面板数据与随机效应

张雪菲

2019.10.31

面板数据

- 面板数据的每个样本观测值至少有截面i和时间t两个维度：如地区-年份，企业-年份
  - 或者更高的维度：(i,j,t)，如地区-银行-年份
- 截面单位个数：$i = 1, \ldots, N$；
  时间单位个数：$t = 1, \ldots, T$
- 面板数据分类：大N小T，小N大T，大N大T
  - 大部分研究中碰到的情况都是大N小T：中国地市面板、中国银行面板；跨国面板

面板数据分类

- 平衡面板：

  每个截面单位的样本时间长度均为T，样本总数为$N \times T$

- 非平衡面板：

  截面i的样本时间长度为$T_i \leq T$，样本总数为$\sum_{i=1}^{N} T_i$

- 大部分面板回归的方法对平衡面板和非平衡面板均适用

面板回归模型

- 给定一组面板数据$\{y_{it}, x_{it}\}$，y_{it}是被解释变量，x_{it}是解释变量
- 研究中最常用的面板模型：

$$y_{it} = x_{it}'\beta + v_{it}$$

β：回归系数

v_{it}：均值为0的残差项

个体效应与时间效应

- 残差项v_{it}一般可以分解为三部分：

$$v_{it} = u_i + \delta_t + \varepsilon_{it}$$

  - u_i：个体效应
  - δ_t：时间效应
  - ε_{it}：假设为在i和t两个维度均相互独立且与其他变量(x_{it}, u_i, δ_t)相独立

- 引入个体效应和时间效应，可以控制无法观测到的个体特征和时间变化对y_{it}的影响

面板回归模型与OLS估计

- 通常情况下，OLS方法是估计回归系数β最简单的方法
- OLS估计值$\hat{\beta}_{OLS}$具有一致性的前提假设：x_{it}和v_{it}不具有相关性
  - β的值本质上由理论确定
  - 由于样本是随机的，$\hat{\beta}_{OLS}$也是一个随机变量
  - $\hat{\beta}_{OLS}$具有一致性：当样本足够大时，其概率极限逼近理论值β，

$$\operatorname*{plim}_{N\to\infty}\hat{\beta}_{OLS}=\beta$$

单一解释变量的例子

- 考虑单一解释变量x_{it}的情形，此时有

$$\hat{\beta}_{OLS}=\frac{\sum_{t=1}^{N}\sum_{t=1}^{T}x_{it}y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^{2}}=\frac{\sum_{t=1}^{N}\sum_{t=1}^{T}x_{it}(x_{it}\beta+v_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^{2}}$$

$$=\beta+\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}v_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^{2}}=\beta+\frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}v_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^{2}}$$

- 通常假定$\text{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^{2}$是一个常数，则$\hat{\beta}_{OLS}$的一致性取决于$\text{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}v_{it}=\text{cov}(x_{it},v_{it})=0$是否成立：

关键$\text{cov}(x_{it},u_i)=0$及$\text{cov}(x_{it},\delta_t)=0$

OLS估计下引入虚拟变量

- 绝大多数面板数据都是大N小T型，在此情况下，时间效应不会带来任何问题，原因：
  - 此情形下一般假设T固定，$N \to \infty$，因此可以设置固定数量的时间虚拟变量D_t，直接估计出δ_t，避免其与x_{it}之间的相关性干扰OLS估计
- 对于个体效应，无法使用虚拟变量来解决问题：D_i的数目随N的增大而增大

个体效应的问题

- 通常情况下无法保证$cov(x_{it}, u_i) = 0$，因此OLS估计很可能出现不一致
- 此外，即便有$cov(x_{it}, u_i) = 0$，$\hat{\beta}_{OLS}$的标准差——即$\hat{\beta}_{OLS}$作为随机变量的标准差——也需要考虑到个体效应u_i的影响

随机效应

- 只考虑个体效应，则回归模型残差项为

$$v_{it} = u_i + \varepsilon_{it}$$

- 若$cov(x_{it}, u_i) = 0$，则u_i称为随机效应
  - 在此情况下，面板回归模型的OLS估计$\hat{\beta}_{OLS}$具有一致性
- 若$cov(x_{it}, u_i) \neq 0$，称为个体固定效应
  - 在此情况下，面板回归模型的OLS估计$\hat{\beta}_{OLS}$不具有一致性
  - 原因在于：

$$\text{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it}$$

$$= \text{plim}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(u_i + \varepsilon_{it}) \neq 0$$

随机效应模型OLS估计的参数推断

- 单一解释变量，此时$\hat{\beta}_{OLS}$的渐进方差可以写作：

$$AV\left[\hat{\beta}_{OLS}-\beta\right]$$

$$=\operatorname*{plim}_{N\to\infty}\frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}v_{it}\right)^2\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2\right)^{-2}$$

$$=\operatorname*{plim}_{N\to\infty}\frac{1}{N^2}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}x_{it}^2v_{it}^2+2\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}x_{it}x_{is}v_{it}v_{is}\right)\times\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2\right)^{-2}$$

- 残差v_{it}具有序列相关性：$cov(v_{it},v_{is})\neq 0$

对标准误的说明

- 此处没有任何附加的同方差假设；允许异方差，$E[v_{it}] \neq E[v_{jt}]$
  - 允许异方差的标准误估计成为异方差稳健标准误，或White稳健标准误
- 用到的假设：v_{it}，实质上是u_i，在截面上相互独立
  - 上页第二个等号用到这个假设

聚类标准误

- 这一类的标准误，又称为聚类标准误，或截面个体层面聚类标准误
  - 也可以不在截面个体层面上进行聚类，比如地区-银行-年份面板，可以再地区-银行层面聚类（相当于上面的截面个体具体），也可以在地区层面聚类（此时允许同一地区不同银行残差项相关）
- 聚类标准误具有很好的稳健性；用此标准误进行的统计推断最为准确

  这类标准误又称为面板稳健标准误

个体固定效应

张雪菲

2019.10.31

固定效应估计法：一个例子

- 引入“固定效应”的一个例子：工会成员的身份与工资之间的关系
- $E[Y_{0it}|A_i, X_{it}, t, D_{it}] = E[Y_{0it}|A_i, X_{it}, t]$

  A_i：不可观察但是固定的干扰因素构成的向量
- **固定效应估计法**的关键假设：
  - A_i没有时间因素：$E[Y_{0it}|A_{i,}X_{it}, t] = \alpha + \lambda_t + A_i'\gamma + X_{it}\beta$
  - 因果效应不变且可加：$E[Y_{1it}|A_{i,}X_{it}, t] = E[Y_{0it}|A_{i,}X_{it}, t] + \rho$
- $E[Y_{it}|A_i, X_{it}, t, D_{it}] = \alpha + \lambda_t + \rho D_{it} + A_i'\gamma + X_{it}'\beta$

  ρ：因果效应

固定效应模型

- 固定效应模型：

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}$$

$$\alpha_i = \alpha + A_i'\gamma$$

$$\varepsilon_{it} = Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$$

- 需要估计的固定因素（虚拟变量）：

  α_i：个体效应，λ_t：年份效应

- 给定面板数据，不可观察的个体效应和年份效应就是相应虚拟变量前的系数。

随机效应模型

- 假设α_i与回归元不相关：

  因为随机效应模型中**被遗漏的变量与回归元不相关**，所以这些变量的遗漏不会导致估计值的偏误

- 遗漏变量成为残差的一部分：对于给定个人的各期残差是相关的

偏离均值法

- 将个体效应视为待估参数，等同于估计个体对均值的偏离程度
- 估计个体均值：

$$\bar{Y}_i = \alpha_i + \bar{\lambda} + \rho\bar{D}_i + \bar{X}_i'\beta + \bar{\varepsilon}_i$$

- 减去相应均值：

$$Y_{it} - \bar{Y}_i = \lambda_t - \bar{\lambda} + \rho(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- 不可观察的个体效应 α_i 消失了
- 估计值：内估计量（within estimator）、协方差分析（analysis of covariance）
- 进行估计的过程：吸收固定效应

差分模型

- 差分模型：

$$\Delta Y_{it} = \Delta\lambda_t + \rho\Delta D_{it} + \Delta X_{it}'\beta + \Delta\varepsilon_{it}$$

- 差分模型的残差是序列相关的，计算标准误时需要进行调整
- 两种方法的选择：
  - 在同方差及ε_{it}不存在序列相关且考虑的时期大于两期时，偏离均值法更有效
  - 必须手动计算时，差分模型更有效

固定效应模型估计值与截面估计值的比较

- 假设人们基于不可观测但是固定的个体特征选择是否加入工会，研究工会身份对工资的影响

表 5.1　估计出的工会身份对工资的影响

调　　查	截面估计值	固定效应估计值
May CPS，1974—1975	0.19	0.09
National Longitudinal Survey	0.28	0.19
Michigan PSID，1970—1979	0.23	0.14
QES，1973—1977	0.14	0.16

固定效应模型的度量偏差

- 观察到固定效应模型的估计值低于截面估计值
- 可能是由于截面模型中存在正的选择偏误
- 可能是由于固定效应模型中存在着度量偏差：
  - 诸如工会身份这样的经济变量倾向于长期稳定，而度量误差每年变化着
  - 虽然消除了一些遗漏变量偏误，同时也丢掉了我们感兴趣变量的很多信息

    例子：控制家庭固定效应，用双胞胎估计教育水平对工资的影响
- 处理方法：工具变量、外部信息

双重差分

张雪菲

2019.10.31

双重差分（difference-in-difference）

- 研究政策问题：使用群体层面的固定效应来解决地域或年份层面上出现的遗漏变量偏误
- 最早提出：物理学家John Snow（1855）
- 19世纪中期伦敦市的霍乱传染问题是由受污染的水传染而来的
- 相比由水厂Southwark & Vauxhal供水的地区，由迁往上游的水厂Lambeth供水地区的霍乱死亡率剧降

双重差分：一个例子

- 最低工资对就业的影响作用：新泽西州（处理组）、宾夕法尼亚州（控制组）
- 核心假设：没有收到处理的那个州潜在结果可以写成两部分相加的形式

$$E(Y_{0ist}|s,t) = \gamma_s + \lambda_t$$

- γ_s：不随时间变化的州效应（不可观察的个体效应）

  λ_t：对两个州都相同的年份效应

估计因果效应

- 假设$E(Y_{1ist} - Y_{0ist}|s,t) = \delta$，那么

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}$$

由于$E(\varepsilon_{ist}|s,t) = 0$，可得

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)$$
$$= \lambda_{Nov} - \lambda_{Feb}$$
$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)$$
$$= \lambda_{Nov} - \lambda_{Feb} + \delta$$

- 进行双重差分，得到因果效应：

$$\{E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)\}$$
$$-\{E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)\}$$
$$= \boldsymbol{\delta}$$

估计因果效应

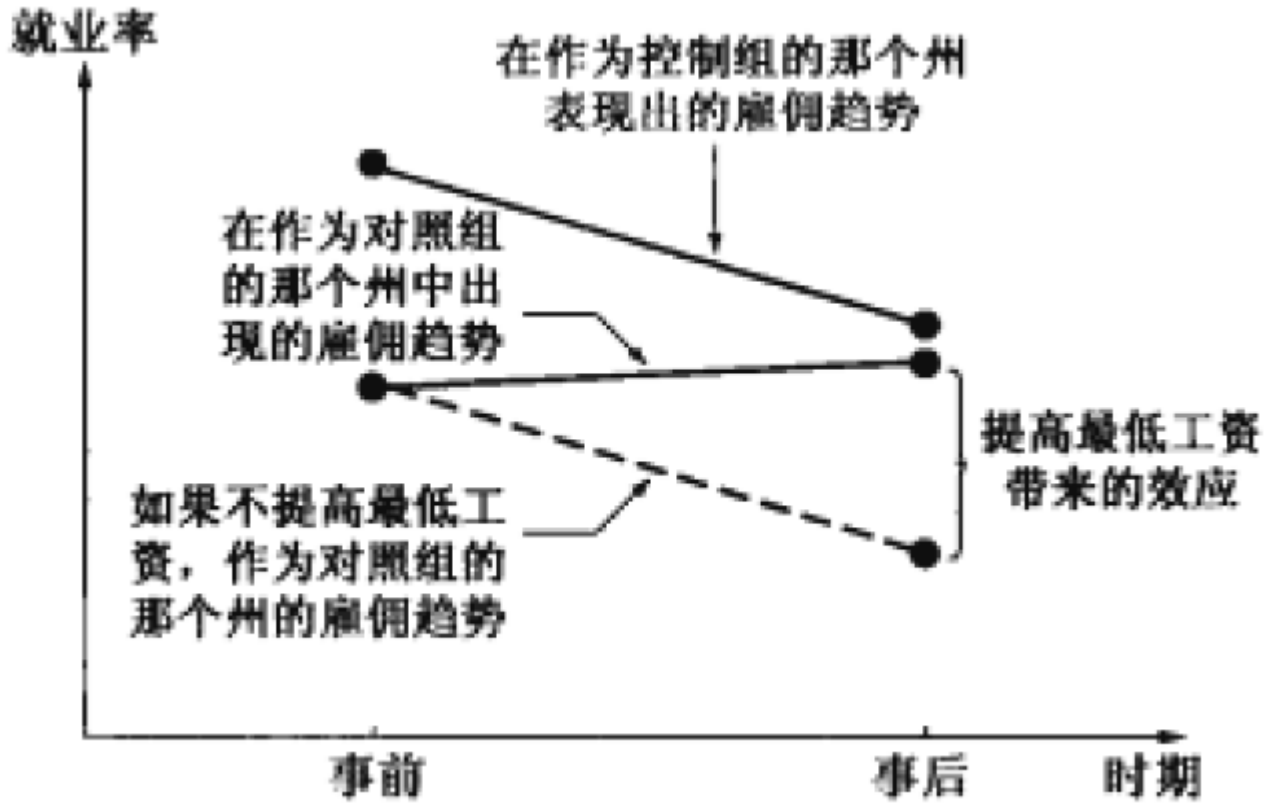
- 正的双重差分估计值：与我们所期待的情况相反

表 5.2　在新泽西州最低工资上升前后快餐店的平均雇员数

变　　量	宾夕法尼亚州(i)	新泽西州(ii)	差分，NJ－PA(iii)
1. 最低工资上升前的全职雇员，使用所有可用的观察值	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. 最低工资上升后的全职雇员，使用所有可用的观察值	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. 全职雇员平均数量的变化	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

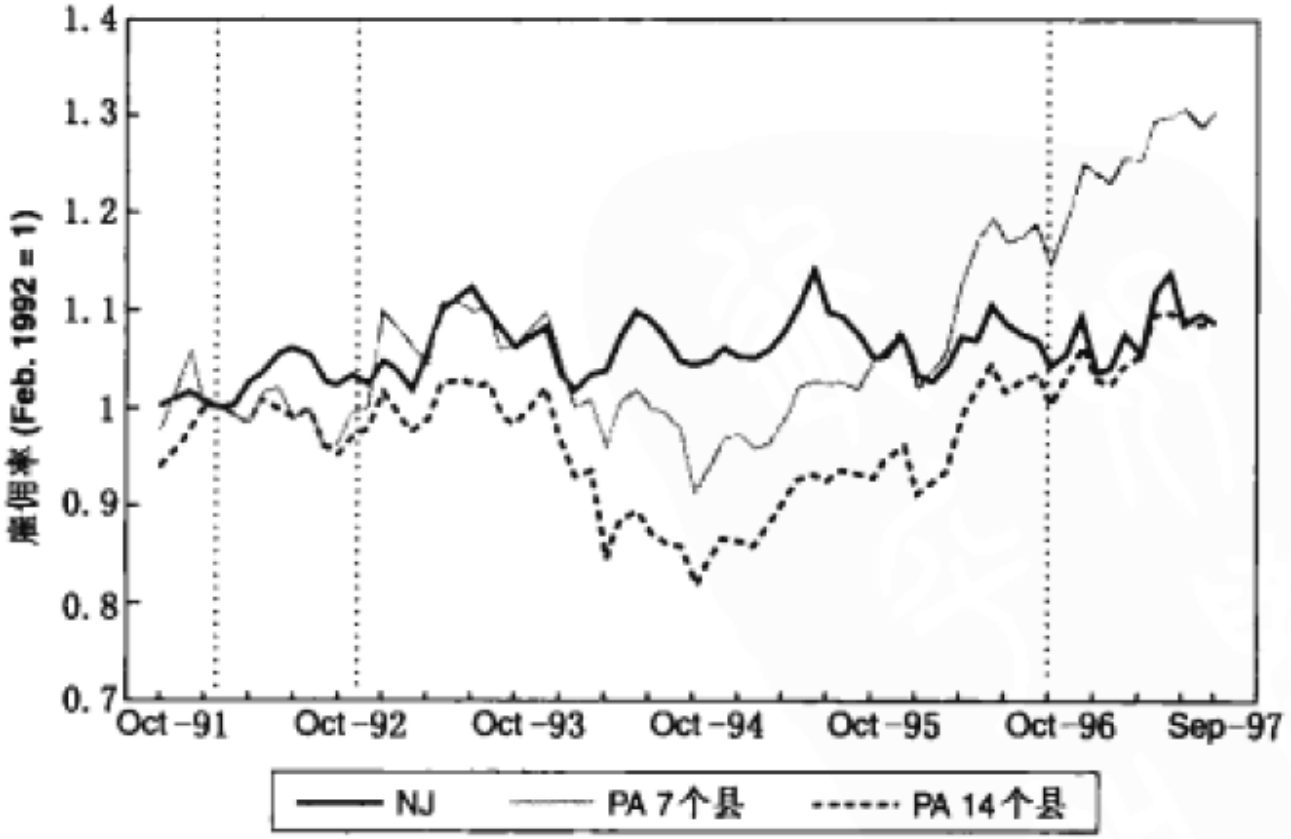
估计因果效应

- 关键性假设：如果新泽西州没有收到处理（没有提高最低工资）那么这两个的就业趋势应该是相同的



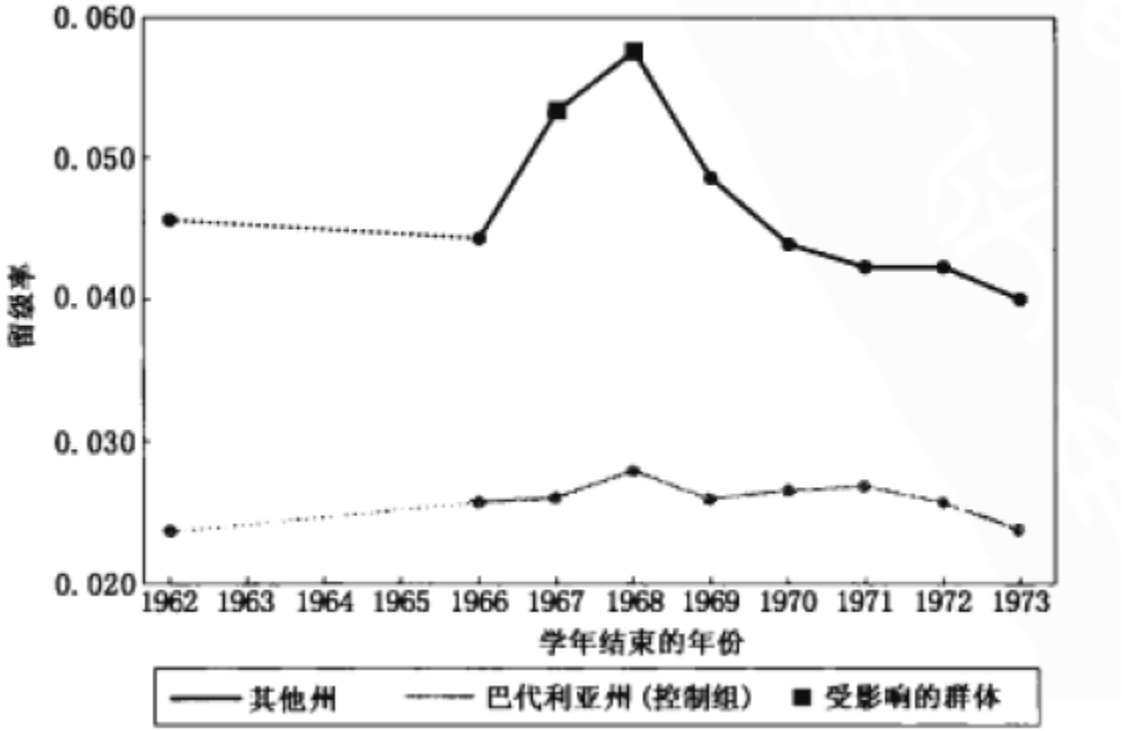
检验共同趋势假设

- 宾夕法尼亚快餐店就业水平不是度量反事实情况下新泽西州快餐店就业水平的良好指标



一个例子

- 德国各州（除巴伐利亚州）由春季开学变为秋季开学：考察学校学期长度对学生成绩的影响
- 一个处理引起处理组剧烈短暂的反应，偏离潜在趋势



双重差分回归

- NJ_s：处在新泽西州的餐馆（表示地域的虚拟变量）

  d_t：表示时间的虚拟变量，时间哑变量（time-dummy）

- $$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \cdot d_t) + \varepsilon_{ist}$$

  $$NJ_s \cdot d_t = D_{st}$$

- 饱和模型

待估参数

- $Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \cdot d_t) + \varepsilon_{ist}$
  - $\alpha = E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$
  - $\gamma = E(Y_{ist}|s = NJ, t = Feb) - E(Y_{ist}|s = PA, t = Feb) = \gamma_{NJ} - \gamma_{PA}$
  - $\lambda = E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$
  - $\boldsymbol{\delta} = \{E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)\} - \{E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)\}$

一个例子

- 最低工资产生的影响中存在的地区差异

$$Y_{ist} = \gamma_s + \lambda_t + \delta(FA_s \cdot d_t) + \varepsilon_{ist}$$

- 一阶差分方程：$\Delta\bar{Y}_s = \lambda^* + \delta FA_s + \Delta\bar{\varepsilon}_s$

表 5.3　最低工资对年轻人影响的双重差分回归估计值，1989—1990 年

解释变量	工作对数值的平均值的变化		年轻雇员在总人口中的比例	
	(1)	(2)	(3)	(4)
1. 受影响的青少年比例(FA_s)	0.15 (0.03)	0.14 (0.04)	0.02 (0.03)	−0.01 (0.03)
2. 就业人数占总人口比重的变化	—	0.46 (0.60)	—	1.24 (0.60)
3. R^2	0.30	0.31	0.01	0.09

双重差分回归的好处

- 即使政策变化无法使用虚拟变量来描述，我们也能对政策变化进行研究
- 很容易添加更多的表示地点和时期的虚拟变量以及协变量

$$E(Y_{0ist}|s,t,X_{st}) = \gamma_s + \lambda_t + X'_{st}\beta$$

X_{st}：随着州与时间变化而变化的协变量组成的向量，包括成年人就业量

- 注意：分析的是州层面的平均值而不是个体值

$$Y_{ist} = \gamma_s + \lambda_t + \delta(FA_s \cdot d_t) + X'_{ist}\beta + \varepsilon_{ist}$$

X_{ist}可以包括种族等个体层面上的特性，也可以是随时间变化而变化的州层面上的变量

包含多年观测值时的因果检验

- 观察原因是否发生在结果之前，而不是相反：

  D_{st}引起了Y_{ist}的变化，而不是Y_{ist}引起了D_{st}的变化

- $Y_{ist} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau} D_{s,t-\tau} + \sum_{\tau=1}^{q} \delta_{+\tau} D_{s,t+\tau} + X'_{ist}\beta + \varepsilon_{ist}$

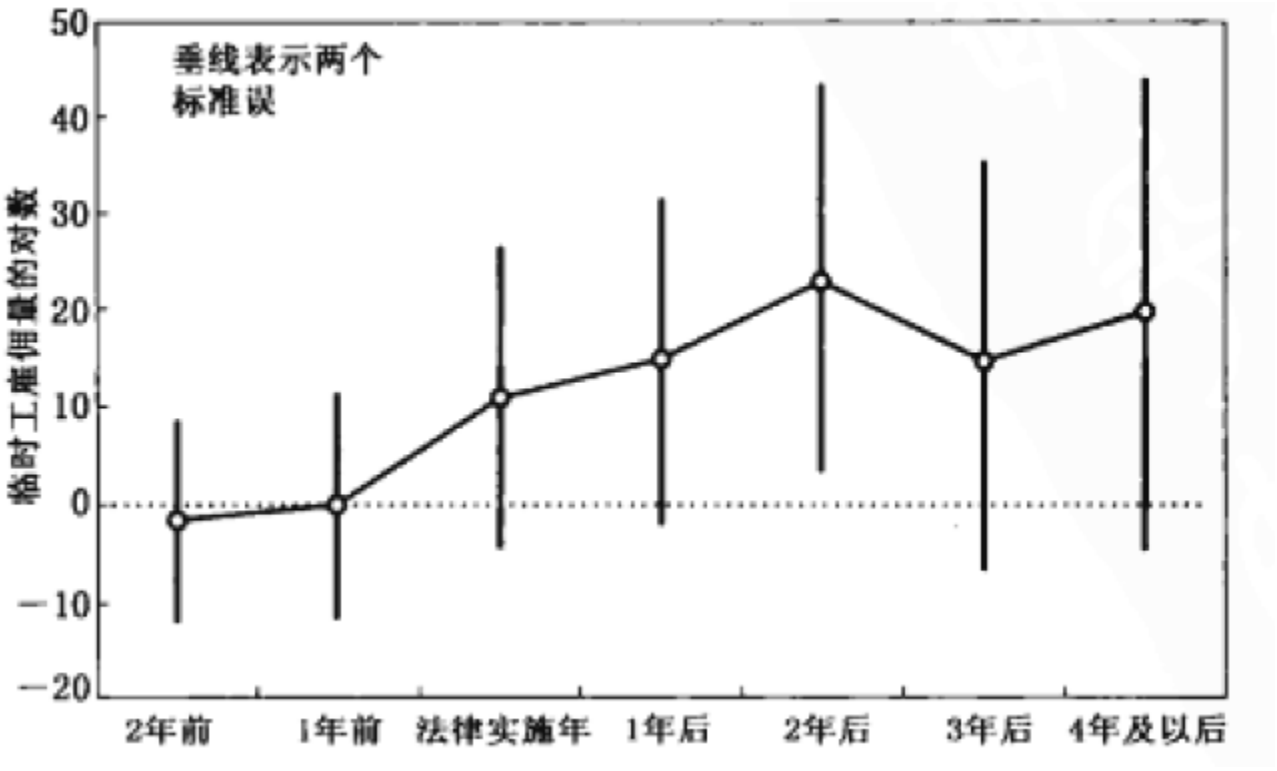
- m阶滞后效应（$\delta_{-1}, \delta_{-2}, \cdots, \delta_{-m}$）

  q阶提前效应（$\delta_{+1}, \delta_{+2}, \cdots, \delta_{+q}$）

- 滞后效应：随着时间的推移，因果效应会减弱或者加强

因果检验：一个例子

- 考察雇佣保护措施如何对企业使用临时工造成影响
- 两期提前和四期滞后



稳健性检验：加入时间趋势项

- 在控制变量中加入与每个州相联系的时间趋势项：

$$Y_{ist} = \gamma_{0s} + \gamma_{1s}t + \lambda_t + \delta D_{st} + X'_{ist}\beta + \varepsilon_{ist}$$

- γ_{1s}：对每个州求出的时间趋势系数
- 好处：
  - 允许处在处理组的州和处在控制组的州沿着不同的趋势发展
  - 更加稳健
- 局限：
  - 至少需要三期数据来估计时间趋势系数

稳健性检验：一个例子

- 研究印第安纳州劳动管制对企业绩效的影响

表 5.4　在印度的各个州中估计出的劳动力管制对企业绩效的影响

	(1)	(2)	(3)	(4)
劳动力管制(滞后)	−0.186	−0.185	−0.104	0.000 2
	(0.064)	(0.051)	(0.039)	(0.020)
log(个人发展上的人均花费)		0.240	0.184	0.241
		(0.128)	(0.119)	(0.106)
log(人均已建成电力设备数)		0.089	0.082	0.023
		(0.061)	(0.054)	(0.033)
log(州人口)		0.720	0.310	−1.419
		(0.96)	(1.192)	(2.326)
国会中的多数席位			−0.000 9	0.020
			(0.01)	(0.010)
极左势力在国会中的席位			−0.050	−0.007
			(0.017)	(0.009)
Janata 政党在国会中的席位			0.008	−0.020
			(0.026)	(0.033)
地区代表在国会中的席位			0.006	0.026
			(0.009)	(0.023)
州趋势项	无	无	无	无
R^2	0.93	0.93	0.94	0.95

挑选控制变量

- 最具代表性的两个维度：“州”、“时间”
- “州”：任何人口统计意义上的组

  其中一些组受到政策影响，另外一些不受影响
- “时间”：出生年份或者不同个体特征

如：研究州堕胎发的变化对青少年的影响，使用了州和出生年份变量

双重差分法的缺点

- 缺点：随着处理结果的不同，控制组和比较组的组成人员可能发生着变化
- 如：研究不同程度公共救助对劳动力供给的影响

但是无论怎样都会成为就业市场上的弱势群体的穷人可能会移居到福利更为优厚的州

- 可以使用工具变量来解决：用出生地所在州或者之前居住地来构造当前居住地的工具变量

更高阶的差分模型：一个例子

- 研究全美医疗补助的覆盖范围扩大后，对母亲的劳动参与状态和收入的影响
- 三重差分模型：

$$Y_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \delta D_{ast} + X'_{iast}\beta + \varepsilon_{iast}$$

- a：该家庭中最小孩子的年龄

D_{ast}：在医疗救助覆盖的相应州和相应时期，且孩子处在接受医疗补助的年龄段的家庭