

小白学统计|面板数据分析与Stata应用笔记（九）

本期内容： 双重差分模型

一、双重差分模型的介绍

双重差分模型（difference-in-differences）主要被用于社会学中的**政策效果评估**。使用双重差分方法对各大改革的效果进行评估，从而更好地为政府制定政策提供建议。

双重差分方法的**原理**是基于一个反事实的框架来评估政策发生和不发生这两种情况下被观测因素 y 的变化。如果一个外生的政策冲击将样本分为两组—受政策干预的 *Treat* 组和未受政策干预的 *Control* 组，且在政策冲击前，*Treat* 组和 *Control* 组的 y 没有显著差异，那么我们就可以将 *Control* 组在政策发生前后 y 的变化看作 *Treat* 组未受政策冲击时的状况（反事实的结果）。通过比较 *Treat* 组 y 的变化(ΔY_t)以及 *Control* 组 y 的变化 (ΔY_c)，我们就可以得到政策冲击的实际效果($\Delta\Delta Y = \Delta Y_t - \Delta Y_c$)。

具体而言就是，利用政策的准自然实验将研究对象随机的分成**处理组**和**对照组**，其中受到政策影响的个体称为处理组，反之是对照组。为了估计政策效应，先比较处理组在政策发生前后的变化，**需要注意的是**，这种变化的部分可能是**时间效应**所导致的，即没有政策的发生处理组也会随着时间变动而发生变化。为了剔除这种随着时间变动的影响我们引入了对照组，由于对照组不会受到政策的影响，因而对照组在政策实施前后发生的变化就是由于时间效应带来的变化。我们使用对照组在政策发生前后的结果变量的变化剔除时间效应。

直观上讲，我们只不过是通过对对照组构造了一个没有受到政策影响的处理组。

通常情况下，我们可以通过三种方式来理解双重差分方法，即**表格法**、**画图法**和**回归法**。

1、表格法

通过表格法我们可以直观的理解双重差分方法。

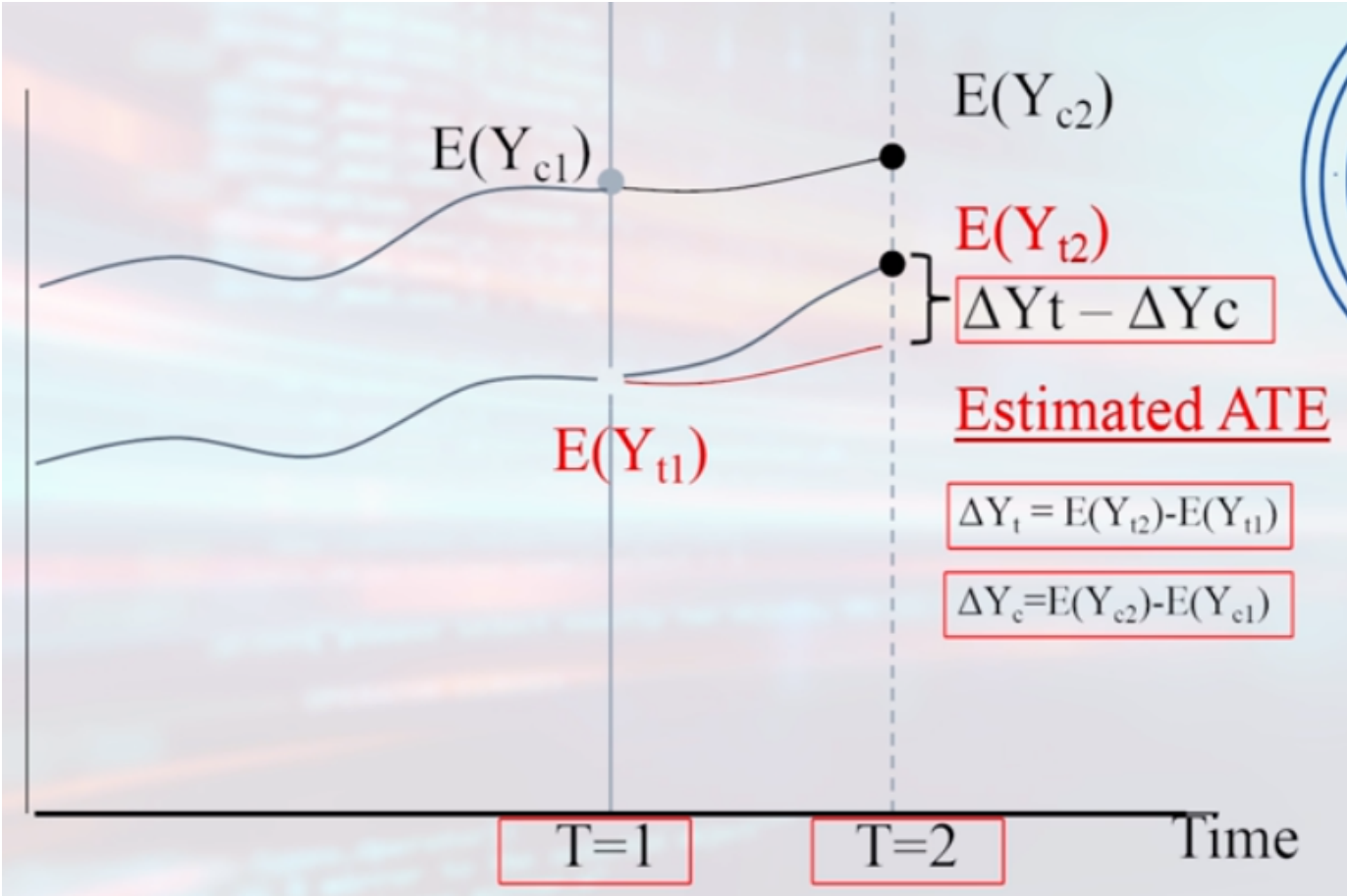
首先，我们分别计算处理组和对照组在政策发生前后结果变量的**均值**；然后，用处理组政策发生后的均值减去政策发生前的均值得到处理组政策前后的变化，将对照组也进行同样的操作，从而得到对照组在政策发生前后的变化（对照组的变化即为由于**时间效应**的存在对结果变量造成的影响）；最后，用处理组的变化减去对照组的变化，**剔除掉时间效应**，就可以得到政策效应，这两次相减的过程就体现了**双重差分**的思想。

	Before Change	After Change	Difference
Group 1 (Treat)	$E(Y_{t1})$	$E(Y_{t2})$	$\Delta Y_t = E(Y_{t2}) - E(Y_{t1})$
Group 2 (Control)	$E(Y_{c1})$	$E(Y_{c2})$	$\Delta Y_c = E(Y_{c2}) - E(Y_{c1})$
Difference			$\Delta\Delta Y = \Delta Y_t - \Delta Y_c$

表格中**Group 1**为处理组，**Group 2**为对照组。 $E(Y_{t1})$ 表示处理组政策发生前的结果变量均值； $E(Y_{t2})$ 表示处理组政策发生后的结果变量均值； $E(Y_{c1})$ 表示对照组政策发生前的结果变量均值； $E(Y_{c2})$ 表示对照组政策发生后的结果变量均值； ΔY_t 表示处理组政策发生前后结果变量均值的差分； ΔY_c 表示对照组政策发生前后结果变量均值的差分； $\Delta\Delta Y$ 是处理组政策发生前后结果变量均值的差减去对照组政策发生前后结果变量均值的差，即我们所关注的**政策效应**。

2、画图法

我们可以将对照组与处理组结果变量的均值随时间的变化绘制成时序图，这样可以直观的理解双重差分方法。



在上图中我们可以看到，上面的曲线是对照组结果变量随时间变动的轨迹；下面的曲线是处理组随时间变动的曲线。政策发生前为 $T=1$ ，政策发生后为 $T=2$ ，我们需要估计的是政策发生前后结果变量发生了多大的变化。与表格法一样， $E(Y_{t1})$ 表示处理组政策发生前的结果变量均值； $E(Y_{t2})$ 表示处理组政策发生后的结果变量均值； $E(Y_{c1})$ 表示对照组政策发生前的结果变量均值； $E(Y_{c2})$ 表示对照组政策发生后的结果变量均值； ΔY_t 表示处理组政策发生前后结果变量均值的差分； ΔY_c 表示对照组政策发生前后结果变量均值的差分； $\Delta\Delta Y$ 为剔除了时间效应之后的政策效应。

通过上图的这一曲线图可以发现，结果变量的数值高低是不重要的，重要的是它的变化方向以及变化的大小。

###由画图法可以看到，我们需要特别注意的一点是共同趋势假定：政策发生前，对照组与处理组的结果变量呈现出共同的变化趋势。只有满足这一假定，才能使用对照组来模拟处理组在未受到政策冲击时的时间效应。

除了共同趋势检验，在DID方法中还有一个经常被提起的检验，即安慰剂检验。共同趋势检验是安慰剂检验，而反过来则不成立，因为安慰剂检验可以包括除了共同趋势检验以外的检验。在现有的政策评估的文献中，有很多种进行安慰剂检验的方法。

- 第一种，可以采用政策发生之前的数据，将政策实施前的除第一年之外的所有年份"人为地"设定成为处理组的政策实施年份。然后，根据DID模型逐年回归，当所有回归中的交互项系数都不显著时，说明通过了安慰剂检验，表明之前识别的政策平均效应是可靠的，否则就是不可靠的。如果政策实施前有n年数据，那么就要做n-1次上述回归。

- 第二种, "人为地"随机选择政策实施对象(处理组), 然后使用全样本做DID回归, 如果交互项系数不显著, 则判断政策对随机选择的处理组都不存在政策效应, 可以进一步证明之前识别的政策平均效应结果是可靠稳健的。
- 第三种, 改变被解释变量, 通常选择理论上不受政策影响的其他变量, 保持真实的对照组和处理组、真实的政策实施时间, 重新进行DID回归, 理想的结果是, 该政策的实施对其他被解释变量都不存在政策效应。

3、回归法

我们考虑**两组两期**的回归模型, 两组指处理组和对照组, 两期是指政策发生前和政策发生后。

回归法设定的模型如下:

$$Y_{it} = \beta_0 + \beta_1 treat_i + \beta_2 P_t + \beta_3 treat_i * P_t + \varepsilon_{it}$$

其中, Y_{it} 是被解释变量; $treat_i$ 是组别虚拟变量: 如果个体属于处理组, 则 $treat_i = 1$, 如果属于对照组, 则 $treat_i = 0$; P_t 是时间虚拟变量: 政策实施之后 $P_t = 1$, 否则 $P_t = 0$; β_3 的大小和方向反映了**双重差分政策效应**。

因为回归法具备很多其他方法不具备的优点, 所以我们通常采用回归法估计政策效应。

回归法的优点:

- 能够计算政策效应的标准误和显著性;
- 回归过程中可以控制其他影响被解释变量的变量;
- 回归法允许包含多期的数据使结果更加准确;
- 在评估政策时还可以将政策强度考虑在内(比如: 研究关税政策变动的影响)。

在现实中, 很多政策的实施都是渐进式的, 通常会有试点区然后再逐步推广至其他地区, 针对这种现实情况, 我们就需要构造**两组多期的DID模型**和**多组多期的DID模型**。

(1) 两组多期DID模型的设定

$$Y_{it} = \beta_0 + \beta_1 treat_i + \beta_2 P_t + \beta_3 treat_i * P_t + \lambda_t + \varepsilon_{it}$$

其中, λ_t 为时间效应, 其他变量与上面两组两期的模型的变量解释相同。

(2) 多组多期DID模型的设定

$$Y_{ist} = \beta_1 D_{st} + \beta_2 X_{ist} + u_s + \lambda_t + \varepsilon_{it}$$

其中的下标 i 表示个体, t 表示时间, s 表示组别; u_s 为多组的估计效应, λ_t 为多重的时间效应; D_{st} 为虚拟变量, 表示组别在 t 时是否受到了政策的影响, 如果受到政策的影响则为1, 否则为0。

此外, 在现有的政策性评估的文章中, 除了以上的DID模型外, 还有控制个体效应的DID模型。模型设定如下所示:

$$Y_{it} = \beta_1 treat_i * P_t + u_i + \lambda_t + \varepsilon_{it}$$

其中, u_i 是个体效应, 其他变量解释与上述相同。

如果政策发生前后, 处理组和对照组中的**个体个数**没有发生变化, 那么在方程中加入个体效应, 可以估计出处理组和对照组中的**所有的**个体效应。在两次相减的过程中, 个体效应就会抵消, β_1 仍然能够反映政策效应。此时, 这种控制个体效应的DID模型就是可行的, 否则, 就不能真正的识别政策效应。

二、做DID需要注意的若干问题

在构建双重差分模型时，我们需要注意一些问题。

第一，DID方法不仅仅能够估计出政策效应，还能识别出组别固定效应和时间效应。

第二，双重差分模型反映的是政策冲击的"即时效应"，也就是政策发生时所能带来的平均效应是多少，但是有时候政策冲击具有滞后效应，今年实施的政策可能明年或者更之后的年份才能展现出效果。

对于滞后效应的衡量：
通常情况下可以用 D_{st} 的滞后一期来估计政策冲击的平均效应。在科研实践中，如果某一个政策不存在即时效应一定要尝试考察是否存在滞后效应。

第三，使用DID方法必须保证处理组和对照组在政策实施前的时间趋势是一致的，这个假设只能通过足够长的时间序列数据才能检验，必须要保证对照组和处理组在政策发生之前至少有两期的数据。如果不满足共同趋势的假定，那么估计得到的交互项的系数就可能存在偏误。

如果对照组和处理组确实不满足共同趋势的假定，我们可以通过如下三种思路进行解决：

第一个，是寻找更好的对照组（这里我们需要用到合成控制法），把多个对照组加权构成一个虚拟的对照组，使得虽然每个对照组都与处理组的时间趋势不一样，但加权后的虚拟对照组的时间趋势与处理组的一样。这样虚拟对照组与处理组就满足了共同趋势的假定。这种方法常用于案例分析中。

第二个，是估算出因时间趋势不同而带来的偏差，然后从双重差分结果中减去这个偏差，这种方法也被称作三重差分方法（Difference-in-differences-in-differences,DDD）。

三重差分法的思路是，既然两个地区的时间趋势不一样，那么可以分别在两个地区寻找一个没有受到干预影响的人群或行业，通过对这两组的双重差分估算出时间趋势的差异，然后再从原来实验组和对照组的双重差分估算值中减去这个时间趋势差异。

对于三重差分模型的设定，可以参考陈强教授的《高级计量经济学及stata应用》第343页的例子：

假设美国B州针对65岁及以上老年人（用E表示）引入了一项新的医疗保健政策，而该政策不适用于65岁以下的人群（用N表示），欲考察此政策对于健康状况 y 的影响。

如果运用双重差分模型，我们有两种方法可以参考：

- 以B州65岁及以上人群作为实验组，65岁以下人群作为对照组。这种方法的缺陷是，年轻人和老年人的健康状况随着时间可能发生不同的变化；
- 以相邻A州65岁及以上的老年人作为对照组，这种方法虽然避免了上述方法中的问题，但是不能保证正常实施之前，两个州的老年人健康状况有着同样的时间趋势。

所以，最好的办法是将上述两种方法综合起来。

首先，基于第一种方法，将B州65岁及以上人群作为实验组，65岁以下人群作为对照组；

然后，使用A州的数据衡量年轻人和老年人的健康状况随时间变化的不同；

最后，再从双重差分的结果中把这一不同减去，就能够估计出政策效应。这就是三重差分法。

我们引入三重差分模型：

$$y = \beta_0 + \beta_1 B + \beta_2 E + \beta_3 B \times E + \gamma_0 D + \gamma_1 D \times B + \gamma_2 D \times E + \delta D \times B \times E + \varepsilon_{it}$$

其中， B 表示是否为B州（B州为1，A州为0）， E 表示是否为老年人（是为1，否为0）， D 为时间虚拟变量（政策实施后为1，实施前为0）。交互项 $D \times B \times E$ 前面的系数 δ 就是政策效果。

第三个，是增加不同组别不同的时间趋势 λ_{st} ，即控制同一年份不同组别不同的时间趋势。需要注意的是，如果在基准回归中加入了这一趋势变量，那么在安慰剂检验时也需要加入，二者必须保持一致，通过这种控制，之前提到的三种安慰剂检验方法，跑出来的交互项系数可能就不显著，从而可能通过安慰剂检验。

$$Y_{ist} = \beta_0 + \beta_1 D_{st} + \beta_2 X_{ist} + u_s + \lambda_{st} + \varepsilon_{it}$$

三、双重差分模型的Stata操作

我们通过Card和Krueger发表在1994年AER上的一篇文章关于最低工资调整的案例来理解双重差分模型的stata基本步骤和程序。

他们研究了最低工资调整对就业的影响，文章以快餐业为考察对象，处理组和对照组分别为新泽西州和宾夕法尼亚州。选择快餐业的原因是因为快餐业属于低技能行业，此类行业的从业人员对于最低工资的反映最为强烈。

文章选择的政策事件背景是，从1992年4月1日起，新泽西州最低工资由每小时4.25美元上升到每小时5.05美元，而宾夕法尼亚州最低工资保持不变。新泽西州的快餐业为处理组，宾夕法尼亚州的快餐业为对照组。

	id	t	treated	fte	bk	kfc	roys	wendys	treatedt
1	1	0	NJ	31	1	0	0	0	0
2	1	1	NJ	40	1	0	0	0	1
3	2	0	NJ	13	1	0	0	0	0
4	2	1	NJ	12.5	1	0	0	0	1
5	3	0	NJ	12.5	0	1	0	0	0
6	3	1	NJ	7.5	0	1	0	0	1
7	4	0	NJ	16	0	0	1	0	0
8	4	1	NJ	20	0	0	1	0	1
9	5	0	NJ	20	0	0	1	0	0
10	5	1	NJ	25	0	0	1	0	1
11	6	0	NJ	3	0	0	1	0	0
12	6	1	NJ	6	0	0	1	0	1
13	8	0	NJ	.	0	0	0	1	0
14	8	1	NJ	27.5	0	0	0	1	1
15	9	0	NJ	32	1	0	0	0	0
16	9	1	NJ	16	1	0	0	0	1
17	10	0	NJ	25	1	0	0	0	0
18	10	1	NJ	22.5	1	0	0	0	1
19	11	0	NJ	25	1	0	0	0	0
20	11	1	NJ	24	1	0	0	0	1
21	12	0	NJ	18.25	1	0	0	0	0
22	12	1	NJ	22.5	1	0	0	0	1
23	13	0	NJ	12.5	0	1	0	0	0
24	13	1	NJ	7.5	0	1	0	0	1
25	14	0	NJ	14	0	1	0	0	0
26	14	1	NJ	12.5	0	1	0	0	1
27	15	0	NJ	15.5	0	0	1	0	0
28	15	1	NJ	15	0	0	1	0	1
29	16	0	NJ	17	0	0	1	0	0
30	16	1	NJ	15.5	0	0	1	0	1

根据上述数据集，我们设定双重差分模型。

$$Y_{it} = \beta_0 + \beta_1 NJ_i + \beta_2 after_t + \beta_3 NJ_i * after_t + \varepsilon_{it}$$

设定分组虚拟变量和时间虚拟变量：如果是新泽西州则设定 $NJ = 1$ ，否则 $NJ = 0$ ；对于时间虚拟变量，如果是1992年4月1日之后，则 $after = 1$ ，否则 $after = 0$ 。

Stata中可以通过【diff】命令对双重差分模型进行估计。需要注意的是：diff命令只能用于两组两期的方法。

#首次使用diff命令，需要通过命令【ssc install diff】进行安装。

diff命令的基本格式是：

【diff outcome_var [if] [in] [weight] [, options]】

其中，outcome_var是被解释变量；必选项treat(varname)用来指定处理变量；必选项period(varname)用来指定时间虚拟变量；选项cov(varlist)用来指定其他的控制变量；选项cluster(varname)和robust用来

设定标准差，cluster是聚类标准误，robust是稳健标准误；选项test是进行平衡性检验，比较处理组和对照组在政策发生之前在控制变量方面是否存在显著性的差异，一般来说，如果没有显著性的差异，就说明处理组和对照组是高度相似的，是符合使用DID方法的。

在Stata中调用数据集

```
1 use http://fmwww.bc.edu/repec/bocode/c/CardKrueger1994.dta
```

使用des命令查看数据集，结果如下：

. des

Contains data from http://fmwww.bc.edu/repec/bocode/c/CardKrueger1994.dta

obs: 820 Dataset from Card&Krueger (1994)

vars: 8 27 May 2011 20:36

variable name	storage type	display format	value label	variable label
id	int	%8.0g		Store ID
t	byte	%8.0g		Feb. 1992 = 0; Nov. 1992 = 1
treated	long	%8.0g	treated	New Jersey = 1; Pennsylvania = 0
fte	float	%9.0g		Output: Full Time Employment
bk	byte	%8.0g		Burger King == 1
kfc	byte	%8.0g		Kentuky Fried Chicken == 1
roys	byte	%8.0g		Roy Rogers == 1
wendys	byte	%8.0g		Wendy's == 1

Sorted by: id t

通过sum命令查看各变量的描述统计的情况。

. sum

Variable	Obs	Mean	Std. Dev.	Min	Max
id	820	246.5073	148.1413	1	522
t	820	.5	.5003052	0	1
treated	820	.8073171	.3946469	0	1
fte	801	17.59457	9.022517	0	80
bk	820	.4170732	.4933761	0	1
kfc	820	.195122	.3965364	0	1
roys	820	.2414634	.4282318	0	1
wendys	820	.1463415	.3536639	0	1

使用命令【diff fte,t(treated) p(t)】进行双重差分模型估计（命令中没有加入控制变量）。

```
1 diff fte,t(treated) p(t)
```

```
. diff fte,t(treated) p(t)
```

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 801

Before		After		
Control:	78	77		155
Treated:	326	320		646
	404	397		

Outcome var.	fte	S. Err.	t	P> t
Before				
Control	19.949			
Treated	17.065			
Diff (T-C)	-2.884	1.135	-2.54	0.011**
After				
Control	17.542			
Treated	17.573			
Diff (T-C)	0.030	1.143	0.03	0.979
Diff-in-Diff	2.914	1.611	1.81	0.071*

R-square: 0.01

* Means and Standard Errors are estimated by linear regression

Inference: * p<0.01; ** p<0.05; * p<0.1

从上表中的估计结果可以看到，政策实施前处理组和对照组的差值Diff(T-C)是-2.884，政策实施后的差值是0.030，二者的差值也就是双重差分的系数为2.914，其对应的p值为0.071，表示在10%的显著性水平下通过了检验。

#需要注意的是在政策实施前，处理组是326家企业，政策实施后是320家企业，这是一个非平衡面板数据。

此外，我们也可以构建treated*t的交互项进行OLS回归，两者的结果是一致的。

在上述模型的基础上，我们加入控制变量对双重差分模型进行估计（report为报告控制变量的系数结果）。

```
1 diff fte,t(treated) p(t) cov(bk kfc roys) report
```


DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 801

	Before	After	
Control:	78	77	155
Treated:	326	320	646
	404	397	

Report - Covariates and coefficients:

Variable(s)	Coeff.	Std. Err.	t	P> t
bk	0.917	0.889	1.032	0.303
kfc	-9.205	1.006	-9.154	0.000
roys	-0.897	0.967	-0.927	0.354

Outcome var.	fte	S. Err.	t	P> t
Before				
Control	21.161			
Treated	18.837			
Diff (T-C)	-2.324	1.031	-2.25	0.024**
After				
Control	18.758			
Treated	19.369			
Diff (T-C)	0.611	1.037	0.59	0.556
Diff-in-Diff	2.935	1.460	2.01	0.045**

R-square: 0.19

* Means and Standard Errors are estimated by linear regression

Inference: * p<0.01; ** p<0.05; * p<0.1

最后，我们进行平衡性检验来查看各变量在处理组和控制组之间是否存在差异。

```
1 diff fte,t(treated) p(t) cov(bk kfc roys wendys) test
```

```
. diff fte,t(treated) p(t) cov(bk kfc roys wendys) test
```

TWO-SAMPLE T TEST

Number of observations (baseline): 404

	Before	After	
Control:	78	-	78
Treated:	326	-	326
	404	-	

t-test at period = 0:

Variable(s)	Mean Control	Mean Treated	Diff.	t	Pr(T > t)
fte	19.949	17.065	-2.884	2.44	0.0150**
bk	0.436	0.408	-0.028	0.45	0.6538
kfc	0.154	0.209	0.055	1.09	0.2769
roys	0.218	0.252	0.034	0.62	0.5368
wendys	0.192	0.132	-0.060	1.37	0.1726

*** p<0.01; ** p<0.05; * p<0.1

从结果中可以看到，被解释变量在处理组和对照组之间存在显著性的差异，而其余的控制变量的平均效应不存在显著性的差异，这就说明对照组和控制组选择是恰当的。



长按二维码关注