Federate Report

19331053 纪传宇

2023 年 8 月 3 日

目录

1 DID

$$\begin{aligned}\mathbb{E}\left[Y_1(1)-Y_1(0)\mid T=1\right] &= \mathbb{E}\left[Y_1(1)\mid T=1\right]-\mathbb{E}\left[Y_1(0)\mid T=1\right]\\ &= \mathbb{E}\left[Y_1\mid T=1\right]-\mathbb{E}\left[Y_1(0)\mid T=1\right]\end{aligned}$$

Common trends: $\mathbb{E}\left[Y_1(0)\mid T=1\right]-\mathbb{E}\left[Y_0(0)\mid T=1\right]=\mathbb{E}\left[Y_1(0)\mid T=0\right]-\mathbb{E}\left[Y_0(0)\mid T=0\right]$

$$\begin{aligned}\mathbb{E}\left[Y_1(0)\mid T=1\right] &= \mathbb{E}\left[Y_0(0)\mid T=1\right]+\mathbb{E}\left[Y_1(0)\mid T=0\right]-\mathbb{E}\left[Y_0(0)\mid T=0\right]\\ &= \mathbb{E}\left[Y_0(0)\mid T=1\right]+\mathbb{E}\left[Y_1\mid T=0\right]-\mathbb{E}\left[Y_0\mid T=0\right] \quad \text{No pretreatment effect:}\\ &= \mathbb{E}\left[Y_0(1)\mid T=1\right]+\mathbb{E}\left[Y_1\mid T=0\right]-\mathbb{E}\left[Y_0\mid T=0\right] \quad \mathbb{E}\left[Y_0(1)\mid T=1\right]-\mathbb{E}\left[Y_0(0)\mid T=1\right]=0\\ &= \mathbb{E}\left[Y_0\mid T=1\right]+\mathbb{E}\left[Y_1\mid T=0\right]-\mathbb{E}\left[Y_0\mid T=0\right]\end{aligned}$$

$\mathbb{E}\left[Y_0(1)\mid T=1\right]-\mathbb{E}\left[Y_0(0)\mid T=1\right]=0$

$$\begin{aligned}\mathbb{E}\left[Y_1(1)-Y_1(0)\mid T=1\right] &= \mathbb{E}\left[Y_1\mid T=1\right]-\left(\mathbb{E}\left[Y_0\mid T=1\right]+\mathbb{E}\left[Y_1\mid T=0\right]-\mathbb{E}\left[Y_0\mid T=0\right]\right)\\ &= \left(\mathbb{E}\left[Y_1\mid T=1\right]-\mathbb{E}\left[Y_0\mid T=1\right]\right)-\left(\mathbb{E}\left[Y_1\mid T=0\right]-\mathbb{E}\left[Y_0\mid T=0\right]\right)\end{aligned}$$

2 DRDID

2.1 Basic assumptions of DD

Assumption 1: Assume panel data or repeated cross-sectional data

Assumption 2: Conditional parallel trends If you were putting covariates into your DD regression, then you were assuming conditional parallel trends

$$E\left[Y_1^0-Y_0^0\mid X,D=1\right]=E\left[Y_1^0-Y_0^0\mid X,D=0\right]$$

Assumption 3: Common support or overlap

For some e>0 , the probability of being in the treatment group is greater than e and the probability of being in the treatment group conditional on $Xis \leq 1-e$.

Intuition of assumption 3: Called”overlap or common support. Means there is at least a small fraction of the population that is treated and that for every value of the covariates X there is at least a small chance that the unit is not treated. It’s called common support when it’s a propensity score but it’s just about the distribution of treatment and control across values of X .

2.2 doubly robust

Outcome regression

This is the Heckman, et al. (1997) approach where the outcome evolution is modeled with a regression

$$\widehat{\delta}^{OR} = \bar{Y}_{1,1} - \left[\bar{Y}_{1,0} + \frac{1}{n^T} \sum_{i \mid D_i = 1} \left(\widehat{\mu}_{0,1}\left(X_i\right) - \widehat{\mu}_{0,0}\left(X_i\right) \right) \right]$$

where $\bar{Y}$ is the sample average of Y among units in the treatment group at time t and $\widehat{\mu}(X)$ is an estimator of the true, but unknown, $m_{d,t}(X)$ which is by definition equal to $E\left[Y_t \mid D = d, X = x\right]$.

Inverse probability weighting

This is the Abadie (2005) approach where we use weighting

$$\widehat{\delta}^{ipw} = \frac{1}{E_N[D]} E\left[\frac{D - \widehat{p}(X)}{1 - \hat{p}(X)} \left(Y_1 - Y_0\right) \right]$$

where $\hat{p}(X)$ is an estimator for the true propensity score. Reduces the dimensionality of X into a single scalar.

one can combine them to form doubly robust (DR) moments/estimands for the ATT. Here, double robustness means that the resulting estimand identifies the ATT even if either (but not both) the propensity score model or the outcome regression models are misspecified. As so, the DR DID estimand for the ATT shares the strengths of each individual DID method and, at the same time, avoids some of their weaknesses.

Before describing how we exactly combine the OR and the IPW approaches to form our DR DID estimand, we need to introduce some additional notation. Let $\pi(X)$ be an arbitrary model for the true, unknown propensity score. When panel data are available, let $\Delta Y = Y_1 - Y_0$ and define $\mu_{d,\Delta}^p(X) \equiv \mu_{d,1}^p(X) - \mu_{d,0}^p(X), \mu_{d,t}^p(x)$ being a model for the true, unknown outcome regression $m_{d,t}^p(x) \equiv \mathbb{E}\left[Y_t \mid D = d, X = x\right], d, t = 0, 1$.

For the case in which panel data are available, we consider the estimand

$$\tau^{dr,p} = \mathbb{E}\left[\left(w_1^p(D) - w_0^p(D, X; \pi)\right) \left(\Delta Y - \mu_{0,\Delta}^p(X) \right) \right],$$

where, for a generic g ,

$$w_1^p(D) = \frac{D}{\mathbb{E}[D]}, \quad \text{and} \quad w_0^p(D, X; g) = \frac{g(X)(1-D)}{1-g(X)} / \mathbb{E}\left[\frac{g(X)(1-D)}{1-g(X)} \right].$$

When panel data are available, $\tau^{dr,p} = \tau$ if either (but not necessarily both) $\pi(X) = p(X)$ a.s. or $\mu_{\Delta}^p(X) = m_{0,1}^p(X) - m_{0,0}^p(X) a.s.$;

Case 1: Panel data are available and propensity score model is correctly specified

In this case, we have that $\pi(X) = p(X) a.s.$. In order to show that $\tau^{dr,p} = \tau \equiv ATT$, first note that, by the law of iterated expectations,

$$\mathbb{E}\left[\frac{\pi(X)(1-D)}{1-\pi(X)} \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)} \mid X \right] \right] = \mathbb{E}[p(X)] = \mathbb{E}[D],$$

which yields

$$w_0^p(D, X; \pi) = \frac{\pi(X)(1-D)}{1-\pi(X)} / \mathbb{E}\left[\frac{\pi(X)(1-D)}{1-\pi(X)} \right] = \frac{1}{\mathbb{E}[D]} \frac{p(X)(1-D)}{1-p(X)}.$$

Therefore, we have that

$$\begin{aligned}
\tau^{dr,p} &= \frac{1}{\mathbb{E}[D]}\mathbb{E}\left[\left(D-\frac{p(X)(1-D)}{(1-p(X))}\right)\left(\Delta Y-\mu_{0,\Delta}^{p}(X)\right)\right] \\
&= \frac{1}{\mathbb{E}[D]}\mathbb{E}\left[\left(D-\frac{p(X)(1-D)}{(1-p(X))}\right)\Delta Y\right]-\frac{1}{\mathbb{E}[D]}\mathbb{E}\left[\left(D-\frac{p(X)(1-D)}{(1-p(X))}\right)\mu_{0,\Delta}^{p}(X)\right] \\
&= \tau-\mathbb{E}\left[(p(X)-p(X))\cdot\mu_{0,\Delta}^{p}(X)\right] \\
&= \tau
\end{aligned}$$

where the second to last equality follows from Lemma 3.1 and equation (10) in Abadie (2005) and the law of iterated expectations.

Case 2: Panel data are available and outcome regression models are correctly specified.

In this case, we have that $\mu_{0,\Delta}^{p}(X)=m_{0,\Delta}^{p}(X)a.s.$, i.e. the outcome regression models are correctly specified. Let $B=\mathbb{E}\left[\frac{\pi(X)(1-D)}{(1-\pi(X)}\right]^{-1}$. Therefore,

$$\begin{aligned}
\tau^{dr,p} &= \mathbb{E}\left[\frac{D}{\mathbb{E}[D]}\left(\Delta Y-m_{0,\Delta}^{p}(X)\right)\right]-\mathbb{E}\left[w_{0}^{p}(D,X;\pi)\left(\Delta Y-m_{0,\Delta}^{p}(X)\right)\right] \\
&= \mathbb{E}\left[\Delta Y-m_{0,\Delta}^{p}(X)\mid D=1\right]-B\cdot\mathbb{E}\left[\frac{1-\pi(X)}{\pi(X)}\left(\Delta Y-m_{0,\Delta}^{p}(X)\right)\mid D=0\right](1-p(X)) \\
&= \mathbb{E}\left[m_{1,\Delta}^{p}(X)-m_{0,\Delta}^{p}(X)\mid D=1\right]-B\cdot\mathbb{E}\left[\frac{1-\pi(X)}{\pi(X)}\left(m_{0,\Delta}^{p}(X)-m_{0,\Delta}^{p}(X)\right)\mid D=0\right](1-p(X)) \\
&= \tau
\end{aligned}$$

where the third equality follows from the law of iterated expectations, and the last one from the conditional parallel trends assumption, i.e., Assumption 2.

Here's the TWFE specification:

$$Y_{it}=\alpha_1+\alpha_2 T_t+\alpha_3 D_i+\delta\left(T_i\times D_t\right)+\varepsilon_{it}$$

Just add in covariates then right?

$$Y_{it}=\alpha_1+\alpha_2 T_t+\alpha_3 D_i+\delta\left(T_i\times D_t\right)+\theta\cdot X_{it}+\varepsilon_{it}$$

Collecting terms

$$\begin{aligned}
&E\left[Y_1^1\mid D=1,X\right]=\alpha_1+\alpha_2+\alpha_3+\delta+\theta_1 X \\
&E\left[Y_1^0\mid D=1,X\right]=\alpha_1+\alpha_2+\alpha_3+\theta_2 X \\
&E\left[Y_1^1\mid D=1,X\right]-E\left[Y_1^0\mid D=1,X\right] \\
&=(\alpha_1+\alpha_2+\alpha_3+\delta+\theta_1 X)-(\alpha_1+\alpha_2+\alpha_3+\theta_2 X) \\
&=\delta+(\theta_1 X-\theta_2 X)
\end{aligned}$$

By allowing for the possibility that $\theta_1 X\neq\theta_2 X$, we open up the possibility of bias from TWFE which is zero under three additional assumptions.

Assumption 4 The implications of that TWFE regression with assumptions 1-3 gave us those previous expressions which then require placing further restrictions on treatment effects and trends when estimating with TWFE. TWFE Assumption 4: Homogenous treatment effects in X

$$E\left[Y_1^1-Y_1^0 \mid X, D=1\right]=E\left[Y_1^1-Y_1^0 \mid D=1\right]$$

This is because when you difference out those previous equations, you need θX to cancel to leave you with δ which implies homogeneity in X .

X-specific trends : TWFE places restrictions on trends for the two groups too. Take conditional expectations of our TWFE equation.

$$\begin{aligned}
& E\left[Y_1 \mid D=1\right]=\alpha_1+\alpha_2+\alpha_3+\delta+\theta X_{11} \\
& E\left[Y_0 \mid D=1\right]=\alpha_1+\alpha_3+\theta X_{10} \\
& E\left[Y_1 \mid D=0\right]=\alpha_1+\alpha_2+\theta X_{01} \\
& E\left[Y_0 \mid D=0\right]=\alpha_1+\theta X_{00}
\end{aligned}$$

Eliminating terms, we get:

$$\begin{aligned}
\delta^{DD} = & \delta+ \\
& \left(\theta X_{11}-\theta X_{10}\right)-\left(\theta X_{01}-\theta X_{00}\right)
\end{aligned}$$

Assumption 5 and 6

For D=0,1 , we need ”no X-specific trends in both groups”:

$$E\left[Y_1 \bullet -Y_0 \mid D=d, X\right]=E\left[Y_1-Y_0 \mid D=d\right]$$

Intuition: Sant’Anna and Zhao (2020) say in footnote 4 ”[this] follows from analogous arguments” which is the previous slides’ manipulation of terms. Key is to remember these are time-varying covariates so they don’t cancel out within treatment category, so you need the trends in X to cancel out.

3 DML

假设一个简单的 **Partial Linear Regression(PLR) Model**：

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

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，通过假设参数模型（例如常见的线性模型），或者利用一定非参方法（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 $\hat{\theta}_0$：

The two strategies rely on very different moment conditions for identifying and estimating θ_0 :

$$
\begin{aligned}
\mathrm{E}\left[\left(Y-D \theta_0-g_0(Z)\right) D\right]=0 \\
\mathrm{E}\left[\left(Y-D \theta_0\right)(D-E[D \mid Z])\right]=0 \\
\mathrm{E}\left[\left((Y-E[Y \mid Z])-(D-E[D \mid Z]) \theta_0\right)(D-E[D \mid Z])\right]=0
\end{aligned}
$$

- (1) - Regression adjustment;

- (2) - "propensity score adjustment"

- (3) - Neyman-orthogonal (semi-parametrically efficient under homoscedasticity).

Both approaches generate estimators of θ_0 that solve the empirical analog of the moment conditions above, where instead of unknown nuisance functions $g_0(Z), \quad m_0(Z):=\mathrm{E}[D \mid Z], \quad \ell_0(Z)=\mathrm{E}[Y \mid Z]$

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

研究者们很自然地会对 $\hat{\theta}_0$ 的性质进行探讨，例如这个估计量是否无偏。遗憾的是 $\hat{\theta}_0$ 往往是有偏的：

(1)

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

可以看出误差分为两项。a 项来自于 U 和 D 的独立性，若二者不独立则会造成偏误，即 $\theta_0=\frac{\operatorname{cov}\left(D, Y-g_0(X)\right)}{\operatorname{var}(D)}-\underbrace{\frac{\operatorname{cov}(D, U)}{\operatorname{var}(D)}}_{\neq 0}$。

然而 b 项的偏差并不能被简单消除。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 $m_0\left(X_i\right)\left(g_0\left(X_i\right)-\hat{g}_0\left(X_i\right)\right)$ 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，$m_0\left(X_i\right)$ 是数据本身的性质，因此 $m_0\left(X_i\right)$ 会决定偏误的大小而无法缓解，导致估计非常不稳健。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。

为了尽可能消掉 $m_0\left(X_i\right)\left(g_0\left(X_i\right)-\hat{g}_0\left(X_i\right)\right)$ 项，一个实际的考虑是消除 $m_0\left(X_i\right)$。

Consider estimation based on (3)

$$
\check{\theta}_0=\left(\frac{1}{n} \sum_{i=1}^n \widehat{V}_i^2\right)^{-1} \frac{1}{n} \sum_{i=1}^N \widehat{V}_i \widehat{W}_i
$$

- $\widehat{V} = D - \widehat{m}_0(Z), \widehat{W} = Y - \widehat{\ell}_0(Z)$, Under mild conditions, can write

$$\begin{aligned}\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) = & \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} V_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i U_i}_{:=a^*} \\ & + \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} V_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(m_0\left(Z_i\right) - \hat{m}_0\left(Z_i\right)\right)\left(\ell_0\left(Z_i\right) - \hat{\ell}_0\left(Z_i\right)\right)}_{:=b^*}\end{aligned}$$

Key difference between estimation based on (1) and estimation based on (3) is that (3) satisfies the Neyman orthogonality condition: Let

$$\eta_0 = (\ell_0, m_0) = (\mathrm{E}[Y \mid Z], \mathrm{E}[D \mid Z]), \quad \eta = (\ell, m).$$

The Gateaux derivative operator of the moment condition (3) with respect to η vanishes:

$$\left.\partial_\eta \mathrm{E}[\underbrace{((Y - \ell(Z)) - (D - m(Z))\theta_0)\,(D - m(Z))}_{\psi(W,\theta_0,\eta)}]\right|_{\eta=\eta_0} = 0$$

Heuristically, the moment condition remains "valid" under "local" mistakes in the nuisance function.

**Neyman 正交化和矩条件 **

τ_0 可以被视为以下估计方程的解：

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

其中，φ 是一个已知的评分函数（score function），$\hat{g}0$ 是干扰参数 $g0$ 的估计。例如，在偏回归模型中，评分函数定义为 $\varphi(W; \tau, g) = (Y - \tau W - g(\mathbf{X}))W$。显然，评分函数 φ 对于 g 的估计是否有偏非常敏感。具体而言，关于 g 的 Gateaux 导数算子不等于 0：

$$\partial_g \mathbb{E}\left[\varphi\left(W; \tau_0, g_0\right)\right]\left[g - g_0\right] := \lim_{r\to 0}\frac{\partial \mathbb{E}\left[\phi\left(W; \tau_0, g_0 + r\left(g - g_0\right)\right)\right]}{\partial r} \neq 0$$

这个条件是确保估计量良好性能的关键。相比之下，$\check{\tau}_0 = \left(\frac{1}{n_m}\sum_{i\in\mathbf{I}}\hat{Z}_i W_i\right)^{-1}\frac{1}{n_m}\sum_{i\in\mathbf{I}}\hat{Z}_i\left(Y_i - \hat{g}_0\left(\mathbf{X}_i\right)\right)$ 中的 DML 估计 $\check{\tau}_0$ 是以下方程的解：

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

其中，$\hat{\eta}0$ 是干扰参数 $\eta 0 = (g_0, e_0)$ 的估计，ψ 是一个满足 Gateaux 导数算子为 0 的正交评分函数：

$$\partial_\eta \mathbb{E}\left[\psi\left(W; \tau_0, \eta_0\right)\right]\left[\eta - \eta_0\right] := \lim_{r\to 0}\frac{\partial \mathbb{E}\left[\psi\left(W; \tau_0, \eta_0 + r\left(\eta - \eta_0\right)\right)\right]}{\partial r} = 0$$

该条件被称为 Neyman 正交性，而函数 ψ 被称为 Neyman 正交评分函数，这个概念是由 Neyman 首次提出的。直观来说，Neyman 正交性条件意味着用于识别 τ_0 的矩条件对干扰参数的值局部不敏感，这允许我们在估计 τ_0 时使用干扰参数的有偏估计而不会强烈违反矩条件。在偏线性模型中，估计量

$\check{\tau}_0$ 使用评分函数 $\psi(W;\tau,\eta)=(Y-W\tau-g(\mathbf{X}))(W-e(\mathbf{X}))$，其中干扰参数为 $\eta=(e,g)$。很容易看出，这些评分函数 ψ 对 η_0 的有偏估计不敏感，即性质 (2-15) 成立。这个性质和样本划分是双重机器学习方法中的两个关键，使得我们能够建立一个 τ_0 的有效估计。

因此，推断 τ_0 的方法是使用评分函数和 DML (Double/Debiased Machine Learning) 方法：

$$\psi(D;\tau,\eta):=Y-W\tau-g(X)(W-e(X)),\quad \eta=(e,g)$$

其中 $D=(Y,W,X)$，g 和 e 是从 $\mathcal{X}$ 映射到 $\mathbb{R}$ 的函数。可以看出，τ_0 满足矩条件 $\mathbb{E}\left[\psi\left(D;\tau_0,\eta_0\right)\right]=0$，并且满足正交条件 $\partial_\eta\mathbb{E}\left[\psi\left(D;\tau_0,\eta_0\right)\right]\left[\eta-\eta_0\right]=0$。

Now, addressing your second point, let $\delta_\ell(X),\delta_m(X)$ be two test functions respectively perturbing ℓ and m . Then Neyman orthogonality states that for choices of δ_ℓ and δ_m, we have

$$\frac{\mathrm{d}E\left[\psi\left(W;\theta,\eta_0+r\left(\delta_\ell,\delta_m\right)\right)\right]}{\mathrm{d}r}=0$$

where the derivative is taken around the point where r=0 . To prove this, we can simply expand out the definition of ψ to obtain

$$\begin{aligned}E\left[\psi\left(W;\theta,\eta_0+r\left(\delta_\ell,\delta_m\right)\right)\right]=&\;E\left[\left(Y-\ell_0(X)-r\delta_\ell(X)\right)\left(D-m_0(X)-r\delta_m(X)\right)\right]\\&-\theta E\left[\left(D-m_0(X)-r\delta_m(X)\right)^2\right]\end{aligned}$$

Let us first check that the derivative of the first term is mean 0. To do so, we note that by differentiating under the expectation sign around r=0 , we have

$$\begin{aligned}&=E\left[\tfrac{\mathrm{d}}{\mathrm{d}r}\left(Y-\ell_0(X)-r\delta_\ell(X)\right)\left(D-m_0(X)-r\delta_m(X)\right)\right]\\&=-E\left[\left(Y-\ell_0(X)\right)\delta_m(X)+\delta_\ell(X)\left(D-m_0(X)\right)\right]\\&=-E[\underbrace{E\left[Y-\ell_0(X)\mid X\right]}_{=0}\delta_m(X)]-E\left[E[\delta_\ell(X)\underbrace{E\left[D-m_0(X)\mid X\right]}_{=0}]\right.=0\end{aligned}$$

In light of the above, all that remains to be checked is that the limit of the third term goes to 0 as $r\to0$. Specifically, we must show

$$\frac{\theta\mathrm{d}E\left[\left(D-m_0(X)-r\delta_m(X)\right)^2\right]}{\mathrm{d}r}=0$$

Now, differentiating under the integral sign again, we have

$$\begin{aligned}\frac{\mathrm{d}\theta E\left[\left(D-m_0(X)-r\delta_m(X)\right)^2\right]}{\mathrm{d}r}&=\theta E\left[\frac{\mathrm{d}}{\mathrm{d}r}D-m_0(X)-r\delta_m(X)\Bigg)^2\right]\\&=-2\theta E\left[\left(D-m_0(X)\right)\delta_m(X)\right]\\&=\theta E\left[E\left[\left(D-m_0(X)\right)\delta_m(X)\mid X\right]\right]\\&=\theta E[\underbrace{E\left[\left(D-m_0(X)\right)\mid X\right]}_{=0}\delta_m(X)]\\&=\theta E[0\mid X]=0\end{aligned}$$

So once again, after some manipulation, this third term equalling 0 is due to the definition of m_0 as the conditional expectation function of D . 这个条件保证了估计量的无偏性和一致性。通过使用正交评分函数 ψ，DML 估计 $\check{\tau}_0$ 能更好地处理干扰参数 g_0 和 e_0 的估计。

单个数据集中平均处理效应 (ATE) 的估计量 $\hat{\tau}_{\text{DML}}$ 具有以下形式：

$$\begin{aligned}\tau_{\text{DML}} &= \left(\tfrac{1}{n_m}\textstyle\sum_{i\in I} W_i\left(W_i-\hat{e}_0\left(\mathbf{X_i}\right)\right)\right)^{-1}\tfrac{1}{n_m}\textstyle\sum_{i\in I}\left(Y_i-\hat{g}_0\left(\mathbf{X}_i\right)\right)\left(W_i-\hat{e}_0\left(\mathbf{X_i}\right)\right)\\ &= \left(\tfrac{1}{n_m}\textstyle\sum_{i\in I} W_i\hat{Z}_i\right)^{-1}\tfrac{1}{n_m}\textstyle\sum_{i\in I}\hat{Z}_i\left(Y_i-\hat{g}_0\left(\mathbf{X}_i\right)\right)\end{aligned}$$

其中干扰参数 η_0 使用辅助样本估计。这里的 $\hat{Z}_i = W_i - \hat{e}_0(\mathbf{X}_i)$ 是从 W_i 中移除 $\mathbf{X}_i$ 影响后的正交化的回归量。通过使用 $\hat{Z}_i$ 和从辅助样本得到的 $\hat{g}_0$ 估计来消除混淆直接影响，并使用样本均值来估计干扰参数，τ_{DML} 是对 ATE 的一种有效估计。

4 DML and DRDID

在考虑处理效应完全异质且处理分配变量为二元变量 $W \in 0,1$ 的情况下，我们可以使用以下模型来描述三元组 $D = (Y, W, \mathbf{X})$：

$$\begin{aligned}Y = g_0(W,\mathbf{X}) + U, &\quad \mathbb{E}_P[U \mid \mathbf{X}, W] = 0,\\ W = e_0(\mathbf{X}) + Z, &\quad \mathbb{E}_P[Z \mid \mathbf{X}] = 0.\end{aligned}$$

在这个模型中，处理分配变量 W 在结果模型中是不可分的，相比于处理分配变量为二元情况下的偏回归模型，这个模型更为通用 [21]。在这个模型中，我们感兴趣的目标参数是平均处理效应（ATE）：

$$\tau_0 = \mathbb{E}_P\left[\mu_0(1,\mathbf{X}) - \mu_0(0,\mathbf{X})\right]$$

其中，$\mu_0(W,\mathbf{X}) = \mathbb{E}[Y \mid W, \mathbf{X}]$。另外一个常见的目标参数是个体处理效应（ATT）：

$$\tau_0 = \mathbb{E}_P\left[\mu_0(1,\mathbf{X}) - \mu_0(0,\mathbf{X}) \mid W = 1\right]$$

混杂因素 $\mathbf{X}$ 通过倾向得分 $e_0(\mathbf{X})$ 影响处理分配变量 W，并通过函数 $g_0(D,\mathbf{X})$ 影响结果变量。这些函数是未知的，并且可能具有复杂的形式，以灵活地刻画处理效应的异质性。因此，使用机器学习方法来学习这些函数是更为合适的选择。这样可以通过机器学习算法来探索变量之间的复杂关系，并从中学习到处理效应的异质性模式。

我们同样采用具有 Neyman 正交评分的矩条件进行推断。对于 ATE 的估计，影响函数的形式如下：

$$\psi(D;\tau,\eta) := g(1,\mathbf{X}) - g(0,\mathbf{X}) + \frac{W(Y-g(1,\mathbf{X}))}{e(\mathbf{X})} - \frac{(1-W)(Y-g(0,\mathbf{X}))}{1-e(\mathbf{X})} - \tau$$

其中，干扰参数 $\eta = (g, e)$ 由 P 次可积函数 g 和 e 组成，分别将 $(W, \mathbf{X})$ 的支撑集映射到 $\mathbb{R}$ 和将 $\mathbf{X}$ 的支撑集映射到 $(\epsilon, 1-\epsilon)$，其中 $\epsilon \in (0, 1/2)$。干扰参数 η 的真实值为 $\eta_0 = (g_0, e_0)$。这个正交矩条件基于 Robins 和 Rotnitzky 提出的缺失数据的平均值的影响函数 [50]。通过使用这个影响函数，我们可以获得基于评分的估计量 $\tilde{\tau}0$，该估计量满足矩条件 $\mathbb{E}[\psi(D;\tau_0,\eta_0)] = 0$ 和正交化条件 $\partial\eta\mathbb{E}[\psi(D;\tau_0,\eta_0)][\eta-\eta_0] = 0$。

对于 ATT 的估计，使用的影响函数如下：

$$\psi(D;\tau,\eta)=\frac{W(Y-\bar{g}(\mathbf{X}))}{p}-\frac{e(\mathbf{X})(1-W)(Y-\bar{g}(\mathbf{X}))}{p(1-e(\mathbf{X}))}-\frac{W\tau}{p}$$

其中，干扰参数 $\eta=(g,e,p)$ 由 P 次可积的函数 g 和 e 组成，分别将 $(W,\mathbf{X})$ 的支撑集映射到 $\mathbb{R}$ 和将 $\mathbf{X}$ 的支撑集映射到 $(\epsilon,1-\epsilon)$，其中 $\epsilon\in(0,1/2)$。干扰参数 η 的真实值为 $\eta_0=(g_0,e_0,p_0)$，其中 $\bar{g}_0(X)=g_0(0,X)$，并且 $p_0=\mathbb{E}[W]$。需要注意的是，估计 ATT 并不需要估计 $g_0(1,X)$

由于 p 是一个常数，在基于评分 ψ（公式（2-22））得到的 $\tilde{\tau}_0$ 中不会产生影响。然而，这简化了 $\tilde{\tau}_0$ 的方差形式。通过使用这些评分函数，我们可以观察到 ATE 或 ATT 的真实参数值 τ_0 遵循以下矩条件和正交化条件：

Doubly Robust 公式

$$\mathrm{A\hat{T}}=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{T_i}\left(\mathrm{Y_i}-\hat{\mu}_1\left(\mathrm{X_i}\right)\right)}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}+\hat{\mu}_1\left(\mathrm{X_i}\right)\right)-\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\left(1-\mathrm{T_i}\right)\left(\mathrm{Y_i}-\hat{\mu}_0\left(\mathrm{X_i}\right)\right)}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}+\hat{\mu}_0\left(\mathrm{X_i}\right)\right)$$

$$\begin{aligned}&\hat{\mu}_1\left(\mathrm{X_i}\right)\ :\text{ estimation of }\mathrm{E}[\mathrm{Y}\mid\mathrm{X},\mathrm{T}=1]\\&\hat{\mu}_1\left(\mathrm{X_0}\right)\ :\text{ estimation of }\mathrm{E}[\mathrm{Y}\mid\mathrm{X},\mathrm{T}=0]\\&\hat{\mathrm{e}}\left(\mathrm{X_i}\right)\ :\text{ estimation of }\mathrm{E}[\mathrm{T}=1\mid\mathrm{X}]\end{aligned}$$

Doubly Robust 推导

需证明: $\mathrm{A\hat{T}E}=\mathrm{E}\left(\mathrm{Y_1}\right)-\mathrm{E}\left(\mathrm{Y_0}\right)$

首先证明: $\hat{\mathrm{E}}\left(\mathrm{Y_1}\right)=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{T_i}(\mathrm{Y_i}-\hat{\mu}_1(\mathrm{X_i}))}{\hat{\mathrm{e}}(\mathrm{X_i})}+\hat{\mu}_1\left(\mathrm{X_i}\right)\right)$

假设 $\hat{\mu}_1(\mathrm{X})$ 正确，则 $\mathrm{E}\left[\mathrm{T_i}\left(\mathrm{Y_i}-\hat{\mu}_1\left(\mathrm{X_i}\right)\right]=0\right.$ 因为乘以 $\mathrm{T_i}$ 只选择了观测值为 $\mathrm{T}=1$ 的样本，这些样本由 $\hat{\mu}_1$ 估计的残差的期望是 0 (根据定义)，同样根据定义，$\hat{\mathrm{E}}\left(\mathrm{Y_1}\right)=\mathrm{E}\left(\hat{\mu}_1\left(\mathrm{X_i}\right)\right)$

假设 $\hat{e}(X)$ 正确，则 $\mathrm{E}\left[\mathrm{T_i}-\hat{\mathrm{e}}\left(\mathrm{X_i}\right)\right]=0$ ，且根据逆概率加权，$\hat{\mathrm{E}}\left(\mathrm{Y_1}\right)=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{\mathrm{T_i}\mathrm{Y_i}}{\hat{\mathrm{e}}(\mathrm{X_i})}$

$$\begin{aligned}\hat{\mathrm{E}}\left(\mathrm{Y_1}\right)&=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{T_i}\left(\mathrm{Y_i}-\hat{\mu}_1\left(\mathrm{X_i}\right)\right)}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}+\hat{\mu}_1\left(\mathrm{X_i}\right)\right)\\&=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\frac{\mathrm{T_i}\mathrm{Y_i}}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}-\frac{\mathrm{T_i}\hat{\mu}_1\left(\mathrm{X_i}\right)}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}+\hat{\mu}_1\left(\mathrm{X_i}\right)\right)\\&=\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left[\frac{\mathrm{T_i}\mathrm{Y_i}}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}-\left(\frac{\mathrm{T_i}-\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}{\hat{\mathrm{e}}\left(\mathrm{X_i}\right)}\right)\hat{\mu}_1\left(\mathrm{X_i}\right)\right]\end{aligned}$$

5 Goodman DID

What is Difference-in-Differences?

$$y_{it}=\gamma_0\text{ TREAT }_i+\gamma_1\text{ POST }_t+\hat{\beta}^{DD}\text{ TREAT }_i\text{ POST }_t+u_{it}$$

$$\hat{\beta}^{DD}=\left(\bar{y}_{POST}^{TREA}-\bar{y}_{PRE}^{TREAT}\right)-\left(\bar{y}_{POST}^{CONTROL}-\bar{y}_{PRE}^{CONTROL}\right)$$

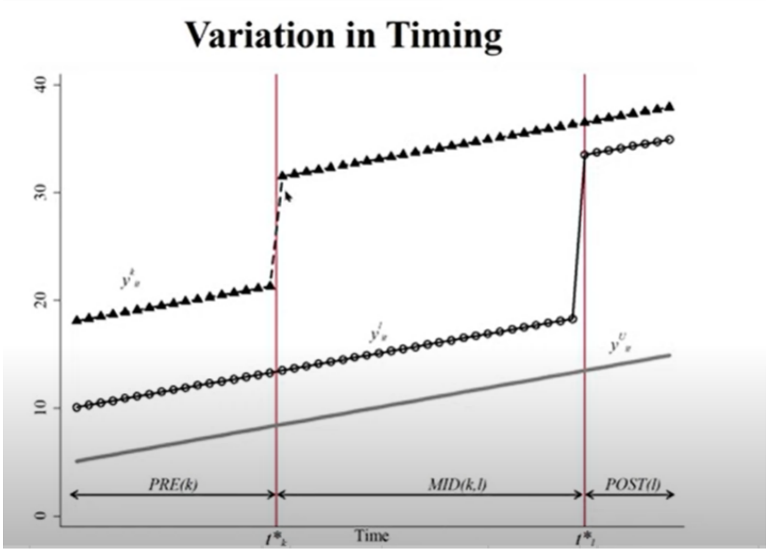

Variation in Timing

图 1

What is $\widehat{\boldsymbol{\beta}}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

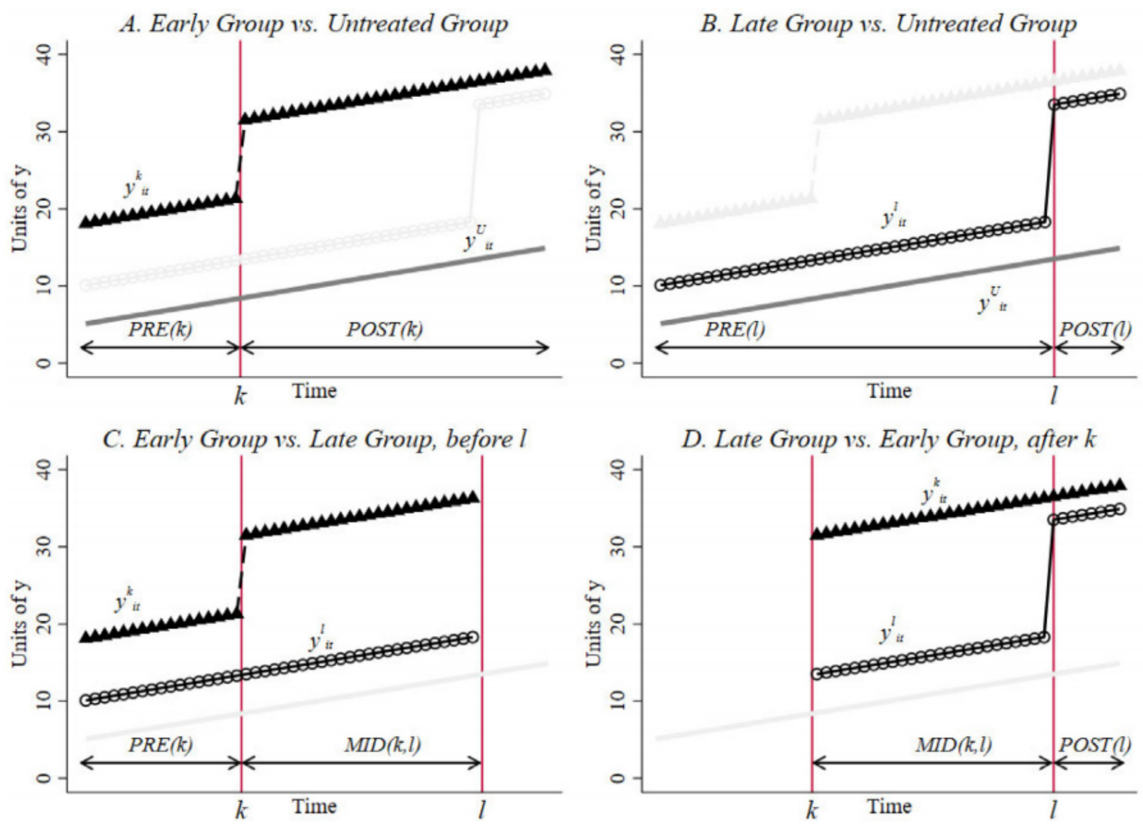
1. Partial out fixed effects (Frisch-Waugh):

$$\widetilde{D}_{it} = \left(D_{it} - \overline{\overline{D}}\right) - \left(\bar{D}_i - \overline{\overline{D}}\right) - \left(\bar{D}_t - \overline{\overline{D}}\right)$$

$$\tilde{y}_{it} = (y_{it} - \overline{\overline{y}}) - (\bar{y}_i - \bar{y}) - (\bar{y}_t - \overline{\overline{y}})$$

2. Calculate univariate coefficient by brute force:

$$\hat{\beta}^{DD} = \frac{\widehat{Cov}\left(\widetilde{D}_{it}\tilde{y}_{it}\right)}{\bar{\nabla}\left(\widetilde{D}_{it}\right)} = \frac{\frac{1}{NT}\sum_i \sum_t (y_{it} - \overline{\overline{y}})\left(D_{it} - \overline{\overline{D}}\right)}{\frac{1}{NT}\sum_i \sum_t \left(D_{it} - \overline{\overline{D}}\right)}$$


A. Early Group vs. Untreated Group

B. Late Group vs. Untreated Group

C. Early Group vs. Late Group, before l

D. Late Group vs. Early Group, after k

图 2

What is $\widehat{\boldsymbol{\beta}}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}_{kU}^{DD} + s_{\ell U}\hat{\beta}_{\ell U}^{DD} + \left[s_{k\ell}^{k}\hat{\beta}_{k\ell}^{DD,k} + s_{k\ell}^{\ell}\hat{\beta}_{k\ell}^{DD,\ell}\right]$$

其中 2 × 2 的 DD 的估计器为:

$$\begin{aligned}
\hat{\beta}_{kU}^{2\times2} &\equiv \left(\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_U^{POST(k)} - \bar{y}_U^{PRE(k)}\right), \\
\hat{\beta}_{k\ell}^{2\times2,k} &\equiv \left(\bar{y}_k^{\mathrm{MID}(k,\ell)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_\ell^{\mathrm{MID}(k,\ell)} - \bar{y}_\ell^{PRE(k)}\right), \\
\hat{\beta}_{k\ell}^{2\times2,\ell} &\equiv \left(\bar{y}_\ell^{POST(\ell)} - \bar{y}_\ell^{\mathrm{MID}(k,\ell)}\right) - \left(\bar{y}_k^{\mathrm{POST}\ (\ell)} - \bar{y}_k^{\mathrm{MID}(k,\ell)}\right).
\end{aligned}$$

权重为:

$$s_{kU} = \frac{(n_k + n_U)\overbrace{n_{kU}\left(1 - n_{kU}\right)\bar{D}_k\left(1 - \bar{D}_k\right)}^{\hat{V}^D}}{\hat{V}_{kU}^D},$$

$$s_{k\ell}^{k} = \frac{\left((n_k + n_\ell)\left(1 - \bar{D}_\ell\right)\right)^2 \overbrace{n_{k\ell}\left(1 - n_{k\ell}\right)\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}^{\hat{V}^D}}{\hat{V}_{k\ell}^{D,k}}$$

$$s_{k\ell}^{\ell}=\frac{\left((n_k+n_\ell)\bar{D}_k\right)^2\overbrace{k\ell}(1-n_{k\ell})\frac{\bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_k-\bar{D}_\ell}{\bar{D}_k}}{\hat{v}_{k\ell}^{D,\ell}}$$

$$\text{and } \textstyle\sum_{k\neq U}s_{kU}+\sum_{k\neq U}\sum_{\ell>k}\left[s_{k\ell}^{k}+s_{k\ell}^{\ell}\right]=1.$$

差法分解定理完全描述了 TWFEDD 估计器中可识别变异的来源及其重要性。假设数据包含按照接受二元治疗的时间排序的 K 个时机组（$k=1,\ldots,K$），其中可能存在一个包含从未接受治疗的单位的时机组 U。在双向固定效应回归模型 $y_{it}=\alpha_{i\cdot}+\alpha_{\cdot t}+\beta^{DD}D_{it}+e_{it}$ 中，OLS 估计量 $\hat{\beta}^{DD}$ 是所有可能的 2×2 差异法估计量的加权平均值。

Difference-in-Differences Decomposition Theorem Assume that there are $k=1,\ldots,K$ groups of treated units ordered by treatment time t_k^* and one control group, U , which does not receive treatment in the data. The share of units in group k is n_k , and the share of periods that group k spends under treatment is $\bar{D}_k$. The DD estimate from a two-way fixed effects model is a weighted average all two-group DD estimators:

$$\hat{\beta}^{DD}=\sum_{k\neq U}s_{kU}\hat{\beta}_{kU}^{DD}+\sum_{k\neq U}\sum_{\ell>k}\left[s_{k\ell}^{k}\hat{\beta}_{k\ell}^{D,k}+s_{k\ell}^{\ell}\hat{\beta}_{k\ell}^{DD,\ell}\right]$$

With weights equal to:

$$\begin{aligned}
s_{kU}&=\frac{(n_k+n_U)^2\,\hat{V}_{kU}^{D}}{V\left(\widetilde{D}_{it}\right)}\\
s_{k\ell}^{k}&=\frac{\left((n_k+n_\ell)\left(1-\bar{D}_\ell\right)\right)^2\hat{V}_{k\ell}^{D,k}}{V\left(\widetilde{D}_{it}\right)}\\
s_{k\ell}^{\ell}&=\frac{\left((n_k+n_\ell)\,\bar{D}_k\right)^2\hat{V}_{k\ell}^{D,\ell}}{V\left(\widetilde{D}_{it}\right)}
\end{aligned}$$

$$\sum_{k\neq U}s_{kU}+\sum_{k\neq U}\sum_{\ell>k}\left[s_{k\ell}^{k}+s_{k\ell}^{\ell}\right]=1.$$

6 Sun DID

TWFE Regression

$$Y_{i,t}=\alpha_i+\delta_t+\sum_{g\in G}\mu_g 1\left\{t-E_i\in g\right\}+\varepsilon_{i,t}$$

They say E_i is the initial time of a binary variable absorbing treatment for unit i . Fixed effects should be obvious. μ_g is the population regression coefficient on the leads and lags that we want to estimate. We estimate this using OLS and get $\widehat{\mu_g}$.

We are interested in the properties of μ_g under differential timing as well as whether there are any never-treated units

首先对“事件研究设计”进行形式化，具体而言，我们考虑一个随机样本，其中观察了 $T+1$ 个时间期内的 N 个单位，其中 T 是固定的。对于每个 $i\in 0,\ldots,N$ 和 $t\in 0,\ldots,T$，我们观察到结果 $Y_{i,t}$ 和处理状态 $D_{i,t}\in 0,1$：如果在时间期 t 中单位 i 受到处理，则 $D_{i,t}=1$，如果单位 i 在时间期

t 中未受到处理，则 $D_{i,t} = 0$。在整个分析过程中，我们假设观测值 $\{Y_{i,t}, D_{i,t}\}_{t=0}^{T}$ 是独立同分布的（i.i.d.）。

我们可以通过初始处理的时间期来唯一确定一个处理路径，记为 $E_i = \min\{t : D_{i,t} = 1\}$ 。如果单位 i 从未接受过处理，即对于所有的 t ，有 $D_{i,t} = 0$，我们将 $E_i = \infty$ 。根据首次接受处理的时间，我们还可以将单位唯一地分为不相交的群体 e ，其中 $e \in 0, \ldots, T, \infty$，群体 e 中的单位在相同的时间首次接受处理，即 $\{i : E_i = e\}$ 。

Specifying the leads and lags How will we specify the $1\{t - E_i \in g\}$ term? SA considers a couple:

(3) Static specification:

$$Y_{i,t} = \alpha_i + \delta_t + \mu_g \sum_{1 \geq 0} D_{i,t}^{l} + \varepsilon_{i,t}$$

(2) Dynamic specification:

$$Y_{i,t} = \alpha_i + \delta_t + \sum_{l=-K}^{-2} \mu_{i\cdot} D_{i,t}^{\prime} + \sum_{l=0}^{L} \mu_l D_{i,t}^{l} + \varepsilon_{i,t}$$

Definition 1:

The cohort-specific ATT I periods from initial treatment date e is:

$$\mathrm{CATT}_{e,l} = E\left[Y_{i,e+l} - Y_{i,e+1}^{\infty} \mid E_i = e\right]$$

Assumption 1: Parallel trends in baseline outcomes: $E\left[Y_{i,t}^{\infty} - Y_{i,s}^{\infty} \mid E_i = e\right]$ is the same for all $e \in \operatorname{supp}(E_i)$ and for all s , t and is equal to $E\left[Y_{i,t}^{\infty} - Y_{i,s}^{\infty}\right]$

Assumption 2: No anticipator behavior in pre-treatment periods: There is a set of pre-treatment periods such that $E\left[Y_{i,e+1}^{e} - Y_{i,e+l}^{\infty} \mid E_i = e\right] = 0$ for all possible leads.

Basically means that potential outcomes prior to treatment at baseline by on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

Assumption 3: Treatment effect homogeneity: For each relative time period I , the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_1$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

Multicollinearity

Dynamic specification requires deciding which leads to drop. They recommend dropping two: I=-1 and somè other one (they seem to favor I=-4) . The reason is twofold. You drop one of them to avoid multicollinearity in the relative time indicators. You drop a second one because of the multicollinearity coming from the linear relationship between TWFE and the relative period indicators.

6.1 Interpreting $\widehat{\mu_g}$ under no to all assumptions

Proposition 1 (no assumptions): The population regression coefficient on relative period bin g is a linear combination of differences in trends from its own relative period $l \in g$, from relative periods $l \in g'$ of other bins $g' \neq g$, and from relative periods excluded from the specification (e.g., trimming).

$$\begin{aligned}\mu_g = &\underbrace{\sum_{l\in g}\sum_{e} w^g_{e,l}\left(E\left[Y_{i,e+l}-Y^{\infty}_{i,0}\mid E_i=e\right]-E\left[Y^{\infty}_{i,e+l}-Y^{\infty}_{i,0}\right]\right)}_{\text{Good stuff}}\\ &+\underbrace{\sum_{g'\neq g}\sum_{l\in g'}\sum_{e} w^g_{e,l}\left(E\left[Y_{i,e+l}-Y^{\infty}_{i,0}\mid E_i=e\right]-E\left[Y^{\infty}_{i,e+l}-Y^{\infty}_{i,0}\right]\right)}_{\text{Bleh - Other included relative time}}\\ &+\underbrace{\sum_{l\in g^{\text{exc}}}\sum_{e} w^g_{e,\lambda}\left(E\left[Y_{i,e+1}-Y^{\infty}_{i,0}\mid E_i=e\right]-E\left[Y^{\infty}_{i,e+1}-Y^{\infty}_{i,0}\right]\right)}_{\text{More bleh - Excluded}}\end{aligned}$$

Superscript g associates the weight with coefficient μ_g . The weight associated with cohort e in relative period I is equal to the population regression coefficient on the $1\left\{t-E_i \in g\right\}$ from regression $D'_{i,t} \times 1\left\{E_i=e\right\}$ on all bin indicators included in the regression and TWFE. Just the mechanics of double demeaning from TWFE

Weight $\left(w^g_{e,l}\right)$ summation cheat sheet

(1) For relative periods of $\mu_g own l \in g, \sum_{l\in g}\sum_e w^g_{e,l}=1$

(2) For relative periods belonging to some other bin $l \in g'$ and $g' \neq g,\mathrm{t}\sum_{l\in g'}\sum_e w^g_{e,l}=0$

(3) For relative periods not included in G , $\sum_{l\in g^{\text{excl}}}\sum_e w^g_{e,l}=-1$

Proposition 2: Under the parallel trends only, the population regression coefficient on the indicator for relative period bing g is a linear combination of $CATT_{e,l\in g}$ as well as $CATT_{d,l'}$ from other relative periods $I' \notin g$ with the same weights stated in Proposition 1:

$$\begin{aligned}\mu_g = &\underbrace{\sum_{l\in g}\sum_{e} w^g_{e,l}CATT_{e,l}}_{\text{Desirable}}\\ &+\underbrace{\sum_{g'\neq g, g'\in G}\sum_{l'\in g'}\sum_{e} w^g_{e,l'}CATT_{e,l'}}_{\text{Undesirable - other specified bins}}\\ &+\underbrace{\sum_{l'\in g^{\text{excl}}}\sum_{e} w^g_{e,l'}CATT_{e,l'}}_{\text{Undesirable - excluded relative time indicators}}\end{aligned}$$

Proposition 3: If parallel trends holds and no anticipation holds for all I<0 (i.e., no anticipatory behavior pre-treatment), then the population regression coefficient μ_g for g is a linear combination of post-treatment $CATT_{e,l'}$ for all $I' \geq 0$.

$$\begin{aligned}&\mu_g=\textstyle\sum_{l'\in g, l'\geq 0}\sum_e w^g_{e,l'}\ \mathrm{CATT}_{e,l'}\\ &+\textstyle\sum_{g'\neq g, g'\in G}\sum_{l'\in g', l'\geq 0}\sum_e w^g_{e,l'}\ \mathrm{CATT}\ _{e,l'}\\ &+\textstyle\sum_{l'\in g^{\text{excl}}, l'\geq 0}\sum_e w^g_{w,l'}\ \mathrm{CATT}\ _{e,l'}\end{aligned}$$

Notice how once we impose zero pre-treatment treatment effects, those terms are gone (i.e., no $l \in g, l < 0$). But the second term remains unless we impose treatment effect homogeneity (homogeneity causes terms due to weights summing to zero to cancel out). Thus μ_g may be non-zero for pre-treatment periods even though parallel trends hold in the pre period.

Proposition 4: If parallel trends and treatment effect homogeneity, then $CATT_{e,l} = ATT_1$ is constant across e for a given I , and the population regression coefficient μ_g is equal to a linear combination of $ATT_{l \in g}$, as well as $ATT_{l' \notin g}$ from other relative periods

$$\begin{aligned} \mu_g = & \sum_{l \in g} w_l^g ATT_l \\ & + \sum_{g' \neq g} \sum_{l' \in g'} w_{l'}^g ATT_{l'} \\ & + \sum_{l' \in g_{\text{excl}}} w_{l'}^g ATT_{l'} \end{aligned}$$

6.2 toy example

Balanced panel T=2 with cohorts $E_i \in \{1, 2\}$. We drop two relative time periods to avoid multicollinearity, so we will include bins $\{-2, 0\}$ and drop $\{-1, 1\}$.

$$\begin{aligned} \mu_{-2} = & \underbrace{\text{ CATT }_{2,-2}}_{\text{own period}} + \underbrace{\frac{1}{2} \text{ CATT }_{1,0} - \frac{1}{2} \text{ CATT }_{2,0}}_{\text{other included bins}} \\ & + \underbrace{\frac{1}{2} \text{ CATT }_{1,1} - \text{ CATT }_{1,-1} - \frac{1}{2} \text{ CATT }_{2,-1}}_{\text{Excluded bins}} \end{aligned}$$

- Parallel trends gets us to all of the CATT
    - No anticipation makes CATT =0 for all I<0 (all I<0 cancel out)
    - Homogeneity cancels second and third terms
    - Still leaves $\frac{1}{2} CATT_{1,1}$ - you chose to exclude a group with a treatment effect

** 步骤 1. ** 使用线性的双向固定效应规范来估计从初始处理开始的 $C_{ATT_{e,\ell}}$，该规范将相对期间指标与同组指标进行交互，并排除来自某个集合 C 的同组指标：

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{e \notin C} \sum_{\ell \neq -1} \delta_{e,\ell} \left(\mathbf{1} \{E_i = e\} \cdot D_{i,t}^{\ell} \right) + \epsilon_{i,t}$$

具体规范取决于给定应用中的同组份额。如果存在从未接受处理的同组，即 $\infty \in \operatorname{supp}\{E_i\}$，那么我们可以设置 $C = \infty$ 并对所有观测估计回归方程。如果不存在从未接受处理的单位，即 $\infty \notin \operatorname{supp}\{E_i\}$，那么我们可以设置 $C = \{\max\{E_i\}\}$，即最晚接受处理的同组，并对观测值 $t = 0, \ldots, \max\{E_i\} - 1$ 估计回归方程。最后，如果存在一个始终接受处理的同组，即 $0 \in \operatorname{supp}\{E_i\}$，那么我们需要从估计中排除该同组。

回归中的系数估计量 $\widehat{\delta}_{e,\ell}$ 是 $CATT_{e,\ell}$ 的差分中差（DID）估计量，具体的选择取决于前期和控制群体。

** 步骤 2. ** 通过样本中每个群体在相关期间 $\ell \epsilon g$ 的份额来估计权重 $\operatorname{Pr}\{E_i=e \mid E_i \in[-\ell, T-\ell]\}$

** 步骤 3. ** 为了构建我们的 IW 估计量，我们将步骤 1 中对 $CATT_{e,\ell}$ 的估计结果按照步骤 2 中得到的权重进行加权平均。更具体地说，IW 估计量为：

$$\widehat{v}_g=\frac{1}{|g|} \sum_{\ell \in \sigma} \sum_\rho \widehat{\delta}_{e, \ell} \widehat{\operatorname{Pr}}\{E_i=e \mid E_i \in[-\ell, T-\ell]\}$$

$$\hat{\delta}_{e, \ell}=\frac{\mathbb{E}_N\left[\left(Y_{i, e+\ell}-Y_{i, s}\right) \cdot \mathbf{1}\left\{E_i=e\right\}\right]}{\mathbb{E}_N\left[\mathbf{1}\left\{E_i=e\right\}\right]}-\frac{\mathbb{E}_N\left[\left(Y_{i, e+\ell}-Y_{i, s}\right) \cdot \mathbf{1}\left\{E_i \in C\right\}\right]}{\mathbb{E}_N\left[\mathbf{1}\left\{E_i \in C\right\}\right]}$$

如果平行趋势假设和无预期假设成立，那么使用任何小于 e 的预处理期间 s 和非空的对照组 C 的 DID 估计量是对 $C_{ATT_{e,\ell}}$ 的无偏一致估计量。

7 CSDID

- On the other hand, our paper has some unique features on it:

- We allow for covariates in a flexible form;

-We propose different estimation procedures based on outcome regression, IPW and doubly robust methods;

- We discuss different aggregation schemes to further summarize the effects of the treatment;

- We cover both panel and (stationary) repeated-cross section cases;

- We attempt to make minimal parallel trends assumptions to identify the ATT(g, t).

Framework for the panel data case - Consider a random sample

$$\left\{\left(Y_{i, 1}, Y_{i, 2}, \ldots, Y_{i, \mathcal{T}}, D_{i, 1}, D_{i, 2}, \ldots, D_{i, \mathcal{T}}, X_i\right)\right\}_{i=1}^n$$

where $D_{i, t}=1$ if unit i is treated in period t , and 0 otherwise

- $G_{i, g}=1$ if unit i is first treated at time g , and zero otherwise (”Treatment start-time dummies”)

- C=1 is a ”never-treated” comparison group (not required, though)

- Staggered treatment adoption: $D_{i, t}=1 \Longrightarrow D_{i, t+1}=1, fort=1,2, \ldots, \mathcal{T}$.

- No Treatment Anticipation (in the paper we relax this a bit):

$$\mathbb{E}\left[Y_t(g) \mid X, G_g=1\right]=\mathbb{E}\left[Y_t(0) \mid X, G_g=1\right] \text { a.s.. }$$

for all $g \in \mathcal{G}, t \in 1, \ldots, \mathcal{T}$ such that $\underbrace{t<g}_{\text{"pre-treatment periods"}}$.

- Generalized propensity score uniformly bounded away from 1:

$$p_{g, t}(X)=P\left(G_g=1 \mid X, G_g+\left(1-D_t\right)\left(1-G_g\right)=1\right) \leq 1-\epsilon \text { a.s. }$$

- Parameter of interest:

$$\operatorname{ATT}(g, t)=\mathbb{E}\left[Y_t(g)-Y_t(0) \mid G_g=1\right], \text { for } t \geq g .$$

Parallel trend assumption based on a ”never treated” group

Assumption (Conditional Parallel Trends based on a ”never-treated” group)
For each $t \in \{2, \ldots, \mathcal{T}\}, g \in \mathcal{G}$ such that $t \geq g$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, C = 1\right] \text{ a.s. } .$$

Parallel Trends based on not-yet treated groups

Assumption (Conditional Parallel Trends based on ”Not-Yet-Treated” Groups)
For each $(s, t) \in \{2, \ldots, \mathcal{T}\} \times \{2, \ldots, \mathcal{T}\}, g \in \mathcal{G}$ such that $t \geq g, s \geq t$

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) \mid X, D_s = 0, G_g = 0\right] \text{ a.s..}$$

- In the case where covariates do not play a major role into the DiD identification analysis, and one is comfortable using the ”never treated” as comparison group,

$$\mathrm{ATT}_{\text{unc}}^{\text{nev}}(g, t) = \mathbb{E}\left[Y_t - Y_{g-1} \mid G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} \mid C = 1\right].$$

- This looks very similar to the two periods, two-groups DiD result without covariates. - The difference is now we take a ”long difference”.

- If one prefers to use the ”not-yet treated” as comparison groups,

$$ATT_{\text{unc}}^{ny}(g, t) = \mathbb{E}\left[Y_t - Y_{g-1} \mid G_g = 1\right] - \mathbb{E} Y_t - Y_{g-1} \mid D_t = 0].$$

Identification results - never treated as comparison group

When covariates play an important role and we use the ”never treated” units as comparison group, we have that

$$\mathrm{ATT}_{dr}^{nev}(g, t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\frac{p_g(X) C}{1 - p_g(X)}}{\mathbb{E}\left[\frac{p_g(X) C}{1 - p_g(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{nev}(X)\right)\right].$$

where $m_{g,t}^{nev}(X) = \mathbb{E}\left[Y_t - Y_{g-1} \mid X, C = 1\right]$

When covariates play an important role and we use the ”not-yet treated” units as comparison group, we have that

$$\mathrm{ATT}_{dr}^{ny}(g, t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}\left[G_g\right]} - \frac{\frac{p_{g,t}(X)(1 - D_t)}{1 - p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1 - D_t)}{1 - p_{g,t}(X)}\right]}\right)\left(Y_t - Y_{g-1} - m_{g,t}^{ny}(X)\right)\right].$$

where $m_{g,t}^{ny}(X) = \mathbb{E}\left[Y_t - Y_{g-1} \mid X, D_t = 0, G_g = 0\right]$

Summarizing ATT (g, t)

- We propose taking weighted averages of the ATT(g, t) of the form:

$$\sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} w_{gt} ATT(g, t)$$

- The two simplest ways of combining ATT (g, t) across g and t are, assuming no-anticipation,

$$\theta_M^O := \frac{2}{\mathcal{T}(\mathcal{T} - 1)} \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} \, \mathrm{ATT}(g, t)$$

and

$$\theta_W^O := \frac{1}{\kappa}\sum_{g=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}} 1\{g \leq t\} ATT(g,t) P(G = g \mid C \neq 1)$$

- Problem: They "overweight" units that have been treated earlier.

Event-study / dynamic treatment effects

- The effect of a policy intervention may depend on the length of exposure to it.

- Average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly e time periods

$$\theta_D(e) = \sum_{g=2}^{\mathcal{T}} 1\{g + e \leq \mathcal{T}\} ATT(g, g+e) P(G = g \mid G + e \leq \mathcal{T}, C \neq 1)$$

Cohort-heterogeneity

- Average effect of participating in the treatment that units in group g experienced:

$$\theta_S(g) = \frac{1}{\mathcal{T} - g + 1}\sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t)$$

8 RIPW

Two-way fixed effect (TWFE) regression model and estimator

TWFE model : $\underbrace{Y_{it}}_{\text{outcome}} = \underbrace{\alpha_i}_{\text{unit FE}} + \underbrace{\lambda_t}_{\text{time FE}} + \underbrace{\tau}_{\text{effect}} \cdot \underbrace{W_{it}}_{\text{treatment}} + \beta \cdot \underbrace{X_{it}}_{\text{covariates}} + \epsilon_{it}$

TWFE estimator : $\hat{\tau}_{\text{TWFE}} \leftarrow OLS\,(Y_{it} \sim unitdummy + timedummy + W_{it} + X_{it})$

DiD estimator $\Longleftrightarrow TWFE(withT = 2)$

$\checkmark\,\hat{\tau}_{\text{TWFE}}$ is unbiased for τ under the TWFE model

- Biased with heterogeneous treatment effect or violation of parallel trend Borusyak et al '17, Goodman-Bacon '17, de Chaisemartin and d'Haultfoeuille '18, Athey and Imbens '18, Sun and Abraham '18

Has TWFE been fully understood? NO!

- A class of estimands: doubly average treatment effects (DATE) ✓

A new estimator: reshaped inverse probability weighting (RIPW)-TWFE estimator

Valid design-based inference: time- and unit-varying effects (finite population framework) many dependent designs: sampling without replacement, two-stage randomization, ...

Double robustness: RIPW $\xrightarrow{p}$ DATE if either the treatment assignment model is known/well estimated or the TWFE model is correct

8.1 Potential Outcome

Balanced panel: n units and T time periods; fixed T

harder than large T with unmeasured confounders: unit FE can't be consistently estimated

Binary treatment: $\mathbf{W}_i = (W_{i1}, \ldots, W_{iT})\,; \mathbf{W}_i \sim \boldsymbol{\pi}_i$ generalized propensity score

Potential outcomes: $(Y_{it}(1), Y_{it}(0))_{t=1}^{T}$

SUTVA: observed outcome $Y_{it} = Y_{it}\left(W_{it}\right)\left(Y_{it}(1), Y_{it}(0)\right)$ are fixed with essentially no restriction in this part of the talk.

$$\hat{\tau}_{\text{IPW}} \triangleq \arg\min_{\tau} \sum_{i=1}^{n} \sum_{t=1}^{T} \underbrace{\left(Y_{it} - \alpha_i - \lambda_t - W_{it}\tau\right)^2}_{\text{TWFE objective}} \qquad \underbrace{\frac{1}{\pi_i\left(\mathbf{W}_i\right)}}_{\text{generalized propensity score}} \qquad \xrightarrow{p} ?$$

8.2 Example

Transient treatments

$$\begin{aligned} &\mathbf{W}_i \in \{(0,0,0),(0,0,1),(0,1,0),(1,0,0)\} \\ &\hat{\tau}_{\text{IPW}} \xrightarrow{p} \tfrac{1}{3}\tau_1 + \tfrac{1}{3}\tau_2 + \tfrac{1}{3}\tau_3 = \tau_{\text{eq}} \end{aligned} \tag{1}$$

Staggered rollouts

$$\begin{aligned} &\mathbf{W}_i \in \{(0,0,0),(0,0,1),(0,1,1),(1,1,1)\} \\ &\hat{\tau}_{\text{IPW}} \xrightarrow{p} 0.3\tau_1 + 0.4\tau_2 + 0.3\tau_3 \end{aligned} \tag{2}$$

Theorem (Arkhangelsky, Imbens, L., and Luo '21) Under regularity conditions (overlap, limited dependence, bounded moments), as $n \to \infty$,

$$\hat{\tau}_{\text{IpW}} \xrightarrow{p} \sum_{t=1}^{T} \xi_t \tau_t \tag{3}$$

where $\mathbb{S} = \bigcup_i \operatorname{Supp}\left(\mathbf{W}_i\right)$ and

$$\xi_t \propto \eta_t\left(1 - \eta_t\right), \qquad \text{where } \eta_t = \frac{|w \in \mathbb{S} : w_t = 1|}{|\mathbb{S}|} \tag{4}$$

don's restrict the heterogeneity pattern of τ or y_{it}

Transient treatments

$$W_{i1} + W_{i2} + \ldots + W_{iT} \leq 1$$

$$\hat{\tau}_{\text{IPW}} \xrightarrow{p} \frac{1}{T}\sum_{t=1}^{T}{}^{1}\, t = \tau_{\text{eq}}$$

Staggered rollouts

$$W_{i1} \leq W_{i2} \leq \ldots \leq W_{iT}$$

$$\hat{\tau}_{\text{IFW}} \xrightarrow{p} \sum_{t=1}^{T} \frac{(T+1-t)t}{\sum_{t=1}^{T}(T+1-t)t}\tau_t$$

What if we want DATE with pre-specified weights (e.g., τ_∞) ?

8.3 Reshaped IPW estimator

Given a data-independent distribution Π on $\mathbb{S}=\bigcup_i \operatorname{Supp}\left(\mathbf{W}_i\right)$:

$$\text{RIPW estimator: } \hat{\tau}_{\text{RIPW}}(\Pi) \triangleq \underset{\tau}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T\left(Y_{it}-\alpha_i-\lambda_t-W_{it} \tau\right)^2 \frac{\Pi\left(\mathbf{W}_i\right)}{\pi_i\left(\mathbf{W}_i\right)}$$

The IPW-TWFE estimator is the RIPW-TWFE estimator with $\Pi \sim \operatorname{Unif}(\mathbb{S})$✓ When $\pi_i=\Pi$,

the RIPW-TWFE estimator reduces to the TWFE estimator For what Π does $\hat{\tau}_{\text{RPPW}}(\Pi) \xrightarrow{p} \tau_{\text{DATE}}(\xi)$?

how to choose ipw estimator that the our ipw aspirator will converge to a date with user specified weights

DATE equation

Theorem (Arkhangelsky, Imbens, L., and Luo '21)

Given $\mathbb{S}$ and Π with $\operatorname{Supp}(\Pi)=\mathbb{S}, \hat{\tau}_{\text{TWFE}} \xrightarrow{p} \tau_{\text{DATE}}(\xi)$ if and "only if"

$$\mathbb{E}_{\mathbf{W} \sim \Pi}\left[\left(\operatorname{diag}(\mathbf{W})-\xi \mathbf{W}^{\top}\right) J\left(\mathbf{W}-\mathbb{E}_{\mathbf{W} \sim \Pi}[\mathbf{W}]\right)\right]=0 \quad \text{(DATE equation)},$$

where $J=I-\mathbf{1}_T \mathbf{1}_T^{\top} / T$.

Only depends on $\mathbb{S}$

D Quadratic equations on $(\Pi(w): w \in \mathbb{S})$ with linear constraints (simplex, positivity)

- Closed-form solutions exist in many examples (DiD, cross-over, staggered rollouts, transient, ...)

✓Generic solver based on nonlinear programming (BFGS algorithm)

Transient treatments

$$\mathbf{W}_i \in\{(0,0,0),(0,0,1),(0,1,0),(1,0,0)\}$$

$$\begin{aligned}&(\Pi(0,0,0), \Pi(0,0,1), \Pi(0,1,0), \Pi(1,0,0)) \\ &=\lambda \cdot(1,0,0,0)+(1-\lambda) \cdot\left(0, \tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right)\end{aligned}$$

$\lambda \in(0,1), \quad Unif$is a solution

Staggered rollouts

$$\mathbf{W}_i \in\{(0,0,0),(0,0,1),(0,1,1),(1,1,1)\}$$

$$\begin{aligned}&(\Pi(0,0,0), \Pi(0,0,1), \Pi(0,1,1), \Pi(1,1,1)) \\ &=\lambda \cdot\left(\tfrac{2}{9}, \tfrac{1}{3}, 0, \tfrac{4}{9}\right)+(1-\lambda) \cdot\left(\tfrac{4}{9}, 0, \tfrac{1}{3}, \tfrac{2}{9}\right)\end{aligned}$$

$\lambda \in(0,1), \quad Unif$ is NOT a solution

An interpretation of DATE equation

RIPW estimator: $\hat{\tau}_{\text{RIPW}}(\Pi) \triangleq \underset{\tau}{\arg\min} \sum_{i=1}^n \sum_{t=1}^T\left(Y_{it}-\alpha_i-\lambda_t-W_{it} \tau\right)^2 \frac{\Pi\left(\mathbf{W}_i\right)}{\pi_i\left(\mathbf{W}_i\right)}$

When $\boldsymbol{\pi}_i=\boldsymbol{\Pi}, \hat{\tau}_{\text{TWFE}}=\hat{\tau}_{\text{RIPW}}(\boldsymbol{\Pi}) \xrightarrow{p} \tau_{\text{DATE}}(\xi)$

DATE equation gives all completely randomized experiments for which TWFE "works"!

8.4 Doubly Robust

RIPW estimators with covariates

Covariates: $\mathbf{X}_i = (X_{i1}, \ldots, X_{iT})$ (satisfying a latent ignorability assumption)

Use $\mathbf{X}_i$ to fit an assignment model $\hat{\boldsymbol{\pi}}_i(\cdot)$:

Staggered rollouts: duration models (e.g., Cox proportional hazard model)

General designs: discrete Markov model, conditional logit model ...

Use $\mathbf{X}_i$ to fit an outcome model $\hat{\mathbf{m}}_i = (\hat{m}_{i1}, \ldots, \hat{m}_{iT})$ for effects varying with units and time

Under TWFE $Y_{it} = \alpha_i + \lambda_t + m_{it} + \epsilon_{it}$ where $m_{it} = X_{it}^{\top}\beta$, then $\hat{m}_{it} = X_{it}^{\top}\hat{\beta}_{\text{TWFE}}$

- No need to estimate FE; crucial since α_i cannot be consistently estimated for fixed T

$$\hat{\tau}(\Pi) \triangleq \arg\min_{\tau} \sum_{i=1}^{n} \sum_{t=1}^{T} (\underbrace{(Y_{it} - \hat{m}_{it})}_{\text{modified outcome}} -\alpha_i - \lambda_t - W_{it}\tau)^2 \frac{\Pi(\mathbf{W}_i)}{\hat{\pi}_i(\mathbf{W}_i)}$$

$$\hat{\tau}(\Pi) \triangleq \arg\min_{\tau} \sum_{i=1}^{n} \sum_{t=1}^{T} (\underbrace{(Y_{it} - \hat{m}_{it})}_{\text{regression adjustment}} -\alpha_i - \lambda_t - W_{it}\tau)^2 \frac{\overbrace{\Pi(\mathbf{W}_i)}_{\text{assignment modeling}}}{\hat{\pi}_i(\mathbf{W}_i)}$$

Double robustness: RIPW^p → $DATE$ if either the assignment model is well estimated or the TWFE model is correct

Fundamentally different from the double robustness discussed in Sant'Anna and Zhao ('20)

- When the assignment model is well estimated, they still require (conditional) parallel trends

- Their outcome model is more general than ours

our method when assignment model is correct,we don't need other conditional parallel and we can allow all the units to be non-dentically distributed, we only consider standard two-way model.

9 Fedearted casual effect

9.1 动机

在很多情况下，同一种处理方法（治疗药物或政策等）会被应用于不同的环境中，例如不同医院使用同一种医疗方案或者医疗药物，政府对不同地区实施相同的政策干预等等，而每个应用环境的数据都是单独收集和存储的。如果可能的话，将不同环境的数据集中起来估计处理效应往往更具优势（例如，当任何一个数据集的样本量太小而无法获得精确的估计时，使用集中的数据集可以得到更准确稳健的估计）。然而，现实中常存在一些限制因素阻碍数据的收集与整合（例如，法律限制、隐私问题、专有利益或竞争壁垒等）。因此，开发一套分析工具使得我们可以在不实际汇集数据的情况下获得数据整合的优势便十分具有实用价值。那些在数据集之间只共享汇总级信息 (summarylevel information) 的方法被称为“联邦”学习 (federated learning) 方法。2021 年 Xiong 等人 [1] 针对因果推断问题结合联邦学习方法提出了联邦因果推断方法。该方法允许跨数据集异质的处理效应和结果模型，并调整处理组样本和控制组样本之间协变量分布的不平衡。

9.2 参数模型介绍

9.2.1 参数模型及模型假设

假设有 K 个数据集, 其中 K 是有限的。对于每个数据集 $k \in \{1, \cdots, K\}$, 有 n_k 个观测 $\left(Y_i^{(k)}, W_i^{(k)}, \mathbf{X}_i^{(k)}\right) \in \mathcal{Y} \times \{0,1\} \times \mathcal{X}_k$, 每个观测独立同分布, 来自某个分布 $\mathbb{P}^{(k)}$ 。记 $n_{\text{pool}} = \sum_{i=1}^{K} n_k$ 为总的观测数量, $i \in \{1, \cdots, n_k\}$ 标记数据集 k 的个体, $\mathbf{X}_i^{(k)}$ 表示数据集 k 中长度为 d_k 的协变量向量, $Y_i^{(k)}$ 表示感兴趣的结果变量, $W_i^{(k)}$ 表示处理分配变量。对于每个数据集 k 而言, 协变量向量的长度 d_k 可以不同, 即使 d_k 相同, 协变量向量自身包含的协变量也可以不同。根据 Neyman-Rubin 的潜在结果模型 [3] 与 Imbens 和 Rubin 的个体处理值稳定假设 (stable unit treatment value assumption, SUTVA) , 令 $Y_i(1)$ 表示个体 i 如果接受干预将得到的潜在结果, $Y_i(0)$ 表示个体如果没有接受干预将得到的潜在结果。对于每个数据集 k , 假定可忽略性假设 (ignorability assumption, 也称为 nonunmesure confounders assumption 成立,

假设 2.1 (可忽略性)

$$\left\{Y_i^{(k)}(0), Y_i^{(k)}(1)\right\} \perp\!\!\!\perp W_i^{(k)} \mid \mathbf{X}_i^{(k)}, \forall i \in n, \forall k \in K$$

并且对于倾向得分 $e^{(k)}(\mathbf{x}) = \Pr\left(W_i^{(k)} = 1 \mid \mathbf{X}_i^{(k)} = \mathbf{x}\right)$ 的重叠假设 (overlap assumption) 成立,

假设 2.2 (重叠)

对某个 $\epsilon > 0$, 有 $\epsilon < e^{(k)}(\mathbf{x}) < 1 - \epsilon, \forall \mathbf{x} \in \mathcal{X}_k$.

那么, 对于每个数据集 k , 可以定义平均处理效应 (average treatment effect, ATE), 记为 $\tau_{\text{ate}}^{(k)}$, 以及处理组的平均处理效应 (average treatment effect on the treated, ATT), 记为 $\tau_{\text{att}}^{(k)}$, 有如下形式,

$$\tau_{\text{ate}}^{(k)} := \mathbb{E}\left[Y_i^{(k)}(1) - Y_i^{(k)}(0)\right], \quad \tau_{\text{att}}^{(k)} := \mathbb{E}\left[Y_i^{(k)}(1) - Y_i^{(k)}(0) \mid W_i^{(k)} = 1\right].$$

本节主要关注以下条件 (2.1) 和条件 (2.2) 所述的参数化结果模型和参数化倾向模型。参数化结果模型在医学领域中被广泛应用, 例如, 用于流行病学研究中估计优势比的逻辑回归, 以及用于评估医疗费用的广义线性模型 (generalized linear model, GLM)。为了估计倾向模型, 最常见的方法之一是使用逻辑模型 (例如, Imbens 和 Rubin)。此外, 估计的参数结果模型和 (或) 倾向模型可以作为估计 ATE 和 ATT 的输入。

条件 2.1 (参数结果模型) 对于任意数据集 k , 在给定 $\mathbf{x}$ 与 w 下, 结果变量 y 的条件密度函数服从一个参数模型, 记为 $f_0(y \mid \mathbf{x}, w, \boldsymbol{\beta})$, 其中参数真值记为 $\boldsymbol{\beta}_0^{(k)}$ 。

条件 2.2 (参数倾向性评分模型) 对任意数据集 k , 在给定 $\mathbf{x}$ 下接受干预的条件概率 $\Pr(w = 1 \mid \mathbf{x})$ 服从一个参数模型, 记为 $e_0(\mathbf{x}, \gamma)$, 其中参数真值记为 $\gamma_0^{(k)}$.

给定条件 (2.1) 和条件 (2.2), 我们可以通过最大化 (加权) 似然函数来估计结果模型和倾向模型。由于参数模型 $f_0(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e_0(\mathbf{x}, \gamma)$ 是先验未知的, 在估计结果模型和倾向模型时选择的分布族, 分别记为 $f(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e(\mathbf{x}, \boldsymbol{\gamma})$, 可能包含也可能不包含真实模型结构 $f_0(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e_0(\mathbf{x}, \gamma)$ 。联邦估计量考虑到了模型误设的可能性, 并且在模型误设的情况下, 联邦估计仍然可以得到合并数据集中特定参数 (例如 ATE 或 ATT) 的一致估计.

9.2.2 MLE

在参数化的结果模型下，我们定义在联合数据上，基于协变量和处理分配的结果的对数似然函数为：

$$\boldsymbol{\ell}_{n_{\text{pool}}}(\boldsymbol{\beta})=\sum_{k=1}^{D}\underbrace{\sum_{i=1}^{n_k}\log f\left(Y_i^{(k)}\mid\mathbf{X}_i^{(k)},W_i^{(k)},\boldsymbol{\beta}\right)}_{\boldsymbol{\ell}_{n_k}(\boldsymbol{\beta})},$$

其中 $\boldsymbol{\ell}_{n_k}(\boldsymbol{\beta})$ 是数据集 k 上的对数似然函数。假设 $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{cb}}$ 是最大化对数似然函数 $\boldsymbol{\ell}_{n_{\text{pool}}}(\boldsymbol{\beta})$ 的解，那么 $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{cb}}$ 是 $\boldsymbol{\beta}^{\text{cb}}$ 的估计量。

极大似然估计量是下面优化问题的解向量 $\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{cb}}$,

$$\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{cb}}=\arg\max_{\boldsymbol{\beta}}\boldsymbol{\ell}_{n_{\text{pool}}}(\boldsymbol{\beta}).$$

在参数倾向模型（条件 (2.2)）下, 可以类似地使用 MLE 估计倾向模型中的参数,

$$\boldsymbol{\ell}_n(\boldsymbol{\gamma})=\textstyle\sum_{i=1}^{n}\log\left[e\left(\mathbf{X}_i,\gamma\right)^{W_i}\left(1-e\left(\mathbf{X}_i,\boldsymbol{\gamma}\right)\right)^{1-W_i}\right]$$
$$\hat{\gamma}_{\text{mle}}=\arg\max_{\gamma}\boldsymbol{\ell}_n(\boldsymbol{\gamma}).$$

9.2.3 基于 IPW-MLE 的模型参数

在结果模型中，估计参数的另一种方法是使用 IPW-MLE，通过倒数倾向得分调整对数似然函数，以估计在数据非随机缺失情况下的总体均值。

$$\boldsymbol{\ell}_{n_{\text{pool}}}(\boldsymbol{\beta},\hat{e})=\sum_{k=1}^{D}\underbrace{\sum_{i=1}^{n_k}\varpi_{i,\hat{e}}^{(k)}\log f\left(Y_i^{(k)}\mid\mathbf{X}_i^{(k)},W_i^{(k)},\boldsymbol{\beta}\right)}_{\ell_{n_k}(\boldsymbol{\beta},\hat{e})},$$

其中下标”ê” 是在联合数据上估计的倾向得分的简写，$\boldsymbol{\ell}_{n_k}(\boldsymbol{\beta},\hat{e})$ 是数据集 k 上加权的对数似然函数，$\varpi_{i,\hat{e}}^{(k)}$ 是单位 i 的权重，可以通过以下方式计算：

$$\varpi_{i,\hat{e}}^{(k)}=\begin{cases}W_i^{(k)}/\hat{e}\left(\mathbf{X}_i^{(k)}\right)+\left(1-W_i^{(k)}\right)/\left(1-\hat{e}\left(\mathbf{X}_i^{(k)}\right)\right) & ATEweighting\\ W_i^{(k)}+\hat{e}\left(\mathbf{X}_i^{(k)}\right)\left(1-W_i^{(k)}\right)/\left(1-\hat{e}\left(\mathbf{X}_i^{(k)}\right)\right) & ATTweighting.\end{cases}$$

设 $\hat{\boldsymbol{\beta}}_{\text{ipw}cble}^{cb}$ 是最大化加权对数似然函数 $\boldsymbol{\ell}_{n_{\text{pool}}}(\boldsymbol{\beta},\hat{e})$ 的估计量。该估计量可用于估计处理组和对照组的结果，并形成 ATE 和 ATT 的双重稳健估计量

当给定密度函数 $f\left(Y_i\mid\mathbf{X}i,Wi,\boldsymbol{\beta}\right)$ 后，我们可以使用参数化的条件结果模型 $\mu_{(w)}\left(\mathbf{X}i,\boldsymbol{\beta}\right)$ 来表示 $\mathbb{E}\left[Yi\mid\mathbf{X}i,Wi=w\right]$，其中 $\boldsymbol{\beta}$ 是参数向量。通过最大似然估计方法，我们可以使用 $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ 来估计平均处理效应 τ_{ate}。

具体而言，我们可以通过以下公式来计算估计的平均处理效应 $\hat{\tau}_{\text{ate}}$：

$$\hat{\tau}_{\text{ate}}=\frac{1}{n}\sum_{i=1}^{n}\left[\mu_{(1)}\left(\mathbf{X}_i,\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}\right)-\mu_{(0)}\left(\mathbf{X}_i,\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}\right)\right]$$

这种估计方法具有双重稳健性质即使结果模型或倾向模型中存在模型误设，只要其中一个模型正确，对 τ_{ate} 的估计仍然是一致的。如果结果模型的假设正确，那么不论倾向模型的假设正确与否，$\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ 都是条件 $\boldsymbol{\beta}_0$ 的一致估计量。因此，$\frac{1}{n}\sum_i \mu_{(w)}\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_{\text{ipw-mle}}\right)$ 也是 $\mathbb{E}\left[Y_i(w)\right]$ 的一致估计，从而使得估计量 $\hat{\tau}_{\text{ate}}$ 也是一致的。

另一方面，如果结果模型存在误设而倾向模型被正确指定，那么 $\hat{\boldsymbol{\beta}}_{\text{ipw-mle}}$ 是 $\boldsymbol{\beta}$ 的一致估计量，其中 $\boldsymbol{\beta}$ 是最大化 $\mathbb{E}\left[\log f\left(Y_i \mid \mathbf{X}_i, W_i, \boldsymbol{\beta}\right)\right]$ 的唯一解，但与参数真值 $\boldsymbol{\beta}_0$ 可能不相等。如果条件结果模型满足 $\mathbb{E}\left[\mu(w)\left(\mathbf{X}_i, \boldsymbol{\beta}\right)\right] = \mathbb{E}\left[Y_i(w)\right]$，那么 $\hat{\tau}$ate 仍然是一致的。特别地，如果 $\mu(w)\left(\mathbf{X}i, \boldsymbol{\beta}^*\right)$ 是关于 $\mathbf{X}_i$ 和 w 的带截距项的线性函数或 logistic 函数，那么 $\hat{\tau}_{\text{ate}}$ 也是一致的

9.2.4 AIPW

我们可以使用 AIPW 估计量在联合数据上估计 ATE：

$$\hat{\tau}_{\text{ate}}^{\text{cb}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \cdot \underbrace{\frac{1}{n_k}\sum_{i=1}^{n_k} \hat{\phi}\left(\mathbf{X}_i^{(k)}, W_i^{(k)}, Y_i^{(k)}\right)}_{\hat{\tau}_{\text{ate}}^{(k)}},$$

其中按样本量加权平均的 ATE 可以写为各数据集的加权平均，其中 $\hat{\phi}(\cdot)$ 是在联合数据上的估计得分，定义为：

$$\hat{\phi}(\mathbf{x}, w, y) = \hat{\mu}_{(1)}(\mathbf{x}) - \hat{\mu}_{(0)}(\mathbf{x}) + \frac{w}{\hat{e}(\mathbf{x})}\left(y - \hat{\mu}_{(1)}(\mathbf{x})\right) - \frac{(1-w)}{1-\hat{e}(\mathbf{x})}\left(y - \hat{\mu}_{(0)}(\mathbf{x})\right),$$

其中，$\hat{\mu}_{(1)}(\mathbf{x})$ 和 $\hat{\mu}_{(0)}(\mathbf{x})$ 是在联合数据上估计的条件处理组和对照组结果模型。如果估计对象是 ATT，我们也可以使用最开始的公式，但是估计得分 $\hat{\phi}(\cdot)$ 的定义如下：

$$\hat{\phi}(\mathbf{x}, w, y) = w\left(y - \hat{\mu}(1)(\mathbf{x})\right) - \frac{\hat{e}(\mathbf{x})(1-w)}{1-\hat{e}(\mathbf{x})}\left(y - \hat{\mu}(0)(\mathbf{x})\right)$$

AIPW 具有两个显著的特性：双重稳健性（Robins 等，1994）和半参数效率。

9.3 联邦学习的条件

我们先在联邦中需要考虑的条件，以获得目标参数的有效点和方差估计量。

-**Condition3（已知倾向得分）** 对于所有数据集，真实的倾向得分是已知并被使用的。

当真实的倾向得分是已知并被使用时，我们在联邦 IPW-MLE 中不需要联合倾向模型。

-**Condition4（稳定的倾向模型）** 倾向模型中的协变量集合和参数在所有数据集中是相同的，即对于任何 j 和 k，$\gamma_0^{(j)} = \gamma_0^{(k)}$。

-**Condition5（稳定的结果模型）** 结果模型中的协变量集合和参数在所有数据集中是相同的，即对于任何 j 和 k，$\boldsymbol{\beta}0^{(j)} = \boldsymbol{\beta}0^{(k)}$。

-**Condition6（稳定的协变量分布）** 协变量集合及其联合分布在所有数据集中是相同的。即对于任意两个数据集 j 和 k，$d_j = d_k \ \mathbb{P}^{(j)}(\mathbf{x}) = \mathbb{P}^{(k)}(\mathbf{x})$。

这里需要注意，在违反条件 4、5 或 6 的情况下，我们将数据集称为“异质的”。如果条件 5 成立（对于条件 4 也是类似的），那么组合数据上的参数 $\boldsymbol{\beta}_0^{\text{cb}}$ 等于任何 k 上的参数 $\boldsymbol{\beta}_0^{(k)}$；否则，我们将

参数 $\boldsymbol{\beta}^{(k)} = \left(\boldsymbol{\beta}\mathrm{s}, \boldsymbol{\beta}_{\mathrm{uns}}^{(k)}\right)$ 分为共享参数 $\boldsymbol{\beta}_{\mathrm{s}}$ 和数据集特定参数 $\boldsymbol{\beta}_{\mathrm{uns}}^{(k)}$（对于任何 k ），并将组合数据上的参数定义为 $\boldsymbol{\beta}^{\mathrm{cb}} = \left(\boldsymbol{\beta}\mathrm{s}, \boldsymbol{\beta}_{\mathrm{uns}}^{(1)}, \boldsymbol{\beta}_{\mathrm{uns}}^{(2)}, \cdots, \boldsymbol{\beta}_{\mathrm{uns}}^{(D)}\right)$。这里根据是否违反条件进行的联邦学习估计的方法是不同的。

Hessian Weighting

Hessian 加权方法用于估计结果模型中的目标参数: $\boldsymbol{\beta}_0^{\mathrm{cb}}$ 和倾向性模型中的目标参数 $\boldsymbol{\gamma}_0^{\mathrm{cb}}$，其定义如下：

$$\hat{\boldsymbol{\beta}}^{\mathrm{fed}} = \left(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}\right)^{-1} \left(\sum_{k=1}^{D} \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} \hat{\boldsymbol{\beta}}^{(k)}\right), \quad \text{where } \hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)} = \frac{\partial^2 \boldsymbol{\ell}_{n_k}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}{\partial \boldsymbol{\beta}^{(k)} \left(\partial \boldsymbol{\beta}^{(k)}\right)^{\top}}$$

对于倾向性模型，我们只需将 $\hat{\boldsymbol{\beta}}^{(k)}$ 替换为 $\hat{\boldsymbol{\gamma}}^{(k)}$ ，将 $\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$ 替换为 $\hat{\mathbf{H}}_{\gamma}^{(k)}$ 即可.

Sample Size Weighting

样本量加权用于获得方差估计量 (详见表 2、3 和 4)，并用于估计不稳定倾向或结果模型下的 ATE 和 ATT。对于某些一般的标量或矩阵 $\mathbf{M}$，我们称样本大小权重为：

$$\mathbf{M}^{\mathrm{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\mathrm{pool}}} \mathbf{M}^{(k)}, \quad \text{where } n_{\mathrm{pool}} = \sum_{k=1}^{D} n_k.$$

Inverse Variance Weighting

在稳定的倾向模型和结果模型下，使用逆方差加权 (IVW) 估计 ATE 和 ATT 及其方差。对于某些一般的点估计量 $\hat{\boldsymbol{\nu}}$，我们将方差逆加权记为:

$$\begin{aligned} \hat{\boldsymbol{\nu}}^{\mathrm{fed}} &= \left(\sum_{k=1}^{D} \left(\mathrm{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1}\right)^{-1} \left(\sum_{k=1}^{D} \left(\mathrm{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1} \boldsymbol{\nu}^{(k)}\right), \\ \widetilde{\mathrm{Var}}\left(\hat{\boldsymbol{\nu}}^{\mathrm{fed}}\right) &= n_{\mathrm{pool}} \left(\sum_{k=1}^{D} \left(\mathrm{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1}\right)^{-1}, \end{aligned}$$

其中 $\mathrm{Var}(\hat{\boldsymbol{\nu}})$ 是 $\hat{\boldsymbol{\nu}}$ 的方差，$\widetilde{\mathrm{Var}}(\hat{\boldsymbol{\nu}})$ 是 $\mathrm{Var}(\hat{\boldsymbol{\nu}})$ 乘以样本量。

统计学中，逆方差加权 (英语: inverse-variance weighting) 是一种对随机变量测量值进行加权平均的方法。每个随机变量被其方差的倒数加权。该方法可使平均值的方差最小。若随机变量的一系列独立测量值为 y_i ，其方差为 ${\sigma_i}^2$ ，则这些测量值的逆方差加权平均为

$$\hat{y} = \frac{\sum_i y_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}.$$

在所有加权平均方法中，逆方差加权平均的方差最小，为

$$D^2(\hat{y}) = \frac{1}{\sum_i 1/\sigma_i^2}.$$

若各测量值的方差相等，则逆方差加权平均与简单平均相同。逆方差加权通常在元分析中用来整合独立测量的结果。

对于每个类别，我们从倾向和结果模型稳定的简单情况开始。在这种情况下，我们将联邦估计器称为受限联邦估计器。在处理倾向模型和结果模型都稳定的简单情况之后，我们进一步考虑更具挑战

性的情形，其中倾向模型和结果模型至少有一个是不稳定的。针对这种情况，我们引入了无限制联邦估计器，它是基于相应受限联邦估计器的进一步发展。

无限制联邦估计器是一种应对不稳定模型的方法，用于联合估计倾向模型和结果模型。它的设计目的是克服模型不稳定性可能带来的估计偏差和方差问题。相比之下，受限联邦估计器只考虑了模型都稳定的简单情况。

通过采用适当的稳定性条件和估计方法，无限制联邦估计器能够在模型不稳定的情况下提供一致的估计结果。

联邦 MLE

通过使用结果模型来说明联邦最大似然估计（MLE），但联邦 MLE 也可以应用于参数倾向模型。在稳定模型的情况下，我们可以使用限制性联邦 MLE 方法（满足条件（2.4）或条件（2.5））来获得联邦的点估计。

为了得到联邦的点估计，首先我们对每个数据集使用 MLE 估计参数 $\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}$ ，然后利用海塞加权（Hessian weighting）进行联合。

我们以满足最大似然估计的一阶条件为目标，提出了这个联合估计。当我们使用黑森加权时，可以通过以下关键步骤来实现这一目标:

$$\begin{aligned}\frac{\partial \sum_{k=1}^{D} \boldsymbol{\ell}_{n_k}\left(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}\right)}{\partial \boldsymbol{\beta}} &= \sum_{k=1}^{D} \frac{\partial \boldsymbol{\ell}_{n_k}\left(\boldsymbol{\beta}_0\right)}{\partial \boldsymbol{\beta}} + \sum_{k=1}^{D} \mathbf{H}_{\boldsymbol{\beta}}^{(k)}\left(\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} - \boldsymbol{\beta}_0\right) \\ &= \sum_{k=1}^{D} \frac{\partial \boldsymbol{\ell}_{n_k}\left(\boldsymbol{\beta}_0\right)}{\partial \boldsymbol{\beta}} + \sum_{k=1}^{D} \mathbf{H}_{\boldsymbol{\beta}}^{(k)}\left(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} - \boldsymbol{\beta}_0\right) \quad \left(\text{Hessian weighting of } \hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}}\right) \\ &= \sum_{k=1}^{D} \frac{\partial \boldsymbol{\ell}_{n_k}\left(\hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}\right)}{\partial \boldsymbol{\beta}} = 0 \quad \left(\text{gradient at } \hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} \text{ is zero for all } k\ \right)\end{aligned}$$

联邦 MLE 的估计值为:

$$\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} = \left(\sum_{k=1}^{K} \hat{H}_{\boldsymbol{\beta}}^{(k)}\right)^{-1}\left(\sum_{k=1}^{K} \hat{H}_{\boldsymbol{\beta}}^{(k)} \hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)}\right), \quad \text{其中} \hat{H}_{\boldsymbol{\beta}}^{(k)} = \frac{\partial^2 \ell_{n_k}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}{\partial \boldsymbol{\beta}^{(k)}\left(\partial \boldsymbol{\beta}^{(k)}\right)^{\top}}.$$

联邦方差估计量的构造是基于单个数据集的模型误设稳健的方差形式，即 $\mathbf{V}_\beta = \mathbf{A}_\beta^{-1}\mathbf{B}_\beta\mathbf{A}_\beta^{-1}$。联邦方差估计包括三个步骤:

第一步，在每个数据集上估计 $\mathbf{A}_\beta^{(k)}$ $\mathbf{B}_\beta^{(k)}$。

第二步，使用样本量加权方式分别联合每个数据集的 $\hat{\mathbf{A}}_\beta^{(k)}$ 和 $\hat{\mathbf{B}}_\beta^{(k)}$ $\hat{\mathbf{A}}_\beta^{\text{fed}}$ 和 $\hat{\mathbf{B}}_\beta^{\text{fed}}$

$$\hat{\mathbf{A}}_{\boldsymbol{\beta}}^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \hat{\mathbf{A}}_{\boldsymbol{\beta}}^{(k)} \quad \text{and} \quad \hat{\mathbf{B}}_{\boldsymbol{\beta}}^{\text{fed}} = \sum_{k=1}^{D} \frac{n_k}{n_{\text{pool}}} \mathbf{B}_{\boldsymbol{\beta}}^{(k)}.$$

第三步,将估计的 $\mathbf{A}\beta^{\text{fed}}$ $\mathbf{B}\beta^{\text{fed}}$ 代入计算公式中,得到联邦方差估计量 $\hat{\mathbf{V}}_\beta^{\text{fed}} = \left(\hat{\mathbf{A}}_\beta^{\text{fed}}\right)^{-1} \hat{\mathbf{B}}_\beta^{\text{fed}} \left(\hat{\mathbf{A}}_\beta^{\text{fed}}\right)^{-1}$。

其中联邦方差估计量的设计对结果模型的误设是稳健的。通过上述步骤,我们可以获得联邦 MLE 的点估计以及联邦方差估计量。这种方法在联邦学习中的实际应用中非常有用，因为它能够处理模型不稳定性和模型误设，从而提供准确可靠的估计结果。

不稳定模型的无限制联邦 MLE(违反条件 4/ 5)

当结果模型不稳定时，只要存在一些跨数据集共享的参数，通过联合估计结果模型可以提高共享参数的估计精度。我们将每个数据集 k 的参数集合划分为共享参数集 $\boldsymbol{\beta}_s$ 和每个数据集特定的参数集 $\boldsymbol{\beta}_{\text{uns}}^{(k)}$ ，即 $\boldsymbol{\beta}^{(k)} = \left(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{\text{uns}}^{(k)}\right)$ 。其中，共享参数集 $\boldsymbol{\beta}_s$ 包括感兴趣的处理分配变量 W 的系数。

非限制性联邦估计量建立在限制性联邦估计量的基础上，旨在跨数据集联合共享参数 $\hat{\boldsymbol{\beta}}_s$ ，同时保持每个数据集特定参数集 $\hat{\boldsymbol{\beta}}_{\text{uns}}^{(k)}$ 的原始值。其中 $\left(\hat{\boldsymbol{\beta}}_s, \hat{\boldsymbol{\beta}}_{\text{uns}}^{(k)}\right)$ 是通过 MLE 估计得到的。为了实现这一目标，我们定义合并数据的参数为 $\boldsymbol{\beta}^{\text{bm}} = \left(\boldsymbol{\beta}_s, \boldsymbol{\beta}_{\text{uns}}^{(1)}, \boldsymbol{\beta}_{\text{uns}}^{(2)}, \cdots, \boldsymbol{\beta}_{\text{uns}}^{(K)}\right)$ ，并用零进行填充（zero-pad）来使 $\hat{\boldsymbol{\beta}}^{(k)}$ 的维度与 $\boldsymbol{\beta}^{\text{bm}}$ 一致，记为 $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}$ ，如公式（2-8）所示。类似地，对 $\hat{\mathbf{H}}\beta^{(k)}, \hat{\mathbf{A}}_{\boldsymbol{\beta}}^{(k)}$ 和 $\hat{\mathbf{B}}_{\beta}^{(k)}$ 进行零填充，使其维度与合并数据相应的矩阵一致，分别记为 $\hat{\mathbf{H}}_{\beta}^{\text{pad},(k)}, \hat{\mathbf{A}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$ 和 $\hat{\mathbf{B}}_{\beta}^{\text{pad},(k)}$ 。

接下来的非限制性联邦 MLE 估计方法与限制性联邦 MLE 的过程类似，使用 $\hat{\boldsymbol{\beta}}^{\text{pad},(k)}, \hat{\mathbf{H}}_{\beta}^{\text{pad},(k)}, \hat{\mathbf{A}}_{\boldsymbol{\beta}}^{\text{pad},(k)}$ 和 $\hat{\mathbf{B}}_{\beta}^{\text{pad},(k)}$ 来联合点估计量和方差估计量:

$$\hat{\boldsymbol{\beta}}^{\text{pad},(k)} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\text{s}}^{(k)} \\ \mathbf{0}_{S_1^{k-1}} \\ \hat{\boldsymbol{\beta}}_{\text{uns}}^{(k)} \\ \mathbf{0}_{S_{k+1}^K} \end{pmatrix}, \quad \hat{\mathbf{H}}_{\beta}^{\text{pad},(k)} = \begin{pmatrix} \hat{\mathbf{H}}_{\boldsymbol{\beta},\text{s,s}}^{(k)} & \mathbf{0}_{s_0 \times S_1^{k-1}} & \hat{\mathbf{H}}_{\boldsymbol{\beta},\text{s,uns}}^{(k)} & \mathbf{0}_{s_0 \times S_{k+1}^K} \\ \mathbf{0}_{S_1^{k-1} \times s_0} & \mathbf{0}_{S_1^{k-1} \times S_1^{k-1}} & \mathbf{0}_{S_1^{k-1} \times S_k^k} & \mathbf{0}_{S_1^{k-1} \times S_{k+1}^K} \\ \hat{\mathbf{H}}_{\boldsymbol{\beta},\text{uns,s}}^{(k)} & \mathbf{0}_{S_k^k \times S_1^{k-1}} & \hat{\mathbf{H}}_{\boldsymbol{\beta},\text{ uns, uns}}^{(k)} & \mathbf{0}^K_{S_k^k \times S_{k+1}^K} \\ \mathbf{0}_{S_{k+1}^K \times s_0} & \mathbf{0}_{S_{k+1}^K \times S_1^{k-1}} & \mathbf{0}_{S_{k+1}^K \times S_k^k} & \mathbf{0}^K_{S_{k+1}^K \times S_{k+1}^K} \end{pmatrix},$$

其中 s_0 和 s_k 分别是 $\boldsymbol{\beta}_s$ $\boldsymbol{\beta}_{\text{uns}}^{(k)}$ 的维数, 而 $S_{j_1}^{j_2} = \sum_{j=j_1}^{j_2} s_j$, 对于 $x, y \in \{\text{s}, \text{uns}\}, \hat{\mathbf{H}}_{\boldsymbol{\beta},x,y}^{(k)} = \frac{\partial^2 \ell_{n_k}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}{\partial \boldsymbol{\beta}_x^{(k)} \left(\partial \boldsymbol{\beta}_y^{(k)}\right)^\top}, \mathbf{0}_{n_1 \times n_2}$ 表示一个 $n_1 \times n_2$ 的零矩阵。

这里注意：即使某些参数是稳定的，也可以将它们视为特定于数据集的参数。这种方法不影响联合评估的一致性然而，随着组合数据上的参数数量的增加，联邦估计器的效率会弱于使用最简洁规范的估计器。

这个表格也适用于倾向性模型。第二行对应于条件 5。$\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$ 表示估计的海森矩阵。$\mathbf{A}_{\boldsymbol{\beta}}$ 和 $\mathbf{B}_{\boldsymbol{\beta}}$ 在表格 1 中有定义。$\hat{\mathbf{H}}_{\boldsymbol{\beta}}^{(k)}$ 随着样本量 n_k 的增加而增加，而 $\mathbf{A}_{\boldsymbol{\beta}}$ 和 $\mathbf{B}_{\boldsymbol{\beta}}$ 不会。对于一般的向量或矩阵 $\mathbf{x}$，$\mathbf{x}^{\text{pad}}$ 表示在 $\mathbf{x}$ 后面填充了零。

稳定模型的受限联邦 IPW-MLE(条件 4 和条件 5)

在估计结果模型的参数时，IPW-MLE 使用了倾向得分，因此与联邦 MLE 相比，需要额外考虑联邦 IPW-MLE 中倾向模型的参数联合。这里我们关注的是在单个数据集上，倾向模型和结果模型都是通过 MLE 估计得到的情况。然后，我们可以利用联邦 MLE 的构造和渐近特性来建立具有理论保证的联邦 IPW-MLE 方法。

首先，考虑联邦 IPW-MLE 的点估计量。从倾向模型开始。如果倾向模型是通过 MLE 估计得到的，那么我们可以利用联邦 MLE 方法来联合各个数据集的倾向模型参数。然而，如果真实的倾向得分是已知且被使用的（满足条件 (2.3)），那么可以跳过倾向模型的参数联合步骤。接下来是估计结果模型中的联合系数。首先，使用联合的（或真实的）倾向得分来估计每个单独数据集中结果模型的系数。然后，利用海塞加权法将各数据集中结果模型的系数估计值结合起来。

因此，联邦 IPW-MLE 的步骤包括倾向模型参数的联合估计（如果倾向模型是通过 MLE 估计得到的）和结果模型中系数的联合估计。在倾向模型的参数联合步骤中，可以使用联邦 MLE 方法;

在结果模型系数的联合估计步骤中，使用联合的或真实的倾向得分来估计各数据集中的系数，并通过海塞加权法进行合并。这样可以得到具有理论保证的联邦 IPW-MLE 方法，与联邦 MLE 相似，该联邦点估计器满足 IPW-MLE 的一阶条件。

同样我们需要考虑方差估计量：

联邦 IPW-MLE 的方差估计量基于单个数据集的 IPW-MLE 的渐近方差 V_β 构建。当真实的倾向得分已知且被使用（满足条件 (2.3)），则方差估计量为 $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}\mathbf{D}_{\boldsymbol{\beta},\varpi}\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$。

如果倾向得分是通过 MLE 估计得到的，

则对于 ATE，方差估计量为 $V_\beta = \mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}\left(\mathbf{D}_{\beta,\varpi} - \mathbf{C}_{\beta,\varpi}\mathbf{V}_\gamma\mathbf{C}_{\boldsymbol{\beta},\varpi}^\top\right)\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$ ，

对于 ATT,方差估计量为 $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{A}_{\beta,\varpi}^{-1}\left(\mathbf{D}_{\beta,\varpi} - \mathbf{C}_{\boldsymbol{\beta},\varpi,1}\mathbf{V}_\gamma\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^\top - \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_\gamma\mathbf{C}_{\boldsymbol{\beta},\varpi,1}^\top + \mathbf{C}_{\boldsymbol{\beta},\varpi,2}\mathbf{V}_\gamma\mathbf{C}_{\boldsymbol{\beta},\varpi,2}^\top\right)\mathbf{A}_{\boldsymbol{\beta},\varpi}^{-1}$
。

其中 $\mathbf{V}_\gamma = \mathbf{A}_\gamma^{-1}\mathbf{D}_\gamma\mathbf{A}_\gamma^{-1} \cdot \mathbf{A}_{\beta,\varpi}, \mathbf{D}_{\beta,\varpi}, \mathbf{C}_{\beta,\varpi}, \mathbf{C}_{\boldsymbol{\beta},\varpi,1}, \mathbf{C}_{\beta,\varpi,2}, \mathbf{A}_\gamma, \mathbf{B}_\gamma$ 的定义可见表)。

为了得到联合的方差估计, 我们需要先估计每个数据集的 $\mathbf{A}_{\beta,\varpi}$ 和 $\mathbf{D}_{\beta,\varpi}$ (如果有需要的话, 还需估计 $\mathbf{C}_{\beta,\varpi}$, (或者 $\mathbf{C}_{\beta,\varpi,1}, \mathbf{C}_{\boldsymbol{\beta},\varpi,2}$), $\mathbf{A}_{\boldsymbol{\gamma}}$, 和 $\mathbf{B}_\gamma$), 然后使用样本量加权法跨数据集联合 $\mathbf{A}_{\beta,\varpi}$ 以及 $\mathbf{D}_{\beta,\varpi}$ (如果有需要的话, 还有 $\mathbf{C}_{\boldsymbol{\beta},\varpi}$ (或者 $\mathbf{C}_{\boldsymbol{\beta},\varpi,1}, \mathbf{C}_{\boldsymbol{\beta},\varpi,2}$) , $\mathbf{A}_\gamma$ 和$\mathbf{B}_\gamma$)。

因此，联邦 IPW-MLE 的步骤包括估计每个数据集的 $\mathbf{A}_{\beta,\varpi}$ 和 $\mathbf{D}_{\beta,\varpi}$, 然后通过样本量加权方法将它们联合起来得到联合方差估计量。这样可以得到具有理论保证的联邦 IPW-MLE 的方差估计量。

不稳定模型的无限制联邦 IPW-MLE(违反条件 4 或 5)

与不受限制的联邦 MLE 类似，在非限制性联邦 IPW-MLE 中，倾向模型和（或）结果模型中的参数被划分为共享参数和数据集特定参数。非限制性联邦 IPW-MLE 通过联合估计各数据集的共享参数来提高估计精度，同时保持数据集特定参数在联合中保持其原始值。我们为合并数据集指定了模型参数，其维度通常高于单个数据集的维度。接下来，我们对每个数据集的估计参数和相关矩阵进行适当的零填充，以匹配合并数据的参数维度。然后，非限制性联邦 IPW-MLE 的过程与限制性联邦 IPW-MLE 相同，但使用了零填充的参数和矩阵。

这种非限制性联邦 IPW-MLE 的方法允许在联合估计中利用各数据集的共享信息，以提高参数估计的准确性。通过合并数据集并联合估计共享参数，我们可以更好地捕捉整体数据的特征和模式。同时，数据集特定参数保持其原始值，这允许我们保留数据集间的差异性和个体特征。

第二行和第三行对应于条件 4 和条件 5。”是”或”否”表示解决方案是否随条件是否满足而变化。$\mathbf{A}_{\boldsymbol{\beta},\varpi}, \mathbf{D}_{\boldsymbol{\beta},\varpi}, \mathbf{C}_{\boldsymbol{\beta},\varpi}, \mathbf{C}_{\boldsymbol{\beta},\varpi,1}, \mathbf{C}_{\boldsymbol{\beta},\varpi,2}, \mathbf{A}_\gamma,$ **和**$\mathbf{B}_\gamma$ 的定义可以在表 1 中找到。

当估计倾向性模型时（条件 3 不满足），系数联合过程对于所有情景是相同的，但在使用真实倾向性时（条件 3 满足）会简化。方差联合过程取决于是否使用真实倾向性以及是否使用 ATE 或 ATT 加权。

对于一般的向量或矩阵 $\mathbf{x}$，$\mathbf{x}^{\mathrm{pad}}$ 表示在 $\mathbf{x}$ 后面填充了零。

稳定模型和稳定协变量分布的受限 AIPW 估计量 (条件 4,5 和 6 成立)

由于 AIPW 估计器同时使用结果模型和倾向模型，我们需要联合倾向模型和结果模型。当协变量分布、倾向模型和结果模型稳定时，我们建议使用限制联合 AIPW。该方法分为三个步骤：

首先，我们使用联邦 MLE 得到联邦倾向模型和联邦结果模型。

其次，我们使用 AIPW 与联合倾向和结果模型来估计每个数据集上的 ATE。

最后，我们通过对每个数据集上估计的 ATE 进行方差逆加权来得到联邦 ATE，

$$\hat{\boldsymbol{\nu}}^{\text{fed}} = \left(\sum_{k=1}^{D}\left(\operatorname{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1}\right)^{-1}\left(\sum_{k=1}^{D}\left(\operatorname{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1}\boldsymbol{\nu}^{(k)}\right),$$

$$\widetilde{\operatorname{Var}}\left(\hat{\boldsymbol{\nu}}^{\text{fed}}\right) = n_{\text{pool}}\left(\sum_{k=1}^{D}\left(\operatorname{Var}\left(\hat{\boldsymbol{\nu}}^{(k)}\right)\right)^{-1}\right)^{-1},$$

需要注意的是，在稳定的协变量分布和稳定的倾向和结果模型下，所有数据集的 ATE 和 ATE 的渐近方差是相同的。在这种情况下，我们可以应用任何加权方案来将估计的 ATE 组合在一起。

这里我们选择 IVW 方法是因为它在所有加权方案中方差最小。

不稳定模型或不稳定协变量分布的无限制 AIPW 估计 (条件 4,5 或 6 都不满足)

当倾向模型、结果模型或协变量分布不稳定时，不同数据集的 ATE 可能不相同。针对这种情况，我们建议使用无限制的联合 AIPW。

对于无限制的估计器，我们首先估计每个数据集上的 ATE 及其渐近方差，然后使用样本量加权来组合这些估计的 ATE 和方差。

$$\hat{\tau}_{\text{aipw}}^{\text{fed}} = \sum_{k=1}^{D}\frac{n_k}{n_{\text{pool}}}\hat{\tau}_{\text{aipw}}^{(k)} \quad \hat{\mathbf{V}}_{\tau}^{\text{fed}} = \sum_{k=1}^{D}\frac{n_k}{n_{\text{pool}}}\hat{\mathbf{V}}_{\tau}^{(k)}$$

其中 $\hat{\tau}_{\text{aipw}}^{(k)}$ 是数据集 k 上的估计 ATE , $\hat{\mathbf{V}}_{\tau}^{(k)}$ 是 $\hat{\tau}_{\text{aipw}}^{(k)}$ 的方差估计. 通过考虑每个数据集的样本量，我们能够更准确地反映每个数据集对于联邦 ATE 估计的贡献。

具体而言，我们可以针对每个数据集估计 ATE 和其渐近方差，并根据数据集的样本量进行加权组合。这样可以更好地平衡不同数据集的贡献，并考虑到样本量较大的数据集对于联邦 ATE 估计的影响更大。

这个联合 AIPW 估计器非常通用，具有多个优势。首先，它对于倾向模型或结果模型的错误描述具有鲁棒性，即使这些模型在不同的数据集之间存在变化或误差，估计器也能保持有效性。其次，它允许倾向模型和/或结果模型在数据集之间的任意变化，因此具有很大的灵活性。第三，该方法不要求一种方法来联合估计倾向和结果模型，因此可以使用不同的机器学习方法（如随机森林）来估计倾向得分。

然而，需要权衡的是，在协变量分布稳定且倾向模型和结果模型稳定的情况下，无限制的估计器相对于限制的估计器来说可能效率较低。这是因为限制的估计器可以利用模型的稳定性假设来获得更高效的估计。因此，在稳定的时候选择该模型还是上一个模型需要综合考虑模型的稳定性和估计的效率。