从双重差分法到事件研究法

黄炜　张子尧*　刘安然

摘　要：近年来双重差分法在政策评估领域得到了广泛应用，然而由于对双重差分法的识别假设等基本问题理解不够准确或存在误解，部分研究出现了随意添加控制变量、错误解释平行趋势检验等一系列问题。本文试图对双重差分法进行系统性的归纳梳理，以厘清在双重差分法实践应用中的一些相关基本问题。本文分析了双重差分法的识别假设及其经济含义，归纳了研究中常见的几类双重差分法的设定方式，详细分析了控制变量的选取、平行趋势检验以及组间线性时间趋势的控制等应用中的常见问题。针对近年来使用逐渐增多的交错双重差分法及其可能存在的偏误，本文建议使用动态双重差分法和事件研究法作为基准识别策略，并详细说明了二者的使用方法、相互关系和注意事项。最后，本文强调了使用双重差分法进行实证研究的其他问题，包括重视真实的制度背景、对政策外生的理解、溢出效应的处理以及一般均衡视角下的成本收益分析等。

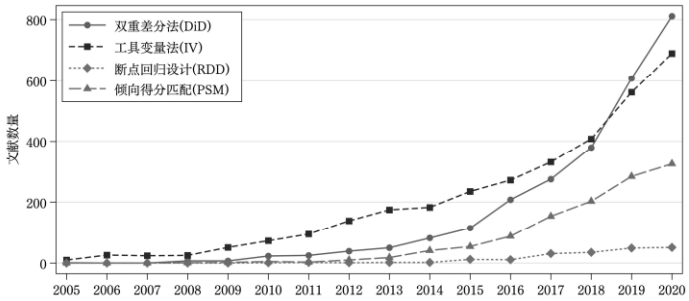
关键词：因果推断；双重差分法；事件研究法；实证研究

中图分类号：F224.0

DOI：10.19313/j.cnki.cn10-1223/f.20211227.002

一、引　言

随着计量经济学“可信性革命”（credibility revolution）席卷经济学的各个领域，基于潜在因果模型的因果效应识别策略，如匹配法（matching）、工具变量法（instrumental variable）、双重差分法（difference-in-differences）和断点回归设计（regression discontinuity design）等，逐渐成为了经济学等社会科学领域实证研究的通行研究范式。上述几种方法的使用要求和适应场景各不相同，双重差分法由于其直观清晰、易于理解，并且实际操作难度较低、上手简单等特点而广为应用。图 1 展示了中文期刊经济管理类学术论文各类方法的使用数量变化，可以看到自 2015 年后使用双重差



注：检索范围为 2005-2020 年间发表在核心期刊和 CSSCI 期刊上的经济与管理科学类学术论文。检索条件为主题、篇名、摘要或关键词出现“双重差分”或“倍差”“工具变量”“断点回归”“倾向得分匹配”。

图 1　2005-2020 年间中文期刊经济管理类学术论文计量方法使用简况

* 黄炜，美国埃默里大学（Emory University）助理教授；张子尧（通信作者，zzy2018@ruc.edu.cn），中国人民大学财政金融学院博士研究生；刘安然，中国人民大学农业与农村发展学院硕士研究生。项目来源：中国人民大学农业与农村发展学院学生科研训练计划项目（批准号：2021A1）。作者感谢中国人民大学国家发展与战略研究院刘瑞明教授的评论与建议。感谢匿名审稿人的宝贵意见。文责自负。

分法的国内经济管理类研究数量急剧上升，在 2019 年超越工具变量法成为了目前使用最为广泛的计量方法，并且其上升趋势仍有进一步加强的倾向。由此或可推测，在未来的一段时期内，双重差分法仍然将是经济管理类实证研究的主流方法之一。

双重差分法在实证研究中主要用于评估政策效应。与其他方法相比，双重差分法的识别方法非常直观：先观察受政策影响的个体在政策前后的变化，再观察未受政策影响的个体在政策前后的变化，两个变化之间的差异就是政策干预对个体的影响。同时，双重差分法可以非常方便地使用最小二乘法来实现。直观理解加上简单易行使得双重差分法得到了广泛应用，学者们使用双重差分法评估了许多重要的政策效应，例如，刘瑞明和赵仁杰（2015）发现国家高新区的建设显著地促进了地区经济增长；吕越等（2019）发现“一带一路”倡议促进了中国企业海外绿地投资；Liu 和 Mao（2019）发现增值税转型改革显著提高了企业投资并改善了生产率；宋弘等（2021）发现社保法定费率下降使得企业社保缴费参与率提高，但削弱了企业的劳动力需求。以上文献只是大量使用双重差分法实证研究中的沧海一粟。

然而，伴随着双重差分法的广泛使用，一些对于双重差分法的不精确理解甚至是错误认识也逐渐开始出现。常见的一些问题包括：双重差分法的基本识别假设是什么？双重差分法需要政策是完全随机分配的吗？平行趋势假设是什么？通常所说的平行趋势检验真的是在检验平行趋势假设吗？控制变量应该如何选取？什么样的变量必须控制，什么样的变量必须不能控制，什么样的变量可以控制也可以不控制？当政策干预时点不一致时双重差分法应该如何实现？这种实现方法有什么问题，应该如何改进？等等。实证研究者在研究过程中或多或少都曾遇到或将要遇到上述问题，但是标准的计量经济学教材中很少直接回应这些实践方面的疑问，研究者不得不根据自身的理解来处理上述问题，这是对双重差分法产生不正确理解的原因之一。基于此，本文结合国际上关于双重差分法的最新研究，试图对双重差分法应用中的一系列问题进行初步探讨，希望能够帮助廓清一些疑惑，为我国经济学界研究与国际前沿接轨提供些微贡献。

本文的结构如下：第二部分描述双重差分法的计量实现，对研究中常用的几种双重差分法进行归纳总结，而后着重强调了双重差分法的识别假设及其直观含义。第三部分分析双重差分法使用中的控制变量选取、平行趋势检验的实现和理解，以及组间线性时间趋势是否控制三个常见易混淆的问题。第四部分讨论了近年来广泛应用的交错双重差分法的实现和潜在问题，以及如何尝试使用动态双重差分法和事件研究法来克服交错双重差分法的不足。第五部分讨论了双重差分法评估政策效应时常见的几个问题，包括需要重视真实的制度背景、政策干预是否需要完全随机、溢出效应以及一般均衡视角下的成本收益分析。最后是总结性评论。

二、双重差分法的计量实现和识别假设

（一）标准 DiD（standard DiD）

双重差分是一种尝试采用控制组实际未经处理的结果变化作为处理组倘若未经处理的结果变化的反事实来分析因果效应的方法，通常包括冲击事件、处理组、控制组和时期这四个要素，其经典构造可以表示为如下形式：

$$Y_{it} = \alpha + \delta D_i + \lambda T_t + \beta \left(D_i \times T_t \right) + \varepsilon_{it} \tag{1}$$

其中，Y_{it} 为结果变量，D_i 为政策分组虚拟变量，T_t 为政策时间虚拟变量，$D_i \times T_t$ 为两者交互项，δ、λ 和 β 为各项前的系数，ε_{it} 为随机误差项。对上式取条件期望后，可得到表 1 所示的估计效

应，其中β表示文章所关注的因果效应。双重差分法通常涉及两组人群与两个时期。其中一组人群在第一个时期未接受处理，在第二个时期则受到处理或干预；另一组人群则在两个时期都未接受处理。将个体i在时期t接受处理定义为定义$D_{it}=1$，未接受处理定义为$D_{it}=0$。一般将在处理组接受处理前的时期（pre-treatment period）记为$T=0$，处理后的时期（post-treatment period）记为$T=1$。其中，对处理组个体有$D_{i1}=1$，对控制组个体有$D_{i1}=0$，对所有个体i有$D_{i0}=0$。

表1　　双重差分效应示意表

$E(Y\mid D,T)$	$T=0$	$T=1$	Δ
$D=0$	α	$\alpha+\lambda$	λ
$D=1$	$\alpha+\delta$	$\alpha+\delta+\lambda+\beta$	$\lambda+\beta$
Δ	δ	$\delta+\beta$	β

双重差分的核心是通过构造交互项来识别政策冲击对受影响个体（处理组）的平均处理效应（average treatment effect on the Treated，ATT），[①]即基于一个反事实框架来评估政策冲击发生与不发生这两种情况下处理结果Y_{it}的变化。真实的因果效应需要通过比较处理组接受处理与不接受处理的状态得出，然而在现实生活中，当冲击发生后，我们仅能观察到处理组受到冲击后的情况，无法真正知晓其未受冲击的情况。而在双重差分方法中，控制组提供了一个可供研究的反事实，即可将未受到处理的控制组在观察时期内的“变化”近似于处理组倘若未受到冲击将发生的变化。从处理组前后时期的变化中减去控制组前后时期的变化，即可得到因果效应β。上述分析的数学表达式如下式所示，第一个中括号内为处理组前后时期的差分效应，第二个中括号内为控制组前后时期的差分效应，两个一次差分再相减后，得到双重差分处理效应：

$$\beta=\left[\mathrm{E}(Y|D=1,T=1)-\mathrm{E}(Y|D=1,T=0)\right]-\left[\mathrm{E}(Y|D=0,T=1)-\mathrm{E}(Y|D=0,T=0)\right]$$

在实际应用中，双重差分方法经常与面板数据联系起来使用，此时多采用双向固定效应（Two-way fixed effects）模型，因此双重差分法有时会表述为如下形式：

$$Y_{it}=\alpha+\beta\left(D_i\times T_t\right)+\mu_i+\gamma_t+\varepsilon_{it}$$

其中，μ_i、γ_t分别为个体固定效应（individual fixed effects）和时间固定效应（time fixed effects），通过在回归时加入个体虚拟变量和时间虚拟变量便可控制个体固定效应和时间固定效应，而此时如果再放入处理组虚拟变量会带来严格多重共线性。μ_i、γ_t是对个体层面和每期时间的控制，比原本模型中的政策分组虚拟变量D_i（控制至组别层面）和政策时间虚拟变量T_t（控制处理期前后的效应）更为精细，包含了更多的信息。

（二）双重差分法的其他形式拓展

1. 交错双重差分法（staggered DiD）。标准双重差分法模型和双向固定效应双重差分法模型涉及的政策实施时点或冲击发生时点为同一时期。然而，现实生活中诸多政策实施未必发生在某一时点，而是先有试点再逐步推广，在渐进的过程中推而行之，如增值税转型、土地确权、新农保实施、高铁修建等。交错双重差分法为处理这类情形提供了方法。[②]当个体接受政策冲击的时间不同时，

① 更严格地说是干预发生后时期内的处理组平均因果效应（ATT at the post-treatment period）。

② 在中文文献中交错双重差分法有许多其他称谓，如时变双重差分法、异时双重差分法、多期双重差分法、渐进双重差分法等。

政策分组虚拟变量 D_i 变为 D_{it}，此时 D_{it} 即可用来表示个体 i 在时间 t 处是否受到政策冲击，而无需再生成交互项。不过在实际应用中，交错双重差分法可能会遇到难以找到控制组、部分样本始终为处理组、异质性处理效应等问题。由于交错双重差分法适用面较广且使用时又有诸多需要注意的事项，本文将在第四部分详细讨论这一方法的应用与利弊。

2. 广义双重差分法（generalized DiD）。当所有研究对象均或多或少同时受到了政策干预，即仅有处理组而无控制组时，仍然能够考虑应用双重差分法。对此，可以根据研究对象受到的具体冲击情况来构建处理强度（treatment intensity）指标来进行分析，此时个体维度并不是从 0 到 1 的改变，而是连续的变化。因此，可以将个体维度的政策分组虚拟变量替换为用以表示不同个体受政策影响程度的连续型变量，该种方法被称为广义双重差分法。[①]Nunn 和 Qian（2011）研究了一个经典的例子，他们研究了土豆种植扩散对欧洲人口增长的影响。欧洲几乎所有地区都种植了土豆，不存在未种植土豆的地区，因此没有标准意义上的控制组。他们的选择是将地区间土豆种植适宜度作为处理强度，以 1700 年前后为处理时点，使用广义双重差分法估计了引入土豆对人口增长的影响。

3. 队列双重差分法（cohort DiD）。队列双重差分法也被称为截面双重差分法，即使用横截面数据来评估某一历史事件对个体的长期影响。队列双重差分法同样是比较两个维度上的差异大小：一个维度为地区间差异，标识该地区是否受干预政策影响或干预强度；另一个维度为出生队列间差异，标识个体是否受到了干预政策的影响。队列双重差分法本质上是使用未受政策干预的出生队列作为受到政策干预的出生队列的反事实结果。Duflo（2001）是早期应用队列双重差分法的经典研究，近年来使用这一方法的代表性文献有 Chen 等（2020）的研究文献。

4. 模糊双重差分法（fuzzy DiD）。在标准双重差分法等方法的应用情境中，处理组和控制组之间通常泾渭分明，因此可以通过分组差分得到较为“干净”的处理效应。但是，有时冲击并未带来急剧（sharp）变化，所谓的“处理组”中虽然受冲击率高于其他组别，但并没有完全被干预或受政策冲击，而所谓的“控制组”中也并非完全没有受到冲击，即处理组和控制组之间没有明确的分野，不存在“干净”的处理组与控制组。模糊双重差分法为处理此类情形提供了可能，de Chaisemartin 和 d’Haultfoeuille（2018）在文章中详细介绍了该种方法，并利用该方法重新评估了印度尼西亚的教育回报。

5. 三重差分法（triple differences）。顾名思义，三重差分法引入了第三个维度“组别”（group），通过比较不同组别间的处理组和控制组在干预政策前后结果变量变化的差异来识别因果效应。[②]三重差分法的应用场景通常有两个：一是在平行趋势假设不满足时引入第三个维度的差分来帮助消除处理组和控制组间的时间趋势差异；二是在平行趋势满足时，用于识别干预政策在不同群体间的异质性处理效应。三重差分法是一个典型的实践先于理论的方法，其使用最早可以追溯到 Gruber（1994），近年来在顶级期刊使用越来越频繁，不过直到 Olden 和 Møen（2020）才较为完整地讨论了三重差分法的识别假设和使用条件。

6. 其他双重差分法。纵观上述各种类型的双重差分法，其基本思路是寻找观测样本在两个维度上的差异，其中一个维度用于控制不可观测的时间趋势，另一个维度用于测度政策效应的变化。如果从更加一般化的角度理解双重差分法背后的直觉和思想，可以发现事实上几乎任何两个维度的差异之差异都可以从双重差分的角度去理解。也就是说，几乎所有的交互项模型都可以理解为一

① 也有文献称之为强度双重差分法或连续双重差分法。

② 即组别（group）—处理状态（statement）—处理前后（time）三个维度的差异。

种双重差分法。一个典型的例子是 Mayzlin 等（2014）的研究，他们研究了造假成本对在线旅店预定网站的消费者评论的影响。两家在线酒店预订网站中，Expedia 网站只有实际完成订单的消费者可以评价服务质量，而 TripAdvisor 网站则是任何人都可以评价服务质量，所以两个网站的造假成本是不同的。他们发现，当一家旅店周围没有其他旅店存在时，该旅店在 TripAdvisor 上的好评率显著高于在 Expedia 上的好评率，这是因为该旅店试图操纵评论提高本店评分。当一家旅店旁边存在另一家邻近的旅店时，该旅店在 TripAdvisor 上的差评率显著高于 Expedia 上的差评率，这是各旅店试图打压竞争对手，为对手恶意评低分。他们的识别策略事实上和双重差分法不谋而合：Expedia 和 TripAdvisor 的造假成本构成了一个维度差异，旅店邻近范围内是否存在直接竞争者构成了另一个维度的差异，通过二者之差就能够识别出造假成本对网站消费者评论操纵的影响。[①]另一个例子是 Rajan 和 Zingales（1998）的研究，该研究试图论证金融发展对经济增长的影响。他们使用的一个识别策略是交互项模型：被解释变量是 k 国 j 行业的增长率，解释变量是 j 行业的外部融资依赖度和 k 国的金融发展程度的交互项。其背后的直觉如下：若金融发展确实能够促进经济增长，那么 j 行业的外部融资依赖度越高，金融发展对其经济增长的激励越强。因此，如果不同行业间的外部融资依赖程度差异（第一个维度）和金融发展水平的跨国差异（第二个维度）能够解释不同国家行业间的增长率差异，就能够论证金融发展对经济增长的影响。[②]

（三）双重差分法的识别假设

双重差分法的应用需要满足一定的假设条件，倘若违背了这些前提假设，估计结果可能会严重偏离真实的因果效应。本部分对双重差分法的识别假设内容及可能违背假设的情景、后果进行讨论。

1. 平行趋势假设。双重差分法最基本的假设是平行趋势假设（parallel trend assumption），又称共同趋势假设（common trend assumption），是指倘若处理组个体未接受干预或冲击，则其结果变动趋势与控制组个体结果变动趋势相同。该假设数学表达如下：

$$\mathrm{E}(Y^0|D=1,T=1)-\mathrm{E}(Y^0|D=1,T=0)=\mathrm{E}(Y^0|D=0,T=1)-\mathrm{E}(Y^0|D=0,T=0)$$

其中，Y^0 表示未受干预或冲击的结果变量。在该假设下，双重差分法的估计结果正是处理组接受处理后的平均处理效应（ATT at the post-treatment period）：

$$\begin{aligned}\beta&=\left[\mathrm{E}(Y|D=1,T=1)-\mathrm{E}(Y|D=1,T=0)\right]-\left[\mathrm{E}(Y|D=0,T=1)-\mathrm{E}(Y|D=0,T=0)\right]\\&=\underbrace{\left[\mathrm{E}(Y^1|D=1,T=1)-\mathrm{E}(Y^0|D=1,T=1)\right]}_{\text{处理组因果效应}}\\&\quad+\underbrace{\left[\mathrm{E}(Y^0|D=1,T=1)-\mathrm{E}(Y^0|D=1,T=0)\right]-\left[\mathrm{E}(Y^0|D=0,T=1)-\mathrm{E}(Y^0|D=0,T=0)\right]}_{\text{组间趋势差异}}\\&=\mathrm{E}\left(Y^1-Y^0|D=1,T=1\right)\end{aligned}\tag{2}$$

由上述分析可知，双重差分法要求在没有干预或处理的情况下，处理组和控制组的平均结果随

① 他们在原文第 2423 页第二段写到“Our main empirical analysis is akin to a differences in differences approach (although, unconventionally, neither of the differences is in the time dimension). Specifically, we examine differences in the reviews posted at TripAdvisor and Expedia for different types of hotels.”。

② 也可以从广义双重差分法的角度理解，这里的处理单元是行业，外部融资依赖度类似于每个行业的处理强度，各国金融发展程度的差异类似于政策干预前和干预后，金融发展程度低相当于政策干预前，金融发展程度高相当于政策干预后（虽然金融发展程度是连续变量，但也可以从离散化的角度理解）。

时间变化的趋势相同。该识别假设可以记为更简便的形式：$\mathrm{E}(\Delta Y^0|D{=}1)=\mathrm{E}(\Delta Y^0|D{=}0)$。双重差分法背后隐含着“准自然实验”的思想，并不严格要求处理组与控制组之间满足随机分组条件。实际上，双重差分法所要求的“随机分组”，是指结果变量的变动趋势独立于政策冲击，即关于 ΔY^0 满足随机分组条件。需要强调的是，这一识别假设和我们通常所说的随机分组是不同的，一般意义上的随机分组要求处理状态和潜在结果不相关，即 $\mathrm{E}(Y^0|D{=}1)=\mathrm{E}(Y^0|D{=}0)$，显然，该识别假设和双重差分法要求的潜在结果差分意义上的随机分组有区别。假使处理组与控制组满足随机分组原则，那么便近似于随机对照试验（randomized controlled trial，RCT），处理组与控制组的结果对比便是处理效应，无需再使用双重差分法。

这里需要说明一个问题：双重差分法作为一种计量模型，其本身解决内生性问题吗？答案应该是否定的。事实上，双重差分法是一个估计量，更是一种研究设计。作为估计量的双重差分法，估计的是处理组和控制组的结果变量在干预前的组间均值差异和干预后的组间均值差异，即差异之差异。然而，这个估计量是否能够正确识别我们关心的因果效应，取决于识别假设式（2）是否成立。更为严谨的说法是，在满足识别假设的前提下双重差分法能够正确识别因果效应，而式（2）经过简单的变形可以发现，它实际上就是双重差分环境下的外生性假设 $\mathrm{E}\left(\Delta\varepsilon|D\right)=0$。所以，作为估计量的双重差分本身并没有解决内生性问题，而是“假设”不存在内生性问题。而作为一种研究设计，双重差分法可以追溯至 19 世纪中期物理学家 John Snow 对伦敦霍乱成因的研究（Snow，1855），[①]Card 和 Krueger（1994）关于最低工资的早期研究也采用类似的设计思想。[②]如果没有研究设计的“双重比对”的想法，是不会产生双重差分法这一估计量的。事实上，是在有了双重对比的研究设计后，我们使用双重差分这一估计量来捕捉所关心的具体的因果效应。然而，当下一些使用双重差分法的实证研究将估计量与研究设计二者等同起来，似乎有了这个估计量，就自然而然有了对应的研究设计，就可以直接避开内生性问题，这是不正确的。双重差分法解决内生性问题，本质上仍然依赖于干预或政策冲击本身的外生性。

从处理组前后两期结果的变化中减去控制组的两期结果的变化，其实质是去除共同趋势的影响，从而得到“干净”的政策效果。需要注意的是，严格来说，共同趋势假设是无法被完全检验的。文章中的做法通常是检验处理组和控制组的事前平行趋势，然而，冲击发生前变化平行并不能保证今后依然平行。倘若政策冲击并不随机，而是会被某因素 X 所影响，那么 X 在决定干预是否发生的同时，也很有可能会影响共同趋势的变化。因此，尽管双重差分法不要求处理组与控制组在各方面相似，但如果一些与结果变量相关的预处理特征在处理组和控制组之间不平衡，那么研究对象很有可能不满足共同趋势假设。通常我们仍然希望处理组和控制组之间较为相似，此时可以去检验关键控制变量的差异，或者尝试与匹配方法相结合等。其中，匹配方法可作为非参数估计手段，也可以作为一种数据预处理手段。双重差分法本身近似于一种差分意义上的匹配方法。[③]倘若处理组和

① Snow（1855）比较了由不同水厂供水的地区的霍乱死亡率的差异。Southwark & Vauxhal 水厂和 Lambeth 水厂原本均从同一卫生条件较差的地区汲水以供家庭使用，但后来，Lambeth 水厂改为从另一卫生条件较好的地区汲水。之后，相比 Southwark & Vauxhal 水厂供水的地区，Lambeth 水厂供水的地区霍乱死亡率大幅降低。据此，Snow 指出霍乱可能是通过受污染的水传播，而非当时普遍认为的由糟糕的空气传播。

② 最低工资问题与双重差分法的“缘份”并未结束。Cengiz 等（2019）结合聚束（bunching）的思想与双重差分法再议了最低工资这一经典题目，感兴趣的读者可进一步阅读其论文。

③ Heckman et al.（1997, 1998）较早地讨论了双重差分意义下的匹配方法。

控制组之间存在明显差异，那么通常要选取不同的控制组来进行稳健性检验。此外，如果处理组和控制组在处理前后存在成分变化（compositional changes），这意味着政策可能具有很强的内生性，通常难以满足共同趋势假设，在该种情况下，要慎重使用双重差分法。

双重差分法中除去政策、时间等两个维度的变量外，还可以再加入其他变量进行控制，即在模型中加入控制变量W_{it}。然而，在实际回归操作中，具体应当加入什么控制变量、哪些变量不能被控制、是否要加入线性趋势等问题需要格外留意。本文将在第三部分对此展开讨论。

2. 单位处理变量值稳定假设（SUTVA）。单位处理变量值稳定假设（stable unit treatment values assumption，SUTVA）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。直观理解，不满足 SUTVA 意味着控制组个体也受到了干预政策的影响，因而不再是事实上未受干预影响的“真实”控制组，也就无法使用控制组时间趋势来构建处理组时间趋势的反事实。在理想情况下，处理组和控制组被严格区分开来，彼此互不干涉，然而，在现实生活中，相当多的政策冲击具有一定的外部性，例如加强上游省份水污染企业的环境督查也会有利于改善下游省份水质。此外，个体的行为也往往具有一定的策略性和选择性，如处理组地区得到了较好的政策帮扶，那么原本控制组地区的个体可能会自发从控制组地区迁移至处理组地区，意味着宏观上非政策目标地区也受到了干预政策的影响，这就是通常所说的一般均衡效应（general equilibrium effect）或溢出效应（spillover effect）。一般均衡效应或溢出效应会使得 SUTVA 不再成立，进而导致双重差分法无法正确识别因果效应。Butts（2021）在 Callaway 和 Sant’Anna（2020）研究基础上采用事件分析法对这类溢出情况进行了处理。

三、双重差分方法中需要注意的具体问题

（一）控制变量

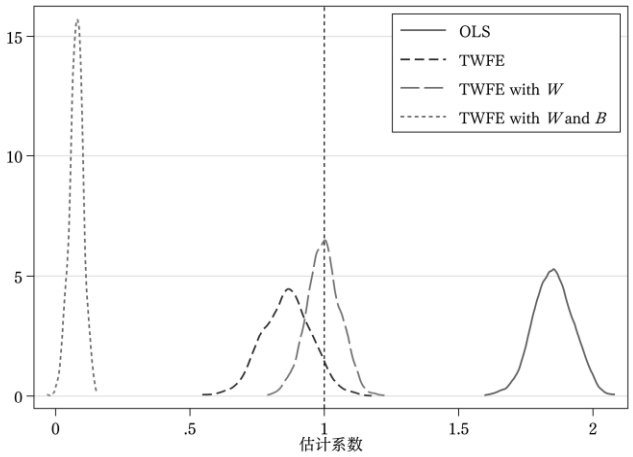
在回归方程中加入控制变量起到两个作用。第一，保证条件独立假设（conditional independence assumption，CIA）成立。[①]条件独立假设成立意味着给定控制变量时处理变量D_i与误差项ε_{it}不相关，从而保证了 OLS 估计量b是我们所关心的因果效应β的一致估计。这是观测性研究的因果推断中控制变量所发挥的最核心作用。第二，减小误差，提高估计精度。如果处理变量D_i与误差项ε_{it}已经不相关，无论是否加入控制变量，b都是因果效应β的一致估计。此时加入合理的控制变量可以降低误差从而提高估计精度。

Cinelli 等（2021）将控制变量分为三类。第一类控制变量是为了保证 CIA 成立而控制的变量（称为好控制变量，good control），必须在回归方程中加以控制。由于这类变量既影响Y_{it}又影响D_i，不控制这类变量会导致明显的“遗漏变量”问题，从而使得 OLS 估计系数b不是因果效应β的一致估计，这是观测性实证研究面临的最大挑战。以常用的面板数据为例，首先，通常个体固定效应和时间固定效应必须加以控制，其次是既影响Y_{it}又影响D_i的可观测变量$\mathbf{X}_{it}$。不过这里需要强调的是，发生在处理时点之后（$t \geqslant T_D$）的$\mathbf{X}_{it}$作为事后变量，很有可能是一个“坏”控制变量（见下文），对其加以控制会导致估计系数b不一致。为了避免这类问题，一般的做法是控制事前某一

① 本文主要从实证方法应用的角度进行阐述和讨论。一些最新的文献对于包含控制变量的双重差分法背后对应的假想实验进行了更为深入的理论分析和讨论，例如区分了严格外生性（strict exogeneity）和序列可忽略性（sequential ignorability）两种不同的识别假设。感兴趣的读者可进一步阅读 Xu（2021）等相关文献。

期的前定变量 $\mathbf{X}_{i,T_D-k}$（$k \geqslant 1$）与时间趋势 $f(t)$ 的交互项 $\mathbf{X}_{i,T_D-k} \times f(t)$，本质上是控制处理前特征不同的个体间可能存在的时间趋势差异。文献中通常取期初变量 $\mathbf{X}_{i,0}$ 或干预前一期变量 $\mathbf{X}_{i,T_D-1}$。时间趋势 $f(t)$ 可以设定为线性时间趋势 $f(t)=t$，更一般的做法是时间固定效应 $f(t)=\{f_1,f_2,\ldots,f_T\}$，后者可以控制更为灵活的时间趋势形式，因而在实践中更为常用。

第二类控制变量是可能导致 CIA 不成立的变量（称为坏控制变量，bad control），必须排除在回归方程之外。受到 D_i 影响的结果变量一般都是坏控制变量，加入回归方程会使得估计系数 b 不再具有因果解释力。坏控制变量问题可能对因果效应的估计产生极大的影响，图 2 是一个模拟估计的例子：添加了合理控制变量的双向固定效应模型能够很好地估计真实因果效应，然而一旦继续加入坏控制变量，估计系数会产生极大的偏误。判断控制变量是否合理的一个经验法则是考虑控制变量的决定时间：在处理时点之后产生变化的变量都可能受到 D_i 的影响，很可能是坏控制变量。[①]在过去相当长一段时期内有一种看法认为“凡是与 Y_{it} 和 D_i 相关的变量均应该作为控制变量纳入回归方程”，这种看法忽略了坏控制变量的存在。对控制变量的选择直接决定了实证研究的可信性，需要研究者更加谨慎地对待。[②]



注：真实系数等于 1。上图分别使用最小二乘法（OLS）、双向固定效应模型（TWFE）、加入好控制变量 W 的双向固定效应模型以及同时加入好控制变量 W 和坏控制变量 B 的双向固定效应模型进行了 100 次模拟。

图 2　坏控制变量对估计结果的影响

第三类控制变量是不影响 CIA 是否成立的变量（称为中性控制变量，neutral control），在回归方程中可加可不加。从因果效应识别的角度而言，这类变量是否加入回归方程并不影响对因果效应估计的一致性，控制或不控制均可。从统计推断的角度来看，合理地控制这类变量有助于减小残差从而提高估计精度，但是与坏控制变量问题类似，选取不当的中性控制变量反而会使得估计偏误增加。判断中性控制变量是否应该控制的一个经验法则是：影响被解释变量 Y_{it} 的中性控制变量可以加入回归方程中以减小误差，提高估计精度；影响 D_i 的中性控制变量一般不控制，因为若控制则

① 当然，作为经验法则而言这种判断标准有些过于严苛了，在一些极少见的特殊情况下控制事后变量可能有助于纠正选择性偏误（Cinelli，2021），不过绝大多数情形下控制事后变量都需要非常谨慎。

② 作者的建议是对控制变量采取较为保守的策略，除非有很强的理由认为某变量会导致严重的选择性偏误而必须加以控制，否则一般不加入回归方程。即使是必须控制的变量，也要注意选择干预时点之前的前定变量。

会减小 $D_i \times T_t$ 的变动性（variation），降低估计精度。

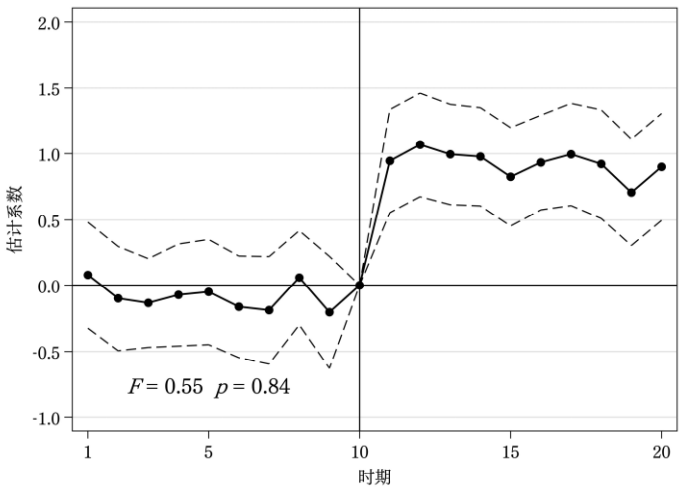
（二）平行趋势与事前趋势检验

平行趋势（parallel trend）又称共同趋势（common trend），指处理组个体的 Y_{it} 在没有接受处理的状态下拥有和控制组个体 Y_{it} 相同的时间变动趋势，它是双重差分法能够正确识别因果效应的前提条件。由于处理组个体在处理时点后的反事实结果（处理组没有接受处理的 Y_{it} ）无法观察到，平行趋势假设本质上是无法直接检验的。因此，研究者通常退而求其次，通过检验可观察的处理组和控制组事前趋势是否相同来间接地检验平行趋势假设。如果处理组和控制组的事前趋势平行，那么研究者就有一定的信心认为事后趋势也是平行的。

对于一般的双重差分法（处理时点相同），一般通过如下方程对事前平行趋势进行检验：

$$Y_{it} = \alpha + \sum_{s=1}^{T_D-2} \beta_s^{pre} \left(D_i \times T_t^s \right) + \sum_{s=T_D}^{T} \beta_s^{post} \left(D_i \times T_t^s \right) + \theta W_{it} + \mu_i + \gamma_t + \varepsilon_{it} \qquad (3)$$

式（3）中的 D_i 是分组变量，T_t^s 是第 s 期的时间虚拟变量，β_s^{pre} 和 β_s^{post} 可以直观地理解为在处理发生前和处理发生后的第 s 期处理组和控制组被解释变量 Y_{it} 的差异相对于基期（这里是处理发生前一期）处理组和控制组被解释变量 Y_{it} 的差异。[①②]事前平行趋势满足意味着在处理时点 T_D 之前的各个时期组间差异没有发生明显变化，因此可以通过检验 β_s^{pre} 是否显著异于 0 来间接地检验事前平行趋势是否成立。图 3 是一个模拟的例子，可以看到在处理发生前各个时期的 β_s^{pre} 均不显著，联合检验结果也无法拒绝处理前系数都为 0 的原假设，因此可以认为事前平行趋势得到了满足。



注：模拟使用的数据与图 2 相同，虚线为 95%置信区间，下同。处理发生在第 11 期，这里以处理前一期（10 期）作为基期。F 值对应的原假设为基期之前的 9 个系数联合不等于 0。

图 3　平行趋势检验

式（3）不仅能够检验事前平行趋势，还能够观察到处理效应的动态变化。注意，β_s^{post} 代表了处理时点 T_D 之后的各个时期组间差异相对于基期的差异，如果处理效应确实存在，我们应该期望得

① 为了避免共线性问题，不能加入全部时间虚拟变量。通常以处理发生前一期（$s=T_D-1$）作为基期。

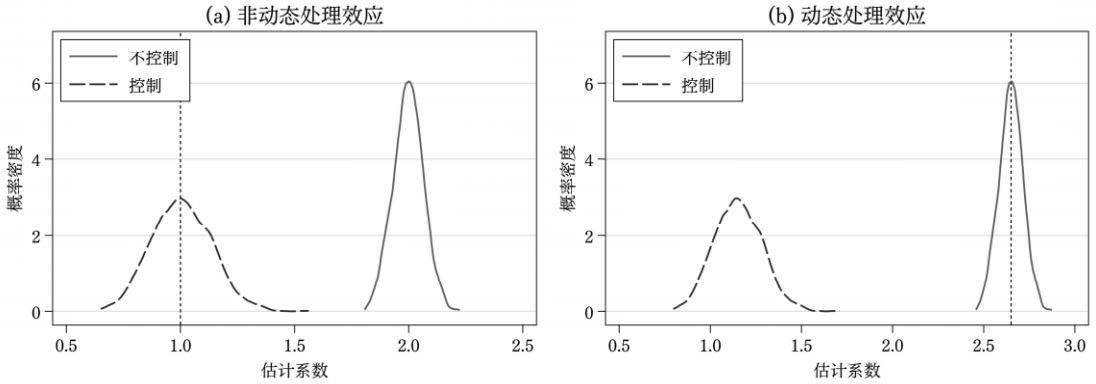
② 即两个时期的差异之差异，这是双重差分法的基本思想。

到 β_s^{post} 显著不为 0。图 3 中从处理后第 1 期（11 期）开始估计系数 β_s^{post} 显著不为 0，并且基本等于真实因果效应 1。因此式（3）实际上发挥着检验事前平行趋势与处理动态效应的双重作用。

需要强调的是，事前平行趋势通过检验并不意味着平行趋势假设一定成立。正如前文强调的，平行趋势假设本身不可检验，而事前平行趋势只是整个平行趋势假设的一部分，即使事前平行趋势通过检验也只是表明处理组和控制组在处理发生前保持相同时间趋势，并不能确保事后趋势也一定平行，所以“事前平行趋势检验通过，平行趋势假设成立”说法并不准确。①

（三）组别时间趋势的进一步分析

使用双重差分法评估政策效应的可靠性依赖于平行趋势假设，因此，在实证研究中最为担心的一点就是干预分配的过程可能使得平行趋势假设不成立。例如研究贫困县政策对经济发展的影响时，由于贫困县依据人均 GDP 等经济指标来认定，被划为贫困县的地区经济发展速度很可能原本就比非贫困县更慢，处理组（贫困县）和控制组（非贫困县）之间的经济发展状况很难满足平行趋势。一个可能的选择是加入组间线性趋势 $D_i \times Trend_t$ 以控制组间线性时间趋势的差异，从而缓解这一问题。②图 4a 给出了数值模拟的证据，当处理组和控制组存在明显的时间趋势差异时，直接使用双重差分法估计出的处理效应存在明显偏误，但控制组间线性时间趋势后就能准确地估计处理效应。事实上，根据上述的分析，在双重差分法中额外地控制住组间线性趋势可以作为一种稳健性检验：若平行趋势假设满足，那么是否加入组间线性时间趋势不会对估计结果产生明显影响；反之，若估计结果发生了明显改变，则预示着组间时间趋势可能存在差异，平行趋势假设可能并不满足。



注：图 a 和 b 分别使用不同的数据生成过程进行了 1000 次模拟，虚线是真实的平均处理效应。图 a 的数据生成过程不存在动态处理效应，真实的平均处理效应为 1，但是组间时间趋势存在差异。图 b 的数据生成过程存在动态处理效应，真实的平均处理效应为 2.65，但组间时间趋势不存在差异。

图 4　组间线性时间趋势与估计结果分布

然而，控制组间时间趋势也是一把双刃剑，可能会产生一些不合意的后果。第一，组间线性时间趋势 $D_i \times Trend_t$ 和双重差分的核心解释变量 $D_i \times Post_t$ 的构造方式相似，因此二者存在比较明显的共线性，控制组间线性时间趋势会大大减少核心解释变量的变动程度从而降低估计效率、提高标

① 值得注意的是，Sun & Abraham（2018）等研究指出当处理效应存在异质性时，平行趋势检验可能存在偏误。Liu 等（2021）提出了一种应对异质性处理效应的稳健作图方法并提供了相应的开源软件包。由于异质性处理效应并非本文讨论重点，故不再展开论述。感兴趣的读者可以进一步阅读相关文献。

② Moser 和 Voena（2012）的研究提供了一个经典的例子。

准误。从图 4a 中可以发现加入线性时间趋势后的估计系数分布明显更加分散，这表明估计量效率降低、标准误变得更大了。第二，如果处理效应不是一次性的，而是随着时间推移逐步显现出来，那么组间线性时间趋势会吸收一部分处理效应，导致双重差分法会低估真实效应。图 4b 的模拟结果说明了这一点：在处理效应存在动态变化时，加入组间线性时间趋势会大大低估真实的处理效应。因此，是否控制组间时间趋势需要研究者结合具体的研究情景仔细斟酌。

从本质上看，组间时间趋势存在差异的根本原因是存在某些可观测或不可观测的前定变量在处理组和控制组之间存在差异或者是存在随时间变化的混淆因素。比如前面提到的贫困县的例子，贫困县和非贫困县的经济发展趋势差异是由当地的初始经济发展水平、地理条件、文化等一系列因素综合造成的。对于可观测的因素，可以通过添加控制变量的方法加以控制，但对于不可观测的因素则一般很难直接处理，通过控制组间线性趋势差异可以部分缓解这一问题，然而当组间时间趋势差异和动态处理效应同时存在时也无法完全解决这一问题。针对这种复杂情况，目前主要有两种处理思路。一种思路是在双重差分的框架下，通过使用未受处理的样本来更为干净地估计和剔除掉时间趋势。[①]另一种思路可能需要超越双重差分法，寻找工具变量或使用空间断点回归设计等方法，不过这些问题超出了本文的范围，这里不再加以讨论。

四、动态双重差分法和事件研究法

（一）交错双重差分法

在标准的双重差分法中处理组在同一个时间点受到干预，然而现实中有相当多的政策并非是一次性全面实施，而是先在某些地区试点后再分批逐步推广，处理时点并不一致。一个典型的例子是增值税转型改革：2004 年 7 月首先在东北地区开始试点，2007 年 7 月扩大至中部 6 省，2008 年 7 月推广至内蒙古以及汶川地震受灾地区，2009 年 1 月 1 日起覆盖全国。标准的双重差分法并不适用于这样的政策。一个常用的方法是交错双重差分法（staggered DiD），“交错”一词表明该方法适用于干预时点有前后差异的政策。交错双重差分法的回归方程设定为如下形式：

$$Y_{it}=\alpha+\beta D_{it}+\theta W_{it}+\mu_i+\gamma_t+\varepsilon_{it} \tag{4}$$

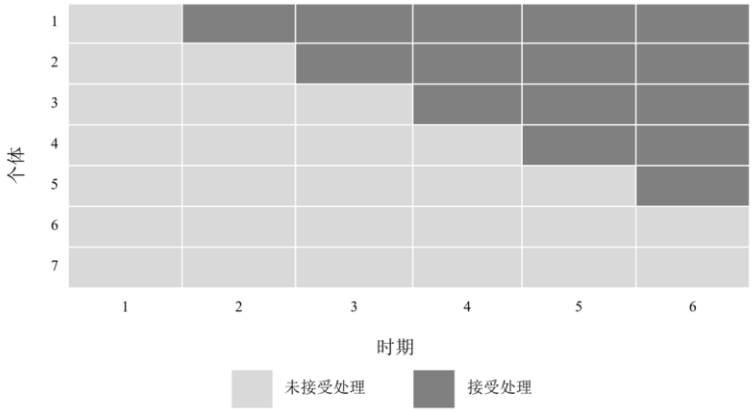
式（4）中的 D_{it} 表示个体 i 在 t 期的处理状态，接受处理时取 1，未接受处理取 0。图 5 是一个典型的干预时点交错发生时的 D_{it} 取值示例。可以发现，标准双重差分法是交错双重差分法的一个特例：当处理组受到干预影响的时点全部相同时，D_{it} 可以分解为 $D_{it}=D_i\times Post_t$，当干预时点不同时则无法做上述分解。

交错双重差分法在政策评估领域得到了广泛的应用，[②]然而最近一些理论计量学者发现交错双重差分法可能存在一些比较严重的问题（Callaway 和 Sant’Anna, 2020；de Chaisemartin 和 d’Haultfœuille, 2020；Goodman-Bacon , 2021）。最主要的问题在于，当政策效应随着时间改变时，交错双重差分法估计的结果［即式（4）的 β］并不是一个定义良好的平均处理效应，而是多个标

① 本质上说，该问题的核心原因是动态处理效应的存在会使得时间固定效应中混杂入一部分处理效应。使用未受处理的样本可以正确地估计时间趋势从而避免偏误。一些研究沿着该思路提出了新的因果效应估计方法，例如 Xu（2017）和 Liu 等（2021）提出的反事实估计量（counterfactual estimatior），以及 Gardner（2021）提出的两阶段双重差分法（two-stage DiD）。

② 例如 Liu 和 Mao（2019）使用交错双重差分法评估了增值税转型改革对企业投资和全要素生产率的影响。

准双重差分法估计的平均处理效应的加权平均，并且权重可能是负的（de Chaisemartin 和 d'Haultfœuille, 2020）。这意味着即使干预本身对所有时点的处理组都是正效应，但交错双重差分法的估计系数仍然可能为负。也就是说，在异质性处理效应的前提下交错双重差分法的单一系数估计结果不再可信。[①]而交错双重差分法的动态效应检验——本文称之为动态双重差分法（dynamic DiD）——则是一种可能应对该种情形的分析工具。



注：接受处理个体在当期受到干预影响，取值为 1；未接受处理个体没有受到影响，取值为 0。
上图使用 stata 中的第三方命令 panelView 制作。

图 5　交错双重差分法的个体处理状态

（二）从动态双重差分法到事件研究法

动态双重差分法可以被视作交错双重差分法的动态效应检验。与标准双重差分法检验动态效应的基本思路一致，也是通过检验处理组和控制组在干预前和干预后的组间均值差异变化来识别政策的动态效应。与标准双重差分法不同的是，在干预时点交错发生的情境下无法定义一个绝对的时间参照点作为处理前和处理后的分界线。因此，动态双重差分法不再以绝对时间为参照系，而是以干预发生时点作为相对时间参照系（图 6）。动态双重差分法的计量方程设定形式为：

$$\begin{aligned} Y_{it} = {} & \alpha + \beta_s^{precut}\left[D_i \times \mathbf{1}\left(t - T_D < \underline{EW}\right)\right] + \sum_{s=\underline{EW}}^{-2} \beta_s^{pre}\left[D_i \times \mathbf{1}\left(t - T_D = s\right)\right] \\ & + \sum_{s=0}^{\overline{EW}} \beta_s^{post}\left[D_i \times \mathbf{1}\left(t - T_D = s\right)\right] + \beta_s^{postcut}\left[D_i \times \mathbf{1}\left(t - T_D > \overline{EW}\right)\right] + u_i + \eta_t + \varepsilon_{it} \end{aligned} \tag{5}$$

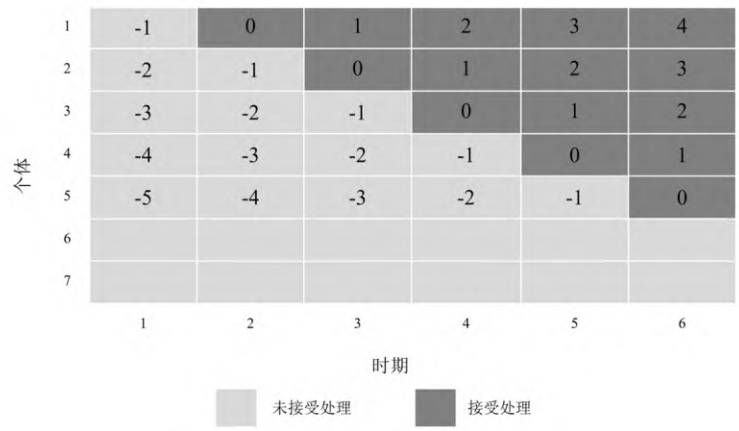
式（5）中的 $\mathbf{1}(\cdot)$ 是示性函数，T_D 是政策发生当期，$\underline{EW}$ 和 $\overline{EW}$ 是事件窗口（event window）的开始期和结束期[②]。对比式（4）和式（5），可以看到动态双重差分法和标准双重差分法动态效应的计量模型设定结构是非常相似的，其差别在于时间坐标系的选择：标准双重差分法以绝对时间为参照系（T_t^s），动态双重差分法以距离干预发生时点的相对时间为参照系（$t - T_D = s$）。因此，虽然计量模型结构有一些区别，但两者的核心思想是基本一致的：比较干预发生前和发生后的处理

① 对于这一问题，已有学者提出了一些初步的解决方案，不过这并非本文讨论的重点。感兴趣的读者可以参考 Callaway 和 Sant'Anna（2020）、de Chaisemartin 和 d'Haultfœuille（2020）、Goodman-Bacon（2021）和 Sun 和 Abraham（2021）等的文献。

② 与前文一致，式（5）也以政策发生前一期为基期。

组和控制组组间差异变化趋势。

不过，研究者可能会感到疑惑的一点是，对于从未接受处理的控制组个体，由于无法定义受到干预影响的时点 T_D，也就无法定义相对于 T_D 的时间坐标，也就是说控制组个体 $\mathbf{1}(t-T_D=s)$ 一项无法定义，控制组个体应该如何取值？对于这一问题，可以尝试从动态效应的核心思想去理解：β_s^{pre} 和 β_s^{post} 估计的是 s 期受到干预影响的个体和未受到干预影响的个体之间的平均差异（相对于基期），受影响的个体取值为 1，从未接受处理的控制组个体自然应该取值为 0。事实上这一问题背后隐藏的更为本质的问题是：动态双重差分法的控制组到底是谁？从图 6 可以清晰地看到，第 s 期的控制组除了包含从未接受处理的控制组个体外，还包括当期未受处理但在未来会受到处理的处理组个体。



注：图中单元内标注数字表示距离干预发生时间。0 表示个体处于接受处理的当期，-1 表示个体处于接受处理的前一期，其他单元类似。没有标注数字的个体是从未接受处理的控制组个体，无法定义相对时间。

图 6　动态双重差分法的个体处理状态和变量赋值

那么，一个自然延伸出的问题是，既然可以使用当期未受处理但在未来会受到处理的处理组个体作为控制组，那么是否可以在没有从未接受处理的控制组样本的情形下使用动态双重差分法？答案是可以，这种情形就是经典的事件研究法（event study）。事实上事件研究法在公司金融、资产定价等领域的应用要远早于双重差分法，早期的代表性文献有 Fama 等（1969）的研究。事件研究法的计量模型设定为

$$\begin{aligned} Y_{it} = {} & \alpha + \beta_s^{precut}\mathbf{1}\left(t-T_D<\underline{EW}\right) + \sum_{s=\underline{EW}}^{-2}\beta_s^{pre}\mathbf{1}\left(t-T_D=s\right) \\ & + \sum_{s=0}^{\overline{EW}}\beta_s^{post}\mathbf{1}\left(t-T_D=s\right) + \beta_s^{postcut}\mathbf{1}\left(t-T_D>\overline{EW}\right) + u_i + \eta_t + \varepsilon_{it} \end{aligned} \tag{6}$$

式（6）中的符号定义与式（5）相同。比较式（5）和式（6）可以发现二者本质上是一致的：如果所有个体都会受到处理（但处理时点不同）、没有从未受到处理的控制组，那么样本中全部观测值的 D_i 都等于 1，式（5）就会变化为式（6）。因此，事件研究法本质上可以近似为去除了控制组的动态双重差分法。图 7 使用了同一组模拟数据分别应用动态双重差分法和事件研究法，可以看到两种方法的系数估计结果几乎完全一致，只不过由于事件研究法剔除了控制组样本使得样本量偏小、估计系数的标准误更大。从计量方法的发展历程看，事件研究法出现的时间要更早，动态双

重差分法是事件研究法在样本包含未接受干预的处理组情形下的自然拓展。

使用动态双重差分法或事件研究法需要注意事件窗口的选择，这里主要指窗口时间宽度的选择。一般来说，干预交错发生的数据结构涉及到的事件窗口宽度要更长一些。比如若数据集包含 10 期的观测值，其中既有第 1 期就接受干预的个体，也有到第 10 期才接受干预的个体，那么该样本涉及到的窗口宽度为干预前 9 期、干预发生当期以及干预发生后 9 期，共 19 期。[①]由于窗口宽度大于样本时间跨度，观测值在干预前后各期的分布是不平衡的，一般而言距离干预时点越远的样本越少。不平衡样本可能带来样本选择偏误（selection bias）和样本消耗（attrition）问题的困扰。选择的事件窗口越宽，样本不平衡现象越严重，会愈发加剧上述担忧。此外，事件窗口越长，越有可能受到同时期发生的其他事件和混杂因素的干扰。如果从时间断点（time cut-off）回归设计的角度理解事件研究法，可以将时间视为驱动变量（running variable），一般来说窗宽选择越宽则样本规模越大、估计越有效（efficient），但可能会有更大的偏误（bias）。总体来看事件窗口的宽度不宜过长。由于事件研究法的估计结果对事件窗口的选择较为敏感，在实际研究中通常需要更换事件窗口宽度来做一些稳健性检验。目前，学界仍在不断完善这一方法，Sun 和 Abraham（2020）、Borusyak 等（2021）的研究围绕事件分析法中的异质性处理等问题进一步进行了拓展与讨论。

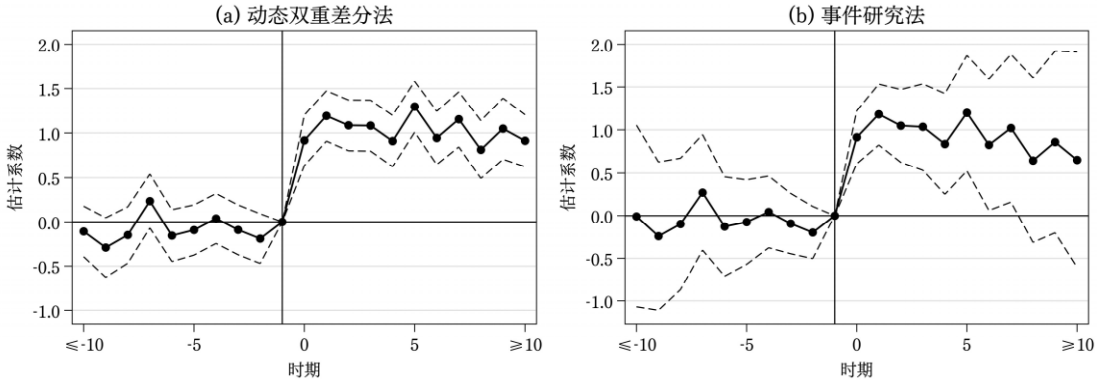


图 7　动态双重差分法和事件研究法估计结果

五、双重差分法研究中的其他问题

至此，本文从标准双重差分法出发，讨论了各类双重差分法的计量模型设定和识别假设，并且详细说明了干预交错发生（staggered adoption）情况下的动态双重差分法和事件研究法的使用方法。上述问题主要集中于方法本身，接下来本文尝试说明使用双重差分法进行实证研究时需要注意的一些重要问题。

（一）制度背景和政策实施真实情况

双重差分法应用最多的场景是评估政策效应。对于制度背景的清晰梳理和政策真实实施情况的正确观察应该是政策评估类实证研究的基石。一项政策可能发布了却没有很好地实施，也可能受政策影响的个体采取了“上有政策，下有对策”的策略式行动影响了政策实施真实效果，如果研究

① 如果是干预发生时间相同的标准双重差分法，涉及到的时点仅为 10 期。更一般地说，对于时间维度为 T 的面板数据，动态双重差分法和事件研究法涉及的窗口宽度最多为 $2T$-1 期。

者没有很好地厘清这些制度背景和政策实施的真实情况，就不可能准确地评估政策效应，甚至可能得到误导性的研究结论。

这里举一个实例。相当多的研究发现地方政府的财政补贴相当低效，企业获得了大量的财政补贴却并没有激励企业的研发创新能力，甚至会引起企业寻租（王红建等，2014；张杰等，2015）。然而，范子英和王倩（2019）通过对地方政府税收征管实务的观察，发现财政补贴实施过程中存在相当明显的“列收列支”问题：地方政府为了增加名义上的税收收入，会先向企业多征收一部分税款，再以财政补贴的名义返还回去。所以，相当一部分名义上为财政补贴的资金实际上是企业自有资金，而这部分“虚假”的财政补贴自然不会对企业经营行为产生影响。因此，财政补贴的低效率很可能是由于对政策实施真实情况的把握不够深入导致的错误结果。总体而言，使用双重差分法评估政策效应要求对政策的具体实施情况有深入、清晰的了解：政策什么时候开始真正实施？政策是否按照要求得到了准确执行？行为主体是否采取了一些应对措施？等等。这一系列问题与双重差分法是否合理、可行程度密切相关，也是进一步深入分析政策机制效应的良好开端。因此，政策评估类的实证研究有必要高度重视制度背景和政策实施情况。

（二）干预政策需要严格外生或随机分配吗？

在第二部分双重差分法的识别假设部分，我们强调了双重差分法本身并没有解决内生性问题，而是“假设”干预政策是外生，内生性问题的解决仍然依赖于干预政策本身的外生性。然而，这里的外生性是什么意义上的外生性？换言之，双重差分法下需要干预政策和谁之间是外生的？一种看法认为干预政策必须是完全随机（自然实验）或者近似随机分配（准自然实验），即干预政策和模型未考虑的所有因素（扰动项）之间不相关，只有在这种情况下才适用双重差分法（陈林和伍海军，2015）。但是，现实中的任何一项政策几乎都有特定的政策目标和政策对象，完全随机分配的政策几乎并不存在，那么这类政策是否完全不适用双重差分法呢？本文认为并非如此。第二部分对识别假设的讨论清楚地表明，双重差分法所需要的外生性是干预政策和扰动项在差分意义上的外生性，这与水平意义上的外生性显然并非是等价的。[①]

我们以贫困县政策的经济发展效应评估为例。水平意义上的外生性要求贫困县名额的分配过程要近似完全随机，无论是贫困地区还是富裕地区都有差不多的机会入选贫困县，显然这并不符合现实——贫困县的选取标准主要是人均 GDP、人均财政收入等指标，被选为贫困县的地区都是经济发展十分落后的县域，因此贫困县政策并不满足水平意义上的外生性。但是，差分意义上的外生性是有可能满足的，即贫困县可能和非贫困县有相同的经济发展趋势。如果研究设计能够尽量满足这一识别假设，就可以使用双重差分法。例如黄志平（2018）的做法是首先使用倾向得分匹配法（PSM）对数据预处理，在非贫困县中尽量选取与贫困县的各方面禀赋条件类似的控制组，从而尽可能地使得平行趋势假设成立（等价于差分意义上的外生性），而后使用双重差分法估计因果效应。

① 变量水平意义上的外生性和差分意义上的外生性不等价隐含着平行趋势是否成立取决于变量构造方式。例如地区的人均收入不满足平行趋势，但取自然对数后可能满足平行趋势。换言之，平行趋势假设是否成立与具体的函数形式有关，这被称为平行趋势假设的尺度依赖（scale-dependent）、感兴趣的读者可进一步阅读 Roth 和 Sant’Anna（2021）的研究文献。

（三）溢出效应

双重差分法的另一个核心识别假设是 SUTVA，即干预不存在一般均衡效应或溢出效应。然而，现实中的各项政策几乎或多或少都会存在一定的一般均衡效应，例如前文提到的上游省份加强水质环境规制会影响下游省份水质的例子。特别是在长期中，当处理组个体的决策发生变化时，控制组个体一定会随之调整自身的行为决策。因此，干预政策是否存在溢出效应是任何一个使用双重差分法的实证研究必须考虑的潜在威胁。

不过，检验溢出效应是否存在并非一项简单的工作，研究者需要根据制度背景仔细识别可能受到溢出效应影响的控制组个体，而后检验溢出效应。Lu 等（2019）研究中国经济开发区对当地经济发展的影响，其对溢出效应的讨论和处理是一个较为成功的范例。他们采取了两种识别策略检验溢出效应，第一种是检验与经济开发区所属村庄邻近的同县其他村庄经济发展是否也得到了提高，第二种是检验经济开发区对经济发展的激励效应是否随着村庄离经济开发区越来越远而减弱。第一种方法的结果表明同县其他村庄的总产出、就业等仅有略微的提高且统计上不显著，第二种方法的结果表明距离经济开发区 2 千米之外的村庄基本上不受经济开发区的影响，两种方法都提供了证据表明经济开发区政策的溢出效应并不显著。

还需要强调的一点是，如果研究重点本身就是政策的溢出效应的话，那么是不适用双重差分法的。例如一些研究试图探讨地区产业政策对企业选择效应和集聚效应的影响：本地拥有更加优惠的产业政策（如税收优惠）会吸引相邻地区的企业迁移到本地区，产生选择效应和集聚效应。这里的选择效应和集聚效应就是溢出效应的一个典型表现：本地区的政策对邻近地区的企业产生了影响，因此该话题显然不适合使用双重差分法。研究者需要注意避免类似的问题。

（四）一般均衡视角下的成本收益分析

双重差分法广泛应用于各类公共政策的评估，如果估计得到了政策效应符合预期，是否就意味着政策达到了初始目标或是政策本身就是有效的呢？不是。一般而言，双重差分法只能评估干预政策对研究者感兴趣的结果变量的影响，但研究者并不清楚政策本身的机会成本有多大，也不清楚政策的净收益到底是多少。评估政策效应整体上是否符合预期或是政策是否有效率，并不能仅根据估计结果就判断政策是否有效，而是需要从更广泛的一般均衡角度，从整体上对政策进行成本收益分析。

Duflo（2001）是在政策效用评估类文献中成功应用成本收益分析的早期经典代表，她研究了印度尼西亚修建学校对当地儿童的长期劳动力市场的影响。根据双重差分法的基准结果，她估计了印度尼西亚政府投资学校建设的成本和对儿童未来的工资收益，发现投资学校建设的内部回报率为 8.8%-12%，远高于当地实际利率，因此投资教育是一个非常高收益的投资项目。[①]Lu 等（2019）对中国经济开发区的政策效应同样进行了成本收益分析，他们根据双重差分法的估计结果计算得到 2006-2008 年间经济开发区为当地居民和企业提高的工资和利润总额约为 1 807 亿元，付出的税收成本则为 558 亿元，净收益高达 1 249 亿元。上述例子都体现了研究者在一般均衡的视角下，从

① Duflo（2001）还强调了成本收益分析依赖的关键假设和因素，她特别强调了该结果依赖于印度尼西亚在 20 世纪 70 年代到 90 年代的高速经济增长。如果经济增长速度下降，投资学校建设的收益率会大幅下降，净收益甚至会变为负值。该结论从另一方面强调了一般均衡视角在成本收益分析时的重要作用，即必须要考虑到政策效应对其他核心因素的变动的影响。

机会成本和政策收益两个角度对政策效果进行完整的评估。研究者在完成双重差分法的估计后，通常需要对政策进行成本收益分析，在此基础上才能更为完整地回答政策是否达到预期目标、是否有效率等问题，并提供合理、可行的政策建议。否则，若研究者过于关注政策的直接效果而忽略了潜在的政策成本，就可能对政策的整体效果产生错误判断，将整体上无效率的政策判定为有效政策，最终导致错误的政策建议。

六、总结性评论

本文结合近年来国内外关于双重差分法的理论和实证研究文献，系统梳理了双重差分法的基本计量设定、识别假设和双重差分法的各个类型变体，着重分析了双重差分法实际应用中面临的控制变量选择、平行趋势检验和组间时间趋势差异等容易混淆或理解不准确的问题。特别是近年来交错双重差分法逐渐得到广泛使用，但最新的一些理论计量研究成果表明交错双重差分法在异质性处理效应下存在着一系列不合意之处，可能导致错误的因果效应估计结果，因此，本文建议研究者可以考虑使用动态双重差分法或事件研究法来替代交错双重差分法作为基准识别策略和实证结果展示方法。本文详细介绍了动态双重差分法和事件研究法的计量实现以及两者的区别和联系，通过数值模拟方法揭示了二者本质上的等价性。本文还强调了实践中使用动态双重差分法和事件研究法时对窗宽选择的重要性。最后，本文从政策评估实证研究的角度提出了研究者在使用双重差分法进行实证研究时需要注意的几个重要问题，包括重视制度背景和政策真实效应的梳理和确认、对于政策干预随机性的准确理解、重视对溢出效应的处理和讨论，以及从一般均衡视角对政策效应的收益和成本进行全面评估等。

近年来使用双重差分法进行的实证研究呈现爆发式增长，近乎泛滥，但若深究其中，许多研究并没有正确地理解双重差分法基本识别假设和需要注意的问题，产生了各式各样的偏差与错误。并且，许多学术期刊的匿名审稿人也出现了这些错误和问题，使得一些匿名审稿人提出没有意义甚至是错误的修改建议，而论文作者多数时候只能将错就错去迎合匿名审稿人，甚至将原本正确的做法被迫修改为错误的做法，可谓是见笑于大方之家。长期来看这种错误会极大阻碍我国经济学研究与国际一流研究接轨的脚步，产生的伤害不可谓不严重。本文试图对上述错误和问题在一定程度上进行归纳、总结、厘清和解决，如果能对未来的研究者提供一些参考，为我国经济学研究进步提供些微助力，本文的目的就完全达到了。

当然，本文的观点均是由作者从自身的理解和实践经验中提取总结而来，作为一家之言，必定有谬误或不足之处，仅为抛砖引玉。期待后续学界同行的进一步研究，促成我国经济学界的共同进步。

参考文献

[1] 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，第 133-148 页。

[2] 范子英、王倩：《财政补贴的低效率之谜：税收超收的视角》[J]. 《中国工业经济》，2019 年第 12 期，第 23-41 页。

[3] 黄志平：《国家级贫困县的设立推动了当地经济发展吗?——基于 PSM-DID 方法的实证研究》[J]. 《中国农村经济》，2018 年第 5 期，第 98-111 页。

[4] 刘瑞明、赵仁杰：《西部大开发：增长驱动还是政策陷阱——基于 PSM-DID 方法的研究》[J]. 《中国工业经济》，2015 年第 6 期，第 32-43 页。

[5] 吕越、陆毅、吴嵩博、王勇：《“一带一路”倡议的对外投资促进效应——基于 2005—2016 年中国企业绿地投资的双重差分检验》[J]. 《经济研究》，2019 年第 9 期，第 187-202 页。

[6] 宋弘、封进、杨婉彧：《社保缴费率下降对企业社保缴费与劳动力雇佣的影响》[J]. 《经济研究》，2021 年第 1 期，第 90-104 页。

[7] 王红建、李青原、邢斐：《金融危机、政府补贴与盈余操纵——来自中国上市公司的经验证据》[J]. 《管理世界》，2014 年第 7 期，第 157-167 页。

[8] 张杰、陈志远、杨连星、新夫：《中国创新补贴政策的绩效评估：理论与证据》[J]. 《经济研究》，2015 年第 10 期，第 4-17+33 页。

[9] Borusyak K., Jaravel X., Spiess J., “Revisiting Event Study Designs: Robust and Efficient Estimation”, *Papers*, 2021.

[10] Butts K., “Difference-in-Differences Estimation with Spatial Spillovers”, *Papers*, 2021.

[11] Callaway B., Sant’Anna P. H. C., “Difference-in-differences with multiple time periods”, *Journal of Econometrics*, 2020(5).

[12] Card D., Krueger A. B., “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”, *American Economic Review*, 1994, 84(4): 772-93.

[13] Cengiz D., Dube A., Lindner A., et al., “The effect of minimum wages on low-wage jobs”, *The Quarterly Journal of Economics*, 2019, 134(3): 1405-1454.

[14] Chen Y., Fan Z., Gu X., et al., “Arrival of young talent: The send-down movement and rural education in China”, *American Economic Review*, 2020, 110(11): 3393-3430.

[15] Cinelli M., Morales G. D. F., Galeazzi A., et al., “The echo chamber effect on social media”, *Proceedings of the National Academy of Sciences*, 2021, 118(9).

[16] De Chaisemartin, C., d’Haultfoeuille X., “Fuzzy differences-in-differences”, The Review of Economic Studies, 2018, 85(2), 999-1028.

[17] De Chaisemartin C., d’Haultfoeuille X., “Two-way fixed effects estimators with heterogeneous treatment effects”, *American Economic Review*, 2020, 110(9): 2964-96.

[18] Duflo E., “Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment”, *American economic review*, 2001, 91(4): 795-813.

[19] Fama E. F., Fisher L., Jensen M, et al., “The adjustment of stock prices to new information”, *International economic review*, 1969, 10(1).

[20] Gardner J., “Two-Stage Difference-in-Differences, *Working paper*, 2021.

[21] Goodman-Bacon A., “Difference-in-differences with variation in treatment timing”, *Journal of Econometrics*, 2021, 225(2): 254-277.

[22] Gruber J., “The incidence of mandated maternity benefits”, *The American economic review*, 1994, 622-641.

[23] Heckman J. J., Ichimura H., Todd P E., “Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme”, *The review of economic studies*, 1997, 64(4): 605-654.

[24] Heckman J. J., Ichimura H., Todd P., "Matching as an econometric evaluation estimator", *The review of economic studies*, 1998, 65(2): 261-294.

[25] Liu L., Wang Y., Xu Y., "A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data", *arXiv preprint arXiv*:2107.00856, 2021.

[26] Liu Y., Mao J., "How Do Tax Incentives Affect Investment and Productivity? Firm-Level Evidence From China", *American Economic Journal: Economic Policy*, 2019,11(3): 261-291.

[27] Lu Y., Wang J., Zhu L., "Place-based policies, creation, and agglomeration economies: Evidence from China's economic zone program", *American Economic Journal: Economic Policy*, 2019, 11(3): 325-60.

[28] Mayzlin D., Dover Y., Chevalier J., "Promotional reviews: An empirical investigation of online review manipulation", *American Economic Review*, 2014, 104(8): 2421-55.

[29] Moser P., Voena A., "Compulsory licensing: Evidence from the trading with the enemy act", *American Economic Review*, 2012, 102(1): 396-427.

[30] Nunn N., Qian N., "The potato's contribution to population and urbanization: evidence from a historical experiment", *The quarterly journal of economics*, 2011, 126(2): 593-650.

[31] Olden A., Møen J., "The triple difference estimator, *NHH Dept. of Business and Management Science Discussion Paper*, 2020.

[32] Rajan R., Zingales L., "Financial development and growth", *American Economic Review*, 1998, 88(3): 559-586.

[33] Roth J., Sant'Anna P H C., "Efficient estimation for staggered rollout designs", *Papers*, 2021.

[34] Snow J., "On the mode of communication of cholera", *John Churchill*, 1855.

[35] Sun L., Abraham S., "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects", *Journal of Econometrics*, 2020.

[36] Xu, Y., "Causal Inference with Time-Series Cross-Sectional Data: A Reflection", *SSRN working paper*, 2021.

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

From Difference-in-Differences to Event Study

WEI HUANG

(*Emory University*)

ZIYAO ZHANG

(*Renmin University of China*)

ANRAN LIU

(*Renmin University of China*)

Abstract: Difference-in-differences (DiD) has been widely used in policy evaluation. However, due to the misunderstanding of DiD's identification hypothesis and other fundamental problems, problems like adding unnecessary

control variables or incorrectly interpreting the parallel-trend test occurred in studies. This paper attempts to systematically summarize the method to clarify relevant issues in the application of DiD. We analyze DiD's identification assumptions and economic implications, summarize DiD's typical setting modes, and discuss problems such as control variable selection, parallel trend test, and inter-group linear time trend control. Given the increasing use of the staggered DiD and its possible errors in recent years, we suggest using the dynamic DiD and event study as the benchmark identification strategy. We interpret the two methods, their relationship, and points for attention in detail. In the end, we emphasize other common problems in empirical research, including the critical role of the factual institutional background, the understanding of policy externalities, the treatment of spillover effects, and the cost-benefit analysis in general equilibrium.

Keywords: causal inference; difference-in-differences; event study; empirical research

JEL Classification: F224.0

执行编辑［刘自敏］