更多资料，请关注微信公众号【草莓科研服务网】

双重差分法 | PSM - DID

本文主要是介绍选择偏差及其导致的内生性问题，以及缓解这种内生性问题的倾向得分匹配法（Propensity Score Matching，PSM），并且用一实例介绍一下如何将PSM与DID结合，即PSM-DID在Stata中的具体操作。

一、选择偏差与内生性

1.1 结论可信吗？

例1：进行一项调查，调查内容是去不去医院是否会影响个人健康，因此向医院里的各类人员发放问卷并得出其健康状况，最后发现去医院不利于个人健康。

例2：评估一项污染防治政策的政策效果，选择期初污染程度基本一致的地区作为样本，并**根据各地区意愿决定其是否实施该项政策**，3年后政策实施地的污染指标明显低于未实施该政策的地区，结论是这项政策有效。

以上问卷调查和政策评估的结论可信吗？

例1的问题很明显，去医院的大多是健康状况不佳的患者，因此从逻辑上来说正是人们健康状况不佳才选择去医院，而不是因为他们去了医院才导致健康状况不佳。为了得到更令人信服的结论，应该将调查人群（样本）扩大至所有场合，而不仅仅是医院。与之类似的案例还有，在健身房调查健身对脱发的影响等。

例2的问题不易看出，从DID的角度来看，**“期初污染程度基本一致”说明平行趋势检验基本通过**，因此直观上政策实施地（处理组）的污染指标（结果变量y）与未实施政策地区（控制组）的污染指标的差值就是政策的处理效应，差值为负（假设指标越小污染程度越低）说明政策有效。但是，还应该看到的是，**“根据各地区意愿决定其是否实施该项政策”**，也就是说各地区是否实施该项政策是自己选择的。**实际上，某地是否实施政策更多基于地理区位、经济发展和产业结构等因素的考量。**这也与现实情况吻合，政策一般都在先行示范区率先实施，而这些示范区经济发展水平都较高。

那为什么说例2的政策评估结论不可信呢？

从计量的角度分析，影响某地是否实施政策（即处理组虚拟变量du的取值）的因素也会影响该地区的污染指标（y），这些因素可以分为可观测因素与不可观测因素。

**首先，可观测说明可度量，这些可观测变量与y相关，因此可以将其引入到我们的模型中，虽然可能与du存在共线性，但一定程度的共线性不是问题。问题是我们无法找出所有的可观测变量（或者可观测变量以非线性的形式影响结果变量），而这部分没有引入模型中的可观测变量（或者可观测变量的非线性形式）就被放到扰动项中，造成扰动项与du相关，即存在内生性，最后导致did项的估计系数存在偏误。**

**其次，不可观测说明不可度量，因此这部分不可观测因素必然就在扰动项中，同样会存在内生性问题，造成估计偏误。**

1.2 样本选择性偏差与自选择偏差

例1和例2存在的问题本质上是不同的。

**例1存在的问题是样本的选择不随机**，即我们只获取了医院的样本，而没有随机调查其他场所。**某些影响样本选择过程的因素也会影响到y，因此这些因素被放到扰动项中，造成选择过程与干扰项相关，存在内生性问题，这被称为样本选择性偏差**（Sample Selection Bias）。然而，当我们重新审视样本选择性偏差导致内生性的逻辑时，可以发现所谓的“样本选择过程”并不是一个特定的变量，也没有放到我们的模型中，那为什么会导致解释变量与扰动项相关？Hansen在《ECONOMETRICS（V 2021）》第27章第9小节给出了推导，并且还介绍了一种流行的解决方法——Heckman两步法。

**例2存在的问题是变量的选择不随机**，换言之，各地区对是否实施政策有着自己的“小九九”。du的取值不随机，而是和其他因素相关，而这些因素又被放到扰动项中，造成解释变量与干扰项相关，存在内生性问题，这被称为自选择偏差（Self-Selection Bias）。若是由可观测变量导致的自选择偏差，可以使用PSM方法予以解决，这也是本次推文的主要内容之一。

综合来看，样本选择性偏差和自选择偏差都属于选择偏差（Selection Bias），只是侧重的角度不同，一个侧重的是样本的选择不随机，一个侧重的是变量的选择不随机，但都表明一个观点：**非随机化实验将导致内生性**。**那么问题来了，什么是“非随机化实验”？先从为什么需要随机化实验说起，最后引出PSM - DID。**

1.3 随机分组与依可测变量选择

还是以一个例子来分析。

假设要评估一项实验的效果，也就是考察个体A在接受试验（D=1）后他的结果变量y的变化。最理想也是最不现实的一种评估方法是找出这个人在平行时空中的个体A1，并让A1不接受实验（D=0），相同的一段时间后分别得到结果变量并将其作差，这个差值就是此次实验的**参与者平均处理效应（Average Treatment Effect on the Treated，ATT）**，也就是我们要考察的实验效果。

但是，平行时空中的个体一般是无法找到的（至少现在不能），现实中，我们更多的是用参与实验个体的结果变量与未参与实验个体的结果变量作差（即全部样本的平均处理效应，Average Treatment Effect，ATE），以ATE来代替ATT。**也就是说，当我们将处理组的与控制组的直接作差时，这个差值并不能代表纯粹的政策处理效应，即ATE不等于ATT（即便平行趋势检验通过）。**

而为了使ATE等于ATT，其中一个思路就是消掉选择偏差这一部分，怎么消除？假设条件期望等于无条件期望；怎么实现条件期望等于无条件期望？假设y均值独立于D；怎么实现y均值独立于D？随机分组。

也就是说，**如果处理组的选择是随机的，就能实现处理组虚拟变量与结果变量均值独立，从而推出ATE等于ATT。**换言之，只要处理组的选择是随机的，两组间的之差就是我们需要的参与组平均处理效应ATT。问题是，现实中处理组的选择是非随机的，即存在前文所说的自选择偏差，各个样本做出是否参与实验的决策是一种内生化的行为。那么，如果我们找到决定个体是否参与实验的因素，然后在控制组中匹配到这些因素与处理组相等的样本，最后将这些样本作为我们真正参与评估的控制组，这样是否就能说明处理组的选择是近似随机的呢？答案是可以的，Rosenbaum & Rubin（1983）给出了证明。

简单捋一下上文的逻辑，**估计ATT最理想的方法是找到参与实验的个体在平行时空的自己，并假设平行时空的自己没有参与实验，最后作差得出最纯粹的ATT，但是找到平行时空的自己不现实；退而求其次，我们可以使用随机分组的处理组与控制组，作差得到ATT，但现实中个体是否参与实验的选择不随机；为了得到随机化分组的样本，找出影响个体是否参与实验的因素，控制两组间因素的取值相等，最后利用处理后的分组样本作差得到ATT。**

其中，当决定个体是否参与实验的因素是可观测因素时，个体的决策就是依可观测变量选择；当决定个体是否参与实验的因素是不可观测因素时，个体的决策就是依不可观测变量选择。**由于不可观测因素不可度量，而将控制组样本与处理组样本匹配时需要准确识别出这些因素，因此以下匹配方法都是基于可观测变量（协变量）来设计的。**

1.4 倾向得分匹配

关于怎么匹配又是一个问题。

如果决定个体是否参与实验的可观测变量是一个单一协变量，那么我们只要在控制组中找到与处理组协变量取值相等的样本作为我们的被匹配对象。然而，决定个体决策的协变量并非单一，而是一个由多个协变量构成的多维向量，直接使用进行匹配可能遇到数据稀疏的问题。一种可行的思路是将多维向量进行降维，降维的方法解决数据稀疏问题的同时还保留了足够多的信息。实际使用中的方法主要有以下两种。

**一是使用距离函数，如马氏距离。**

**二是倾向得分匹配PSM。**

由于距离函数不是本文关注的重点，并且距离函数有其固有缺陷，因此这次不做赘述，详情参阅陈强（2014）《高级计量经济学及Stata应用（第二版）》第542页。

PSM主要有以下三个步骤。

**选择协变量。**实际应用中，多数论文直接将基础回归所使用的控制变量作为协变量，这种做法的基本逻辑在于要求协变量与y相关，但问题在于这些控制变量不一定与处理组变量D相关，即便相关，也可能是协变量的高次项或交互项与D相关。**更稳妥的做法是根据相关文献找出同时影响y与D的变量，但相关数据可能很难找到，因此直接使用控制变量作为协变量其实是一种妥协，但还是要仔细甄别每一个控制变量是否在理论上真正影响到D，并且也要做协变量高次项或交互性的敏感性分析。**

**估计倾向得分值。**D作为被解释变量，协变量作为解释变量，由于D是一个二元哑变量，因此使用logit模型或probit模型，应用中更多使用logit模型。logit回归之后会得到各个样本的倾向得分值（在psmatch2中将生成\_pscore变量，\_pscore介于之间），之后就是根据\_pscore进行第三步的匹配。

**根据\_pscore进行匹配。**最常用的是**卡尺最近邻匹配**，其他匹配方法参阅陈强（2014）《高级计量经济学及Stata应用（第二版）》第545页。卡尺最近邻匹配基于最近邻匹配，最近邻匹配需要设置邻居数（neighbor(#1)），也就是说根据倾向得分值最接近的原则，每一个处理组样本匹配到#1个控制组样本，但如果控制组样本的倾向得分值与处理组样本差距太大，这样的匹配是没有多大意义的，**因此需要还设置一个卡尺（caliper(#2)），即控制一个阈值界限，界限内的控制组样本可以作为匹配对象，界限外的样本则被忽视**。

1.5 PSM + DID

PSM和DID是天生绝配！

**因为现实中的政策本质上是一种非随机化实验（或称，准自然实验），因此政策效应评估所使用的DID方法难免存在自选择偏差，而使用PSM方法可以为每一个处理组样本匹配到特定的控制组样本，使得准自然实验近似随机，注意是近似，因为影响决策的不可观测因素在两组间仍然存在差异。**

如，石大千等（2018）为了使实验组和控制组城市在各方面特征上尽可能地相似，消除选择偏差，选择PSM - DID方法以便更准确地评估智慧城市建设降低环境污染的效应。王雄元和卜落凡（2019）认为开通“中欧班列”的枢纽城市自身交通基础设施较好，且上市公司分布存在区域集聚现象，为了消除选择偏差采用PSM - DID方法进行稳健性检验。**丁宁等（2020）考虑到绿色信贷和银行成本效率之间可能存在内生性，为了解决这个问题，采用了PSM - DID方法。**郭晔和房芳（2021）为了缓解实验组和对照组企业之间的其他差异对研究结果的干扰，同样采用了PSM - DID方法。其他采用PSM - DID方法缓解因选择偏差导致的内生性问题的文献还有孙琳琳等（2020）、陆菁等（2021）和余东升等（2021)等。

然而，PSM - DID也并非是解决选择偏差的灵丹妙药，除了PSM本身不能控制因不可观测因素导致的组间差异，在与DID结合时还存在一个更为关键的问题。

**从本质上来说，PSM适用于截面数据，而DID仅仅适用于时间-截面的面板数据。**

对于PSM，每一个处理组样本匹配到的都是同一个时点的控制组样本，相应得到的ATT仅仅是同一个时点上的ATT。下文psmatch2的输出结果中，ATT那一行结果就仅仅代表同一个时点上的参与者平均处理效应。

**对于DID，由于同时从时间与截面两个维度进行差分，所以DID本身适用的条件就是面板数据。因此，由PSM匹配到的样本原本并不能直接用到DID中做回归。**

面对两者适用数据类型的不同，现阶段的文献大致有两种解决思路。

**第一，将面板数据视为截面数据再匹配。**如上文参考文献中的绝大多数。

**第二，逐期匹配。**如，Heyman et al.（2007）、Bockerman & Ilmakunnas（2009）等。

而，谢申祥等（2021）指出了这两种方法的不足。

**第一，将面板数据转化为截面数据进行处理存在“自匹配”问题。**

**第二，逐期匹配将导致匹配对象在政策前后不一致。**

虽然传统的研究思路存在一定程度的不足，但本次推送还是按照传统的设计思路介绍一下PSM - DID如何在Stata中实现。

具体来说，本次推送将**使用李青原和章尹赛楠（2021）公布在《中国工业经济》官网的原始数据，首先，分别采用截面PSM与逐年PSM两种方法获得匹配样本，之后将匹配样本进行DID回归，并比较匹配后的回归结果与匹配前的基准回归结果**，进而验证结论的稳健性。

**事实上，这篇文章也注意到了选择偏差的问题，但却是采用自助法（Bootstrapping）进行重复随机抽样**（**抽取500次，每次抽取1,000个样本），然后将抽取到的样本进行回归，最后重复随机抽样的结果与基准回归结果无实质性差异，从而证明了结论的稳健性。因此，本次推送将是从PSM的角度对基准回归结果进行复盘和二次验证。**

二、PSM - DID的实现

**2.1 数据初步处理**

在PSM和DID之前先定义路径、设置图片输出格式（Stata自带的图片主题太丑）、定义控制变量（协变量）和回归命令选择项的全局暂元以及生成处理组虚拟变量，然后保存好初步处理的原始数据。

其中，处理组虚拟变量原始数据集中是没有的，因为多期DID直接结合时间虚拟变量与分组虚拟变量构成did项（数据集中的变量名为FB）。由于官网放出的原始代码给出了三段政策时间节点，以及各个时点政策实施地区的城市代码或省份代码，我们根据这个信息可以生成处理组虚拟变量treated。

\*定义路径

cd "C:\Users\林豸\Desktop\计量代码二次整理\01 DID\05 PSM-DID"

\*设置图片输出样式

graph set window fontface "Times New Roman"

graph set window fontfacesans "宋体"

set scheme s1color

use "C:\Users\林豸\Desktop\计量代码二次整理\01 DID\05 PSM-DID\行业数据.dta"

\*定义全局暂元

全局元定义的目的是为了方便省事，可以不同频繁的把所需要的变量一个个敲出来。

global xlist "ADM PPE ADV RD HHI INDSIZE NFIRMS FCFIRM MARGIN LEVDISP SIZEDISP ENTRYR EXITR"

global regopt "absorb(city ind3) cluster(city#ind3 city#year) keepsing"

\*生成处理组虚拟变量

gen treated = ( city == 5101 | city == 5000 | city == 2102 | city == 3501| city == 4401 | city == 3701 | city == 3201 | city == 3702 | city == 3101 | city == 4403 | city == 1200 | city == 4201| city == 4404 | prov == 44 | prov == 45 | prov == 43| prov == 32 | prov == 33 | city == 1100 | city == 5301| city == 2101 | city == 3502 | city == 6101 | city == 2201 | city == 2301 | city == 6201 | city == 6401 )

save psmdata.dta, replace

**2.2 截面PSM - DID**

由于psmatch2为外部命令，因此第一次使用该命令时需要键入如下代码进行安装。

ssc install psmatch2, replace

在使用psmatch2命令前，首先设置了一个随机种子并将样本进行随机排序，这么做的原因是在匹配控制组样本时，如果有几个样本的倾向得分值相同，系统会优先选择排序靠前的样本。当然，不进行随机排序直接使用原始数据集的默认样本顺序问题也不大。

\*# 一、截面匹配

use psmdata.dta, clear

\*\*# 1.1 卡尺最近邻匹配（1:2）

set seed 0000

gen norvar\_1 = rnormal()

sort norvar\_1

上面就是生成随机序列数，然后对随机序列数进行排序。

psmatch2 treated $xlist , outcome(TFPQD\_OP) logit neighbor(2) ties common ate caliper(0.05)

save csdata.dta, replace

treated是本例中的分组变量。

$xlist是协变量的全局暂元，纯粹为了演示，这里直接使用基础回归中的控制变量，实际使用中一定要结合其他文献及敏感性分析挑选出适合的协变量。

outcome(TFPQD\_OP)括号里面填的是结果变量（这里是TFPQD\_OP），即基准回归中的被解释变量y，括号里可以填入多个结果变量。

logit表明使用logit模型估计倾向得分值，默认使用probit模型。

neighbor(2)表明使用最近邻匹配，括号里填入2说明使用1:2的匹配方法，即一个处理组样本最多匹配2个控制组样本。至于为什么是2？因为使用1:1的配对方式最后的结果不好看。

ties说明如果多个控制组样本的倾向得分值相同并且处于这个得分值的样本应该作为被匹配对象，那么匹配样本的结果变量取它们的均值，默认按照样本的排列顺序进行选择。这同时说明，如果加上ties，那么数据集是否提前排序就不怎么重要了。

common表示只对倾向得分值共同取值范围内的样本进行匹配。也就是说，两组样本的倾向得分值如果不在共同取值范围内，则直接被排除，从一开始就没有匹配资格。默认对所有样本进行匹配。

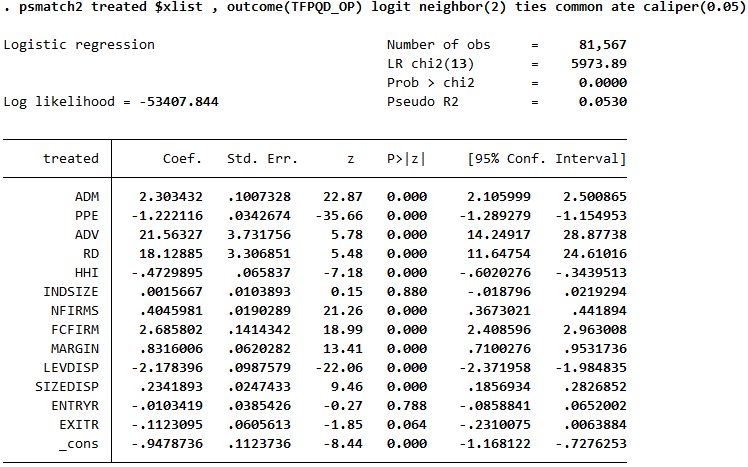
**ate表示同时汇报ATE、ATU和ATT，默认只汇报ATT。**ATE和ATT前文有介绍，ATU是非参与者的平均处理效应（Average Treatment Effect on the Untreated），我们重点关注的是ATT。

caliper(0.05)表示卡尺设置为0.05，即控制组样本的倾向得分值如果在处理组样本倾向得分值±0.05以外，直接没有被匹配的资格。因此，如果选择项同时包括neighbor(#)和caliper(#)，则说明使用卡尺最近邻匹配。至于为什么是0.05？道理同neighbor(2)。

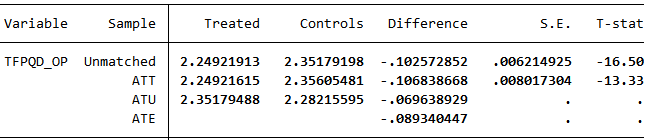
关于psmatch2其他选择项的详细说明，请键入如下代码进行了解。

help psmatch2

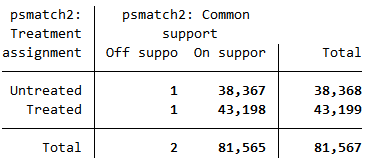
psmatch2的运行结果包括三部分的内容。



第一部分是logit回归结果，在这里没有多大意义。



**第二部分则重点关注ATT的大小与显著性。**在未匹配前（Unmatched那一行），平均处理效应为-0.1026，且在1%的水平下显著（t值的绝对值大于2.58）；匹配后，参与者的平均处理效应ATT为-0.1068，同样在1%的水平下显著。结果虽好，但不能开心太早，因为这里的ATT仅仅是参与者实施政策对结果变量的影响，双重差分只在截面维度进行了差分，时间维度并未考虑，换言之，这不是我们需要的DID的处理效应。反之，就算ATT不显著关系也不大，因为我们PSM的目的是获得匹配样本，真正的DID回归在之后。



第三部分考察两组在共同取值范围内（On support那一列）的样本量。可以看到，处理组和控制组都分别只有一个样本在共同取值范围外，这说明绝大多数样本（特别是控制组样本）都有资格参与匹配。

接下来是平衡性检验（Balance Test）。即检验匹配后协变量取值在两组间是否存在显著差异，如果差异不明显，则说明匹配效果好，使用这样的匹配样本进行DID回归就比较合适。平衡性检验主要有两种。

**一是度量两组间协变量的标准化均值的偏差%bias。如果匹配后协变量的%bias小于10%，且明显小于未匹配前的%bias，则说明对于这个协变量来说，两组间并无差距。**

**二是通过t检验来判断各个协变量的取值在两组间是否存在系统性偏差。t检验的原假设（H0）是“两组间协变量的取值不存在系统性偏差”，因此我们的目标是最终接受H0。也就是要求检验结果不显著。**

psmatch2自带两个估计后检验命令，一个是pstest，用于进行平衡性检验；一个是psgraph，用柱状图的方式直观呈现出两组间满足共同支撑假设（即倾向得分值在共同取值范围内）样本的分布情况。

\*\*# 1.2 平衡性检验

pstest, both graph saving(balancing\_assumption, replace)

graph export "balancing\_assumption.emf", replace

psgraph, saving(common\_support, replace)

graph export "common\_support.emf", replace

both表示同时显示匹配前后两组间各变量的平衡情况。

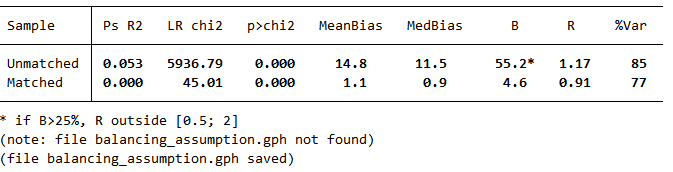
graph表示图示各变量匹配前后的平衡情况。

saving表示将图形保存为Stata可以识别的\*.gph格式，方便后续修改。

graph export表示将图形输出为\*.emf矢量图，方便插入到论文中。



第一张表是我们关注的重点。可以看到，所有协变量匹配后的%bias均小于10%，且都明显小于匹配前的%bias，%bias的绝对值较匹配前大幅下降了56.1%~98.7%；除FCFIRM、MARGIN和LEVDISP这三个协变量外，其余协变量均拒绝“两组间协变量的取值不存在系统性偏差”的原假设。



第二张表是匹配前后logit回归的情况，其中Unmatched那一行是匹配前的回归结果，与psmatch2运行后第一张表保持一致，Matched那一行是匹配后的logit回归情况。**可以看到，匹配后回归结果中伪R方（Ps R2）明显变小，这说明匹配后两组间的所有协变量取值差异性都不大，从而对logit回归中的被解释变量的变动没有太大的解释力。**



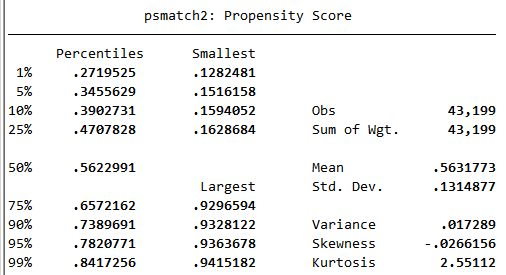
图 1和第一张表的结果保持一致，即所有协变量的%bias均小于10%，并且都明显小于匹配前的%bias。

psgraph的运行结果如图 2。同样，图 2和前文psmatch2运行结果中的第三张表保持一致，即处理组和控制组的绝大多数样本都在共同取值范围内，而不在共同取值范围内的样本的倾向得分值比较极端（从图片中暂时无法看出是极端大还是极端小）。



**然后我们还可以考察一下两组倾向得分值在匹配前后是否存在差异，这里可以使用核密度图来直观体现。如果匹配前两组间的核密度曲线偏差比较大，而匹配后核密度曲线比较接近，说明匹配效果好。**

\*\*# 1.3 倾向得分值的核密度图

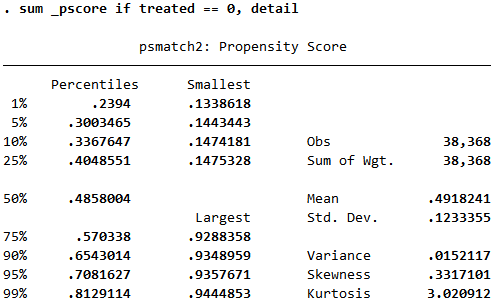


sum \_pscore if treated == 1, detail // 处理组的倾向得分均值为0.5632

\*匹配前

sum \_pscore if treated == 0, detail

匹配前控制组的倾向匹配得分均值为0.4918



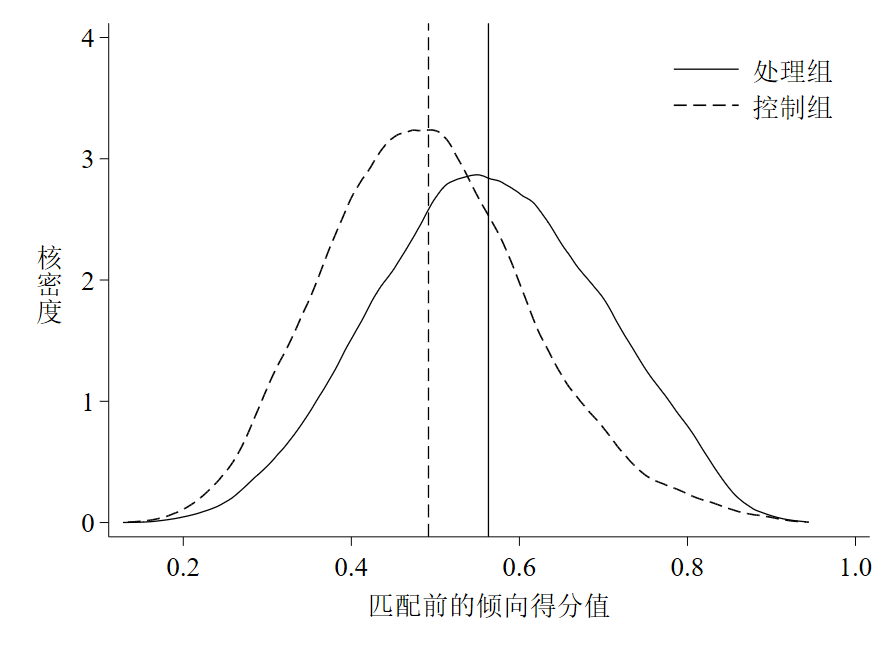
匹配前

sum \_pscore if treated == 0, detail

twoway(kdensity \_pscore if treated ==1, lpattern(solid) lcolor(black) lwidth(thin) scheme(qleanmono) ytitle("{stSans:核}""{stSans:密}""{stSans:度}",size(medlarge) orientation(h)) xtitle("{stSans:匹配前的倾向得分值}",size(medlarge)) xline(0.5632, lpattern(solid) lcolor(black)) xline(`r(mean)', lpattern(dash) lcolor(black)) saving(kensity\_cs\_before, replace)) (kdensity \_pscore if treated == 0, lpattern(dash)),xlabel( , labsize(medlarge) format(%02.1f)) ylabel(0(1)4, labsize(medlarge)) legend(label(1"{stSans:处理组}") label(2"{stSans:控制组}") size(medlarge) position(1) symxsize(10))

graph export "kensity\_cs\_before.emf", replace

discard

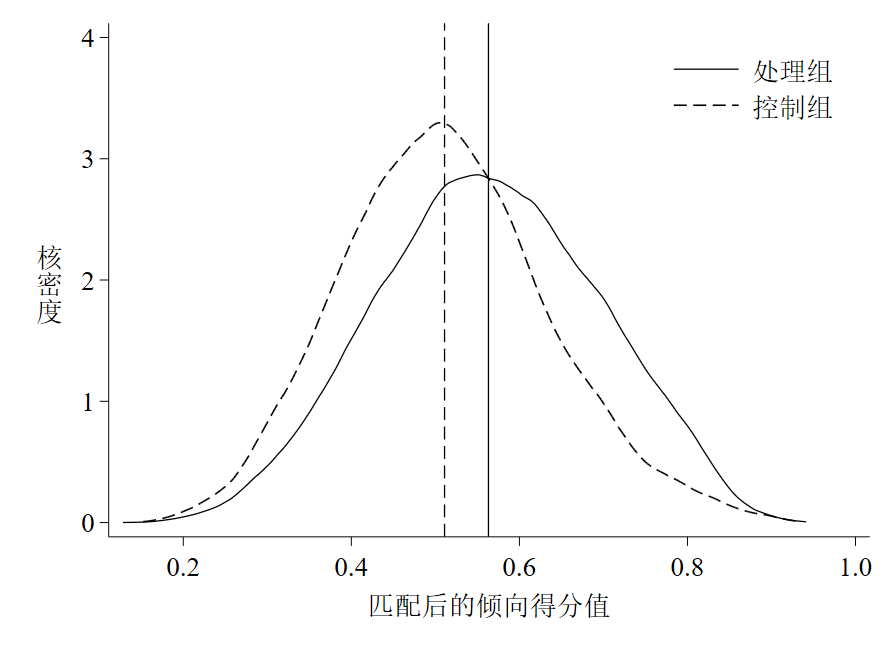


sum \_pscore if treated == 0 & \_weight != ., detail

twoway(kdensity \_pscore if treated == 1, lpattern(solid) lcolor(black) lwidth(thin) scheme(qleanmono) ytitle("{stSans:核}""{stSans:密}""{stSans:度}", size(medlarge) orientation(h)) xtitle("{stSans:匹配后的倾向得分值}", size(medlarge)) xline(0.5632, lpattern(solid) lcolor(black)) xline(`r(mean)', lpattern(dash) lcolor(black)) saving(kensity\_cs\_after, replace)) (kdensity \_pscore if treated == 0 & \_weight != ., lpattern(dash)), xlabel( , labsize(medlarge) format(%02.1f)) ylabel(0(1)4, labsize(medlarge)) legend(label(1"{stSans:处理组}") label(2"{stSans:控制组}") size(medlarge) position(1) symxsize(10))

graph export "kensity\_cs\_after.emf", replace

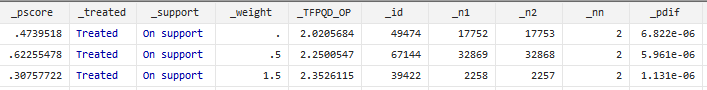
discard



可以看到，匹配前后两条核密度曲线偏差都较大，但匹配后两条曲线更为接近了，这点可以从均值距离缩小看出（图中垂直于横轴的实线是处理组样本倾向得分值的均值线，虚线是控制组的均值线）。因此，一定程度上可以说明我们的匹配还是有效果的（虽然效果不太明显）。

最后就是利用PSM后的样本进行DID回归，并将回归结果与基准回归结果进行比较。回归需要样本，那么如何获取匹配后的样本呢？

打开Stata的数据编辑器，可以观察到psmatch2在我们的数据集中生成了几个新变量，如下图。



\_pscore是样本的倾向得分值。

\_treated是处理组虚拟变量，和我们手动生成的treated保持一致。

\_support是样本是否满足共同支撑假设的虚拟变量，即如果样本处于共同取值范围内，该变量取值为1，否则为0。

\_weight是一个关键变量。

对于处理组样本来说，该变量为空说明没有参与匹配（即未满足共同支撑假设），或者参与匹配了但在卡尺范围内暂未匹配到控制组样本；如果匹配成功了，该变量取值为1。

对于控制组样本来说，该变量为空说明样本不满足共同支撑假设或在卡尺范围内没有匹配成功；对于匹配成功的样本，该变量可能取整也可能不取整，这取决于样本匹配成功的次数。具体来说，如果是1:1无放回匹配，则每个处理组样本最多匹配成功一次，并且控制组样本也最多被匹配一次，因此对于控制组中匹配成功的样本来说，\_weight只可能等于1；如果是1:1有放回匹配，则每个处理组样本最多匹配成功一次，而控制组样本可能被匹配k次，因此对于控制组中匹配成功的样本来说，\_weight的取值为大于等于1的整数；如果是1:m的有放回匹配（psmatch2只允许在1:1匹配中设置无放回选项），则一个处理组样本最多可以匹配到m个对象，因此被匹配到的对象所占的权重为1/m，那么对于控制组样本来说，如果它被匹配到了k次，那么它的总权重就是k/m。如果我们想要知道在1:m有放回匹配中，控制组样本被匹配的次数k，那么就只要将它们的\_weight乘以m即可。因此，\_weight的取值表明了样本是否参与匹配以及样本的重要性程度。

\_TFPQD\_OP代表样本匹配到对象的结果变量（这里是TFPQD\_OP）的均值。

\_id是psmatch2自动为每一个样本赋予的唯一识别编码。

\_n1和n2代表样本匹配对象在\_id中的编码，由于这里是1:2匹配，所以有两个。

\_nn表示处理组样本匹配到的控制组样本的个数。

\_pdif表示样本的倾向得分值与\_n1样本倾向得分值之差的绝对值。

由此可知，只要样本的\_weight不为空，就代表该样本参与了匹配，就可以将这些样本代入到DID回归模型中进行参数估计，而且一定程度上缓解了基准回归中存在的选择偏差问题，进一步还可以与基准回归结果进行比较以检验其稳健性。事实上，多数文献都是按照这个思路进行的，当然，也有部分文献直接选择满足共同支撑假说的样本代入到DID回归模型中，如李青原和肖泽华（2020）。但是，这样可能忽视了一个关键问题。无论是使用权重不为空的样本还是使用满足共同支撑假设的样本进行回归，都忽视了一个基本事实，那就是被匹配的控制组样本可能作为多个处理组样本的匹配对象，因此不同权重的控制组样本在总体的控制组样本中的重要性程度是不一样的。权重越大，说明被匹配上的次数越多，参与回归时越应该被重视，因此一种可行的办法是根据权重来复制控制组中被匹配上的样本，也就是所谓的频数加权回归。

所以，在进行回归结果对比时，除了需要包括两个基准回归、一个使用权重不为空的样本进行的回归和一个使用使用满足共同支撑假设的样本进行的回归，还需要包括一个考虑样本重要性的频数加权回归。

\*\*# 1.4 回归结果对比

use csdata.dta, clear

\*- 基准回归1（混合OLS）

qui: reg TFPQD\_OP FB $xlist , cluster(city)

est store m1

\*- 基准回归2（固定效应模型）

qui: reghdfe TFPQD\_OP FB $xlist , $regopt

est store m2

\*- PSM-DID1（使用权重不为空的样本）

qui: reghdfe TFPQD\_OP FB $xlist if \_weight != ., $regopt

est store m3

\*- PSM-DID2（使用满足共同支撑假设的样本）

qui: reghdfe TFPQD\_OP FB $xlist if \_support == 1, $regopt

est store m4

\*- PSM-DID3（使用频数加权回归）

gen weight = \_weight \* 2

计算权重，假设每个样本都被匹配了两次，这样的权重为 \_weight \* 2

replace weight = 1if treated == 1 & \_weight != .

处理组只能匹配1次，因此需要将处理组前权重不为0的样本处理权重为1。

qui: reghdfe TFPQD\_OP FB $xlist [fweight = weight], $regopt

est store m5

local mlist\_1 "m1 m2 m3 m4 m5"

下面是将回归结果输送到word里面，以便于进行比较。

reg2docx `mlist\_1' using 截面匹配回归结果对比.docx, b(%6.4f) t(%6.4f) scalars(N r2\_a(%6.4f)) noconstant replace mtitles("OLS""FE""Weight!=.""On\_Support""Weight\_Reg") title("基准回归及截面PSM-DID结果")

基准回归及截面PSM-DID结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) |
|  | OLS | FE | Weight!=. | On\_Support | Weight\_Reg |
| FB | -0.1243\*\*\* | -0.0496\*\*\* | -0.0545\*\*\* | -0.0497\*\*\* | -0.0422\*\*\* |
|  | (-5.2480) | (-3.6531) | (-3.9823) | (-3.6550) | (-2.9228) |
| ADM | 1.0682\*\*\* | 0.8719\*\*\* | 0.9287\*\*\* | 0.8719\*\*\* | 0.9839\*\*\* |
|  | (17.6388) | (18.0260) | (16.1842) | (18.0262) | (12.0535) |
| PPE | -0.1183\*\*\* | -0.0845\*\*\* | -0.0784\*\*\* | -0.0845\*\*\* | -0.0813\*\*\* |
|  | (-4.3488) | (-5.4869) | (-4.3384) | (-5.4864) | (-3.6168) |
| ADV | 3.4239\* | 0.4029 | 1.6045 | 0.3975 | 2.3618 |
|  | (1.9284) | (0.2547) | (0.8734) | (0.2513) | (0.9525) |
| RD | -1.1993 | -3.5116\*\*\* | -3.4516\*\* | -3.4852\*\*\* | -4.2360\* |
|  | (-0.8120) | (-2.6572) | (-2.1919) | (-2.6371) | (-1.8977) |
| HHI | -0.0284 | 0.1391\*\*\* | 0.1347\*\*\* | 0.1393\*\*\* | 0.0702 |
|  | (-0.5123) | (3.7371) | (3.2368) | (3.7409) | (1.3515) |
| INDSIZE | 0.1061\*\*\* | 0.0398\*\*\* | 0.0425\*\*\* | 0.0398\*\*\* | 0.0541\*\*\* |
|  | (6.4798) | (5.1231) | (5.1566) | (5.1227) | (5.1725) |
| NFIRMS | -0.2011\*\*\* | -0.1012\*\*\* | -0.1048\*\*\* | -0.1012\*\*\* | -0.1324\*\*\* |
|  | (-7.0272) | (-8.2637) | (-7.9029) | (-8.2608) | (-7.5215) |
| FCFIRM | 0.2561\*\* | -0.0343 | 0.0229 | -0.0335 | 0.0774 |
|  | (2.5595) | (-0.5604) | (0.2952) | (-0.5472) | (0.6958) |
| MARGIN | 0.2333\*\*\* | 0.2377\*\*\* | 0.2472\*\*\* | 0.2377\*\*\* | 0.2808\*\*\* |
|  | (6.6555) | (8.2641) | (7.5753) | (8.2641) | (6.3765) |
| LEVDISP | 1.8228\*\*\* | 1.1270\*\*\* | 1.1118\*\*\* | 1.1271\*\*\* | 1.0811\*\*\* |
|  | (15.3788) | (20.1912) | (17.9101) | (20.1920) | (13.5927) |
| SIZEDISP | 0.1373\*\*\* | 0.0678\*\*\* | 0.0667\*\*\* | 0.0678\*\*\* | 0.0785\*\*\* |
|  | (5.2810) | (4.5861) | (4.0912) | (4.5819) | (3.8797) |
| ENTRYR | 0.0735\*\*\* | 0.0732\*\*\* | 0.0573\*\* | 0.0732\*\*\* | 0.0684\*\*\* |
|  | (2.9277) | (3.2806) | (2.4976) | (3.2805) | (2.6353) |
| EXITR | 0.0560\* | 0.0491\* | 0.0420 | 0.0492\* | -0.0326 |
|  | (1.7687) | (1.7444) | (1.3206) | (1.7474) | (-0.8330) |
| N | 81567 | 81567 | 60078 | 81565 | 116882 |
| r2\_a | 0.0646 | 0.1613 | 0.1548 | 0.1613 | 0.1592 |

其中，频数加权回归除了可以使用Stata自带的fweight方法，还可以考虑直接将样本复制进而回归，这种思路来源于黄河泉（2017）在经管之家论坛的回答，最终结果与fweight方法完全一致。

gen weight = \_weight \* 2

replace weight = 1if treated == 1 & \_weight != .

keep if weight != .

expand weight

reghdfe TFPQD\_OP FB $xlist , $regopt

est store m5

可以看到，核心解释变量外资银行进入（FB）在五个回归中均显著为负，且第(3)(4)(5)列FB系数值大小与第(2)列相差不大，其余控制变量的系数也符合预期，这说明当考虑到选择偏差问题后，基准回归结果依旧稳健。

此外，观察各回归中实际参与回归的样本数N可以发现，因为第(3)列使用的是权重不为空的样本，即PSM匹配成功的样本，因此样本量相比基础回归有所减少，且减少幅度较大；第(3)列使用的是满足共同支撑假设的样本，由于只有两个样本不满足假设，因此参与回归的样本只比基础回归少两个；第(4)列由于是根据权重进行的频数加权回归，实际参与回归的样本会根据权重进行复制，因此最终有116,882个样本参与回归。

更多资料，请关注微信公众号【草莓科研服务网】