Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1473

Returns to Education in the United States: A Comparison of OLS and Double Machine Learning Methods

Al Mansor Helal
University of Arkansas

Ryotaro Hiraki
University of Arkansas

Harry Anthony Patrinos
University of Arkansas

This study examines the economic returns to education in the U.S. using 2024 CPS data and compares Ordinary Least Squares (OLS) regression with a Double Machine Learning (DML) framework incorporating models such as random forests, boosted trees, lasso, GAMs, and neural networks (MLP). Results show consistent returns of 8 to 9 percent per additional year of schooling across methods. Simulations reveal that all predictors perform well under linear assumptions if hyperparameters are optimally adjusted, while OLS/Lasso suffer from nonlinearity. Findings suggest that OLS remains robust in low-dimensional, near-linear contexts, offering practical guidance for economists and policymakers balancing model complexity and interpretability in education research.

VERSION: May 2026

Suggested citation: Helal, Al Mansor, Ryotaro Hiraki, and Harry Anthony Patrinos. (2026). Returns to Education in the United States: A Comparison of OLS and Double Machine Learning Methods. (EdWorkingPaper: 26-1473). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/tmac-c636

Returns to Education in the United States: A Comparison of OLS and Double Machine Learning Methods

Al Mansor Helal

University of Arkansas

Ryotaro Hiraki

University of Arkansas

Harry Anthony Patrinos

University of Arkansas

APRIL 2026

ABSTRACT

Returns to Education in the United States: A Comparison of OLS and Double Machine Learning Methods

This study examines the economic returns to education in the U.S. using 2024 CPS data and compares Ordinary Least Squares (OLS) regression with a Double Machine Learning (DML) framework incorporating models such as random forests, boosted trees, lasso, GAMs, and neural networks (MLP). Results show consistent returns of 8 to 9 percent per additional year of schooling across methods. Simulations reveal that all predictors perform well under linear assumptions if hyperparameters are optimally adjusted, while OLS/Lasso suffer from nonlinearity. Findings suggest that OLS remains robust in low-dimensional, near-linear contexts, offering practical guidance for economists and policymakers balancing model complexity and interpretability in education research.

JEL Classification: I20, J31, J24, D62, O15
Keywords: Returns to education; Machine learning

Corresponding author:
Harry Patrinos
University of Arkansas
Fayetteville, AR 72701
USA
E-mail: patrinos@uark.edu

1. Introduction

Understanding the returns to education, that is, how much additional income individuals earn from spending another year in school, is a foundational question in the economics of education. A robust estimate of this return informs both individual educational decisions and public policy. Traditionally, the research has been conducted using Ordinary Least Squares (OLS) regression (Harmon et al. 2003; Psacharopoulos and Patrinos 2018), often based on the Mincer earnings function (Mincer 1958, 1974), to estimate the effect of education on earnings. However, the validity of OLS hinges on stringent assumptions about functional form and the absence of omitted variable bias (Buscha and Dickson 2023; Card 1999).

In response to these challenges, researchers have developed quasi-experimental designs, such as Instrumental Variables (IV) (Angrist and Imbens 1996; Card 1999) and Regression Discontinuity (RD) designs (Thistlethwaite and Campbell 1960), to more reliably estimate causal effects. More recently, advances in machine learning have introduced Double Machine Learning (DML) as a flexible tool for estimating treatment effects in high-dimensional and complicated contexts. DML attempts to address biases in OLS by modeling and removing the influence of control variables using machine learning methods such as Random Forests, Boosted Trees, and Neural Networks (Multi-Layer Perceptron, MLP) (Gardner and Dorling 1998).

If machine learning outperforms conventional parametric methods, then why is its use not more widespread? The key reason is the different goals between the two: traditional methods aim to draw a causal inference while the goal of machine learning is to make a prediction. Indeed, a machine learning model itself has been introduced into the field of education for the purpose of prediction. The review from Luan and Tsai (2021) identified 40 papers that leverage a machine learning model to improve precision in education research, but most of the studies focus on the prediction of learning outcomes such as performance, dropout,

and retention. However, DML enables researchers to deliver the advantages of machine learning to causal inference.

DML has been implemented by researchers in diverse disciplines. Chernozhukov et al. (2018) outlined a method for applying DML in empirical research in econometrics. It has since been applied not only to economics and econometrics, but also to healthcare, sociology, and politics. Fuhr et al. (2024) identified 46 published papers using DML. Despite its growing popularity in economics, DML has thus far been employed in a limited number of studies in education, and few of which tuned the parameters and number of folds and repetitions accordingly (Fuhr et al. 2024).

The DML model is particularly well suited to high-dimensional settings, where a large number of covariates are included and effective variable selection (i.e., identifying those variables that are systematically related to the outcome) is required. It alleviates the assumptions about functional form of covariates and model specification. Even when all confounding variables are observed, using an incorrect functional form to include them in a parametric regression model can result in biased estimates of the treatment effect (Wooldridge 2012). Chan and Meunier (2022) implemented DML to estimate the effect of technological intensity on support for the European Union's Foreign Direct Investment screening mechanisms with 29 variables for 28 EU member countries. Even with such a case where the number of features (variables) exceeds the number of sampled units, DML can still provide estimates with the use of non-parametric machine learning models such as random forest or boosted trees. Random forest is a machine learning algorithm that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting risks by growing diverse trees that are independent of each other. This method is used in classification or regression tasks. Boosted trees is another tree-based method that builds decision trees

sequentially, where each new tree focuses on correcting the errors of the previous ones to improve prediction accuracy.

This paper aims to contribute to this methodological discussion by comparing OLS and DML estimates which entail five different machine learning (ML) algorithms (random forest, boosted trees, lasso, generalized additive models, and neural nets (MLP)) for a relatively simple setting. Specifically, the analysis is based on the Mincerian equation used to estimate the returns to education using 2024 CPS data for the United States (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2024). It will also use adjusted DML parameters (number of folds in cross-fitting (K) and number of repetitions of the framework (S)) as recommended by Chernozhukov et al. (2018). Although the setting in education is usually able to obtain a large sample, it still suffers from nonlinearity or model selection. The goal is to evaluate whether DML offers an advantage over OLS in the context of a relatively homogeneous labor market and whether machine learning improves causal inference in this setting. We estimate the return to education not only with real world data, but also with simulated demographic data to evaluate each DML prediction model as well as comparison with the performance of classic OLS.

DML holds promises for advancing the field of the economics of education by enabling a less biased estimate of complex relationships, potentially nonlinear relationships between educational attainment and wages. Evaluating the performance of DML in this context is therefore critical not only to improve the estimation methodology but also to inform policy decisions with reliable evidence. This study contributes to the literature by assessing whether DML yields more accurate and more robust estimates than conventional approaches (OLS) in estimating returns to education as it does not rely on parametric functional form assumptions.

To preview the results, there is no significant difference between OLS and DML models observed with CPS data, both of which indicate an 8 to 9 percent rate of return from an additional year of schooling. The implementation of the models with simulated data, whose

effect on education is linear, reveals that linear based models such as OLS fit well as it is based on linear assumption. Also, DML with flexible methods such as Generalized additive models (GAMs), Neural Network and the tree-based ensemble models – random forest and boosted trees – performed well. However, Lasso suffered from its own regularization (i.e., decrease variance to avoid overfitting) as shrinkage (i.e., shrink coefficients towards zero) of variables were not truly effective in our setting. As model complexity increased, linear-based method such as OLS and DML with Lasso could not perform the prediction task while other flexible models, especially Neural Networks, exhibited a good prediction power.

2. Literature Review

Research over the past few decades has found positive returns to education. A seminal review by Card (1999) found returns ranging from 6 to 15 percent per additional year of schooling. More recent evidence by Deming (2022) and Oreopoulos (2006) emphasizes this finding while suggesting the heterogeneity of returns depends on methodology and population.

A particularly important distinction is that quasi-experimental methods such as IV and RD often yield higher estimates than OLS, suggesting that OLS may underestimate returns, particularly at the margin (Patrinos and Psacharopoulos 2026). Quasi-experimental research shows that policy-driven natural experiments often produce more robust causal estimates than purely observational approaches (Angrist and Krueger 1991; Card 2001; Duflo 2001; Patrinos and Sakellariou 2005). Nevertheless, the returns estimated with OLS and causal methods are very similar.

In many studies, researchers rely on observational data because experimental interventions are often infeasible, unethical, or prohibitively expensive (Athey and Imbens 2017). However, it is essential to make some assumptions without the experimental variations. For instance, a common assumption made is that all variables that affect both treatment and outcome variables – so-called confounders – are observed and adequately adjusted for (Imbens

2004). In real-world applications of causal inference, one of the primary obstacles is providing justification for the required assumptions (Hernan and Robins 2020). For this reason, researchers seek to develop and implement a method that can rely on weaker and more realistic assumptions.

Recent research has shown that machine learning (ML) methods can allow us to relax functional form assumptions. While causal assumptions remain the same (e.g., unconfoundedness, no omitted confounders), ML methods bring about flexibility in modeling complex relationships between covariates, treatment, and outcomes. Supervised ML (ML methods with an associated response y for each predictor x), which is typically designed to make accurate predictions in complex or nonlinear settings, excels in handling high-dimensional datasets—those in which the number of variables may exceed the number of observations (Hastie et al. 2005). However, the primary goal of ML is prediction, not causal inference, which makes its application to causal analysis nontrivial (Shmueli 2010). Still, as Mullainathan and Spiess (2017) argue, machine learning can play a useful role in causal inference by improving the estimation stages that underlie many causal methods. They claim that although ML is utilized to obtain $\hat{y}$ (predict outcome with X), it can be applied if the task is to estimate β when estimating the first stage of linear instrumental variable regression (this is effectively predicting X with Z), estimating heterogeneous treatment effects, and flexibly controlling for observed confounders. For example, Dube et al. (2022) employed ML methods to predict minimum wage workers with their demographic information to estimate the effect of minimum wage increase, then they incorporated difference-in-difference and two-way fixed effects approaches to draw a causal inference.

Unlike traditional parametric models like OLS, which assume a linear relationship between predictors and the outcome, ML algorithms – such as random forests or gradient-boosted trees – can automatically capture nonlinearities, interactions, and discontinuities

without the need for manual specification. For example, whereas a linear regression assumes a functional form:

$$\hat{Y} = \hat{f}(X) + \varepsilon,$$

where Y is outcome and X is treatment variable with $\hat{f}$ constrained to be linear, ML approaches allow $\hat{f}$ to be learned directly from the data, offering greater flexibility (James et al. 2023).

This flexibility enables ML to capture complex, nonlinear relationships between inputs and outcomes that traditional linear models may miss. However, this leads to less interpretability. For a linear model, it is relatively easy to understand the relationship between response Y and predictors $X_1, X_2, X_3 \ldots X_p$, though for a highly flexible ML methods such as random forest or neural nets, their complicated estimates of f do not let us understand the association of each individual predictor with the response (James et al. 2023). Also, increased flexibility comes at the cost of increased model complexity, potential overfitting, and reduced interpretability. The flexibility in nonparametric methods in ML typically fits every observed data point, including errors or noise, perfectly, which leads to such a case where a model performs well with the training data but poorly in new samples, described as a biased estimate.

Researchers have developed a method called regularization to prevent overfitting the data by reducing the variance, but this effort creates regularization bias (Chernozhukov et al., 2018). Because ML methods avoid perfectly fitting the data so that each trained model looks somewhat similar, there is bias within each model as a result. For example, Lasso (least absolute shrinkage and selection operator) regression, introduced by Tibshirani (1996), employs a restriction to the Residual Sum of Square (RSS) minimization:

$$(\hat{\alpha}, \hat{\beta}) = argmin\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\} \qquad subject\ to \sum_j |\beta_j| \leq t$$

where $t \geq 0$ is the tuning parameter, $\hat{\alpha}, \hat{\beta}$ are lasso estimates, and $(x_i, y_i), i = 1,2, \ldots, N$. Or minimizing

$$\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2 + \lambda\sum_j |\beta_j| = RSS + \lambda\sum_j |\beta_j|$$

where λ is the tuning parameter (James et al. 2023). The L1 penalty, which is absolute value of the coefficient added for regularization, denoted as $\sum_j |\beta_j|$ prevents the coefficients to take an extreme value and shrinks them towards zero. In Lasso, when *t* is sufficiently small or λ is sufficiently large, it forces some coefficients to exactly equal to zero (Tibshirani 1996). Other examples of regularization are early stopping, weight decay, and dropout in neural network ML methods (Hestie et al. 2009).

Tree methods are an example of non-parametric machine learning algorithms. Decision Trees, which are applied to random forest or boosted trees, divide the data into subsets by applying a series of decision rules based on the input variables. At each node, the algorithm selects the variable and corresponding split point that minimizes the prediction error. This process continues until terminal nodes, or *leaves*, are formed. The predicted value for each observation is then given by the average outcome of the training samples within the corresponding leaf. Random forest and boosted trees create or use multiple trees from the same sample in different ways to obtain an accurate prediction (Athey and Imbens 2019).

In this paper, we examine the Double/Debiased Machine Learning (DML) framework. The *double* comes from its procedure that estimates the treatment and outcome individually with ML methods (double ML). This method is developed from the *double selection* procedure proposed by Belloni et al. (2014) which selects covariate predictors for outcome and for treatment with Lasso, then uses OLS regression with all covariates selected with Lasso. DML generalizes this idea and implemented sample splitting, which enables application to other modern ML methods not limited to Lasso (Fuhr et al. 2024). The fundamental promise of DML is its ability to leverage flexible machine learning methods to effectively adjust for observed confounders. This approach enables researchers to obtain unbiased estimates even in the

presence of numerous confounders and complex relationships. By employing DML with nonparametric machine learning models, rather than relying on a fixed parametric specification, researchers can relax assumptions regarding both the selection of variables and the functional forms used to account for confounding (Fuhr et al. 2024), which is important because the ability to flexibly adjust for a large number of covariates enhances the plausibility that all relevant confounding variation has been accounted for (Chan and Meunier 2022; Qiu et al. 2022; McConnell and Lindner 2019).

Chernozhukov et al. (2018) are one of the outstanding examples which used DML for causal inference. They established the method to implement machine learning in econometrics. The promise of DML is to overcome regularization bias and overfitting with the use of Neyman orthogonality and cross-fitting. First, Neyman orthogonality handles regularization bias, as explained by Chernozhukov et al. (2018), W is sample, θ_0 is a parameter of interest, $\eta_0 = (g_0, m_0)$ is nuisance parameter where $g_0 = E[Y|X]$ and $m_0 = E[D|X]$, and the moment condition is $E[\psi(W; \theta_0, \eta_0)] = 0$. Neyman orthogonality stands when the directional derivative with respect to the estimation error of η is 0. This error is the bias made by regularization of machine learning methods. Therefore, Neyman orthogonality allows marginal error of estimation of $E[Y|X]$ and $E[D|X]$ in DML. Second, cross-fitting is used to handle overfitting. The number of folds represents the number of sets the data is divided to train the data, which is called cross-fitting. In cross-fitting, the averaged coefficients from $K - 1$ number of folds are used to train the model and estimate the coefficient in the Kth fold. By splitting into K number of folds and averaging the effects, a substantial bias from overfitting can be recovered (Chernozhukov et al. 2018).

Two parameters in DML to consider are the number of folds in cross-fitting, *K,* and the number of repetitions, *S.* The number of repetitions, *S*, is how many times you train the data and report the median of the repeated estimates, which is, how many times an entire algorithm

is repeated. Chernozhukov et al. (2018) recommend $K = 4$ or 5 and $S = 100$ whereas much research is done by $K = 10$ and $S = 1$ since it is the Stata default (Fuhr et al. 2024).

The DML method is employed in education research, but only a few studies exist. McNamara (2020) estimated the effects of attending higher education on the labor market outcomes for dropouts. Female dropouts have a higher occupational status than those who never participated in higher education, but do not experience a wage premium, while male dropouts do experience a wage premium. The results are similar to OLS estimates in both low and high-dimensional setting. Osman and Rubb (2025) applied DML to confirm the robustness of their estimates of the returns to online education for nurses with linear regression. They find a wage penalty for online education and the OLS estimates are confirmed with DML. Interestingly, DML has also revealed counterintuitive patterns, such as wage premiums among certain groups of dropouts. Also, Spano (2024) implemented the DML approach to survey the effect of change in family structure to educational performance, showing facing family structure change between age of 9 and 13 has harmful impact on education outcome at the age of 20.

This study builds on this body of literature by implementing both OLS and DML approaches on recent labor and education data from the U.S., offering insights into the applicability of machine learning methods to the study of the returns to education. Most papers do not implement DML in the way recommended by Chernozhukov et al. (2018), which suggests 4-5 folds (K = 4 or 5), and repetitions S (they use S = 100) to obtain median estimates (Fuhr et al. 2024). According to Fuhr et al. (2024), among 46 published papers that implemented DML, 15 papers used K (<4) and two employed K (=1) which means that they did not split the sample for cross-fitting. For the number of repetition S, only 22 percent of papers repeated the algorithm more than once. Thus, we report a considerably reliable result

to estimate a return to education using the Mincerian equation by implementing DML with K = 5 and S = 100, with real-world data.

3. Data and Methods

Data and Descriptive Statistics

The dataset used in this analysis comes from the 2024 Current Population Survey (CPS), a household survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics. The data was filtered to include only individuals residing in the United States. Key variables include wage: hourly wage (calculated from reported weekly earnings and hours worked); education: years of completed schooling; demographics: age, gender, race (Hispanic, Asian, and African American), and marital status; and experience: proxy variables such as age minus years of education minus 6.

Table 1: Descriptive Statistics

Variable	Mean	Std.dev.
Education	14.31	2.8
Hourly Wage	32.85	18.7
Experience	23.09	13.0
Female	0.45	0.5
African American	0.11	0.3
Asian	0.08	0.3
Hispanic	0.21	0.4
Married	0.60	0.5
Labor Union	0.02	0.1
N	49,109	

Table 1 presents the sample description. Since the hourly wage variable in the original CPS data includes only those who are paid hourly, we created the value for most others by this formula: Annual Wage / (Hours worked per week * Weeks worked per year). In order to handle the outliers, we removed the observations whose hourly wage is more than $100 or less than $7.5, which is the federal minimum wage. We included only full-time workers defined as those working more than 35 hours a week. We excluded those who work less than 10 weeks a year to avoid seasonal workers or dependent workers. Finally, the earning weight is considered. CPS (2024). CPS survey is conducted and sample is collected in first 4 months, then it allows 8 months of interval, then collects the sample for 4 months (8 samples total). The questions regarding earnings are asked in 4^{th} and 8^{th} months. Because of this survey structure of CPS survey, it sets earning weights so that only the samples from 2 months are able to make an estimate about the entire U.S. population. Therefore, earning weight is set to 0 for most

observations in the data, which significantly reduces the sample size used for estimation from 69180 to 8145.

Methodology

To analyze the relationship between years of education and earnings, we propose following the linear regression model and double machine learning (DML) model. The OLS model follows the Mincerian equation:

$$\log Wage_i = \beta_0 + \beta_1 education_i + \gamma X_i + u_i$$

where: $logwage_i$ is the hourly wage of an individual i, $education_i$ is the treatment variable measured by years of schooling of individual i, and X_i represents the control variables which are experience and experience-squared in the parsimonious specification, but including demographic and other factors such as sex, race, marriage, unionization of i described above.

We use the heteroskedasticity-robust standard error for dealing with a potential variance difference between error terms and get a valid estimate with such a large dataset as CPS data.

One disadvantage of the OLS model is that it assumes that the relationship is linear, and there is a risk of mis-specifying a model with several variables including interactions. Thus, we propose a double machine learning model. With this method, we analyze not only a linear relationship but nonlinear ones including interactions between control variables.

Traditionally, researchers handle the nonlinearity observed in the context of returns to education such as diminishing marginal returns to education (Heckman, 2006), and signaling (Spence, 1978). We extend the OLS models to the one with education-squared variable and college dummy variable to fit forms of nonlinearity- diminishing returns and signaling, respectively.

The DML model works in the following ways. First, with all the confounders, X, the model predicts the wage and education with only the use of the control variables by a machine learning model introduced below. Second, using the predicted wage, it finds the residuals

between the predicted wage and education and the true values by simply subtracting the two. The residuals here are the pure variation that the model is able to exclude from the effect of control variables ($\widehat{V_{wage}}$ and $\widehat{V_{edu}}$). Finally, we fit the OLS between two residuals. The coefficient here means the pure effect of education on the hourly wage. The DML method allows us to make an estimate while capturing the complex nonlinear relationships and interaction effects within the variables.

We implement cross-fitting in the DML model to avoid overfitting the data which might include some outliers. Cross-fitting separates the data into a certain number of folds (K) and trains the data with all other folds to predict the other fold and finally average the prediction. By using this method, the prediction from the DML model becomes more solid and consistent without any bias. As Chernozhukov et al. (2018) recommend, the number of folds is set to K = 5, and the number of repetitions (S) is set to S = 100. Thus, the reported estimates are the median coefficient of 100 estimates with different cross-fitting folds. (Chernozhukov et al. 2018).

The DML algorithm can be summarized as follows (Fuhr et al. 2024):

1. Train two machine learning models on K-1 folds:

   (a) outcome $wage_i$, features X_i

   (b) outcome $education_i$, features X_i
2. Use the models to make predictions ($\widehat{wage}$ and $\widehat{education}$) on the held-out fold.
3. Compute the residuals as follows: $\widehat{V_{wage}} = wage - \widehat{wage}$, $\widehat{V_{edu}} = education - \widehat{education}$
4. Use a linear regression to estimate coefficient from residuals: Regress $\widehat{V_{wage}}$ on $\widehat{V_{edu}}$, obtain the coefficient on $\widehat{V_{edu}}$.
5. Repeat for all folds (K = 5), average resulting coefficients to obtain the final causal estimate

For more robustness, repeat the algorithm S =100 times for different splits, then report the median estimate.

Prediction models and implementation

Fuhr et al. (2024) reports that researchers often use lasso or random forest for the prediction part, which is due to the early implementation of DML, for which in Stata lasso was used. They review each machine learning prediction model suggested by Chernozhukov et al. (2018) and group them by three dimensions. Tree-based methods like random forests and boosted trees, as well as neural networks- specifically, Multilayer Perceptron (MLP) with nonlinear activation functions, can naturally capture complex, nonlinear relationships in data, unlike linear regression and lasso, which are limited to linear effects unless researchers manually add transformations, which is a process that is both challenging and theoretically uncertain. Generalized additive models (GAMs) flexibly model smooth, additive relationships but struggle in high-dimensional settings and require manual specification of interactions. Regarding variable selection, lasso and tree-based methods can automatically focus on predictive features even with many variables, whereas linear regression, GAMs, and neural networks do not, unless additional techniques are used to prevent overfitting. Finally, although highly flexible models like random forests, boosted trees, and neural networks can model a wide variety of functional forms, they generally require large sample sizes, while linear models do not become more flexible with increased data (Fuhr et al. 2024).

We also implement the prediction models suggested by Fuhr et al. (2024) to observe if there is a difference between the models to predict the return to education with CPS 2024 data, which is a large sample, but because of our model structure has neither a high-dimensional setting nor non-linear relationship, the model is kept relatively simple. Here are the implemented prediction models within the DML framework:

- Random Forest

- Boosted Trees (XG Boost)
- Lasso regression
- Generalized Additive Models (GAMs)
- Neural Nets (Multilayer Perceptron- MLP)

Random Forest

Random Forest is an ensemble method of decision trees. Decision trees split data into branches based on rules. For each split, they choose a variable and a cutpoint that minimize the residual squared error (RSS). The prediction is the mean among the data that falls in the non-overlapping region R. One of the problems of this method is that when you grow multiple trees, each tree looks similar to the others if there is a strong predictor. Random forest gives it diversity by considering only $\sqrt{p}$ predictors (p is the number of predictors). This approach enables lowering the correlation between trees, resulting in more variance and reliable results (James et al. 2023). We implemented random forest with the *ranger* package (Wright and Ziegler 2017) as a wrapper with 500 trees, and *mtry* (number of variables considered for each split), minimum node size, and the fraction of sample to be considered are autotuned by random search cross validation (3 folds, 30 times).

XGBoost

XGBoost is another ensemble method of decision trees. XGBoost was developed by Chen et al. (2016) based on the so-called boosting technique in decision trees. Boosting technique also grows a number of trees, but unlike random forest, each tree fits on a modified version of the original dataset and learns slowly. Set $\hat{f}(X) = 0$ and $r_i = y_i$, for all i. For $b = 1,2, \dots, B$, repeat: fit a tree $\widehat{f^b}$ with d split to the training data, update $\hat{f}$ by adding a shrunken version of the new tree $\hat{f}(x) = \hat{f}(x) + \lambda\widehat{f^b}(x_i)$, then update the residuals$r_i = r_i - \lambda\widehat{f^b}(x_i)$. Then outputted model is $\hat{f}(x) = \sum_{b=1}^{B} \lambda\, \widehat{f^b}(x)$ (James et al., 2023). XGBoost-eXtreme Gradient Boosting adds a penalty term that represents the complexity of the tree including

number of nodes and weights in each node. In other words, XGBoost balances residual minimization and model simplicity. For its implementation, we use *xgboost* package by Chen et al. (2023) as a wrapper in *mtry package*. The learning rate (eta), the maximum depth of each tree, the proportion of subsamples, the proportion of features considered, and the number of rounds is autotuned by random search cross validation (3 folds, 30 times).

Lasso Regression

Lasso regression is a shrinkage regularization method applied to regression. It prevents overfitting by shrinking less predictive coefficients to zero. This method is particularly worthwhile in high-dimensional settings where the number of predictors exceeds the number of observations and conventional methods cannot be applied. Lasso balances model complexity and generalizability (Strittmatter, 2025). We implemented lasso regression with *glmnet* package (Friedman et al. 2010) as a wrapper, and lambda is auto-determined by leveraging the default values of 10 for cross-validation inside the lasso regression, and K = 5 folds are applied for DML. For example, if the data consists of 1000 samples, it is divided into 5 folds for DML (200 rows each), for cross-fitting. Fold 1 is used for test data, and folds 2-5 are used to train with lasso (800 rows total), then the training data is divided into 10 folds for cross validation (80 rows each).

Generalized Additive Models (GAMs)

GAMs is an additive model that allows non-linear relationships to fit each variable. To be specific, it calculates different f_j for X_j, and it adds them together in the model (James et al. 2023). e.g. $wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon$

Since we do not need to specify a functional form, it allows us to make a nonlinear relationship and an accurate prediction of the outcome variable. As it is an additive model, the effect of each variable on Y is measured holding other variables fixed. For generalized additive models (GAMs), we employ *mgcv* (Wood 2017) with a smooth nonlinear function estimated

with restricted maximum likelihood (REML). The REML approach is recommended for its robustness and ability to avoid overfitting.

Neural Nets (Multilayer Perceptron- MLP)

Neural Nets is an advanced model inspired by the human brain structure where signals are passed on from cell to cell. The term itself is an umbrella term that represents so-called deep learning models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). For this paper, we use a Multilayer Perceptron (MLP), which consists of an input layer, one or more hidden layers, and an output layer. In the hidden layers, the model assigns weights and biases to each input, applies an activation function, and passes the result to the next layer until it reaches the output layer (Gardner, 1998). It is very flexible and capable of fitting nonlinear relationships at the cost of a larger sample size requirement, sensitivity to scaling, and difficulty of tuning several hyperparameters. In this paper, Neural Nets/Neural Networks represents MLP. For Neural Nets (MLP), we implemented *nnet* package by Venables and Ripley (2002). We set the number of units in one hidden layer is 10, the number of iterations is 200, and the sigmoid activation function. The values for variables: education, experience, and experience-squared are scaled (value - mean/SD) for neural nets to obtain reliable results. To avoid overfitting, the maximum number of iteration until it converges is set to 500, and the maximum number of parameters is 5000. Furthermore, we need to make Neural Nets capable of learning a linear relationship. We added a skip passage from the input layer directly to the output layer to enable the model to learn a linear relationship. The weight decay rate (L2- regularization term) and the number of units in hidden layer is autotuned by random search cross validation (3 folds, 30 times).

Double Machine Learning (DML)

For the implementation of DML, we use the *mlr3* package (Lang et al. 2019) in R (version 4.5.0) (R. Core Team, 2024) to efficiently run DML framework with a selected prediction model. We set the number of folds to K = 5 and repeat the process for S = 100 times and report the median estimate (Lang et al. 2019). All the predictions are automatically made with *mlr3* package in R and wrapper packages introduced above, except GAMs and Neural Nets. We implemented DML process with GAMs and Neural Nets manually because GAMs were not available from *mlr3*, and some parameter settings of *nnet* package were not available to access from *mlr3* package as of today.

4. Results

This section presents estimates for the OLS and DML model. We first show the results of the OLS and DML and compare them to each other. Next, we present some robustness checks and discuss the limitations of the research. Table 2 illustrates the estimates with OLS and Table 3 shows the estimate of the returns to education on wage by Double Machine Learning methods.

Table 2: OLS results

Variable	(1)	(2)	(3)	(4)
Education	0.090***	0.087***	0.031**	
	(0.002)	(0.003)	(0.011)	
$Education^2$			0.002***	
			(0.000)	
College				0.450***
Degree				(0.011)
Experience	0.024***	0.019***	0.020***	0.022***
	(0.002)	(0.002)	(0.001)	(0.001)
$Experience^2$	-0.000***	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Female		-0.156***	-0.157***	-0.148***
		(0.011)	(0.010)	(0.010)
Married		0.102***	0.097***	0.093***
		(0.012)	(0.010)	(0.011)
African		-0.086***	-0.087***	-0.089***
American		(0.019)	(0.015)	(0.015)
Asian		0.049	0.044*	0.028
		(0.024)	(0.020)	(0.020)
Hispanic		-0.093***	-0.104***	-0.152***
		(0.016)	(0.013)	(0.013)
Labor union		0.027	0.030	0.044**
		(0.017)	(0.016)	(0.016)
Constant	1.776***	1.906***	2.290***	2.962***
	(0.038)	(0.040)	(0.079)	(0.017)
Adjusted R^2	0.240	0.282	0.284	0.271
N	8,145	8,145	8,145	8,145

Note: *** p < 0.001, ** p < 0.01, * p < 0.05. The value in the parenthesis shows the HC1 standard error.

Table 3: The education coefficient from DML model

Variable	Random Forest	XGBoost	Lasso	GAMs	Neural Nets
Education	0.0856***	0.0856***	0.0859***	0.0871***	0.0850***
	(0.0025)	(0.0026)	(0.0025)	(0.0025)	(0.0025)
N	8,145	8,145	8,145	8,145	8,145

Note: Note: *** p < 0.001, ** p < 0.01, * p < 0.05. The outcome variable is log hourly wage; The coefficient shown is the median estimate of S = 100 times repetition with different folds. The reported estimate from Neural Nets is already unscaled (the value was scaled in the implementation) The median standard error of S = 100 repetition is reported in the parenthesis.

The OLS estimates of the coefficient on schooling are 0.087 to 0.090 ((1) and (2)), indicating that an additional year of education increases the hourly wage by 8.8 to 9.1 percent, which is also highly statistically significant. The OLS estimate with college degree dummy variable (4) shows a 45 percent increase in wage with a college degree, indicating approximately 11 percent increase per year for four-year college, which is mostly the case in the United States. (The other OLS estimate with education square variable (3) showed only a 3 percent increase in wage for additional year of schooling.) The coefficients of the control variables are estimated as expected. The wages rise with experience and are lower for females and non-whites and also rise with marital status. DML model produces a coefficient of approximately 0.086, suggesting an 8.6 percent increase in wages for an additional year in school, again statistically significant.

In order to find out if the difference in estimates between the two models is statistically significant, we perform a t-test and calculate p-values. The t-statistics and p-value tell if the difference could just be due to sampling variability or whether it is large enough to be considered statistically meaningful. The *t-statistics* and *p*-value are calculated by the following formula:

$$t = \frac{\hat{\beta}_{OLS} - \hat{\beta}_{DML}}{SE(\hat{\beta}_{OLS} - \hat{\beta}_{DML})}$$

To obtain the standard error of $\hat{\beta}_{OLS} - \hat{\beta}_{DML}$, we consider following formula as $\hat{\beta}_{OLS}$ and $\hat{\beta}_{DML}$ are estimated from the same sample and correlated to each other.

$$SE(\hat{\beta}_{OLS} - \hat{\beta}_{DML}) = \sqrt{Var(\hat{\beta}_{OLS} - \hat{\beta}_{DML})}$$

$$= Var(\hat{\beta}_{OLS}) + Var(\hat{\beta}_{DML}) - 2Cov(\hat{\beta}_{OLS}, \hat{\beta}_{DML})$$

We calculated the standard error with a bootstrap technique 500 times and constructed a variance-covariance matrix. However, for $\hat{\beta}_{OLS}$ and $\hat{\beta}_{DML}$ on numerator, we used point estimate coefficients in Table 3 and Table 4 to calculate t-statistics due to a computational

difficulty. The p-value is calculated as follows:

$$p = 2 \times (1 - \Phi(|t|))$$

Table 5: Difference between OLS coefficient and DML coefficient across different ML methods

	Random forest	XG Boost	Lasso	GAMs	Neural Nets
T-statistics	0.987	1.124	2.864*	-0.369	1.741
p-value	(0.324)	(0.261)	(0.004)	(0.712)	(0.082)

Note: p = 0, ***, p < 0.001, **, p < 0.01, *

Table 5 shows the results, including the t-stats and the associated p-value. It shows a *p*-value far above the 5 percent or even 10 percent thresholds indicating that there is no statistically significant difference between OLS and DML models in this sample except for DML with Lasso predictor at 99% level. This suggests that the estimated return by lasso was significantly lower than the one from the OLS estimator.

5. Robustness and Simulations

As discussed before, the OLS model is estimated with a robust standard error and the earn weight, and the DML model implements a cross-fitting method to avoid overfitting and some possible bias in the estimate. Since the findings do not follow what McNamara (2020) assert that the DML model is higher or more nuanced estimates when the dataset is complex and interacts in nonlinear ways, we discuss a potential factor that explains our findings.

Even if our model does include non-linear terms such as education and experience and is low-to-middle-dimensional with 8 control variables, it might not be yet complex enough and still relatively linear compared to other settings such as economic activity or healthcare field to create a difference between OLS and DML or between machine learning predictors

The labor data for a single year may be homogeneous regarding wages and education levels. This may reduce the variance of the data which results in less bias from omitted

variables and nonlinearities, enabling the OLS model to perform decently well. The DML model and the OLS model can be different if we include more variables with more observations and years.

The performance of DML is assessed usually by simply comparing different machine learning methods, and others do so by evaluating the methods with simulated data or comparing the prediction with the results from causal inference studies as Gordon et al. (2022) tested the performance of DML for the estimate of advertising effectiveness at Facebook.

Fuhr et al.'s (2024) simulation framework to test the accuracy of OLS and machine learning models includes non-linear and high-dimensional terms such as square, cube, interaction term and step function which is a non-linear as discontinuous form to make their simulation model complex enough to observe any differences between the algorithms.

We evaluate our OLS and DML models by Monte-Carlo simulation by using the simulated data as the true coefficient of education in unknown from the observed data. We follow the proposed equation by Fuhr et al. (2024) as follows:

Simulation 1

$$W = \alpha_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \epsilon_0$$

$$Y = \alpha_1 + \beta W + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \epsilon_1$$

Simulation 2

$$W = \alpha_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \epsilon_0$$

$$Y = \alpha_1 + \beta W + \gamma_1 X_1 + \gamma_2 X_2^2 + \gamma_3 X_1 X_2 + \gamma_4 step(X_3) + \gamma_5 X_4^3 + \epsilon_1$$

Simulation 3

$$W = \alpha_0 + \delta_1 X_1 + \delta_2 X_2^2 + \delta_3 X_1 X_2 + \delta_4 step(X_3) + \delta_5 {X_4}^3 + \epsilon_0$$

$$Y = \alpha_1 + \beta W + \gamma_1 X_1 + \gamma_2 X_2^2 + \gamma_3 X_1 X_2 + \gamma_4 step(X_3) + \gamma_5 X_4^3 + \epsilon_1$$

where β is the true effect of education on wage which is set to (β = 0.1), X_1 to X_4 are the covariates that we consider in the model, and W is treatment variable (the returns to education

in this context) and Y is an outcome variable (wage in this context). We set confounding strength δ_i, γ_i (i = {1, 2, 3, 4}) between 0 and 1. $step()$ is the implementation of a step function. Also, W value is adjusted between 8 and 20 so that it better represents years of education. In simulation 1, both W and Y are linear function of confounders, whereas in simulation 2, Y includes nonlinear terms such as square, cube, interaction, and step function keeping the W still a linear function of covariates. In simulation 3, W and Y are now nonlinear function of covariates.

All simulations are implemented with number of observations (N = 1000) and the same number of folds and repetition for DML (K =5, S = 100). We report on the estimated effect of education, standard deviation, bias, and Root-Mean-Squared-Error (RMSE) for each model (Table 6).

In a setting where the true data-generating process is linear and as simple as simulation 1, OLS and GAMs recover the true treatment effect with minimal bias and variance. Machine learning methods such as Neural Nets, random forests, and XGBoost also perform well, though with slightly higher variance. Lasso, however, exhibits large bias and variance in this context, which underscores the need to tailor the choice of method to the complexity of the underlying data-generating process. The result from simulation 2, where Y is a nonlinear function of X's, exhibits minimal bias and variance for machine learning models such as Neural Nets, random forest, and GAMs. OLS and XGBoost performed fairly well, but they show higher variance. Lasso suffered from nonlinearity and a low-dimensional setting, which turned out to be of high variance. In simulation 3, where both W and Y are nonlinear function of X's, Neural Nets outperformed other algorithms with a minimal bias, tree-based machine learning algorithms follow, and models that assume a linear form, such as OLS and Lasso, suffered from nonlinearity the most.

Table 6: Results from simulations

Simulation 1				
Model	Mean	SD	Bias	RMSE
OLS	0.103	0.0319	0.0031	0.0319
GAMs (DML)	0.103	0.0319	0.0033	0.0319
Neural Nets (DML)	0.103	0.0338	0.0034	0.0338
XG Boost (DML)	0.095	0.0369	-0.0053	0.0371
Random Forest (DML)	0.121	0.0331	0.0206	0.0388
Lasso (DML)	0.139	0.0306	0.0392	0.0496
Simulation 2				
Model	Mean	SD	Bias	RMSE
Neural Nets (DML)	0.087	0.0322	-0.0129	0.0345
GAMs (DML)	0.101	0.0381	0.0011	0.0380
Random Forest (DML)	0.114	0.0374	0.0141	0.0398
OLS	0.102	0.0420	0.0018	0.0418
XG Boost (DML)	0.096	0.0422	-0.0039	0.0421
Lasso (DML)	0.160	0.0420	0.0600	0.0732
Simulation 3				
Model	Mean	SD	Bias	RMSE
Neural Nets (DML)	0.102	0.0315	0.0019	0.0314
XG Boost (DML)	0.130	0.0386	0.0295	0.0484
Random Forest (DML)	0.188	0.0343	0.0878	0.0942
GAMs (DML)	0.232	0.0303	0.1320	0.1350
OLS	0.544	0.0377	0.4440	0.4460
Lasso (DML)	0.575	0.0385	0.4750	0.4770

Note. The true Average Treatment Effect (ATE) coefficient is $\beta = 0.1$ in all simulations. Models are ordered from the smallest RMSE to the largest. The effect of education on wage is completely linear in simulation 1 and simulation 2, and nonlinear in simulation 3. Number of observations in each sample (N) is 1000, number of simulation (S) is 100, and number of cross-fitting folds for DML (K) is 5.

To summarize the findings, DML with Neural Nets predictor performed well in all settings with minimal error at the cost of the necessity of tuning several hyperparameters. GAMs did very well in Simulation 1 and 2 for their strength of additive functional forms but suffered in simulation 3 as GAMs is limited to smooth functional forms. Tree-based methods: Random Forest and XGBoost performed fairly well in all simulations for their flexibility to fit in nonlinear functional forms; however, not as good as Neural Nets. We could recover this by increasing the sample size, as very flexible methods, such as ensemble tree-based, typically require a large number of observations. OLS, which is limited to a linear functional form, outperformed others in simulation 1, but suffered from nonlinearity in simulations 2 and 3.

Interestingly, when the variable of interest is a linear function, OLS is able to estimate well with the existence of nonlinearity in the model seen in simulation 2. Finally, Lasso could not recover the true treatment effect in any setting. Lasso assumes linear relationships and also gives full play to its ability in a high-dimensional setting for its shrinkage optimization.

Figures 2, 3, and 4 visualize the results from each simulation. W denotes the variable of interest (treatment) and g(X) represents the function of Y.

Figure 2: boxplot from simulation 1

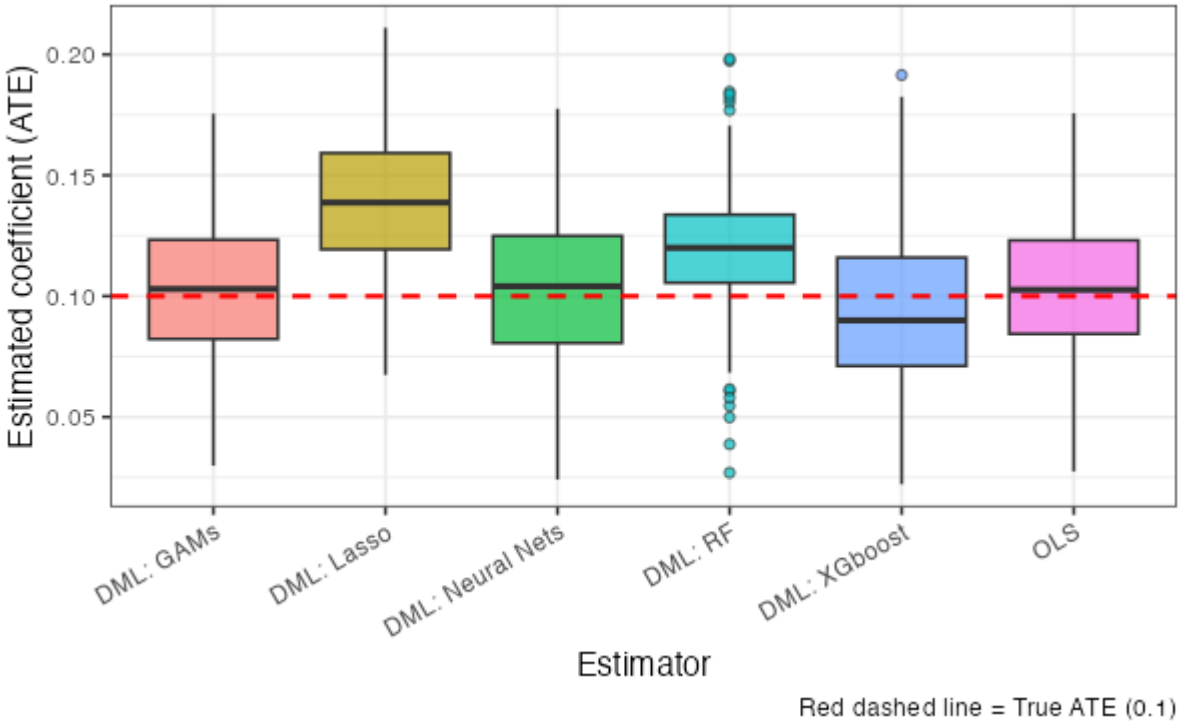


Figure 3: Boxplot from simulation 2

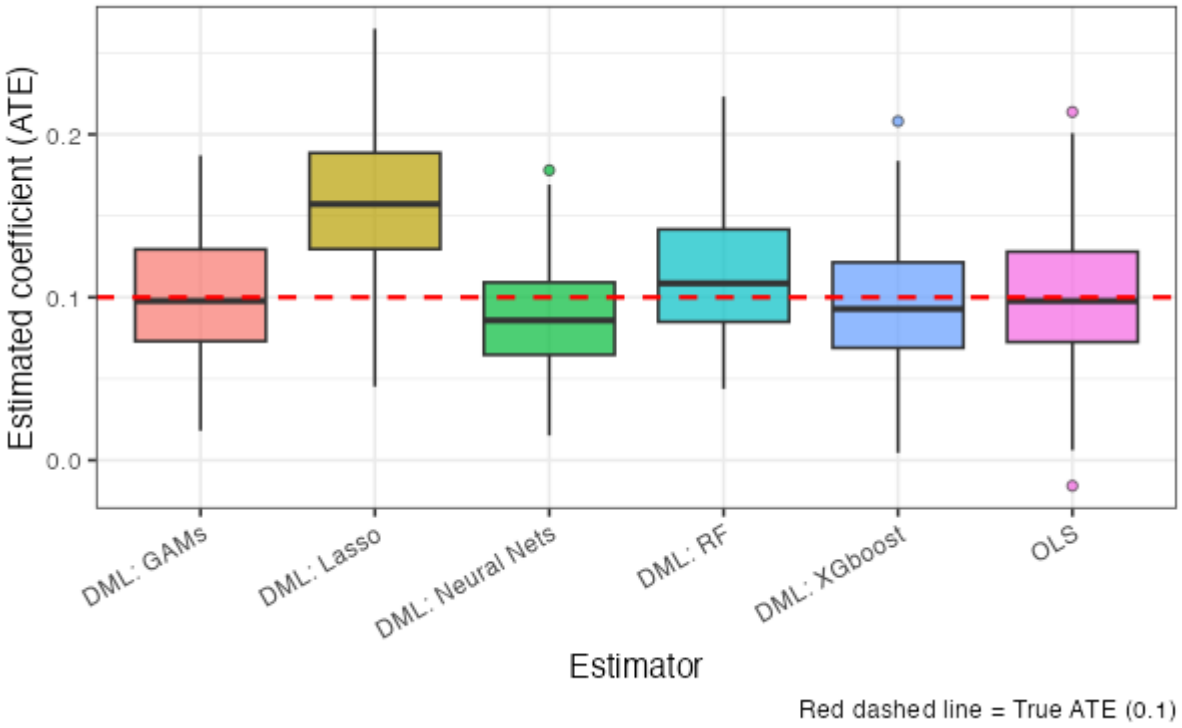
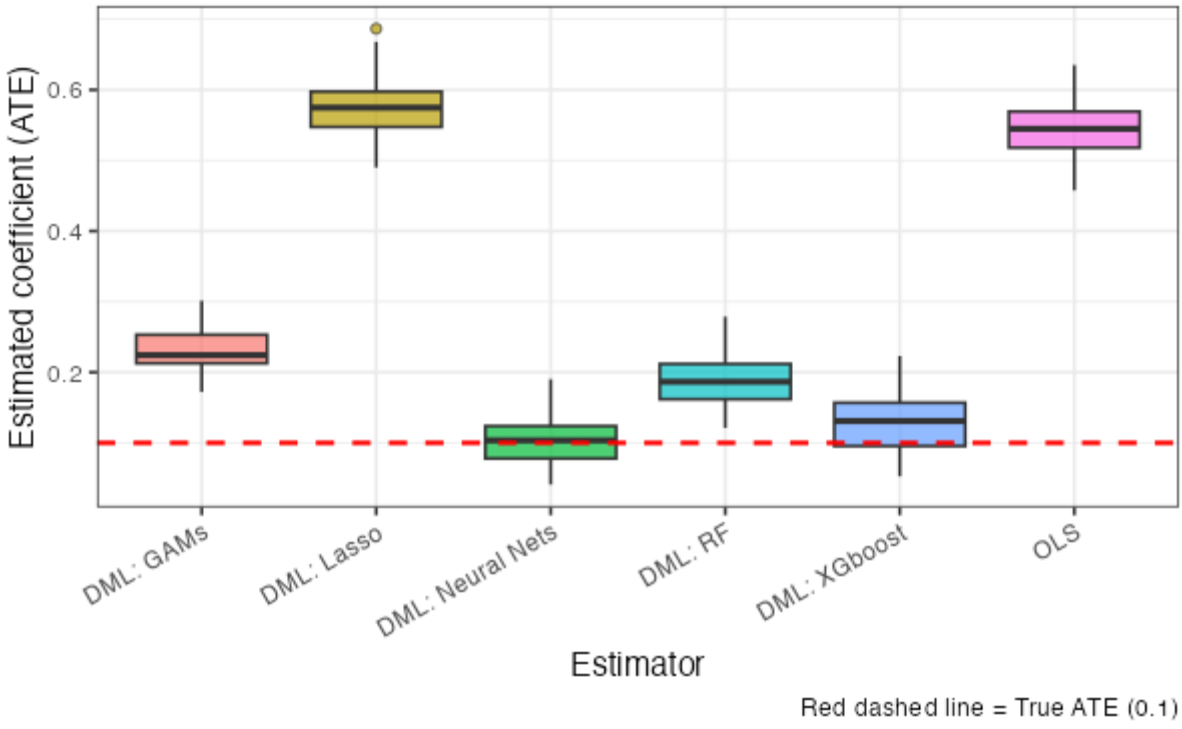


Figure 4: Boxplot from simulation 3



6. Discussion

This study compared Ordinary Least Squares (OLS) and Double Machine Learning (DML) in estimating the returns to education in the United States using 2024 CPS data. The results demonstrate remarkable consistency: both methods yield returns of roughly 8 to 9 percent per additional year of schooling. Despite DML's theoretical advantages in handling high-dimensional, nonlinear relationships (McNamara 2020; Fuhr et al. 2024), estimates in this setting do not differ significantly from those produced by OLS. However, DML estimate with Neural Nets provides significantly lower returns to education compared to OLS estimator.

The simulations provide insight into why this is the case. When the true data-generating process is linear – as in the Mincerian framework – linear estimators such as OLS and Lasso, and machine learning approaches such as GAMs or Neural Nets accurately recover the treatment effect with minimal bias or variance. Tree-based models (random forest and boosted trees) also perform reasonably well. As nonlinearity grows, linear-based methods began to suffer, while a very flexible method, Neural Nets kept producing an accurate estimate, followed by tree-based models and GAMs. These findings underscore an important point: the suitability of advanced machine learning approaches depends on the complexity of the underlying data structure. In relatively low-dimensional, linear contexts such as the US wage–schooling relationship, traditional econometric models remain highly reliable.

The evidence also suggests that the 2024 CPS labor market data may be too homogeneous to reveal meaningful differences between OLS and DML. In contexts with more heterogeneity, nonlinearities, or complex interactions—such as international comparisons, policy shocks, or long-run panel data—DML's flexibility may prove more advantageous, especially with Neural Nets as a predictor. This highlights the importance of aligning the estimation method with the complexity of the data-generating process rather than defaulting to more sophisticated models in every application. Also, another thing that needs to be considered

is that flexible methods, such as Neural Nets or Tree-based methods, require a large sample size, the necessity of hyperparameter tuning, and computational burden.

From a methodological perspective, the results caution against the uncritical adoption of machine learning methods in applied economics. While DML offers robustness in high-dimensional settings and can reduce bias from model misspecification, its advantages do not automatically translate into better estimates in simpler empirical contexts. For policy purposes, where interpretability and transparency are crucial, OLS continues to provide reliable and easily communicable results.

Several limitations should be acknowledged. First, the analysis relies on a single year of CPS data (2024), which constrains external validity. A broader temporal or international scope could reveal greater heterogeneity in returns and potentially different performance of OLS versus DML. Second, like all observational studies, the models assume that the included control variables capture all relevant confounders, i.e. exogeneity. If important unobserved factors influence both education and wages, estimates may still be biased. Third, while DML helps relax functional form assumptions, the low interpretability of some machine learning algorithms—particularly neural networks—limits their usefulness for applied policy discussions. Finally, the simulations conducted here were restricted to relatively linear data-generating processes; results may differ under more complex or nonlinear structures.

7. Conclusion

This study compared OLS and Double Machine Learning (DML) in estimating the returns to education using 2024 CPS data. Across methods, the estimated return was consistently 8 to 9 percent per additional year of schooling, with no statistically significant differences between OLS and the various machine learning predictors, except for Lasso, which exhibited a marginal difference between OLS estimate. Simulations confirmed that when the underlying data-generating process is linear, OLS, Lasso, GAMs, and Neural Nets perform

with minimal bias and variance; tree-based models perform reasonably well. As the data-generating process becomes complex and nonlinearity increases, Neural Nets consistently produced an accurate estimate while linear based methods such as OLS and Lasso started to fail. These findings underscore the importance of aligning estimation methods with the complexity of the data rather than assuming more advanced algorithms will always yield superior results. In relatively simple settings, OLS remains a dependable tool, but as data complexity increases, DML and related approaches may become essential for accurate and policy-relevant estimates. This evidence implies that OLS remains a robust tool for estimating returns to education for its performance and its interpretability, and implementation of DML could help handle high dimensionality and nonlinear/complex functional form, but sample size, hyperparameter tuning, computational capability, and optimal model selection are essential.

At the same time, several limitations must be noted. The analysis is based on one year of US data, assumes that observed controls account for all confounders, and relies on simulations of relatively simple linear structures. In more complex or heterogeneous contexts, different results may emerge.

Future research should extend this comparison to multi-year datasets, international samples, and settings with stronger nonlinearities or richer covariate structures. Combining DML with quasi-experimental approaches, such as instrumental variables or regression discontinuity designs, may also enhance causal inference. In instrumental variable (IV) approaches, the first stage involves predicting treatment assignment from one or more instruments. ML methods can enhance this step by modeling complex relationships flexibly and nonparametrically, reducing the need for strong functional form assumptions. DML techniques exploit this by incorporating ML algorithms to estimate nuisance functions, such as treatment and outcome models, thereby improving robustness in high-dimensional contexts.

Doing so could clarify where machine learning adds value beyond conventional econometric approaches and provide stronger evidence for policymakers weighing educational investments.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444-455.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The quarterly journal of economics*, *106*(4), 979-1014.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, *31*(2), 3-32.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29-50.

Buscha, F., & Dickson, M. (2023). Returns to education: Individuals. In *Handbook of Labor, Human Resources and Population Economics* (pp. 1-39). Cham: Springer International Publishing.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127-1160.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, *3*, 1801-1863.

Chan, Z. T., & Meunier, S. (2022). Behind the screen: Understanding national support for a foreign investment screening mechanism in the European Union. *The Review of International Organizations*, *17*(3), 513-541.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J. (2023). xgboost: Extreme Gradient Boosting. R package version 3.0.2.1, https://github.com/dmlc/xgboost.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives*, *36*(3), 75-102.

Cengiz, D., Dube, A., Lindner, A., & Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. *Journal of Labor Economics*, *40*(S1), S203-S247.

Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American economic review*, *91*(4), 795-813.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*, 1-22.

Fuhr, J., Berens, P., & Papies, D. (2024). Estimating Causal Effects with Double Machine Learning--A Method Evaluation. *arXiv preprint arXiv:2403.14385*.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close enough? A large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*, *42*(4), 768-793.

Harmon, C., Oosterbeek, H., & Walker, I. (2003). The returns to education: Microeconomics. *Journal of economic surveys*, *17*(2), 115-156.

Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900-1902.

Hernan, M. A., & Robins, J. M. (2020). *Causal inference: What if chapman hall/crc.* Boca Raton.

Hastie, T., Tibshirani, R., and Friedman, J. (2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, 2 edition.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4-29.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 103). New York: springer.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., ... & Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903.

Luan, H., & Tsai, C. C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, *24*(1), 250-266.

McConnell, K. J., & Lindner, S. (2019). Estimating treatment effects with machine learning. *Health services research*, *54*(6), 1273-1282.

McNamara, S. (2020). Returns to higher education and dropouts: A double machine learning approach. *ZEW Discussion Papers*, *20*.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, *66*(4), 281-302.

Mincer, J. (1974). Schooling, experience, and earnings. Human behavior & social institutions no. 2.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review*, *96*(1), 152-175.

Osman, S., & Rubb, S. (2025). Education, online education, and the earnings of nurses. *Applied Economics Letters*, 1-6.

Patrinos, H.A., & Psacharopoulos, G. (2026). Causal returns to education. *International Journal of Educational Development*, 122, p.103565.

Patrinos, H., & Sakellariou, C. (2005). Schooling and labor market impacts of a natural policy experiment. *Labour*, *19*(4), 705-719.

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics*, *26*(5), 445-458.

Qiu, M., Zigler, C., & Selin, N. E. (2022). Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions. *Atmospheric chemistry and physics*, *22*(16), 10551-10566.

Shmueli, G. (2010). To explain or to predict?. *Statistical science*, 289-310.

Spano, I. F. (2024). Family Structure Change and Children's Education: A Double-Machine Learning Approach. *Available at SSRN 4887361*.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281-306). Academic Press.

Strittmatter, A. (2025). Machine learning for causal inference in economics. *IZA World of Labor*.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

U.S. Census Bureau and U.S. Bureau of Labor Statistics. (2024). Current Population Survey, Annual Social and Economic Supplement (ASEC) 2024 [Data set]. U.S. Department of Commerce. https://www.census.gov/cps

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S*. Springer Science & Business Media.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Wooldridge, J. M. (2016). *Introductory econometrics a modern approach*. South-Western cengage learning.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of statistical software*, *77*, 1-17.