

双重机器学习的理论与应用

——从“黑箱”到“工具箱”的实践指南

胡诗云 江弘毅 解海天*

摘要: 基于线性回归的因果效应估计难以捕捉经济数据中的高维特征与非线性关系。双重机器学习允许研究者在灵活控制高维复杂协变量的同时,获得对因果效应的稳健估计和推断。本文旨在为实证研究者提供关于双重机器学习的理论指引与实践指南。本文系统性地回顾了双重机器学习的理论脉络,并阐述了其与工具变量、双重差分、断点回归等经典识别策略的结合。进一步通过数值模拟讨论了双重机器学习的性能表现与适用边界,并以评估“精准扶贫”对农村居民增收的政策效应为例,展示了双重机器学习在实证研究中的实施流程以及在模型选择、诊断性检验等方面的注意事项。本文为实证研究者正确理解和应用双重机器学习方法提供了参考。

关键词: 双重机器学习 因果推断 高维数据 计量方法 实证研究

一、引言

识别、估计因果效应并给出推断,是计量经济学的核心任务之一。传统的计量经济学方法发展了控制变量、工具变量、双重差分以及断点回归等识别策略,为因果推断提供了严谨而有效的理论框架。然而,在估计过程中,研究者往往需要额外施加对函数形式的线性假定。随着大数据时代的到来,研究者能够获取的微观数据日益丰富,控制变量(协变量)的维度急剧增加,变量间潜在的非线性与交互关系也变得愈发复杂。在此背景下,传统线性模型不仅面临着模型误设风险(林梦芸等,2025),还可能在处理高维协变量时陷入多重共线性,导致因果效应的估计量不

* 胡诗云,博士研究生,北京大学国家发展研究院,电子邮箱:syhu2017@nsd.pku.edu.cn;江弘毅,博士研究生,北京大学国家发展研究院,电子邮箱:hyjiang2017@nsd.pku.edu.cn;解海天(通讯作者),助理教授,北京大学光华管理学院,电子邮箱:xht@gsm.pku.edu.cn。本文获得国家自然科学基金项目(72403008,72495123)的资助。本文未使用AI。感谢匿名审稿专家的宝贵意见,文责自负。

稳定甚至不一致,影响研究结论的可靠性。

面对这一挑战,双重机器学习(Double Machine Learning, DML)应运而生,为在高维、非线性世界中进行稳健的因果推断提供了解决方案。自从Chernozhukov等(2018)的开创性论文以来,越来越多的实证研究开始采用双重机器学习。在中文社会科学引文索引(Chinese Social Sciences Citation Index, CSSCI)期刊上,使用双重机器学习的文献数量自2021年起以接近年均100%的速度增加。此外,随着中国研究者越来越多地关注这一方法,陈茁和陈云松(2025)对其在社会科学中的应用做了初步的概述。这一新方法的引入,一方面为研究中国问题提供了更多高质量证据,另一方面也引发了盲目崇拜、误解和误用。如何正确地理解和使用双重机器学习,如何更好地将双重机器学习与已有数据和实证策略结合,以及如何认识双重机器学习在因果推断中扮演的角色,这些问题都亟待澄清。

本文旨在系统性地回答上述问题,为实证研究者提供一份关于双重机器学习的理论指引与实践指南,核心观点是:双重机器学习并非一种与工具变量、双重差分等相提并论的新型因果识别策略,而是一种服务于既有识别策略的、强大的高维非线性统计工具。在潜在结果(Potential Outcome)框架中,因果效应是不同处理状态下的潜在结果之差(Rubin, 2005)。定义 Y 为个体的可观测结果, D 为个体的处理变量, $D = 1$ 表示个体接受处理,属于处理组; $D = 0$ 表示个体未接受处理,属于控制组。 $Y(1)$ 和 $Y(0)$ 分别为个体接受处理或不接受处理对应的潜在结果,从而有 $Y = DY(1) + (1 - D)Y(0)$, $Y(1) - Y(0)$ 为个体的异质性因果效应。识别(Identification)是指通过对反事实的数据生成过程进行一定的假设(“识别假设”),找到从可观测总体分布唯一地还原出因果效应的方法。随后的估计(Estimation)则是给定从总体分布中获得的样本,通过统计分析得到因果效应的估计量。双重机器学习可以在存在高维、非线性以及非结构化数据时,构造表现更好的统计量,而非通过改变或弱化识别假设。它在保留机器学习的灵活性、降低估计波动的同时,通过额外的去偏手段避免了机器学习的偏误向因果系数估计量的传导,从而将机器学习这一强大的预测“黑箱”转化为估计因果效应过程中的有力“工具箱”。

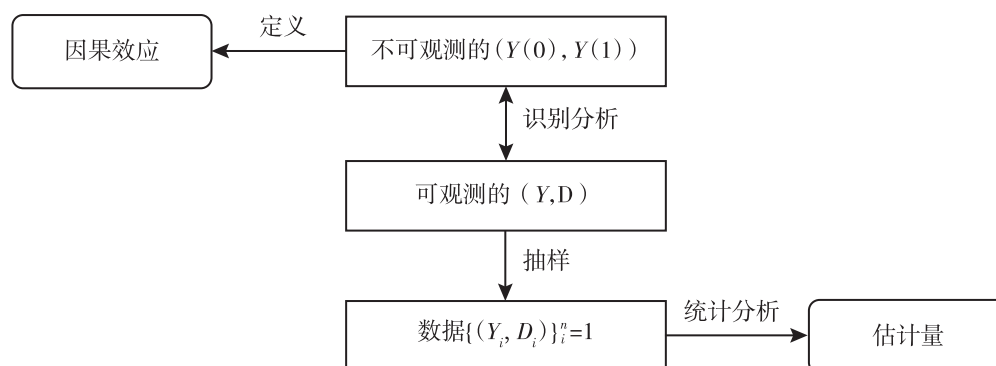


图1 “识别—估计分离原则”

本文首先在第二部分系统回顾了双重机器学习的理论脉络,从部分线性模型(Partial Linear Model, PLM)拓展至更灵活的交互回归模型(Interactive Regression Model, IRM),最终在矩估计的框架下讨论其一般性质。在第三部分,本文探讨了双重机器学习如何与工具变量(Instrumental Variable, IV)、双重差分(Difference-in-differences, DID)、断点回归(Regression Discontinuity Design, RDD)等经典识别策略结合,在不改变核心识别假设的前提下,增强模型的灵活性与稳健性。第四部分总结已有实证文献中的常见操作。第五部分通过数值模拟和“精准扶贫”政策评估的实证案例,演示双重机器学习的适用边界和实施流程。附录详细说明了具体算法以及在模型选择、诊断性检验中的注意事项,并提供可复制的代码。

本文的主要贡献在于:第一,厘清定位与祛除误用。在双重机器学习方法日益普及但易被误解的背景下,本文系统梳理其理论基础,明确其核心作用在于改善估计而非提供识别,旨在帮助研究者建立“先识别、后估计”的严谨范式,避免将其误用为解决内生性问题的“万能钥匙”。第二,搭建理论与实践的桥梁。本文不仅介绍了双重机器学习的理论框架,更通过数值模拟和完整的实证案例,提供了一套具有可操作性的实践指南,覆盖了从模型选择、软件实现到结果解读的全过程,为研究者在复杂的中国情境下正确、高效地应用这一方法提供了参考。

二、双重机器学习的理论基础

双重机器学习是一种将机器学习用于因果系数估计的统计技术。机器学习泛指一套用于高维数据预测的统计方法,通过正则化等技术自动进行模型选择并防止过拟合(Gu等,2020)。在绝大多数情况下,机器学习模型的系数本身并无因果解释。为了将机器学习用于因果推断,双重机器学习将因果推断中统计模型的参数区分为两类:因果系数(Causal Parameters),如平均处理效应(Average Treatment Effect, ATE);以及冗余参数(Nuisance Parameters),如控制变量对处理和结果的影响。双重机器学习利用机器学习的强大预测能力去拟合冗余参数,再基于“内曼正交性”(Neyman Orthogonality)和“交叉拟合”(Cross-fitting)两大理论支柱,阻止其拟合过程中的偏误向因果系数传导,从而对机器学习“取灵活性之长、避方差大与过拟合之短”,实现对因果系数的稳健估计和推断。

(一)部分线性模型

含控制变量的线性回归假定了控制变量对结果的影响是线性的。一种最直接的拓展,就是让控制变量以任意函数形式进入回归方程的同时,依然假设 D 与 X 的效应是线性可加的。与常见的线性回归一样,识别条件为误差项的零条件均值:

$$Y = \beta D + g(X) + u, E(u|X, D) = 0 \quad (1)$$

一个自然的估计思路是用机器学习模型近似 $g(\cdot)$,再使用最小二乘法估计 β 。

但是,由于 g 可能存在高度非线性,直接估计可能导致严重的偏误,而这会进一步传导至对于 β 的估计,导致其无法按一般的速率收敛至真实值。对此的解决方案是首先去除掉 D 和 Y 中能够被 X 解释的部分,然后使用 Y 关于 X 的残差 \check{Y} 对 D 关于 X 的残差 \check{D} 回归,得到 β 的估计。将等式两边分别减去关于 X 的条件期望,得到:

$$Y - E(Y|X) = \beta(D - E(D|X)) + u - \underbrace{E(u|X)}_{=0} \quad (2)$$

记 $m(X) = E(Y|X)$, $l(X) = E(D|X)$, $\check{Y} = Y - m(X)$, $\check{D} = D - l(X)$, 则:

$$\check{Y} = \beta\check{D} + u \quad (3)$$

本文只关心 β 的数值而并不关心 m 和 l 的具体形式,这两个函数就属于冗余参数^①。可以使用机器学习对这两个函数进行估计,然后计算拟合的残差 \check{Y} 和 \check{D} 。由于仍然有 $E(u\check{D}) = 0$, 利用这个矩条件,用 \check{Y} 和 \check{D} 做线性回归,即可得到 β 的估计。

然而,上述想法仍然存在不足,原因在于机器学习可能对数据过拟合,使得残差估计存在偏误。考虑一个极端情形:假如机器学习完美地拟合了 (Y, D) 每一个数据点,那么残差将全部等于 0,第二步中残差对残差的回归将毫无意义。因此,必须引入交叉拟合以缓解这种情况。具体的做法如算法 2-1-1 所述。

算法 2-1-1(部分线性模型的双重机器学习)

1.(交叉拟合)将数据划分为 K 个互不相交的子样本。对于第 k 个子样本,使用其余 $K - 1$ 个子样本:

第一利用机器学习算法学习 $E(Y|X)$, 得到 $\hat{m}_{[k]}$ 。

第二利用机器学习算法学习 $E(D|X)$, 得到 $\hat{l}_{[k]}$ 。

2.对于每个样本 $i \in \{1, 2, \dots, n\}$, 其所处的子样本为 $k(i)$, 计算交叉拟合残差

$$\check{Y}_i = Y_i - \hat{m}_{[k(i)]}(X_i), \check{D}_i = D_i - \hat{l}_{[k(i)]}(X_i)$$

3.使用 \check{Y}_i 对 \check{D}_i 做线性回归, β 的标准误、置信区间和统计推断与普通的线性回归相同。

Chernozhukov 等(2018)证明,在一定的条件下^②, 算法 2-1-1 所得到的估计量具有一致性和渐近正态性,并且可以按照一般流程进行统计推断。

① 在计量经济学理论中,函数也可以视为一种高维参数。

② 这些条件包括: $\{Y_i, D_i, X_i\}_{i=1}^n$ 是独立同分布的样本; Y 和 D 的高阶矩(大于 4)有界且 $\text{Var}(Y|X)$ 和 $\text{Var}(D|X)$ 有界; $Y - m(X)$ 和 $D - l(X)$ 相关且 $D - l(X)$ 存在波动; \hat{m} 和 \hat{l} 是对真实函数 m 和 l 足够好的近似:对于 $(\hat{\eta}, \eta) \in \{(\hat{m}, m), (\hat{l}, l)\}$, 随着样本量增大,有 $\sqrt{E\left[\int (m(x) - \hat{m}(x))^2 f_X(x) dx\right]} + \sqrt{E\left[\int (l(x) - \hat{l}(x))^2 f_X(x) dx\right]} = o\left(\frac{1}{n^{1/4}}\right)$ 。

(二)交互回归模型

部分线性模型允许了 $g(X)$ 的高度非线性,但仍然假设了 D 与 $g(X)$ 的效应是线性可加的;无论 X 的取值如何,处理变量对于结果的边际因果效应都是 β 。交互回归模型基于引言中的潜在结果框架,放松了这一假设。

一般而言,因果识别会关心平均处理效应, $ATE = E(Y(1) - Y(0))$,以及处理组的平均处理效应(Average Treatment Effect for the Treated, ATT), $ATT = E(Y(1) - Y(0)|D = 1)$ 。但由于不能同时观测到同一个个体的 $Y(1)$ 和 $Y(0)$,因此在不施加任何假设的情况下,无论是 ATE 还是 ATT 一般都不能被识别。如果条件于协变量 X ,处理变量 D 在不同个体之间的分配是“随机”的;且对于不同的 X 的取值,个体都会存在处理和控制两种状态,那么 ATE 就可以被识别,具体的识别假设为:

假设 1(交互回归模型的识别假设)

1. 可忽略性假设(条件独立性假设): $(Y(1), Y(0)) \perp D | X$
2. 共同支撑假设(重叠性假设): $P(D = 1|X) \in (0, 1)$ 几乎处处成立

可忽略性假设允许在 X 的取值范围上识别条件平均处理效应(Conditional Average Treatment Effect, CATE), $CATE(x) = E(Y(1) - Y(0)|X = x)$,而共同支撑假设则表明 X 的取值范围不会随着 D 的改变而改变,这允许在 X 的整个取值范围上识别 CATE,从而对 CATE 积分得到 ATE。在假设 1 之下,ATE 有两种识别方式:基于结果模型(Outcome Model)的因果识别,以及基于逆概率加权(Inverse Probability Weighting, IPW)的因果识别。

如果基于结果模型进行因果识别,定义 $m_1(x) := E(Y|D = 1, X = x)$ 和 $m_0(x) := E(Y|D = 0, X = x)$,分别为不同处理状态下结果变量关于协变量的条件期望,那么 ATE 可以写作:

$$ATE = \int (m_1(x) - m_0(x)) f_X(x) dx \tag{4}$$

其中, f_X 为 X 的概率密度函数。给定样本可以首先使用机器学习的方式估计出 \hat{m}_1 和 \hat{m}_0 ,再代入式(4),得到 $\widehat{ATE} = n^{-1} \sum_{i=1}^n (\hat{m}_1(X_i) - \hat{m}_0(X_i))$ 。

另一种方式基于逆概率加权。定义 $p(x) = P(D = 1|X = x)$ 为给定协变量 X 的取值,个体接受处理的概率(“倾向值”, Propensity Score, PS),那么:

$$ATE = E\left(\frac{DY}{p(X)}\right) - E\left(\frac{(1-D)Y}{1-p(X)}\right) = E\left(\frac{(D-p(X))Y}{p(X)(1-p(X))}\right) \tag{5}$$

可以用机器学习的方式估计出 \hat{p} ,再代入式(5),得到 $\widehat{ATE} = n^{-1} \sum_{i=1}^n \frac{(D_i - \hat{p}(X_i))Y_i}{\hat{p}(X_i)(1 - \hat{p}(X_i))}$ 。

假如如同“先知”一般获得了“上帝的神谕(Oracle)”,知道了 m_1 、 m_0 或者 p 的真

实函数形式,那么无论是基于式(4)还是式(5)进行估计,其结果都是一致且渐进正态的^①。但在现实中,这些冗余参数都要首先使用机器学习进行估计,而这就会产生偏误,而这又会直接影响到ATE估计量,使其无法以一般的速率收敛到真实值。事实上,本文并不关心 m_1 、 m_0 或者 p 的具体形式,因为这部分冗余参数的估计误差影响了真正关心的ATE估计是得不偿失的。

有没有一种识别式能够将式(4)和式(5)两种识别式结合起来,使得冗余参数估计量的偏误对ATE估计量的概率极限和渐近方差没有影响,从而降低模型误设和过拟合的风险?答案就是双重稳健估计量(Doubly Robust Estimator)。定义

$$DR_1(Y, D, X; p, m_1) = m_1(X) + \frac{D}{p(X)}(Y - m_1(X)) \quad (6)$$

$$DR_0(Y, D, X; p, m_0) = m_0(X) + \frac{1-D}{1-p(X)}(Y - m_0(X)) \quad (7)$$

可以证明,在假设1之下:

$$ATE = E(DR_1(Y, D, X; p, m_1) - DR_0(Y, D, X; p, m_0)) \quad (8)$$

简述式(6)、式(7)的直觉。以式(6)为例,考虑它关于 X 的条件期望。它的第一项 $m_1(X)$ 就是结果方程。如果 m_1 是正确的,而且第二项的条件期望为0,结果不会产生影响;但如果 m_1 是错误的,偏离了实际的条件期望,这一偏差还可以被第二项修正。第二项中 $Y - m_1(X)$ 就是实际值与(可能错误的)结果方程之差,然后用倾向值的倒数加权;分子上的 D 保证了该式只讨论处理组。

双重稳健估计量的优良性质具体体现为下面的定理。

定理1:令 $\hat{m}_1, \hat{m}_0, \hat{p}$ 为关于 X 的任意函数, $m_1(X) = E(Y|D=1, X)$, $m_0(X) = E(Y|D=0, X)$, $p(X) = P(D=1|X)$,皆为真实函数,在假设1下,双重稳健估计量具有如下性质:

$$\begin{aligned} E(Y(1)) &= E(DR_1(Y, D, X; p, m_1)) = E(DR_1(Y, D, X; \hat{p}, m_1)) = \\ &E(DR_1(Y, D, X; p, \hat{m}_1)) \\ E(Y(0)) &= E(DR_0(Y, D, X; p, m_0)) = E(DR_0(Y, D, X; \hat{p}, m_0)) = \\ &E(DR_0(Y, D, X; p, \hat{m}_0)) \end{aligned}$$

定理1的证明见附录一。定理1表明,倾向值和结果模型两者只要其中一个正确,最终的因果识别就是正确的,这也是“双重稳健”名称的由来。从数学形式上看,这是因为它的构造保证了任何冗余参数的扰动对估计量都不存在一阶影响。

然而,仅仅使用双重稳健估计量是不够的,还需要交叉拟合来进一步避免偏误。在真实数据中,双重稳健估计量的不确定性来自于两方面:一是在函数拟合过

^① 这种假设冗余参数已知的估计量也被称作“神谕估计量”(Oracle Estimator)。

程中产生的不确定性,即 $\hat{m}_1, \hat{m}_0, \hat{p}$ 估计的不确定性;二是给定真实的 m_1, m_2, p 之后,将数据点代入双重稳健估计量,因为数据 (Y, D, X) 随机抽样而带来的不确定性。对于机器学习算法而言, $\hat{m}_1, \hat{m}_0, \hat{p}$ 极有可能对数据产生过拟合,甚至因拟合噪音而无法收敛到真实函数。如果用全样本数据来估计函数,再将相同的数据代入估计得到的函数计算 \widehat{ATE} ,那么两部分不确定性将具有相同的来源并导致偏差。交叉拟合可以使两种不确定性独立开来,具体的做法如下:

算法 2-2-1(交互回归模型的双重机器学习)

1. 将数据划分为 K 个互不相交的子样本。对于第 k 个子样本,使用其余 $K - 1$ 个子样本:

第一,利用机器学习算法学习 $E(Y|X, D = 1)$,得到 $\hat{m}_{1[k]}$ 。

第二,利用机器学习算法学习 $E(Y|X, D = 0)$,得到 $\hat{m}_{0[k]}$ 。

第三,利用机器学习算法学习 $P(D = 1|X)$,得到 $\hat{p}_{[k]}$ 。

2. 对于每个样本 $i \in \{1, 2, \dots, n\}$,其所处的子样本为 $k(i)$,计算可得

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \left(DR_1(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{1[k(i)]}) - DR_0(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{0[k(i)]}) \right)$$

3. 计算标准误:

$$se(\widehat{ATE}) = \frac{1}{n} \sqrt{\sum_{i=1}^n \hat{\psi}_i^2}$$

其中, $\hat{\psi}_i = DR_1(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{1[k(i)]}) - DR_0(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{0[k(i)]}) - \widehat{ATE}$ 。而 ATE 估计量置信区间的构造方式与一般渐近正态统计量的构造方式相同。

在满足一定条件时,双重稳健估计量具有一致性和渐近正态性,并且估计量随着样本量 n 的增大以 $1/\sqrt{n}$ 的速率收敛到真实值(Chernozhukov 等, 2018),这与常见的 OLS、TWFE 等参数化估计量的收敛速度相同,快于核回归(Kernel Regression)、筛分法(Sieve Method)等常见非参数方法。这些条件包括:

假设 2(交互回归模型的估计假设)

第一, $\{Y_i, D_i, X_i\}_{i=1}^n$ 是独立同分布的样本;

第二,因变量不存在大的异常值,即 $Var(Y|X, D)$ 有界;

第三,共同支撑假设成立,倾向值函数严格处于 0 和 1 之间,即存在 $\epsilon > 0$,使得 $\epsilon \leq p(X) \leq 1 - \epsilon$ 。

第四,倾向值函数的估计量严格处于 0 和 1 之间,即存在 $\epsilon > 0$,使得 $\epsilon \leq \hat{p}(X) \leq 1 - \epsilon$ 。

第五, $\hat{m}_1, \hat{m}_0, \hat{p}$ 是对真实函数 m_1, m_2, p 足够好的近似,即对于 $(\hat{\eta}, \eta) \in \{(\hat{m}_1, m_1), (\hat{m}_0, m_0), (\hat{p}, p)\}$,随着样本量增大,有以下条件成立:

$$E\left[\int(p(x) - \hat{p}(x))^2 f_X(x) dx\right] E\left[\int(m_1(x) - (\hat{m}_1(x))^2 f_X(x) dx\right] = o\left(\frac{1}{n}\right)$$

$$E\left[\int(p(x) - \hat{p}(x))^2 f_X(x) dx\right] E\left[\int(m_0(x) - (\hat{m}_0(x))^2 f_X(x) dx\right] = o\left(\frac{1}{n}\right)$$

第一、第四、第五尤其需要研究者注意。首先,独立同分布是交叉拟合的前提。其次,为了保证倾向值函数的估计量严格处于0和1之间,往往需要在算法2-2-1第一步之后对数据进行截尾,舍弃掉 $\hat{p}_{[k(i)]}(X_i)$ 过大或过小的样本,或通过缩尾将其限制在一定范围内。最后,需要选取好的机器学习算法来保证近似的有效性。

在交互回归模型中,可以进一步考察处理效应的异质性。考虑协变量 X 的一个子集 X^s ,条件于该子集取值的平均处理效应可以识别为:

$$\begin{aligned} CATE(X^s) &= E(Y(1) - Y(0)|X^s) \\ &= E\left(DR_1(Y_i, D_i, X_i; p, m_1) - DR_0(Y_i, D_i, X_i; p, m_0)|X^s\right) \end{aligned} \quad (9)$$

即条件处理效应为括号内表达式的条件期望。由于这一条件期望难以准确逼近,在实际操作中,一般以最佳线性预测(Best Linear Prediction, BLP)作为退而求其次的近似。具体的做法如下:

算法2-2-2(交互回归模型的异质性因果效应估计)

第一,执行算法2-2-1。

第二,对于每个样本 $i \in \{1, 2, \dots, n\}$,其所处的子样本为 $k(i)$,计算得分(Score):

$$\hat{h}_i = DR_1\left(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{1[k(i)]}\right) - DR_0\left(Y_i, D_i, X_i; \hat{p}_{[k(i)]}, \hat{m}_{0[k(i)]}\right)$$

第三,用 \hat{h}_i 对异质性变量 X_i^s 做线性回归,得到处理效应基于 X^s 的最佳线性预测。也可以先使用样条函数等对 X_i^s 进行非线性变换,再将变换后的变量作为自变量放入线性回归,从而拟合得到关于 X_i^s 非线性的异质性效应。

对比部分线性模型和交互回归模型,部分线性模型对数据施加了更多的函数形式假设:式(1)的设定要求处理变量 D 和协变量 X 的效应是线性可加的,并且处理效应对所有个体都相同,而交互回归模型无此要求。因果效应存在异质性,如何解读部分线性模型(1)式得到的回归系数 β ? 此时,如果处理变量 D 为二元变量,并且算法2-1-1估计的 $l(x)$ 处于 $[0, 1]$ 区间内,则可以将其视作是对倾向值的估计,然后根据Frisch-Waugh定理,可以证明 β 是条件因果效应的凸组合:

$$\begin{aligned} \beta &= E[w(X)CATE(X)] \\ w(X) &= \frac{p(X)(1 - p(X))}{E[p(X)(1 - p(X))]} \geq 0 \end{aligned} \quad (10)$$

这一因果效应的凸组合又叫作交叠加权平均处理效应(Average Treatment Effect on the Overlap, ATO):它给处理组、对照组分布较为平衡的区域赋予较高权重,而给不

太平衡的区域赋予较低权重。虽然它不是平均因果效应,但可以看作一种“无伤大雅”(Mostly Harmless)的简化(Angrist和Pishke, 2009; Li等, 2018; 林梦芸等, 2025)。若希望通过部分线性模型进一步研究因果效应的异质性,则可以仿照线性回归,构造异质性维度与处理变量的交互项作为近似替代。

(三) 双重机器学习的一般形式

进一步在矩估计的框架下讨论双重机器学习的性质,从而使其应用于更多模型,特别是结构式估计当中。令 W 为数据, θ 为研究者关心的参数,真值为 θ^0 。 η 为可以使用机器学习估计的冗余参数,真值为 η^0 , 矩条件为 $M(\theta, \eta) = E(\psi(W; \theta, \eta))$, 在冗余参数为真值时满足对关心参数的可识别性条件,即 $M(\theta^0, \eta^0) = E(\psi(W; \theta^0, \eta^0)) = 0$ 。在双重机器学习中,首先会使用机器学习估计 η , 得到 $\hat{\eta}$ 。然后,代入矩条件,利用 $M(\theta, \hat{\eta}) = 0$ 解得目标参数 $\hat{\theta}$ 。

问题的关键在于,如何避免第一步中 $\hat{\eta}$ 的扰动导致第二步 $\hat{\theta}$ 的偏误。考虑对 η 的一个扰动方向 Δ , 扰动程度为 $t \in R$ 。如果对于任意的扰动方向 Δ , 都有^①:

$$\frac{\partial}{\partial t} M(\theta^0, \eta^0 + t\Delta) \Big|_{t=0} = 0 \tag{11}$$

那么冗余参数带来的一阶小扰动就不会对矩条件以及 θ 的估计产生影响。这一条件就是内曼正交性,而 $\psi(W; \theta, \eta)$ 称为内曼正交性分数(Neyman Orthogonal Score)。部分线性模型中, $\theta = \beta$, $\eta = (m, l)$, 内曼正交性分数为 $\psi(W; \theta, \eta) = (Y - m(X) - \theta(D - l(X)))(D - l(X))$ 。交互回归模型中, $\theta = ATE$, $\eta = (m_1, m_0, p)$, 内曼正交性分数为 $\psi(W; \theta, \eta) = DR_1(Y, D, X; p, m_1) - DR_0(Y, D, X; p, m_0) - \theta$ 。不难验证这两个矩条件都满足内曼正交性,具体证明过程见附录一。对于更一般的情形,内曼正交性分数需要研究者自己构造。

内曼正交性条件通过排除冗余参数一阶扰动对矩条件的影响,只留下二阶和更高阶的扰动,极大放宽了对于冗余参数估计量收敛速度的要求。一般来说,只需要满足:

$$\sqrt{E\left[\int \|\hat{\eta}(x) - \eta^0(x)\|^2 f_X(x) dx\right]} = o_p\left(n^{-\frac{1}{4}}\right) \tag{12}$$

双重机器学习估计量 $\hat{\theta}$ 就能以常规的 $1/\sqrt{n}$ 速率收敛至真实值 θ^0 。大部分机器学习算法在相当宽泛的条件下,都能够实现这一收敛速率。例如线性和参数稀疏性假设下的 LASSO 模型(Bickel 等, 2009; Buhlmann 和 van de Geer, 2011; Belloni 等, 2011、2012; Belloni 和 Chernozhukov, 2011、2013)、一大类决策树与随机森林(Wager 和 Walther, 2016), 以及一大类神经网络(Chen 和 White, 1999; Schmidt-Hieber, 2020)。

① 左侧的导数称作泛函 M 关于参数 η 沿 Δ 方向的 Gateaux 导数。

与前文所述的部分线性模型以及交互回归模型一样,一般情况下的双重机器学习也依赖于交叉拟合来避免过拟合问题,具体做法如下:

算法 2-3-1(一般形式的双重机器学习)

输入:独立同分布的数据 $\{W_i\}_{i=1}^n$,满足内曼正交性的分数 $\psi(W; \theta, \eta)$,其中 θ 为目标参数而 η 为冗余参数。

第一,将数据划分为 K 个互不相交的子样本。对于第 k 个子样本,使用其余 $K-1$ 个子样本利用机器学习得到 $\hat{\eta}_{[k]}$ 。

第二,利用交叉拟合计算样本矩:

$\hat{M}(\theta, \hat{\eta}) = \frac{1}{n} \sum_{i=1}^n \psi(W_i; \theta, \hat{\eta}_{[k(i)]})$, 并利用矩条件解出目标参数 $\hat{\theta}$ 为满足 $\hat{M}(\hat{\theta}, \hat{\eta}) = 0$ 的解。

第三,计算标准误:根据矩方法的标准误公式, $\hat{\theta}$ 的渐近方差估计量为:

$$\hat{V} = \hat{J}^{-1} \hat{\Omega} \hat{J}'$$

其中

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]}) \psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]})' - \frac{1}{n} \sum_{i=1}^n \psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]}) \frac{1}{n} \sum_{i=1}^n \psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]})'$$

$$\hat{J} = \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta} \psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]})$$

若 $\hat{\theta}$ 为标量,则对 $\frac{\hat{V}}{n}$ 开平方即可得到标准误。

概括下来,双重机器学习的三要素分别为:足以识别目标参数且满足内曼正交性的矩条件,交叉拟合,以及足够好的用于估计冗余参数的机器学习算法。这三者共同保证了双重机器学习估计量的一致性和渐近正态性。

三、双重机器学习的拓展

线性回归以外,双重机器学习还可以和研究者所熟知的其他方法相结合,本部分简要介绍双重机器学习如何与工具变量、双重差分、断点回归等经典实证研究工具相结合,并提供对应的算法。

(一)与工具变量结合

部分线性模型的工具变量。假设模型仍然与式(1)相同,但此时 D 和 u 条件于 X 可能相关。在使用工具变量部分线性模型时,工具变量 Z 应当满足:(1)条件工具相关性,即 $Cov(Z, D|X) \neq 0$; (2)条件工具外生性,即 $Cov(Z, u|X) = 0$ 。基于条件工具外生性这一矩条件,可以构造双重机器学习估计量。其基本思路是首先通过机器学习,剥离 Y, D, Z 当中可以被控制变量 X 解释的部分;其次再利用剩余的变动使用工具变量估计因果参数。具体算法见附录二算法 A2-1。

交互回归模型的工具变量。交互回归模型允许异质性处理效应的存在,识别了依从者的(条件)局部平均处理效应,识别假定包括条件排他性、条件外生性、条件单调性、条件相关性,以及共同支撑假设。当工具变量 Z 和处理变量 D 都为取值为0、1的虚拟变量时,根据个体对工具变量的反应不同,可以把个体分为四类:“永远接受者”(Always Taker)、“永远拒绝者”(Never Taker)、“依从者”(Complier)和“叛逆者”(Defier)。通过Wald估计量 $LATE(x) = \frac{E(Y|Z=1, X=x) - E(Y|Z=0, X=x)}{E(D|Z=1, X=x) - E(D|Z=0, X=x)}$ 可以识别 $X=x$ 处依从者的平均处理效应,对 X 积分后可以得到局部平均处理效应(Local Average Treatment Effect, LATE)(胡诗云等,2025)。然而,在关于 X 的条件期望存在明显非线性时,直接将条件期望估计值代入Wald估计量可能造成严重偏误。

使用双重机器学习可以应对条件期望函数的非线性问题。定义 $m(Z, X) = E(Y|Z, X)$, $p(Z, X) = E(D|Z, X)$, 以及 $r(X) = E(Z|X)$, 内曼正交性分数可以写为:

$$\psi(Y, D, Z, X; \theta, m, p, r) = m(1, X) - m(0, X) + \frac{Z - r(X)}{r(X)(1 - r(X))} (Y - m(Z, X)) - \left(p(1, X) - p(0, X) + \frac{Z - r(X)}{r(X)(1 - r(X))} (D - p(Z, X)) \right) \theta \quad (13)$$

其中, θ 为待识别的因果参数, m, p, r 三个条件期望函数都可以通过机器学习得到, 然后就可以代入双重机器学习的一般框架。具体算法见附录二算法 A2-2。

当工具变量 Z 和处理变量 D 都为多值变量时, Xie(2024)将上述交互回归模型的识别拓展到广义局部平均处理效应(Generalized Local Average Treatment Effect, GLATE), 并讨论了如何在弱工具变量情形下修正双重机器学习算法。

(二)与双重差分结合

双重差分和事件研究法(Event Study)已经成为当今经济学、管理学最流行的研究方法之一(张子尧和黄炜,2023), 而双重机器学习可用于在双重差分研究中灵活地控制变量, 基于条件平行趋势假设识别因果效应, 增强研究的可信度。

在 2×2 的双重差分设定下, 时期 $t = 0, 1, \{Y_i(d)\}_{d=0,1}$ 为 t 期的潜在结果, D 为标识处理组的虚拟变量, X 为控制变量, 核心识别假设为条件平行趋势(Conditional Parallel Trend):

$$E(Y_1(0) - Y_0(0)|D = 1, X) = E(Y_1(0) - Y_0(0)|D = 0, X) \quad (14)$$

即给定控制变量不变, 处理组和对照组未受处理的潜在结果具有平行趋势; 换句话说, 两组之间未受处理的潜在结果如果存在趋势差异, 完全是由于 X 的差异导致的。在条件平行趋势假设以及无预期效应假设之下, 可以在共同支撑的范围内识别处理组的平均处理效应: $ATT(X) = E(Y_1(1) - Y_1(0)|D = 1, X)$; 进一步施加共同支

撑假设: $\exists \varepsilon > 0, 0 \leq P(D = 1|X) \leq 1 - \varepsilon$, 保证每一个处理组样本都能找到倾向值相似的对照组样本, 则可以进一步识别 $ATT = E(Y_1(1) - Y_1(0)|D = 1)$ 。

在传统的线性回归框架中, 往往通过加入 X 与时间虚拟变量的交互项, 来控制处理组和对照组因为 X 而导致的潜在趋势差异。这种做法假定了 X 对于潜在结果的影响是线性的。一些研究会直接加入 X 的高阶项与时间虚拟变量的交互项, 来控制非线性影响, 但这一部分的估计误差会传导至因果效应估计, 导致更大的偏误。使用双重机器学习可以较好地处理 X 对潜在结果的非线性影响。理论表明, 在面板数据双重差分中, 内曼正交性分数为 (Sant' Anna 和 Zhao, 2020; Chang, 2020):

$$\psi(Y_1, Y_0, D, X; \alpha, \pi, p, g) = \frac{D - p(X)}{\pi(1 - p(X))} (\Delta Y - g(0, X)) - \frac{D}{\pi} \alpha \quad (15)$$

其中, α 为 ATT, 是需要识别的因果参数, $\Delta Y = Y_1 - Y_0$ 为个体结果的一阶差分, $\pi = P(D = 1)$ 为接受处理的无条件概率, $p(X) = P(D = 1|X)$ 为给定控制变量后个体接受处理的概率(倾向得分), $g(0, X) = E(\Delta Y|D = 0, X)$ 为给定控制变量后对照组的前后之差, 反映了由 X 带来的未受处理的潜在结果趋势的差异。

实际上, 和横截面数据中用于识别 ATT 的内曼正交性分数相比, 式(15)就是将 Y 替换为了 ΔY 。因此, 只需要首先将因变量进行事后与事前差分, 再运用横截面数据估计 ATT 的双重机器学习方法。具体算法见附录二算法 A2-3。

需要注意的是, 双重差分不是双向固定效应回归。将时间和个体虚拟变量作为高维协变量直接放入部分线性模型, 并将处理状态虚拟变量前的系数解释为双重差分估计量, 是不规范的。若研究者希望在 DML 中估计带有固定效应的静态面板模型, 应参照 Clarke 和 Polselli (2025) 提出的方法。

近年来, 有关渐进双重差分 (Staggered DID, 又称多时点双重差分) 与事件研究的讨论方兴未艾, 如何将机器学习算法应用于渐进处理的情形也成为了计量经济学的前沿问题。Callaway 和 Sant' Anna (2021) 提出, 渐进双重差分可以划归为多个 2×2 双重差分问题。具体来说, Callaway 和 Sant' Anna (2021) 定义了从时刻 g 开始接受处理的个体在时刻 $t (t \geq g)$ 的处理效应, 记作 $ATT(g, t)$ 。他们的研究指出, 不同的 $ATT(g, t)$ 应当分别估计, 再根据事件后的相对期数进行加总。

基于类似的框架, Hatamyar 等 (2023) 在 Lu 等 (2019) 两期半参数 DID 算法的基础上, 开发了针对渐进双重差分的双重机器学习算法。目前, 基于算法 3-2-2 的双重机器学习渐进 DID 可以通过 DoubleML 软件包实现一站式估计 (Bach 等, 2022)。具体算法见附录二算法 A2-4。

(三) 与断点回归结合

断点回归因为其接近自然实验的特点, 在越来越多的实证研究中得到应用

(黄炜等, 2025)。断点回归的基本识别假设是潜在结果函数在断点处的连续性, 从而可观测结果在断点附近的任何不连续性都可以归结为断点两侧政策的差异。因此, 从识别的角度看, 断点回归不需要控制变量。但是, 从提升估计精度的角度考虑, 加入协变量有助于纳入更多先验信息并减小因变量中的误差项 (Frölich 和 Huber, 2019; Calonico 等, 2019)。此外, 协变量也可用于研究不同子群体处理效应的异质性 (Cattaneo 等, 2023)。利用双重机器学习可以去除结果变量 Y 当中能被已知协变量 X 解释的部分。其做法可以简述为使用机器学习算法利用 X 拟合 Y , 再将预测残差作为新的因变量进行断点回归。具体算法见附录二算法 A2-5。

值得注意的是, 在断点回归的理论框架中, 处理变量 D 完全由驱动变量决定, 不存在通常意义上的“内生性”, 因此也不必剔除协变量 X 对处理变量的影响。Noack 等 (2021) 证明了只要 X 能帮助解释结果变量, 上述做法几乎总是能够减小断点估计量的方差。

(四)小结: 旧的识别假设, 新的估计方法

通过讨论如何将 DML 运用于工具变量、双重差分以及断点回归等研究者常用的实证策略, 发现 DML 并没有改变原有的研究设计、识别假设与目标参数, 而是为估计提供了新的方法。以控制变量为例, 其研究设计 (Research Design) 的核心假设是“给定可观测协变量相同, 处理的分配应该同潜在结果是独立的”。无论是线性回归、倾向值匹配还是双重机器学习, 都是在实现与一个控制变量的研究设计。不能因为使用了双重机器学习, 而忽视了对条件独立性假设的检验。一组未能充分捕捉混淆因素的控制变量, 不会因为使用了双重机器学习而自动变成充分的控制变量。

表 1 总结了不同的研究设计中, DML 与传统估计方法的对比。总之, 将 DML 视作一种新型识别策略是不妥当的。正如 Angrist 和 Pischke (2009) 在《基本无害的计量经济学》第三章的标题所述, “让回归变得有意义”, 只有明确了研究设计与识别假设, DML 的估计结果才会变得有意义。

表 1 不同研究设计中 DML 与传统估计方法的对比

研究设计	共同的核心识别假设	共同的识别目标	不同的估计方法
控制变量	条件可忽略性假设	平均处理效应 (ATE)、处理组处理效应 (ATT)	传统方法: 线性回归、倾向值匹配等 DML: 算法 2-2-1
工具变量	条件于协变量的: 排他性、外生性、相关性、单调性假设	依从者局部平均处理效应 (LATE)	传统方法: 在两阶段最小二乘中, 将控制变量作为外生变量纳入方程 DML: 算法 A2-2

(续)

研究设计	共同的核心识别假设	共同的识别目标	不同的估计方法
双重差分	条件平行趋势假设	处理组处理效应(ATT)	传统方法:包含控制变量的双向固定效应、事件研究回归 DML:算法 A2-3、A2-4
断点回归	潜在结果连续性假设	断点处局部平均处理效应(LATE)	传统方法:在参数化断点回归中将协变量纳入方程 DML:算法 A2-5

注:本表对比了DML与传统方法在识别假设和估计上的关系。为了突出因果识别本身与函数形式的无关性,叙述简洁起见,本表第2、3列的假设与因果识别均在非参数框架下进行,最后一列仅包含了无任何函数形式假定的DML算法。

四、双重机器学习的应用技术细节

在已有利用双重机器学习的实证文献中,还有三个常见的实践细节值得注意。第一,使用部分线性模型时,估计 $E(Y|X)$ 和 $E(D|X)$ 两个条件期望函数的机器学习算法可以不同。研究者可以针对 Y 和 X 分别尝试LASSO、随机森林、神经网络等逼近方法,选取各自交叉验证(Cross-validation)^①表现最好的机器学习模型。例如孙传旺等(2024)就尝试了随机森林和LASSO两种方法。而在使用交互回归模型时,估计倾向得分函数 p 和结果方程 m 的机器学习方法也可以不同。实践中可能存在多种满足估计要求的机器学习算法,可以通过最小化交叉拟合误差来选择合适的算法,具体算法见附录二的算法A2-6。

第二,在交叉拟合的过程中,已有模拟实验表明选择折数 $K = 4$ 或 5 通常能取得较好效果(Chernozhukov等,2024)。对于较小规模的数据集,增加折数能增大训练集规模,可采用 $K \geq 10$ 的交叉验证。为了避免特定的交叉拟合方式对结果产生影响,研究者可以采用多种样本划分方式进行 K 折交叉拟合,并取多次估计的中位数。根据Chernozhukov等(2018)的推荐,其一般做法是:采用 S 种不同的方式划分样本,记每次得到的估计量为 $\hat{\theta}_s$ 。最终汇报的估计量为:

$$\hat{\theta}^{median} = \text{median} \{ \hat{\theta}_s \}_{s=1}^S \quad (16)$$

标准误也需要针对样本划分方式的不确定性进行调整,具体为:

$$\hat{\sigma}^{2, median} = \text{median} \{ \hat{\sigma}_2^2 + \left(\hat{\theta}_s - \hat{\theta}^{median} \right) \left(\hat{\theta}_s - \hat{\theta}^{median} \right)' \} \quad (17)$$

在面板数据应用中,若使用聚类稳健推断,则在交叉拟合时应保证同一聚类永

^① 交叉验证的目的是评估机器学习模型的样本外预测能力。其第一步与交叉拟合相同;第二步是计算交叉拟合值与真实数据的差异,并计算整个样本上的平均损失函数。好的机器学习模型应当有较小的交叉验证损失。

远处于同一子样本当中(Chernozhukov等,2024)。这是因为聚类推断允许同一聚类内存在任意形式的自相关,聚类之间相互独立。

第三,当面临多维复杂参数推断时,乘数自助法(Multiplier Bootstrap)有独特的优势。它直接利用算法2-3-1中已经算得的影响函数(Influence Function) $\psi(W_i; \hat{\theta}, \hat{\eta}_{[k(i)]})$,将其乘以方差为1的随机变量,模拟抽样随机性的影响。通过生成多组随机变量,可以得到多组扰动后的影响函数,并模拟得到估计量 $\hat{\theta}$ 的抽样分布。与传统自助法相比,乘数自助法避免了重复训练机器学习模型 $\hat{\eta}_{[k(i)]}$ 的巨大算力消耗,又巧妙地解决了复杂相关结构下的同时推断难题。乘数自助法已经可以通过软件包实现,具体公式参见Belloni等(2018)。

五、数值模拟与实证案例

(一)何时选择双重机器学习?基于数值模拟的讨论

双重机器学习虽然能极大地放松对于线性函数形式的严格假设,但它也不是解决一切计量问题的“灵丹妙药”。作为一种非参数方法,它在避免函数形式误设的同时,也对数据量提出了更高的要求。接下来以当前中国经济学研究最常用的实证策略之一——双重差分为背景,利用数值模拟,从三个角度考察双重机器学习方法在不同场景下的适用性:一是协变量系数是否随时间变化;二是处理效应是否有异质性;三是满足共同支撑条件的样本量。张征宇等(2025)指出,当存在连续型控制变量时,即便是单时点双重差分,使用双向固定效应模型依然有可能造成“负权重”问题,其根源就在于错误地假定了控制变量线性的函数形式。这正是双重机器学习有可能发挥作用的地方。

参照张征宇等(2025)的研究,首先,按照如下规律生成潜在结果:

$$Y_{it}(d) = \alpha_i + t + dTE(X_{it}) + \beta_i X_{it} + u_{it} \quad (18)$$

$$TE(X_{it}) = 0.7 + cX_{it}^2 \quad (19)$$

$$X_{i0} \sim N(1, 1), X_{i1} \sim -2 + 3X_{i0} + v_{i1}, v_{i1} \sim N(0, 1) \quad (20)$$

其中, $i \in \{1, 2, \dots, N\}$, $t \in \{0, 1\}$, $d \in \{0, 1\}$, $TE(X_{it})$ 代表了处理效应,参数 c 为代表处理效应异质性程度的参数,若 $c \neq 0$,则代表存在异质性处理效应。参数 β_i 为协变量对于潜在结果的影响,若 $\beta_0 \neq \beta_1$,则协变量对于潜在结果的影响是随时间变化的。处理的分配以及从潜在结果到可观测结果的过程如下:

$$D_{i0} = 0, D_{i1} = 1 \left(X_{i1}^{\frac{1}{3}} \geq (u - l)U + l \right), U \sim \text{Uniform}(0, 1) \quad (21)$$

$$Y_{it} = (1 - D_{it})Y_{it}(0) + D_{it}Y_{it}(1) \quad (22)$$

其中,参数 l, u 为调整共同支撑样本占比的参数。表2展示了1000次数值模拟、每

次模拟 1000 位个体的结果,以估计 ATT 为目标。

当数据生成过程与传统模型的假设完全吻合时,即不存在协变量时变效应和异质性处理效应(过程 1),双向固定效应模型表现最佳。然而,一旦双向固定效应模型是错误设定,例如存在协变量时变效应(过程 2)或异质性处理效应(过程 3、过程 4)时,就会出现严重的偏差。部分线性模型表现优于双向固定效应模型,具有较小的偏差和 MSE。尽管如此,部分线性模型假设了处理效应与协变量的可分离性,仍无法完全捕捉异质性处理效应带来的复杂性。在过程 5、过程 6 当中,部分线性模型的表现弱于交互回归模型。此外,模拟结果也表明了双重机器学习对数据质量的依赖。比较共同支撑样本占比低(过程 4)和占比高(过程 6)的两种情景可以看出,当共同支撑范围更广时,交互回归模型的表现明显提升。

表 2 数值模拟结果——以 ATT 为估计目标

数据生成过程	协变量时变系数	异质性处理效应	共同支撑样本占比	真实 ATT	估计量	偏差	MSE
1	否	否	低	0.700	双向固定效应	-0.002	0.020
					部分线性模型	-0.070	0.110
					交互回归模型	-1.725	3.259
2	是	否	低	0.700	双向固定效应	-0.515	0.293
					部分线性模型	-0.063	0.115
					交互回归模型	-6.664	44.795
3	否	是	低	3.421	双向固定效应	-5.010	25.273
					部分线性模型	-2.779	7.840
					交互回归模型	-1.725	3.259
4	是	是	低	3.421	双向固定效应	-5.524	30.677
					部分线性模型	-2.741	7.650
					交互回归模型	-6.664	44.795
5	否	是	高	4.521	双向固定效应	-1.102	1.263
					部分线性模型	-1.091	1.283
					交互回归模型	0.063	0.112
6	是	是	高	4.521	双向固定效应	-1.316	1.786
					部分线性模型	-1.075	1.251
					交互回归模型	0.286	0.710

注:本表展示了 1000 次数值模拟结果,每次模拟 1000 位独立同分布个体,各参数的具体数值见附录三。“双向固定效应”为因变量对处理状态、协变量以及时间和个体固定效应的线性回归。“部分线性模型”是指对因变量做一阶差分后使用算法 2-3-1。“交互回归模型”使用算法 A2-3。所有机器学习模型均为随机森林。“共同支撑样本占比”是指数据生成过程决定的总体中,重叠处理组样本占所有处理组样本的比重。“共同支撑样本占比低”的模拟中,该比例为 16.7%;“共同支撑样本占比高”的模拟中,该比例为 100%。

从式(4)、式(5)可以看出,交互回归模型对于接近于0或1的倾向值极其敏感;在表2中,交互回归模型在共同支撑样本占比低时(过程2、过程4)也会出现大方差。在实际研究当中,若估计的倾向值大量接近于0和1,则是出现数据弱重叠的信号。这一现象可能是由两种原因造成的:一是机器学习在估计中的随机性,导致出现了极端的拟合值;二是总体数据本身就不满足重叠性假设。

对于前一种情况,实证研究中最常用的实践是通过将极端的倾向值进行缩尾来缓解,例如将所有小于0.01的拟合值归并至0.01,所有大于0.99的拟合值归并至0.99^①。除此之外,倾向值截断也是一种可能的方法(Crump等,2009),但Ma等(2023)指出这一截断可能导致双重稳健估计量失去双重稳健的特性,并提出了一种截断误差的纠偏方式。Heiler和Kazak(2021)提出了一种在不进行倾向值截断的情况下改进的Bootstrap推断方法。对于后一种情况,由于找不到相似的对照组,无法识别平均处理效应。此时可退而求其次,识别式(10)定义的ATO。ATO不仅可以解读为“有足够信息进行对比的那部分个体的处理效应”,还可以解读为“在是否接受处理之间不太确定、位于边际上的个体的效应”,有明确的政策含义。附录三展示了以ATO为识别目标的模拟结果。

数值模拟结果对实证应用者的启发如下:第一,在模型选择上,若有充分的理论或先验知识支撑线性与同质性假设,传统的精简模型依然是高效的选择。第二,当研究者怀疑政策效果可能因个体特征而异,或控制变量的影响随时间变化时,双重机器学习是获得更可信、更稳健因果估计的强大工具。第三,在应用双重机器学习时要关注共同支撑条件,在共同支撑条件较弱时,PLM比IRM更有优势。

(二)实证案例:精准扶贫的增收效应再评估

为具体展现双重机器学习在实证研究中的应用价值,本部分借鉴Li等(2025),评估精准扶贫这一大规模、多维度扶贫政策对农村人均纯收入的影响。使用与Li等(2025)相同的数据和识别策略^②,首先利用经典的双向固定效应模型复现其核心结果,然后使用双重机器学习重新分析,从而展示其在实证研究中的应用流程,并突出其相对于传统方法能对数据特征、实证设计提供的新洞见。

精准扶贫开始于2014年,而2012年调整确定的“国家扶贫开发工作重点县”则是受到精准扶贫政策影响较大的地区。一个县是否选入扶贫开发工作重点县,主要取决于以下7项指标:人均收入是否小于2300元、是否已入选“八七扶贫攻坚计划”、是否属于少数民族自治县、是否为革命老区县、是否为山区县、农村人口占比以及人均财政收入。由于政策的分配主要基于这些可观测的期前特征,只要能够充分控制这些期前特征,就有理由相信,在政策未实施的情况下,作为处理组的重

① 本部分的数值模拟结果均采用了倾向值缩尾方法。

② 通过<https://www.openicpsr.org/openicpsr/project/201661/version/V1/view>下载该数据。

点县与作为对照组的非重点县的农村收入会遵循平行的变化趋势(Li等,2025)。

本文基于此识别策略,比较了五种不同模型的估计结果,以考察模型设定的灵活性对估计结果以及事前平行趋势检验的影响。第一种是未加入任何控制变量的简单双重差分模型。第二种是遵循原文设定的双向固定效应模型,不仅控制了县级个体固定效应和省份与年份的交互固定效应,还控制了7项入选标准(期前特征)与年份虚拟变量的交互项,以尽可能满足条件平行趋势假设。其余三种机器学习模型分别是广义线性模型、神经网络和随机森林,高维协变量包括省虚拟变量和7项入选标准。模型具体设定详见附录四。基准估计均使用交互回归模型,部分线性模型作为稳健性检验。

表3和图2分别展示了静态效应和动态效应的估计结果。首先,表3中五种模型均一致地表明,自2014年政策实施之后,国家扶贫开发工作重点县的农村居民收入相对于对照组出现了显著的增长,说明政策取得了积极成效。表3同时报告了各机器学习算法对于结果方程和倾向值方程的样本外拟合效果。可以看出,在三种机器学习模型当中,随机森林实现的样本外拟合效果总体上是最好的。它对于倾向值方程和处理组结果方程的样本外拟合误差显著更低,对于对照组结果方程的样本外预测效果与神经网络接近并优于广义线性模型。因此,随机森林的估计结果(0.076)更可能接近于真实效应,这略低于双向固定效应的估计值。

表3 精准扶贫对农村增收的静态效应估计

模型	系数估计	标准误	RMSE ΔY (对照组)	RMSE ΔY (处理组)	RMSE D
简单DID	0.137	0.008	0.100	0.140	0.492
双向固定效应	0.109	0.005	-	-	-
DML:广义线性模型	0.121	0.012	0.100	0.138	0.328
DML:神经网络	0.099	0.020	0.085	0.114	0.296
DML:随机森林	0.076	0.011	0.087	0.106	0.269

注:本表展示了不同模型下精准扶贫政策对农村人均收入影响的静态效应估计。因变量为农村人均纯收入的对数。双向固定效应的估计方式与Li等(2025)的主回归设定相同。参照Bertrand等(2004)对多期面板数据DID的处理,表中简单DID与三种DML的估计方法为:首先将因变量去除横截面均值,然后计算每个个体政策前和政策后的因变量均值,从而形成两组一两时点(2×2)DID,随后利用2×2DID算法进行估计。标准误均在县级层面聚类。所有DML均为交互回归模型。对于DML方法,表格汇报了99次不同交叉拟合所得系数的中位数,并依据式(17)对标准误进行了调整。RMSE ΔY (对照组)、RMSE ΔY (处理组)以及RMSE D 分别为机器学习方法对对照组结果方程、实验组结果方程以及倾向值方程的交叉拟合误差。双向固定效应因未使用交叉拟合,故略去相关数值。

在图2的动态效应中,各模型在细节上有一定差异。在政策实施前的时期,无控制变量的双重差分模型事前系数显著异于零。在控制变量后,事前平行趋势得到了更好的满足。但是,相比起双向固定效应,双重机器学习发现即便在控制入选标准后,处理组样本仍有一定程度的“阿森费尔特沉降”(Ashenfelter's Dip)(Ashenfelter, 1978; 张子尧和黄炜, 2023),且在2010年最为显著。这说明在控制Li等(2025)所列举的入选标准后,那些在政策前几年遇到短期经济困难的县,更有可能进入政策处理组。与此同时,在2017年之后,使用双重机器学习得到的事后处理效应估计量略低于双向固定效应的结果,其对于2020年农村人均纯收入对数的政策效应估计值从双向固定效应的0.15下降到随机森林的0.11。以上对比表明,机器学习方法可以灵活和非参数化地控制事前特征,最大限度地减少线性外推带来的偏误,同时更好地侦测事前可能存在的非平行趋势,从而提示事后可能存在的潜在趋势差异。从以上结果来看,精准扶贫政策是卓有成效的,但政策的长期效果可能略小于已有文献中的双向固定效应估计值。

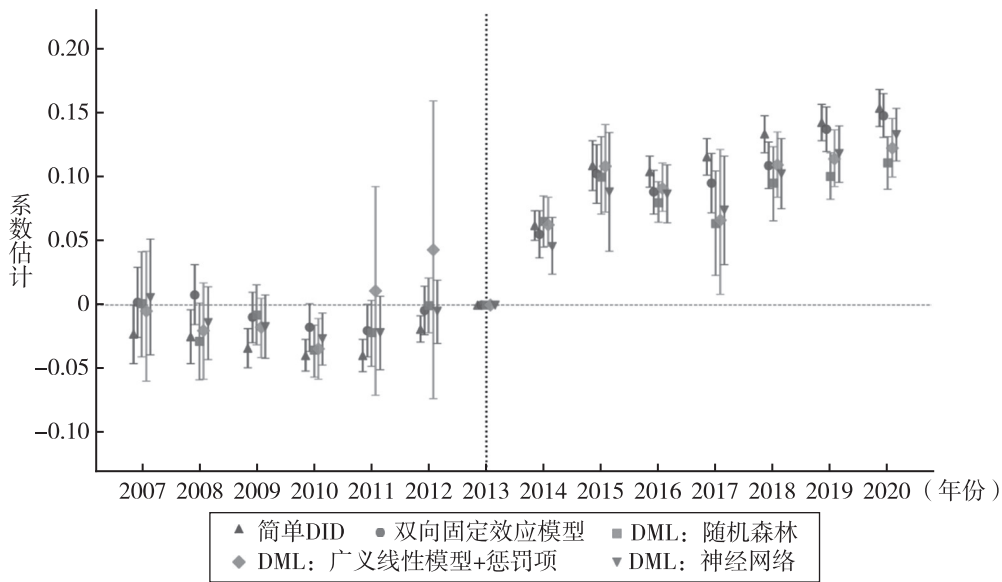


图2 精准扶贫对农村增收的动态效应估计

注:本图展示了不同模型下精准扶贫政策对农村人均收入影响的动态效应估计。因变量为农村人均纯收入的对数。图中报告了五种估计方法(无控制变量的简单DID、双向固定效应模型、广义线性模型加惩罚项、随机森林、神经网络)下,处理组(国家扶贫开发工作重点县)相对于对照组在各年份的系数估计值。所有系数均以政策实施前一年(2013年)为基准。对于DML方法,表格汇报了10次不同交叉拟合所得系数的中位数,并依据式(17)对标准误进行了调整。误差棒代表95%置信区间,标准误均在县级层面聚类。

在基准结果之外,还可以进行异质性分析,并进行以下稳健性和诊断性检验:一是通过绘制处理组和对照组的倾向值分布,观察样本交叠情况;同时绘制不同倾向值下处理组和对照组结果方程的拟合值,排除潜在异常值的影响。二是检验因果系数估计结果对于机器学习超参数的敏感性。三是使用事前变量进行平衡性检验,观察在使用双重机器学习控制协变量后,其他事前特征是否存在明显差异。异质性分析和稳健性、诊断性检验的具体过程、结果和解读见附录四。

(三)实践中的双重机器学习:一个经验总结

随着双重机器学习理论框架的成熟与应用的拓展,目前已经出现了比较成熟的软件包,使用双重机器学习进行因果推断在编程技术上已经不是难事。表4列举了Python、Stata、R语言中的双重机器学习命令。

表4 当前主流编程语言对双重机器的支持情况

语言	软件包	横截面数据 控制变量	IV	单时点 DID	多时点 DID	RDD
Python	软件包 DoubleML, 机器学习需与 Scikit-learn 结合	支持	支持	支持 2×2 设计, 动态效应需自 行编程	支持,基于 CSDID 框架(Callaway 和 Sant'Anna, 2021)	支持
Stata	命令 ddml	支持	支持	不支持	不支持	不支持
R 语言	软件包 DoubleML, 机器学习需与 mlr3 结合	支持	支持	不支持	不支持	不支持

然而,应用计量经济学并不只是掌握几个编程命令那么简单。只有“知其然又知其所以然”,才能正确地应用双重机器学习。结合理论结果、数值模拟与实证案例,双重机器学习在实践中的经验可以总结为以下六点:第一,应遵循“先识别、后估计”原则,将双重机器学习与现有识别策略结合而非视其为独立研究设计;第二,需根据样本量和重叠性决定是否使用双重机器学习以及选择交互回归或部分线性模型;第三,基于样本量和特征维度选择合适的机器学习模型,如带惩罚项的线性模型、随机森林或神经网络;第四,观察预测结果如倾向值分布和边界拟合值以确定超参数;第五,进行诊断性检验并对比不同估计量以评估稳健性;第六,采用适应双重机器学习的结果汇报范式,重点关注模型选择和冗余参数拟合效果,而不必汇报模型整体 R^2 。

附录五对于以上每一点在具体操作中的考虑进行了详细分析,附录六进一步总结了近年来使用双重机器学习的一系列实证文献,供进一步参考。

六、总结性评论

随着相关理论的成熟与开源软件包的普及,双重机器学习正从前沿的计量理论研究,逐渐“飞入寻常百姓家”。本文旨在系统地梳理双重机器学习的理论基础、主要拓展与应用场景,并为其恰当、有效的使用提供一份清晰的指引。

本文面向实证研究者,将双重机器学习这一看似复杂的“黑箱”拆解为一个逻辑清晰的“工具箱”。首先阐明了“内曼正交性”与“交叉拟合”这两大理论支柱如何协同作用,既利用了机器学习的灵活性,降低了冗余参数估计时的模型误设风险,又保证了关键因果系数估计的稳健性。在此基础上,展示了双重机器学习如何与工具变量、双重差分、断点回归等经典识别策略对接,在不改变核心识别假设的前提下,显著提升估计的稳健性。

这一系列讨论最终导向一个根本原则:应用双重机器学习时,必须遵循“先识别,后估计”的规范,其价值在于优化估计过程,而非替代研究设计本身。双重机器学习的真正力量,在于将研究者的精力从“如何设定函数形式”的困境中解放出来,从而能够更专注于识别策略本身是否可靠。当模型简单、线性关系明确时,传统方法依然高效;而当面对非线性、高维、异质性等复杂现实时,双重机器学习则能更稳健地逼近真相。

展望未来,双重机器学习的应用边界仍在不断拓宽。一个特别值得关注的新兴方向,是将其与大语言模型(Large Language Models, LLMs)相结合。在许多场景中,LLM生成的内容(如建议、信息、摘要等)以及文本嵌入本身可以被视为一种高维、复杂的变量。然而,其生成结果可能受到预测误差、特定提示词或训练数据偏见的影响,双重机器学习则为LLM输出结果用于下游因果效应估计提供了理论框架(Egami等,2024),拓宽了识别策略的边界。辩证地看,尽管DML本身不改变识别策略,但确实使得一些原本无法进入估计层面的识别策略变得可行。例如,在估计商品的需求曲线时,理想的识别策略应当是控制住商品的所有特征(如文本形式的商品描述),比较不同价格商品的销量。线性回归等传统的估计方法难以控制这些高维特征,而DML使得原本在实际中不可行的识别策略变得可行了。在这个意义上说,DML有帮助改善识别的作用。此外,在网络数据分析中,面对个体间相互影响带来的挑战,双重机器学习方法也可以用于估计网络环境下的平均处理效应,有效控制了高维的节点和连接特征(Chen等,2024)。

双重机器学习赋予了灵活性来处理复杂的数据结构,但并未降低对严谨研究设计的要求。希望本文能够为双重机器学习祛魅,破除盲目崇拜的迷思。掌握其原理、明晰其边界、审慎其应用,将是未来实证研究者从海量数据中挖掘可靠因果关系的关键一步。

参考文献

- [1]陈茁,陈云松.双重机器学习在社会科学因果推断中的应用[J].浙江社会科学,2025,(6):72~85.
- [2]胡诗云,江弘毅,孙振庭.异质性因果效应模型中的工具变量有效性统计检验[J].数量经济技术经济研究,2025,(3):155~176.
- [3]黄炜,向科谚,袁洛琪.断点回归设计的实证指南:操作规范、应用误区与实践拓展[J].数量经济技术经济研究,2025,(6):111~131.
- [4]林梦芸,徐阳,郭汝飞,等.在模型误设的统一框架下理解双重差分方法的最新发展[J].管理世界,2025,(6):227~264.
- [5]张征宇,吴路遥.双重差分法下固定效应估计量的负权重问题——新的机制与解决方案[J].数量经济技术经济研究,2025,(4):197~218.
- [6]张子尧,黄炜.事件研究法的实现、问题和拓展[J].数量经济技术经济研究,2023,(9):71~92.
- [7] Angrist J. D., Pischke J.S., 2009, *Mostly Harmless Econometrics: An Empiricist's Companion* [M]. Princeton: Princeton University Press.
- [8] Ashenfelter O., 1978, *Estimating the Effect of Training Programs on Earnings* [J], *The Review of Economics and Statistics*, 60(1), 47~57.
- [9] Bach P., Chernozhukov V., Kurz M.S., Spindler M., 2022, *DoubleML—an Object-oriented Implementation of Double Machine Learning in Python* [J], *Journal of Machine Learning Research*, 23(53), 1~6.
- [10] Belloni A., Chernozhukov V., 2011, *L1-penalized Quantile Regression for High Dimensional Sparse Models* [J]. *The Annals of Statistics*, 39(1), 82~130.
- [11] Belloni A., Chernozhukov V., 2013, *Least Squares After Model Selection in High-dimensional Sparse Models* [J]. *Bernoulli*, 19(2), 521~547.
- [12] Belloni A., Chen D., Chernozhukov V., Hansen C., 2012, *Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain* [J]. *Econometrica*, 80(6), 2369~2429.
- [13] Belloni A., Chernozhukov V., Wang L., 2011, *Square-root-lasso: Pivotal Recovery of Sparse Signals Via Conic Programming* [J]. *Biometrika*, 98(4), 791~806.
- [14] Belloni A., Chernozhukov V., Chetverikov D., Wei Y., 2018, *Uniformly Valid Post-regularization Confidence Regions for Many Functional Parameters In Z-estimation Framework* [J]. *The Annals of statistics*, 46(6B), 3643~3675.
- [15] Bertrand M., Duflo E., Mullainathan S., 2004, *How Much Should We Trust Differences-in-differences Estimates?* [J]. *The Quarterly Journal of Economics*, 119(1), 249~275.
- [16] Bickel P. J., Ritov Y., Tsybakov A. B., 2009, *Simultaneous Analysis of Lasso and Dantzig Selector* [J]. *The Annals of Statistics*, 37(4), 1705~1732.
- [17] Buhlmann P., van de Geer S., 2011, *Statistics for High-dimensional Data* [M], Berlin:

Springer Verlag.

[18] Callaway B., Sant'Anna P.H.C., 2021, *Difference-in-differences with Multiple Time Periods* [J], *Journal of Econometrics*, 225(2), 200~230.

[19] Calonico S., Cattaneo M.D., Farrell M.H., Titiunik R., 2019, *Regression Discontinuity Designs Using Covariates* [J], *The Review of Economics and Statistics*, 101(3), 442~451.

[20] Cattaneo M. D., Keele L., Titiunik R., 2023, *Covariate Adjustment in Regression Discontinuity Designs* [C], *Handbook of Matching and Weighting Adjustments for Causal Inference*, 153~168.

[21] Chang N.C., 2020, *Double/Debiased Machine Learning for Difference-in-differences Models* [J], *The Econometrics Journal*, 23(2), 177~191.

[22] Chen W., Cai R., Yang Z., et al., 2024, *Doubly Robust Causal Effect Estimation under Networked Interference via Targeted Learning* [R], arXiv preprint, No. 2405.03342.

[23] Chen X., White H., 1999, *Improved rates and asymptotic normality for nonparametric neural network estimators* [J], *IEEE Transactions on Information Theory* 45(2), 682~691.

[24] Chernozhukov V., Chetverikov D., Demirer M., et al., 2018, *Double/Debiased Machine Learning for Treatment and Structural Parameters* [J], *The Econometrics Journal*, 21(1), C1~C68.

[25] Chernozhukov V., Hansen C., Kallus N., et al., 2024, *Applied Causal Inference Powered by ML and AI* [R], arXiv preprint, No. 2403.02467.

[26] Clarke P. S., Polsell A., 2025, *Double machine learning for static panel models with fixed effects* [J], *The Econometrics Journal*, utaf011.

[27] Crump R.K., Hotz V.J., Imbens G.W., Mitnik O.A., 2009, *Dealing With Limited Overlap in Estimation of Average Treatment Effects* [J], *Biometrika*, 96(1), 187~199.

[28] Egami N., Hinck M., Stewart B. M., Wei H., 2024, *Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models* [R], arXiv preprint, No. 2306.04746.

[29] Frölich M., Huber M., 2019, *Including Covariates in the Regression Discontinuity Design* [J], *Journal of Business & Economic Statistics*, 37(4), 736~748.

[30] Gu S., Kelly B., Xiu D., 2020, *Empirical Asset Pricing via Machine Learning* [J], *The Review of Financial Studies*, 33(5), 2223~2273.

[31] Hatamyar J., Kreif N., Rocha R., Huber M., 2023, *Machine Learning for Staggered Difference-in-differences and Dynamic Treatment Effect Heterogeneity* [R], arXiv preprint, No. 2310.11962.

[32] Heiler P., Kazak E., 2021. *Valid Inference for Treatment Effect Parameters Under Irregular Identification and Many Extreme Propensity Scores* [J], *Journal of Econometrics*, 222(2), 1083~1108.

[33] Li F., Morgan K. L., Zaslavsky A. M., 2018, *Balancing Covariates Via Propensity Score Weighting*. *Journal of the American Statistical Association*, 113(521), 390~400.

[34] Li R., Song H., Zhang J., Zhang J., 2025, *The Effects of a Multifaceted Poverty Alleviation*

Program on Rural Income and Household Behavior in China [J], *American Economic Journal: Economic Policy*, 17(2), 319~357.

[35] Lu C., Nie X., Wager S., 2019, *Robust Nonparametric Difference-in-differences Estimation* [R], arXiv e-prints, No. arXiv-1905.

[36] Ma Y., Sant'Anna P.H.C., Sasaki Y., Ura T., 2023, *Doubly Robust Estimators with Weak Overlap* [R], arXiv preprint, No. 2304.08974.

[37] Noack C., Olma T., Rothe C., 2021, *Flexible Covariate Adjustments in Regression Discontinuity Designs* [R], arXiv preprint, No. 2107.07942.

[38] Rubin D.B., 2005, *Causal inference using potential outcomes: Design, modeling, decisions* [J], *Journal of the American statistical Association*, 100(469), 322~331.

[39] Sant'Anna P.H., Zhao J., 2020, *Doubly Robust Difference-in-differences Estimators* [J], *Journal of Econometrics*, 219(1), 101~122.

[40] Schmidt-hieber J., 2020, *Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function* [J], *Annals of Statistics*, 48(4), 1875~1897.

[41] Wager S., G. Walther, 2016, *Adaptive Concentration of Regression Trees, with Application to Random Forests* [R]. Preprint (arXiv:1503.06388).

[42] Xie H.T., 2024, *Efficient and Robust Estimation of the Generalized LATE Model* [J], *Journal of Business & Economic Statistics*, 42(3), 1053~1065.

The Theory and Practices of Double Machine Learning: From “Black Box” to “Tool Box”

HU Shiyun¹ JIANG Hongyi¹ XIE Haitian²

(1.National School of Development, Peking University;

2.Guanghua School of Management, Peking University)

Summary: Double machine learning (DML) is a significant methodological advancement in causal inference, particularly when high-dimensional confounders and complex nonlinear relationships are involved. Despite its growing adoption in empirical economics, misconceptions persist regarding its role and proper application. This study provides a comprehensive overview of DML, clarifying its theoretical foundations, dispelling common misunderstandings, and offering practical guidance for empirical researchers. We emphasize that DML is not a novel identification strategy but rather a sophisticated estimation framework that enhances robustness within established causal inference paradigms.

We begin by systematically reviewing the theoretical underpinnings of DML, which

rest on two fundamental principles—neyman orthogonality and sample splitting (cross-fitting). Neyman orthogonality renders the estimator of the causal parameter asymptotically insensitive to first-order estimation errors in nuisance parameters, while cross-fitting prevents overfitting bias by using separate subsamples for nuisance parameter estimation and structural parameter inference. Our theoretical exposition progresses from the partially linear model (PLM), which accommodates homogeneous treatment effects, to the interactive regression model (IRM), which allows for fully flexible treatment effect heterogeneity, and culminates in a general method of moments framework. We pay particular attention to the interpretation of the PLM coefficient, revealing that it corresponds to an overlap-weighted average treatment effect when treatment effects are heterogeneous and covariate overlap is imperfect.

We then demonstrate how DML integrates with standard identification strategies in applied econometrics, including instrumental variables, difference-in-differences, and regression discontinuity designs. For each strategy, we review the relevant identifying assumptions, explain how DML can be employed for parameter estimation while maintaining identification, and provide detailed implementation algorithms. Through systematic comparison with conventional parametric approaches, we clarify that DML's value lies in strengthening estimation procedures under existing identification frameworks rather than supplanting them.

To assess DML's performance characteristics and practical limitations, we conduct Monte Carlo simulations comparing PLM and IRM against two-way fixed effects estimators in difference-in-differences settings across various data-generating processes. The simulation evidence demonstrates that DML methods deliver substantial gains in estimation accuracy when the true data-generating process features nonlinear confounding relationships and heterogeneous treatment effects. However, we document that PLM exhibits markedly superior performance relative to IRM when covariate overlap is limited, highlighting important practical considerations for method selection.

We illustrate DML's empirical utility through a reanalysis of China's targeted poverty alleviation policy and its effects on rural household income. Contrasting DML estimates with those from conventional fixed effects specifications, we employ random forests for nuisance parameter estimation. The DML analysis confirms a positive and statistically significant policy effect, although the magnitude is more conservative than linear model estimates, reflecting DML's ability to flexibly account for nonlinear confounding trends that may be imperfectly controlled in parametric specifications. Moreover, the flexibility of

DML facilitates detailed heterogeneity analysis, revealing that the policy effects were substantially higher in counties with revolutionary historical legacies and in mountainous regions.

Drawing on theoretical insights, simulation evidence, and empirical applications, we offer practical recommendations for implementing DML in applied research. These guidelines follow the natural research workflow as follows: (1) Prioritize identification over estimation—establish credible identifying assumptions through standard econometric reasoning before employing DML for estimation. (2) Base the decision to use DML and the choice between IRM and PLM on sample size considerations and the extent of covariate overlap. (3) Select machine learning algorithms (e.g., penalized regression, random forests, and neural networks) based on the dimensionality of the feature space relative to sample size. (4) Conduct careful diagnostics of nuisance parameter estimates, examining propensity score distributions and checking boundary predictions to guide hyperparameter tuning. (5) Implement comprehensive specification checks, including covariate balance tests and sensitivity analyses, and compare estimates across alternative specifications. (6) Adopt transparent reporting practices that document machine learning model selection and nuisance parameter fit quality, without unnecessarily reiterating the DML theory or reporting aggregate model fit statistics that may be misleading.

In conclusion, DML is a tool for enhancing estimation precision and robustness but not a substitute for rigorous research design. It equips researchers with principled methods for controlling high-dimensional confounding and accommodating complex functional form flexibility. By enabling more robust estimation and credible inference in the presence of model misspecification concerns, DML is a valuable addition to the applied econometrician's toolkit for drawing reliable causal inferences from the increasingly complex and high-dimensional data characteristic of modern economic research.

Keywords: Double Machine Learning; Causal Inference; High-dimensional Data; Econometrics; Empirical Strategy

JEL Classification: C14; C21; C55; C23

(责任编辑:李兆辰;数据编辑:元风行)