Double Machine Learning

2023 年 2 月 3 日

目录

1 简介

1.1 Why：为什么我们需要 Double Machine Learning

实证研究往往会面临一个质疑：模型设定是正确的吗？例如，如果希望研究班级人数对教学质量的影响，常见的方法是构造回归方程：

$$Score = \beta_0 + \beta_1 Number + \sum_j \beta_j \times Control + \varepsilon$$

其中，$Score$ 代表成绩；$Number$ 代表班级人数；$Control$ 为控制变量，可能包括：每天学习时间、作业完成率、出勤率等。那么这些特征的关系真的是线性的吗？显然不是。例如随着学习时间增加，成绩自然会提高，然而学习时间过长很可能由于疲惫、睡眠不足等因素造成学习效率下降，反而使得成绩下降。需要注意的是，我们实际上并不关心学习时间对成绩的影响，我们只希望研究 β_1，但同时，我们需要处理控制变量对 β_1 造成的影响。

接下来用更严谨的方法描述上述问题。考虑因果模型：

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \quad \mathrm{E}[U \mid X, D] = 0 \\ D &= m_0(X) + V, \qquad \mathrm{E}[V \mid X] = 0 \end{aligned}$$

其中 Y 是模型的 Outcome，D 是因果模型的 treatment。这里，我们关注 θ，即 treatment 的因果效应。一种常见的思路是，通过假设（例如常见的线性假设），或者利用一定方法（通常是机器学习）估计，得到 $\hat{g}_0$，随后就可以利用线性回归得到 $\hat{\theta}_0$：

$$\hat{\theta}_0 = \frac{\operatorname{cov}\left(D, Y - \hat{g}_0(X)\right)}{\operatorname{var}(D)} = \frac{\frac{1}{n}\sum_{i\in I} D_i \left(Y_i - \hat{g}_0\left(X_i\right)\right)}{\frac{1}{n}\sum_{i\in I} D_i^2} \tag{1}$$

接下来，很自然的想要研究这个估计量是否无偏。遗憾的是 $\hat{\theta}_0$ 往往是有偏的：

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_0 - \theta_0\right) &= \sqrt{n}\frac{\frac{1}{n}\sum_{i\in I} D_i \left(Y_i - \hat{g}_0\left(X_i\right)\right)}{\frac{1}{n}\sum_{i\in I} D_i^2} \\ &\quad - \left(\sqrt{n}\frac{\frac{1}{n}\sum_{i\in I} D_i \left(Y_i - g_0\left(X_i\right)\right)}{\frac{1}{n}\sum_{i\in I} D_i^2} - \sqrt{n}\frac{\frac{1}{n}\sum_{i\in I} D_i U_i}{\frac{1}{n}\sum_{i\in I} D_i^2}\right) \\ &= \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i \left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right)}_{:=b} \end{aligned}$$

可以看出误差分为两项。a 项来自于 U 和 D 的独立性，即 $\frac{\operatorname{cov}(D,U)}{\operatorname{var}(D)}$，若二者不独立则会造成偏误。

然而问题来源于 b 项，我们将其展开为以下形式：

$$b = \left(E\left[D_i^2\right]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} m_0\left(X_i\right)\left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right) + o_P(1)$$

注意到 $m_0(X_i)(g_0(X_i)-\hat{g}_0(X_i))$ 项。首先，g_0 的估计往往存在误差，例如对于高维数据，往往会采用正则项处理，造成正则化误差，此时 b 项发散；此外，$m_0(X_i)$ 是数据本身的性质，因此数据会决定偏误的大小而无法改变，导致估计非常不稳健。

综合以上推论，可以说因果模型 treatment effect 的传统估计方法并不完美。因此，我们引入 *Double Machine Learning* 的概念，为因果估计提供更为稳健的方法。

1.2 What：Double Machine Learning 的本质与关键

1.2.1 Neyman Orthogonality

为了尽可能消掉 $m_0(X_i)(g_0(X_i)-\hat{g}_0(X_i))$ 项，一个实际的考虑是消除 $m_0(X_i)$。其出现原因在于用于回归的 D 实际上包含了 X 的信息如图，注意到 V 实际上可以看作工具变量，此时可以构造估计：

$$\theta_0=\frac{\operatorname{cov}(V,Y-g_0(X))}{\operatorname{cov}(V,D)}=\frac{\operatorname{cov}(D-m_0(X),Y-g_0(X))}{\operatorname{cov}(D-m_0(X),D)}=\frac{\operatorname{cov}(D-m_0(X),D\theta_0+U)}{\operatorname{cov}(D-m_0(X),D)}$$

为了求 V，可以采用：

$$\hat{V}=D-\hat{m}_0(X)$$

其中 $\hat{m}_0(X)$ 可以通过 X 对 D 回归得到，因此我们得到一种新的估计：

$$\breve{\theta}_0=\frac{\frac{1}{n}\sum_{i\in I}\hat{V}_i\left(Y_i-\hat{g}_0(X_i)\right)}{\frac{1}{n}\sum_{i\in I}\widehat{V}_iD_i} \tag{2}$$

在这个估计下，新的 b 项变为：

$$b^*=\left(E\left[D_i^2\right]\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}\left(\hat{m}_0(X_i)-m_0(X_i)\right)\left(g_0(X_i)-\hat{g}_0(X_i)\right)$$

此时偏误仅仅取决于回归误差，因此这个估计更为稳健。

此时我们已经在因果模型中构造了稳健的估计量，接下来，我们希望从中总结出一套普适性更强的方法论，以用于更多场景。

我们对上述过程进行总结：

- 存在目标函数 θ_0 和其他不太关心的回归函数 g_0 和 m_0；
- θ_0 的传统估计方法中，如果拟合的函数 $\hat{g}_0$ 存在误差，则误差项不收敛；
- 需要构造估计量使得 g_0 和 m_0 出现误差时，偏误依然足够小。

这里我们用 *Gateaux Dervative* 描述 g_0 估计误差对偏误的影响：

$$\partial_f E\left[\varphi\left(W;\theta_0,f_0\right)\right]=\lim_{r\to 0^+}\frac{E\left[\varphi\left(W;\theta_0,f_0+r\left(f-f_0\right)\right)\right]-E\left[\varphi\left(W;\theta_0,f_0\right)\right]}{r}$$

我们将（1）式的传统估计方法重新写成：

$$\frac{1}{n}\sum_{i\in I}\varphi\left(W;\hat{\theta}_0,\hat{g}_0\right)=0$$

其问题在于 Gateaux Dervative 不为零：

$$\partial_g E\left[\varphi\left(W;\theta_0,g_0\right)\right]\left[g-g_0\right]\neq 0$$

换言之，g_0 的微小扰动会导致 ϕ 发生较大的变化，因此估计不准时存在较大误差。

我们将（2）式的估计方法重新写成：

$$\frac{1}{n}\sum_{i\in I}\psi\left(W;\check{\theta}_0,\hat{\eta}_0\right)=0$$

其中 $\hat{\eta}_0=(\hat{m}_0,\hat{g}_0)$ 称为 *Nuisance Parameter*。可以证明这个估计量的 Gateaux Dervative 为零：

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

直观来说，Gateaux Dervative 事实上刻画了 ϕ 对 g_0 变化的敏感程度。如果 Gateaux Dervative 等于零，意味着这个估计量更为稳健。我们称这个性质为 *Neyman Orthogonality*。这种构造就是 *Double Machine Learning* 的思想基础

1.2.2 样本分割

当使用高度复杂的拟合方法如：boosting,random forests,ensemble 等机器学习方法的时候，会由于过度拟合而产生偏差，我们使用 empirical moments 的交叉拟合形式，去估计 θ_0，这样可以完全消除由过拟合引起的偏差。

1.3 How：如何构造 Double Machine Learning

接下来，我们希望找到通用的方法构造 ψ 满足 Neyman Orthogonality 不妨设 $\max_{\theta,\beta}E_p[l(W;\theta,\beta)]$，我们的目标可以通过最大化某个函数得到，通常为对数似然，故在极值处导数为 0：

$$E_p\left[\partial_\theta l\left(W;\theta_0,\beta_0\right)\right]=0, E_p\left[\partial_\beta l\left(W;\theta_0,\beta_0\right)\right]=0$$

因此，一个自然的想法是令 $\phi(W;\theta,\beta)=\partial_\theta l(W;\theta,\beta)$。然而，为了满足 Neyman Orthogonality，我们还需要自己构造估计量：

$$\psi(W;\theta,\beta)=\partial_\theta l(W;\theta,\beta)-\mu\partial_\beta l(W;\theta,\beta)$$

此时只需要找到合适的 μ 使 ψ 满足 Neyman Orthogonality：

$$\partial_\beta\psi(W;\theta,\beta)=\partial_{\theta\beta} l(W;\theta,\beta)-\mu\partial_{\beta\beta} l(W;\theta,\beta)=0$$

可以得到解析解：$\mu=J_{\theta\beta}J_{\beta\beta}^{-1}$

2 DML 在 PLR 模型中的应用

2.1 Partially linear regression models(PLR) 简介

部分线性回归模型 (PLR) 包括标准线性回归模型，在数据分析中发挥着重要作用，其形式如同简介中介绍：

$$\begin{aligned} Y&=D\theta_0+g_0(X)+\zeta, \quad \mathbb{E}(\zeta\mid D,X)=0\\ D&=m_0(X)+V, \qquad\quad \mathbb{E}(V\mid X)=0 \end{aligned} \tag{3}$$

2.2 在 PLR 模型中应用 DML 的步骤

PLR 模型可以被写成

$$\begin{aligned} Y-D\theta_0-g_0(X)&=\zeta, \quad \mathbb{E}(\zeta\mid D,X)=0\\ D-m_0(X)&=V, \qquad\quad \mathbb{E}(V\mid X)=0 \end{aligned} \tag{4}$$

如何去利用 DML 去对 θ_0 进行估计的思想可分为三步：

(1) 利用不同类型的机器学习方法如:boosting，random forest，ensemble 去估计 g(x)，m(x) 以及 θ，得到他们的估计值 $\hat{g}(x)$，$\hat{m}(x)$ 和 $\hat{\theta}$ 。其中 $\hat{g}(x)$ 和 $\hat{m}(x)$ 即为侵扰函数的估计 $\hat{\eta}$

(2) 为了消除用 $\hat{g}(x)$ 和 $\hat{m}(x)$ 去代替 $\hat{g}(x)$,$\hat{m}(x)$ 而产生的误差，我们构造关于 PLR 的 Neyman-orthogonal score：

$$\psi(W;\theta,\eta):=(Y-D\theta-g(X))(D-m(X)), \tag{5}$$

利用 Neyman-orthogonal score 的特点即当 $\hat{g}_0$ 作为 g_0 的估计是有偏的，θ 的近似分布将不取决于由侵扰函数产生的误差，即也可以得到 θ_0 的无偏估计。只需在满足：

$$\frac{1}{n}\sum(Y-D\check{\theta}_0-\hat{g}(X))(D-\hat{m}(X))=0 \tag{6}$$

的条件下同样利用机器学习的方法求解出 $\check{\theta}_0$ 即可。

(3) 由于大部分机器学习都是高度复杂的拟合方法，所以在进行估计的时候会因为过拟合产生偏差，所以我们利用交叉拟合 (cross fitting) 的形式来对其 θ 进行估计，即先将数据集分为 N 份，每次都用其中的 N-1 份来估计侵扰函数 $\hat{g}(x)$ 和 $\hat{m}(x)$ 得到 $\hat{g}(x)$ 和 $\hat{m}(x)$ 再用剩下的一份与前面所得的侵扰函数的估计做 Neyman orthogonal，得到 $\check{\theta}$，最后将 N 次的结果取平均即得到最终关于 θ 的估计 $\tilde{\theta}$。

利用 DML 对 PLR 模型进行因果推断估计的算法步骤可以表示为：

(1) 提供数据集 $(W_i)_{i=1}^N$ 和 Neyman-orthogonal score 函数：$\psi(W;\theta,\eta) := (Y - D\theta - g(X))(D - m(X)) \quad \eta = (g,m), \quad \eta_0 = (g_0,m_0)$ ，并指定对 η 进行估计的机器学习方法

(2) 利用机器学习对每组进行训练：将样本随机分为 K 组，$(I_k)_{k=1}^K$ 每组样本量的大小为 $I_kisn = N/K$. 对每一组 $k \in [K] = \{1,\ldots,K\}$, 都构造一个高质量的机器学习估计器

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i\notin I_k}\right) \tag{7}$$

其中 $x \mapsto \hat{\eta}_{0,k}(x)$ 的映射只依赖 $(W_i)_{i\notin I_k}$

(3) 对于每个 $k \in [K]$, 通过求解方程：

$$\frac{1}{n}\sum_{i\in I_k}\psi\left(W_i;\check{\theta}_{0,k},\hat{\eta}_{0,k}\right) = 0 \tag{8}$$

得到点估计 $\check{\theta}_{0,k}$

(4) 通过聚合得到因果效应的估计值

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^K \check{\theta}_{0,k} \tag{9}$$

$\tilde{\theta}_0$ 即为对因果效应的估计

2.3 实例：衡量税收和公共服务对房价的影响

2.3.1 背景与设定

对于一般人来说，住房的选择和购置往往是个人一生中最重要的财务决策之一，除了可以直接居住外，房子的价值还与在社区或地区一级提供的一系列地方公共产品相关联。所以住房的价格不仅取决于其物理属性，还取决于其周围环境的特点及其提供当地服务的质量，在一个地区或社区内为支付此类服务而征收的税款也会影响当地的房地产价格，其

影响程度一直是关于税收资本化的大量文献的主题。然而由于遗漏了重要的地方控制措施，税收资本化估算存在严重偏差，而且也没有充分考虑地方税收的内生性。

Semenova 等人（2020）认为，有两种主要方法可用于解决内生回归问题：

(1) 解释内生性来源

(2) 找到工具变量

我们遵循第一种方法，将 Double Machine Learning 用到瑞典 2010-2016 年 947 个时变市政级控制的独特且详尽的数据中。

数据涵盖 2010-2016 年期间，包括 947 个潜在协变量, 其中不仅有标准控制措施，如当地经济、住房和劳动力市场条件，还有一些不太常用的控制措施，如地方公共服务、人口统计、学校教育、政治、基础设施、移民和公共财政。控制地方政治结果和地方公共财政状况使我们能够可信地处理税收内生性问题。

由于数据集十分庞大，所以我们能够消除各种混杂因素的影响，包括税收变化的政治和经济驱动因素，从而产生一种可以被视为随机分配的 treatment variable。因此，我们的结果在无根据假设下具有因果解释。

2.3.2 计算方法与仿真

首先，我们仿照模型设定进行数据生成

我们的模型设定如下：

$$p_{jt} = \tau_{jt}\theta_0 + g_0\left(\zeta_g\left(x_{jt}\right) + \gamma_j + \eta_t\right) + u_{jt}$$

$$\tau_{jt} = m_0\left(\zeta_m\left(x_{jt}\right) + \delta_j + \xi_t\right) + v_{jt}$$

其中 g_0 和 m_0 的选择是任意的，这里选择线性函数。另外，这里 Treatment Effect 设定为 -5。同时，为了引入非线性，对 treatment 采用以下设定：

$$\zeta_g\left(x_{jt}\right) = \frac{1}{1 + exp^{-\left(g_0^0 x_{jt}^0 + \ldots + g_0^{k-1} x_{jt}^{k-1}\right)}}$$

$$\zeta_m\left(x_{jt}\right) = \frac{1}{1 + exp^{-\left(m_0^0 x_{jt}^0 + \ldots + m_0^{k-1} x_{jt}^{k-1}\right)}}$$

我们将使用 DML 在存在高维和潜在非线性干扰参数的情况下估计税收资本化，即将税收看成 treatment variable, 房价看成因变量，对 treatment effect 进行估计。

更正式地说，我们将税收资本化估计问题视为部分线性模型，即 PLR 模型，从而运用 Double Machine Learning 的方法对 treatment effect 进行估计

$$p_{jt} = \tau_{jt}\theta_0 + g_0\left(x_{jt}\right) + u_{jt}, \quad E\left[u_{jt} \mid x_{jt}, \tau_{jt}\right] = 0$$

$$\tau_{jt} = m_0\left(x_{jt}\right) + v_{jt}, \qquad E\left[v_{jt} \mid x_{jt}\right] = 0$$

p_{jt} 是每平方米的房价，τ_{jt} 是市政税税率水平, x_{jt} 是 K 维的混杂向量，$\left(x_{jt}^1, \ldots, x_{jt}^k\right)$, u_{jt} 和 v_{jt} 是干扰项。$g_0\left(x_{jt}\right)$ 表示允许混杂因子的高维向量对 p_{jt} 产生直接和潜在的非线性影响，第二个式子表示混杂变量中的一些因素会对市政水平 τ_{jt} 产生影响，例如，我们可以预期，x_{jt} 中城市和时间的人口变化将影响 τ_{jt} 和 p_{jt}。

我们将数据集分成两个相等的部分：I 和 I^c。我们使用 I^c 来估计 $\hat{m}_0$ 和 $\hat{g}_0$。然后我们可以估计分割后感兴趣的参数：

$$\hat{\theta}_0\left(I^c, I\right)=\left(\frac{1}{\bar{n}\bar{t}} \Sigma_{jt \in I} \hat{v}_{it} \tau_{jt}\right)^{-1} \frac{1}{\bar{n}\bar{t}} \sum_{jt \in J} \hat{v}_{jt}\left(p_{jt}-\hat{g}_0\left(x_{jt}\right)\right)$$

$\bar{n}$ 和 $\bar{t}$ 是 I 中的城市数量和时间段, $\hat{\theta}_0\left(I^c, I\right)$ 使用了辅助样本 I^c 来估计 $\hat{m}_0$ 和 $\hat{g}_0$, p_{jt} 和 x_{jt} 来自主样本 I。最后再交换辅助样本与主样本，从而估计出 $\hat{\theta}_0\left(I, I^c\right)$，再计算均值，得到 θ_0 的估计值：

$$\hat{\theta}_0=\frac{1}{2}\left[\hat{\theta}_0\left(I^c, I\right)+\hat{\theta}_0\left(I, I^c\right)\right]$$

估计结果为 -0.4926

3 DML 在 IRM 模型中的应用

3.1 IRM 模型的简介

当结果完全异质且 treatment variable 为二元 0，1 变量时，我们考虑模型：

$$\begin{aligned} Y &= g_0(D, X)+U, \quad \mathbb{E}(U \mid X, D)=0 \\ D &= m_0(X)+V, \quad \mathbb{E}(V \mid X)=0 \end{aligned} \tag{10}$$

由于 D 与 X 是不可线性分离开的，所以在 D 为二元变量的情况下，该模型比部分线性模型的应用广泛性更好。我们在模型中感兴趣的目标参数是平均治疗效果 (ATE)：

$$\theta_0=\mathbb{E}\left[g_0(1, X)-g_0(0, X)\right] \tag{11}$$

另一个常见的目标参数是被治疗者的平均治疗效果 (ATTE)

$$\theta_0=\mathbb{E}\left[g_0(1, X)-g_0(0, X) \mid D=1\right] \tag{12}$$

3.2 在 IRM 模型中应用 DML 的步骤

我们讨论如何用 DML 去估计 IRM 模型中的 ATE：

(1) 由于 D 与 X 是没有办法线性分开的，所以 $\theta_0=\mathbb{E}\left[g_0(1, X)-g_0(0, X)\right]$ 看作是对因果参数求解的过程，即在该模型中应该把 $g_0(0, X)$, $g_0(1, X)$ 以及 $m_0(X)$ 都看作是侵扰

函数 η_0. 同样利用不同类型的机器学习的方法估计出 $\hat{g}_0(0,X)$, $\hat{g}_0(1,X)$ 以及 $\hat{m}_0(X)$ 作为 $\hat{\eta}_0$

(2) 构造关于 IRM 模型中求解 ATE 的 Newman Orthogonal score 函数：

$$\begin{aligned}&\psi(W;\theta,\eta) := (g(1,X)-g(0,X)) + \tfrac{D(Y-g(1,X))}{m(X)} - \tfrac{(1-D)(Y-g(0,X))}{1-m(X)} - \theta,\\ &\eta(X) = (g(0,X), g(1,X), m(X)), \quad \eta_0(X) = (g_0(0,X), g_0(1,X), m_0(X)),\end{aligned} \tag{13}$$

其中 $\eta(X)$ 为侵扰函数，其真值表示为 $\eta_0(X)$

在满足：

$$\frac{1}{n}\sum(\hat{g}(1,X)-\hat{g}(0,X)) + \frac{D(Y-\hat{g}(1,X))}{\hat{m}(X)} - \frac{(1-D)(Y-\hat{g}(0,X))}{1-\hat{m}(X)} - \theta = 0 \tag{14}$$

的条件下利用机器学习的方法求出 $\check{\theta}_0$ 即可。

(3) 同时也因为要考虑过拟合的问题，所以需要通过交叉拟合的方式对 θ_0 进行估计，从而增加估计的稳健性。

利用 DML 对 IRM 模型的 ATE 进行因果推断估计的算法步骤可以表示为：

(1) 提供数据集 $(W_i)_{i=1}^N$ 和 Neyman-orthogonal score 函数：$\psi(W;\theta,\eta) := (g(1,X) - g(0,X)) + \frac{D(Y-g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)} - \theta$ ，并指定对 η 进行估计的机器学习方法

(2) 利用机器学习对每组进行训练：将样本随机分为 K 组，$(I_k)_{k=1}^K$ 每组样本量的大小为 $I_kisn = N/K$. 对每一组 $k \in [K] = \{1,\ldots,K\}$, 都构造一个高质量的机器学习估计器

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i\notin I_k}\right) \tag{15}$$

其中 $x \mapsto \hat{\eta}_{0,k}(x)$ 的映射只依赖 $(W_i)_{i\notin I_k}$

(3) 对于每个 $k \in [K]$, 通过求解方程：

$$\frac{1}{n}\sum_{i\in I_k}\psi\left(W_i;\check{\theta}_{0,k},\hat{\eta}_{0,k}\right) = 0 \tag{16}$$

得到点估计 $\check{\theta}_{0,k}$

(4) 通过聚合得到因果效应的估计值

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^K \check{\theta}_{0,k} \tag{17}$$

$\tilde{\theta}_0$ 即为对因果效应的估计

3.3 实例：401（k）计划对净金融资产的影响

3.3.1 背景

401k 是美国的一项养老金计划，即若公司有这样一个计划，且员工有资格加入这个计划，则公司会给员工设立一个养老金账户，员工可以把一部分的年收入放入该账户中，当把一部分年收入转移到养老金账户中后对于收入所应纳的税额就减少，而当退休后，必须要固定从 401(k) 账户中取出前，这时取出的钱是要计入当年应该缴纳的所得税的收入，但是由于退休后收入是减少的，所以很有可能为了得到这笔钱所应该缴纳的税人让路 i 在一个比较低的税率档，所以缴纳的总额是减少的。

要探究 401(k) 计划是否对净金融资产的影响，首先要了解到能否参加 401(k) 计划是受到个人的自身条件的影响的，而净金融资产同时也受到这些自身条件如：年龄、收入、家庭规模等的影响。

混杂效应可以通过研究人员事先选择的少量变量进行充分控制，即其中一个人是否有资格获得 401(k) 养老金，在适当调整收入和其他与工作选择相关的控制变量后，可以被视为外生变量。所以在这样的条件下我们可以通过因果模型探究 401（k）计划对净金融资产的影响

3.3.2 计算方法与仿真

首先，我们仿照模型设定进行数据生成

我们的模型设定如下：

$$\begin{aligned} y_i &= g_0\left(d_i, x_i\right) + \xi_i \\ d_i &= m_0\left(x_i\right) + v_i \end{aligned} \tag{18}$$

其中 g_0 和 m_0 的选择是任意的，这里 g_0 选择具有交互项的线性函数，m_0 选择生成二元变量的非线性函数。即：

$$\begin{aligned} d_i &= 1\left\{\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} > v_i\right\} \\ y_i &= \theta d_i + c_y x_i \beta d_i + \zeta_i \end{aligned} \tag{19}$$

其中 $x_i = \left(x_i^1, \ldots, x_i^k\right)$ 为 K 维的混杂向量，$\beta = \left(\beta^1, \ldots, \beta^k\right)^{\mathrm{T}}$ 为系数向量的转置。另外，这里 Treatment Effect 设定为 5。

我们将使用 DML 在存在高维和潜在非线性干扰参数的情况下估计 401(k) 对净金融资产的影响，即将有无参与 401(k) 计划看成 treatment variable(这是个二元变量), 净金融资产看成因变量，对 treatment effect 进行估计。

更正式地说，我们将探究 401(k) 计划对净金融资产影响的问题视为交互式回归模型，

即 IRM 模型，从而运用 Double Machine Learning 的方法对 treatment effect 进行估计

$$\begin{aligned} y_i &= g_0\left(d_i, x_i\right) + \xi_i, \qquad E\left[\xi_i \mid x_i, d_i\right] = 0 \\ d_i &= m_0\left(x_i\right) + v_i, \qquad E\left[v_i \mid x_i\right] = 0 \\ \theta_0 &= \mathbb{E}\left[g_0(1, x_i) - g_0(0, x_i)\right] \end{aligned}$$

这里将净金融资产作为结果变量 y_i, 将是否参加 401(k) 计划看作 treatment effect d_i, 将年龄、收入、家庭规模、教育年限、已婚指标、双职工状态指标、固定福利养老金状态指标、个人退休账户参与指标和房屋所有权指标看作混杂的原始协变量 x_i。

4 DML 在其他模型中的应用

4.1 部分线性工具变量回归模型 (PLIV)

$$\begin{aligned} Y - D\theta_0 &= g_0(X) + \zeta, \quad \mathbb{E}(\zeta \mid Z, X) = 0 \\ Z &= m_0(X) + V, \qquad \mathbb{E}(V \mid X) = 0 \end{aligned} \tag{20}$$

在这个模型中，Y 和 D 之间存在结构或因果关系，其关系由 θ 来表示，$g_0(X) + \zeta$ 可以看作是随机误差，其中 $g_0(X)$ 可以被协变量 X 所解释，由于 Y 和 D 是共同确定的，所以我们引入一个外部因素，称为工具 Z，以解释 D 的外生变化。注意，Z 应该影响 D。这里的 X 再次作为 confounding factors，因此我们可以认为 Z 中的变化是外生的，仅以 X 为条件。

一个来自生物统计学的例子：Y 为健康指标，D 为吸烟的指标，则 θ_0 代表着抽烟对健康的影响，健康指标 Y 和吸烟行为 D 是同时被定下来的。X 代表患者特征，Z 是医生建议不要吸烟 (或其他行为治疗)，这一行为由 X 决定，Z 只能通过改变行为 D 去影响结果 Y。

在该模型中，也可构造出 Neyman-orthogonal score 函数去消除用 $\hat{g}_0(x)$ 和 $\hat{m}_0(x)$ 去代替 $g_0(x)$ 和 $m_0(x)$ 进行估计而产生的误差，其 Neyman-orthogonal score 函数为：

$$\begin{aligned} &\psi(W; \theta, \eta) := (Y - \ell(x) - \theta(D - r(X)))(Z - m(X)) \\ &\eta = (\ell, m, r), \quad \eta_0 = (\ell_0, m_0, r_0) \end{aligned} \tag{21}$$

4.2 交互式工具变量模型 (IIVM)

我们考虑 D 是二元 0，1 变量，$Z \in \{0, 1\}$, 利用 D 和 Z 对局部平均治疗效果 (LATE) 进行估计，模型为：

$$\begin{aligned} Y &= \ell_0(D, X) + \zeta, \quad \mathbb{E}(\zeta \mid Z, X) = 0 \\ Z &= m_0(X) + V, \quad \mathbb{E}(V \mid X) = 0 \end{aligned} \tag{22}$$

其中 g_0 将 (Z, X) 映射到 $\mathbb{R}$ 上，r_0 和 m_0 分别将 (Z, X) 和 X 映射到 $(\epsilon, 1-\epsilon)$, $\epsilon \in (0, 1/2)$。

$$\begin{aligned} Y &= g_0(Z, X) + \nu, \quad \mathbb{E}(\nu \mid Z, X) = 0 \\ D &= r_0(Z, X) + U, \quad \mathbb{E}(U \mid Z, X) = 0 \\ Z &= m_0(X) + V, \quad \mathbb{E}(V \mid X) = 0 \end{aligned} \tag{23}$$

其中我们感兴趣的估计为:

$$\theta_0 = \frac{\mathbb{E}\left[g_0(1, X)\right] - \mathbb{E}\left[g_0(0, X)\right]}{\mathbb{E}\left[r_0(1, X)\right] - \mathbb{E}\left[r_0(0, X)\right]} \tag{24}$$

θ_0 是编译器的平均处理效果

与上面三种模型相同，只需要找到满足 IIRM 模型的 Neyman-orthogonal score 函数，也可消除估计侵扰函数的误差，从而对 treatment effect 作出准确的估计，其 Neyman-orthogonal score 函数为:

$$\begin{aligned} \psi :=& g(1, X) - g(0, X) + \frac{Z(Y - g(1, X))}{m(X)} - \frac{(1 - Z)(Y - g(0, X))}{1 - m(X)} \\ & - \left(r(1, x) - r(0, X) + \frac{Z(D - r(1, x)}{m(X)} - \frac{(1 - Z)(D - r(0, X)}{1 - m(X)} \right) \times \theta \\ \eta &= (g, m, r), \quad \eta_0 = (g_0, m_0, r_0) \end{aligned} \tag{25}$$

4.3 Double Machine Learning 用于估计 treatment effect 的通用算法

Algorithm 1 DML1. (Generic double machine learning with cross-fitting)

Require:

Choose a model (PLR, PLIV, IRM, IIVM)

provide data $(W_i)_{i=1}^N$

a Neyman-orthogonal score function $\psi(W;\theta,\eta)$

specify machine learning methods for η

Ensure:

The estimate of the causal parameter $\tilde{\theta}_0$;

1: **Train ML predictors on folds** : Take a K -fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1,\ldots,N\}$ such that the size of each fold I_k is n=N / K . For each $k \in [K] = \{1,\ldots,K\}$, construct a high-quality machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i\notin I_k}\right) \tag{26}$$

2: For each $k \in [K]$, construct the estimator $\check{\theta}_{0,k}$ as the solution to the equation

$$\frac{1}{n}\sum_{i\in I_k}\psi\left(W_i;\check{\theta}_{0,k},\hat{\eta}_{0,k}\right) = 0 \tag{27}$$

3: The estimate of the causal parameter is obtained via aggregation

$$\tilde{\theta}_0 = \frac{1}{K}\sum_{k=1}^{K}\check{\theta}_{0,k} \tag{28}$$

4: **return** $\tilde{\theta}_0$ and the values of the evaluated score function

Algorithm 2 DML2. (Generic double machine learning with cross-fitting)

Require:

Choose a model (PLR, PLIV, IRM, IIVM)

provide data $(W_i)_{i=1}^N$

a Neyman-orthogonal score function $\psi(W;\theta,\eta)$

specify machine learning methods for η

Ensure:

The estimate of the causal parameter $\tilde{\theta}_0$;

1: **Train ML predictors on folds** : Take a K -fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1,\ldots,N\}$ such that the size of each fold I_k is n=N / K . For each $k \in [K] = \{1,\ldots,K\}$, construct a high-quality machine learning estimator

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i\notin I_k}\right) \tag{29}$$

2: Construct the estimator for the causal parameter $\tilde{\theta}_0$ as the solution to the equation

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi\left(W_i;\tilde{\theta}_0,\hat{\eta}_{0,k}\right) = 0 \tag{30}$$

3: **return** $\tilde{\theta}_0$ and the values of the evaluated score function

5 DML 在分配政策中的应用

5.1 分配政策学习的必要性

分配政策的学习或从个人特征到分配的映射问题是公司、企业甚至政府都会遇到的问题。而需要进行政策学习的原因是因为资源不是无限的，基于政策的分配是要在尊重一定约束的限制下（如资源预算是有限的）进行资源的分配。

例如：在医学领域中，当疫苗供不应求时 (限制)，医生或者政府必须决定给具有怎样特征的人提供疫苗才能使疫苗的效益最大化、在市场营销中，公司在有限的资源或者时间中向具有哪些特征的客户发送目标产品的信息才能使公司收益最大化。

5.2 分配政策学习的理论

从形式上看，我们拥有一些基于观察得到的数据，每个人的数据可看作 $X_i \in \mathcal{X}$ i 为第 i 为被观察者，我们希望通过这些数据来学习一个从个体特征 $X_i \in \mathcal{X}$ 到一个 01 决定 $\pi : \mathcal{X} \to \{0, 1\}$ 的政策，同时这个政策要满足一定的越苏条件，即在类 Π 里面

我们进行分配政策学习的方法是先利用 Double Machine Learning 的方法先对平均治疗效果 θ 进行估计：

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\Gamma}_i, \tag{31}$$

其中 $\widehat{\Gamma}_i$ 是经过 Double Machine Learning 估计出来的在有混杂变量干预下的关于目标估计的双重稳健得分。

同时只要 $\widehat{\Gamma}_i$ 是具有双重稳健性的，则可以预测治疗每个人的平均效用，同时也可以通过简单的程序来筛选出需要进行治疗或者干预的对象。即在给定约束类 Π 的情况下，利用

$$\hat{\pi} = \operatorname{argmax} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(2\pi\left(X_i\right) - 1\right) \widehat{\Gamma}_i : \pi \in \Pi \right\}, \tag{32}$$

去构造分配的规则。

与 Double Machine Learning 的符号定义相似，我们定义：独立不相关样本 (X_i, Y_i, W_i) , 其中 $Y_i \in \mathbb{R}$ 是我们想要干预的结果, W_i 是已经观察到的治疗分配的方案，X_i 是个体的一系列特征。

同时我们根据潜在结果模型定义干预的因果效应 $\pi(\cdot)$, 其中 $\{Y_i(w)\}$ 对应于如果处理设置为 $W_i = w$, 和 $Y_i = Y_i\left(W_i\right)$ 在第 i 个样本中观察到的效用。在 treatment effect 是二元变量的时候，我们将实行了分配政策在 treatment 为 0 的个体的效用定义为：

$$V(\pi) = \mathbb{E}\left[Y_i\left(\pi\left(X_i\right)\right) - Y_i(0)\right] \tag{33}$$

在 treatment effect 是连续变量的时候，我们将这种无限小干预的效用定义为：

$$V(\pi) = \left[\frac{d}{d\nu} \mathbb{E}\left[Y_i\left(W_i + \nu\pi\left(X_i\right)\right)\right]\right]_{\nu=0}, \tag{34}$$

在两种不同的 treatment effect 的情形下，都将在类 Π 的约束下最佳分配政策与当前分配政策的差值定义为：

$$R(\pi) = \max\left\{V\left(\pi'\right) : \pi' \in \Pi\right\} - V(\pi). \tag{35}$$

无论是二元的 treatment 还是连续的 treatment，政策 π 的 $R(\pi)$ 都可以用条件平均处理效应函数表示:

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right] \text{ or } \tau(x) = \left[\frac{d}{d\nu} \mathbb{E}\left[Y_i\left(W_i + \nu\right) \mid X_i = x\right]\right]_{\nu=0}, \tag{36}$$

使得 $V(\pi) = \mathbb{E}\left[\pi\left(X_i\right)\tau\left(X_i\right)\right]$, $R(\pi)$ 与 (27) 所得式子相同。