 8 break, 2 degree spline on family income because we believed a strong causal relationship between this variable and others such as GPA, degree and could even be a potential explainer for income. The same rationale was applied to the rest of the variables, as intuition leads us to believe that a lot of the variables have strong causal relationships with each other. Doing this also allows us to capture any non-linear relationships that we hypothesize to exist between variables. For instance, the family income variable represents how much the participants parents earned when they still shared the same roof, and while a positive relationship between this and future earnings was hypothesized, we also expected diminishing returns in future earnings with an increase in family income, because moving from low to higher income brings a significantly increased level of access to better education, extracurriculars, etc, while at high income these are already present, leading to smaller increasing increments of future earnings when family income increased. This was our rationale behind breaking family income up so heavily, to explore how far the relationship goes.

For our purposes, we have decided to use two DDML methods; DDML linear regression with a LASSO constraint and DDML random forest to check them against each other. For LASSO, we will be using the DDML2 procedure, selecting the optimal penalty in each stage using 10-fold cross validation, and splitting the sample 5 times while running DDML. For Random Forest, we use a random forest with 500 trees, a maximal tree depth of 7, minimal node size of 5 and with 4 variables considered at each split. The driving principle behind DDML is Neyman Orthogonality, which in summary ensures that our estimator is robust and our estimator is independent of any influence from our vector of confounders or nuisance parameters. It does so by setting a requirement that reduces the influence of confounders on treatment and outcome. The general formula being used here follows; $earnings = Degree + g(Xi) + errorterm$ where $g(X)$ is our confounder vector. When working with DDML linear regression, it is better summarized and understood through an algorithm, which follows;

1. Split sample using cross validation into training and test
2. Using Supervised Machine Learning (SML) methods, both LASSO and Random Forest separately, run Future Earnings/Outcome(Yi) on Xi and obtain residuals, labelling h(.)
3. Using the same SML methods, run Degree/Treatment(Di) on Xi and obtain residuals, labelling m(.)
4. Residualize to obtain V_i^Y by computing $Y_i - h(X)$ and V_i^D by computing $D - m(X)$

5. Regress V_i^Y on V_i^D using OLS and obtain double de-biased estimator, labelling θ.

STEP 1: The dataset is split to ensure that our chosen SML methods do not overfit. If the same dataset is used for both training and obtaining residuals, then the residuals become biased especially in the presence of high dimensionatl confounders.

Step 2 and 3: The formulae used for DDML LASSO and Random Forest are as follows;

LASSO: $\beta = argmin(1/n^* \sum_{i=1}^{n}(A_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^{j}|\beta_j|)$ where Ai is either Y_i or D_i depending on what stage of DDML is currently underway and,

RF: $1/B^* \sum_{i=1}^{B} Y_t$ where B is our number of trees. A random forest is done by drawing random samples with replacement from our dataset, training a decision tree on a random sample, then averaging all the predictions from all the decision trees.

Step 4: Partialling Out is the process which removes the influence of the confounders from the Treatment and Outcome respectively by subtracting the vector of confounders from the treatment and outcome respectively. This process is also known as residualizing.

STEP 5: We use the residuals V_i^D and V_i^Y, both free from the influence of confounders and regress V_i^D on V_i^Y using OLS to get estimator θ.

Directed Acyclic Graphs (DAGs)

Understanding the causal relationship between education and income is a fundamental question in economics. Traditional regression models often fail to account for complex interactions between variables, leading to biased estimates. To address these challenges, we employ Directed Acyclic Graphs (DAGs) to visually represent assumptions about causal relationships and use Double De-biased Machine Learning (DDML) to estimate causal effects in a data-driven manner.

In this study, we construct 4-5 DAGs to explore different causal structures affecting income, focusing on the role of education as a treatment variable. Each DAG incorporates relevant socioeconomic, demographic, and academic factors that may influence income, allowing us to systematically test different model specifications. In order to make sure we maximize the efficiecy of this method, each teammate came up with 1-2 DAGs of their own, so that in case one of us missed out an important causal connection, it will most likely be covered by our other teammates. By applying DDML, we mitigate biases arising from high-dimensional confounding and improve the robustness

of our estimates.

Our approach allows us to compare how different model assumptions impact the estimated returns to education. The insights from these analyses will contribute to the broader discussion on education policy and labor market outcomes. The following sections outline our methodology, the structure of our DAGs, and the empirical findings from our DDML estimations.

DAG 1

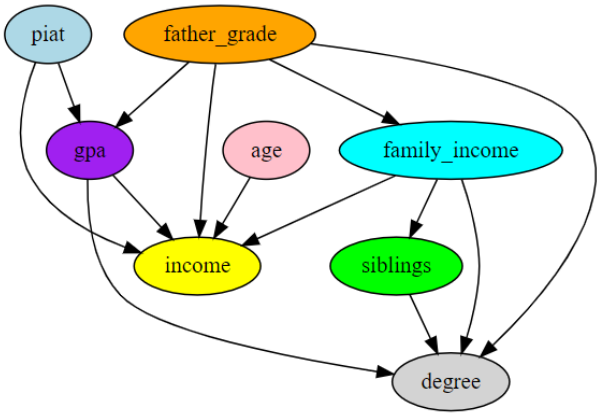

Figure 2: DAG 1

This DAG considers the causal effect of a college degree on earnings, with some of the core control variables. Setting the degree variable as the exposure variable and income as outcome variable, we find that the adjustment sets for this DAG are (family income, father grade, gpa, piat) and (father grade, gpa, piat, siblings).

We then run a DDML Regression using LASSO, and the results are as follows:

Coefficient	**Estimate**	**Std. Error**	**T Value**	**P Value**	**Significance**
Degree	5375	1031	5.214	0.000000185	Highly Significant

Table 1: Statistical Analysis Results

DAG 2

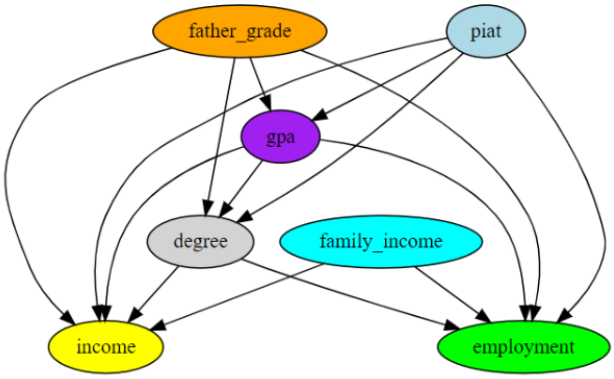

Figure 3: DAG 2

In this DAG, we include employment as one of the core variables. As you can see from the graph, employment is a key variable affected by many of the other variables. After runnning a DDML Regression using LASSO, and the results are as follows:

Coefficient	**Estimate**	**Std. Error**	**T Value**	**P Value**	**Significance**
Degree	4979	1022	4.874	0.00000109	Highly Significant

Table 2: Regression Analysis Results

Although the estimate is higher than the previous DAG and is also highly significant, the previous DAG had a smaller p value, suggesting a relatively more significant regression.

DAG 3

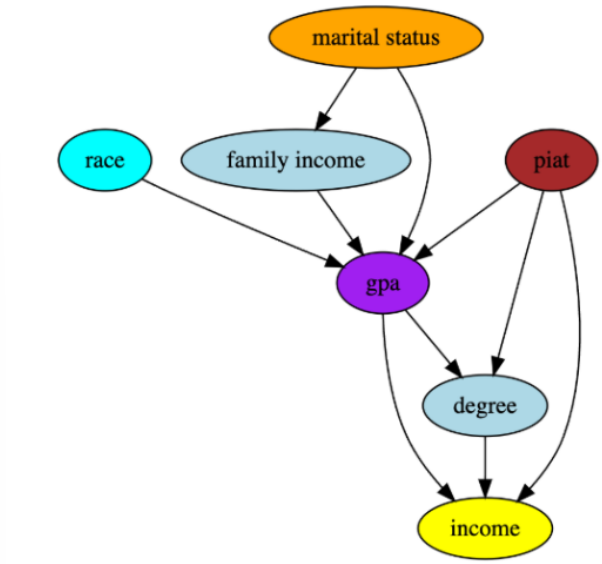

Figure 4: DAG 3

With marital status acting as an indicator of a stable household, the participants grade(GPA) would benefit from a positive household environment, and family income can be better used directly in the participants education rather than to other issues such as logistics, divorces etc.

Race has an influence on grades due to the fact that some cultures place heavy value on academic success.

The PIAT in this sense is a measure of ability, and people with higher ability tend to have more technical options available to them, an argument can be made that they work hard so they can have these options available to them. Higher ability in itself can also be an explainer of high income regardless of college attendance, and can be an explainer for higher grades.

Families that earn more are able to provide either children with greater resources such as tutors, better schooling, extracurriculars etc, which would have a positive relationship with higher grades. Because the PIAT and GPA explain both treatment and outcome, they serve as backdoor paths, which offer non causal explanations for the relationship between treatment and outcome, and therefore have to be controlled for.

After running a DDML LASSO Regression, the results are as follows:

Coefficient	Estimate	Std. Error	T Value	P Value	Significance
Degree	5808	1138	5.105	0.000000331	Highly Significant

Table 3: Statistical Analysis Results

DAG 4

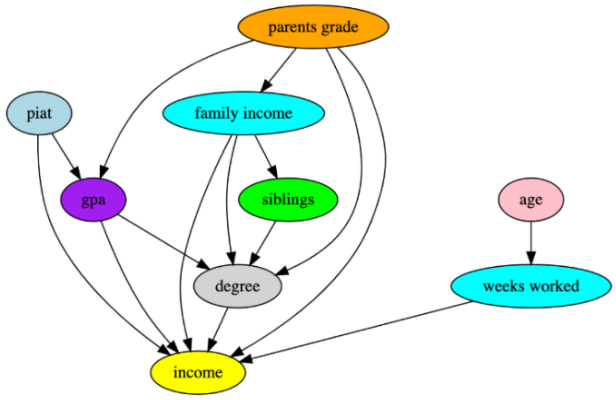

Figure 5: DAG 4

The older someone is, the more they tend to have worked, and with this comes experience, which could have a possible explanatory relationship with income.

Families that earn more can afford more children, which in turn dictates how their resources are distributed amongst their children, leading to either higher or lower grades depending on how these resources are distributed.

Parents who have done through college education have realized its value and are therefore more likely to influence their children to do the same a, therefore serving as a possible positive explainer of their children going to college. It could also serve as an explanation of higher income due to them having resources outside of money, such as connections that could provide their children with jobs, leading to higher paying jobs, or a head start along that path.

Family income and the parents' grades influence both treatment and outcome in non causal ways, therefore they must be controlled for.

Finally, running a DDML on this DAG, we get:

Coefficient	**Estimate**	**Std. Error**	**T Value**	**P Value**	**Significance**
Degree	4832	1026	4.711	0.00000246	Highly Significant

Table 4: Statistical Analysis Results

Results

Variable: Degree:- binary variable for college attendance

Method	Coefficient (2010 USD)	P-value	Statistical Significance
Naïve OLS	12113.4	¡2e-16	Yes
Controlled OLS	5715	1.14e-07	Yes
Flexible OLS (Polynomials)	4206	7.78e-05	Yes
DDML LASSO (All Ages)	4688	1.57e-05	Yes
DDML Random Forest (All ages)	5151	2.35e-06	Yes
DDML LASSO Ages 25 to 26	4184	0.0315	Yes
DDML Random Forest Ages 25 to 26	4855	0.01	Yes
DDML LASSO Ages 27 to 28	4820	0.00064	Yes
DDML Random Forest Ages 27 to 28	4796	0.000932	Yes
DDML LASSO Ages 29 to 31	8034	0.0336	Yes
DDML Random Forest Ages 29 to 31	12214	0.00267	Yes

Table 5: Statistical Analysis Results (OLS And DDML Methods by age group)

Our results show that our naive OLS model grossly overestimated the effects of college education on income, but as more complex models are used, the true treatment is shown to be a more conservative figure. Our flexible OLS gives us a lower estimate than both DDML methods, however, there seems to be a trade of for statistical significance, so it is possible that the true effect of treatment is between $4206 and $5151. These results are in line with our expectations, as well as being backed by the literature.

In order to study the effects of college on college attendance, we decided to group the participants into 3 bins, ages 25 to 26 had 568 participants, ages 27 to 28 had 1061 observations, and ages 29 to 31 had 195 observations. From our data, we see that our highest age group had a significantly higher treatment effect estimate in both DDML LASSO and Random Forest SML methods. This supports our decision to use the year 2010 for our cross-sectional analysis; to give the participants time to realize the full value of their education.

Variable: Age

Method	Coefficient	P-value	Statistical Significance
Naive OLS Age as treatment	995.5	0.0329	Almost insignificant
Naive OLS DegreeXAge	2158.9	0.0255	Almost insignificant
Controlled OLS DegreeXAge	2145	0.011793	Yes
Controlled OLS	290.7	0.835409	No
Flexible OLS Band 1	1872	0.975517	No
Flexible OLS Band 2	-14,620	0.409074	No

Table 6: Summary of OLS Methods and Results

To examine how age and college attendance interact, that is, to see if age and college attendance go hand in hand as explainers of future earnings to any degree, a comparison to age as the treatment was made. Our coefficients show that while age has both positive effects on college and non college goers, the effect is more pronounced with those who did attend college, although out P-values show little statistical significance and the lack of controls on both models show that these effects are more than likely overestimated. The coefficients of the flexible OLS suggest a concave shape, implying income increases with age but plateaus then falls sharply at some point, from a more conservative $1872 at first and the sharp fall by $14620, implying diminishing returns over age. This would be plausible is the maximum age in our dataset was not 30 but closer to retirement age. These results are, however, not statistically significant.

The following graph shows the average income of the different subgroups included in our dataset:

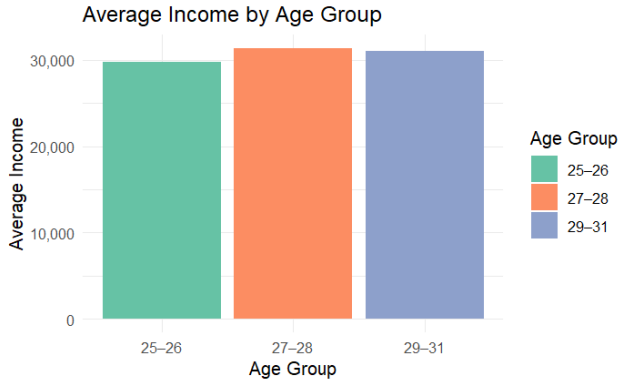

Figure 6: Average Income by Age Groups: Average age group of Ages 29 to 31 who had the highest estimated treatment effect have seen a normalization to average incomes matching the other age groups, suggesting the presence of outliers in that age group.

Variable: Weeks Worked (measure of experience)

Method	**Coefficient (weeks)**	**P-value**	**Significance**
Naive OLS (Weeks Worked as treatment)	225.44	6.77e-12	Yes
Naive OLS (DegreeXWeeks worked)	18.02	0.8133	No
Controlled OLS	1.224e+02	1.72e-05	Yes
Controlled OLS (DegreeXWeeks worked)	2.467e+01	0.714767	No
Flexible OLS Band 1	122.4	0.000848	Yes
Band 2	123.700	1.04e-12	Yes
Band 3	14.910	0.385170	No
Band 4	15.720	0.362579	No

Table 7: Summary of OLS Methods and Results

Our results show that weeks worked has a positive, albeit small, interaction with all participants regardless of college attendance, but this effect is significantly reduced with those who attended college, but as similarly to our comparison in the degree age interaction, the lack of controls points towards an overestimation of the the effect of weeks worked. Although the interaction term gives a larger estimate in the presence of controls, the coefficient does not show much of an effect as well. The flexible OLS coefficients showed a positive, although diminishing, relationship between the weeks worked and future earnings.

Variable: Family Income

Method	**Coefficient**	**P-value**	**Significance**
Naive OLS (Family Income as a control)	1.678e-01	$< 2e-16$	Yes
Naive OLS (FamilyIncomeXdegree)	-7.203e-03	0.668	No
Controlled OLS	1.400e-01	$< 2e-16$	Yes
Controlled OLS (FamilyIncomeXdegree)	-2.187e-04	0.989193	No
(Flexible OLS)			
Band 1	4152	0.45604	No
Band 2	21,740	9.24e-10	Yes
Band 3	22,730	4.75e-08	Yes
Band 4	27,300	2.49e-12	Yes
Band 5	30,870	6.03e-15	Yes
Band 6	37,350	$< 2e-16$	Yes
Band 7	29,670	9.23e-08	Yes
Band 8	45,210	$< 2e-16$	Yes

Table 8: Summary of OLS Analysis

After applying family income as treatment and interacting it with college attendance in both our naive and controlled models, we can see that family income has a small but statically significant effect on future earnings when not interacted with college attendance. When interacted with college

attendance, it is inconclusive to say that family income has any influence on future earnings when participants went to college.

Our flexible OLS coefficients prove our earlier hypotheses concerning diminishing marginal returns of future earnings with increased family income. Studying the figures shows a non linear increase in the estimated effects, as the $17,588 jump from band 1 to band 2 is the largest jump, with estimated effects decreasing before a dip in band 7 and a jump in band 8.

3 Conclusion

In conclusion, we analyzed the causal relationship between college education and future earnings using Double Debiased Machine Learning (DDML) with LASSO and Random Forest methods. Our results estimate an increase between $4,206 and $5,151 in income, aligning with existing literature that indicates college education positively affects earnings. Initially, simpler models suggested a much larger impact of education, highlighting the importance of appropriately controlling for key confounders. By adjusting our models to account for critical variables such as family income, GPA, age, and parental education, we were able to obtain more accurate and realistic estimates. We first utilized controlled Ordinary Least Squares (OLS) regression to manage these confounders and later transitioned to DDML methods to further refine our estimates.

Further analysis demonstrated that age held limited explanatory power within our dataset, likely due to its narrow age range. Conversely, experience, measured through weeks worked, consistently showed a significant and positive influence on earnings, highlighting the importance of work experience irrespective of educational attainment. Additionally, by employing polynomial terms and spline functions, we effectively identified and captured complex, non-linear relationships. A prominent example was the diminishing returns observed with increasing family income, suggesting that initial increments in family income have a substantial impact on future earnings potential, whereas additional increases at higher income levels produce progressively smaller impacts. These findings align closely with expectations derived from prior studies, reinforcing the validity of our methodological approach and analytical strategies. Future analysis would benefit from the use of panel data to study trends in the impact and significance of college education in more modern times.

References

Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings?*. *The Quarterly Journal of Economics*, 106(4):979–1014, 11 1991. ISSN 0033-5533. doi: 10.2307/2937954. URL https://doi.org/10.2307/2937954.

Gary S. Becker. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition*. Number beck-5 in NBER Books. National Bureau of Economic Research, Inc, 1964. URL https://ideas.repec.org/b/nbr/nberbk/beck-5.html.

Kristin F. Butcher and Anne Case. The effect of sibling sex composition on women's education and earnings*. *The Quarterly Journal of Economics*, 109(3):531–563, 08 1994. ISSN 0033-5533. doi: 10.2307/2118413. URL https://doi.org/10.2307/2118413.

David Card. Chapter 30 - the causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pages 1801–1863. Elsevier, 1999. doi: https://doi.org/10.1016/S1573-4463(99)03011-4. URL https://www.sciencedirect.com/science/article/pii/S1573446399030114.

Zvi Griliches. Estimating the returns to schooling: Some econometric problems. *Econometrica*, 45 (1):1–22, 1977. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913285.

James J. Heckman, John Eric Humphries, and Gregory Veramendi. Returns to education: The causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126 (S1):S197–S246, 2018. doi: 10.1086/698760. URL https://doi.org/10.1086/698760.

Jacob Mincer. Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy*, 66(4):281–281, 1958. doi: 10.1086/258055. URL https://ideas.repec.org/a/ucp/jpolec/v66y1958p281.html.