Market Research and Analysis

Lecture 5: Double Machine Learning for Causal Inference

- Zhenyu Zhao
- Nankai Institute of International Economics
- Feishu: 2120253538
- Email: zzynankai@outlook.com
- Website: xishanyu2.github.io

01

双重机器学习：原理、实现与应用

> 就像2SLS是做两次OLS，“双重机器学习”(Double Machine Learning) 就是做两次机器学习。

1. 简介

在实证研究中，我们经常面临这样的困境：一方面，传统的线性回归模型可能过于简化，无法捕捉数据中复杂的非线性关系；另一方面，直接使用机器学习方法 (如随机森林、神经网络) 进行因果推断又会引入严重的偏误。那么，能否既利用机器学习的灵活性，又保证因果推断里参数估计的无偏性和有效性呢?

Double/Debiased Machine Learning(DDML，双重机器学习)正是为解决这一问题而生。这一方法由MIT经济学家Chernozhukov及其合作者系统提出，已成为因果推断领域最重要的方法之一。DDML 的核心思想可以概括为：通过**正交化**(orthogonalization)消除正则化偏误，通过**交叉拟合**(cross-fitting)消除过拟合偏误，从而实现对因果参数的一致估计和渐近正态推断。

本文旨在以通俗易懂的方式介绍DDML的基本原理、实现步骤和应用场景，帮助读者理解这一强大工具如何将机器学习与因果推断完美结合。

2. 问题的提出：为什么需要DDML?

2.1 传统方法的局限

考虑一个典型的因果推断问题：我们想要估计处理变量 D 对结果变量 Y 的因果效应 θ_0，同时需要控制一组协变量 X ，在传统的计量经济学中，通常使用线性回归：

$$Y = D\theta_0 + X'\beta + U$$

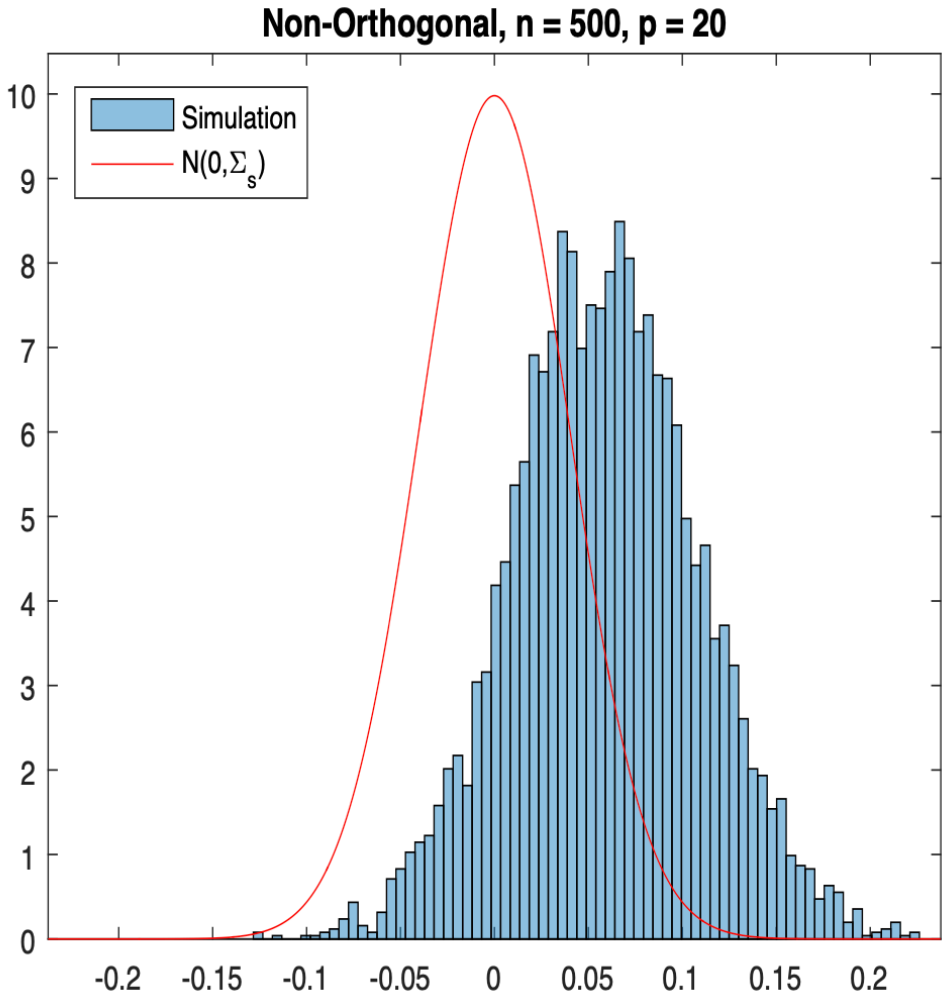
这一方法简单直观，但存在两个关键问题：

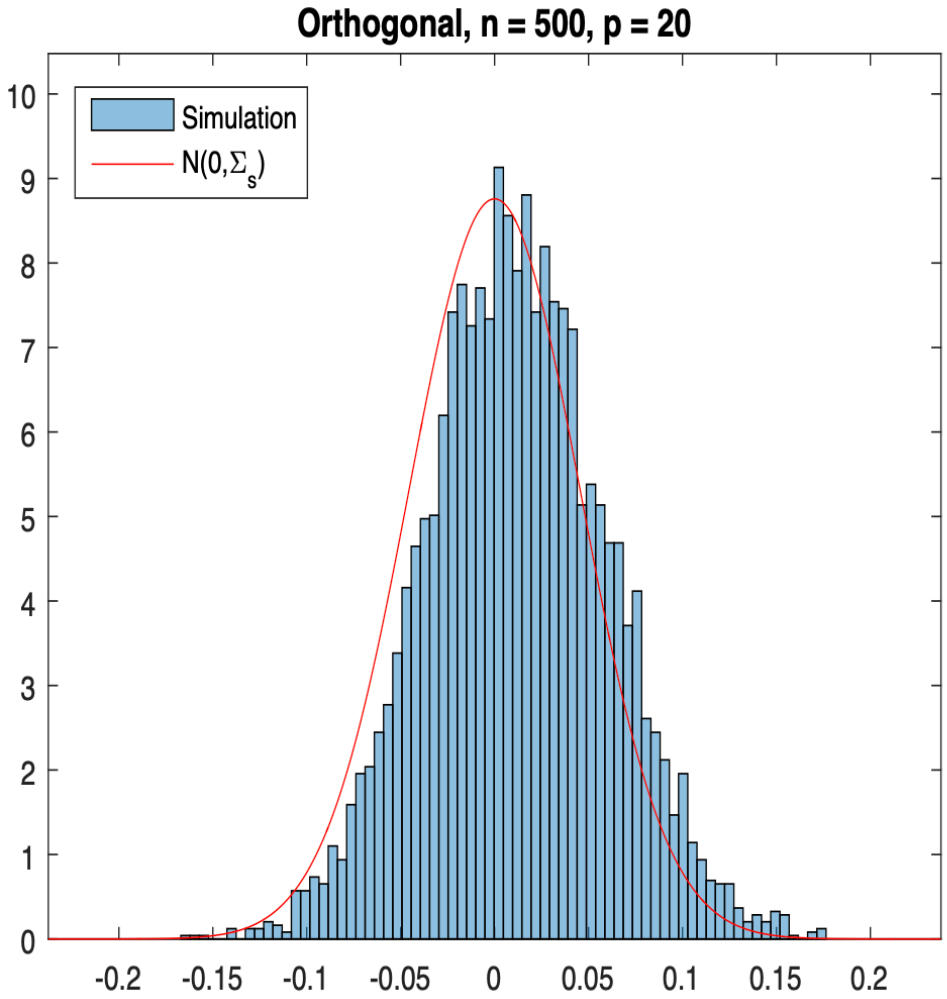
1. **模型误设**：模型假设 X 对 Y 的影响是线性的，但现实中这种关系往往是高度非线性的。这种函数形式假设过强，如果函数形式设定错误，θ_0 的估计就会有偏。
2. **高维问题**：当控制变量维度很高 (协变量个数 p 接近甚至超过样本量 n) 时，传统OLS会失效。

2.2 直接使用机器学习的困境

那么，能否直接用机器学习方法 (如Lasso、RandomForest) 来估计上述模型呢?
答案是否定的。直接使用机器学习会引入两种偏误：

1. **正则化偏误**：大多数机器学习方法为了防止过拟合，会对参数施加惩罚 (如Lasso的 L_1 惩罚)。这种正则化虽然提高了预测精度，却会导致参数估计有偏。
2. **过拟合偏误**：如果在同一份数据上既训练机器学习模型又估计因果参数，模型会过度拟合训练数据，导致方差膨胀和偏误。
   这两种偏误使得直接用机器学习得到的 $\hat{\theta}$ 既不无偏，也无法进行有效的统计推断 (比如构造置信区间、进行假设检验)。


Non-Orthogonal, n = 500, p = 20
Simulation
N(0,Σs)
10
9
8
7
6
5
4
3
2
1
0
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

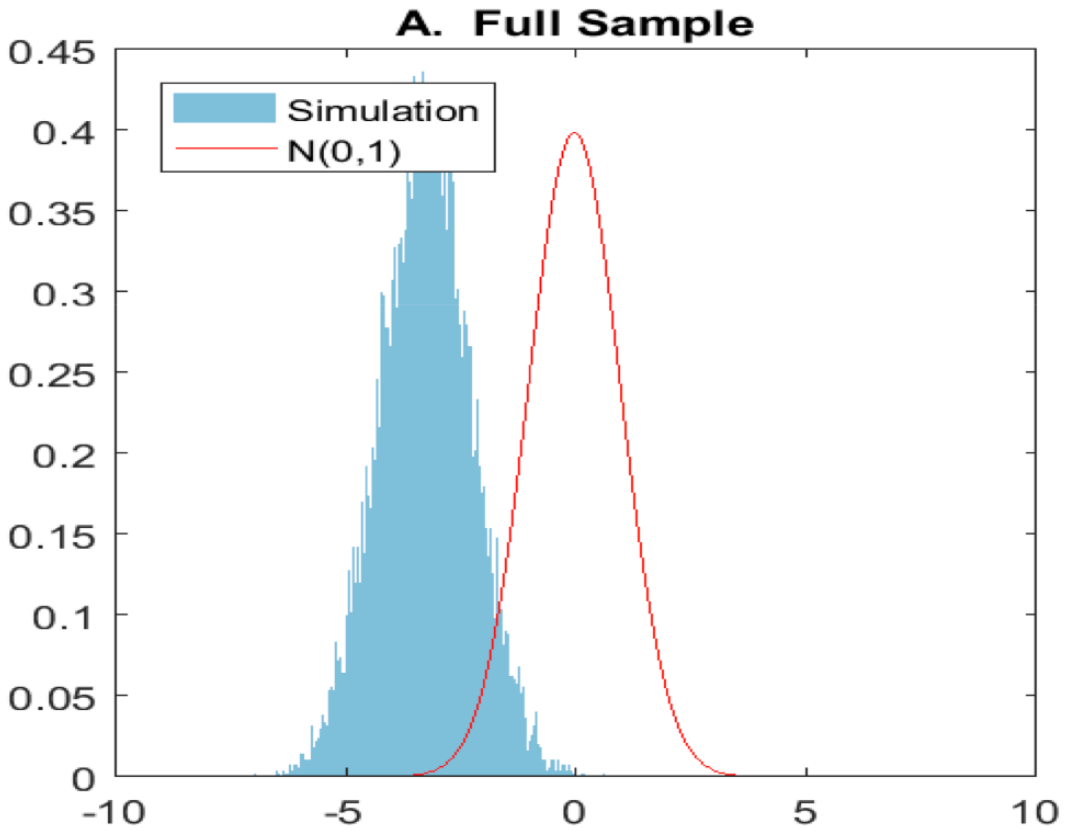

Orthogonal, n = 500, p = 20
Simulation
N(0,Σs)
10
9
8
7
6
5
4
3
2
1
0
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

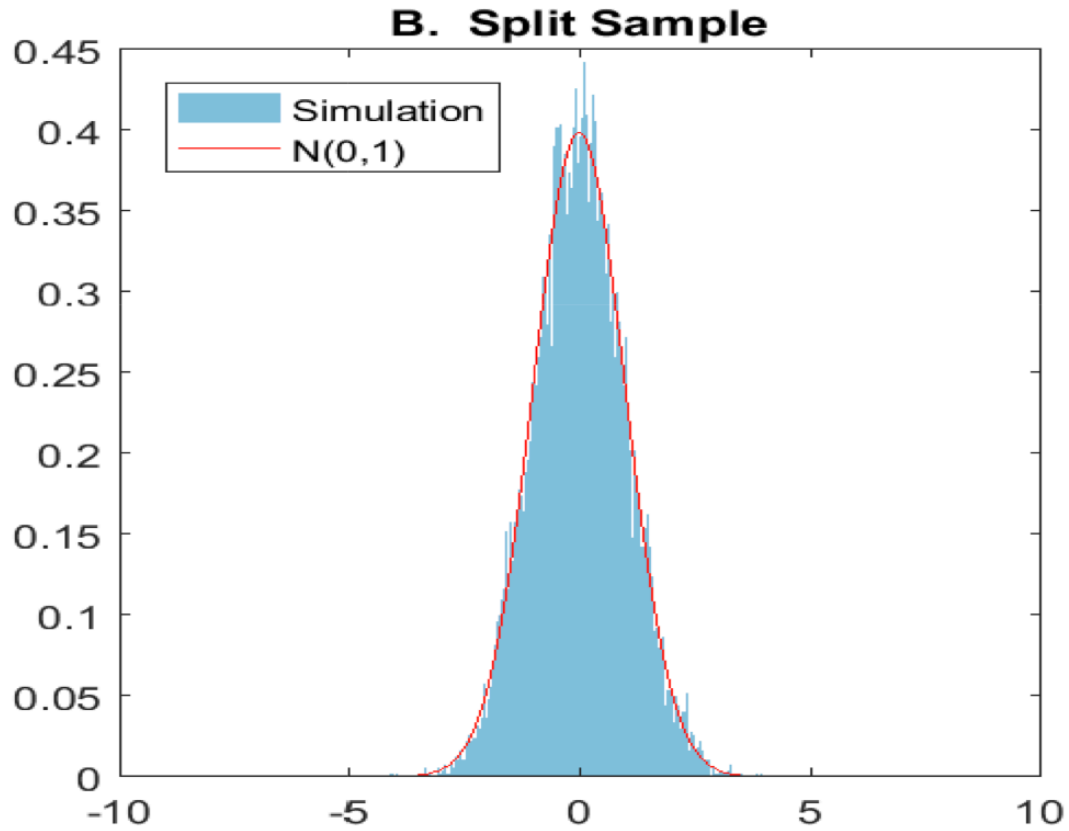
2.3 DDML的解决方案

DDML通过两个关键步骤巧妙地解决了上述问题：

1. **正交化**：借鉴*Frisch-Waugh-Lovell*(FWL)定理的思想，先用机器学习分别预测 Y 和 D 关于 X 的条件期望，然后用残差进行因果推断。这一步消除了正则化偏误。
2. **交叉拟合**：将数据分成 K 折，在 $K-1$ 个子集上训练机器学习模型，然后在剩余折上进行预测和估计。这一步消除了过拟合偏误。

通过这两步，DDML 实现了对因果参数 θ_0 的一致估计，并且在弱正则性条件下，$\hat{\theta}$ 渐近正态，可以进行标准的统计推断。


A. Full Sample
Simulation
N(0,1)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-5
0
5
10


B. Split Sample
Simulation
N(0,1)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-5
0
5
10

3. DDML的核心原理

3.1 部分线性模型(Partially Linear Model)

为了理解DDML的工作原理，我们先从最简单的部分线性模型开始：

$$Y = D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0$$

$$D = m_0(X) + V, \quad \mathbb{E}[V|X] = 0$$

其中：

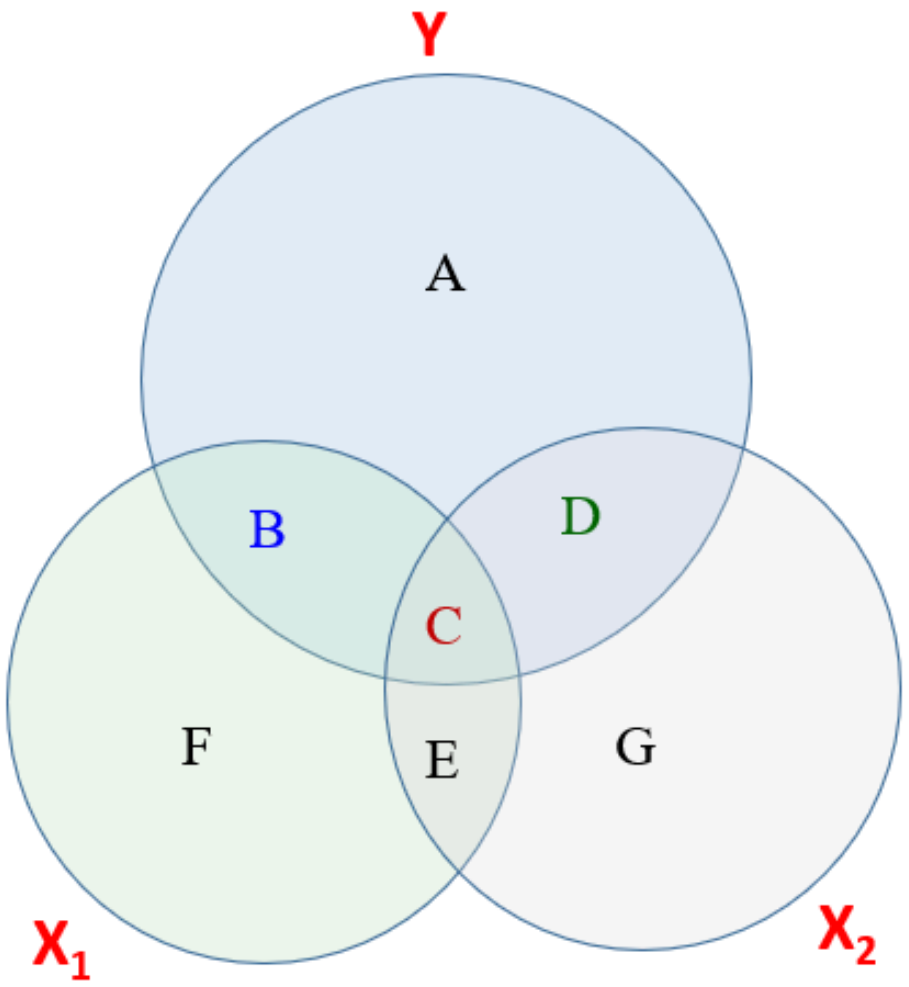
- θ_0 是我们感兴趣的因果参数 (处理效应)
- $g_0(X) = \mathbb{E}[Y|X, D = 0]$ 是结果变量关于协变量的条件期望
- $m_0(X) = \mathbb{E}[D|X]$ 是处理变量关于协变量的条件期望

部分线性体现在：D 对 Y 的影响是线性的，但 X 对 Y 和 D 的影响可以是任意非线性函数 $g_0(\cdot)$ 和 $m_0(\cdot)$

3.2 正交化：FWL定理的机器学习版本

FWL定理告诉我们，在线性模型 $Y = \theta_0 D + X'\beta + U$ 中，以下两种估计 θ_0 的方法是等价的：

1. **直接回归**：将 Y 对 D 和 X 一起回归
2. **分步回归**：
   - 第1步：将 D 对 X 回归，得到残差 $\tilde{D} = D - X'\hat{\gamma}$
   - 第2步：将 Y 对 X 回归，得到残差 $\tilde{Y} = Y - X'\hat{\delta}$
   - 第3步：将 $\tilde{Y}$ 对 $\tilde{D}$ 回归，得到 $\hat{\theta}_0$



令A+B+C+D=1

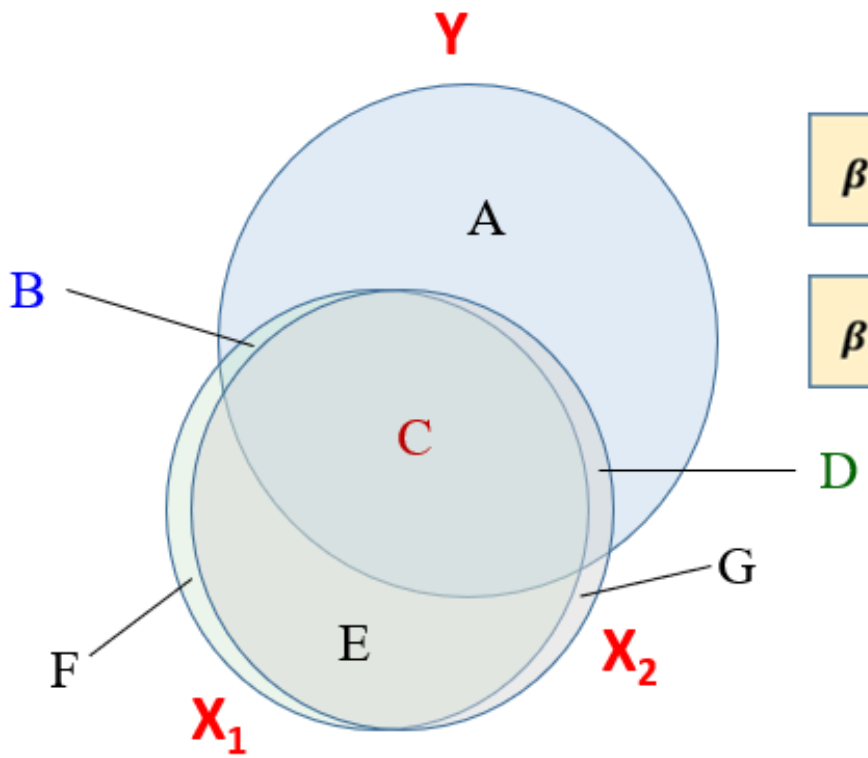
$R^2 = \frac{B+C}{A+B+C+D}$

reg Y X1 → R^2 = B+C

reg Y X2 → R^2 = C+D

reg Y X1 X2 → R^2 = B+C+D

Case



$\beta_1 = \frac{\partial Y}{\partial X1}|_{X2}$

$\beta_2 = \frac{\partial Y}{\partial X2}|_{X1}$

reg Y X1 X2 → R^2 = B+C+D

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

reg Y X1 → $\hat{u}$ = A+D

reg X2 X1 → $\hat{e}$ = D+G

reg $\hat{u}$ $\hat{e}$ → R^2 = D

干净的Y　干净的X2

$D = \widehat{\beta_2}|_{X1}$

$B = \widehat{\beta_1}|_{X2}$

DDML将这一思想推广到非线性情形：用机器学习方法估计 $g_0(X)$ 和 $m_0(X)$，然后用残差估计 θ_0 。
具体而言，DDML的估计量为：

$$\hat{\theta}_0 = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{V}_i (Y_i - \hat{g}_0(X_i))$$

其中 $\hat{V}_i = D_i - \hat{m}_0(X_i)$

Neyman正交性

估计方程对冗余参数(nuisance parameter) η_0 的一阶导数为零：

$$\frac{\partial}{\partial \eta} \mathbb{E}[\psi(W; \theta_0, \eta)]_{\eta=\eta_0} = 0$$

这保证了nuisance函数估计误差对因果参数估计的影响是二阶小量，从而实现了去偏。

3.3 交叉拟合

如果在同一份数据上既训练机器学习模型 ($\hat{g}_0$ 和 $\hat{m}_0$) 又估计 θ_0，会出现过拟合问题：第1、2步中的噪声会进入到第3步估计中，导致估计量有偏且方差很大。

交叉拟合通过样本分割解决这一问题。具体步骤如下：

1. 将数据随机分成 K 折 (通常 $K = 5$)
2. 对于第 k 折：
    - 在其它 $K - 1$ 折上训练机器学习模型，得到 $\hat{g}_0^{(-k)}$ 和 $\hat{m}_0^{(-k)}$
    - 在第 k 折数据上用这些模型进行预测，得到残差 $\hat{Y}_i^{(k)}$ 和 $\hat{V}_i^{(k)}$
3. 合并所有折的残差，计算最终的 $\hat{\theta}_0$

交叉拟合的关键在于：**预测和估计使用的是不同的数据**。这样，即使机器学习模型过拟合了训练数据，也不会影响在测试折上的估计。

4. DDML的理论性质

4.1 一致性

DDML理论的优雅之处在于，它对机器学习方法的要求相对宽松。只要nuisance函数的估计满足：

$$\|g_0 - \hat{g}_0\|_{L_2} = o_P(n^{-1/4})$$

$$\|m_0 - \hat{m}_0\|_{L_2} = o_P(n^{-1/4})$$

即 $g_0 - \hat{g}_0$ 和 $m_0 - \hat{m}_0$ 的收敛速度快于 $n^{-1/4}$时，DDML估计量就是$\sqrt{n}$-一致的。这一条件对大多数常用的机器学习方法 (Lasso, Random Forest, Gradient Boosting, Neural Networks) 都能满足。

4.2 渐近正态性

在适当的正则性条件下，DDML估计量满足：

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$$

其中 V 是渐近方差。

5. DDML的应用场景

1. 部分线性模型(partially linear model)

$$Y = D\theta + g(X) + U, D = m(X) + V$$

2. 交互模型(interactive model)

$$Y = g(X, D) + U, D = m(X) + V$$

3. 部分线性工具变量模型(partially linear IV model)

$$Y = D\theta + g(X) + U, Z = m(X) + V$$

4. 灵活部分线性工具变量模型(flexible partially linear IV model)

$$Y = D\theta + g(X) + U, D = m(Z) + g(X) + V$$

5. 交互工具变量模型(interactive IV model)

$$Y = g(Z, X) + U, D = h(Z, X) + V, Z = m(X) + E$$

6. 实证案例

A. 失业保险金对失业持续时间的影响

美国劳工部于1980年代开展的宾夕法尼亚再就业奖金实验，旨在检验不同失业保险(UI)补偿方案的激励效应。在这个实验中，失业保险申领者被随机分配到对照组或五个处理组之一：对照组适用标准的失业保险规则；处理组中的个体若在规定的资格期内找到工作并保持一定工作时长则可获得现金奖金。不同处理组在奖金水平、资格期长短以及奖金是否随时间递减等方面存在差异(Bilias & Koenker, 2002)。

在本实证分析中，我们仅关注最慷慨的补偿方案——处理组4，并剔除了接受其他处理的个体。在该处理组中，奖金金额较高，资格期较长，且申领者有资格参加职业培训。处理变量D为是否被分配到处理组4的二值变量，结果变量Y为失业持续时间的对数，协变量X包括年龄组哑变量、性别、种族、受抚养人数、实验所在季度、州内地理位置、是否存在召回预期以及职业类型。

基于五种方法估计用于构建正交估计方程的扰动函数,分别为：随机森林(Random Forest)、单棵回归树(Reg. Tree)、提升树(Boosting)、Lasso(ℓ_1惩罚线性回归)和神经网络(Neural Net)。此外，我们还考虑了两种混合方法：Ensemble和Best。Ensemble通过加权平均四种机器学习方法(Lasso、Boosting、Random Forest、Neural Net)来实现最优组合，权重通过5折交叉验证下的平均均方预测误差最小化确定。Best则是从五种方法和Ensemble中，为每个扰动函数选择在对应目标变量上具有最低样本外预测误差的方法。

表1报告了基于DML的平均处理效应估计结果，采用中位数方法，基于100次不同的样本划分，分别使用2折和5折交叉拟合。对于异质性效应模型(Panel A)和部分线性模型(Panel B)，ATE估计值在-0.073至-0.085之间，标准误约为0.035–0.036。所有方法在5%显著性水平下均显示奖金对失业持续时间具有显著负向影响，与先前研究结论一致。

Table 1. Estimated Effect of Cash Bonus on Unemployment Duration

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Regression Model							
ATE (2 fold)	-0.081	-0.084	-0.074	-0.079	-0.073	-0.079	-0.078
	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.081	-0.085	-0.074	-0.077	-0.073	-0.078	-0.077
	[0.036]	[0.036]	[0.036]	[0.035]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
B. Partially Linear Regression Model							
ATE (2 fold)	-0.080	-0.084	-0.077	-0.076	-0.074	-0.075	-0.075
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.080	-0.084	-0.077	-0.074	-0.073	-0.075	-0.074
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Note: Estimated ATE and standard errors from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

B. 401(k)资格及参与对净金融资产的影响

确定401(k)资格对净金融资产影响的关键难点在于：为提供401(k)计划的公司工作并非随机分配。为克服这一问题，遵循Poterba et al. (1994a, 1994b)的策略，使用1991年收入与参与计划调查(SIPP)数据。在401(k)计划初期，个体不太可能基于雇主是否提供该计划来决定就业，而更关注收入及其他工作特征，因此在适当控制收入等变量后，401(k)资格可视为外生。

本例中，结果变量Y为净金融资产（IRA余额、401(k)余额、支票账户、美国储蓄债券、其他利息账户、个人持有的债券、股票、共同基金等减去非抵押债务），处理变量D为是否具有401(k)资格资格，协变量X包括年龄、收入、家庭规模、教育年限、婚姻状况、双职工状态、确定性养老金状态、IRA参与状态及住房拥有状态。

表2报告了DML对401(k)资格平均处理效应的估计，采用与前述相同的机器学习方法。为减少交互模型中极端倾向得分权重的过大影响，我们将倾向得分在0.01和0.99处截断。基于100次样本划分的中位数估计显示：交互模型ATE约为6830–8105美元（标准误约1134–1533）；部分线性模型ATE约为7717–9247美元（标准误约1294–1749）。相较于不控制任何协变量时的基准估计值19559美元，加入灵活控制后效应显著减弱，表明存在较强的混淆效应。不同机器学习方法的结果高度一致，与Poterba等人的简单线性控制结果也基本吻合。

Table 2. Estimated Effect of 401(k) Eligibility on Net Financial Assets

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Regression Model							
ATE (2 fold)	6830	7713	7770	7806	7764	7702	7546
	[1282]	[1208]	[1276]	[1159]	[1328]	[1149]	[1360]
	(1530)	(1271)	(1363)	(1202)	(1468)	(1170)	(1533)
ATE (5 fold)	7170	7993	8105	7713	7788	7839	7753
	[1201]	[1198]	[1242]	[1155]	[1238]	[1134]	[1237]
	(1398)	(1236)	(1299)	(1177)	(1293)	(1148)	(1294)
B. Partially Linear Regression Model							
ATE (2 fold)	7717	8709	9116	8759	8950	9010	9125
	[1346]	[1363]	[1302]	[1339]	[1335]	[1309]	[1304]
	(1749)	(1427)	(1377)	(1382)	(1408)	(1344)	(1357)
ATE (5 fold)	8187	8871	9247	9110	9038	9166	9215
	[1298]	[1358]	[1295]	[1314]	[1322]	[1299]	[1294]
	(1558)	(1418)	(1328)	(1328)	(1355)	(1310)	(1312)

Note: Estimated ATE and standard errors from a linear model (Panel B) and heterogeneous effect model (Panel A) based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

进一步地，将参与401(k)计划作为内生处理变量，以资格作为工具变量，估计局部平均处理效应。表3结果显示，参与401(k)对净金融资产的LATE估计在8944–11764美元之间，均显著为正，与线性IV基准估计13102美元（标准误1922）相比略有减弱，但方向与显著性一致。

Table 3. Estimated Effect of 401(k) Participation on Net Financial Assets

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
LATE (2 fold)	8978	11073	11384	11329	11094	11119	10952
	[2192]	[1749]	[1832]	[1666]	[1903]	[1653]	[1657]
	(3014)	(1849)	(1993)	(1718)	(2098)	(1689)	(1699)
LATE (5 fold)	8944	11459	11764	11133	11186	11173	11113
	[2259]	[1717]	[1788]	[1661]	[1795]	[1641]	[1645]
	(3307)	(1786)	(1893)	(1710)	(1890)	(1678)	(1675)

Note: Estimated LATE based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

C. 制度对经济增长的影响

为展示部分线性结构方程模型结合工具变量的DML估计，沿用Acemoglu et al. (2001)的框架，重新估计制度对总产出的影响。结果变量Y为对数人均GDP，内生解释变量D为对个人财产的保护程度（作为制度质量代理），工具变量Z为早期欧洲定居者死亡率。控制变量X包括距赤道距离以及非洲、亚洲、北美、南美洲的哑变量。

表4报告了DML估计结果，系数在0.73–1.00之间，均显著为正。尽管略小于AJR原文基线估计1.10（标准误0.46），但定性结论一致。DML通过机器学习方法放松了文章对地理因素仅作线性控制的假设，结果显示制度对经济增长具有显著正向影响。

Table 4. Estimated Effect of Institutions on Output

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
2 fold	0.85	0.81	0.84	0.77	0.94	0.8	0.83
	[0.28]	[0.42]	[0.38]	[0.33]	[0.32]	[0.35]	[0.34]
	(0.22)	(0.29)	(0.3)	(0.27)	(0.28)	(0.3)	(0.29)
5 fold	0.77	0.95	0.9	0.73	1.00	0.83	0.88
	[0.24]	[0.46]	[0.41]	[0.33]	[0.33]	[0.37]	[0.41]
	(0.17)	(0.45)	(0.4)	(0.27)	(0.3)	(0.34)	(0.39)

Note: Estimated coefficient from a linear instrumental variables model based on orthogonal estimating equations. Column labels denote the method used to estimate nuisance functions. Results are based on 100 splits with point estimates calculated the median method. The median standard error across the splits are reported in brackets and standard errors calculated using the median method to adjust for variation across splits are provided in parentheses. Further details about the methods are provided in the main text.

7. DDML的优势与局限

7.1 主要优势

1. **灵活性**：无需指定 $g_0(\cdot)$ 和 $m_0(\cdot)$ 的函数形式，机器学习自动学习
2. **稳健性**：对模型误设定稳健，只要ML估计收敛速度足够快
3. **高维友好**：可以处理 $p \gg n$ 的情形 (协变量维度远大于样本量)
4. **有效推断**：提供渐近正态的估计量，可以进行标准的假设检验
5. **通用性**：适用于多种因果推断模型(PLM, IRM, IV等)

7.2 主要局限

1. **仍需因果识别假设**：DDML不能解决遗漏变量偏误、反向因果等问题。它只是提供了更灵活的估计方法，但因果识别仍然依赖于标准假设 (无混淆、工具变量有效性等)
2. **黑箱问题**：机器学习模型的可解释性较差，难以理解 X 如何影响 Y 和 D
3. **计算成本**：交叉拟合和多次训练ML模型增加了计算负担，尤其在大数据集上
4. **调参复杂性**：机器学习方法通常有很多超参数需要仔细调参
5. **小样本表现未知**：DDML的理论性质是渐近的，在小样本下的表现缺乏充分研究

8. 软件实现

8.1 Stata

- ddml (Ahrens et al., 2024)：功能最全面的Stata包，支持多种模型，可以与 `cvlasso`, `rforest`, `pystacked` 等ML命令结合使用。链接

8.2 Python

- DoubleML (Bach et al., 2024)：基于 `scikit-learn` 的Python包，支持多种DDML模型和ML算法。链接
- EconML (Microsoft Research)：微软开发的因果推断工具包，包含DDML及其它方法。链接

8.3 R

- DoubleML (Lang et al., 2019)：基于 `mlr3` 生态系统的R包。链接

9. 总结

Double/Debiased Machine Learning是近年来因果推断领域最重要的进展之一。它巧妙地将机器学习的预测能力与计量经济学的因果推断框架结合，实现了鱼与熊掌兼得：既利用了ML的灵活性处理高维、非线性问题，又保证了因果参数的无偏估计和有效推断。

参考文献

- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." American Economic Review, 107 (5): 261-265. Link, PDF, Replication
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey,J. Robins, 2018, Double/debiased machine learning for treatment and structural parameters, The Econometrics Journal, 21 (1): C1-C68. Link, PDF
- Ahrens, A., Hansen, C. B., Schaffer, M. E., & Wiemann, T. (2024). ddml: Double/debiased machine learning in Stata. The Stata Journal, 24(1), 3–45. Link, PDF, Google
- Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2023). pystacked: Stacking generalization and machine learning in Stata. The Stata Journal, 23(4), 909–931. Link, PDF, Google
- Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. The Stata Journal, 20(1), 176–235. Link, PDF, Google

相关推文

- 李梦玉, 2025, DML-CER：使用双重机器学习克服面板数据中的不可观测异质性, 连享会 No.1687.
- 张逸林, 2025, 异质性分析的新视角：政策效应分解, 连享会 No.1684.
- 赵俊, 2025, 静态面板数据下的双重机器学习模型（下）——R代码实操, 连享会 No.1665.
- 赵俊, 2025, 静态面板数据下的双重机器学习模型（上）—— 理论基础, 连享会 No.1664.
- 李俊奇, 2025, Python-EconML包：快速上手动态双重机器学习, 连享会 No.1577.
- 李金桐, 2023, 因果推断：双重机器学习-ddml, 连享会 No.1221.
- 王卓, 2023, Python：从随机实验到双重机器学习, 连享会 No.1204.
- 董洁妙, 2022, Stata：双重机器学习-多维聚类标准误的估计方法-crhdreg, 连享会 No.1036.

02

双重机器学习在经济学中的研究进展

[1]张涛,李均超. 网络基础设施、包容性绿色增长与地区差距——基于双重机器学习的因果推断[J].数量经济技术经济研究,2023,40(04):113-135.

[2]蔡运坤,周京奎,袁旺平. 数据要素共享与城市创业活力——来自公共数据开放的经验证据[J].数量经济技术经济研究,2024,41(08):5-25.

[3]蒋金荷,黄珊. 贸易新业态对绿色技术创新的影响研究——来自跨境电商综合试验区政策的证据[J].数量经济技术经济研究,2024,41(12):133-154.

[1]孙传旺,占妍泓,徐梦洁. 电力需求响应信号与新能源制造企业绩效[J].管理世界,2024,40(12):185-203.

[2]王渊,李牧南,梁彦希. 政策取向一致性与企业高质量绿色转型[J].管理世界,2025,41(07):108-139.

[3]杨丹,朱珠,刘自敏,等. 共同富裕目标下农产品区域公用品牌的收入效应研究——来自原国家级贫困县的经验证据[J].管理世界,2025,41(07):149-175.

先导性文章：王茹婷等（2022.4）

> 王茹婷,彭方平,李维,等.打破刚性兑付能降低企业融资成本吗？[J].管理世界,2022,38(04):42-64.

- **关键词**：刚性兑付；融资成本；风险溢价；双重机器学习
- **摘要**：本文结合债券市场和上市公司数据，应用双重机器学习方法，检验了打破刚性兑付这一事件对中国债券市场利率的影响。本文发现，打破刚性兑付不仅没有有效降低企业融资成本，反而引起市场整体融资成本上升，且对于不同企业的影响表现出显著的异质性。具体来讲，打破刚性兑付使得低信用企业，特别是高风险民营企业的融资成本显著提高；而对高信用民营企业融资成本没有显著影响。上述研究结论说明在无法有效降低风险溢价的情况下，寄希望于打破刚性兑付来降低企业融资成本是不现实的。

开创性文章：张涛和李均超（2023.4）

> 张涛,李均超. 网络基础设施、包容性绿色增长与地区差距——基于双重机器学习的因果推断[J].数量经济技术经济研究,2023,40(04):113-135.

- 双重机器学习做基准回归、稳健性检验、机制检验、重抽样检验（中位数调整）

```
*常见报错：Cross-fitting fold 1 unrecognized command
```

```
ssc install ddml
ssc install pystacked
set python_userpath "D:\StataNow19\utilities", permanently
set python_exec "D:\anaconda3\python.exe", permanently
python query
```

```
///基准回归
global Y y
global X x1 x2 x3
global D d
set seed 42
ddml init partial, kfolds(5)
ddml E[D|X]: pystacked $D $X, type(reg) method(rf)
ddml E[Y|X]: pystacked $Y $X, type(reg) method(rf)
ddml crossfit
/*
Cross-fitting E[y|X] equation: y
Cross-fitting fold 1 2 3 4 5 ...completed cross-fitting
Cross-fitting E[D|X] equation: d
Cross-fitting fold 1 2 3 4 5 ...completed cross-fitting
*/
ddml estimate, robust
```

```
///稳健性检验

//改变样本分割比例
ddml init partial, kfolds(3)
*ddml init partial, kfolds(8)

//更换机器学习方法
ddml E[D|X]: pystacked $D $X, type(reg) method(lassocv)
ddml E[Y|X]: pystacked $Y $X, type(reg) method(lassocv)
*ddml E[D|X]: pystacked $D $X, type(reg) method(gradboost)
*ddml E[Y|X]: pystacked $Y $X, type(reg) method(gradboost)
*ddml E[D|X]: pystacked $D $X, type(reg) method(nnet)
*ddml E[Y|X]: pystacked $Y $X, type(reg) method(nnet)

//交互式模型
ddml init interactive, kfolds(5)

//工具变量法
ddml init iv, kfolds(5)
ddml E[Y|X]: pystacked $Y $X, type(reg) method(rf)
ddml E[D|X]: pystacked $D $X, type(reg) method(rf)
ddml E[Z|X]: pystacked $Z $X, type(reg) method(rf)
```

双重机器学习处理内生性问题

> 蔡运坤,周京奎,袁旺平. 数据要素共享与城市创业活力——来自公共数据开放的经验证据[J].数量经济技术经济研究,2024,41(08):5-25.

（6）双重机器学习因果检验。选择性**内生性问题**仍然可能冲击回归结果的稳健性。此部分，借鉴Bodory等（2022）提出的双重机器学习法重新检验。与传统的因果推断方法相比，双重机器学习方法的优势体现在变量选择和模型设定上。一方面，企业家创业动机是经济社会互动与共生的复杂环境关系交互的结果，本文控制了多维度影响因素，但设定回归模型时可能因**“维度诅咒”与多重共线性问题**而导致估计结果偏误。通过机器学习及其正则化算法，双重机器学习方法将在预选项中自动筛选出精度较高的控制变量集合，由此避免高维控制变量冗余导致的“维度诅咒”，得到可靠的估计量。另一方面，双重机器学习的优势还体现在基于机器学习方法**处理变量非线性问题**上，由此可以缓解经济系统中变量非线性关系引致的模型设定偏误。

基于此，采用双重机器学习方法重新检验公共数据开放与城市创业活力间的因果关系。具体借鉴**张涛和李均超（2023）** 思路，将双重机器学习方法**样本分割比例设定为1:4**，为了避免模型设定上的主观选择性，采用**一般性交互式**双重机器学习模型检验，使用**随机森林算法**预测方法，在控制城市与年份固定效应的基础上，逐步纳入控制变量集合的一次项、二次项与三次项。附录表7第(1)~(3)列结果显示，公共数据开放的回归系数均在1%水平上显著为正。此外，将双重机器学习方法**样本分割比例设定为1:7**，附录表7第(4)~(6)列结果显示，公共数据开放的回归系数均在1%水平上显著为正。上述结果表明**采用双重机器学习方法缓解可能存在的内生性问题后，研究结论仍然稳健**。

方法延申

蒋金荷,黄珊. 贸易新业态对绿色技术创新的影响研究——来自跨境电商综合试验区政策的证据[J].数量经济技术经济研究,2024,41(12):133-154.

- **关键词**：贸易新业态；绿色技术创新；跨境电商综试区；双重机器学习；广义随机森林模型
- **摘要**：加快发展贸易新业态下，跨境电商的迅速发展能否有效提升中国绿色技术创新水平，对于抢抓外贸发展新机遇、助推产业绿色转型、服务构建新发展格局具有重要意义。本文以跨境电商综合试验区政策为准自然实验，采用双重机器学习模型，检验了贸易新业态对城市绿色技术创新的影响效应及内在机制。研究发现，跨境电商综合试验区政策的实施能够显著提高城市绿色创新水平。跨境电商综合试验区政策存在显著的绿色技术创新偏向性，且地处沿海、绿色金融发展程度高、地区政府环境注意力水平高的城市，政策的影响效果越明显。就内在机制而言，跨境电商综合试验区政策通过提高本土企业绿色管理能力、激发绿色产业创业活跃度、增强知识产权保护力度，可以提升城市绿色技术创新水平。基于上述研究结论，应完善跨境电商综合试验区创新制度顶层设计，畅通跨境电商综试区政策绿色创新效应发挥渠道，加快推进技术统一大市场建设。