交错双重差分：处理效应异质性与估计方法选择

刘　冲　沙学康　张　妍*

摘要：双重差分法是社会科学中进行因果推断和政策评估时最广泛采用的研究手段。然而，近年来不断涌现的前沿文献发现，对于交错双重差分的情形，因存在处理效应异质性，采用传统双向固定效应模型可能会造成严重的估计偏误。为此，理论计量领域诞生了多种异质性—稳健的估计方法，但这也让应用者在实践中对如何选取合适的估计方法、如何验证前提假设产生困惑。本文阐释了处理效应异质性导致潜在偏误的根源，总结了三类异质性—稳健估计方法的经济学直觉。本文对比了这些方法的核心假设、应用场景和估计量性质，通过模拟数据检验了估计效果，并对验证“平行性趋势”假设进行了深入讨论。最后，针对国内当前的使用现状，本文结合应用案例和现有综述文章，为应用研究者提供了操作建议。

关键词：交错双重差分　处理效应异质性　异质性—稳健估计量　应用场景

中图分类号：F064.1　**文献标识码：**A　**文章编号：**1000-3894(2022)09-0177-28

DOI:10.13653/j.cnki.jqte.20220805.001

引　言

双重差分法（Difference-in-Differences，简称 DID）是社会科学研究中使用最为广泛的一种研究手段，其清晰直观、易于操作，因而深受政策效应评估者喜爱（Baker 等，2022；Roth 等，2022；黄炜等，2022）。在我国，周黎安和陈烨（2005）最早引入了双重差分法。他们利用我国农村税费改革先后在安徽、江苏、湖南、湖北等区域的试点，最后推向全国的“准实验”，检验了农村税费改革的政策效果。此后，国内利用双重差分方法，尤其是“先行先试”政策特点进行“渐进式”双重差分的研究设计如雨后春笋般涌现，这也深刻地改变了国内经验研究的图景。

多数情况下，经验研究者并没有直接参与到政策的设计和实施过程，无法按照科学评估方法的要求来构造和生产数据，因此利用观测数据的政策评估就必须非常小心和科学论证，针对不同的数据结构和政策类型采用不同的评估方法，并反复检验该方法的适用性（范子英，2018）。就双重差分方法而言，根据政策实施时点（或处理时点）的不同，一般分为单时点和多时点两种情况。如图 1 所示，类型一为单时点 DID，标注黑色的个体表示接受处理（Treatment），其特点是政策在同一时间实施。类型二和三展示了多时点 DID 的情形，其特

* 刘冲（通讯作者），长聘副教授，北京大学经济学院，电子邮箱：pkuliuchong@pku.edu.cn；沙学康，博士研究生，北京大学经济学院，电子邮箱：1901110879@pku.edu.cn；张妍，博士研究生，北京大学光华管理学院，电子邮箱：yan-zhang@pku.edu.cn。本文是国家社会科学基金重大项目（21&ZD097）的阶段性成果，并得到北京大学经济学院种子基金的资助。作者感谢匿名审稿人和编辑部的宝贵意见，文责自负。

点是政策发生于不同时间点，类型二在文献中通常被称作交错 DID 或者渐进 DID，类型三相较于二的区别是存在政策处理停止的情形（也称为退出情形）。在本文中，为了和理论文献统一，我们使用交错 DID（Staggered DID）作为讨论基准，并就退出情形进行额外说明。

在推行政策前由于无法准确预判经济影响，许多国家或部门都愿意推行“先行先试”的政策，通过挑选一些区域或行业做政策试点来提升政策容错空间。试点类政策的推行为交错 DID 方法带来了丰富的应用场景，近年来使用该方法的文章数量很多且呈现上升趋势（Baker 等，2022；王鹏超和韩立彬，2022）。以 2000～2019 年的金融和会计顶级期刊为例，使用 DID 方法的文章中有 54.7% 采用了交错 DID 方法（Baker 等，2022）；在 2017～2021 年的 Top 5 期刊上[①]，交错 DID 方法占全部 DID 文章的 33%，而在 2017～2021 年中文权威期刊上[②]，交错 DID 方法占全部 DID 文章的 53%。

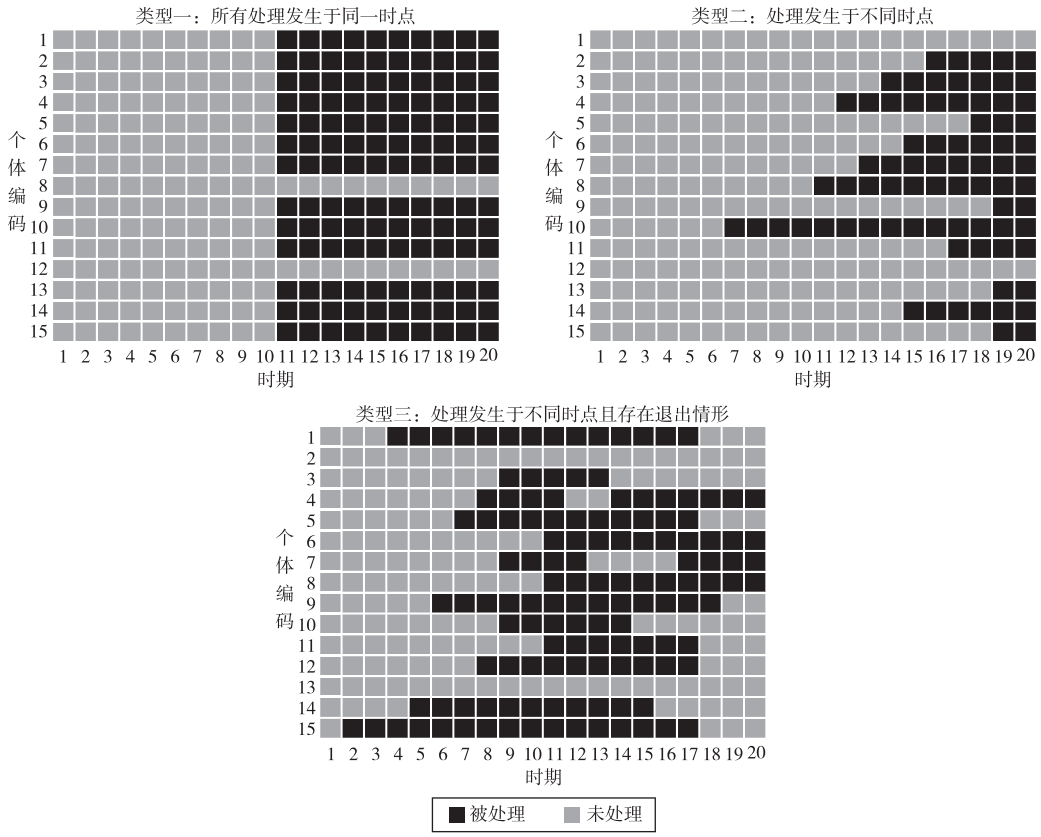

图 1　双重差分方法的三种常见类型

然而，随着交错双重差分在经济学各个领域的普遍使用（陈林和伍海军，2015；石华军和楚尔鸣，2017；胡日东和林明裕，2018），包括 Borusyak 和 Jaravel（2017）、de Chaisemartin 和 D'Haultfœuille（2020a）、Goodman-Bacon（2021）在内的很多研究者都关注到，交错 DID 有

① 即 Econometrica、American Economic Review、Journal of Political Economy、Quarterly Journal of Economics 和 Review of Economic Studies。

② 我们统计了《经济研究》《管理世界》《世界经济》《经济学（季刊）》《中国工业经济》《金融研究》这几种权威期刊。

一个重要的潜在问题是存在异质性处理效应（Heterogeneous Treatment Effects），即同一处理对于不同个体产生的效果存在差异，这种差异可能表现在接受处理后的时长或者不同时点接受处理的组别两个维度。在此背景下，如果利用传统双向固定效应模型估计量（Two-Way Fixed Effect Estimator，以下简称 TWFE）进行估计，无论是静态还是动态估计结果，都存在潜在偏误。一方面，在静态情形下，如 Goodman-Bacon（2021）所指出的，即使“平行性趋势”假设满足，TWFE 也会存在“坏的控制组”问题，即由于处理时点的差异，较早接受处理的样本会成为较晚处理样本的控制组，从而可能带来估计偏误。另一方面，在动态情形下，TWFE 不仅存在“坏的控制组”问题，其每一期的估计系数还会受到跨期交叉污染而变得难以解释，甚至还会面临平行性趋势检验失效的风险（Sun 和 Abraham，2021）。为了修正交错 DID 研究设计中 TWFE 估计量的潜在偏误，大量理论计量工作不断涌现，并提出了多种“异质性—稳健”估计量（Heterogeneity-Robust Estimator）。我们从中选取了文献中重点关注的三类解决方案进行总结对比，发现尽管刚刚过去几年时间，这一系列研究已被广泛引用，其中 Goodman-Bacon（2021）、de Chaisemartin 和 D'Haultfœuille（2020a）和 Callaway 和 Sant'Anna（2021）三篇研究的引用量甚至已经超过了 1000 次，表明学术界对交错 DID 传统估计方法存在的问题和解决方案高度关注①。

理论的飞速进展也给应用研究者带来诸多困扰，即如何根据应用场景选取合适的估计方法、如何理解和验证相应的前提假设、这些估计量的性质有何差异等。鉴于此，本文系统梳理和总结了三类解决方案，试图对多种“异质性—稳健”估计量在应用场景、前提假设和统计量的性质等方面进行详细阐释和对比。总的来说，交错 DID 需要解决的核心问题是处理效应异质性，这种异质性体现在接受处理后的时长以及不同时点接受处理的组别这两个维度。针对这一问题，文献中有三类解决思路，一是计算组别 - 时期平均处理效应并进行加权平均。其核心是避免使用已处理的个体作为“坏的控制组”，只选取“好的控制组”计算组别 - 时期平均处理效应，再通过组别、时期两个维度进行加权平均得到平均处理效应。二是利用插补的方法插补出合理的反事实结果。其核心是从控制组样本中估计出每个处理组个体每个时期的反事实结果变量，从而避免“坏的控制组”问题的出现。三是利用堆叠回归的方式。其核心是给每个处理组选择“好的控制组”组成数据集，按照相对事件时间（而不是日历时间）堆叠数据集并进行回归估计。

三类解决方案各有千秋，核心思想都在于寻找一个合理控制组或者利用控制组计算出合理的反事实结果变量。然而，这些解决方案对于函数形式设定、是否依赖“从未接受处理组”的存在、进入估计的数据量等方面存在差异，因此研究者需要根据数据特点和政策实施情况进行合理选择。本文梳理了多种方法的适用场景和前提假设的特点，结合文献的使用现状，给出相应的操作建议。本文还特别指出了这些研究方法的优势和存在的局限，以期引发更广泛的讨论，避免方法误用，希望能够为我国经济学研究与国际前沿接轨提供一些有益的帮助。

相比已有综述类文献，本文的边际贡献可能体现在以下方面：第一，本文详尽梳理了交错 DID 的前沿文献，阐释并对比了三类解决处理效应异质性问题的思路、核心假设和估计量性质。特别在适用场景上，本文给应用研究者提供了详细的操作建议，有助于研究者更加科学地进行估计方法选择。第二，本文结合多种数值模拟方式，从多个维度探讨了三类解决

① 更详尽的统计展示在了《数量经济技术经济研究》杂志网站的论文附录五图 D.1 中，详见《数量经济技术经济研究》杂志网站的论文附录。

方案之间的差异。这将有助于应用者更全面地了解三类估计量的特点，以便在后续研究中进行合理选择。本文希望达到两个目的：一是引发应用研究者对交错 DID 产生的“异质性 – 稳健估计量”科学使用的探讨；二是让研究者和读者都能够拨开数据和方法的“迷雾”，更好地洞察真实的因果关系，也更合理地检验前提假设，进而提高国内经济学经验研究的可信性和规范性。

一、经典双重差分法回顾

我们首先回顾经典双重差分方法的基本原理和重要假设。最经典的双重差分仅包含两个组（处理组和控制组）和两个时期（政策处理前和处理后），即所谓的 2 ×2 – DID。图 2 左侧展示了 2 ×2 – DID 的基本原理，在这个例子中，研究者通常会采集政策处理前和后的两期的数据，计算出两个组在政策前后的结果变量的变化，再计算出处理组相较于控制组的相对变化，得到对政策效果的评估。

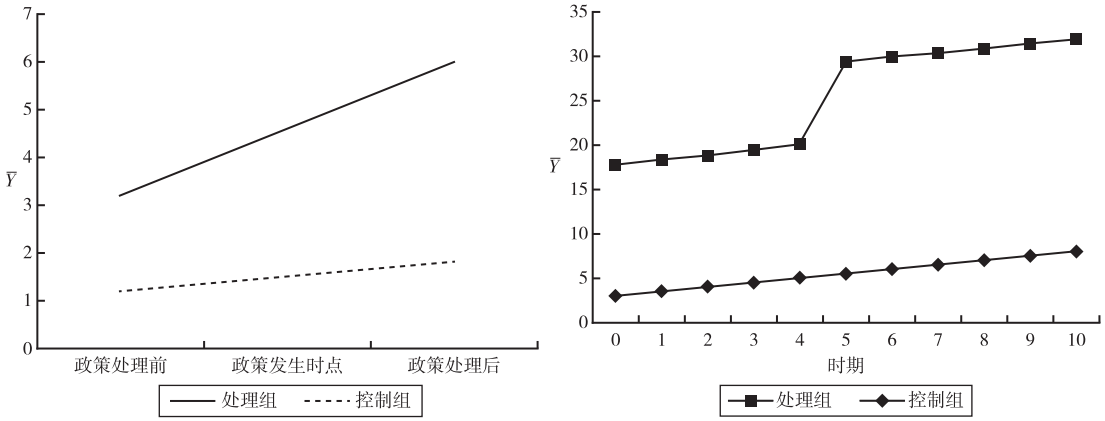

图 2　2 ×2 – DID 和多期 DID 示意图

上述思想可以通过线性回归的方式实现，基于 2 ×2 – DID 的设定，研究者通常采用下面的模型估计：

$$y_{it} = \sigma + \mu D_i + \rho T_t + \delta D_i \times T_t + \varepsilon_{it} \tag{1}$$

其中 i 表示个体，t 表示时期。y_{it} 是结果变量，D_i 是政策处理的分组虚拟变量（$D_i = 1$ 表示处理组，$D_i = 0$ 表示控制组），T_t 是政策处理的时间虚拟变量（$T_t = 1$ 表示政策处理后，$T_t = 0$ 表示政策处理前），ε_{it} 是随机误差项，δ 为研究者关心的平均处理效应。在实际应用中，样本数据会含有多个时期，此时双重差分法往往与面板数据联系起来（如图 2 右侧所示，称为多期 DID），研究者通常会使用双向固定效应模型进行估计：

$$y_{it} = \sigma + \alpha_i + \lambda_t + \delta D_i \times T_t + \varepsilon_{it} \tag{2}$$

其中 α_i 表示个体固定效应，λ_t 表示时间固定效应。在诸多研究设计中，政策的推行往往不在同一时点进行，而是随着时间的推移在样本内不断推开，例如我国的经济开发区就是在全国不同省份、不同年份逐步建成的。此时，多期 DID 就拓展成为交错 DID。过往大多数文献会使用双向固定效应模型估计，一般是通过重新定义虚拟变量 D_{it}，代替（2）式中 $D_i \times T_t$ 这一交互项得到：

$$y_{it} = \sigma + \alpha_i + \lambda_t + \beta_{fe} D_{it} + \varepsilon_{it} \quad (3)$$

回归（3）式中，我们通常关注的是 TWFE：$\hat{\beta}_{fe}$，将它视为平均处理效应的一致估计量。

值得注意的是，无论是经典的多期 DID 还是交错 DID，使用双向固定效应模型进行因果推断时，需要满足以下四个重要假设：

第一，严格外生假设（Strict Exogeneity Assumption）。该假设是使用面板数据进行因果推断的关键假设，它要求：不能存在随时间变化的混杂因素；过去的结果变量不能对当期的结果变量产生影响；过去的结果变量或协变量不能对当期和未来的处理状态产生影响；当期的处理状态不能对未来的结果变量产生影响。数学上，严格外生假设强于平行趋势假设；而实际研究中两者差别不大（Xu，2022）。

第二，无预期效应假设（No Anticipation Assumption）。该假设指的是个体在当期的结果变量不会受到个体在未来的接受政策处理状态的影响。也即个体并不能预知其在未来是否会接受政策处理，从而根据这种预期改变其行为。

第三，单位处理变量值稳定假设（Stable Unit Treatment Values Assumption）。其是指不同个体是否受到政策冲击是相互独立的，某一个体受政策冲击的情况（Treatment Status）不影响任何其他个体的结果（黄炜等，2022）①。

第四，处理效应同质性假设（Homogeneous Treatment Effect Assumption）。它要求处理效应满足两个维度的同质性：第一，处理效应在不同的组别间是同质的，即同一政策对于不同处理组的影响是相同的。第二，处理效应在时间维度上是同质的，即对于同一时间受到政策处理的所有个体，随着时间推移，处理效应的大小不变。需要特别指出的是，在过往的研究中，使用 TWFE 进行交错 DID 估计时，往往忽视了这一重要的隐含假设。以 Goodman-Bacon（2021）等为代表的一众学者指出，忽视这一假设将可能产生较为严重的估计偏误。

二、交错双重差分使用双向固定效应估计量的潜在问题

为了清晰地展示交错 DID 情境下处理效应同质性假设的重要意义，我们首先通过一个简单的例子进行说明。如表 1 所示，我们假设研究者观察到 3 个个体（A，B 和 C），每个个体有三期数据（0，1 和 2）。α_i 是指示个体的虚拟变量（$i=$A，B，C），λ_t 是指示时期的虚拟变量（$t=0$，1，2）。D 是表示处理状态的虚拟变量，假设个体 A 在第 2 期接受处理，而个体 B 和个体 C 在第 1 期就接受处理。

为了阐释方便，我们首先对所有接受处理的个体给定一个反事实结果 Y（0），即如果该个体未受处理的潜在结果，而个体观测到的真实结果记为 Y；Y 和 Y（0）的差值就是每个个体的处理效应 Δ。我们假定有两种情形：Δ^1 假设处理效应是同质的，即处理效应在各个体之间以及在不同时期是一个相同的常数；Δ^2 则假设处理效应是异质的，即个体 A、B 和 C 的处理效应大小不同并且个体 B 和 C 各自的处理效应在不同时期也是不同的。在实证研究中，研究者由于无法观测到接受处理个体的反事实结果，因而需要经过一定假设来构造出反事实。而与实际研究的逻辑不同，在此案例中我们首先给出了反事实结果，以便直接计算出个体处理效应，进而比较双向固定效应估计量距离真实处理效应的偏差。

① 该假设涉及研究中对于政策溢出效应的认识，以 Butts（2021）、Huber 和 Steinmayr（2021）等为代表的一众学者提出了新的估计量可以在放松该假设下得到一致的估计。

表 1　　处理效应异质性问题示例

ID	T	α_B	α_C	λ_1	λ_2	D	Y^1	Y^2	Y（0）	Δ^1	Δ^2
A	0	0	0	0	0	0	0	0	—	—	—
A	1	0	0	1	0	0	1	1	—	—	—
A	2	0	0	0	1	1	2	2	1	1	1
B	0	1	0	0	0	0	1	1	—	—	—
B	1	1	0	1	0	1	3	2	2	1	0
B	2	1	0	0	1	1	3	7	2	1	5
C	0	0	1	0	0	0	2	2	—	—	—
C	1	0	1	1	0	1	4	3	3	1	0
C	2	0	1	0	1	1	4	11	3	1	8

针对表 1 中的示例数据，研究者一般会使用下面的双向固定效应模型进行估计：

$$Y_{it} = \beta_0 + \beta_{fe} D_{it} + \beta_{\alpha_B}\alpha_B + \beta_{\alpha_C}\alpha_C + \beta_{\lambda_1}\lambda_1 + \beta_{\lambda_2}\lambda_2 + \varepsilon_{it}$$

为了说明问题所在，我们将双向固定效应模型的估计过程打开，即分别看个体 A、B 和 C 的平均处理效应。对于个体 A，由于此例中个体 B 和 C 均在第 1 期接受处理且两者在数据中出现的次数相同（均有三期观测），因此个体 B 和 C 分别以$\frac{1}{2}$的权重加总构成个体 A 的控制组，从而 A 的平均处理效应是 $DID_A = [E(Y_{A,2}) - E(Y_{A,1})] - \left\{\frac{1}{2}[E(Y_{B,2}) - E(Y_{B,1})] + \frac{1}{2}[E(Y_{C,2}) - E(Y_{C,1})]\right\}$，结合表 1 中已给出的处理效应$\Delta_{i,t}$可以计算出 $DID_A = \Delta_{A,2} - \frac{1}{2}(\Delta_{B,2} - \Delta_{B,1}) - \frac{1}{2}(\Delta_{C,2} - \Delta_{C,1})$。以此类推，个体 B 的平均处理效应可以表示为 $DID_B = [E(Y_{B,1}) - E(Y_{B,0})] - [E(Y_{A,1}) - E(Y_{A,0})] = \Delta_{B,1}$；个体 C 的平均处理效应可以表示为 $DID_C = [E(Y_{C,1}) - E(Y_{C,0})] - [E(Y_{A,1}) - E(Y_{A,0})] = \Delta_{C,1}$。

若处理效应是同质的（即表 1 中第一种情形，$\Delta^1_{B,2} = \Delta^1_{B,1}$，$\Delta^1_{C,2} = \Delta^1_{C,1}$），那么 $DID_A = \Delta_{A,2}$；若处理效应是异质性的（即表 1 中第二种情形，$\Delta^2_{B,2} \neq \Delta^2_{B,1}$，$\Delta^2_{C,2} \neq \Delta^2_{C,1}$），则 $DID_A \neq \Delta_{A,2}$。进一步地，β_{fe}是三个平均处理效应的加权平均，因此我们可以将其表示为：

$$\beta_{fe} = \frac{1}{3}(DID_A + DID_B + DID_C) = \frac{1}{3}\Delta_{A,2} + \frac{1}{2}\Delta_{B,1} + \frac{1}{2}\Delta_{C,1} + \left(\frac{-1}{6}\right)\Delta_{B,2} + \left(\frac{-1}{6}\right)\Delta_{C,2}$$

我们可以结合具体数值来计算处理效应同质性和异质性两种情况下的结果。在处理效应同质性假定（第一种情形）之下，$\beta_{fe} = 1$；而在处理效应异质性情况（第二种情形）下，$\beta_{fe} = \frac{-11}{6}$。在第一种情形下由于不存在处理效应异质性，因此$\beta_{fe}$与平均处理效应在数值上是相等的，即不存在偏误。而在第二种情形下，由于存在处理效应异质性，β_{fe}与平均处理效应在数值上是不相等的。更为明显的是，当异质性问题较为严重时（例如不同时点接受处理的个体之间处理效应差异较大），尽管所有 $\Delta_{i,t}$都是正的，但由于负权重

的存在，β_{fe}为负值，与平均处理效应的符号完全相反。

1. TWFE 在估计静态模型时的潜在问题

在实证研究中，对于交错 DID，过往的工作通常采用双向固定效应模型估计，Goodman-Bacon（2021）从理论上指出，当处理效应存在异质性时，同一处理对于不同个体产生的效果存在差异，将会导致 TWFE 产生潜在的估计偏误。

接下来，我们对同质性处理效应假设违背时产生的问题进行阐释。在图 3 交错 DID 示意图中，假设存在 k，l，u 三组个体，它们在接受处理时点上存在差异。其中 k 组个体最早接受政策处理（第 5 期），l 组个体相对较晚（第 9 期），u 组个体是从未接受处理组（Never-Treated Group①）。传统的单时点 DID 中仅存在 k 组或 l 组中的一个，此时使用双向固定效应模型估计量（TWFE）不存在偏误。对于交错 DID 的情形而言，Goodman-Bacon（2021）指出，个体接受处理的多时点特征使得样本中产生了多个组别，TWFE 是不同的 2×2 – DID 组合估计系数的加权平均。如图 4 所示，$\hat{\beta}_{fe}$可分解为四种类似 2×2 – DID 估计系数$\hat{\beta}_{ku}$，$\hat{\beta}_{lu}$，$\hat{\beta}_{kl}$，$\hat{\beta}_{lk}$②。我们发现，交错 DID 相较于传统 2×2 – DID 的关键不同在于，较早受处理的个体会作为较晚受处理个体的控制组进入估计中，如 $\hat{\beta}_{lk}$。Goodman-Bacon（2021）指出，较早接受处理组并不是很好的控制组（也称作“坏的控制组”），因为相较于较晚接受处理组或从未接受处理组（也称作“好的控制组”），它们的事前趋势已经发生了变化。而正是由于这种“坏的控制组”的存在，才导致了 TWFE 在进行交错 DID 估计时，会产生潜在偏误。

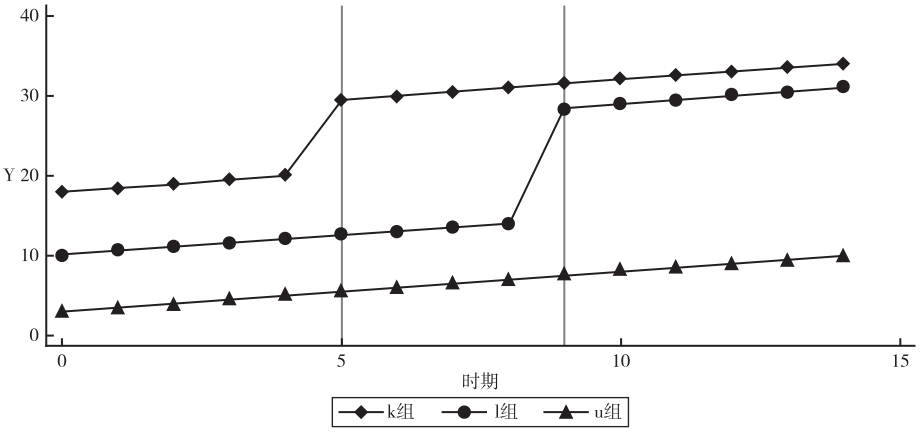

图 3　交错 DID 示意图

Goodman-Bacon（2021）将图 4 展示的分解逻辑进行了一般化分析，从理论上指出了 TWFE 的潜在偏误来源。我们假定样本有 T 期，在截面上包含 N 个个体，令 e 表示个体接受处理的时间。根据处理时间可将其分为 $e=1$，…，E 个处理组，另有一个“从未接受处理组”U 包含所有样本期内均未接受处理的个体。使用双向固定效应模型（3）估计出的回归

① 从未接受处理组（Never-Treated Group）在这一领域的文献中是一个重要概念，该组个体在样本期内从未接受政策处理，因此仅作为控制组出现在研究设计之中。

② 其中$\hat{\beta}_{ku}$是以 k 组为处理组，u 组为控制组使用双向固定效应模型进行 2×2 – DID 估计得到的系数，以此类推。需要指出的是，Goodman-Bacon 分解中提到的 2×2 – DID 与第二节提到的 2×2 – DID 略有差异。经典双重差分法语境下的 2×2 – DID 仅包含两个时期，但 Goodman-Bacon 分解中的 2×2 – DID 可能不仅仅包含两个时期。

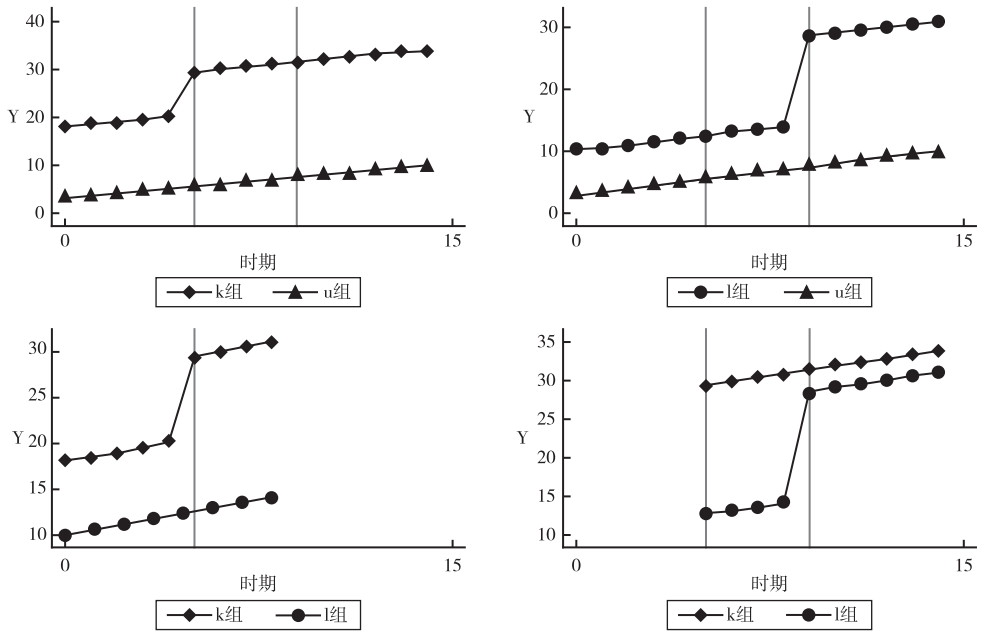

图4　交错 DID 的分解示意图

系数可以表示为：

$$\hat{\beta}_{fe} = \sum_{e \neq U} w_{eU} \hat{\beta}_{eU}^{2\times2} + \sum_{e \neq U} \sum_{l > e} \left[w_{el}^{e} \hat{\beta}_{el}^{2\times2,e} + w_{el}^{l} \hat{\beta}_{el}^{2\times2,l} \right] \tag{4}$$

其中 w_{eU}，w_{el}^{e}，w_{el}^{l} 分别为使用不同处理组和控制组进行比较时的权重，具体可以表示为子样本的样本量 n 与虚拟变量 D_{it} 方差的函数①。$\hat{\beta}_{eU}^{2\times2}$ 是以从未接受处理组作为控制组进行 2×2－DID 估计的系数，$\hat{\beta}_{el}^{2\times2,e}$ 是以较晚接受处理组（类似图 4 中的 l 组）为控制组进行 2×2－DID 估计的系数，$\hat{\beta}_{el}^{2\times2,l}$ 是以较早接受处理组（类似图 4 中的 k 组）为控制组进行 2×2－DID 估计的系数。上述分解过程表明：当存在多个处理时点时，TWFE 是所有可能的 2×2－DID 估计系数的加权平均。可能出现问题的地方在于，这些 2×2－DID 之中包含将较早接受处理的样本用作控制组与处理时间较晚的样本对比的结果。如果处理效应存在异质性，那么其系数 $\hat{\beta}_{el}^{2\times2,l}$ 不再是无偏估计。

在大样本情形下，Goodman-Bacon（2021）证明了（4）式的分解结果可以写为：

$$plim_{N\to\infty} \hat{\beta}_{fe} = VWATT + VWCT - \Delta ATT \tag{5}$$

① $w_{eU} = \dfrac{(n_e + n_U)^2 n_{eU} (1 - n_{eU}) \bar{D}_e (1 - \bar{D}_e)}{\hat{V}^D}$，$w_{el}^{e} = \dfrac{((n_e + n_l)(1 - \bar{D}_l))^2 n_{el} (1 - n_{el}) \dfrac{\bar{D}_e - \bar{D}_l}{1 - \bar{D}_l} \dfrac{1 - \bar{D}_e}{1 - \bar{D}_l}}{\hat{V}^D}$，$w_{el}^{l} = \dfrac{((n_e + n_l) \bar{D}_e)^2 n_{el} (1 - n_{el}) \dfrac{\bar{D}_l}{\bar{D}_e} \dfrac{\bar{D}_e - \bar{D}_l}{\bar{D}_e}}{\hat{V}^D}$，并且 $\sum_{e \neq U} w_{eU} + \sum_{e \neq U} \sum_{l > e} \left[w_{el}^{e} + w_{el}^{l} \right] = 1$。这里我们可以看出权重 w_{eU}，w_{el}^{e}，w_{el}^{l} 由两部分组成：第一部分是相对应的 2×2－DID 子样本的样本容量（如 $n_e + n_U$）；第二部分则是衡量相对应的 2×2－DID 子样本中个体接受政策处理状态的离散程度的方差。

（5）式表明估计系数包含三个部分：第一部分 *VWATT*，即“方差加权平均处理效应”（Variance-Weighted ATT），是正权重加权的 ATT。第二部分 *VWCT*，即“方差加权共同趋势”（Variance-Weighted Common Trend），是由所有可能的 2×2－DID 的处理组和控制组之间的平行趋势加权平均得到的。第三部分 *ΔATT* 代表“加权的处理效应变化”，这一项在使用较早接受处理组作为控制组时会出现。Goodman-Bacon（2021）认为，即使平行趋势假设得到满足（即 VWCT 为 0），TWFE 仍可能是有偏且不一致的。其原因在于：第一，若处理效应在个体间是异质性的，那么 *ΔATT* 为 0，但此时 VWATT（方差加权 ATT）与平均处理效应（样本加权 ATT）存在差异；第二，若处理效应在时间维度上是异质的，那么 *ΔATT* 不等于 0。此时尽管 VWATT 与平均处理效应相同，但由于 *ΔATT* 的存在，TWFE 仍然是有偏且不一致的；第三，若处理效应在个体和时间两个维度上均存在异质性，那么 VWATT 不等于平均处理效应和 *ΔATT* 不等于 0 这两个问题均存在，都可能导致 TWFE 有偏且不一致。VWATT含有个体异质性，存在个体异质性VWATT不等于ATT；ΔATT含有时间异质性，存在时间异质性，则其不等于0

2. TWFE 在估计动态模型时的潜在问题

在一些文献中，研究者会采用（6）式的动态双向固定效应模型（Dynamic TWFE）进行事件研究。常见的做法是，在样本期 $[\underline{T}, \bar{T}]$ 内加入一系列表示相对于接受处理时点时长的虚拟变量，采用如下设定回归：

$$y_{it} = \alpha_i + \lambda_t + \sum_{\substack{r \neq -1 \\ -\underline{T} \leqslant r \leqslant \bar{T}}} 1[R_{it} = r]\beta_r + \epsilon_{it} \tag{6}$$

其中，$1[R_{it}=r]$ 是一个指示变量，指示个体 i 观测时点 t 距离接受处理时点是否为 r 期，实际操作中，研究者一般会根据样本总时期长短对 r 进行部分归并（Bin）处理，并且为了避免多重共线性问题常会将“－1”期排除在外作为基期。β_r 则是我们感兴趣的估计系数，即动态平均处理效应（Dynamic Treatment Effect）。

对于这类动态模型的估计，Sun 和 Abraham（2021）指出，当存在处理效应异质性时，该模型估计系数可能会存在偏误并很难解释，原因在于：第一，类似于静态估计，β_r 自身在估计的过程中可能会由于使用了“坏的控制组”而产生偏误。① 第二，系数 β_r 可能会赋予一些不属于 r 期的样本以正权重，从而使得每一个 β_r 可能受到来自其他时期的系数 β_s（$s \neq r$）的“交叉污染”（Cross-Lag Contamination）。特别地，处理效应异质性还可能会导致平行趋势检验的方法面临失效。即便是所有时期在事实上都满足平行趋势，使用传统的动态双向固定效应估计量也可能得到不满足平行趋势的结论。正是存在上述潜在风险，当面临交错 DID 中的处理效应异质性问题时，采用 Dynamic TWFE 方法要十分谨慎。

三、三类异质性—稳健估计量

在交错 DID 的设定下，处理效应异质性是导致 TWFE 产生潜在偏误的重要原因。为此，理论界提出了“异质性－稳健”估计量，主要表现为三种解决思路。第一种思路是计算组别－时期平均处理效应（Cohort-Specific Average Treatment Effects on the Treated，CATT），再加权加总。例如 de Chaisemartin 和 D'Haultfœuille（2020a，2022a）、Callaway 和 Sant'Anna（2021）以及 Sun 和 Abraham（2021）等均是基于这一思想；第二种思路是使用插补估计量

① 例如本节例子中的“负权重”带来的偏误。

（Imputation Estimator）构造反事实结果进行估计。以 Borusyak 等（2021）、Liu 等（2022）和 Gardner（2021）等为代表的学者构造了新估计量；第三种思路是 Cengiz 等（2019）采用的堆叠回归估计量（Stacked Regression Estimator）。本文作为重要尝试，探讨各类方法的优势和局限，以方便研究者基于数据特点和研究设计进行恰当的方法选取。

1. 计算组别－时期平均处理效应

这一类异质性－稳健估计量的解决思路比较直接，即首先计算特定组别－特定时期的平均处理效应，再在组别、时期两个维度进行合理的加权加总。这种方法避免使用较早接受处理组（“坏的控制组”）作为控制组，从而直接避免估计偏误，目前在实证研究领域已经得到了广泛的应用。接下来，我们将具体介绍三种常见的估计量。

（1）de Chaisemartin 和 D'Haultfœuille 提出的估计量。de Chaisemartin 和 D'Haultfœuille（2020a，2022a）认为，双向固定效应模型的关键问题在于错误地将那些已经在当期之前接受过处理的样本作为控制组和当期接受处理的样本进行对比，造成了“禁止对比（Forbidden Comparison）”问题。因此一个解决方案是构造避免进行这种比较的参数，然后使用 DID 方法来估计它。他们给出的参数是 δ_s，对应的估计量为 DID_M。从直觉上说，参数 δ_s 仅考虑政策发生时点前后的观测，并将政策发生时点前后处理状态发生变化的个体视作处理组，比较这些处理组个体实际接受处理后的结果与其反事实结果，从而得到处理效应。由于处理组个体在未接受政策处理状态下的反事实结果不可得，DID_M 估计量的任务就是利用样本构造一个对应于 δ_s 的估计量。在思想上，DID_M 与传统的 DID 估计量非常相似，只不过其处理组限制为政策发生时点前后政策处理状态发生变化的个体，而控制组则限制为政策发生时点前后政策处理状态未发生变化的个体。

具体来说，我们令 i 表示个体，g 表示个体所属组别，t 表示时期。$Y_{i,g,t}(1)$ 表示组别 g 的个体 i 在时期 t 接受处理的产出，$Y_{i,g,t}(0)$ 则表示对应的反事实产出。$N_{g,t}$ 表示（g，t）单元内的样本数，$D_{g,t}$ 表示接受处理的样本比例。基于以上标记，δ_s 可以写为：

$$\delta_S = E\left[\frac{1}{N_{S(i,g,t):t\geq 2,D_{g,t}\neq D_{g,t-1}}}\sum [Y_{i,g,t}(1) - Y_{i,g,t}(0)]\right]$$

其中 $N_S = \sum_{(g,t):t\geq 2,D_{g,t}\neq D_{g,t-1}} N_{g,t}$，也即处理状态发生转换的样本数。在满足严格外生假设、平行趋势假设、组间独立假设和存在稳定组假设的前提下①，de Chaisemartin 和 D'Haultfœuille（2020a）指出估计量 DID_M 是 δ_S 的无偏估计量：

$$DID_M = \sum_{t=2}^{T}\left(\frac{N_{1,0,t}}{N_S}DID_{+,t} + \frac{N_{0,1,t}}{N_S}DID_{-,t}\right)$$②

其中 $N_{1,0,t}$ 指的是 t 期接受处理而 $t-1$ 期未接受处理的样本量，$N_{0,1,t}$ 则是指 t 期未接受处理而 $t-1$ 期接受处理的样本量。$DID_{+,t}$ 刻画的是处理状态从 0 到 1 的情形，而 $DID_{-,t}$ 刻画

① 为了节省篇幅，我们将本文涉及的主要方法的假设的详细描述放到了附录一，供读者查阅。需要指出的是，这里的严格外生假设与第二节提到的严格外生假设略有差异，详见附录一中对 de Chaisemartin 和 D'Haultfœuille（2020a）的严格外生假设的介绍，详见《数量经济技术经济研究》杂志网站的论文附录。

② 其中 $DID_{+,t} = \sum_{g:D_{g,t}=1,D_{g,t-1}=0}\frac{N_{g,t}}{N_{1,0,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=D_{g,t-1}=0}\frac{N_{g,t}}{N_{0,0,t}}(Y_{g,t} - Y_{g,t-1})$，$DID_{-,t} = \sum_{g:D_{g,t}=D_{g,t-1}=1}\frac{N_{g,t}}{N_{1,1,t}}(Y_{g,t} - Y_{g,t-1}) - \sum_{g:D_{g,t}=0,D_{g,t-1}=1}\frac{N_{g,t}}{N_{0,1,t}}(Y_{g,t} - Y_{g,t-1})$，$DID_M$ 可以同时考虑政策实施与退出的情况。

的是处理状态从 1 到 0 的退出情形。需要指出的是，DID_M 估计量也可被应用于处理变量为非二元虚拟变量的情形。

需要特别注意的是，DID_M 与传统的平均处理效应存在区别：DID_M 考察的是处理组和控制组在 $t-1$ 期和 t 期之间的差异，因此它估计出的是即时处理效应（Instantaneous Treatment Effect）。传统的平均处理效应实际上是将政策带来的即时处理效应和动态处理效应进行加权平均后得到的结果（Borusyak 和 Jaravel，2017；Roth 等，2022）①，因此 DID_M 并不是平均处理效应，应用者在解读时需要格外小心。为了计算动态处理效应，de Chaisemartin 和 D'Haultfœuille（2022a）拓展了原有的分析框架，首先给出对应于动态处理效应的参数 $\delta_{+,l}$，其次将这些动态处理效应参数进行加权平均获得参数 δ_+，该参数衡量的是政策处理发生一单位变化所带来的平均累积处理效应的变化②。最后构造与之相对应的估计量 $DID_{+,l}$ 和 $\hat{\delta}_+$，从而实现参数估计。

总的来说，在 de Chaisemartin 和 D'Haultfœuille 的框架中，DID_M 可以提供政策即时处理效应的无偏估计，$DID_{+,l}$ 提供了政策动态处理效应的无偏估计，而 δ_+ 则提供了政策平均处理效应的无偏估计。

（2）Sun 和 Abraham（2021）提出的估计量。Sun 和 Abraham（2021）针对处理效应异质性问题给出的 IW 估计量（Interaction-Weighted Estimator），尤其适用于事件研究。从直观上想，既然处理效应可能因为第一次接受处理的时点不同以及接受处理时长不同而具有异质性，那么对于同一时点接受处理并且接受处理后时长相同的样本处理效应则是一样的，因此作者定义了“组别 - 时期平均处理效应”，在估计时则选取“尚未接受处理”和“从未接受处理”的样本作为对照，而研究者关心的动态处理效应和平均处理效应就可以在 CATT 的基础上通过加权求和得到。

具体来说，他们根据样本第一次接受处理的时间差异定义了组别 - 时期平均处理效应 CATT（e，l），其中 e 表示个体第一次接受政策处理的时点，l 表示距离政策发生时点的期数。需要说明的是，传统上，使用双向固定效应模型估计的动态处理效应不仅包含 CATT（e，l），还包含来自其他期数的 CATT，从而对该期的估计造成了污染。为了解决这一问题，Sun 和 Abraham（2021）提出了基于以下回归模型的 IW 估计量：

$$y_{it} = \alpha_i + \lambda_t + \sum_e \sum_{l \neq -1} \delta_{e,l}(1\{E_i = e\} \cdot D_{it}^l) + \varepsilon_{it}$$

其中 D_{it}^l 是进行传统的事件研究时定义的表示处理状态的虚拟变量，$1\{E_i = e\}$ 是新加入的关键虚拟变量，指示个体是否属于第一次接受处理时间为 e 的组别。$\delta_{e,l}$ 即为 CATT（e，l）对应的估计量，以接受政策处理时长大于等于 l 的组别的占比为权重，再将估计出的 $\delta_{e,l}$ 进行加权平均即可获得平均处理效应的估计。

值得指出的是，使用 Sun 和 Abraham（2021）提出的 IW 估计量进行估计时，要求样本

① Borusyak 和 Jaravel（2017）指出平均处理效应 $\gamma = \sum_{k=0}^{\infty} \omega_k \gamma_k$，其中 k 表示距离接受政策处理的时期长度，γ_k 表示相对应的动态处理效应，ω_k 为权重。权重 ω_k 实际上是时间指示变量 $1\{K_{it} = k\}$ 对组别、时期固定效应以及表示是否接受政策处理的虚拟变量 D_{it} 回归时，变量 D_{it} 前的系数。

② de Chaisemartin 和 D'Haultfœuille（2022a）指出在交错 DID 的研究设计下，若样本中不存在纯处理组（Always-Treated Group），且至少存在一组从未接受处理组，那么 δ_+ 与平均处理效应是等价的。此外，δ_+ 还可以解释为一系列意向处理效应（Intention to Treat Effect）的加权平均。

中存在从未接受处理的个体，否则会存在识别不足（Underidentification）的问题①。如果样本中不存在这样的个体，那么需要将样本中最后接受处理的个体作为从未接受处理的个体来对待，同时删除最后接受处理的个体在处理之后的全部样本。

（3）Callaway 和 Sant'Anna（2021）提出的估计量。Callaway 和 Sant'Anna（2021）与 Sun 和 Abraham（2021）的思想有一致之处。他们提出的估计量同样需要先计算每个组别－时期内的平均处理效应即 ATT（e，t），随后对 ATT（e，t）进行加总从而获得平均处理效应的估计，而标准误则通过 Bootstrap 的方式获取。具体来说，Callaway 和 Sant'Anna（2021）定义指示变量 E_e，即当个体第一次接受处理的时间点是 e 时该变量取 1。同时定义虚拟变量 C，如果个体从未接受处理则该变量取值为 1。对于样本中的每一个个体，在虚拟变量集合 $\{E_1, \ldots, E_T, C\}$ 中必有一个虚拟变量取值为 1。我们令 $p_e(X) = P(E_e = 1 \mid X, E_e + C = 1)$ 表示在条件于协变量 X 和受处理状态下，个体在时点 e 接受处理的概率。基于以上设定，我们可以将 $ATT(e,t)$ 表示为：

$$ATT(e,t) = E\left[\left(\frac{E_e}{E[E_e]} - \frac{\frac{p_e(X)C}{1-p_e(X)}}{E\left[\frac{p_e(X)C}{1-p_e(X)}\right]}\right)(Y_t - Y_{e-1})\right]$$

该参数背后的直觉在于计算 $ATT(e,t)$ 仅考虑组别第一次接受处理时期为 e 的样本和控制组的样本，而忽略所有其他样本。同时给予控制组中的那些与处理组内出现更频繁的个体特征更相似的样本以更大的权重，反之则给予较小的权重。这种做法确保了处理组和控制组的个体在特征方面的平衡性。值得注意的是，控制组个体可以是从未接受处理的个体，也可以是在样本期 t 之前尚未接受处理的个体②。

Callaway 和 Sant'Anna（2021）与 Sun 和 Abraham（2021）的方法之间的主要区别在于：（1）两者计算组别－时期平均处理效应的方式不同，前者使用非参方法进行计算而后者则采用线性回归进行计算；（2）两者对于组别－时期平均处理效应的加权方式有所不同。值得一提的是，当样本中存在从未接受处理组时，Callaway 和 Sant'Anna（2021）提出的估计量 $ATT(e,t)$ 与 Sun 和 Abraham（2021）提出的估计量计算得到的 $CATT$ 在数值上是等价的③。两者的关键区别在于当不存在从未接受处理组时，Callaway 和 Sant'Anna（2021）提出的估计量可以使用尚未接受处理组（Not-Yet-Treated Group）作为控制组，而 IW 估计量则需要对样本进行删减并使用最后接受处理的样本作为控制组。

在得到 $ATT(e,t)$ 的基础上，将其进行加权平均即可获得平均处理效应 θ：

① Borusyak 和 Jaravel（2017）详细阐释了识别不足的问题，相关内容可见《数量经济技术经济研究》杂志网站的论文附录二。

② Callaway 和 Sant'Anna（2021）提到，某些情况下研究者可能怀疑“从未接受处理组”在特征上与接受处理的样本存在较大差别，出于这种顾虑研究者可以删去数据中所有从未接受处理的样本，并在估计时选择“尚未接受处理组”作为控制。此外需要指出的是，Callaway 和 Sant'Anna（2021）在原文中指出以从未接受处理组和以尚未接受处理组为控制组时计算 $ATT(e,t)$ 的公式有所不同，且当研究者采用结果变量回归（Outcome Regression，OR）、逆概率加权法（Inverse Probability Weighting，IPW）和双重稳健（Doubly Robust，DR）这三种不同的估计方法来估计 $ATT(e,t)$ 时相应的公式也存在差异。为了节约篇幅，本文仅列举了以从未接受处理组为控制组且使用逆概率加权法计算 $ATT(e,t)$ 的相应公式，其他情形下的公式请参见 Callaway 和 Sant'Anna（2021）原文中的阐述。

③ 事实上，当使用尚未接受处理组作为控制组且无控制变量时，de Chaisemartin 和 D'Haultfœuille（2020a，2022a）提出的 DID_M 和 $DID_{+,l}$ 与 Callaway 和 Sant'Anna（2021）提出的估计量估计出的短期和长期平均处理效应在数值上相等。

$$\theta = \sum_{e} \sum_{t=2}^{T} w(e,t) \cdot ATT(e,t)$$

其中 w（e，t）是用以加权平均的权重。Callaway 和 Sant'Anna（2021）指出可以使用接受处理时间为 e 的组别的占比 P（$E=e \mid E \leqslant T$）为权重，从而定义出平均处理效应对应的参数：

$$\theta_W^O = \frac{1}{\kappa} \sum_{e} \sum_{t=2}^{T} 1\{t \geqslant e\} ATT(e,t) P(E = e \mid E \leqslant T)$①

其中 $\kappa = \sum_e \sum_{t=2}^{T} 1\{t \geqslant e\} P(E=e \mid E \leqslant T)$。

（4）“组别 – 时期平均处理效应”系列方法的使用建议。计算组别 – 时期平均处理效应再加权加总，得到的“异质性 – 稳健”估计量已经在学术界得到了较为广泛的使用，一些经济学顶级期刊上的研究已经将这类方法应用在文章中。不过这类方法仍有几个问题未能充分解决：第一，计算组别 – 时期平均处理效应的过程中丢失了大量样本（Borusyak 等，2021），这可能会影响估计效率；第二，这类方法并非都能应对政策存在退出的情形。第三，部分方法会依赖“从未接受处理组”的存在。

对于这些问题，本文认为研究者在应用这类方法时有以下几个方面需要格外注意：第一，当样本量不充足时，这类方法的估计效率将会受到较大影响，因此要审慎解读估计结果，与此同时建议研究者汇报这类方法在估计时使用到的样本量；第二，当政策存在退出情形时，Sun 和 Abraham（2021）、Callaway 和 Sant'Anna（2021）提出的方法将不再适用，因此建议研究者采用 de Chaisemartin 和 D'Haultfœuille（2020a，2022a）提出的估计量进行估计。第三，当样本中不存在“从未接受处理的个体”时，建议谨慎采用 Sun 和 Abraham（2021）的方法。尽管研究者可以通过将样本中最后接受处理的个体作为从未接受处理的个体来对待，即删除最后接受处理的个体在处理之后的全部样本，但可能会带来样本筛选问题。

2. 插补估计量

除了计算组别 – 时期平均处理效应外，有一支文献提出使用插补估计量（Imputation Estimator）来解决交错 DID 的处理效应异质性问题。插补估计量的直觉是：首先，利用从未接受处理的样本或尚未接受处理的样本估计出每个处理组个体每个时期的反事实结果。此后，计算处理组个体的处理效应，即真实结果与反事实结果的差。最后，将个体层面的处理效应进行加总，即得到平均处理效应的估计。这类做法在给定模型设定的前提下，通过构造合理的反事实，同样避免了使用“坏的控制组”问题。事实上，这一类解决思想得到了很多学者的青睐，包括 Borusyak 等（2021）、Xu（2017）②，Liu 等（2022）、Gardner（2021）和 Wooldridge（2021）都给出了思想类似的“异质性 – 稳健”估计量。受篇幅所限，我们重点选取 Borusyak 等（2021）以及 Liu 等（2022）两篇较为全面的工作进

① Callaway 和 Sant'Anna（2021）指出这种加权方式避免了负权重问题，但其缺陷在于会给予接受政策处理时间更长的组别以更高的权重。Callaway 和 Sant'Anna（2021）根据处理效应异质性形式的不同提出了不同的加权方式：第一，处理效应的大小与接受政策处理的时长有关；第二，处理效应的大小与接受政策处理的时点有关；第三，计算政策的累积处理效应。

② Xu（2017）的研究虽未明确提出适用于解决双向固定效应模型中的处理效应异质性问题，但其研究中实际上使用了插补估计量来计算反事实结果作为控制组进行估计。其估计量构造的核心思想与 Borusyak 等（2021）等研究有相似之处。

行阐述。

（1）Borusyak 等（2021）提出的估计量。与前文介绍的估计量相似，该估计量同样依赖于平行趋势假设和无预期效应假设①。不过需要注意的是，插补估计量所依赖的平行趋势假设与前文提到的 de Chaisemartin 和 D'Haultfœuille（2020a，2022a）、Callaway 和 Sant'Anna（2021）、Sun 和 Abraham（2021）这几个估计量所依赖的平行趋势假设存在差异。de Chaisemartin 和 D'Haultfœuille（2020a，2022a）等提出的估计量所依赖的平行趋势假设建立在组别的层面，也即不同组别未接受处理时的因变量的变化趋势应当保持平行。而 Borusyak 等（2021）提出的插补估计量所依赖的平行趋势假设为：$E[Y_{it}(0)] = \alpha_i + \beta_t$，即该平行趋势假设是建立在个体的维度。尤其需要强调的是，Borusyak 等（2021）对平行趋势假设的形式有参数化的要求，即隐含了该估计方法依赖于模型的正确设定，因此这是一种更强的平行趋势假设。

在具体的估计方面，该方法是依据如下步骤计算：第一步，使用从未接受处理的样本或尚未接受处理的样本估计出处理组个体对应的反事实结果 $\hat{Y}_{it}(0)$。第二步，将因变量 Y_{it} 与 $\hat{Y}_{it}(0)$ 相减获得处理组个体 i 在时期 t 的处理效应估计量 $\hat{\tau}_{it}$。第三步，对 $\hat{\tau}_{it}$ 进行加权平均即可获得平均处理效应的估计量 $\hat{\tau} = \sum_{it \in \Omega_1} \omega_{it} \hat{\tau}_{it}$。其中 Ω_1 指的是样本中接受政策处理的观测（Treated Observations）。

Borusyak 等（2021）证明了该估计量在满足同方差假设的前提下是一种有效估计量，他们进一步指出，即使误差项存在自相关和异方差，这种插补估计量的效率仍然要高于 Callaway 和 Sant'Anna（2021）等提出的方法。不过 de Chaisemartin 和 D'Haultfœuille（2022b）指出，同方差假设在大多数情形下都难以满足，而一旦不能满足，Borusyak 等（2021）提出的估计量就未必更有效。更严重的是，相较于其他方法，一旦平行趋势不能得到满足，那么 Borusyak 等（2021）提出的估计量的偏误将会更大。

Borusyak 等（2021）还指出其提出的估计量并不适用于高频面板数据（如金融领域常用的股价日度数据）进行因果推断的研究设计。使用高频面板数据可能面临两个潜在的威胁：首先，若样本中个体数较少（即 N 较小），那么可能不存在足够的“好的控制组”用以进行插补估计；其次，高频面板数据时期较多（也即 T 较大），因此更容易面临潜在的各时期间的相互干扰等问题，从而违背无预期效应假设，进而导致估计偏误。基于此，在使用高频面板数据进行因果分析时应谨慎采用 Borusyak 等（2021）提出的方法。

（2）Liu 等（2022）提出的估计量②。Liu 等（2022）同样基于插补的思想来构建估计量，但不同于 Borusyak 等（2021）的做法，他们提供了更为丰富的三种插补估计量：固定效应反事实估计量（FEct），交互固定效应反事实估计量（IFEct）和矩阵补全估计量（Matrix Completion，MC）。其中，固定效应反事实估计量（FEct）与 Borusyak 等（2021）提出的估计量是等价的；而存在不可观测的随时间变化的混杂因素时，固定效应反事实估计量（FEct）将不再适用，交互固定效应反事实估计量（IFEct）和矩阵补全估计量（MC）则能处理这种情况。

① 受篇幅所限，有关具体假设的详细说明，见《数量经济技术经济研究》杂志网站的论文附录一。

② Liu 等（2022）包含三个估计量：他们给出了 FEct 估计量，并将其与 Gobillon 和 Magnac（2016）、Xu（2017）提出的 IFEct 估计量，以及 Kidziński 和 Hastie（2018）、Athey 等（2021）提出的 MC 估计量纳入到统一的框架中进行分析和比较。尽管有些估计量并不是他们提出的，为表述方便，我们在本文中统一称作 Liu 等（2022）提出的估计量。

这一框架本质上是通过假设一种函数形式 $Y_{it}(0) = f(X_{it}) + h(U_{it}) + \epsilon_{it}$（其中 X_{it} 是控制变量，U_{it} 是不可观测因素）对处理组的反事实进行插补。换言之，该假设意味着模型需要正确设定。在应用过程中，研究者则需要根据具体情况选择 $f(\cdot)$ 和 $h(\cdot)$ 的形式。

具体而言，作者给出了三种插补模型：（1）在双向固定效应模型的基础上进行插补，即模型设定为 $Y_{it}(0) = X'_{it}\beta + \alpha_i + \lambda_t + \epsilon_{it}$（且有 $\sum_{D_{it}=0}\alpha_i = \sum_{D_{it}=0}\lambda_t$）；（2）结合交互固定效应模型使用因子增强模型进行插补，即模型设定为 $Y_{it}(0) = X'_{it}\beta + \alpha_i + \lambda_t + \varphi'_i f_t + \epsilon_{it}$；（3）将反事实的插补估计问题视为补全一个有 $N \times T$ 缺失值的矩阵，即使用矩阵补全估计量（MC）进行插补。IFEct 和 MC 的主要区别在于分解矩阵时对奇异值进行正则化的方式，两种估计量的优劣也视具体情况而定：一般而言，只有少量时变混杂因素存在并且每个时变混杂因素都表现出相对强的信号时，IFEct 的表现优于 MC；而存在大量弱时变混杂因素时，MC 的表现更好。

（3）插补估计量的使用建议。相较于计算组别 - 时期的平均处理效应，插补估计量在计算过程中由于没有造成大量的样本丢弃，因而估计效率更高。不过也需要特别强调的是，无论是 Borusyak 等（2021）还是 Liu 等（2022）提出的插补方法，反事实结果的构造依赖于模型的正确设定，其估计量的有效性也依赖于选择正确的模型设定。

一般而言，Liu 等（2022）的方法适用于更为丰富多样的面板数据结构，且他们给出了针对强外生性假设等一系列重要假设的检验方式。因此，我们给出以下建议：当研究者使用的数据样本结构较为简单（例如只有简单的政策退出情形，不存在反复进入退出）且不存在不可观测的随时间变化的混杂因素时，使用 Borusyak 等（2021）或 Liu 等（2022）提出的 FEct 方法均可。当研究者使用的数据存在较为复杂的政策进入退出情形或存在不可观测的随时间变化的混杂因素的干扰，建议研究者采用 Liu 等（2022）提出的 IFEct 方法或 MC 方法。

3. 堆叠回归估计量

（1）Cengiz 等（2019）提出的估计量。Cengiz 等（2019）提出了一种堆叠回归估计量（Stacked Regression Estimator）。从直觉上看，这种方法为每一个处理组的观测都匹配了从未接受处理或尚未接受处理的观测，进而形成一个数据集，随后将这些数据集堆叠在一起，通过进一步加入组别 - 个体、组别 - 时间固定效应进行线性回归。与前两类方法相似，从本质上讲这种做法也是通过避免使用较早接受处理组作为控制组来解决处理效应异质性问题。

具体来说，首先为每一个处理组 m 匹配从未接受处理或尚未接受处理的样本作为控制组，以此形成一个数据集，再将这些数据集堆叠。定义组别 m 的固定效应并将该固定效应与个体固定效应和时期固定效应交乘。最后，使用下式进行回归即可估计平均处理效应。

$$y_{itm} = \alpha_{im} + \lambda_{tm} + \sum_k \delta_k 1[t - E_i = k] + \varepsilon_{itm}$$

其中 E_i 即个体 i 接受处理的时点，k 为时点 t 距离政策发生时点的时间间隔，$1[t - E_i = k]$ 即为传统的事件研究中的时间虚拟变量。

需要指出的是，堆叠回归估计量在应用中会出现数据重复或嵌套的问题，其原因在于部分样本可能在不同的子数据集中作为控制组被重复使用，因此在应用该方法时应格

外注意样本量的变化以及聚类问题，此外 Cengiz 等（2019）在其研究工作中尚未提及这种堆叠估计量所依赖的前提假设和统计量的性质，因此这些方面的内容有待研究者进一步挖掘。

（2）堆叠回归估计量的使用建议。堆叠回归估计量目前面临的主要问题有两个：第一，该方法提供的估计量的统计性质并没有给出，也未经过严格证明；第二，该估计量在估计的过程中可能会造成数据重复使用的问题。此外，现有堆叠回归估计量的软件包不够完善，只能计算动态效应中各期的系数，无法直接实现加权平均，目前研究者应用时大多手工堆叠数据后再进行回归，因此没有形成统一且规范的做法。基于以上问题，本文建议研究者谨慎采用堆叠估计量进行交错 DID 估计，当然，可以将其作为一种稳健性检验的方式。

基于以上探讨，本文对三类估计量的多个维度进行了总结和对比，具体内容总结为下面的表 2。为了方便读者的使用和操作，我们也结合一些文献，汇总了相关的软件命令，详见附录四[①]。

表 2　　不同估计量的对比

估计量	核心假设	统计量性质	估计动态处理效应	是否解决退出情形	优势	局限
计算组别－时期平均处理效应（CATT）加权平均						
de Chaisemartin 和 D'Haultfœuille（2020a，2020b，2022a）	（1）平行趋势假设 （2）严格外生假设 （3）处理组和控制组组间独立 （4）存在稳定组	无偏，但不满足有效性	是	是	估计不依赖于线性模型设定；可适用于处理变量非二元虚拟变量的情形；可适用于同时存在多个政策处理变量的情形（de Chaisemartin 和 D'Haultfœuille，2020b）	估计过程中会丢失较多样本，可能会造成估计无效
Sun 和 Abraham（2021）	（1）平行趋势假设 （2）无预期效应假设	满足无偏性、一致性和有效性	是	否	使用线性回归估计，估计速度较快	估计过程中会丢失较多样本，可能会造成估计无效；依赖从未接受处理组的存在，若不存在则需对样本删减

① 详见《数量经济技术经济研究》杂志网站的论文附录。

（续）

估计量	核心假设	统计量性质	估计动态处理效应	是否解决退出情形	优势	局限
计算组别 – 时期平均处理效应（CATT）加权平均						
Callaway 和 Sant'Anna（2021）	（1）政策处理不可逆 （2）随机抽样 （3）政策效应有限预期 （4）平行趋势假设 （5）重叠性假设	满足无偏性、一致性	是	否	估计不依赖于线性模型设定；除计算平均处理效应、动态处理效应外，还提供了多种加权方式来测算政策的累积效应	估计过程中会丢失较多样本，可能会造成估计无效
插补估计量						
Borusyak 等（2021）	（1）平行趋势假设（隐含了函数形式的设定） （2）无预期效应假设 （3）残差同方差且无自相关或至少要知道异方差的形式	满足一致性和有效性	是	是	使用更充足的样本进行估计，估计更有效率	依赖于正确的模型设定；应用于高频面板数据时应保持谨慎；存在不可观测的随时间变化的混杂因素时，谨慎使用该方法
Liu 等（2022）	（1）函数形式； （2）严格外生性假设（同时隐含了“平行趋势假设”与“无预期效应假设”）	满足无偏性、一致性	是	是	使用更充足的样本进行估计，估计更有效率；能够较好处理时间序列横截面（TSCS）数据或长面板（Long-Panel）形式的数据，并且能够处理数据中包含随时间变化的协变量的情形	依赖于正确的模型设定；IFEct 和 MC 方法的调参过程较为主观，缺乏一般性标准
堆叠回归估计量						
Cengiz 等（2019）	（1）平行趋势假设 （2）无预期效应假设	–	是	–	–	统计量性质不明确；存在数据重复使用的问题

四、数值模拟示例

这一节将通过数值模拟的方式，首先向研究者展示如何呈现处理效应异质性问题的严重性；接下来，我们计算了三类“异质性－稳健”估计量并与 TWFE、真值进行对比，就各方法的估计效果进行了全面比较；最后，我们特别强调了部分方法在平行趋势检验时的注意要点。

我们参考 Borusyak（2021）[①] 的数值模拟方法，假设个体数量为 500（$i \in [1, 500]$），每个个体有 8 期（$t \in [1, 8]$），形成了一个 4000 样本量的平衡面板数据。首先，我们按照随机的均匀分布为这些个体分配一个处理时间，并保留一些“从未接受处理”个体，即生成一个在 2～9 之间的随机数，处理时间为 9 的个体则标记为“从未接受处理组”；其次，根据处理状态，随机为个体生成处理效应，这里我们假定处理效应表现为随时间增长的形式[②]，即 $\tau_{it} = (t - 4.5) \times 1\{D_{it} = 1\}$，其他更多种不同形式的异质性详见附录三[③]（图 A.1～图 A.3）；此后，我们生成一个服从标准正态分布的随机扰动项 ϵ_{it}（$\epsilon_{it} \sim N(0, 1)$），在此基础上加总个体的处理效应、时间趋势和随机扰动项计算出最终的结果变量，即 $Y_{it} = i + 1.5 \times t + \tau_{it} \times D_{it} + \epsilon_{it}$。

在进行检验和估计之前，我们可以画出样本处理状态的分布图（如附录五[④]图 D.2 所示），如果政策处理时点较多且不集中，则潜在的异质性问题风险较大，应引起重视。

1. 处理效应异质性检验：Goodman-Bacon 分解法

Goodman-Bacon（2021）检验的核心是将双向固定效应估计量拆分为若干个 2×2－DID 组合，画出每一个 2×2－DID 对应的处理效应和权重，再对处理效应异质性问题是否严重进行判断。接下来，我们使用上面的模拟数据来介绍这种检验方法。

Goodman-Bacon 的检验结果如表 3 所示。若使用“较早接受处理组”为控制组的比重较大，且对应的平均处理效应符号与使用“从未接受处理组”为控制组的符号相反，则异质性问题可能较为严重。在本文数值模拟的例子中，使用这类使用“较早接受处理组”为控制组的 2×2－DID 占整体的 38.2%，并且其平均处理效应与其他两类符号相反，说明异质性问题较为明显，应当考虑使用其他稳健的方法加以估计。事实上，Goodman-Bacon（2021）还提供了分解图以便于更为清晰地展示相关结果，但由于篇幅所限本文将其收录于附录五图 D.3 中。

表 3　　Goodman-Bacon 分解结果

2×2－DID 控制组类型	权重	平均处理效应
以“尚未接受处理组”为控制组	0.373	0.477
以“较早接受处理组”为控制组	0.382	－0.575
以“从未接受处理组”为控制组	0.245	1.981

2. TWFE、三类“异质性－稳健”估计量的估计效果对比

我们沿用前面生成的模拟数据，在下面的图 5 中展示了 TWFE 以及三类“异质性－稳

① 详见 Borusyak 的 GitHub 主页：https://github.com/borusyak/did_imputation。
② 只是为了说明方便，我们按照文献惯例进行了这一假设。如果我们改变处理效应的形式也不会影响结论。
③ 详见《数量经济技术经济研究》杂志网站的论文附录。
④ 详见《数量经济技术经济研究》杂志网站的论文附录。

健”估计量的估计效果。作为参照依据，我们首先计算出各期真实的平均处理效应，即将数据中模拟生成的处理效应按照距离处理的时长分别加总，得到图 5 中 0 期至 +6 期的真实值，而在处理发生前的 -7 至 -1 期，真实值为 0。受篇幅所限，我们选取了六个主要估计量进行比较，分别为 TWFE、de Chaisemartin 和 D'Haultfœuille（2022a）、Sun 和 Abraham（2021）、Callaway 和 Sant'Anna（2021）、Borusyak 等（2021）①、Cengiz 等（2019）。

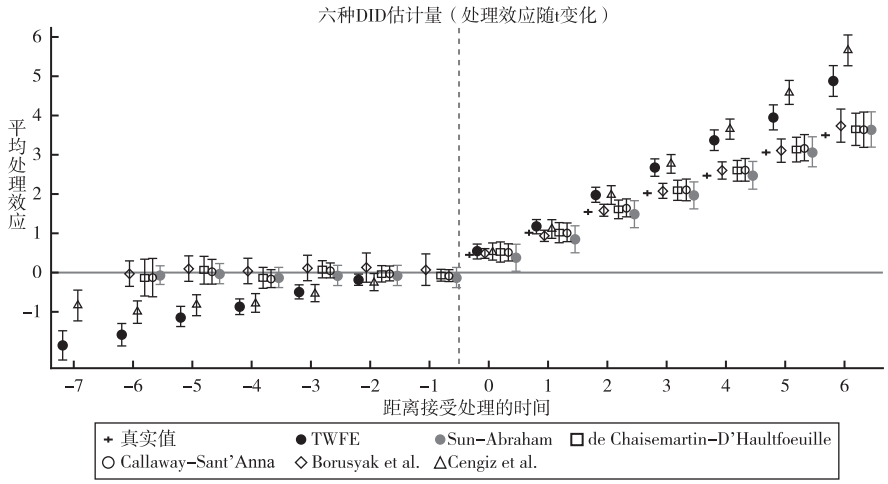

图 5　事件研究：六种估计量的对比

在图 5 中，我们对比真实值发现，处理效应异质性比较严重时，TWFE 会错误地估计 -7至 -2 期的动态效应，并且 0 至 +6 期的系数估计也会随着时间增加而偏误愈发明显。相比之下，其他四种估计量，即 Chaisemartin 和 D'Haultfœuille（2022a）、Callaway 和 Sant'Anna（2021）、Sun 和 Abraham（2021）和 Borusyak 等（2021）的置信区间都涵盖了真实值，并且点估计的结果也都在真实值附近，明显优于 TWFE。需要指出的是，Cengiz 等（2019）提出的估计量与真实值相差较远，这可能也与其存在的多方面局限有关。

这里需要特别说明的是，图 5 中各方法基期的选择略有差别，但并不影响结论。通常情况下，研究者使用 TWFE 时选取接受处理前的“ -1 期”作为基准，我们在图 5 中也依照惯例将“ -1 期”作为 TWFE 和 Cengiz 等（2019）的基准。而 Chaisemartin 和 D'Haultfœuille（2022a）以及 Callaway 和 Sant'Anna（2021）提出的估计量则没有使用固定基期的方式，特别是 Chaisemartin 和 D'Haultfœuille 提供的估计量目前的软件包尚不支持使用者选择固定的基期。为了方便同一类方法内部的对比，我们选取了 -7 期为其他四种方法的基期。作为稳健性检验，我们在附录三中的第二部分（详见图 B. 1 ~ 图 B. 4），提供了将“ -7 期”作为基期的对比，尽管趋势有所差异，但结论并未发生变化。

总的来说，由于处理效应异质性的存在，TWFE 的估计会存在潜在偏误；“计算组别 - 时期平均处理效应（CATT）加权平均”的三类方法的估计效果普遍较好，彼此也十分接近；插补估计量在某些情况下更有效率（如第 0 期 ~ 第 4 期的置信区间明显小于其他方

① Liu 等（2022）的 FEct 和 Borusyak 等（2021）是等价的，受篇幅所限，我们选取了 Borusyak 等（2021）进行呈现。在《数量经济技术经济研究》杂志网站的论文附录三第三部分（图 C. 1），我们也呈现了 Liu 等（2022）的估计结果。

法），但这种有效性基于的假设并不总是成立；堆叠估计量甚至有可能会出现偏误，正如本文之前的建议所言，研究者应当谨慎使用这一方法。

为了进一步对比六种估计量的估计效果，本文还尝试了另一种数值模拟的方式。我们构造了一个包含200个体和5期数据的平衡面板，随机模拟了处理时间和处理效应的大小，与上一个模拟不同之处在于改变了处理效应异质性的形式，我们假定同一时点接受处理的个体处理效应相同，不同时点接受处理的个体处理效应不同，即 $\tau_{it} = (0.3 \times E_i + 1) \times 1\{D_{it} = 1\}$。在此假设下，本文进行了500次数值模拟，每次模拟都随机分配处理时间和处理效应的大小，通过观察不同方法估计出的各期系数分布来比较其估计效果。

图6展示了500次模拟得到的各期系数分布箱线图，受篇幅限制，我们这里只展示了6期的分布信息。对于每一个箱体，中间的线代表中位数，箱体上下两端分别代表75%和25%的四分位数，虚线的上下两端表示最大值和最小值。我们发现，图6与图5呈现的结果一致，由于处理效应异质性的存在，TWFE的估计会存在潜在偏误；“计算组别－时期平均处理效应加权平均”的三种估计量的整体表现较好，也十分接近；插补估计量表现较好，但并非在所有时期都具备有效性；堆叠估计有可能存在较大偏差，应当谨慎使用。

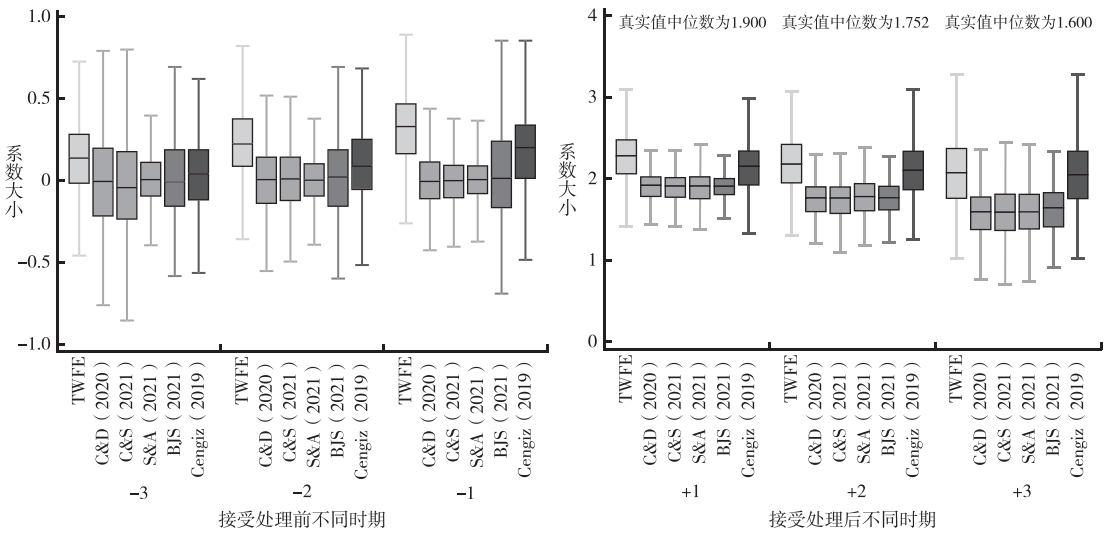

图6　事件研究：六种估计量系数分布对比

3. 平行趋势检验的对比和探讨

平行趋势是双重差分方法所依赖的关键假设，上一小节中，我们通过事件研究检验了TWFE以及6种异质性－稳健估计量的“平行性趋势”是否满足。对于这些新方法而言，它们大多提供了各自独特的检验方式，我们接下来对其进行简要介绍和对比，并对需要注意的问题进行探讨。

就计算组别－时期平均处理效应的三种方法而言，de Chaisemartin 和 D'Haultfœuille（2020a，2022a）提出了安慰剂检验用以检验平行趋势。Sun 和 Abraham（2021）使用IW估计量绘制事件研究图用以判断平行趋势是否满足，不过并没有给出单独的检验估计量。

Callaway 和 Sant'Anna（2021）则建议对每一个组别－时期平均处理效应（$ATT(e, t)$）的平行趋势是否成立进行检验，即分别绘制每一个组别－时期平均处理效应的事件研究图。

在附录五图 D.4 中，我们以“样本首次受处理的时点”为划分依据，将样本标示为 2 ~ 8 组①，进而呈现每个组的事件研究图进行判断。

特别值得注意的是，Callaway 和 Sant'Anna（2021）在平行趋势检验和事件研究时默认使用了“变化基期（Varying Base Period）”，即估计每一期系数时假设政策冲击发生在上一期，而 TWFE 以及 Sun 和 Abraham（2021）则使用了选定的“统一基期（Universal Base Period）”。针对这两种方式的选择，Callaway 在其个人主页②中进行了讨论，他指出：当（1）研究者认为潜在的预期效应可能有影响，且/或（2）接受处理前时期数相对较小时，使用“变化基期”更好；当（1）研究者认为组间趋势上有长期的差异，且/或（2）接受处理前时期数相对较大时，使用“统一基期”更好。

就插补估计量而言，Borusyak 等（2021）采用了 F 检验的方式，具体来说，他们将样本限制为从未接受处理的观测和尚未接受处理的观测。然后使用双向固定效应模型进行回归分析，其中需要加入一系列政策处理前的指示变量。通过检验这些指示变量的联合显著性来判断平行趋势假设是否成立③。Liu 等（2022）则采用了“等效检验（Equivalence Test）”的方法，即构造出与 F 检验具有同等效力的“等效区间”，借助安慰剂检验的思想不断去掉一期（Leave-One-Period-Out），然后通过其他数据来估计该期的值，得到冲击发生前每一期的平均样本外预测误差，如果接受处理前的时间趋势落在等效区间内则认为通过了检验④。

就堆叠回归估计量而言，虽然 Cengiz 等（2019）未在文中明确指出进行平行趋势检验的方式，但其方法可以通过绘制事件研究图来检验。

五、“异质性 - 稳健”估计量的应用现状与建议

1. 异质性—稳健估计方法的应用

目前，有一些学者使用“异质性 - 稳健”估计量对以往发表的文献进行了重估，发现部分文章与原结果存在较大差异。例如，de Chaisemartin 和 D'Haultfœuille（2020a）在其文章中对 Vella 和 Verbeek（1998）有关工会身份对工资的影响的研究进行了回顾。他们发现 TWFE 估计下，会得出工会身份会显著带来正的工资溢价（β_{fe} 为 0.107）。而当考虑到处理效应异质性问题后，则发现工会身份仅会导致工资较小幅度的提升（DID_M 为 0.041），且并不显著。Baker 等（2022）在其文章中回顾了 Beck 等（2010）有关放松银行洲际支行建设管制对经济不平等的影响的研究，同样发现了与原始结果截然不同的结论。他们使用 Callaway 和 Sant'Anna（2021）的方法重新估计了原文的基准结果，发现政策效应变为 0.001 且并不显著。

对于交错 DID 的情形，处理效应异质性会带来 TWFE 的潜在估计偏误，但并不绝对。在以《美国经济评论》为代表的经济学顶级刊物中，越来越多的学者以稳健性检验的方式

① 由于在估计过程中不能包含“纯处理组”（Always-Treated）的样本，因此该数据中不存在“首次接受处理的时点为 1”的样本，即不存在 Group 1。Callaway 和 Sant'Anna（2021）提供的平行趋势检验方式可能面临估计效率的问题，如果处理组中观测量较少，那么基于这些样本进行估计得到的系数的标准误可能较大，从而影响了平行趋势检验的效率。

② 详见：https://bcallaway11.github.io/posts/event-study-universal-v-varying-base-period。

③ 值得注意的是，Borusyak 等（2021）的这种做法通过使用未接受政策处理的样本进行平行趋势检验检验，避免了传统平行趋势检验中无法分离识别（Identification）和检验（Validation）的问题。更多细节可以参考 Borusyak 等（2021）在原文中的论述，本文不做过多探讨。

④ 利用模拟数据进行“等效检验”的结果参见《数量经济技术经济研究》杂志网站的论文附录三第三部分（图 C.2）。

将新估计量应用到研究工作中，从而对 TWFE 的结果进行佐证。例如，Guriev 等（2021）探讨了 3G 网络建设对政府腐败行为的影响，他们在稳健性检验部分利用 de Chaisemartin 和 D'Haultfœuille（2020a）的估计量，发现核心结论并不受处理效应异质性的干扰。Cantoni 和 Pons（2021）检验同事退休对法官决策的影响中，采用了 Sun 和 Abraham（2021）的 IW 估计量。Kim（2022）则使用了 Callaway 和 Sant'Anna（2021）提出的估计量来解决处理效应异质性问题。而 Biasi 和 Sarons（2022）更为全面，他们研究了灵活工资制度对性别不平等的影响，同时使用了 Sun 和 Abraham（2021）、Borusyak 等（2021）和 Cengiz 等（2019）提出的估计量，验证了其结果的稳健性。

总之，这些最新发表的文献涉及了环境经济学、劳动经济学、财政金融等多个领域，充分说明处理效应异质性问题已经受到学界广泛重视。

2. 国内应用现状与建议

我国出台的很多政策都具有“先行先试”的特点，国内经济学领域采用交错 DID 研究设计的文章十分普遍（王鹏超和韩立彬，2022）。然而，在估计方法使用的科学性方面，还需要进一步增强。我们以国内某五种权威期刊近年来发表的文章为例，总结了目前国内交错 DID 相关研究存在的一些特点。第一，目前几乎没有研究关注到处理效应的异质性，也未对处理效应同质性假设进行过探讨。第二，随着一些试点政策的全面铺开（例如“营改增”政策），数据样本期内不再存在“从未接受处理组”，绝大多数研究忽略了这一基本事实。第三，一些政策冲击存在“退出”情形（如政治关联的建立与消失），现有中文的实证工作基本未能给出一个合理的估计框架。第四，个别研究关注到了“处理状态发生转换”样本量较少的问题（如研究重庆和上海推行房产税的影响），但是尚未给出合理的解决方案。

鉴于此，我们除了在第四节针对各种方法的适用场景给出建议外，还结合了一系列综述文章（Baker 等，2022；Liu 等，2022；Roth 等，2022）和应用性工作（例如 Guriev 等，2021；Cantoni 和 Pons，2021；Kim，2022），给应用研究者提供了一份更为全面的操作建议：

（1）研究者在进行实证分析之前尽量提供样本中个体接受处理时间的分布图。交错 DID 引发偏误的关键原因就在于样本接受政策处理的时点存在差异，进而导致“坏的控制组”进入实证估计中。在开展研究前绘制处理时间分布图有利于直观地判断样本中较早接受处理组，较晚接受处理组和从未接受处理组的分布情况，从而初步判断异质性的严重程度。

（2）研究者应首先使用双向固定效应模型进行估计。TWFE 可以作为一个比较基准，用以判断是否存在潜在的处理效应异质性问题及其严重性。建议研究者不仅用 TWFE 估计平均处理效应，还应采用事件研究法绘制系数图像，获得动态效应的基准估计。

（3）研究者应结合适用场景，使用多种估计量进行实证分析和稳健性检验。在做研究方法选择时，研究者应当结合政策和观测数据的特点重点关注以下几个维度：第一，试点政策是不是全面铺开，即数据中是否存在“从未接受处理组”；第二，是否存在政策“退出情形”；第三，“处理状态发生转换”的样本数量是否过少；第四，是否存在不可观测的随时间变化的混杂因素；第五，是否是高频面板数据等。研究者在进行实证分析的过程中，应同时结合政策和数据的特点，综合判断估计方法的选择，并呈现方法的稳健性。当然，必须强调的是，由于各种方法的前提假设不尽相同，研究者在稳健性检验中，不必要过分追求过多使用新方法，而是要根据数据和政策的特点决定。

（4）研究者应根据不同估计量的特点，谨慎解释回归结果。由于包括 de Chaisemartin 和 D'Haultfœuille（2020a，2022a）在内的多个估计量在估计时会丢失较多样本，因此在应用中，我们可能会发现不同方法在点估计以及置信区间上存在较大差异。此时，我们建议研究者不仅仅要关心估计系数的符号和显著性与 TWFE 的差异，还要关注估计系数大小的差别，并谨慎解读回归结果。

（5）研究者在使用新的估计方法时要格外关注平行趋势检验的问题。首先，研究者应注意并非所有异质性 - 稳健估计量都采用绘制事件研究图的方式进行平行性趋势检验，如 Borusyak 等（2021）采用了 F 检验。其次，若使用绘制事件研究图的方式检验平行趋势则应注意基期选择，Freyaldenhoven 等（2021）建议选择政策发生的前一期作为基期。再次，使用绘制事件研究图的方式检验平行趋势时，研究者应谨慎应对时间窗口的选择问题，Baker 等（2022）建议尽量避免对时间窗口进行归并处理。最后，研究者也可以参照 Roth 等（2022）的建议，在事件研究基础上增加效力分析和敏感性分析，提升平行趋势检验的稳健性。

六、拓展与展望

1. 假设与推断

首先，本文介绍的各种方法均建立在平行趋势假设下。与新的估计量相伴而生的检验方法也大多延续传统思路，并未解决传统方法中的潜在风险。近年来，部分学者致力于开发出新的检验方法。例如，Roth 和 Sant'Anna（2021b）提出传统的平行趋势检验方法多数情况下可能依赖于研究者选择的函数形式进而表现出不稳健性。在此基础上，Rambachan 和 Roth（2021）提出了更为稳健的检验平行趋势假设的方式，并且给出了敏感性分析来说明其方法对于函数形式等影响因素的变化是稳健的。然而，这些新的检验方式尚未被大量应用到实证研究中，未来有待研究者进一步完善。

其次，溢出效应是使用双重差分法时难以规避的问题，其对单位处理变量值稳定假设产生重要威胁。而在实际情况下，研究者往往难以避免溢出效应的产生，例如相邻地区在空间上的相互干扰、个体通过社会网络形成的互动极有可能导致处理效应的溢出。目前已有学者关注同时存在处理效应异质性和溢出效应时，应如何准确估计政策的处理效应，例如 Butts（2021）基于 Callaway 和 Sant'Anna（2021）的研究框架考察了存在溢出效应时的平均处理效应的估计问题，部分解决了局部空间溢出问题；Huber 和 Steinmayr（2021）也提出了一个从溢出效应中分离出个体层面处理效应的框架。这些研究尚不足以形成完整的、充分解决溢出问题的体系，仍需要研究者未来进行更加深入的探索。

此外，在使用上述新的估计量进行统计推断时，往往会采用聚类稳健标准误，而这是建立在有大量聚类的前提假设下，在此基础上才能借助中心极限定理给出相应的置信区间。而在现实情况下，彼此独立的聚类数量往往不是足够的，例如以省为单位推行的政策，省的数量往往是相对较少的、不足以支撑得到准确的估计。学界针对这一问题，形成了若干种不同的解决思路。第一种是基于模型的解决方法，其核心思想是利用模型来描述聚类内的联系，通常做法是对聚类层面的普遍冲击施加一定的约束条件，而不同方法的约束条件一般存在差异。代表性文献包括 Donald 和 Lang（2007），Conley 和 Taber（2011）和 Ferman 和 Pinto（2019）。第二种是集群自助法（Cluster Wild Bootstrap），Cameron 等（2008）利用模拟结果说明该方法在仅有 5 个聚类的设定下能够有出色表现；Canay 等（2021）则正式提出了集群

自助法得到有效估计的若干条件，并特别指出这一方法依赖于较强的同质性假设。此外，还有较多文献考虑借鉴 Fisher 随机检验（Fisher Randomization Tests，FRTs）也即置换检验（Permutation Tests）的思路解决该问题，如 Roth 和 Sant'Anna（2021a）在渐进式 DID 框架下的拓展，表明 FRTs 可以提供有限样本下的有效估计。这些新的推断思路尚未被广泛运用，与新的估计方法之间的融合也有待探索。

2. 研究设计

本文所介绍的“异质性－稳健”估计方法从本质上讲都属于利用面板数据进行因果分析的研究范畴。在这一研究领域中有关合成控制双重差分法（Synthetic Difference-in-Differences）的研究正逐渐兴起，并有与异质性－稳健估计方法相融合的趋势。Xu（2017）在其研究中首次将合成控制法与固定效应模型结合在一起，提出了广义合成控制法（Generalized Synthetic Control Method）。双重差分法是这一方法的应用场景之一。Arkhangelsky 等（2021）的研究正式提出了合成控制双重差分方法。较之于传统的双重差分法，这种方法实现了对控制组的进一步筛选使得估计结果更为精准。Arkhangelsky 等（2021）还指出合成控制双重差分法也能解决交错 DID 中的处理效应异质性问题。Ben-Michael 等（2021）则将合成控制双重差分法的思想聚焦于交错 DID 领域，提出了专门应对交错 DID 的部分混合合成控制法（Partially Pooled SCM）。可以预见的是，在未来的研究中合成控制双重差分法可以与异质性－稳健估计方法实现彼此融合，从而优化使用面板数据的因果分析。

需要指出的是，本文研究的政策情景主要涉及非随机的政策处理。事实上，当样本中的个体接受政策处理为随机事件时，采用恰当的估计量可能会获得更为有效的估计（Roth 和 Sant'Anna，2021a）。这一研究视角延续了 McKenzie（2012）的工作，他强调了 TWFE 估计量在随机控制实验（RCT）的背景下是没有效率的。Shaikh 和 Toulis（2021）则在此基础上提出了一种适用于观测数据且处理时间在控制某些固定特征后具有随机性的情况下的估计方法。基于观测数据、具有准自然实验特征的研究一直以来是实证领域的重点，这些应用方法上的进展也将会得到越来越多的关注和运用。

七、结　　语

众所周知，试点是中国政策创新与制度建设的一个重要机制，它体现了中央政府“尊重地方和基层的经验、智慧和首创精神”的基本理念（江小涓，2020），也是理解中国政策过程的重要研究视角（赵慧，2019）。近年来，来自经济学、政治学等社会科学领域的很多实证研究工作都主要是在政策试点的研究框架下展开讨论的，进而将其视为解释中国渐进式经济改革取得成功的原因之一。因此，如何结合政策的特征以及观测数据的特点，选取科学的研究方法对政策进行评估，就显得尤为重要。

目前，国内经济学领域大量文献利用双重差分设计的框架进行经验研究，尤其是在试点型政策的评估中，交错 DID 研究框架得到了广泛使用。然而，交错 DID 有一个重要特点就是存在处理效应的异质性，它可能会导致传统的 TWFE 估计出现偏误。在此背景下，为了给国内应用研究者提供一套科学、行之有效的解决方案建议，本文系统总结了前沿计量文献中三类新的“异质性－稳健”估计量，就前提假设、估计量性质、适用场景和使用局限给出了深入探讨，并结合每一类方法的特点给出了操作建议和估计方法的选择建议，以期待引发更加广泛的探讨，避免方法误用。

本文是提升国内学界利用交错 DID 进行政策试点评估和因果推断经验研究规范性的一

项努力。由于观测数据背后的因果问题十分复杂，研究方法的合理选取和数据结构的特点都将影响经验研究结论的科学性和可信性。鉴于此，本文试图帮助应用研究者拨开研究方法和数据的“迷雾”，让更多科研工作者在实践中不被研究方法所困，也充分体现出科学地应用国际前沿方法的“初心”，从而更好地洞察真实世界，同时也让经验研究成果的可信性和科学性进一步提升，更好地为学术界和政策界服务。

参 考 文 献

[1] Angrist J. D. , Pischke J. S. , 2008, *Mostly Harmless Econometrics* [M], Princeton University Press.

[2] Arkhangelsky D. , Athey S. , Hirshberg D. A. , Imbens G. W. , Wager S. , 2021, *Synthetic Difference-in-Differences* [J], American Economic Review, 111 (12), 4088 ~ 4118.

[3] Athey S. , Bayati M. , Doudchenko N. , Imbens G. , Khosravi K. , 2021, *Matrix Completion Methods for Causal Panel Data Models* [J], Journal of the American Statistical Association, 116 (536), 1716 ~ 1730.

[4] Baker A. C. , Larcker D. F. , Wang C. C. , 2022, *How Much Should We Trust Staggered Difference-in-Differences Estimates*? [J], Journal of Financial Economics, 144 (2), 370 ~ 395.

[5] Beck T. , Levine R. , Levkov A. , 2010, *Big Bad Banks*? *The Winners and Losers from Bank Deregulation in the United States* [J], The Journal of Finance, 65 (5), 1637 ~ 1667.

[6] Biasi B. , Sarsons H. , 2022, *Flexible Wages*, *Bargaining*, *and the Gender Gap* [J], The Quarterly Journal of Economics, 137 (1), 215 ~ 266.

[7] Ben-Michael E. , Feller A. , Rothstein J. , 2021, *Synthetic Controls with Staggered Adoption* [J], National Bureau of Economic Research.

[8] Borusyak K. , Jaravel X. , 2017, *Revisiting Event Study Designs* [R], Working Paper.

[9] Borusyak K. , Jaravel X. , Spiess J. , 2021, *Revisiting Event Study Designs*: *Robust and Efficient Estimation* [R], Working Paper.

[10] Butts K. , 2021, *Difference-in-Differences Estimation with Spatial Spillovers* [R], Working Paper.

[11] Callaway B. , Sant'Anna P. H. , 2021, *Difference-in-Differences with Multiple Time Periods* [J], Journal of Econometrics, 225 (2), 200 ~ 230.

[12] Cameron A. C. , Gelbach J. B. , Miller D. L. , 2008, *Bootstrap-Based Improvements for Inference with Clustered Errors* [J], The Review of Economics and Statistics, 90 (3), 414 ~ 427.

[13] Canay I. A. , Santos A. , Shaikh A. M. , 2021, *The Wild Bootstrap with a "Small" Number of "Large" Clusters* [J], Review of Economics and Statistics, 103 (2), 346 ~ 363.

[14] Cantoni E. , Pons V. , 2021, *Strict ID Laws Don't Stop Voters*: *Evidence from a US Nationwide Panel*, 2008 ~ 2018 [J], The Quarterly Journal of Economics, 136 (4), 2615 ~ 2660.

[15] Cengiz D. , Dube A. , Lindner A. , Zipperer B. , 2019, *The Effect of Minimum Wages on Low-Wage Jobs* [J], The Quarterly Journal of Economics, 134 (3), 1405 ~ 1454.

[16] Conley T. G. , Taber C. R. , 2011, *Inference with 'Difference in Differences' with a Small Number of Policy Changes* [J], The Review of Economics and Statistics, 93 (1), 113 ~ 125.

[17] de Chaisemartin C. , D'Haultfoeuille X. , 2020a, *Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects* [J], American Economic Review, 110 (9), 2964 ~ 96.

[18] de Chaisemartin C. , D'Haultfoeuille X. , 2020b, *Two-way Fixed Effects Regressions with Several Treatments* [R], Working Paper.

[19] de Chaisemartin C. , D'Haultfœuille X, 2022a, *Difference-in-Differences Estimators of Intertemporal Treatment Effects* [R], NBER Working Paper, No. 29873.

[20] de Chaisemartin C. , D'Haultfœuille X. , 2022b, *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects*: *A Survey* [R], NBER Working Paper, No. 29734.

[21] Donald S. G., Lang K., 2007, *Inference with Difference-in-Differences and Other Panel Data* [J], The Review of Economics and Statistics, 89 (2), 221 ~ 233.

[22] Ferman B., Pinto C., 2019, *Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity* [J]. Review of Economics and Statistics, 101 (3), 452 ~ 467.

[23] Freyaldenhoven S., Hansen C., Shapiro J. M., 2019, *Pre-event Trends in the Panel Event-Study Design* [J], American Economic Review, 109 (9), 3307 ~ 3338.

[24] Freyaldenhoven S., Hansen C., Pérez J. P., Shapiro J. M., 2021, *Visualization, Identification, and Estimation in the Linear Panel Event-Study Design* [R], NBER Working Paper, No. 29170.

[25] Gardner J., 2021, *Two-Stage Differences in Differences* [R], Working Paper.

[26] Gobillon L., Magnac T., 2016, Regional policy evaluation: Interactive fixed effects and synthetic controls [J], Review of Economics and Statistics, 98 (3), 535 ~ 551.

[27] Goodman-Bacon A., 2021, *Difference-in-differences with Variation in Treatment Timing* [J], Journal of Econometrics.

[28] Guriev S., Melnikov N., Zhuravskaya E., 2021, *3G Internet and Confidence in Government* [J], The Quarterly Journal of Economics, 136 (4), 2533 ~ 2613.

[29] Huber M., Steinmayr A., 2021, *A Framework for Separating Individual-Level Treatment Effects from Spillover Effects* [J], Journal of Business & Economic Statistics, 39 (2), 422 ~ 436.

[30] Imai K., Kim I. S., 2021, *On the Use of Two-Way fixed Effects Regression Models for Causal Inference with Panel Data* [J], Political Analysis, 29 (3), 405 ~ 415.

[31] Jakiela P., 2021, *Simple Diagnostics for Two-Way Fixed Effects* [R], Working Paper.

[32] Kidziński Ł., Hastie T., 2018, *Modeling Longitudinal Data Using Matrix Completion* [R], Working Paper.

[33] Kim W., 2022, *Television and American Consumerism* [J], Journal of Public Economics, 208 (222), 104609.

[34] Liu L., Wang Y., Xu Y., 2022, *A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data* [J], American Journal of Political Science, forthcoming.

[35] Marcus M., Sant'Anna P. H., 2021, *The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics* [J], Journal of the Association of Environmental and Resource Economists, 8 (2), 235 ~ 275.

[36] McKenzie D., 2012, *Beyond Baseline and Follow-up: The Case for More T in Experiments* [J], Journal of development Economics, 99 (2), 210 ~ 221.

[37] Rambachan A., Roth J., 2021, *An Honest Approach to Parallel Trends* [R], Working Paper.

[38] Roth J., 2022, *Pre-Test with Caution: Event-Study Estimates After Testing for Parallel Trends* [J], American Economic Review: Insights, forthcoming.

[39] Roth J., Sant'Anna P. H., 2021a, *Efficient Estimation for Staggered Rollout Designs* [R], Working Paper.

[40] Roth J., Sant'Anna P. H., 2021b, *When is Parallel Trends Sensitive to Functional Form?* [R], Working Paper.

[41] Roth J., Sant'Anna P. H., Bilinski A., Poe J., 2022, *What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature* [R], Working Paper.

[42] Shaikh A. M., Toulis P., 2021, *Randomization Tests in Observational Studies with Staggered Adoption of Treatment* [J], Journal of the American Statistical Association, 116 (536), 1835 ~ 1848.

[43] Strezhnev A., 2018, *Semiparametric Weighting Estimators for Multi-Period Differencein-Differences Designs* [R], Working Paper.

[44] Sun L., Abraham S., 2021, *Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects* [J], Journal of Econometrics, 225 (2), 175 ~ 199.

［45］ Vella F. , Verbeek M. , 1998, *Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men* ［J］, Journal of Applied Econometrics, 13 （2）, 163 ~ 183.

［46］ Wooldridge J. , 2021, *Two-Way Fixed Effects, The Two-Way Mundlak Regression, and Dfference-in-Differences Estimators* ［R］, Working Paper.

［47］ Xu Y. , 2017, *Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models* ［J］, Political Analysis, 25 （1）, 57 ~ 76.

［48］ Xu Y. , 2022, *Causal Inference with Time-Series Cross-Sectional Data: A Reflection* ［R］, Working Paper.

［49］ 陈林、伍海军：《国内双重差分法的研究现状与潜在问题》［J］，《数量经济技术经济研究》2015 年第 7 期。

［50］ 范子英：《如何科学评估经济政策的效应》［J］，《财经智库》2018 年第 3 期。

［51］ 胡日东、林明裕：《双重差分方法的研究动态及其在公共政策评估中的应用》［J］，《财经智库》2018 年第 3 期。

［52］ 黄炜、张子尧、刘安然：《从双重差分法到事件研究法》［J］，《产业经济评论》2022 年第 2 期。

［53］ 江小涓：《江小涓学术自传》［M］，广东经济出版社，2020。

［54］ 石华军、楚尔鸣：《政策效果评估的双重差分方法》［J］，《统计与决策》2017 年第 7 期。

［55］ 王鹏超、韩立彬：《交错双重差分法的潜在问题与解决措施》［J］，《东北财经大学学报》2022 年。

［56］ 赵慧：《政策试点的试验机制：情境与策略》［J］，《中国行政管理》2019 年第 1 期。

［57］ 周黎安、陈烨：《中国农村税费改革的政策效果：基于双重差分模型的估计》［J］，《经济研究》2005 年第 8 期。

Staggered Difference-in-differences Method: Heterogeneous Treatment Effects and Choice of Estimation

Liu Chong[1]　Sha Xuekang[1]　Zhang Yan[2]

（1. School of Economics, Peking University;
2. Guanghua School of Management, Peking University）

Abstract: Difference-in-differences （DID） is one of the most popular methods in social sciences for estimating causal effects. However, recent econometrics research on staggered DID documents that the traditional two-way fixed effect estimator （TWFE） may not provide a valid estimation due to the existence of heterogeneous treatment effects. To solve such a problem, a variety of heterogeneity-robust estimators have been raised. This paper provides a literature review of new trends on staggered DID designs and gives practical suggestions for practitioners. We first survey a fast-growing literature and explain the reasons for the potential estimation bias of static/dynamic TWFE in a staggered DID setting. We then synthesize the intuition of three types of solutions for heterogeneity-robust estimation. Based on the simulation data, we apply these new methods and find that these alternative estimators helpful identifying the true treatment effects under their respective assumptions. We further discuss in detail how to conduct tests for "parallel trend" and how to choose base period in event study. In addition to traditional event-study plots, we also introduce the "Equivalence Test" provided by Liu et al. （2022） and "F-test" proposed by Borusyak et al. （2021）. The origin of the pitfalls with TWFE in a staggered DID design is "the forbidden comparisons", i. e. the previously

treated groups compared to newly treated groups. Given this issue, several alternative heterogeneity-robust estimators have been proposed to capture ATT effectively. To compare the differences among various estimators, we further categorize them into three types according to their key ideas and then give application suggestions respectively. (1) We suggest researchers pay attention to three aspects before carrying on "CATT" methods: whether the sample size is large enough, whether the treatment has ever turned off and whether there exists never-treated samples. (2) We remind researchers that imputation estimators may rely on correct model specification, and if there are unobservable time-varying confounding factors, it is recommended that researchers use the IFEct or MC method proposed by Liu et al. (2022). (3) We recommend that researchers choose "stacked regression estimator" with caution since its statistical properties haven't been rigorously proved and that data replication issues may arise. Our simulation results also support the suggestions above. We find the "CATT" estimators are similar but less efficiency. Imputation estimators can lead to more efficient estimate, but this is not always the case. Stacked regression estimator may lead to bias but researchers can use this method as robustness check. First, this paper reviews the recent advances in the econometrics of staggered DID and summarizes these theoretical works from an applied researcher's perspective. Our discussion on these heterogeneity-robust estimators highlights the different applicable scenarios of each method and helps to clarify when and how to use these new approaches. Second, this paper explores the differences in the three types of heterogeneity-robust estimation methods from their core assumptions to statistical properties with numerical simulation results. This will help the applied researchers' better understanding of the characteristics of each method, so as to make reasonable choices in their research. Commonly used TWFE DID specification is susceptible to biased estimates. This paper provides a guidance for applied researchers on how to select an appropriate heterogeneity-robust estimator in combination with the application scenarios, and how to understand and verify the corresponding premise assumptions.

Key Words: Staggered DID; Heterogeneous Treatment Effects; Heterogeneity-robust Estimator; Application Scenarios

JEL Classification: C13; C21; C22; C23

(责任编辑：李兆辰)