

# 第八讲 实验和自然实验

许文立\*<sup>1,2</sup>

<sup>1</sup>安徽大学经济学院

<sup>2</sup>安徽生态与经济发展研究中心

January 20, 2018

在许多领域，实验是最常用的因果效应估计方法，例如心理学和医药学。研发了一种新药，投放市场前，必须要经过临床实验检验其效力。而这种临床试验就是随机的选择一些病人来服用这种新药，而另一些病人服用无害的无效替代品（也称为“安慰剂”，这就是为什么我们经常在论文中看到“安慰剂检验”）。只有这种随机控制实验表明新药是安全有效的（可信的统计证据），新药才会投放市场。

从目前来看，随机控制实验也是经济学中最重要的关注点，现在很多学者都“绞尽脑汁”去寻找各种政策的“随机控制实验（自然实验）”。如果你找到了，恭喜你成功了一大半了！

当你幸运地发现了一个别人还没有做过的政策”随机控制实验“的时候，接下来你可以就要思考，用什么方法把政策的效应给估计出来。那么，本讲的内容对你可能是最实用的。我们在进行项目评估，也就是估计一个项目、政策、干预或”处理“的效应时，最常用的估计方法有：**双重差分（DID）、断点回归设计（RDD）、倾向性匹配得分（PSM）和合成控制法**。而其中前两

---

\*E-mail: xuweny87@163.com。非常欢迎大家给我们提出有益意见和建议。个人和机构可以利用本讲稿进行教学活动，但请不要用于商业目的。版权和最终解释权归许文立所有。当然，文责自负。

种是最流行、最常用的方法。

## 1 理想实验

假设我们正在进行医药试验，随机选择一部分病患服用试验药品，被称为**处理组**；另一部分病患服用无害无效的安慰剂，称为**控制组**。这样，我们就可以得到两种结果。这两种**潜在结果**的差异就是试验药品对病患的（因果）效应。

在理想实验中，处理组与控制组的病患选择完全是随机的。因此，病患个体究竟被分到哪一组或是否服用试验药品，与个体的特征或其他可能影响潜在结果的因素是完全独立的。因此，解释变量（是否服用试验药品）与扰动项不相关，即 $cov(X_i, u_i) = 0$ 。这样，无论是否有遗漏变量，都不会出现遗漏变量偏误。这就是理想实验的最大优点。

### I. 差分估计量

假设 $X_i = 1$ 表示服用新药， $X_i = 0$ 表示未服用新药， $Y_i$ 表示病人的病情。那么，服用新药后的潜在结果为 $\bar{Y}_{i,treat} = E(Y_i | X_i = 1)$ ，而服用安慰剂的潜在病情为 $\bar{Y}_{i,control} = E(Y_i | X_i = 0)$ 。那么，**平均处理效应（ATE）**就是两种潜在结果之差，即 $E(Y_i | X_i = 1) - E(Y_i | X_i = 0)$ 。如果我们写出回归方程

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad (1)$$

那么，（1）式的OLS估计量为

$$\tilde{\beta}_1^{OLS} = \bar{Y}_{i,treat} - \bar{Y}_{i,control} \quad (2)$$

由此，可以看出， $\beta_1$ 的OLS估计量是处理组样本均值与控制组样本均值之差，故也成为“差分估计量”。

注意：虽然，在理想实验中，服用新药的病人是随机分配的，无论是否存在遗漏变量，服用新药（处理）对病情影响的估计量总是无偏、一致的。但是，在实践中，我们通常还是会加入如控制变量W，因为在回归中加入控制变量W可以提高 $\beta_1$ 的OLS估计量的效率（Stock and Watson, 2015）。如果控制变量W对解释Y的变动有帮助，那么，回归中加入控制变量W可以降低 $\beta_1$ 的OLS估计量的标准误。

## II. 小班教学实验

20世纪80年代初期，美国田纳西州进行了一项围棋4年的初级教育阶段班级缩减的实验，STAR项目。这项实验主要是想评价小班教学的效果。这次小班教学实验划分了三种班级规模：普通规模——22-25名学生/班/教师，无助教；小班——13-17名学生/班/教师，无助教；普通规模，有助教。学生和老师在入学时都是**随机分配**到不同班级规模。这项实验的第一年，大约有6400名学生，108个小班，101个无助教普通班，99个有助教普通班。4年期间，共有11600名学生和80所学校参与小班教学实验。

这个实验与上面论述的有点差异：两个**处理组**——小班（ $SC_i$ ）和有助教普通班（ $RA_i$ ），一个**控制组**——无助教普通班。因此，小班教学实验的基准回归模型为

$$Y_i = \beta_0 + \beta_1 SC_i + \beta_2 RA_i + W_i \gamma_i + u_i \quad (3)$$

其中， $Y_i$ 表示考试成绩；如果第i个学生在小班，那么 $SC_i = 1$ ，否则等于0；如果第i个学生在有助教普通班，那么 $RA_i = 1$ ，否则等于0； $W_i$ 表示控制变量向量。根据前面对平均处理效应的论述，相对于无助教普通

班来说，小班的效应就是 $\beta_1$ ，有助教普通班的效应是 $\beta_2$ 。回归方程（3）

中， $\beta_1$ ， $\beta_2$ 的OLS估计量可以通过差分估计量计算得到。如表1所示。

Table 1: 小班教学实验的效应：无控制变量

解释变量	(1)	(2)	(3)	(4)
小班	13.90** (2.45)	29.78** (2.83)	19.39** (2.71)	19.59** (2.40)
有助教普通班	0.31 (2.27)	11.96** (2.65)	3.48 (2.54)	-0.29 (2.27)
常数项	918.04** (1.63)	1039.39** (1.78)	1157.81** (1.82)	1228.51** (1.68)
Obs.	5786	6379	6049	5967

注：(1)1-4列分别为全样本，以及1-3年级样本；  
 (2)括号中为标准误；  
 (3)\*\*\*、\*\*、\*表示1%、5%、10%的显著性水平。

下面，我们加入一些控制变量。回归结果如表2所示。

[插入表2]

表1和表2中估计的教学实验效果应该如何理解呢？

有两种方式回答这个问题：

第一，将每一行的考试分数估计值转换成单位考试得分标准差的效应，这就使得每个年级的估计效应可以进行比较。例如，学生考试成绩的全样本标准差为73.7，从表1中可以知道，小班教学的效应估计值是13.9，那么，小班教学的单位得分标准差效应为 $13.9/73.7=0.19$ ，标准误为 $2.45/73.7=0.03$ 。

第二，将班级规模的效应系数与另外一些估计系数进行比较。

### III. 准实验或自然实验

这项实验花费了田纳西州1200万美元。因此，实验经济学成本实在太高，对于我们普通人来说几乎不可能实施。而理想实验的最大优势在于处理组和控制组的随机分配。因此，我们要紧紧抓住这个随机性。

Table 2: 小班教学实验的效应：控制变量

解释变量	(1)	(2)	(3)	(4)
小班	13.90** (2.45)	14.00** (2.45)	15.93** (2.24)	15.89** (2.16)
有助教普通班	0.31 (2.27)	-0.60 (2.25)	1.22 (2.04)	1.79 (1.96)
教师执教年数		1.47** (0.17)	0.74** (0.17)	0.66** (0.17)
男孩				-12.09** (1.67)
免费午餐				-34.70** (1.99)
黑人				-25.43** (3.50)
其他种族				-8.50 (12.52)
常数项	918.04** (1.63)	904.72** (2.22)		
学校指标	no	no	yes	yes
$\bar{R}^2$	0.01	0.02	0.22	0.28
Obs.	5786	5766	5766	5748

注：(1)1-4列分别为全样本，以及1-3年级样本；  
(2)括号中为标准误；  
(3)\*\*\*、\*\*、\*表示1%、5%、10%的显著性水平。

在实践中，我们通常见到的是一种非实验情形，但它又具有某些随机性。我们把这种情形称为**准实验或自然实验**。在这些自然实验情形下，对个体的处理通常是“似乎”是随机分配的。在我们国家，最常见的自然实验就是各种改革措施或政策的试点，例如“营改增”、“省直管县”、“开发区”等等。

两种类型的自然实验：

(1) 个体是否受到“处理”，似乎是随机决定的。这种情形就可以利用全面的“差分估计量”来计算处理效应；

(2) 随机变动“似乎”只是部分决定了是否被处理。这个时候，因果效应就可以利用IV回归。回忆一下第七讲中构造IV的方法就是分离出与误差项无关的成分，而这种自然实验中的随机变动部分就提供了工具变量。

下面的问题就是，我们如何估计自然实验中的因果效应。

## 2 双重差分(DID)

在自然实验中，个体接受处理与否似乎是随机分配的，但是我们并不能控制这种随机性，即使我们控制了那些影响随机性的变量 $W$ ，处理组和控制组之间的某些效应仍然不能估计出来。

有一种方法可以消除上述问题：我们不去比较处理组和控制组之间产出水平 $Y$ 的差异，而是去比较两组产出 $Y$ 变化的差异，和处理前后 $Y$ 变化的差异。这个估计量是处理组和控制组之间效应变化的差异，或者时间上的差异，因此，这就是我们熟知的**双重差分（DID）估计量**。

### 2.1 DID

下面，我们来看看上海对外经贸大学的司继春博士在“知乎”上举的一个例子：<https://www.zhihu.com/question/24322044>

现在要修一条铁路，铁路是条线，所以必然会有穿过的城市 and 没有被穿过的城市。记 $D_i = 1$ ，如果铁路穿过城市 $i$ ； $D_i = 0$ ，如果城市 $i$ 没有被穿过。现在我们感兴趣的问题是：铁路修好以后，被铁路穿过的城市是不是经济增长更快了？我们该怎么做呢？一开始的想法是，我们把 $D_i = 1$ 的城市的GDP加总，减去 $D_i = 0$ 的城市的GDP加总，然后两者一减，即 $E(Y_i | D_i = 1) - E(Y_i | D_i = 0)$ ，这样我们就算出了两类城市GDP的平均之差。如果大家还记得，这就是我们前面所讲的理想实验中的差分估计量。

那么，这样做是不是就得到了我们感兴趣的铁路效应呢？不用说这样肯定有问题。如果没有问题，那我们讲完第一节就可以打铃下课，我收工回家了。

我们想想，万一被铁路穿过的城市在建铁路之前GDP就高呢？为了解决这个问题，我们需要观察到至少两期，第一期是建铁路之前，第二期是建铁路之后。我们先把两类城市的GDP做铁路修建前后两期之差，即：

$$\Delta Y_i = \frac{1}{N} \sum (Y_{i,after} - Y_{i,before}) \quad (4)$$

(4) 式就是第一次差分。它计算的实际上是城市*i*建铁路前后平均GDP的增长（如果是GDP取对数，就是增长率）。接下来，我们再来计算GDP变化的平均处理效应，也就是

$$ATE = E(\Delta Y_i | D_i = 1) - E(\Delta Y_i | D_i = 0) \quad (5)$$

这是第二次差分。这一步就把两类城市在修建铁路之前和之后的GDP增长的差异给算出来了，这就是我们要的处理效应，即修建铁路之后对城市经济的促进作用。

还可以将DID估计量换一个写法。记T=1，如果时间为建铁路之后；T=0，如果时间为建铁路之前。然后，结合上面城市修建铁路与否的虚拟变量，我们可以得到下面的表3：

Table 3: DID估计量

Treated	D=1	D=0
T=1	1	0
T=0	0	0

Treated表示在某一时期，某城市是否修建了铁路。从表3可以看出，在T=0时期，没有城市修建铁路，而在T=1期，也只有D=1的城市修建了铁

路。因此， $Treated = D_i \times T$ 。我们可以写出下列回归方程：

$$Y_{it} = \alpha D_i + \beta T + \gamma D_i \times T + u_{it} \quad (6)$$

其中， $Y_{it}$ 表示城市*i*在第*t*期的GDP。我们感兴趣的是系数 $\gamma$ 。

首先，我们将（5）式在时间维度上做一次差分：

$$\Delta Y_i = \beta + \gamma D_i + \Delta u_i \quad (7)$$

然后，再对（6）式在个体层面做一次差分，并取期望：

$$E(\Delta Y_1 - \Delta Y_0) = \gamma \quad (8)$$

到此，我们得到了建铁路的经济增长效应DID估计量 $\tilde{\gamma}$ 。

## 2.2 平行趋势假设

下面，我们用图1来看看。

由图1可以清晰地看出，DID最关键的假设是common trend，也就是两个组别在不处理的情况下，Y的趋势是一样的。那么你仍会说，铁路穿过的城市可能本身GDP也高，而GDP 高的城市按照理论GDP 增长率可能更高可能更低，所以common trend的假设可能是不对的，那怎么办？如果这个问题存在，我们可以进一步假设在控制了某些外生变量之后，common trend 是对的，比如上个问题，我们可以控制城市在*t*=0期的GDP level。当我们控制其他变量之后，自然不能直接减两次了，我们需要用上面说的回归式子，即对下列回归方



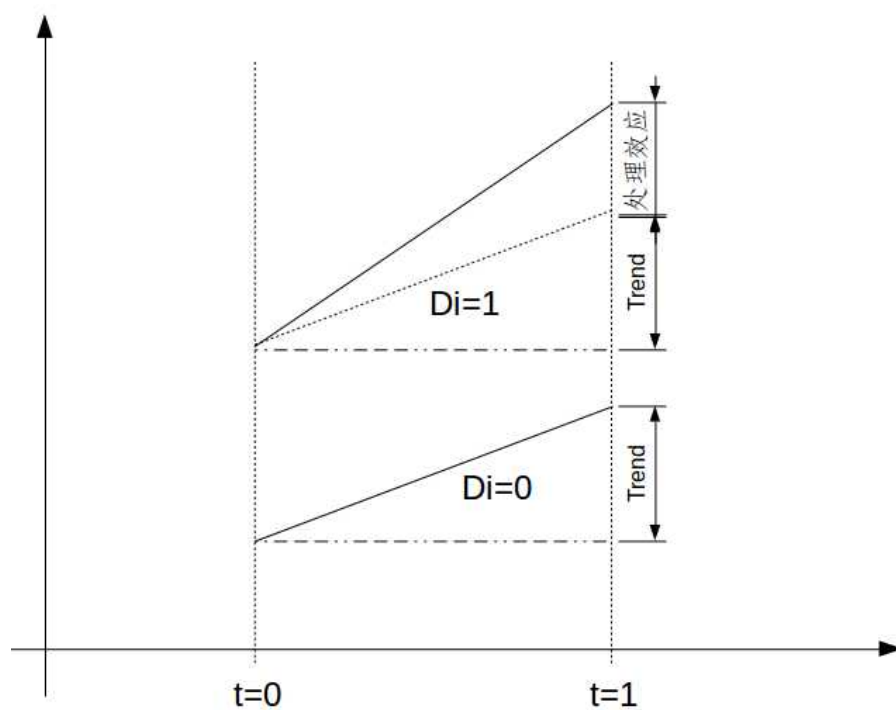


Figure 1: DID估计量-平行趋势

程run OLS:

$$Y_{it} = \alpha D_i + \beta T + \gamma D_i \times T + X' \delta + u_{it} \quad (9)$$

其中，X是控制变量向量。

既然common trend是DID最关键的假设，那么，我们如何检验和处理非平行趋势呢？下面我们来看看陈强老师于2016-10-25在微信公众号“计量经济学及stata应用”上给出的讲解：

### 方法一、画时间趋势图

如果在政策干预前有多期数据，则可分别画处理组与控制组的时间趋势图（类似于上图），并直观判断这两组的时间趋势是否平行（比如，考察是否存在Ashenfelter's dip）。如果二者大致平行，则可增强对平行趋势假定的信心。然而，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（后者本质上不可观测，因为时间效应已与处理效应混合在一起）。另外，如果只有两期数据，则无法使用此法。

### 方法二、加入更多的控制变量

从上文的讨论可知，非平行趋势可能由于遗漏变量所导致，故在回归方程中加入更多控制变量，或可缓解内生性。但此法在实践中不易实施。

### 方法三、假设线性时间趋势

如果假设时间趋势为线性函数，则可加入每位个体的时间趋势项：

在具体回归时，加入个体虚拟变量与时间趋势项 $t = 1, 2, \dots, T$ 的交互项即可。然而，线性时间趋势毕竟是比较强的假定，不一定能成立。故此法也不完全解决问题，但可作为稳健性检验。

### 方法四、三重差分法

在一定条件下，可通过引入两个控制组，进行三次差分，称为“三重差分法”（difference-in-differences-in-differences，简记DDD），这样可以更好地控制时间趋势的差异，使得平行趋势假定更易成立。有关DDD的进一步介绍，参见陈强（2014，第343页）。

## 2.3 DID在Stata中的实现

要估计自然实验中的平均处理效应，如果直接在stata中run（9）式，那么，直接使用普通的面板数据命令`xtreg`即可。而DID则有专门的命令估计。厦门大学赵西亮老师的书里介绍了一种DID的命令，`diff`，其语法和基本选项为：

```
diff outcome-var [if] [in] [weight],period(varname)
    treated(varname) [cov(varlist)
    kernel id(varname) bw(#) ktype(kernel) rcs
    qdid(quantile) pscore(varname) logit
    support addcov(varlist) cluster(varname)
    robust bs reps(int) test report noustar export(filename)]
```

`outcome-var`是结果变量，`period(varname)`告诉软件时期变量，`treated(varname)`告诉软件处理变量。其他命令（也就是中括号里的命令）都是可选择的。参见赵西亮（2017）第177页。

**操作实例：**下面，我们利用Card and Krueger（1994，AER）的数据为例，估计新泽西州最低工资调整对新泽西州快餐业就业的影响，数据为两期面板数据，主要变量有：`id`为快餐；`t`为时间，最低工资调整前为0，调整后为1；`treated`为分组变量，1为新泽西，0为宾夕法尼亚；`fte`为全职就业人数，协变量有`bk`、`kfc`、`roys`、`wendys`。如下图2所示：


Variables <span>⌵</span> <span>⌶</span> <span>✕</span>		
	Filter variables here	
	Name	Label
	id	Store ID
	t	Feb. 1992 = 0; N...
	treated	New Jersey = 1; ...
	fte	Output: Full Tim...
	bk	Burger King == 1
	kfc	Kentucky Fried C...
	roys	Roy Rogers == 1
	wendys	Wendy's == 1

Figure 2: 变量

首先，安装DID命令： `ssc install diff, replace`

然后，我们就可以在stata中输入DID命令估计回归系数。

不控制任何协变量时的结果：

`diff fte, period(t) treated(treated) robust`

```
. diff fte,period(t) treated(treated) robust
DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
Number of observations in the DIFF-IN-DIFF: 780
      Baseline      Follow-up
Control: 76      76      152
Treated: 314    314      628
      390      390
Outcome var.    fte    S. Err.    t    P>|t|
Baseline
Control        20.013
Treated        17.069
Diff (T-C)     -2.944    1.440    -2.04    0.041**
Follow-up
Control        17.523
Treated        17.518
Diff (T-C)     -0.005    1.037    -0.00    0.996
Diff-in-Diff    2.939    1.774    1.66    0.098*
R-square:      0.01
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

Figure 3: DID结果

控制协变量时的结果：

`diff fte, period(t) treated(treated) robust cov(bk kfc roys)`

```
. diff fte,period(t) treated(treated) robust cov(bk kfc roys)
DIFFERENCE-IN-DIFFERENCES WITH COVARIATES
DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
Number of observations in the DIFF-IN-DIFF: 780
      Baseline      Follow-up
Control: 76      76      152
Treated: 314    314      628
      390      390
Outcome var.    fte    S. Err.    t    P>|t|
Baseline
Control        21.342
Treated        19.003
Diff (T-C)     -2.339    1.282    -1.83    0.068*
Follow-up
Control        18.852
Treated        19.452
Diff (T-C)     0.600    0.912    0.66    0.511
Diff-in-Diff    2.939    1.873    1.87    0.062*
R-square:      0.19
* Means and Standard Errors are estimated by linear regression
**Robust Std. Errors
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

Figure 4: 控制协变量的DID结果

### 3 断点回归设计（RDD）

如果大家关注了微信公众号“香樟经济学术圈”的话，肯定记得2016年的时候，“满天都是RD”——各种RD经典文献解读，RD原理介绍。社科院付明卫老师写了一篇“断点回归（RD）的规定动作”的推文。里面写道：

订阅了各种经济学类公号的小伙伴们，最近有没有断点回归（RD）设计满天飞的感觉？作为同道中人，我感觉，被推送的RDD论文数量，在今年六七月份明显存在一个断点：从那以后，开始井喷！看着这些推文，多少人心默念：

“论文发表不轻松，要把断点为我用！”

RDD确实是个好方法。它等于是在断点附近的局部随机试验。这一点赖以成立的前提条件，并不难以满足。此外，跟随机试验中全域（global）随机性可以被检验一样，RDD等于局部随机试验的假设，也可以通过观察前定变量的分布是否平衡来检验。从这个意义上讲，RD方法比IV、DiD更接近于随机试验。随机试验是因果识别的终极杀招，越接近随机试验的方法当然越好！

#### 3.1 断点回归估计量

在自然实验中，还可能出现一种情形：个体接受处理完全或部分依赖于某个可观测变量 $W$ 是否超过某一阈值（门槛）。例如，一个学生是否要参加“短学期”依赖于他期末平均绩点（GPA）是否在规定阈值以下。根据前面的理想实验中平均处理效应的idea，估计参加“短学期”的效应也是要比那些GPA在阈值以下（参加短学期）的学生成绩与那些GPA在阈值以上（不参加短学期）的学生成绩。这个有阈值限制的可观测变量 $W$ 称为**参考变量**。

另外一个例子是Lee(2008, JoE)对美国各地区众议员选举中在位党在竞选是否具有优势的分析。美国两大党派在选举中获得的选票份额超过对手时，

该党就是在位党。Lee以民主党选票份额与共和党选票份额之差作为参考变量 $W$ ，间断点为0，只要上次选举中参考变量大于0，即意味着民主党在位，否则共和党在位。图5展示了数据集的散点图，如果 $W > 0$ ，民主党在位。

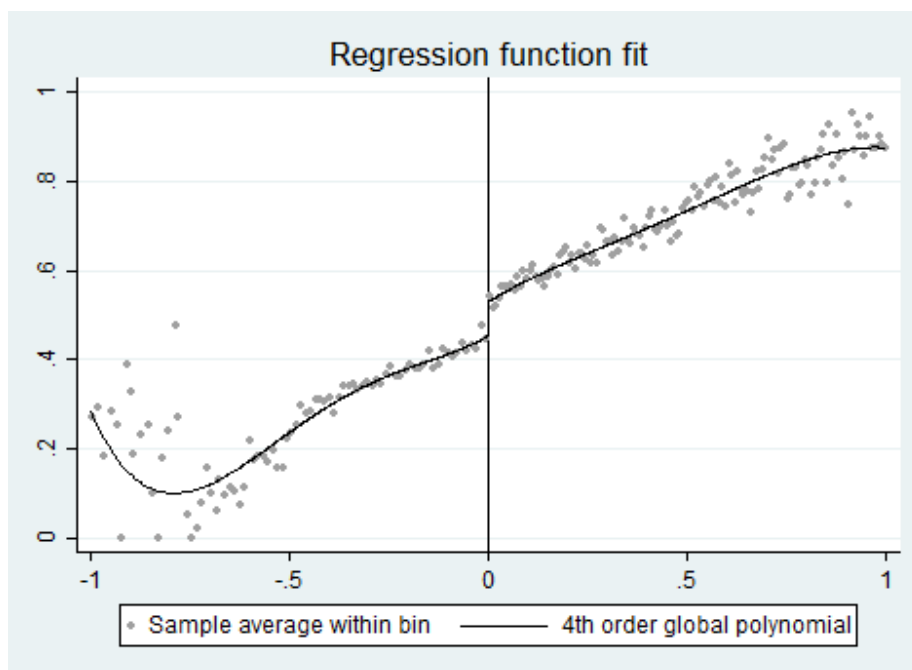


Figure 5: 断点回归设计的散点图

图5显示了，下一次的选举得票份额是现在两党得票之差 $W$ 的函数。如果阈值 $w_0$ 的唯一作用只是识别在位党派，那么，下一次选举得票份额在阈值处的“跳跃”就是竞选中的在位效应估计值。

也就是说，更一般化的分组规制是

$$D_i = \begin{cases} 1 & \text{if } x_i \geq w_0 \\ 0 & \text{if } x_i \leq w_0 \end{cases} \quad (10)$$

假设在选举前，各党派的得票份额的结果 $y_i$ 与 $x_i$ 之间存在如下线性关系：

$$y_i = \alpha + \beta x_i + \epsilon_i \quad (11)$$

我们从图5可以看出，在 $x_i = w_0$ 处， $y_i$ 与 $x_i$ 的线性关系存在一个向上跳跃（jump）的断点。但是，得票率(%)为49.8、49.9、50、50.1、50.2等，可以认为党派在各个方面没有系统差异，因此，这个跳跃发生的唯一原因只可能是 $D_i$ 的处理效应，也就是在位党的优势。

图5也是一个分段函数，因此，我们可以引入虚拟变量来表示具有不同截距的分段函数。因此，我们可以将（11）式重新写成：

$$y_i = \alpha + \beta(x_i - w_0) + \delta D_i + \gamma(x_i - w_0)D_i + \epsilon_i \quad (12)$$

引入交互项 $(x_i - w_0)D_i$ 是为了允许在断点两侧的回归线斜率不同。对方程（12）进行OLS回归，得到的 $\tilde{\delta}$ 就是**断点回归估计量**，也称为局部平均处理效应（LATE）。

在估计断点回归时，**要特别注意两点：**

1、方程（12）中包含了交互项。如果断点两侧的回归线斜率相同，则可不包含交互项。但在实践中，一般断点两侧斜率会不同，因此，如果不包含交互项，则可能导致断点右（左）侧的观测值影响对左（右）侧截距的估计，从而引起偏误；

2、在有交互项的情形下，如果方程中没有 $(x_i - w_0)$ ，而是使用的 $x_i$ ，那么，虽然 $\tilde{\delta}$ 还是断点两侧的距离之差，但是并不等于这两条回归线在 $(x_i = w_0)$ 处跳跃的距离。

由于在参考变量的阈值处，结果变量的跳跃或断点，那些探讨在某一



阈值处接受处理的概率的非连续性的研究被称为**断点回归(RD)**设计。它又分为精准断点回归 (sharp RD) 和模糊断点回归 (fuzzy RD)。SRD在断点 $x_i = w_0$ 处, 个体接受处理的概率从0跳跃到1, 而FRD在断点 $x_i = w_0$ 处, 我们只知道个体接受处理的概率从a跳跃到b, 而 $0 \leq a \leq b \leq 1$ 。

### I. 精准断点回归

在Sharp RDD中, 接受处理完全由参考变量W是否超过某一阈值决定: 当 $W \geq 0$ 时, 民主党是在位党, 当 $W < 0$ 时, 共和党是在位党; 即用D表示民主党是否在位, 当 $W \geq 0$ 时,  $D_i = 1$ , 当 $W < 0$ 时,  $D_i = 0$ 。在这种情形下, 下一次获得选票份额Y在 $W = 0$ 处的跳跃就等于 $W = 0$ 时子样本的处理效应。

由此, 我们可以看到, 上述例子是一个精准断点回归。那么, 我们是否还可以利用 (12) 式进行OLS估计呢? 可以是可以, 但是这存在两个问题:

- 1、可能存在遗漏变量偏误, 例如如果回归中还有高次项 $(x_i - w_0)^2$ ;
- 2、断点回归可以看作是“局部随机实验”, 因此从原理上看, 我们应该只是用断点附近的观测值样本, 但我们在实践中却是用全部样本进行回归。

为了解决上述问题, 我们可以引入高次项, 并限定x的范围 $w_0 - h \leq x_i \leq w_0 + h$ 。这里的h就是最优带宽。回归方程变为

$$y_i = \alpha + \beta_1(x_i - w_0) + \beta_2(x_i - w_0)^2 + \delta_{D_i} + \gamma_1(x_i - w_0)D_i + \gamma_2(x_i - w_0)^2D_i + \epsilon_i, \quad w_0 - h \leq x_i \leq w_0 + h \quad (13)$$

可是, 现在我们不能确定最优带宽h, 还是不能估计 (13) 式呀。在确定h时, 一般是采用非参数回归来最小化均方误差 (MSE)。直观来说, h越

小，偏差越小，但是估计方差会变大；反之亦然。

针对断点回归，我们一般使用两种核回归（kernel regression）：三角核（triangle kernel）与矩形核（rectangle kernel）。

### 关于协变量问题

1、我们可以在（13）式中加入影响Y的协变量。虽然断点回归是局部随机实验，包不包括协变量并不影响断点回归估计量的一致性，**但是加入协变量的好处为：加入协变量可以解释被解释变量Y，那么，就可以减低方差。使得估计更准确。但坏处是：如果加入的协变量是内生变量，与误差项相关，那么就会影响估计量。**

2、实际上，断点回归有个隐含假设：协变量在断点处不存在跳跃，是连续的。如果协变量在断点处也存在跳跃，那么，我们就不能把 $\tilde{\delta}$ 全部归于处理效应。因此，在实践中，我们要现将所有的协变量作为被解释变量，进行断点回归，考察其分布是否在断点处存在跳跃。

此外，我们还应该注意“内生分组”问题。如果个体事先知道分组规则，并可通过自身行为来完全控制分组变量，那么，就可以自行选择进入处理组还是控制组，这就导致了随机分组失败，从而断点回归失灵。

**小贴士:在实践中，我们建议同时汇报出以下情形，以确保结果稳健：**

- 1、分别汇报三角核与矩形核的回归结果；
- 2、分别汇报使用不同带宽的结果；
- 3、分别汇报包含协变量与不包含协变量的结果；
- 4、进行模型设定检验时，包括检验分组变量与协变量的条件密度在断点处是否存在跳跃。

## II. 模糊断点回归

在Fuzzy RDD中，参考变量超过阈值会影响到是否接受处理，但这不是决定处理的唯一影响因素。例如，假设有些GPA在阈值以下的学生并没有参加短学期，而有些GPA超过阈值的学生又参加了短学期。如果临界值规则是一个决定treated非常复杂的过程的一部分，那么上述情况就可能会出现。在模糊断点回归中， $X_i$ 一般与误差项 $u_i$ 相关。

### 3.2 断点回归的规定动作

下面的内容来源于“香樟经济学术圈”的推文，付明卫（2016）：

#### 第1步

检查配置变量（assignment variable，又叫running variable、forcing variable）是否被操纵。画出配置变量的分布图。最直接的方法，是使用一定数量的箱体（bin），画出配置变量的历史直方图（histogrm）。为了观察出分布的总体形状，箱体的宽度要尽量小。频数（frequencies）在箱体间的跳跃式变化，能就断点处的跳跃是否正常给我们一些启发。从这个角度来说，最好利用核密度估计做出一个光滑的函数曲线。McCrary（2008）为判断密度函数是否存在断点提供了一个正规的检验（命令是DCdensity，介绍见陈强编著的《高级计量经济学及Stata应用》（第二版）第569页）。

#### 第2步

挑选出一定数目的箱体，求因变量在每个箱体内的均值，画出均值对箱体中间点的散点图。一定要画每个箱体平均值的图。如果直接画原始数据的散点图，那么噪音太大，看不出潜在函数的形状。不要画非参数估计的连续统，因为这个方法自然地倾向于给出存在断点的印象，尽管总体中本来不存在这样的断点。需要报告由交叉验证法（Cross-validation, CV）挑选的带宽。一般而

言，为了看出潜在函数的形状，不要挑选过大的带宽。但是，带宽太小也会导致看不出潜在函数的形状。比较因变量均值在断点两边的两个箱体间的变化，可以预判处理效应的大小。如果图形中都看不出因变量在断点处有跳跃，那么回归方程也不可能得到显著的结果。

### 第3步

将Y在每个箱体内的均值作为因变量，用处理变量、配置变量的多次项作为自变量，在断点两边分别跑回归，得到因变量的拟合值。将这些拟合值画在第2步的图中，并用光滑的曲线连接起来。在推文人读过的RD论文中，多次项一般都使用1到4次项，但没有论文解释为什么只用到4次项。

### 第4步

检验前定变量在断点处是否跳跃。此步和第1步是RD方法的适用性检验。此步的检验包括两项内容：1. 像前三步那样画前定变量的图。无论参数还是非参数，RD研究都要大把的图！这些图在正式发表的论文中都必不可少！原文中说了这么句话：用RD做的论文，如果缺乏相关的图，十有八九是因为图显示的结果不好，作者故意不报告。2. 将前定变量作为因变量，将常数项、处理变量、配置变量多次项、处理变量和配置变量多次项的交互项作为自变量，跑回归。一个前定变量有一个回归，看所有回归中处理变量的系数估计是否都为0。检验这种跨方程的假设，需要用似不相关回归（Seemingly Unrelated Regression, SUR）（命令是sureg，用法见陈强编著的《高级计量经济学及Stata应用》（第二版）第471-474页）。在推文人读过的RD实证论文中（尤其是AER2015-2016年所有用RD做的论文中），均没用SUR，只是简单的看每个回归中处理变量的系数估计均为0。

### 第5步

检验结果对不同带宽、不同多项式次数的稳健性。尝试的其它带宽，一般是最优带宽的一半和两倍。挑选多项式的最优次数，可用赤池信息准则（Akaike's Information Criterion, AIC）。在我们尝试的包含配置变量1次方、2次方、……N次方的众多方程中，AIC取值最小的那个就是我们想要的。实操时，试到多少次为好？原文中至少试到了6次。我们做研究时需要试到10次还是100次呢？Gelman和Imbens（2014）解除了我们的这个烦恼，详见“江湖上的新动作”这一部分。

### 第6步

检验结果对加入前定变量的稳健性。如上所述，如果不能操控配置变量的假设成立，那么无论前定变量与因变量的相关性有多高，模型中加入前定变量都不应该影响处理效应的估计结果。如果加入前定变量导致处理效应的估计结果变化较大，那么配置变量可能存在排序现象，前定变量在断点处也很可能存在跳跃。实操时在确定多项式的次数后，直接在回归方程中加入前定变量。如果这导致处理效应估计值大幅变化或者导致标准误大幅增加，那么可能意味着函数中多项式的次数不正确。另外一个检验是残差化，看相同次数的多项式模型对残差的拟合好不好。

### 江湖上的新动作

Thistlethwaite和Campbell1960年首次用RD方法做政策评估。经过近40年的沉寂后，20世纪90年代末以来，经济学关于RD方法的性质、局限性等方面的理论研究有了巨大进展。关于RD方法本身的研究，并没有因为Lee和Lemieux（2010）的发表而停止。我把Lee和Lemieux（2010）发表后的进展称作“新招式”。据我的不完全了解，“新招式”有这些：

1. 多项式次数的选择。根据Lee和Lemieux（2010），配置变量的次数

要试到N次。但是，Gelman和Imbens（2014）的NBER工作论文说，试到N次的做法要不得，最多只能搞到2次。至于原因，他们讲了三条，感兴趣的请参考原文。尽管他们的论文还未正式发表，但学界都已乖乖听他们的啦。AER2015-2016年间所有用RD做的论文（共6篇）里，5篇都只用1次或2次。

**2. 最优带宽。** Lee和Lemieux（2010）介绍了两种确定最优带宽的方法：拇指规则法（rule of thumb）和交叉验证法（CV）。现在，江湖上有另外两种比较受关注的方法：IK法和CCT法。IK法以Imbens和Kalyanaraman两个人命名，对应着论文Imbens和Kalyanaraman（2012）。这篇论文发表在Review of Economic Studies, Lee和Lemieux（2010）文中提到过此文2009年的NBER工作论文版。CCT法以Calonico、Cattaneo和Titiunik三个人命名，对应着论文Calonico、Cattaneo和Titiunik（2014a）。用非参数法做断点回归估计时的stata命令rd，就是用IK法确定最优带宽。stata命令rdrobust、rdbwselect，提供CV、IK、CCT三种不同的最优带宽计算方法选项。然而，尽管Calonico、Cattaneo和Titiunik（2014a）2014年发表在牛刊Econometrica上，AER2015-2016年上的文章没有买它的账。AER2015-2016年的6篇相关文章中，仅有1篇提到过CCT，其他5篇就像不知道Calonico、Cattaneo和Titiunik（2014a）这篇文章。我甚为不解！难道是因为CCT非牛人？

**3. 核密度检验。** Lee和Lemieux（2010）介绍了McCrary（2008）的核密度检验方法。Frandsen（2013）提出了一种新的检验方法，感兴趣的请参考原文。

### 3.3 例子与stata操作

下面，我们使用Lee(2008)的数据来演示一下断点回归的stata操作。这个数据集中包括两个变量：vote表示民主党的选票份额；margin表示民主党在上次

竞选获得的选票与共和党选票份额之差。因此，margin就是参考变量，如果margin大于0，民主党就是在位党，这是一个SRD。我们感兴趣的问题是，在位党是否会获得优势。我们将样本限制在 $margin \pm 0.5$ 之间，样本共有4900个。

第一步，下载安装断点回归命令。Calonico et al. (2014) 提供了一个专门进行断点回归分析的程序包rdrobust，里面包含三个命令：rdplot——断点回归图形；rdbwselect——选择最优带宽；rdrobust——估计断点回归估计量。

**findit rdrobust** （查找、安装rd程序）

stata会出来下列界面

```

search for rdrobust

-----
Search of official help files, FAQs, Examples, SJs, and STBs

SJ-14-4 st0366 . . . Robust data-driven inference in reg.-discontinuity design
. . . . . S. Calonico, M. D. Cattaneo, and R. Titiunik
(help rdrobust, rdwbselect, rdplot if installed)
Q4/14 SJ 14(4):909--946
conducts robust data-driven statistical inference in
regression-discontinuity designs

Web resources from Stata and other users

(contacting http://www.stata.com)

2 packages found (Stata Journal and STB listed first)
-----

st0366_1 from http://www.stata-journal.com/software/sj17-2
SJ17-2 st0366_1. Update: Local polynomial... / Update: Local polynomial
regression-discontinuity / estimation with robust bias-corrected
confidence / intervals and inference procedures / by Sebastian Calonico,
University of Miami, / Miami, FL / Matias D. Cattaneo, University of

st0366 from http://www.stata-journal.com/software/sj14-4
SJ14-4 st0366. Robust data-driven inference... / Robust data-driven
inference in the regression- / discontinuity design / by Sebastian
Calonico, University of Miami, / Coral Gables, FL / Matias D. Cattaneo,
University of Michigan, / Ann Arbor, MI / Rocio Titiunik, University of

(click here to return to the previous screen)

(end of search)

```

Figure 6: 查找rd程序

点击“st0366\_1 from <http://www.stata-journal.com/software/sj17-2>”，进入页面再点击“click here to install”进行安装。

第二步，画断点图。输入

`rdplot vote margin, c(0) nbins(50)`（画断点图）

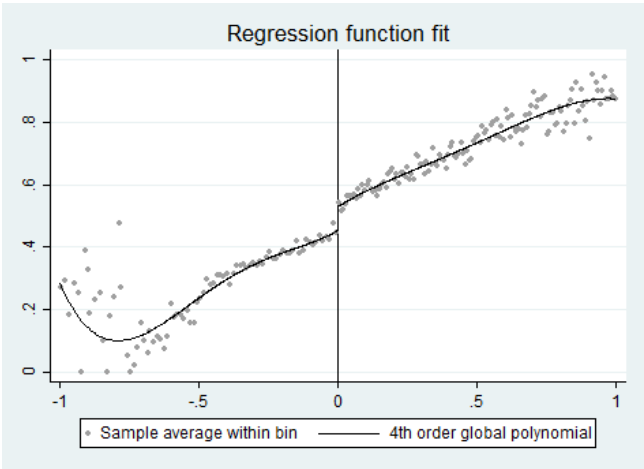


Figure 7: 断点回归设计的散点图

第三步，选择最优带宽。输入

`rdbwselect vote margin, c(0) kernel(uni) all`（选择最优带宽）

上述命令中，`kernel()`是设置核估计方法。此处选择的是矩形核。得到的结果是

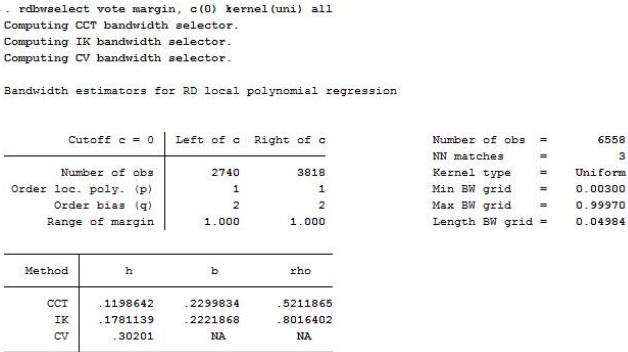


Figure 8: 最优带宽选择结果

第四步，估计断点回归估计量。输入：



`rdrobust vote margin, c(0) kernel(uni) all`（估计断点回归估计量）

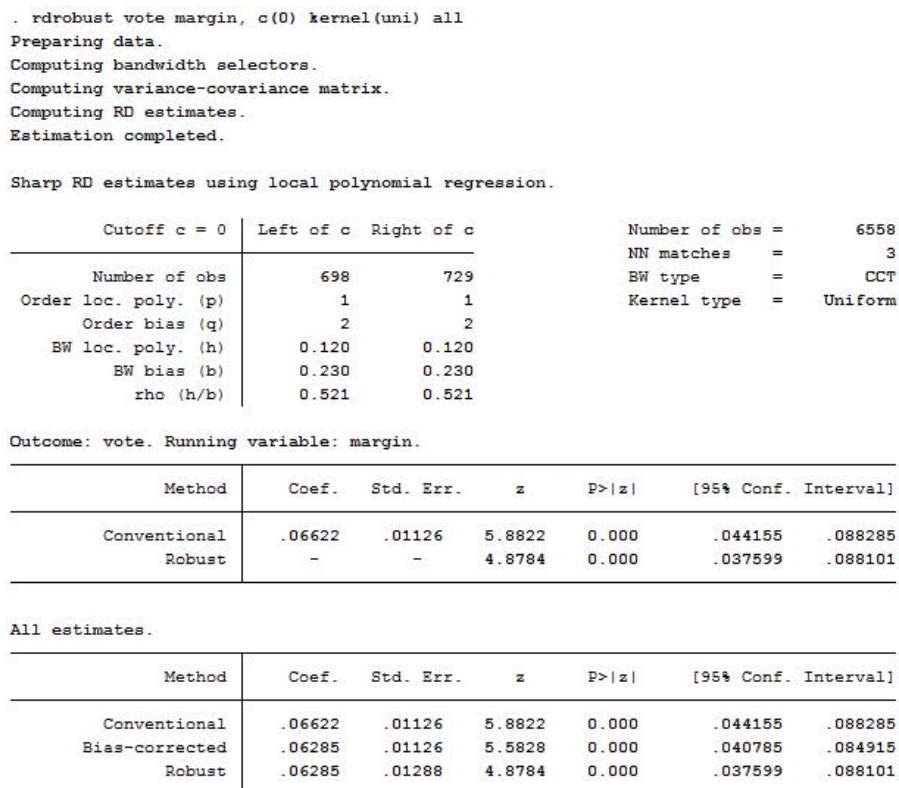


Figure 9: 断点回归估计结果

以上四步就是断点回归的基本操作步骤。

下面，我们来详细介绍一下上述三个rd命令的基本语法格式：

```
rdplot depvar indepvar [if] [in][,c(#) p(#) kernel(#) weights() h(# #)
nbins(# #) binselect() scale(# #) ci() shade generate(id_var meanx_var meany_var cil_var cir_var)
graph_options(gphopts) hide]
```

两个必选项：

- 1、depvar是结果变量、原因变量或其他协变量；
- 2、indepvar是参考变量；

中括号里的全部为可选命令：

3、c(#)用于设置断点的位置；

4、p(#)设定多项式的阶数；

5、kernel(#)设定核估计类型，有三种：三角核triangular、Epanechnikov核epanechnikov、矩形核uniform；

6、h( # #)设置断点左右的带宽；

7、nbins( # #)设定划分的区间数；

8、binselect()设定带宽的选择方法；

9、ci() shade画出每个区间拟合点的置信区间，shade表示置信区间用阴影表示。

```
rdbwselect depvar indepvar [if] [in][,c(#) p(#) q(#) deriv(#) fuzzy(fuzzyvar[sharpbw])
covs(#) kernel(#) bwselect() scaleregul(#) vce(vcetype[vceopt1 vceopt2]) al-
l]
```

最优带宽选择命令中与画图命令中有很多相同命令。需要注意的是：

1、q(#)为偏差修正的多项式阶数；

2、deriv(#)可以用于估计弯折回归（RKD），0为断点回归，1为弯折回归；

3、fuzzy(fuzzy-var[sharpbw])用于模糊断点回归或模糊弯折回归，fuzzyvar是原因变量，sharpbw表示使用结果变量的最优带宽；

4、covs(#)引入协变量；

5、bwselect()最优带宽的估计方法。

```
rdrobust depvar runvar [if] [in][,c(#) p(#) q(#) deriv(#) fuzzy(fuzzyvar[sharpbw])
covs(#) kernel(#) h( # #) b( # #) rho(#) bwselect() scaleregul(#)
```

scalepar(#) vce(vctype[vceopt1 vceopt2]) level(#) all]

估计命令与带宽估计命令相似。

## 4 其他因果效应识别方法——匹配与合成控制法

### 4.1 匹配法

我们回忆一下在理想实验中，平均处理效应（ATE）等于服用新药之后的病情（ $E(Y_{1i}|D_i = 1)$ ）与没有服用新药病情（ $E(Y_{0i}|D_i = 0)$ ）之间的差异。

即  $ATE = E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0)$ 。

下面，我们来做一个简单的数学变换（不要看到“数学”两个字就怕，它们很多时候都是“纸老虎”，例如，此处只要学过中学的加减移项即可看懂）。

我们将ATE计算式右边加上一项 $E(Y_{0i}|D_i = 1)$ ，又减去一项 $E(Y_{0i}|D_i = 1)$ ：

$$\begin{aligned} ATE &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\ &= [E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)] + [E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)] \quad (14) \end{aligned}$$

(10)式显示，我们可以把ATE分解成两个部分（由中括号表示）：第一个部分是“参与者平均处理效应（ATT）”——项目实际参与者的平均处理效应；第二部分是选择偏差——假设参与者和未参与者的处理效应之差（实际上两类人都未参与项目）。

从政策制定者的角度来看，他们可能更关心ATT，因为这是政策或项目实施后的毛收益，我们只需要将这个毛收益与政策成本进行比较，就能判定这个政策是否值得。但从（10）式可知，ATE 与ATT 之间是存在一个选择偏差。

（需要注意的是，这里的选择偏差与第五讲中的样本选择偏误有所不同。样本选择偏误是所选取的样本不是总体中的代表性样本所引起的偏误，这时通常不考虑政策或项目效应。

那么，如何解决这个问题呢？我们前面已经提到过，随机实验最大的优势在于其随机分配。那么，完全随机选择处理与否自然可以消除上述选择偏差。但实践中，很难执行这种随机实验。在这种情况下，我们有两种方法可以尽量消除选择偏误：

### I. 匹配估计量

**使用条件：假设个体根据可观测变量来选择是否参与项目**

下面，我们用一个就业培训项目为例。在对这个项目进行效应评估时，我们除了能观测到人们是否参与了该项目 $D_i$ 和项目实施前后的收入 $Y_i$ ，我们还能观测到参与者的一些个体特征，例如年龄、受教育程度、肤色、婚否、性别等——协变量。

如果个体是否参与项目完全由某些协变量 $X$ 决定的，那么，我们就可以利用**匹配估计方法**来估计出处理效应。

匹配估计的**思想**其实非常简单：实践中，个体 $i$ 参与培训了（处理组），他就不可能又“穿越”回到过去不参加培训。这个时候，我们就需要在没有参加培训的那些人（控制组）找到某个或某些人 $j$ ，那么怎么找呢？上面说过，参与项目 $D_i$ 完全取决于可观测变量 $X_i$ ，那么，自然就是找那些与参与者 $i$ 有相近 $X$ 的未参与者 $j$ 。我们选择到的 $X_j$ 与 $X_i$ 越接近， $j$ 参与培训的概率就越接近 $i$ 。那么，我们就可以把 $j$ 的收入 $Y_j$ 近似当作 $i$ 在没有参与培训情形下的收入，然后把 $i$ 的实际收入 $Y_i$ 减去这个近似收入 $Y_j$ ，即可得到培训的处理效应，即匹配估计量。

我们来看看表3中的一个简单匹配例子，参见陈强（2014）：541页。

Table 4: 匹配估计

i	$D_i$	$X_i$	$Y_i$	匹配对象	$\tilde{Y}_{0i}$	$\tilde{Y}_{1i}$
1	0	2	7	5	7	8
2	0	4	8	4,6	8	7.5
3	0	5	6	4,6	6	7.5
4	1	3	9	1,2	7.5	9
5	1	2	8	1	7	8
6	1	3	6	1,2	7.5	6
7	1	1	5	1	7	5

根据匹配估计的思想，是否参加项目D只依赖于X，而且我们要从未参加组（D=0）里为参加者（D=1）找到X相近的那些人，例如，我们要从1-3中为4-7找到X相近的人，第7个人的X=1，而1-3中X最接近1的是第一个人的X=2，因此，与第7个人匹配的对象就是第1个人。那么，我们就应该把第1个人的Y=7当做第7个人没有参加项时的 $\tilde{Y} = 7$ ，而实际参与项目的Y=5，因此，该项目的效应就是 $5 - 7 = -2$ 。其他人也这样匹配。

#### 两个技术细节需要特别注意：

第一，在寻找匹配对象时是否允许匹配对象放回。放回就是说，当表3中的第4个人与第2个人匹配了，但是在对第6个人进行匹配时，第2个人仍然在备选之列，也就是第2个人匹配之后又放回备选对象行列，可以进行下一次匹配。而不放回，就意味着2与4匹配之后，2就不能进行下次匹配，那么6只能与1进行匹配。

第二，是否允许匹配对象并列。也就是说，4与1、2的X都比较接近，那么，在允许并列的情形下，我们会将1、2的Y的均值作为4的 $\tilde{Y}$ 。如果不允许并列，软件就会根据数据排列的顺序来选择匹配对象及其 $\tilde{Y}$ ，例如，在不允许并列时，根据表3数据的排列顺序，与4匹配的就是1，那么 $Y_1$ 就会作为4的 $\tilde{Y}_4$ 。

一般来说，匹配估计量会存在偏差，因为 $X_i$ 不可能与 $X_j$ 完全相同。那么，

在非精确匹配的情形下：（1）一对一匹配，偏差较大，方差较小；（2）一对多匹配，偏差较小，方差加大。**经验法则：最好进行一对四匹配，这样能使均方误差(MSE)最小。**

上面就是匹配估计法的最基本思想：就是找到两组中X最接近的对象进行匹配。虽然原理很简单，但是实际操作起来可就难了。因为在实践中，我们通常不会像表3那样，D依赖的X只有单一变量，而是X中会包含很多个变量。也就是说，我们要根据多个协变量同时进行比较，例如对不同人的年龄、受教育程度、性别等同时进行比较，这个时候就可能会遇到，两个人年龄相仿，但受教育程度差距很大，受教育程度相同，但年龄差距有很大，这个时候我们要这么比较这匹配对象是否接近呢？

这个时候，我们就需要拿出一种有效的武器——**倾向得分匹配（PSM）**。

那么，PSM的思想是什么呢？说简单也简单，说难也难！

简单是因为，我们就是要找到一个批判不同对象之间是否相似，但在多个X情况下，我们无从下手。那么，我们想法办把多个X转换成一个指标，即通过某种函数 $f(X)$ ，把多维变量变成一维变量，这个一维变量就是**倾向得分（PS）**。然后，我们就可以根据这个倾向得分来进行上述匹配。

难是因为，这个转换函数 $f(X)$ 到底是什么？这个问题我们就不展开了。有兴趣者可自行查阅相关资料。

#### **PSM计算处理效应的步骤：**

（1）选择协变量X。尽量将影响D和Y的相关变量都包括在协变量中。如果协变量选择不当或太少，就会引起效应估计偏误；

（2）计算倾向得分，一般用logit回归；

（3）进行倾向得分匹配。如果倾向得分估计较为精确，那么，X在匹配后

的处理组和控制组之间均匀分布，这就是**数据平衡**。那么我们检验得分是否准确就需要计算X 中每个分量的“标准化偏差”。**经验法则：一般来说，标准化偏差不能超过10%，如果超过10%，我们就要重新返回第（2）步重新计算，甚至第（1）步重新选择匹配协变量，或者改变匹配方法。**

（4）根据匹配后的样本计算处理效应。

第三步中，得分匹配效果不好，可能要改变匹配方法：一、k邻近匹配；二、卡尺匹配或半径匹配；三、卡尺内最近邻匹配；四、核匹配；五、局部线性回归匹配；六、样条匹配。在实践中，并没有明确准则来限定使用哪种匹配方法。但有一些**经验法则可作为参考：**

（1）如果控制组个体不多，则应该选择又放回匹配；

（2）如果控制组有较多个体，则应该选择核匹配；

（3）**最常用的方法：尝试不同的匹配方法，然后比较它们的结果，结果相似说明很稳健。结果差异较大，就要去深挖其中的原因。**

**PSM的局限性：**

（1）大样本；

（2）要求处理组和控制组有较大的共同取值范围；

（3）只控制了可观测的变量，如果存在不可观测的协变量，则会引起“隐性偏差”。

## **II. DID-PSM估计量**

**使用条件：假设个体根据不可观测变量来选择是否参与项目**

上面提到，如果存在根据不可观测变量进行选择时，会引起“隐性偏差”。而消除这种影响的方法很多，其中之一就是利用面板数据，且结合DID-PSM来计算处理效应。DID和PSM原理我们在上面均详细讲过，因此，下面直接给出

其stata操作。

除了DID-PSM之外，断点回归和工具变量法都可以尽量消除“隐性偏差”。

#### 4.1.1 PSM的Stata应用演示

下面，我们用Dehejia and Wahba（1999）职业培训的数据来演示stata的匹配操作。

从stata13.0开始，就提供匹配命令`teffects`命令，我们在stata中输入“`help teffects`”就可以看到命令描述：

```
Title
[TE] teffects — Treatment-effects estimation for observational data

Syntax
teffects subcommand ... [, options]

subcommand      Description
-----
aipw             augmented inverse-probability weighting
ipw             inverse-probability weighting
ipwra           inverse-probability-weighted regression adjustment
nrmatch         nearest-neighbor matching
overlap         overlap plots
psmatch         propensity-score matching
ra             regression adjustment

Description
teffects estimates potential-outcome means (POMs), average treatment effects (ATEs), and average treatment
Regression-adjustment, inverse-probability-weighted, and matching estimators are provided, as are doubly r:
weighting. teffects overlap plots the estimated densities of the probability of getting each treatment lev

The outcomes can be continuous, binary, count, fractional, or nonnegative. The treatment model can be bins

For a brief description and example of each estimator, see Remarks in teffects intro.
```

Figure 10: 匹配命令描述

下面，我们采用1:1最近邻匹配，估计培训对个人收入的效应。输入如下命令：



## 4.2 合成控制法

在上面的讲解中，我们反复多次强调，随机实验中的“随机性”最为关键，因为它可以很好地识别出我们所要进行比较的“处理组”和“控制组”。一旦“处理组”和“控制组”的outcomes被我们观测到，我们就可以利用处理组的结果减去控制组的结果得到我们感兴趣的`处理效应`，例如`产业政策效应`。

但困难之处在于，实践中，我们往往只能观测到“处理组”的结果，而观测不到“如果处理组没有接受处理”的结果。这个时候，我们就需要去选择或“假象”一个或多个“控制组”（注：这里的假象，也是有理有据地假象。我们还有一个专业术语叫做“构造”）。

陈强老师说“选择控制组是一门艺术。确实，寻找适当的控制组（control group），即在各方面都与受干预地区相似却未受干预的其他地区，以作为处理组（treated group，即受到干预的地区）的反事实替身（counterfactuals）。但通常不易找到最理想的控制地区（control region），在各方面都接近于处理地区（treated region）。

比如，要考察仅在北京实施的某政策效果，自然会想到以上海作为控制地区；但上海毕竟与北京不完全相同。或可用其他一线城市（上海、广州、深圳）构成北京的控制组，比较上海、广州、深圳与北京在政策实施前后的差别，此方法也称“比较案例研究”（comparative case studies）。但如何选择控制组通常存在主观随意性（ambiguity），而上海、广州、深圳与北京的相似度也不尽相同。

因此，在上面一小节，我们简单介绍了匹配方法——通过能体现个体特征的协变量的相似度来构造出一个“控制地区”的结果。除了上述PSM方法之外，还有另一种构造“控制组”的方法——**合成控制法（Synthetic Control**

Method)。

合成控制法是由Abadie and Gardeazabal (2003)提出来研究西班牙巴斯克地区 (Basque country) 恐怖活动的经济成本。其基本思想为：虽然无法找到巴斯克地区的最佳控制地区，但通常可对西班牙的若干大城市进行适当的线性组合，以构造一个更为贴切的“合成控制地区” (synthetic control region)，并将“真实的巴斯克地区”与“合成的巴斯克地区”进行对比，故名“合成控制法”。合成控制法的一大优势是，可以根据数据 (data-driven) 来选择线性组合的最优权重，避免了研究者主观选择控制组的随意性。

Abadie et al.(2010)用合成控制法研究了美国加州香烟控制99法案对加州香烟消费的影响。为了限制香烟消费，1988年11月，加州政府通过了99法案，主要内容是将香烟消费税每包提高25美分，该法案1989年1月正式生效，作者主要考察这一政策对香烟消费的抑制作用到底有多大。由此，我们知道，香烟消费税在加州地区改变（提高）了，而在美国其他州并没有变化。因此，加州是“处理地区”，美国其他州是可能的“控制地区”。

此时，可能我们立马就能想起来，可以使用PSM呀。我们找出各个州的典型特征作为协变量，然后计算匹配得分，进行对比处理地区和匹配地区的香烟消费，得到香烟消费税提高的控烟效应。如果大家还记得上一节最后我们给出的PSM方法局限性——要求大样本。大家可能就会意识到：哎呀，PSM用在加州控烟税上可能有问题，因为这个样本中就只有加州和其他40多个州多年数据，这个样本似乎也不大。这种情况也是我们在写论文或者进行政策评估时经常会遇到的情况，所以用什么方法已经要深思熟虑。

我们接着来看加州控烟税的例子。为了避免其他州类似的政策对控制组产

生影响，Abadie等剔除了研究期内出台相似控烟政策的州，最后潜在控制组里剩下38个州。也就是说样本包括美国39各州，1970-2000年的面板数据，变量包括：州年度人均香烟消费量、香烟平均零售价格、州人均收入对数、州人口结构（15-24岁比例）、州人均啤酒消费量等等。

记 $Y_{it}$ 为地区 $i$ 在 $t$ 期实际观测到的结果变量，即香烟消费量。

记 $Y_{it}^I$ 的上标 $I$ 表示地区 $i$ 接受政策干预，即这个变量表示加州在提高香烟消费税后的香烟消费量。同理， $Y_{it}^N$ 的上标 $N$ 表示没有受到政策干预。根据上文将的处理效应，我们实际上感兴趣的是：

$$ATE = Y_{it}^I - Y_{it}^N$$

现在的问题是， $Y_{it}^N$ 观测不到，因此，我们要估计出它。那么，怎么估计呢？即书本上的“因子模型”来估计 $Y_{it}^N$ ：

$$Y_{it}^N = \delta_t + \theta_t Z_i + \lambda_t \mu_i + \epsilon_{it} \quad (15)$$

(11)式右边第一项表示时间固定效应(time fixed effects)。第二项表示可观测的向量（不受政策干预影响，也不随时间而变；比如，干预之前的预测变量之平均值），由于它对 $Y$ 的效应随时间可变，因此，其系数 $\theta$ 有时间下标。第三项表示不可观测的“交互固定效应”（Interactive Fixed Effects），即个体固定效应 $\mu_i$ 与时间固定效应 $\lambda_t$ 的乘积（Bai, 2009）。第(4)项为随机扰动项。

我们咋一看，这不就是面板回归吗？但是请注意，第三项“交互固定效应”是不可观测的变量。这一点很重要。

我们回忆一下合成控制法的基本思想：虽然无法找到加州地区的最佳控制地区，但通常可对美国其他的若干州进行适当的线性组合，以构造一个更为

贴切的“合成控制地区”（synthetic control region），并将“真实的加州地区”与“合成的加州地区”进行对比。也就是说，我们要利用其他州的线性组合来拟合出加州的 $Y_{it}^N$ ，线性组合就是每个州的Y前面乘以一个权重W之和，假设每个州的权重用 $w_j$ 表示，那么，我们将线性组合表示为

$$\sum_{j=1}^{j=N} w_j Y_{jt} = \delta_t + \theta_t \sum_{j=1}^{j=N} w_j Z_j + \lambda_t \sum_{j=1}^{j=N} w_j \mu_j + \sum_{j=1}^{j=N} w_j \epsilon_{jt} \quad (16)$$

接下来，我们用（11）式减去（12）式：

$$Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = \theta_t (Z_i - \sum_{j=1}^{j=N} w_j Z_j) + \lambda_t \sum_{j=1}^{j=N} (\mu_i - w_j \mu_j) + \sum_{j=1}^{j=N} (\epsilon_{it} w_j \epsilon_{jt}) \quad (17)$$

我们的目的是用其它38个州的线性组合 $\sum_{j=1}^{j=N} w_j Y_{jt}$ 来代替加州的 $Y_{it}^N$ 。因此，我们想要 $Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = 0$ 。那么，我们只要使得（13）式右边的第一个括号为0，第二个括号为0，那么整个（13）式的期望就等于0。此时，用 $\sum_{j=1}^{j=N} w_j Y_{jt}$ 合成的结果就是无偏的。但是，我们要注意， $\mu_i$ 是不可观测的变量，因此， $\lambda_t \sum_{j=1}^{j=N} (\mu_i - w_j \mu_j)$ 估计不出来，也即是说，上述估计行不通。

这怎么办呢？

我们仔细观察（13）式，里面可观测的变量有干预前的 $Y_{it}$ 、 $Y_{jt}$ 、 $Z_i$ 、 $Z_j$ ，那么，我们只需要找到最优的权重 $w_j$ 使得：

$$Y_{it}^N - \sum_{j=1}^{j=N} w_j Y_{jt} = 0, Z_i - \sum_{j=1}^{j=N} w_j Z_j = 0$$

即根据可观测的经济特征与干预前结果变量所选择的合成控制w，也会使得合成控制的不可观测特征接近于处理地区。反之，如果无法找到w，使得合成控制能很好地复制（reproduce）处理地区的经济特征以及干预之前的结果变

量，则不建议使用合成控制法。Abadie et al. (2010) 已经证明了，当干预前的时期数趋向于无穷，那么，合成控制估计量就是无偏的。

小贴士：合成控制法对样本数据的要求为政策干预以前需要很多期数据，有人认为至少需要15年的数据，而政策干预后需要有5年以上的数据。同时地区最好超过10个，但是又不会太多。最为重要的是，接受政策干预的地区个数极少。

#### 4.2.1 例子及stata操作

我们来继续看看加州控烟税政策的效果。

第一步，我们在stata中输入下列命令：

`ssc install synth, replace` （下载并安装synth程序）

其中，选择项“replace”表示如有此命令更新版本，可以新命令覆盖旧命令。

命令synth的基本句型为：

`synth y x1 x2 x3, trunit(#) trperiod(#) counit(numlist) xperiod(numlist) mspeperiod() figure  
resultspanel() nested allopt keep(filename)`

其中，“y”为结果变量（outcome variable），“x1 x2 x3”为预测变量（predictors）。

必选项“trunit(#)”用于指定处理地区（trunit 表示treated unit）。

必选项“trperiod(#)”用于指定政策干预开始的时期（trperiod 表示treated period）。

选择项“counit(numlist)”用于指定潜在的控制地区（即donor pool，其中counit 表示control units），默认为数据集中的除处理地区以外的所有地区。

选择项 “`xperiod(numlist)`” 用于指定将预测变量（predictors）进行平均的期间，默认为政策干预开始之前的所有时期（the entire pre-intervention period）。

选择项 “`mspeperiod()`” 用于指定最小化均方预测误差（MSPE）的时期，默认为政策干预开始之前的所有时期。

选择项 “`figure`” 表示将处理地区与合成控制的结果变量画时间趋势图，而选择项 “`resultsperiod()`” 用于指定此图的时间范围（默认为整个样本期间）。

选择项 “`nested`” 表示使用嵌套的数值方法寻找最优的合成控制（推荐使用此选项），这比默认方法更费时间，但可能更精确。在使用选择项 “`nested`” 时，如果再加上选择项 “`allopt`”（即 “`nested allopt`”），则比单独使用 “`nested`” 还要费时间，但精确度可能更高。

选择项 “`keep(filename)`” 将估计结果（比如，合成控制的权重、结果变量）存为另一Stata数据集（`filename.dta`），以便进行后续计算。更多选择项，详见`help synth`。

第二步，打开数据集之后，输入下列命令：

```
xtset state year（声明面板数据）
```

第三步，在stata中输入下列合成控制法估计命令：

```
synth cigsale retprice lnincome age15to24 beer cigsale(1975) cigsale(1980) cigsale(1988),  
trunit(3) trperiod(1989) xperiod(1980(1)1988) figure nested keep(smoking_synth)
```

估计结果如下：

图6显示，大多数州的权重为0，而只有以下五个州的权重为正，即Colorado (0.161)，Connecticut (0.068)，Montana (0.201)，Nevada (0.235)与Utah (0.335)，此结果与Abadie et al. (2010)汇报的结果非常接近（细微差别或由于计算误

州↵	权重↵	州↵	权重↵
Alabama↵	0↵	Nevada↵	.235↵
Arkansas↵	0↵	New Hampshire↵	0↵
Colorado↵	.161↵	New Mexico↵	0↵
Connecticut↵	.068↵	North Carolina↵	0↵
Delaware↵	0↵	North Dakota↵	0↵
Georgia↵	0↵	Ohio↵	0↵
Idaho↵	0↵	Oklahoma↵	0↵
Illinois↵	0↵	Pennsylvania↵	0↵
Indiana↵	0↵	Rhode Island↵	0↵
Iowa↵	0↵	South Carolina↵	0↵
Kansas↵	0↵	South Dakota↵	0↵
Kentucky↵	0↵	Tennessee↵	0↵
Louisiana↵	0↵	Texas↵	0↵
Maine↵	0↵	Utah↵	.335↵
Minnesota↵	0↵	Vermont↵	0↵
Mississippi↵	0↵	Virginia↵	0↵
Missouri↵	0↵	West Virginia↵	0↵
Montana↵	.201↵	Wisconsin↵	0↵
Nebraska↵	0↵	Wyoming↵	0↵

Figure 11: 权重

差)。

下图显示了加州和其他38个州的合成结果：

	Treated	Synthetic
retprice	89.42222	89.41464
lnincome	10.07656	9.858694
age15to24	.1735324	.1735444
beer	24.28	24.21326
cigsale(1975)	127.1	127.0633
cigsale(1980)	120.2	120.4545
cigsale(1988)	90.1	91.6356

Figure 12: 合成结果比较

从图7可知，加州与合成加州的预测变量均十分接近，故合成加州可以很好地复制加州的经济特征。然后比较二者的人均香烟消费量在1989年前后的表现：

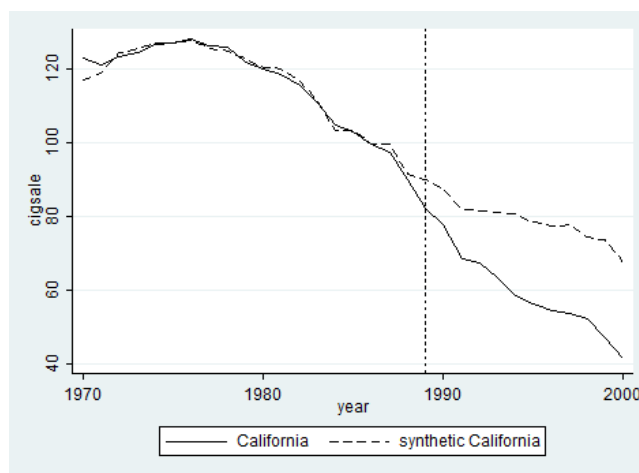


Figure 13: 合成控制结果

从上图可知，在1989年控烟法之前，合成加州的人均香烟消费与真实加州几乎如影相随，表明合成加州可以很好地作为加州如未控烟的反事实替身。在控烟法实施之后，加州与合成加州的人均香烟消费量即开始分岔，而且此效应越来越大。



更直观地，可打开另一Stata程序，调用已存的数据集smoking\_synth.dta，计算加州与合成加州人均香烟消费之差（即处理效应），然后画图。

```
use smoking_synth.dta, clear
```

（如不打开另一Stata程序，则此数据集将覆盖原有的数据集smoking.dta）

```
gen effect = _Y_treated - _Y_synthetic
```

（定义处理效应为变量effect，其中“\_Y\_treated”与“\_Y\_synthetic”分别表示处理地区与合成控制的结果变量）

```
label variable _time "year"
```

```
label variable effect "gap in per-capita cigarette sales(in packs)"
```

（为了画图更漂亮，加上时间变量与处理效应的标签，可使用变量管理器(variable manager)来方便地加标签）

```
line effect _time, xline(1989, lp(dash)) yline(0,lp(dash))
```

（画处理效应的时间趋势图，并在横轴1989年处与纵轴0处分别画虚线，结果见下图）

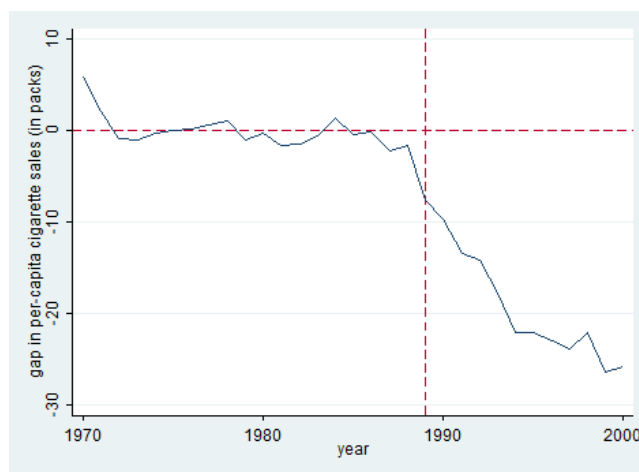


Figure 14: 趋势图

图9显示，加州控烟法对于人均香烟消费量有很大的负效应，而且此效应随着时间推移而变大。具体来说，在1989-2000年期间，加州的人均年香烟消费减少了20多包，大约下降了25%之多，故其经济效应十分显著（economically significant）。

到此，我们关心的合成控制法主要结果都呈现出来了。但是在使用合成控制法时，如何进行稳健性检验与统计推断？

为了检验上述合成控制估计结果的稳健性，Abadie et al. (2010)加入了更多的预测变量，比如失业率、收入不平等、贫困率、福利转移、犯罪率、毒品相关的逮捕率、香烟税、人口密度等；发现结果依然稳健。

另外一个担心是，地区之间无互相影响（no interference between units）的假定可能不满足，比如加州的反烟运动可能波及其他州，烟草行业或将其他州的香烟广告预算投入到加州，甚至从其他州走私便宜香烟到加州。Abadie et al. (2010)根据史实对此进行了探讨，认为这些效应均不大，至少不可能导致上文图中如此大的处理效应。

### **安慰剂检验**

上述结果为对控烟法处理效应的点估计。此点估计是否在统计上显著（statistically significant）？Abadie et al. (2010)认为，在比较案例研究中，由于潜在的控制地区数目通常并不多，故不适合使用大样本理论进行统计推断。

为此，Abadie et al. (2010)提出使用“安慰剂检验”（placebo test）来进行统计检验，这种方法类似于统计学中的“排列检验”（permutation test），适用于任何样本容量。

“安慰剂”（placebo）一词来自医学上的随机实验，比如要检验某种新药

的疗效。此时，可将参加实验的人群随机分为两组，其中一组为实验组，服用真药；而另一组为控制组，服用安慰剂（比如，无用的糖丸），并且不让参与者知道自己服用的究竟是真药还是安慰剂，以避免由于主观心理作用而影响实验效果，称为“安慰剂效应”（placebo effect）。

安慰剂检验借用了安慰剂的思想。具体到加州控烟法的案例，我们想知道，使用上述合成控制法所估计的控烟效应，是否完全由偶然因素所驱动？换言之，如果从donor pool随机抽取一个州（而不是加州）进行合成控制估计，能否得到类似的效应？

为此，Abadie et al. (2010)进行了一系列的安慰剂检验，依次将donor pool中的每个州作为假想的处理地区（假设也在1988年通过控烟法），而将加州作为控制地区对待，然后使用合成控制法估计其“控烟效应”，也称为“安慰剂效应”。通过这一系列的安慰剂检验，即可得到安慰剂效应的分布，并将加州的处理效应与之对比。

在此有个技术细节，即在对某个州进行安慰剂检验时，如果在“干预之前”其合成控制的拟合效果很差（均方预测误差MSPE很大），则有可能出现在“干预之后”的“效应”波动也很大，故结果不可信。类似地，如果合成加州在干预前对于加州的拟合很差，则我们也不会相信干预之后的合成控制估计结果。

### 注意事项

在使用合成控制法时，需要特别注意以下几点：

1、我们是将没有实行政策的地区作为备选的合成控制组，但是如果*i*地区实行政策对*j*地区也产生了影响，那么，这个时候，我们就应该将*j*地区从备选控制组中提出掉，例如加州提高控烟税，对威斯康辛州有很大影响，那么，我们

就应该将威斯康辛州从38个备选里去掉；

2、如果在研究期间，有一些地区受到非常大的特殊冲击，那么，这时候我们也要将其剔除；

3、尽量使得控制地区与处理地区具有相似的特征；

4、合成控制法对样本数据的要求为政策干预以前需要很多期数据，有人认为至少需要15年的数据，而政策干预后需要有5年以上的数据。同时地区最好超过10个，但是又不会太多。最为重要的是，接受政策干预的地区个数极少。

5、如果政策冲击的效应需要一段时间才会显现（滞后效应），则也要求干预后的期数足够大。

## 5 克服计量方法选择困难症

### DID与PSM

我曾经看过一个最简单的描述：

DID是比较四个点，Treated before, treated after; control before, control after;

Matching是比较两个点：Treated, control;

DID+Matching是用matching的方法来确定treated和control。

### 断点回归

1、断点回归最大的优势就是在断点附近接近随机实验，也就是说，断点回归可以认为是一种“局部随机实验”；

2、但是，正因为断点回归只在断点附近随机性强，因此，仅能推断断点处的因果关系，不一定能推广到其他样本，也就是说结论在理论上的一般性存

疑。

### 合成控制法与DID

首先，根据Abadie et al. (2010)的因子模型（factor model），合成控制法对双向固定效应模型作了推广。具体来说，双重差分法仅允许个体固定效应与个体时间效应以相加（additive）的形式存在，隐含假设所有个体的时间趋势都相同（parallel trend assumption）；而合成控制法的因子模型，则允许“互动固定效应”（interactive fixed effects），即可以存在多维的共同冲击（common shocks），而每位个体对于共同冲击的反应（factor loading）可以不同，故允许不同个体有不同的时间趋势。

其次，Abadie et al. (2015)指出，回归法也可以视为对控制地区作了线性组合，且权重之和也为1；而不同之处在于，合成控制法的权重必须非负，但回归法的权重可能出现负值，即出现过分外推（extrapolation）而离开了样本数据的取值范围（support of the data）。比如，在跨国研究中，将很不相同的国家放在一起进行回归，就可能出现过分外推，而导致“外推偏差”（extrapolation bias）。由于合成控制法的权重必须非负，故避免了过分外推。