arXiv:2603.01378v1 [stat.ME] 2 Mar 2026

数据集偏移下个体参与者数据与汇总数据的整合：汇总统计量比较与可扩展计算

Ming-Yueh Huang

中央研究院统计科学研究所，台北*11529*，台湾

Jing Qin

美国国立卫生研究院美国国家过敏症和传染病研究所生物统计学研究处，马里兰州贝塞斯达*20892*，美国

Chiung-Yu Huang

加利福尼亚大学旧金山分校流行病学与生物统计学系，
美国加利福尼亚州旧金山 *94158*

摘要

个体参与者数据与汇总数据整合分析将个体参与者数据（IPD）与汇总数据（AD）相结合，日益被视为从异质性研究中得出更可靠、更具可推广性推断的有效策略。尽管现有研究大多聚焦于算法方法，本文探讨一个互补但尚未充分探索的问题：不同形式的汇总数据如何影响数据整合的效率。在约束最大似然估计框架下，我们比较了通常报告的汇总统计量，并表明亚组特异性汇总数据可显著提高估计效率。具体而言，我们发现，来自按结局分层的亚组（例如病例与对照）的汇总数据所带来的效率增益，始终大于基于按协变量分层的亚组（例如年龄或暴露类别）的汇总数据，尤其是在结果变量为连续型时。尽管按结局分层的汇总通常会针对离散型结局进行报告，但当结果变量为连续型时却很少提供。因此，我们的研究结果支持在试验报告和公共数据存储库中常规纳入连续终点的按结局分层的汇总，以促进更高效的证据综合。我们进一步扩展约束最大似然框架以适应数据集偏移，并开发了一种快速的非迭代估计程序，以提高数值稳定性和可扩展性。我们通过两个应用说明所提出的方法：协变量偏移下的收入数据分析，以及在先验

概率偏移。

关键词：约束最大似然估计；协变量偏移；效率；先验概率偏移。

1 引言

将个体参与者数据（IPD）与汇总数据（AD）整合，已成为综合不同研究证据的一项重要策略。尽管IPD提供了详细的参与者层面信息，有助于灵活开展亚组分析并更好地控制潜在混杂，但出于隐私、监管或实际操作方面的约束，获取此类数据往往受到限制（Ahmed等人，2012）。相比之下，汇总数据由研究水平汇总统计量构成，通常较易获取，但缺乏开展深入调查所需的细致程度（Sutton和Higgins，2008）。个体参与者数据与汇总数据整合分析通过在单一建模框架中结合IPD的详细信息与研究水平的AD，最大限度地利用数据，使研究人员能够发挥两种数据类型各自的优势，从而提高统计推断的效率和普适性。

IPD–AD整合面临的一项核心方法学挑战是，AD可以采用不同形式报告，从简单的描述性统计量到基于模型的估计值不等；这些数据均存在抽样变异，且各自包含关于底层参数的不同信息。约束最大似然估计（CMLE）为将这类异质性汇总数据与IPD相结合提供了一个通用框架（Qin，2000；Chatterjee等人，2016）。在这一框架中，AD被视为满足无偏总体估计方程的参数估计值，随后将这些方程作为约束条件施加于基于IPD的似然，以提高效率。对于AD中的不确定性，可以假设汇总估计值近似服从正态分布，并将相应的似然贡献纳入受约束的目标函数中加以处理（Zhang等人，2020）。

个体参与者数据与汇总数据整合分析面临的另一项关键挑战是不同数据集之间存在异质性，尤其是当个体参与者数据和汇总数据来自不同来源、研究总体或时间段时（Chen 等人，2021年）。尽管荟萃分析文献已广泛承认研究间异质性，并且通常通过随机效应框架对其建模，但对异质性具体形式的系统评估仍然有限。本文借鉴机器学习中的数据集偏移框架（Storkey，2008年；Finlayson 等人，2021年），从概念上刻画研究间异质性，并考虑一种通用的数据集偏移模型，其中涵盖多种重要的偏移类型，例如协变量偏移（Sheng 等人，2022年；Cheng 等人，2023年）和先验概率偏移（Chen 等人，2024年）。在这一框架下，我们采用通用密度比

模型（Qin 和 Zhang，1997年；Qin，1998年），以关联个体参与者数据和汇总数据的数据生成机制，推导汇总数据的无偏估计方程，并开发统一的约束最大似然估计量程序，从而在存在这些偏移时仍能进行有效推断。

尽管现有关于 IPD–AD 荟萃分析的大量文献着重于开发整合算法，但对于不同形式的汇总数据如何影响统计推断的效率，关注较少。为弥补这一空白，我们在统一的 CMLE 框架下比较三种常用的汇总类型：结果变量或协变量的边际均值（Imbens and Lancaster，1994年；Qin，2000年）、按协变量分层的结局汇总（Huang等人，2016年； Shenget al.，2022年；Cheng et al.，2023年），以及按结局分层的协变量汇总。我们评估了它们在不存在数据集偏移以及一般偏移模型下的表现，并表明，相较于另外两类汇总，按结局分层的协变量汇总能够显著提高估计效率。尽管病例对照研究（即针对二元结局的研究）通常会报告此类汇总，但连续结局研究中较少提供这类数据。因此，我们的研究结果有力支持在试验论文和公共数据存储库中常规报告连续终点的按结局分层汇总，以促进更高效的证据综合。

直观而言，随着更多约束条件被纳入 CMLE 公式，潜在的效率提升也会随之增加。然而，当使用丰富的汇总数据时，例如包含许多协变量且结局经过细致划分的按结局分层的协变量汇总，约束条件的数量可能变得非常庞大，进而导致高维拉格朗日乘子系统和显著的数值不稳定性。因此，尽管标准 CMLE 算法灵活，且在理论上适用于整合个体参与者数据和汇总数据，但在实际应用中，其计算开销可能很大，也难以扩展。为应对这一挑战，我们基于Gao和Chan (2023年)的非迭代策略，开发了一种快速算法，可在单步内求得约束最大似然估计量，同时兼顾汇总数据的不确定性和数据集偏移。该方法保留了约束最大似然估计量的理论优势，并显著提高了其在高维整合问题中的数值稳定性和可扩展性。

本文其余部分安排如下。第2 节概述了在考虑数据不确定性的同时利用汇总数据的方法。第3节进一步扩展了协变量偏移和先验概率偏移下的约束最大似然估计量。第4节介绍了一种快速的非迭代算法，以提高计算效率。第5 节展示了全面的模拟研究，第6 节分析了两个实证示例，以说明所提方法的实用价值。第7节给出了简要总结和进一步讨论。

2 个体参与者数据-汇总数据综合分析

2.1 简要回顾

我们首先回顾用于整合个体参与者数据-汇总数据分析的约束最大似然估计框架（Qin，2000年；Chatterjee等人，2016年）。设 Y 为感兴趣的结果变量，X为 p维协变量。假设个体参与者数据 $\{(Y_i, \mathrm{X}_i): i = 1, \ldots, n\}$是$(Y, \mathrm{X})$的独立同分布副本。我们假设 Y在给定X $=$x时的条件密度服从参数模型$f(y|\mathrm{x}; \boldsymbol{\beta})$，其中 f是已知函数， $\boldsymbol{\beta}$是一个 d维参数。协变量的边缘分布 $G(\mathrm{x})$不作具体设定，并视为干扰参数。基于个体参与者数据的完整对数似然函数为$\ell(\boldsymbol{\beta}, G) = \sum_{i=1}^{n} \log\ f(Y_i \,|\, \mathrm{X}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} \log\ \Delta G(\mathrm{X}_i)$。根据标准的似然论证，当 G 仅在观测到的协变量取值 $\{\mathrm{X}_i: i = 1, \ldots, n\}$处赋予概率质量时， $\ell(\boldsymbol{\beta}, G)$达到最大值。令 p_i 表示 G 在X_i, $i = 1, \ldots, n$处的跳跃大小，并定义$\mathrm{p} = (p_1, \ldots, p_n)$。于是，对数似然可改写为 $\ell(\boldsymbol{\beta}, \mathrm{p}) = \sum_{i=1}^{n} \log\ f(Y_i \,|\, \mathrm{X}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} \log\ p_i$，其中$p_i \geq 0$ 且 $\sum_{i=1}^{n} p_i = 1$。记$(\boldsymbol{\beta}, \tilde{\mathrm{p}})$为最大似然估计量；在没有汇总数据信息的情况下， $\tilde{\mathrm{p}} = (\tilde{p}_1, \ldots, \tilde{p}_n)$由 $\tilde{p}_i = n^{-1}$给出。

正如包括 Chatterjee等人（2016年）、Zhang等人（2020年）和 Huang等人（2016年）在内的多位作者所指出的，汇总数据通常可表示为满足总体估计方程 $\mathbb{E}\{\boldsymbol{\varphi}(Y, \mathrm{X}; \boldsymbol{\phi}, \boldsymbol{\theta})\} = \mathbf{0}$ 的 q 维参数 $\boldsymbol{\phi}$ 的估计值 $\boldsymbol{\phi}$ ；其中，r 维函数 $\boldsymbol{\varphi}$ 也可能取决于另一个未知的 s 维参数$\boldsymbol{\theta}$。第 2.2 节将讨论 $\boldsymbol{\phi}$的示例，包括亚组特异性汇总统计量。通过二重期望，感兴趣参数 $\boldsymbol{\beta}$ 可通过总体估计方程 $\mathbb{E}\{\boldsymbol{\Psi}(\mathrm{X}; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})\} = \mathbf{0}$ 与汇总数据参数$\boldsymbol{\phi}$ 建立联系，其中

$$\boldsymbol{\Psi}(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}\{\boldsymbol{\varphi}(Y, \mathbf{X}; \boldsymbol{\phi}, \boldsymbol{\theta}) \,|\, \mathbf{X}\} = \int \boldsymbol{\varphi}(y, \mathbf{X}; \boldsymbol{\phi}, \boldsymbol{\theta}) f(y \,|\, \mathbf{X}; \boldsymbol{\beta}) \mathrm{d}y.$$

当个体参与者数据和汇总数据服从相同分布时，汇总数据总体矩可由个体参与者数据样本矩近似，从而得到样本矩约束 $\sum_{i=1}^{n} p_i \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \tilde{\mathbf{0}}$。当汇总数据样本量足够大时， $\boldsymbol{\phi}$ 的不确定性可忽略不计，因此可将其视为 $\boldsymbol{\phi}$的真实值。为整合汇总数据，可通过在约束条件下最大化完整对数似然 $\ell(\boldsymbol{\beta}, \mathbf{p})$，求得约束最大似然估计（$\boldsymbol{\beta}, \hat{\mathbf{p}}\ \boldsymbol{\theta}$），其中对 $(\boldsymbol{\beta}, \mathbf{p}\ \boldsymbol{\theta})$ 进行最大化，约束条件为

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta}) = \mathbf{0}. \tag{1}$$

应用标准的经验似然论证（Qin 和 Lawless，1994年；Qin，2000年），我们得到

$$p_i = \frac{1}{n} \times \frac{1}{1 + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta})}, \quad i = 1, \ldots, n, \tag{2}$$

约束对数似然（相差一个常数）为，

$$\sum_{i=1}^{n} \log f(Y_i \mid \mathbf{X}_i; \boldsymbol{\beta}) - \sum_{i=1}^{n} \log\{1 + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta})\},$$

其中，$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)$ 是满足以下条件的拉格朗日乘子

$$\sum_{i=1}^{n} \frac{\boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta})}{1 + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta})} = \mathbf{0}. \tag{3}$$

可以证明，约束最大似然估计比其无约束对应估计更有效率。直观来说，这是因为汇总数据带来的额外约束进一步将参数估计限制在同时符合个体参与者数据和汇总数据的取值范围内，从而降低了估计变异性。

当汇总数据的不确定性不可忽略时，我们采用Zhang等人（2020年）提出的策略，将汇总数据视为随机向量的一次实现。具体而言，当汇总数据样本量 N 足够大时，$N^{1/2}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi})$ 近似服从均值为零、方差-协方差矩阵为 $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$的正态分布。为纳入汇总数据，我们在个体参与者数据对数似然中加入一个基于汇总数据正态近似的二次项：

$$\ell(\boldsymbol{\beta}, \mathbf{p}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \log f(Y_i \mid \mathbf{X}_i; \boldsymbol{\beta}) + \sum_{i=1}^{n} \log p_i - \frac{N}{2}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi})^{\mathrm{T}} \mathbf{V}^{-1} (\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}), \tag{4}$$

其中，$\mathbf{V}$ 是一个工作方差-协方差矩阵。通过在约束条件下最大化增广对数似然，可以得到约束最大似然估计

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbf{0}, \tag{5}$$

关于（$\boldsymbol{\beta}$，$\mathbf{p}$ $\boldsymbol{\phi}$，$\boldsymbol{\theta}$）。与前文类似，忽略常数项后，约束对数似然可改写为

$$\sum_{i=1}^{n} \log f(Y_i \mid \mathbf{X}_i; \boldsymbol{\beta}) - \frac{N}{2}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi})^{\mathrm{T}} \mathbf{V}^{-1} (\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) - \sum_{i=1}^{n} \log\{1 + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})\}, \tag{6}$$

其中 $\boldsymbol{\eta}$ 由满足以下条件的拉格朗日乘子组成

$$\sum_{i=1}^{n} \frac{\boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})}{1 + \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})} = \mathbf{0}.$$

正如Zhang等人（2020年）所示，任意正定的 $K \times K$ 选择$\mathbf{V}$ 都会得到一个 $n^{1/2}$-相合估计量。最优选择为$\mathbf{V} = \boldsymbol{\Sigma}_{\boldsymbol{\phi}}$，由此得到有效性最高的估计量$\widehat{\boldsymbol{\beta}}$。以下定理总结了 $\boldsymbol{\beta}$ 的渐近性质。

定理 1. 设 $(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \mathbf{0})$ 为 $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\eta})$ 的真实值，而 $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}, \widehat{\boldsymbol{\eta}})$ 为约束最大似然估计。假设补充材料中的正则条件 *A1–A6* 均成立，并且

假设 $N/n \to \kappa \in (0,\infty)$。那么， $n^{1/2}\{(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\eta}}) - (\boldsymbol{\beta}_0,\boldsymbol{\phi}_0,\boldsymbol{\theta}_0,\mathbf{0})\}$ 收敛于均值为零、方差-协方差矩阵为 $\mathbf{J}^{-1}\boldsymbol{\Sigma}\mathbf{J}^{-1}$ 的正态分布，其中

$$\boldsymbol{\Sigma} = \begin{pmatrix} -\mathbf{H}_{\boldsymbol{\beta}_0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \kappa\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\beta}_0,\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)^{\otimes 2}\} \end{pmatrix},$$

$$\mathbf{J} = \begin{pmatrix} -\mathbf{H}_{\boldsymbol{\beta}_0} & \mathbf{0} & \mathbf{0} & \dot{\boldsymbol{\Psi}}_{\boldsymbol{\beta}} \\ \mathbf{0} & \kappa\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1} & \mathbf{0} & \dot{\boldsymbol{\Psi}}_{\boldsymbol{\phi}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dot{\boldsymbol{\Psi}}_{\boldsymbol{\theta}} \\ \dot{\boldsymbol{\Psi}}_{\boldsymbol{\beta}}^{\mathrm{T}} & \dot{\boldsymbol{\Psi}}_{\boldsymbol{\phi}}^{\mathrm{T}} & \dot{\boldsymbol{\Psi}}_{\boldsymbol{\theta}}^{\mathrm{T}} & -\mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X};\beta_0,\boldsymbol{\phi}_0,\boldsymbol{\theta}_0)^{\otimes 2}\} \end{pmatrix},$$

$\mathbf{H}_{\boldsymbol{\beta}} = \mathbb{E}\{\partial^2 \log\ f(Y \mid \mathbf{X};\ \boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\otimes 2}\}$，以及 $(\dot{\boldsymbol{\Psi}}_{\boldsymbol{\beta}}^{\mathrm{T}},\ \dot{\boldsymbol{\Psi}}_{\boldsymbol{\phi}}^{\mathrm{T}},\ \dot{\boldsymbol{\Psi}}_{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}} = \mathbb{E}\{\partial\boldsymbol{\Psi}(\mathbf{X};\ \boldsymbol{\beta}_0,\ \boldsymbol{\phi}_0,\ \boldsymbol{\theta}_0)/\partial(\boldsymbol{\beta}^{\mathrm{T}},\ \boldsymbol{\phi}^{\mathrm{T}},\ \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}\}$。

如第 4 节将要证明的那样，定理 1 中的渐近结果可用于构造一个渐近等价的估计量，并且可通过非迭代算法高效地求得。

2.2 协变量分层和结局分层汇总数据

本文聚焦于最常见的汇总数据形式：按暴露或结局分层的子组汇总统计量，例如均值或比例。这些汇总数据常规见于生物医学研究，通常以所谓的“表 1”形式报告，用于描述不同亚群体中协变量或结局的分布，因此可便捷地用于个体参与者数据-汇总数据分析。

一个重要的例子是汇总由年龄、性别或暴露状态等基线协变量定义的亚组内结果。亚组可以由单个协变量或多个协变量的组合定义，从而形成一组（可能相互重叠的）子集 $\Omega 1,\ldots,\Omega_K$。报告的亚组均值近似于总体参数 $\phi_k = \mathbb{E}(Y \mid \mathbf{X} \in \Omega_k),\ k = 1,\ldots,K$，其中每个参数都可以重新表示为估计方程 $\mathbb{E}\{\varphi_k(Y,\mathrm{X};\ \phi_k)\} = 0$ with $\varphi_k(Y,\mathrm{X};\ \phi_k) = (Y-\phi_k)I(\mathrm{X} \in \Omega_k)$ 的解。根据迭代期望， ϕ_k 还可以通过 $\mathbb{E}\{\Psi_k(\mathrm{X};\ \boldsymbol{\beta},\phi_k)\} = 0$ 与结局模型 $f(y \mid \mathrm{x};\ \boldsymbol{\beta})$ 进一步关联，其中 $\Psi_k(\mathrm{X};\ \boldsymbol{\beta},\phi_k) = \{\int y f(y \mid \mathrm{X};\ \boldsymbol{\beta})\mathrm{d}y - \phi_k\}I(\mathrm{X} \in \Omega_k)$。对于连续结局，亚组均值通常还会附有方差估计，这些估计可以作为额外的约束条件纳入其中。直观而言，增加亚组数量可以提供更丰富的辅助信息，并提高效率。然而，亚组之间的大量重叠会导致汇总统计量之间存在较强相关性，可能造成数值不稳定性。因此，谨慎

地选择至关重要，以便在选择用于整合的汇总信息时，平衡信息量与冗余度。

另一种常见的汇总数据形式是，按结果变量定义的子组汇总协变量。典型例子是病例对照研究，其中分别报告病例（$Y = 1$）和对照（$Y = 0$）的协变量均值或比例，以突出潜在风险因素。按结局分层的汇总统计量近似于 $\boldsymbol{\phi}_k = \mathbb{E}(\mathbf{X} \mid Y = k)$、 $k = 0$、1，这些量对应于总体估计方程 $\mathbb{E}\{\boldsymbol{\varphi}_k(Y, \mathbf{X}; \boldsymbol{\phi}_k)\} = \mathbf{0}$ 的解，其中$\boldsymbol{\varphi}_k(Y, \mathbf{X}; \boldsymbol{\phi}_k) = (\mathbf{X} - \boldsymbol{\phi}_k) I(Y = k)$。通过二重期望， $\boldsymbol{\phi}_k$ 可通过估计方程 $\mathbb{E}\{\boldsymbol{\Psi}_k(\mathbf{X};\, \boldsymbol{\beta}, \boldsymbol{\phi}_k)\} = \mathbf{0}$ 与结局模型 $\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = f(\mathbf{x};\, \boldsymbol{\beta})$ 相联系，其中 $\boldsymbol{\Psi}_k(\mathbf{X};\, \boldsymbol{\beta}, \boldsymbol{\phi}_k) = (\mathbf{X} - \boldsymbol{\phi}_k) f(\mathbf{X};\, \boldsymbol{\beta})^k \{1 - f(\mathbf{X};\, \boldsymbol{\beta})\}^{1-k}$。这一框架可自然扩展至连续型结局，即将结局范围划分为若干区间， $-\infty = y_0^* < y_1^* <$

$\cdots < y_{K-1}^* < y_K = \infty$。对于每个区间（$y_{k-1}^*, y_k^*$], 条件协变量均值 $\boldsymbol{\phi}_k = \mathbb{E}(\mathbf{X} \mid y_{k-1}^* < Y \le y_k^*)$， $k = 1, \ldots, K$ 满足估计方程 $\mathbb{E}\{\boldsymbol{\varphi}_k(Y, \mathbf{X}; \boldsymbol{\phi}_k)\} = \mathbf{0}$，其中 $\boldsymbol{\varphi}_k(Y, \mathbf{X};\, \boldsymbol{\phi}_k) = (\mathbf{X} - \boldsymbol{\phi}_k) I(y_{k-1}^* < Y \le y_k^*)$。在给定 $\mathbf{X}$ 时，若 Y 遵循参数模型 $f(y \mid \mathbf{x};\, \boldsymbol{\beta})$，则参数 $\boldsymbol{\phi}_k$ 满足总体估计方程 $\mathbb{E}\{\boldsymbol{\Psi}_k(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\phi}_k)\} = \mathbf{0}$，其中 $\boldsymbol{\Psi}_k(\mathbf{X};\, \boldsymbol{\beta}, \boldsymbol{\phi}_k) = (\mathbf{X} - \boldsymbol{\phi}_k) \int_{y_{k-1}^*}^{y_k^*} f(y \mid \mathbf{X};\, \boldsymbol{\beta}) \mathrm{d}y$。与按协变量分层的汇总数据相比，按结局分层的汇总通常更透明、冗余更少，因为结局是一维的，因而可以进行系统划分。鉴于这些方法在整合个体参与者数据-汇总数据分析中的概念优势，连续结局的按结局分层汇总鲜少被报告，这使得更有效的实践错失了机会。

3 存在数据集偏移时的个体参与者数据-汇总数据分析

当个体参与者数据和汇总数据来自不同总体时，可能会出现数据集偏移，导致结果变量与协变量的联合分布存在差异。为便于讨论，我们聚焦于汇总数据来自单项研究的情形，不过，所提出的框架也可扩展至多个来源。令 $f(y, \mathbf{x})$ 表示个体参与者数据研究中随机变量（$Y, \mathbf{X}$）的联合密度，并令 $f^*(y, \mathbf{x})$ 表示汇总数据研究中（$Y^*, \mathbf{X}^*$）的联合密度。当 $f(y, \mathbf{x}) \neq f^*(y, \mathbf{x})$ 时，就会出现数据集偏移。

为刻画数据集偏移，我们引入一个涵盖两种主要偏移类型的一般模型：协变量偏移和先验概率偏移。在协变量偏移下，不同研究中的$\mathbf{X}$ 边缘分布不同，但 Y 给定$\mathbf{X}$ 的条件分布保持不变。令 $f(y \mid \mathbf{x})$ 表示 Y给定$\mathbf{X}$ 的条件密度，并令 $G(\mathbf{x})$ 表示$\mathbf{X}$的边缘分布函数；此外，以类似方式定义 $(Y^*, \mathbf{X}^*)$ 的 $f^*(y \mid \mathbf{x})$ 和 $G^*(\mathbf{x})$。记 $\mathrm{d}G^*(\mathbf{x}) = w_{\mathbf{X}}(\mathbf{x}) \mathrm{d}G(\mathbf{x}) / \int w_{\mathbf{X}}(\mathbf{u}) \mathrm{d}G(\mathbf{u})$。这里，$\mathrm{d}G^*(\mathbf{x}) / \mathrm{d}G(\mathbf{x}) \propto w_{\mathbf{X}}(\mathbf{x})$ 可以

视为刻画协变量分布中异质性的偏移函数。由 $f(y\,|\,\mathbf{x}) = f^*(y\,|\,\mathbf{x})$ 和 $\int f(y\,|\,\mathbf{x})\mathrm{d}y = 1$ 可得

$$f^*(y,\mathbf{x}) = \frac{f(y\,|\,\mathbf{x})w_{\mathbf{X}}(\mathbf{x})\mathrm{d}G(\mathbf{x})/\mathrm{d}\mathbf{x}}{\iint f(v\,|\,\mathbf{u})w_{\mathbf{X}}(\mathbf{u})\mathrm{d}G(\mathbf{u})\mathrm{d}v} = \frac{w_{\mathbf{X}}(\mathbf{x})f(y,\mathbf{x})}{\iint w_{\mathbf{X}}(\mathbf{u})f(v,\mathbf{u})\mathrm{d}v\mathrm{d}\mathbf{u}}.$$

接下来，我们考虑先验概率偏移；在分类结局的情形下，这也称为标签偏移。在这种情况下，Y 的边缘分布会因研究而异，而给定 Y 时 $\mathbf{X}$ 的条件分布保持不变。稍微滥用一下符号，令 $f(\mathbf{x}\,|\,y)$ 表示给定 Y 时 $\mathbf{X}$ 的条件密度，$H(y)$ 表示个体参与者数据研究中 Y 的边缘分布函数。对于汇总数据研究中的 $(Y^*, \mathbf{X}^*)$，相应地定义 $f^*(\mathbf{x}\,|\,y)$ 和 $H^*(y)$。写作 $\mathrm{d}H^*(y) = w_Y(y)\mathrm{d}H(y) / \int w_Y(v)\mathrm{d}H(v)$；这里，$w_Y(y)$ 是偏移函数，用于刻画不同研究间 Y 分布的异质性。根据 $f(\mathbf{x}\,|\,y) = f^*(\mathbf{x}\,|\,y)$ 和 $\int f(\mathbf{x}\,|\,y)\mathrm{d}\mathbf{x} = 1$，先验概率偏移下的联合密度函数为

$$f^*(y,\mathbf{x}) = \frac{f(\mathbf{x}\,|\,y)w_Y(y)\mathrm{d}H(y)/\mathrm{d}y}{\iint f(\mathbf{u}\,|\,v)w_Y(v)\mathrm{d}H(v)\mathrm{d}\mathbf{u}} = \frac{w_Y(y)f(y,\mathbf{x})}{\iint w_Y(v)f(v,\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}v}.$$

注意，在这种情况下，我们有

$$f^*(y\,|\,\mathbf{x}) = \frac{w_Y(y)f(y\,|\,\mathbf{x})\mathrm{d}G(\mathbf{x})}{\int w_Y(v)f(v\,|\,\mathbf{x})\mathrm{d}G(\mathbf{x})\mathrm{d}v} = \frac{w_Y(y)f(y\,|\,\mathbf{x})}{\int w_Y(v)f(v\,|\,\mathbf{x})\mathrm{d}v} \neq f(y\,|\,\mathbf{x}), \tag{7}$$

这表明，给定$\mathbf{X}$时，Y 的条件分布因研究而异。最后，为了同时适应协变量偏移和先验概率偏移，我们采用有偏抽样表述：

$$f^*(y,\mathbf{x}) = \frac{w_{\mathbf{X}}(\mathbf{x})w_Y(y)f(y,\mathbf{x})}{\iint w_{\mathbf{X}}(\mathbf{u})w_Y(v)f(v,\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}v}. \tag{8}$$

在此框架下，汇总数据研究中的$(Y^*, \mathbf{X}^*)$可视为来自个体参与者数据研究总体的$(Y, \mathbf{X})$的有偏样本，其抽样权重为 $w_{\mathbf{X}}(\mathbf{x})w_Y(y)$。当 $w_Y(y)$为常数时，抽样权重仅取决于协变量，对应于协变量偏移；当 $w_{\mathbf{X}}(\mathbf{x})$为常数时，抽样权重仅取决于结果变量，对应于先验概率偏移。

在一般的数据集偏移模型（8）下，我们有

$$\begin{aligned}\mathbb{E}\{\boldsymbol{\Psi}(\mathbf{X}^*;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})\} &= \mathbb{E}[\mathbb{E}\{\boldsymbol{\varphi}(Y^*,\mathbf{X}^*;\boldsymbol{\phi},\boldsymbol{\theta})\,|\,\mathbf{X}^*\}] \\ &= \frac{\int\{w_{\mathbf{X}}(\mathbf{x})\int\boldsymbol{\varphi}(y,\mathbf{x};\boldsymbol{\phi},\boldsymbol{\theta})w_Y(y)f(y\,|\,\mathbf{x};\boldsymbol{\beta})\mathrm{d}y\}\mathrm{d}G(\mathbf{x})}{\iint w_{\mathbf{X}}(\mathbf{x})w_Y(y)f(y\,|\,\mathbf{x};\boldsymbol{\beta})\mathrm{d}y\mathrm{d}G(\mathbf{x})}.\end{aligned} \tag{9}$$

定义 $\boldsymbol{\psi}(\mathbf{X};\ \boldsymbol{\beta}、\boldsymbol{\phi}、\boldsymbol{\theta}) = w_{\mathbf{X}}(\mathbf{X})\int\boldsymbol{\varphi}(y,\mathbf{X};\ \boldsymbol{\phi}、\boldsymbol{\theta})w_Y(y)f(y\,|\,\mathbf{X};\ \boldsymbol{\beta})\mathrm{d}y$。为在数据集偏移下获得约束最大似然估计量，可以在约束条件中用 $\boldsymbol{\psi}(\mathbf{X};\ \boldsymbol{\beta}、\boldsymbol{\phi}、\boldsymbol{\theta})$ 替换 $\boldsymbol{\Psi}(\mathbf{X};\ \boldsymbol{\beta}、\boldsymbol{\phi}、\boldsymbol{\theta})$，从而最大化（6）。在协变量偏移下，$\boldsymbol{\psi}(\mathbf{X};\ \boldsymbol{\beta}、\boldsymbol{\phi}、\boldsymbol{\theta})$ 可化为

$$w_{\mathbf{X}}(\mathbf{X})\int\boldsymbol{\varphi}(y,\mathbf{X};\boldsymbol{\phi},\boldsymbol{\theta})f(y\,|\,\mathbf{X};\boldsymbol{\beta})\mathrm{d}y = w_{\mathbf{X}}(\mathbf{X})\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta}). \tag{10}$$

由于通常假设边际密度不包含关于 $\boldsymbol{\beta}$的信息， $w_{\mathbf{X}}$ 是 $\boldsymbol{\beta}$中的常数函数。当汇总数据仅包含总体协变量汇总时， $\boldsymbol{\psi}(\mathrm{X};\ \boldsymbol{\beta}、\ \boldsymbol{\phi}、\ \boldsymbol{\theta})$ 不依赖于 $\boldsymbol{\beta}$。在这种情况下，关于 $\boldsymbol{\beta}$ 最大化（6）的极大化解等价于$\sim\boldsymbol{\beta} = \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log\ f(Y_i \mid \mathrm{X}_i;\ \boldsymbol{\beta})$。因此，结合此类汇总数据相比无约束最大似然估计不会带来效率提升。另一方面，当汇总数据涉及与结果变量相关的汇总（例如总体结局汇总、按协变量分层的结局汇总以及按结果变量-$\widetilde{}$分层的协变量汇总）时， $\boldsymbol{\psi}(\mathrm{X};\ \boldsymbol{\beta}、\ \boldsymbol{\phi}、\ \boldsymbol{\theta})$ 会随 $\boldsymbol{\beta}$ 变化，并且相比 $\boldsymbol{\beta}$可提高效率。接下来，在先验概率偏移下， $\boldsymbol{\psi}(\mathrm{X};\ \boldsymbol{\beta}、\ \boldsymbol{\phi}、\ \boldsymbol{\theta})$ 可化为

$$\int \boldsymbol{\varphi}(y, \mathbf{X}; \boldsymbol{\phi}, \boldsymbol{\theta}) w_Y(y) f(y \mid \mathbf{X}; \boldsymbol{\beta}) \mathrm{d}y. \tag{11}$$

因此，大多数类型的汇总数据都会产生依赖于 $\boldsymbol{\beta}$ 的 $\boldsymbol{\psi}(\mathbf{X};\ \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})$，而由此带来的效率提升会随着 $H^*(y)$ 与 $H(y)$ 之间偏移程度的增大而增加，该程度由 $w_Y(y)$ 表征。

在实践中，在数据集偏移下实施约束最大似然估计量面临两大挑战。首先，偏移函数 $w_{\mathbf{X}}(\mathrm{x})$ 和 $w_Y(y)$ 通常未知，必须进行估计。遵循 Qin（1998），我们采用半参数密度比模型，并通过最大化约束对数似然，联合估计模型参数与 $(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta})$。例如，在协变量偏移下，我们设定 $w_{\mathbf{X}}(\mathrm{x}; \boldsymbol{\theta}^*) = \exp\{\mathrm{h}(\mathrm{x})^{\mathrm{T}} \boldsymbol{\theta}^*\}$，其中 $\mathrm{h}(\cdot)$ 是预先设定的 s 维函数，$\boldsymbol{\theta}^*$ 是一个 s 维参数向量；例如，令 $\mathrm{h}(\mathrm{x}) = \mathrm{x}$ 即得到指数倾斜密度比模型（Qin 和 Zhang，1997年）。类似地，在先验概率偏移下，可以考虑 $w_Y(y; \boldsymbol{\theta}^*) = \exp\{\mathrm{h}(y)^{\mathrm{T}} \boldsymbol{\theta}^*\}$。如果 H^* 与 H 之间的对数密度比是平滑的，则可利用泰勒定理通过多项式展开对其进行近似，从而启发我们设定 $\mathrm{h}(y) = (y, y^2, \ldots, y^s)^{\mathrm{T}}$。尽管密度比模型引入了一个额外参数 $\boldsymbol{\theta}^*$，但它只出现在约束条件（5）中，可视为约束最大似然估计量中的一个干扰 $\boldsymbol{\theta}$。因此，最大化约束对数似然（6）的算法可直接用于估计参数 $\boldsymbol{\beta}$、 $\boldsymbol{\phi}$ 和 $\boldsymbol{\theta}$，包括 $\boldsymbol{\theta}^*$。正如 Sheng等（2022年）所指出的，选择纳入协变量偏移函数 $w_{\mathbf{X}}(\mathrm{x}; \boldsymbol{\theta}^*)$ 的协变量时，可通过比较不同研究中的汇总统计量，并结合对研究设计的了解和专业领域知识来进行指导。类似的方法也可用于指导先验概率偏移下 $w_Y(y; \boldsymbol{\theta}^*)$ 的模型设定。这些边际汇总为识别偏移模型参数 $\boldsymbol{\theta}$提供了重要约束。

第二个挑战是计算 $\boldsymbol{\psi}(\mathbf{X}; \boldsymbol{\beta},\ \boldsymbol{\phi},\ \boldsymbol{\theta})$，其中涉及一个额外的积分项，

$$\int \boldsymbol{\varphi}(y, \mathbf{x}; \boldsymbol{\phi}, \boldsymbol{\theta}) w_Y(y) f(y \mid \mathbf{x}; \boldsymbol{\beta}) \mathrm{d}y. \tag{12}$$

在协变量偏移下，由（10）可知，积分可简化为 $\boldsymbol{\Psi}(\mathbf{X};\ \boldsymbol{\beta},\ \boldsymbol{\phi},\ \boldsymbol{\theta})$。因此，第2.1节所述的CMLE算法可通过将权重 $w_{\mathbf{X}}(\mathbf{X}_i)$ 直接纳入约束 $\sum_{i=1}^{n} p_i \boldsymbol{\Psi}(\mathbf{X}_i;\ \boldsymbol{\beta},\ \boldsymbol{\phi},\ \boldsymbol{\theta}) = \mathbf{0}$，在无需计算新积分的情况下处理协变量偏移。另一方面，在先验概率偏移下，利用（7），额外的积分项（12）可改写为

$$\begin{aligned}&\int \boldsymbol{\varphi}(y,\mathbf{x};\boldsymbol{\phi},\boldsymbol{\theta}) f^*(y\,|\,\mathbf{x};\boldsymbol{\beta})\mathrm{d}y \times \int w_Y(y) f(y\,|\,\mathbf{x};\boldsymbol{\beta})\mathrm{d}y \\ &= \mathbb{E}\{\boldsymbol{\varphi}(Y^*,\mathbf{X}^*;\boldsymbol{\phi},\boldsymbol{\theta})\,|\,\mathbf{X}^*=\mathbf{x}\} \times \int w_Y(y) f(y\,|\,\mathbf{x};\boldsymbol{\beta})\mathrm{d}y, \end{aligned} \tag{13}$$

其中，$f^*(y\,|\,\mathbf{x};\ \boldsymbol{\beta}) = w_Y(y) f(y\,|\,\mathbf{x};\ \boldsymbol{\beta}) / \int w_Y(y) f(y\,|\,\mathbf{x};\ \boldsymbol{\beta})\mathrm{d}y$ 是偏移后的条件密度。虽然可以对这些积分进行数值近似或蒙特卡洛近似，但计算量很大。然而，当 $f(y\,|\,\mathbf{x};\ \boldsymbol{\beta})$ 属于指数族，且 $w_Y(y)$ 遵循密度比模型 $\exp\{\mathrm{h}(y)^{\mathrm{T}}\boldsymbol{\theta}^*\}$ 时，偏移后的条件密度也属于指数族。具体而言，当 $f(y\,|\,\mathbf{x};\ \boldsymbol{\beta})$ 遵循广义线性模型，且 $\mathrm{h}(y) = y$ 时，$f^*(y\,|\,\mathbf{X};\ \boldsymbol{\beta})$ 和缩放因子 $\int w_Y(y) f(y\,|\,\mathbf{X};\ \boldsymbol{\beta})\mathrm{d}y$ 都有简单的闭式表达式。表 1 给出了一些示例。在这些情形下，偏移后的条件分布与 $f(y\,|\,\mathbf{x};\ \boldsymbol{\beta})$ 属于同一类分布，其自然参数根据 θ^* 的简单函数进行偏移。因此，对于大多数类型的汇总数据，例如结果变量和协变量的总体汇总、按结果变量分层的协变量汇总，以及按协变量分层的结果变量汇总，偏移后的期望 $\mathbb{E}\{\boldsymbol{\varphi}(Y^*,\mathbf{X}^*;\boldsymbol{\phi},\boldsymbol{\theta})\,|\,\mathbf{X}^*\}$ 也具有简单的闭式表达式。因此，对于这些情形，无需进行计算量很大的数值近似。

4A 快速非迭代算法

在实践中，可以通过求解以下嵌套优化问题来获得约束最大似然估计量：

$$\max_{\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \log f(Y_i\,|\,\mathbf{X}_i;\boldsymbol{\beta}) - \frac{N}{2}(\widetilde{\boldsymbol{\phi}}-\boldsymbol{\phi})^{\mathrm{T}}\mathbf{V}^{-1}(\widetilde{\boldsymbol{\phi}}-\boldsymbol{\phi}) - \min_{\boldsymbol{\eta}} \sum_{i=1}^{n} \log\{1+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})\} \right].$$

这一优化问题带来了两大计算挑战。首先，对于给定的 $(\boldsymbol{\beta},\ \boldsymbol{\phi},\ \boldsymbol{\theta})$，内层最小化会得到拉格朗日乘子 $\widehat{\boldsymbol{\eta}}$，该乘子通过牛顿-拉夫森迭代计算，作为 $\{\widehat{\boldsymbol{\eta}}^{(m)}\ \text{的极限：}\ m\in\mathbb{N}\}$：

$$\widehat{\boldsymbol{\eta}}^{(m+1)} = \widehat{\boldsymbol{\eta}}^{(m)} + \left[\sum_{i=1}^{n} \frac{\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})^{\otimes 2}}{\{1+\widehat{\boldsymbol{\eta}}^{(m)\mathrm{T}}\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})\}^2} \right]^{-1} \sum_{i=1}^{n} \frac{\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})}{1+\widehat{\boldsymbol{\eta}}^{(m)\mathrm{T}}\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta})}.$$

当000位于 $\{\boldsymbol{\psi}(\mathbf{X}_i;\boldsymbol{\beta},\ \boldsymbol{\phi},\ \boldsymbol{\theta}):i=1,\ \ldots\ldots,\ n\}$的凸包内时，这些迭代保证收敛。然而，梯度和海森矩阵的维度可能相当大，尤其是在整合多个AD 数据源时。其次，外层最大化

表 1：不同分布假设下结局模型 $f_{Y|\mathbf{X}}(y|\mathbf{x})$ 及其指数倾斜版本 $f^*(y|\mathbf{x}) \propto f(y|\mathbf{x};\boldsymbol{\beta})\exp(\theta^* y)\mathrm{d}y$，以及相应的归一化常数 $\int f(y|\mathbf{x};\boldsymbol{\beta})\exp(\theta^* y)y$；$\tilde{\mathbf{x}} = (1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$。

$f_{Y\|\mathbf{X}}(y\|\mathbf{x})$	$f^*(y\|\mathbf{x})$	$\int f(y\|\mathbf{x};\boldsymbol{\beta})\exp(\theta^* y)\mathrm{d}y$
伯努利分布$\{p(\mathbf{x};\boldsymbol{\beta})\}$ $p(\mathbf{x};\boldsymbol{\beta}) = \text{逻辑函数}(\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta})$	伯努利分布$\{\text{逆逻辑函数}(\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}+\theta^*)\}$	$\frac{1+\exp(\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}+\theta^*)}{1+\exp(\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta})}$
泊松分布$\{\lambda(\mathbf{x};\boldsymbol{\beta})\}$ $\lambda(\mathbf{x};\boldsymbol{\beta}) = \exp(\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta})$	泊松分布$\{\lambda(\mathbf{x};\boldsymbol{\beta})\exp(-\theta^*)\}$	$\exp[\lambda(\mathbf{x};\boldsymbol{\beta})\{\exp(-\theta^*) - 1\}]$
正态分布$\{\mu(\mathbf{x};\boldsymbol{\beta})\ \sigma^2\}$ $\mu(\mathbf{x};\boldsymbol{\beta}) = \tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}_{(1)}$	正态分布$\{\mu(\mathbf{x};\boldsymbol{\beta}) + \sigma^2\theta^*, \sigma^2\}$	$\exp\{(2\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}_{(1)} + \sigma^2\theta^*)\theta^*/2\}$
伽马分布$\{\nu, \lambda(\mathbf{x};\boldsymbol{\beta})\}$ $\lambda(\mathbf{x};\boldsymbol{\beta}) = \nu\,\text{指数}(-\tilde{\mathbf{x}}^{\mathrm{T}}\boldsymbol{\beta}_{(1)})$	伽马分布$\{\nu, \lambda(\mathbf{x};\boldsymbol{\beta}) - \theta^*\}$	$\left\{\frac{\lambda(\mathbf{x};\boldsymbol{\beta})}{\lambda - \theta^*}\right\}^{\nu}$

也可以通过牛顿–拉夫森算法求解，但其收敛到全局最大值并不保证（Han和Wang，2013年）。为应对这些计算困难，我们采用Gao和Chan（2023年）提出的框架，并开发一种非迭代算法，以避免显式计算拉格朗日乘子，并绕过对牛顿-拉夫森迭代的需求。

该算法通过比较无约束最大似然估计 $\tilde{\boldsymbol{\beta}}$ 与约束最大似然估计量 $\boldsymbol{\beta}$ 的影响函数推导而来。令 $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta})$ 表示式 (6) 中的约束对数似然，且 $\mathbf{S}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta}) = \partial\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\eta})/\partial(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\phi}^{\mathrm{T}},\boldsymbol{\theta}^{\mathrm{T}},\boldsymbol{\eta}^{\mathrm{T}})^{\mathrm{T}}$。正如 Zhang等人 (2020年) 所示，约束最大似然估计量具有如下渐近线性展开

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (\mathbf{I}_d, \mathbf{0}, \mathbf{0}, \mathbf{0})\mathbf{J}^{-1} \cdot n^{-1/2}\mathbf{S}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \mathbf{0}) + o_p(1),$$

其中 $\mathbf{I}_d$ 是 $d \times d$ 单位矩阵。另一方面，仅基于个体参与者数据的最大似然估计具有如下渐近线性展开

$$n^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (-\mathbf{H}_{\boldsymbol{\beta}_0}^{-1}, \mathbf{0}, \mathbf{0}, \mathbf{0}) \cdot n^{-1/2}\mathbf{S}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \mathbf{0}) + o_p(1).$$

比较这两个渐近线性展开，可推导出 $\boldsymbol{\beta}$ 的渐近等价形式：

$$\widetilde{\boldsymbol{\beta}} + (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{H}_{\boldsymbol{\beta}_0}^{-1}\dot{\boldsymbol{\psi}}_{\boldsymbol{\beta}})\mathbf{J}^{-1} \cdot n^{-1}\mathbf{S}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \boldsymbol{\theta}_0, \mathbf{0}).$$

最后，将真实参数替换为初始估计量（$\tilde{\boldsymbol{\beta}}$，$\tilde{\boldsymbol{\phi}}$，$\tilde{\boldsymbol{\theta}}$），并注意到 $\mathbf{S}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\theta}}, \mathbf{0}) = (\mathbf{0}, \mathbf{0}, \mathbf{0}, -n^{-1}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{X}_i; \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\theta}})^{\mathrm{T}})^{\mathrm{T}}$，所提出的非迭代估计量

可表示为

$$\widehat{\boldsymbol{\beta}}_{\text{fast}} = \widetilde{\boldsymbol{\beta}} + (\mathbf{0}, \mathbf{0}, \mathbf{0}, \widetilde{\mathbf{H}}_{\boldsymbol{\beta}_0}^{-1} \widetilde{\dot{\boldsymbol{\psi}}}_{\boldsymbol{\beta}}) \widetilde{\mathbf{J}}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -n^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{X}_i; \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}) \end{pmatrix}, \tag{14}$$

其中 $\widetilde{\mathbf{H}}_{\boldsymbol{\beta}_0} = n^{-1} \sum_{i=1}^{n} \partial^2 \log f(Y_i \mid \mathbf{X}_i; \widetilde{\boldsymbol{\beta}}) / \partial \boldsymbol{\beta}^{\otimes 2}$ 和 $\dot{\boldsymbol{\psi}}_{\boldsymbol{\beta}} = n^{-1} \sum_{i=1}^{n} \partial \boldsymbol{\psi}(\mathbf{X}_i; \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}) / \partial \boldsymbol{\beta}$。与 Gao 和 Chan（2023）的方法相比，我们的方法同时考虑了 AD 不确定性以及个体参与者数据与 AD 数据源之间可能存在的数据集偏移。尽管这一推广需要使用包含涉及 $\kappa \boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{-1}$ 和 $\dot{\boldsymbol{\psi}}_{\boldsymbol{\phi}}$的额外分块的增广矩阵 $\widehat{\mathbf{J}}$，但只需计算一次该扩展矩阵，即可获得更新后的估计量 $\boldsymbol{\beta}_{\text{fast}}$。因此，使用该算法时，我们无需耗费太多额外的计算时间。

对于 $\boldsymbol{\beta}_0$ 和 $\boldsymbol{\phi}_0$，自然的初始估计量分别是无约束最大似然估计 $\widetilde{\boldsymbol{\beta}}$ 和观测到的汇总数据 $\widetilde{\boldsymbol{\phi}}$。对于冗余参数 $\boldsymbol{\theta}_0$，可以通过广义矩估计法（GMM）获得初始估计量 $\widetilde{\boldsymbol{\theta}}$：在 $\boldsymbol{\theta}$ 上最小化 $\left\| \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{X}_i; \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta}) \right\|^2$。值得注意的是，只要初始估计量满足 $n^{1/2}$-一致性，$\boldsymbol{\beta}_{\text{fast}}$ 的渐近分布就不依赖于初始估计量的具体选择。理论上，GMM 程序只要求 $\dim(\boldsymbol{\psi}) \geq \dim(\boldsymbol{\theta})$，即可确保所得估计量具有 $n^{1/2}$-一致性。然而，当 $\dim(\boldsymbol{\psi}) = \dim(\boldsymbol{\theta})$ 时，初始估计量 $\widetilde{\boldsymbol{\theta}}$ 满足 $\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{X}_i; \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\theta}) = 0$，因此 $\boldsymbol{\beta}_{\text{fast}}$ 相较于 $\widetilde{\boldsymbol{\beta}}$没有改进。由此，我们提出的算法要求 $\dim(\boldsymbol{\psi}) > \dim(\boldsymbol{\theta})$ 严格成立。在存在数据集偏移的情形下，如果对偏移函数采用灵活的建模方式，使得 $\dim(\boldsymbol{\theta}_0)$ 较大，相应的参数维度 $\dim(\boldsymbol{\theta})$ 也会增加。在这种情况下，需要更多汇总摘要统计量，才能确保 $\dim(\boldsymbol{\psi}) > \dim(\boldsymbol{\theta})$ 成立。

根据 $n^{-1}\mathrm{S}(\boldsymbol{\beta}_0, \boldsymbol{\phi}_0, \boldsymbol{\theta}_0, 0)$的渐近正态性以及 $\widetilde{\mathbf{H}}_{\boldsymbol{\beta}_0}$, $\dot{\boldsymbol{\psi}}_{\boldsymbol{\beta}}$和 $\widehat{\mathbf{J}}$的一致性，斯卢茨基定理保证了 $\boldsymbol{\beta}_{\text{fast}}$ 与 $\widetilde{\boldsymbol{\beta}}$具有相同的渐近正态性，具体见定理 1。采用所提出的估计量 $\boldsymbol{\beta}_{\text{fast}}$后，渐近方差-协方差矩阵可估计为 $\widehat{\mathbf{J}}^{-1} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{J}}^{-1}$，其中 $\widehat{\mathbf{J}}$是 $\widetilde{\mathbf{J}}$的对应形式，将 $\widetilde{\boldsymbol{\beta}}$ 替换为 $\widehat{\boldsymbol{\beta}}_{\text{fast}}$，并且$\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} \mathrm{S}(\widehat{\boldsymbol{\beta}}_{\text{fast}}, \widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\theta}}, 0)^{\otimes 2}$。对于 $\boldsymbol{\beta}_0$，$100(1-\alpha)\%$置信集构造为 $\{\boldsymbol{\beta} : (\widehat{\boldsymbol{\beta}}_{\text{fast}} - \boldsymbol{\beta})^{\mathrm{T}} \widehat{\mathbf{J}} \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\mathbf{J}} (\widehat{\boldsymbol{\beta}}_{\text{fast}} - \boldsymbol{\beta}) \leq \chi_d^2(1-\alpha)\}$，其中 $\chi_d^2(1-\alpha)$是 χ_d^2分布的第$(1-\alpha)$分位数。所提出的估计方法及相应推断可通过我们开发的 R软件包 GLMaggregate 实现。

5 模拟研究

5.1 不存在数据集偏移时的集成 IPD-AD 分析

我们首先评估在不存在数据集偏移的情况下，约束最大似然估计量用于整合个体参与者数据-汇总数据分析的表现。对于个体参与者数据和汇总数据，协变量 X_1 服从标准正态分布， X_2 服从成功概率为$\mathbb{P}(X_2 = 1) = 0.6$的伯努利分布。响应变量 Y服从均值为$\beta_{00} + \beta_{01}X_1 + \beta_{02}X_2$、方差为 σ_0^2的正态分布，其中$(\beta_{00}, \beta_{01}, \beta_{02}, \sigma) = (0.5, -0.5, 0.5, 1)$。我们考虑不同的汇总数据集，用于估计以下量：

$\boldsymbol{\phi}^Y$: $\mathbb{E}(Y)$;$\boldsymbol{\phi}^{\mathbf{X}|Y}$中位数 : $\mathbb{E}(\mathbf{X} \mid Y \leq q_{0.5})$ $\mathbb{E}(\mathbf{X} \mid Y > q_{0.5})$;$\boldsymbol{\phi}^{\mathbf{X}|Y}$四分位数 : $\mathbb{E}(\mathbf{X} \mid Y \leq q_{0.25})$ $\mathbb{E}(\mathbf{X} \mid q_{0.25} < Y \leq q_{0.5})$ $\mathbb{E}(\mathbf{X} \mid q_{0.5} < Y \leq q_{0.75})$ $\mathbb{E}(\mathbf{X} \mid Y > q_{0.75})$;$\boldsymbol{\phi}_1^{Y|\mathbf{X}}$: $\mathbb{E}(Y \mid -1 < X_1 \leq 1)$; $\boldsymbol{\phi}_2^{Y|\mathbf{X}}$: $\mathbb{E}(Y \mid -1 < X_1 \leq 1)$ $\mathbb{E}(Y \mid X_2 = 1)$;$\boldsymbol{\phi}_3^{Y|\mathbf{X}}$: $\mathbb{E}(Y \mid -1 < X_1 \leq 1)$ $\mathbb{E}(Y \mid X_2 = 1)$ $\mathbb{E}(Y \mid X_2 = 0)$,

其中， q_p 表示 Y的第 p个分位数。汇总数据样本量设为 $N = 1000$。以下模拟结果基于1000次重复模拟。

图 1 展示了约束最大似然估计量的偏差，以及其相对于无约束最大似然估计的相对效率。正如预期，偏差接近于零，并随着个体参与者数据样本量的增加而减小。此外，约束最大似然估计量通常比约束最大似然估计具有更小的标准差，这表明利用汇总数据能够提高效率。此外，比较纳入按协变量分层的结果变量均值 $\widetilde{\boldsymbol{\phi}}_1^{Y|\mathbf{X}}$ 、 $\widetilde{\boldsymbol{\phi}}_2^{Y|\mathbf{X}}$ 和 $\widetilde{\boldsymbol{\phi}}_3^{Y|\mathbf{X}}$ 的约束最大似然估计量的标准差，可以发现，整合更多汇总数据能够提高效率，这与理论结果一致。对于按结果变量分层的协变量均值，基于 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数的约束最大似然估计量优于基于 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数的估计量。虽然 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数 不包含在 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数中，但对结局范围进行更精细的划分，能提供有关协变量与结果变量之间关系的更多详细信息，因此，结合 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 比结合 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数具有更高的效率。

一般而言，结合按结局分层的协变量均值 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数所带来的效率提升，大于结合按协变量分层的亚组结局均值 $\widetilde{\boldsymbol{\phi}}_3^{Y|\mathbf{X}}$ ，但估计离散协变量的系数时除外。效率降低可能是由于 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数的变异性较高，这源于亚组样本量较小。然而，当个体参与者数据样本量足够大（ $n = 400$）且汇总数据样本量增加时，图2

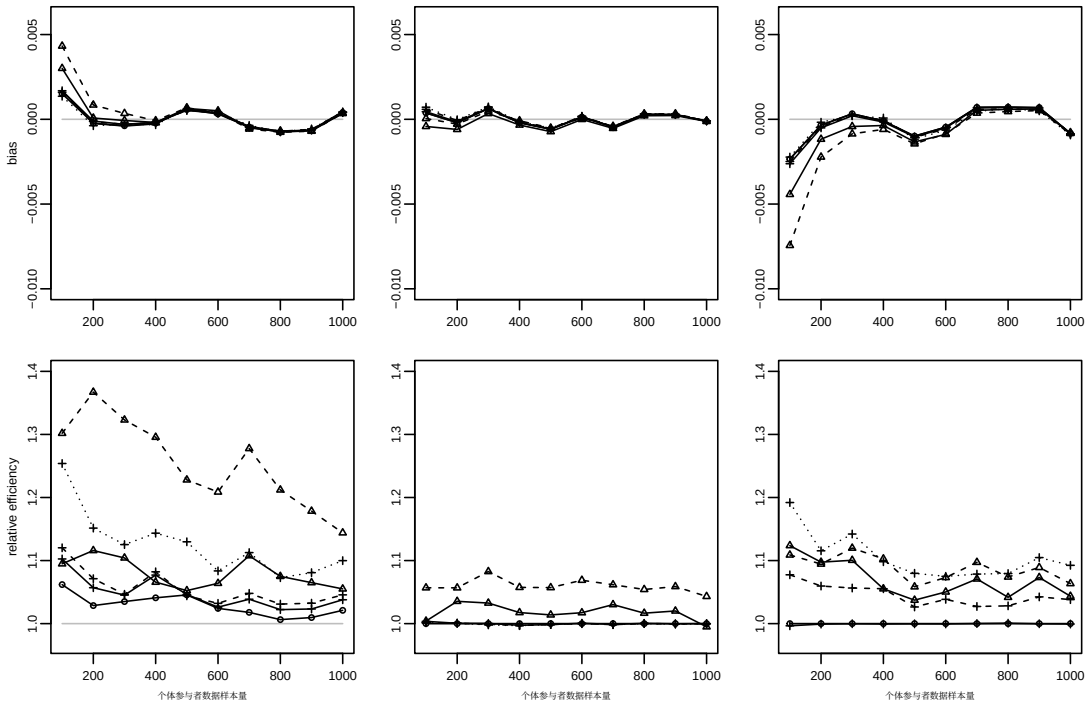

图 1：在不同汇总数据条件下，$\widetilde{\boldsymbol{\phi}}^{Y}$ （带有 ∘的实线）、 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数 （带有 △的实线）、 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 （带有 △的虚线）、 $\widetilde{\boldsymbol{\phi}}_1^{Y|\mathbf{X}}$ （带有 ＋的实线）、 $\widetilde{\boldsymbol{\phi}}_2^{Y|\mathbf{X}}$ （带有 ＋的虚线）和 $\widetilde{\boldsymbol{\phi}}_3^{Y|\mathbf{X}}$（带有 ＋的点线）的约束最大似然估计量在 β_{00} （左）、 β_{01} （中）和 β_{02} （右）下的偏差（上图）和相对效率（下图），汇总数据样本量为$N = 1000$。

显示，纳入 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数的估计量最终优于其他汇总数据。

5.2 数据集偏移下的集成 IPD-AD 分析

接下来，我们考察约束最大似然估计量在协变量偏移或先验概率偏移下的表现。在协变量偏移情形中，外部协变量 $\mathbf{X}^*$ 由密度比模型 $\mathrm{d}G^*(\mathbf{x})/\mathrm{d}G(\mathbf{x}) = \exp(\theta_{01}x_1 + \theta_{02}x_2)/\int\exp(\theta_{01}x_1 + \theta_{02}x_2)\mathrm{d}G(\mathbf{x})$生成，其中 $(\theta_{01}, \theta_{02}) = (0.5,0.5)$。在先验概率偏移情形中，外部结局 Y^* 由密度比模型 $\mathrm{d}F^*(y) = \exp(\theta_0 y)\mathrm{d}F(y)/\int\exp(\theta_0 y)\mathrm{d}F(y)$生成，其中$\theta_0 = 0.5$。此外，我们还考虑与目标参数相对应的汇总数据：

$$\boldsymbol{\phi}^{\mathbf{X}} : \mathbb{E}(\mathbf{X}^*).$$

如第3节所述， $\widetilde{\boldsymbol{\phi}}^{X}$ 并未提供关于 $\boldsymbol{\beta}$ 在协变量偏移下的任何信息，并且几乎不提供关于 $\boldsymbol{\beta}$ 在先验概率偏移较小时的信息。

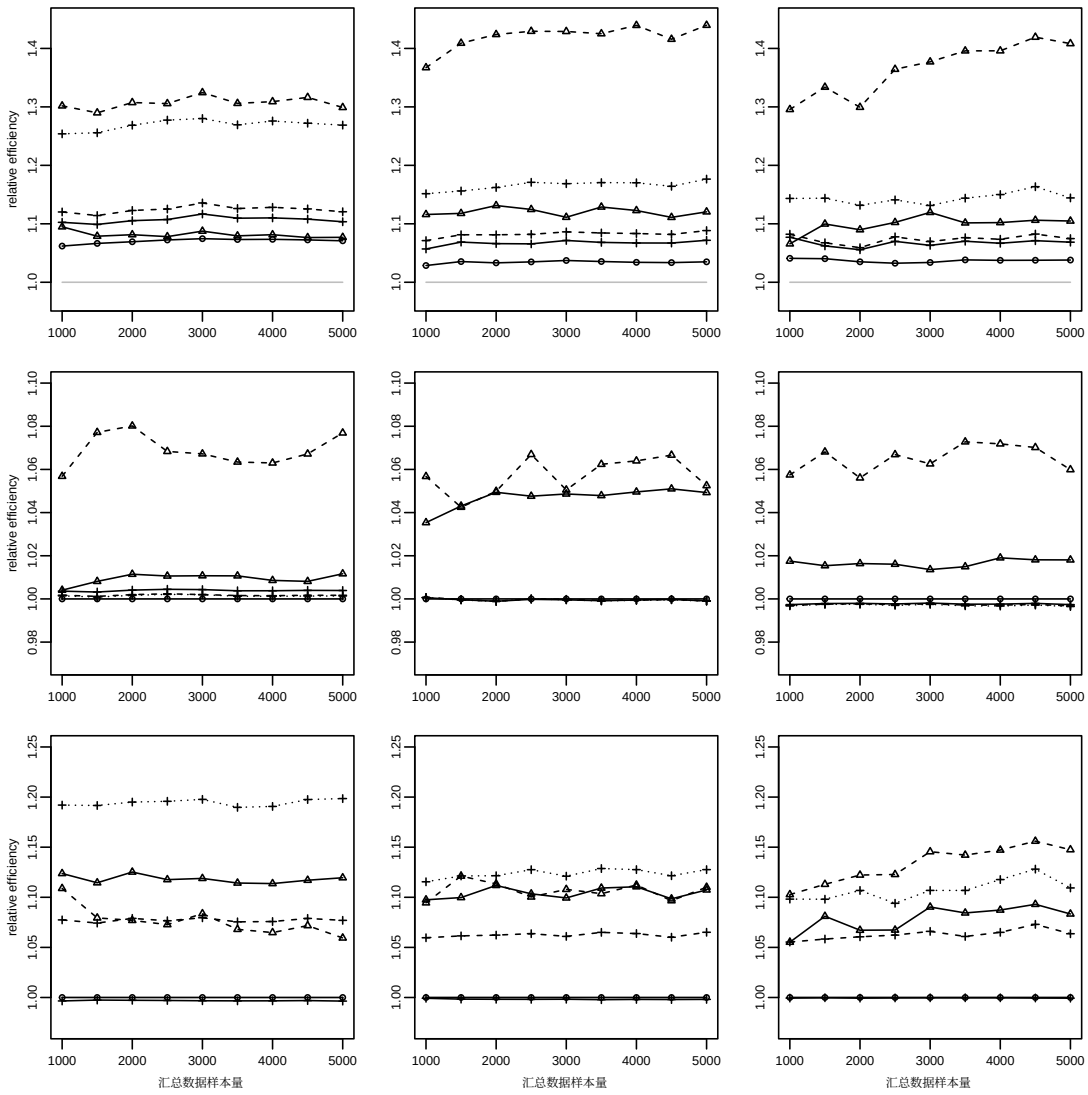

图2：约束最大似然估计量在 β_{00} （顶行）、 β_{01} （中间行）和 β_{02} （底行）下的相对效率，个体参与者数据样本量为 $n = 100$ （左列）、$n = 200$ （中间列）和 $n = 400$ （右列），以及各种汇总数据： $\widetilde{\boldsymbol{\phi}}^{Y}$ （实线，带○），$\widetilde{\boldsymbol{\phi}}\mathbf{x}_{|Y\text{中位数}}$ （实线， △），$\widetilde{\boldsymbol{\phi}}\mathbf{x}_{|Y\text{四分位数}}$ （虚线， △），$\widetilde{\boldsymbol{\phi}}_1^{Y|}\mathbf{x}$ （实线， +），$\widetilde{\boldsymbol{\phi}}_2^{Y|}\mathbf{x}$ （虚线， +），以及 $\widetilde{\boldsymbol{\phi}}_3^{Y|}\mathbf{x}$ （点线， +）。

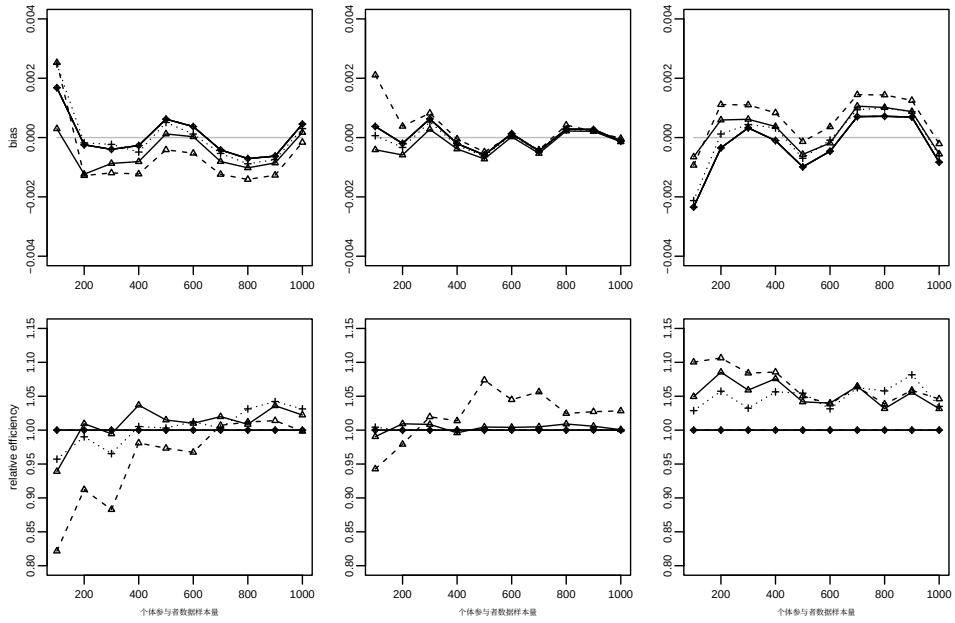

图 3：在协变量偏移下，对于 β_{00} （左）、 β_{01} （中）和 β_{02} （右），采用约束最大似然估计量得到的偏差（上图）和相对效率（下图）；其中汇总数据有多种形式： $\widetilde{\boldsymbol{\phi}}^{Y}$ （带有 ∘的实线）、 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数 （带有 △的实线）、 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 （带有△的虚线）、 $\widetilde{\boldsymbol{\phi}}_1^{Y|}\mathbf{x}$ （带有 +的实线）、 $\widetilde{\boldsymbol{\phi}}_2^{Y|\mathbf{X}}$ （带有 +的虚线）以及 $\widetilde{\boldsymbol{\phi}}_3^{Y|}\mathbf{x}$ （带有 +的点线）；汇总数据样本量为 $N = 1000$。

图 3 （协变量偏移情形）和图4 （先验概率偏移情形）展示了约束最大似然估计量（相对于无约束最大似然估计）的偏差和相对效率。与未发生数据集偏移的情形所得结论一致，大多数偏差接近于零，并且随着样本量增加而减小。然而，在纳入 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}}$ 先验概率偏移情形时，我们观察到显著偏差。尽管矩条件是根据 $\boldsymbol{\beta}$推导出来的，但在这些模拟设置下它非常平坦，导致 $\widetilde{\mathbf{J}_{\mathbf{V}}}$ 近乎奇异且估计值不稳定。因此，即使存在先验概率偏移， $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}}$ 所提供的信息也有限。

在存在数据集偏移的情况下，纳入按结局分层的亚组协变量均值与按协变量分层的亚组结局均值可获得相当的效率，并且优于纳入边际结局均值。当样本量足够大（$n \geq 800$）时，汇总数据 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 和 $\widetilde{\boldsymbol{\phi}}_3^{Y|\mathbf{X}}$ 通常分别在按结局分层的亚组协变量均值和按协变量分层的亚组结局均值中带来更大的效率提升。这证实了约束最大似然估计量的渐近性质。此外，在估计连续协变量的系数时， $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 优于 $\widetilde{\boldsymbol{\phi}}_3^{Y|\mathbf{X}}$ 。

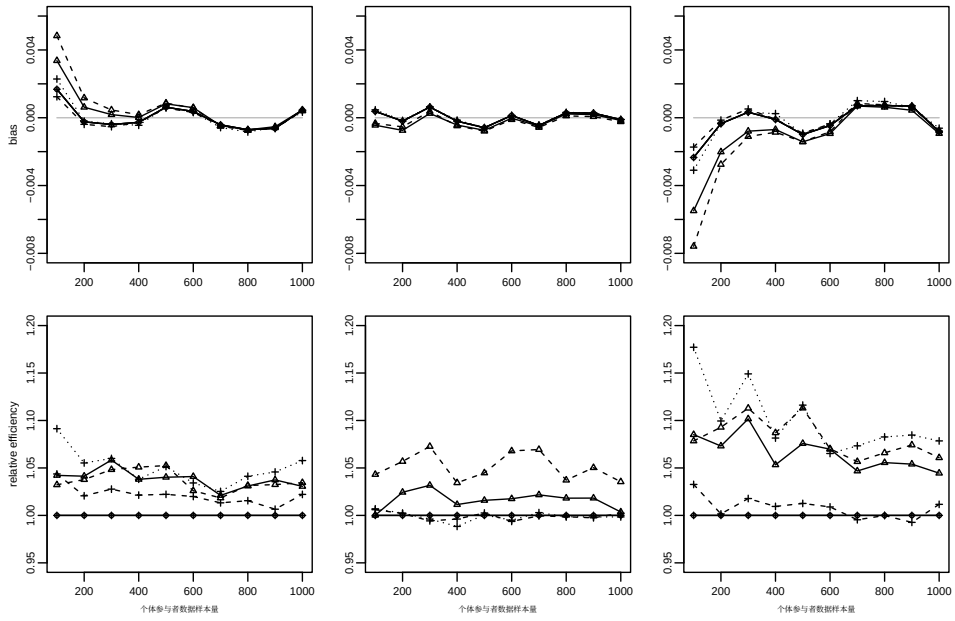

图4：在先验概率偏移下，针对不同汇总数据的约束最大似然估计量的偏差（上图）和相对效率（下图）：β_{00}（左）、β_{01}（中）和β_{02}（右）；$\widetilde{\boldsymbol{\phi}}^{Y}$（带有 ∘的实线）、$\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数（带有 △的实线）、$\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数（带有 △的虚线）、$\widetilde{\boldsymbol{\phi}}_1^{Y|\mathbf{X}}$（带有 +的实线）、$\widetilde{\boldsymbol{\phi}}_2^{Y|\mathbf{X}}$（带有 +的虚线）以及 $\widetilde{\boldsymbol{\phi}}_3^{Y|}\mathbf{X}$（带有 +的点线）；汇总数据样本量为 $N = 1000$。

6 实证示例

6.1 全国纵向调查数据分析

全国青年纵向调查（NLSY）是一项长期研究，追踪一批出生于 1980 年至 1984 年的美国青年。本小节使用该队列中 5095 名受访者的数据，考察人口统计特征对收入的影响（Y，单位为千美元）。对中心化和标准化后的结果变量进行线性回归模型拟合：

$$\tilde{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + \varepsilon,$$

其中，$\tilde{Y} = (Y - 51.5)/33.1$ 和 $\varepsilon \sim N(0, \sigma_S^2)$，识别出若干显著因素：性别（男性对应 $X_1 = 1$，女性对应 $X_1 = 0$）、种族（$(X_2, X_3) = (1, 0)$ 表示西班牙裔，$(X_2, X_3) = (0, 1)$ 表示黑人，$(X_2, X_3) = (0, 0)$ 表示其他种族）、教育程度（大学学位对应 $X_4 = 1$，其他情况对应 $X_4 = 0$）以及居住地（城市对应 $X_5 = 1$，非城市对应 $X_5 = 0$）。为提高估计效率，我们纳入 NLSY 1979 年调查波次的外部辅助数据，并评估哪些类型的汇总数据对效率提升贡献最大。汇总数据样本量为 4052。与模拟研究一致，我们考虑以下类型的汇总数据：

1. $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}}$：协变量的边际均值；2. $\widetilde{\boldsymbol{\phi}}^{Y}$：结果变量（收入）的边际均值；3. $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 中位数：协变量的亚组均值，亚组按结果变量的中位数划分；4. $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数：协变量的亚组均值，亚组按结果变量的四分位数划分；5. $\widetilde{\boldsymbol{\phi}}^{Y|\mathbf{X}}$：结果变量的亚组均值，亚组按协变量的中位数划分。

这些汇总统计量的数值列于表 2。

整合这些汇总数据的一大挑战是存在协变量偏移。在 NLSY97 队列中，男性受访者的比例较高（0.512），高于 NLSY79 队列（0.478）。同样，拥有大学学位的受访者比例在 NLSY97 队列中（0.652）高于 NLSY79 队列（0.546）。我们在 CMLE 框架中对这种协变量偏移进行了处理。分析结果见表3，包括估计值（Est.）、相应的标准误（S.E.）以及相对于最大似然估计的相对效率（R.E.）。所有系数估计值均具有可比性，且相对效率均大于一

表 2：1979年全国纵向调查数据的汇总统计量。 Q_1、 Q_2和 Q_3是结果变量的四分位数， $q_\alpha(\cdot)$ 是第 α分位数函数。

	协变量均值							结果变量均值	
子组	所有	$(0, Q_2]$	(Q_2, ∞)	$(0, Q_1]$	$(Q_1, Q_2]$	$(Q_2, Q_3]$	(Q_3, ∞)	全部	$(0, \quad q_{0.5}(X_j^*)]$ $(q_{0.5}(X^*) \quad \infty)$
								54.82	
性别	0.48	0.38	0.58	0.36	0.40	0.54	0.62	63.97	46.45
西班牙裔	0.19	0.20	0.18	0.18	0.22	0.18	0.17	52.28	55.42
黑人	0.28	0.33	0.23	0.35	0.31	0.26	0.20	46.82	57.97
大学	0.55	0.43	0.67	0.41	0.45	0.59	0.74	64.72	42.91
城市	0.79	0.78	0.80	0.77	0.80	0.79	0.81	55.62	51.80

表3：全国纵向调查数据分析中的回归系数估计值（Est.）、相应的标准误（S.E.）和相对效率（R.E.）

	$\tilde{\boldsymbol{\phi}}^{\mathbf{X}}$			$\tilde{\boldsymbol{\phi}}^{Y}$			$\tilde{\boldsymbol{\phi}}^{\mathbf{X}\|Y}_{\text{中位数}}$			$\tilde{\boldsymbol{\phi}}^{\mathbf{X}\|Y}_{\text{四分位数}}$			$\tilde{\boldsymbol{\phi}}^{Y\|\mathbf{X}}$		
	估计值	标准误	随机效应	估计值	标准误	随机效应	估计值	标准误	随机效应	估计值	标准误	随机效应	估计值	标准误	随机效应
截距	-0.64	0.03	1.23	-0.62	0.04	1.00	-0.64	0.03	1.23	-0.63	0.03	1.36	-0.55	0.02	4.08
性别	0.52	0.02	1.27	0.49	0.03	1.00	0.52	0.02	1.27	0.50	0.02	1.37	0.47	0.02	1.15
西班牙裔	-0.18	0.03	1.26	-0.18	0.03	1.00	-0.18	0.03	1.26	-0.16	0.03	1.38	-0.19	0.03	1.03
黑人	-0.38	0.03	1.30	-0.39	0.03	1.00	-0.38	0.03	1.30	-0.37	0.03	1.35	-0.40	0.03	1.04
大学	0.62	0.02	1.31	0.60	0.03	1.00	0.62	0.02	1.31	0.62	0.02	1.46	0.58	0.02	1.21
城市	0.13	0.03	1.31	0.13	0.03	1.00	0.13	0.03	1.31	0.12	0.03	1.61	0.10	0.03	1.39

在纳入不同类型的汇总数据时。具体而言，结合 $\tilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}_{\text{四分位数}}$ 可使标准误在各种汇总数据类型中最小，体现了所提出方法的有效性。

根据纳入 $\tilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}_{\text{四分位数}}$的分析结果，男性受访者的平均收入比女性受访者高 1.65 万美元（95% 置信区间 [CI]: 15.1–18.0）。拥有大学学位的受访者比没有学位的受访者多赚 2.04 万美元（95% CI: 18.9–21.9）。居住在城市地区的受访者比非城市地区的受访者多赚 4100 美元（95% CI: 2.3–5.8）。西班牙裔受访者的收入少 5400 美元（95% CI: 3.6–7.3），黑人受访者的收入比其他人少 1.22 万美元（95% CI: 10.4–13.9）。

6.2 日本房地产价格分析

在本小节中，我们分析由日本国土交通省收集的大阪房地产交易记录。我们的目标是评估各种因素对大阪房地产市场中房价（Y，单位为千万日元）的影响。所考虑的协变量包括到最近车站的时间（X_1，单位为分钟）、土地面积（X_2，单位为

m^2），建筑面积（X_3，单位为m^2），房龄（X_4，单位为年），宽度（X_5，单位为m），建筑覆盖率（X_6），以及容积率（X_7）。国土交通省定期发布与房价及其相关属性有关的汇总统计量和图表。为探究哪些类型的汇总统计量能够提高评估这些属性效应的效率，我们以2019年包含1565条记录的个体层面数据作为个体参与者数据，并将其与由2018年数据得到的汇总统计量相结合。汇总数据样本量为2506。与我们的模拟研究一致，所考虑的汇总统计量包括响应变量和协变量的边际均值与亚组均值。亚组依据响应变量和协变量的分位数定义。这些统计量的取值列于表4。

正如Huang等人（2024年）所指出的，交易记录仅反映已售出的房产，并不构成整个住房存量的随机样本。具体而言，这些记录可能偏向价格较低的交易，因为高价房屋因潜在买家群体较小而较少出售。因此，在不同时期收集的数据集之间自然会出现先验概率偏移。我们首先对 2019年 的集中化和标准化成交价格拟合线性模型

$$\widetilde{Y} = \theta_0 + \theta_1 X_1 + \cdots + \theta_7 X_7 + \varepsilon,$$

其中 $\widetilde{Y} = (Y - 25.5)/16.3$。这里，$\varepsilon$ 服从均值为零、方差为θ_8^2 的正态分布，并假设其与$\mathbf{X}$相互独立。随后，我们按照第6.1 节所述，在 CMLE 框架中对五组汇总数据中的这种偏移进行处理。分析结果列于表5，包括估计值（Est.）、相应的标准误（S.E.）以及相对于最大似然估计的相对效率（R.E.）。合并不同汇总数据后，各回归系数估计值彼此相近。值得注意的是，整合 $\widetilde{\boldsymbol{\phi}}^{\mathbf{X}|Y}$ 四分位数 通常能在各种类型的汇总数据中得到最小的标准误，但截距以及房龄和建筑覆盖率的系数除外。这一结果凸显了按结果变量分层的协变量均值作为汇总数据形式的实用价值，与我们的模拟研究结果一致。

拟合的模型表明，房屋到最近车站的距离、房龄和容积率会对房价产生负面影响。值得注意的是，步行距离每增加一分钟，平均房价就会降低 31.14 万日元（95% 置信区间 [CI]: 25.72–36.58 万日元）；房龄每增加一年，房价就会降低 48.27 万日元（95% 置信区间: 45.84–50.70 万日元）。相反，土地面积、建筑面积、宽度和建筑覆盖率会对房价产生正面影响。有趣的是，建筑面积每增加一平方米，房价平均上涨 860 万日元（95% 置信区间：650–1,080 万日元）；而土地面积增加相同数值，房价涨幅则较低，为 710 万日元（95% 置信区间: 620–800 万日元）。

表4：2018年日本房价数据的汇总统计量。 Q_1、 Q_2和 Q_3 是结果变量的四分位数， $q_\alpha(\cdot)$是第 α 分位数函数。

	协变量均值							结果变量均值	
子组	全部	$(0, Q_2]$	(Q_2, ∞)	$(0, Q_1]$	$(Q_1, Q_2]$	$(Q_2, Q_3]$	(Q_3, ∞)	全部	$(0, q_{0.5}(X_j^*)]$ $(q_{0.5}(X_j^*) \infty)$
								2.68	
Time	15.30	16.22	14.36	15.86	16.58	15.40	13.31	2.87	2.50
土地面积	1.23	1.11	1.35	0.92	1.30	1.23	1.47	2.23	3.17
建筑面积	1.00	0.92	1.09	0.81	1.02	1.03	1.15	2.20	3.38
Age	18.70	30.76	6.42	39.56	21.99	6.39	6.45	3.59	1.77
宽度	5.51	5.30	5.73	5.11	5.49	5.62	5.84	2.47	2.91
建筑覆盖率	0.60	0.59	0.61	0.60	0.59	0.60	0.61	2.63	3.11
容积率	1.85	1.84	1.86	1.90	1.78	1.86	1.85	2.66	3.07

表5：日本房价数据分析中的回归系数估计值（Est.）、相应的标准误（S.E.）和相对效率（R.E.）

	$\widetilde{\boldsymbol{\phi}}^{\mathbf{X}}$			$\widetilde{\boldsymbol{\phi}}^{Y}$			$\widetilde{\boldsymbol{\phi}}^{\mathbf{X}\|Y}_{\text{中位数}}$			$\widetilde{\boldsymbol{\phi}}^{\mathbf{X}\|Y}_{\text{四分位数}}$			$\widetilde{\boldsymbol{\phi}}^{Y\|\mathbf{X}}$		
	估计值	标准误差	随机效应	估计值	标准误差	随机效应	估计值	标准误差	随机效应	估计值	标准误差	随机效应	估计值	标准误	随机效应
截距	-0.58	0.16	1.01	-0.58	0.16	1.00	-0.65	0.15	1.15	-0.67	0.14	1.21	-0.69	0.14	1.24
Time	-0.02	0.00	1.01	-0.02	0.00	1.00	-0.02	0.00	1.22	-0.02	0.00	1.27	-0.02	0.00	1.20
土地面积	0.44	0.03	1.01	0.44	0.03	1.00	0.43	0.03	1.14	0.44	0.03	1.19	0.44	0.03	1.08
建筑面积	0.53	0.07	1.00	0.53	0.07	1.00	0.51	0.07	1.16	0.53	0.07	1.20	0.54	0.07	1.13
Age	-0.03	0.00	1.01	-0.03	0.00	1.00	-0.03	0.00	1.21	-0.03	0.00	1.25	-0.03	0.00	1.43
宽度	0.03	0.01	1.01	0.03	0.01	1.00	0.03	0.01	1.19	0.03	0.01	1.26	0.03	0.01	1.14
建筑覆盖率	0.80	0.30	1.01	0.80	0.31	1.00	0.89	0.29	1.13	0.87	0.28	1.21	0.93	0.27	1.30
容积率	-0.12	0.05	1.01	-0.12	0.05	1.00	-0.09	0.05	1.15	-0.09	0.05	1.22	-0.11	0.05	1.19

7 结论与讨论

整合式 IPD–AD 荟萃分析为综合多项研究的证据提供了强大的框架，但其效率在很大程度上取决于如何汇总和纳入汇总数据。在统一的 CMLE 框架下，我们比较了几种常见的汇总数据形式，包括结果变量或协变量的边际均值、按协变量分层的结果汇总，以及按结局分层的协变量汇总；比较既涵盖 IPD 与汇总数据来自同一分布的情形，也涵盖二者因数据集偏移而来自不同分布的情形。结果表明，与另外两类汇总方式相比，按结局分层的协变量汇总能够显著提高效率，尤其是在结果变量为连续型时。此外，我们展示了如何通过通用的数据集偏移模型扩展约束最大似然估计量，以处理 IPD 与汇总数据之间的异质性，并开发了一种快速的非迭代算法，可在汇总数据丰富且约束条件众多的高维应用中提升数值稳定性和可扩展性。

这些发现对临床和流行病学研究的设计、报告与整理具有重要意义。在病例对照设计中，二元结局通常会报告按结局分层的协变量汇总，但连续终点的此类汇总则少得多。我们的结果表明，针对连续结局提供此类汇总，可以显著提高后续 IPD–AD 分析的效率。因此，我们建议试验报告和公共数据存储库常规纳入连续终点的按结局分层的协变量汇总，从而支持信息量更丰富、效率更高的证据综合。

本文聚焦于参数模型。未来研究的一个重要方向是将约束似然方法扩展到更一般的半参数设定，例如Delecroix 等人（2003年）提出的单指标模型，以及Li（1991年）提出的有效降维模型。除有限维参数外，这些模型还包含函数型干扰成分，给估计和计算都带来了新的挑战。这些挑战表明，有必要进一步发展适用于半参数模型的约束似然方法，既涵盖存在数据集偏移的情形，也涵盖不存在数据集偏移的情形，并对其理论和数值性质进行严格考察。

参考文献

Ahmed, I., Sutton, A. J. 和 Riley, R. D.（2012）。利用个体参与者数据评估荟萃分析中的发表偏倚、选择偏倚和不可用数据：一项数据库调查。《英国医学杂志》，344(1):d7762。

Chatterje e, N., Chen, Y.-H., Maas, P., 和 Carroll, R. J. (2016年). 约束最大似然 e-

似然估计用于利用外部大数据源的汇总级信息进行模型校准。美国统计学会会刊, 111(513):107–117. Chen, C., Han, P., Chen, S., Shardell, M., 和 Qin, J. (2024年). 在存在先验概率偏移的情况下整合外部汇总信息：在评估原发性高血压中的应用。《*Biometrics*》, 80(3):ujae090. Chen, Z., Ning, J., Shen, Y., 和 Qin, J. (2021年). 在不假设可比性的情况下将主要队列数据与外部汇总信息相结合。《*Biometrics*》, 77(3):1024–1036. Cheng, Y.-J., Liu, Y.-C., Tsai, C.-Y., 和 Huang, C.-Y. (2023年). 在存在总体异质性的情况下利用外部汇总数据对变换模型进行半参数估计。《*Biometrics*》, 79(3):1996–2009. Delecroix, M., H¨ardle, W., 和 Hristache, M. (2003年). 条件单指标回归中的有效估计。多元分析杂志, 86(2):213–226. Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., 和 Saria, S. (2021年). 临床医生与人工智能中的数据集偏移。新英格兰医学杂志, 385(3):283–286. Gao, F. 和 Chan, K. C. G. (2023年). 利用基于总体的辅助信息对半参数模型的回归估计量进行非迭代调整。《*Biometrics*》, 79(1):140– 150. Han, P. 和 Wang, L. (2013年). 缺失数据下的估计：超越双重稳健性。*Biometrika*, 100(2):417–430. Huang, C.-Y., Qin, J., 和 Tsai, H.-T. (2016年). 利用辅助亚组生存信息对 {C}ox模型进行有效估计。美国统计学会会刊, 111(514):787–799. Huang, M.-Y., Qin, J., 和 Huang, C.-Y. (2024年). 先验概率偏移下的高效数据整合。《*Biometrics*》, 80(2):ujae035. Imbens, G. W. 和 Lancaster, T. (1994年). 在微观计量经济学模型中结合微观和宏观数据。经济研究评论, 61(4):655–680. Li, K.-C. (1991年). 用于降维的切片逆回归。美国统计学会会刊, 86(414):316–327.

Qin, J. (1998). 病例对照数据和半参数双样本密度比模型的推断。*Biometrika*, 85(3):619–630. Qin, J. (2000). 结合参数似然与经验似然。*Biometrika*, 87(2):484–490. Qin, J. and Lawless, J. (1994). 经验似然与广义估计方程。《*The Annals of Statistics*》, 22(1):300–325. Qin, J. and Zhang, B. (1997). 基于病例对照数据的逻辑回归模型拟合优度检验。*Biometrika*, 84(3):609–618. Sheng, Y., Sun, Y., Huang, C., and Kim, M. (2022). 总体异质性存在时外部汇总信息的综合：一种惩罚经验似然方法。《*Biometrics*》, 78(2):679–690. Storkey, A. (2008). 当训练集和测试集不同时：表征迁移学习。载于《机器学习中的数据集偏移》，第 3–28 页。麻省理工学院出版社。Sutton, A. J. and Higgins, J. P. T. (2008). 荟萃分析的最新进展。《*Statistics in Medicine*》, 27(5):625–650. Zhang, H., Deng, L., Schiffman, M., Qin, J., and Yu, K. (2020). 利用外部汇总数据改进统计推断的广义整合模型。*Biometrika*, 107(3):689– 703.