# Nonparametric Machine Learning and Average Treatment Effect: The Improvement for Estimator by Reduced-Form at the First Stage

Zhenghai Chi, Zeping Liu, Yifan Wu, Huijun Yang

Thesis Advisor: Associate Professor Honghao Zhao

Macau University of Science and Technology

March 2023

## Abstract

Recent progress in the improvement of the instrumental variables (IV) method demonstrated the potential future of the combination of machine learning (ML) and econometrics. Inspired by prior scholars' contributions, we propose a new alternative based on ML named Backward Inference Instrumental Variable (BIIV) to estimate the causality within social science. Our works make great contributions to several different strands of extant research. First, we come up with a new idea that refines and extracts the unknown confounders from the estimated random errors, which relaxes the strict assumption of Double/debiased Machine Learning (DML) that assures the performance of estimator via the necessary omniscience of all confounders. Second, BIIV is data-driven and relies on nonparametric algorithms, where we thus do not require the specification of functions for estimation, and namely, the requirement of ex-ante assumption is milder than before. Third, although we give concessions to several previous rigid assumptions, the backward inference within our methodology still guarantees the basic requirement of consistency under the final OLS estimation. More importantly, the process of backward inference within our methodology can be logically inferential, distinguishing from the conventional idea of the "black box heuristic" that exists in the realm of ML for a long time. We employ extensive Monte Carlo simulations to attest to the performance of BIIV, in which we find that our estimator outperforms the DML in obedience to the requirement of consistency as the sample size shrinks.

**Keywords**: Causal Inference, Instrumental Variable, Double Machine Learning, Nonparametric

## 1 Introduction

These days, scholars affiliating with social science emphasize causality between various factors rather than shallow correlation, which is especially reflected in the upsurge of empirical studies. Whereas for a long time, restricted by human cognition, the concept of causality was once misunderstood, and such blur induced much ridiculous abuse of causal tests. For example, people once took the equivalence between temporal ordering and causality for granted. Under this premise, early scholars even concluded an absurd causal relation that eggs came first before chicken (because there is a temporal ordering) via Granger Causality (Thurman et al., 1988). Later scholars tried to set a rigorous definition for causality (Granger, 1988) as follows:

(a) The cause occurs before the effect.
(b) The causal series contains special information about the series being caused that is not available in the other available series.

Although the definition of causality has been clarified more clearly, methodologies that aim at examining this relation still remain messy for further improvement. Currently, methodologies such as Rubin Causal Model (RCM) and the Structural Causal Model (SCM) (Rubin, 1974; Rubin, 1978; Pearl, 2009) have helped us to anatomize causal relations effectively, but an essential problem in practice makes it difficult for the diagnosis of causality, which is the inherent flaw of data record and observation. For example, we try to figure out the impact of treatment X (hospitalization for cure) on outcome Y (pneumonia), of which the confounder Z (Covid-19, and it cannot be known or diagnosed before it is discovered) cannot be observed. As a result, we may conclude that the cure does not make a difference to pneumonia, and we can clearly see that it is wrong because the unobservable confounder Covid-19 is not taken into account. Also, similar difficulty deters the economists' assertion in many aspects. Another famous instance in social science is the investigation of causality between education (cause) and income (effect), yet for which some unobservable confounders (e.g., IQ, parent's social rank, etc.) may also affect the cause and the effect to some extents and thus blur the inference for this causality (Angrist & Krueger, 1991).

The core of the aforementioned can be summarized as the problem of unobservable confounding. And it will distort our estimation of the objective causal relation because many statistical methodologies posit unconfoundedness, which indicates that all confounders are observable and measurable (Angrist & Pischke, 2009). In practice, one way to follow the above method is to add the observable confounders into models as controls. However, the reality is that we do not confirm whether pertinent confounders are captured, or in other depressing words, we cannot determine all confounders in certitude (Imbens & Rubin, 2015). Such a problem above is usually referred to endogeneity problem, including omitted variables, measurement error, and simultaneity.

To circumvent this limitation, economists achieve consensus in the employment of quasi-experiments, which try to acquire more exact causal inference via identification strategy design (Wooldridge, 2007; Angrist & Pischke, 2010). Among this framework, three techniques leverage naturally randomized experiments to estimate the causal impact, which are Regression Discontinuity Designs (RDD), Difference-in-Differences (DID), and Instrumental Variables (IV). DID is created based on the counterfactual framework, where economists try to find a control group that is as same as the treated group except for the treatment variable. And later development generalized the DID framework to nonlinear models with more flexible applications (Athe & Imbens, 2006). As for RDD, the classical sharp RDD finds a cutoff of a continuous running variable to confirm a quasi-randomized point to diagnose the treatment effect as well as the causal relation (Imbens & Lemieux, 2008). These two methodologies emphasize the mechanism design, while IV seems different from them. The core of IV is to seek effectively correlated instrumental variables of the treatment variable, and it can be seen as a two-stage estimation with different objectives. The first stage is to "filter" the treatment variable out of confounders by instrumental variables, and the second stage is to regress the outcome variable to the "filtered" treatment variable that is exogenous after being filtered. In other words, IV method could be regarded as two tasks: one is a prediction task in stage one, and another is inferential statistics in stage two. Under this context, it provides a perfect combination with machine learning (ML) for better causal inference, which makes it distinguished from the other two and offers us the opportunity to improve it based on machine learning.

In view of the function of the prediction of ML, scholars have developed a series of interdisciplinary methodologies. Earlier employment of ML in causal inference emphasizes the dichotomous classification, therefore apprehending the causality from the difference between two groups (Neyman, 1923). Namely, the heterogeneity here is regarded as the causal relation. One instance of RCM may make it simple to understand: Given a perfect but impossible context, we hope to observe a heterogeneous condition, which includes the outcome that a patient is treated as well as the outcome that the same patient is not treated so that we can summarize the effect (namely, the causality) of the cure on patient's illness. It is explicitly impossible, and Wager & Athey (2018) suggested seeking the analogous groups of treated and control for comparison, rendering the priority of asymptotic theory over the standard prediction context. In the same year as Athey Wager's work was published, a milestone work by Chernozhukov et al. (2018) proposed Double/debiased Machine Learning (DML). Their commendable contributions extremely propel the fusion of ML and econometrics. Especially the application of Neyman-orthogonal moments, which overcame the influence incurred by the high-dimensional nuisance parameters, relaxing the assumption that limits the complexity of the parameter space. This work contributed greatly to the methodologies for causal inference.

Although great success was achieved in causal inference relative to previous absurd causal conclusions, some challenging work remained to be extended for further progress. Standing on the shoulders of giants, this paper endeavors to seek a breakthrough to address the classical problem in this field. Specifically, we proposed an estimator called Backward Inference Instrumental Variable (BIIV) to boost the estimated treatment effect when the failure of conventional assumptions occurs. In totality, our works make several contributions to the strand of current methodologies and extant literature as follows:

First, relative to current ML methodologies for causal inference, BIIV relaxes the requirement for the omniscience of all latent confounders. Under the premise that orthogonality is not broken, the disturbance of confounders should be considered carefully. However, limited by the unobservable complexity of reality, almost all known ML methodologies for causal inference have to commence their inferential work with a weakly robust assumption that all confounders are incorporated (Mooij et al., 2016; Huang et al., 2020; Schölkopf, 2022). To tackle this problem, the BIIV admits the unobservable confounders in default and then revisits the nominal 'random errors' in the first stage to estimate a ratio that evaluates and controls this influence in the second stage. Our framework for causal inference can extremely mitigate the necessity of the assumption that omniscience of confounders. More importantly, this progress is logically tenable and mathematically reasonable, and we will offer fruitful evidence for both intuition and mathematics.

Second, although we relax the classical assumption as mentioned above, we still keep the advantage of ML, which is related to the function choice. In detail, traditional IV requires the empirically ex-ante specification of the model, namely, the reduced-form estimation. However, there is a contentious between the structural model and the reduced model over the past decades (Duffie & Lando, 2001; Jarrow & Protter, 2012). Timmins & Schlenker (2009) elucidated that the reduced-form model requires a strongly restricted assumption (e.g., quasi-experiment) to construct an evident causal path for inference. In fact, a direct aspect related to the controversy is whether the complexity of the model can fit reality well. The conventional estimator of econometrics relies on strong logical reasoning, and economists need to debate the viability of their reduced-form model (Arora et al., 2012). But BIIV could overcome this problem when keeping the generalizability of the regression model. Our methodology identifies the final causal relation by the traditional OLS model, while we let the data "explain themselves" before, meaning that we do not make a stance on either the reduced-form even or the choice of instruments. Meanwhile, the final OLS estimator assures that BIIV is on par with the 2SLS IV.

Finally, we employ extensive Monte Carlo simulations to compare the performance of different models, including the BIIV, the latest orthogonal Double Machine Learning IV (DMLIV), and IV based on Two Stage Least Square (2SLS). Admittedly, the estimated variance is apparently larger than orthogonal DMLIV given the randomly generated samples. However, BIIV outperforms DMLIV in terms of both interpretability and consistency. Especially, violating the assumption of the omniscience of confounders (unconfoundedness), BIIV is the only model that can approximate consistency of estimation. Furthermore, beyond our expectations, the conventional theory guarantees consistency of estimation under the large simple, but BIIV exhibits robust consistency as the sample size shrinks. Collectively, BIIV dominates over traditional IV in any case, whereas the tug of war between BIIV and DMLIV is nip and tuck.

## 2 Literature Review

Our work mainly builds upon two important cornerstones: one is the development of ML, and another is statistical progress in orthogonality theory as well as its application of it. In addition, the intuition from economics also inspires the prosperity of this field.

Miscellaneous confounders raise the problem of high-dimensional issues. Theoretically, the high-dimensional problem induces a plummet in the degree of freedom, which finally causes the invalidity of the OLS estimator. Given a multivariable regression with independent variable vector $X$, dependent variable $Y$, n indicates the number of observations, p is the dimension of independent variables and $\beta$ is the coefficient vector, due to the high-dimensional characteristic, $rank(X) \leq n < p$, and thus the $(X'X)^{-1}$ does not exist so that the $\hat{\beta} = (X'X)^{-1}X'Y$ is unsolvable.

To tackle this problem, Hoerl & Kennard (1970) proposed the ridge regression for strict multicollinearity between tremendous independent variables, where they added the $L_2\, norm$ as penalty term to conduct penalized regression. Following their works, Tibshirani (1996) proposed the Least absolute shrinkage and selection operators (Lasso), which inherits the idea from the ridge regression and imposes a severe punishment via $L_1\, norm$ to restrict the regression coefficients. Given a compromise for above two methodologies, Zou & Hastie (2005) developed the Elastic Net estimator with both $L_1\, norm$ and $L_2\, norm$, which then enlightened the rapid development of algorithms, including adaptive Lasso, grouped Lasso, smoothly clipped absolute deviation (SCAD) penalized term, and minimax concave penalty (Zou, 2006; Simon et al., 2016; Wang et al., 2007; Zhang, 2010). However, the laggard relevant statistical inference makes the inferential interpretability ambiguous, e.g., there is a lack of uniform standard errors in penalized regression (Efron & Hastie, 2016).

Compared to penalized regression algorithms (it is still a parametric approach), other ML algorithms gave up the interpretability of the statistical model and focused on the ability of prediction. For those algorithms, without any empirical ex-ante setting of parameters, algorithms seek patterns from data themselves, which can alleviate the risk of dimensions to some extent.

Here, we revisit three classical nonparametric approaches. The first classical method is K nearest neighbors (KNN), proposed by Fix & Hodges (1951). This method fully incorporates the information gained from independent variables but ignores the response variable. Although KNN is the simplest learner, it still plays a role in data processing like adjusting the covariates accurately (Wager & Athey, 2018). Demanding the efficient use of information from datasets, decision tree algorithm incorporates both the dependent variable and independent variables into model construction, which was first created by Breiman et al. (1984). Decision trees are nominally the first algorithm that escapes from the risk dimension even though it adopts the identical strategy of "divide and conquer" to distinguish the feature space with KNN, while

KNN leaves the consideration of response variable out so that it is vulnerable to being disturbed by confounders and high dimensions. Later scholars from computer science came up with further improvements such as ID3 (Quinlan, 1986; Quinlan, 1996). As a transition, decision tree algorithms make great contributions to the development of ML, which sparks the next stage – ensemble learning. Bagging (Breiman, 1996), Random Forest (Breiman, 2001), and Gradient Boosting (Friedman, 2001), those ensemble learning algorithms propelled the application and created a blue ocean for further development. Relying on the statistical progress and mathematical evidence, they all exhibit outstanding performance than any other before, which has been the cornerstone of current causal inference based on ML.

With the prosperity in the ML field, economists also commence exploring the interdisciplinary attempt. Contingent to the need for investigation of heterogeneous treatment effects, scholars have explored the fusion of econometrics and ML, including nearest-neighbor matching, and kernel methods (Crump et al., 2008; Lee, 2009). Furthermore, little but growing literature has paid attention to the use of forest-based algorithms for estimating heterogeneous treatment effects (Green & Kern, 2012; Hill & Su, 2013). Additionally, Foster et al. (2011) took advantage of the regression forest to estimate the treated and control group's impact of covariates on the outcome, respectively. And the nuance from estimation will train the new decision tree to project the treatment effects on the unit's attributes. Though it also indirectly compares the treatment's heterogeneity among different groups, a lack of plausible statistical inference made it difficult to extrapolate. Following this, Athey & Imbens (2016) innovated estimating the treatment within each partitioned subgroup, of which the "honest" estimation could be assured via splitting the sample. Subsequent works (Wager & Athey, 2018) give a much more direct causal estimation called causal forest based on ML, which is pointwise consistency relying on generic Gaussian theory for a large family of random forest. But it is limited by the strict assumption of confoundedness, the same as before.

At the same time, another question that traditional ML never gives a straight answer to is the risk of dimension, which is not only about the extremely large feature spaces but also the latent sparsity.

The division of research fields prolongs this confusion and leaves them till now. Apparently, this problem is related to the classical semi-parametric estimators, which focus on the approximation of $\sqrt{n}$-consistent and asymptotically normal estimates for low-dimensional components with nuisance parameters estimated by the conventional nonparametric estimator, such as the generic application estimators contingent on kernel method. Levit (1976) gave vital evidence showing that:

$$\sum_{t=1}^{n} \psi(X_i, T_n) = 0$$

Given the sufficiency of admissible unknown distribution F, $T_n$ exists as the locally asymptotically minimax estimates of corresponding functional $\psi(F)$, which can be summarized as above, where $X_i$ are observations and $T_n$ is an ex-post defined class in the preceding function. Furthermore, as indispensable guidance for estimation in nonparametric models, efficiency bounds play a functionally important role. Newey (1990) defined the nature of it and provided the method of calculation, including the construction of semiparametric estimators and their limiting distribution. The cornerstone has been consolidated in the 1990s, Andrews (1994), Newey(1994), and van der Vaart (1998) presented a general set of results for low-dimensional parameters. Their breakthrough explained the presence of nonparametric estimators with the feature of nuisance factors and defined a set of regularity conditions to confirm the validity of the primitive under the $\sqrt{n}$-consistency and asymptotic normality for functions of the estimator of projected series. Meanwhile, what should be highlighted is that Andrews testified the key equicontinuity condition based on Neyman Orthogonality and Donsker Conditions:

$$(1/\sqrt{n})\sum_{i\in I}(\psi(W_i;\theta_0,\hat{\eta}) - \int \psi(w;\theta_0,\hat{\eta})\, dP(w) - \psi(W_i;\theta_0,\eta_0)) \to 0$$

Where $\theta_0$ is the low-dimensional parameter in the presence of nuisance parameters $\eta_0$, $\psi$ indicates an orthogonalized score function that satisfies the property that the Gateaux derivative operator, which is based on by Neyman (1959, 1979) and can be used to evaluate the true parameter properties as the nuisance vanishes, see below:

$$\partial_\eta E[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$$

Invoking the improvement above, the objective of conventional Partially Linear Regression (Robinson, 1988) can be reduced to below, which involves $1/\sqrt{n} - normalized\ sums$ of products the structural unobservable, where $g_0(\cdot)$ indicates the impact of confounders on outcomes:

$$(1/\sqrt{n})\sum_{i\in I} V_i(\hat{g}_0(X_i) - g_0(X_i))$$

In this case, the prediction of nuisances cannot distort the limiting distribution of parameters of interest. Related works include the contribution of van der Vaart (1988), who reset the averages of semi-parametrically efficient scores for better estimation.

Another polar thought was proposed by Scharfstein et al. (1999) and developed by Laan & Rubin (2006). The idea is to estimate the least favorable direction with the targeted maximum likelihood, such as imposing constraints inherent in the data. Van der Laan et al. (2007) referred this as to "super learner", and the subsequent works emphasized the importance of ML in the estimation of parameters within "super learner" (Luedtke & van der Laan, 2016; Zheng et al., 2018).

Such efforts enlarge our cognition to circumvent the confounders, notwithstanding the risk of high dimension. All these contributions are based on Donsker Conditions, which is a classical but strong setting that validates for and only for abundant structures of fixed function classes $\mathcal{G}$. We use a linear model to demonstrate it, and this model generates a parameter space given by the Euclidean ball with the unit radius:

$$\mathcal{G}_n = \{x \mapsto g(x) = x'\theta; \theta \in \mathbb{R}^{Pn}: \|\theta\| \leq 1\}$$

Here we see clearly that the Donsker Conditions require bounded complexity, which is easily violated in machine learning because the dimension of X will be modeled as increasing with the sample size so that the estimators must live in high dimension, even the high-dimensional linear model. In a word, Donsker Conditions cannot assure the validity of estimators as the dimension increases. Extant literature focuses on the degree of complexity, and researchers employed extensive simulations to observe to what extent the outcome keeps continuous. Belloni et al. (2015; 2016) adopted this strategy and proposed found that the increase of entropy (complexity) must be restricted by strict control, or the overfitting is severe.

Ultimately, our works are developed upon the Frisch-Waugh-Lovell (FWL) theorem (Frisch & Waugh, 1933; Lovell, 1963; Lovell, 2008). Here, we give a brief recall of the statement of this theorem. Given D as the treatment variable, X as a confounder, and Y as an outcome:

$$Y = \beta D + \alpha X + \varepsilon$$

Where the coefficient (treated effect) $\beta$ is numerically equivalent to the $\delta$ as below:

$$r_Y = \delta r_D + \mu$$

$$Note: where\ r_Y = Y - E(Y|X)\ and\ r_D = D - E(D|X)$$

This simple yet crucial theorem confirms the validity and consistency of our DML and our methodology. Furthermore, the limitation of current methodologies for causal machine learning in dichotomous treatment drove our attention to the estimator for continuous treatment.

## 3 Discussion of Invalidity of Instrumental Variables (IV)

In this section, we will introduce the application of Instrumental Variables (IV) in extant economic literature, and then give a brief introduction to the principle of IV. Finally, our discussion will mainly focus on the contentious problems of IV and the cause that may incur the invalidity of IV.

### 3.1 Extant Application of IV Method

In this part, we give a brief qualitative review of the history of instrumental variables (IV). Defined by Goldberger (1972) and Angrist et al. (1996), the structural equation models are "stochastic models in which each equation represents a causal link, rather than a mere empirical association", and the IV is a special case (it is a simplified specification) of structural equation model. These kinds of variables can affect the outcome through their exclusively restricted effect on other variables, which shed light on that they are out of the structure, but their effect can enter this system via some unique variables. Due to this characteristic, IV is then widely used for causal inference.

Some explorations for problems in social science demonstrate its strong ability of identification, where an example could be assistance of better understanding of the power of IV. Angrist & Krueger (1991) once solved a contentious problem that is to which degree education can affect income. Proverbially, education dose positively affects income, intuitively, in which some confounders (e.g., individual unobservable IQ) can distort the real causal effect. Therefore, one exclusively restricted instrumental variable is required to obtain the real causal effect of education on income. And authors use the season of birth and the regulatory rule of admission and quit of American senior high to design a causal inferential model. The resulting conclusion indicates the overestimation of the return of education on income for previous research. Although other scholars further show the flaw of this instrumental variable (Jackson et al., 2015), it is still regarded as an excellent work that inspires the latter much. Other works also contribute to this topic in a variety of research gist, including corporates reform, political election, the influence of institutions, and even psychological health status such as autism (Groves et al., 1994; Levitt & Snyder, 1997; Acemoglu et al., 2001; Waldman et al., 2006).

In fact, the relevant doubt for works that used IV is tenable and positively propels the development of IV as well as its standard test. The test for the exclusion restriction deserves our attention because it is essential to the basic assumption of IV. There is no quantitatively identical method to test this ex-ante assumption, and the qualitative explanation dominates. For example, Miguel et al. (2004) posited four potential situations in their work that may violate the exclusion restriction and ultimately defended the robustness of their research design. Notwithstanding, the authors finally admit that their evidence cannot eliminate the possibility that the exogenous variable affects the outcome variable through other channels. Another idea (Berkowitz et al., 2012) focuses on examining the correlation between error terms in regression models and instrumental variables. They base an assumption to test exogeneity of instrumental variables, which is that this correlation will gradually disappear with the sample size increasing, namely, the asymptotic orthogonality. And then extending the implementation of Anderson-Rubin to purely exogenous conditions via fractional resampling, to which Riquelme et al. (2013) make functional contributions.

Next, we have to draw our attention to the problem of weak identification of instrumental variables. Even Angrist & Krueger (1991) once complained about the hardship of finding a suitable instrumental variable, one of whom was the recipient of the 2021 Nobel Prize in Economics. But more disappointingly, we cannot guarantee the degree to which it can identify the endogenous variables even if we find one. This weak correlation may bring an unexpectedly large estimation of standard error. Furthermore, Bound et al. (1995) pointed out that the bias is extremely close to the OLS estimator with such a correlation weakened. Stock & Yogo (2002) suggested the implementation of Cragg-Donald (Cragg & Donald, 1993) statistics to search the potential weak instrumental variables.

Finally, the Generalized Method of Moments (GMM) can mitigate the invalidity of IV due to heteroscedasticity and self-correlation (Hansen, 1982). But for the flaw incurred by weak instrumental variables, we still need to employ other procedures to handle it in nonlinear GMM because this problem leads to GMM statistics with nonnormal distributions even in large sample sizes (Stock et al., 2002). Although this remains to be solved, we leave it to others because it is out of the gist of our paper.

### 3.2 Principle of IV Method

Next, we introduce the principle of the IV estimator, compared to the OLS estimator, which is usually called 2SLS. This conceptual framework is contributed by Basmann (1957) and Theil (1953). We first give the simple setting, all these settings are based on matrix form: given $y_1$ indicates outcome, $X_1$ is a set of exogenous variables, $Y_1$ represents the endogenous variables, $X_2$ is as a symbol of excluded exogenous variables (namely, the instrumental variables, and we assume that the number of instrumental variables is enough to identify this equation). The first structural equation is written as, where $Z_1 = [Y_1, X_1]$:

$$y_1 = Y_1\alpha_1 + X_1\beta_1 + \mu_1 = Z_1\delta_1 + \mu_1$$

We first regress all exogenous variables within this system on $X = [X_1 X_2]$ and the dimension of the column will increase, and finally we get the estimation of $\hat{Y}_1$ as below:

$$\hat{Y}_1 = X(X'X)^{-1}X'Y_1$$

In the resulting stage, we employ the estimator obtained from regressing $y_1$ on $\hat{Y}_1$ and $X_1$, where $\hat{Z}_1 = [\hat{Y}_1, X_1] = X(X'X)^{-1}X'Z_1$ and $w_1 = \mu_1 + (Y_1 - \hat{Y}_1)\alpha_1$:

$$y_1 = \hat{Y}_1\alpha_1 + X_1\beta_1 + w_1 = \hat{Z}_1\delta_1 + w_1$$

The estimator of the second stage can be calculated as below:

$$\hat{\delta}_1 = (\hat{Z}_1{}'\hat{Z}_1)^{-1}\hat{Z}_1 y_1 = (Z_1{}'X(X'X)^{-1}X'Z_1)^{-1}Z_1{}'X(X'X)^{-1}X'y_1$$

Given the projection matrix of $PJ_X = X(X'X)^{-1}X'$, the following equation can be rewritten as:

$$\hat{\delta}_1 = (\hat{Z}_1{}'\hat{Z}_1)^{-1}\hat{Z}_1 y_1 = (Z_1{}'PJ_X Z_1)^{-1}Z_1{}'PJ_X y_1$$

### 3.3 Invalidity of IV Method

Invoking the idea of DMLIV, as aforementioned, it can reduce the intervention of confounders, but it also faces the predicament.

DMLIV can take confounders into account, but the intermediate process is not perfectly statistically inferential, which induces confusion about the degree to the consideration of each variable. Taking the random forest algorithm as an example, the tremendous workloads of sample partition are contingent on the information gain from the diminution of MSE or other loss functions. However, including the visualization tool such as Partial Dependence Plot, little is known about the real interpretability and contribution of explanation of each variable. Namely, to some extreme, under the condition of high dimensions, each leaf of all trees in the forest can use any variables as a classification basis but excludes only one variable, which means that this variable is not taken into account. In a regression problem, it is equivalent to which this variable loses the matching parameter to explain it from statistical consideration. More generally, the regression model does not include it.

Due to the characteristic of “black box heuristic” of ML, predominantly, the sample partition is unobservable and out of our control, so that it may cause three results for two stage estimation of IV. To begin with, the first stage does not take some variables into account, and the second stage estimates perfectly from limited known variables. Second, the first stage estimates perfectly, but the second stage happens omission of some variables. Second, there is the omission in both the first and second stages, but it is too extreme to happen, and we leave it out.

Now revisiting the first two latent problems. The first one will induce the invalidity of the

2SLS estimator. Given $X^*$ as the vector of independent variables. An estimator for endogenous variable can be rewritten as $\hat{Y}_1^* = PJ_{X*}Y_1$, and the below condition will be satisfied:

$$\hat{Y}_1^{*\prime}\hat{Y}_1^* = \hat{Y}_1^{*\prime}PJ_{X*}Y_1 = \hat{Y}_1^{*\prime}Y_1$$

Also, the condition $X_1'\hat{Y}_1^* = X_1'Y_1$ holds if and only if $X_1$ has to be the subset of independent variables within $X^*$. Therefore, $X^*$ should include $X_1$ and $X_2$ that satisfies the requirement of identification for instrumental variables. Given the definition of $W = \hat{Z}_1 = [\hat{Y}_1, X_1]$, and in this case, the IV estimator using $W^* = \hat{Z}_1^* = [\hat{Y}_1^*, X_1]$ reveals the same result obtained from the 2SLS. There are two important points for our careful attention. First, this also requires the consistency of an IV estimator. Second, the requirement for just-identified is imposed, then there is an exact parameter of our interest of the X's excluded from that equation. But the problem of the lack of dimension shows that the 2SLS may reduce to an IV estimator with $W = X$, i.e.,

$$\hat{\delta}_{1,2SLS} = \hat{\delta}_{1,IV} = (X'Z_1)^{-1}X'y_1$$

But note that $X'Z_1$ is not square and the estimator above cannot be calculated if the first stage is over-identified. This degeneration of 2SLS indicates a severe outcome.

Furthermore, we can understand this severe outcome from the mutual independence between independent variables. Similar to the aforementioned, $y_1$ indicates outcome, $z_1$ is a set of exogenous variables, $Y_1$ represents the endogenous variables, $z_2$ is instrumental variable, and the model is set as follows:

$$y_1 = Y_1\alpha_1 + z_1\delta_1 + \mu_1$$

If we only regress $Y_1$ on $z_2$ in the first stage as shown:

$$Y_1 = z_2\gamma_2 + \mu_2$$

We can conclude the final estimator as below, where we cannot testify that $\mu_2$ is not correlated with $z_1$, rendering the violation of the assumption of IV and distorting the estimator. Overall, the first problem remained to be tackled.

$$y_1 = z_2\gamma_2\alpha_1 + \mu_2\alpha_1 + z_1\delta_1 + \mu_1$$

We exhibit this problem by Simulations, in which we show the invalid 2SLS estimator compared to the opposite. Subfigure a show the validity of 2SLS, in which OLS estimates 6.98 of ATE and IV estimates 5.11 of ATE, given the true ATE is 5. In such a case, the IV exhibits the absolute dominant estimation of causal relation than the OLS. But the episode twists with the correlation between the instrumental variable and error term. Under that condition, given the true ATE of 5 as well, however, Both OLS and IV demonstrate the overestimation (7.90 and 7.39, respectively), far away from the expected true value. And our methodology will try to work this out.

The second condition seems easier to mitigate, in which the first stage estimates perfectly but the second stage is disturbed by the omission of some variables. Although the "black box" exists in DMLIV estimator, 2SLS can assure the validity of statistic inference and is simple to control the manipulation of variables. Therefore, the adoption of both 2SLS and DMLIV as estimator can mitigate the risk to some extent and promise the robustness. But honestly, there is no utterly effective method to eliminate this problem. However, logically, this risk is not obvious because in most cases the ML can estimate the second stage well provided the perfect estimation in the first stage because the confounders are the same set.

## 4 Methodology

In this chapter, we will give a related discussion about the previous contributions of DML and conventional inverse mill's ratio. And finally, we will give a brief and concise summary of our innovative methodology - Backward Inference Instrumental Variables (BIIV).

### 4.1 Double Machine Learning (DML)

This section consists of two components, one is the introduction of DML and its application, and the other will elucidate the relatively complex theoretical background behind the DML.

#### 4.1.1 Related Works of DML

As aforementioned, our works build upon the achievements of Frisch & Waugh (1933), Lovell (1963, 2008), and Chernozhukov et al. (2018). In this part, we will review their contributions and the relevant cornerstone of our methodology.

Revisiting the naïve OLS estimator for causal inference, we set the regression as below:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Where Z indicates the confounders (in empirical research, they are usually regarded as control or covariates), X represents the independent variable (the variable of our interest and we interpret it as the cause), and Y is the dependent variable (we interpret it as the effect). In such a naïve regression via linear estimation in reduced form, $\beta_1$ is the causal effect of X on Y.

However, as we discussed above, some violations for assumptions of OLS make $\beta_1$ invalid, e.g., the relationship between Y and Z is nonlinear or there is a correlation between X and Z. Therefore, we want to extract the pure causal relationship between X and Y by eliminating the impact of Z on X and Y respectively.

Invoking the Partially Linear Model (PLM) as written below:

$$Y = X\theta_0 + g_0(Z) + \varepsilon$$

$$Note\!: where\ E[\varepsilon|Z, X] = 0$$

$$X = f_0(Z) + \mu$$

$$Note\!: where\ E[\mu|Z] = 0$$

And we relax the specification from naïve regression, where $\theta_0$ is the causal effect of X on Y that is of our interest, X is still the variable of the cause, Y is the outcome of the effect, $\varepsilon$ and $\mu$ are error terms, and $Z$ is the P-dimension confounders (control variables or covariates). In PLM framework, we take the breakthrough via relaxing the dimension of confounders (Z, as aforementioned, it could be any P dimensions from the extended real number, namely, $Z \in \mathbb{R}^P$). Furthermore, two functions merit our extra attentions, $g_0(\cdot)$ and $f_0(\cdot)$. These two functions are called "nuisance functions", as seen here, we only give the definition of them instead of the specification form inserted within the functions themselves. This kind of setting eliminates the subjective and ex-ante empirical assumption of function, reducing the risk of model misspecification. For those, we adopt the ML for purpose of prediction. The intrinsic nature of these functions are high-dimensional nuisance functions, and we thus ensure the specification of PLM by permitting the nonlinear impacts of Z. Finally, this effort can ensure the validity of $\theta_0$ – the causal relation of our interest.

An outright method is adopting any ML algorithms to estimate the $g_0(\cdot)$ and keeping the residuals for final regression, which is usually completed in the dataset that is regarded as "auxiliary part" (we will explain next). In totality, we partial out the effect of Z on Y and employ the regression. However, as discussion inspired by Chernozhukov et al. (2018), This straightforward abelite naïve leads to the overfitting and regularization bias, triggered by deviation from $\sqrt{n}$-consistency. And they proposed the orthogonalized formulation of PLM to correct the "inferior" convergence rate. Then, the above equation can be rewritten as follows:

$$Y - \ell_0(Z) = (X - f_0(Z))\theta_0 + \varepsilon$$

Where $\ell_0(Z) = E[Y|Z]$ and $f_0(Z) = E[X|Z]$, which reveals that they are the conditional expectation of Y and X respectively given by Z – confounding variables. And that is the reason why we call this method "Double Machine Learning" because we learn both two functions $\ell_0(\cdot)$ and $f_0(\cdot)$ via ML algorithm. And we learn the function $\ell_0(\cdot)$ to get $E[Y|Z]$ (a forecasting task for Y based on Z, which generates Y_hat), and we learn the function $f_0(\cdot)$ And finally, the estimator for causal relation ($\theta_0$, also the ATE) is acquired by the regression model.

Thereinto, the secret that circumvents the distortion of estimator $\theta_0$ incurred by overfitting is the sample partition. Usually, the n-to-n partition (one is the main part, and another one is the auxiliary) yields the $\sqrt{n}\ consistent$. The efficiency of estimation can be recovered by flipping the main and auxiliary samples. As for the base learner to learn function $\ell_0(\cdot)$ and $f_0(\cdot)$, which is contingent on the pattern of primitive data. In some cases, the standardization required by unique ML algorithms is necessary.

### 4.1.2 Theory behind DML

Related to the detail of the rationale behind DML, we will explain in this section, including the improvement for valid causal inference via the alteration of moment conditions. All the symbol and structure of model specification inherits from the above. Under the aforementioned PLM specification, the application of machine learning for approximation to the unknown function $g_0(\cdot)$, which reduces convergence rate to $1/\sqrt{n}$ – the regularization bias occurs. To illustrate it, splitting sample equally into two parts, including a main part and an auxiliary part, denoted by $i \in I^c$. Collectively, DML first approximates the $g_0(\cdot)$ in the auxiliary sample and then estimates $\hat{\theta}_0$ by OLS as follows:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} X_i^2\right)^{-1} \left(\frac{1}{n}\sum_{i\in I} X_i\,(Y_i - \hat{g}_0(Z_i))\right)$$

Decomposing the scaled estimated error, the slower convergence rate is shown as below. Under the relatively mild conditions, $a = \left(\frac{1}{n}\sum_{i\in I} X_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} X_i\varepsilon_i$ obeys the asymptotically normal distribution.

$$\sqrt{n}\big(\hat{\theta}_0 - \theta\big) = \left(\frac{1}{n}\sum\nolimits_{i\in I} X_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum\nolimits_{i\in I} X_i\varepsilon_i + \left(\frac{1}{n}\sum\nolimits_{i\in I} X_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum\nolimits_{i\in I} X_i(g_0(Z_i) - \hat{g}_0(Z_i))$$

Thereby, we desire the approximately 0 that $\left(\frac{1}{n}\sum_{i\in I} X_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} X_i(g_0(Z_i) - \hat{g}_0(Z_i))$ converges to. However, a problem hinders that:

$$b = (E[X_i^2])^{-1}\left(\frac{1}{n}\sum\nolimits_{i\in I} X_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum\nolimits_{i\in I} f_0(Z_i)\big(g_0(Z_i) - \hat{g}_0(Z_i)\big) + O_p$$

The sum of n terms that do not have mean zero and hence, by diving with $\sqrt{n}$ and it will not converge in probability. Furthermore, proverbially, regularization is necessary for process in high-dimensional datasets. The key success factor that drives the ML algorithm achievement is the balance between variance and bias. However, depending on the regularized process, a converge rate of $n^{-\varphi}$ with $\varphi_g < \frac{1}{2}$ is often triggered. In this instance, it causes the "regularized bias" due to the difference between $g_0(Z_i)$ and $\hat{g}_0(Z_i)$, in which the order is $O_p(n^{-\varphi_g})$ and $\varphi_g < \frac{1}{2}$. Finally, it deteriorates the expectation for b with stochastic order $\sqrt{n}n^{-\varphi_g} \to +\infty$. Bothered by this rate of convergence of $\theta_0$, researchers have to consider another structure to overcome this.

To cope with the inferior rate of convergence, Chernozhukov et al. (2018) propose an alternative called DML for estimator $\theta_0$. By innovating the loss function to take the regularized bias into account, a consistent estimate of $\theta_0$ can be guaranteed. The following will highlight their breakthrough and main idea.

Invoking the below equation in form of PLM again:

$$Y - \ell_0(Z) = (X - f_0(Z))\theta_0 + \varepsilon$$

Defining that $M_i = X_i - f_0(Z_i) = m_i$ and $K_i = Y_i - \ell_0(Z_i)$, then the estimator $\theta_0$ can be calculated by:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum\nolimits_{i\in I} \widehat{M}_i^2\right)^{-1}\left(\frac{1}{n}\sum\nolimits_{i\in I} \widehat{M}_i\widehat{K}_i\right)$$

Via the orthogonalization, the estimate of $\theta_0$ follows the $\sqrt{n} - consistent$ and approximately Gaussian distribution under a mild condition. In accordance with the naïve approach, the estimated error of $\hat{\theta}_0$ can be decomposed as well:

$$\sqrt{n}\big(\hat{\theta}_0 - \theta_0\big) = \left(\frac{1}{n}\sum\nolimits_{i\in I} M_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum\nolimits_{i\in I}\big[\widehat{M}_i\widehat{K}_i - M_i(L_i - \varepsilon_i)\big] = a^* + b^* + c^*$$

As per the naïve approach, the resulting term given by $a^*$ under the mild conditions will be asymptotically normally distributed. The regularized bias term $b^*$ is:

$$b^* = (E[X_i^2])^{-1}\frac{1}{\sqrt{n}}\sum\nolimits_{i\in I}(\hat{f}_0(Z_i) - f_0(Z_i))\big(\hat{g}_0(Z_i) - g_0(Z_i)\big)$$

Distinguishing from the naïve estimate, it depends on the product of the estimator errors of both $\hat{f}_0(Z_i)$ and $\hat{g}_0(Z_i)$. Consistent with the arguments above, the convergence rates of $\hat{f}_0(\cdot)$ and $\hat{g}_0(\cdot)$ are respectively $n^{-\varphi_f}$ and $n^{-\varphi_g}$, causing $b^*$ to have an upper bound of $\sqrt{n}n^{-(\varphi_f+\varphi_g)}$. Although both $\varphi_f$ and $\varphi_g$ are usually under 1/2, the product of estimator errors is typically found to be bigger than so that $\hat{\theta}_0$ approximates good properties even if $f_0$ and $g_0$ show the relatively slow rates of convergence. As for the $c^*$, it includes terms as follows:

$$\left(\frac{1}{n}\sum\nolimits_{i\in I} M_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum\nolimits_{i\in I} M_i(\hat{g}_0(Z_i) - g_0(Z_i))$$

Where the correlation may lead to the problem of overfitting. As suggested, the partition of sample can ensure the $c^*$ vanishing in probability limit. By estimating $\hat{g}_0$ in the auxiliary sample, and $\varepsilon_0$ in the main sample, the equation would vanish in probability by Chebyshev's inequality.

The application of orthogonalized formulation and the separation of data make the distortion of regularized bias and overfitting bias be accounted for, which forces $\sqrt{n}(\hat{\theta}_0 - \theta_0)$ being asymptotically normally distributed. Notwithstanding we must admit the loss of efficiency when estimating the $\hat{\theta}_0$ because of the loss of data. In our example, we divide dataset into two parts equally, which indicates the loss is half of the dataset. But this problem can be alleviated by flipping the role of the main part and auxiliary samples, where the full efficiency can be restored, especially when it is averaging under the condition that these two estimators are approximately independent.

### 4.2 Odds Weighted Pseudo Inverse Mill's Ratio

In this section, we propose the improvements for traditional IV, which is a key innovation in our method, and make it distinguished from any prior one. Specifically, we come up with an idea of reverse estimation for the unknown impact, which is incurred by the omission of some variables in the first stage's estimates. Furthermore, we adopt the weighted method based on odds instead of conventional probability density estimation.

#### 4.2.1 Introduction to Heckman's Inverse Mill's Ratio

Our improvements are inspired by previous commendable works by Heckman (1979), and we will give a brief introduction to his remarkable achievements.

This is a problem related to sample selection bias. A classical instance can help us better understand. When we investigate the relationship between women's education and their labor returns, it is obvious that we can only collect samples from those who have jobs, and we cannot collect data from those who do not have jobs, which induces the possibility of the problem of underestimation.

In fact, this phenomenon of truncation appears widely. Another example is that the Bureau of Statistics may only collect the information of firms given some conditions. For example, only firms whose income is bigger than c (a unique threshold ) will be included in official statistics. Therefore, there is some lack of statistical data in official surveys if some firms' income is less than c.

Considering a more stochastic condition - the incidental truncation, similarly, we define $Y_i$ as outcome (dependent variable) and define $Z_i$ as a decisive variable. Invoking the question of women's labor returns, we postulate the labor hours follow the below:

$$Hours = \alpha_0 + \alpha_1 wage + \alpha_2 marriage + \epsilon$$

The labor returns follow:

$$w^0 - w^r = \beta_0 + \beta_1 age + \beta_2 education + \beta_3 marriage + \zeta$$

Where $w^0$ indicates the offered wage decided by the market and $w^r$ is the reservation wage. Apparently, women choose to work if and only if $w^0 - w^r \geq 0$, and we cannot observe their labor hours if they do not work, which thus leads to the incidental truncation.

Given random variables (Y, Z), the expectation is $(\mu_Y, \ \mu_Z)$, the standard variation is $(\sigma_Y, \ \sigma_Z)$, the coefficient of correlation is $\rho$, and the Joint probability density function is $f(Y, Z)$. Assuming that whether an individual is selected by selection mechanism of the sample is decided by whether decisive variable Z is larger than a constant c. For example, in the example of women's labor returns, $Z = w^0 - w^r$ and the constant is c. Based on this, we can obtain the conditional expectation of Y under the incidental truncation:

$$E(Y|Z > c) = \mu_Y + \rho\sigma_Y\lambda[(c - \mu_Z)/\sigma_Z]$$

It is clear that the selection process of Z does not have any impact on Y if $\rho = 0$ (Y is independent of Z). Resetting the regression model of $Y_i = X_i'\beta + \varepsilon_i$ and simplifying the Z into a dummy variable, the selection mechanism is written as below:

$$Y_i = \begin{cases} observable, & Z_i = 1 \\ unobservable, & Z_i = 0 \end{cases}$$

And the specification model of $Z_i$ is below, where $Z_i^* = w_i'\gamma + \vartheta_i$:

$$Z_i = \begin{cases} 1, & Z_i^* > 0 \\ 0, & Z_i^* \leq 0 \end{cases}$$

Assuming $\vartheta_i$ obeys normal distribution, the $Z_i$ follows the Probit model, and thus $P(Z_i = 1|w_i) = \Phi(w_i'\gamma)$. The conditional expectation of observable samples is:

$$E(Y_i|Y_i = observable) = E(Y_i|\ Z_i^* > 0) = E(X_i'\beta + \varepsilon_i|w_i'\gamma + \vartheta_i > 0)$$
$$= E(X_i'\beta + \varepsilon_i|\vartheta_i > -w_i'\gamma) = X_i'\beta + E(\varepsilon_i|\vartheta_i > -w_i'\gamma) = X_i'\beta + \rho\sigma_\varepsilon\lambda(-w_i'\gamma)$$

Where $E(\varepsilon_i) = E(\mu_i) = 0$, and we standardize the disturbance error of Probit model to 1. Obviously, the direct estimation via OLS will omit the nonlinear term $\rho\sigma_\varepsilon\lambda(-w_i'\gamma)$ so that we have to follow Heckman two stage to capture this effect. In the first stage of the Heckman method, we estimate the Inverse Mill's Ratio (IMR) and then control it in the second regression. Based on the above instance of dichotomous decisive variable $Z_i$, given the assumption that $Y \sim N(0,1)$, it is easy to see that:

$$E(Y|Y > c) = \frac{\phi(c)}{1 - \Phi(c)}$$

Thereby, given any constant $c \in \mathbb{R}$:

$$\lambda(c) \equiv \frac{\phi(c)}{1 - \Phi(c)}$$

### 4.2.2 Odds and Pseudo Inverse Mill's Ratio

As we discussed pertaining to the invalidity of instrumental variables before, we try to contribute a viable method to tackle it.

Now, revisiting the key factor rendering the invalidity of instrumental variables, which is due to the lack of information on confounders in the first stage even if the second estimate is perfectly captured. This may be inherently caused by the weak features themselves (e.g., a feature may be never considered as a leaf to split samples in random forest because the information gain is much too small), but such an omission of information from this unique feature induces invalidity of instrumental variables. And ultimately, the disturbance error term in the first stage is correlated to confounders in the second stage, for detail please refer to the explanation in section 3.3.

Our innovation lies in the reverse estimation of the impact influenced by such an omission of ignored information. And we rewrite the regression model to better explain. The character $y_i$ indicates the outcome, $Y_i$ indicates the endogenous regressor, $X_i$ represents the vector of confounders, and $Z_i$ is the instrumental variables. We posit a simplified regression model as the final estimator in the first stage obtained by ML algorithm with the omission of the $i^{th}\ X$ :

$$Y_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 X_{j:j=1,2,\dots,i-1} + \varepsilon_i$$

And the perfect captured second stage is:

$$y_i = \beta_0 + \beta_1 \hat{Y}_i + \beta_2 X_i + \vartheta_i$$

Due to the lack of information stemming from $i^{th}\ X$, the $\varepsilon_i$ is associated with $X_i$, and namely, $E(X_i|\varepsilon_i) \neq 0$. And our idea of reverse estimation focuses on the $E(X_i|\varepsilon_i)$, especially the $E(X_i|X_i < \varepsilon_i)$. The logic is not hard to understand. Intuitively, the more valid the estimate of the first stage is, the smaller the $\varepsilon_i$ is, and so does the higher propensity of condition $X_i > \varepsilon_i$, *vice versa*. Logically, the invalid estimates of the first stage swell the $\varepsilon_i$ and cause the condition that $X_i < \varepsilon_i$ so that we should pay more attention to the $E(X_i|X_i < \varepsilon_i)$ because the invalidity is more likely to be more serious in this case.

Based on our intuition, given the assumption that $z \sim N(0,1)$ and $x \sim N(0,\sigma^2)$, we give the derivative as follows, where we assume that z is exogenous random variable following the standard normal distribution and x is exogenous confounders with standard variance of $\sigma$:

$$E(x|x < \varepsilon) = \int_{-\infty}^{\varepsilon} xg(x|x < \varepsilon)dz = \int_{-\infty}^{\varepsilon} x\frac{g(x)}{Prob(x < \varepsilon)}dz = \frac{1}{Prob(x < \varepsilon)}\int_{-\infty}^{\varepsilon} xg(x)dz$$

$$= \frac{1}{\int_{-\infty}^{\varepsilon} g(x)dx}\int_{-\infty}^{\varepsilon} xg(x)dz \overset{z=\frac{x}{\sigma}}{=} \frac{1}{\int_{-\infty}^{\varepsilon/\sigma} \phi(z)dz}\int_{-\infty}^{\varepsilon/\sigma} \sigma z\phi(z)dz$$

$$= \frac{1}{\Phi\left(\frac{\varepsilon}{\sigma}\right)}\sigma\left(-\phi(z)|_{-\infty}^{\frac{\varepsilon}{\sigma}}\right) = -\sigma\frac{\phi\left(\frac{\varepsilon}{\sigma}\right)}{\Phi\left(\frac{\varepsilon}{\sigma}\right)}$$

And we propose and define the Pseudo Inverse Mill's Ratio (PIMR) as follows (we use ξ as the symbol indicating it), which can be used to correct the invalidity IV incurred by the information loss of some features:

$$\xi(\varepsilon) \equiv \phi(\varepsilon)/\Phi(\varepsilon)$$

Finally, our empirical intuition reckons the potential of weighted method. Therefore, we also propose the matching weight method. Proverbially, the most direct idea is to estimate the probability density function of $\varepsilon$ and use it as a weighted base. However, the functionality of the naïve weighted method cannot capture the nuance and the effect is not significant, especially for the potentially minimal residuals. Therefore, we propose the replacement by odds:

$$Odds = \frac{Prob(\cdot)}{1 - Prob(\cdot)}$$

Here we want to share an explanation from the dictionary: odds is the connection between two numbers that shows how much money somebody will receive if they win a bet, e.g., odds of ten to one means that ten times the amount of money that has been bet by somebody will be paid to them if they win. In fact, interestingly, under the premise that we do not know whether invalidity occurs, and the reverse estimation is like a bet where the weighted application is kind a of payment for a bet with high risk, namely just like the word – odds. For this "bet", we match the higher risk (one with the higher probability, and it is more likely due to the omission because most error terms include it of which it may be the systematical omission of some variables) to the higher odds, and match the lower-risk to the lower odds because there is not too much extra benefit deserved to be focused on.

Finally, we defined the odds weighted PIMR as $\Xi(\cdot)$:

$$\Xi(\varepsilon) = \frac{\phi(\varepsilon)}{1 - \phi(\varepsilon)} \xi(\varepsilon)$$

### 4.3 Backward Inference Instrumental Variables

In this section, we summarize and give our methodology, which is similar to the traditional 2SLS, including two stages. To clarify it more straightforwardly, we define some variables for better comprehension as well. Similarly, $y_i$ indicates the outcome, $Y_i$ indicates the endogenous regressor, $X_i$ represents the vector of confounders, and $Z_i$ is the instrumental variables. In view of the 2SLS, the idea can be understood as below:

In the first stage, we employ any presetting ML algorithm to estimate the $Y_i$, where $F_i(\cdot)$ indicates the fitting function via ML:

$$Y_i = F_0(Z_i) + F_1(X_i) + \varepsilon_i$$

Obtaining the $\hat{Y}_i$ and $\hat{\varepsilon}_i$ from the first stage estimates. And then calculating the odds weighted PIMR, where the odds is calculated based on estimated probability given the assumption of normal distribution.

In the second stage, we employ the orthogonal DML algorithm to complete the final estimate of the causal effect of $Y_i$ on $y_i$, in which we add the odds weighted PIMR as confounders as below. As for the detail, we will elaborate on in the next chapter.

$$y_i - \ell_0(X_i, \Xi(\varepsilon_i)) = \left(\hat{Y}_i - f_0(X_i, \Xi(\varepsilon_i))\right)\theta_0 + \mu_i$$

## 5 Backward Inference Instrumental Variables Algorithm

In this chapter, we introduce the Backward Inference Instrumental Variables Algorithm in detail.

Table 1 The BIIV Algorithm Procedure

**Algorithm Procedure: Backward Inference Instrumental Variables (BIIV)**

**1.** Inputting the dictionary of potential parameters for the hyper-parameter space searching. We suggest the base-learner of Gradient Boosting Regressor.

**2.** Predicting the closely pure exogenous treatment $\hat{T}_i$ based on the instrumental variables $Z_i$ and observable confounders $X_i$ as follows:

$$T_i = F(Z_i, X_i) + \varepsilon_i$$

And then obtaining the residuals $\hat{\varepsilon}_i$ for the basis of backward inference. The final specification of model is contingent on the best parameters, and the criterion is MSE.

**3.** Calculating the Odds Weighted Pseudo Inverse Mill's Ratio, which is consistent of two components. First, we calculated the PIMR (namely, the $\xi(\cdot)$) as shown:

$$\xi(\varepsilon_i) = \phi(\varepsilon_i)/\Phi(\varepsilon_i)$$

Then, estimating the Probability Density Function of $\varepsilon_i$ and calculating the odds for weights:

$$Odds = \frac{\phi(\varepsilon_i)}{1 - \phi(\varepsilon_i)}$$

And the Odds Weighted Pseudo Inverse Mill's Ratio is:

$$\Xi(\varepsilon) = \frac{\phi(\varepsilon_i)}{1 - \phi(\varepsilon_i)} \xi(\varepsilon_i)$$

**4. For** iteration in (0, M):

> a. Splitting the sample into two equal sized parts randomly: the main part (sample size N/2) and the auxiliary part (sample size N/2).
>
> b. Applying any ML to estimate $\ell_0$ and $f_0$ on the auxiliary part sample.
>
> $$y_i = \ell_0\big(X_i, \Xi(\varepsilon_i)\big) + \vartheta_i$$
>
> $$\hat{Y}_i = f_0\big(X_i, \Xi(\varepsilon_i)\big) + \tau_i$$
>
> c. Employing on the main part sample and get the coefficient $\theta'_0$
>
> $$y_i - \ell_0\big(X_i, \Xi(\varepsilon_i)\big) = \Big(\hat{Y}_i - f_0\big(X_i, \Xi(\varepsilon_i)\big)\Big)\theta'_0 + \mu_i$$
>
> d. Flipping the main part and auxiliary part and repeat the step b to c, and then obtaining the coefficient $\theta''_0$ via cross-fitting strategy. The final $\theta_0$ is the average of $\theta'_0$ and $\theta''_0$

**End**

**5.** Stopping after the iteration M-1, and M is suggested of 100 (Hansen & Siggaard, 2022). For M estimates ($\theta^{(0)}{}_0$, $\theta^{(1)}{}_0$, $\theta^{(2)}{}_0$, …, $\theta^{(M-1)}{}_0$), the median of these is the final estimate of $\theta_0$

**Note**: given $y_i$ indicates the outcome, $T_i$ indicates the endogenous regressor (treatment ), $X_i$ represents the vector of confounders, $Z_i$ is the instrumental variables, and Sample Size is N.

In totality, the whole procedure can be broken down into three stages:

1. Predicting the endogenous treatment variable based on the instrumental variables via machine learning, which is a task aiming at prediction. Simultaneously, calculating the residuals of the treatment variable.
2. Calculating the Pseudo Inverse Mill's Ratio (PIMR) based on the residuals acquired in the first stage. And then, we calculate the Probability Density Function so that we can obtain the odds for weighted process.
3. Following the orthogonal DML method, estimating the Average Treatment Effect (ATE).

## 6 Monte Carlo Simulation and Results

In this chapter, we employ extensive Monte Carlo Simulations to attest to the functionality of BIIV, in contrast to the direct OLS, IV based on 2SLS, and Orthogonal DMLIV. This section is composed of two sections, one is the description of the generation of simulated datasets, and another one is the explanation of Monte Carlo simulated results.

### 6.1 Generation of Simulated Datasets

In view of the various aims of tests, we design two procedures for different tests, including one based on a single observable confounder (namely, the satisfaction of unconfoundedness), and another based on multiple unobservable confounders (violation of unconfoundedness so that the DMLIV may be invalid).

For datasets with a single observable confounder, we set two key parameters to indicate the extension of confoundedness and to which degree the violation of the assumption of conventional assumption: the covariance between the treatment variable (T) and instrumental variable (Z) as well as the covariance between the treatment variable (T) and the confounder (X). In the classical IV method, the strict assumption requires the independence of instrumental variables relative to the confounders. In this dataset with a single observable confounder, we mainly examine the validity of BIIV facing the correlation between confounders and instrumental variables, where we can also compare it directly to orthogonal DMLIV.

Next, for datasets with multiple unobservable confounders, we simulate the causal path and finally randomly select fewer confounders as observable so that simulating the condition where the unobservable confounders exist. As aforementioned, one of the flaws of orthogonal DMLIV is the rigid assumption of unconfoundedness, which requires the omniscience of all confounders even if it is clearly seen that is impossible. Neither the expensive field experiments following the RCM causal framework nor the exquisite design of quasi-experiments can identify all confounders in causal inference, which induces the doubt for credence behind these methodologies. As we emphasized above, in the belief of the inevitable existence of unobservable confoundedness, we propose the backward inference from the residuals in the first stage, which is the base of BIIV methodology. Facing the condition of confoundedness, the comparison between BIIV and other methodologies will be conducted on this simulated dataset.

### 6.2 Simulated Results

In this part, we report the results of our Monte Carlo simulation, which is conducted in two types of datasets. To keep the comparability of our results before and after, we set the gradient boosting regressor as the base learner of BIIV, and there is no change in case of no extra explanation.

#### 6.2.1 Simulated Results in the Datasets with a Single Observable Confounder

Figure 2 shows the simulated results, where the horizontal axis is the estimated average treatment effect and the vertical axis is the estimated density calculated by Kernel Density Estimation, and thus, the estimated probability is the product of the estimated density and the length of each bar in the histogram.

Without violating the basic assumption of the IV method, we can see that: the direct OLS cannot obtain the exact estimate of average treatment effect (ATE) due to the endogeneity of treatment variable; compared to BIIV and orthogonal DMLIV (thereafter as DMLOIV), the traditional estimate of ATE obtained from the 2SLS IV also shows the deviation to some extent, but in general, it can guarantee the credible consistency; ultimately, related to methodologies that are in equipped by ML algorithm, both DMLOIV, and BIIV highlight the functionally strong estimates of ATE. One that may deserve our attention is that the ATE for those two methodologies using machine learning for prediction is the conditional average treatment effect (CATE) because the impact of confounders is taken into account as an expectation given unique conditions.

Overall, under the premise that the traditional IV method, the Monte Carlo simulated results indicate that the BIIV and DMLOIV are on par with the effective estimates by 2SLS IV, whereas the within-group variance of BIIV seems too large for better convergence except for the decent consistency.

The picture below in figure 2 demonstrates the simulated results when the 2SLS IV is ineffective. In detail, the data violates the basic assumption that the instrumental variable is independent of confounders, namely, the exogenous instrumental variable.

First, the conventional econometric methodologies (OLS and 2SLS IV) cannot identify the ATE at all. According to the KDE, OLS gives estimates of about 6.7, and 2SLS IV causes extremely deviated estimates of about 8.8, too far away from the true ATE of 5. Our simulated results show the terrible fact that 2SLS estimates may not only mitigate the distortion of confounders on real causal relation but also enhance this trend in cases where the identification is not logically reasonable or ignored.

Second, also in these simulated results, methodologies related to the application of ML perfectly outperform any conventional econometric one. Our estimates of BIIV and DMLOIV demonstrate great consistency close to the true ATE. Admittedly, BIIV still shows the weakness of huge within-group variance relative to DMLOIV, but we can see that the huge within-group variance has been improved. In fact, notwithstanding, we believe this is an appropriate advantage. For the balance between the variance and bias, the contentious discussion focuses on the generalized error and the final models determined. A consensus has been achieved that the increase of variance to some extent can alleviate the overfitting problem, which may be a testimony that why our model can finally outperform the DMLOIV even if the convergence of variance is relatively worse than DMLOIV.

In totality, the simulated results in the second exhibition show that DMLOIV and BIIV perfectly outperform the conventional econometric methodologies. The "black box heuristic" fusion could assure that the methodologies related to ML are on par with any econometric one, and almost outperform those.

### 6.2.2 Simulated Results in the Datasets with Unobservable Confounders

This section examines the power of BIIV under the premise that not all confounders are observable. In other words, the omission of confounders is inevitable, which opposes the strict assumption of DMLOIV, and so do the OLS and 2SLS IV to some extent.

In figure 3, we obtain the ATE from the simulated dataset based on the unobservable confounders. As per the KDE results, in general, BIIV best approximates the true presetting ATE of 5, though the final estimates leftward deviate from the true effect. In contrast, other methodologies are obviously rightward deviated from the true effect, where DMLOIV and OLS exhibit the analogous fitting and IV seems worse than the prior two.

Also, in this figure, the within-group variance of BIIV further decreases. Similar to the aforementioned, we believe the difference of variance between BIIV and others is not a problem, which can increase the generalizability of BIIV model and guarantee an estimate of out-of-sample causal relation. In fact, we hope the trained model can acquire a strong ability to infer the cause and effect in the real world.

In addition, in figure 4, we also report the violin plot to directly show the distribution of simulated ATE. As we mentioned above, BIIV has the proclivity to underestimate the causal intensity whereas others are inclined to overestimate. But such an inherent characteristic of BIIV estimates seems less risky yet admittedly "observative" compared to the overestimation. Furthermore, BIIV demonstrates strong consistency of median and average, which means robustness to some degree. Ultimately, the estimated interval cannot efficiently converge as other methodologies, which induces a relatively longer and overlapped tail distributed on both two sides.

Last, compared to the sample size in research about causal inference, which in our simulated datasets is only 1000. However, our algorithm still can provide the robust causal inferential estimator. Relative to the consistency guaranteed by large sample size, our algorithm offer the potential innovation in small a small one.

## 6.3 Pertinent Discussion

Figures 5 and 6 help us with a further investigation related to the MSE in the first stage and the Odds Weighted Pseudo Inverse Mill's Ratio. We record both of them and explore the latent relationship between them.

According to figure 5, we scatter the MSE, Odds Weighted Pseudo Inverse Mill's Ratio (we calculate the average in each dataset for comparison, thereafter as PIMR_mean), and the deviation from the true ATE in a three-dimensional space. We conjecture a connection between those, which may enlighten us pertaining to the impact of the fitted model on the estimates for ATE. However, the empirical evidence twists, and there is no apparent symptom indicating that ex-ante conjecture. Also, from figure 6, we fabricate the two-dimensional connection between MSE and PIMR_mean, and then show them with the deviation from ATE in a three-dimensional plot. The final plot of the wireframe figure shows a mess with tremendous haphazard overlap.

Our investigation shows that the model specification in the first stage has little impact on the final estimation, namely, the mess does not matter. Therefore, this may be evidence that the specification of the model in econometrics is worse than ML algorithm regarding the above discussion and exhibition.

## 7 Conclusion

In this paper, we propose an innovative algorithm called Backward Inference Instrumental Variables (BIIV) for causal inference. Our works build upon the contributions of conventional Instrumental Variables as well as orthogonal Double Machine Learning. Given the extensive Monte Carlo simulations, in general, the BIIV is on par with the orthogonal DML and outperforms traditional econometric methodologies.

On the one hand, our work follows the advanced application of machine learning algorithms for causal inference in social science. Specifically, we draw on the advantages of nonparametric algorithms, which relax the traditional model specification and deeply mine the "truth" from data themselves. Simultaneously, the adoption of the idea that orthogonal Double Machine Learning can mitigate the loss due to the deviation from $\sqrt{n}$-consistent and asymptotically normal estimates incurred by high-dimensional risk.

On the other hand, we innovate the idea that backward inference to capture the omission of unobservable confounders, which is a breakthrough that relaxes the strict assumption for almost all econometric methodologies as well as machine learning methodologies. The process of backward inference can still guarantee the required consistency under a concession for classical restriction, and this idea is logically intuitive for the explanation.

For future direction, currently, our algorithm focuses on the estimates of continuous variables, and we are devoid of knowledge about the effect of dichotomous variables. And we believe the spirits from Athey et al. can inspire the further development of BIIV. This would be worth doing in the future.

## References

Acemoglu, D., Johnson, S. and Robinson, J.A., 2001. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5), pp.1369-1401.

Andrews, D.W., 1994. Asymptotics for semiparametric econometric models via stochastic equicontinuity. Econometrica: Journal of the Econometric Society, pp.43-72.

Angrist, J.D. and Krueger, A.B., 1991. Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, *106*(4), pp.979-1014.

Angrist, J.D. and Pischke, J.S., 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J.D. and Pischke, J.S., 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, *24*(2), pp.3-30.

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

Arora, N., Bohn, J.R. and Zhu, F., 2012. Reduced-Form versus Structural Models of Credit Risk: A Case Study of Three Models. *The Credit Market Handbook: Advanced Modeling Issues*, pp.132-164.

Athey, S. and Imbens, G.W., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), pp.7353-7360.

Athey, S. and Imbens, G.W., 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), pp.431-497.

Basmann, R.L., 1957. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica: Journal of the Econometric Society*, pp.77-83.

Belloni, A., Chernozhukov, V. and Wei, Y., 2016. Post-selection inference for generalized linear models with many controls. Journal of Business & Economic Statistics, 34(4), pp.606-619.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C., 2015. Program evaluation with high-dimensional data (No. CWP55/15). cemmap working paper.

Berkowitz, D., Caner, M. and Fang, Y., 2012. The validity of instruments revisited. *Journal of Econometrics*, *166*(2), pp.255-266.

Bound, J., Jaeger, D.A. and Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), pp.443-450.

Breiman, L., 1996. Bagging predictors. Machine learning, 24, pp.123-140.

Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

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

Cragg, J.G. and Donald, S.G., 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, *9*(2), pp.222-240.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A., 2008. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics, 90(3), pp.389-405.

der Vaart, V. and Statistics, A.A., 1998. Cambridge University Press: New York. NY, USA.

Duffie, D. and Lando, D., 2001. Term structures of credit spreads with incomplete accounting information. *Econometrica*, *69*(3), pp.633-664.

Efron, B. and Hastie, T., 2021. Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press.

Fix, E. and Hodges, J.L., 1951. Discriminatory analysis, nonparametric discrimination.

Frisch, R. and Waugh, F.V., 1933. Partial time regressions as compared with individual trends. Econometrica: Journal of the Econometric Society, pp.387-401.

Goldberger, A.S., 1972. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pp.979-1001.

Granger, C.W., 1988. Some recent development in a concept of causality. *Journal of econometrics*, *39*(1-2), pp.199-211.

Green, D.P. and Kern, H.L., 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public opinion quarterly, 76(3), pp.491-511.

Groves, T., Hong, Y., McMillan, J. and Naughton, B., 1994. Autonomy and incentives in Chinese state enterprises. *The Quarterly Journal of Economics*, *109*(1), pp.183-209.

Hansen, J. and Siggaard, M., DOUBLE MACHINE LEARNING: EXPLAINING THE POST-EARNINGS ANNOUNCEMENT DRIFT. *Cecilie. Thank you for everything.*, p.55.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pp.1029-1054.

Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp.153-161.

Hill, J. and Su, Y.S., 2013. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. The Annals of Applied Statistics, pp.1386-1420.

Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), pp.55-67.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C. and Schölkopf, B., 2020. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, *21*(1), pp.3482-3534.

Imbens, G.W. and Lemieux, T., 2008. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), pp.615-635.

Imbens, G.W. and Rubin, D.B., 2015. *Causal inference in statistics, social, and biomedical* *sciences*. Cambridge University Press.

Jackson, C.K., Johnson, R.C. and Persico, C., 2015. *The effects of school spending on educational and economic outcomes: Evidence from school finance reforms* (No. w20847). National Bureau of Economic Research.

Jarrow, R.A. and Protter, P., 2012. Structural versus Reduced-Form Models: A New Information-Based Perspective. *The Credit Market Handbook: Advanced Modeling Issues*, pp.118-131.

Lee, M.J., 2009. Non-parametric tests for distributional treatment effect for randomly censored responses. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(1), pp.243-264.

Levit, B.Y., 1976. On the efficiency of a class of non-parametric estimates. Theory of Probability & Its Applications, 20(4), pp.723-740.

Levitt, S.D. and Snyder Jr, J.M., 1997. The impact of federal spending on House election outcomes. *Journal of political Economy*, *105*(1), pp.30-53.

Li, B., Friedman, J., Olshen, R. and Stone, C., 1984. Classification and regression trees (CART). Biometrics, 40(3), pp.358-361.

Lovell, M.C., 1963. Seasonal adjustment of economic time series and multiple regression analysis. Journal of the American Statistical Association, 58(304), pp.993-1010.

Lovell, M.C., 2008. A simple proof of the FWL theorem. The Journal of Economic Education, 39(1), pp.88-91.

Luedtke, A.R. and Van Der Laan, M.J., 2016. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Annals of statistics, 44(2), p.713.

Miguel, E., Satyanath, S. and Sergenti, E., 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of political Economy*, *112*(4), pp.725-753.

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J. and Schölkopf, B., 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, *17*(1), pp.1103-1204.

Newey, W.K., 1990. Semiparametric efficiency bounds. Journal of applied econometrics, 5(2), pp.99-135.

Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica: Journal of the Econometric Society, pp.1349-1382.

Neyman, J., 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, *10*(1), pp.1-51.

Neyman, J., 1959. Optimal asymptotic tests of composite hypotheses. Probability and statsitics, pp.213-234.

Neyman, J., 1979. C (α) tests and their use. Sankhyā: The Indian Journal of Statistics, Series A, pp.1-21.

Pearl, J., 2009. *Causality*. Cambridge university press.

Quinlan, J.R., 1986. Induction of decision trees. Machine learning, 1, pp.81-106.

Quinlan, J.R., 1996, August. Bagging, boosting, and C4. 5. In Aaai/Iaai, vol. 1 (pp. 725-730).

Riquelme, A., Berkowitz, D. and Caner, M., 2013. Valid tests when instrumental variables do not perfectly satisfy the exclusion restriction. *The Stata Journal*, *13*(3), pp.528-546.

Robinson, P.M., 1988. Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society, pp.931-954.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), p.688.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp.34-58.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M., 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448), pp.1096-1120.

Schölkopf, B., 2022. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (pp. 765-804).

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R., 2013. A sparse-group lasso. Journal of computational and graphical statistics, 22(2), pp.231-245.

Stock, J.H. and Yogo, M., 2002. Testing for weak instruments in linear IV regression.

Stock, J.H., Wright, J.H. and Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), pp.518-529.

Theil, H., 1953. Repeated least squares applied to complete equation systems. *The Hague: central planning bureau*.

Thurman, W.N. and Fisher, M.E., 1988. Chickens, eggs, and causality, or which came first. *American journal of agricultural economics*, *70*(2), pp.237-238.

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

Timmins, C. and Schlenker, W., 2009. Reduced-form versus structural modeling in environmental and resource economics. *Annu. Rev. Resour. Econ.*, *1*(1), pp.351-380.

Van Der Laan, M.J. and Rubin, D., 2006. Targeted maximum likelihood learning. The international journal of biostatistics, 2(1).

Van der Laan, M.J., Polley, E.C. and Hubbard, A.E., 2007. Super learner. Statistical applications in genetics and molecular biology, 6(1).

Wager, S. and Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), pp.1228-1242.

Waldman, M., Nicholson, S. and Adilov, N., 2006. Does television cause autism?.

Wang, H., Li, R. and Tsai, C.L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, 94(3), pp.553-568.

Wooldridge, J., 2007. What's new in econometrics? Lecture 10 difference-in-differences estimation. *NBER Summer Institute, available at: www. nber. org/WNE/Slides7–31–07/slides_10_diffindiffs. pdf, accessed April*, *9*(2011), p.85.

Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty.

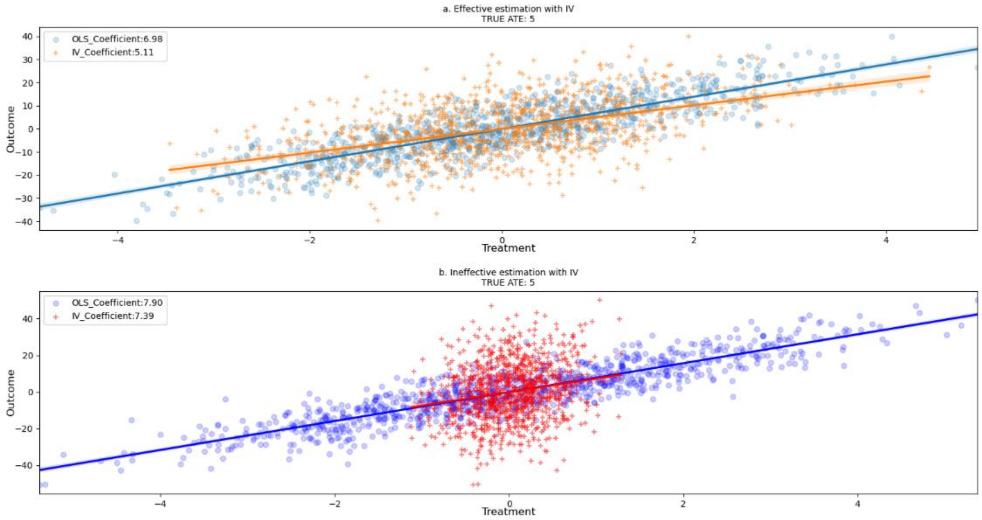
Zheng, W., Luo, Z. and van der Laan, M.J., 2018. Marginal structural models with counterfactual effect modifiers. The international journal of biostatistics, 14(1).

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), pp.301-320.

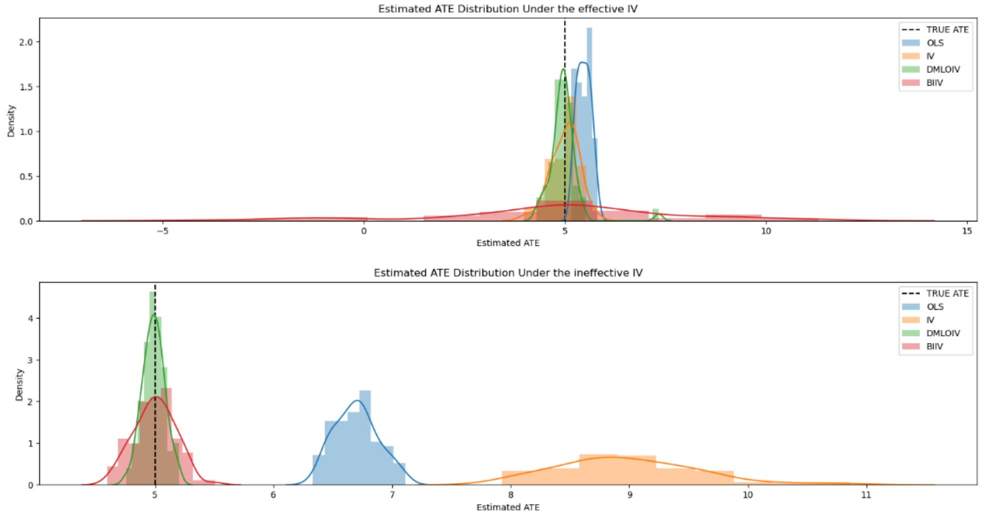
Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), pp.1418-1429.

# Appendix

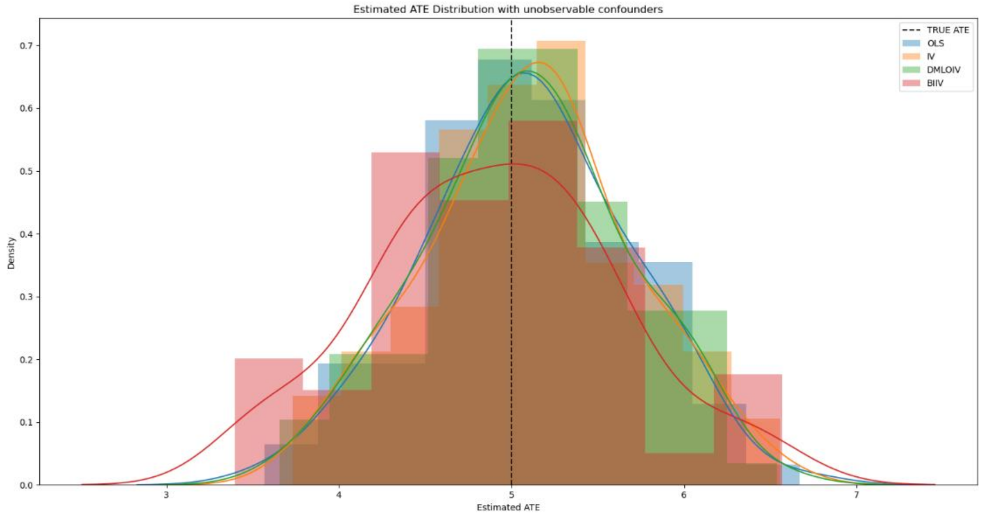
## Figure 1 Comparison between IV and OLS



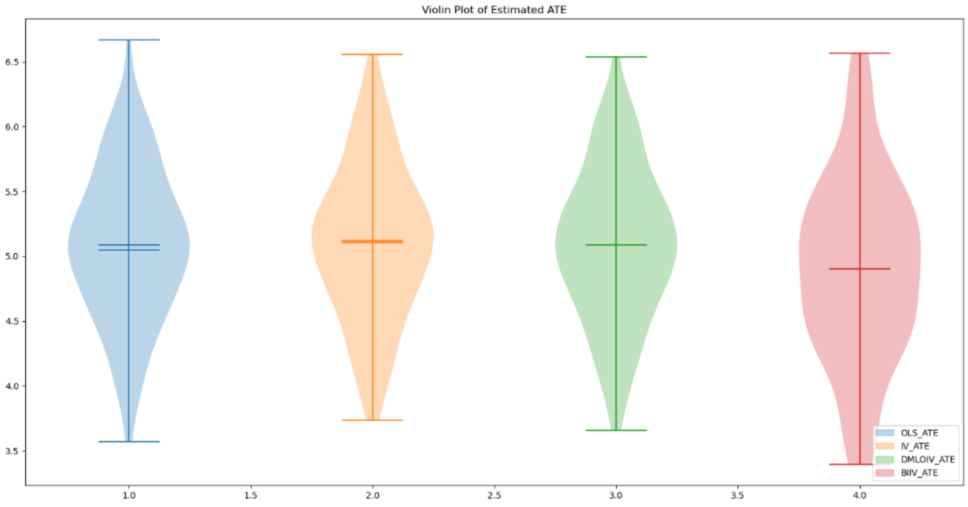
## Figure 2 Simulated Distribution of Estimated ATE with Observable Confounders



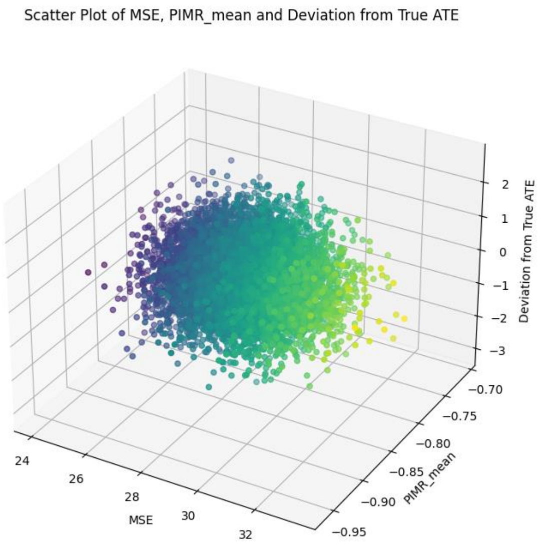
## Figure 3 Simulated Distribution of Estimated A TE with Unobservable Confounders



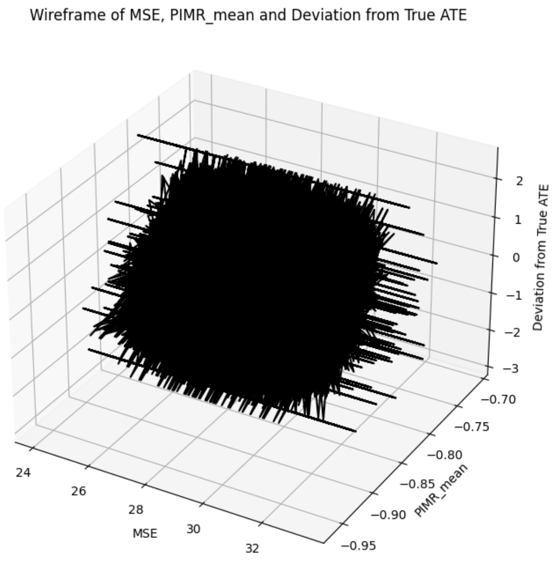
## Figure 4 Violin Plot of Estimated ATE



## Figure 5 Scatter of MSE, Estimated Odds Weighted PIMR, and Deviation from ATE



## Figure 6 Wireframe Plot of MSE, Estimated Odds Weighted PIMR, Deviation from ATE


Wireframe of MSE, PIMR_mean and Deviation from True ATE
MSE
PIMR_mean
Deviation from True ATE