第 17 章　双重差分法

17.1　平行趋势假定

由于政策冲击的效果一般需要一段时间才能显现出来，故常使用面板数据进行因果推断。

样本中的个体可以是个人、企业、地区，甚至国家。

受到政策冲击的个体构成**处理组**(treatment group)

始终未受政策冲击的个体则构成**控制组**(control group)。

记我们关心的**结果变量**(outcome variable)为 y。

暂时假设面板数据只有两期，即 $t=1$ 表示政策实施前(before)，而 $t=2$ 表示政策实施后(after)。

如何评估处理组的政策效应？

一种简单(天真)的做法是比较处理组均值的前后差异，比如

$$\Delta \bar{y}_{\text{treat}} \equiv \bar{y}_{treat,\,2} - \bar{y}_{treat,\,1} \qquad (17.1)$$

其中，$\bar{y}_{treat,\,2}$ 为处理组政策实施后的样本均值，而 $\bar{y}_{treat,\,1}$ 为处理组政策实施前的样本均值。

上式被称为**差分估计量**(difference estimator)。

由于经济环境也随时间而变，故处理组的前后差异未必就是处理效应，因为可能已经混杂了“时间效应”(time effects)或“时间趋势”(time trend)。

常用解决方法是寻找适当的控制组，由始终未受政策冲击的个体所构成，作为处理组的反事实(counterfactual)参照系。

由于控制组始终未受政策冲击，故在一定条件下可将其均值的前后变化视为纯粹的时间效应，即

$$\Delta \bar{y}_{\text{control}} \equiv \bar{y}_{control,\,2} - \bar{y}_{control,\,1} \qquad (17.2)$$

其中，$\bar{y}_{control,2}$为控制组政策实施后的样本均值，而$\bar{y}_{control,1}$为控制组政策实施前的样本均值。

若想剔除(17.1)式中所包含的时间效应，可综合以上两个差分，即将“处理组均值的前后变化”减去“控制组均值的前后变化”，得到对于处理效应更可靠估计：

$$\Delta\bar{y}_{treat}-\Delta\bar{y}_{\text{control}}=(\bar{y}_{treat,2}-\bar{y}_{treat,1})-(\bar{y}_{control,2}-\bar{y}_{control,1}) \quad (17.3)$$

这就是**双重差分估计量**(Difference in Differences，简记 DID)，因为它是处理组差分与控制组差分之差。

DID 的反事实推断之所以成立，其基本前提是，处理组若未受政策干预，其时间趋势应与控制组一样(故可以后者来控制时间效应)。

这就是所谓的**平行趋势假定**(parallel trends assumption)或**共同趋势假定**(common trends assumption)假定。

图 17.1 直观地展示了 DID 的思想与前提。

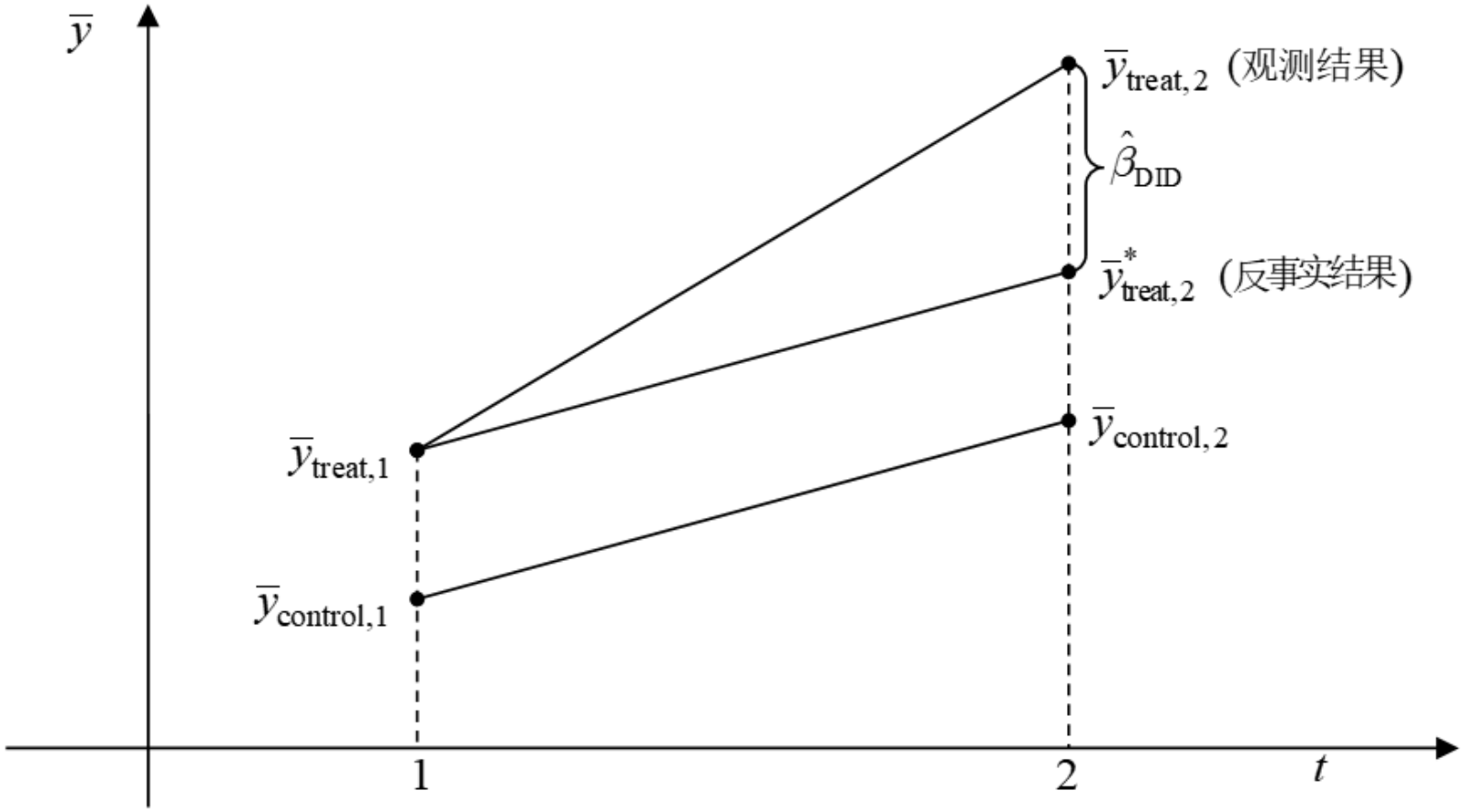

图 17.1 双重差分示意图

$t = 1$ 表示政策实施前，而 $t = 2$ 表示政策实施后。$\bar{y}_{\text{treat},2}$为处理组第 2 期的观测结果(observed outcome)。

根据平行趋势假定，若处理组未受政策冲击，则其第 2 期的反事实结果(counterfactual outcome)为$\bar{y}^{*}_{\text{treat},2}$。

DID 估计量为$\hat{\beta}_{\text{DID}} = \bar{y}_{\text{treat},2} - \bar{y}^{*}_{\text{treat},2}$。

双重差分(17.3)式也可等价地写为

$$(\bar{y}_{treat,2} - \bar{y}_{control,2}) - (\bar{y}_{treat,1} - \bar{y}_{control,1}) \qquad (17.4)$$

DID 估计量已经控制了处理组与控制组在第 1 期初始条件的差别，即剔除了处理组与控制组“处理前差异”(pretreatment differences)的影响。

记 $y_{it}(0)$ 为个体 i 在第 t 期若未受政策冲击的“潜在结果”(potential outcome)，而 $y_{it}(1)$ 为个体 i 在第 t 期若受政策冲击的潜在结果。

记处理组虚拟变量为 $treat_i$：

$$treat_i = \begin{cases} 1 & \text{若个体 } i \text{ 属于处理组} \\ 0 & \text{若个体 } i \text{ 属于控制组} \end{cases} \tag{17.5}$$

$treat_i$ 就是第 15-16 章的 D_i。

假定 17.1（平行趋势假定）如果没有受到政策冲击，则处理组的时间趋势与控制组相同，即

$$\mathrm{E}\left[y_{i2}(0)-y_{i1}(0)\,|\,treat_i=1\right]=\mathrm{E}\left[y_{i2}(0)-y_{i1}(0)\,|\,treat_i=0\right] \quad (17.6)$$

上式左边为处理组($treat_i=1$)若未受政策冲击，其前后结果的期望变化

右边为控制组($treat_i=0$)若未受政策冲击，其前后结果的期望变化。

若均未受政策干预，则处理组与控制组的时间趋势完全相同。

上式仅意味着二者的总体期望相同，而实际的样本均值仍可能存在差异。将上式移项即可得：

$$\mathrm{E}\left[y_{i2}(0)\,|\,treat_i=1\right]=\mathrm{E}\left[y_{i1}(0)\,|\,treat_i=1\right]+\mathrm{E}\left[y_{i2}(0)-y_{i1}(0)\,|\,treat_i=0\right] \quad (17.7)$$

左边 $\mathrm{E}\left[y_{i2}(0)\,|\,treat_i=1\right]$ 为处理组在第 2 期的反事实结果(counterfactual outcome)，因为处理组在第 2 期事实上受到了政策冲击。

处理组在第 2 期的反事实结果等于，“处理组在第 1 期的结果”加上“控制组前后结果的变化”。因此，

$$\tau_{\text{ATT}} \equiv \mathrm{E}\left[y_{i2}(1) \mid treat_i = 1\right] - \mathrm{E}\left[y_{i2}(0) \mid treat_i = 1\right] \qquad \text{(ATT的定义)}$$

$$= \mathrm{E}\left[y_{i2}(1) \mid treat_i = 1\right]$$

$$- \left\{\mathrm{E}\left[y_{i1}(0) \mid treat_i = 1\right] + \mathrm{E}\left[y_{i2}(0) - y_{i1}(0) \mid treat_i = 0\right]\right\} \qquad \text{(17.7式)}$$

$$= \left\{\mathrm{E}\left[y_{i2}(1) \mid treat_i = 1\right] - \mathrm{E}\left[y_{i1}(0) \mid treat_i = 1\right]\right\}$$

$$- \left\{\mathrm{E}\left[y_{i2}(0) \mid treat_i = 0\right] - \mathrm{E}\left[y_{i1}(0) \mid treat_i = 0\right]\right\} \qquad \text{(整理合并)}$$

$$(17.8)$$

上式就是总体期望的双重差分，而与之相应的样本估计值正是 $(\bar{y}_{treat,2} - \bar{y}_{treat,1}) - (\bar{y}_{control,2} - \bar{y}_{control,1})$，即 DID 估计量。

例　最低工资对就业的影响。一般认为提高法定最低工资(minimum wage)将降低企业对低技能工人的需求。

Card and Krueger (1994)通过一个自然实验发现了相反的结果。

在 1992 年，美国新泽西州通过法律将最低工资从每小时$4.25提高到$5.05，但在相邻的宾夕法尼亚州最低工资却保持不变。

这两个州的快餐店仿佛被随机地分配到处理组(新泽西州)与控制组(宾夕法尼亚州)。

样本数据包括两个州的快餐店在实施新法前后的全职雇佣人数(full-time equivalent employment)。

全职雇佣人数的计算公式为“全职员工人数 +0.5×兼职员工人数”。

根据 Card and Krueger (1994) Table 3 的数据，DID 估计结果为

$$
\begin{aligned}
&(\bar{y}_{treat,2} - \bar{y}_{treat,1}) - (\bar{y}_{control,2} - \bar{y}_{control,1}) \\
&= (21.03 - 20.44) - (21.17 - 23.33) \\
&= 0.59 - (-2.16) \\
&= 2.75
\end{aligned}
\tag{17.9}
$$

新法实施后处理组(新泽西快餐店)的平均雇佣人数增加了 0.59 人(从 20.44 人增加到 21.03 人)，而同期控制组(宾夕法尼亚快餐店)的平均雇佣人数则下降了 2.16 人(从 23.33 人减少到 21.17 人)。

DID 估计结果显示，提高最低工资的处理效应为增加全职雇佣人数 2.75 人。这与当时理论预期的负效应在符号上相反。

17.2 双重差分估计量

更一般地，可以将双重差分法推广至多期面板($t=1,\cdots,T$)，即所谓“多期 DID”。为此，定义如下处理期虚拟变量 $post_t$：

$$post_t=\begin{cases}1 & \text{若时期 } t \text{ 属于处理后}\\ 0 & \text{若时期 } t \text{ 属于处理前}\end{cases} \tag{17.10}$$

根据处理组虚拟变量$treat_i$与处理期虚拟变量 $post_t$的不同取值组合，可将样本分为四个子样本，并计算结果变量y的相应样本均值，参见表 17.1：

表 17.1 多期的双重差分法

		$post_t$	
		0	1
$treat_i$	0	$\overline{y}_{control,\ pre}$	$\overline{y}_{control,\ post}$
	1	$\overline{y}_{treat,\ pre}$	$\overline{y}_{treat,\ post}$

$\overline{y}_{control,\ pre}$：控制组个体在处理前各期$(treat_i = 0, post_t = 0)$的均值

$\overline{y}_{control,\ post}$:控制组个体在处理后各期$(treat_i = 0, post_t = 1)$的均值

$\overline{y}_{treat,\ pre}$： 处理组个体在处理前各期$(treat_i = 1, post_t = 0)$的均值

$\overline{y}_{treat,\ post}$： 处理组个体在处理后各期$(treat_i = 1, post_t = 1)$的均值

多期的双重差分估计量可写为

$$\Delta\bar{y}_{treat} - \Delta\bar{y}_{\text{control}} \equiv (\bar{y}_{treat,\,post} - \bar{y}_{treat,\,pre}) - (\bar{y}_{control,\,post} - \bar{y}_{control,\,pre}) \tag{17.11}$$

直接通过双重差分估计 ATT 不易计算估计量的标准误，也无法加入控制变量。在实践中，一般通过回归的方法进行 DID 估计。

处理组个体在处理前($post_t = 0$)未受政策冲击，但在处理后($post_t = 1$)则受到政策冲击；而控制组个体始终未受冲击。

样本中个体i在第t期的处理状态(treatment status)可用交互项$treat_i \times post_t$来表示，也称为**处理变量**(treatment variable)或**政策虚拟变量**(policy dummy)。

将此交互项放入双向固定效应模型可得：

$$y_{it} = \alpha + \beta treat_i \times post_t + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T) \tag{17.12}$$

其中，u_i为个体固定效应，λ_t为时间固定效应，而扰动项ε_{it}称为“个殊性冲击”(idiosyncratic shock)或“暂时性冲击”(transitory shock)。

对于参数β的双向固定效应估计量$\hat{\beta}$即为 DID 估计量。

只有处理组个体($treat_i = 1$)到了处理期($post_t = 1$)，其结果变量y_{it}才会加上β。

给定$treat_i = 1$与$post_t = 0$，处理组在处理前的条件期望为

$$\mathrm{E}\left(y_{it} \mid treat_i = 1, post_t = 0\right) = \alpha + \mathrm{E}(u_i \mid treat_i = 1) + \lambda_1^* \quad (17.13)$$

其中，$\lambda_1^* \equiv \mathrm{E}(\lambda_t \mid post_t = 0)$为处理前的平均时间效应。

处理组在处理后的条件期望为

$$\mathrm{E}\left(y_{it} \mid treat_i = 1, post_t = 1\right) = \alpha + \beta + \mathrm{E}(u_i \mid treat_i = 1) + \lambda_2^* \quad (17.14)$$

其中，$\lambda_2^* \equiv \mathrm{E}(\lambda_t \mid post_t = 1)$为处理后的平均时间效应。

直接将(17.14)式减去(17.13)式，则为差分估计量，其结果为$\beta + (\lambda_2^* - \lambda_1^*)$。

差分估计量混杂了时间趋势$(\lambda_2^* - \lambda_1^*)$，故并不可靠。

考察控制组在处理前的条件期望：

$$\mathrm{E}\left(y_{it} \mid treat_i = 0, post_t = 0\right) = \alpha + \mathrm{E}(u_i \mid treat_i = 0) + \lambda_1^* \quad (17.15)$$

而控制组在处理后的条件期望为

$$\mathrm{E}\left(y_{it} \mid treat_i = 0, post_t = 1\right) = \alpha + \mathrm{E}(u_i \mid treat_i = 0) + \lambda_2^* \quad (17.16)$$

将(17.16)式减去(17.15)式，可得控制组的差分为$(\lambda_2^* - \lambda_1^*)$。

总体中的双重差分估计量[(17.14) – (17.13)] – [(17.16) – (17.15)]的结果正是β，相应的样本估计量等价于(17.11)式的 DID 估计量。

更一般地，可在方程(17.12)中引入协变量$\mathbf{w}_{it}$，即为在实践中常用的经典 DID 模型：

$$y_{it} = \alpha + \beta treat_i \times post_t + \boldsymbol{\gamma}'\mathbf{w}_{it} + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T) \tag{17.17}$$

上式只是在双向固定效应模型中加入了交互项$treat_i \times post_t$，故可照常使用双向固定效应的估计方法；比如，使用组内估计量或 LSDV 法。

17.3 平行趋势图

DID 模型的关键前提为平行趋势假定，故对此假定须进行检验。

一种方法是画时间趋势图，直观地考察处理组与控制组结果变量的时间趋势在处理前是否平行。

DID 方法最早由 Ashenfelter (1978)引入经济学，研究就业培训对于个体收入的影响。

Ashenfelter (1978)发现，处理组个体在参加培训的 1964 年以及之前的 1963 年，其平均收入不仅相对于控制组下降，而且绝对地下降，称为“阿氏沉降”(Ashenfelter's dip)，参见图 17.2。

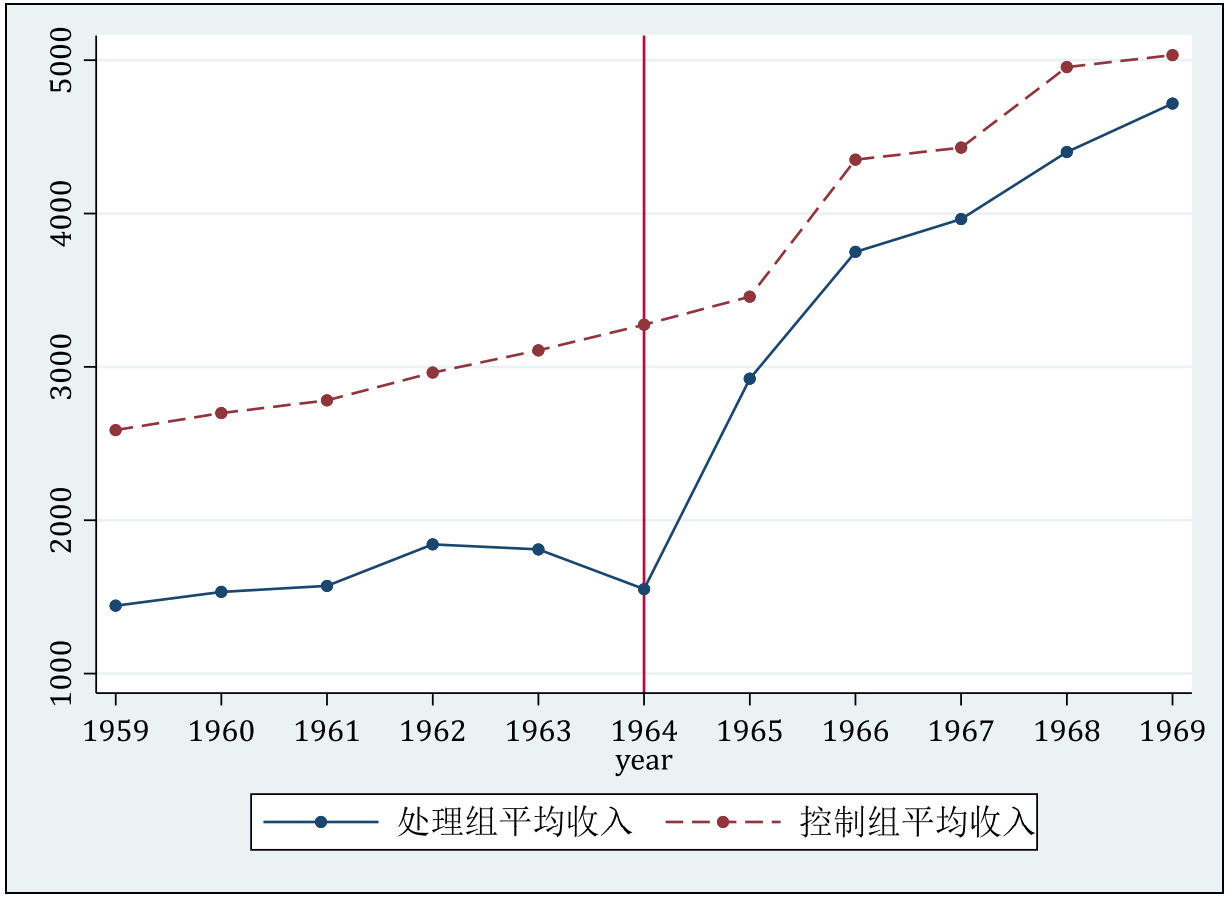

图 17.2 存在阿氏沉降的时间趋势图

注：数据来自 Ashenfelter (1978, p. 51)。

所有个体均为男性白人(white male)，实线为处理组(参与项目者)的平均收入，而虚线为控制组(未参与项目者)的平均收入。

在处理前(pretreatment)，处理组与控制组的时间趋势并不平行。

可能原因是，在 1963-1964 年运气欠佳的收入下降者中(ε_{it}为绝对值较大的负向冲击)，较多人“自我选择”(self select)参加就业培训，从而导致阿氏沉降。

个殊性冲击(idiosyncratic shock)ε_{it}与分组变量$treat_i$存在相关性(在此例中为负相关)，导致内生性偏差。

由于y_{it}的时间趋势可视为每期ε_{it}的累积结果，故若ε_{it}与$treat_i$

相关(存在内生性)，则平行趋势假定必然无法满足。

由于平行趋势假定不满足，故 Ashenfelter (1978)得出结论，此样本数据并不适用 DID，建议以随机实验研究就业培训的效应。

如果在处理前，处理组与控制组的时间趋势大致平行，则可增强对平行趋势假定的信心。

但即使在政策干预前两组的时间趋势相同，也无法保证二者在干预后的时间趋势也相同。

由于处理组在干预后的时间效应不可观测(已与处理效应混杂在一起)，故平行趋势检验在本质上不可检验。

实践中的平行趋势检验一般仅针对处理前的时间趋势而进行。

对于两期面板，由于处理前只有一期数据，故无法画时间趋势图。

例如，Card and Krueger (1994)只有两期数据，故无法进行平行趋势检验；但由于可视为自然实验，故基本不用担心内生性。

17.4 平行趋势检验

对于政策干预前有多期的面板数据，可通过画时间趋势图初步考察平行趋势假设，但通过画图所得结论毕竟带有主观性。

即使平行趋势假定(17.6)在总体中成立，抽样所得的时间趋势也未必完全平行。

处理组与控制组的时间趋势究竟要多么平行，才算足够平行？

严格的结论有赖于统计检验。

将交互项$treat_i \times post_t$进一步细分为$treat_i$与各期虚拟变量的乘积。

记D_{2t}为第 2 期的虚拟变量(若为第 2 期，则取值为 1；否则取值为 0)，D_{3t}为第 3 期的虚拟变量；以此类推，直至D_{Tt}为第T期的虚拟变量。考虑以下模型：

$$y_{it} = \alpha + \sum_{k=2}^{T} \beta_k treat_i \times D_{kt} + \boldsymbol{\gamma}' \mathbf{w}_{it} + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T)$$

(17.18)

其中，未放入的第 1 期交互项$treat_i \times D_{1t}$被作为基准的参照系。

上式依然是双向固定效应模型。

假定第$1, \cdots, T_0$期为处理前(pretreatment)，而第$T_0 + 1, \cdots, T$期为处理后(posttreatment)。

给定处理前的某个时期$k \in \{2, \cdots, T_0\}$，从方程(17.18)可知，控制组的时间趋势为λ_k，而处理组的时间趋势为$\beta_k + \lambda_k$。

平行趋势假定意味着，$\beta_2 = \cdots = \beta_{T_0} = 0$。

平行趋势假定要求，所有处理前的β_k，即$\left\{\beta_2, \cdots, \beta_{T_0}\right\}$均不显著，或联合不显著。

以β_2为例，若β_2显著不等于 0，则处理组在第 2 期的时间趋势为$\beta_2 + \lambda_2$；而控制组在第 2 期的时间趋势为λ_2，故二者的时间趋势在第 2 期不相同。

处理后的系数$\left\{\beta_{T_0+1}, \cdots, \beta_T\right\}$也包含信息，表示**动态处理效应**(dynamic treatment effects)，允许处理后各期的处理效应各不相同。

还可将$\left\{\hat{\beta}_2,\cdots,\hat{\beta}_T\right\}$的点估计与置信区间画图，更直观地呈现平行趋势检验与动态处理效应的结果。

这种平行趋势的检验方法类似于金融学与会计学中常用的**事件分析**(event study)；比如研究上市公司并购对于股价的影响。

17.5 交叠 DID

以上介绍的双重差分法均隐含地假设所有处理组个体在同一时期受到政策冲击。

但现实中的政策实施也可能交错进行，即所谓**交叠处理**(staggered adoption)。

这意味着，每位个体受到政策冲击的时间不完全相同。

比如，增值税改革试点在不同城市分批推出。进一步，一旦个体受到政策冲击，则在样本期间无法退出。

此时，由于每位个体受到处理的时间不尽相同，故无法使用交互项$treat_i \times post_t$表示个体i在第t期的处理状态(treatment status)；

但仍可使用如下“处理变量”(treatment variable)来表示：

$$D_{it} = \begin{cases} 1 & \text{若个体 } i \text{ 在第 } t \text{ 期受到处理} \\ 0 & \text{若个体 } i \text{ 在第 } t \text{ 期未受处理} \end{cases} \tag{17.19}$$

只要将处理变量 D_{it} 替代经典 DID 模型(17.17)的交互项 $treat_i \times post_t$，即可得到**交叠 DID**(staggered DID)模型：

$$y_{it} = \alpha + \beta D_{it} + \boldsymbol{\gamma}'\mathbf{w}_{it} + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T) \tag{17.20}$$

这依然是双向固定效应模型。

由于不同个体受政策冲击的时间不尽相同，故无法画交叠 DID 的时间趋势图。

另一方面，平行趋势检验则仍可进行(在此从略)。

交叠 DID 的更严重问题是，若存在**异质性处理效应**(heterogeneous treatment effects)，则以双向固定效应模型估计交叠 DID 将导致偏差。

在异质性处理效应的情况下，每位个体的处理效应可以各不相同，此时上式可写为：

$$y_{it} = \alpha + \beta_{it} D_{it} + \boldsymbol{\gamma}' \mathbf{w}_{it} + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T) \tag{17.21}$$

其中，β_{it}为个体i在第t期的处理效应，既可随个体i而变，也可随时期t而变(即动态处理效应)。在异质性处理效应情况下，如何估计交叠 DID 模型是目前活跃的研究前沿。

17.6 双重差分法的 Stata 案例

使用 Cao and Chen (2022)的数据 cao_chen.dta 进行双重差分法的演示。

该文使用清朝 1650-1911 年 575 个县的面板数据，研究 1826 年“漕粮海运”政策冲击导致大运河逐渐失修废弃，并因主要贸易路线中断而引发社会动荡乃至叛乱。

样本数据包含大运河流经的六个省份共 575 个县，其中运河流经的县称为“运河县”(以虚拟变量 *canal* 表示)，构成处理组(*canal*=1)；而运河未流经的县称为“非运河县”，构成控制组(*canal*=0)。

长期以来，大运河一直是清政府从南向北运输漕粮的主要通道，同时也承载着重要的贸易功能。由于 1825 年暴雨洪灾导致大运河在黄河交汇处决堤，1826 年清政府开启首次“漕粮海运”试验并取得成功。

由于海运的成本优势，漕粮海运逐渐成为主导模式，而大运河则渐渐失修、堵塞乃至废弃。因此，从 1826 年开始，由运河县构成的处理组开始受到漕粮海运的政策冲击。

被解释变量的原始数据来自《清实录》。作者通过搜索关键字“匪”，并阅读原文，以确定叛乱发生的县与年份，由此得到第i个县在第t年的叛乱发生(rebellion onset)数目$rebel_num_{it}$(不包括从去年延续到今年的叛乱)。

首先，打开数据集 cao_chen.dta，以 *county* 为面板变量而 *year* 为时间变量，设为面板数据，并考察变量 *rebel_num* 的取值分布。

```
. use cao_chen.dta, clear
. xtset county year
```

```
Panel variable: county (strongly balanced)
 Time variable: year, 1650 to 1911
         Delta: 1 unit
```

```
. tab rebel_num
```

number of rebellion onset	Freq.	Percent	Cum.
0	149,541	99.26	99.26
1	1,075	0.71	99.98
2	33	0.02	100.00
3	1	0.00	100.00
Total	150,650	100.00	

样本容量为 150,650，但绝大多数取值为 0；而最大值为 3，但仅出现一次。

考虑到一个县发生叛乱的概率应与该县人口正相关，故将此叛乱次数除以 *pop1600* (1600 年的县人口)，以此定义 *rebel_density*，并考察其统计特征：

```
. gen rebel_density = rebel_num / (pop1600/1000000)
. sum rebel_density
```

Variable	Obs	Mean	Std. dev.	Min	Max
rebel_dens~y	140,432	.4603983	7.117167	0	395.9637

“pop1600/1000000”将 1600 年县人口的单位变为百万人。

变量 *rebel_density* 有很多零值，但最大值却高达 395.96。

通过直方图考察其分布。

```
. hist rebel_density,fraction
```

其中，选择项“fraction”表示直方图的纵轴为“比重”。

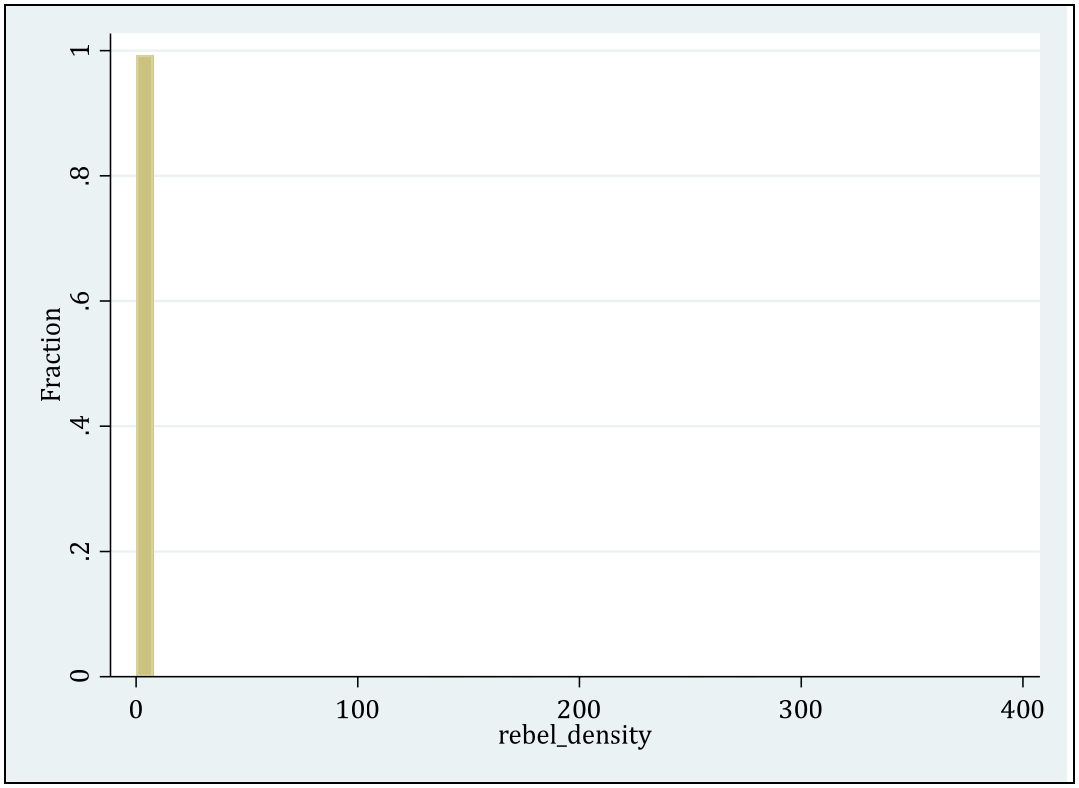

图 17.3 变量 *rebel_density* 的直方图

尽管变量 *rebel_density* 有大量零值，但其分布明显右偏(right skewed)，存在极端值(右侧有长尾)。

解决右偏分布极端值问题的常用方法为取对数，但该变量有很多零值，故无法直接取对数。

Cao and Chen (2022)使用**逆双曲正弦函数**(inverse hyperbolic sine function)对 *rebel_density* 进行变换。

此函数为对数的推广，函数形式为

$$\operatorname{asinh}(x) \equiv \ln\left(x + \sqrt{x^2 + 1}\right) \qquad (17.22)$$

该函数对于任何实数都有定义，且是关于原点对称的奇函数，在 Stata 中用 `asinh()`来表示。显然，$\operatorname{asinh}(0) = 0$。

若自变量x较大，则$\text{asinh}(x) \approx \ln(2x) = \ln 2 + \ln x$，约等于对数$\ln x$加上常数$\ln 2$，故可近似地以弹性或半弹性解释回归系数。

若x较小(本例 *rebel_density* 的均值仅为 0.46，且包含大量零值)，则此近似有较大误差，不宜用弹性或半弹性解释回归系数。

在 Stata 中同时画逆双曲正弦函数与对数函数：

```
. twoway function asinh=asinh(x),range(-20 20) xline(0,lp(dot)) yline(0,lp(dot)) || function log=log(x),range(-20 20) lp(dash)
```

其中，选择项“`range(-20 20)`”限制画图范围从-20 到 20。

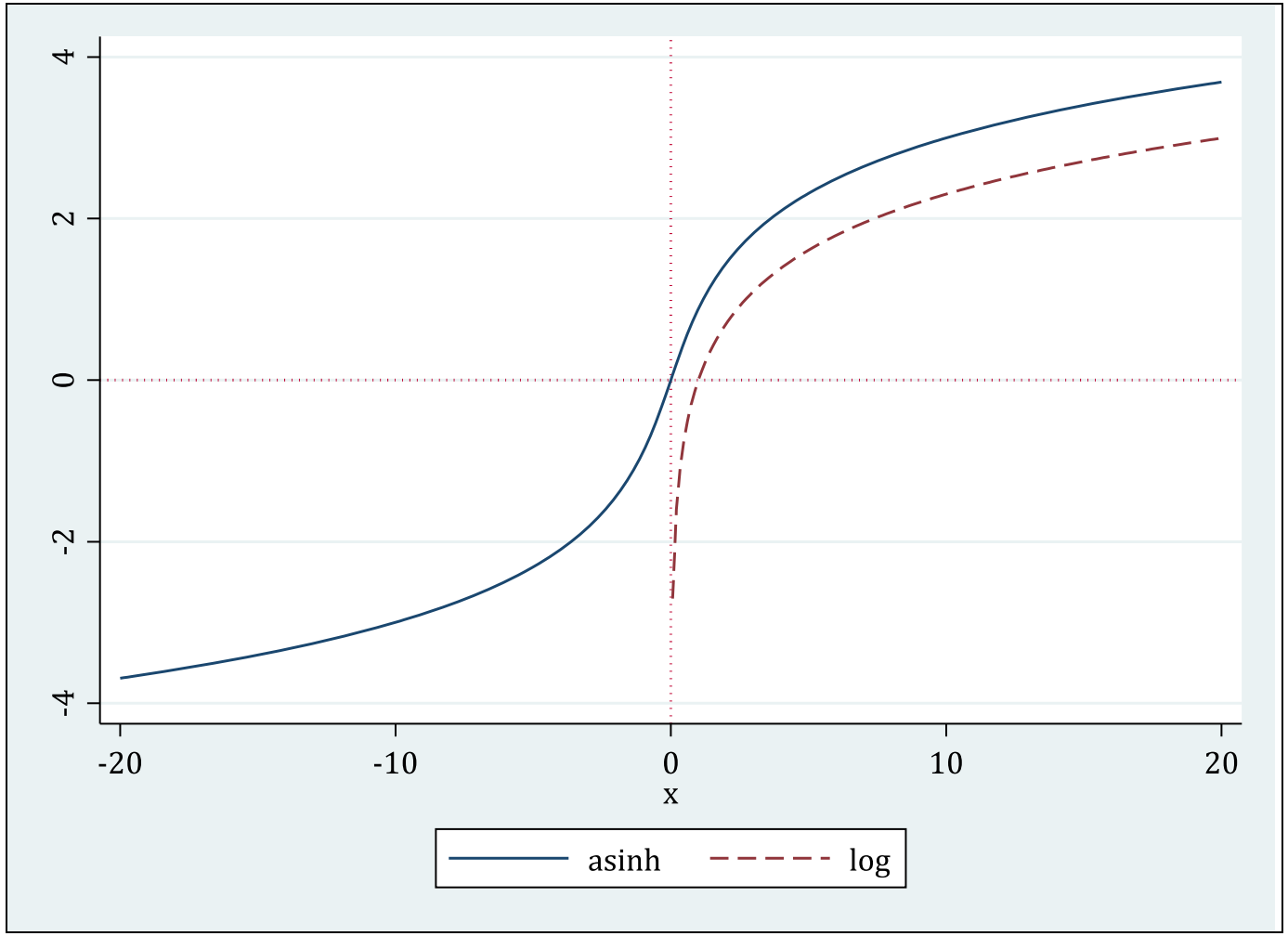

图 17.4 逆双曲正弦函数与对数函数

对变量 *rebel_density* 进行逆双曲正弦变换，记为 *rebel*，并考察其统计特征：

```
. gen rebel = asinh(rebel_density)
. sum rebel
```

Variable	Obs	Mean	Std. dev.	Min	Max
rebel	140,432	.0329779	.389762	0	6.674471

经逆双曲正弦变换后，均值为 0.033；最大值下降为 6.67，极端值问题已缓解；而大量的零值经变换后依然是零值，因为 $\text{asinh}(0)=0$。

Cao and Chen (2022)主要使用变换后的 *rebel* 作为被解释变量；而直接以 `rebel_density` 为被解释变量的回归结果依然类似。

在使用双重差分法之前，首先画平行趋势图，以初步考察平行趋势假定是否成立。

根据处理组虚拟变量 *canal* 与时间变量 *year* 将数据集压缩，以计算变量 *rebel* 在处理组(*canal*=1)与控制组(*canal*=0)的每年平均值。

为避免损失原数据集，可使用命令 `preserve`，将原数据集暂时存放于内存，在完成画图后再以命令 `restore` 将原数据集调回：

```
. preserve

. collapse (mean) mean_year=rebel,by(canal year)
```

```
. twoway (connect mean_year year if
canal==1,msize(small)) (connect mean_year year if
canal==0,lp(dash) msize(small)
xline(1825,lp(dash)) legend(label(1 Canal
Counties) label(2 Non-canal Counties)))

. restore
```

命令 `collapse` 将数据根据 *canal* 与 *year* 的不同取值分割为若干子样本，在每个子样本中计算变量 *rebel* 的均值，并记为 *mean_year*。

命令 `twoway` 则分别画变量 *mean_year* 在处理组(`canal==1`)与控制组(`canal==0`)的时间趋势图。

选择项“msize(small)”指定散点的尺度为“小”(msize 表示 marker size)。

选择项“xline(1825,lp(dash))”指示在 1825 年处画一条垂直虚线。

选择项“legend()”用于指定图例，结果参见图 17.5。

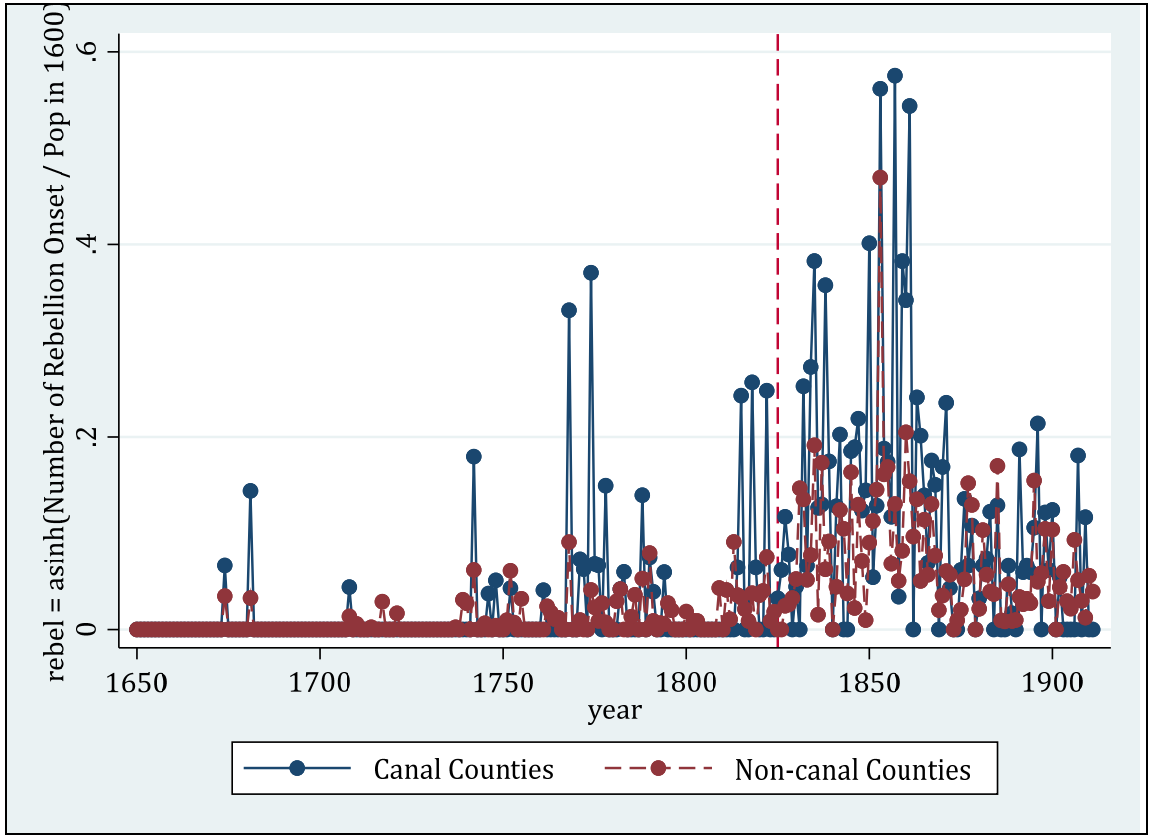

图 17.5 每年的平行趋势图

每年的平行趋势图噪音很大，因为某县在某年发生叛乱毕竟是小概率事件。

下面以每十年为单位画平行趋势图。以政策冲击开始的 1826 年为基准，将数据划分为每十年的区间，并以变量 *period* 表示。

```
. gen period=floor((year-1826)/10)*10

. replace period=-60 if period<-60
```

其中，函数 `floor(x)`表示小于或等于 `x` 的最大整数，

例如 `floor(1.6)`的取值为 1，而 `floor(-1.6)`的取值为-2。

对于$1826 \le year \le 1835$，$period = 0$；对于$1836 \le year \le 1845$，$period = 10$；而对于$1816 \le year \le 1825$，则 $period = -10$；以此类推。

遵照 Cao and Chen (2022)的做法，对于距政策冲击 50 年以上的观测值，将其变量 *period* 统一记为-60，以便将年代较久远的观测值均作为参照系。

使用命令 `tabstat` 考察在 *period* 取值不同的子样本中变量 *year* 的最小值与最大值。

```
. tabstat year,by(period) stat(min max) nototal
```

Summary for variables: year
Group variable: period

period	Min	Max
-60	1650	1775
-50	1776	1785
-40	1786	1795
-30	1796	1805
-20	1806	1815
-10	1816	1825
0	1826	1835
10	1836	1845
20	1846	1855
30	1856	1865
40	1866	1875
50	1876	1885
60	1886	1895
70	1896	1905
80	1906	1911

根据 *period* 的不同取值生成一系列虚拟变量，以备后续使用。

```
. tab period, gen(period)
. describe period*
```

Variable name	Storage type	Display format	Value label	Variable label
period	float	%9.0g		
period1	byte	%8.0g		period== -60.0000
period2	byte	%8.0g		period== -50.0000
period3	byte	%8.0g		period== -40.0000
period4	byte	%8.0g		period== -30.0000
period5	byte	%8.0g		period== -20.0000
period6	byte	%8.0g		period== -10.0000
period7	byte	%8.0g		period== 0.0000
period8	byte	%8.0g		period== 10.0000
period9	byte	%8.0g		period== 20.0000
period10	byte	%8.0g		period== 30.0000
period11	byte	%8.0g		period== 40.0000
period12	byte	%8.0g		period== 50.0000
period13	byte	%8.0g		period== 60.0000
period14	byte	%8.0g		period== 70.0000
period15	byte	%8.0g		period== 80.0000

压缩数据，以 *canal* 与 *period* 的不同取值分割子样本，并根据各子样本的 *rebel* 均值画平行趋势图。

```
. preserve
. collapse (mean) mean_decade=rebel if period<80,by(canal period)
. twoway (connect mean_decade period if canal==1,xtitle(Number of years since the 1826 Reform)) (connect mean_decade period if canal==0,lp(dash) xline(-5,lp(dash)) xlabel(-60 "-60" -50 "-50" -40 "-40" -30 "-30" -20 "-20" -10 "-10" 0 "10" 10 "20" 20 "30" 30 "40" 40 "50" 50 "60" 60 "70" 70 "80") legend(label(1 Canal Counties) label(2 Non-canal Counties)))
. restore
```

其中，命令 `collapse` 所用的条件“`if period<80`”限制观测值不超过政策冲击 80 年(因 *period* = 80 不包含完整的十年数据，去掉此限制不影响结果)。

选择项 `xlabel()`用于指定横轴的标签，并将 0 记为 10(横轴不显示 0)，10 记为 20，以此类推。

结果参见图 17.6。

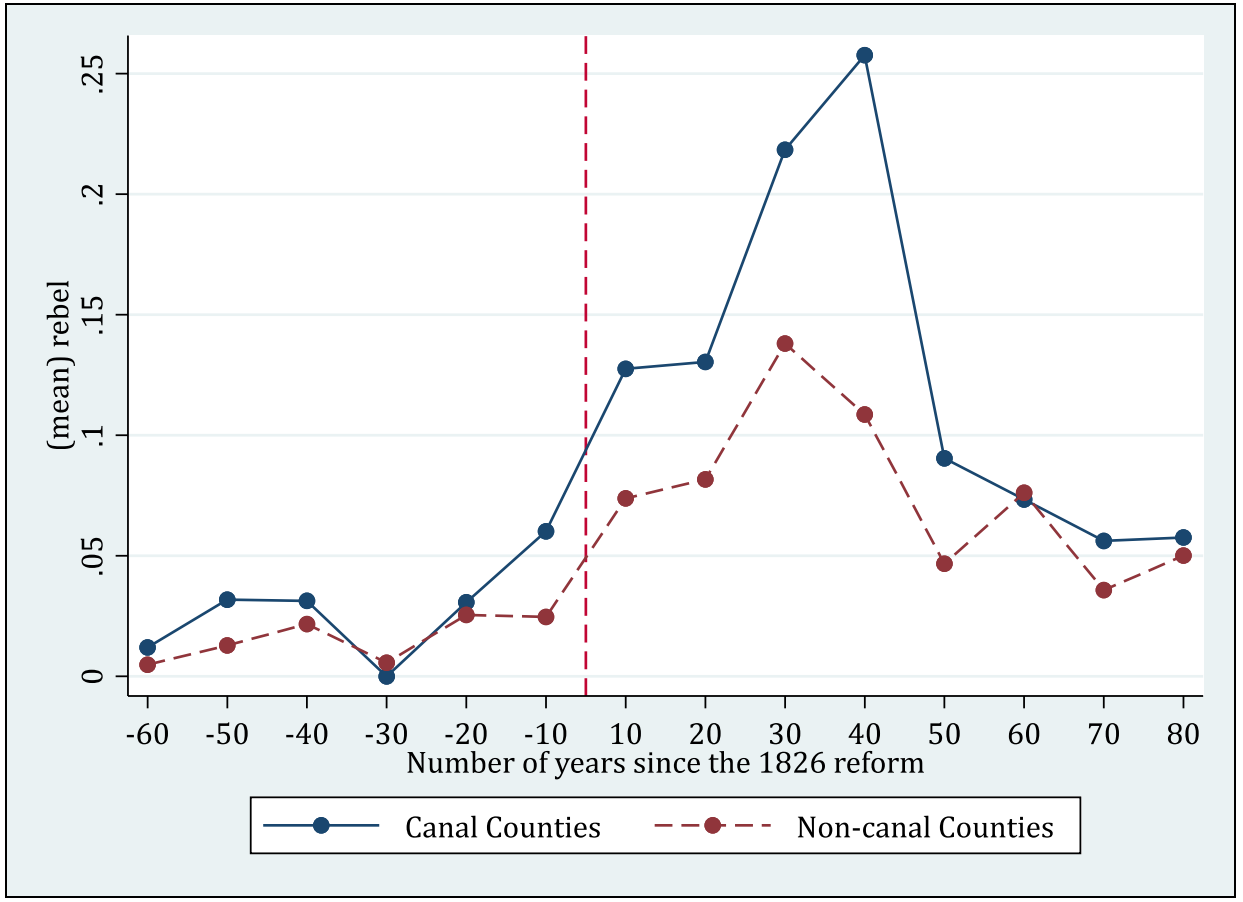

图 17.6 每十年的平行趋势图

似乎很难判断平行趋势假定是否成立，下面进行平行趋势检验。

在进行平行趋势检验时，Cao and Chen (2022)还使用了一系列的控制变量(若不用控制变量结果也类似)，包括 *drought*(是否有旱灾)，*flood*(是否有水灾)，*disaster*(是否出现反常气温)，*maize*(是否已引种玉米)，*sweetpotato*(是否已引种红薯)。

定义处理期虚拟变量为 *post*，若 *year* ≥1826，则取值为 1；反之，取值为 0。

考虑到有些因素不随时间而变，比如土地面积对数(*larea*)、土地崎岖指数(*rug*)、1600 年人口密度(*popden1600*)、小麦种植适宜指数(*wheat*)、稻米种植适宜指数(*rice*)，故将这些变量乘以 *post*，以其交互项作为控制变量，即 *larea_after* (*larea*×*post*)，*rug_after* (*rug*×*post*)，*popden1600_after* (*popden1600*×*post*)，*wheat_after* (*wheat*×*post*)，*rice_after* (*rice*×*post*)。

其他控制变量还包括 *drought_after* (*drought*×*post*)，*flood_after* (*flood*×*post*)，*disaster_after* (*disaster*×*post*)，*maize_after* (*maize*×*post*)，*sweetpotato_after* (*sweetpotato*×*post*)。

由于协变量较多，下面以命令 `global` 定义一个名为“*cov*”(表示 covariates)的“全局宏”(global macro)，指代上述所有协变量：

```
. global cov drought flood disaster maize
sweetpotato larea_after rug_after
popden1600_after wheat_after rice_after
flood_after drought_after disaster_after
maize_after sweetpotato_after
```

定义好全局宏 *cov* 后，只需使用`$cov`(在之前加上美元符号)，即可调用这些控制变量。

由于时间维度较长(共 262 年)，若使用命令“`xtreg,fe`”进行双向固定效应的估计，则须加入很多年度虚拟变量。

Cao and Chen (2022)使用非官方命令 `reghdfe`，可方便地“吸收”(absorb)大量虚拟变量。`hdfe` 表示“高维固定效应” (high dimensional fixed effects)，即存在多个方向的固定效应(可能不止双向固定效应)，且有些固定效应需要引入大量的虚拟变量才能控制。

为了说明命令 `reghdfe` 的算法，考虑如下模型：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{d}_{it}'\boldsymbol{\gamma} + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, \cdots, T) \quad (17.23)$$

其中，$\mathbf{d}_{it}$ 由大量虚拟变量所组成(虚拟变量可随个体变、随时间变，或同时随个体与时间变)。

由于$\mathbf{d}_{it}$的维度可能很高，故命令 reghdfe 的算法分两步进行。

第一步，将y_{it}与$\mathbf{x}_{it}$分别对所有虚拟变量$\mathbf{d}_{it}$进行回归，将所得残差分别记为$\tilde{y}_{it}$与$\tilde{\mathbf{x}}_{it}$。

第二步，把$\tilde{y}_{it}$对$\tilde{\mathbf{x}}_{it}$进行回归，则根据 Frisch-Waugh-Lovell 定理，所得估计量$\hat{\boldsymbol{\beta}}$等价于对方程(17.23)直接进行回归。

使用此算法的方便之处在于，第一步回归很容易，因为对虚拟变量进行回归类似于计算子样本的样本均值。

第二步回归已不包含高维虚拟变量。命令 reghdfe 还优化了具体算法，以提升运行速度。

安装 reghdfe 的命令为：

```
. ssc install reghdfe,replace
```

平行趋势检验的命令为：

```
. reghdfe rebel c.canal#(c.period2-period14) $cov if period<80, absorb(i.county i.year c.prerebel#i.year i.prov#i.year) cluster(county)
```

其中，“c.canal#(c.period2-period14)”表示 *canal* 与 *period2-period14* 的系列交互项(前缀 c.表示将这些变量视为连续变量)；

“$cov”表示调用上述系列协变量；

“if period<80”限制 *period* 小于 80(去掉此限制结果类似)；

“cluster(county)”表示以 *county* 为聚类变量，计算聚类稳健标准误。

选择项 absorb()包含了一系列需要“吸收”的虚拟变量，其中 i.county 为各县的虚拟变量，i.year 为各年的虚拟变量，c.prerebel#i.year 为变量 *prerebel* (1826 年前该县人均累计叛乱次数的逆双曲正弦变换)与各年虚拟变量的乘积，而 i.prov#i.year 为各省虚拟变量(*prov* 为省 ID)与各年虚拟变量的乘积。

HDFE Linear regression — Number of obs = 137,216
Absorbing 4 HDFE groups — F(28, 535) = 4.08
Statistics robust to heteroskedasticity — Prob > F = 0.0000
R-squared = 0.0666
Adj R-squared = 0.0503
Within R-sq. = 0.0031
Number of clusters (county) = 536 — Root MSE = 0.3781

(Std. err. adjusted for 536 clusters in county)

rebel	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
c.canal#c.period2	.0068628	.014053	0.49	0.626	-.0207431	.0344687
c.canal#c.period3	-.0078396	.0158541	-0.49	0.621	-.0389836	.0233045
c.canal#c.period4	-.0051483	.0029215	-1.76	0.079	-.0108873	.0005907
c.canal#c.period5	-.0156022	.0149164	-1.05	0.296	-.0449041	.0136997
c.canal#c.period6	.0005762	.0222107	0.03	0.979	-.0430546	.0442071
c.canal#c.period7	.0702413	.0304988	2.30	0.022	.0103293	.1301533
c.canal#c.period8	.0328635	.0349834	0.94	0.348	-.0358581	.1015851
c.canal#c.period9	.0684217	.0507336	1.35	0.178	-.0312398	.1680832
c.canal#c.period10	.1092208	.0532055	2.05	0.041	.0047036	.2137381
c.canal#c.period11	.0316668	.0445372	0.71	0.477	-.0558223	.119156
c.canal#c.period12	.0038247	.032332	0.12	0.906	-.0596885	.0673379
c.canal#c.period13	-.0062874	.0283768	-0.22	0.825	-.062031	.0494563
c.canal#c.period14	-.0137393	.0252731	-0.54	0.587	-.0633861	.0359075

larea_after	.0313119	.0075697	4.14	0.000	.0164419	.0461819
rug_after	-.0002663	.0000456	-5.84	0.000	-.0003559	-.0001766
disaster	-.0037653	.0042262	-0.89	0.373	-.0120672	.0045366
disaster_after	-.0054553	.01176	-0.46	0.643	-.0285568	.0176462
flood	.0022678	.00375	0.60	0.546	-.0050987	.0096342
drought	.0062989	.0031899	1.97	0.049	.0000327	.0125651
flood_after	.0008002	.0191872	0.04	0.967	-.0368912	.0384917
drought_after	.0036165	.0135557	0.27	0.790	-.0230123	.0302454
popden1600_after	-.0232826	.0049565	-4.70	0.000	-.0330193	-.013546
maize	-.0045284	.0045506	-1.00	0.320	-.0134675	.0044108
maize_after	.031228	.015369	2.03	0.043	.0010371	.061419
sweetpotato	-.006068	.0035705	-1.70	0.090	-.0130819	.000946
sweetpotato_after	.0024114	.0106115	0.23	0.820	-.0184339	.0232568
wheat_after	.0333789	.011927	2.80	0.005	.0099494	.0568083
rice_after	-.0042705	.0150319	-0.28	0.776	-.0337993	.0252582
_cons	-.0181251	.0190115	-0.95	0.341	-.0554714	.0192212

Absorbed degrees of freedom:

Absorbed FE	Categories	- Redundant	= Num. Coefs	
county	536	536	0	*
year	256	0	256	
year#c.prerebel	256	0	256	?
prov#year	1536	256	1280	

? = number of redundant parameters may be higher
* = FE nested within cluster; treated as redundant for DoF computation

在政策冲击前，*canal* 与 *period2-period6* 的交互项均不显著。

在政策冲击后，*canal*×*period7* 与 *canal*×*period10* 的系数均在 5% 的水平上显著。

上表下部的“Absorbed degrees of freedom”可以忽略(下文将不再展示)，只是汇报了正确计算标准误所需的自由度调整。

检验 *canal* 与 *period2-period6* 的交互项的联合显著性：

```
. test c.canal#c.period2 c.canal#c.period3
c.canal#c.period4 c.canal#c.period5
c.canal#c.period6
```

```
 ( 1)  c.canal#c.period2 = 0
 ( 2)  c.canal#c.period3 = 0
 ( 3)  c.canal#c.period4 = 0
 ( 4)  c.canal#c.period5 = 0
 ( 5)  c.canal#c.period6 = 0

       F(  5,   535) =    1.08
            Prob > F =    0.3714
```

F 检验的 p 值为 0.37，政策冲击前的所有交互项联合不显著，故可接受平行趋势假定。

将所有各期交互项的点估计与置信区间(已汇报于回归结果中)，以画图形式呈现。为此，下载非官方命令 `coefplot`。

```
. ssc install coefplot,replace
```

```
. coefplot,vertical keep(c.canal*)
msymbol(circle_hollow) ciopts(lp(dash)
recast(rcap)) addplot(line @b @at) xtitle(Number
of years since the 1826 Reform)  xlabel(1 "-50" 2
"-40" 3 "-30" 4 "-20" 5 "-10" 6 "10" 7 "20" 8 "30"
9 "40" 10 "50" 11 "60" 12 "70" 13 "80") xline(5.5,
lp(dash) lwidth(vthin)) ytitle(Coefficients)
ylabel(-0.1(0.05)0.3) yline(0,lp(dash)
lwidth(vthin))
```

选择项“vertical”表示画纵向的置信区间。

“keep(c.canal*)”表示只画以“c.canal”开头的交互项系数;“msymbol(circle_hollow)”以空心圆表示回归系数的点估计。

“ciopts(lp(dash) recast(rcap))”表示以虚线画置信区间，但两端为小短横的“帽子”。

“addplot(line @b @at)”将回归系数的点估计连结成线。

“xtitle()”指定横轴的标题

“xlabel()”指定横轴的标签

“xline(5.5, lp(dash) lwidth(vthin))”指定以很细(very thin)的虚线在$x=5.5$处画一条垂直线(以区分政策冲击前后)。

结果参见图 17.7。

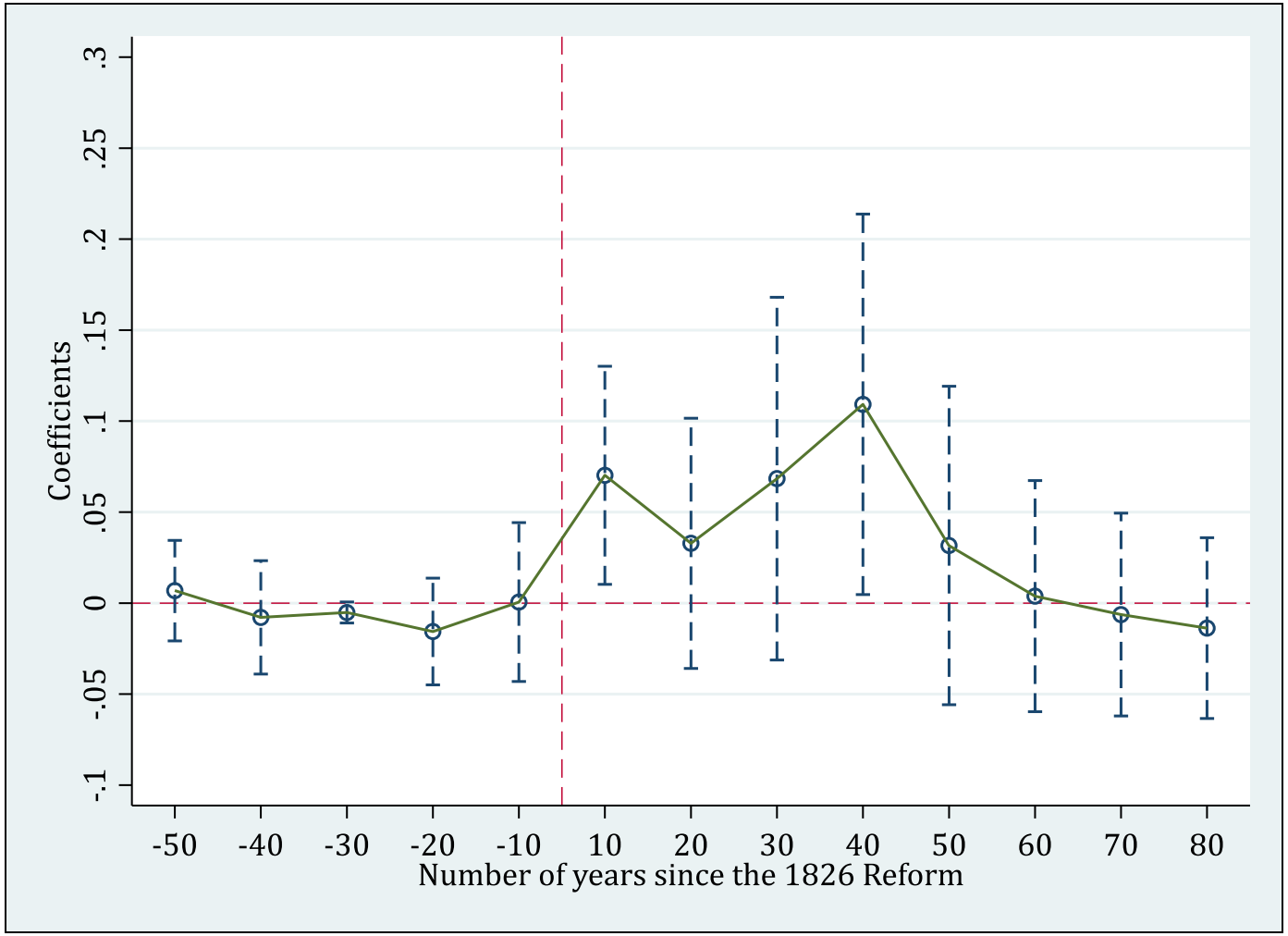

图 17.7 平行趋势检验的可视化

在政策冲击前，各期回归系数的 95%置信区间均包含 0，故这些交互项均不显著，因此平行趋势假定成立。

政策冲击后第 1 个十年与第 4 个十年的回归系数的置信区间均不包含 0，故在 5%的水平上显著。

下面正式进行 DID 回归。

在数据集中，作为处理变量的交互项记为 *canal_post* (即 *canal*×*post*)。

第一，估计常规的双向固定效应模型，并将估计结果存为 fe。

```
. reghdfe rebel canal_post, absorb(i.county i.year) cluster(county)
. est sto fe
```

```
HDFE Linear regression                            Number of obs   =    140,432
Absorbing 2 HDFE groups                           F(   1,    535) =       5.23
Statistics robust to heteroskedasticity           Prob > F        =     0.0226
                                                  R-squared       =     0.0308
                                                  Adj R-squared   =     0.0253
                                                  Within R-sq.    =     0.0002
Number of clusters (county)  =        536         Root MSE        =     0.3848

                              (Std. err. adjusted for 536 clusters in county)
------------------------------------------------------------------------------
             |               Robust
       rebel |  Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  canal_post |    .0380143    .016621     2.29   0.023     .0053639    .0706647
       _cons |    .0313251   .0007227    43.35   0.000     .0299054    .0327447
------------------------------------------------------------------------------
```

第二，在上述回归中加入变量 *prerebel* 与各年虚拟变量的交互项，并将估计结果存为 prerebel。

```
. reghdfe rebel canal_post, absorb(i.county i.year c.prerebel#i.year) cluster(county)

. est sto prerebel
```

HDFE Linear regression — Number of obs = 140,432
Absorbing 3 HDFE groups — F(1, 535) = 4.62
Statistics robust to heteroskedasticity — Prob > F = 0.0321
R-squared = 0.0395
Adj R-squared = 0.0322
Within R-sq. = 0.0002
Number of clusters (county) = 536 — Root MSE = 0.3834

(Std. err. adjusted for 536 clusters in county)

rebel	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
canal_post	.0368671	.0171607	2.15	0.032	.0031565	.0705776
_cons	.0313749	.0007461	42.05	0.000	.0299092	.0328407

第三，加入各省虚拟变量与各年虚拟变量的交互项，并将估计结果存为prov。

```
. reghdfe rebel canal_post, absorb(i.county i.year c.prerebel#i.year i.prov#i.year) cluster(county)

. est sto prov
```

```
HDFE Linear regression                            Number of obs   =    140,432
Absorbing 4 HDFE groups                           F(   1,    535) =       6.90
Statistics robust to heteroskedasticity           Prob > F        =     0.0088
                                                  R-squared       =     0.0632
                                                  Adj R-squared   =     0.0471
                                                  Within R-sq.    =     0.0003
Number of clusters (county)  =        536         Root MSE        =     0.3805

                                (Std. err. adjusted for 536 clusters in county)
```

rebel	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
canal_post	.0453479	.0172577	2.63	0.009	.0114467	.0792491
_cons	.0310062	.0007504	41.32	0.000	.0295322	.0324802

第四，加入各府虚拟变量与时间趋势(*year*)的交互项(变量 *pref* 表示 prefecture，即府的 ID)，并将估计结果存为 `pref`。

```
. reghdfe rebel canal_post, absorb(i.county i.year c.prerebel#i.year i.prov#i.year i.pref#c.year)
```

```
. est sto pref
```

“i.year”表示根据 *year* 的不同取值，生成系列年度虚拟变量。

“c.year” 为时间趋势项(将 *year* 视为连续变量)。

```
HDFE Linear regression                            Number of obs   =    140,432
Absorbing 5 HDFE groups                           F(   1, 137987) =      30.01
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.0663
                                                  Adj R-squared   =     0.0497
                                                  Within R-sq.    =     0.0002
                                                  Root MSE        =     0.3799

------------------------------------------------------------------------------
       rebel | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  canal_post |   .0426956   .0077943     5.48   0.000     .0274188    .0579723
       _cons |   .0311215    .001069    29.11   0.000     .0290262    .0332168
------------------------------------------------------------------------------
```

第五，在上述回归中加入以$cov 所代表的一系列协变量，并将估计结果存为 cov。

```
. reghdfe rebel canal_post $cov, absorb(i.county i.year c.prerebel#i.year i.prov#i.year i.pref#c.year) cluster(county)

. est sto cov
```

HDFE Linear regression — Number of obs = 140,432
Absorbing 5 HDFE groups — F(16, 535) = 5.35
Statistics robust to heteroskedasticity — Prob > F = 0.0000
R-squared = 0.0675
Adj R-squared = 0.0509
Within R-sq. = 0.0015
Number of clusters (county) = 536 — Root MSE = 0.3797

(Std. err. adjusted for 536 clusters in county)

rebel	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
canal_post	.0340131	.0165679	2.05	0.041	.0014669	.0665592
larea_after	.0353451	.0074733	4.73	0.000	.0206644	.0500258
rug_after	-.0002408	.0000457	-5.27	0.000	-.0003305	-.000151
disaster	-.0041912	.0040244	-1.04	0.298	-.0120968	.0037145
disaster_a~r	-.0173285	.0116611	-1.49	0.138	-.0402357	.0055787
flood	.0020658	.0037689	0.55	0.584	-.0053379	.0094695
drought	.0060161	.0031977	1.88	0.060	-.0002654	.0122976
flood_after	-.0064245	.0185975	-0.35	0.730	-.0429575	.0301086
drought_af~r	-.0014186	.0128029	-0.11	0.912	-.0265687	.0237315
popden1600~r	-.0195797	.0044988	-4.35	0.000	-.0284171	-.0107423
maize	-.0078825	.0040151	-1.96	0.050	-.0157698	4.78e-06
maize_after	.0239267	.015651	1.53	0.127	-.0068182	.0546716
sweetpotato	-.0054499	.0048869	-1.12	0.265	-.0150498	.0041501
sweetpotat~r	-.0033825	.010557	-0.32	0.749	-.0241208	.0173559
wheat_after	.0268425	.0105067	2.55	0.011	.006203	.0474819
rice_after	.0124387	.0140688	0.88	0.377	-.0151983	.0400757
_cons	-.0267575	.0195447	-1.37	0.172	-.0651513	.0116363

下面将以上五个 DID 回归结果以表格形式汇总。

```
. esttab fe prerebel prov pref cov,se r2 mtitle
nogap star(* 0.1 ** 0.05 *** 0.01)
```

选择项“nogap”表示在表中每行之间不留空行，以节省空间。

	(1) fe	(2) prerebel	(3) prov	(4) pref	(5) cov
canal_post	0.0380**	0.0369**	0.0453***	0.0427***	0.0340**
	(0.0166)	(0.0172)	(0.0173)	(0.00779)	(0.0166)
larea_after					0.0353***
					(0.00747)
rug_after					-0.000241***
					(0.0000457)
disaster					-0.00419
					(0.00402)
disaster_a~r					-0.0173
					(0.0117)
flood					0.00207
					(0.00377)
drought					0.00602*
					(0.00320)
flood_after					-0.00642
					(0.0186)
drought_af~r					-0.00142
					(0.0128)
popden1600~r					-0.0196***
					(0.00450)
maize					-0.00788*
					(0.00402)
maize_after					0.0239
					(0.0157)
sweetpotato					-0.00545
					(0.00489)
sweetpotat~r					-0.00338
					(0.0106)
wheat_after					0.0268**
					(0.0105)
rice_after					0.0124
					(0.0141)
_cons	0.0313***	0.0314***	0.0310***	0.0311***	-0.0268
	(0.000723)	(0.000746)	(0.000750)	(0.00107)	(0.0195)
N	140432	140432	140432	140432	140432
R-sq	0.031	0.040	0.063	0.066	0.067

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

此表复现了 Cao and Chen (2022) Table 3 的结果。

无论使用何种模型设定，漕粮海运政策冲击对叛乱发生均具有显著的正效应。

Cao and Chen (2022)还进行了一系列的稳健性检验，包括使用其他计量方法、改变被解释变量的度量、考察政策冲击的力度(县内运河长度、运河十公里内城镇数目等)、处理效应的异质性(运河北段与南段的不同效应)，以及运河县对非运河县的溢出效应等，在此从略。