Federate Report

19331053 纪传宇

2023 年 6 月 16 日

目录

1 动机

在许多情况下，我们会对不同环境中的人群采用相同的政策或者治疗方法，但数据却分别存储在各自的环境中，即每个地方的数据有可能因为环境的不同而不同。然而，由于某些数据集中的样本量太小，无法得到各个地方的准确的治疗效应估计，因此如果有可能的话，将不同环境的数据进行组合往往会很有益。然而，由于法律限制、隐私顾虑、专有利益或竞争壁垒等原因，个人级数据不能盲目的组合。

我们今天所介绍的分析工具，可以在不合并个人级数据的情况下获得数据组合的好处。实现这一目标的方法被称为“联邦学习”方法。

1.1 例子

Koenecke 等人在 2021 年研究了两个独立的医疗索赔数据集，MarketScan 和 Optum。这两个数据集明显不同：Optum 的数据包含更多年长的患者，并覆盖了比 MarketScan 更多的年份。他们从这两个数据集中找到了证据，表明接触　受体阻滞剂（一类常用的处方药物）可以降低急性呼吸窘迫患者不良结果的风险。然而，现有的联邦学习方法不足以在考虑处理组和对照组之间以及两个独立数据集之间人群异质性的情况下对药物效应进行推断。

1.2 方法

我们首先介绍了两种类别的联邦推断方法，用于解决参数估计问题，特别是处理效应的估计。这两种方法分别是基于逆概率加权最大似然估计器（IPW-MLE）和增强逆概率加权估计器（AIPW）。

我们关注 IPW-MLE 和 AIPW 有两个主要原因：

(1) 这两种估计方法都使用倾向得分来平衡处理组和对照组之间的协变量分布。倾向得分方法能够通过估计处理组和对照组之间的概率差异，消除由于选择性引起的潜在偏差。

(2)IPW-MLE 和 AIPW 都具有双重鲁棒性，即它们对于倾向得分模型和结果模型中的一个被错误规定的情况下仍然能够保持稳健性。这种稳健性使得这两种方法能够在模型假设不完全满足的情况下提供一致的估计结果。

其次，我们指出参数模型具有解释性和简易性的优势，但在处理复杂数据时可能受到限制。非参数回归模型提供了更自由形式和较少约束的建模方法，但在高维数据情况下可能遭受“维度灾难”，估计变得困难不稳定。半参数回归模型结合了参数模型的解释性和非参数模型的灵活性，解决了维度灾难问题，提供了处理复杂数据的综合建模框架。因此，我们可以利用半参数估计器对单个数据集进行估计，同时也可以应用于多个数据集的联邦估计。在这里，我们介绍了偏回归模型和双重机器学习（DML）方法，用于估计单个数据集的处理效应（ATE）及其方差。最后，我们使用了反方差加权 IVW 和样本量加权两种加权方法来汇总每个数据集的估计值。

此外，传统的回归模型通过引入交互项来分析处理效应的异质性，但难以处理大量交互项和复杂交互关系。近年来，倾向得分方法逐渐成为研究处理效应异质性的重要工具，通过研究处理效应如何随个体倾向得分的变化而变化来解决这一问题。然而，倾向得分方法存在模型和系数的不确定性。结合机器学习方法和因果推断分析，可以更好地探索处理效应的异质性，并提供更准确和全面的结果，推动处理效应研究的进展。因此，我们使用了 Chernozhukov 等人提出的双重机器学习方法和 Athey 等人提出的广义随机森林方法（GRF）的非参数方法，来计算单个数据集中的因果效应估计值及其方差的估计值，并使用反方差加权法和样本量加权法两种加权方法来汇总每个数据集的估计值。

2 符号说明

Notation	Explanation
D	数据集个数
n_k	$k \in \{1, \cdots, D\}$ 的样本个数
$\left(\mathbf{X}_i^{(k)}, Y_i^{(k)}, W_i^{(k)}\right) \in \mathcal{X}_k \times \mathbb{R} \times \{0,1\}$	每个个体的特征即观测值从某个分布 $\mathbb{P}^{(k)}$ 取得
$i \in \{1, \cdots, n_k\}$	观察个体即样本.
$\mathbf{X}_i^{(k)}$	d_k 个观测协变量的向量
$Y_i^{(k)}$	感兴趣的结果变量
$W_i^{(k)}$	treatment 的分配
$n_{\text{pool}} = \sum_{i=1}^{D} n_k$	观测样本的总数

两个最重要的假设:

在 Neyman-Rubin 潜在结果模型和稳定单元处理值假设（Imbens 和 Rubin，2015）下，设 $Y_i^{(k)}(1)$ 为第 i 个个体接受处理时的结果，而 $Y_i^{(k)}(0)$ 为相反情况下的结果。

对于每个数据集 k ，假设以下**标准无混淆性假设**（Rosenbaum 和 Rubin，1983）成立：

$$\left\{Y_i^{(k)}(0), Y_i^{(k)}(1)\right\} \perp W_i^{(k)} \mid \mathbf{X}_i^{(k)}$$

对于倾向得分 $e^{(k)}(\mathbf{x}) = \text{pr}\left(W_i^{(k)} = 1 \mid \mathbf{X}_i^{(k)} = \mathbf{x}\right)$ 有**重叠即有观测值假设**成立:

$$\eta < e^{(k)}(\mathbf{x}) < 1 - \eta \quad \forall \mathbf{x} \in \mathcal{X}_k$$

其中 $\eta > 0$

对于每个数据集 k，我们定义平均处理效应（ATE），表示为 $\tau_{\text{ate}}^{(k)}$，以及处理组的平均处理效应（ATT），表示为 $\tau_{\text{att}}^{(k)}$

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

3 构建参数模型

在 MLE 中必须要有参数的密度函数才能进行极大似然估计, 而且在医学应用中普遍使用参数结果模型，例如，在流行病学研究中使用逻辑回归来估计比值比（Sperandei，2014），在临床试验中使

用 Cox 回归进行生存分析（Singh 和 Mukhopadhyay，2011），以及使用广义线性模型（GLM）评估医疗成本（Blough 等，1999 年；Blough 和 Ramsey，2000 年）。

所以我们需要先对所需要估计的系数进行参数化模型的构造，下面将重点放在 **Condition 1** 和 **Condition 2** 中所述的参数结果和倾向模型上。

条件 1（参数化结果模型）：对于任何数据集 k，结果变量 y 在给定 **x** 和 w 的条件下，服从一个参数化模型，记为 $f_0^{(k)}(y \mid \mathbf{x}, w, \boldsymbol{\beta})$，其中 $\boldsymbol{\beta}_0^{(k)}$ 是真实的参数值。

条件 2（参数化倾向模型）：对于任何数据集 k，处理概率 $pr(w = 1 \mid \mathbf{x})$ 在给定 **x** 的条件下，服从一个参数化模型，记为 $e_0^{(k)}(\mathbf{x}, \boldsymbol{\gamma})$，其中 $\gamma_0^{(k)}$ 是真实的参数值。

根据条件 1 和条件 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。由于参数化模型 $f_0^{(k)}(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e_0^{(k)}(\mathbf{x}, \gamma)$ 事先是未知的，因此在估计结果模型和倾向模型时选择的分布族，记为 $f^{(k)}(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e^{(k)}(\mathbf{x}, \boldsymbol{\gamma})$，可能包含或不包含真实结构 $f_0^{(k)}(y \mid \mathbf{x}, w, \boldsymbol{\beta})$ 和 $e_0^{(k)}(\mathbf{x}, \boldsymbol{\gamma})$。

3.1 参数模型在联邦估计的作用

由于我们旨在将不同数据集之间的估计结合, 所以先定义了联邦方法旨在估计的目标参数 (符号中的上标"(k)" 表示使用数据集 k 估计的对象；上标"cb" 表示针对合并的个体级数据的对象；上标"fed" 表示联邦估计器。)

目标参数是在将 D 个数据集的个体数据串联在一起形成的合并数据上定义的。第一组目标参数是真实条件结果密度函数 $f_0^{\mathrm{cb}}(\cdot)$ 上的参数，记为 $\boldsymbol{\beta}_0^{\mathrm{cb}}$，其中 $f_0^{\mathrm{cb}}(\cdot)$ 定义为

$$f_0^{\mathrm{cb}}\left(Y_i^{(k)} \mid \mathbf{X}i^{(k)}, Wi^{(k)}, \boldsymbol{\beta}_0^{\mathrm{cb}}\right) := \prod j = 1^K \left[f_0^{(j)}\left(Y_i^{(j)} \mid \mathbf{X}i^{(j)}, Wi^{(j)}, \boldsymbol{\beta}_0^{(j)}\right)\right]^{\mathbf{1}(j=k)}, \quad \forall k,$$

其中当观测来自数据集 k 时，它等于数据集 k 上的真实条件结果密度。$\boldsymbol{\beta}_0^{\mathrm{cb}}$ 定义为 $\boldsymbol{\beta}_0^{(1)}, \cdots, \boldsymbol{\beta}_0^{(K)}$ 的联合。例如，如果 $\boldsymbol{\beta}_0^{(1)} = \cdots = \boldsymbol{\beta}_0^{(K)}$，那么对于任何 k $\boldsymbol{\beta}_0^{\mathrm{cb}} = \boldsymbol{\beta}_0^{(k)}$；如果 $\boldsymbol{\beta}_0^{(1)}, \cdots, \boldsymbol{\beta}_0^{(K)}$ 完全不同，则 $\boldsymbol{\beta}_0^{\mathrm{cb}} = \left(\boldsymbol{\beta}_0^{(1)}, \cdots, \boldsymbol{\beta}_0^{(K)}\right)$。

第二组目标参数是合并数据上真实倾向度 $e_0^{\mathrm{cb}}(\cdot)$ 的参数，记为 γ_0^{cb}，其中 $e_0^{\mathrm{cb}}(\cdot)$ 定义为

$$e_0^{\mathrm{cb}}\left(W_i^{(k)} \mid \mathbf{X}i^{(k)}, \gamma 0^{\mathrm{cb}}\right) := \prod_{j=1}^{K} \left[e_0^{(j)}\left(W_i^{(j)} \mid \mathbf{X}i^{(j)}, \gamma 0^{(j)}\right)\right]^{\mathbf{1(j=k)}}, \quad \forall k,$$

其中当观测来自数据集 k 时，它等于数据集 k 上的真实倾向度。类似于 $\boldsymbol{\beta}_0^{\mathrm{cb}}$ $\boldsymbol{\gamma}_0^{\mathrm{cb}}$ 定义为 $\boldsymbol{\gamma}_0^{(1)}, \cdots, \gamma_0^{(K)}$ 的联合。

第三组目标参数是合并数据上的平均处理效应 (ATE) 和平均处理效应对待受试者 (ATT)，分别记为 $\tau_{\mathrm{ate}}^{\mathrm{cb}}$ 和 $\tau_{\mathrm{att}}^{\mathrm{cb}}$ ，定义如下：

$$\tau_{\mathrm{ate}}^{\mathrm{cb}} := \sum_{k=1}^{D} p_k \tau_{\mathrm{ate}}^{(k)}, \quad \tau_{\mathrm{att}}^{\mathrm{cb}} := \sum_{k=1}^{D} p_k \tau_{\mathrm{att}}^{(k)}$$

其中，$\tau_{\mathrm{ate}}^{\mathrm{cb}}$ 和 $\tau_{\mathrm{att}}^{\mathrm{cb}}$ 是根据 p_k 权重下的 $\tau_{\mathrm{ate}}^{(k)}$ 和 $\tau_{\mathrm{att}}^{(k)}$ 的平均值，其中 p_k 是数据集 k 中观测的人口比例。无论样本大小如何，$\tau_{\mathrm{ate}}^{\mathrm{cb}}$ 和 $\tau_{\mathrm{att}}^{\mathrm{cb}}$ 都不依赖于样本大小。

4 组合个人数据的估计方法

首先我们介绍当个人水平的数据可以合并时如何利用 MLE、IPW-MLE 和 AIPW 估计我们想要的目标参数

4.1 模型参数的 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,

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

4.2 基于 IPW-MLE 的模型参数和平均处理效应估计

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：

$$\varpi_{i,\hat{e}}^{(k)} = \begin{cases} W_i^{(k)} / \hat{e}\left(\mathbf{X}_i^{(k)}\right) + \left(1 - W_i^{(k)}\right) / \left(1 - \hat{e}\left(\mathbf{X}_i^{(k)}\right)\right) & \text{ATE weighting} \\ W_i^{(k)} + \hat{e}\left(\mathbf{X}_i^{(k)}\right)\left(1 - W_i^{(k)}\right) / \left(1 - \hat{e}\left(\mathbf{X}_i^{(k)}\right)\right) & \text{ATT weighting.} \end{cases} \tag{2}$$

设 $\hat{\boldsymbol{\beta}}_{\text{ipw cble}}^{\text{cb}}$ 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 $\hat{\boldsymbol{\beta}}$ipw-mle 来估计平均处理效应 τate。

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

4.3 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。

4.4 半参数偏回归模型

对于每个数据集 k，考虑 Robinson 于 1988 年提出的偏回归模型。该模型可以表示为：

$$\begin{aligned} Y &= W\tau_0 + g_0(X) + U, \quad \mathbb{E}[U \mid X, W] = 0 \\ W &= e_0(X) + Z, \quad \mathbb{E}[Z \mid X] = 0 \end{aligned}$$

其中，Y 是结果变量，W 是感兴趣的处理分配变量（或经济学中的政策变量），X 是其他协变量。U 和 Z 是干扰变量（在此处考虑为标量）。

第一个公式是结果模型，其中 τ_0 是需要进行推断的最主要的参数。如果给定 X 的条件下，W 是外生的，即给定 X 的条件下 W 的分配机制完全随机，那么参数 τ_0 就可以解释为处理效应。第二

个公式反映了混淆情况，即处理分配变量对协变量的依赖。这个公式本身并不是本文所关心的，但它对描述和消除正则化偏差起着重要作用。混杂因素 X 通过函数 $e_0(X)$ 影响处理分配变量 W，通过函数 $g_0(X)$ 影响结果变量。

在许多应用中，协变量向量 X 的维度 d_x 相对于样本量 n 来说是比较大的。为了刻画 d_x 相对于样本量来说并不小的特点，现代分析将 d_x 构建为随着样本量的增加而增加，这使得限制干扰参数 $\eta_0 = (e_0(X), g_0(X))$ 参数空间复杂性的传统假设失效。

在高维数据情境下进行因果推断面临着诸多挑战，其中之一是干扰参数空间的复杂性随着样本量的增加而增加，这使得我们脱离了传统半参数研究中考虑的经典框架。在传统框架中，干扰参数空间的复杂性通常被假设为较小的程度。

特别是在高维数据中，估计参数面临着正则化偏差的问题。正则化是一种常用的方法，用于在高维数据中控制过拟合，通过对模型进行惩罚，减少模型的复杂性。然而，正则化可能导致估计值偏离真实参数值，产生正则化偏差。

为了应对这一问题，双重机器学习（DML）方法采用了 Neyman 正交化技术来消除正则化偏差。Neyman 正交化是一种基于高维向量空间的数学工具，它能够将原始参数估计与正则化项分离，从而减少因正则化而引起的偏差。

这里简单介绍一下正则化偏差的来源：

为了使用机器学习来估计因果效应 τ_0，一种简单直接的方法是构建一个复杂的机器学习估计量 $W\hat{\tau}0 + \hat{g}0(\mathbf{X})$，用于学习回归模型 $W\tau_0 + g_0(\mathbf{X})$。为了清晰起见，假设我们将样本量为 n 的样本随机划分为两部分，以避免过拟合问题：一个样本量为 n_m 的主样本，其中观测个体用 $i \in I$ 表示；一个样本量为 $n - n_m$ 的辅助样本，其中观测个体用 $i \in I^c$ 表示。为了简化问题，暂时假设 $n_m = n/2$，然后我们将讨论更一般的情况，包括不等分割规模和使用多个分割的情况，以实现与使用完整样本进行估计 τ_0 相同的效率。假设 $\hat{g}0$ 是使用辅助样本获得的，而在给定 $\hat{g}0$ 的情况下，通过主样本得到 τ_0 的最终估计为：

$$\hat{\tau}_0 = \left(\frac{1}{nm} \sum_{i \in \mathbf{I}} W_i^2 \right)^{-1} \frac{1}{n_m} \sum_{i \in \mathbf{I}} W_i \left(Y_i - \hat{g}_0 \left(\mathbf{X}i \right) \right)$$

这个估计量 $\hat{\tau}_0$ 通常会以比 $1/\sqrt{n}$ 更慢的速度收敛，即:

$$|\sqrt{n_m} \left(\hat{\tau}0 - \tau 0 \right)| \to_P \infty$$

为了更清晰地说明学习 g_0 的偏差的影响，可以将 τ_0 的估计误差分解为：

$$\sqrt{n_m} \left(\hat{\tau}_0 - \tau_0 \right) = \underbrace{\left(\frac{1}{n_m} \sum_{i \in \mathbf{I}} W_i^2 \right)^{-1} \frac{1}{\sqrt{n_m}} \sum_{i \in \mathbf{I}} W_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n_m} W_i^2 \right)^{-1} \frac{1}{\sqrt{n_m}} \sum_{i \in \mathbf{I}} W_i \left(g_0 \left(\mathbf{X}_i \right) - \hat{g}_0 \left(\mathbf{X}_i \right) \right)}_{:=b}$$

第一项在温和的条件下表现良好，它服从均值为 0、方差为某个 $\bar{\Sigma}$ 的正态分布，即 $a \rightsquigarrow N(0, \bar{\Sigma})$。这意味着在适当的条件下，估计量 a 可以以较快的速度收敛到真实值 τ_0。然而，第二项 b 是正则化偏差项，它在一般情况下不会收敛。

我们可以将 b 写成以下形式：

$$b = \left(\mathbb{E} \left[W_i^2 \right] \right)^{-1} \frac{1}{\sqrt{n_m}} \sum_{i \in \mathbf{I}} e_0 \left(\mathbf{X}_i \right) \left(g_0 \left(\mathbf{X}_i \right) - \hat{g}_0 \left(\mathbf{X}_i \right) \right) + o_P(1)$$

其中，b 是 n_m 个项 $e_0\left(\mathbf{X}i\right)\left(g0\left(\mathbf{X}i\right)-\hat{g}0\left(\mathbf{X}i\right)\right)$ 的和，再除以 $\sqrt{nm}$。这些项的均值不为零，因为在高维或其他高度复杂的环境中，我们必须采用正则化的估计方法（如 LASSO、ridge、boosting 或惩罚性神经网络）才能使信息学习变得可行。正则化方法在控制估计方差的同时，也会引入估计 g_0 的实质性偏差。换言之，为了降低估计方差，我们不可避免地会引入估计的偏差。

具体来说，$\hat{g}0$ 对 $g0$ 的收敛速率通常为 $n^{-\varphi_g}$，其中 $\varphi_g<\frac{1}{2}$。此外，由于权重 W_i 并非随机分配，意味着 $e_0\left(\mathbf{X}_i\right)\neq 0$，所以 b 不会收敛到 0，这导致了公式 (2-12) 的出现。

这些说明了在高维数据情形下进行因果推断时，我们面临的正则化偏差问题。尽管正则化可以控制方差，但我们需要认识到它引入了一定的偏差，从而影响了因果估计的准确性。

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 (2-9)-(2-10) 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，$\check{\tau}_0=\left(\frac{1}{n_m}\sum_{i\in\mathbf{I}}\hat{Z}_iW_i\right)^{-1}\frac{1}{n_m}\sum_{i\in\mathbf{I}}\hat{Z}_i\left(Y_i-\hat{g}_0\left(\mathbf{X}_i\right)\right)$ 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：

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

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。

4.5 非参数交互式回归模型

在这里我们介绍 Chernozhukov 等人提出的非参数模型，并讨论了在存在处理效应异质性的情况下如何联合估计平均处理效应（ATE）。本节首先介绍了交互式回归模型（Interactive Regression Model, IRM），然后利用广义随机森林框架中的因果随机森林方法 [42] 来获得单个数据集中的 ATE 估计。该方法基于 Chernozhukov 等人提出的双重机器学习（DML）框架，并具有双重稳健性，可以在存在处理效应异质性的情况下得到准确的估计。

交互式回归模型 (IRM) 是一种非参数模型，用于估计处理效应。它通过引入交互项来捕捉处理效应的异质性。这种模型的优势在于不需要对处理效应的函数形式进行假设，可以更好地适应实际情况中复杂的异质性。

在单个数据集中，我们利用广义随机森林框架中的因果随机森林方法来估计 ATE。因果随机森林是一种基于机器学习的非参数方法，可以通过组合多个决策树来估计处理效应。这种方法具有灵活性和高预测准确性，能够有效地处理高维协变量和非线性关系。

然后，我们通过样本量加权和反方差加权的方法来联合各个数据集中的估计。样本量加权方法将每个数据集的估计结果按照其样本量进行加权平均，以得到联合的估计结果。反方差加权方法则根据各个数据集的估计结果的方差来进行加权平均，以得到具有更小方差的联合估计结果。

这样的联合估计方法能够综合各个数据集的信息，提高估计的准确性和稳健性。通过将不同数据集的估计结果进行合理的组合，我们可以获得更全面和可靠的处理效应的估计结果。

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：

矩条件：

$$\mathbb{E}\left[\psi\left(D;\tau_0,\eta_0\right)\right]=0$$

正交化条件：

$$\partial_\eta\mathbb{E}\left[\psi\left(D;\tau_0,\eta_0\right)\right]\left[\eta-\eta_0\right]=0$$

这些条件是在估计处理效应时的关键要求。通过满足这些条件，我们可以获得一致性的估计量，并且估计结果具有良好的渐近性质。正交评分的引入使得我们能够通过寻找解 $\tilde{\tau}_0$ 的方程组来获得估计值，并通过矩条件和正交化条件对参数进行推断。

之后我们采用了 Athey 等人在 2019 年提出的广义随机森林（Generalized Random Forest，GRF）方法来估计干扰参数以及 ATE 或 ATT，而不是 DML 中的随机森林方法。GRF 是一种基于随机森林的自适应非参数方法，可以看作是随机森林的一种推广。

与 Breiman 在 2001 年提出的随机森林相比，GRF 保留了随机森林的几个关键要素，如递归划分、子采样和随机分裂选择。然而，GRF 在以下两个方面有所不同：

自适应性加权平均：GRF 放弃了对每棵树的输出进行简单平均的思想，而是采用加权平均的方式。每个样本观测值的权重衡量了新变量 $\mathbf{X}_{\text{new}}$ 与每棵树中的每个样本观测值 X_i 的相似度。这样，GRF 可以根据相似度调整每个样本对预测值的影响，从而实现自适应性。

异质性分裂：在每个回归树的节点分裂方法上，GRF 不采用标准的 CART 算法，而是更加关注目标参数的异质性。它从最大化异质性的角度来选择每个节点的分裂方式，以更好地捕捉处理效应的异质性。

通过采用这种广义随机森林方法，我们可以克服传统回归模型中的线性假设限制，并更好地应对处理效应的异质性。GRF 能够通过灵活的非参数建模来捕捉复杂的关系，同时利用自适应性加权平均和异质性分裂来提高预测性能。因此，GRF 成为处理处理效应异质性和进行因果推断的一种前沿方法。

使用 GRF 获取单个数据集的双重稳健 ATE 一致估计量的流程可以分为三个步骤：

第一步，使用 GRF 的回归森林（regression forest）估计 $\mathbb{E}\left[Y\mid\mathbf{X}_i\right]$ 和 $\mathbb{E}\left[W\mid\mathbf{X}_i\right]$。

第二步，使用 GRF 的因果森林（causal forest）估计 $g_0(1,\mathbf{X})$ 和 $g_0(0,\mathbf{X})$。

第三步，将公式 (2-21) 中的 $g_0(1,\mathbf{X})$，$g_0(0,\mathbf{X})$ 和 $e_0(\mathbf{X})$ 分别替换为 $\hat{g}(1,\mathbf{X})$，$\hat{g}(0,\mathbf{X})$ 和 $\hat{e}(\mathbf{X})$。

通过以上三步，可以得到 ATE 的估计量 $\tau_{\text{GRF}}=\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i$，其中 $\hat{\tau}_i$ 为：

$$\hat{\tau}_i=\hat{g}(1,\mathbf{X}_i)-\hat{g}(0,\mathbf{X}_i)+\frac{W_i}{\hat{e}(\mathbf{X}_i)}(Y_i-\hat{g}(1,\mathbf{X}_i))-\frac{1-W_i}{1-\hat{e}(\mathbf{X}_i)}(Y_i-\hat{g}(0,\mathbf{X}_i))$$

ATE 的方差 V_τ 可以通过以下形式计算得到：

$$V_\tau=\frac{1}{n^2}\sum_{i=1}^{n}\left(\tau_i-\frac{1}{n}\sum_{j=1}^{n}\tau_j\right)^2$$

这个流程利用了 GRF 的非参数自适应性和异质性分裂的特点，能够在处理效应存在异质性的情况下得到稳健的因果估计量，并提供了对 ATE 估计的方差估计。

5 联邦学习的前置知识

5.1 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 $\boldsymbol{\beta}_0^{\mathrm{cb}}$ 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。

5.2 联邦学习加权的工具介绍

我们介绍联邦估计中使用的三种加权方法，这里每个联合估计器中选择不同的加权方法是基于相应单个数据集估计器的函数形式不同而定的.

5.2.1 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.

5.2.2 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：

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

5.2.3 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:

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

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

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

6 联邦学习

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。

6.1 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。

联邦 MLE 的估计值为：

$$\hat{\boldsymbol{\beta}}_{\text{mle}}^{\text{fed}} = \left(\sum_{k=1}^{K} \hat{H}_{\boldsymbol{\beta}}^{(k)} \right)^{-1} \left(\sum_{k=1}^{K} \hat{H}_{\boldsymbol{\beta}}^{(k)} \hat{\boldsymbol{\beta}}_{\text{mle}}^{(k)} \right), \quad \text{其中} \hat{H}_{\boldsymbol{\beta}}^{(k)} = \frac{\partial^2 \ell_{n_k}\left(\hat{\boldsymbol{\beta}}^{(k)} \right)}{\partial \boldsymbol{\beta}^{(k)} \left(\partial \boldsymbol{\beta}^{(k)} \right)^{\top}}. \tag{3}$$

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（参考 [54]），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。

6.2 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:

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

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 $x, y \in \{\text{s}, \text{uns}\}, \hat{\mathbf{H}}_{\boldsymbol{\beta},x,y}^{(k)} = \dfrac{\partial^2 \ell_{n_k}\left(\hat{\boldsymbol{\beta}}^{(k)}\right)}{\partial \boldsymbol{\beta}_x^{(k)} \left(\partial \boldsymbol{\beta}_y^{(k)}\right)^\top}, \mathbf{0}_{n_1 \times n_2}$ 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。

6.3 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 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 (2-1))。

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。

6.4 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。

6.5 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，

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

为了获得联邦方差，我们首先估计每个数据集上估计的 ATE 的方差，然后使用方差逆加权将所有数据集上估计的方差组合在一起

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

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。

6.6 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。

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

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。

6.7 偏回归模型的联邦 DML 估计

在这里我们不具体区分模型是否稳定与不稳定, 因为在这类模型下稳定和不稳定都能用样本量加权的方法去进行联邦估计, 而在稳定模型下可以用 IVW 估计即与 AIPW 模型类似, 下面介绍两种估计方法下的具体公式:

如果倾向模型或结果模型中的参数随数据集的变化而变化，那么在 DML 估计的定义中的评分函数 $\psi(\mathbf{X}_i, W_i, Y_i)$（作为倾向模型和结果模型的函数）也会随数据集而变化。因此，我们建议使用样本量加权来联合单个数据集的 ATE 或 ATT 估计，以得到联合的 ATE 或 ATT 估计。即：

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

其中，$\hat{\tau}^{(k)}_{\text{DML}}$ 与 $\hat{\mathbf{V}}\tau^{(k)}$ 由数据集 k 的倾向模型和结果模型估计得到。在这里，n_k 表示数据集 k 的样本量，n_{pool} 表示所有数据集的总样本量。通过样本量加权，我们将每个数据集的估计按照其相对样本量的比例进行组合，得到联合的估计和估计的方差。这样做可以更准确地反映不同数据集之间的贡献，并得到更稳健的估计结果。

DML 方法在倾向模型或结果模型中允许参数在不同数据集中任意变化，从而提供了很大的灵活性。然而，使用样本量加权的联合估计方法相对于反方差加权（IVW）的效率较低。IVW 方法通过最小化方差来实现加权平均。因此，在实践中，我们可以选择使用 IVW 方法来联合单个数据集的平均处理效应估计 $\hat{\tau}^{(k)}_{\text{DML}}$ 和方差估计 $\hat{V}\tau^{(k)}$:

$$\hat{\tau}^{\text{fed}}_{\text{DML}} = \left(\sum_{k=1}^{K} \left(\hat{V}^{(k)}_{\tau} \right)^{-1} \right)^{-1} \sum_{k=1}^{K} \left(\hat{V}^{(k)}_{\tau} \right)^{-1} \hat{\tau}^{(k)}_{\text{DML}}, \quad \hat{\mathbf{V}}^{\text{fed}}_{\tau} = \left(\sum_{k=1}^{K} \left(\hat{\mathbf{V}}^{(k)}_{\tau} \right)^{-1} \right)^{-1}$$

这种联合估计方法的一个优点是可以灵活地适应不同数据集中的参数变化。在高维协变量的情况下，我们采用 DML 中的 LASSO 方法来估计结果模型的干扰参数 g_0 和倾向模型的 e_0，以实现变量选择的目的。当然，如果对协变量的影响和模型的可解释性不感兴趣，也可以考虑使用 DML 框架中的其他方法，如随机森林或神经网络。

6.8 交互式模型的联邦估计

与半参数的偏回归模型以及 AIPW 模型类似, 当模型稳定的时候利用 IVW 加权方法对单个数据集的 ATE 估计进行联合：

$$\hat{\tau}^{\text{fed}}_{\text{GRF}} = \left(\sum_{k=1}^{K} \left(\hat{V}^{(k)}_{\tau} \right)^{-1} \right)^{-1} \sum_{k=1}^{K} \left(\hat{V}^{(k)}_{\tau} \right)^{-1} \hat{\tau}^{(k)}_{\text{GRF}}, \quad \hat{V}^{\text{fed}}_{\tau} = \left(\sum_{k=1}^{K} \left(\hat{V}^{(k)}_{\tau} \right)^{-1} \right)^{-1}$$

当模型不稳定的时候利用样本加权法对单个数据集的 ATE 估计进行联合：

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