arXiv:2303.03497v2 [stat.ME] 6 Feb 2024

利用外部数据的汇总信息，对部分协变量具有复杂非线性效应的整合数据分析

Jia Liang 生物统计学系，圣裘德儿童研究医院；Shuo Chen 生物统计学与生物信息学部，马里兰大学医学院；Peter Kochunov、L. Elliot Hong，马里兰精神病学研究中心，马里兰大学医学院；Chixiang Chen*生物统计学与生物信息学部，马里兰大学医学院 联系邮箱 *: chixiang.chen@som.umaryland.edu

2024年2月7日

摘要

在探索复杂特征的研究中，例如考察与年龄相关的脑功能能力变化时，完整的参数化线性设定可能不足以捕捉复杂模式。另一种选择是采用由参数元素和非参数元素构成的部分线性模型（PLM），其拟合效果可能更好。该模型已广泛应用于经济学、环境科学和生物医学研究。本文提出一种新颖的统计推断框架，通过将外部数据中的汇总信息有效整合到主要分析中，使PLM具备较高的估计效率。这种整合方案能够灵活地吸收来自外部研究的各种简化模型。理论分析表明，所提方法有效且数值计算简便，并且

与PLM中的经典方法相比，该方法可确保获得较高的效率增益。我们还通过两项数据应用进一步验证了该方法，评估了脑成像测量指标和血压的风险因素。

关键词： 信息整合；脑衰老；部分线性模型；效率提升；经验似然。

1 引言

整合相似研究中的信息一直是、并将继续成为统计推断中的一项重要策略（Qin 等人（2022））。当收集丰富且细致变量的个体/本地研究样本量较小时，信息整合方案尤其有用，因为这会导致分析效能不足、准确性降低。鉴于隐私保护和数据保密方面的要求，无需获取原始数据库、借助真实世界数据/研究中的信息，是一种便捷且前景广阔的解决方案。在过去几十年中，当无法获取原始外部数据时，研究人员采用荟萃分析来综合多个数据源的汇总信息（Haidich 2010，Lin 和 Zeng（2010），Liu 等人（2015），Chen 等人（2020））。然而，传统荟萃分析的局限在于，外部研究采用的模型必须与内部研究使用的模型相同（即协变量和参数化方式相同）。为允许不同研究采用不同模型，研究人员从两个方面开发了多种方法：一类基于频率学派方法，例如基于经验似然的估计量（Qin & Lawless 1994，Qin（2000），Qin 等人（2015），Han & Lawless 2019，Sheng 等人（2022），Chen 等人（2022），Chatterjee 等，2016，Zhang 等，2020，Zhai 和 Han（2022））和广义荟萃估计量（Kundu 等人（2019））；另一类基于贝叶斯思想，使用信息先验（Ibrahim 等人（2015），Jiang 等人（2021））。此外，He 等人（2019）和 Zheng 等人（2021）探讨了将信息整合扩展至生存模型。

尽管付出了大量努力，现有的信息整合方法均假设所有协变量效应均处于（广义）线性结构中，因此无法应用于部分协变量具有复杂非线性效应的情境。一个典型例子是实际年龄的非线性效应。具体而言，年龄已被证实是发达国家常见疾病的主要风险因素（Niccoli 和 Partridge（2012）），并且可能与脑结构和生理指标的变化呈非线性关联，例如白质完整性和血压（Bethlehem 等人（2022））。在脑衰老研究中，研究人员关注探究年龄和心血管指标等风险因素对脑结构状况（即白质的微观结构完整性）的影响，而脑结构状况与认知功能衰退相关。在

现有文献中（Lee等人，2022年，Bethlehem 等人（2022）），随着年龄增长，检测到各向异性分数（FA）呈非线性下降趋势。另一个例子可见于高血压研究，其中年龄可能在收缩压（SBP）的进展中发挥复杂作用（Brummett等人，2019年）。因此，采用简约模型来量化这种关系并不明智；相比之下，允许年龄对白质完整性或血压产生非线性效应的半参数模型更为理想（Fan，1993年，Fan& Li 2004）。对于同时包含线性和非线性协变量效应的模型，当内部研究的样本量较小时，统计功效可能会降低。因此，信息整合变得更加不可或缺。然而，尚无文献探索其可行性和实用性。

本文提出了一种新颖的信息整合框架，该框架在部分线性模型（PLM）的背景下，理论上高效且实际易用（H¨ardle 等，2000，Liang，2006年），这是信息整合文献中的首创。所提整合框架建立在剖面最小二乘法的基础上（Fan与Li（2004），Fan等人，2007年），并允许模型中存在非线性效应。仅利用外部研究的汇总数据，即可为内部分析提供大量信息；该外部研究所用模型可以是参数模型（可能存在设定错误），且不同于我们关注的主要PLM。我们证明，信息整合后的估计量在以下任一情况下均能带来显著的效率增益：1）汇总信息可视为真实数据生成机制；或2）汇总信息来自某项外部研究，且其汇总估计值的方差-协方差矩阵已知，或可根据内部数据进行估计。与现有方法不同（Chatterjee 等，2016，Zhang 等，2020），所提方法无需两项研究具有相同的协变量分布即可确保无偏估计，因此在实践中对异质数据集更具稳健性。此外，我们展示了如何最好地利用外部信息，并提出一种渐近等价估计量，从而实现计算负担较轻的稳定算法；当内部研究的样本量较小时，这一点在实践中尤为重要。为证明所提方法相较于其他方法的优势，我们基于计算机模拟开展了广泛的数值评估，并将所提方法应用于两个现实世界问题；其中一个问题的情形是，原始外部数据可能不便共享，但汇总数据可以在两个研究团队之间轻松传递。

但汇总数据可以在两个研究团队之间轻松传递。

本文其余部分安排如下。第 2 节介绍所提出的信息整合框架。第3 节通过计算机模拟对所提出的估计量进行数值评估。第 4节展示基于两个真实世界数据对所提出估计量的评估。第 5节讨论我们方法的扩展。所有技术证明、数值计算过程和额外模拟均见补充材料。

2 方法

2.1 符号与方法流程

在介绍方法之前，我们先说明内部研究和外部研究的基本设定。假设内部研究中有 n 名独立受试者。对于每位受试者 $i = 1, \ldots, n$，令 $(Y_i, \tilde{\mathbf{X}}_i^T, \mathbf{O}_i^T)^T$ 表示个体层面数据，其中 Y_i 是连续尺度上的结果变量，$\mathbf{X}_i$ 是文献中已有充分认识的协变量向量，$\mathbf{O}_i$ 是文献中尚未探讨的额外协变量向量。内部研究的主要目标是拟合结果变量 Y_i 的条件均值模型，即 $E(Y_i|\mathbf{X}_i, \mathbf{O}_i)$。此外，我们假设存在一个 d 维向量 $\tilde{\mathbf{Z}}_i$，它是向量 $\mathbf{X}_i$ 或向量 $\mathbf{O}_i$ 的子集（见补充材料中的图S1），对结果变量具有复杂的非线性效应，而其余协变量具有线性效应，即 $E(Y_i|\tilde{\mathbf{X}}_i, \mathbf{O}_i) = m_0(\mathbf{Z}_i) + \mathbf{X}_i^T\boldsymbol{\beta}_0$，其中 $\mathbf{X}_i$ 是 $(\tilde{\mathbf{X}}_i^T, \mathbf{O}_i^T)^T$ 中排除向量 $\mathbf{Z}_i$ 所含协变量后的 p 维向量，$\boldsymbol{\beta}_0$ 是 $\mathbf{X}_i$ 的真实系数向量，$m_0(\cdot)$ 是一个真实但未知的连续有界可测函数，且其海森矩阵有界。由于非参数估计存在维数灾难，在许多应用中，我们会考虑标量 Z_i（Wand & Jones 1994）。例如，在第4.2节的真实数据示例中，Z_i 变量为年龄，它可能与血压存在复杂关联。此外，$\mathbf{X}_i$ 向量包含脑部39个区域的平均FA、体质指数、性别和种族变量，这些变量在两个数据集中均可获得；而向量 $\mathbf{O}_i$ 包含极低密度脂蛋白胆固醇，该变量仅在内部数据中可获得。对于

为开发方法，我们定义堆叠变量$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$、$\mathbf{X} = (\mathbf{X}_1^T \ldots, \mathbf{X}_n^T)^T$和$\mathbf{Z} = (\mathbf{Z}_1^T \ldots, \mathbf{Z}_n^T)^T$。

汇总信息可以是真实数据生成机制，也可以是从外部研究中估计得到的。对于后一种情况，假设存在一项外部研究，关注相同的结果变量 Y_j 以及 $\mathbf{X}j$ 中公认的协变量（这里，$j = 1, \ldots, N$ 表示外部研究中的受试者索引）。此外，假设该外部研究已为结果变量拟合了一个参数条件均值模型，即 $E(Y_j|\mathbf{X}j)$。令 $\mathbf{H}(Y_j, \mathbf{X}j; \boldsymbol{\theta})$ 表示外部研究中用于拟合由 $\boldsymbol{\theta}$ 参数化的均值模型的估计函数（例如得分函数）。本文考虑外部研究采用广义线性模型（GLM）（Liang & Zeger 1986）的情形，该模型在科学文献中得到广泛应用。函数形式的细节将在第 2.3.1 节介绍。令由此得到的估计值为 $\boldsymbol{\theta}$，其估计的方差-协方差矩阵为 V。为便于说明，我们进一步假设外部数据和内部数据均来自同一总体，因此 $E\{\mathbf{H}(Y, \mathbf{X}; \boldsymbol{\theta}_0)\} = \mathbf{0}$，其中期望是针对内部数据计算的，而 $\boldsymbol{\theta}_0$ 是 $\boldsymbol{\theta}$ 的极限值。关于外部模型和异质性总体的更多讨论见第 2.3.2 节和第 5 节。我们注意到，由于内部 PLM 是真实模型，因此基于 GLM 的外部模型存在错误设定。然而，我们认为，错误设定的外部模型仍有助于改进主要分析；这一观点受到信息整合现有文献的启发（Chatterjee 等，2016，Han & Lawless 2019，Zhang 等，2020）。本文的主要目标是利用外部研究的汇总信息，提高条件均值模型 $E(Y_i|\mathbf{X}_i, \mathbf{O}_i)$ 中估计的效率。

2.2 剖面最小二乘（PLS）

我们首先介绍基于 PLS 的众所周知的估计方法。对于来自内部研究的受试者 $i = 1, \ldots, n$，考虑以下 PLM：

$$Y_i = m(\mathbf{Z}_i) + \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{2.1}$$

其中，ε_i 是均值为零且方差有限的残差。我们不对残差施加任何分布假设。与主文施加的线性效应不同

协变量，$\mathbf{X}_i$，PLM 中的协变量$\mathbf{Z}_i$ 对结果变量具有非线性效应。我们假设 $m(\mathbf{Z}_i)$ 的确切形式未知，因此，由于干扰参数具有无限维数，传统回归会遇到困难（Fan 1992）。另一方面，PLS 是一种广受欢迎的估计方法，用于处理（2.1）中的 PLM（Fan & Li2004）。具体而言，PLS 方法首先假设参数 $\boldsymbol{\beta}$ 已知。在此假设下，通过定义 $Y_i^* = Y_i - \mathbf{X}_i^T\boldsymbol{\beta}$，模型（2.1）可改写如下：

$$Y_i^* = m(\mathbf{Z}_i) + \varepsilon_i. \tag{2.2}$$

在参数已知 $\boldsymbol{\beta}$的情况下，式（2.2）中唯一未知的量变为 $m(\mathbf{Z}_i)$，可通过非参数方法进行估计。本文采用局部线性回归（LLR）技术来估计该未知函数（Fan & Gijbels1996），但并不限于此。给定$\mathbf{Z}$ 的取值后，LLR 通过最小化以下平方和来估计参数 a_0 和$\mathbf{a}_1$：

$$\sum_{i=1}^{n}\{Y_i^* - a_0 - \mathbf{a}_1^T(\mathbf{Z}_i - \mathbf{z})\}^2 K_{\mathbf{B}}(\mathbf{Z}_i - \mathbf{z}), \tag{2.3}$$

其中，$K_{\mathbf{B}}(\mathrm{u}) = |\mathrm{B}|^{-1/2}K(\mathrm{B}^{-1/2}\mathrm{u})$。这里，带宽矩阵 B 的形式为$Diag\{b^2\}_{d\times d}$，而 $K(\cdot)$ 是满足补充材料中假设（M1）的单峰光滑函数。这里我们指出，参数 a_0 是 $m(\mathbf{z})$的有效近似（Fan，1993年）。因此，通过求解（2.3），可得估计向量 $\hat{}\mathrm{m}(\mathrm{z}\ \boldsymbol{\beta}) = \mathrm{S}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta})$，其中 S 是局部线性平滑器，且仅取决于矩阵 Z。S 的显式形式见补充材料中的（2.11）。我们还指出，只有在已知 $\boldsymbol{\beta}$ 的情况下，才能计算 m(Z) 的估计值。为此，我们将 $\boldsymbol{\beta}$替换为 $\hat{}\mathrm{m}(\mathrm{Z}\ \boldsymbol{\beta})$，从而改写式（2.1）中的模型，即：

$$(\mathbf{I} - \mathbf{S})\mathbf{Y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.4}$$

其中 $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$；I 是 n 阶单位矩阵。由于 S 不含未知参数 $\boldsymbol{\beta}$，我们可以通过应用最小二乘法轻松计算出 $\boldsymbol{\beta}$ 的估计值。最后，我们推导出 PLS估计量的显式形式，即 $\boldsymbol{\beta}_{pls} = \{\mathrm{X}^T(\mathrm{I} - \mathrm{S})^T(\mathrm{I} - \mathrm{S})\mathrm{X}\}^{-1}\mathrm{X}^T(\mathrm{I} - \mathrm{S})^T(\mathrm{I} - \mathrm{S})\mathrm{Y}$。利用 $\hat{}\ \boldsymbol{\beta}_{pls}$ 估计量，我们还可以进一步得到 m(Z) 的估计值，即 $\mathrm{S}(\mathrm{Y} - \mathrm{X}\boldsymbol{\beta}_{pls})$。

在结束本节之前，我们在此指出，平滑矩阵S 所用的带宽 b 可能对$\mathbf{m}(\mathbf{Z}\ \boldsymbol{\beta}_{pls})$的精度影响较大，但对 $\boldsymbol{\beta}_{pls}$的精度影响较小。

在实际应用中，我们建议使用基于差分的估计（Fan& Huang2005），并通过交叉验证选择带宽。更多数值评估见第 3节和补充材料。

2.3 半参数信息整合

从上一节的讨论中，我们了解到，PLS估计量涉及由非参数核函数构造的平滑矩阵 $\mathbf{S}$ ，在小样本量下会面临效率低的问题。作为PLS的一种潜在改进方法，整合外部数据中的信息有望显著降低估计不确定性并使推断更加稳定。本节旨在从两个方面提出一种有效且快速的PLS框架下的信息整合方案：(1) 假设汇总信息 $\boldsymbol{\theta}_0$ 是固定的，并视为真实数据生成机制，或 (2) 汇总信息 $\boldsymbol{\theta}$ 由外部研究估计得到，且可获得估计的方差-协方差矩阵 $\mathbf{V}$ 。

2.3.1 给定真实汇总信息

在本小节中，我们考虑这样一种情形：所获得的汇总信息是 $\boldsymbol{\theta}_0$真实数据生成机制。这种情形可能出现在汇总信息提取自一个完整的全国性数据库时，因此我们相信它能够准确代表研究人群的真实数据生成机制。许多数据整合研究中都能见到这种设定（Chatterjee 等，2016，Han&Lawless2019）。在此设定下，我们考虑一种新的加权PLS估计方法，通过基于伪结局 $(\mathbf{I}-\mathbf{S})\mathbf{Y}$ 对（2.4）中的估计方程进行重新加权。假设我们有一个对角权重矩阵 $\mathbf{P}_1$，其构造将在下文说明。那么，（2.4）中估计方程的加权形式可写为

$$\mathbf{X}^T(\mathbf{I}-\mathbf{S})\hat{\mathbf{P}}_1\{(\mathbf{I}-\mathbf{S})\mathbf{Y}-(\mathbf{I}-\mathbf{S})\mathbf{X}\boldsymbol{\beta}\}=\mathbf{0}.$$

通过求解 通过上述方程，我们由此得到一个新的闭式估计量 ：

$$\hat{\boldsymbol{\beta}}_{ib1}=\{\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\hat{\mathbf{P}}_1(\mathbf{I}-\mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\hat{\mathbf{P}}_1(\mathbf{I}-\mathbf{S})\mathbf{Y}. \tag{2.5}$$

显然，要成功利用外部数据传递信息，权重矩阵 $\hat{\mathbf{P}}_1$ 不能是任意的。理想的权重应包含来自外部数据的额外信息。为此，我们借鉴了经验似然的思想，该方法最初旨在通过非参数似然提高估计效率（Qin& Lawless 1994，Chen 等人（2022））。具体而言，我们要求 $\hat{}$ i在对角线上的第 th 个元素 $\hat{}p_i$ ，即 $\hat{}\mathbf{P}_1$ 的该元素，通过最大化 $\prod_{i=1}^{n} p_i$ 关于 p_i 并满足以下三个约束来计算：

$$p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{H}(Y_i, \tilde{\mathbf{X}}_i; \boldsymbol{\theta}_0) = \mathbf{0}. \tag{2.6}$$

前两个约束是似然的基本特征，而第三个约束涉及一个包含真实信息的“工作”估计函数 $\boldsymbol{\theta}_0$。例如，如果$\boldsymbol{\theta}_0$ 是线性条件均值模型中的真实参数，则该“工作”函数可以是 $\mathbf{H}(Y_i, \tilde{\mathbf{X}}_i; \boldsymbol{\theta}_0) = \tilde{\mathbf{X}}_i(Y_i - \tilde{\mathbf{X}}_i^T \boldsymbol{\theta}_0)$。如果 θ_0 是外部研究中的结果均值，则该“工作”函数可以是 $H(Y_i, \mathbf{X}_i; \theta_0) = (Y_i - \theta_0)$。

现在，我们直观地说明为什么所提出的权重可能有效。直观而言，由于 $\boldsymbol{\theta}_0$ 是来自外部来源的真实值，因此“工作”函数 $\mathbf{H}(\cdot; \boldsymbol{\theta}_0)$ 在（2.6）中变为与参数无关的函数，因而在额外自由度的作用下形成过度识别。因此，基于经验似然理论（Qin & Lawless 1994、Chen 等人（2022）、2023），$\hat{}$ 矩阵 $\mathbf{P}_1$ 中的结果权重 p_i 将意味着对数据$\tilde{}$分布 $\{Y_i, \mathbf{X}_i\}$ 进行比经验分布更高效的估计。与简单的等权重（1 或 $1/n$）相比，将更高效的估计值作为权重纳入主要估计方程，可能带来额外信息，从而有助于主要参数估计。

我们在第2.4 节已证明， $\boldsymbol{\beta}_{ib1}$ 的估计变异性小于 $\boldsymbol{\beta}_{pls}$ 的估计变异性，同时保持估计一致性。获得效率更高的估计量 $\boldsymbol{\beta}_{ib1}$后，非参数成分$\mathbf{m}(\mathbf{Z})$ 的估计值可更新为$\mathbf{m}(\mathbf{Z}\ \boldsymbol{\beta}_{ib1}) = \mathbf{S}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{ib1})$。

2.3.2 使用估计的汇总信息

式（2.5）中的估计量是通过假设汇总信息等于底层真实值 $\boldsymbol{\theta}_0$构造的。然而，当汇总信息

ˆ需要估计（即 $\boldsymbol{\theta}$），且两项研究的样本量相近时，这种估计量可能无法奏效（数值证据参见第3节）。理想的信息整合方案应将 $\boldsymbol{\theta}$ 的估计不确定性考虑在内。为此，除了考虑˜数据 Y_i 和 $\mathbf{X}_i$ 的分布（见式（2.6））之外，自然还应考虑纳入 $\boldsymbol{\theta}$ 分布的联合分布。具体而言，将观测数据 $\mathbf{Y}$、$\mathbf{X}$ 和 $\boldsymbol{\theta}$ 的联合分布记为 $F(\mathbf{Y}, \mathbf{X}\ \boldsymbol{\theta}) = F(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta})F(\boldsymbol{\theta})$。与式（2.6）类似，条件分布 $F(\mathbf{Y}, \mathbf{X}|\boldsymbol{\theta})$ 可通过概率质量 $\prod_{i=1}^{n} p_i$ 的乘积进行半参数建模；$F(\boldsymbol{\theta})$ 是均值为 $\boldsymbol{\theta}$ˆ、方差矩阵为 $\mathbf{V}$ 的多元正态分布。这一设定的依据是 $N^{0.5}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ 的渐近正态性（McCulloch与Searle（2004））。这里，$\boldsymbol{\theta}$ 是 $\hat{\boldsymbol{\theta}}$ 的极限值。因此，对于 $i = 1, \ldots, n$，Y_i、$\mathbf{X}_i$ 和 $\boldsymbol{\theta}$ 的对数似然函数可定义为

$$l = \sum_{i=1}^{n} log(p_i) - (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T (\hat{\mathbf{V}})^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})/2, \tag{2.7}$$

其中 p_i 可通过对 p_i 和 $\boldsymbol{\theta}$最大化 l 求解，并且需满足以下三个约束：

$$p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \mathbf{H}(Y_i, \tilde{\mathbf{X}}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{2.8}$$

˜ ˆ 函数 $\mathbf{H}(Y_i, \mathbf{X}_i; \boldsymbol{\theta})$ 在第 2.3.1 节中定义。此外，令 $\mathbf{P}_2$ 为一个对角矩阵，其第 i个元素为通过上述优化求得的 $\hat{p}_i$。那么，$\boldsymbol{\beta}$ 的新估计量定义为

$$\hat{\boldsymbol{\beta}}_{ib2} = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T \hat{\mathbf{P}}_2 (\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1} \mathbf{X}^T (\mathbf{I} - \mathbf{S})^T \hat{\mathbf{P}}_2 (\mathbf{I} - \mathbf{S})\mathbf{Y}. \tag{2.9}$$

详细的数值解请参阅补充材料p_i。直观而言，新估计量继承了式（2.5）中估计量的性质，并通过指定一个半参数联合似然，进一步考虑 $\boldsymbol{\theta}$ 的不确定性，该似然涉及˜ 内部数据 $\{\mathbf{Y}、\mathbf{X}\}$ 和 $\boldsymbol{\theta}$。这一做法颇具新意，不同于Zhang 等人（2020）提出的框架。例如，在我们的真实数据应用（第4.1节）中，汇总数据 $\boldsymbol{\theta}$ 是通过将各向异性分数对脉搏率、体质指数和性别进行回归而得到的估计系数。随后，我们使用估计系数 $\boldsymbol{\theta}$，通过应用式（$\boldsymbol{\beta}_{ib2}$ 2.9）中的公式得到。当外部研究中的汇总信息需要估计时，我们预计 $\boldsymbol{\beta}_{ib2}$ 相比 $\boldsymbol{\beta}_{ib1}$ 表现更好，接下来的两节将展示这一点。

作为 $\boldsymbol{\beta}_{ib2}$性能更优所付出的代价是，在计算 $\boldsymbol{\theta}$ 之前，必须先通过约束优化估计参数向量 $\boldsymbol{\beta}_{ib2}$。第2.4节中，我们给出一种计算上更简便的方式来估计 $\boldsymbol{\theta}$，与标准的经验似然估计流程相比，这种方法更方便（Qin &Lawless 1994）。本节假设整个协方差矩阵$\mathbf{V}$均可用以作说明，也允许将其推广至仅部分可用的$\mathbf{V}$ 。更多扩展请参阅第5 节。

再次强调其独特性。 在进行理论研究之前，我们再次强调，所提出的重新加权估计量在现有文献中具有独特性，并且具备若干优势。该方法不同于经典荟萃分析（Haidich 2010），后者假设外部研究中的模型与内部研究中的模型具有相同形式；该方法也不同于广义整合模型（GIM）（Chat-terjee et al. 2016、Zhang 等，2020）、广义荟萃方法（Kundu 等人（2019））以及信息先验方法（Jiang 等人（2021）），因为这些方法仅考虑了广义线性模型。我们的方法也不同于经典经验似然框架：它将主要估计与经验似然解耦为两个步骤，从而得到主要估计量的解析形式。这种便捷且稳定的计算尤其适用于样本量较小的内部研究以及复杂的模型结构，例如 PLM。此外，与现有方法相比，所提出的估计量对异质总体不那么敏感，因为从理论上讲，我们的方法仅要求两个数据集之间具有~相同的条件均值，即可确保 $E\{\mathbf{H}(Y,\ \boldsymbol{\theta}_0\mathbf{X};\ )\} = \mathbf{0}$，而一些方法（例如 GIM）则要求更强的假设，即协变量分布相同。因此，所提出的方法在文献中具有独特性，并能够在 PLM（一种半参数框架）中实现信息整合，同时具备快速且稳定的计算性能。

2.4 渐近性质

本节详细阐述了式（2.5）和（2.9）中提出的两个估计量的渐近性质。基本上，我们证明了两点：1）两个估计量（$\boldsymbol{\beta}_{ib1}$ 和 $\boldsymbol{\beta}_{ib2}$）均具有预言机收敛速率，即 O_P（$\mathbf{n}^{-0.5}$）；2）当汇总信息等于真实数据生成机制时，式（ $\boldsymbol{\beta}_{ib1}$ 2.5）中的估计量比 PLS 估计量更有效率；而式（ $\boldsymbol{\beta}_{ib2}$ 2.9）中的估计量则能保证方差降低，即使汇总信息

是由存在不可忽略估计变异性的外部研究估计得到的。记号$m_0(\mathbf{z})$和 $\boldsymbol{\beta}_0$ 分别表示向量$\mathbf{z}$的真实回归函数和 $\boldsymbol{\beta}$的参数向量。技术性证明见补充材料。

ˆ 我们首先在汇总信息等于真实数据生成机制 $\boldsymbol{\theta}_0$ 的设定下，展示估计量 $\boldsymbol{\beta}_{ib1}$ 的渐近性质。

定理 2.1 在补充材料中的正则条件下，且给定ˆ真实汇总信息 $\boldsymbol{\theta}_0$，估计量 $\boldsymbol{\beta}_{ib1}$ 在（2.5）中被证明是 $\boldsymbol{\beta}_0$ 的一致估计量，并且满足

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ib1} - \boldsymbol{\beta}_0) \to \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}^{-1}(\mathbf{G} - \mathbf{M}\boldsymbol{\Omega}\mathbf{M}^T)(\boldsymbol{\Sigma}^T)^{-1}) \tag{2.10}$$

依分布收敛，其中 $\boldsymbol{G} = E[\{\boldsymbol{X}_i - E(\boldsymbol{X}_i|\boldsymbol{Z}_i)\}E\{\boldsymbol{X}_i - E(\boldsymbol{X}_i|\boldsymbol{Z}_i)\}^T\varepsilon_i^2]$; $\boldsymbol{M} = E[\{\boldsymbol{X}_i - E(\boldsymbol{X}_i|\boldsymbol{Z}_i)\}\varepsilon_i\boldsymbol{H}_i^T]$; $\boldsymbol{\Omega} = \{E(\boldsymbol{H}_i\boldsymbol{H}_i^T)\}^{-1}$; $\boldsymbol{\Sigma} = E[\{\boldsymbol{X}_i - E(\boldsymbol{X}_i|\boldsymbol{Z}_i)\}\{\boldsymbol{X}_i - E(\boldsymbol{X}_i|\boldsymbol{Z}_i)\}^T]$, $b = sn^{-a}$，其中常数 a 介于 *1/8* 和 *1/(2d)* 之间。

注意，ˆ $\boldsymbol{\beta}$ 的收敛速度为 $O_P(\mathbf{n}^{-0.5})$，且 $\mathbf{M}\boldsymbol{\Omega}\mathbf{M}^T$ 通常为正定。因此，从外部研究借用信息后，所得估计量的估计效率有所提高。方差降低源于纳入了真实汇总信息 $\boldsymbol{\theta}_0$。我们在此强调，上述定理在汇总信息等于真实数据生成机制 $\boldsymbol{\theta}_0$时成立；若不相等，该估计量仍然一致，但（2.10）中推导出的方差将低估真实值（Han & Lawless 2019, Zhang 等，2020）。下一个定理表明，当需要估计ˆ汇总信息时，（2.9）中的估计量始终能够保证效率提高，即使外部数据的样本量小于内部样本数据的样本量也是如此。

定理 2.2 在与定理 *2.1* 相同的条件下，给定估计得到的汇总信息ˆ ˆ $\boldsymbol{\theta}$，估计量 $\boldsymbol{\beta}_{ib2}$ 在（*2.5*）中是 的相合估计量，并且服从 $\boldsymbol{\beta}_0$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ib2} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma}^{-1}(\mathbf{G} - \rho\mathbf{M}\boldsymbol{\Omega}\mathbf{M}^T)(\boldsymbol{\Sigma}^T)^{-1}) \tag{2.11}$$

在分布中，其中 $\rho = \text{极限}_{n,\ N\to\infty}\, N/(n+N)$ 是满足以下条件的常数 $0 < \rho < 1$；$\mathbf{G}$、$\mathbf{M}$、$\boldsymbol{\Omega}$和 $\boldsymbol{\Sigma}$ 在定理 *2.1*中定义； $b = sn^{-a}$ 其中某个常数 $1/8 < a < 1/(2d)$。当 n和 N 均趋于无穷大时，上述结论成立。

如果外部样本量 N 远大于内部样本量 n，则有$\rho \approx 1$；因此，（2.11）中的方差将降至（2.10）中的方差。因此，估计量 $\hat{\boldsymbol{\beta}}_{ib1}$ 也可以看作外部样本量远大于内部样本量时的一种特殊情形。另一方面，当外部样本量较小时，$\hat{}$ 在数值上不能忽略 $\boldsymbol{\theta}$的不确定性，而使用定理 2.1 将导致低估真实变异性。

$\hat{}$ 尽管该估计量 $\boldsymbol{\beta}_{ib2}$ 在理论上具有优势，但它需要在经验似然框架中联合估计 $\hat{}p_i$ 和$\boldsymbol{\theta}$，这需要复杂的计算策略。作为替代方案，我们建议使用 $\boldsymbol{\theta}$ 的代入估计，以避免繁重的计算；该估计由以下推论定义并验证。

推论 *2.1* 通过式 *(2.8)* 中的约束优化得到的干扰参数 $\boldsymbol{\theta}$ 的经验似然估计值，在渐近意义下等价于基于两项研究的荟萃分析所得的估计值，即通过最小化 $\sum_{i\in\{0,1\}}(\boldsymbol{\theta}-\hat{}\boldsymbol{\theta}_{(i)})^T\hat{}V^{-1}{}_{(i)}(\boldsymbol{\theta}-\hat{}\boldsymbol{\theta}_{(i)})$ 得到，$\hat{}\hat{}\hat{}\hat{}\hat{}\hat{}$其中 $\boldsymbol{\theta}_{(1)}=\boldsymbol{\theta}$，$V^{(1)}=V=Cov(\boldsymbol{\theta})$；$\boldsymbol{\theta}_{(0)}$ 是根据内部数据，通过估计方程$\sum_{i=1}^{n}H(Y_i,\tilde{}X_i;\boldsymbol{\theta})=0$ 求得的估计值，而 $\hat{}V^{(0)}$ 是 $\boldsymbol{\theta}_{(0)}$的方差-协方差矩阵的一致估计。

该推论意味着，我们可以通过汇总汇总信息，而非通过式（2.8）中的约束优化求解经验似然，来降低计算 $\boldsymbol{\theta}$ 估计值的计算负担。该估计量计算简单便捷，无需按照经验似然的标准流程联合估计 $\boldsymbol{\theta}$ 和 p_i 。因此，通过荟萃分析使用此类估计量可以实现可扩展计算，进一步拓宽我们方法的应用价值和适用性。其实现十分简单：通过荟萃分析获得 $\boldsymbol{\theta}$ 估计值后，我们将该估计值视为固定值，并代入式（2.7）和（2.8）中，以计算 $\hat{}p_i$（参见补充材料第 2.1 节）。我们还指出，如果可以获得原始外部数据，则可以将外部数据与内部数据合并，以估计 $\boldsymbol{\theta}$；并且很容易证明，该估计值在渐近意义上也等价于经验似然估计值 $\boldsymbol{\theta}$。此处省略证明。不过，从仅需外部研究的汇总信息这一点来看，荟萃分析这一替代方法更可取；这类信息易于获取、通常会在文献中加以描述，并且可在出版物或

从合作者处获取。

ˆ 以下推论表明，在所有估计值类别中，我们的估计量 $\boldsymbol{\beta}_{ib2}$ 对外部知识的利用最充分。

ˆ**推论 2.2** 估计量 $\boldsymbol{\beta}_{ib2}$ 是所有一致估计量类别中的最佳估计量 ˆ $\boldsymbol{\beta} = \hat{}\boldsymbol{\beta}_{pls} - \mathbf{C}(\hat{}$ $\boldsymbol{\theta}_{(0)} - \hat{}\boldsymbol{\theta}) + o_p(\mathbf{n}^{-1/2})$，其中常数矩阵 $\mathbf{C}$；ˆ$\boldsymbol{\theta}$ 是来自外部研究的估计ˆ参数向量； $\boldsymbol{\theta}_{(0)}$ 定义见推论 *2.1*。

除主参数向量 $\boldsymbol{\beta}$的估计值外，我们还研究了由 $\mathbf{S(Y} - \mathbf{X}\boldsymbol{\beta})$ 得到的非参数拟合 ˆ $\mathbf{m(Z}\ \boldsymbol{\beta})$ 的表现，其中 $\boldsymbol{\beta}$ 可以是 $\boldsymbol{\beta}_{pls}$、 $\boldsymbol{\beta}_{ib1}$、ˆ ˆ 或 $\boldsymbol{\beta}_{ib2}$中的任意一个。然而，我们已证明，对于向量 $\mathbf{Z} = \mathbf{z}$ 中的任意有限取值，只要 ˆ $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(\mathbf{n}^{-0.5})$，ˆ$m(\mathbf{z}\ \boldsymbol{\beta})$ 的渐近方差始终相同。结果总结如下推论。

推论 2.3 如果 $b \to 0$ 且 $nb^d \to \infty$，并满足补充材料中的正则条件，则有

$$\sqrt{nb^d}\left\{\hat{m}(\mathbf{z}, \hat{\boldsymbol{\beta}}) - m_0(\mathbf{z}) - \frac{b^2}{2} tr\{\underline{Hess}_m(\mathbf{z})\} \int u^2 K(u) du\right\} \to N\left(0, \frac{\sigma^2}{f_{\mathbf{Z}}(\mathbf{z})} \int K^2(\mathbf{u}) d\mathbf{u}\right)$$

依分布成立，其中 $tr\{Hess_m(\mathbf{z})\}$ 是回归函数 $m(\cdot)$在 $\mathbf{z}$ 处的海森矩阵的迹，而 $f_{\mathbf{Z}}(\mathbf{z})$ 是 $\mathbf{Z}$ 的密度函数。

从上述推论可以看出，该估计量渐近相合，且收敛速度为 $\sqrt{nb^d}$，慢于 $\sqrt{n}$。这些发现与核回归结论中的大多数现有文献（Ruppert & Wand1994）十分吻合。另一方面，方差分量可通过以下量计算：$\mathbf{SCS}^T$，其中 $\mathbf{C} = Diag\{\hat{}\varepsilon_i^2\}_{i=1}^n$。我们再次强调，上述结果对任意 $\sqrt{n}$-相合的 $\boldsymbol{\beta}$估计量均成立，这意味着信息整合渐近而言未必能降低$\mathbf{m(Z)}$估计的方差。然而，我们在第3 节确实观察到，在有限样本量下，信息整合后$\mathbf{m(Z)}$的平均均方误差（MSE）显著降低。因此，在实际应用中，当内部样本量较小时，我们仍建议采用信息整合，以提升非参数成分$\mathbf{m(Z)}$的估计。

3 模拟

我们总共考虑了四种情形，以体现我们的方法在信息借用方面的灵活性（补充材料中的图 S1）；下文展示了两个典型情形，另外两种外部模型设置不同的情形则见补充材料。外部数据基于样本量N 进行模拟，该样本量较大或较小。内部数据基于样本量 n进行模拟，取值为 50、200 和 500。进行非参数估计时，我们采用 Epanichkov核（$K(\cdot) = 0.75(1-t^2)_+$）。为评估 $\boldsymbol{\beta}$ 估计量对带宽的敏感程度，我们将定理 2.1 和 2.2 中的 a 分别设为 0.65、0.8、1 和 1.25；对于非参数成分的估计，我们采用最小二乘交叉验证（CV）（Wand与Jones（1994））为 m(Z) 估计选择最优带宽。为统一真实数据应用中的带宽选择过程，我们将采用基于差分的估计（DBE）（Fan & Huang 2005、Fan与Li（2004）），并结合 CV 同时选择带宽。

3.1 数据生成

情形I（不考虑变量 O，且在外部模型中错误设定变量 Z，给定真实汇总信息）。以下设置模拟了我们的真实数据应用。我们通过真实数据生成机制 $Y = sin(5Z) + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 O + \varepsilon$生成内部数据和外部数据，其中$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 1, 1)^T$。在此，我们独立地通过 Uniform $[0, 1]$分布生成 Z 和 X_3。为考虑混杂结构，我们从正态分布生成 X_1 ，其均值等于每个样本的 Z 值，方差等于 1。类似地，我们从伯努利分布生成 X_2 ，其成功概率形式为 $exp(Z)/\{1 + exp(Z)\}$。我们使用伯努利分布生成外部不可观测变量 O，其成功概率等于$exp(Z)/\{2 + exp(Z)\}$。残差 ε 服从标准正态分布。该设置涵盖了一般情形：内部研究考虑包含非线性成分 Z、X 中的共同协变量以及 O 中的额外协变量的模型，而外部研究仅考虑包含共同协变量 X 和 Z 的简化（错误设定）模型，

即，$\mathbf{Y} = \gamma\mathbf{Z} + \theta_1\mathbf{X}_1 + \theta_2\mathbf{X}_2 + \theta_3\mathbf{X}_3 + \tilde{\varepsilon}$。上述设置用于评估我们的 ˆ ˆ 估计量 $\boldsymbol{\beta}_{ib1}$ 在给定真实汇总信息（通过计算机模拟计算）的情况下，与金标准PLS估计量 $\boldsymbol{\beta}_{pls}$相比的表现。

情形II（给定估计的汇总信息）。除以下情况外，我们保持情形I中的所有设置不变：现在需要利用小样本量的外部数据估计汇总信息：1）外部 $N = 1000$，内部 $n = 200$；2）外部 $N = 200$，内部 $n = 200$；3）外部 $N = 200$，内部 $n = 500$。同样，我们使用情形I中定义的线性回归来计算估计量 $\boldsymbol{\theta}$ 及其对应的方差-协方差矩阵 $\boldsymbol{\theta}$，即$\mathbf{V}$；假设二者均可从外部研究中获得。此设置用于评估在外部样本量较小的情况下，通过纳入可用的方差-协方差信息，我们的估计量 $\boldsymbol{\beta}_{ib2}$ 相较于 $\boldsymbol{\beta}_{pls}$ 和 $\boldsymbol{\beta}_{ib1}$ 的表现。

如补充材料所示，鉴于变量$\mathbf{O}$ 的可用性以及外部研究模型中变量$\mathbf{Z}$ 的设定，我们还考虑了另外两种情形（有关更多数据生成细节和结果讨论，请参阅补充材料中的第3节）。

3.2 评估

ˆ ˆ对 $\boldsymbol{\beta}_{ib1}$ 和 $\boldsymbol{\beta}_{ib2}$的评估。我们采用基于1000次蒙特卡洛ˆ ˆ模拟运行的以下指标来评估估计量 $\boldsymbol{\beta}_{ib1}$ 和 $\boldsymbol{\beta}_{ib2}$的性能：偏差、蒙特卡洛标准差（MCSD）、渐近标准误（ASE）、95%覆盖概率（CP），以及相对效率（RE）。RE定义为ˆPLS估计量 $\boldsymbol{\beta}_{pls}$ 与信息整合后我们提出的估计量的经验均方误差之比，因此比率大于1更理想。情形I下的结果汇总于表1中。ˆ在此情形下，由于缺少方差—协方差ˆ矩阵$\mathbf{V}$的信息，我们仅评估了 $\boldsymbol{\beta}_{ib1}$ 。总体而言，我们观察到以下几点：（1）无论内部样本量如何，估计量 $\boldsymbol{\beta}_{ib1}$的偏差都很小；（2）随着内部样本量 n 增加，ASE越来越接近MCSD，CP也越来越接近其95%的名义水平；（3）RE显著大于1，这表明整合外部研究的汇总信息带来了可观的效率增益。有趣的是，我们还

在模拟中观察到，不同带宽的选择对估计量 $\boldsymbol{\beta}$ 的影响很小，这表明 $\boldsymbol{\beta}$ 估计对带宽选择并不十分敏感。

以上结果基于真实汇总信息。表2 总结了情形II下外部研究的样本量相较于内部研究较小时的结果。手稿中仅展示了带宽等于1时的结果。此时，估计量 $\boldsymbol{\beta}_{ib1}$ 无法有效利用外部信息（即相对效率小于1），且渐近标准误远小于蒙特卡洛标准差。尽管偏差很小，但不准确的渐近标准误导致第一类错误率大幅膨胀，覆盖概率远低于95%的名义水平。另一方面，估计量 $\boldsymbol{\beta}_{ib2}$ 对带宽并不敏感（见表1以及补充材料中的表S3），即使外部样本量较小，仍能带来显著的效率增益。我们还观察到，渐近标准误令人满意，与蒙特卡洛标准差接近；覆盖概率也令人满意，接近其95%的名义水平。这些观察结果表明，当外部研究的样本量较小时，纳入 $\boldsymbol{\theta}$的方差-协方差信息至关重要；因此，在情形II中，估计量 $\boldsymbol{\beta}_{ib2}$ 优于 $\boldsymbol{\beta}_{ib1}$ 。随着外部样本量增加，我们观察到估计量 $\boldsymbol{\beta}_{ib1}$ 的表现有所改善，并逐渐接近估计量 $\boldsymbol{\beta}_{ib2}$的表现。上述所有观察结果均验证了我们在第2.4节中的理论发现。

ˆ ˆm(Z $\boldsymbol{\beta}$) 的评估。我们采用以下指标，根据 1000 次蒙特卡洛运行评估估计量 ˆ ˆ m(Z $\boldsymbol{\beta}$) 的表现，其中 $\boldsymbol{\beta}$ 可以是 $\boldsymbol{\beta}_{pls}$、ˆ ˆ$\boldsymbol{\beta}_{ib1}$ 和 $\boldsymbol{\beta}_{ib2}$ 中的任意一个：平均蒙特卡洛方差（AMCV）、平均渐近方差（AV）、平均相对效率（ARE）（定义为信息整合后，基于PLS的估计量与我们提出的估计量在 ˆm(Z $\boldsymbol{\beta}$) 的经验均方误差平均值之比），以及平均总体相对效率（AORE）（定义为信息整合后，基于PLS的估计量与我们提出的估计量在 X$\boldsymbol{\beta}$ + ˆm(Z $\boldsymbol{\beta}$) 的经验均方误差平均值之比）。比值大于 1 为理想结果。表 3 汇总了 ˆm(Z $\boldsymbol{\beta}$) 的结果。我们发现，与使用PLS估计量 $\boldsymbol{\beta}_{PLS}$ 计算估计量 ˆ ˆm(Z $\boldsymbol{\beta}$) 相比，使用估计量 $\boldsymbol{\beta}_{ib1}$ 或 $\boldsymbol{\beta}_{ib2}$ 计算时，ˆm(Z $\boldsymbol{\beta}$) 的相对效率大于 1，表现有所提升。

$\hat{}\,\hat{}$ 计算估计量 $\hat{}\mathbf{m}(\mathbf{Z}\ \boldsymbol{\beta})$。我们还发现，随着样本量增加，$\hat{}\mathbf{m}(\mathbf{Z}\ \boldsymbol{\beta})$ 的平均 AV 越来越接近平均 MCV。

在其他数据设置中，$\boldsymbol{\beta}$ 和$\mathbf{m}(\mathbf{Z})$的估计值也呈现出类似的结果和模式（表 S1 和 S2）。此外，我们评估了两个数据集中的协变量分布不同的情形，并观察到偏差很小（表 S7）。这验证了所提方法的稳健性。更多详情请参阅补充材料中的第 3 节。对于带宽的基于差分的估计，我们在情形II下进行了评估，结果见表 S8 和 S9。其中，偏差、覆盖概率和相对效率方面的结果均令人满意。

与 GIM 的比较。如前文所述，我们的方法在文献中独具特色，因为它将信息整合到含有非参数成分的模型中。因此，现有方法无法直接应用于本文所讨论的情境。为说明这一点，我们将所提方法与知名的 GIM（Zhang 等，2020）进行了比较。尽管 GIM 具有先进的理论性质，但它侧重于采用（广义）线性模型的似然框架。因此，我们在 GIM 中将所有协变量均视为具有线性效应。补充材料中的表 S6 总结了 GIM 在情形II下的表现。与所提方法（$\boldsymbol{\beta}_{ib2}$）相比，我们发现 GIM 对非线性成分的估计存在偏差（结果未显示），且其线性成分的估计变异性更大。值得注意的是，由于采用了所提出的解耦估计，我们的方法所需的计算时间远少于现有方法。

4 数据应用

4.1 基于英国生物银行（UKB）影像数据的评估。

在本节中，我们介绍如何使用真实世界数据验证我们的方法。在此应用中，我们重点关注中低收入总体，研究全脑白质完整性（白质纤维束上的平均FA值）与脉搏率、年龄、体质指数和性别等关键变量之间的关联。在模型中，年龄被视为非线性趋势，且未预先指定参数形式（Lee等人，2022年，Bethlehem 等人（2022））。我们总共纳入了

14851名参与者。为评估我们的方法，我们采用了以下抽样策略：从拥有完整变量数据的参与者中（共11397名）随机抽取200条记录，并将这200条记录作为小样本量的内部数据。整个队列（11397名参与者）则作为真实数据生成机制，并用作方法评估的金标准。随后，我们将包含参与者人口统计信息（年龄、体质指数和性别）的额外样本视为外部数据来源（共3454名参与者）。两组数据的人口统计特征汇总于表 S4，且两队列之间相似。为评估外部研究样本量的影响，我们考虑了两个外部数据集，其样本量分别为400和1，200，均从外部数据来源中随机抽取。基于外部数据，对平均FA（乘以1，000）与三个人口统计变量（年龄、体质指数和性别）进行线性回归，得到估计值 $\boldsymbol{\theta}$ 及其估计的方差-协方差矩阵$\mathbf{V}$ 。由于外部数据的样本量与内部数据相当，我们仅考虑第 $\boldsymbol{\beta}_{ib2}$ 2.5节提出的估计量，并评估其表现。作为基准，我们使用包含完整人口统计信息和脉搏率记录的全部数据集（11，397个样本）计算PLS估计量，并将其视为真实数据生成机制（Oracle估计量）。从图1中可以观察到，基于该Oracle估计量，年龄效应呈非线性。这一发现验证并支持了部分线性模型的使用。所提出的估计量 $\boldsymbol{\beta}_{ib2}$、仅基于内部数据的PLS估计量 $\boldsymbol{\beta}_{pls}$ 以及Oracle PLS估计量 $\boldsymbol{\beta}_{opls}$ 的结果见表4。

从表 4中我们发现，利用外部信息后，体质指数和性别效应估计值的变异性显著降低。外部样本量越大，所提出的估计量就越接近 Oracle PLS 估计量。我们还观察到，所提出的估计量显示体质指数（P值=0.0022，外部样本量为 $\boldsymbol{\beta}_{ib2}$ 时 $N = 1200$）和性别（P值=0.0359， $\boldsymbol{\beta}_{ib2}$ 时 $N = 400$）具有显著效应，而使用 PLS 估计量时未能观察到这些显著效应。基于所提出的估计量得到的体质指数负效应和男性 FA 测量值较高的发现，与近期研究结果一致；这些研究表明，体重较重以及女性均与白质完整性降低显著相关（Stanek 等，2011，Poletti

等，2020）。在图 1中，我们观察到，与仅基于内部样本的 PLS 估计量相比，采用信息整合的估计量更接近 Oracle 估计量。尽管准确性更高，信息整合后的估计值并未呈现明显的非线性趋势。这可能是由于内部数据的样本量较小，而且非参数估计的效率增益并无渐近保证（第2.4节）。总之，上述结果表明，借用外部信息能够提高估计精度，因此在实际应用中可能提供更准确、可靠的统计推断。

4.2 基于 UKB 和 ARIC 数据研究收缩压的应用

在本节中，我们应用所提出的方法，研究老年患者（$\geq$ 65 岁）中收缩压（SBP）与脑影像数据、胆固醇测量值等临床因素之间的关联。我们还将年龄、性别（男性=1/女性=0）、种族（黑人=1/白人=0）和体质指数（BMI）作为感兴趣的协变量纳入分析。需要注意的是，第一作者可以访问ARIC数据（Wright等，2021年），而其他作者可以访问UKB数据。由于数据保密和数据使用协议，共享原始数据可能需要付出大量文档整理工作。因此，我们决定共享ARIC数据中的汇总信息，以增强基于UKB数据的分析。由于处理后的ARIC数据库（包含798名老年人）中没有完整的核磁共振（NMR）代谢组学和白质纤维束完整性记录，第一作者对SBP关于年龄、性别和体质指数（BMI）变量进行了线性回归，并与其他作者共享了汇总信息（估计值和方差矩阵）。随后，我们利用汇总信息，将所提出的整合工具应用于UKB数据（295名拥有完整NMR和神经影像学数据的老年人），并进行部分线性回归：允许年龄具有潜在的非线性效应，同时允许所有其他感兴趣变量具有线性效应。比较PLS和IB2方法的结果总结于表5中。我们观察到，通过整合ARIC研究中的信息，估计效率得到提高（ASE更小），并成功检测出BMI的显著效应，这与文献中的发现一致（Bann等，2021年）。

我们还从图 S2中观察到，随着年龄增加，IB2和PLS均显示出年龄效应呈较为明显的线性递增趋势。

5 讨论

我们开发了一种实用工具，可将外部研究的信息用于内部分析；在这种分析中，人们认为部分协变量对结果变量具有复杂的非线性效应。基于计算机模拟的数值评估结果，我们的方法能够很好地拟合样本量甚至不足一百的真实数据；信息整合后，所有协变量（尤其是参数成分）在数值上均有益处。我们提出的估计量的优越性也通过神经影像学和收缩压数据的应用得到了证明，其研究发现与现有文献高度一致。

基于当前框架，可以进行若干扩展。第一项扩展是整合多个外部研究的信息。在 K 不同研究中的总体同质这一假设下，通过将（2.7）中的损失函数修改为 $l = \sum_{i=1}^{n} log(p_i)$ $- \sum_{j=1}^{K} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)^T (\hat{\mathrm{V}}j)^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j)/2$，即可成功汇总这些信息，其中 $\boldsymbol{\theta}_j$ 和 $\mathrm{V}j$ 分别是第 j项外部研究中 $j = 1$、.....、 K的估计值和协方差矩阵。第2.3.2节中的其他估计程序保持不变。第二项扩展是考虑只能获得部分协方差矩阵$\hat{}$V信息的情况。当汇总信息摘自现有文献，而文献仅报告估计值的方差、未提供完整协方差矩阵时，就会出现这种情况。在这种情形下，可类似地采用Zhang等人（2020）提出的迭代算法，首先基于内部数据恢复协方差矩阵$\hat{}$V，然后通过迭代算法在数值上实现最优效率。第三项扩展是考虑研究之间同质性假设不成立的情况。尽管我们的方法对不同的协变量分布并不敏感，但异质条件分布会导致内部分析产生有偏估计。可以考虑更先进的技术，例如密度倾斜技术（Sheng 等人（2022）），但这些方法都需要付出大量努力，值得在未来进一步研究。第四项扩展是允许外部研究采用更复杂的模型结构。大多数信息

整合方法都聚焦于某种参数模型。尽管科学研究中常见的做法是采用广义线性模型，但衰老研究中非参数模型也受到了广泛关注（Lee等人，2022年）。然而，我们当前的理论框架无法确保对非参数成分进行信息整合；未来研究中，我们将致力于实现这一扩展。

参考文献

Bann, D., Scholes, S., Hardy, R. & O'Neill, D. (2021),《体质指数与血压关联随时间的变化：来自多项横断面研究和队列研究的证据》，《预防医学》**153**, 106825。

Bethlehem, R. A., Seidlitz, J., White, S.R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A. et al. (2022),《人类生命周期的脑图谱》，《自然》**604**(7906), 525–533。

Brummett, B. H., Babyak, M. A., Jiang, R., Huffman, K. M., Kraus, W. E., Singh, A., Hauser, E. R., Siegler, I. C. & Williams, R. B. (2019),《大型多研究总体中的收缩压与社会经济地位》，SSM-人口健康**9**, 100498。

Chatterjee, N., Chen, Y.-H., Maas, P. &Carroll, R. J. (2016), '利用外部大数据来源的汇总级信息进行模型校准的约束最大似然估计'，《美国统计学会杂志》**111**(513), 107–117。

Chen, C., Han, P. & He, F. (2022), '借助辅助数据中的信息改进主要分析'，医学统计学 41(3),567–579。

Chen, C., Wang, M. & Chen, S. (2023), '一种高效的数据整合方案，用于综合多个次级数据集的信息以对主要分析进行参数推断'，Biometrics。

Chen, D.-G., Liu, D., Min, X. & Zhang, H. (2020)，《在随机效应荟萃分析中使用汇总数据与个体数据的相对效率》，Biometrics **76**(4),1319–1329.

Fan, J. (1992), 'Design-adapti非参数 tri回归 ion', J l of the Ameri 期刊
美国统计协会 87(420), 998–1004. Fan, J. (1993),《局部线性回归平滑器及其极小极大效率》，《统计年刊》第 196–216 页。Fan, J. & Gijbels, I. (1996),《局部多项式建模及其应用》，伦敦：Chapman and Hall/CRC。Fan, J. & Huang, T. (2005),《半参数变系数部分线性模型的剖面似然推断》，Bernoulli 11(6),1031–1057。

Fan, J., Huang, T. & Li, R. (2007),《采用协方差函数半参数估计分析纵向数据》，《美国统计学会杂志》102(478), 632–641。

Fan, J. & Li, R. (2004),《纵向数据分析中半参数建模的新估计与模型选择程序》，《美国统计学会杂志》99(467)，710–723。Haidich, A.-B. (2010),《医学研究中的荟萃分析》，Hippokratia 14(Suppl 1)，29。Han, P. & Lawless, J. F. (2019),《利用具有不同协变量分布的辅助汇总信息进行经验似然估计》，Statistica Sinica 29(3), 1321–1342。H¨ardle, W., Liang, H. & Gao, J. (2000)，部分线性模型，Springer Science & Business Media。He, J., Li, H., Zhang, S. & Duan, X. (2019),《带有辅助亚组生存信息的加性风险模型》，Lifetime Data Analysis 25(1)，128–149。Ibrahim, J. G., Chen, M.-H., Gwon, Y. & Chen, F. (2015),《幂先验：理论与应用》，Statistics in medicine 34(28)，3724–3749。Jiang, L., Nie, L. & Yuan, Y. (2021),《临床试验中利用幂先验从历史数据动态借用信息》，Biometrics。

Kundu, P., Tang, R. & Chatterjee, N. (2019), 《利用广义荟萃分析对具有不同协变量信息的研究开展多重回归模型分析》, Biometrika106(3), 567–585。Lee, H., Chen, C., Kochunov, P., Elliot Hong, L. & Chen, S. (2022), 《对具有依赖关系的多变量年龄相关影像变量进行建模》, Statistics in Medicine。Liang, H. (2006), 《部分线性模型中的估计与数值比较》, 计算统计与数据分析 50(3), 675–687。Liang, K.-Y. & Zeger, S. L. (1986), 《利用广义线性模型进行纵向数据分析》, Biometrika 73(1), 13–22。Lin, D.-Y. & Zeng, D. (2010), 《荟萃分析中使用汇总统计量相对于个体层面数据的相对效率》, Biometrika 97(2), 321–332。Liu, D., Liu, R. Y. & Xie, M. (2015), 《仅利用汇总统计量对异质性研究进行多变量荟萃分析：效率与稳健性》, 《美国统计学会杂志》110(509), 326–340。McCulloch, C. E. & Searle, S. R. (2004), 广义、线性与混合模型, 约翰·威立父子出版公司。Niccoli, T. & Partridge, L. (2012), 《衰老作为疾病的风险因素》, 《当代生物学》22(17), R741–R752。Poletti, S., Melloni, E., Mazza, E., Vai, B. & Benedetti, F. (2020), 《暴露于轻度压力的健康成人中白质微观结构的性别特异性差异》, 《压力》23(1), 116–124。Qin, J. (2000), 《杂记：结合参数似然与经验似然》, Biometrika87(2), 484–490。Qin, J. & Lawless, J. (1994), 《经验似然与广义估计方程》, 《统计年刊》22(1), 300–325。

Qin, J., Liu, Y. & Li, P. (2022), ‘一篇利用相似研究中的校准信息的统计方法综述’, 统计理论与相关领域 pp. 1–16.

Qin, J., Zhang, H., Li, P., Albanes, D. & Yu, K. (2015), ‘利用协变量特异的疾病患病率信息提高病例对照研究的检验效能’, Biometrika102(1), 169–180.

Ruppert, D.& Wand, M. P. (1994), ‘多变量局部加权最小二乘回归’, 《统计年刊》 pp. 1346–1370.

Sheng, Y., Sun, Y., Huang, C.-Y. & Kim, M.-O. (2022), ‘在总体异质性存在时综合外部汇总信息：一种惩罚经验似然方法’, Biometrics 78(2), 679–690.

Stanek, K. M., Grieve, S. M., Brickman, A. M., Korgaonkar, M. S., Paul, R. H., Cohen, R. A. & Gunstad, J. J.(2011), ‘肥胖与其他方面健康的成人白质完整性降低有关’, Obesity 19(3), 500–504.

Wand, M. P. & Jones, M. C. (1994),《核平滑》，CRC出版社。

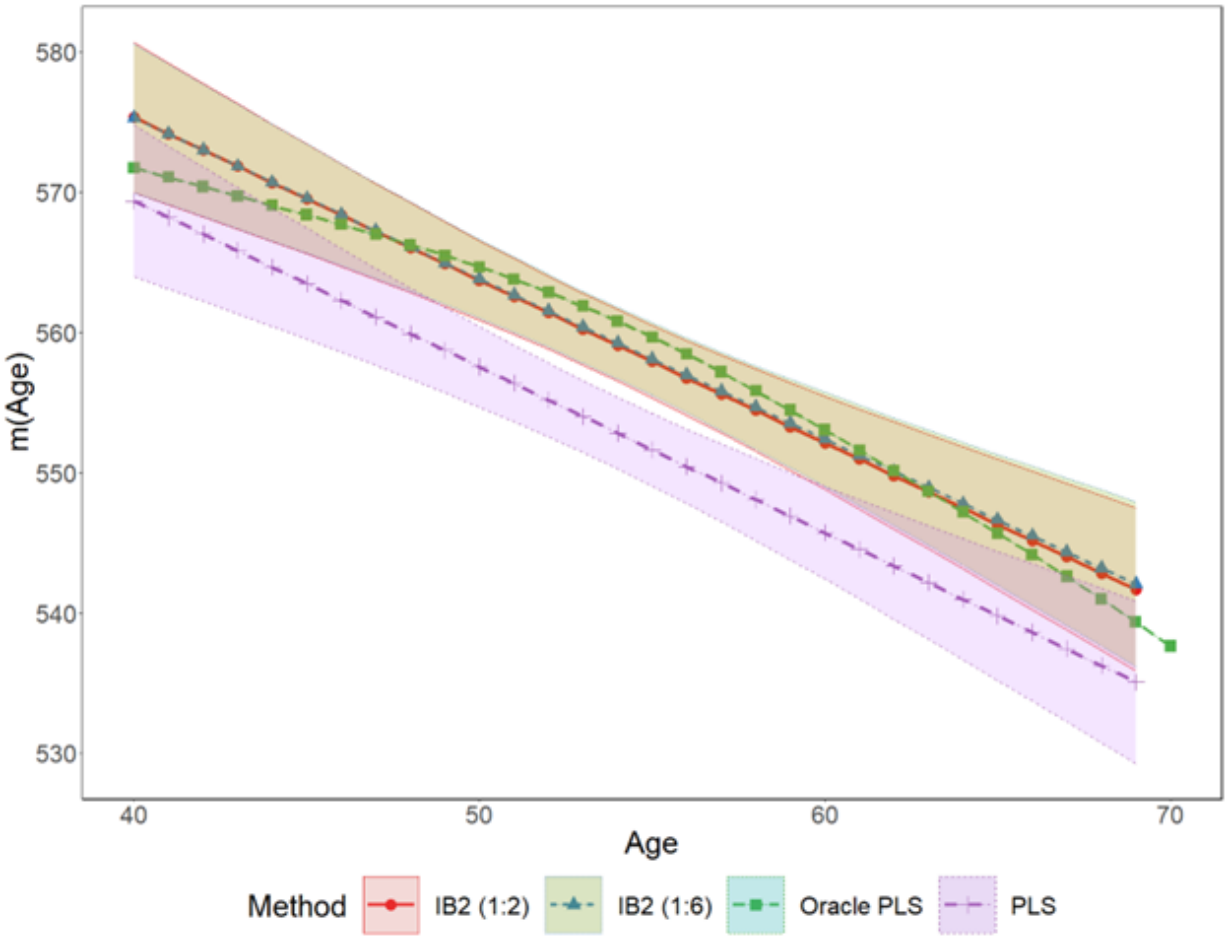
Wright, J. D., Folsom, A. R., Coresh, J., Sharrett, A. R., Couper, D., Wagenknecht, L. E., Mosley Jr, T. H., Ballantyne, C. M., Boerwinkle, E. A., Rosamond, W. D. 等 (2021), ‘ARIC（社区动脉粥样硬化风险）研究：JACC专题研讨会 3/8’, 《美国心脏病学会杂志》77(23), 2939–2959.

Zhai, Y. & Han, P. (2022), ‘利用来自异质总体的外部信息进行预言式使用的数据整合’, 《计算与图形统计学杂志》 pp. 1–12.

Zhang, H., Deng, L., Schiffman, M., Qin, J. & Yu, K. (2020), ‘通过利用外部汇总数据改进统计推断的广义整合模型’, Biometrika107(3), 689–703.

Zheng, J., Zheng, Y. 和 Hsu, L.（2021），《利用来源个体层面数据的汇总统计量进行事件发生时间结局的风险预测》，《美国统计学会杂志》第 1–13 页。

图1：UKB 神经影像学研究中年龄效应非线性拟合的表现。IB2 (1:r) 表示估计量 $\hat{\mathbf{m}}(\mathbf{Z}\ \boldsymbol{\beta}_{ib2})$，其中 $1:r$ 表示内部样本量与外部样本量的比率（$r = 2,6$）；PLS 表示剖面最小二乘估计量 $\hat{\hat{\mathbf{m}}}(\mathbf{Z}\ \boldsymbol{\beta}_{pls})$；Oracle PLS 是使用全部主数据的估计量（作为基准），即 $\hat{\mathbf{m}}(\mathbf{Z}\ \boldsymbol{\beta}_{opls})$。彩色区域为置信区域。



ˆ 表1：评估估计量 $\boldsymbol{\beta}_{ib1}$，内部样本量为 50、200、500，并使用真实汇总信息。所有结果均基于 1000 次蒙特卡洛运行。

		50			200			500		
s		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.650	Bias	0.000	-0.006	0.020	-0.001	0.004	-0.002	-0.000	-0.001	0.003
	MCSD	0.097	0.208	0.346	0.042	0.085	0.144	0.026	0.057	0.093
	ASE	0.084	0.172	0.290	0.039	0.080	0.133	0.024	0.051	0.084
	RE	2.412	2.604	2.399	2.875	3.162	2.927	3.237	2.750	3.043
	CP	0.909	0.904	0.903	0.927	0.936	0.928	0.932	0.923	0.920
0.800	Bias	0.000	0.008	-0.001	0.001	0.001	0.000	0.001	-0.005	-0.005
	MCSD	0.100	0.201	0.339	0.042	0.090	0.139	0.026	0.054	0.089
	ASE	0.084	0.170	0.286	0.039	0.080	0.133	0.024	0.050	0.083
	RE	2.309	2.462	2.464	3.273	2.794	3.129	2.925	3.144	3.053
	CP	0.913	0.922	0.918	0.922	0.928	0.937	0.932	0.936	0.936
1.000	Bias	-0.004	0.007	0.008	0.001	0.002	0.008	0.002	-0.000	0.004
	MCSD	0.099	0.186	0.321	0.040	0.084	0.140	0.026	0.053	0.091
	ASE	0.082	0.167	0.280	0.038	0.080	0.133	0.024	0.050	0.083
	RE	2.371	2.734	2.718	3.208	3.182	2.952	3.026	2.971	2.802
	CP	0.895	0.944	0.909	0.942	0.938	0.930	0.920	0.934	0.923
1.250	Bias	-0.001	0.012	-0.003	0.003	0.003	0.002	-0.000	0.001	-0.001
	MCSD	0.092	0.183	0.309	0.038	0.084	0.135	0.025	0.053	0.088
	ASE	0.080	0.166	0.277	0.038	0.080	0.132	0.024	0.050	0.083
	比率	2.810	2.993	2.873	3.370	3.396	3.494	3.067	3.031	3.050
	CP	0.915	0.940	0.920	0.945	0.936	0.946	0.940	0.945	0.940

s 表示带宽公式中的常数：$s * n^{-0.2}$；MCSD 表示蒙特卡洛标准差；ASE 表示渐近标准误；RE 表示相对效率，即 $\boldsymbol{\beta}_{pls}$ 的均方误差与 $\boldsymbol{\beta}_{ib1}$的均方误差之比；CP 表示 95%覆盖概率。

ˆˆ 表 2：在有限内部样本量 (n) 和有限外部样本量 (N) 下，使用估计的汇总信息评估 $\boldsymbol{\beta}_{ib1}$ 和 $\boldsymbol{\beta}_{ib2}$ ，其中 $s = 1$ （带宽中的一个常数）。

		IB2			IB1		
		β_1	β_2	β_3	β_1	β_2	β_3
n=200，N=1000	Bias	0.001	0.003	-0.004	0.002	0.006	-0.011
	MCSD	0.044	0.091	0.156	0.046	0.094	0.164
	ASE	0.045	0.094	0.156	0.038	0.079	0.133
	RE	2.843	2.939	2.744	2.582	2.704	2.505
	CP	0.953	0.957	0.954	0.891	0.898	0.888
n=200，N=200	Bias	-0.002	0.007	-0.000	-0.001	0.009	0.016
	MCSD	0.056	0.118	0.195	0.075	0.153	0.261
	ASE	0.056	0.117	0.195	0.039	0.081	0.135
	RE	1.593	1.668	1.527	0.946	1.058	0.889
	CP	0.942	0.945	0.949	0.671	0.69	0.672
n=500，N=200	Bias	-0.001	0.000	-0.001	-0.003	0.012	-0.003
	MCSD	0.041	0.084	0.134	0.068	0.142	0.234
	ASE	0.040	0.083	0.138	0.025	0.052	0.085
	RE	1.286	1.258	1.247	0.442	0.458	0.429
	CP	0.947	0.949	0.957	0.492	0.525	0.531

MCSD 表示蒙特卡洛标准差；ASE 表示渐近标准误；ˆ RE 表示相对效率，即 $\boldsymbol{\beta}_{pls}$ 的均方误差除以 $\hat{\boldsymbol{\beta}}_{ib}$ 的均方误差；CP 表示 95% 覆盖概率。

ˆ ˆ 表3：基于两种提出的数据整合方案（即 $\boldsymbol{\beta}_{ib1}$ˆ ˆ ˆ 和 $\boldsymbol{\beta}_{ib2}$），采用 $s=1$ 对 $\boldsymbol{\beta}_{ib}$ 进行评估，并使用 GCV 为 ˆm(Z $\boldsymbol{\beta}_{ib}$) 中的带宽选择带宽。

		均方误差(pls)	均方误差(ib)	渐近相对效率	平均交叉验证 平均方差		AORE
ˆ m(Z; $\hat{\boldsymbol{\beta}}_{ib1}$)	n=50, N=10,000	0.245	0.175	1.398	0.606	0.751	3.011
	n=200, N=10, 000	0.069	0.045	1.531	0.533	0.579	2.559
	n=500, N=10,000	0.031	0.020	1.545	0.518	0.551	2.250
ˆ m(Z; $\hat{\boldsymbol{\beta}}_{ib2}$)	n=500; N=200	0.030	0.023	1.31	0.514	0.552	1.297
	n=200; N=200	0.067	0.053	1.261	0.538	0.593	1.646
	n=200; N=1000	0.069	0.046	1.495	0.533	0.580	1.923

MSE 表示均方误差；ARE 表示 ˆm(Z, $\hat{\boldsymbol{\beta}}_{ib}$)的平均相对效率，即 ˆm(Z, $\hat{\boldsymbol{\beta}}_{pls}$) 在各样本上的均方误差平均值与 ˆm(Z; $\hat{\boldsymbol{\beta}}_{ib}$) 平均值 over AORE 表示总体相对效率，即 MSE{ˆm(Z; $\hat{\boldsymbol{\beta}}_{pls}$) + X $\hat{\boldsymbol{\beta}}_{pls}$}对.....取平均样本中的均方误差{ˆm(Z; $\hat{\boldsymbol{\beta}}_{ib}$) + X $\hat{\boldsymbol{\beta}}_{ib}$} 对样本取平均；AMCV 是蒙特卡洛方差的样本平均值；AV 是渐近方差的样本平均值。

ˆ ˆ 表 4：基于 UKB 数据对 $\boldsymbol{\beta}_{ib2}$ 和 $\boldsymbol{\beta}_{pls}$ 进行评估。

		脉搏率	BMI	性别
$\hat{\boldsymbol{\beta}}_{ib2}$ (n=200, N=400)	估计值	-0.003	-0.3473	3.895
	ASE	0.1158	0.1656	1.649
	Z统计量	-0.0259	-2.0972	2.362
	P值	0.9793	0.0359	0.0182
$\hat{\boldsymbol{\beta}}_{ib2}$ (n=200, N=1,200)	估计值	0.00154	-0.338	2.135
	ASE	0.114	0.11	1.101
	Z统计量	0.0135	-3.05	1.928
	P值	0.989	0.0022	0.0537
$\hat{\boldsymbol{\beta}}_{pls}$ (n=200)	估计值	0.0011	-0.1448	5.541
	ASE	0.11675	0.284	2.8128
	Z统计量	0.0095	-0.506	1.96849
	P值	0.9923	0.6125	0.049
$\hat{\boldsymbol{\beta}}_{opls}$ (n=11,307)	估计值	0.0042	-0.37672	1.2036

渐近标准误为 ˆ e 渐近标准误； $\boldsymbol{\beta}_{opls}$ 是预言机轮廓最小二乘法估计量 r.

ˆ ˆ 表5：基于ARIC（外部）和UKB数据（内部）对 $\boldsymbol{\beta}_{ib2}$ 和 $\boldsymbol{\beta}_{pls}$ 的评估。

		估计值	ASE	Z统计量	P值
$\hat{\boldsymbol{\beta}}_{pls}$	胆固醇	2.733	5.072	0.539	0.590
	BMI	0.144	0.337	0.428	0.668
	性别	5.443	2.39	2.277	0.023
	平均FA	-123.288	62.01	-1.988	0.046
$\hat{\boldsymbol{\beta}}_{ib2}$	胆固醇	3.088	5.064	0.609	0.542
	BMI	0.656	0.114	5.763	¡0.001
	性别	1.386	0.966	1.435	0.151
	平均FA	-100.289	59.442	-1.687	0.0916

ASE 是渐近标准误。