DID操作stata代码展示

1.主要含义

双重差分法，英文名Differences-in-Differences，别名“倍差法”，小名“差中差”。作为政策效应评估方法中的一大利器，双重差分法受到越来越多人的青睐，概括起来有如下几个方面的原因：（1）可以很大程度上避免内生性问题的困扰：政策相对于微观经济主体而言一般是外生的，因而不存在逆向因果问题。此外，使用固定效应估计一定程度上也缓解了遗漏变量偏误问题。（2）传统方法下评估政策效应，主要是通过设置一个政策发生与否的虚拟变量然后进行回归，相较而言，双重差分法的模型设置更加科学，能更加准确地估计出政策效应。（3）双重差分法的原理和模型设置很简单，容易理解和运用，并不像空间计量等方法一样让人望而生畏。（4）尽管双重差分法估计的本质就是面板数据固定效应估计，但是DID听上去或多或少也要比OLS、FE之流更加“时尚高端”，因而DID的使用一定程度上可以满足“虚荣心”。

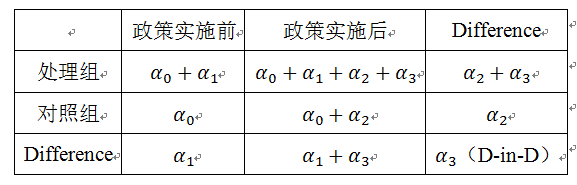
在细致介绍DID之前首先强调一点，一般而言，DID仅适用于面板数据，因此在只有截面数据时，还是不要浪费心思在DID上了。不过，事无绝对，在某些特殊的情景下，截面数据通过巧妙的构造也是可以运用DID的，大神Duflo曾经就使用截面数据和DID研究了南非的养老金计划项目对学前儿童健康的影响，感兴趣的可以去搜搜大神的文章。

具体来说，基准的DID模型设置如下：



其中，du为分组虚拟变量，若个体i受政策实施的影响，则个体i属于处理组，对应的du取值为1，若个体i不受政策实施的影响，则个体i属于对照组，对应的du取值为0。dt为政策实施虚拟变量，政策实施之前dt取值为0，政策实施之后dt取值为1。du·dt为分组虚拟变量与政策实施虚拟变量的交互项，其系数α3就反映了政策实施的净效应。

从DID的模型设置来看，要想使用DID必须满足以下两个关键条件：一是必须存在一个具有试点性质的政策冲击，这样才能找到处理组和对照组，那种一次性全铺开的政策并不适用于DID分析；二是必须具有一个相应的至少两年（政策实施前后各一年）的面板数据集。为什么交互项du·dt的系数α3就能够体现出政策的净效应呢？这一点可以通过下表来体现（下表也反映了双重差分法五个字的真正含义）



双重差分法的基本思想就是通过对政策实施前后对照组和处理组之间差异的比较构造出反映政策效果的双重差分统计量，将该思想与上表的内容转化为简单的模型（1），这个时候只需要关注模型（1）中交互项的系数，就得到了想要的DID下的政策净效应。

更进一步地，DID的思想与上表的内容可以通过下图来体现：

图中红色虚线表示的是假设政策并未实施时处理组的发展趋势。事实上，该图也反映出了DID最为重要和关键的前提条件：共同趋势（Common Trends），也就是说，处理组和对照组在政策实施之前必须具有相同的发展趋势。DID的使用不需要什么政策随机以及分组随机，只要求CT假设，因此用DID做论文时必须对该假设进行验证，至于具体怎么验证，后面再说。

很多时候，大家在看使用DID的文献时，会发现别人的基准模型和模型（1）并不完全一致，别人的模型如下：

7ed047db8c0fd7f8293e913ba5d011cd

别人的模型里只有交互项du·dt，而缺失了du和dt，是哪里出问题了么？并没有，模型（1）和（2）本质还是一样的，且模型（2）在多年面板数据集里更为常见。模型（2）中，v为个体固定效应，更为精确地反映了个体特征，替代了原来粗糙的分组变量du；V为时间固定效应，更为精确地反映了时间特征，替代了原来粗糙的政策实施变量dt。因而，du和dt并未真正从模型中消失，只是换了个马甲。模型（2）事实上就是一个双向固定效应模型，这也是为什么DID方法一定程度上可以减轻遗漏变量偏误的原因（主要是消除那些不可观测的非时变因素，为了使估计结果尽可能准确，模型中还是要加入控制变量）。

在介绍完DID的基本思想和模型设定后，现在要开始强调同等重要的内容，那就是稳健性检验——必须证实所有效应确实是由政策实施所导致的。很多人对这一点并不重视，认为DID很简单，随便跑几个回归就可以写出一篇大作了。关于DID的稳健性检验，主要表现在两个方面：

1. 共同趋势的检验。这个假设是比较难验证的，看文献时会发现别人经常没有做该检验，比如，很多人做DID时只有政策实施前后各一年的数据，这样的的话根本无法验证政策实施前的趋势问题。不过，如果是多年的面板数据，可以通过画图来检验CT假设。

（2）即便处理组和对照组在政策实施之前的趋势相同，仍要担心是否同时发生了其他可能影响趋势变化的政策，也就是说，政策干预时点之后处理组和对照组趋势的变化，可能并不真正是由该政策导致的，而是同时期其他的政策导致的。这一问题可以概括为处理变量对产出变量作用机制的排他性，对此，可以进行如下的检验：

A. 安慰剂检验，即通过虚构处理组进行回归，具体可以：a）选取政策实施之前的年份进行处理，比如原来的政策发生在2008年，研究区间为2007-2009年，这时可以将研究区间前移至2005-2007年，并假定政策实施年份为2006年，然后进行回归；b）选取已知的并不受政策实施影响的群组作为处理组进行回归。如果不同虚构方式下的DID估计量的回归结果依然显著，说明原来的估计结果很有可能出现了偏误。

B. 可以利用不同的对照组进行回归，看研究结论是否依然一致。

C. 可以选取一个完全不受政策干预影响的因素作为被解释变量进行回归，如果DID估计量的回归结果依然显著，说明原来的估计结果很有可能出现了偏误。等等。

2.命令安装

1.安装命令：

ssc instann diff

2.diff的基本格式

diff y,treat(VAR) period(VAR) cov(Z1 Z2) robust report test

y为被解释变量（结果变量），treat(VAR)为分组虚拟变量（即控制住记为0，处理组记为1），period(VAR) 用来制定时间虚拟变量，实验前和实验后，实验前记为0，试验后记为1，cov(Z1 Z2)用来表示其它变量，report为汇报其它控制变量系数，robust为稳健标准误回归，test用于检验在基期控制组和被控制组系数是否相等（为了确定样本的无差异性）。

Options:

kernel选择默认的核密度估计；

pscore(varname )将核密度的结果保存，以varname命名；

id(varname )使用核密度估计时候，需要用这个确定个体的标识；

bw(#)选择核密度估计的带宽、默认的是0.06；

ktype(kernel ) 选择核密度估计的类型主要用这几种：epanechnikov (the default), gaussian , biweight , uniform and tricube .

rcs表明核密度估计被用来设置重复的截面数据，这里可以不适用id作为标识；同时使用该命令意味着假定控制变量不随着时间发生改变；

qdid(quantile )执行分位数DID，通常跟核密度kernel和协变量cov一起使用；

logit采用logit估计倾向匹配得分，默认是采用probit形式；

addcov(varlist )核密度估计时候，用于添加协变量之外的变量，常用来控制时间效应；

ddd(varname )三重差分时候的二次分组虚拟变量；

test测试控制住和处理组协变量的差异；

nostar不展示P值的星号；

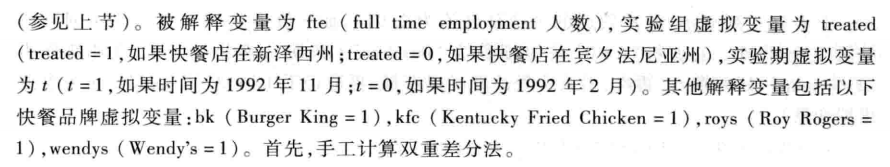
report当使用kernel时候，展示倾向匹配汇报结果；

bs采用自助法；

reps(int)采用自助法的时候重复的次数，默认是50次；

support表示仅使用共同范围的观测值进行匹配。

3.数据展示汇报



首先引入数据

use "C:\接单\DID操作演示数据.dta"

destring t,replace

destring treated,replace

destring fte,replace

destring bk,replace

destring kfc,replace

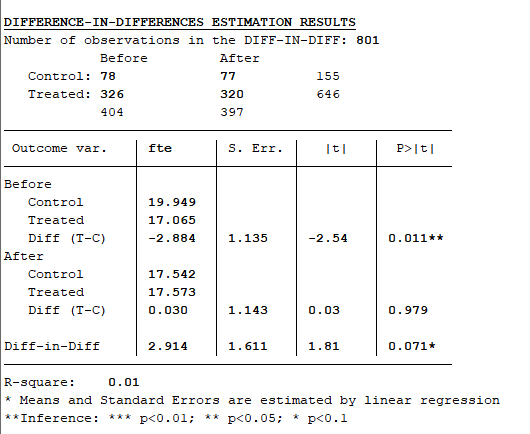
destring roys,replace

destring wendys,replace //引入数据，并转换为stata可识别的数值型数据。

diff fte, t(treated) p(t) //是不是在新泽西州、是否实施了该项政策。

进行没有协变量回归

diff fte, t(treated) p(t)

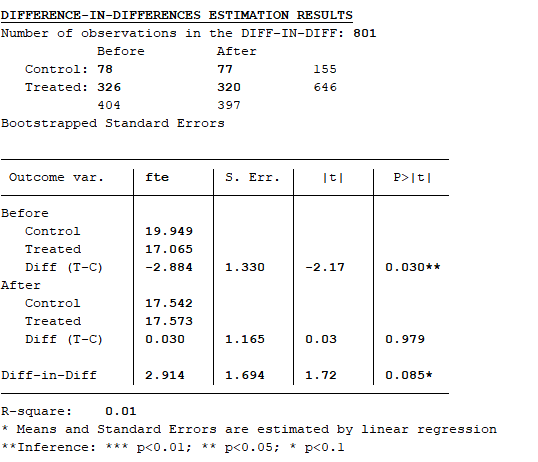


可以看到上面的回归结果中：number of observation in DIFF—IN—DIFF表示实观测的样本数量为801人。接下来就是控制组和处理组在政策实施前和实施后的具体对象个数。

接下来展示了政策实施前控制组和处理组的差异，可以看出，处理除明显的小于控制住，在5%的水平上显著。政策实施后控制组和处理的差异就很微小了。同时也是不显著的。进一步可以看出处理效应的系数为2.914，并且是显著的，进一步，采用自助抽取法进行回归：

diff fte, t(treated) p(t) bs rep(50)

表示采用自助法，同时是自助50次。



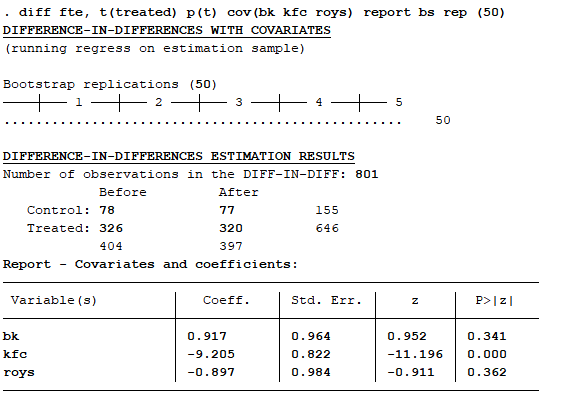
可以看出采用自助法计算出来的结果没有明显差别，处理效应的系数仍是2.914。进一步，我们加入协变量进行分析：

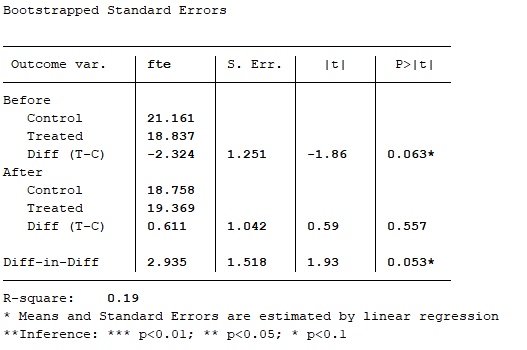
diff fte, t(treated) p(t) cov(bk kfc roys)

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

diff fte, t(treated) p(t) cov(bk kfc roys) report bs rep (50)

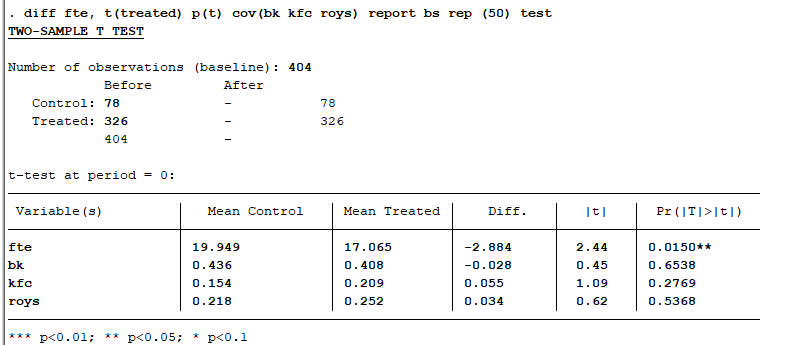
第一行采用不报告协变量系数，第二行报告协变量系数，第三行采用自助法并且报告协变量系数,采用第三个进行回归。





可以看出值有一定变大，并且显著性有了一定提高。进一步检验系数各均值在处理组和控制组是不是相等：

diff fte, t(treated) p(t) cov(bk kfc roys) report bs rep (50) test



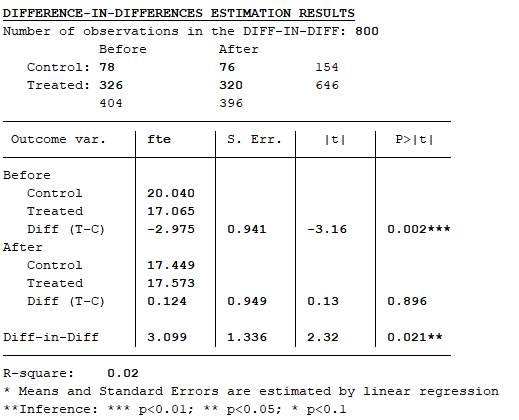
显示fte在两者之间有明显差异，即系数不是相等的，说明其余系数没有明显的差异，基本可以认为系数是相等的。

进一步，执行kernel检验：

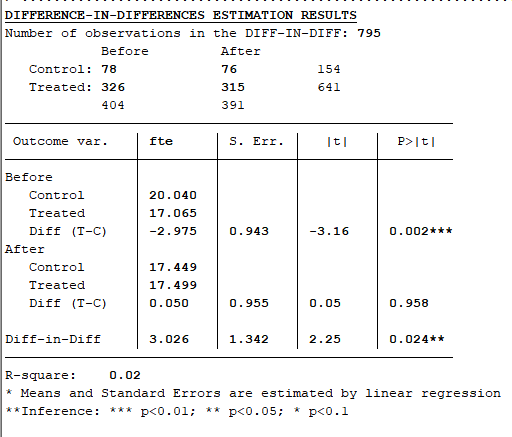
采用核密度估计，并标识每个个体：注意这里严格来说执行的是PSM——DID

diff fte, t(treated) p(t) cov(bk kfc roys) kernel id(id)

注意：采用核密度估计需要用id来确定个体。

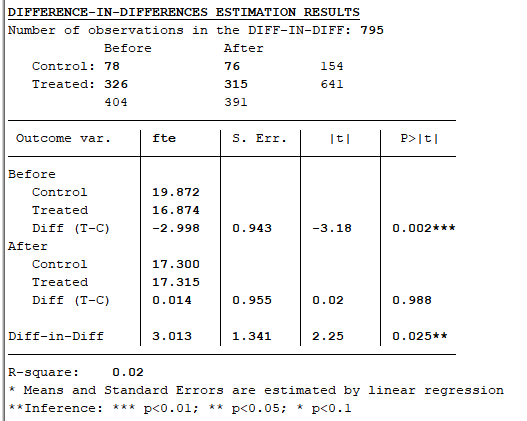


diff fte, t(treated) p(t) cov(bk kfc roys) kernel id(id) support



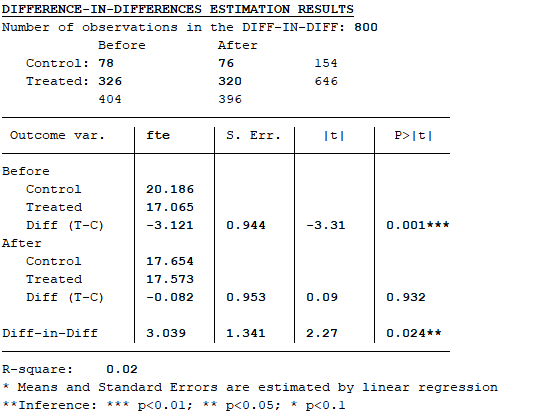
diff fte, t(treated) p(t) cov(bk kfc roys) kernel id(id) support addcov(wendys)

在上面的基础上，加入时间固定效应后的估计，addcov可以加入除了cov声明的协变量之外的其它协变量，也可以是时间固定效应；



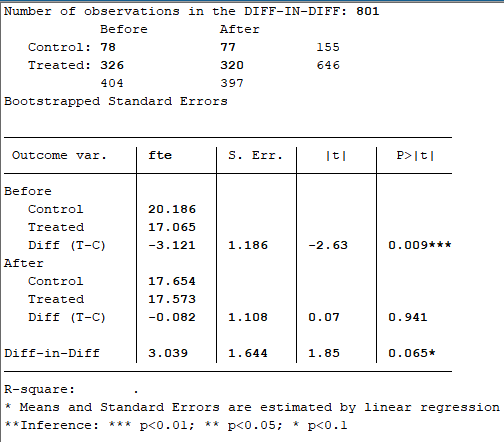
diff fte, t(treated) p(t) kernel id(id) ktype(gaussian) pscore(\_ps) report

变换核密度估计的类型，并汇报



diff fte, t(treated) p(t) kernel id(id) ktype(gaussian) pscore( \_ps) bs reps(50)

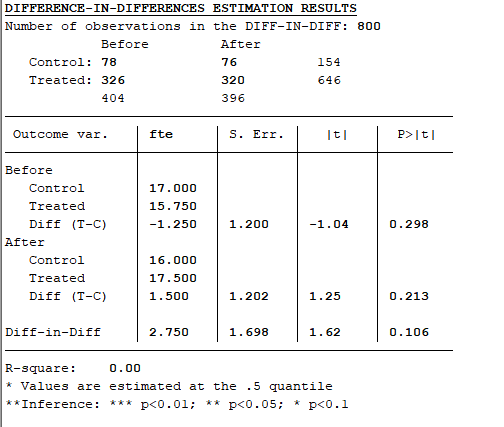
采用自助法汇报相关结果，变化核密度估计类型，并汇报



回归中，考虑到分位数进行回归：

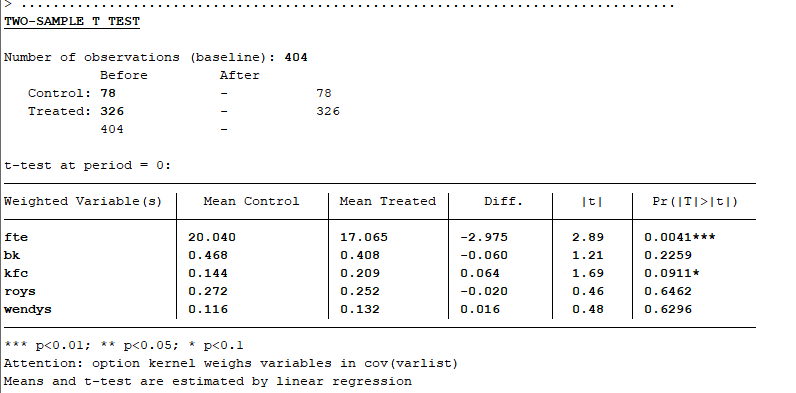
diff fte, t(treated) p(t) qdid(0.50) cov(bk kfc roys) kernel id(id)

这里是使用前50%数据进行回归。



diff fte, t(treated) p(t) cov(bk kfc roys wendys) test id(id) kernel

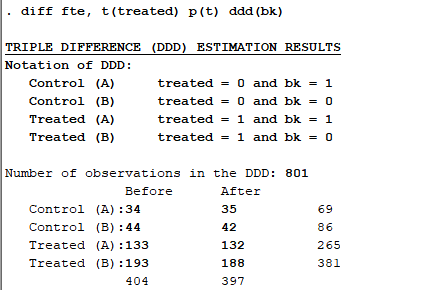
核密度估计，并假设初期的系数是相等的：显示fte和kfc具有明显的系数差异。

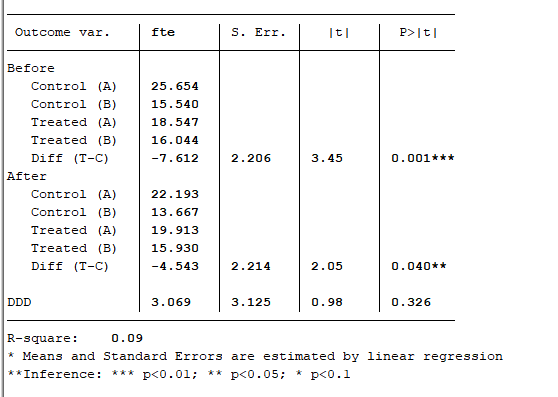


显示有两个初期是有明显差异的：

diff fte, t(treated) p(t) ddd(bk)

采用三重DID进行测度，t表示分组虚拟变量1，ddd表示分组虚拟变量2，p表示分组实施的时间虚拟变量;





更多资料，请关注微信公众号【草莓科研服务网】