

Federate Report

19331053 纪传宇

2023 年 8 月 3 日

目录

1	DRDID	1
1.1	Basic assumptions of DD	1
1.2	doubly robust	2
1.3	TWFE specification	2
2	DML	4
2.1	Neyman 正交	5
3	DML and DRDID	5
4	RIPW	7
4.1	Potential Outcome	7
4.2	Example	8
4.3	Reshaped IPW estimator	8
4.4	Doubly Robust	9

1 DRDID

1.1 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

$$E[Y_1^0 - Y_0^0 | X, D = 1] = E[Y_1^0 - Y_0^0 | X, D = 0]$$

Assumption 3: Common support or overlap

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

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 .

1.2 doubly robust

Outcome regression

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

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

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

Inverse probability weighting

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

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

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

Combined:

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 $m_{d,t}^p(x) \equiv E[Y_t | D = d, X = x]$, $d, t = 0, 1$.

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

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

where, for a generic g ,

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

Theorem 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.;

1.3 TWFE specification

Here's the TWFE specification:

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

Just add in covariates then right?

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

Collecting terms

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

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

$$E[Y_1^1 - Y_1^0 | X, D = 1] = E[Y_1^1 - Y_1^0 | D = 1]$$

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.

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

Eliminating terms, we get:

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

Assumption 5 and 6

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

$$E[Y_1 - Y_0 | D = d, X] = E[Y_1 - Y_0 | D = d]$$

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.

2 DML

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

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

其中 Y 是模型的 Outcome, D 是因果模型的治疗。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 :

$$\begin{aligned} E[(Y - D\theta_0 - g_0(Z)) D] &= 0 \\ E[(Y - D\theta_0)(D - E[D | Z])] &= 0 \\ E[((Y - E[Y | Z]) - (D - E[D | Z])\theta_0)(D - E[D | Z])] &= 0 \end{aligned}$$

- (1) - Regression adjustment;
- (2) - "propensity score adjustment"
- (3) - Neyman-orthogonal (semi-parametrically efficient under homoscedasticity).

Based on (1),

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

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

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 $\theta_0 = \frac{\text{cov}(D, Y - g_0(X))}{\text{var}(D)} - \underbrace{\frac{\text{cov}(D, U)}{\text{var}(D)}}_{\neq 0}$ 。

Consider estimation based on (3)

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

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

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

2.1 Neyman 正交

τ_0 可以被视为以下估计方程的解：

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

其中, φ 是一个已知的评分函数 (score function), \hat{g}_0 是干扰参数 g_0 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:

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

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

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

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：

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

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 的矩条件对于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 $\tilde{\tau}_0$ 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。

3 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）：

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

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

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

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 $\hat{\tau}_0$ ，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 $\hat{\tau}_0$ 中不会产生影响。然而，这简化了 $\hat{\tau}_0$ 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

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

$\hat{\mu}_1(X_i)$: estimation of $\mathbb{E}[Y | X, T = 1]$

$\hat{\mu}_0(X_i)$: estimation of $\mathbb{E}[Y | X, T = 0]$

$\hat{e}(X_i)$: estimation of $\mathbb{E}[T = 1 | X]$

DRDID ATT:

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

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

DiD estimator $\iff TWFE(\text{with } T = 2)$

✓ $\hat{\tau}_{\text{TWFE}}$ 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: $\text{RIPW} \xrightarrow{p} \text{DATE}$ if either the treatment assignment model is known/well estimated or the TWFE model is correct

4.1 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: $\mathbf{W}_i = (W_{i1}, \dots, W_{iT})$; $\mathbf{W}_i \sim \pi_i$ generalized propensity score

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

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

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

4.2 Example

Transient treatments

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

Staggered rollouts

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

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

$$\hat{\tau}_{\text{IPW}} \xrightarrow{p} \sum_{t=1}^T \xi_t \tau_t \quad (3)$$

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

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

don't restrict the heterogeneity pattern of τ or y_{it}

Transient treatments

$$\begin{aligned} W_{i1} + W_{i2} + \dots + W_{iT} &\leq 1 \\ \hat{\tau}_{\text{IPW}} &\xrightarrow{p} \frac{1}{T} \sum_{t=1}^T \tau_t = \tau_{\text{eq}} \end{aligned}$$

Staggered rollouts

$$\begin{aligned} W_{i1} &\leq W_{i2} \leq \dots \leq W_{iT} \\ \hat{\tau}_{\text{IPW}} &\xrightarrow{p} \sum_{t=1}^T \frac{(T+1-t)t}{\sum_{t=1}^T (T+1-t)t} \tau_t \end{aligned}$$

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

4.3 Reshaped IPW estimator

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

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

The IPW-TWFE estimator is the RIPW-TWFE estimator with $\Pi \sim \text{Unif}(\mathbb{S})$ ✓ 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 $\text{Supp}(\Pi) = \mathbb{S}$, $\hat{\tau}_{\text{TWFE}} \xrightarrow{P} \tau_{\text{DATE}}(\xi)$ if and "only if"

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

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

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

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

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

Staggered rollouts

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

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

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

An interpretation of DATE equation

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

When $\pi_i = \Pi$, $\hat{\tau}_{\text{TWFE}} = \hat{\tau}_{\text{RIPW}}(\Pi) \xrightarrow{P} \tau_{\text{DATE}}(\xi)$

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

4.4 Doubly Robust

RIPW estimators with covariates

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

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

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 $\hat{\mathbf{m}}_i = (\hat{m}_{i1}, \dots, \hat{m}_{iT})$ 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

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

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

Double robustness: $\text{RIPW}^p \rightarrow \text{DATE}$ 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.

5 Fedearated casual effect

5.1 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 (summary level 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。

5.2 参数模型介绍

5.2.1 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 $k \in \{1, \dots, K\}$ ，有 n_k 个观测 $(Y_i^{(k)}, W_i^{(k)}, \mathbf{X}_i^{(k)}) \in \mathcal{Y} \times \{0, 1\} \times \mathcal{X}_k$ ，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， $i \in \{1, \dots, n_k\}$ 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)

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

并且对于倾向得分 $e^{(k)}(\mathbf{x}) = \Pr(W_i^{(k)} = 1 \mid \mathbf{X}_i^{(k)} = \mathbf{x})$ 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,

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

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 $f_0(y \mid \mathbf{x}, w, \beta)$, 其中参数真值记为 $\beta_0^{(k)}$ 。

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 $f_0(y \mid \mathbf{x}, w, \beta)$ 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 $f(y \mid \mathbf{x}, w, \beta)$ 和 $e(\mathbf{x}, \gamma)$, 可能包含也可能不包含真实模型结构 $f_0(y \mid \mathbf{x}, w, \beta)$ 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。

5.2.2 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:

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

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

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

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

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，

$$\begin{aligned} \ell_n(\gamma) &= \sum_{i=1}^n \log \left[e(\mathbf{X}_i, \gamma)^{W_i} (1 - e(\mathbf{X}_i, \gamma))^{1-W_i} \right] \\ \hat{\gamma}_{\text{mle}} &= \arg \max_{\gamma} \ell_n(\gamma). \end{aligned}$$

5.2.3 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。

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

其中下标“ \hat{e} ”是在联合数据上估计的倾向得分的简写， $\ell_{n_k}(\beta, \hat{e})$ 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：

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

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

当给定密度函数 $f(Y_i | \mathbf{X}_i, W_i, \beta)$ 后，我们可以使用参数化的条件结果模型 $\mu_{(w)}(\mathbf{X}_i, \beta)$ 来表示 $\mathbb{E}[Y_i | \mathbf{X}_i, W_i = w]$ ，其中 β 是参数向量。通过极大似然估计方法，我们可以使用 $\hat{\beta}_{\text{ipw-mle}}$ 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 $\hat{\tau}_{\text{ate}}$ ：

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

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， $\hat{\beta}_{\text{ipw-mle}}$ 都是条件 β_0 的一致估计量。因此， $\frac{1}{n} \sum_i \mu_{(w)}(\mathbf{X}_i, \hat{\beta}_{\text{ipw-mle}})$ 也是 $\mathbb{E}[Y_i(w)]$ 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 $\hat{\beta}_{\text{ipw-mle}}$ 是 β 的一致估计量，其中 β 是最大化 $\mathbb{E}[\log f(Y_i | \mathbf{X}_i, W_i, \beta)]$ 的唯一解，但与参数真值 β_0 可能不相等。如果条件结果模型满足 $\mathbb{E}[\mu(w)(\mathbf{X}_i, \beta)] = \mathbb{E}[Y_i(w)]$ ，那么 $\hat{\tau}_{\text{ate}}$ 仍然是一致的。特别地，如果 $\mu(w)(\mathbf{X}_i, \beta^*)$ 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

5.2.4 AIPW

我们可以使用 AIPW 估计量在联合数据上估计 ATE:

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

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:

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

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:

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

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

5.3 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 , $\beta_0^{(j)} = \beta_0^{(k)}$ 。

-**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 β_0^{cb} 等于任何 k 上的参数 $\beta_0^{(k)}$; 否则, 我们将参数 $\beta^{(k)} = (\beta_s, \beta_{\text{uns}}^{(k)})$ 分为共享参数 β_s 和数据集特定参数 $\beta_{\text{uns}}^{(k)}$ (对于任何 k), 并将组合数据上的参数定义为 $\beta^{\text{cb}} = (\beta_s, \beta_{\text{uns}}^{(1)}, \beta_{\text{uns}}^{(2)}, \dots, \beta_{\text{uns}}^{(D)})$ 。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: β_0^{cb} 和倾向性模型中的目标参数 γ_0^{cb} , 其定义如下:

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

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

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

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 $\hat{\nu}$ ，我们将方差逆加权记为：

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

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

其中 $\text{Var}(\hat{\nu})$ 是 $\hat{\nu}$ 的方差， $\widetilde{\text{Var}}(\hat{\nu})$ 是 $\text{Var}(\hat{\nu})$ 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 σ_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 $\hat{\beta}_{\text{mle}}^{(k)}$ ，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：

$$\begin{aligned} \frac{\partial \sum_{k=1}^D \ell_{n_k}(\hat{\beta}_{\text{mle}}^{\text{fed}})}{\partial \beta} &= \sum_{k=1}^D \frac{\partial \ell_{n_k}(\beta_0)}{\partial \beta} + \sum_{k=1}^D \mathbf{H}_{\beta}^{(k)} (\hat{\beta}_{\text{mle}}^{\text{fed}} - \beta_0) \\ &= \sum_{k=1}^D \frac{\partial \ell_{n_k}(\beta_0)}{\partial \beta} + \sum_{k=1}^D \mathbf{H}_{\beta}^{(k)} (\hat{\beta}_{\text{mle}}^{(k)} - \beta_0) \quad (\text{Hessian weighting of } \hat{\beta}_{\text{mle}}^{\text{fed}}) \\ &= \sum_{k=1}^D \frac{\partial \ell_{n_k}(\hat{\beta}_{\text{mle}}^{(k)})}{\partial \beta} = 0 \quad (\text{gradient at } \hat{\beta}_{\text{mle}}^{(k)} \text{ is zero for all } k) \end{aligned}$$

联邦 MLE 的估计值为：

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

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 $\hat{\mathbf{B}}_{\beta}^{\text{fed}}$

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

第三步，将估计的 $\hat{\mathbf{A}}_{\beta}^{\text{fed}} \hat{\mathbf{B}}_{\beta}^{\text{fed}}$ 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 β_s 和每个数据集特定的参数集 $\beta_{\text{uns}}^{(k)}$ ，即 $\beta^{(k)} = (\beta_s, \beta_{\text{uns}}^{(k)})$ 。其中，共享参数集 β_s 包括感兴趣的分配变量 \mathbf{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 $\hat{\beta}_s$ ，同时保持每个数据集特定参数集 $\hat{\beta}_{\text{uns}}^{(k)}$ 的原始值。其中 $(\hat{\beta}_s, \hat{\beta}_{\text{uns}}^{(k)})$ 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 $\beta^{\text{bm}} = (\beta_s, \beta_{\text{uns}}^{(1)}, \beta_{\text{uns}}^{(2)}, \dots, \beta_{\text{uns}}^{(K)})$ ，并用零进行填充 (zero-pad) 来使 $\hat{\beta}^{(k)}$ 的维度与 β^{bm} 一致，记为 $\hat{\beta}^{\text{pad},(k)}$ ，如公式 (2-8) 所示。类似地，对 $\hat{\mathbf{H}}_{\beta}^{(k)}, \hat{\mathbf{A}}_{\beta}^{(k)}$ 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 $\hat{\mathbf{H}}_{\beta}^{\text{pad},(k)}, \hat{\mathbf{A}}_{\beta}^{\text{pad},(k)}$ 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 $\hat{\beta}^{\text{pad},(k)}$, $\hat{\mathbf{H}}_{\beta}^{\text{pad},(k)}$, $\hat{\mathbf{A}}_{\beta}^{\text{pad},(k)}$ 和 $\hat{\mathbf{B}}_{\beta}^{\text{pad},(k)}$ 来联合点估计量和方差估计量:

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

其中 s_0 和 s_k 分别是 β_s $\beta_{\text{uns}}^{(k)}$ 的维数, 而 $S_{j_1}^{j_2} = \sum_{j=j_1}^{j_2} s_j$, 对于 $x, y \in \{s, \text{uns}\}$, $\hat{\mathbf{H}}_{\beta,x,y}^{(k)} = \frac{\partial^2 \ell_{n_k}(\hat{\beta}^{(k)})}{\partial \beta_x^{(k)} (\partial \beta_y^{(k)})^\top}$, $\mathbf{0}_{n_1 \times n_2}$ 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。 $\hat{\mathbf{H}}_{\beta}^{(k)}$ 表示估计的海森矩阵。 \mathbf{A}_{β} 和 \mathbf{B}_{β} 在表格 1 中有定义。 $\hat{\mathbf{H}}_{\beta}^{(k)}$ 随着样本量 n_k 的增加而增加, 而 \mathbf{A}_{β} 和 \mathbf{B}_{β} 不会。对于一般的向量或矩阵 \mathbf{x} , \mathbf{x}^{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 $\mathbf{V}_{\beta} = \mathbf{A}_{\beta,\omega}^{-1} \mathbf{D}_{\beta,\omega} \mathbf{A}_{\beta,\omega}^{-1}$ 。

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 $V_{\beta} = \mathbf{A}_{\beta,\omega}^{-1} \left(\mathbf{D}_{\beta,\omega} - \mathbf{C}_{\beta,\omega} \mathbf{V}_{\gamma} \mathbf{C}_{\beta,\omega}^{\top} \right) \mathbf{A}_{\beta,\omega}^{-1}$,

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

。

其中 $\mathbf{V}_\gamma = \mathbf{A}_\gamma^{-1} \mathbf{D}_\gamma \mathbf{A}_\gamma^{-1} \cdot \mathbf{A}_{\beta, \varpi}, \mathbf{D}_{\beta, \varpi}, \mathbf{C}_{\beta, \varpi}, \mathbf{C}_{\beta, \varpi, 1}, \mathbf{C}_{\beta, \varpi, 2}, \mathbf{A}_\gamma, \mathbf{B}_\gamma$ 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 $\mathbf{C}_{\beta, \varpi, 1}, \mathbf{C}_{\beta, \varpi, 2}$), \mathbf{A}_γ , 和 \mathbf{B}_γ), 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 $\mathbf{C}_{\beta, \varpi}$ (或者 $\mathbf{C}_{\beta, \varpi, 1}, \mathbf{C}_{\beta, \varpi, 2}$), \mathbf{A}_γ 和 \mathbf{B}_γ)。

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。 $\mathbf{A}_{\beta, \varpi}, \mathbf{D}_{\beta, \varpi}, \mathbf{C}_{\beta, \varpi}, \mathbf{C}_{\beta, \varpi, 1}, \mathbf{C}_{\beta, \varpi, 2}, \mathbf{A}_\gamma$, 和 \mathbf{B}_γ 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}^{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,

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

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

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。