交叠的秘密：经济学研究领域的交叠 DID 导读

许文立

xuweny87@hotmail.com

安徽大学经济学院

内容提要：双重差分(DID)设计是 2021 年诺贝尔经济学奖的主要贡献(自然实验)中最重要的因果识别方法。越来越多的学者应用交叠 DID 的设定解决经济学领域的因果问题，但最新的 DID 计量经济学理论文献表明，在交叠处理情形下使用双向固定效应(TWFE)估计量可能会由于异质性处理效应和负权重问题致使平均处理效应产生偏误，甚至得到与真实效应相反的结果。基于此，本文回顾了传统 DID 设计的基本原理，简要阐述交叠 DID 估计量分解、偏误诊断原理与方法，并比较了最新的异质性处理效应稳健估计量。基于此，以一套模拟数据和两篇已发表的经济学论文（Beck, Levine, and Levkov (2010, Journsl of Finance)；曹清峰(2020，中国工业经济)）为例，提出了在使用交叠 DID 进行因果识别过程中的一些实践研究建议。

关键词：双重差分；时变处理；异质性处理效应；培根分解；稳健估计量

中图分类号：F064.1

文献标识码：A

The Secret behind Staggered: A Primer to Staggered DID in the Field of Empirical Economics

Wenli Xu

Anhui University，China

Simon Fraser University， Canada

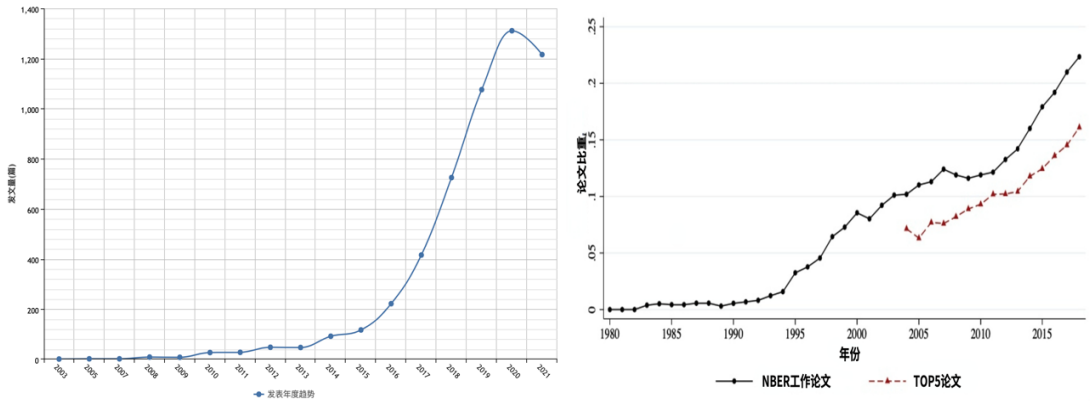
Abstract: Difference in Differences (DID) design is the most important causal identification method in the main contribution of the Nobel Prize in Economics in 2021—natural experiment. And more and more researchers use the setting of staggered DID, but the recent DID econometric literature shows that the use of two-way fixed effects (TWFE) estimator under time-varying treatment may produce bias, or even get the opposite Causal effect. Based on this, this article reviews the traditional DID design, briefly describes the decomposition of staggered DID estimators, bias diagnosic, the latest robust DID estimator, and then uses a simulation data and two published economic papers (Beck, Levine, and Levkov (2010, JF); Cao Qingfeng (2020, China Industrial Economy)) to illustrate some necessary/best elements of staggered DID in practice.

Keywords: Difference in Differences; time-varying treatment; heterogeneous treatment effect; Bacon decomposition; robust estimator

一、引言

“人们无时无刻不在面临着权衡取舍”——这是 Mankiw 经典教材《经济学原理》开篇的经济学十大原理之一。如果人们要做出一个好的决策，他必须要理解这个决策产生的结果/效果。这不仅适用于个体决策，也同样适用于政策制定者。政策实施的效果通常是学界、业界和政策制定者最关心的事情，但政策的效果评估非常具有挑战性（诺贝尔经济学奖委员会，2021）。政策评估中最主流的方法就是寻找到一部分参与这项政策/项目的个体（家庭、企业或地区等，被称为处理组）和一部分没有参与其中的个体（被称为控制组），用政策/项目实施前后处理组的结果差异与控制组的结果差异进行比较，比较的“净差”即为政策的实施效果，这种方法也被称为“双重差分(difference in Differences,DID)法”。该方法主要应用于经验研究领域，在自然实验（Card 等 2021 年诺奖的贡献）设计框架提出之前就已经被研究者广泛地使用，例如，Ashenfelter（1974, 1978）、Ashenfelter 和 Card（1985）用双重差分法评估了美国政府资助的一些就业培训项目，而 Card 和 Krueger（1994）研究最低工资的就业效应的成果可能是利用双重差分法评估政策效应最著名的文献之一。

以“difference-in-differences”为关键词在谷歌学术（Google Scholar）中可以搜索出超过 50000 篇文章。de Chaisemartin 和 D’ Haultfoeuille (2021)统计了美国经济评论（the American Economic Review）2015-2019 年间各年谷歌学术引用率最高的 20 篇论文，共 100 篇，其中有 26 篇使用了 DID 估计量。[1]图（*1*）显示了 2000 年以来中文期刊上以“双重差分”为研究方法的发文数量，以及 NBER 和经济学 top5 英文期刊上以“双重差分”为研究方法的发文比例。目前，中文社科类期刊每年发表的 DID 文献超过 1200 篇，而 NBER 上的 DID 工作论文占比约为 23%，经济学 top5 英文期刊上的 DID 文献占比约为 16%。由此可见，DID 已成为经济学研究领域最常用、最流行的准实验研究设计（Currie et al.，2020）。



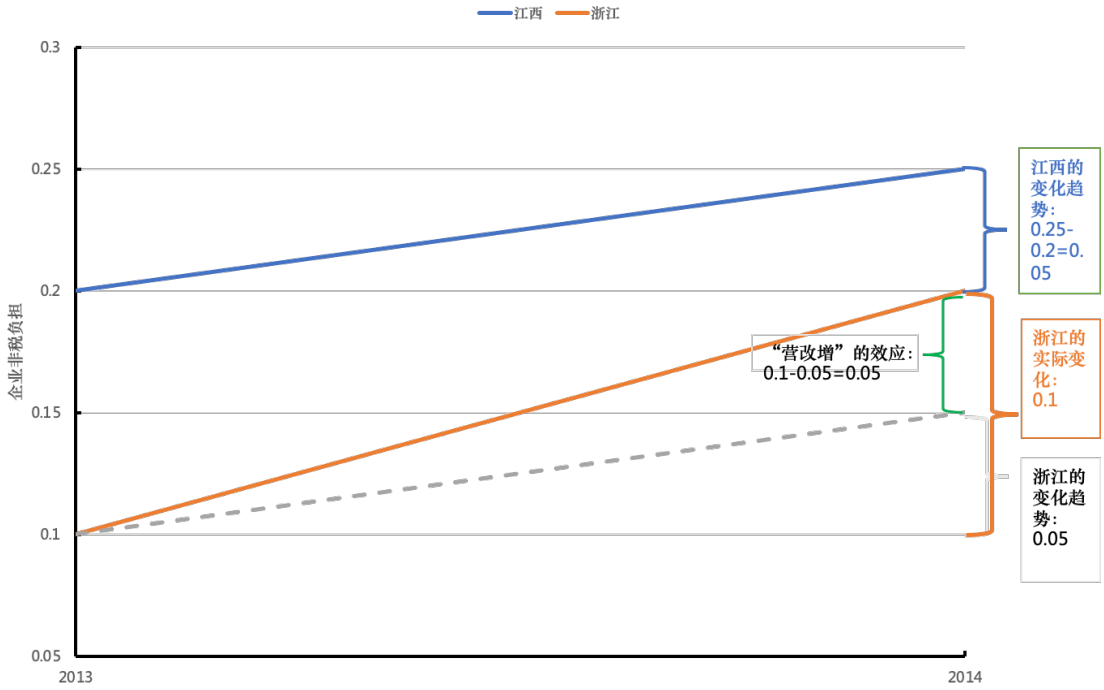
[2]图（*1*） “双重差分”的论文数量与比例变化趋势

二、经典 DID 设计：从 2×2 到 N×T

[1] 长期以来，研究者将双向固定效应等价于 DID 估计量（de Chaisemartin 和 D’ Haultfoeuille，2021）。

[2]图（*1*）左边的数据来自于中国知网，2021 年 11 月 15 日以“双重差分”为关键词搜索到的社科类文献数量；右图来源于 Currie et al.(2020)的图（4）“准实验方法”。

最经典的一种 DID 设计仅仅只包含两个(类)组群和两个时期：一个组群在第二期接受处理——称为“处理组”，另一个组群在两期内都未受到处理——称为“控制组”。经典的例子就是 Card 和 Krueger（1994）研究最低工资的就业效应。他们研究了美国新泽西州最低工资水平上升对就业的影响，以与新泽西州相邻的宾夕法尼亚州东部地区——最低工资水平没有变化——作为控制组，进而比较了最低工资水平变化前后新泽西州就业规模的差异与宾夕法尼亚州东部地区就业规模差异之间的变动，从而识别了最低工资水平变化的就业效应。下面，本文使用“营改增”（例如，彭飞和许文立等，2020）的例子来简要说明经典 2×2 DID 的基本原理。图（*2*）展示了其中的含义。



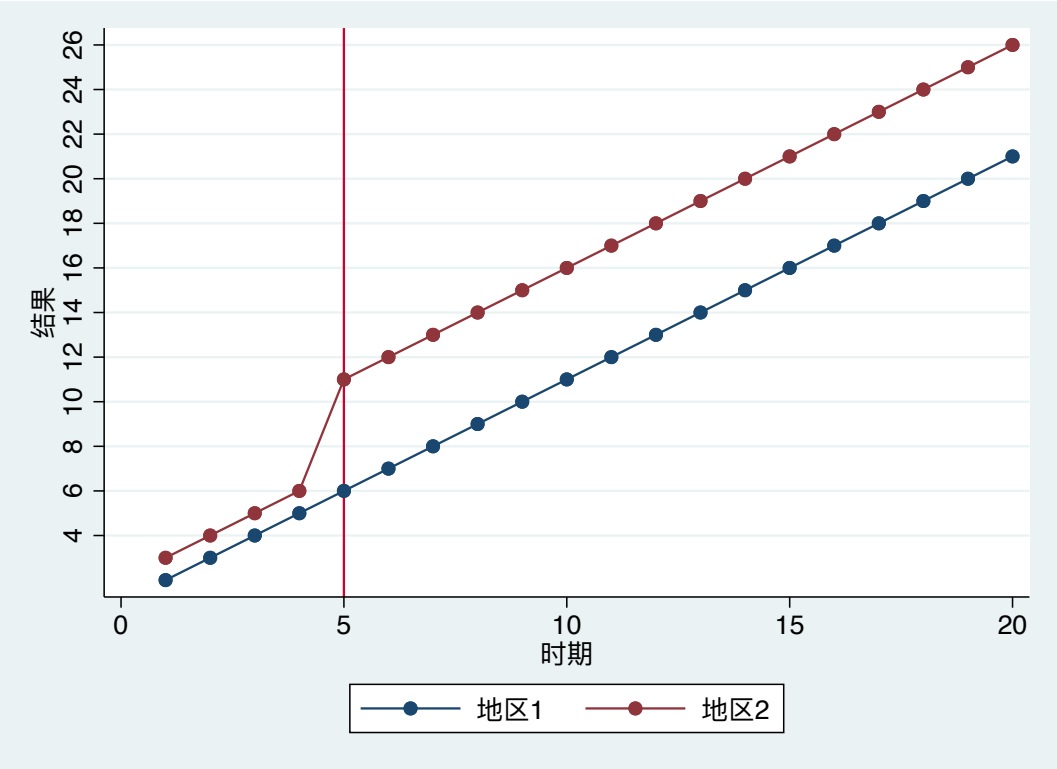
图（*2*） 假设的两地区-两期“营改增”效应

浙江于 2013 年下半年开始“营改增”试点，那么 2013 年为“营改增”前，2014 年为“营改增”后，而江西在 2013 和 2014 年均未进行这项试点工作。Goodman-Bacon（2021）将这种两地区和两期的研究设计称为 2×2 DID。2×2 DID 有一个处理组 z 和一个控制组 j。现假设浙江“营改增”前的平均非税负担为$\bar{Y}_z^{pre}=0.1$，“营改增”后上升到$\bar{Y}_z^{post}=0.2$，而江西“营改增”前的平均非税负担为$\bar{Y}_j^{pre}=0.2$，“营改增”后上升到$\bar{Y}_j^{post}=0.25$。则“营改增”试点对处理组的平均处理效应（ATT）$\hat{\beta}$为：

$$\hat{\beta}=(\bar{Y}_z^{post}-\bar{Y}_z^{pre})-(\bar{Y}_j^{post}-\bar{Y}_j^{pre}) \quad (1)$$

式（1）表明了双重差分的计算过程：首先，计算处理组在政策实施前后的结果差分（$\bar{Y}_z^{post}-\bar{Y}_z^{pre}$=0.1），控制组在政策实施前后的差分（$\bar{Y}_j^{post}-\bar{Y}_j^{pre}$=0.05）；然后，将得到的两个差分再次差分（$(\bar{Y}_z^{post}-\bar{Y}_z^{pre})-(\bar{Y}_j^{post}-\bar{Y}_j^{pre})$=0.1-0.05=0.05）。也就是说，“营改增”企业非税负担的简化 2×2 DID 估计效应为$\hat{\beta}=0.05$。

若将 2×2 DID 在时期数上进行延展，即每一个地区在政策实施前后均有多期（T）数据，就组成了一个 2×T DID 设计。用模拟数据做出这类研究设计的图像，如图（*3*）所示。存在两个地区 1 和 2，每个地区都有 20 期数据，处理发生在第 5 期初（如红色垂直线所示）。



图（*3*） 2×T DID 研究设计

下面，用更加正式的潜在结果的条件期望$E[Y_z|post]$、$E[Y_z|pre]$、$E[Y_j|post]$和$E[Y_j|pre]$来替代处理组和控制组的结果变量样本均值。列出更一般情形下的式（1）：

$$\hat{\beta} = (E[Y_z^1|post] - E[Y_z^0|pre]) - \left(E\left[Y_j^0|post\right] - E\left[Y_j^0|pre\right]\right) \quad (2)$$

式（2）中，上标 1 表示处理后的结果变量，上标 0 表示未处理的结果变量。采用 Cunningham（2020）的转换方法，可以将式（2）变形为：

$$\hat{\beta} = (E[Y_z^1|post] - E[Y_z^0|pre]) - \left(E\left[Y_j^0|post\right] - E\left[Y_j^0|pre\right]\right) + (E[Y_z^0|post] - E[Y_z^0|post]) = \underbrace{(E[Y_z^1|post] - [Y_z^0|post])}_{\text{平均处理效应ATT}} + \underbrace{\left[\left(E\left[Y_j^0|post\right] - E\left[Y_j^0|pre\right]\right) - (E[Y_z^0|post] - E[Y_z^0|pre])\right]}_{\text{平行趋势偏误}} \quad (3)$$

DID 研究设计实质上就是估计处理组的平均处理效应 ATT，而式（3）表明，只有当第二项大括号中内容等于 0 的情况下，才可能干净地识别出 ATT。而第二项内容为 0 有另一个熟知的叫法——平行趋势假设。即，只有假设处理组和对照组满足处理前的平行趋势，才能干净识别处理组的平均处理效应 ATT。此外，平行趋势假设还隐含着另一层含义：确保找到一个好的对照组。

再回到图（*2*）中的模拟数据。可以清晰地看到，处理前地区 1 和地区 2 的结果变量变化趋势相同，且$[Y_i^0|post]-E[Y_i^0|pre]=3$，$E[Y_z^0|post]-E[Y_z^0|pre]=3$，两项之差等于 0，因此可以干净地识别 DID 估计量为 4。

若将上述情形推广至多个（N）组群、多时期（T），只要处理时点相同，都可以得到式（3）的结果，并在满足平行趋势假设下干净地识别处理组的平均处理效应。

三、双向固定效应估计量与交叠 DID

1 双向固定效应回归

在大多数经济学研究中，研究者通常采用更加灵活的回归方法来获得 DID 估计量，例如，2×2 DID 可以当做是一种固定效应估计量，而且很容易在回归里加入组群/个体和时间虚拟变量，以及其它的协变量（Angrist 和 Pischke，2009）。因此，在实践中，大部分学者都会采用式（4）形式的双向固定效应（TWFE）模型来获得 DID 估计量：

$$Y_{i,t}=\alpha_i+\alpha_t+\beta^{DID}D_{i,t}+\epsilon_{i,t} \tag{4}$$

其中，$Y_{i,t}$表示结果变量，α_i表示组群/个体固定效应，α_t表示时间固定效应，$D_{i,t}$表示示性变量，即处理变量——处理期的处理个体的值，在二值虚拟变量的情形下，$D_{i,t}=I_i*T_t$，其中，如果个体接受处理，I_i=1，否则为 0；如果在处理期后，T_t=1，否则为 0。β^{DID}就是研究者感兴趣的平均处理效应 ATT 或者 ATE。基于双向固定效应回归的 DID 有许多优势：（1）为平均处理效应β^{DID}提供了点估计值和标准误；（2）可以加入协变量、时间趋势、动态处理效应，变换处理变量形式等来适用于形式更加一般化、多样化的 DID 研究设计，例如连续型 DID（Nunn 和 Qian，2011）、混合截面 DID（Kiel 和 McClain，1995）、队列 DID（Chen et al.，2020）等。为了更直观、清晰展示回归原理，本文将上述 TWFE 模型写成下列交乘回归形式，同时用表格展现计算过程：

$$Y_{i,t}=\beta_0+\beta_1 I_i+\beta_2 T_t+\beta^{DID}(I_i*T_t)+\epsilon_{i,t} \tag{5}$$

2×2 DID 估计量计算过程的表格形式如下：

表（*1*）2×2 DID 估计量计算过程

	I = 0	I = 1	差分
T = 0	β_0	$\beta_0+\beta_1$	β_1

	I = 0	I = 1	差分
T = 1	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta^{DID}$	$\beta_1 + \beta^{DID}$
差分	β_2	$\beta_2 + \beta^{DID}$	β^{DID}

上述表格右下角的β^{DID}就是两次差分后的结果，也是研究者感兴趣的 TWFE 估计量。

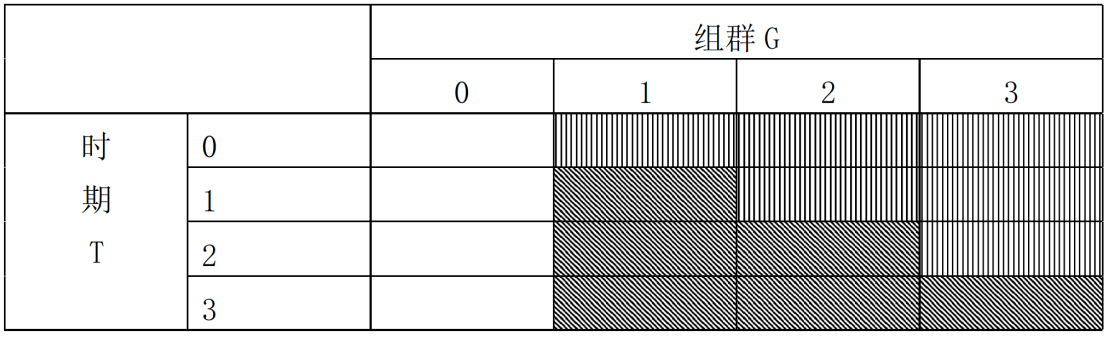
2 交叠 DID

下面，将 2×2 DID 推导到更一般化的多组群-多时期 N×T DID。假设有四个组群 G=0、1、2、3，四个时期 T=0、1、2、3。其中 1-3 号组群会在不同的时间点接受处理，而 0 号组群在全部时期内都不会受到处理。这种时变处理时点的 DID 称为交叠 DID（staggered DID）（Goodman-Bacon，2021）。

从研究设计的视角来看，时变处理时点并不是一个简单问题，研究者只需要将有效的 DID 堆叠在一起就可以得到有效的估计量。而且，相比于同一处理时点 DID 来说，交叠 DID 会带来更多合意的理论性质。在单一处理时点情形下，研究者最关心的是其它因素造成的当期趋势会混淆平均处理效应，即打破平行趋势假设。而交叠 DID 包含多个不同的处理时点，因此可以有效地消除当期趋势带来的混淆处理误差，因此，研究者们认为交叠 DID 是一种更可信、更稳健的研究设计（Baker et al.，2022）。

然而，DID 计量经济理论的最新研究表明，时变处理时点会导致双向固定效应回归不再适用，TWFE 估计量也会产生偏误，甚至出现与事实相反的因果效应，即使随机配置处理仍会出现问题（Sun 和 Abraham, 2020; Borusyak 和 Jaravel, 2021; Callaway 和 Sant'Anna, 2021; Goodman-Bacon, 2021; Imai 和 Kim, 2020; Strezhnev, 2018; Athey 和 Imbens, 2018）。

回到上面的多组群-多时期、不同处理时点的例子。用一个多组群-多时期矩阵图来说明交叠 DID 出现问题的原因，矩阵图如图（*4*）所示。



图（*4*） 多组群-多期、时变处理时点矩阵图。

G=0 号组群在矩阵图中每一期都是空白，表示从来没有处理，G=1-3 号组群的竖条栅格表示还未受到处理的时点，而斜线栅格表示受到处理。例如，G=1 的组群 T=0 期没有接受处理，在 T=1-3 期接受处理，G=2-3 号处理组依次类推。且 G=1-3 号组群开始接受处理的时点分别为 T=1、2、3 期。

图（*4*）中，所有的空白栅格都表示组群从来没有处理，因此可以作为控制组；竖线栅格表示尚未接受处理，可以作为潜在的控制组；斜线栅格则已经接受处理，是处理组。处理

组和控制组有以下类别：（1）从上文的 DID 对照组内容可知，只要不存在溢出效应[1]，研究者总是可以用 G=0 作为一个好的控制组，即所有的处理组和 G=0 控制组构成的 2×2 DID 都可以得到有效的平均处理效应。（2）如果样本中不存在从未处理的对照组，也存在其它的 2×2 DID 配对。例如，（G(1,2),T(0,1)）、（G(1,3),T(0,1)）、（G(2,3),T(1,2)）等等，在这些 2×2 DID 中，G=2、3 的组群在时期 T=0、1 时还未受到处理，因此在 T=0、1 时期，G=2 和 3 可以作为处理组 G=1 的控制组。同理，在 T=1、2 时期，G=3 可以作为 G=2 的控制组。因为这种情形下，控制组还未受到处理，不会受到处理变量的影响。（3）除上述两种情形之外，双向固定效应模型还会通过将已经在其它时点处理过的组群作为控制组来识别平均处理效应。例如，（G(1,2),T(1,2)）、（G(1,2),T(1,3)）、（G(1,3),T(2,3)）、（G(2,3),T(2,3)）等。在 T=1 和 2 两期，G=1 都收到了处理，其结果变化趋势不变，而 T=1 期，G=2 并没有受到处理，T=2 期受到了处理，那么，这个 2×2 DID 中，确实只有 G=2 在 T=2 的处理状态发生了变化。如果处理效应是同质的，这也不是大的问题，但如果处理变量对 G=1 和 G=2 的影响存在异质性，即存在异质性处理效应——不满足平行趋势假设，那么这类 2×2 DID 并不能很好地识别出平均处理效应（Huntington-Klein，2022），这正是交叠 DID 研究设计中 TWFE 估计量的缺陷（Goodman-Bacon，2021）。

2018 年后，许多 DID 的计量经济理论文献都开始关注并解决 TWFE 对异质性处理效应不稳健的问题，例如，Borusyak and Jaravel (2018), Goodman-Bacon (2021), de Chaisemartin and D’Haultfoeuille (2020), and Sun and Abraham (2020)。需要注意的是，并不是所有的 2×2 DID 的 TWFE 估计量都会出现这个问题：在一个 TWFE 回归中，那些处理状态并没有随时间变化的个体被当做了处理状态随时间变化的个体的对照组。在多时期、时变处理时点的情形下，这些对照组可以分为以下三类：

- 新处理个体（处理组）vs 从未处理个体（对照组）（好对照组！）
- 新处理个体（处理组） vs “还未处理”的个体（对照组）（好对照组！）
- 新处理个体（处理组）vs 已经处理过的个体（对照组）（坏对照组！）

前两个对照组是好对照组：用还未处理过的结果趋势来“当作”处理个体在未处理这种反事实情形下的结果趋势，进行 DID 估计。但第三组不是一个好的对照组——结果变量里已经包含了处理效应。传统 DID 研究设计均忽略了对此的解释，直到 Goodman-Bacon（2021）指出，总的 TWFE 估计量是上述三类 DID 处理效应的加权平均值。因此，问题严重性也依赖于第三类 DID 估计量的权重和大小，；如果第三类 2×2 DID 占比较少，那么这个问题也不会太严重，在实践中正好可以作为 TWFE 估计量的一种稳健性检验；相反，如果第三组平均估计量的权重较大，就会影响 TWFE 估计量的结果以及对其因果效应的解释，甚至可能产生非常严重的后果。本来所有个体和所有时间的处理效应为正，但是 TWFE 估计量却得到了一个负的处理效应。甚至在一些情形下，TWFE 估计量出现一些“负权重”，而它们又难以解释，这样就得不到准确的处理效应，也不能较好地解释现实世界决策的因果关系。

3 估计量分解和潜在偏误的诊断

已经有一些文献为交叠 DID 设计的 TWFE 估计量潜在偏误提供了诊断工具，例如，Tymon Słoczyński（2020）、Goodman-Bacon（2021）、Pamela Jakiela（2021）。重新考察一下式（3）：

[1] 空间溢出效应的 DID 设计参见 Clarke, D. (2017)、Butts, K. (2021)。

$$\hat{\beta}^{DID} = \underbrace{(E[Y_z^1|post] - [Y_z^0|post])}_{\text{平均处理效应ATT}} + \underbrace{[(E[Y_j^0|post] - E[Y_j^0|pre]) - (E[Y_z^0|post] - E[Y_z^0|pre])]}_{\text{控制组选择偏误}}$$

（4）

式（4）意味着 2×2 DID 的双向固定效应估计量可以表达成处理组平均处理效应 ATT 与控制组选择偏误之和。如果研究者选择一个好的控制组，即平行趋势满足，那么控制组选择偏误就为 0，TWFE 估计量完美地匹配了真实的处理组平均处理效应 ATT。但是，按照上文的划分的三类控制组，用 pre 表示处理前的时期，mid 表示不同处理时点之前的中间处理时期，用 post 表示处理后的时期，那么，写出各自的处理效应：

类型一：新处理个体（treated，t）vs 从未处理个体（untreated，ut）

$$\hat{\beta}_{t,ut}^{DID} = \underbrace{(E[Y_t^1|post] - [Y_t^0|post])}_{\text{平均处理效应ATT}} + \underbrace{[(E[Y_{ut}^0|post] - E[Y_{ut}^0|pre]) - (E[Y_t^0|post] - E[Y_t^0|pre])]}_{\text{控制组选择偏误}}$$

类型二：新处理个体（early-treated，et） vs “还未处理”的个体（yet-untreated，yut）

$$\hat{\beta}_{et,yut}^{DID} = \underbrace{(E[Y_{et}^1|mid] - [Y_{et}^0|mid])}_{\text{平均处理效应ATT}} + \underbrace{[(E[Y_{yut}^0|post] - E[Y_{yut}^0|pre]) - (E[Y_{et}^0|mid] - E[Y_{et}^0|pre])]}_{\text{控制组选择偏误}}$$

类型三：新处理个体（later-treated，lt）vs 已经处理过的个体（early-treated，et）

$$\hat{\beta}_{lt,et}^{DID} = \underbrace{(E[Y_{lt}^1|post] - [Y_{lt}^0|post])}_{\text{平均处理效应ATT}} + \underbrace{[(E[Y_{lt}^0|post] - E[Y_{lt}^0|mid]) - (E[Y_{et}^0|post] - E[Y_{et}^0|mid])]}_{\text{控制组选择偏误}} + \underbrace{[(E[Y_{et}^1|mid] - E[Y_{et}^0|mid]) - (E[Y_{et}^1|post] - E[Y_{et}^0|post])]}_{\text{时间异质性偏误}}$$

（1）-（2）两种情形下，只要控制组选择偏误为 0，就可以干净识别处理组平均处理效应 ATT。但是在（3）中，即使控制组选择偏误为 0，只要存在时间异质性处理效应，TWFE 估计量也是有偏的。

根据 Frisch－Waugh－Lovell 定理（(Frisch and Waugh, 1933; Lovell, 1963），$\hat{\beta}^{DID}$等于结果变量$Y_{i,t}$对去均值的处理虚拟变量的单变量 OLS 估计系数：

$$\hat{\beta}^{DID} = \frac{\frac{1}{NT}\sum_{it} Y_{i,t}\tilde{D}_{it}}{\frac{1}{NT}\sum_{it} \tilde{D}_{it}^{2}} \quad (5)$$

其中，N 为组群/个体的总个数；T 为时期数。$\tilde{D}_{it} = (D_{it} - D_i) - (D_t - \bar{D})$表示二值处理变量去个体和时间均值，$\bar{D} = \frac{\sum_{it}\tilde{D}_{it}}{NT}$表示所有观测值的均值。

式（5）表明，DID 研究设计的双向固定效应估计量$\hat{\beta}^{DID}$是所有样本结果变量的加权和。且 TWFE 估计量的权重与$\tilde{D}_{it}$符号相同、余值成比例，参见 Goodman-Bacon（2021）、de Chaisemartin and D’Haultfoeuille(2021)、 Jakiela（2021）。当平行趋势满足的时候，处理前个体均值和时间层面的冲击就会被固定效应差分掉，$\hat{\beta}^{DID}$的期望就是所有 2×2 DID 的处理效应的线性组合。$\tilde{D}_{it}$与D_i成反比，即权重与D_i成反比，也就是说，那些平均处理效应最大的组群和平均处理效应最小的组群存在显著差异；而 $\tilde{D}_{it}$与D_i成反比，即权重与D_i成反比，也就是说，那些平均处理效应最大的时期和平均处理效应最小的时期存在显著差异。此外，有一些处理个体可能会存在负的权重，这是因为双向固定效应将一个二值处理变量D_{it}转换成了一个连续型处理强度指标$\tilde{D}_{it}$，而$\tilde{D}_{it}$又未被固定效应所解释。正如在所有的结果变量对连续型处理强度指标的单变量 OLS 回归中，小于平均处理强度的观测样本都会获得负权重，因此这些样本就被当作控制组的一部分。这说明，在双向固定效应模型中，在余值化处理强度均值水平以下的结果才会有负权重。

如果接受处理的个体也获得负权重，那么，它更可能发生在类型三“先处理个体 vs 后处理个体”的情形中。只要从未处理个体数量足够大，且处理前的时期数据足够多就可以保证处理个体不会获得负权重（Pamela Jakiela，2021）。然而，当样本数据的处理前时期有限，或者所有大部分个体都会接受处理时，双向固定效应估计量就会对类型三的平均处理效应施加负权重。

若平均处理效应是同质的情形，双向固定效应模型会由于结果变量和处理变量之间的线性关系而得以正确地声明。此时，OLS 估计会调整来刻画真实处理效应，因此负权重并不是问题。但若平均处理效应存在异质性，尤其在处理个体内随时间变化的样本时，“负权重”会使得双向固定效应估计量产生严重的偏误（de Chaisemartin and D’Haultfoeuille，2020；Goodman-Bacon，2021）。以刘守义、夏璋煦和许文立（2021）的最低工资-消费效应研究为例，A 地区的最低工资提振消费 10%，获得的权重是 0.2，B 地区的最低工资提振消费 8%，而获得的权重是-0.3，那么我们估计的平均处理效应可能是 0.2×10%-0.3×8%=-0.4%。也就是说，无论在 A 地区还是在 B 地区，最低工资原本均可以提振消费，但最后估计得到的总消费效应反而是抑制了 0.4%。

需要注意的是，在二值型 DID 研究设计中，权重也可能都为正。因为$D_i + D_t \leq 1$，进入求和中的 group-time 的$D_{it} = 1$。因此，当没有 group 在大部分时间里被处理，且不存在特定的时期里大部分 group 被处理，这个时候所有的权重就都为正。

4 交叠 DID 的模拟结果

下面使用一套模拟数据作为例子来说明交叠 DID 研究设计的 TWFE 估计量产生偏误的问题。本文模拟了 300 个个体（i）、60 期（t），共 18000 个样本的面板数据。用 Y 表示结果变量，D 表示二值处理变量。所有的个体分为六类，同一类的个体在相同的时点接受处理，但不同类别个体间的处理时点不相同。在 t=5 到 t=55 的整数时期间随机配置每一类个体的处理时点（timing），且每一类的效应规模τ^*也为(5,20)之间的随机整数。在这个例子中，所有的个体都会受到处理，也就是说没有“从未处理”的组群。

结果变量 Y 的数据生成过程（DGP）为：

$$Y_{i,t} = \alpha_i + \alpha_t + \tau D_{i,t} + \epsilon_{i,t}$$

其中，α_i为个体固定效应，数据产生过程为$\alpha_i = i$；α_t为时间固定效应，数据产生过程为$\alpha_t = t$，$\epsilon_{i,t} \sim N(0,1)$表示误差项。$D_{i,t}$表示处理变量，当个体 i 在 t 期接受处理时，$D_{i,t} = 1$，否则为 0。$\tau$表示处理效应参数，其值为：

$$\tau = \begin{cases} \tau^*(t - timing)，\ 如果 D_{i,t} = 1 \\ 0，\ 如果 D_{i,t} = 0 \end{cases}$$

从上述数据产生过程可以看出，无论处理发生在哪个时点，模拟数据的真实处理效应都为正[1]，只是每一类的处理效应大小存在差异。表（2）呈现了双向固定效应模型的回归结果。

表（2） 双向固定效应估计结果

	Y
D	-10.2***
	（2.21）
个体固定效应	是
时间固定效应	是
R^2	0.873
样本量	18000

从上述回归结果可以看出，双向固定效应模型得到的处理效应为-10.2，且在 1%的水平下显著。双向固定效应估计量为负的结果，与在模拟样本数据时人为设置的正处理效应明显不符。而这就是交叠 DID 下，双向固定效应估计量产生的严重偏误。

[1] 模拟数据的 dofile 可以去 github.com/wenddymacro 下载。

Goodman-Bacon（2021）提出了一种分解双向固定效应估计量的方法，利用该方法得到的分解结果如下：

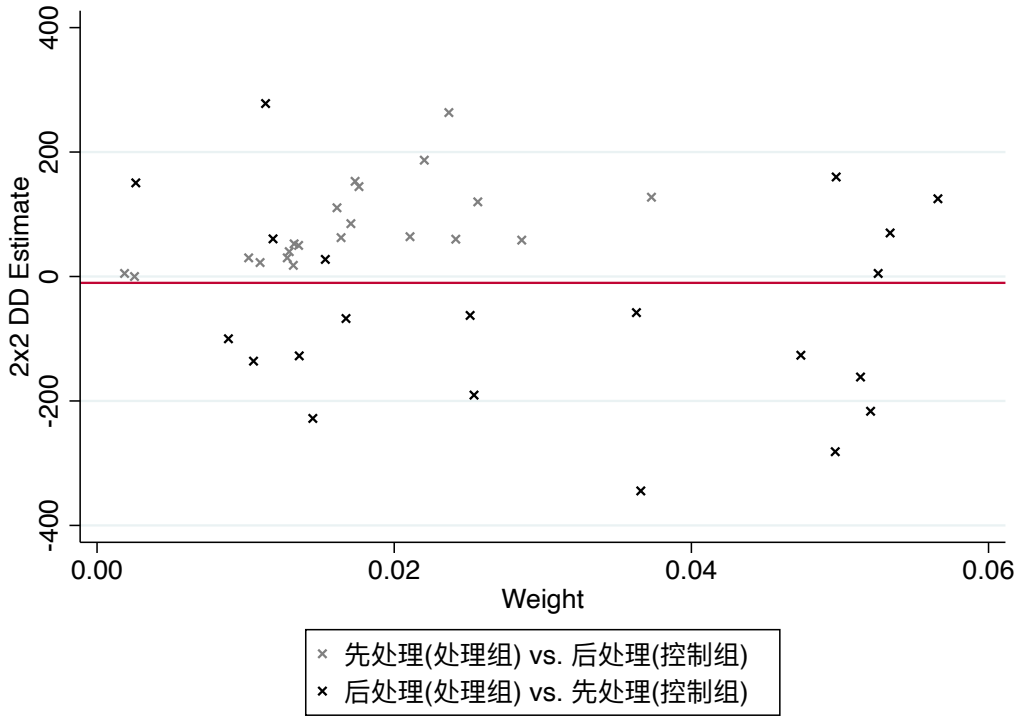
[1]表（3） 培根分解结果

双向固定效应估计量	-10.2	
DID 对比	权重	平均 DID 估计量
先处理组 vs. 后处理组	0.358	98.2
后处理组 vs. 先处理组	0.642	-70.8

从培根分解的结果可以看出（第一行），DID 效应估计量也是-10.2。表中第二行分别展示了 DID 对比、权重、平均 DD 估计量；第三行和四行的第一列分别表示 2*2DID 的分组为“先受到处理的个体 vs 后处理个体还未处理时期”、“后受到处理的个体 vs 先处理个体处理后的时期”，且“vs.”前的为处理组，后未、为对照组。如果我们把权重✖对应的平均估计量：

$$0.358\times 98.2+0.642\times(-70.8)=-10.2$$

也就是说，TWFE 估计量是每个 2*2DID 估计量的加权平均。如前文所述，TWFE 估计量的偏误主要来自于“后受到处理的个体 vs 先处理个体处理后的时期”这一类 2*2 DID，这一类处理组与对照组的平均处理效应为-70.8，占 TWFE 估计量的 64.2%，即平均处理效应中一大半来自于这类 2*2DID。同时，培根分解还提供了一张更为详细的分解图：



图（5） 模拟数据的培根分解

从图（5）可以更清晰地看出，横轴表示每个 2*2DID 的权重，纵轴表示估计量，红色的水平线表示双向固定效应估计量。深色“×”代表的“后受到处理的个体 vs 先处理个体处理后的时期”这一类 2*2 DID 中大部分都落在了负效应区间，而且有 4 对 2*2DID 估计量落

[1] vs.前为处理组，vs.后为控制组

入了-200 以下的区域，且权重较大，这种分布直接导致了 TWFE 估计量最终的平均处理效应为负。

综上所述，多期异质性处理时点的 TWFE 将已经接受处理的组群作为控制组，会使得处理前的平行趋势不再满足。（1）如果效应本身就具有动态性，或者处理效应在组群间变动，那么我们设定的回归不再满足平行趋势。（2）如果处理效应会随着时间而增强，那么“已受处理的控制组”会具有向上变动的趋势，而“刚刚受到处理”的组群则没有这个趋势，因而平行趋势不再成立，识别失败。正如 Sun and Abraham (2020)指出：不同时期的效应会彼此交叠和影响。这也是被称为“交叠 DID”的原因。

四、偏误诊断与稳健估计量

1 偏误诊断

那么，如何诊断 DID 研究设计 TWFE 估计量是否存在偏误呢？

Pamela Jakiela（2021）认为实践中可以做两件事：第一，用处理变量对个体和时间固定效应回归来获得处理变量的余值，进而获得权重，并画出权重的分布图，分析是否所有的处理个体都获得负权重；第二，直接检验同质处理效应假设，即用μ_i表示 t=1 时个体 i 的结果变量$Y_{i,t}$，μ_t表示没有处理时 t-1 和 t 之间结果变量的变化（在平行趋势假设下，个体间是恒定的）。那么，个体 i 在 t 期的结果变量就可以表示成：

$$Y_{i,t} = \mu_i + \sum_{\tau=1}^{t} \mu_t + \delta D_{i,t}$$

其中，δ表示同质处理效应。因此在同质处理效应和平行趋势假设下，余值化结果变量$\tilde{Y}_{i,t}$是$\tilde{D}_{i,t}$的线性函数，其斜率在处理组和控制组之间并无差异。如果回归结果存在差异，说明同质处理效应假设并不成立。

Goodman-Bacon（2021）提出了一种简单、易用，且更为直接地分解 TWFE 估计量和偏误诊断的方法，Cunningham（2020）将该诊断方法称为培根分解（Bacon decomposition）[1]，并将培根分解定理总结为：双向固定效应估计量是所有潜在的 2×2 DID 估计量的加权平均，其权重依赖于组群规模和处理变量的方差。在方差加权共同趋势（VWCT）和时间不变处理效应假设下，方差加权的 ATT 就是所有可能的 2×2 DID 的 ATT 的加权平均。且上文三种类型的 DID 估计量的权重分别为s_t、s_{et}、s_{lt}。这些权重有如下性质：

$$\sum_{t} s_t + \sum_{et \neq ut} \sum_{lt > et} (s_{et} + s_{lt}) = 1$$

[1] de Chaisemartin and D' Haultfoeuille (2020)也提出了一种 TWFE 估计量分解方法。此外，Tymon Słoczyński（2020，REStat forthcoming）也提出了一种分解方法，不过他是更一般化的平均处理效应 OLS 估计量的分解。

$$\hat{\beta}^{DID} = \sum_{t} s_t \hat{\beta}^{DID}_{t,ut} + \sum_{et \neq ut} \sum_{lt > et} \left(s_{et} \hat{\beta}^{DID}_{et,yut} + s_{lt} \hat{\beta}^{DID}_{lt,et} \right)$$

2 稳健估计量

为了修正异质性处理效应情形下双向固定效应估计量的偏误问题，许多学者提出了一些修正的方法，主要有：事件研究设计和一些稳健 DID 估计量。

（1）事件研究设计

事件研究设计广泛应用于 DID 等识别方法中，但通常只作为平行趋势假设的检验方法来使用。自从发现交叠处理时点下双向固定效应估计量存在偏误后，就有学者提出使用事件研究法来纠正估计量的偏误。Goodman-Bacon（2021）指出在一些环境中，可以用面板数据事件研究设计来解决交叠 DID 估计量的偏误问题。当异质性处理效应出现在同一处理个体的不同时期时，面板数据事件研究设计可以较好地应对异质性处理效应带来的估计量偏误。但不同处理时点的个体存在不同“形状”的处理效应时，面板数据事件研究可能不会起作用。

（2）其它一些稳健估计量

所有稳健 DID 估计量都专注于通过避免使用已处理的个体作为对照组，来估计处理组的处理效应。然而要避免使用已处理组作为对照组，就需要引入样本选择，因此，我们必须要解释它。交叠 DID 估计的作者们分别采用不同的策略来实现这一点，其中大多数策略涉及通过基于估计子样本的 ATT 来加权得到整体 ATT，例如组群-时间 ATT 或估计组群处理效应本身。Scott Cunningham（2021）将这些稳健 DID 估计量分为以下类别：

- 加权组群-时间的 ATT
- 通过相对事件时间的平衡来堆叠
- 插补(imputation)方法

Callaway and Sant'Anna (2021)就是采用的加权组群-时间的平均处理效应。Callaway and Sant'Anna (2021)所做的事情就是，估计样本数据中所有“好”的 2×2 DID 组群-时间配对。这个过程非常的耗费时间，因为 2×2 DID 的数量会随着组群数量和时期数量递增。例如，样本数据有 5 个组群（不同时间接受处理）和 10 个时期，那么使用 Callaway and Sant'Anna (2021)方法就要估计 50 个不同的平均处理效应。只要获得所有单个平均处理效应估计量，就可以根据组群、时期、事件类型等等来平均这些平均处理效应以获得最终的交叠 DID 估计量。此外，Sun and Abraham (2020) 估计量与 Callaway and Sant'Anna (2020)类似，这两个估计量是彼此的嵌套的，这个加权平均的过程使它们“感觉”更有可能是“正确”的方法。

但是，使用从不或尚未处理的组群作为对照组来加权 ATT 并不是解决交叠处理的唯一方法。例如，堆叠（Stacking）也是可行的替代方案。堆叠是通过将数据集重组为相对事件时间，而不是日历时间，将时序差分问题重新转换为两个组群的研究设计，从而解决了该问题。这样做是因为两组设计实际上不会遇到交叠处理 TWFE 的问题。一旦数据被重建为相对事件时间的平衡面板，其中处理以相同的“相对处理日期”为中心，然后可以估计传统的 TWFE 模型——控制组群和时间固定效应，以得到处理效应的加权平均值。因此，与堆叠法最相关的文章是 Cengiz et al.（2019）。

第三种方法是一种估算方法，它在一个多步骤过程中估计缺失的反事实，该过程利用平行趋势假设估计未处理组群中的动态效应。这方面的文献有 Borusyak、Jaravel and Spiess（2021）及其插补估计量。尽管在很多方面，Athey et al.（2021）关于使用面板数据完成矩阵的文章也是用了类似的方法，但在技术上，他们并不是 DID 估计量，而是合成控制估计量。

此外，今年的一篇 NBER 工作论文中，Gardner（2021）提出了一个两阶段 DID 估计量，它应该介于上述三类方法之间。从技术上讲，Gardner 确实从一个相同的加权组群时间 ATT 的目标参数开始，将其与 Callaway and Sant'Anna (2021) 等放在一起，但它不会使用双重稳健方法或逆概率权重来估计整体 ATT。相反，正如他所说的那样，两阶段 DID(2sDiD) 最终将是我们最熟悉的双向固定效应 (TWFE) 回归的解决方案的一种扩展，如堆叠。但它也是一个多步骤过程，仅使用控制组来估计拟合值，这使它与 Borusyak、Jaravel and Spiess（2021）的插补方法比较类似。

Borusyak、Jaravel and Spiess（2021）和 Gardner（2021）都采用多步骤来规避“已处理”组群进入控制组的的问题。即：

第一步，用还未处理的观测样本来识别潜在结果（假设没有发生处理效应）：

$$y_{i,t} = \alpha_i + \alpha_t + e_{i,t}$$

第二步，得到个体水平的处理（处理结果下的观测值与未处理结果下的预测值之间的差分）：

$$y_{i,t} = \widehat{\alpha_i} - \widehat{\alpha_t} = ATT_{i,t}$$

第二步要求加总，我们可以按照感兴趣的一些组群来平均所有的$ATT_{i,t}$。此外，Gardner(2021)用 GMM 来估计该模型，而 BJS(2021)用了其它方法来得到矫正的标准误。

五、应用

下面，我们以两篇已经公开发表的论文为例，来说明如何诊断交叠 DID 研究设计中的 TWFE 估计量的偏误，以及如何应对异质性处理效应。交叠 DID 研究设计目前已广泛用于经济、金融、会计、法律、历史等社会科学领域，本文选取了分别发表在金融学英文 top 期刊（Beck, T., R. Levine, and A. Levkov，2010，下文简称“BLL（2010）”)和经济学中文权威期刊（曹清峰，2020，下文简称“曹清峰（2020）”）上的两篇文章。需要说明的是，这两篇文章已经公开发表，无论在研究设计、论证与解释等等方面都非常的完善，本文以它们作为例子，并非表明它们存在重大缺陷，而是为阐明在交叠 DID 设计中使用 TWFE 估计量来推断因果效应可能产生的问题及应对之策。

对于每篇文献，我们首先复现其主要的实证结果。然后利用上文提出的方法来诊断处理效应是否存在异质性，并利用培根分解来说明有偏 2×2 DID 估计量与无偏估计量对总平均处理效应是否有影响，程度几何？再然后，运用最新 DID 计量经济理论文献提出的矫正 TWFE 偏误的估计量来检验文献的实证结果是否稳健。最后对文献实证进行更多稳健性检验，尤其是处理效应异质性检验。

1 BLL（2010）的“金融管制-收入分配的效应”

金融业是最热门的行业，也是收入最高的行业之一。但是，围绕金融机构扩张好坏的争论一直持续了几百年。由美国次贷危机引发的 2008 年全球金融又将金融业及其管制措施推到了风口浪尖。20 世纪 70 年代至 90 年代，美国多数州都取消了对银行分支的限制，这一措施加强了银行业的竞争、提高了银行的运行效率和绩效，研究者们也围绕这一政策带来的一些经济后果进行了探索，包括经济增长、创业、经济波动等。BLL（2010）则从收入分配的视角来评估美国放松银行分支机构管制措施的效应。作者们在这篇论文中研究的问题是银行放松管制对美国收入分配差距的影响，即 20 世纪 70 年代到 90 年代，美国大多数州取消了对州内银行分支机构的限制，这一政策加剧了银行竞争，降低了费用，扩展了低收入群体获得银行信贷的渠道，从而缩小了收入分配差距。

BLL（2010）收集了银行分支管制放松的实施时点，收入分配和其它一些州层面特征的数据来评估银行分支管制放松对收入分配的影响。样本包括 1976-2006 年美国 48 个州和哥伦比亚特区的数据，共计 1519 个观察样本，包含 1859411 个个体，主要为 25-54 岁收入为正的公民。这段时期，各州也解除了跨州银行的分支机构限制，但作者同时控制州内核州际银行分支管制解除时，发现只有州内管制解除是显著的，因此他们只关注于州内管制解除措施，即选择州内解除管制的日期作为国家允许进行并购的时点。有关收入分配的信息来自对美国各地约 6 万个家庭进行的年度调查《Current Population Survey》(CPS)。且收入分配的测量指标有四种方法：（1）基尼系数；（2）泰尔指数；（3）第 90 百分位和第 10 百分位之间自然对数的差异；（4）第 75 百分位和第 25 百分位之间自然对数的差异。此外，还包括一些控制变量。[1]

（1）双向固定效应模型

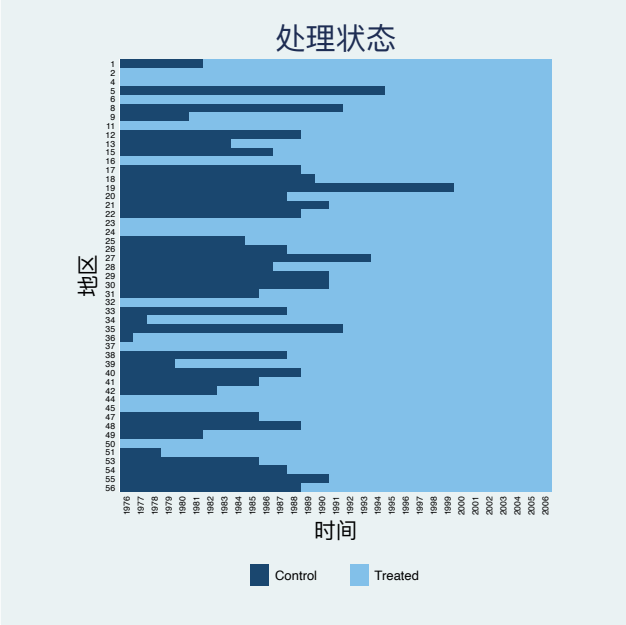
BLL（2010）声明了一个二值型处理变量的双向固定性效应模型来评估银行分支管制解除对收入分配的效应，模型设置如下：

$$Y_{s,t} = \alpha + \beta D_{s,t} + \delta X_{s,t} + A_s + B_t + \epsilon_{s,t}$$

其中，下标 s 表示 48 个州和 1 个特区，t 表示 1976-2006 年。$Y_{s,t}$表示 s 州 t 年的收入分配测量指标。A_s和B_t分别刻画了州和年份固定效应，$X_{s,t}$表示时变的州层面控制变量，$\epsilon_{s,t}$表示误差项。作者感兴趣的变量是二值虚拟变量$D_{s,t}$——州 s 实施了去分支管制后的年份为 1，否则为 0。系数β就表示去分支管制对收入分配的效应。如果β显著为正，意味着去管制对收入分配的不平等程度有正向影响，反之则会降低收入分配不平等。

BLL 还指出，TWFE 的 DID 研究设计允许他们控制一些遗漏变量。例如，包括年份虚拟变量来控制国家层面的冲击和收入分配的变化趋势；包括州虚拟变量允许控制不随时间变化、不可观测的州层面的特征等。样本中 49 个地区的处理状态变化——处理时点如图(*6*)所示。从图中可以看出，放松管制政策在各州实施的时点不相同，且在 1977 年前就有一些州实施了银行分支机构管制的放松政策，也就是说，这些州的处理状态一直为“处理”。且大部分地区放松银行分支机构管制的时点处于整个样本期的前期阶段。这可能会产生很多的“后处理组 vs.先处理组”的 2×2 DID，从而造成 TWFE 的偏误。

[1] 数据来源于 https://dataverse.nl/dataset.xhtml?persistentId=hdl:10411/15996。复制 BLL(2010)结果的 stata 代码下载地址：github.com/wenddymacro



图（*6*） 州内银行分支管制放松政策实施时点图

双向固定效应回归结果如表（*4*）所示。注意，我们复制的结果与 BLL（2010）表（*5*）的原始结果数值稍有差异，具体为 90 分位/10 分位自然对数的结果与标准误的差异[1]。

[1]表（*4*） 解除银行分支管制对收入分配的影响

	基尼系数逻辑斯谛克转换	基尼系数自然对数	泰尔指数自然对数	90/10 分位自然对数	75/25 分位自然对数
Panel A: 无控制变量					
银行分支管制	-0.039***	-0.022***	-0.041**	-0.134**	-0.077***
	{0.013}	{0.008}	{0.016}	{0.058}	{0.019}
R-squared	0.54	0.54	0.56	0.76	0.7
样本量	1519	1519	1519	1519	1519
Panel B: 有控制变量					
银行分支管制	-0.031***	-0.018***	-0.032**	-0.100**	-0.065***
	{0.011}	{0.006}	{0.013}	{0.050}	{0.017}
R-squared	0.57	0.56	0.58	0.77	0.73
样本量	1519	1519	1519	1519	1519

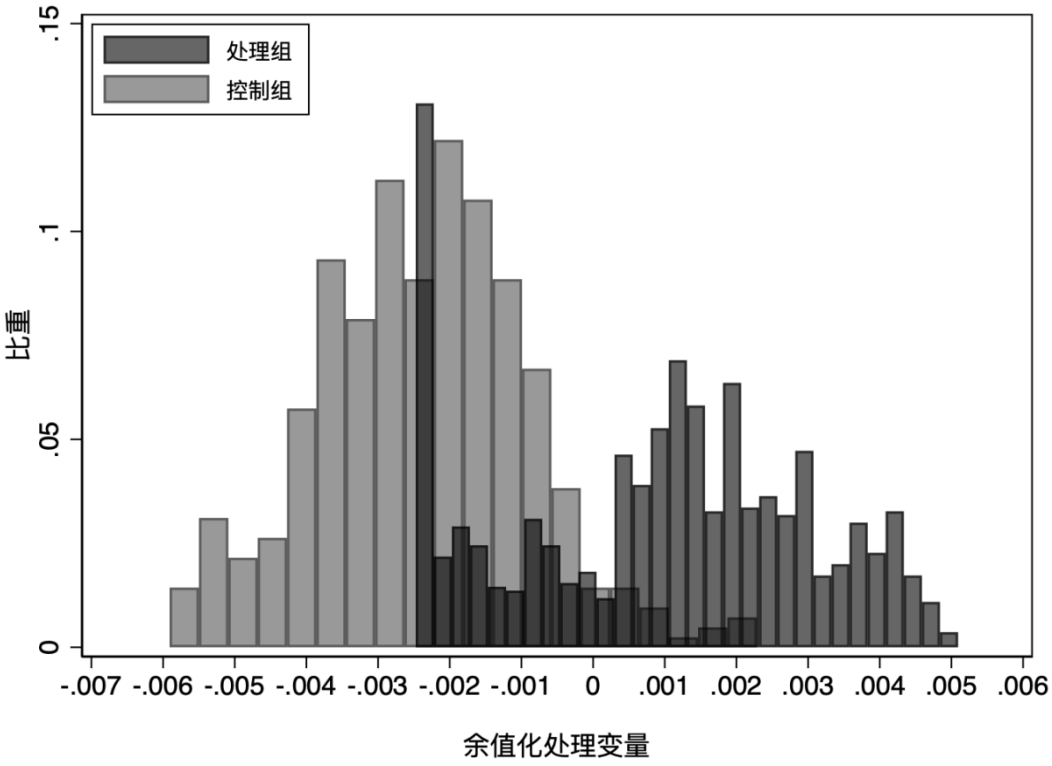
从表（4）的复现结果可以看出，无论是否加入控制变量，银行分支机构管制放松都显著为负，且十分稳健。因此，作者得到的结论是，放松对银行跨洲际分支机构管制会降低收入分配不平等水平。以“基尼系数逻辑斯谛克转换 logit(gini)”为例，放松管制使得

[1]括号中为标准误；* p<0.10, ** p<0.05, *** p<0.01

logit(gini)下降 3.9%，这一数值在经济意义上非常大。作者进一步将回归系数估计值与logit(gini）系数的标准差进行比较，得出放松管制政策解释了收入不平等变化的60%。

（2）偏误诊断

图（7）呈现了计算银行分支管制放松对收入不平等的双向固定效应估计量时地区-时间层面观测值的权重分布直方图。正如上文所述，这些权重与处理变量对地区和时间固定效应回归后的余值成比例关系。图（5）显示，处理组和控制组的权重之和为零，但有一些处理后的地区-时间观测值被赋予了负的权重，而一些控制组的地区-时间观测值则被赋予了正的权重。且在总的处理效应估计过程中，15%左右的处理组具有负的权重。

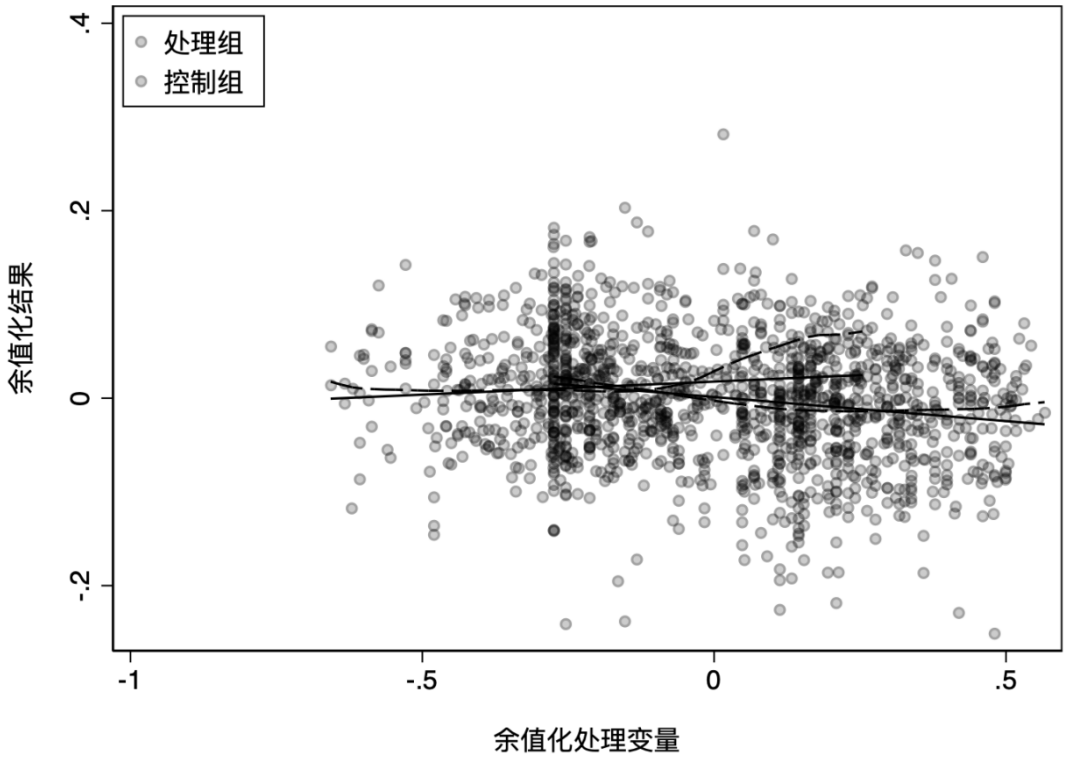


图（7） 负权重分布图

表（5）给出了余值化的Gini 系数与余值化处理变量的OLS 回归。结果表明，余值化结果与余值化处理线性关系的斜率估计量在处理组和控制组之间存在显著的差异。即放松银行分支管制的处理变量与余值化处理变量的交乘项系数为-0.079，且在1%的置信水平下显著。而余值化处理变量的回归系数为0.028，且不显著。图（8）的余值化Gini 系数与余值化处理变量散点图也印证了两者之间并不存在线性关系。这说明，BLL（2010）的双向固定效应并不满足同质性处理效应假设，因此，双向固定效应估计量可能存在偏误。

[1]表（5） 余值化结果与余值化处理变量之间线性关系的检验

	双向固定效应因变量
	logistics_gini
余值化处理变量	0.028
	(0.019)
处理组	-0.016***
	(0.006)
处理组×余值化处理变量	-0.079***
	(0.021)



图（8） 余值化 Gini 系数与余值化处理变量的线性关系

（3）稳健估计量

下面，用面板数据事件研究设计来估计放松银行分支机构管制对地区不平等的效应。最常采用的事件研究设计就是线性动态效应面板数据模型：

$$Y_{s,t} = \sum_{m=-G}^{M} \beta_m D_{s,t-m} + \delta X_{s,t} + A_s + B_t + \epsilon_{s,t}$$

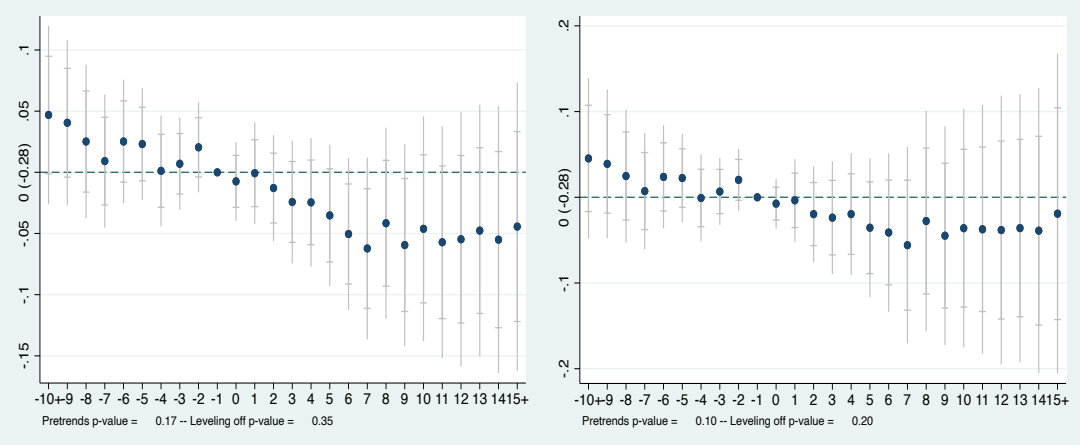
[1] ***表示 1%，**表示 5%，*表示 10*，括号中为标准误

其中，$D_{s,t-m}$表示地区 s 是否在时点 t 前 m 期放松了银行分支机构管制的二值变量。$\sum_{m=-G}^{M}\beta_m D_{s,t-m}$表示放松银行分支机构管制的动态效应，时点 t 的地区不平等最多只能被 t 前的 M 期和 t 后的 G 期的管制放松政策所影响。参数集$\{\beta_m\}_{m=-G}^{M}$包含了这些动态效应的大小。

通常，经济学家更加关心政策的累积效应，即不同时期 k 的$\sum_{m=-G}^{k}\beta_m$，以及政策影响时期外的累积政策效应。因此，采用 Simon Freyaldenhoven et al.(2021)对于面板数据事件研究设计的模型设定与事件研究图的建议。将上述动态处理效应回归模型变形如下：

$$gdpr_{i,t} = \sum_{k=-G-L_G}^{M+L_M-1}\beta_k D_{s,t-k} + \beta_{M+L_M} D_{s,t-M-L_M} + \beta_{-G-L_G-1}(1 - D_{s,t+G+L_G}) + \delta X_{s,t} + A_s + B_t + \epsilon_{s,t}$$

其中，$D_{s,t-k}$表示地区 s 是否在时点 t 前 m 期放松了银行分支机构管制的二值变量，$(1 - D_{s,t+G+L_G})$表示地区 s 在 t 时点后是否仍放松银行分支管制，$D_{s,t-M-L_M}$表示地区 s 在时点 t 前至少$M + L_M$期就放松了银行分支机构管制。



[1]图（9） 事件研究图

事件研究结果如图（9）所示。左图为使用 BLL（2010）全部样本的事件研究图，右图为删掉了 1977 年以前就已经放松银行分支管制地区的样本后的事件研究图。从图（9）（左）可以看出，在放松银行分支机构管制前的时期，放松管制前时间虚拟变量的系数均不显著，且 90%的置信区间下不能拒绝“没有处理前的趋势”这个原假设。这意味着处理组和控制组并没有显著的差异化趋势。且放松管制后的 6、7、9 年地区不平等有所缓解。

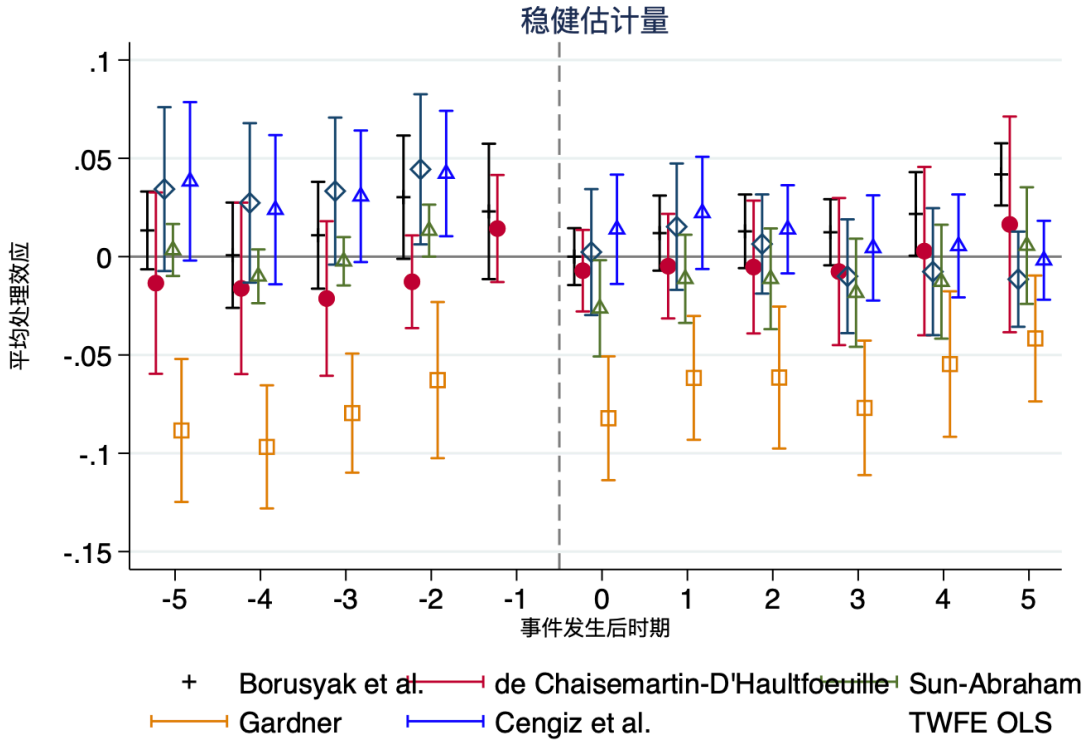
正如上文所述，1977 年前就有一些州放松了银行分支机构管制，这些州在研究样本期内的处理状态一直未变，它们作为“控制组”时，就会使得对应的控制组差分量包含处理效

[1]纵轴表示动态处理效应估计量，纵轴 0 点处的括号和数值表示处理时点前一期结果变量的均值；横轴表示事件时间，且设置初次处理时点为 0。实心圆点表示点估计量，点估计量上下的横杠表示 95%的置信区间，而横杠外的线条表示 95%的均匀置信区间带。而图中左下角的两个 p 值分别表示拒绝两个原假设“没有处理前的趋势”、“所有的动态效应都已经显示”的概率。

应，从而产生偏误。因此去掉这些一直接受处理的州，重新进行事件研究，如图（*9*）（右）所示。虽然，放松管制前的事件研究系数也不显著，但在 90%置信区间下可以拒绝原假设“没有处理前的趋势”，这可能表明平行趋势假设不满足。更为重要的是，放松管制后的动态处理效应在 95%的置信区间下也都不显著。这就意味着，并没有证据显示放松银行分支管制会降低地区不平等。

下面，使用最近几年 DID 计量经济学理论文献提出的一些稳健估计量来估计放松银行分支机构管制对地区不平等效应。这些稳健估计量分别为 Borusyak et al.（2021）、de Chaisemartin 和 D’ Haultfoeuille（2019）、Sun 和 Abraham（2021）、Gardner（2021）和 Cengiz et al.（2019）等，如图（*10*）所示。

虽然，大部分稳健估计量的事件研究图显示，放松管制前的估计系数不显著，平行趋势满足，但放松管制后的估计量也在 95%的置信区间下也不显著，再次表明没有证据显示放松银行分支机构管制会降低地区的不平等。



图（*10*） 稳健估计量

2 曹清峰（2020）的“国家级新区对经济增长的效应”

建党百年来，尤其是改革开放 40 多年来，中国正在实现中华民族的伟大复兴，而国家级新区的设立起着不可替代的作用。曹清峰（2020）研究了国家级新区对区域经济增长的带动作用。他认为，国家级新区的发展历程大致可以划分为三个阶段：第一阶段是在中国国内改革面临诸多不确定性、探索建立中国特色社会主义经济体制的关键节点上，于 1992 年设立了首个国家级新区——上海浦东新区，树立了中国进一步扩大改革开放的一面旗帜；第二个阶段则是在中国特色社会主义市场经济体制初步建立后，为在新形势下特别是加入世界贸易组织后探索改革开放的新经验，于 2006 年设立了第二个国家级新区——天津滨海新区；第三个阶段则是国家级新区的扩容阶段，主要为了应对中国经济进入“新常态”以及改革进入“深水区”后面临的新挑战，国家级新区设立不断加速，于 2010 年后相继设立了重庆两江新区、甘肃兰州新区等一系列国家级新区，基本上覆盖了中国主要经济板块。因此，最直

接、最重要的问题可能就是：国家级新区是否能促进所在地区的经济增长？如果能，效应有多大？

为了实证上述问题，曹清峰（2020）选取了中国 70 个大中城市作为研究样本，时间跨度为 2003-2017 年，且浦东新区早在 1992 年就已经设立，因此，在样本时期内一直属于“处理组”，包含了处理效应，因此，将浦东新区从样本中剔除。其他变量指标还有全市 GDP 实际增长率、全市 GDP、市辖区 GDP、全市第二产业增加值、城市总人口、全市固定资产投资总额、全市全社会商品零售总额、政府财政支出总额、城市出口总额、城市专利授权总量等。这些数据来源于《中国城市统计年鉴》、各省市统计年鉴、中国研究数据服务平台（CNRDS）。

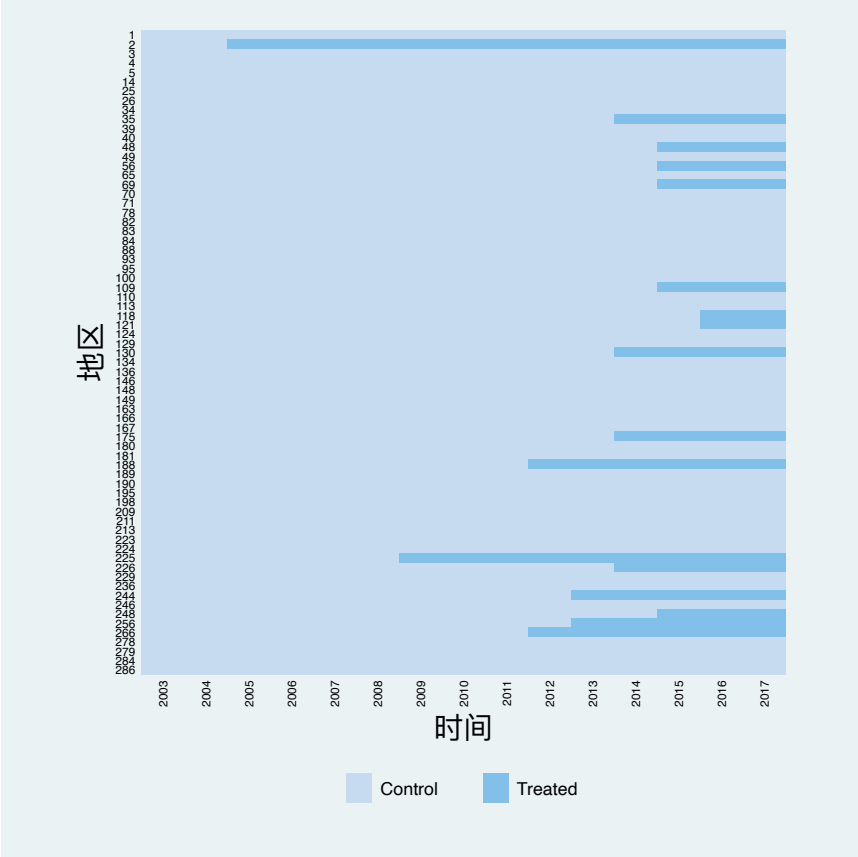
（1）结果复现

曹清峰（2020）使用了双重差分研究设计，回归方程设定为如下双向固定效应模型：

$$gdpr_{i,t} = \beta_0 + \beta_1 did_{i,t} + \lambda Z_{i,t} + v_i + \mu_t + \epsilon_{i,t}$$

其中，$gdpr_{i,t}$为城市经济增长率，即全市 GDP 实际增长率；$did_{i,t}$表示二值型处理变量，即城市设立新区为 1，否则为 0；v_i、μ_t分别表示个体和时间固定效应；$Z_{i,t}$为协变量。国家级新区对区域经济增长的平均处理效应为回归系数β_1。需要说明的是，国家新区设立前，地方政府已经知晓了是否设立新区，已经提前部署、开展相关工作，因此，曹清峰（2020）将处理变量“是否设立新区”的时间提前 1 年，例如，2006 年天津滨海新区获得国务院批复，那么滨海新区 2005 年的 did=1。

图（*11*）呈现了国际级新区设立的时间状态。从图中可以看到，样本期内，最早设立国际级新区的时间为 2006 年，且大部分设立时间位于样本期的后半段。



图（*11*） 国家级新区设立时间

表（*6*）呈现了国家级新区设立对区域经济增长的拉动作用。（1）列表示无控制变量的TWFE 估计量，（2）列表示有控制变量的估计量。表 3 的估计结果显示，二值处理变量 did 的 TWFE 回归系数为 1.16（无协变量）、1.51（有协变量），且在 10%的置信水平下显著。这意味着，国家级新区的设立会促进城市 GDP 实际增长率提升 1.16-1.51 个百分点。这一结果对于不断增大的中国经济下行压力无疑会起到非常巨大的稳增长作用。

[1]表（*6*） 国家级新区对区域经济增长的效应

	（1）	（2）
国际级新区	1.16*	1.51***
	（0.59）	（0.48）
协变量	否	是
R-squared	0.62	0.72
样本量	1053	1035

（2）TWFE 估计量偏误诊断

下面，用 Goodman-bancon（2021）提出的诊断方法来将总的 DID 估计量分解为三组：（1）“先设立国家级新区的城市 vs 后设立国家级新区的城市”；（2）“后设立国家级新区的城市 vs 先设立国家级新区的城市”；（3）“设立国家级新区的城市 vs 从未设立国家级新区的城市”。表（*6*）中得到的总的 DID 估计量等于每一组的平均 DID 估计量乘以各自权重之和。如表（*7*）所示，培根分解给出的无控制变量时总的 DID 估计量与表（*6*）的结果相同，且 1.163=0.057×1.571+0.031×1.659+0.912×1.120。从表（*7*）结果可以进一步看出，“后设立新区的城市 vs 先设立新区的城市”这一类坏对照组的 2×2 DID 估计量所占权重仅为 3.1%，这个比重并不大，且这一类 DD 估计量为 1.659 与 TWFE 的估计量 1.163 相差也不大，因此，这类 2×2 DID 对总的 TWFE 估计量的影响也不大。对 TWFE 估计量影响最大的组别是“设立新区的城市与从未设立新区的城市”，其权重为 91.2%。因此，尽管曹清峰（2020）的研究中也存在“Later T vs Earlier C”这样的坏对照组的影响（所有的交叠 DID 都会存在），但其对总 TWFE 估计量的影响不大。但是要注意的是，它会拉高 TWFE 估计量，即高估国家级新区对区域经济增长的拉动效应。

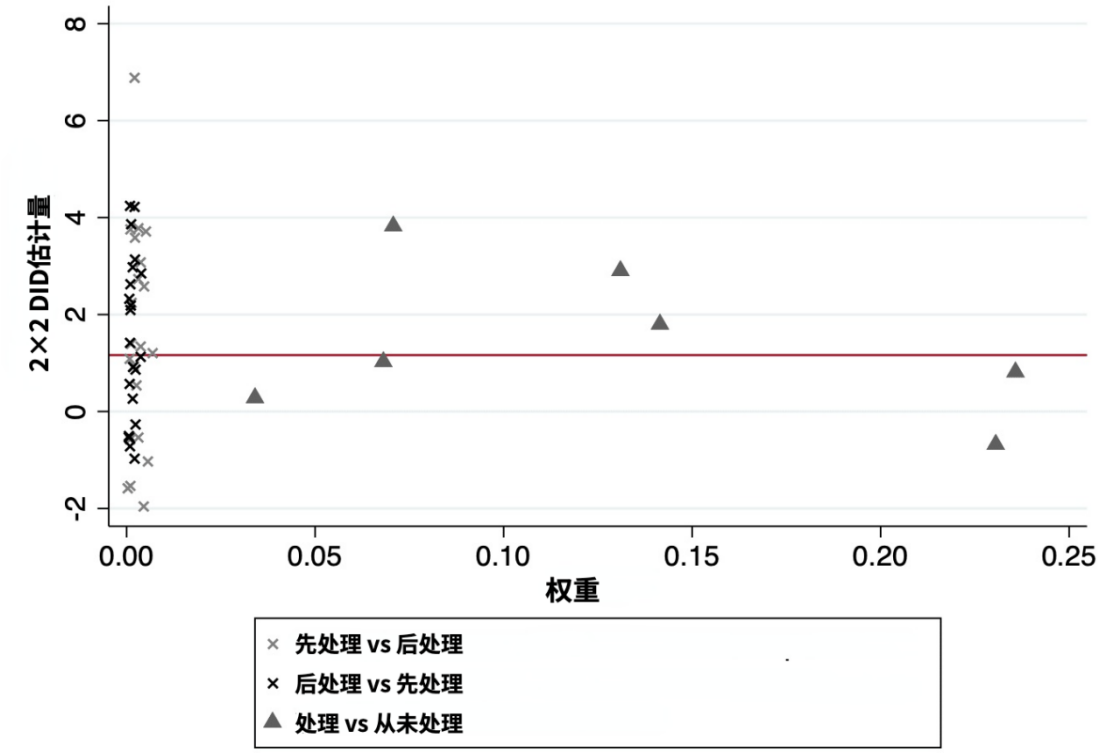
[2]表（*7*） 无控制变量的培根分解

总的 DID 估计量	1.163	
类别	权重	平均 DID 估计量
先处理 vs 后处理	0.057	1.571
后处理 vs 先处理	0.031	1.659
处理 vs 从未处理	0.912	1.120

[1]括号中为标准误；* p<0.10, ** p<0.05, *** p<0.01

[2] vs 前后分别表示处理组和控制组

下面来看看 Bacon 分解图，如图（*12*）所示。图中，每个点都代表这一个 2*2DID。横轴表示权重，纵轴表示单个 DID 估计量。红色水平线表示 TWFE 估计量 1.163。因此，越靠近右边的点就表示其对 TWFE 估计量的影响越大。从图中还可以看出，并非所有的 2×2 DID 的效应都是正，且最右边（权重最大）的两个 2×2 DD 估计量都在红色线条的下方，而且权重都接近四分之一，这两个平均处理效应均拉低了总的 TWFE 估计，尤其是最右下方这个三角点的效应为负，进一步拉低了最终的平均处理效应。



图（*12*） 无控制变量的培根分解

（3）稳健估计量

下面，用面板数据事件研究设计来估计国家级新区设立对城市经济增长的拉动效应。最长采用的事件研究设计就是线性动态效应面板数据模型：

$$gdpr_{i,t} = \sum_{m=-G}^{M} \beta_m did_{i,t-m} + \lambda Z_{i,t} + v_i + \mu_t + \epsilon_{i,t}$$

其中，$did_{i,t-k}$表示城市 i 是否在时点 t 前第 k 期设立国家级新区的二值变量。$\sum_{m=-G}^{M} \beta_m did_{i,t-m}$表示国家级新区设立的动态效应。时点 t 的城市 GDP 实际增长率最多只能被 t 前的 M 期和 t 后的 G 期的政策所影响。参数集$\{\beta_m\}_{m=-G}^{M}$包含了这些动态效应的大小。

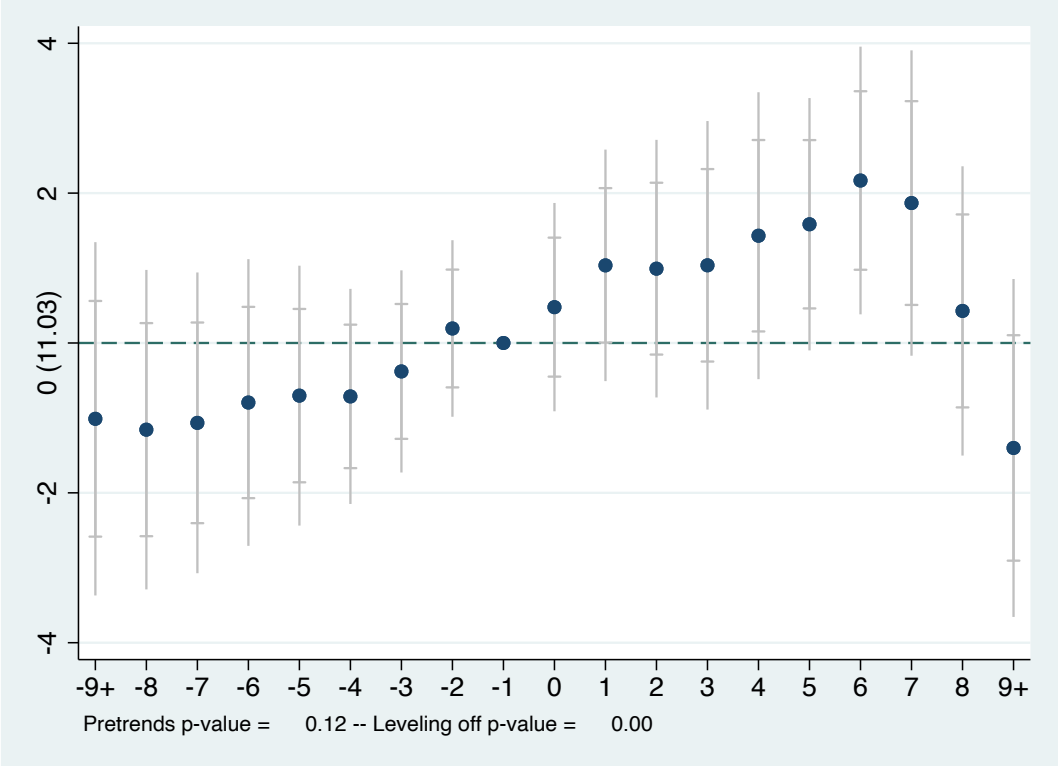
通常，经济学家更加关心政策的累积效应，即不同时期 k 的$\sum_{m=-G}^{k}\beta_m$，以及政策影响时期外的累积政策效应。因此，采用 Simon Freyaldenhoven et al.(2021)对于面板数据事件研究设计的模型设定与事件研究图的建议。将上述动态处理效应回归模型变形如下：

$$gdpr_{i,t} = \sum_{k=-G-L_G}^{M+L_M-1} \beta_k did_{i,t-k} + \beta_{M+L_M} did_{i,t-M-L_M} + \beta_{-G-L_G-1}(1 - did_{i,t+G+L_G}) + \lambda Z_{i,t} + v_i + \mu_t + \epsilon_{i,t}$$

其中，$did_{i,t-k}$表示城市 i 是否在时点 t 前第 k 期设立国家级新区的二值变量，$(1 - did_{i,t+G+L_G})$表示城市 i 在 t 时点后是否仍有国家级新区，$did_{i,t-M-L_M}$表示城市 i 在时点 t 前至少$M + L_M$期就设立了国家级新区。

国家级新区对区域经济增长效应的事件研究结果如图 13 所示。结果显示，国家级新区对城市 GDP 实际增长率有促进作用。在设立国家级新区前的时期，事件相对时间虚拟变量的系数在 95%的置信区间均不显著，这意味着没有证据显示在设立了国家级新区的城市与未设立新区的城市之间存在差异化趋势，这一点也可以从（90%的置信区间下）不能拒绝“没有处理前的趋势”的原假设得到证实。在设立国家级新区后的时期，国家级新区的经济拉动效应立即开始显现，但是在设立新区后的最初 4 年并不显著，直到第 5-8 年才开始显著拉动城市经济增长。

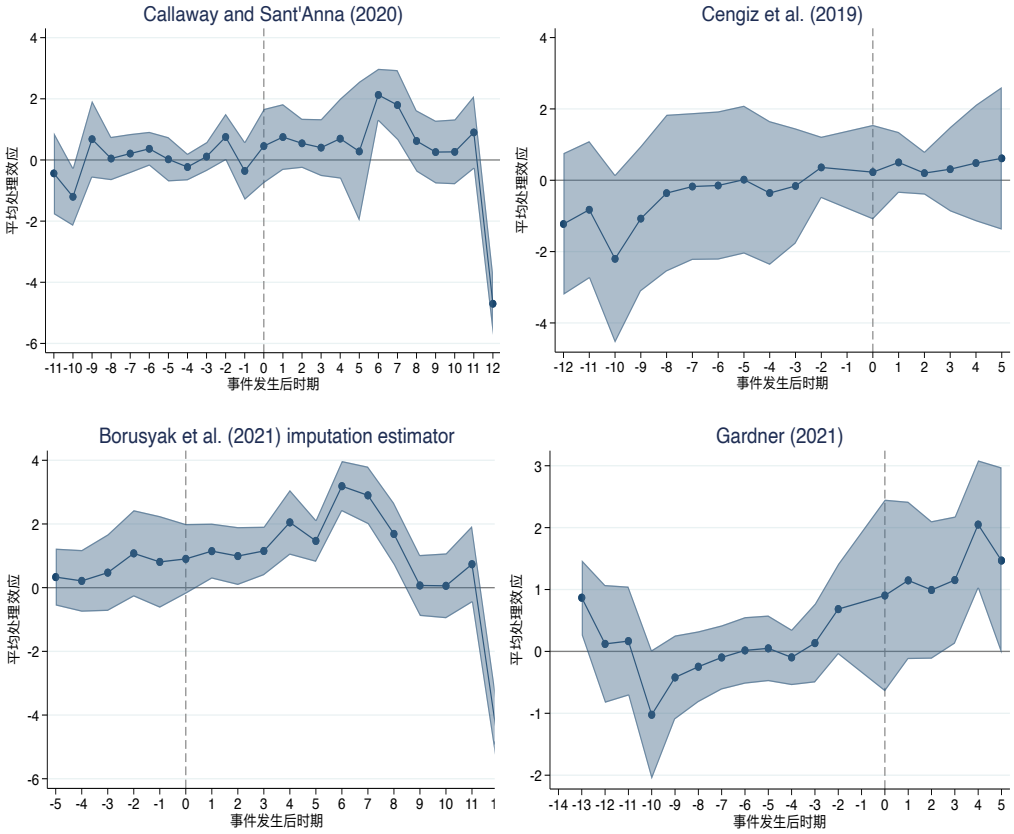
此外，从图 13 中还可以看出，在设立新区后的第 6 年经济增长效应达到最高，即使得 GDP 实际增速提高 2%以上，从而让城市 GDP 实际增速达到 13%（11.02%+2%）以上。



[1]图（*13*） 国家级新区对区域经济增长的动态拉动效应

下面，使用最近几年 DID 计量经济学理论文献提出的稳健估计量来对国家级新区对城市经济增长效应进行更多稳健性检验。这些稳健估计量的事件研究结果如图 14 所示。从左到右，从上到下依次为 Callaway 和 Sant‘Anna（2020）、Cengiz et al.（2019）、Borusyak et al.（2021）、Gardner（2021）的估计量。图中的点线表示点估计量，阴影部分表示 95%置信区间。

从这些稳健估计量结果来看，大部分的事件研究结果均显示了国家级新区的设立确实可以显著促进城市经济增长，且具有持久的拉动作用。



图（*14*） 国家级新区对区域经济增长拉动效应的稳健估计量

综上所示，可以得到结论，曹清峰（2020）利用双向固定效应估计量对国家级新区拉动城市经济增长的估计较为稳健，即使在考虑了最新的稳健 DID 估计量后，结果依然稳健。因此，上述经验证据表明了，国家级新区可以显著促进区域经济增长。

六、总结和实践建议

双重差分法是应用经济学领域最常用的识别方法。学者们通常采用更灵活的双向固定效应模型来呈现双重差分法。传统的双重差分设计更多的关注于平行趋势假设，即平行趋势假设成立的情况下可以得到无偏的平均处理效应。但最新的 DID 理论计量经济学文献表明，在时变处理时点下，如果存在时间异质性处理效应，那么，传统的双向固定效应估计量可能存在偏误，甚至得到完全相反的结论。在某些情形下，2×2 DID 估计量还存在不合意的权重。

[1]纵轴表示动态处理效应估计量，纵轴 0 点处的括号和数值表示处理时点前一期结果变量的均值；横轴表示事件时间，且设置初次处理时点为 0。实心圆点表示点估计量，点估计量上下的横杠表示 95%的置信区间，而横杠外的线条表示 95%的均匀置信区间带。而图中左下角的两个 p 值分别表示拒绝两个原假设“没有处理前的趋势”、“所有的动态效应都已经显示”的概率。

为此，近几年学者们提出了诊断偏误的方法，以及修正偏误的估计方法和稳健估计量。从经验研究的实践来看，可以采用如下的建议性步骤来推进经济学经验研究设计。

第一，画出处理时点图。处理时点图可以清晰地描绘出是否存在时变处理时点，而且也可以从初次处理时间在样本期内的分布来获取更多关于可能的“处理组与控制组”信息。

第二，进行双向固定效应的估计。D. Powell（2021）的研究显示，即使不存在交叠处理时点，一些协变量也可能导致双向固定效应估计量存在偏误。因此，在进行双向固定效应回归时，既要汇报出不包括协变量的回归结果，也要汇报出包括协变量的结果。

第三，平行趋势假设检验。双重差分法的关键在于构造出处理组的“反事实”，因此，控制组要与处理组在处理前满足平行趋势假设。

第四，诊断双向固定效应估计量是否存在偏误或者负的权重。可以采用的诊断方法有：（1）用处理变量与固定效应的回归的余值来检验处理组和控制组在总的平均处理效应中的权重分布信息；（2）用余值化结果变量和余值化处理变量的线性回归来检验“同质性处理效应假设”；或者（3）培根分解。

第五，稳健性检验。稳健性检验的目的在于对 DID（尤其是交叠 DID）的一些关键假设做更多的检验。通常可以从以下几个方面进行：（1）更多的平行趋势检验。（2）更多的“同质性处理效应假设”检验，例如，在同质处理效应假设下，舍弃一些处理个体-时期应该不会影响处理效应估计量的预期值（平行趋势假设满足），包括刀切法（jackknife）估计[1]。还可以逐渐增加样本时期，观察处理效应和权重的变化，如果处理效应很稳定，意味着存在同质处理效应；改变每个个体处理后的时期数量；舍弃一些个体样本。（3）其它稳健性。

第六，事件研究设计和更多稳健性估计量。

第七，与研究主题相关的一些问题的扩展分析，例如，异质性分析。

第八，其它必要的分析。

虽然双向固定效应估计量确实可能存在一些偏误的问题，但是这并不意味着要在经验研究中舍弃它。正如 Wooldrige（2021）在他的工作论文“Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators”中指出的那样，只要恰当地实施双向固定效应模型，仍然可以使用它，并得到类似于 BJS（2021）和 Gardner（2021）处理效应一样的有效估计量。Wooldridge 使用的方法非常类似于两步插值（imputation）法。Wooldridge 提出一个回归方程来估计两步，尤其是估计下列的方程：

$$y_{i,t} = \alpha_i + \alpha_t + \sum_{g=g_0}^{G} \sum_{t=g}^{T} \lambda_{g,t} \times 1(g,t) + e_{i,t}$$

在没有协变量的情形下，他的建议是估计一个带有个体和时间固定效应，并只要组群-时间组合对应一个有效的处理个体，那么就要包含所有可能的组群-时间组合。此时，估计得到的λ就等价于 Callaway 和 Sant‘Anna（2020）的$ATT_{i,t}$。也就是说，Wooldridge（2021）认为，并不是 TWFE 估计量有偏误，而是研究者没有正确使用它。

最后，最新的交叠 DID 理论文献提出了许多的偏误诊断方法和稳健估计量，在实践研究过程中，要想全部通过这些诊断和稳健估计量的推断几乎不可能。但是，目前又没有出现一个公认的、好的稳健估计量。因此，在实践中还需要结合研究目的、研究背景知识等来综合评判估计结果。此外，Scott Cunningham (2020)认为不用双向固定效应估计量可能是更好的方法。例如，Athey et al. (2021)提出了一种基于机器学习的方法论——面板数据“矩

[1] Jackknife 方法由 Quenouille(1949) 提出，并由 Tukey(1958) 创造了 Jackkife 这一术语。Jackknife 是一种再抽样方法，其原始动机是“降低估计的偏差”。

阵完成”（matrix completion）法。他们提出的估计量具备匹配插值法和合成控制法的一些优势，因此未来可能是一种流行的 DID 估计量。[1]

参考文献

[1] Ashenfelter, O (1974), “The Effect of Manpower Training on Earnings: Preliminary Results”, *Proceedings of the 27th Annual Meeting of the Industrial Relations Research Association.*

[2] Ashenfelter, O (1978), “Estimating the Effect of Training Programs on Earnings”, *Review of Economics and Statistics* 60(1): 47-57.

[3] Ashenfelter, O, and D Card (1985), “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs”, *Review of Economics and Statistics* 67(4): 648-60.

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

[5] Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix completion methods for causal panel data models. Journal of the American Statistical Association, 1-15.

[6] Beck, T., R. Levine, and A. Levkov (2010, 10). Big bad banks? The winners and losers from bank deregulation in the United States. Journal of Finance 65 (5), 1637-1667.

[7] Scott Cunningham，(2020). Causal Inference: The Mix Tape.

[8] Nick Huntington-Klein，(2021). The Effect.

[9] Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, Stefan Wager (2021). Synthetic Difference in Differences. American Economic Review.

[10] Dmitry Arkhangelsky , Guido Imbens, Lihua Lei , Xiaoman Luo (2021). Double-Robust Two-Way-Fixed-Effects Regression For Panel Data.

[1] 合成 DID 方法请参见 Dmitry Arkhangelsky et al.（2021）。而矩阵完成法的 stata package 请参见 Licheng Liu et al.（2021）。

[11] Susan Athey, Guido Imbens (2006). Identification and inference in nonlinear difference-indifferences models. Econometrica.

[12] Andrew Baker, David F. Larcker, Charles C. Y. Wang (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? working paper

[13] Kirill Borusyak , Xavier Jaravel , Jann Spiess (2021). Revisiting Event Study Designs: Robust and Efficient Estimation.

[14] Brantly Callaway, Andrew Goodman-Bacon, Pedro H.C. Sant'Anna (2021). Difference-in-Differences with a Continuous Treatment.

[15] Brantly Callaway, Pedro H.C. Sant'Anna (2020). Difference-in-Differences with multiple time periods, Journal of Econometrics.

[16] Clément de Chaisemartin, Xavier D'Haultfoeuille (2018). Fuzzy differences-in-differences. The Review of Economic Studies.

[17] Clément de Chaisemartin, Xavier D'Haultfoeuille (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. American Economic Review.

[18] Clément de Chaisemartin, Xavier D'Haultfoeuille (2021). Two-way fixed effects regressions with several treatments.

[19] Clément de Chaisemartin, Xavier D'Haultfoeuille (2021). Difference-in-Differences Estimators of Inter-temporal Treatment Effects.

[20] Xavier D'Haultfoeuille, Stefan Hoderlein, Yuya Sasaki (2013). Nonlinear difference-indifferences in repeated cross sections with continuous treatments.

[21] Xavier D'Haultfoeuille, Stefan Hoderlein, Yuya Sasaki (2021). Nonparametric Difference-in-Differences in Repeated Cross-Sections with Continuous Treatments.

[22] Bruno Ferman , Cristine Pinto (2021). Synthetic Controls with Imperfect Pre-Treatment Fit. Quantitative Economics.

[23] Simon Freyaldenhoven, Christian Hansen, Jesse M. Shapiro (2019). Pre-event Trends in the Panel Event-Study Design. American Economic Review.

[24] Hans Fricke (2017). Identification based on difference-in-differences approaches with multiple treatments. Oxford Bulletin of Economics and Statistics.

[25] John Gardner (2021). Two-stage differences in differences. NBER WP

[26] Andrew Goodman-Bacon (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics.

[27] Pamela Jakiela (2021). Simple Diagnostics for Two-Way Fixed Effects

[28] Michelle Marcus, Pedro H. C. Sant'Anna (2021). The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics. Journal of the Association of Environmental and Resource Economists.

[29] Jonathan Roth (2021). Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends.

[30] Jonathan Roth , Pedro H.C. Sant'Anna (2021). Efficient Estimation for Staggered Rollout Designs.

[31] Pedro H.C. Sant'Anna , Jun Zhao (2020). Doubly robust difference-in-differences estimators, Journal of Econometrics.

[32] Liyang Sun, Sarah Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics.

[33] Liu, Licheng and Wang, Ye and Xu, Yiqing, A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data (July 12, 2021). Available at SSRN: https://ssrn.com/abstract=3555463 or http://dx.doi.org/10.2139/ssrn.3555463

[34] 曹清峰：《国家级新区对区域经济增长的带动效应研究——基于 70 大中城市的经验证据》，《中国工业经济》2020 年第 7 期

[35] 彭飞,许文立,吕鹏 & 吴华清.(2020).未预期的非税负担冲击:基于"营改增"的研究. 经济研究 (11),67-83.

[36] Currie, Janet, Henrik Kleven, and Esmée Zwiers. 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods." AEA Papers and Proceedings, 110: 42-48.

[37] Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers.MPRA Paper No. 81604。

[38] Butts, K. (2021). Difference-in-Differences Estimation with Spatial Spillovers. arXiv preprint arXiv:2105.03737.

附录

下面，我们来看看《中国工业经济》上一篇文章。作者就是用的时变处理时点的交叠DID，如表 1 所示。且作者的主要回归结果都是用的双向固定效应模型。详细的内容，大家可以去看原文。

表 1 19 个国家级新区，数据来源于百度百科

序号	新区名称	获批时间	主体城市
1	浦东新区	1992 年 10 月 11 日	上海
2	滨海新区	2006 年 05 月 26 日	天津
3	两江新区	2010 年 05 月 05 日	重庆
4	舟山群岛新区	2011 年 06 月 30 日	浙江舟山
5	兰州新区	2012 年 08 月 20 日	甘肃兰州
6	南沙新区	2012 年 09 月 06 日	广东广州
7	西咸新区	2014 年 01 月 06 日	陕西西安、咸阳
8	贵安新区	2014 年 01 月 06 日	贵州贵阳、安顺
9	西海岸新区	2014 年 06 月 03 日	山东青岛
10	金普新区	2014 年 06 月 23 日	辽宁大连
11	天府新区	2014 年 10 月 02 日	四川成都、眉山
12	湘江新区	2015 年 04 月 08 日	湖南长沙

13	江北新区	2015 年 06 月 27 日	江苏南京
14	福州新区	2015 年 08 月 30 日	福建福州
15	滇中新区	2015 年 09 月 07 日	云南昆明
16	哈尔滨新区	2015 年 12 月 16 日	黑龙江哈尔滨
17	长春新区	2016 年 02 月 03 日	吉林长春
18	赣江新区	2016 年 06 月 14 日	江西南昌、九江
19	雄安新区	2017 年 04 月 01 日	河北保定

bacondecomp gdpr did invest consume export gov second agg innov,stub(Bacon_) robust

Computing decomposition across 8 timing groups

including a never-treated group

---

gdpr | Coefficient Std. err. z P>|z| [95% conf. interval]

---+---

did | 1.507396 .4750933 3.17 0.002 .5762297 2.438561

Bacon Decomposition

+---+

| | Beta TotalWeight |

|---+---|

| Timing_groups | 1.693243711 .0907822162 |

| Never_v_timing | 1.21058565 .8640881631 |

| Within | 6.816506863 .0451296207 |