许文立:《双重差分法的最新理论进展与经验研究新趋势》,《广东社会科学》, 2023, 05

《双重差分法的最新理论进展与经验研究新趋势》的附录

未在正文中显示的内容:

附录 A：双向固定效应估计量分解

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：

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

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（A-1）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）和 Sun and Abraham（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（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%。[①]

附录 B：潜在偏误的诊断

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，其实，这两个问题是同一“硬币的两面”。在经验研究中，对于它们的诊断有如下方法：

第一种方法是回归法。Jakiela（2021）提出用回归余值来诊断负权重或者异质性处理效应，并给出两个经验建议：① 根据（A-1）式，权重$\frac{\sum_{it}\tilde{D}_{it}}{\sum_{it}\tilde{D}_{it}^2}$的符号与处理变量的余值$\tilde{D}_{it}$相同，那么，在诊断负权重的时候，可以用处理变量对个体固定效应和时间固定效应或者协变量回归，计算出回归后处理变量的余值。然后将回归后处理变量余值作分布图，观察正负权重的分布。如果负权重的占比较大，那么用双向固定效应模型估计平均处理效应可能带来较大偏误。② 进一步检验同质性处理效应，即除了得到处理变量的余值，还可以用结果变量对个体、时间和协变量进行回归计算结果变量的余值。然后用余值化结果变量对余值化处理变量、控制组虚拟变量、余值化处理变量和控制组虚拟变量交乘项进行回归。我们关注的是交乘项的系数，如果这个系数显著，就拒绝了同质性处理效应假设，说明直接用双向固定效应估计量可能带来偏误。这种方法比较灵活，适用的数据环境非常广泛。

第二种方法是培根分解。Goodman-Bacon（2021）指出，总的双向固定效应估计量等于三类 2×2 DID 估计量加权和。当第三种（“已处理 vs 新处理类型”）的 DID 估计量与另两种差别较大，且权重较大时，总的双向固定效应估计量就越容易受到污染，本文将其称为“培根分解”。Goodman-Bacon 和其他作者一起开发了一个 stata 命令 bacondecomp 来分解三类 DID 估计量和对应的权重，并给出每个 2×2 DID 估计量和权重的散点图。但是这个命令只适合于强平衡面板数据的情形。命令的格式为：

```
bacondecomp outcome treatment,ddetail
```

第三种方法是 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（2020）开发了一个 stata 命令 twowayfeweights 来分解样本中正权重和负权重的样本数量以及对应的权重之和，本文将其称为“CD 分解”。这个命令使用环境也非常广泛，非常灵活，对于重复截面、平衡面板和非平衡面板的数据结构都适用。命令格式为：

```
twowayfeweights outcome groupid timeid treatment, type(feTR)
```

第四种方法是 SA 分解。Sun and Abraham（2021）开发了一个 stata 命令

[①] 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$D_i + D_t \leq 1$，进入求和的个体-时间的$D_{it} = 1$。因此，当没有个体在大部分时间里被处理，且不存在特定的时期里大部分个体被处理的时候，所有的权重就都为正。

eventstudyweights 来计算事件研究回归系数的权重，本文将其称为“SA 分解”。它的基本命令格式为：

eventstudyweights frel time listg, absorb(i.groupid i.timeid) cohort(first treatment) rel time(ry)

第五种方法是静态效应检验。研究者通常会在事件研究图中展示效应以判断政策效应是否具有异质性动态，即异质性处理效应（de Chaisemartin and D’ Haultfoeuille，2021）。Freyaldenhoven et al.(2022)建议使用静态模型约束来检验“政策效应是静态”的假设，即通过一个 Wald 检验的 p 值和静态效应与动态效应系数置信区间的比较来诊断是否存在异质性处理效应。但是，de Chaisemartin and D’ Haultfoeuille（2020）也指出，除了时间维度的异质性处理效应会带来偏误外，个体维度的异质性处理效应也会使得双向固定效应估计量存在很大偏误。因此，SA 分解和静态效应检验法不能诊断个体维度异质性处理效应。

附录 C：更多稳健估计量的比较与应用，如表 C-1 所示：

表 C-1 异质性处理效应稳健估计量的比较及其 stata 命令包

估计量/作者	软件包	可利用的对照组	处理期的时期数	处理类型	数据结构	参数/非参数估计	时变协变量	应用文献
de Chaisemartin 和 D'Haultfoeuille（2018，2020，2021，2022）	Fuzzy_did / did_multiplegt / did_multiplegt_dyn	上一期刚处理	处理前两期	二值型、连续型（w/o stayers）、多个处理、交叠、多个处理值、进入-退出	（非）平衡面板和重复截面	参数	是	Gross et al.（2021，AER）、Cantoni and Pons（2022，QJE）、Braghieri et al.（2022，AER）、许文立和孙磊（2023，数量经济技术经济研究）、田淑英等（2021，财政研究）
Imail 和 Kim(2021)				二值型、交叠、进入-退出		非参数		
Borusyak 等（2023）	did_imputation	从未处理/还未处理	处理前多个时期	二值型、交叠、三重差分	面板和重复截面	参数	是	Biasi and Sarsons（2022,QJE）、

								Braghieri et al.（2022, AER）
Sant'Anna 和 Zhao（2020）	drdid			非交叠	面板/重复截面	参数		
Callaway 和 Sant'Anna（2021）	csdid	从未处理/还未处理	处理前一期	二值型、交叠	平衡面板和重复截面	参数	否	Ang（2021, QJE）、Hu et al.(2023,JRE)、Braghieri et al.（2022, AER）
Sun 和 Abraham（2021）	eventstudy interact	上一期刚处理		二值型、交叠	面板	参数		Miller et al.（2021, QJE）、Cantoni and Pons（2022, QJE）、Braghieri et al.（2022, AER）
Gardner（2021）	did2s			二值型、交叠		参数		
Clarke 和 Tapia-Schythe（2020）	eventdd			二值型、交叠		参数		
Dettmann 等（2022）	flexpaneldid			二值型、交叠、可变处理时期	面板	非参数		
Wooldridge（2021）	jwdid			二值型、交叠、非线性回归		参数	否	

Arkhangelsky 等（2021）	sdid			二值型、交叠	平衡面板		是	
Cengiz 等（2019）	stackedev	“干净”的控制组（每一个都有不同的固定效应）		二值型、交叠				
Freyaldenhoven 等（2022）	xtevent			二值型、交叠、连续型、不可观测混淆因子			是	
Goldsmith-Pinkham 等（2022）	multi			二值型、相互排斥的多处理				
stata 18	xthdidregress							

附录 D：平行趋势假设的类型

正如上文所述，平行趋势假设是 DID 估计量能干净识别平均处理效应的关键。平行趋势假设是不可检验的，但是在不同的处理效应估计量中，对平行趋势假设的理解和使用存在一些差异。也就是说，平行趋势假设存在几种类型：

类型 I：处理前处理组和控制组的结果均值差异在没有发生处理时会在处理后处理组和控制组结果均值间延续。

类型 II：处理前某一时点处理组和控制组的结果差异在没有发生处理时会在处理后某一时点处理组和控制组结果间延续。

类型 III：处理前处理组和控制组的结果均值差异在没有发生处理时会在处理后某一时点处理组和控制组结果间延续。

类型 IV：处理前某些时点处理组和控制组结果差异在没有发生处理时会在处理后某些时点处理组和控制组结果间延续。

类型 V：处理前处理组和控制组的所有结果差异在没有发生处理时都会延续到处理后每个时点。

类型 I 的平行趋势假设是最为人所熟知的，类型 II 则只要求处理前后一个时点的平行趋势延续即可，例如 Callaway and Sant Anna（2021）估计量，类型 III 则要求处理后一期延续处理前多期的平均趋势即可，例如 de Chaisemartin and D'Haultfoeuille

(2020,2021,2022)估计量，类型 IV 与类型 II 的区别在于，类型 II 要求处理前后各一个时点，而类型 IV 则可以放宽至处理前后多个对应时点满足平行趋势即可，例如 Borusyak et al. (2023)估计量，类型 V 是最严格的平行趋势假设，因为它要求处理前后所有时点反事实结果平行地变化。

附录 E：引入协变量的方式

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既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）。从目前的应用研究实践来看，在回归方程中引入协变量的方式如图 E-1 所示。

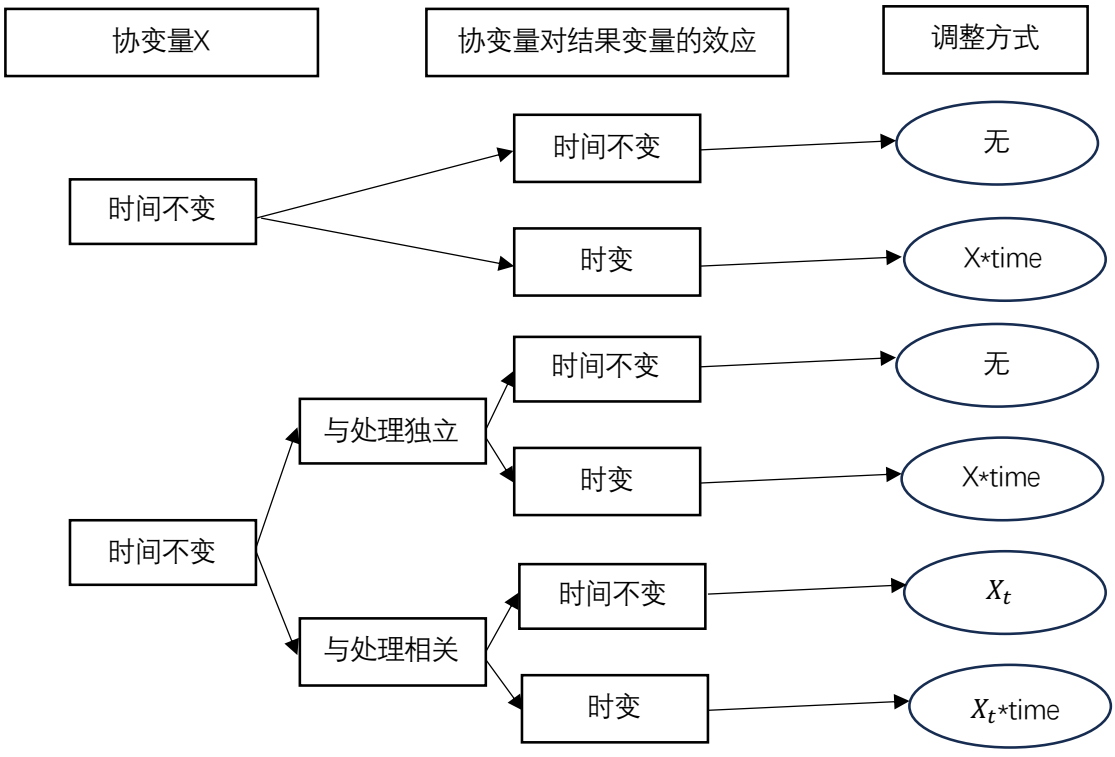


图 E-1 回归方程中引入协变量的方式[2]

附录 F：不可观测遗漏变量的检验

在 DID 回归模型中，研究者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

[2] 来源：https://diff.healthpolicydatascience.org/#confounding。

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。

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）。

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。但在识别因果效应时，研究者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，研究者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。

附录 G：案例应用[3]

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

[3] 相关 stata 数据和代码，请见许文立老师的个人主页：https://wenddymacro.github.io/Wenddy-XU/。

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。[4]

（1）双向固定效应模型

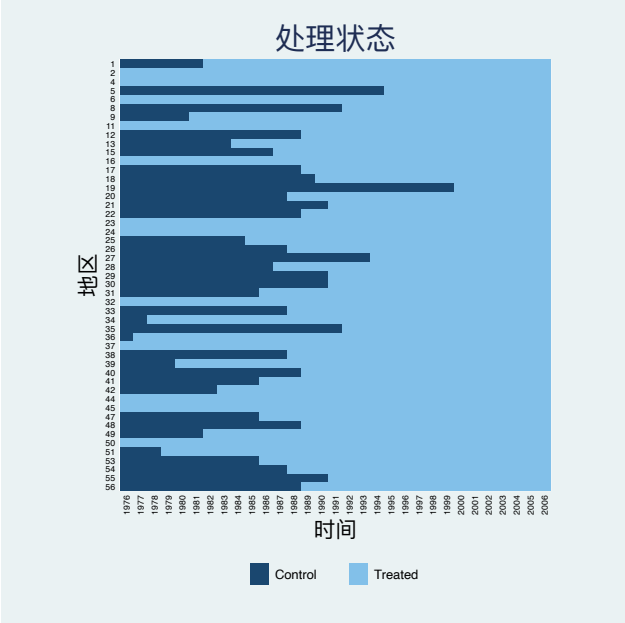
BLL（2010）声明了一个二值型处理变量的双向固定性效应模型来评估银行分支管制解除对收入分配的效应，模型设置如下：

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

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（G-1）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。

[4] 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（G-1） 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（*G-1*）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]。

[5]表（*G-1*） 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

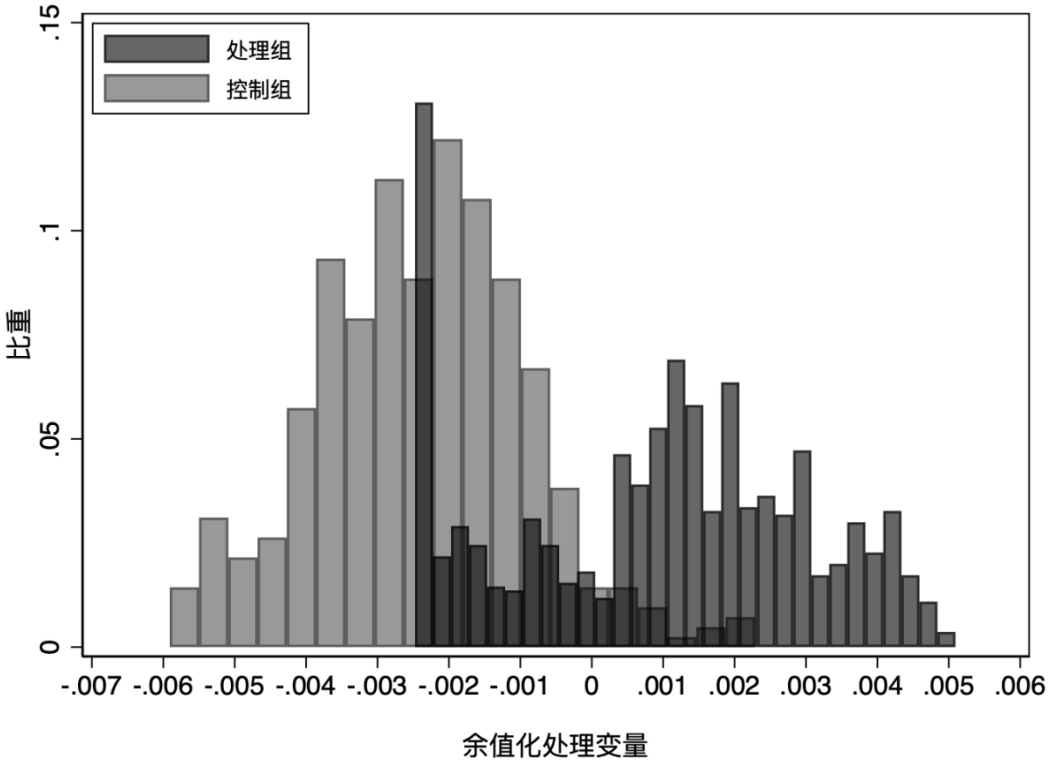
从表（G-1）的复现结果可以看出，无论是否加入控制变量，银行分支机构管制放松都显著为负，且十分稳健。因此，作者得到的结论是，放松对银行跨洲际分支机构管制会降低收入分配不平等水平。以“基尼系数逻辑斯谛克转换 logit(gini)”为例，放松管制使得 logit(gini)下降 3.9%，这一数值在经济意义上非常大。作者进一步将回归系数估计值与 logit(gini）系数

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

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（*G-2*）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（*G-2*）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。



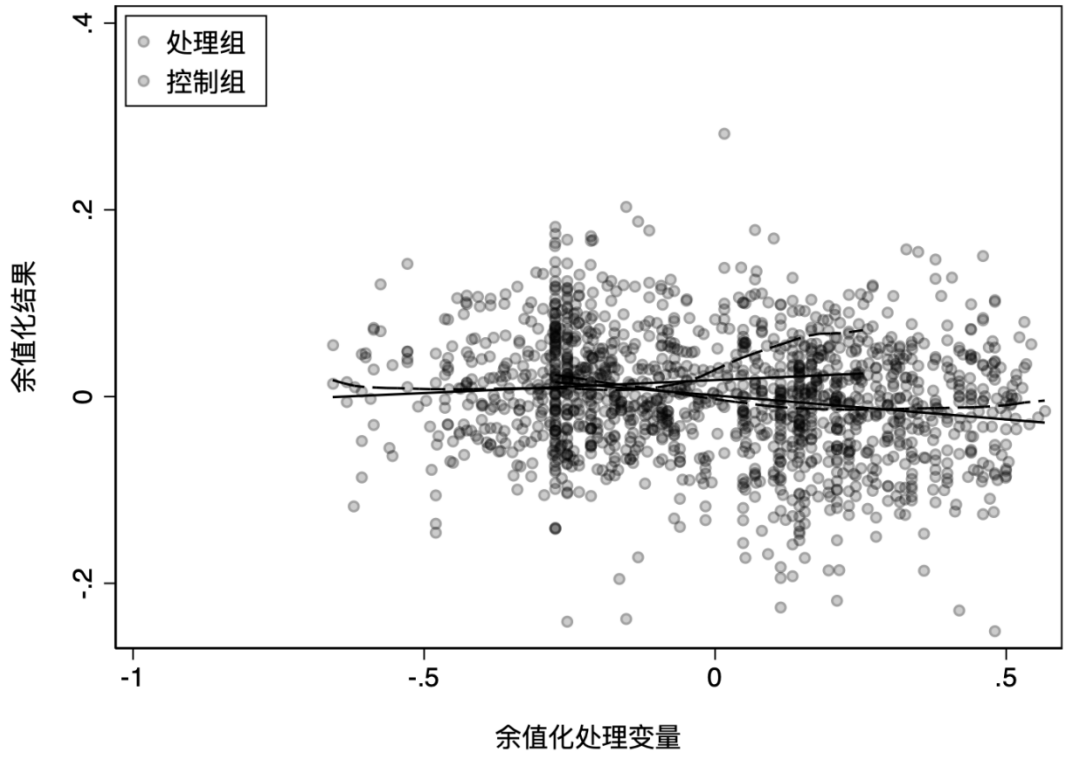
图（*G-2*） 负权重分布图

表（*G-2*）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（*G-3*）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。

[6]表（*G-2*） 余值化结果与余值化处理变量之间线性关系的检验

[6] ***表示 1%，**表示 5%，*表示 10*，括号中为标准误

	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（*G-3*） 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:

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

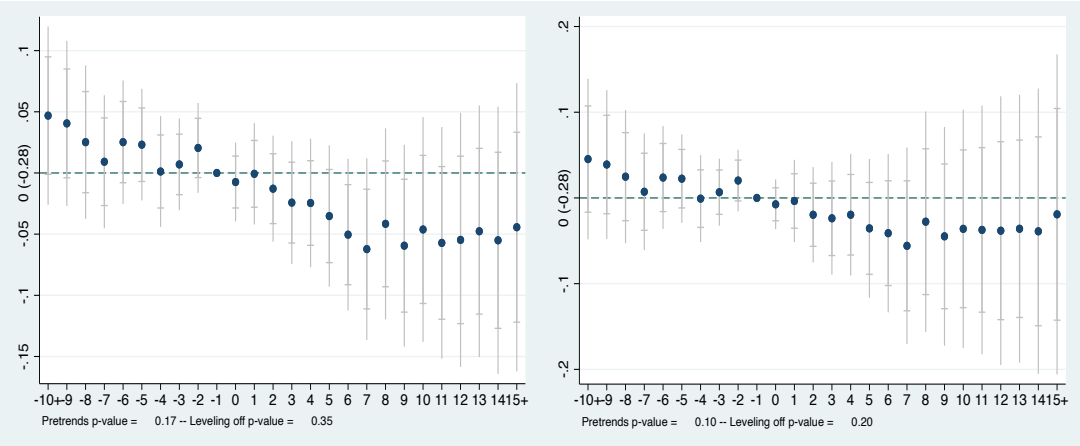
其中，$D_{s,t-m}$表示地区 s 是否在时点 t 前 m 期放松了银行分支机构管制的二值变量。$\sum_{m=-G}^{M} \beta_m D_{s,t-m}$表示放松银行分支机构管制的动态效应，时点 t 的地区不平等最多只能被 t 前的 M 期和 t 后的 G 期的管制放松政策所影响。参数集$\{\beta_m\}_{m=-G}^{M}$包含了这些动态效应的大

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。



[7]图（*G-4*） 事件研究图

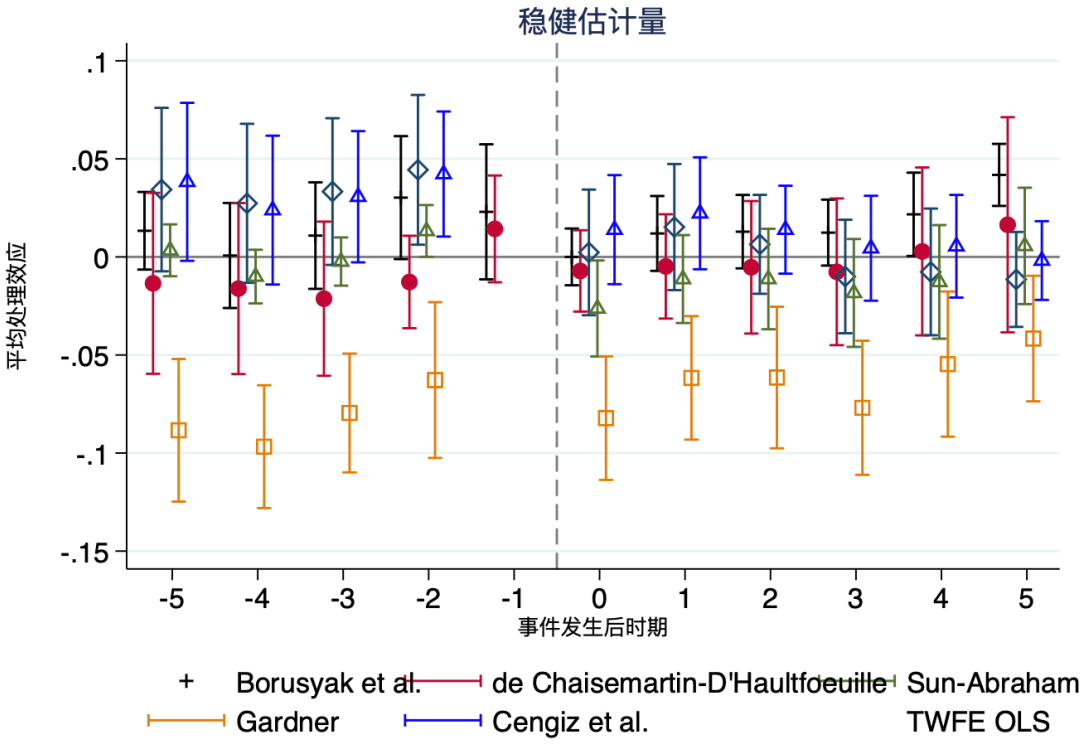
事件研究结果如图（*G-4*）所示。左图为使用 BLL（2010）全部样本的事件研究图，右图为删掉了 1977 年以前就已经放松银行分支管制地区的样本后的事件研究图。从图（*G-4*）（左）可以看出，在放松银行分支机构管制前的时期，放松管制前时间虚拟变量的系数均不显著，且 90%的置信区间下不能拒绝“没有处理前的趋势”这个原假设。这意味着处理组和控制组并没有显著的差异化趋势。且放松管制后的 6、7、9 年地区不平等有所缓解。

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（*G-4*）（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（*G-5*）所示。

[7]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。

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（*G-5*） 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）。

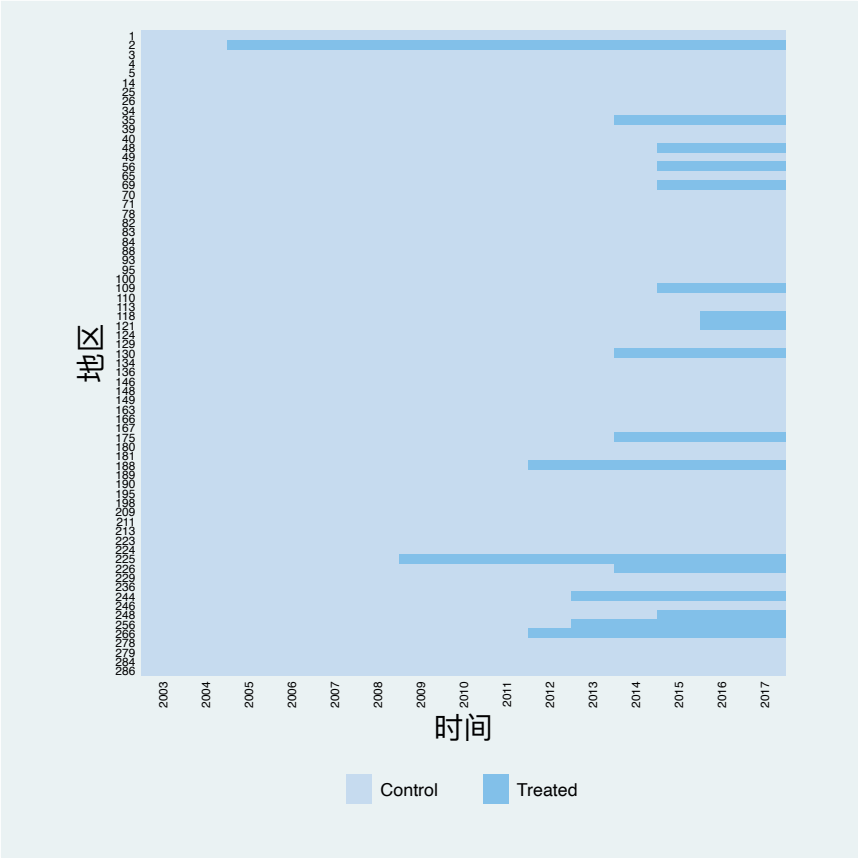
（1）结果复现

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（*G-6*）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（*G-6*） 国家级新区设立时间

表（*G-3*）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。

[8]表（*G-3*） 国家级新区对区域经济增长的效应

	（1）	（2）

[8]括号中为标准误：* p<0.10，** p<0.05，*** p<0.01

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

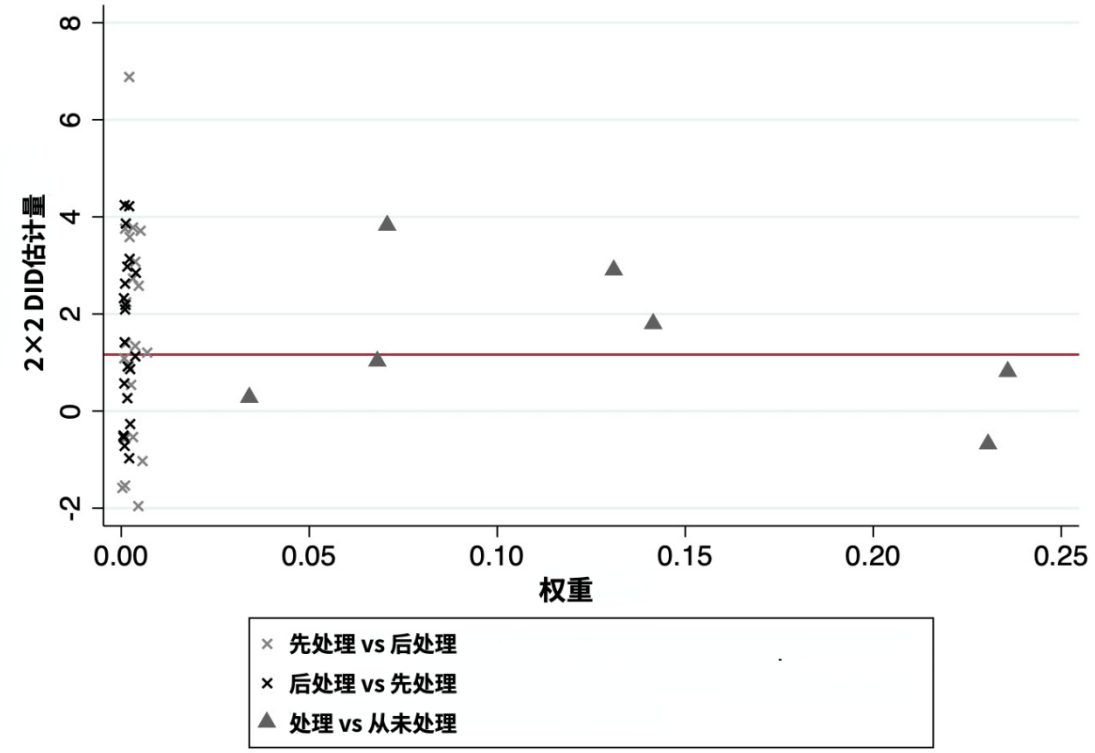
下面，用 Goodman-bancon (2021) 提出的诊断方法来将总的 DID 估计量分解为三组：（1）“先设立国家级新区的城市 vs 后设立国家级新区的城市”；（2）“后设立国家级新区的城市 vs 先设立国家级新区的城市”；（3）“设立国家级新区的城市 vs 从未设立国家级新区的城市”。表（*G-3*）中得到的总的 DID 估计量等于每一组的平均 DID 估计量乘以各自权重之和。如表（*G-4*）所示，培根分解给出的无控制变量时总的 DID 估计量与表（*G-3*）的结果相同，且 1.163=0.057×1.571+0.031×1.659+0.912×1.120。从表（*7*）结果可以进一步看出，“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。

[9]表（*G-4*） 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

下面来看看 Bacon 分解图，如图（*G-7*）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。

[9] vs 前后分别表示处理组和控制组



图（*G-7*） 无控制变量的培根分解

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。

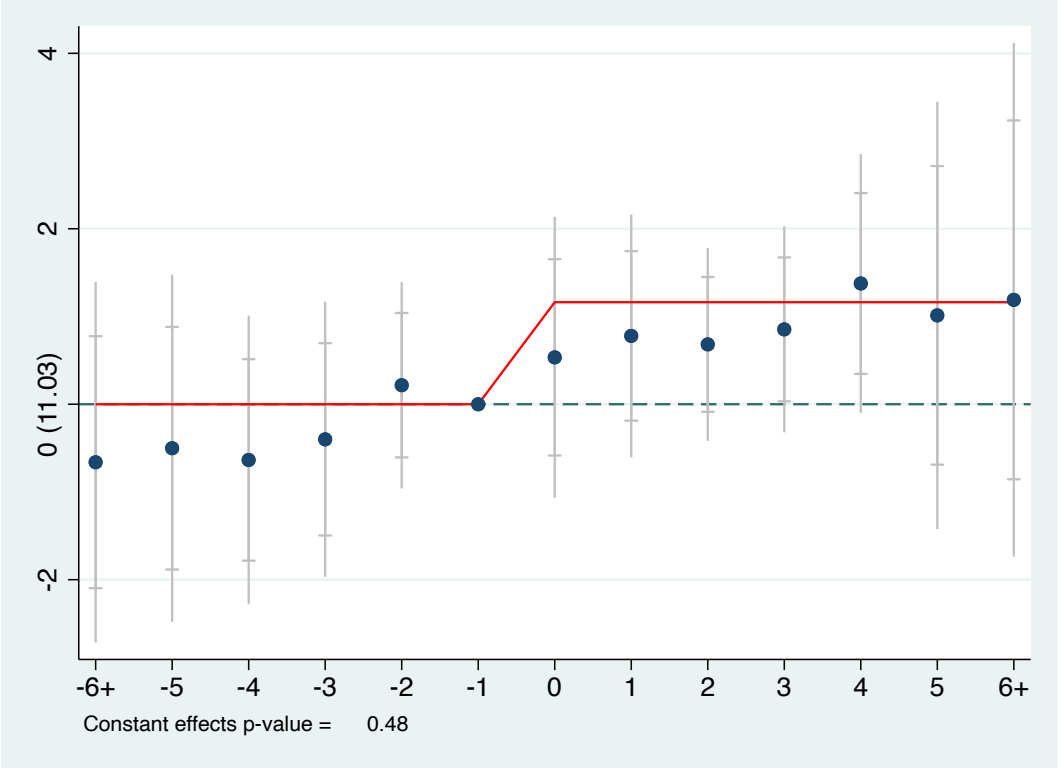


图 G-8 事件研究图：静态效应检验

如图 G-8 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。

(5) 平行趋势检验：预期效应、平行趋势假设检验、平行趋势敏感性检验

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，$\mathrm{did}_{\mathrm{i,t-k}}$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。$\sum_{\mathrm{m=-G}}^{\mathrm{M}} \beta_{\mathrm{m}} \mathrm{did}_{\mathrm{i,t-m}}$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$\{\beta_{\mathrm{m}}\}_{\mathrm{m=-G}}^{\mathrm{M}}$包含了这些动态效应的大小。

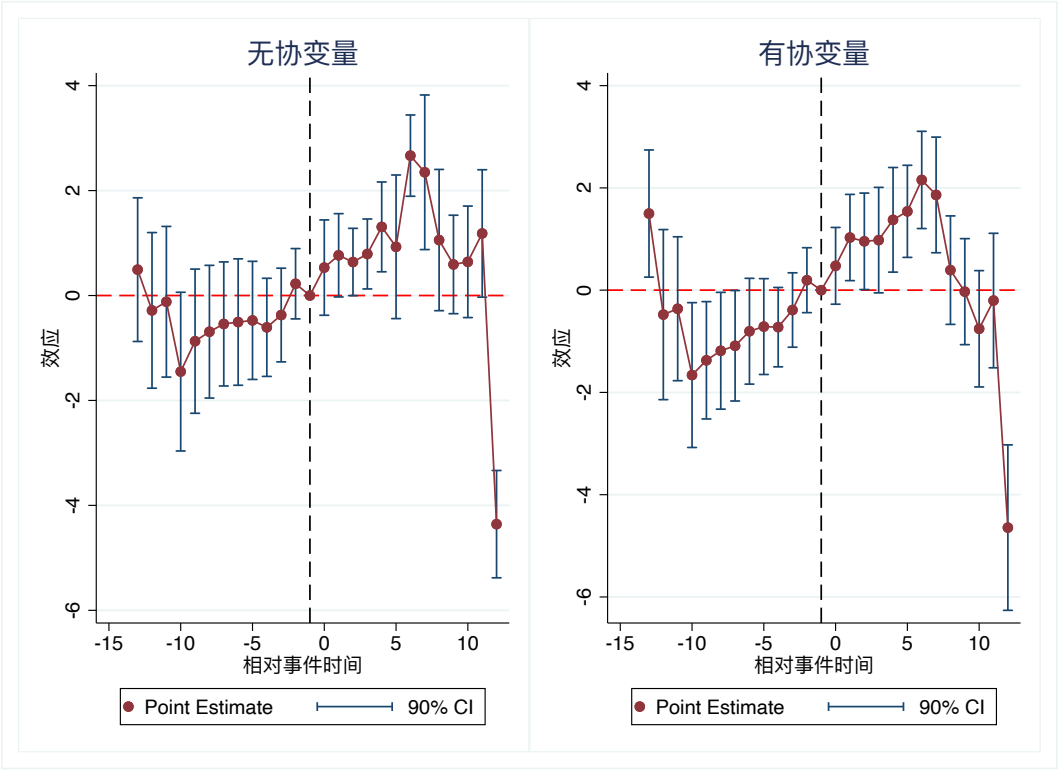
通常，经济学家更加关心政策的累积效应，即不同时期 k 的$\sum_{\mathrm{m=-G}}^{\mathrm{k}} \beta_{\mathrm{m}}$，以及政策影响时期外的累积政策效应。因此，采用 Simon Freyaldenhoven et al.(2021)对于面板数据事件研究设计的模型设定与事件研究图的建议。将上述动态处理效应回归模型变形如下：

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

其中，$\mathrm{did}_{i,t-k}$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，$(1-\mathrm{did}_{i,t+G+L_G})$表示城市 i 在 t 时点后是否仍有国家级新区，$\mathrm{did}_{i,t-M-L_M}$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 G-9 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 G-9 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。



[10]图（*G-9*） 国家级新区对区域经济增长的动态拉动效应

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

[10]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。

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 G-9 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）

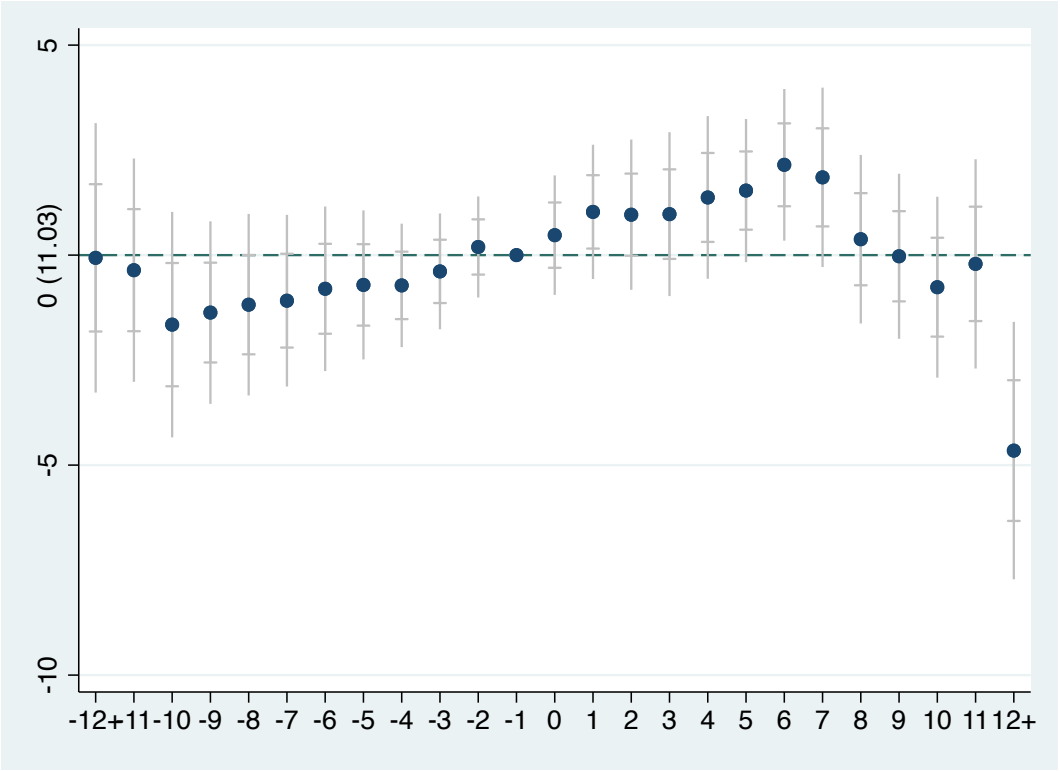


图 G-10 事件研究图：联合显著性检验

如图 G-10 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 G-10 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 G-11 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。

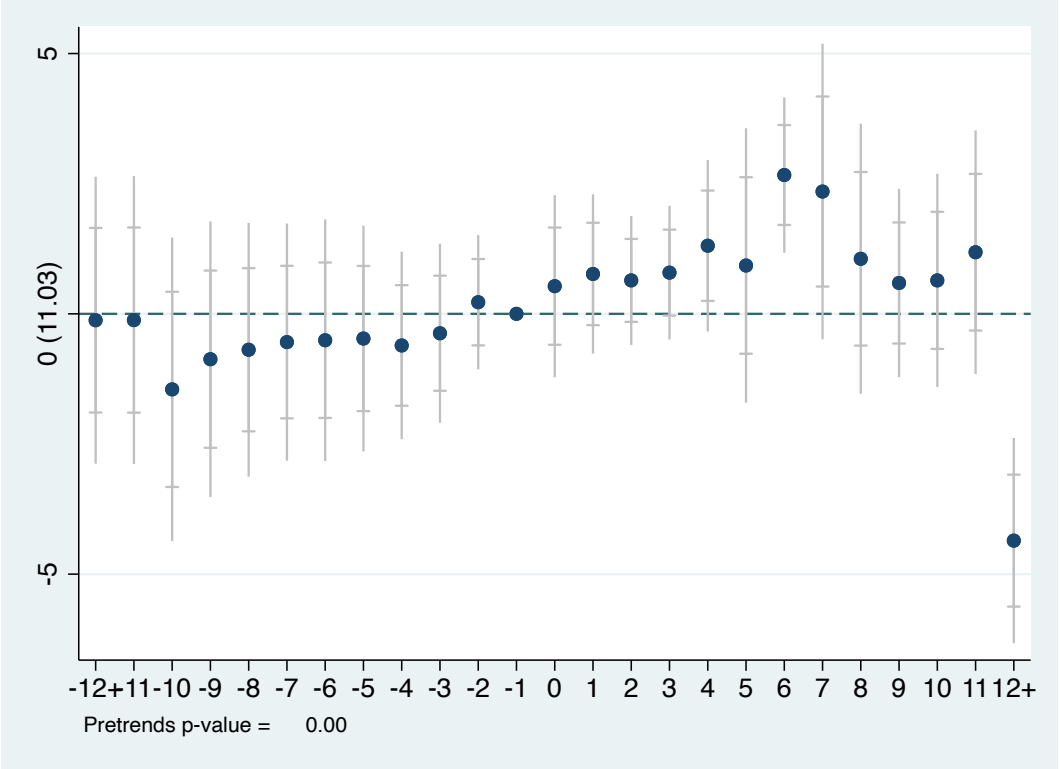
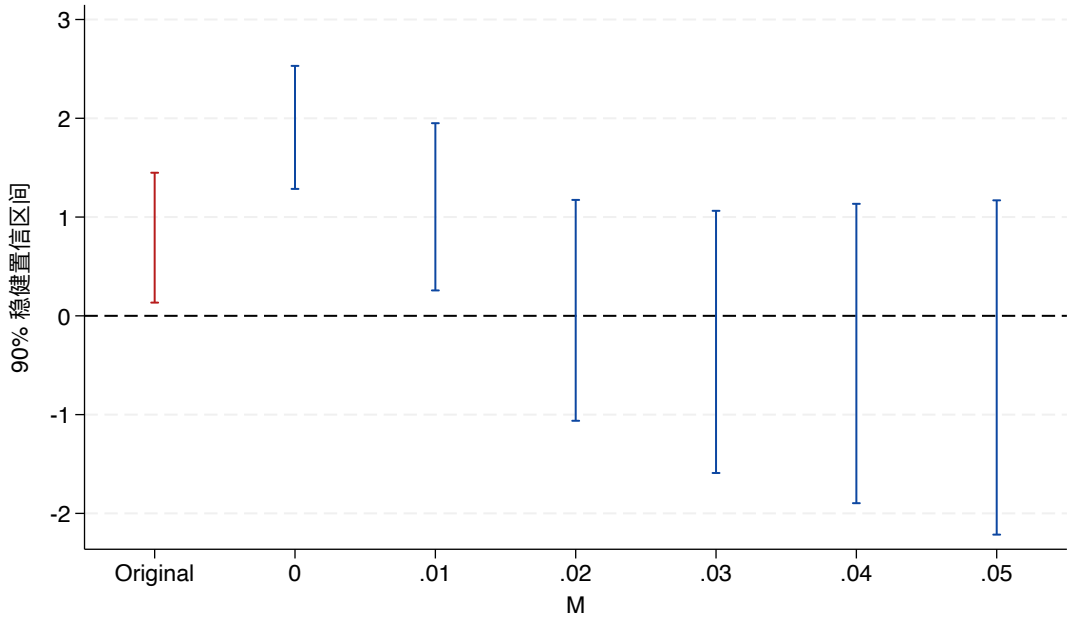


图 G-11 事件研究图：无处理前趋势检验

下面，我们来看看平行趋势的敏感性检验。

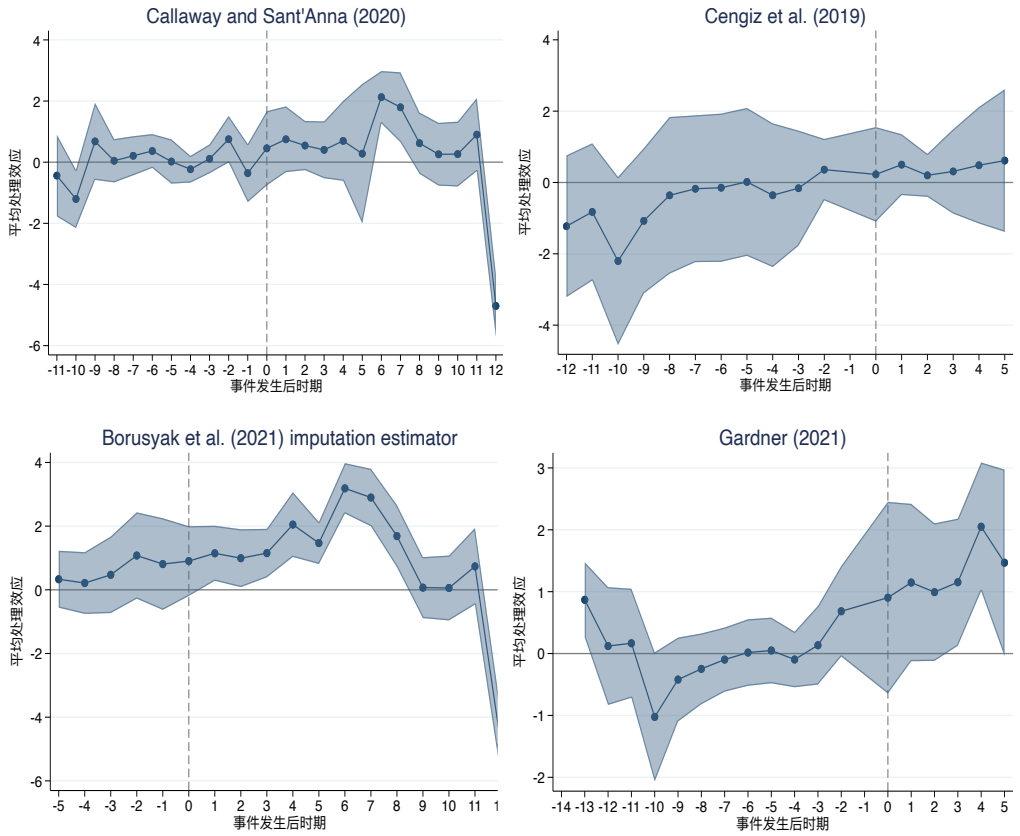


G-12 平行趋势敏感性检验

异质性处理效应稳健估计量

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（*G-13*） 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。

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 G-14 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。

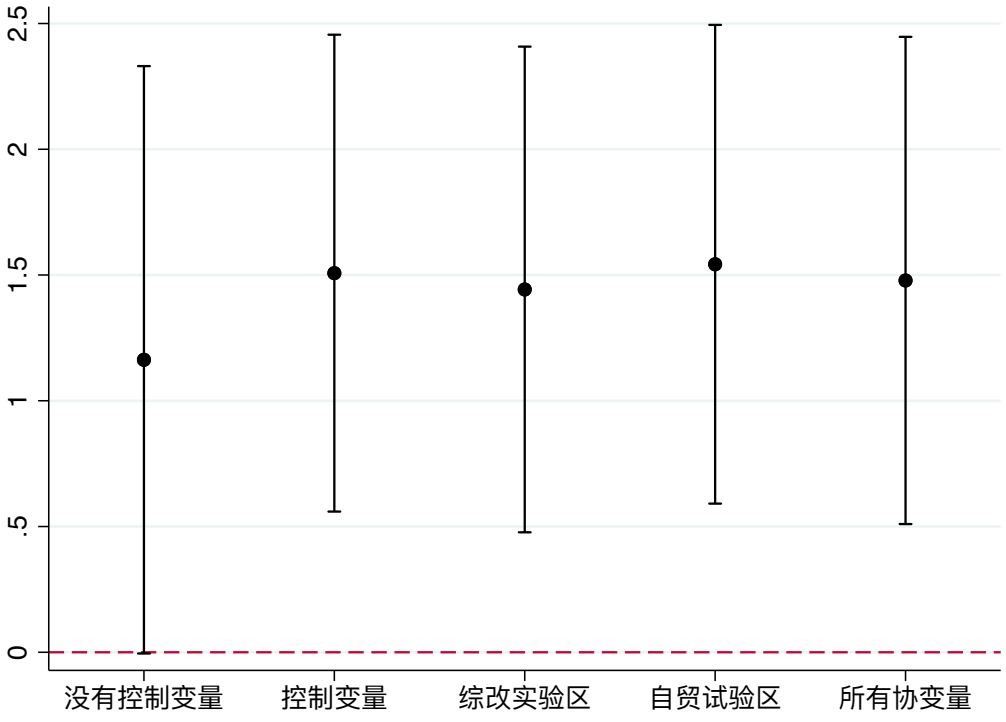


图 G-14 控制可观测协变量

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 G-4 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 G-4 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 G-15 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。

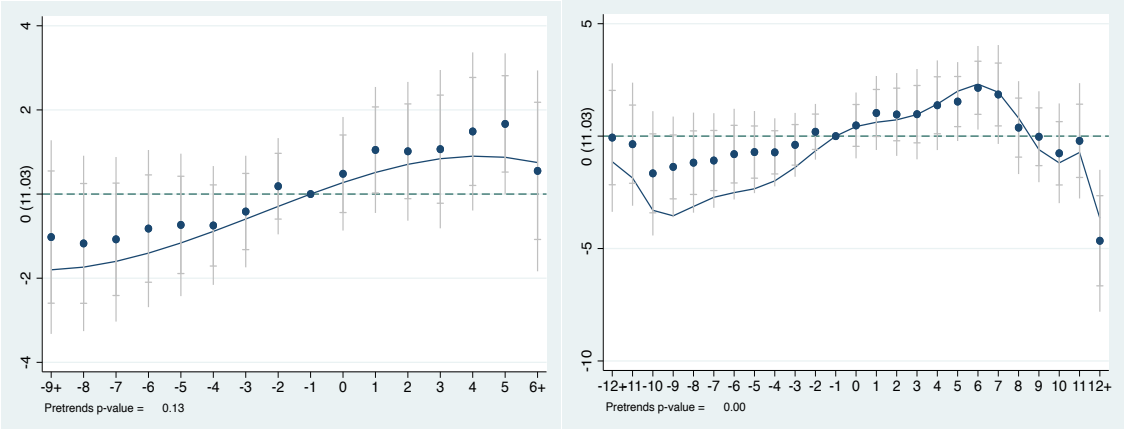


图 G-15 “波浪式”混淆因子检验

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：

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

最简单的时间趋势项就是$f(t) = t$，结果如图 G-16 所示。当然，时间趋势项也可以是时间的多项式。

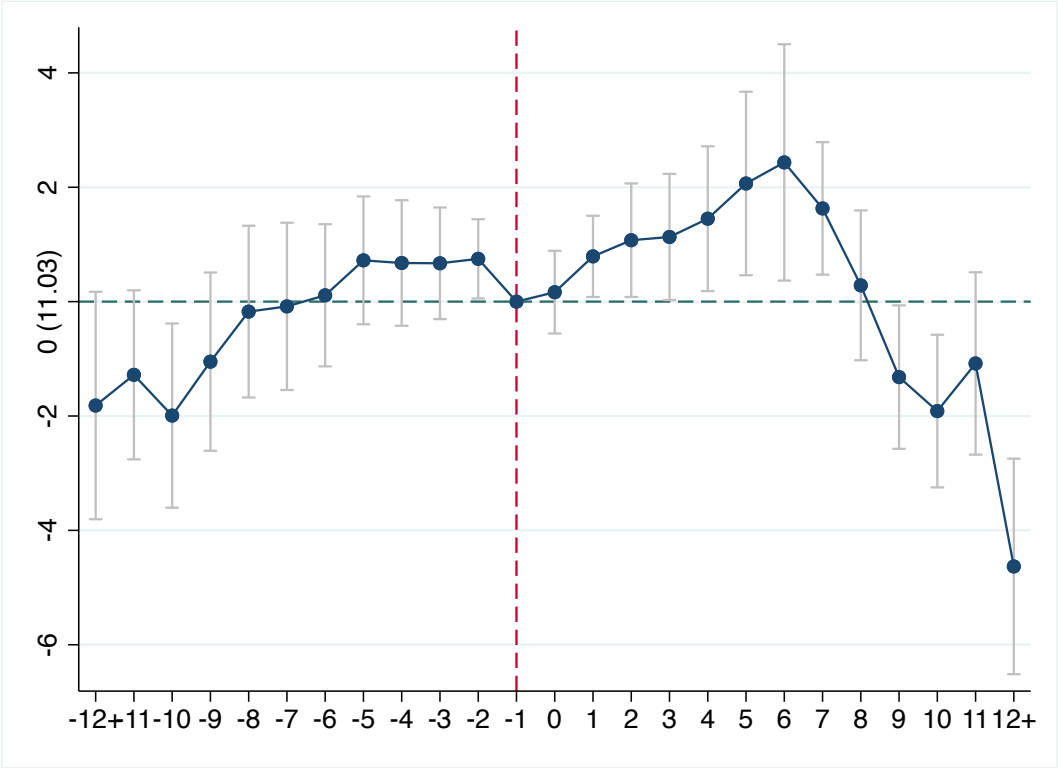


图 G-16 控制线性时间趋势的事件研究图

（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。

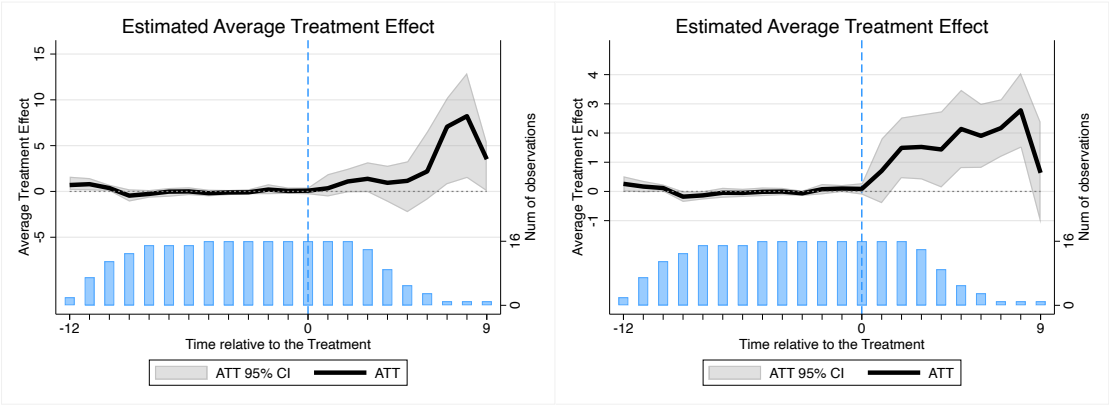


图 G-17 Xu(2017)提出的广义合成控制法和 Athey et al.(2021)提出的矩阵完成法

（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 G-18 所示。

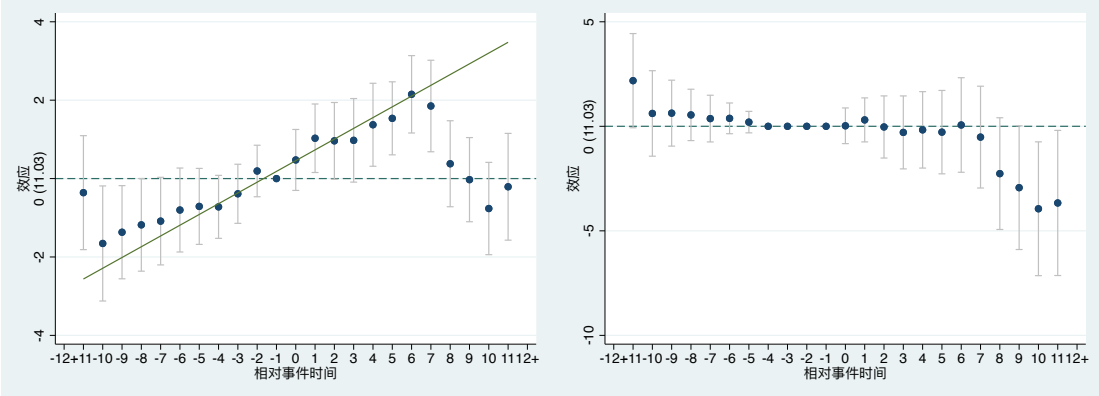


图 G-18 趋势外推后的事件研究图

（4）不可观测混淆因子的代理变量

如图 G-19 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。

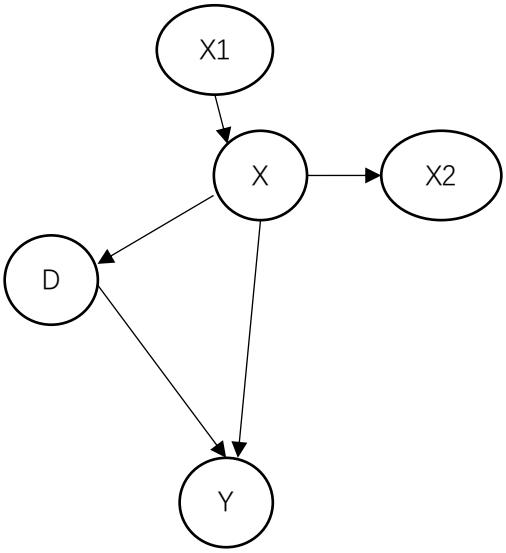


图 G-19 不可观测混淆因子的代理变量

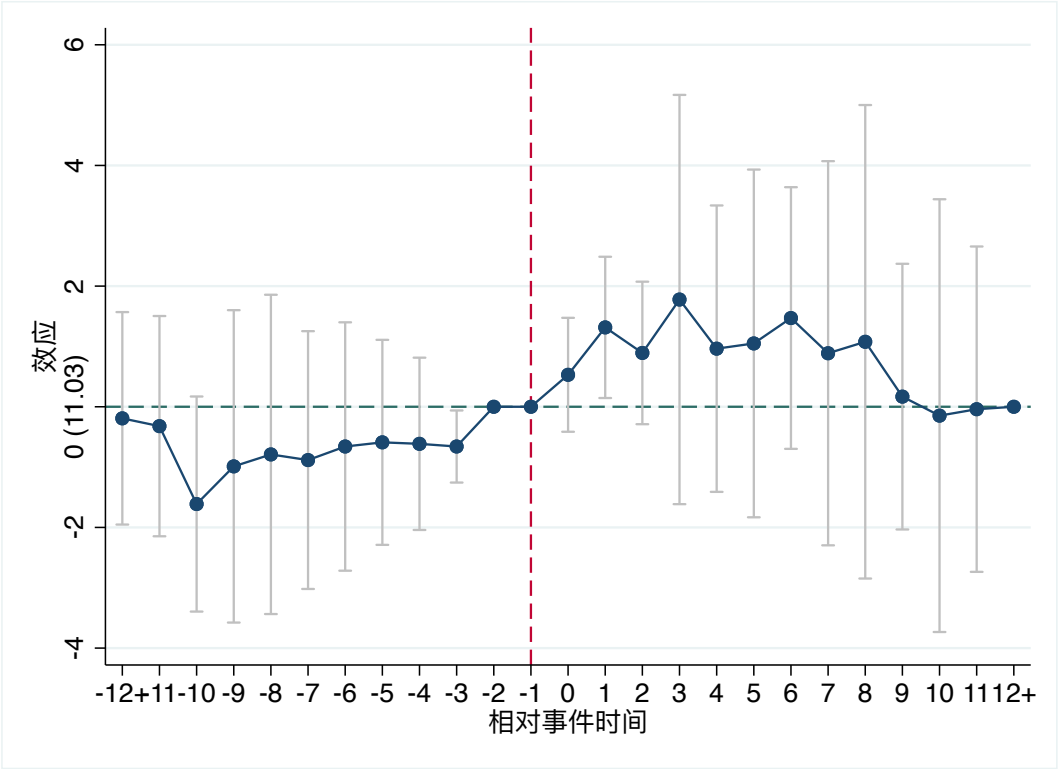


图 G-20 不可观测混淆因子的代理变量

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 G-21。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。”

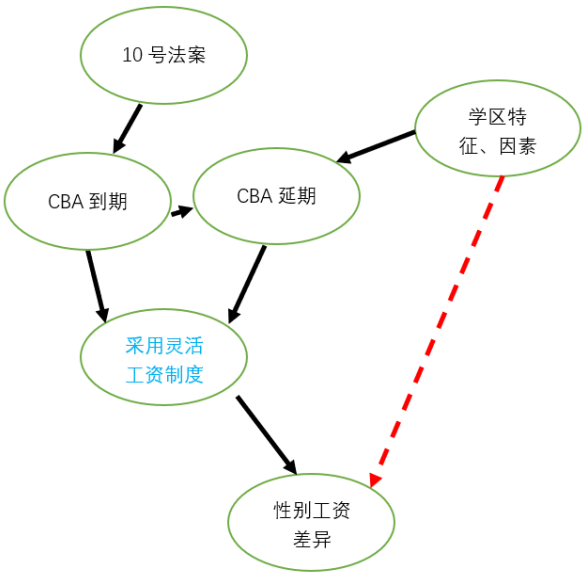


图 G-21

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。