

双重差分Difference in difference

双重差分方法为我们提供了一种寻找反事实结果Potential Outcome的方案，

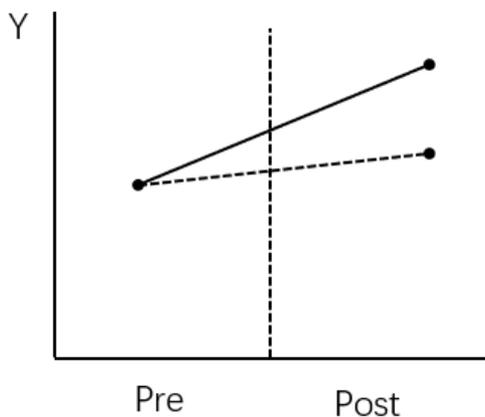
“尽管这个世界上没有两片完全相同的叶子，但是能够找到相似的两片叶子”，

该方法一般被用于评估政策效果、识别处理效应，要求一般有处理组和对照组，及对应组别的处理前后观测值。本专题的主要结构安排分为：

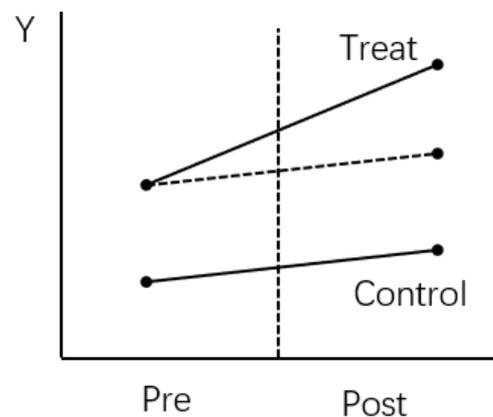
- DID基本逻辑框架
- 经典论文介绍
- 参数估计问题
- 参考文献

DID基本逻辑框架及推演

逻辑简述



a



b

我们仅考虑

两个组别：实验组(Treat)、对照组或称控制组(Control)

两个时期：处理前 (Pre, $t=0$)、处理后 (Post, $t=1$)

为便于描述，我们规定一些标记：

- $Y_{i,t}$ 为观测值 i 在 t 时刻的结果变量 (outcome variable) ；
- $D_{i,t}$ 为 i 在 t 时刻是否受到处理，处理时取值为1，否则为0；
- Y^1 和 Y^0 分别表示该结果变量是否被处理；

- 特别的， $D_{i,1}$ 表示分组变量，因为我们只有在 $t=1$ 的时刻才知道该观测值是否受到了处理，即实验组样本，对应取值为1，否则为控制组样本，记为0。

引入例子

在一个经典的关于“去医院是否会改善病人的身体健康”的估计中，由于我们只能观测到“生病了并且去医院看病的人”，而那些本身就健康的人也不会去医院，那么基于这一样本我们无法识别去医院对身体健康的影响。

要想准确识别我们想要的“平均处理效应(Average Treat Effect of Treatment)”，我们必须找到一个“反事实结果”，即一个病人去医院接受治疗之后的健康状况，和他本应该去医院但是没有去时的健康状况，即图a中Post区域内实线与虚线的差值，以数学表达：

$$ATT = E(Y_{i,1}^1 - Y_{i,1}^0 | D_{i,1} = 1) = E(Y_{i,1}^1 | D_{i,1} = 1) - E(Y_{i,1}^0 | D_{i,1} = 1)$$

由于我们永远无法观测到这一反事实结果，我们就无法真实地识别这一处理效应。

前人想到，或许我们可以找到一个“和这个病人得了相同的病并且身体状况极为相似（一种尽量苛刻的说法），但是确实没有去医院的人”，并以这个人的健康状况变化情况去近似地衡量“反事实结果”，我们的希望是这样的：

$$E(Y_{i,1}^0 | D_{i,1} = 1) - E(Y_{i,0}^0 | D_{i,1} = 1) = E(Y_{i,1}^0 | D_{i,1} = 0) - E(Y_{i,0}^0 | D_{i,1} = 0)$$

该表达式的在DID框架下被叫做“平行趋势假设”。通俗来说，如果我们能找到一个和实验组有着相同变动趋势的控制组观测值，我们可以简单地通过控制组的变动情况求出实验组的“反事实结果Potential Outcome”，进而求得平均处理效应，对应图b。

同时应当注意到，在这一假设中隐含地表达了“控制组的结果变量不会受到处理效应的影响”，我们将会在后续数理推导中对这一问题进行明确表述。

我们将式1.2代入式1.1中，可得：

$$ATT = [E(Y_{i,1}^1 | D_{i,1} = 1) - E(Y_{i,0}^0 | D_{i,1} = 1)] - [E(Y_{i,1}^0 | D_{i,1} = 0) - E(Y_{i,0}^0 | D_{i,1} = 0)]$$

在式1.3中，我们分别在每个组别内在时间维度上进行了“差分”，并进一步地在两个“差分”的基础上进行了一次新的“差分”，此之谓“双重差分”。

基于式1.3，需要注意到DID框架下的求ATT适用的样本观测值：

- 截面重复观测值，即直接套用式1.3进行识别，可以允许实验组、控制组在不同时点下的观测值样本不同，但抽样应当具有代表性（来自于整体分布）；
- 面板数据下，可以观测到同一个个体 i 在不同时间点下的观测值，即可以写作

$$ATT = E(Y_{i,1}^1 - Y_{i,0}^0 | D_{i,1} = 1) - E(Y_{i,1}^0 - Y_{i,0}^0 | D_{i,1} = 0)$$

数理推导

此处以Abadie (2005)中的框架进行数理推导。

重新表述前文中的结果变量：

$$Y_{i,t} = \delta_t + \alpha D_{i,t} + \eta_i + v_{i,t}, \quad t = 0, 1$$

给出识别条件为 $P(D_{i,t} = 1 | v_{i,t}) = P(D_{i,t} = 1)$ ，即要求处理效应是外生的（注意是与 $v_{i,t}$ ）。

引入与个体特征相关的分组标识变量 $E(\eta_i|D_{i,1})$ ，此时有：

$$Y_{i,t} = \delta_t + \alpha D_{i,t} + E(\eta_i|D_{i,1}) + \varepsilon_{i,t}$$

其中 $\varepsilon_{i,t} = v_{i,t} - E(\eta_i|D_{i,1}) + \eta_i$ 。

对式2.2中的部分变量进一步替换，即

$$\delta_t = \delta_0 + (\delta_1 - \delta_0)t$$

$$E(\eta_i|D_{i,1}) = E(\eta_i|D_{i,1} = 0) + [E(\eta_i|D_{i,1} = 1) - E(\eta_i|D_{i,1} = 0)]D_{i,1}$$

因此有：

$$Y_{i,t} = \mu + \sigma t + \tau D_{i,1} + \alpha D_{i,t} + \varepsilon_{i,t}$$

其中 $\mu = E(\eta_i|D_{i,1} = 0) + \delta_0$ 。

基于前面的识别条件，我们有 $E[(1, t, D_{i,1}, D_{i,t}) \cdot \varepsilon_{i,t}] = 0$ ，进而通过OLS估计我们可以得到 $D_{i,t}$ 的系数：

$$\alpha = [E(Y_{i,1}|D_{i,1} = 1) - E(Y_{i,0}|D_{i,1} = 1)] - [E(Y_{i,1}|D_{i,1} = 0) - E(Y_{i,0}|D_{i,1} = 0)]$$

以上推导过程说明，我们可以OLS估计出实验组的平均处理效应。

特别指出的是，识别条件中对于所有的 $t=0, 1$ 而言， $v_{i,1} - v_{i,0}$ 与 $D_{i,1}$ 也是均值独立的，即是说，在没有被处理的情况下，实验组的结果变量（平均）变动会与控制组的结果变量（平均）变动相同，即“平行趋势假定”。

OLS模型表述

为与现有实证研究习惯相一致，以Panel Data为例，我们将式2.3进行一定程度的改写：

$$y_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \beta_3 Post_t \times Treat_i + \gamma' X_{i,t} + \varepsilon_{i,t}$$

在这一模型中：

$Post_t$ 为时间处理变量，当样本处于政策冲击（或处理效应）之后取值为1，否则为0； $Treat_i$ 为分组处理变量，当样本属于实验组时取值为1，否则为0；我们同时引入了控制变量 $X_{i,t}$ ， γ 为对应的系数向量； $\varepsilon_{i,t}$ 为扰动项。

我们以表格形式列示估计结果：

	控制组	实验组	差分
处理前	β_0	$\beta_0 + \beta_2$	β_2
处理后	$\beta_0 + \beta_1$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$	$\beta_2 + \beta_3$
差分	β_1	$\beta_1 + \beta_3$	β_3

如果我们基于面板数据构建了模型3.1，直接通过OLS估计交互项 $Post_t \times Treat_i$ 的系数，即得到实验组的平均处理效应。

DID与双维固定效应（TWFE）

在实证中往往通过双维固定效应模型 (TWFE)识别DID中的系数，控制时间、个体维度的固定效应，即：

$$y_{i,t} = \alpha + \beta Post_t \times Treat_i + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t}$$

此前的推导均假定处理效应是同时发生的，但是在现实观测样本中同一个处理效应并不在同一个时间点发生，例如美国放松银行管制在不同州的时间点并不一致 (Beck et al., 2010)。与之相区别，实证中更多地应用了多时点DID模型 (Multiple Time Periods DID) 用于区别于此前的单期DID模型 (Traditional DID)，该模型仍然基于TWFE形式，即：

$$y_{i,t} = \alpha + \beta D_{i,t} + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t}$$

我们将在参数估计的子节中详细讨论TWFE对DID估计量的影响，以及多时点DID参数估计中可能存在的问题。

其他对于双重差分模型的变形：

- 连续DID，习惯做法是更改 $Treat_i$ 为连续变量；
- Flexible DID，考虑处理的批次不同，处理效应的持续时间不同；
- 队列DID，如上山下乡Chen et al. (2020)；
-

平行趋势检验问题

当观测样本是面板数据的时候，我们可以获取样本的前后观测值。尽管我们无法检验处理效应发生之后实验组样本的反事实结果，但是我们可以检验处理前实验组与控制组样本是否具有相同的变动趋势。我们做以下设定：

- 定义处理效应发生的时间，为简化表述我们只考虑三个时间点， S_{-1}, S_0, S_1 分别代表处理前、处理开始时、处理后，并分别定义虚拟变量，以 S_{-1} 为例：

$$T_s = \begin{cases} 1 & \text{if } s \in S_{-1} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

- 考虑分组处理变量， $Treat_i \in \{0, 1\}$
- 暂不考虑引入虚拟变量与常数项带来的完全共线性问题，实际上我们只能对 T_s 最多引入两个虚拟变量

模型设定如下：

$$y_{i,t} = \alpha + \sum \sigma_s T_s + \beta Treat_i + \sum \delta_s T_s \times Treat_i + \gamma' X_{i,t} + \varepsilon_{i,t} \quad (4.1)$$

或设定为TWFE形式：

$$y_{i,t} = \alpha + \sum \delta_s T_s \times Treat_i + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t} \quad (4.2)$$

对于多时点DID模型，对应的TWFE形式为：

$$y_{i,t} = \alpha + \sum \delta_s D_{i,t}^s + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t} \quad (4.3)$$

其中， $D_{i,t}^s$ 表示观测值 i 在 t 时刻处于 S 状态 ($s \in \{S_{-1}, S_0, S_1\}$) 时取1，否则为0。

类似地，我们以式4.2为例可以得到回归系数表格如下：

	控制组	实验组	差分
S_{-1}	α	$\alpha + \delta_{S_{-1}}$	$\delta_{S_{-1}}$
S_0	α	$\alpha + \delta_{S_0}$	δ_{S_0}
S_1	α	$\alpha + \delta_{S_1}$	δ_{S_1}

对此进行参数检验，由于我们需要满足“平行趋势假定”，即在处理效应发生前实验组与控制组之间没有显著差异（在控制其他条件不变时）。

原假设： $H_0 : \delta_{S_{-1}} = 0$ ，备择假设 $H_1 : \delta_{S_{-1}} \neq 0$

因此，要求对应的回归系数 $\delta_{S_{-1}}$ 不显著为0。

此外，实证中还习惯考虑使用 δ_{S_0} 、 δ_{S_1} 的系数来描述处理效应的后续影响，习惯于称之为“动态DID”。

经典论文介绍

研究动机

使用双重差分最经典的论文之一是Card和Krueger（1994）研究最低工资政策变化对就业率的影响，尤其是对商店企业层面的雇用率影响。从政府角度，本论文的研究可以为以后采取类似措施提供参考，是否能达到政策预期的目的；从企业角度，可以分析该项政策对自身企业是否有影响，以及行业内部是如何针对措施做出反映，利于企业改变经营策略。

传统经济理论预测：最低工资的上升会使处于完全竞争市场的雇主削减雇佣量。有许多实证研究也得到了相似的结论，但另一部分研究则表示并不能找到显著关系。因此Card和Krueger的研究为此提供了另一有力证据。

为了研究这个问题，论文主要采用的是双重差分的方法，以实验组和对照组之差来消除经济周期（平行趋势）所带来的影响，以提升最小工资法案生效前后各项指标之差来消除原有的组别之间的差距，在克服了系统性差异和内生性后，进行回归分析。

实验设计与样本选择

1989年联邦政府颁布法案：

- 1990.4.1从\$3.35提升至\$3.80
- 1991.4.1进一步提升至\$4.25。

NJ州1990颁布法案：

- 1992.4.1提升至\$5.05

1992.3月州立法机构投票，希望将后续工资提高分为2年进行。该次投票险些能越过州长权力直接更改立法，而后州长要求原法案按期执行。在进一步进行两步立法投票之前，考虑到未来最低工资提高后将面临重新下降的问题，州立法机构放弃了。

以上是最低工资政策的背景。新泽西州提高最低工资的法案是在两年前经济健康时制定的，且1992.3州立法机构险些否决了该法案，充分体现了该政策的外生性。

本文的样本选择非常严谨，选择新泽西作为实验组不仅符合实验要求，而且可以剔除很多的外生因素。原因有三个：

- 首先，新泽西当时的经济和政治环境，失业率在上升。可以排除经济上行可能掩盖提高最低工资的效果。
- 其次，新泽西州和宾夕法尼亚的地理位置，导致新泽西既是一个独立体又和宾夕法尼亚经济环境、雇佣周期模式类似，可以作为对比，并且剔除周期效应。
- 第三，收集了上涨之前（1992.2—3）和上涨之后的7—8个月（1992.11—12）的数据，拥有企业的完全的信息并且在分析中考虑关门企业的雇佣量的变化。因此在平均雇佣量的基础上衡量最低工资的所有影响，并且不仅仅是它对现存公司的影响，以此避免幸存者偏差。

调查行业选择快餐业主要因为之所以研究快餐行业，主要有一下五点原因：

- 快餐业是主要的低工资者的雇佣行业
- 行业遵守最低工资标准
- 工资技能要求同质化，且不需要小费
- 组织形式多为连锁店，信息框架统一
- 已有研究显示快餐店对电话调研回复率较高

除了文章列举出的原因，还可能有其他原因：之前学者有针对快餐业的调查，这是一个参考，而且利于本次调查结果和之前结果的对比，得出相反结论时排除行业不同造成的影响。

基于这些条件，选取了新泽西和宾夕法尼亚的汉堡王、KFC、Wendy's和RoyRogers的快餐店作为分析对象。分别进行两次调查，第一次调查是在1992年2月到3月电话调查，包括雇佣量、起薪和商店的其他特征，询问410家餐厅，回应率为87%。第二次调查在政策执行八个月后进行，即1992年11月到12月，电话采访之前410家餐厅，成功调查了371（90%）的店铺。

TABLE 1—SAMPLE DESIGN AND RESPONSE RATES

	All	Stores in:	
		NJ	PA
<i>Wave 1, February 15–March 4, 1992:</i>			
Number of stores in sample frame: ^a	473	364	109
Number of refusals:	63	33	30
Number interviewed:	410	331	79
Response rate (percentage):	86.7	90.9	72.5
<i>Wave 2, November 5–December 31, 1992:</i>			
Number of stores in sample frame:	410	331	79
Number closed:	6	5	1
Number under renovation:	2	2	0
Number temporarily closed: ^b	2	2	0
Number of refusals:	1	1	0
Number interviewed: ^c	399	321	78

^aStores with working phone numbers only; 29 stores in original sample frame had disconnected phone numbers.

^bIncludes one store closed because of highway construction and one store closed because of a fire.

^cIncludes 371 phone interviews and 28 personal interviews of stores that refused an initial request for a phone interview.

样本数据分析

在进行双重差分前，进行了均值检验，鉴别出了均值近似相同和不同的变量。在第一次调查中，两地在平均的营业时间，全职员工的结构和对于雇佣新员工的奖励机制并没有显著的区别全职员工平均雇佣人数稍有不同，员工的起薪相似，显著不同的是全餐的价格。

TABLE 2—MEANS OF KEY VARIABLES

Variable	Stores in:		<i>t</i> ^a
	NJ	PA	
<i>1. Distribution of Store Types (percentages):</i>			
a. Burger King	41.1	44.3	-0.5
b. KFC	20.5	15.2	1.2
c. Roy Rogers	24.8	21.5	0.6
d. Wendy's	13.6	19.0	-1.1
e. Company-owned	34.1	35.4	-0.2
<i>2. Means in Wave 1:</i>			
a. FTE employment	20.4 (0.51)	23.3 (1.35)	-2.0
b. Percentage full-time employees	32.8 (1.3)	35.0 (2.7)	-0.7
c. Starting wage	4.61 (0.02)	4.63 (0.04)	-0.4
d. Wage = \$4.25 (percentage)	30.5 (2.5)	32.9 (5.3)	-0.4
e. Price of full meal	3.35 (0.04)	3.04 (0.07)	4.0
f. Hours open (weekday)	14.4 (0.2)	14.5 (0.3)	-0.3
g. Recruiting bonus	23.6 (2.3)	29.1 (5.1)	-1.0
<i>3. Means in Wave 2:</i>			
a. FTE employment	21.0 (0.52)	21.2 (0.94)	-0.2
b. Percentage full-time employees	35.9 (1.4)	30.4 (2.8)	1.8
c. Starting wage	5.08 (0.01)	4.62 (0.04)	10.8
d. Wage = \$4.25 (percentage)	0.0	25.3 (4.9)	—
e. Wage = \$5.05 (percentage)	85.2 (2.0)	1.3 (1.3)	36.1
f. Price of full meal	3.41 (0.04)	3.03 (0.07)	5.0
g. Hours open (weekday)	14.4 (0.2)	14.7 (0.3)	-0.8
h. Recruiting bonus	20.3 (2.3)	23.4 (4.9)	-0.6

Notes: See text for definitions. Standard errors are given in parentheses.

^aTest of equality of means in New Jersey and Pennsylvania.

在最低工资上涨后，新泽西快餐店行业的平均起薪上升了10%。最低工资的上涨，对于那些最开始处于高工资的饭店并没有产生明显的效果：这些饭店平均工资变动百分比为-3.1%。在两次调查中，只有两个变量相对变动：全职员工的比例和食品价格，新泽西这两个指标相对于宾夕法尼亚都在上涨。

双重差分

论文首先讨论了新泽西和宾夕法尼亚每家店铺的全职工平均就业量的水平和平均就业量的差分，以及这两个州全职工平均就业量的水平变化再差分，（表3的3.4.5行；第4行是平衡样本的变化；将四家暂时关闭的商家的就业量被视为0，而不是作为丢失数据在第5行展示了平衡样本的就业的平均变化）；然后将新泽西商店分成了三组，这三组的最低工资分别为4.25，4.26—4.99，和5.00以上，这三组的平均就业量中，高工资和低工资做差分，中间值和高工资做差分。

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

Variable	Stores by state			Stores in New Jersey ^a			Differences within NJ ^b	
	PA (i)	NJ (ii)	Difference, NJ—PA (iii)	Wage = \$4.25 (iv)	Wage = \$4.26—\$4.99 (v)	Wage ≥ \$5.00 (vi)	Low- high (vii)	Midrange- high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	-2.69 (1.37)	-2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	-2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores ^c	-2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	-2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0 ^d	-2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	-2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

^aStores in New Jersey were classified by whether starting wage in wave 1 equals \$4.25 per hour ($N = 101$), is between \$4.26 and \$4.99 per hour ($N = 140$), or is \$5.00 per hour or higher ($N = 73$).

^bDifference in employment between low-wage (\$4.25 per hour) and high-wage (\geq \$5.00 per hour) stores; and difference in employment between midrange (\$4.26—\$4.99 per hour) and high-wage stores.

^cSubset of stores with available employment data in wave 1 and wave 2.

^dIn this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

双重差分的分析表明，未受到最低工资上涨的影响快餐店的就业量在1992年2月至11月是下降的。出现这种下降可能的原因是1992年经济衰退的原因。加盟快餐点的销售是顺经济周期的，在不受其他因素影响的情况下失业率的上涨也会导致就业率的下降。

调整的回归模型

双重差分比较并没有考虑就业增长的其他变量，这些变量被考虑在回归模型中，模型形式如下：

$$\Delta E_i = a + bX_i + cNJ_i + \epsilon_i \quad \text{或者} \quad (5.1a)$$

$$\Delta E_i = a + bX_i + cGAP_i + \epsilon_i \quad (5.1b)$$

其中, ΔE_i 是商店*i*第一次到第二次的就业改变量, X_i 是商店*i*一系列的特征, NJ_i 是一个虚拟变量, 新泽西商店的该值为1。基于商店*i*的初始工资, 最低工资率对该商店的影响用可替代变量 GAP_i (其实就是分段函数) 来衡量:

$$GAP_i = \begin{cases} 0 & \text{对于宾州的商店} \\ 0 & \text{对于新泽西中 } W_{1i} \geq \$5.05 \text{ 的商店} \\ (5.05 - W_{1i})/W_{1i} & \text{新泽西其他商店} \end{cases}$$

GAP_i 的变化, 既反映了新泽西和宾夕法尼亚的对比, 又反映新泽西内部基于第一次调查的初始工资的差异。

TABLE 4—REDUCED-FORM MODELS FOR CHANGE IN EMPLOYMENT

Independent variable	Model				
	(i)	(ii)	(iii)	(iv)	(v)
1. New Jersey dummy	2.33 (1.19)	2.30 (1.20)	—	—	—
2. Initial wage gap ^a	—	—	15.65 (6.08)	14.92 (6.21)	11.91 (7.39)
3. Controls for chain and ownership ^b	no	yes	no	yes	yes
4. Controls for region ^c	no	no	no	no	yes
5. Standard error of regression	8.79	8.78	8.76	8.76	8.75
6. Probability value for controls ^d	—	0.34	—	0.44	0.40

Notes: Standard errors are given in parentheses. The sample consists of 357 stores with available data on employment and starting wages in waves 1 and 2. The dependent variable in all models is change in FTE employment. The mean and standard deviation of the dependent variable are -0.237 and 8.825 , respectively. All models include an unrestricted constant (not reported).

^aProportional increase in starting wage necessary to raise starting wage to new minimum rate. For stores in Pennsylvania the wage gap is 0.

^bThree dummy variables for chain type and whether or not the store is company-owned are included.

^cDummy variables for two regions of New Jersey and two regions of eastern Pennsylvania are included.

^dProbability value of joint F test for exclusion of all control variables.

表4反映的是模型中变量的参数估计, 括号内是标准误差。比较表中参数值得变化, 可以发现: 由于数据获得受限, 导致对新泽西就业相对增长的估计值略小; 针对连锁和公司直营的控制变量对估计新泽西的虚拟变量的大小没有影响; 使用 GAP_i 变量来衡量最低工资率的影响, 比简单的新泽西虚拟变量稍好; 一个更加强的测试, 增加了虚拟变量来表示新泽西的(北, 中, 南)三个地区和东部宾夕法尼亚的两个地区(阿伦敦—伊斯顿和费城北部郊区), 概率值显示没有证据表明就业增长与区域有关。

表4中还用每个商店的就业比例变化作为因变量进行了模型估计, 但是在约定水平下, 结果不显著。当因变量用变化比例来衡量时, 显示的最低工资的就业影响也更小了。影响大小归因于商店的最低工资影响的异质性。

其他参数检验

表3和表4的结果似乎与传统（最低工资的上升会降低就业）相矛盾。论文中还给出了其他的参数来证明该结论的稳健性，此处不赘述，详细见原文。除了对回归模型中的基本参数检验，还对一些变量进行调整，比如就业被重新定义为不包括管理人员；剔除了在沿着新泽西海岸线的城镇中35个商店，否则季节性效应会导致最低工资影响轻微偏大；第一次差分的合理性等等。

在肯定最低工资提高不会带来就业率的下降之后，作者依次考虑了就业率改变的替代结果：全职工人的聘用率上升、员工福利的减少、在职培训减少以及首次提薪的延迟、价格效应、开店数量的减小。但是上述推断都因为统计不显著、估计系数较小或估计系数正负与实际不相符而被拒绝。

主要发现

- 本文发现最低工资的上涨没有减少快餐店的就业率，这和普遍传统的认识不同。在面对最低工资上涨时，餐厅没有通过减少非工资补偿和工资曲线来抵消最低工资上涨带来的成本增加。
- 最低工资越高，快餐价格越高，价格增长的等级与传统模型对竞争产业的预测一致。
- 高的最低工资率不会影响餐厅的净数量和开店率。相反，所有的估计都体现出有促进的正面影响。

结论

与传统的模型估计不同，本文研究发现新泽西最低工资提高对州内快餐店就业减少没有影响，并且证明结论的稳定性。还比较了新泽西、宾夕法尼亚和纽约在最低工资提高之后的那段时间里快餐行业的青年就业率的改变，进一步说明新泽西低工资的工人就业增加。最后，新泽西快餐价格相对宾夕法尼亚有提高，说明最低工资增长最后的压力转嫁给消费者。但是，价格增长最大的商店是不完全是由于最低工资提高导致的、总之，这些发现很难用标准竞争模型或是其他面对供给限制的模型来解释。

模型识别问题

前文部分中我们给出了DID分析的基础框架，同时展示了最为经典的一篇双重差分论文。本部分我们将主要讨论在模型参数估计过程中可能存在的些许潜在问题，及相应的解决思路。

CIC模型

CIC简述

标准的 DID 模型使用条件较为严苛：

- 需要满足平行趋势假设，每个样本是进入实验组还是控制组是完全随机的，并且实验组和控制组个体的某些特征不会随着时间变化而呈现出不同的变化；
- 需要这个政策干预（或处理效应）只影响实验组，不会对控制组产生交互影响（或者政策干预不会产生外溢效应）；最好，还
- 要求潜在结果变量（反事实结果）与处理变量、时间变量满足线性条件。

在使用 DID 模型时，若没有考虑以上条件，忽略了实验组和控制组之间随时间而变化但却无法观测到的异质性因素，估计出来的政策效果就是有偏误的。

在一个只有两期两组的极端数据场景下，我们很难知道其是否符合平行趋势假设。因为政策（处理效应）在不同阶段下不仅可能改变估计结果的均值和方差，而且在个体之间的影响也会不同。

为了克服经典的 DID 无法解决的异质性处理效应问题，Athey and Imbens (2006) 提出了可以适用于连续型解释变量的非线性双重差分方法(Nonlinear Difference-in-Difference, NL-DID)，也称为双重变换模型 (Changes in Changes, CIC)，它不依赖于函数形式，也允许时间变化与政策干预下对不同个体影响有所不同。

CIC 模型假设实验组和控制组的潜在结果具有异质性分布，群体、时间周期在处理效应前可以是非对称性。通过该假设可以构造处理组的“反事实”分布，识别更加有效的处理效应分布，因此可以用来评估非随机实验的政策效果问题。另外，CIC模型与经典 DID方法不同的是其假设控制组个体的结果变量满足非线性形式而不是简单的线性形式。

识别与推断

CIC模型的具体原理是将控制组经验分布的反函数作为处理组“反事实”的分布函数，进而求出处理组的“反事实”的潜在结果，最后将处理组的可观测结果均值和“反事实”的潜在结果均值的差作为平均处理效应。

具体而言，对于独立同分布的观测值而言，做如下变量设定：

- D_i 为分组变量（实验组为1，控制组为0）， T_t 为时间处理变量；
- $I_i = D_i \times T_t$ ，是个体*i*是否受到处理的标识变量；
- $Y_{i,t}$ 为对应的结果变量，我们使用上标N表示没有受到处理效应而上标I表示接受处理效应；
- $F_{Y_{1,0}}$ 、 $F_{Y_{0,0}}^{-1}$ 、 $F_{Y_{0,1}}$ 分别表示在给定组别和时间下的条件分布函数。

我们观察到的结果变量（在 $t = 1$ 时）为：

$$Y_i = Y_i^N \cdot (1 - I_i) + I_i \cdot Y_i^I$$

当第二个时期的结果变量在个体中是无法观测的部分 u 时，且有 $h(u, 0) = y$ ， $F_{Y_{1,1}}^N$ 分布函数等同于 $k^{CIC}(Y_{1,0})$ 的分布。

因此，具有异质性的平均处理效应：

$$\tau^{CIC} \equiv E[Y_{1,1}^I - Y_{1,1}^N] = E[Y_{1,1}^I] - E[k^{CIC}(Y_{1,0})] = E[Y_{1,1}^I] - E[F_{Y_{0,1}}^{-1}(F_{Y_{0,0}}(Y_{1,0}))] \quad (2)$$

$$\Delta^{CIC} = F_{Y_{0,1}}^{-1}(q') - F_{Y_{0,0}}^{-1}(q') = F_{Y_{0,1}}^{-1}(F_{Y_{0,0}}(y)) - y \quad (3)$$

这一过程的推导详见Athey and Imbens (2006)。

非平行趋势问题

在传统DID的推导中，“平行趋势假定”要求：在没有被处理的情况下，实验组的结果变量（平均）变动会与控制组的结果变量（平均）变动相同。

这一条件过于苛刻，事实上由于协变量的存在（如前文式3.1中引入的控制变量），这些协变量可能是影响实验组、控制组是否受到处理的重要因素。

一个易于理解的例子是，区别于Card and Krueger (2000)中的故事，如果我们突然发布政策改变了某个地区的最低工资水平，可能是考虑了这个地区的特殊性的，例如这个地区的居民普遍生产力水平较低、劳动力人口偏低，而作为控制组的地区在这些特征上与实验组地区存在差异，即协变量会影响处理前实验组与控制组的平衡性。

我们在这里使用来自Abadie (2005)中的数理表达，说明实验组、控制组中可能存在着非平衡趋势问题：

$$D_{i,1} = \begin{cases} 1 & \text{if } Y_{i,1-\kappa} + u_i < \bar{Y} \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

其中， κ 是一个， \bar{Y} 是一个常数， u_i 是一个独立于其他协变量的随机变量。在这种选择过程中，处理前 κ 期表现较低的个体更有可能被处理，我们可以记为 $Y_{i,1-\kappa} = X_i$ 。在更一般的情形中，我们可以将 X_i 视为一系列可观测到的协变量。

单纯地通过引入控制变量进行控制可能是不合适的，例如处理效应会在实验组内的不同的组别中产生差异（基于可观测协变量区分得到的组别），特别是如果我们集中关注某种条件平均处理效应时，这个问题必须引起重视。

可以通过一个两步法策略解决这一问题：

- 基于可观测协变量进行倾向得分估计，并计算拟合值；
- 基于倾向得分进行加权的DID模型估计。

基于以下两个假设：

$$\text{Assumption 1: } E[Y_{i,1}^0 - Y_{i,0}^0 | X_i, D_{i,1} = 1] = E[Y_{i,1}^0 - Y_{i,0}^0 | X_i, D_{i,1} = 0] \quad (5)$$

$$\text{Assumption 2: } P(D_{i,1} = 1) > 0 \text{ and with probability one } P(D_{i,1} = 1 | X_i) < 1 \quad (6)$$

当我们应用了两步法之后，平均处理效应为（不再引入 i ）：

$$E[Y_1^1 - Y_1^0 | D = 1] = E\left[\frac{Y_1 - Y_0}{P(D = 1)} \cdot \frac{D - P(D = 1 | X)}{1 - P(D = 1 | X)}\right] \quad (7)$$

这一做法的本质是对于不同的样本进行了一个权重的调整，该权重为对应进入实验组的倾向得分。

在这一识别过程中，我们可以采用半参数的方法来缓解模型设定可能带来的问题，以求得近似条件平均处理效应。Abadie (2005)进一步讨论了三种情形下的非参数逼近方法，截面重复观测值、观测样本选择问题、多层次的实验组，并给出了详细的估计方法与渐进分布。

TWFE、多时点DID问题

面板数据与DID模型相性十分良好。特别是面板数据中同时包含了个体、时间两个维度的信息，在大多数实证分析中往往采用带有个体、时间双维度固定效应的DID模型进行估计，用于评估政策处理效果（平均处理效应）。

对于双维固定效应模型下的DID模型：

$$y_{i,t} = \alpha + \beta Post_t \times Treat_i + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t}$$

在面板数据背景下，我们可以使用一个变量来描述交互项部分，定义 $D_{i,t} = Post_t \times Treat_i$ ，因此统一与多时点DID模型写作以下形式：

$$y_{i,t} = \alpha + \beta D_{i,t} + \gamma' X_{i,t} + \mu_i + \lambda_t + \varepsilon_{i,t}$$

需要特别指出的是，前文对DID模型进行数理推导时的给出识别条件为

$$P(D_{i,t} = 1 | v_{i,t}) = P(D_{i,t} = 1)$$

即要求处理变量与结果变量 $Y_{i,t}$ 的扰动项 $v_{i,t}$ 独立，但由于此时采用了TWFE模型设定，我们的外生性假定：

$$P(D_{i,t} = 1 | \varepsilon_{i,t}) = P(D_{i,t} = 1)$$

由此，我们的基于双维固定效应对DID模型进行识别依赖于：a. 模型形式正确；b. 严格外生性。

但是这也给处理效应的识别带来了挑战：

- 个体固定效应假定了个体不可观测特征不随时间变化（这是所有个体固定效应模型都存在的问题），但现实中更可能存在的是时变的混合因子，我们无可奈何地忽略了这个问题；
- 对于不同时间点（不同批次）的处理效应而言，已经发生的处理效应可能会对后续结果变量产生影响（feedback from past outcome），而一般这个问题也会被实验者有意无视；
- 处理效应对于不同个体的效果可能是异质性的，这会带来偏差，往往被直接无视；
- 平均处理效应可能是一种加权后的结果，甚至可能出现某些个体的权重为负值(Callaway and Sant'Anna, 2021)，例如：真实处理效应为正，但由于某些重要样本的权重系数为负，最终加权效应为负；
- 关于内生性永远无法绕开的问题，过于严苛的外生性假定；
- 更深层次的问题在于，从实验设计的角度看，固定效应模型意味着什么？详见Athey and Imbens (2022)；
- TWFE下的多时点DID模型，详见Callaway and Sant'Anna (2021)。

标准误问题

当DID的方法被用于多年数据时，尽管许多研究考虑了序列相关问题，但估计结果的标准误会存在不一致的问题，并且这个问题往往被忽视了（Bertrand et al., 2003）。

该文发现，在给定实验组、控制组的情况下，采用安慰剂方法随机生成处理效应，基于OLS估计得到的平均处理效应（即处理变量系数）的标准误是被严重低估了的，并且高达45%的安慰剂结果是显著的（在5%的显著性水平下）。

在运用OLS对DID模型进行估计时，模型设定如下：

$$Y_{ist} = \beta I_{st} + A_s + B_t + cX_{ist} + \varepsilon_{ist}$$

一般地， Y_{ist} 表示个体*i*在第*s*组（如某个美国的州）*t*时刻（如某年）的结果变量， I_{st} 表示是否受到干预的虚拟变量， A_s 和 B_t 分别表示州和年份的固定效应， X_{ist} 表示个体层面的控制变量， ϵ_{ist} 表示误差项。OLS估计的处理效应用 $\hat{\beta}$ 表示。

基于这一模型设定，通过Monte Carlo模拟实验的方式，Bertrand et al.(2003)运用妇女工资的数据（source: the Current Population Survey）比较了不用设定形式下标准误的差异，包括Econometric corrections、Bootstrap、two corrections和one correction，并总结为以下结果：

- 对时间序列过程制定一个特定的参数形式表现不好；
- Bootstrap方法在实验组数量足够大时表现好；
- 基于方差-协方差矩阵的渐进近似估计的相关方法在数量中等的情况下表现好；
- 聚合时间的方法在状态数量很小时也表现好。

解决办法

Parametric Methods

研究采用的是给误差项指定一个自回归相关结构，用估计到的参数计算标准误。对误差项制定的AR(1)过程表述如下：

$$\epsilon_{ist} = \rho\epsilon_{ist-1} + \xi_{ist}$$

当指定自回归系数时，效果要比不指定的情况要好。同样地，在模拟生成回归系数为0.8的AR(1)过程中，其估计的回归系数才不过0.62，这是由于短时间序列造成的OLS估计偏误。

还有一种可能是错误的指定了自回归的形式。当我们分别制定自回归系数为0.55和0.35的AR(2)过程与自回归系数为0.95的AR(1)和noise/signal = 0.13过程时，估计的自回归系数分别为0.44和0.30，拒绝率还高达0.31、0.39。

Block Bootstrap

这种方法的优点在于对误差项几乎没有特别地假定。在实际运用中，t统计量由以下流程计算得出：

- 对于每个安慰剂干预，计算绝对t统计量， $t = \text{abs}(\hat{\beta}/SE(\hat{\beta}))$ ， β 及其标准误由OLS估计得到；
- 通过替代50个 (\bar{Y}_s, V_s) 矩阵来构造一个bootstrap样本， \bar{Y}_s 表示s州的整个时间序列观测值， V_s 表示s州的州、时间和处理虚拟变量矩阵；
- 对样本进行OLS估计。

由此得出 $\hat{\beta}_r$ 和 $t_r = \frac{\text{abs}(\hat{\beta}_r - \hat{\beta})}{SE(\hat{\beta}_r)}$ 。其中， t_r 的样本分布是随机的，并且取决于州的数目。当N趋于无穷时，即使存在州内自相关和异方差，t的样本分布和构造的分布会变得很微小。

但是，由于这种方法的可操作性不强，鲜少在研究中得到应用。

Ignoring Time-Series Information

当计算标准误的时候，我们考虑的一种简单方法是忽视时间序列信息。也就是说，通过简单的平均干预前后的数据进行OLS回归。

当干预发生在不同时期的情形下，我们有一种改进技术：

- 将 Y_{st} 对于州固定效应、年虚拟变量和其他任何相关变量进行回归；
- 将得到控制的州的残差分成两组，即干预以前时间的残差和干预以后时间的残差；
- 由以上两时期面板的OLS得到处理效应的估计值和标准误。

这种方法和简单聚合的方法具有相当的功效。并且，当干预是错开的情形时也表现良好。

The downside of these procedures (both raw and residual aggregation) is that their power is quite low and diminishes fast with sample size.

Empirical Variance-Covariance Matrix & Arbitrary Variance-Covariance Matrix

值得一提的是实证方差-协方差矩阵这个方法的假设：*suppose that the auto-correlation process is the same across all states and that there is no cross-sectional heteroskedasticity*。比如，在一个有50个组的数据中，误差项的方差-协方差矩阵是块对角的，并且每个块由 $T \times T$ 个元素组成，一共有50个这样的块。每个块是对称的， $(i, i + j)$ 元素表示 ϵ_i 和 ϵ_{i-j} 之间的相关关系。因此，可以通过50个组的方差估计矩阵进而计算标准误。

这种方法在无限样本量的情形下良好，但在小样本的情形下功效急剧下降。

当二个假设得不到满足时，我们还是可以通过以上方法计算得到方差-协方差矩阵的估计值，却不能得到方差-协方差矩阵中元素的一致估计量，但是可以运用一般地**White-like formula**计算标准误。这可以由下式得到：

$$W = (V'V)^{-1}(\sum_{j=1}^N u_j' u_j)(V'V)^{-1}$$

其中， N 表示组的总数量， V 表示自变量矩阵（年份虚拟变量，组虚拟变量和处理虚拟变量），每个组的 $u_j = \sum_{t=1}^T e_{jt} v_{jt}$ ， e_{jt} 表示组 i 在 t 时刻估计的残差， v_{jt} 表示因变量的行向量（包含常数）。在组的数目趋于无限时，方差-协方差矩阵的估计值是一致的。

关于标准误的一个小结

- 在小样本的情形下，我们似乎可以优先考虑**Ignoring Time-Series Information**，*Empirical Variance-Covariance Matrix*和*Arbitrary Variance-Covariance Matrix*具有差不多的功效，但差于忽视时间序列信息的方法，*Block Bootstrap*可以不用考虑。
- 在有限样本的情形下，**Arbitrary Variance-Covariance Matrix**和**Empirical Variance-Covariance Matrix**能够胜任。
- 在大样本的情形下，**Block Bootstrap**具有不可替代的优势。

参考文献

- Abadie, A., 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies* 72, 1–19. <https://doi.org/10/bfksdh>
- Athey, S., Imbens, G.W., 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74, 431–497. <https://doi.org/10/dr6h3s>
- Athey, S., Imbens, G.W., 2022. Design-based analysis in Difference-In-Differences settings with staggered adoption. *Journal of Econometrics, Annals Issue in Honor of Gary Chamberlain* 226, 62–79. <https://doi.org/10/gjsmfr>
- Beck, T., Levine, R., Levkov, A., 2010. Big Bad Banks? The Winners and Losers from Bank Deregulation in the United States. *The Journal of Finance* 65, 1637–1667. <https://doi.org/10/bpss37>
- Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* 119, 249–275. <https://doi.org/10/ck2cfm>
- Callaway, B., Sant'Anna, P.H.C., 2020. Difference-in-Differences with multiple time periods. *Journal of Econometrics* S0304407620303948. <https://doi.org/10/ghqtjn>
- Card, D., Krueger, A.B., 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in

New Jersey and Pennsylvania. *The American Economic Review* 84, 772–793.

- Chen, Y., Fan, Z., Gu, X., Zhou, L.-A., 2020. Arrival of Young Talent: The Send-Down Movement and Rural Education in China. *American Economic Review* 110, 3393–3430. <https://doi.org/10/ghj53r>
- Donald, S.G., Lang, K., 2007. Inference with difference-in-differences and other panel data. *The review of Economics and Statistics* 89, 221–233. <https://doi.org/10/cfktcj>
- Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. *Journal of economic literature* 47, 5–86. <https://doi.org/10/cqnhvr>