事件研究的秘密：从 DID 到面板事件研究导读

许文立

安徽大学经济学院，合肥，203601

xuweny87@hotmail.com

摘　要：近年来，面板事件研究是自然实验或政策评估领域最受关注的研究设计方法，它也是传统 DID 研究设计的扩展。但目前，无论是计量经济学教材，还是应用性论文都较少关注 DID 事件研究的假设、时变协变量与估计方法的错配带来的估计量和推断偏误。本文基于已有的 DID 事件研究理论文献成果，首先回顾 DID 事件研究的核心假设和时变协变量可能带来的偏误；然后利用曹清峰（2020）的国家级新区对区域经济增长的拉动效应为例，说明与假设和时变协变量限制相一致的检验方法和估计方法；紧接着，以我国碳排放权交易试点为准自然实验，应用上述检验和估计方法来分析碳排放权交易对企业融资约束的影响。最后，提出一些实践建议和进一步研究方向。

关键词：面板事件研究；DID；时变协变量；假设检验

The Secret Behind The Event Study Design: A Primer on The DID Event Study

Abstract: In recent years, the panel event study is the most concerned research design method in the field of natural experiments or policy evaluation, and it is also an extension of conventional DID research design. However, at present, both econometrics textbooks and applied empirical papers pay less attention to the estimators and inference biases caused by the mismatch of assumptions, time-varying covariates and test-estimation methods in DID event study. Based on the existing theoretical literature results of DID event study, this paper first reviews the core assumptions of DID event study and the possible biases caused by time-varying covariates; Secondly, the test methods and estimation methods consistent with the assumptions and limited time-varying covariate are described through replicating Cao Qingfeng(2020)' s results . Thirdly, taking my country' s carbon emissions trading pilot as a

natural experiment, the above test and estimation methods are applied to analyze the impact of carbon emissions trading on corporate financing condition. Finally, some practical suggestions and further research directions are proposed.

Key words: panel event study; DID; time-varying covariates; hypothesis testing

一、引言

人类不断地探寻科学问题的答案推动了社会的前进。中国战国时期的诗人屈原在《天问》中提出了一百多个问题，其中，许多问题都在探寻自然、社会的因果关系。而大部分的应用科学都在使用基于设计的研究方法（The design-based approach）来进行因果推断。20 世纪 90 年代以来，Angrist (1990), Card (1990), Angrist and Krueger (1991), Card and Krueger (1992, 1994)等系统地使用基于设计的研究方法——自然实验来探讨许多重要的经济和社会政策效应。这些方法的使用极大地改变了应用经济研究的实践（The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel，2021）。在经济学领域，如图 1 和 2 所示，“双重差分（DID）”是使用最多的基于设计的研究方法，在 NBER 工作论文的占比达到 20%以上，占 Top-5 期刊论文的占比也超过 15%，而最近十年增长最快的基于设计的研究方法则是“事件研究（event study）”，在 NBER 和 Top-5 期刊上的论文占比均超过 6%。随着时间的推移，事件研究与双重差分联系越来越紧密，因为在应用经济学研究中，双重差分法几乎总是与事件研究法结合在一起来识别因果效应（Currie et al.，2020）[1]。

[1] Currie et al.(2020)的原话是，几乎很少见到没有画事件研究图的 DID，也很少见到没有控制组的面板事件研究设计。Janet Currie & Henrik Kleven & Esmée Zwiers, 2020. "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," AEA Papers and Proceedings, American Economic Association, vol. 110, pages 42-48, May.

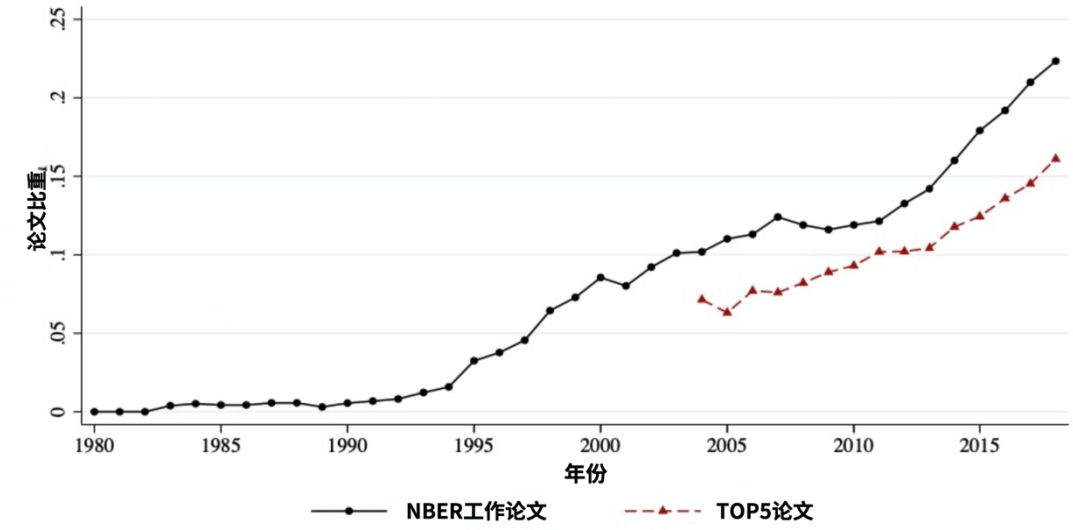

图 1 使用“DID”的论文占比趋势。来源于 Currie et al.(2020)的图 4“准实验方法”。

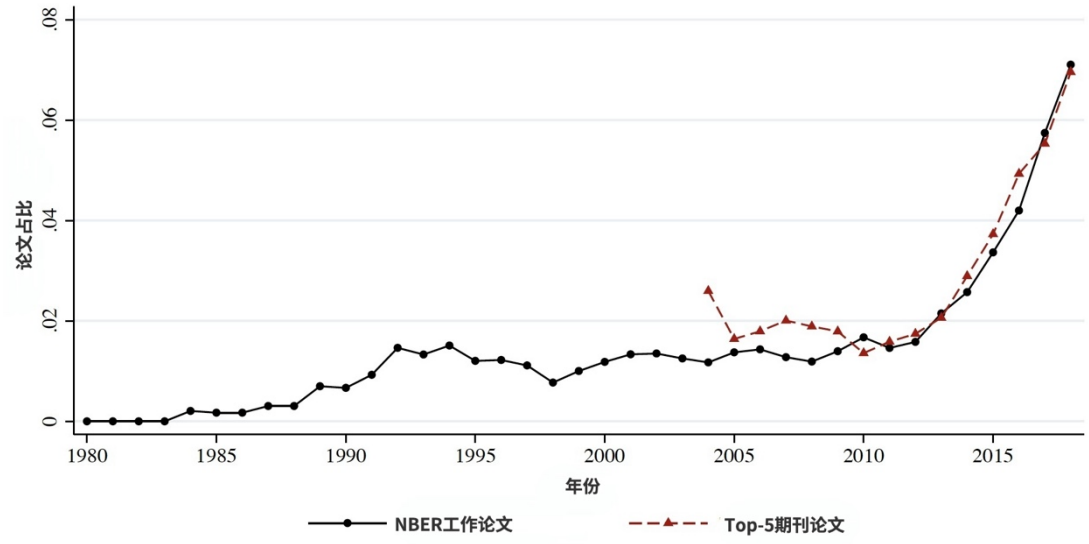

图 2 使用“事件研究设计”的论文占比趋势。来源于 Currie et al.(2020)的图 4“准实验方法”。

关于双重差分的最新理论计量文献表明，在某些情形下，基于传统的双向固定效应（TWFE）模型得到的估计量可能存在较为严重的偏误（de Chaisemartin and D’ Haultfoeuille，2020；Goodman-Bacon，2021； Callaway and Sant’ Anna，2021）。为了解决这些偏误问题，一些学者提出用 DID 事件研究法（DID Event Study）来避免这些偏误（Goodman-Bacon，2021；黄炜等，2022）。正如 Goodman-Bacon（2019）[2]给出的建议：我们可以换一种方式来呈现结果——事件研究。在回答“我应该做事件研究吗？”问题时，他也给出“是的，在

[2] Goodman-Bacon，2019：SO YOU’ VE BEEN TOLD TO DO MY DIFFERENCE-IN-DIFFERENCES THING: A GUIDE

许多情形下，事件研究是对的，且你能相信扁平的处理前效应以及清晰地处理后变化。当你有大量的未处理组个体时，事件研究尤其可信，因为这时给予 ‘有问题’ 2×2 DIDs——用已处理的个体作为控制组——的权重较小。”

Cunningham（2021）问，问题出在静态参数？如果是，那么，我们可以估计动态回归来避免这些问题，即当代 DID 事件研究[3]。但是，Callaway & Sant’ Anna（2021）指出，即使处理组群/类之间的动态处理效应是同质的，有些动态 DID 估计量仍然会存在严重偏误。且 Sun & Abraham（2021）的研究显示，在异质性处理效应的动态模型设定中处理前和后指标系数也会存在偏误。

本文简要回顾一下双重差分和事件研究。然后，给出 DID 和事件研究的识别假设，及其可能造成的偏误问题和一些稳健估计量。再然后，以模拟数据和发表论文作为案例演示传统事件研究估计量的偏误，及其事件研究的实践用法。

二、双重差分事件研究的识别假设

双重方法方法的思想将时间维度维度划分成 “处理前时期” 和 “处理后时期” 的 2 个时期。双向固定效应模型本身允许多个时期，既允许多个处理前时期，也允许多个处理后时期。但是，在处理多时期 DID 时，通常将多个时期加总为两类大的时期段：“处理前” 时段和 “处理后” 时段，然后估计 “处理后” 时段的平均处理效应，即上文的估计参数β^{DID}，或者静态参数。

在实践研究中，除了对 “处理后” 时段的平均处理效应感兴趣外，研究者可能也想估计政策效应如何随时间变化——动态处理效应[4]，例如，“营改增” 对企业非税负担的影响效应随时间不断变化（彭飞等，2020）。用于识别动态处理效应的面板数据模型也被称为 “面板事件研究模型”，或者在本文的环境中称为 “双重差分事件研究模型”（ Nick Huntington-Klein，2021）。

双重差分事件研究设计的本质特征是重新定义时间变量：在上述 DID 研究设计中，时间变量使用日历时间，例如，事件变量 t=2021 年、2022 年等等，但在面板事件研究中，重新定义时间变量——相对事件时间，即处理事件发生前后的时间（Damian Clarke and Kathya

[3] 但是，事件研究法真的可以解决传统 TWFE 估计量的偏误吗？其实答案不能用事件研究来说事，如果仅仅用” 事件研究 event study “这个词来说事，就有点不负责任，甚至误导了。因为时间序列数据（一般在金融领域）也有事件研究（参见 Kothari & Warner（2007，Handbook of Corporate Finance, Volume 1）的文献回顾，或者【应用计量系列 21】金融领域的事件研究），我们这里通常是指的面板数据事件研究或者动态 DID 模型（参见 Clarke & Schythe，2020；Freyaldenhoven et al.,2021）。事件研究既可以应用于事件同时发生的情形，也可以用于事件在不同时点发生在不同个体的情形（Clarke & Schythe，2020），后一种情形被 Athey and Imbens (2018)称为 “交叠采用设计” 。

[4] 需要注意 “动态面板数据模型” 与 “面板数据模型的动态处理效应” 之间的差别：前者是指在面板数据模型中包含结果（因）变量的滞后期作为解释变量，而后者是指面板数据模型中包括处理变量的领先期和滞后期作为解释变量。

Tapia Schythe，2020；Gabor Bekes and Gabor Kezdi,2021）。处理前时期的相对事件时间通常用负整数来表示，例如，将处理前一期的相对事件时间定义为-1，处理前两期定义为-2 等等。处理事件发生的那一期的相对事件时间通常定义为 0。而处理后时期的相对事件时间通常用正整数来表示，例如，处理后一期的相对事件时间定义为 1，处理后两期定义为 2 等等。然而，有些情形下，处理事件发生在两期之间，这时有两种方式定义相对事件时间：第一种方式是将相对事件时间 0 定义为处理事件发生后一期，那么，相对事件时间 1 则表示处理后第二期等等，这种方式的优势可能在于相对时间事件与日历事件频率一致；第二种方式是忽略相对事件时间 0，而直接从相对事件时间-1 跳到 1，这种方式的优势在于更清晰地展现出“处理前”和“处理后”。[5]

我们利用个体 i 和时间 t 的面板数据来估计二值型政策干预$D_{i,t}$对结果$Y_{i,t}$的因果效应，个体一旦接受政策干预，在样本期内一直处于处理状态，E_i表示个体的初次政策干预时点。且定义$Y_{i,t}(0)$表示个体 i 在时期 t 如果没有接受政策干预的潜在结果。在 DID 事件研究中，我们感兴趣的目标估计量为每个时点$t \geq E_i$上的时变平均处理效应ATT_t：

$$
\begin{aligned}
ATT_{t>E_i} &= \mathbb{E}\left[Y_{i,t\geq E_i} - Y_{i,t\geq E_i}(0) \,|D_{i,t}=1\right] = \mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t\geq E_i}(0) \,|D_{i,t}=1\right] \\
&= \left\{\mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i}\,(0)|D_{i,t}=1\right]\right\} \\
&\qquad - \left\{\mathbb{E}\left[Y_{i,t\geq E_i}(0) \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i}\,(0)|D_{i,t}=1\right]\right\} \\
&= \left\{\mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=1\right]\right\} - \left\{\mathbb{E}\left[Y_{i,t\geq E_i}(0) \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=1\right]\right\} \\
&= \left\{\mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=1\right]\right\} - \left\{\mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=0\right] - \mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=0\right]\right\}
\end{aligned}
\tag{1}
$$

为了识别$ATT_{t>E_i}$，需要施加两个假设：（1）无预期效应假设；（2）平行趋势假设。

假设 1：无预期效应假设。从上述潜在结果分析框架可知，$\mathbb{E}\left[Y_{i,t<E_i}\,(0)|D_{i,t}=1\right] = \mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=1\right]$。

无预期效应假设意味着，理性的个体在预期到政策实施前，不会有任何政策效应存在，即理性个体即使预期到政策，在实施前也不会改变其行为。施加这个假设是因为“卢卡斯批判”（Lucas，1978），即理性经济代理人如果预期到政策实施，那么，他们会在政策实施前就改变行为，从而影响结果变量。例如，个体预期到央行会在将来某个时点加息，她会在加息前就减少消费，多储蓄，以使得在加息后获得更多的利息收入，如果我们估计货币政策对个体消费或储蓄的效应，那么，货币政策实施前，消费和储蓄结果就会发生变化。也就是说，政策实施前的反事实结果$\mathbb{E}\left[Y_{i,t<E_i}\,(0)|D_{i,t}=1\right]$与政策实施前的实际结果$\mathbb{E}\left[Y_{i,t<E_i} \,|D_{i,t}=1\right]$可能并不相等。如果它们不相等，即存在理性预期效应，那么，上述潜在结果框架（1）中第二行就得不到第三行的结果。

总之，要使得$\left\{\mathbb{E}\left[Y_{i,t\geq E_i} \,|D_{i,t}=1\right] - \mathbb{E}\left[Y_{i,t<E_i}\,(0)|D_{i,t}=1\right]\right\} - \left\{\mathbb{E}\left[Y_{i,t\geq E_i}(0) \,|D_{i,t}=1\right] -$

[5] 也有学者将这个过程称为“再中心化时间”，即自然时间减去处理事件发生时间，例如，Nick Huntington-Klein（2021）。

$\mathbb{E}\left[Y_{i,t<E_i}(0)|D_{i,t}=1\right]\}=\{\mathbb{E}\left[Y_{i,t\geq E_i}|D_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=1\right]\}-\{\mathbb{E}\left[Y_{i,t\geq E_i}(0)|D_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=1\right]\}$，就要对其施加“无预期效应”假设。

假设 2：平行趋势假设，即$\mathbb{E}\left[Y_{i,t\geq E_i}(0)|D_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=1\right]=\left[Y_{i,t\geq E_i}|D_{i,t}=0\right]-\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=0\right]$。

平行趋势假设有多种理解。按照上述潜在结果框架，平行趋势假设是处理组在没有处理时处理前后结果的变化与控制组结果变化相同。但从这个理解来看，存在“双重假设”，即假设中有假设——假设处理组没有处理的情形。这在潜在结果框架中被称为“反事实”，也就是现实中不存在的情形。因此，这种“平行趋势假设”的理解在实践应用中不可检验。

我们将上述“平行趋势假设”等价变换为$\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=1\right]-\mathbb{E}\left[Y_{i,t<E_i}|D_{i,t}=0\right]=\left[Y_{i,t\geq E_i}(0)|D_{i,t}=1\right]-\left[Y_{i,t\geq E_i}|D_{i,t}=0\right]$。等号左边变成处理组和控制组在处理前的差异，而等号右边则是处理组和控制组在处理后的差异。这种“平行趋势假设”的理解意味着处理组和控制组在处理前的变化趋势在没有处理时会延续到处理事件后。从实践的角度来讲，这个理解仍不可检验。但在实践中，许多经济学者都将这种理解分解成两步：第一步，处理组和控制组处理前的变化趋势，这种变化趋势可以通过数据来检验，被称为“处理前趋势检验（pretrends test）”（Roth，2022）；第二步，在处理后时期，外生政策冲击以相同的方式、程度影响处理组和控制组，被称为“共同冲击(common shocks)”（Dimick and Ryan,2014; Ryan et al., 2015）。

也就是说，真正地检验“平行趋势假设”应该是检验“处理前趋势”和“共同冲击”，缺一不可。从理论上来说，除了“处理前趋势检验”，“共同冲击”也是必要的检验。“共同冲击”有两个核心的要点：第一，外生的政策冲击，也就是说政策干预$D_{i,t}$是外生的，一般来说，经济研究者在评估政策效应时，都是借助“自然实验”或者“准实验”框架（The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel，2021），其核心在于政策实施具有一定的随机性，满足一定的外生性；第二，以相同的方式、程度影响处理组和控制组，这意味着政策干预$D_{i,t}$的效应具有同质性。

假设 3：处理效应同质性假设。每个组群有相同的处理效应。

假设 3 并不要求处理效应在时间上是恒定的，这一点与 Goodman-Bacon（2021）不同。但是，它要求对于所有的类都有相同的处理情况。换言之，假设 3 假设无论是静态还是动态环境下，所有的组群都有相同的处理情况。

三、双重差分事件研究中的时变混淆因子

在 DID 事件研究中，混淆因子的概念从根本上就与其它存在差异。DID 事件研究中的混淆因子使得反事实假设不成立，只要（1）协变量与处理相关；（2）协变量与结果之间存在时变关系，或者处理组和控制组的协变量分布在时间维度上有不同的演化路径（Simon and Belh-Gomez,2018; Zeldow and atfield,2019）。以因果图来说明回归中协变量的选择，如图 3

所示。

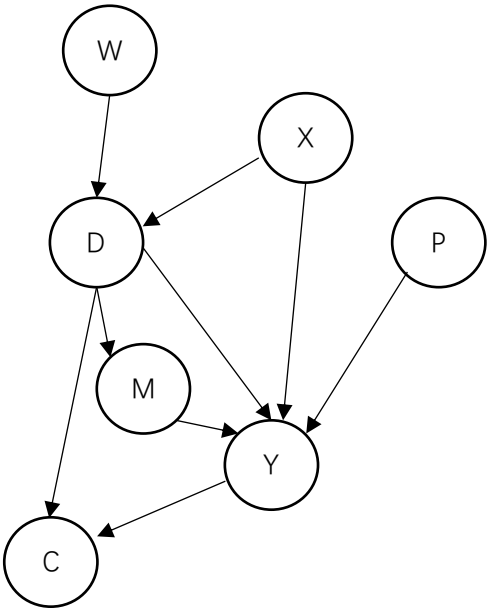

图 3 多种控制变量的因果图模型

上述因果图中包含七种类型的变量：(1) 结果变量 Y；(2) 处理变量 D；(3) 处理变量和结果变量的共同原因 X，也被称为混淆因子；(4) 不影响处理变量，但可以预测结果的变量 P；(5) 处理变量对结果变量的因果链上的中介变量 M；(6) 处理变量和结果变量的共同效应 C；(7) 只影响处理变量的变量 W。在回归方程中，混淆因子 X 是好的控制变量，预测变量 P 是基本无害的控制变量，而 W 是基本有害的控制变，中介变量 M 和共同效应 C 则是“坏的控制变量”（Angrist and Pischke，2009；Matheus Facure Alves，2022），证明过程见附录。

在应用经济学研究文献中，协变量要么不随时间变化，要么只包含处理前协变量的值（Bonhomme and Sauder，2011；Lechner，2011），例如，Li et al.（2016）在回归方程中包含处理前的时变协变量。但当时变协变量可以预测处理状态，紧接着又被处理所影响，反过来又在下一期影响处理状态，这时时变协变量会使得 DID 事件研究估计更加困难，且即使静态 TWFE 估计量也会产生偏误（Caetano et al.，2022）。因为此时时变混淆因子即发挥着共同原因的作用，又充当着中介变量的作用（Hernan and Robins，2020）。

未处理组结果变量$Y_{i,t}^{0}$的经典线性模型可以写成（Athey and Imbens，2006； Angrist and Pischke，2009）：

$$Y_{i,t}^{0} = \alpha_i + \alpha_t + \gamma I + \epsilon_{i,t} \tag{2}$$

因此，我们可以利用下列关系来将可观测的结果变量与未处理组的结果变量联系起来：

$$Y_{i,t} = Y_{i,t}^{0} + \beta D_{i,t} \tag{3}$$

对于一个 2×2 DID 来说，只要我们声明的模型是正确的，那么，平行趋势假设就成立。

这是因为

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=1]=(\alpha_i+\alpha_t+\gamma)-(\alpha_i+\gamma)=\alpha_t \tag{4}$$

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=0]=(\alpha_i+\alpha_t)-\alpha_i=\alpha_t \tag{5}$$

下面，我们引入协变量。我们将未处理组结果变量表示成：

$$Y_{i,t}^0=\alpha_i+\alpha_t+\gamma I+\eta_t X_{it}+\epsilon_{i,t} \tag{6}$$

其中，X_{it}表示时间变的协变量，但其可能会对结果产生时变效应η_t。我们假设协变量对结果变量产生时间不变的效应，即$\eta_{t=2}=\eta_{t=1}=\eta$。

（1）当协变量X_i表示时间不变的协变量，在这种情形下，即使协变量X_i的分布在处理组和控制组间存在差异，平行趋势假设也成立：

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=1]=(\alpha_i+\alpha_t+\gamma+\eta\mathbb{E}[X_i(2)\,|D=1])-(\alpha_i+\gamma+\eta\mathbb{E}[X_i(1)\,|D=1])=\alpha_t \tag{7}$$

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=0]=(\alpha_i+\alpha_t+\eta\mathbb{E}[X_i(2)\,|D=0])-(\alpha_i+\eta\mathbb{E}[X_i(2)\,|D=0])=\alpha_t \tag{8}$$

上式成立，是因为协变量分布在处理组和控制组内不随时间变化，因此，$\mathbb{E}[X_i(2)\,|D=1]=\mathbb{E}[X_i(1)\,|D=1]$，$\mathbb{E}[X_i(2)\,|D=0]=\mathbb{E}[X_i(1)\,|D=0]$。

（2）如果协变量随时间变化X_{it}，即$\mathbb{E}[X_{it}(2)\,|D=1]=\mathbb{E}[X_{it}(1)\,|D=1]$，$\mathbb{E}[X_{it}(2)\,|D=0]=\mathbb{E}[X_{it}(1)\,|D=0]$可能不成立，平行趋势假设就变成：

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=1]=(\alpha_i+\alpha_t+\gamma+\eta_{t=2}\mathbb{E}[X_{it}(2)\,|D=1])-(\alpha_i+\gamma+\eta_{t=1}\mathbb{E}[X_{it}(1)\,|D=1])=\alpha_t+\eta(\mathbb{E}[X_{it}(2)\,|D=1]-\mathbb{E}[X_{it}(1)\,|D=1]) \tag{9}$$

$$\mathbb{E}[Y^0(2)-Y^0(1)\,|D=0]=(\alpha_i+\alpha_t+\eta_{t=2}\mathbb{E}[X_{it}(2)\,|D=0])-(\alpha_i+\eta_{t=1}\mathbb{E}[X_{it}(2)\,|D=0])=\alpha_t+\eta(\mathbb{E}[X_{it}(2)\,|D=0]-\mathbb{E}[X_{it}(1)\,|D=0])$$

此时，处理组的协变量效应演化路径$\eta_{t=2}\mathbb{E}[X_i(2)\,|D=1]-\eta_{t=1}\mathbb{E}[X_i(1)\,|D=1]$和控制组的协变量演化路径$\eta_{t=2}\mathbb{E}[X_i(2)\,|D=0]-\eta_{t=1}\mathbb{E}[X_i(1)\,|D=0]$并不必然相等。因此，只有在协变量在处理前后的演化趋势在处理组和控制组之间无差异时，平行趋势才成立。

四、假设检验与估计方法

下面，我们以曹清峰(2020)的文章《国家级新区对区域经济增长的带动效应》为例来说明，在实证研究实践中，上述假设检验和效应估计的应用方法。

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）。

为此，将个体处理虚拟变量与每一个相对事件时间的虚拟变量交乘，就可以将双向固定效应 DID 模型转换成 DID 事件研究模型[6]：

$$Y_{i,t} = \alpha_i + \alpha_t + \sum_{k=-(L-1)}^{K-1} \beta_k D_{i,t+k} + X_{it}\Gamma' + \epsilon_{i,t} \quad (2)$$

其中，$Y_{i,t}$表示结果变量；α_i、α_t分别表示个体和事件固定效应；X_{it}表示时变协变量，即包括可观测协变量，也包括不可观测协变量，Γ'是协变量的系数向量；$D_{i,t+k}$表示处理事件发生前后 k 期的虚拟变量。β_k是我们关注的动态处理效应估计量，如图 4 所示。L 和 K 分别表示处理事件发生前后最大时期数，即在国家级新区设立对区域经济增长效应的事件研究中，L=13，K=12，即分别考察国家级新区设定前 13 年和设立后 12 年的动态经济增长效应。

[6] DID 事件研究设计的其它形式，参见 Damian Clarke and Kathya Tapia Schythe（2020）、Schmidheiny and Siegloch (2019)、Freyaldenhoven et al. (2022)。

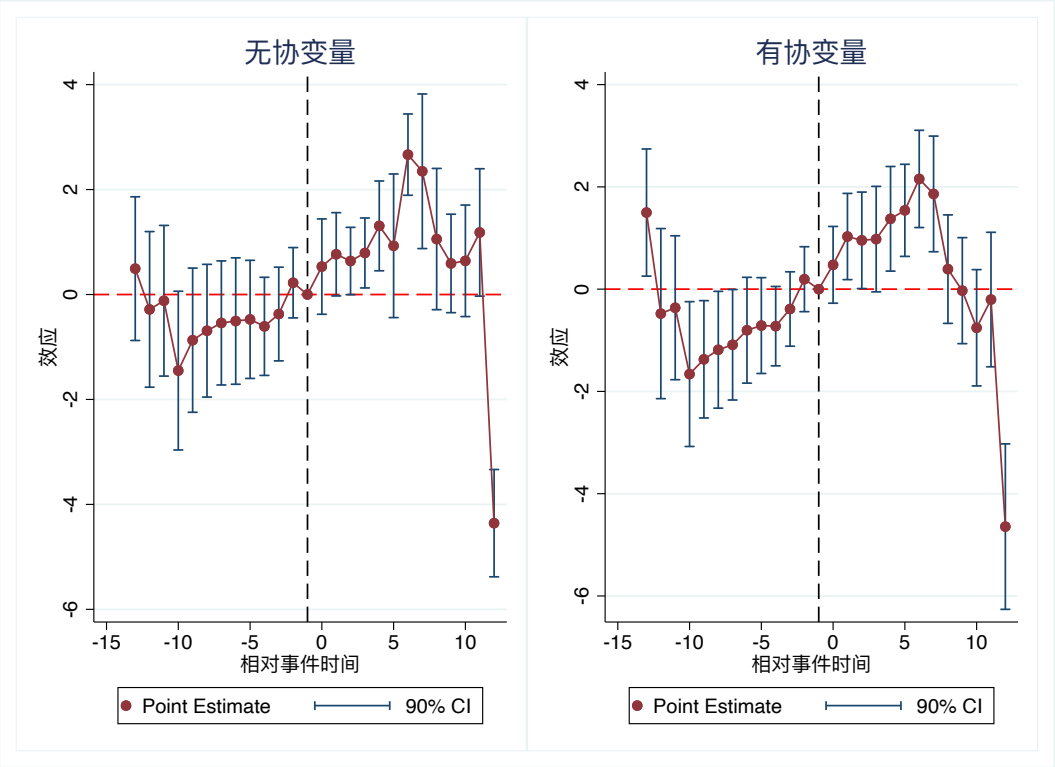

图 4 事件研究图

经济学研究者使用事件研究图来呈现感兴趣的因果效应。从事件研究图中可以观察两类核心信息：第一，处理前时期，处理变量对结果变量的效应估计系数及其置信区间，且理论上来说，处理前时期的系数应该为 0，即使不为 0，也应该不显著，这是因为；第二，处理后时期，处理变量对结果变量的动态处理效应估计系数与置信区间（Nick Huntington-Klein，2021）。

（一）处理前趋势检验：预期与平行趋势假设检验

实践中，大部分的学者都关注于处理前单个时期的估计系数及其显著性，且偶尔会遇少数系数在合适置信水平下显著不为 0。例如，Roth(2022)收集、整理了美国经济学会三本期刊上使用事件研究的 12 篇论文，发现所有的论文都用带有点估计区间的事件研究图来评价单一处理前时期系数的显著性，其中，五篇直接讨论了单个显著性，一篇报告了联合显著性，没有一篇讨论处理前趋势程度，而且有三篇论文的处理前时期至少有一个系数是显著的。因此，研究实践中，并不一定需要所有处理前的系数不显著。这是因为（1）平行趋势有多个版本，即可以理解成处理组和控制组在所有处理前时期均无差异，也可以理解成在所有处理前时期平均无差异，还可以理解成在些处理前时期无差异；（2）静态 DID 利用了样本所有时期来估计因果效应，而 DID 事件研究则只利用一期的数据，因此，估计的系数精度较低（Nick Huntington-Klein，2021）。

图 4 显示，国家级新区设立前，该项改革对区域经济增长没有显著的促进作用。在没有

可观测协变量时，处理前时期系数在 90%置信水平下均不显著，虽然在加入可观测协变量后，有少数处理前时期系数显著为负，但绝大部分系数仍不显著。国家级新区设立后，改革的经济增长作用并没有立即显现，而是在改革后第 4 年开始显现，并持续到第 8 年。

除了单一系数显著性外，研究者可能还对整个时间路径的统计证据感兴趣，例如，处理前后时期系数的联合显著性（Roth，2022；Freyaldenhoven et al.，2022）。一种方式是在事件研究模型估计后，使用 F 统计量来进行联合显著性检验（Clarke and Schythe，2020）；另一种方式是在事件研究图的点估计区间上增加 90%的均匀置信带宽（uniform confidence band），它表示至少在 90%的时间内包含了参数集的真实值（Freyberger and Rai ,2018; Olea and Plagborg-Moller,2019; Freyaldenhoven et al.，2022）

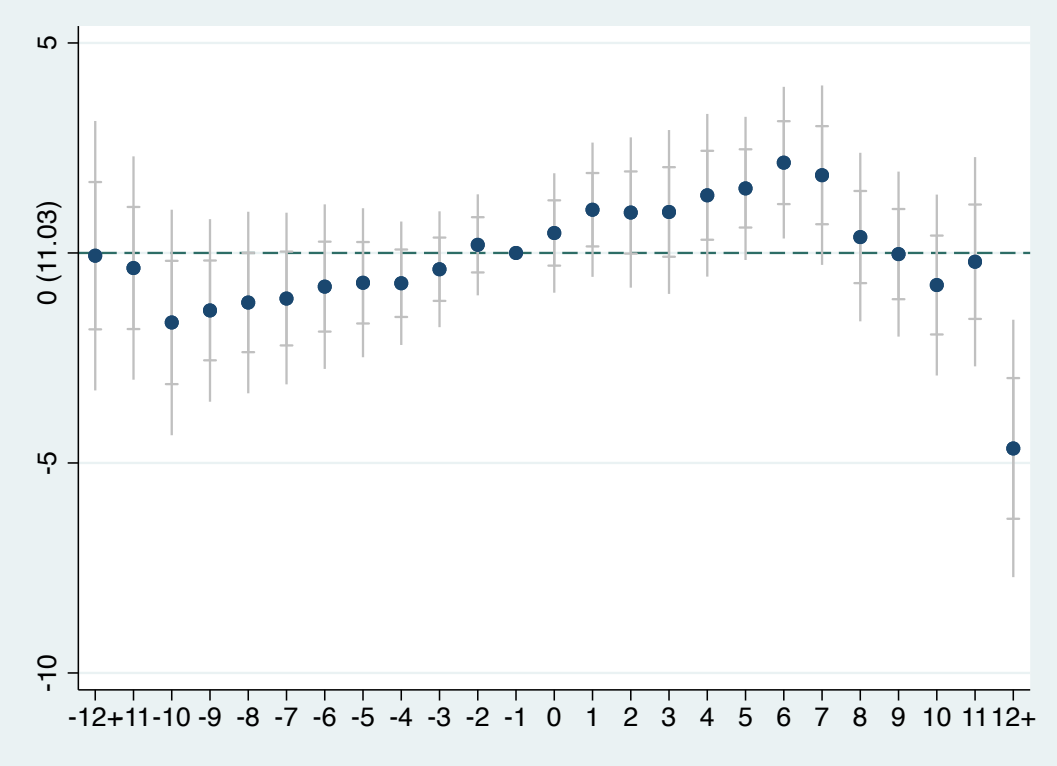

图 5 事件研究图：联合显著性检验

如图 5 所示，在事件研究图中增加了 Freyaldenhoven et al.（2022）建议的 95%均匀置信带宽，也就是点估计置信区间外的线条。我们可以从图 2 中看到，所有的处理前时期系数的均匀置信带宽都包含 0。我们使用均匀置信带宽不能拒绝“国家级新区设立前所有时期的改革效应等于 0“的原假设。但是，从单一时期系数来看，国家级新区设立前第 9 年和第 10 年的点估计分别在 90%水平上显著为负。因此，对于处理前趋势检验，联合显著性检验更加可信。

Freyaldenhoven et al.（2022）指出，如果式（2）的回归方程是正确的，那么，处理前 L 期以外的时期，政策变化并不会引起结果的变化。那么，包含更早的处理前时期意味着事件研究图中包含了关于“更早时期无政策预期效应“假设的信息。我们可以使用 Wald 检验

来得到该假设检验的 p 值。如果拒绝该假设，可能意味着存在预期行为，或者包含混淆因子的效应。

如图 6 所示，事件研究图左下角显示了国家级新区设立前，区域经济增长趋势的 Wald 检验的 p 值。从结果可以看出，在 95%的置信水平下，有显著的统计证据表明，可以拒绝“国家级新区设立前，区域经济增长不存在变化趋势“这个原假设。这时，可能意味着在国家级新区设立前，各地区已经预期到了新区设立，进而改变了地区的相关经济发展行为，区域经济增长已经出现分化。或者有其他混淆因子在新区设立前就拉动了地区经济增长的分化。

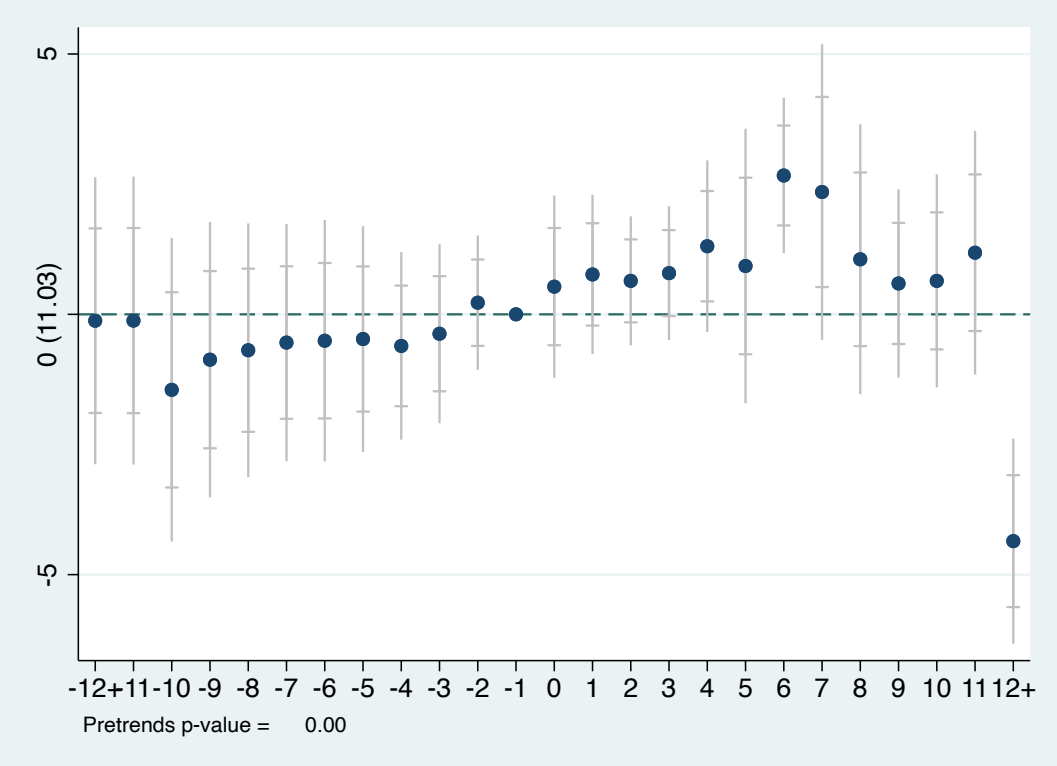

图 6 事件研究图：无处理前趋势检验

（二）交叠处理和异质性处理效应

1、交叠处理下偏误诊断

最新的交叠 DID 理论文献表明，传统 TWFE 估计量的偏误可能由于负权重和异质性处理效应引起。这两个问题是同一硬币的两面，而对于它们的诊断，有如下方法：

（1）回归法

Pamela Jakiela（2021）给了两个建议：第一，画一个权重的分布图，判断有没有受处理个体获得了负权重，即用回归得到余值化的处理变量，进而画出权重分布图。第二，如果处理效应是同质的，处理组中有个体收到负权重并无不妥，但存在异质性处理效应就会产生偏误。为此，可以做一个简易的 OLS 回归：用余值化结果变量对余值化处理变量，是否为控制组（dummy），以及余值化处理变量*是否为控制组进行回归。关注的系数是最后一个交乘

项的系数，如果这个系数显著，就拒绝了处理效应同质性假设，说明直接用 TWFE 估计量是有问题的。

（2）培根分解

为了更简单、易用、直观地发现估计结果中可能存在的偏误，Goodman-Bacon（2021）给出了一种 TWFE 估计量偏误诊断的方法，许文立（2021）将它称为“培根分解”（Bacon decomposition）。这种检验方法呈现的分解结果非常直观，因此，受到了越来越多的应用研究者的关注。例如，Miller et al.（2021）就利用培根分解来检验医疗补助扩围改革对死亡率影响的每个 2×2 DID 在 TWFE DID 估计量中的作用，他们发现了仅仅只使用处理组和未处理组的估计量比基准结构更大。但是，该方法仅限于强平衡面板数据结构，且个体接受处理后就一直处于处理状态。

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”。从表 1 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。

表1 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

注：vs 前后分别表示处理组和控制组

（3）CD 分解

为了诊断交叠 DID 中双向固定效应估计量可能存在的偏误，de Chaisemartin and D’ Haultfoeuille（2018，2022）分解得到了上述加权估计量中的权重，并提出通过一个安慰剂检验——估计权重与处理效应的关系——来诊断权重对处理效应是否有影响。在平行趋势

满足情况下，原假设为“固定效应估计量是平均处理效应的无偏估计”，那么，安慰剂估计量不显著意味着没有证据拒绝原假设。

（4）静态处理效应检验

除了上述方法外，还可以在式（1）中仅仅包含处理变量的当期值来估计静态模型。Freyaldenhoven et al.（2022）建议，评价静态模型对动态效应模型的拟合程度以判断样本数据是否表现出异质性处理效应。

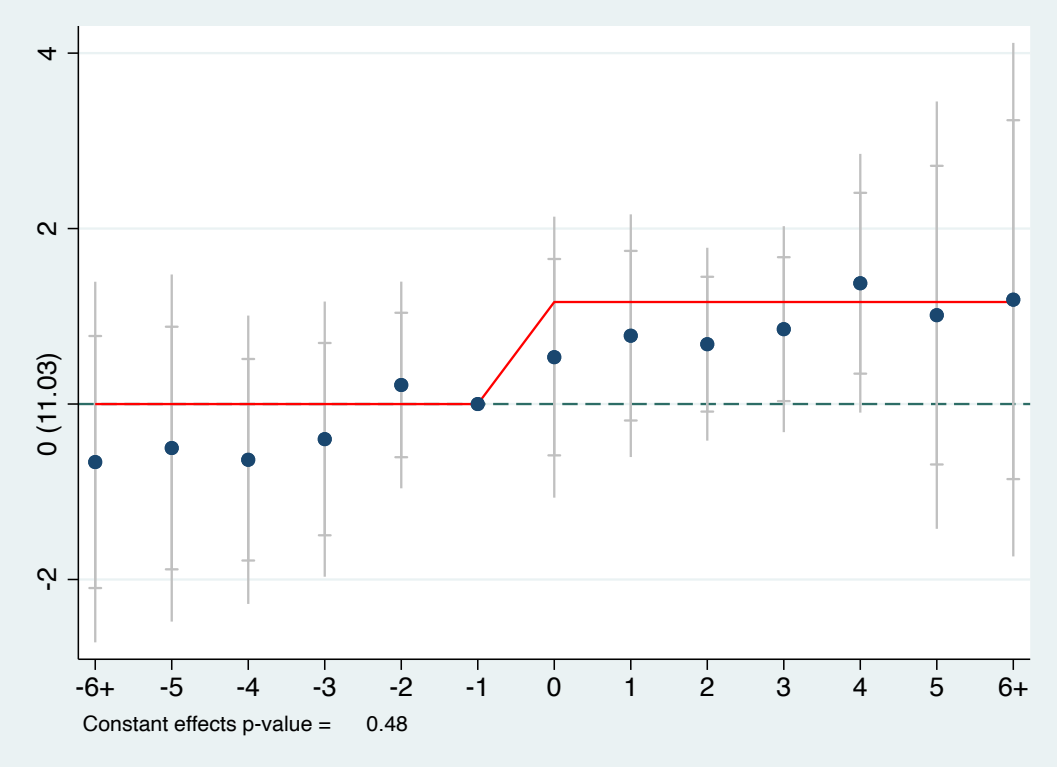

图 7 事件研究图：静态效应检验

如图 7 所示，红色线条表示静态模型的效应估计量，该模型假设国家级新区设立的经济增长效应是静态的。通过均匀置信带宽来比较国家级新区的经济增长静态效应和动态效应的拟合程度。例如，国家级新区的经济增长静态效应全部落入 95%的均匀置信带宽内，这意味着不能拒绝“国家级新区的经济增长效应是静态“的假设。且从图 4 左下角的 Wald 检验 p 值=0.48>0.05 可以看出，不能拒绝静态效应模型。

2、稳健估计量：无预期效应、平行趋势和异质性处理效应

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，传统面板数据事件研究可能不会起作用（Sun and Abraham，2021）。

许多研究者都提出了异质性处理效应稳健估计量，这些估计量采用了不同的策略来规避“禁止的对照组”（ de Chaisemartin and D’ Haultfoeuille，2022）。根据采用的策略，Scott Cunningham（2021）将这些稳健估计量划分为三大类：（1）加权组群-时间的 ATT，例如，Sun and Abraham（2021）、Callaway and Sant’Anna（2021）；（2）通过相对事件时间的平衡来堆叠，例如，Cengiz et al.（2019）；（3）插补(imputation)方法，例如，Borusyak et al.（2022）、Gardner（2021）、Wooldridge（2021）等。

de Chaisemartin and D’ Haultfoeuille（2022）和 Borusyak et al.（2022）从假设、精度与效率等方面比较了这些稳健估计量：相较于 Sun and Abraham（2021）和 Callaway and Sant’ Anna（2021）估计量，Borusyak et al.（2022）估计量更精确，但由于方差更大，可能实践中难以通过统计推断。

在假设检验方面，相较于 Sun and Abraham（2021）和 Callaway and Sant’ Anna（2021）施加的平行趋势假设，Borusyak et al.（2022）施加的平行趋势假设更强，因为 Borusyak et al.（2022）在进行处理前趋势检验时，实际上检验了无预期效应假设和平行趋势假设同时成立。这就使得 Borusyak et al.（2022）估计量可能更难以与数据产生过程相一致。但 Borusyak et al.（2022）将假设检验和估计过程分离，用未处理的样本来进行处理前趋势检验，这就可以避免 Roth（2022）指出的“传统处理前趋势检验并不能侦测出处理前趋势，并由此导致估计的偏误”问题。

在应用经济学研究中，到底哪个稳健估计量更好，还需要结合更多的具体研究背景。例如，国家级新区的设立，从提出想法，审批到设立可能需要一段时间，而在中央批准设立前，地方政府可能就已经在为新区设立做准备（曹清峰，2020）。那么，在研究国家级新区设立对区域经济增长的拉动作用时，新区批复前可能存在明显的预期效应，因此，需要更加关注对预期效应的稳健性，此时 Borusyak et al.（2022）估计量可能更合适，结果如图 8 所示。因为该事件研究图中处理前时期系数使用未处理样本进行估计，因此，可能更加符合无预期效应假设，因此，未处理地区本身就没有申请设立国家级新区。而图 8 的处理前趋势也确实在 95%置信水平下不显著。设立新区后一年开始显现出区域经济拉动作用，并一直持续 8 年。当然，研究实践中，更多学者采用多种稳健估计量混合形式来呈现处理后效应的动态路径，例如，Biasi and Sarsons（2022）等。

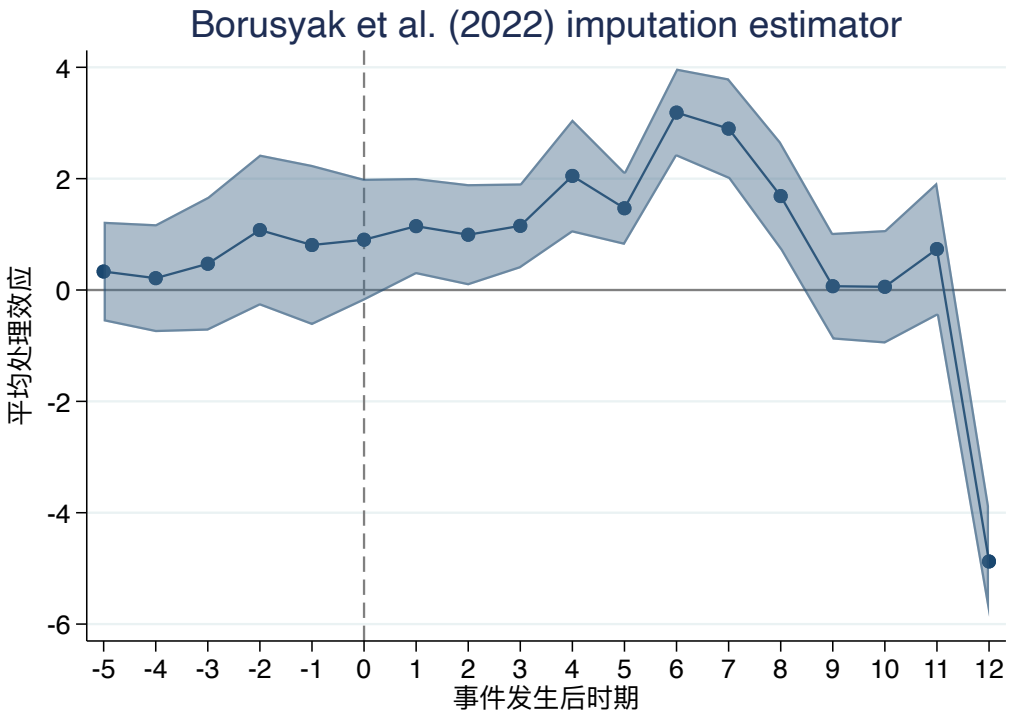

图 8　事件研究图：Borusyak et al.（2022）估计量

（三）混淆因子

在应用经济研究中，遗漏变量偏误是β_k的因果效应识别的主要威胁（Bazzi et al.，2020；Miller et al.，2021）。因此，除了要尽可能去证实识别假设或证伪打破识别假设的条件外，还应该尽可能降低遗漏变量偏误，主要的原则是：尽量控制混淆因子 X 和预测变量 P，尽量避免控制变量 W、中介变量 M 和共同效应 C。

1、排除重要的可观测因素干扰

在事件研究回归模型中，我们需要尽可能地控制混淆因子和预测变量。对于可观测的混淆因子和预测变量，只需要将它们增加到回归方程中即可。在应用经济研究中，通常需要结合研究背景、制度环境和经济理论来初步判断最重要的一些可观测的混淆因子和预测变量。然后再进一步思考一些重要的同期其它政策因素的干扰。

曹清峰（2020）结合现有研究，选取了投资、国内消费、净出口、政府财政支出、经济聚集度、二产比重和创新等作为控制变量，且都为时变协变量。同时，作者还认为“设立国家新区的城市作为国家或者区域中心城市往往受到多项国家层面区位导向性政策的影响”，而且这些区位导向政策也是区域经济增长的重要驱动力。因此，进一步控制“国家综合配套改革试验区政策”和“自由贸易试验区政策”的影响，结果如图 9 所示。控制这些可观测的时变和时间不变协变量后，国家新区设立对区域经济增长的拉动作用在 95%置信水平上依然

显著为正。

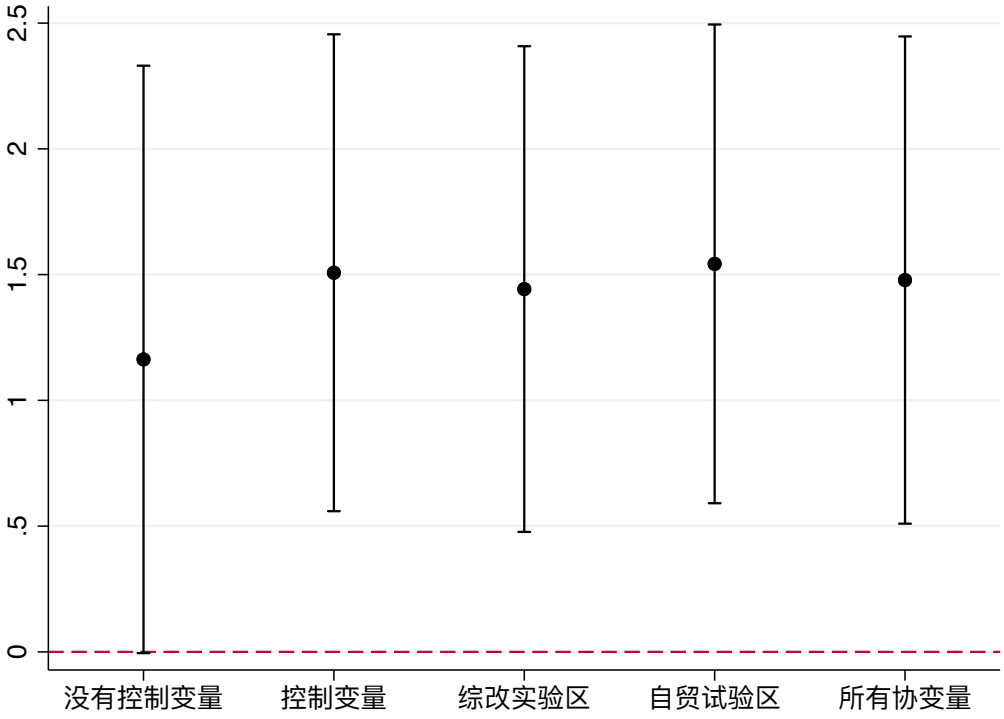

图 9 控制可观测协变量

2、不可观测混淆因子检验

控制了所有应该控制的可观测协变量后，就可以识别出因果效应β_k吗？Angrist and Pischke（2015）指出："仔细考察遗漏变量偏误是应用计量经济研究的必要组成部分。"正因为如此，经济学研究者都希望采用定量方法来评估遗漏变量对其实证结果的影响，尤其是不可观测遗漏变量的重要性（Diegert et al.，2022）。目前，应用经济学研究中最广泛使用的遗漏变量评估方法是 Oster（2019）法。

从经验来看，Oster（2019）建议$|\delta| > 1$时，有少量的不可观测混淆因子可以解释结果的变动（Bazzi，2020）。从表 2 的结果可知无论是加入可观测的协变量，还是排除其它政策试点的干扰，$|\delta| < 1$意味着可能还有一些重要的遗漏变量驱动区域经济增长。

表 2 遗漏变量偏误的定量评估

	无控制变量	控制变量	综改试验区	自贸区	所有协变量
Oster（2019）方法					
Oster δ for $\beta = 0$		-0.49	-0.47	-0.49	-0.48
R^2	0.061	0.692	0.693	0.693	0.694

但是，需要注意的是，Oster（2019）法基于外生控制变量假设——遗漏变量与回归模型中包含的所有控制变量不相关。例如，如果城市的宗族文化是国家新区经济增长效应的重要遗漏变量，那么，Oster（2019）方法就假设城市宗族文化与投资、消费、二产占比、创新、国家层面去为导向政策等控制变量无关。但我们在识别因果效应时，只需要关心遗漏变量是否与处理变量相关，而并不在意遗漏变量是否与控制变量相关。而且在应用经济研究中，我们很难排除遗漏变量与控制变量无关，例如，宗族文化与创新相关（薛胜昔等，2021；朱郭一鸣和尹俊，2021）。最近，Diegert et al.（2022）就放松了外生控制变量的假设，进而评估不可观测遗漏变量的重要性。

除此之外，在 DID 事件研究设计中，还有一种特有的不可观测混淆因子检验方式——“波浪式检验（wiggly test）”。在面对不可证实的因素时，经济学者通常采用证伪的方式来进行推断。在某些情形下，用于完全解释结果的事件时间路径的一些混淆因子从经济现实和理论上均不可信。Freyaldenhoven et al.（2022）提出在事件研究图中增加最可信的混淆因子——统计上与估计的结果变量事件时间路径相一致——的标识线，用来评估混淆因子的可信度。经验上来说，相比于更多“波浪”的曲线，越平滑的趋势线表示更可能存在混淆因子。

如图 10 所示，左图中估计的事件时间路径有更加平滑的趋势线，且接近于线性的趋势，从处理前一直持续到处理后。在许多经济环境中，这可能意味着存在混淆因子，因此，处理后的政策效应可能是由于混淆因子趋势产生的。虽然，左图的处理前趋势检验 Wald p 值为 0.13，在 95%的置信水平下不能拒绝“无处理前趋势”的假设，但这并不意味着没有混淆因子（Roth，2022）。而且正如 Roth（2022）统计的传统事件研究应用文献，没有一篇文献报告、讨论处理前时期系数的大小，即处理前趋势的程度也非常重要。因此，处理前趋势线也可以让经济研究者定量评估处理前是否存在趋势。

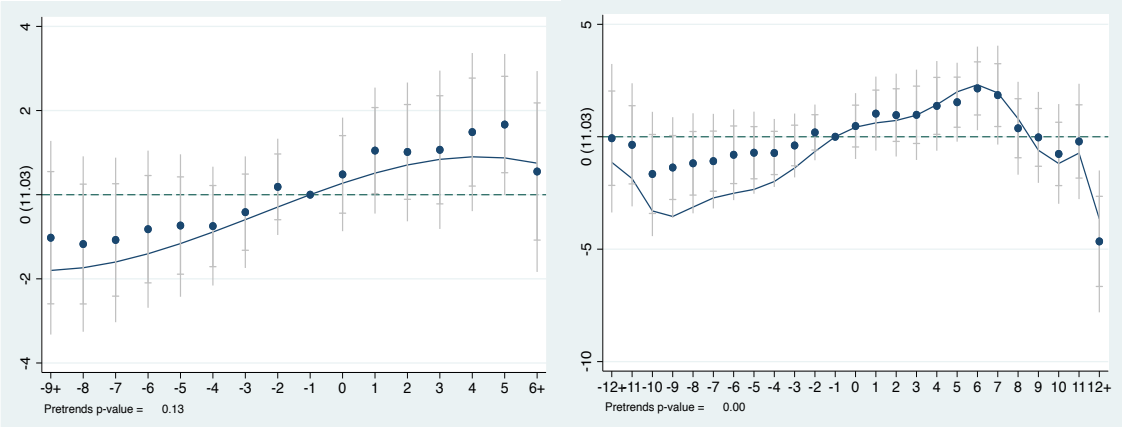

图 10 “波浪式”混淆因子检验

3、不可观测混淆因子的趋势

（1）简单的时间趋势

在 DID 事件研究中忽略时间固定效应，增加时间趋势项：

$$Y_{i,t} = \alpha_i + \alpha_t + \sum_{k=-(L-1)}^{K-1} \beta_k D_{i,t+k} + X_{it}\Gamma' + \phi_i f(t) + \epsilon_{i,t} \quad (2)$$

最简单的时间趋势项就是$f(t) = t$，结果如图 8 所示。当然，时间趋势项也可以是时间的多项式。

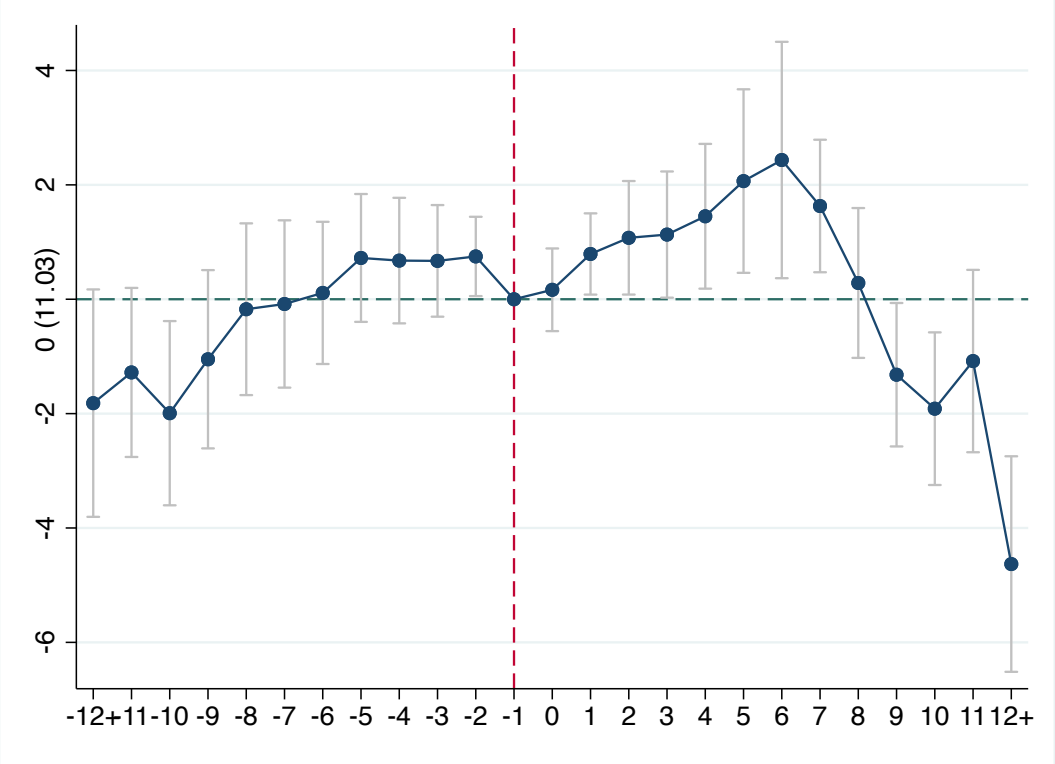

图 11 控制线性时间趋势的事件研究图

（2）未知的时间趋势

时间趋势项$f(t)$的函数形式未知。此时，可以使用 Bai(2009)提出的交互固定效应估计量，Pesaran(2006)提出的共同相关效应估计量，或者 Abadie et al.（2003,2010,2015）提出的合成控制法。

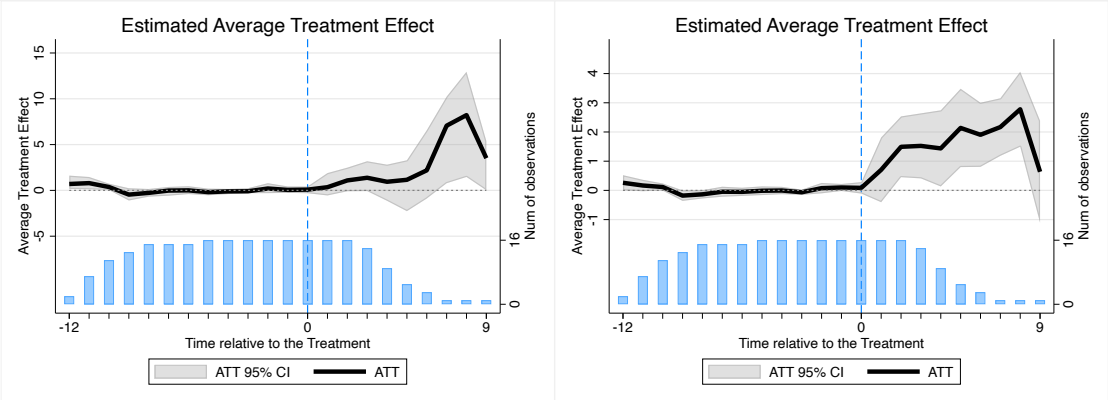

图 12 Xu(2017)提出的广义合成控制法和 Athey et al.(2021)提出的矩阵完成法

（3）相对事件时间趋势：异质性时间趋势

可以使用处理前数据进行趋势外推来降低此类混淆因子的影响，如图 10 所示。

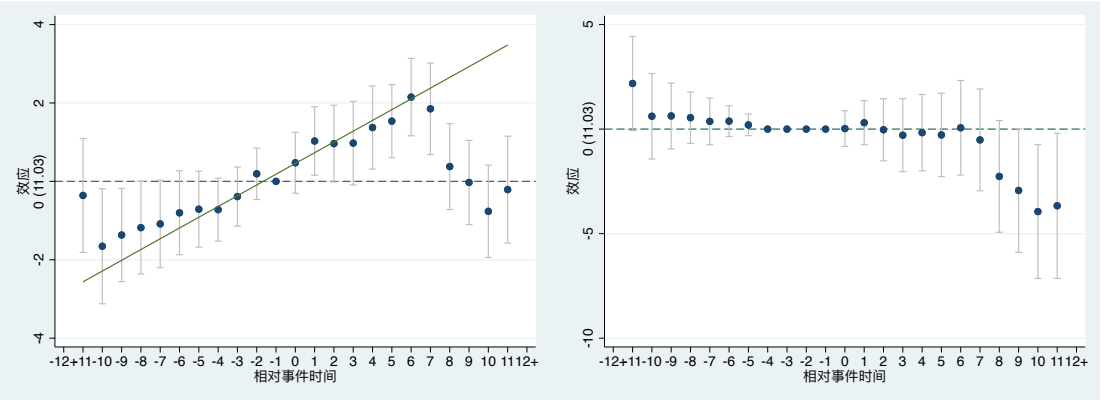

图 13 趋势外推后的事件研究图

（4）不可观测混淆因子的代理变量

如图 14 所示，如果 X 是不可观测的混淆因子，要想控制它，我们可以寻找变量 X 的原因变量 X1 或者结果变量 X2 来作为其代理变量。

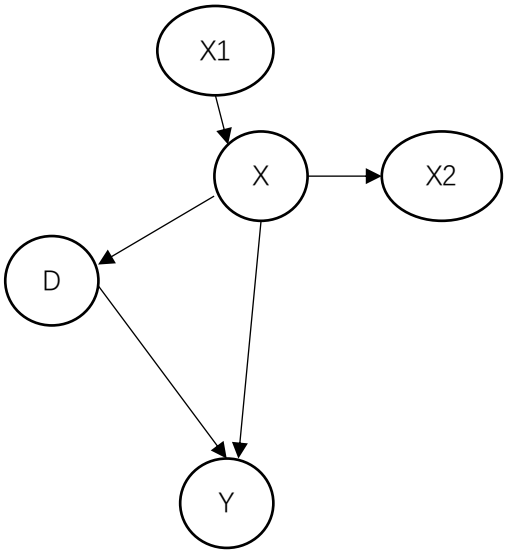

图 14 不可观测混淆因子的代理变量

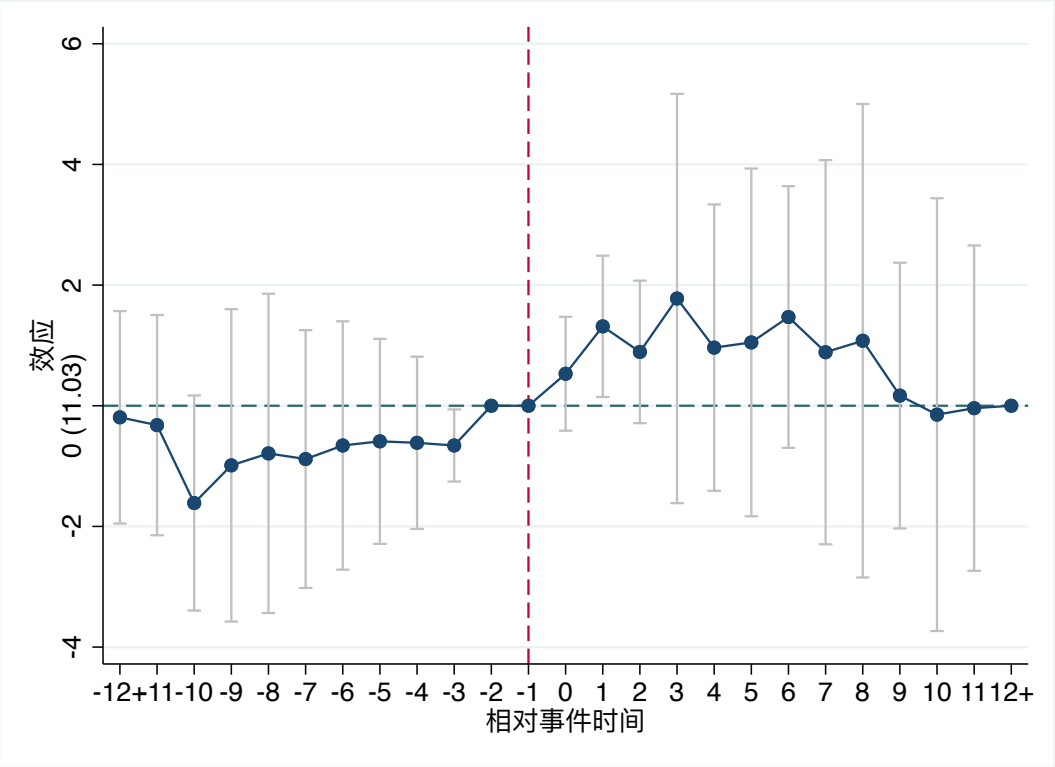

图 15 不可观测混淆因子的代理变量

4、处理变量的工具变量

Biasi and Sarsons（2022）以美国公立学校老师为对象，实证分析了灵活工资制度改革对男女教师工资差异的影响，因果关系图 1。灵活工资制度的实施受到学期 CBAs 到期和 CBAs 延期的影响。而 CBAs 延期是学区自主考量的选择。也就是说，只有当影响 CBAs 延期选择的学区因素不影响学区性别工资差异，即因果关系图 1 中的红色虚线不存在时，作者做出的上述 TWFE 事件研究结果才是可信的。作者在文中使用的原话为：

“在我们的分析中，使用了灵活工资制度的差异化引入试点，而灵活工资制度的引入又是由 CBAs 到期和延期决定的。尽管只有 CBAs 到期日可以视作随机的，但只要引发学区选择 CBAs 延期决策的因素与性别工资差异无关，我们的识别策略仍然可以估计出灵活工资制度的效应。”

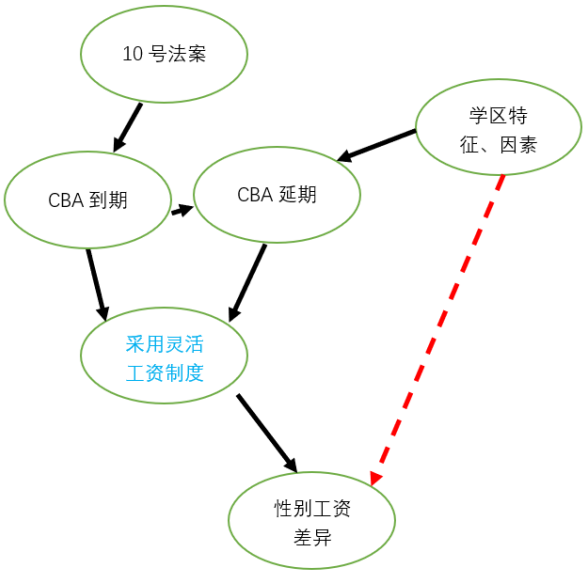

图 15

但要想控制所有与 CBAs 延期决策和性别工资差异同时相关的学区混淆因子谈何容易，即使我们可以控制所有的可观测混淆因子，那不可观测的混淆因子怎么办？这些因素都会导致灵活工资制度并不具有随机性，而 DID 等准自然实验方法都是基于政策/处理的某些随机性。作者也担心这些问题，但是又找不到很好的办法来控制这些混淆因子，因此，他们用了一个技巧：(1) 既然 CBAs 的到期日是随机的，那么，我们可以直接使用 CBAs 到期前后来作为政策变量，重新跑 DID 事件研究，得到的结果如表 1 的第三、四列所示；(2) 既然 CBAs 到期后，学区才能决定是否延期，那么，可以将 CBAs 到期变量作为 CBAs 延期的工具变量来跑 IV 回归，得到的结果均显示，在 90%置信水平下，灵活工资制度确实会恶化性别工资不平等状况，且差异会随着时间以来越大。

五、结论与实践建议

参考文献

附录

With 综合改革实验区

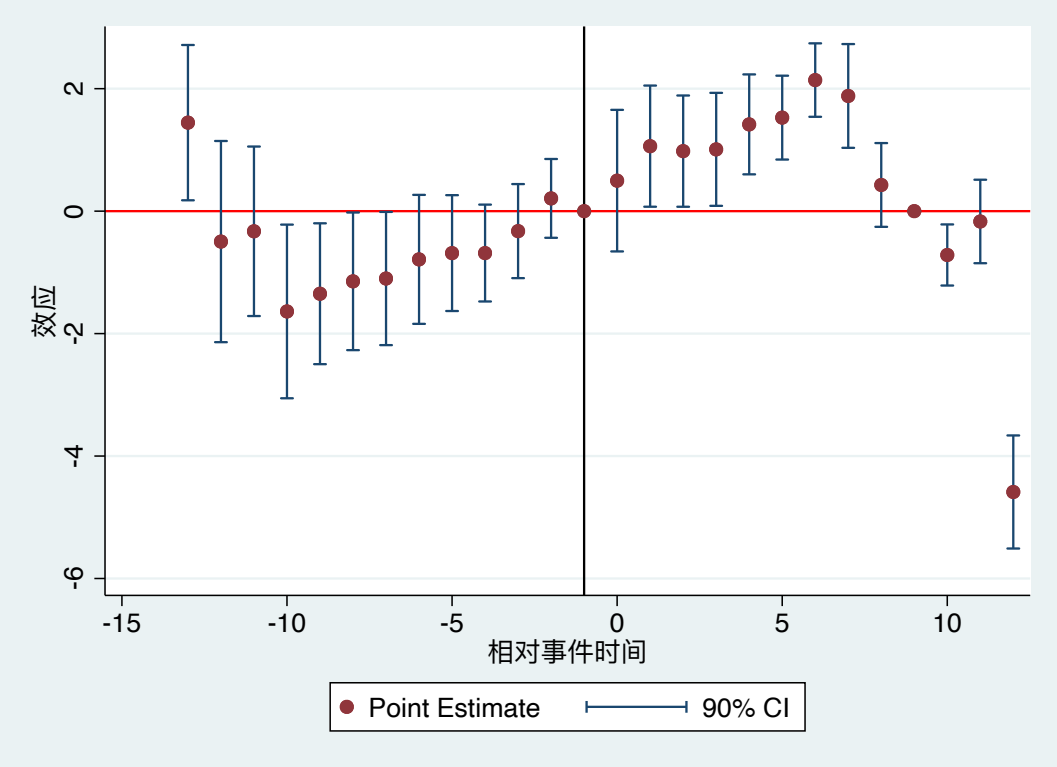

图 7

With 自贸区

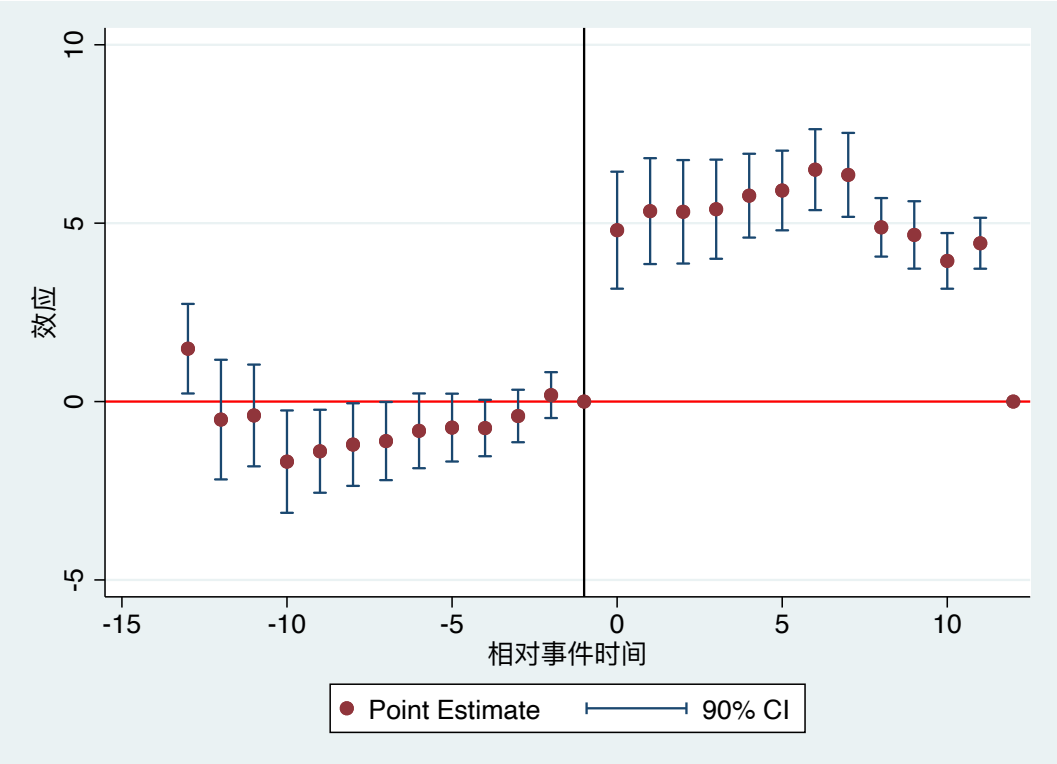

图 8

With 试验区和自贸区

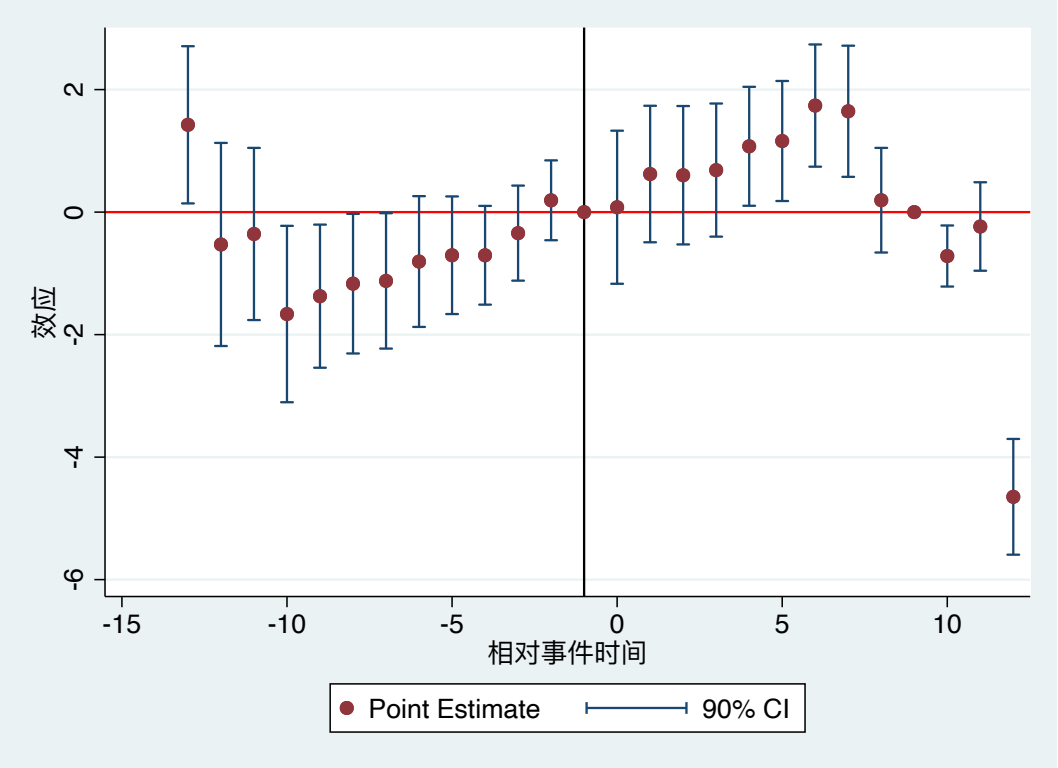

图 9