arXiv:2305.01109v1 [stat.ME] 1 May 2023

在线 A/B 测试中大规模应用协变量调整

Lorenzo Masoero masoerl@amazon.com
Doug Hains dhains@amazon.com
James McQueen jmcq@amazon.com

Amazon.com

2023年5月3日

摘要

提供网络服务的公司通常会开展随机在线实验，以估计采用新功能和政策对关键绩效指标产生的“因果影响”。这些实验用于估计各种效应：例如，重新调整横幅位置带来的点击率提升，或因折扣、特别优惠而产生的订阅率变化等。在这些情境下，即使效应很小，也可能造成重大的后续影响。由于简单易用，简单均值差估计量（Splawa-Neyman 等人，1990）仍是许多在线 A/B 测试平台的标准首选估计量。然而，即使实验包含数千或数百万个观测单位，这种方法也可能无法检测到微小效应。不过，这些实验还会产生大量额外数据（协变量）。本文探讨允许实验人员使用更复杂的估计量、在估计关注的因果效应时利用协变量所带来的益处、成本和风险。我们将近期提出的一种通用算法——用于借助协变量估计因果效应——调整应用于在线 A/B 测试场景。基于这一范式，我们实现了多种协变量调整因果估计量，并全面评估其大规模应用时的表现，重点说明不同方法的优点和不足。我们通过真实实验展示，“协变量调整”估计量如何（i）提高对关注的因果效应的量化精度，以及（ii）解决处理组之间不平衡的问题——这是一个实际问题，却常被文献忽视。进而，（iii）这些更精确的估计可以缩短实验时间、降低成本，并帮助简化决策流程，从而更快地采用有益的干预措施。

1 引言

在线公司的持续增长和产品改进依赖于高效发现新的机遇，并准确衡量决策对客户的影响。为了估计产品或功能变更所产生的因果影响，在线公司高度依赖 A/B 测试（随机对照试验/在线随机实验）。在最小假设条件下，A/B 测试确实能够保证对所测试干预的因果影响给出无偏估计（Splawa-Neyman 等人，1990）。

在下文中，我们假设关注的实验单位是“客户”。实际上，不同的实验可能跟踪不同的单位（卖家、主播、购物任务等）。A/B测试通过随机分配部分客户（通常为一半，即“处理组”）来体验新方案，而其他客户（“对照组”）则体验旧的现状方案。在固定的实验期内，会测量并记录这些客户的各项相关指标。实验人员在A/B测试中量化体验变化对客户影响的一种简单方法，是比较处理组客户某项指标或“关键绩效指标” [KPI] 的平均值与对照组客户同一指标的平均值。这种“均值差” [DIM] 方法可以简单估算体验变化对相关 KPI 的平均因果效应。

借助数百万客户开展 A/B测试既简单又具有较低的边际成本，因此这类测试已在业界广泛应用。A/B测试被用于评估搜索引擎（Google、Bing、Yandex）、在线零售商（Amazon、eBay、Etsy）、流媒体服务（Netflix、Twitch、YouTube）、社交网络（Facebook、LinkedIn、Twitter）、旅游服务（Lyft、Uber、Airbnb、Booking.com）等的前端和后端变更。有关 A/B测试在业界的作用和应用的详细讨论，请参阅 Gupta 等（2019）。

考虑到实验时间的机会成本、较小的处理效应大小以及客户之间的高度异质性，即使实验包含数千或数百万客户，均值差方法也常常无法检测到实际存在的干预效应。然而，在实验之前及实验过程中，人们通常会收集大量与 A/B 测试本身无关的数据（即“协变量”），这些数据描述了实验单位。如此丰富的数据使实验人员有可能采用更复杂的“协变量调整”方法来进行估计。具体而言，任何与干预无关的特征（例如实验前采集的任何测量值）都可用于估计所关注的因果效应。由此得到的协变量调整估计量可以改进对所关注“因果效应”的估计，并降低其变异性。已有实证研究发现，简单的协变量调整方法（例如广受欢迎的“CUPED”（Deng 等，2013））可以显著降低方差，相较于均值差估计量尤为如此。此外，协变量调整也被用作防范措施，以抵御不走运的随机化所带来的影响；在这种情况下，干预可能因偶然因素而显得效果人为地偏好或偏差（参见 Tukey（1991））。

与协变量调整方法相关的文献不断增多（近期研究成果参见 Guo 等（2021）；Jin 和 Ba（2021））。本文探讨了在在线 A/B 测试中，通过采用更多协变量和更复杂的估计量来估计因果效应，从而扩展实验人员工具箱的收益与代价。我们的实验表明，协变量调整方法能够在估计准确性和方差缩减方面带来显著提升。

本文其余部分安排如下：第2节介绍所研究问题的符号表示。接下来，我们在第3节介绍广义 Oaxaca-Blinder 估计量 [GOBEs] ——这是一种灵活的通用方法，可利用任意数量的额外协变量来构造因果效应估计量。第4节讨论这些估计量在实验中的潜在用途，第5节展示实验结果。最后，我们在第6节进行讨论并展望后续工作。

2 潜在结果、随机实验与因果效应

因果推断领域汇集了一系列理论上可靠的工具、方法和程序，可帮助实践者回答有关他们可能希望实施的干预措施影响的问题。尽管对干预措施作出严谨的因果判断很有吸引力，也很有价值，但这需要以通过精心设计的实验收集数据为代价。为了确保因果判断有效，实验人员必须确保以恰当的方式收集数据，避免分析中引入任何偏差或缺陷。确保所收集的数据能够支持我们作出因果判断的标准做法，是开展随机对照试验 （RCT）或 A/B测试。最简单的形式是，将每个实验单位随机分配到对照组（A）或处理（B）体验。随机化是实验人员能够从实验中得出因果结论的关键技术手段。

为了明确讨论内容，我们在此采用潜在结果因果模型（Splawa-Neyman 等人，1990；Rubin，1977）。简而言之，我们假设，在一项观测 N 个单位的实验中，每个个体单位$n \in [N] := \{1, \ldots, N\}$都会接受 $T \geq 2$种不同“处理”或“政策”中的一种。例如， N个单位可能是实验中的不同客户。对于单位 $n \in [N]$ 被分配到政策 t的每一种潜在情形，我们假设存在一个“潜在”结果变量 $Y_n(t)$。也就是说，实验中的每个单位都对应一个（潜在的）潜在结果向量 $[Y_n(0) \ldots, Y_n(T-1)]^\top$，而实验中只能观测到其中一个分量。假设在给定分配的条件下，这些结果变量是固定的。为简便起见，在下文中我们将考虑 $T = 2$ 种处理的情形，并将它们简称为“对照组”($t = 0$)或“处理”($t = 1$)，尽管我们的讨论自然可以扩展到任意 $T > 2$。根据这些定义，我们可以正式定义“因果效应”的含义。我们关注的最重要效应（或目标估计量），也是本文的重点，是平均处理效应 [ATE]。ATE 是总体中单位接受处理($t = 1$)而非另一种对照($t = 0$)所产生的（平均）因果效应。在决策时，我们通常将处理视为相对于标准基准的一种替代政策（其成本或风险可能更高）。ATE 量化了采用这一替代策略对关注结果变量的影响。形式化地，

$$\bar{Y}_k := \sum_{n=1}^{N} \frac{Y_n(k)}{N}, \text{ and } \text{ATE} := \bar{Y}_1 - \bar{Y}_0 = \sum_{n=1}^{N} \frac{[Y_n(1) - Y_n(0)]}{N}.$$

实践中无法直接计算或观测平均处理效应，因为单位要么接受处理，要么处于对照组，但不会同时接受两者。随机对照试验或 A/B 测试用于估计平均处理效应。A/B 测试的基本机制是其随机分配机制（或触发逻辑），它决定每个单位将接受哪种体验。在最简单的情况下，每个单位 n 都被赋予一个均值为 $\pi \in (0, 1)$ 的二元随机变量：

$$J_n \sim \text{Bernoulli}(\pi). \tag{1}$$

如果 $J_n = 0$，则单位 n 会接受对照组。否则，如果 $J_n = 1$，则会呈现处理体验。通过比较对照组和处理组中各单位观察到的结果变量来估计平均处理效应。用$\mathcal{I}_t := \{n \in [N]$ 表示组 $t \in \{0, 1\}$中各单位的 $J_n = t\}$。“均值差” [DIM] 估计量就是各处理组结果变量平均值之间的差：

$$\widehat{\text{ATE}}_{\text{DIM}} := \left[\frac{1}{|\mathcal{I}_1|}\sum_{n\in\mathcal{I}_1} Y_n(1)\right] - \left[\frac{1}{|\mathcal{I}_0|}\sum_{n\in\mathcal{I}_0} Y_n(0)\right]. \tag{2}$$

这个极其简单的估计量的理论性质已得到充分理解（Splawa-Neyman 等人，1990）：它是无偏的，并且在温和条件下，在大样本中服从中心极限定理。有关全面概述和讨论，请参阅 Li 和 Ding（2017）及其中的参考文献。

3 利用协变量：广义 Oaxaca-Blinder 估计量

在收集实验数据时，我们通常还能获取以实验单元为粒度测量的其他协变量，以下记作 $\boldsymbol{z}_n := [z_{n,1}, \ldots, z_{n,K}]^\top \in \mathbb{R}^K$，适用于 $n \in [N]$ 和某个固定的 $K \in \mathbb{N}$。例如，在针对订阅视频流媒体服务的客户参与度开展实验时，我们可能可以获取客户过往的活跃度测量值、客户账户的存续时间、他们是否订阅了按次付费频道等信息。如果这些协变量是：

(C1) 与分配变量 J_n相互独立（C2）与感兴趣的结果变量 Y_n相关

它们可用于构建协变量调整估计量。关于随机实验中回归调整有效性的详细讨论，参见 Imbens 和 Rubin（2015，第 7 章）。

接下来，我们将介绍一种构建“协变量调整”估计量的通用方法。这一方法的核心直觉是，将平均处理效应的估计视为一个“缺失数据”或“插补”问题。对于每个处理t，我们可以利用组 $\mathcal{I}_t$中的观测数据拟合一个回归模型，并使用该回归模型为分配到其他处理组（即 $n \in \mathcal{I}_t^C$）的单位插补“缺失”值。也就是说，对于 $t \in \{0, 1\}$，我们使用相应处理组中的协变量和结果变量拟合回归模型 $Y_n \sim f_t(\boldsymbol{z}_n; \boldsymbol{\theta}_t)$：$\mathcal{D}_t := \{(Y_n, \boldsymbol{z}_n)\}_{n\in\mathcal{I}_t}$。其中，$\boldsymbol{\theta}_t$ 是一个有限维参数，用于表征回归模型（例如，线性回归模型的斜率和截距）。给定 $\mathcal{D}_t$，我们通过最小化在 $\mathcal{D}_t$ 上计算、由 $\boldsymbol{\theta}_t$参数化的损失函数 $\mathcal{L}$ 来估计 $\boldsymbol{\theta}$：

$$\hat{\boldsymbol{\theta}}_t \in \arg\min_{\theta} \mathcal{L}\left\{\mathcal{D}_t; \boldsymbol{\theta}\right\}. \tag{3}$$

由此得到插补算子：

$$\hat{f}_t(Y_n, \boldsymbol{z}_n, J_n; \hat{\boldsymbol{\theta}}_t) = \begin{cases} Y_n & \text{if } J_n = t, \\ f(\boldsymbol{z}_n; \hat{\boldsymbol{\theta}}_t) & \text{otherwise.} \end{cases}$$

这种方法产生了一大类“广义 Oaxaca-Blinder 估计量” [广义 Oaxaca-Blinder 估计量，Guo 和 Basse（2021）]，其形式为

$$\widehat{\text{ATE}}_{\mathcal{M}} = \frac{1}{N}\sum_{n=1}^{N}\left\{\hat{Y}_n(1) - \hat{Y}_n(0)\right\}, \tag{4}$$

ˆ ˆ ˆ 其中 $Y_n(t) = f_t(Y_n, \boldsymbol{z}_n, J_n; \boldsymbol{\theta}_t)$ 和 $\mathcal{M} := \{f_0, f_1\}$ 用于强调该估计量对所用回归函数的依赖性。我们在算法 1 中总结这一过程。

算法 1 得出的置信区间所依据的高斯假设，在大类模型的温和条件下具有渐近合理性（参见 Guo 和 Basse（2021，定理 4））。

算法 1广义 Oaxaca-Blinder 估计量

要求： 数据 $\{(Y_n,\ J_n,\ \boldsymbol{z}_n)\}_{n=1}^N$，回归模型 $\mathcal{M}=\{f_0,f_1\}$，其中 f_0 是针对 T_0 的回归模型，而 f_1针对 T_1。对于 $t\in\{0,1\}$和 $\mathcal{D}_t:=\{(Y_n,\ \boldsymbol{z}_n)\}_{n\in\mathcal{I}_t}$，拟合回归模型 f_t:

$$\hat{\boldsymbol{\theta}}_t\in\arg\min_{\boldsymbol{\theta}}\mathcal{L}\left\{\mathcal{D}_t;\boldsymbol{\theta}\right\}.$$

对于 $n=1,\ldots,N$，插补数值

$$\hat{Y}_n(t)=\hat{f}_t(Y_n,\boldsymbol{z}_n,J_n;\hat{\boldsymbol{\theta}}_t)=\begin{cases}Y_n & \text{if } t=J_n\\ \hat{f}_t(Y_n,\boldsymbol{z}_n,J_n;\hat{\boldsymbol{\theta}}_t) & \text{if } t=1-J_n.\end{cases}$$

对于 $t=0$,1，估计模型的均方误差:

$$\widehat{\mathrm{MSE}}_{\mathcal{M},t}=\frac{1}{|\mathcal{I}_t|-1}\sum_{n\in\mathcal{I}_t}\left\{Y_n-\hat{Y}_n(t)\right\}^2,$$

and

$$\widehat{\mathrm{Var}}_{\mathcal{M}}:=\frac{\widehat{\mathrm{MSE}}_{\mathcal{M},1}}{|\mathcal{I}_1|}+\frac{\widehat{\mathrm{MSE}}_{\mathcal{M},0}}{|\mathcal{I}_0|}.$$

返回 ATE 的估计值及其对应的置信区间$_{\mathcal{M}}$，按照公式 (4) 以及

$$\widehat{\mathrm{CI}}_{\mathcal{M}}(\alpha):=\widehat{\mathrm{ATE}}_{\mathcal{M}}\pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}_{\mathcal{M}}},$$

其中，z_α 是标准正态分布的累积分布函数 [CDF] 的 $100\times\alpha$% 分位数。

均值差作为 GOBE请注意，公式 (2) 中引入的均值差估计量可以视为广义 Oaxaca-Blinder 估计量。事实上，对于选择 $\hat{}f_t(Y_n,\boldsymbol{z}_n,J_n;\hat{}\boldsymbol{\theta}_t)=\frac{1}{|\mathcal{I}_t|}\sum_{n'\in\mathcal{I}_t}Y_{n'}$，平均处理效应$_{\mathrm{DIM}}$ 满足公式 (4)；在这种情况下，我们用组均值填补缺失值，而不考虑协变量的取值。

线性回归作为 GOBE标准的线性回归调整估计量是一个 GOBE。该估计量通过对以下回归进行普通最小二乘法 [OLS] 拟合得到:

$$Y_n\sim\beta_0+\beta_1J_n+\boldsymbol{\gamma}^\top\boldsymbol{z}_n+\boldsymbol{\delta}^\top J_n(\boldsymbol{z}_n-\bar{\boldsymbol{z}}),\tag{5}$$

并使用估计值 平均处理效应$_{\mathrm{LR}}:=\hat{}\beta_1$。这里，$\bar{\boldsymbol{z}}\in\mathbb{R}^K$ 是每个 K 分量在 N 个单位上的平均值。Lin (2013) 表明，渐近地，平均处理效应$_{\mathrm{LR}}$ 无偏，且方差小于平均处理效应$_{\mathrm{DIM}}$。若要将平均处理效应$_{\mathrm{LR}}$ 视为一个“GOBE”，请注意，它也可以通过分别对以下两个线性回归使用 OLS 拟合等价地得到：对于 $n\in\mathcal{I}_t,t\in\{0,1\}$，拟合 $Y_n\sim\boldsymbol{\theta}_t^\top(\boldsymbol{z}_n-\bar{\boldsymbol{z}})$。令 .....即可得到公式 (4) 的 GOBE 表述

$$\hat{f}_t(Y_n,\boldsymbol{z}_n,J_n;\hat{\boldsymbol{\theta}}_t)=\begin{cases}Y_n & \text{if } t=J_n\\ \hat{\boldsymbol{\theta}}_t^\top\boldsymbol{z}_n & \text{if } t\neq J_n.\end{cases}\tag{6}$$

参见，例如 Lin （2013年，引理 3）证明为什么公式（6）和公式（5）得出相同的估计量 .

还可以采用通用的非线性回归模型进行调整。Guo 和 Basse（2021）给出了回归模型产生无偏且渐近正态估计的条件，从而为算法 1 中的高斯近似提供了依据。在这些结果的基础上，Cohen和Fogarty（2020）提出了两步 GOBE。首先拟合一个 GOBE，然后针对每个结果变量 Y_n，使用插补值 $\hat{}\hat{}x_n:=f_{1-J_n}(Y_n,\ \boldsymbol{z}_n,\ J_n;\boldsymbol{\theta}_{1-J_n})$ 作为唯一协变量，利用线性回归模型拟合第二个 GOBE，以生成最终估计值。这种“两步” GOBE 具有渐近无偏性并服从正态分布，且比均值差估计量更高效。关于在线 A/B 测试中开发灵活模型的其他近期方法，请参见 Guo 等（2021）；Jin 和 Ba（2021）。

4 大规模实现广义 Oaxaca-Blinder 估计量

如第 1 节所述，从在线 A/B 测试中得出结论可能颇具挑战：实验通常只涉及与用户体验细节相关的小幅改动。因此，相关效应量可能非常小，即使在大样本中也很难检测到。尽管这些效应很小，却可能对后续环节产生重大影响。正因如此，采用能够精确估计因果效应的模型十分重要：对干预效应进行更精确的估计，可以帮助实践者检测到更小的效应量，并且至关重要的是，缩短获得有关处理有效性的明确结论所需的实验时间。

归根结底，我们希望拥有一个端到端的自动化推断引擎，能够针对每项实验，在无需实验人员指定应采用何种模型和协变量的情况下，给出所研究因果效应的"最佳"估计。实际上，评估哪种估计量最佳绝非易事。一方面，为感兴趣的单项实验开发定制的大型模型并精心挑选协变量，这种做法似乎有利于方差缩减；另一方面，大规模因果推断引擎必须依赖在所有实验中平均表现良好的估计量。因此，所用方法需要满足以下条件：

- 可扩展：公司通常每年会开展大量实验，而其计算资源有限。若分析运行时间过长（例如，求解算法 1 中的最小化问题），导致从业人员不得不等待结果，这是不可取的。
- 可靠：负责维护基础设施的团队相对于其服务的客户群体通常规模较小。所实施的方法需要依赖算法上可靠的程序，以便对我们关注的因果效应得出稳定的估计。
- 可解释：实验结果会被从业者用于制定政策。因此，所得到的估计必须透明、易于理解，且不需要专门知识。

基于这些原因，在第5节所展示的实验中，我们仅使用线性模型、其正则化对应模型（套索回归、ElasticNet、岭回归和主成分回归），以及一种简单的广义线性模型。将分析扩展到更复杂的模型，并评估这些模型在生产环境中的可行性，是正在进行的研究工作的一部分。

这里我们详细介绍用于拟合实验数据的回归模型，以对比不同 GOBE 估计量的表现。我们已在第3节讨论过均值差 [DIM]和简单线性回归 [逻辑回归]估计量，并将其界定为广义 Oaxaca-Blinder 估计量。岭回归、套索回归和弹性网络都是非常流行的简单线性回归模型"正则化"对应模型，它们通过惩罚项对权重向量 $\boldsymbol{\theta}$ 进行"收缩"。具体而言，给定正则化参数 $\gamma > 0$，我们对 $\boldsymbol{\theta}$ 进行优化，使损失函数最小：

$$\mathcal{L}(\mathcal{D}_t; \boldsymbol{\theta}) := \sum_{n \in \mathcal{I}_t} \left(y_n - \boldsymbol{\theta}^\top \boldsymbol{z}_n\right)^2 + \gamma \|\boldsymbol{\theta}\|_\ell^2,$$

其中，$\ell = 1$ 表示套索回归，$\ell = 2$ 表示岭回归（即使用 ℓ-1 范数或 ℓ-2 范数进行正则化）。弹性网络回归通过结合回归系数上的 ℓ_1 和 ℓ_2 惩罚项得到。形式上，在这种情况下，我们最小化损失函数

$$\mathcal{L}(\mathcal{D}_t; \boldsymbol{\theta}) := \sum_{n \in \mathcal{I}_t} \left(y_n - \boldsymbol{\theta}^\top \boldsymbol{z}_n\right)^2 + \gamma\lambda \sum_{k=1}^{K} |\theta_{t,k}| + \frac{\gamma(1-\lambda)}{2} \sum_{k=1}^{K} \theta_{t,k}^2.$$

这里，λ 权衡了 ℓ_1 和 ℓ_2 惩罚项的重要性。与线性回归和简单的均值差不同，这些正则化模型的表现很大程度上取决于某些正则化超参数的调优。为了实现流畅、自动化的模型拟合流程，我们采用标准的交叉验证方法。对于每个模型，我们针对预先设定的多个不同正则化参数值，反复最小化目标函数。对于每个参数值，我们将数据随机划分为 5 折，并通过每次迭代留出其中一折数据，将模型拟合 5 次。对于每一折，我们使用拟合得到的系数计算留出数据的决定系数，并选择在各折上平均决定系数最高的参数值作为最优正则化水平，然后使用完整数据集重新拟合模型。我们还考虑了主成分回归 [PCR] ——首先将回归变量投影到主成分上，以此降低其维度，然后将这些投影作为协变量用于线性回归；此外还考虑了一个使用 Tweedie 分布核的广义线性模型实例。

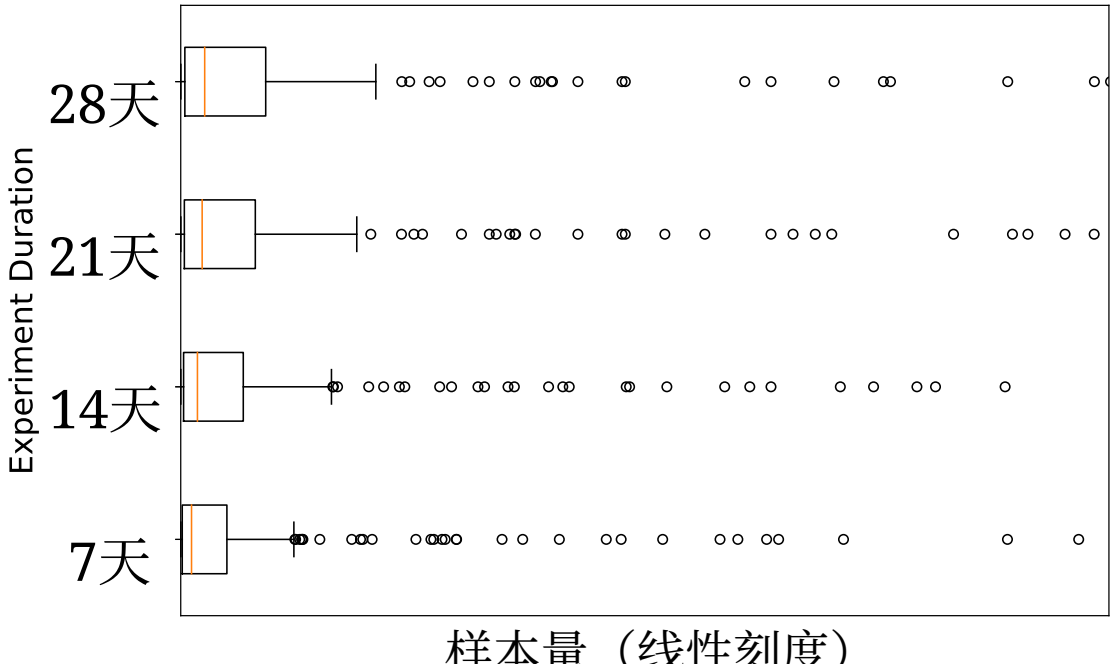

图 1：荟萃分析中所纳入实验样本量的经验分布。

5 实验

5.1 数据描述

在实验中，我们选取了一组具有代表性的 $W = 100$ A/B 测试。这些测试在过去两年中的不同时间段一直在线上运行。每个实验对应一种不同的干预。为简便起见，在分析中，每个实验我们只考虑一组两两比较（T_1 与 T_0）——尽管有些实验可能包含两个以上的处理组。对于每个实验，我们运行数据分析流程，并在累计收集 $D \in \{7, 14, 21, 28\}$天的数据后，计算因果效应估计值。无论分析时长如何，各实验的规模（T_1 组和 T_0 组中的客户总数）都存在很大差异（图 1）。

对于所有这些实验，我们都跟踪同一个关注的关键绩效指标。由于该关键绩效指标的量级在不同实验中有所差异，因此在我们的示例和分析中，我们关注百分比平均处理效应（或提升），其定义为LIFT(:) $=$ ATE$/|Y_0|$。我们使用少于 10 个暴露前的客户指标值作为协变量，这些指标与该关键绩效指标相关。

5.2 方差缩减

直观而言，更精确的估计量（估计方差更低）会带来更准确的估计，并直接促成更快、更好的决策（我们将在第 5.4 节进一步讨论）。我们通过计算模型 $\widehat{\mathcal{M}}$的估计百分比方差缩减 $\widehat{\text{VR}}_{\mathcal{M}}$（相对于基准$\widehat{\text{DIM}}$）来评估模型的精确度。回顾一下，$\text{Var}_{\mathcal{M}}$ 是模型 $\mathcal{M}$ 下平均处理效应的估计方差（根据算法 1），我们定义：

$$\widehat{\text{VR}}_{\mathcal{M}} := 100 \times \left\{ 1 - \frac{\widehat{\text{Var}}_{\mathcal{M}}}{\widehat{\text{Var}}_{\text{DIM}}} \right\}.$$

我们在图2中绘制了荟萃分析中所有实验的方差缩减随分析时长变化的情况。由此可以得出三个主要发现：

- 正如预期（例如，参见 Guo 和 Basse（2021，定理 4）），经过协变量调整的估计值通常具有较小的方差。
- 在较长的分析中观察到更大的方差缩减，这类分析的特点是客户行为更加稳定。

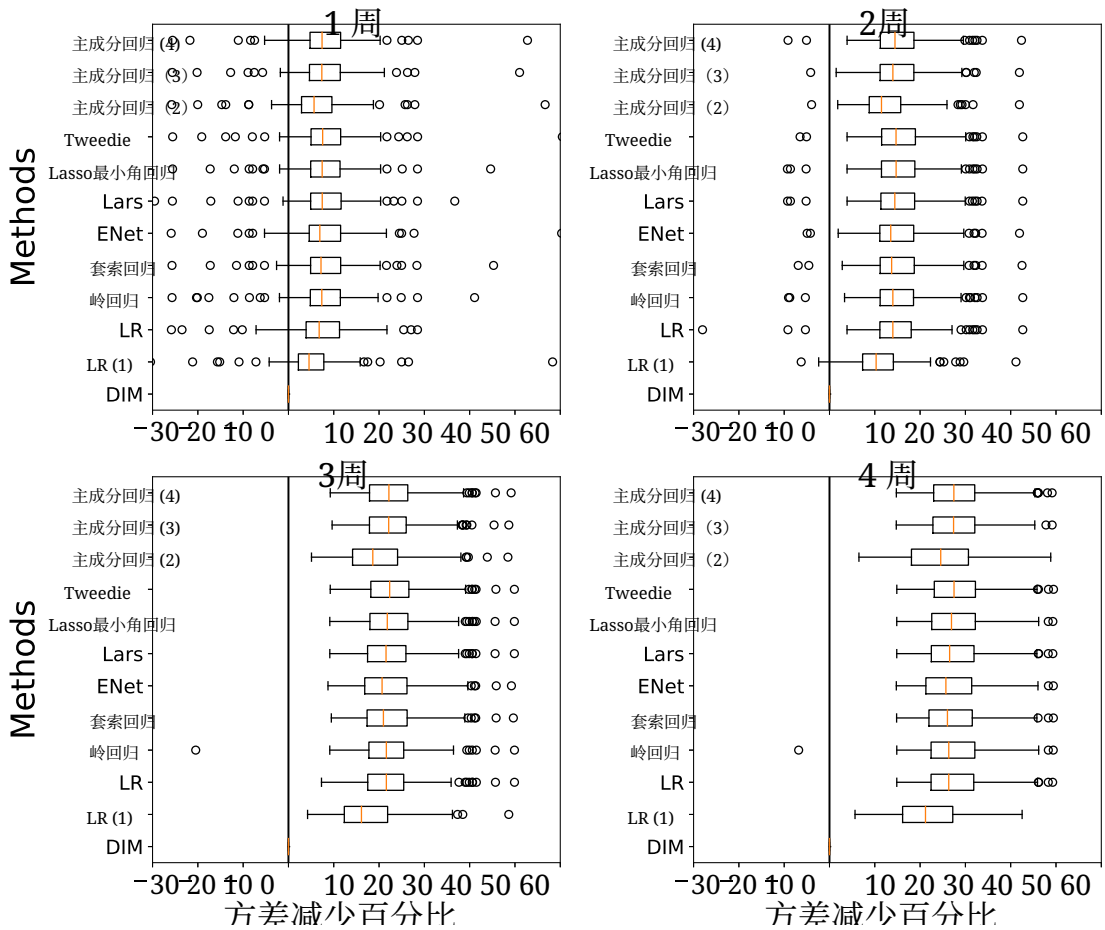

图2：分析中所考虑的各项实验的估计方差缩减箱线图。图中的每个子图对应不同的分析时长，箱线图中的每一行代表不同的模型$\mathcal{M}$。

- 不同协变量调整方法的表现相近。我们在第 5.5 节分析这些方法的表现和计算成本，并考察其与回归变量（可能含噪）数量之间的关系。

接下来，我们进一步尝试理解分析时长、样本量和模型精度之间的关系。对于给定的分析时长（例如 $D = 7$ 天），令 $F_{N,D} : \mathbb{N} \to [0, 1]$ 为实验进行 D 天后的样本量 N 的经验累积分布函数[CDF]，并令 $F_N^{-1}{}_{,D} : [0, 1] \to \mathbb{N}$ 为其逆函数。例如，$F_{N\,7,}^{-1}(0.6)$ 表示在 7天分析中，按样本量从大到小排列的实验中位于前 60% 分位的实验样本量。我们考察第一四分位数内（样本量为 $N \in [F_{N,D}^{-1}(0)\, F_N^{-1}{}_{,D}(0.25)$ 的实验）以及最后四分位数内（样本量为 $N \in [F_{N,D}^{-1}(0.75)\, F_N^{-1}{}_{,D}(1)$ 的实验）的方差缩减收益。我们观察到，在分析时长为 $D = 7$和$D = 28$时，实验表现有所不同。具体而言，对于较短的分析时长（$D = 7$ 天），方差缩减在较大的实验中尤为明显。然而，对于 $D = 28$ 较小的实验似乎从协变量调整中获益最多。更广泛地说，我们预计，协变量调整所带来的方差缩减会随实验规模和时长，以及所使用的协变量和模型而变化。我们建议实践者在采用回归模型之前、实验运行之前，对其数据进行充分分析。

5.3 对偶然不平衡的稳健性

在 A/B测试中，处理组应当在事前具有可比性。也就是说，通过随机化，处理组和对照组中各单位的协变量分布应当一致。这不仅是正确构建的 A/B测试所具备的性质，也是此类实验要产生

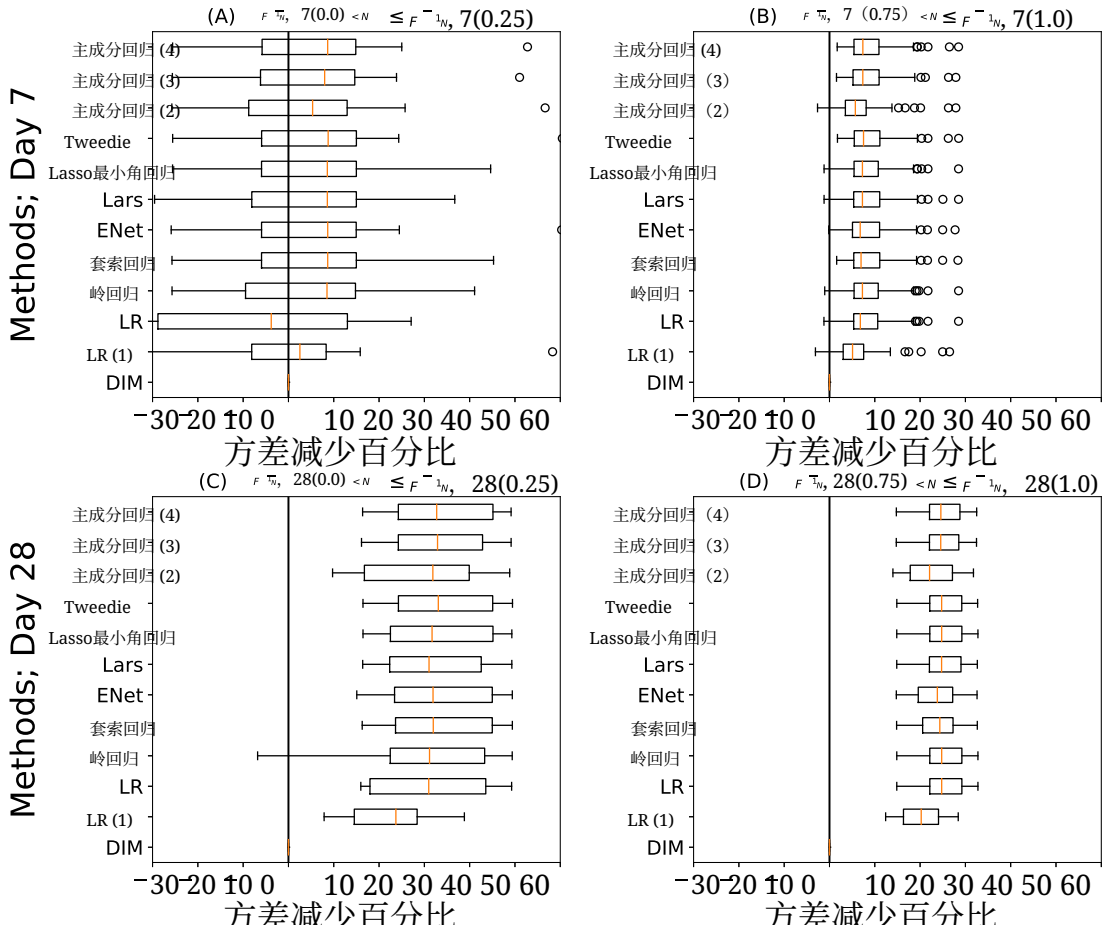

图3：估计方差缩减的箱线图。第一行：7天分析。第二行：28天分析。左侧子图（A、C）：最小的25个实验中的$\mathrm{VR}_{\mathcal{M}}$。右侧子图（B、D）：最大的25个实验中的$\mathrm{VR}_{\mathcal{M}}$。每个箱线图中的每一行对应一个不同的模型 $\mathcal{M}$。

有效推断所必须满足的基本要求。考虑这样一个实验：给定协变量 x 能够预测结果变量 y（例如，x 较高的单位往往具有较高的 y）。如果触发逻辑系统性地以更高（或更低）的概率将 x 值较高的单位分配到处理组，则第3节中的条件（C1）将被违反。因此，从 A/B测试中获得的推断将无效。

然而，即使对于处理分配的触发逻辑设定正确的实验，实际中也可能出现处理组不平衡的情况。例如，触发逻辑遵循方程（1）的实验，可能会导致数据出现一次“不走运”的划分，使处理组中的协变量取值不具可比性。具体而言，分配变量 $J_{1:N} := \{J_1, \ldots, J_N\}$ 可能会定义一个对照组 $\mathcal{I}_0$，其中的单位在实验前时期的关键绩效指标平均值远高于 $\mathcal{I}_1$（或者反过来）。当实验前时期的协变量严重不平衡时，实践者会担心是否能够相信研究结果。我们在此通过实证展示以下内容：

- 在不平衡程度较高时，均值差估计量可能会系统性地导致错误结论。换句话说：DIM 在不考虑协变量不平衡的情况下对平均处理效应的估计是无偏的，但在特定条件下可能存在偏差。参见图4。
- 协变量调整方法能够缓解这一问题，并且对实验前随机不平衡具有稳健性。参见图5。

注意：本节重点讨论简单线性回归——其他方法的定性结论类似，因此不再赘述。首先，我们需要对不平衡给出一个操作性定义，以量化对照组和处理组的实验前可比性。令 x_n 表示单位 Y_n所关注的关键绩效指标的实验前取值

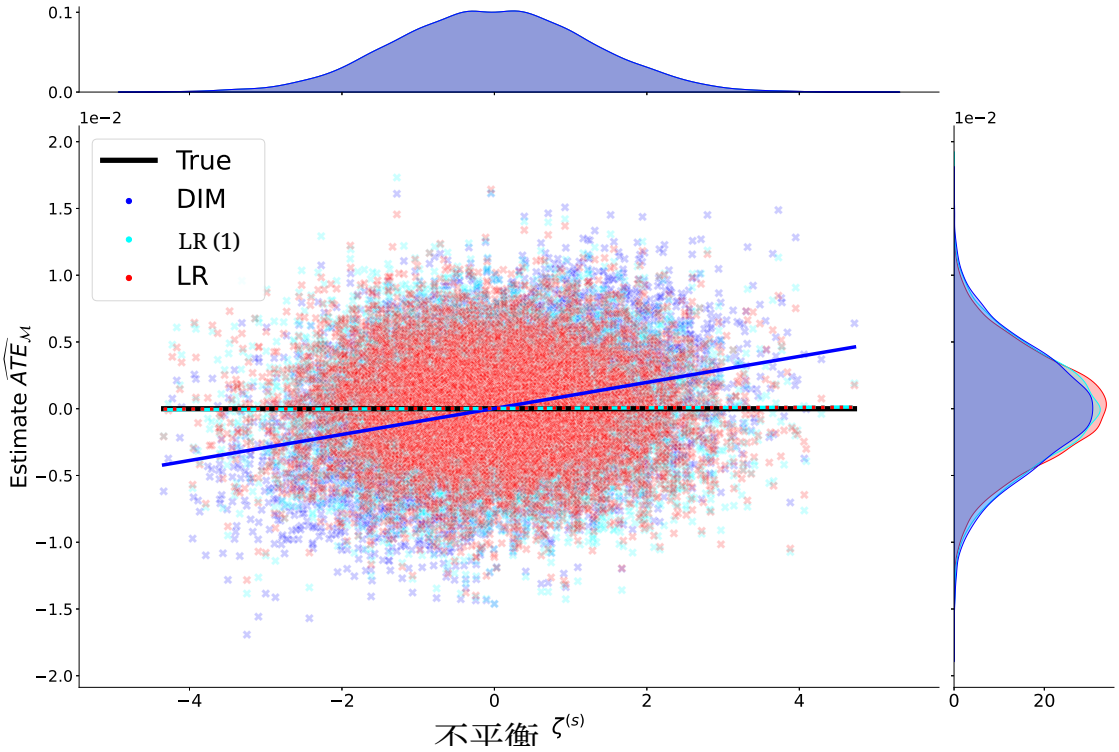

图4：比较单个实验在 $S = 10{,}000$ 次重新随机化下的 DIM 和 LR 估计值。每个点对应一个不同的（$\zeta^{(s)}$, $\widehat{\text{平均处理效应}}_{\mathcal{M}}$）组合。

，并定义不平衡参数 ζ:

$$\zeta := \zeta(x_{1:N}, J_{1:N}) = \bar{x}(1) - \bar{x}(0), \tag{7}$$

其中， $\bar{x}(t) := \sum_{n=1}^{N} x_n \mathbb{1}(J_n = t) / \{\sum_{n=1}^{N} \mathbb{1}(J_n = t)\}$ 表示后来接受处理组 t处理的单位在实验前时期的关键绩效指标平均值。直观地说，当 $|\zeta|$ 较大时，对照组和处理组在事前并不具有可比性。当不平衡足够严重时，实验人员会担心估计值可能不可信。相应地，这通常会导致必须对实验进行重新随机化。这会造成实验流程效率低下：重新随机化成本高昂，因为它需要使用额外的计算资源，并推迟发布决策。因此，推断对随机化偏差不那么敏感的方法更为可取。

现在，我们通过真实数据表明，即使存在严重不平衡，使用协变量调整估计量也能比未调整估计量取得显著更好的结果。进而，这会减少重新随机化的必要性和实验成本。

5.3.1 A/A 分析

为了研究协变量调整对偶然不平衡的稳健性，我们在此采用以下“A/A”方法。给定一项感兴趣的实验，我们将注意力限制在该实验的单个组别（例如对照组）。也就是说，对于固定的 t，我们只考虑暴露于政策 t 的单位子集 $n \in \mathcal{I}_t$，并丢弃所有其他单位及其协变量。为简化符号，在下文中我们用 $\mathcal{I} := \mathcal{I}_t$ 表示处理组 t 中的单位集合。经过这一步预处理后，我们可以将 $\mathcal{I}$ 中的单位视为来自一次 A/A 检验。也就是说，如果我们将它们随机分成两个“虚拟”处理组，那么按照构造，真实平均因果效应已知且等于 0。类似的实验设置也见于 Guo 和 Basse（2021，第4节）。

5.3.2 蒙特卡洛模拟

为评估对不平衡的稳健性，我们采用蒙特卡罗方法。固定一个较大的整数 S，对于每个 s =1、.....、 S，我们将 $\mathcal{I}$随机分成两组，形成 A/A 重新随机分组 $\mathcal{I}_0^{(s)}$， $\mathcal{I}_1^{(s)}$，使得 $\mathcal{I}_0^{(s)} \cup \mathcal{I}_1^{(s)} =$

$\mathcal{I}$ 和 $\mathcal{I}_0^{(s)} \cap \mathcal{I}_1^{(s)} = \varnothing$。随后，我们定义 A/A组指示变量 $J_n^{(s)} := \mathbb{1}(n \in \mathcal{I}_1^{(s)})$，并使用算法 1，根据数据 $\mathcal{D}_t^{(s)} = \{Y_\mathcal{I}, J_\mathcal{I}^{(s)}, X_\mathcal{I}\}$ 为所有待考察的估计量 $\mathcal{M}$ 拟合 $\widehat{\text{平均处理效应}}_\mathcal{M}^{(s)}$。这里 $Y_\mathcal{I} = \{Y_n : J_n \in \mathcal{I}\}$。也就是说，我们假设对照组中的单位索引为 $\mathcal{I}_0^{(s)}$，处理组中的单位索引为 $\mathcal{I}_1^{(s)}$，并据此拟合 $\widehat{\text{平均处理效应}}$。重要的是，请注意，对于每次重新随机化 s，我们都会产生一个特定于该次划分的不平衡程度 $\zeta^{(s)}$，如公式（7）所示。此外，根据构造，平均处理效应为 0，因为此处我们令结果变量 Y_n 保持固定，不受 $J_n^{(s)}$取值影响。我们在算法2中概述了这一过程。

算法2A/A 检验

要求：数据 $\mathcal{D} := \{(y_n, \boldsymbol{z}_n)\}_{n\in\mathcal{I}}$、回归模型集合 $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_W\}$、处理组 t。令 $\mathcal{I} := \{n \in [N] : J_n = t\}$。对于 $s = 1, \ldots, S$，执行：随机拆分 $\mathcal{I}$ 为 $\mathcal{I}_0^{(s)}, \mathcal{I}_1^{(s)}$，使得 $\mathcal{I}_0^{(s)} \cup \mathcal{I}_1^{(s)} = \mathcal{I}$ 且 $\mathcal{I}_0^{(s)} \cap \mathcal{I}_1^{(s)} = \varnothing$。令 $J_n^{(s)} := \mathbb{1}(n \in \mathcal{I}_1^{(s)})$。令 $\bar{x}(\mathcal{I}_t^{(s)}) := \sum_{n\in\mathcal{I}_t^{(s)}} x_n / |\mathcal{I}_t^{(s)}|$并计算：对于 $\mathcal{M} \in \mathfrak{M}$，执行：使用 $\mathcal{D}^{(s)} := \{Y_\mathcal{I}, J_\mathcal{I}^{(s)}, X_\mathcal{I}\}$，通过算法 1 估计 $\widehat{\text{平均处理效应}}_\mathcal{M}^{(s)}$，$\widehat{\text{置信区间}}_\mathcal{M}^{(s)}(\alpha)$。结束循环。结束循环。

$$\zeta^{(s)} = \bar{x}(\mathcal{I}_1^{(s)}) - \bar{x}(\mathcal{I}_0^{(s)}).$$

5.3.3 验证

完成上述蒙特卡洛过程后，我们便能获得估计量 $\text{ATE}_\mathcal{M}$ 随不平衡程度 ζ 变化的经验双变量分布。图4展示了不同估计量的表现：我们在荟萃分析中某个单一实验的对照组上运行了算法2，$\widehat{\text{取}}$ $S = 10,000$。我们绘制了不平衡程度 ζ（横轴）与$\mathcal{M} \in$ {DIM，LR1，LR} 的 $\text{ATE}_\mathcal{M}$（纵轴）散点图。这里，LR1 表示仅使用 KPI 的实验前数值进行回归，而 LR 则使用所有可用的协变量。图的上方绘出了不平衡程度的边缘分布 $\widehat{\zeta}$（如预期，其中心位于 0），右侧则绘出了各估计量 $\text{ATE}_\mathcal{M}$ 的边缘密度。我们还绘制了$\widehat{\text{一条实线}}$，表示 $\text{ATE}_\mathcal{M}$ 对 ζ 的线性拟合。从图4可以明显看出，ATE_DIM 是无条件无偏的（其在重新随机化 $s = 1, \ldots, S$ 下的期望等于真实 ATE 0），但在给定不平衡程度的条件下并非无偏。

由于真实潜在平均因果效应的真实值已知，这一信息可直接转化为估计量准确性关于不平衡程度的联合分布。对不平衡程度不太敏感的估计量，能让实验人员即使在存在此类不平衡时，也对所得结果充满信心。反之，对不平衡程度敏感的估计量，会使实验人员在实验前协变量不平衡时对研究结果产生怀疑。相应地，稳定估计量将使实验流程更加高效，即使来自“运气不佳的分组”的数据仍可用于对所关注的因果效应得出结论。

现在，我们介绍一些指标，以便将这一直觉转化为对估计量质量随不平衡程度变化的定量评估。固定一个值 $\kappa \in \mathbb{N}$，并创建指标集 $\mathcal{S}_1$、……、$\mathcal{S}_\kappa$，其中对于每个 $j = 1$、……、κ，指标集 $\mathcal{S}_j \subset \{1, \ldots, S\}$ 包含与 $\zeta^{(s)}$ 的经验分布中位于第 $100 \times \frac{j-1}{\kappa}$ % 至第 $100 \times \frac{j}{\kappa}$ % 分位数范围内的值相关联的指标。接着，根据 ($\zeta^{(s)}$) 的（排序后）取值，将 ($\widehat{\text{平均处理效应}}_\mathcal{M}^{(s)}$，$\zeta^{(s)}$) 的取值划分为 κ 个等量分组（每组占 κ 分位），对于每个 $j = 1$、……、κ，令$G_{\mathcal{M},\,j} : \mathbb{R} \to [0, 1]$ 为落入第 j 个分桶的估计值 $\widehat{\text{平均处理效应}}_\mathcal{M}^{(s)}$ 的经验累积分布函数，并令 $G_{\mathcal{M},\,j}^{-1} : [0, 1] \to \mathbb{R}$ 为其逆函数（例如，对于 $\kappa = 10$，我们通过$G_{3,\,\mathcal{M}}^{-1}(0.5)$ 得到落在第 30 至第 40 百分位之间的估计平均处理效应的中位数）。在每个 κ 个分桶中，我们计算

- 第 j个分桶中的均方误差估计值：

$$\widehat{\text{MSE}}_{\mathcal{M},j} = |\mathcal{S}_j|^{-1} \sum_{s\in\mathcal{S}_j} \left(\widehat{\text{ATE}}_\mathcal{M}^{(s)} - \text{ATE}\right)^2. \tag{8}$$

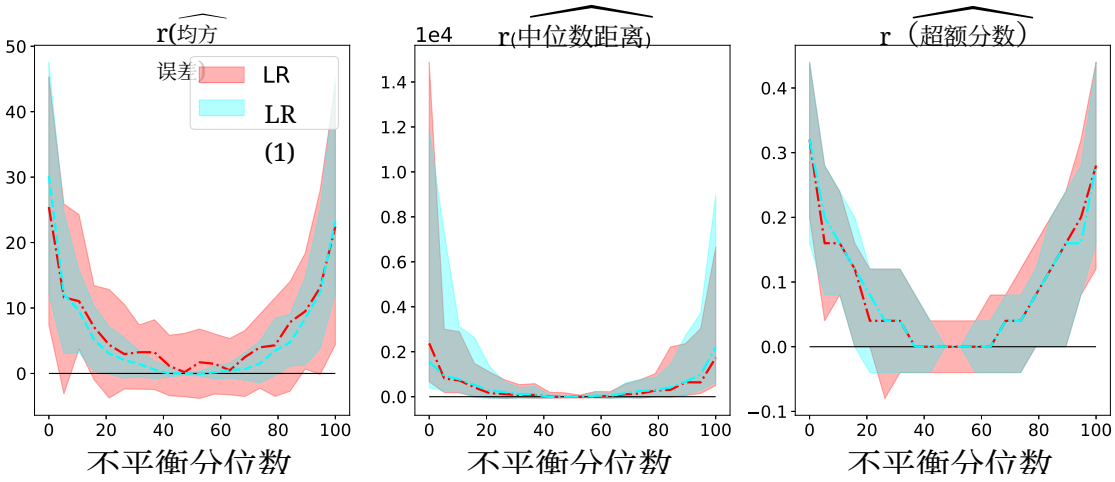

图5：分析第7天各 M 实验中，公式(8)至(10)所引入的稳健性指标的相对对应量（公式(11)）汇总。值越高，表示DIM相对于LR的敏感性越高，LR1。实线：实验间的中位数；阴影区域：（25%–75%）百分位数。

- 第 j个分桶中 $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ 的中位数与真实值ATE之间的平方距离：

$$\widehat{\text{Mediandist}}_{\mathcal{M},j} = (G_{\mathcal{M},j}^{-1}(1/2) - \text{ATE})^2. \tag{9}$$

- 超额 $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ 在第 j个分桶中低估或高估真实处理效应ATE的比例：

$$\widehat{\text{Excessfrac}}_{\mathcal{M},j} = \frac{\max\{\hat{q}_{\mathcal{M},j}^{+}, 1 - \hat{q}_{\mathcal{M},j}^{-}\} - 1/2}{1/2}, \tag{10}$$

其中，$\hat{q}_{\mathcal{M},j}^{+} = \arg\min_\alpha\{G_{\mathcal{M},j}^{-1}(\alpha) \geq \text{ATE}\}$ 是与第 j个分桶中高于ATE的最小值 $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ 相关联的分位数，而$\hat{q}_{\mathcal{M},j}^{-} = \arg\max_\alpha\{G_{\mathcal{M},j}^{-1}(\alpha) \leq \text{ATE}\}$是与第 j个分桶中低于真实ATE的最大值 $\widehat{\text{ATE}}_{\mathcal{M}}^{(s)}$ 相关联的分位数。注意：当估计值的分布以真实值为中心时，公式(10)接近0。在存在较大条件偏差时，它趋近于1。

由于这三个指标都可以视为损失概念（值越低越好），而且我们希望评估协变量调整方法 $\mathcal{M}$ 是否比默认的未调整方法 DIM 带来更低的损失，因此我们也将方法 $\mathcal{M}$相对于 DIM 的“相对”对应指标定义为

$$r(\widehat{\text{Metric}}_{\mathcal{M},j}) := \frac{\widehat{\text{Metric}}_{\text{DIM},j} - \widehat{\text{Metric}}_{\mathcal{M},j}}{\widehat{\text{Metric}}_{\text{DIM},j}}, \tag{11}$$

$\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$ $\widehat{\quad}$对于指标 $\in$ {MSE、Mediandist、Excessfrac}。对于这些相对指标，值越大，表示均值差估计量 ATE_{DIM} 相对于另一种协变量调整估计量 $\text{ATE}_{\mathcal{M}}$ 对不平衡程度 $\zeta^{(s)}$的敏感性越高 $\widehat{\quad}$ $\widehat{\quad}$。我们报告了在分析日 $D = 7$，对照组中所有 $W = 100$ 个实验的这些相对指标值。具体来说，对于每个实验，我们针对 $S = 10,000$ 次重新随机化运行算法2，并在 $\kappa = 20$ 个分桶中计算相对指标（图5）。对于每个实验和每个分桶，我们保留相对指标的中位数，并针对所有不平衡分位数 $\mathcal{S}_j$，即 $j = 1$、.....、κ，绘制一条实线，连接 W个实验中位数的中位数；阴影区域则表示 50% 的中心经验区间。我们发现，各项指标的表现相似：当不平衡程度 $|\zeta^{(s)}|$ 较小时（即接近重新随机化结果中的中位数），这些指标接近于零。然而，随着不平衡程度 $|\zeta^{(s)}|$ 的绝对值增加，相对指标值也会急剧上升，这表明均值差估计量相对于线性调整估计量（使用一个或多个协变量）对不平衡更加敏感。最后，我们检查估计量的校准情况（图6）。给定宽度 $\alpha = 0.95$,

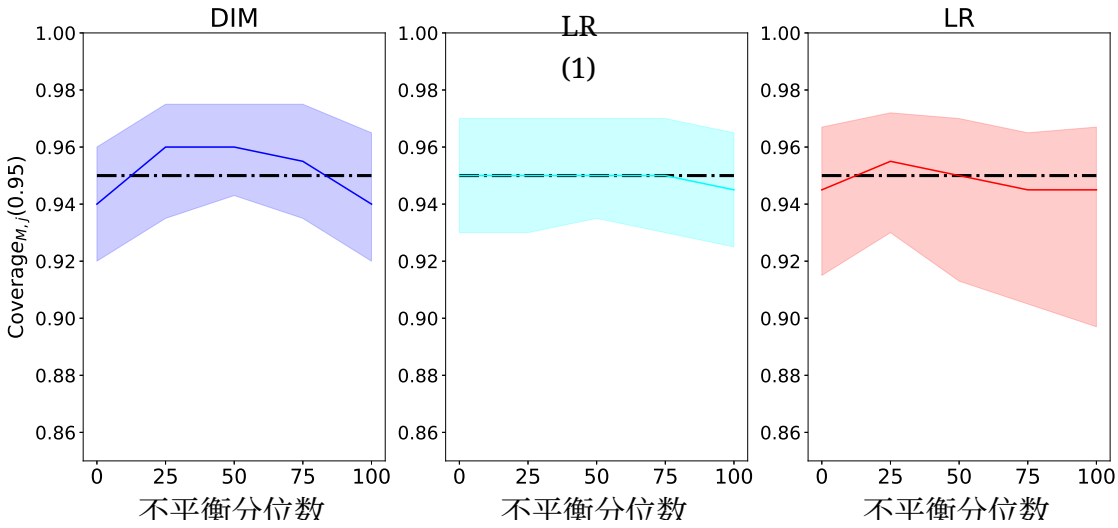

图6：覆盖率摘要$\mathcal{M}, j(\alpha)$，针对 $\mathcal{M} \in \{$DIM、LR1，LR$\}$（纵轴，从左至右），在 $\alpha = 95\%$ 的 $\kappa = 5$ 分桶上（横轴）。实线表示 M 个实验的中位覆盖率，阴影区域覆盖这些实验的第 10% – 90% 百分位数。结果以第 $D = 7$天的分析为基准。黑色实线表示目标标称值 95%。

我们计算出 the 每个分桶中，估计的置信区间包含真实值的次数所占比例 $\mathcal{S}_j$ ：

$$\widehat{\text{Coverage}}_{\mathcal{M},j}(\alpha) = \frac{\sum_{s \in \mathcal{S}_j} \mathbb{1}\left(\text{ATE} \in \widehat{\text{CI}}_\alpha^{(s)}\right)}{|\mathcal{S}_j|}. \tag{12}$$

达到名义覆盖率 α 意味着置信区间校准良好。对于这一稳健性指标，均值差估计量的表现对实验前不平衡的敏感度低于图5中所考察的指标。

5.4 对实验时间的影响

接下来，我们说明较小的估计方差如何缩短实验时间。我们考虑一个假设检验框架，其中 H_0 是处理无效应的零假设—— H_0： $\{$LIFT $= 0\}$ ——而 H_1 是一个固定备择假设，即处理的效应大小固定为 δ%， H_1： $\{$LIFT $= \delta\}$。基于迄今收集的数据（例如，分析的前 $D = 7$ 天），我们预测随着实验进行，未来将触发的单位数量（采用近期的样本量预测方法，参见 Masoero 等人 (2022)；Richardson 等人 (2022)；Camerlenghi 等人 (2022)）。根据这些假设、预测以及 $D = 7$处的估计方差，我们计算出未来的第一个天数 D'： 在该天数，我们预计能够以至少 $\pi = 80\%$的检验功效，在备择假设 H_1 下，以给定的显著性水平 $\alpha = 0.05$拒绝无效应的零假设。如图 1 所示，更高的精度（更小的方差）会直接缩短实验时间。我们再次看到，所考虑的不同协变量调整方法表现非常相似。

我们强调，图 7 中的预测取决于多种因素：既包括实验的属性（例如，关键绩效指标的观测均值和方差），也包括假设的固定效应值 δ。然而，图 7 所示的趋势——即估计方差越小，检验效能越高，因而实验持续时间越短——是符合预期的：估计方差越小，实验时间就会直接缩短。

5.5 大规模权衡：计算、稳健性、可解释性

正如第4节中已经讨论的，大规模推断引擎的设计应考虑其运行规模所带来的约束。我们已经讨论过，线性模型及其正则化具有稳健性（例如，能够抵御偶然不平衡），并且具备可解释性。这里，我们分析计算成本和估计准确性如何随（含噪）附加协变量数量的增加而变化。具体而言，我们通过使用附加伪协变量扩充 K 协变量，测试计算和准确性受到的影响。为此，我们针对每个协变量 k 计算（经验）一阶矩 $\hat{}\mu_k$ 和二阶矩 $\hat{}\sigma_k^2$，并针对每个 $n = 1, \ldots, N$，抽取一组的

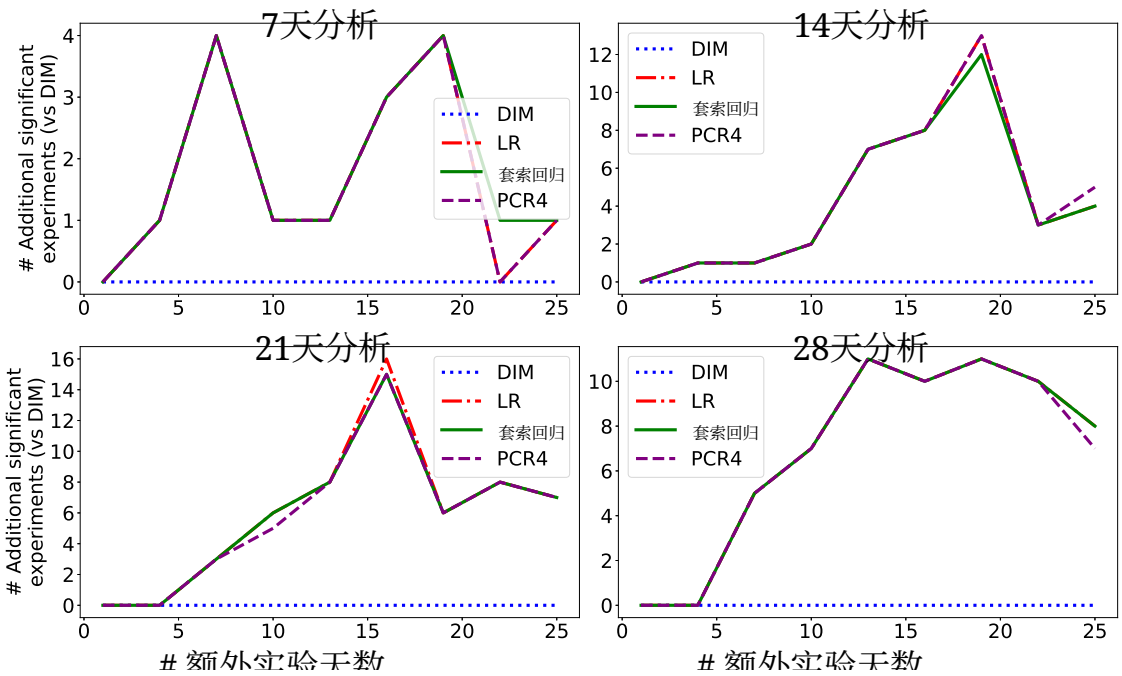

图7：对于给定的额外实验天数（横轴），我们绘制了额外实验的数量（纵轴）。在给定 H_1 的情况下，当使用模型 $\mathcal{M}$ 而非默认 DIM 获取估计值时，对于这些实验， H_0 在显著性水平为 $\alpha = 0.05$ 且统计功效为80%的情况下可以被拒绝。不同子图对应不同的实验持续时间。

伪协变量 $\tilde{z}_{n,k} \sim \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k)$ 独立同分布，对于 $k = 1, \ldots, K$。这会生成一组扩充后的协变量，$\tilde{\boldsymbol{z}}_n := [z_{n,1}, \ldots, z_{n,K}, \tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K}]^\top$。在我们的实验中，我们还考虑了更大的协变量集合，方法是对每个 $n = 1, \ldots, N$，从每个核 $\mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k)$ 中抽取 L 次。在抽取 L 个伪折的一般情况下，所使用的协变量为：

$$\tilde{\boldsymbol{z}}_n := [\overbrace{z_{n,1}, \ldots, z_{n,K}}^{\text{Real Covariates}}, \overbrace{\tilde{z}_{n,1}, \ldots, \tilde{z}_{n,K}}^{\text{First Spurious Fold}}, \overbrace{\tilde{z}_{n,K+1}, \ldots, \tilde{z}_{n,2K}}^{\text{Second Spurious Fold}}, \ldots, \underbrace{\tilde{z}_{n,(L-1)K+1}, \ldots, \tilde{z}_{n,LK}}_{L-\text{th Spurious Fold}}]^\top \in \mathbb{R}^{(L+1)\times K}. \tag{13}$$

公式（13）模拟了一种情形：我们可能会使用一组规模庞大、未经筛选的协变量，其中一些协变量含噪且与结果变量不相关（违反第3节中的条件（C2））。我们在图 8 中分析了运行不同协变量调整方法相对于基准 DIM 的计算成本，并考察其与实验规模及所用协变量数量的关系。在我们的实验中，即使实验规模较大，协变量调整方法的计算成本也适中。我们使用热门的 Python 科学计算库 scipy（Virtanen 等，2020）进行实验，运行环境为 16 核 Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz。计算成本随样本量增长的速度快于随协变量维度增长的速度。

接下来，我们评估不同回归调整方法对含噪协变量的稳健性。由于真实的因果效应未知，我们将使用完整数据拟合线性回归模型得到的估计值视为“真实值”——即，平均处理效应 $:=$ 平均处理效应$_{\text{LR}}$，数据为 $\{y_{1:N}, J_{1:N}, \boldsymbol{z}_{1:N}\}$。对于固定的 $L \geq 1$ 和大量蒙特卡罗随机抽样 S，我们 (i) 抽取伪协变量 $\tilde{\boldsymbol{z}}_{1:N}$，并 (ii) 计算存在伪协变量时估计值相对于真实值的百分比绝对差异（或误差）的经验分布：

$$\widehat{\text{err}}^{(s)}_{w,\mathcal{M},L} := \frac{|\widehat{\text{ATE}}^{(s)}_{w,\mathcal{M},L} - \text{ATE}_w|}{|\text{ATE}_w|}.$$

对于实验 w，$\widehat{\text{ATE}}^{(s)}_w$，$\mathcal{M}$，$L$ 是使用模型 $\mathcal{M}$ 对协变量进行第 s 次随机重抽样（包含 L 个伪折）后得到的真实平均处理效应的估计值。结果展示在图9中，我们按以下方式拆解误差分布

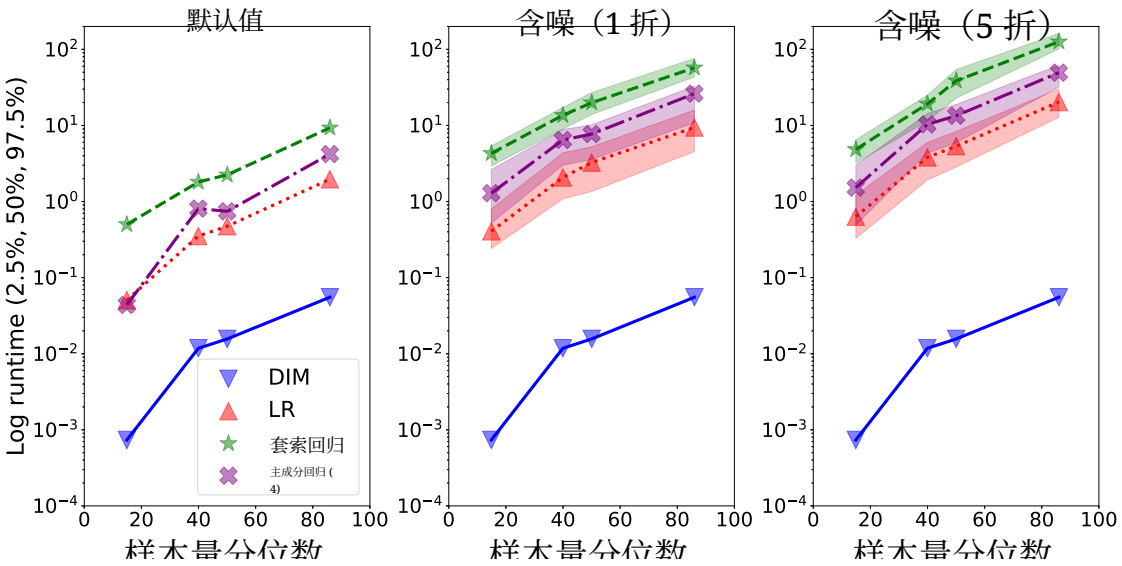

图 8：四个代表性数据集上的对数计算时间（纵轴），这些数据集的样本量为 $N = F^{-1}_{N\,7,}(\alpha)$，$\alpha \in \{0.15, 0.4, .06, 0.85\}$（横轴）。左图使用默认的 K 协变量集。中图加入一折 K 伪协变量，右图则按照公式（13）加入五折伪协变量。在后两种情况下，我们在 100 个随机抽取的不同含噪协变量集上重复拟合 $N_{MC} = 100$ 次。

根据样本量对荟萃分析中的实验进行聚类。具体而言，我们按照 $F_{N,\ 7}$ ((A)–(D)) 将实验分为四个四分位数。在每个四分位数内，我们针对每个实验 w 以及每个
方法 $\mathcal{M}$ 和折 L 在 $s = 1$、.....、S 上的误差中位数（$\text{med}_{1:S}(\widehat{\ }\text{err}^{(1:S)}_{w\ \mathcal{M},\ ,\ L}$）。对于样本量分布的每个四分位数，该过程会为每个 $\mathcal{M}$，L生成一个包含 25 个值的列表。我们在图9中绘制了这些 25 个误差中位数的中位数（纵轴），其横轴为不同方法下的伪折数量。就点估计而言，即使存在多个含噪协变量，所考虑的估计量也对噪声极其稳健。正如图 10 所示，加入伪含噪协变量对方差缩减收益的影响也有限。这表明，在存在含噪协变量时，协变量调整方法的性能是可靠的；对于经过良好优化的库而言，即使样本量很大，计算成本也并不高。

6 讨论

本文探讨了在在线 A/B 测试中采用一大类协变量调整模型（广义 Oaxaca-Blinder 估计量）来估计因果效应的价值与潜力。GOBE 依赖于一种简单而通用的流程，如算法 1 所述。通过利用额外的协变量并采用线性和非线性回归模型，我们在第 5 节基于真实数据开展的大量实验中表明，这些估计量能够对我们关注的因果效应给出精确（第 5.2 节）且稳健（第 5.3 节）的估计，其表现远优于简单的均值差估计量。采用这些估计量有助于从业者在更短时间内了解所测试干预的效果，降低实验成本，并加快有益创新的应用（第 5.4 节）。为了获得这些更精确的估计，需要付出前期成本，既包括计算成本，也包括统计成本。我们在第 5.5 节讨论了这些不足，并在荟萃分析中分析线性模型及其正则化版本的性能和计算成本。具体而言，我们关注此类性能如何随实验规模和协变量维度变化。我们发现，对于所考察的模型，即使实验规模较大，计算成本也不会高到难以承受；即使存在多个伪协变量，推断仍然可靠。

考虑到第4节概述的实际问题和期望，我们选择只考察可解释的简单线性模型及其正则化版本。不过，我们强调，第3节介绍的用于估计因果效应的广义 Oaxaca-Blinder 框架，也可以直接应用于复杂的非线性回归函数（例如神经网络）。拟合灵活的非线性回归模型通常需要求解复杂的非凸优化问题（例如公式 (3) 中的最小化问题），

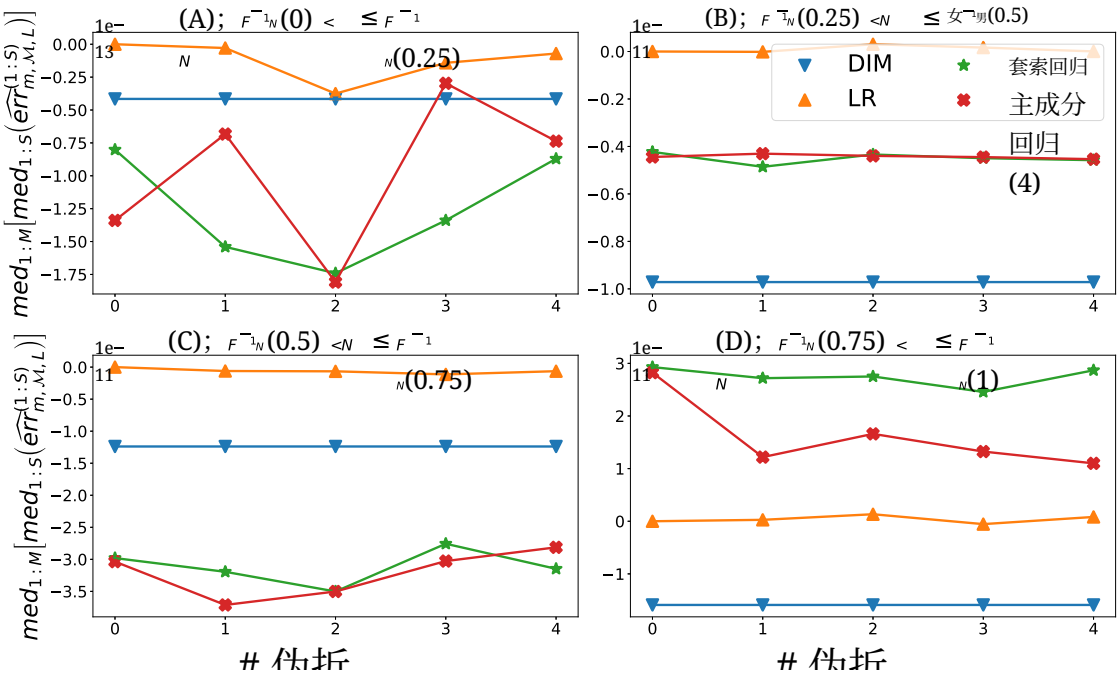

图9：分析第$D = 7$天的误差中位数（纵轴）随伪折数量（横轴）变化的情况。每个子图对应按样本量划分的不同实验四分位组，样本量依据$F_{N,\ 7}$确定。

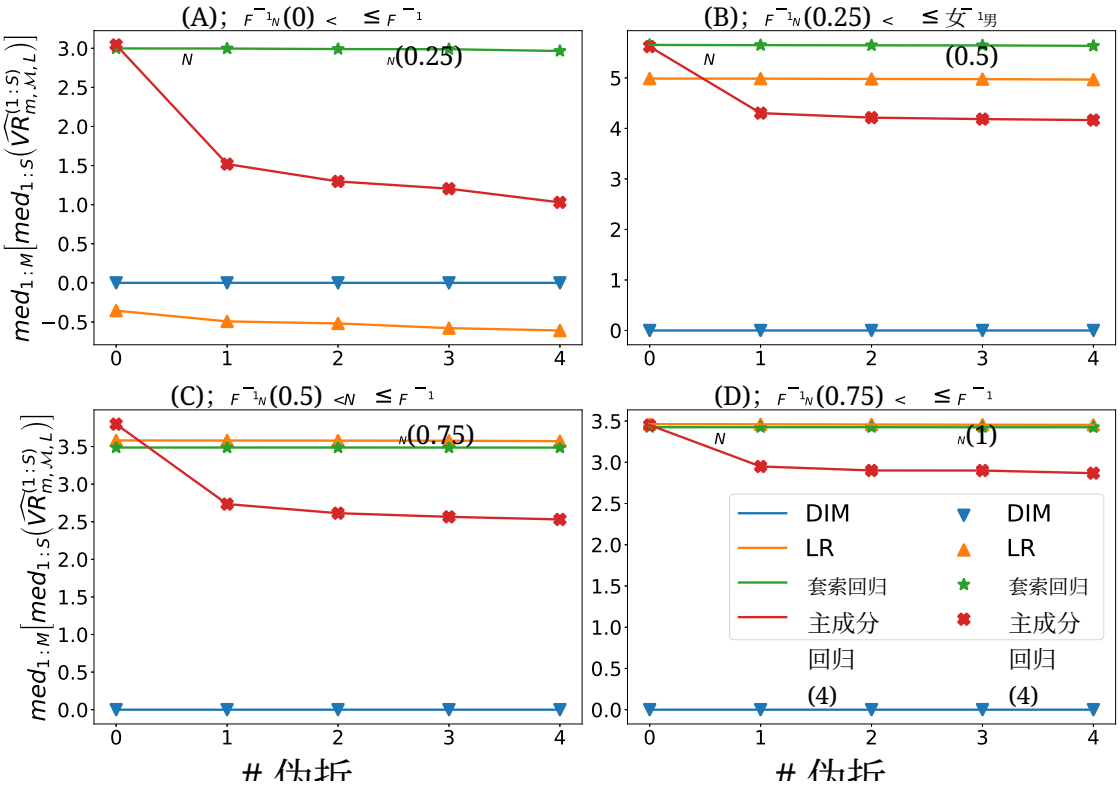

图 10：分析第 $D = 7$ 天，中位方差缩减 $\mathrm{VR}_{\mathcal{M}}$（纵轴）随伪折数（横轴）的变化。每个子图对应按样本量划分的不同四分位数实验，如$F_{N,\ 7}$所示。

并且可能需要采用第4节讨论的交叉拟合方法来调整正则化参数。设计自动化执行这些流程的范式，以及开展相关的成本效益分析，是一个活跃的研究领域。在不同于我们所研究的情境中，灵活的非线性方法有望大幅优于本文讨论的简单线性方法。参见例如 Guo 等（2021）中的讨论。

我们设想未来研究有许多令人振奋的方向。在方法论方面，Guo 和 Basse（2021）为建立一个框架奠定了基础，该框架能够为一大类用于估计因果效应的回归模型提供可证明的保证。拓展这些保证所适用的模型类别，是未来研究的一个令人振奋的方向（Cohen 和 Fogarty，2020；List 等人，2022）。此外，简化保证成立所需的条件，并强化对其特征的刻画，或许能进一步提升这些方法的普及程度。在应用方面，开发自动化流程，以便在规模庞大且异质的数据集中识别最优回归函数，是推动这些方法大规模应用的关键一步。为实现这一目标，深入考察采用规模大、灵活的协变量调整回归方法所带来的益处、成本和风险，将是未来几年从业者面临的重大挑战。

参考文献

Federico Camerlenghi、Stefano Favaro、Lorenzo Masoero 和 Tamara Broderick。2022 年。用于贝叶斯非参数估计未观测遗传变异的缩放过程先验。《美国统计学会杂志》（2022），1–12。Peter L Cohen 和 Colin B Fogarty。2020 年。广义 Oaxaca-Blinder 估计量的无损校准。*arXiv* 预印本 *arXiv:2012.09246*（2020）。Alex Deng、Ya Xu、Ron Kohavi 和 Toby Walker。2013 年。利用实验前数据提高在线对照实验的敏感性。载于 第六届美国计算机协会国际网络搜索与数据挖掘会议论文集。123–132。Kevin Guo 和 Guillaume Basse。2021 年。广义 Oaxaca-Blinder 估计量。《美国统计学会杂志》（2021）。Yongyi Guo、Dominic Coey、Mikael Konutgan、Wenting Li、Chris Schoener 和 Matt Goldman。2021 年。在线实验中的方差缩减机器学习方法。神经信息处理系统大会*2021*（2021）。Somit Gupta、Ronny Kohavi、Diane Tang、Ya Xu、Reid Andersen、Eytan Bakshy、Niall Cardin、Sumita Chandran、Nanyu Chen、Dominic Coey 等。2019 年。首届在线对照实验峰会面临的主要挑战。*ACM SIGKDD*探索通讯 21，1（2019），20–35。Guido W Imbens 和 Donald B Rubin。2015 年。统计学、社会科学与生物医学科学中的因果推断。剑桥大学出版社。Ying Jin 和 Shan Ba。2021 年。在线对照实验中的最优方差缩减。*arXiv* 预印本 *arXiv:2110.13406*（2021）。Xinran Li 和 Peng Ding。2017 年。有限总体中心极限定理的一般形式及其在因果推断中的应用。《美国统计学会杂志》112，520（2017），1759–1769。Winston Lin。2013 年。关于实验数据回归调整的不可知论笔记：重新审视 Freedman的批评。《应用统计年鉴》7，1（2013），295–318。John A List、Ian Muir 和 Gregory K Sun。2022 年。经济实验中高效灵活回归调整的机器学习方法。技术报告。美国国家经济研究局。Lorenzo Masoero、Federico Camerlenghi、Stefano Favaro 和 Tamara Broderick。2022 年。以少博多：通过贝叶斯非参数方法预测并最大化基因组变异发现。《生物计量学》109，1（2022），17–32。Thomas S Richardson、Yu Liu、James McQueen 和 Doug Hains。2022 年。在线活动样本量的贝叶斯模型。载于国际人工智能与统计会议。机器学习研究论文集，1775–1785。Donald B Rubin。1977 年。基于协变量的治疗组分配。《教育统计学杂志》2，1（1977），1–26。Jerzy Splawa-Neyman、Dorota M Dabrowska 和 TP Speed。1923/1990 年。概率论在农业实验中的应用。原理论述。第 9 节。《统计科学》（1923/1990），465–472。John W Tukey。1991 年。在临床试验中使用多个协变量。《国际统计评论/*Revue Internationale de Statistique*》（1991），123–137。Pauli Virtanen、Ralf Gommers、Travis E. Oliphant、Matt Haberland、Tyler Reddy、David Cournapeau、Evgeni Burovski、Pearu Peterson、Warren Weckesser、Jonathan Bright、St´efan J. van der Walt、Matthew Brett、Joshua Wilson、K. Jarrod Millman、Nikolay Mayorov、Andrew R. J. Nelson、Eric Jones、Robert Kern、Eric Larson、C J Carey、 ̇Ilhan Polat、Yu Feng、Eric W. Moore、Jake VanderPlas、Denis Laxalde、Josef Perktold、Robert Cimrman、Ian Henriksen、E. A.

Quintero、Charles R. Harris、Anne M. Archibald、Antˆonio H. Ribeiro、Fabian Pedregosa、Paul van Mulbregt 和 SciPy 1.0 Contributors。2020 年。SciPy 1.0：Python 科学计算的基础算法。*NatureMethods*17 （ 2020 ）， 261–272。 https://doi.org/10.1038/s41592-019-0686-2