doi:10.19920/j.cnki.jmsc.2024.02.003

PSM-DID 在政策评价中的应用现状与改进方法①

蔡 俊[1], 杨 岚[2*], 周亚虹[3]

(1. 华中科技大学管理学院, 武汉 430074; 2. 西南财经大学统计学院, 成都 611130;
3. 上海财经大学经济学院及上海财经大学滴水湖高级金融学院, 上海 200433)

摘要: 倾向得分匹配 – 双重差分模型(PSM-DID)是政策评估及因果推断中最为流行的方法之一. 但是在实际应用中,该方法面临着控制变量在处理组样本和控制组样本之间非平衡性的挑战. 传统基于均值差异 t 检验的平衡性检验容易产生片面和误导性的结论,使得后续因果推断产生偏误. 为克服上述问题,本文对传统的平衡性检验提出以下改进:一是推荐更全面的多维度的平衡性测度指标,便于在匹配后更严谨地比较处理组和控制组的平衡性;二是提出了适用于非平衡样本的新估计方法:倾向得分匹配 – 逆概率加权 – 双重差分(PSM-IPW-DID),该方法结合了倾向得分匹配(PSM)克服样本自选择内生性及对非平衡样本稳健的优势和逆概率加权(inverse probability weighting, IPW)利用全样本信息的长处,在不进一步删除样本的情况下得到一种更稳健的双重差分估计方法. 数据模拟和应用实例显示,本文提出的新方法能更全面、客观地评价宏观、微观政策的作用,得到更为可信的因果推断.

关键词: 倾向得分匹配 – 双重差分; 平衡性; 逆概率加权; 双重稳健性

中图分类号: F064.1 **文献标识码**: A **文章编号**: 1007 – 9807(2024)02 – 0030 – 19

0 引 言

最近十几年来,倾向得分匹配 – 双重差分(propensity score matching-difference in differences, PSM-DID)在政策评价因果效应识别和估计中得到了越来越多的关注. 倾向得分匹配的基本原理是,将原本基于多维控制变量的处理组和控制组的匹配转变为基于一维倾向得分的匹配,从而使匹配维度大大降低,而匹配质量和效果却得到显著提升. 基于匹配后的样本进行分析,能够克服样本自选择偏差,从而使得因果估计更为准确. 而双重差分则可以通过控制时间和个体两个维度的不可观测异质性,在反事实的框架下来评估政策发生和不发生这两种情况下被观测结果的变化,得到因果推断.

本文统计和梳理了自 2012 年—2022 年间国内部分经济学管理学权威期刊,即《经济研究》、《管理世界》、《管理科学学报》、《经济学(季刊)》、《世界经济》、《中国工业经济》中使用了倾向得分匹配的文章. 在这十年间,共有 169 篇文章使用了倾向得分匹配方法,其中 118 篇将其作为主回归或者主回归之一,51 篇使用该方法进行稳健性检验. 通过文献梳理,本文发现:

第一,倾向得分匹配在经济管理各个领域的应用都较为广泛,特别是在政策评估时成为一种重要的因果效应估计方法而广为人知. 由于通常能获取到的数据大多为观测数据,利用观测数据进行政策评价时首先要解决的问题就是样本的自选择性,而倾向得分匹配能够在利用已有数据的

① 收稿日期: 2022 – 08 – 22; 修订日期: 2023 – 05 – 16.
基金项目: 国家自然科学基金资助重点项目(71833004); 国家自然科学基金资助专项项目(72342034); 国家自然科学基金资助项目(72173083); 国家自然科学基金资助青年项目(72303074); 四川省哲学社会科学基金资助青年项目(SCJJ23ND428).
通讯作者: 杨 岚(1993—), 女, 四川广元人, 博士, 副教授. Email: yanglan@swufe.edu.cn

基础上较好的处理样本的自选择性. 例如：孙亮等[1]采用 PSM-DID 方法，系统考察了我国资本市场中政府赋予型声誉的激励效果和作用机理. 黄俊等[2]以第一巡回法庭和第二巡回法庭管辖范围内的上市公司为实验组，采用 PSM 方法构建对照组样本，探究巡回法庭的设立对企业投资的影响. 吴要武[3]基于 PSM 方法估计跨省迁移者相比于省内迁移者真实的收入优势. 贾俊雪和秦聪[4]利用 2 126 个村庄的调查数据进行倾向得分匹配实证检验，识别专业协会建立对处理组农户人均纯收入的平均处理效应.

第二，倾向得分匹配方法常与双重差分法结合使用. 由于原理和模型设置容易理解和运用，双重差分法成为政策效应评估方法中的最流行的方法之一. 十年间，使用倾向得分匹配 – 双重差分法(PSM-DID)进行反事实估计的文章有 107 篇，占到发文总数的 63.3%(其中宏观政策效应评估有 22 篇，占 20.6%；微观政策效应评估有 85 篇，占 79.4%). 例如：宏观政策层面的分析中，万海远和李实[5]采用倾向得分匹配与双重差分的方法来构造反事实，从而在拟实验环境下评估户籍歧视对城乡收入差距产生的影响. 王庶和岳希明[6]使用 PSM-DID 评估退耕还林在农民增收、非农就业和扶贫开发等方面的政策效果. 龙玉等[7]利用 PSM-DID 模型考察在高铁开通前后的高铁沿线各城市风险投资项目和投资金额的变化，来检验高铁对风险投资区域特征的影响. 微观政策层面的分析中，孙文凯和王乙杰[8]采用 PSM-DID 方法估计父母外出务工对留守儿童自评健康的影响. 张杰等[9]利用 PSM-DID 模型系统地检验了出口与生产率的关系. 王桂军和张辉[10]使用 PSM-DID 评估了“一带一路”倡议对中国 OFDI 企业全要素生产率的影响.

第三，倾向得分匹配作为一种数据预处理方法，能让处理组和对照组的可观测特征尽可能接近，从而克服样本自选择带来的估计偏误，但是使用倾向匹配得分需要满足两个前提：一是平衡性假定，在计算倾向得分后，需要评估匹配的质量如何，即检验可观测特征是否均衡；二是共同支撑假设，也即评估控制组和处理组倾向得分的分布，若两组样本没有重合的倾向得分，或者重合的样本量太小，就会导致无法匹配或匹配偏差较大. 在梳理文献时，本文发现仅有 82 篇文章(占比 48.5%)在正文中汇报了平衡性假定检验的结果，有 29 篇文章(占比 17.2%)未汇报具体检验统计指标，仅指出通过该检验或者检验结果备索，有 58 篇文章(占比 34.3%)未进行平衡性检验. 同时，本文还发现，十年间文献进行平衡性检验基本都是基于比较处理组和对照组样本均值是否存在统计意义上的差异，若两组的均值差的 t 统计量显示实验组和对照组均值没有显著差异，则认为通过平衡性检验，使用倾向得分匹配后得到的样本较为均衡. 但是，仅仅基于均值的差异判断两组样本的平衡性较为片面，一种可能的情况是虽然均值较为接近，但是方差差异很大，例如处理组的方差较小，控制组的方差较大，此时，将两组进行比较得出的结论有较大偏差. 因此，本文认为目前文献中广泛使用的平衡性检验方法存在不足，需要进行更多维度平衡性检验指标的检测，只有确保处理组和对照组的样本实现平衡后，基于匹配后的样本所估计的因果效应才有意义[11, 12].

鉴于该方法的广泛应用性和既有研究的不足，本文在梳理国内既有使用倾向得分匹配的文章的基础上，首先对目前广泛使用的平衡性测度指标进行总结，对文献中普遍使用的均值差 t 检验方法进行回顾并说明其缺陷，然后推荐了六种多维度测度指标，分别基于标准化距离和分位数距离的视角对平衡性进行测度. 第二，利用两个实例(一个微观层面的实证分析，一个宏观层面实证分析)说明和验证既有平衡性测度指标的不足，并计算新的平衡性测度指标，基于这些指标判断样本间的平衡性，研究发现均值差 t 检验可能将非平衡的样本判定为平衡，因此单纯基于均值差 t 检验判定平衡性比较片面且可能有误导性. 第三，若样本匹配后在多种平衡性测度指标的判断下显示为不平衡，传统估计方法将有较大偏误，本文创新性地提出了一种新的更稳健的估计方法：倾向得分匹配 – 逆概率加权 – 双重差分(PSM-IPW-DID)，并基于蒙特卡洛模拟比较本文提出的 PSM-IPW-DID 估计量与传统 PSM-DID 估计量的优劣. 最后对两个实例使用该新方法重新进行分析，以进一步说明使用 PSM-DID 和 PSM-IPW-DID 方法时估计结果的差异.

因此，本文的创新性体现在：一是明确指出既

有文献中广泛使用的平衡性测度指标的不足,并给出了更为全面的平衡性测度指标. 二是提出适用于非均衡样本的新的估计方法:倾向得分匹配 – 逆概率加权 – 双重差分(PSM-IPW-DID). 逆概率加权(inverse probability weighting, IPW)由于其基于样本对总体的还原,可以使模型推断结果具有总体代表性,被广泛地应用于缺失数据时的统计分析和因果推断的计量估计中[13-15]. 文献中已有证据表明在倾向得分重合度(overlap)较好时,IPW 方法比 PSM 方法占优(有更小的方差);在倾向得分重合度比较差时,PSM 比 IPW 更稳健[16-18]. 本文提出的新方法结合倾向得分匹配和逆概率加权的长处,规避其短处,在不进一步删除样本的情况下得到一种综合更稳健的双重差分估计方法. 另一方面,已有的文献或是单纯比较倾向得分匹配和倾向得分逆概率加权的联系与区别,如 King 和 Richard[19];或是探讨在一般回归模型中考虑倾向得分的使用,如 Wooldridge[20] 中的逆概率加权 – 回归调整的估计量(IPWRA);本文将 PSM-IPW 与 DID 回归模型结合起来,既考虑倾向得分重合度(overlap)对估计的影响,也为平衡多维控制变量提供了新的思路. ②

1 平衡性的测度

本节首先介绍文献常用的平衡性检验实施方法,然后推荐 6 种本文认为应该更多关注的多维度平衡性测度.

1.1 文献中常用的平衡性检验

基于对既有文献的梳理和总结,本文发现既有文献一般将倾向得分匹配后样本每一个控制变量的处理组和控制组之间的加权均值差异(mean difference)进行 t 检验,以此作为平衡性检验. 文献中该方法的流行得益于 Stata 中 pstest 命令的便利性,该命令输出结果呈现了各变量匹配前后的均值,并输出加权均值差 t 检验的结果(如表 1). ③同时,文献中通常会画出匹配前后的倾向得分核概率密度分布图用以支持共同支撑假设(如图 1 和图 2). 基于这两步操作,既有文献认为验证了倾向得分匹配的有效性. 然而,简单的匹配加权均值差 t 检验只能反映加权总体控制变量分布平衡性的一个维度,比较片面. 而倾向得分概率密度图只是一种简单的图示,可能无法反映变量之间真实的匹配程度.

1.2 值得关注的多维测度

简单的加权均值差异 t 检验,只能反映控制变量的均值在处理组和控制组之间有无显著性差异,且由于目标总体会随着匹配权数的选择而改变,容易忽略处理组和控制组样本数量的差异,忽略对处理组控制变量和控制组处理变量的方差、分位数、分布高阶矩等信息的考察. 而平衡性要求(关键)控制变量的分布尽可能一致,所以仅仅依靠一种均值检验可能并不能说明平衡性的好坏,更无法判断研究设计的优劣. 特别地,在实证研究中,由于倾向得分模型可能被误设,仅依靠倾向得分的平衡性并不足以判断研究设计如“准自然实验”构造的好坏,构造多种多维度的平衡性测度就显得尤为迫切. 基于此,本文参考前沿文献[12],推荐以下六种值得关注的控制变量平衡性测度指标:

1.2.1 标准化均值差异

标准化均值差异是指经过处理组和控制组组间标准误标准化后的组间均值差异,具体公式如下

$$\widehat{\Delta}_{ct} = \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\frac{s_c^2 + s_t^2}{2}}} \tag{1}$$

其中 $\bar{X}_t$、$\bar{X}_c$ 分别为处理组、控制组(一维)控制变量的均值,S_t^2、S_c^2 分别为处理组、控制组的样本方差,如下所示(W_i 为处理状态二值变量)

$$S_c^2 = \frac{1}{N_c - 1} \sum_{i;W_i=0} (X_i - \bar{X}_c)^2 \tag{2}$$

$$S_t^2 = \frac{1}{N_t - 1} \sum_{i;W_i=1} (X_i - \bar{X}_t)^2 \tag{3}$$

使用样本方差的优点是考虑了处理组和控制组之间的样本数量的差异. 在大多数政策评价实

② 最新的相关文献为 Arkhangelsky 和 Imbens[21],他们发现在存在组间异质性时,逆倾向得分加权使固定效应模型估计值更加稳健.

③ Stata 程序 pstest 一般在 psmatch2 后使用,其匹配权重也来自于 psmatch2 中的_weight. 关于匹配权重详情,请见 psmatch2 的帮助文件.

例中，处理组的个体数量一般远小于控制组的个体数量. 文献经验显示，标准化均值差异和平均处理效应(ATE)的估计偏差高度相关，绝对标准化均值差异(即标准化均值差异的绝对值)在百分之十以下才能认为是达到一个比较好的平衡性[22].

1.2.2　马氏距离

马氏距离(Mahalanobis Distance)，也称马哈拉诺比斯距离，是一种有效的衡量两个样本相似度的测度. 与欧氏距离不同的是它考虑到了各个控制变量之间的联系，并且是尺度独立的(scale-invariant). 其表达式为

$$\widehat{\Delta}_{ct}^{mv}=\sqrt{(\bar{X}_t-\bar{X}_c)'\left(\frac{\widehat{\Sigma}_c+\widehat{\Sigma}_t}{2}\right)^{-1}(\bar{X}_t-\bar{X}_c)} \quad (4)$$

其中 $\widehat{\Sigma}_c=\frac{1}{N_c-1}\sum_{i;W_i=0}(X_i-\bar{X}_c)\times(X_i-\bar{X}_c)'$；$\widehat{\Sigma}_t=\frac{1}{N_t-1}\sum_{i;W_i=1}(X_i-\bar{X}_t)\times(X_i-\bar{X}_t)'$. 马氏距离越小，说明控制组样本与处理组样本越接近. 一般来说，对于平衡性检验，马氏距离应控制在 0.1 以内[12].

1.2.3　线性化倾向得分的标准化均值差

类似地，在此也可以考察线性化倾向得分的标准化差值. 线性化倾向得分定义为 $l_{(X)}=\ln\left(\frac{e(x)}{1-e(x)}\right)$，其中 $e(x)$ 为倾向得分. 相比倾向得分而言，线性化倾向得分能更好地测度距离的相对远近，例如同样是倾向得分相差 0.009 的四个样本，倾向得分值为 0.001 和 0.01 两个样本之间的距离比倾向得分值为 0.1 和 0.109 两个样本之间距离大的多，因为前者是十倍的差距而后者只多出 9%；而且在 Logistic 回归估计倾向得分的情况下，$l_{(X)}=\ln\left(\frac{e(x)}{1-e(x)}\right)=\alpha+\beta' x$. 具体而言，线性化倾向得分的标准化均值差的表达式为

$$\widehat{\Delta}_{ct}^{l}=\frac{\bar{l}_t-\bar{l}_c}{\sqrt{\frac{S_{l,c}^2+S_{l,t}^2}{2}}} \quad (5)$$

其中 $\bar{l}_t$、$\bar{l}_c$ 分别为处理组、控制组线性化倾向得分的均值，$S_{l,t}^2$、$S_{l,c}^2$ 分别为处理组、控制组线性化倾向得分的样本方差.

1.2.4　对数样本标准误比

对数样本标准误比(Ln Ratio of Standard Deviations)，即对处理组与控制组控制变量样本标准误比值取对数，主要反映样本间控制变量的分散程度(spread dispersion)差异. 数学表达式为

$$\widehat{\Gamma}_{ct}=\ln(S_t)-\ln(S_c) \quad (6)$$

其中 S_t、S_c 分别为处理组、控制组的样本标准误. 对数样本标准误比是一个控制变量分散度的测量，它也是尺度独立的. 与简单的样本标准误差值或者样本标准误比值相比，对数样本标准误比值更接近于服从正态分布，便于下一步的判定.

1.2.5　分位数差异

针对每一个控制变量，分位数差异分两种：1)以控制组 97.5% 和 2.5% 分位数为基准，处理组样本中处于这两个基准点之外的样本所占比例，定义为 $\widehat{\pi}_t^{0.05}$；2)以处理组 97.5% 和 2.5% 分位数为基准，控制组样本中处于这两个基准点之外的样本所占比例，定义为 $\widehat{\pi}_c^{0.05}$. 在理想状态下(即随机试验中控制变量完全平衡的情况下)，$\widehat{\pi}_t^{0.05}=\widehat{\pi}_c^{0.05}=0.05$. 选择 5% 分位点是由控制组个体特征和处理组个体特征共同决定的. 以倾向得分为例，控制组个体的倾向得分均值一般比较接近 0，且分布集中于均值附近；而处理组个体的倾向得分均值一般比较靠近 1，分布也集中于均值附近. 这两个分布重叠最多的部分往往在 0.5 附近，越往两端重叠概率越小. Imbens 和 Rubin[12]，Imbens[22] 和 Athey 和 Imbens[23] 也推荐使用 5% 分位点.④具体数学表达式如下

$$\widehat{\pi}_t^{0.05}=1-\left(\widehat{F}_t(\widehat{F}_c^{-1}(0.975))+\widehat{F}_t(\widehat{F}_c^{-1}(0.025))\right) \quad (7)$$

$$\widehat{\pi}_c^{0.05}=1-\left(\widehat{F}_c(\widehat{F}_t^{-1}(0.975))+\widehat{F}_c(\widehat{F}_t^{-1}(0.025))\right) \quad (8)$$

其中 $\widehat{F}_t$、$\widehat{F}_c$ 分别为处理组、控制组的经验分布函数. 分位数差异 $\widehat{\pi}_t^{0.05}$ 和 $\widehat{\pi}_c^{0.05}$ 是一种覆盖度的测量

④　% 在这里是类似于 P 值的一个指引，研究者也可以根据实际情况选择 10% 或者 1% 分位点进行测算.

(coverage measurement),它们分别表示处理组中处于控制组 97.5%和 2.5%分位数之外的个体所占的比例和控制组中处于处理组 97.5%和 2.5%分位数之外的个体所占的比例.

1.2.6 q_c & q_t

与分位数差异类似,另外一种直接的测度就是基于线性化倾向得分的分位数差异 q_c & q_t,用以测量控制组和处理组的共同支撑. 若至少能找到一个对照组(i': $W_{i'}=1-W_i$)中样本和其有接近的线性化倾向得分值,也即线性化倾向得分差异小于一个门槛值 l^u(通常设为 0.1 或者 0.05),定义二值变量 ζ_i 为 1,否则为 0. 具体数学表达式如下

$$\zeta_i=\begin{cases}1 & \text{if } \sum_{i':W_{i'}=1-W_i} 1_{|\widehat{l}(X_{i'})-\widehat{l}(X_i)|\leqslant l^u}\geqslant 1\\ 0 & \text{其它}\end{cases} \tag{9}$$

基于此, q_c & q_t 具体表达式为

$$q_c=\frac{1}{N_c}\sum_{i:W_i=0}\zeta_i \tag{10}$$

$$q_t=\frac{1}{N_t}\sum_{i:W_i=1}\zeta_i \tag{11}$$

其中 N_t 是处理组样本数, N_c 是控制组样本数, q_c 和 q_t 反映的是控制组(和处理组)中能找到参照组中类似个体的样本数占该组总样本的比例.

总体上来看,以上推荐的六种测度中,测度(1)、测度(2)和测度(5)是针对每一个控制变量的度量;测度(3)、测度(4)和测度(6)是针对所有控制变量的多维平衡性测度. 测度(1)~测度(4)是标准化距离的度量,测度(5)~测度(6)是控制组和处理组分位数差异的度量. 这些测度方法更注重控制组和处理组整体分布的平衡性,比传统的(加权)均值 t 检验更加全面与客观. 但是,要求所提出的六种平衡性检验全部满足在观测数据实证分析中难以实现,研究者可以根据样本量和研究需要,尽可能满足多种平衡性指标,如 Imbens[22] 和 Athey 和 Imbens[23] 推荐至少满足标准化均值差异、对数样本标准误和分位数差异这三种指标.

2 实例分析: PSM-DID

通过两个实例来说明本文推荐的多种平衡性测度的有效性和实用性.

2.1 "营改增"试点对制造业经营多元化的影响

2.1.1 政策背景

为克服传统税制重复征税的缺陷[24],国务院批准自 2012 年 1 月 1 日起,率先在上海实施了交通运输业和部分现代服务业营改增试点. 在原有 17% 和 13%两档增值税税率下,新增了 11% 和 6%两档较低税率. 在上海试点的基础上,2012 年 9 月 1 日~2012 年 12 月 1 日,"营改增"试点扩大至北京市、天津市、江苏省、安徽省、浙江省、福建省、湖北省与广东省 8 个省份. 一年后,"1+6"行业"营改增"推广至全国所有地区. 并逐步推广至全国和其他服务部门. 下面将基于我国"营改增"税制试点政策这一"准自然实验",将 2012 年进行试点的上海市制造业上市公司作为处理组,使用 PSM 方法匹配其他未进行"1+6"行业"营改增"试点省份的制造企业作为控制组,探究税制改革对制造业经营多元化的影响.

2.1.2 数据与模型

1)数据来源

数据来源 Wind 数据库,使用 2008 年—2014 年全部制造业行业的上市公司相关数据. 从收集的数据中可以看出,非试点省份中有 3 288 个上市制造业公司观测值,政策处理节点 2012 年以前观测值 1 794 个,2012 年以后 1 494 个;处理组(上海市)有 475 个上市制造业公司观测值,其中政策时间节点 2012 年以前观测值 259 个,2012 年后 216 个.

2)实证模型

采用双重差分方法来估计"营改增"试点政策的因果效应,并应用倾向得分匹配(PSM)来构造可比较的处理组和控制组,即在上海的制造业上市公司和非试点省份的上市制造业公司,减少由于样本选择所带来的内生性风险. 具体而言,使用的模型如下

$$\begin{aligned}PilotVAT_i=&\alpha_0+\alpha_1 Lnasset_{it}+\alpha_2 Inexr_{it}+\\&\alpha_3 Lev_{it}+\alpha_4 Profit_{it}+\\&\alpha_5 Market_{it}+\alpha_6 Intasset_{it}+\varepsilon_i\end{aligned} \tag{12}$$

$$\begin{aligned}Revstrura_{ipt}=&\beta_0+\beta_1 Treat_p\times Post_t+\\&\beta_2 X_{ipt}+\eta_i+\gamma_t+\zeta_{it}\end{aligned} \tag{13}$$

其中,下表 i,p,t 分别表示企业,省份和年份. 具体地,方程(12)为估计倾向得分的选择方程(Selec-

tion Equation)，是用于 PSM 匹配的 Pooling Logit 回归模型，被解释变量 $PilotVAT_i$ 为是否加入政策试点的虚拟变量. 若企业 i 位于“营改增”政策试点地区则为 1，否则为 0. 参照已有文献[25]，本文匹配变量选择了可能影响“营改增”试点地区选择的企业特征变量：*Lnasset*(对数资产总额)，*Inexr*(投资支出比，用于购建固定资产、无形资产以及其他长期资产支付的现金与总资产之比)，*Lev*(资产负债率，为企业年末负债总额与资产总额之比)，*Profit*(利润率，营业利润与营业收入之比)，*Market*(市场势力，用勒纳指数衡量，指产品价格与边际成本间的差额，本文采用主营业务收入减主营业务成本之差除以主营业务收入之比获得)，*Intasset*(无形资产占比，企业的无形资产总额与总资产之比).

方程(13)是评估政策的结果方程(Outcome Equation)，是一个双重差分模型. 其中选取 *Revstrura*(主营构成第一名在营业收入中占比)为因变量，反映制造业公司的营业集中度. *Treat* 为政策试点地区的虚拟变量，如果制造业上市公司所在地为上海则 *Treat* 为 1，否则 *Treat* 为 0；*Post* 为政策实施年份前后的虚拟变量，2012 年以前年份为 0，2012 年及以后为 1. 因此，双重差分交互项 $Treat \times Post$ 前的系数为双重差分估计的“营改增”对试点地区制造业上市公司的净效应. X_{ipt} 为控制变量，即方程(12)中的 6 个企业特征(*Lnasset*、*Inexr*、*Lev*、*Profit*、*Market*、*Intasset*). 在结果方程中，本文控制了公司层面个体固定效应 η_i 和年份固定效应 γ_t.

2.1.3　倾向得分匹配结果

如前所述，本文先进行倾向得分匹配，在估计完倾向得分值之后，常用的匹配方法有以下几种：1)有放回的最近临近匹配，通常选择 1∶4[26]或者 1∶1(最优的情况，但是会损失一定样本)同时限制最大卡尺(Caliper)距离为 0.05；2)半径匹配，即选择一个倾向得分匹配所能允许的最大半径值，可以选择 0.05 或者选择一个倾向得分的标准误；3)核加权匹配，需要选择一个匹配的核函数，一般选高斯核函数(Gaussian)或者双权重核函数(Biweight). 本文选择文献中常用的 1∶1 临近匹配方法(用 psmatch2 实现). ⑤匹配前后控制变量的均值差异如表 1 所示.

表 1　匹配前后平衡性检验结果

Table 1 Balance test results before and after matching

变量	Unmatched	Mean		%bias	%reduct	*t*-test		*V*(*T*)/
	Matched	Treated	Control		|bias|	*t*	*p*>*t*	*V*(*C*)
Lnasset	U	21.91	21.83	6.2		1.29	0.197	1.33*
	M	21.92	21.90	1	84.0	0.15	0.884	1.24*
Lev	U	0.423	0.538	−13.7		−2.1	0.036	0.04*
	M	0.423	0.425	−0.2	98.3	−0.13	0.899	0.87
Inexr	U	−0.051	−0.060	10.5		2.11	0.035	1.08
	M	−0.052	−0.047	−4.7	54.7	−0.69	0.489	0.81*
Profit	U	0.049	−0.019	4.7		0.72	0.473	0.04*
	M	0.049	0.088	−2.7	42.9	−1.59	0.112	1.92*
Market	U	0.248	0.240	4.9		0.96	0.335	0.95
	M	0.249	0.249	0	99.7	0	0.998	0.90
Intasst	U	0.520	0.634	−13.5		−2.06	0.039	0.04*
	M	0.521	0.516	0.6	95.3	0.33	0.745	0.98

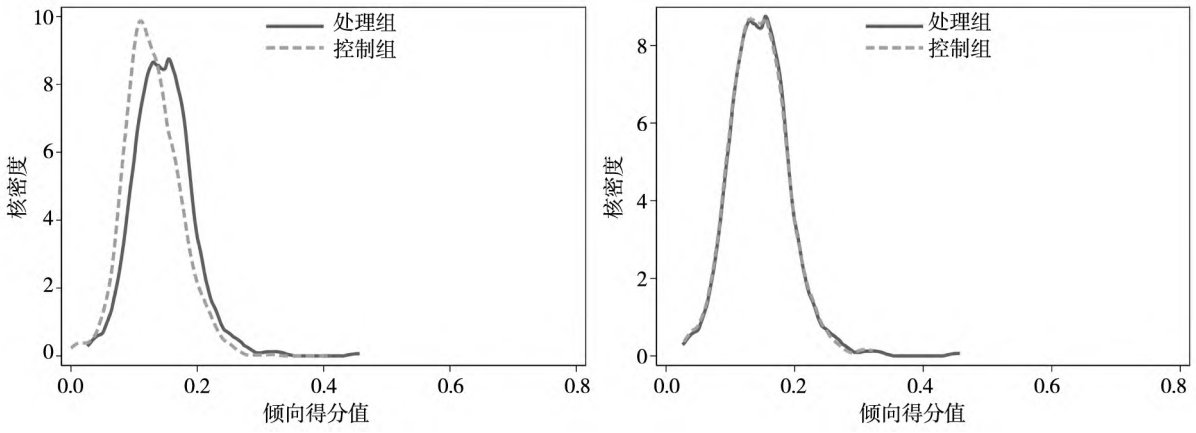
注：“*”含义为对匹配或未匹配的样本方差比大于[0.83；1.20].

表 1 中第二列为匹配前后标识变量，U 代表未匹配(Unmatched)，M 代表匹配后(Matched). 第三列、第四列为匹配权重加权后的均值，第三列为处理组均值，第四列为控制组均值. 第五列%

⑤　考虑到估计的倾向得分所带来的不确定性，1∶1 近邻匹配严格意义上应该使用 teffect psmatch 程序包，但是使用 teffect psmatch 时无法获得控制变量的回归系数及显著性。鉴于此，本文采用了文献中常用的 psmatch2 程序包中 1∶ 1 匹配方法，得到基于均值平衡性检验表及倾向得分重合度检验图，并加上共同支撑假设进行样本删减，以便进行后续回归分析.

bias 是标准化平均值差异. 公式为(Weighted Mean_T-Weighted Mean_UT)/SD,即用表格中处理组与控制组的加权均值之差,除以该变量加权样本的标准误.⑥第六列%reduct |bias|是匹配后标准化平均值差异下降的幅度,其数值是通过前面 %bias 一列得到的,公式为(|UnMatched %bias|-|Matched % bias|)/|UM%bias|,度量匹配之后处理组和控制组间的 bias 减少了多少. 第七列、第八列的 *t*-test,用于判断前述的 %bias 是否显著,若显著则说明针对该变量而言,处理组和控制组的加权均值差异是显著的. 最后一列 *V*(*T*)/*V*(*C*)为控制变量方差比. 星号表示控制变量方差比值超过 F 统计量 2.5%和 97.5%分位值,表示值得关注的控制变量,详见 Austin[27].

从上表中可以看出匹配完后的样本,加权均值差异的 *t* 检验都通过,加权均值差异在统计意义上都不显著,但是方差比值(*Variance ratio*: = *V*(*T*)/*V*(*C*))仍然比较显著,如 *Lnasset*, *Inexr*, *Profit*,而且加权总体可能并不是政策研究所关注的总体. 然而,在实证中这些通常被忽略. 本文将在平衡性检验中使用更多的直观测度来检验控制变量的平衡性.



(a)匹配前　　(b)匹配后

图 1　倾向得分核密度分布图

Fig. 1 Kernel density distribution plot

图 1 展示了匹配前后控制组和处理组倾向得分的核密度分布图. 图 1(a)中匹配前处理组的倾向得分分布和控制组的并不完全重合(特别是众数明显不一致),且有一些倾向得分较小或者较大的个体无法找到匹配个体(匹配奇异值). 但是匹配后,从图 1(b)中可以看出处理组的倾向得分分布和控制组基本重合(众数和分布区间基本一致),共同支撑假设基本成立. 值得注意的是,匹配过程中删掉了 36 个匹配奇异值. 下文实例分析中本文将基于此匹配样本比较常用的 PSM-DID 估计方法和本文提出的新方法.

2.1.4　平衡性测度

此处将第一章中推荐的六种平衡性测度方法应用到“营改增”的数据中,以期从多方面来衡量控制变量的平衡性,并从不同侧面考察控制组和处理组的可比性.

表 2 中前 4 列为控制组均值、控制组标准误、处理组均值和处理组标准误. 从表 2 第 5 列开始,汇报了四种新的平衡性测度:标准化差值、对数标准误比、控制组 5%分位点对应值和处理组 5%分位点对应值. 从表中可以发现,通过倾向得分匹配后,处理组与控制组均值和标准误都比较接近,但是从新的测度指标发现,有些变量并没有如预期的那样完全平衡. 例如,对数资产总额 *lnasset* 标准化差值和处理组 5%分位点对应值都比较高,说明控制组在 2.5%分位数以下及 97.5%分位数以上的制造业上市企业很难找到与其匹配的处理组;资产负债率 *Lev* 对数标准误比较大,说明尽管处理组和控制组均值接近,但是分布比较不均;利润率 *Profit* 和无

⑥ 这里的加权样本标准误为控制组和处理组加权样本方差的均值开根号,详见 Stata 软件中 pstest 帮助文件.

形资产占比 *Intasst* 也存在类似的问题；最后，本文发现即使将所有控制变量单一化为倾向得分，估计的倾向得分在处理组 5%分位点对应值为 0.098，表明控制组的倾向得分在 2.5%分位数以下及 97.5%分位数以上的制造业上市企业很难找到与其匹配的处理组.⑦

表 2　多维度平衡性检验

Table 2 Multidimensional balance test

变量	控制组均值	控制组标准误	处理组均值	处理组标准误	标准化差值	对数标准误比	控制组5%分位点对应值	处理组5%分位点对应值
Lnasset	21.818	1.274	22.072	1.453	**0.186**	0.131	0.010	**0.093**
Lev	0.526	1.853	0.411	0.215	−0.087	**−2.152**	0.064	0.042
Inexr	−0.060	0.086	−0.054	0.089	0.073	0.037	0.039	0.070
Profit	−0.011	1.438	0.026	0.588	0.033	**−0.894**	0.086	0.042
Market	0.242	0.178	0.250	0.178	0.045	0.001	0.070	0.042
Intasst	0.620	1.852	0.518	0.236	−0.078	**−2.062**	0.069	0.061
$e(x)$	0.057	0.028	0.074	0.041	0.492	0.378	0.046	**0.098**
$l(x)$	−2.932	0.561	−2.658	0.575	0.483	0.024	0.046	**0.098**

注： $e(x)$指估计的倾向得分，$l(x)=\ln(e(x)/(1-e(x)))$为线性化的倾向得分.

本文还计算了其他两种平衡性的测度：Mahalanobis 得分和 q_c & q_t. 计算得到的 Mahalanobis 得分是 0.292 1，大于经验值 0.1. 当固定最大倾向得分间距为 0.1 时，得到 q_t = 0.981，q_c = 0.979，即有 1.9%的处理组个体找不到匹配的控制组个体，有 2.1%的控制组个体找不到匹配的处理组的个体. 这些都从不同侧面说明倾向得分匹配后样本的平衡性还有进一步提高的空间或者模型还有待进一步改进.

2.2　智慧城市试点对城市 $PM_{2.5}$ 排放量的影响

2.2.1　政策背景

2010 年开始，中央及地方政府就分别从顶层设计到具体应用不断推出指导和鼓励智慧城市建设的相关政策. 2012 年 12 月 5 日正式发布“关于开展国家智慧城市试点工作的通知”，并印发《国家智慧城市试点暂行管理办法》和《国家智慧城市(区、镇)试点指标体系(试行)》. 首批国家智慧城市试点共涉及 90 个地、县级城市. 本小节将基于我国首批智慧城市试点政策这一“准自然实验”，以首批试点城市作为处理组，采用 PSM 方法选择合适的非试点城市作为对照组，考察智慧城市试点对空气污染物 $PM_{2.5}$ 排放量的影响，从侧面检验数字化发展对环境保护的影响.

2.2.2　数据与模型

1)数据来源

$PM_{2.5}$ 数据来自于哥伦比亚大学国际地球科学信息网络中心(CIESIN)所属的社会经济数据和应用中心(SEDAC)公布的相关数据. 城市层面控制变量数据来自 2006 年—2016 年《中国城市统计年鉴》. 智慧城市名单来自住建部公布名单，将其与中国城市统计年鉴、$PM_{2.5}$ 数据匹配，最终得到 2005 年—2016 年中国 278 个地级市 12 年的面板数据 3 332 个样本.

2)实证模型

为探究智慧城市试点对城市 $PM_{2.5}$ 排放的影响，使用 PSM－DID，基于具有可比性的处理组和控制组分析被处理城市的平均处理效应(ATT)，减少由于样本自选择所带来的内生性风险. 具体来讲，以省内试点城市为处理组，以省内非试点城市为对照组，构建模型如下

$$Smart\_City_c=\alpha_0+\alpha_1\ Pop_{ct}+\alpha_3\ Economic_{ct}+\alpha_4\ Finance_{ct}+\alpha_5\ Urban_{ct}+$$

⑦　这对于估计 ATT 来说问题不大，但是对估计 ATE 或者 ATUT 来说会产生一些偏误. 这从侧面反映匹配后的样本也存在一定程度的不平衡性. 如果在实际测算中发现其值偏离 0.05 较多，则可依据更严格标准进行匹配或者加权.

$$\alpha_6 Open_{ct} + \varepsilon_c \tag{14}$$

$$PM_{2.5ct} = \beta_0 + \beta_1 Treat_C \times Post_t + \beta_2 X_{ct} + \delta_c + \gamma_t + \zeta_{it} \tag{15}$$

其中,方程(14)为估计倾向得分的选择方程(Selection Equation),采用 logit 回归模型,被解释变量 $Smart\_City_C$ 为是否加入政策试点的虚拟变量:若城市 c 为试点城市则为 1,否则为 0. 参考既有文献[28],本文控制了以下城市特征变量(X_{ct}):人口规模(*Pop*),计算方式为 ln(年末总人口);经济发展水平(*Economic*),计算方式为 ln(人均地区生产总值);金融发展水平(*Finance*),计算方式为 ln(年末金融机构人民币各项贷款余额);城市化水平(*Urban*),计算方式为 100 × 非农业人口/年末总人口;市对外开放程度(*Open*),计算方式为外商实际投资额/地区生产总值.

方程(15)为评估政策的结果方程(Outcome Equation),基于匹配后样本进行双重差分回归. 被解释变量为城市年平均 $PM_{2.5}$. $Treat_C$ 为智慧城市二值变量:若为智慧城市试点城市,则为 1,反之则为 0. $Post_t$ 为政策前后虚拟变量,若年份大于 2011 年,则为 1,反之为 0. 因此,核心解释变量为双重差分交互项 $Treat_C \times Post_t$. X_{ct} 为城市层面控制变量. δ_c 为城市层面固定效应,控制了不随时间变化的城市特征. γ_t 为年份固定效应,控制了宏观趋势对回归结果的影响.

表 3 匹配前后平衡性检验结果

Table 3 Balance test results before and after matching

变量	Unmatched	Mean		%bias	%reduct	t – test		$V(T)/V(C)$
	Matched	Treated	Control		|bias|	t	$p>t$	
Pop	U	5.718	5.904	-28.2		-5.26	0.000	1.22
	M	5.718	5.775	-8.7	69.2	-1.17	0.244	1.19
Economic	U	10.569	10.173	57.0		9.91	0.000	0.04*
	M	10.569	10.57	-0.1	99.8	-0.01	0.991	0.87
Finance	U	16.33	16.002	26.3		4.78	0.000	1.08
	M	16.33	16.404	-5.9	77.7	-0.75	0.453	0.81*
Urban	U	80.543	77.889	8.4		1.46	0.145	0.04*
	M	80.543	79.32	3.9	53.9	0.54	0.587	1.92*
Open	U	0.030 46	0.029 68	0.9		0.12	0.901	0.95
	M	0.030 46	0.028 9	1.7	-98.9	0.32	0.752	0.90

注:* 含义为对匹配或未匹配的样本方差比大于[0.81; 1.23].

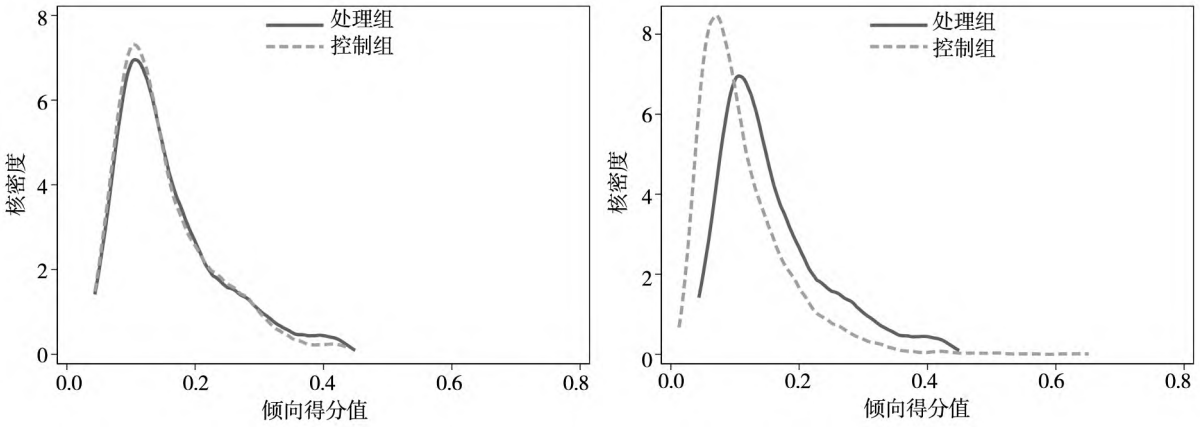
2.2.3 倾向得分匹配结果

由表 3,对比倾向得分匹配完后的各变量的加权样本均值,实验组和对照组的均值差异在统计意义上都不显著(p 值都大于 0.1),因此若按照文献中广泛使用的简单比较均值差异将得出匹配后样本满足平衡性假定的结论. 但是仅从表中信息看,城市化变量方差比值(*Variance ratio*: = $V(T)/V(C)$)部分显著,显示出两组分布的非均衡性,然而在实证中方差比值通常被忽略. 后文将在平衡性检测中使用更多的测度来检验控制变量的平衡性.

更进一步,图 2 展示的是匹配前后控制组和处理组倾向得分的核密度分布图. 图 2(a)中匹配前处理组的倾向得分分布和控制组的并非完全重合,众数和均值表现出明显不一致,且存在匹配奇异值(Outlier),有一些倾向得分较小或者较大的个体无法找到匹配个体. 但是匹配后,从图 2(b)中可以看出处理组的倾向得分分布和控制组基本重合,众数和分布区间基本一致. 值得注意的是,匹配过程中删掉了 255 个的匹配奇异值. 类似地,在后文将基于此匹配样本对常用的 PSM-DID 估计方法与本文提出的新方法进行比较.

2.2.4 平衡性测度

在此将第一章中推荐的六种平衡性测度方法应用到本案例中,以期从多方面来衡量控制变量的平衡性,也从侧面考察控制组和处理组的可比性.



(a)匹配前　　(b)匹配后

图 2　倾向得分核密度分布图

Fig. 2 Kernel density distribution plot

表 4　多维度平衡性检验

Table 4 Multidimensional balance test

变量	控制组均值	控制组标准误	处理组均值	处理组标准误	标准化差值	对数标准误比	控制组 5% 分位点对应值	处理组 5% 分位点对应值
Pop	5.795	0.671	5.669	0.742	−0.178	0.101	0.035	**0.121**
Economic	10.231	0.704	10.955	0.497	1.187	−0.348	**0.469**	**0.176**
Finance	15.964	1.260	16.936	1.036	0.844	−0.196	**0.303**	0.073
Urban	76.854	32.677	99.799	4.548	0.984	**−1.972**	**0.404**	0.049
Open	0.028	0.085	0.029	0.021	0.013	**−1.392**	0.179	0.073
$e(x)$	0.050	0.060	0.136	0.112	0.962	0.620	**0.585**	**0.267**
$l(x)$	−4.769	2.988	−2.101	0.858	1.214	**−1.248**	**0.585**	**0.267**

注： $e(x)$指估计的倾向得分，$l(x)=\ln(e(x)/(1-e(x)))$为线性化的倾向得分.

表 4 中前 4 列为控制组均值、控制组标准误、处理组均值和处理组标准误. 从第 5 列开始，汇报了四种新的平衡性测度：标准化差值、对数标准误比、控制组 5%分位点对应值和处理组 5%分位点对应值. 表 4 显示，经过倾向得分匹配后，处理组与控制组的均值和标准误都比较接近，然而考察新的测度发现，部分变量并未实现完全平衡. 例如，除对外开放水平外，其余控制变量的标准化差值都较高，说明尽管处理组和控制组均值接近，但是分布比较不均. 城市化、对外开放水平的对数标准误差比也都较高. 经济发展水平、金融发展、城市化水平控制组 5%分位点对应值都比较高，说明处理组的 2.5%分位数以下与 97.5%分位数以上的城市很难找到与其匹配的控制组. 人口规模、经济发展水平处理组 5%分位点对应值都比较高，说明控制组的 2.5%分位数以下与 97.5%分位数以上的制造业上市企业很难找到与其匹配的处理组. 最后，表中结果显示即使将所有控制变量单一化为倾向得分，估计的倾向得分在控制组和处理组 5%分位点对应值都较大(相对于理想情况 0.05 而言)，说明处理组和控制组的倾向得分都在 2.5%分位数以下与 97.5%分位数以上的制造业上市企业很难找到与其匹配的处理组.

本案例仍计算了其他两种平衡性的测度：Mahalanobis 得分和 q_c & q_t . 计算得到的 Mahalanobis 得分是 1.108，远高于经验值 0.1. 当最大倾向得分间距固定为 0.1 时，$q_t=0.988$，$q_c=0.546$，即有 1.12%的处理组未能与控制组匹配，而有 45.4%的控制组个体未能与处理组匹配，进一步说明了控制组与处理组之间样本的不平衡性. 以上案例分析结论进一步验证了仅考察均值差异的片面性和误导性，显示出采用更为全面的平衡性

测度指标的重要性和必要性.

3 改进的方法：PSM-IPW-DID

3.1 改进方法

通过前面的实例可以看出,仅仅只对均值差异(Mean Difference)做 t 检验只是一种方便性的选择,远不足以验证控制变量的(分布)平衡性. 对于平衡性的检验,研究者需要从多个角度多个维度来衡量,例如本文推荐的标准化均值差,分位数测度和离算程度差异等等. 但是由于研究中通常使用观测数据,使用多个平衡性测度后常常会发现有些控制变量不能完全平衡甚至分布相差很大,随之而来的问题就是:怎样提升平衡性呢? 怎样提升所估计因果效应(Causal Effect)的可靠性呢?

在引入多种平衡性测度后,可能会出现 PSM 很难满足(绝对的)控制变量平衡性. 为了达到更好的平衡性,一种直接的方法是使用严格的匹配的标准(Criterion),但是这样不仅会使得样本量大量减少,还可能导致后续的双重差分估计结果不显著. 样本量的减少会使得估计结果没有代表性,双重差分结果不显著表明政策效应无法被干净的识别和估计.

为了克服以上问题,本文提出了一种倾向得分匹配 - 逆概率加权 - 双重差分(PSM-IPW-DID)的方法. 逆概率加权(IPW)由 Horvitz 和 Thompson[29]提出来(HT 估计量),随后被计量经济学家广泛引用,如 Hahn[30]、Hirano 等[31]、Frölich[16]、Huber 等[17]、Busso 等[18]等等. 类似地,文献中用逆倾向得分概率加权来计算处理组和控制组的样本均值,这样能有效地去除处理组和控制组由于控制变量的不平衡性和差异性所带来的处理效应估计误差. 具体来讲,在非混淆假设(Confoundedness)下,利用重复期望法则能得到

$$E\left[\frac{D_i Y_i^{obs}}{e(X_i)}\right] = E[Y_i(1)] \tag{16}$$

$$E\left[\frac{(1-D_i) Y_i^{obs}}{1-e(X_i)}\right] = E[Y_i(0)] \tag{17}$$

基于此,两个直接样本的估计量为

$$\widehat{E[Y_i(1)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i Y_i^{obs}}{e(X_i)} \tag{18}$$

$$\widehat{E[Y_i(0)]} = \frac{1}{N}\sum_{i=1}^{N}\frac{(1-D_i) Y_i^{obs}}{1-e(X_i)} \tag{19}$$

因此,所感兴趣的平均处理效应估计量(也是一个 HT 估计量)可写为

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i Y_i^{obs}}{e(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-D_i) Y_i^{obs}}{1-e(X_i)} \tag{20}$$

在实际操作中,可对逆倾向得分权重进行标准化处理,使得其加总和为 1

$$\widehat{\tau} = \frac{\sum_{i=1}^{N}\frac{D_i Y_i^{obs}}{\widehat{e}(X_i)}}{\sum_{i=1}^{N}\frac{D_i}{\widehat{e}(X_i)}} - \frac{\sum_{i=1}^{N}\frac{(1-D_i) Y_i^{obs}}{1-\widehat{e}(X_i)}}{\sum_{i=1}^{N}\frac{(1-D_i)}{1-\widehat{e}(X_i)}} \tag{21}$$

并且通过加权最小二乘回归可以很容易得到 $\widehat{\tau}$

$$Y_i^{obs} = \alpha + \tau D_i + \varepsilon_i \tag{22}$$

其平均处理效应(ATE)权重为 $\widehat{\lambda_i} = \frac{1}{1-\widehat{e}(X_i)}$ if $D_i = 0$; $\widehat{\lambda_i} = \frac{1}{\widehat{e}(X_i)}$ if $D_i = 1$. 类似地,若感兴趣的是处理组的平均处理效应(ATT),则逆概率加权权重为 $\widehat{\lambda_i} = \frac{\widehat{e}(X_i)}{1-\widehat{e}(X_i)}$ if $D_i = 0$; $\widehat{\lambda_i} = 1$ if $D_i = 1$.

值得一提的是,在逆倾向得分概率加权回归中,很容易加入控制变量来拟合观测到的结果变量 Y_i^{obs},即

$$Y_i^{obs} = \alpha + \tau D_i + X_i'\beta + \varepsilon_i \tag{23}$$

这是一种双重稳健(Double Robust)的分析,详见 Sloczynski 和 Wooldridge[32]. 具体体现在

$$\widehat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i^{obs}}{\widehat{e}(X_i)} - \frac{D_i - \widehat{e}(X_i)}{\widehat{e}(X_i)}\widehat{m}_1(X_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(1-D_i) Y_i^{obs}}{1-\widehat{e}(X_i)} - \frac{D_i - \widehat{e}(X_i)}{1-\widehat{e}(X_i)}\widehat{m}_0(X_i)\right) \tag{24}$$

其中 $\widehat{m}_1(X_i) = \widehat{\alpha} + \widehat{\tau} + X_i'\widehat{\beta}$, $\widehat{m}_0(X_i) = \widehat{\alpha} + X_i'\widehat{\beta}$ 分别是处理组和控制组的拟合模型. 注意到 $Y_i^{obs} = D_i Y(1) + (1-D_i) Y(0)$, 式(24)中第一项可表示为

$$E[Y(1)] + E\left[\frac{D - e(X)}{e(X)}(Y(1) - m_1(X))\right] \tag{25}$$

当倾向得分（或 IPW）被正确设定而回归模型不正确或者倾向得分不正确设定而回归模型正确设定时，上式中第二项都等于零. 对 $\widehat{\tau}_{DR}$ 中第二项也同理可推. 这意味着只要其中有一种方法（IPW 或者回归模型）被正确设定且没有不可测的混淆因子（Confounder），双重稳健估计会产生一致的估计量. 因此，相对于单一使用倾向得分匹配模型或者回归模型，双重稳健估计有更高的准确性和稳健性.

3.2　估计步骤

本质上，本文提出的方法结合了倾向得分匹配和逆概率加权的长处，规避了二者短处，得到了一种更为稳健的估计方法. 倾向得分匹配会删除掉一些样本来增加控制变量的平衡性（从而克服样本自选择带来的内生性）和满足共同支撑（common support）的假设，而逆概率加权则可以在不减少样本的情况下控制变量的不平衡性，基于（已匹配）样本还原总体，通过概率加权的方法使得控制组和处理组更加可比，估计结果更有效. 倾向得分匹配和逆概率加权的结合使用既能克服控制变量的不平衡性又能减少删除样本数量，而且克服了单独使用倾向得分匹配（PSM）不足以平衡所有控制变量的短板，也规避了单独使用逆概率加权（IPW）出现极端概率导致估计值方差过大的风险. 具体来讲，本文提出以下估计方法：

第一步，运用 Probit 或者 Logit 估计倾向得分（记为 $\widehat{e}(X_i)$）（如果是面板数据则使用 Pooling Probit 或者 Logit 估计倾向得分），进行倾向得分匹配（1∶4 匹配、半径匹配或者核匹配），得到匹配后样本 S；

第二步，运用本文提出的多种平衡性测度衡量控制变量的平衡性，选出需要特别关注的不能平衡的控制变量 X_{ub}（控制变量 X 的一个子集）；⑧

第三步，基于已经得到的倾向得分，构造 ATT 的逆概率加权 $\widehat{\lambda_i} = \dfrac{\widehat{e}(X_i)}{1-\widehat{e}(X_i)}$ if $D_i = 0$；$\widehat{\lambda_i} = 1$ if $D_i = 1$；

第四步，基于匹配后的样本 S 和得到的逆概率加权，进行包含控制变量 X（包含 X_{ub}）的加权双重差分回归，如下所示

$$Y_{it} = \alpha_0 + \alpha_1 Treat + \alpha_2 Post + \alpha_3 Treat \times Post + X_{it}'\beta + \varepsilon_{it} \tag{26}$$

其中 X_{it} 既包含所有控制变量，也包含不能平衡的控制变量 X_{ub} 及其高阶项. 对于第二步找出的无法平衡的控制变量 X_{ub}，在最后一步（即第四步）回归中可以通过增加高阶项或者交乘项加以重点控制，如 X_{ub}，X_{ub}^2，以及 X_{ub} 与平衡变量 X_b 的交叉项 $X_{ub}X_b$ 等等，以减轻模型误设带来的处理效应估计的偏误.

值得注意的是，本文提出的 PSM-IPW 也可应用于直接估计 ATT 的（均值相减）模型估计中. 如引言中所述，在近十年发表在经济学管理学主要期刊上的 169 篇使用倾向得分匹配的论文中，62 篇是将倾向得分匹配直接应用于均值相减的处理效应 ATT 估计中. 对于这些实例，本文提出的方法只需将第四步的估计方法中双重差分模型改为一个加权均值相减模型使用加权最小二乘法（Weighted OLS）即可.

本文提出的方法既使用了逆概率加权，也加入了（不平衡的）控制变量进行回归，得到的估计量是一个逆概率加权 - 回归调整的估计量（IPWRA，详见 Wooldridge[20]）. 如上文所述，在处理选择模型设定有误而回归模型是正确时，逆概率加权不会影响包含控制变量调整的回归模型分析；另一方面如果处理选择模型设定正确而回归模型有误时，逆概率加权可以纠正回归模型，得到一致性的估计值. 从这个意义上来讲，该估计量沿袭了逆倾向得分加权回归的优点，具有双重稳健的性质（double robustness）[12].

同时，和其他 DID 的改进方法如 Abadie[33] 和 Sant'Anna 等[34] 类似，本文提出的改进方法适用于面板数据，也适用于重复横截面数据（repeated cross sectional data）. 基于 DID 设计，在处理重复横截面数据时，需要构造一个“伪”面板数据（psudo panel data）. 如果不考虑 DID 设计，本文提出的加权方法也适用于截面数据，如文献中

⑧　对于这些通过 PSM 不能平衡的控制变量，需要将其加入到后续的 DID 回归中，甚至考虑其高阶形式. 为了方便起见，后续实例研究中，本文只比较了不加入控制变量和加入所有控制变量两种情况.

Słoczyński 等[35]文章中提出的在截面数据中适用的类似方法. 具体来讲,首先估计倾向得分,以处理组为基础,通过倾向得分来匹配控制组样本,然后构造逆倾向得分权重(IPW),通过加权最小二乘回归(OLS)估计处理组的平均处理效应,即 ATT.

3.3 数值模拟

基于两个蒙特卡洛模拟考察本文提出的 PSM-IPW-DID 估计量与传统 PSM-DID 估计量的优劣. 假设 $W = (W_1, W_2, W_3, W_4)'$, $f_{reg}(W) = 210 + 27.4\,W_1 + 13.7(W_2 + W_3 + W_4)$, $f_{ps}(W) = 0.75(-W_1 + 0.5\,W_2 - 0.25\,W_3 - 0.1\,W_4)$. 假设 $Z = (Z_1, Z_2, Z_3, Z_4)'$, Z 服从一个标准正态分布 $N(0, I_4)$,其中 $\boldsymbol{I}_4$ 是一个 4×4 的单位矩阵. 假设对于 $j = 1,2,3,4$,命 $\widetilde{X_1} = exp\ (0.5\,Z_1)$, $\widetilde{X_2} = 10 + Z_2/(1 + exp\ (Z_1))$, $\widetilde{X_3} = \left(0.6 + \dfrac{Z_1 Z_3}{25}\right)^3$, $\widetilde{X_4} = (20 + Z_2 + Z_4)^2$,标准化后有 $X_j = (\widetilde{X_j} - E[\widetilde{X_j}])/\sqrt{Var(\widetilde{X_j})}$.

与 Kang 和 Schafer[36] 和 Sant' Anna 等[34] 等文献类似,本文考虑以下四种数据生成过程(data generating process)

DGP1: $Y_0(0) = f_{reg}(X) + v(X,D) + \varepsilon_0$,
$Y_1(d) = 2 f_{reg}(X) + v(X,D) + \varepsilon_1(d), d = 0,1$
$p(X) = \dfrac{\exp\ (f_{ps}(X))}{1 + \exp(f_{ps}(X))}$,
$D = 1\{p(X) \geqslant U\}$

DGP2: $Y_0(0) = f_{reg}(X) + v(X,D) + \varepsilon_0$,
$Y_1(d) = 2 f_{reg}(X) + v(X,D) + \varepsilon_1(d), d = 0,1$
$p(Z) = \dfrac{\exp(f_{ps}(Z))}{1 + \exp(f_{ps}(Z))}$,
$D = 1\{p(Z) \geqslant U\}$

DGP3: $Y_0(0) = f_{reg}(Z) + v(Z,D) + \varepsilon_0$,
$Y_1(d) = 2 f_{reg}(Z) + v(Z,D) + \varepsilon_1(d), d = 0,1$
$p(X) = \dfrac{\exp(f_{ps}(X))}{1 + \exp(f_{ps}(X))}$,
$D = 1\{p(X) \geqslant U\}$

DGP4: $Y_0(0) = f_{reg}(Z) + v(Z,D) + \varepsilon_0$,
$Y_1(d) = 2 f_{reg}(Z) + v(Z,D) + \varepsilon_1(d), d = 0,1$
$p(Z) = \dfrac{\exp(f_{ps}(Z))}{1 + \exp(f_{ps}(Z))}$,
$D = 1\{p(Z) \geqslant U\}$

其中 $\varepsilon_0, \varepsilon_1(d), d = 0,1$ 是服从独立标准正态分布的三个随机变量, U 是服从标准均匀分布的随机变量. 对于通用变量 W, $v(W,D)$ 是一个服从均值为 $D \times f_{reg}(W)$,方差为 1 的独立正态随机变量. 观测到的数据集为 $\{Y_{0,i}, Y_{1,i}, D_i, X_i\}_{i=1}^n$,其中, $Y_0 = Y_0(0)$, $Y_1 = D\,Y_1(1) + (1 - D)\,Y_1(0)$. 在上述数据生成过程中,ATT 的真值为零, $v(X, D)$ 代表不随时间变化的不可观测异质性,在 $Y_1(d)$ 方程中 $f_{reg}(X)$ 为时间趋势项(另一个是线性函数部分).

由于本文关注于所观测到的控制变量 X 线性加入到回归模型和 logit 倾向得分模型中的情景,因此 DGP1 中回归模型和倾向得分方程都正确设定,DGP2 中只有回归模型是正确设定的,DGP3 中只有倾向得分方程是正确设定的,DGP4 中回归模型和倾向得分方程设定都不正确. 蒙特卡洛数值模拟的结果如表 5 中所示.

表 5 PSM-IPW-DID 与 PSM-DID 估计值数值模拟分析

Table 5 Comparison between PSM-DID estimator and proposed PSM-IPW-DID estimator

$N = 1\,000, T = 2$	DGP1: 回归模型和倾向得分方程都正确设定			DGP2: 回归模型正确,倾向得分方程不正确设定		
	Bias	RMSE	Cov. rate	Bias	RMSE	Cov. rate
$\hat{\tau}^{psm}$	−9.451	9.661	0.002	−8.735	8.905	0.000
$\hat{\tau}^{psm\_ipw}$	−0.967	1.246	1.000	−0.894	1.168	1.000
	DGP3: 回归模型不正确,倾向得分方程正确设定			DGP4: 回归模型和倾向得分方程都不正确设定		
	Bias	RMSE	Cov. rate	Bias	RMSE	Cov. rate
$\hat{\tau}^{psm}$	−11.294	11.420	0.000	−15.871	15.954	0.000
$\hat{\tau}^{psm\_ipw}$	−2.462	2.644	0.994	−7.823	7.891	0.000

注: 模拟结果来自 1 000 次重复,每次重复中 ATT 的真值为零. $\hat{\tau}^{psm}$ 是方程中提出的 PSM-DID 估计量, $\hat{\tau}^{psm\_ipw}$ 是方程(23)中提出的 PSM-IPW-DID 估计量. Cov. rate 指对应估计量 95%覆盖率.

表 5 中数值模拟结果显示，双重差分估计量在有随控制变量变化的时间趋势项时会引入严重的偏误，且估计值对真值的覆盖率在四种数据生成过程中几乎全部为零（除了 DGP1 中覆盖率有 0.002）. 本质上，在有随控制变量变化的时间趋势项时，双重差分估计量 $\widehat{\tau}^{psm}$ 估计的不再是 ATT，基于 $\widehat{\tau}^{psm}$ 的政策评估可能会产生严重误导. 与之相对的，本文提出的加权双重差分估计量 $\widehat{\tau}^{psm\_ipw}$ 表现出良好的稳健性而且完全符合理论的预期. 在满足双重稳健性质的前三种数据生成过程中，其偏误都比传统双重差分模型估计量低一个数量级；且估计量的覆盖率都是接近 100%（DGP3 中覆盖率为 99.4%）. 即使在第四种数据生成过程中，即回归模型和倾向得分方程都设定错误，偏误也比传统双重差分估计量小一半.

进一步地，为了考察本文估计步骤中筛选不平衡变量在实证分析中的作用，本小节基于数据生成过程 DGP2（回归模型正确，倾向得分方程不正确设定）进行了数值模拟说明. 具体地，假设 X_4 是非平衡控制变量，且回归模型是 X_4 的非线性函数. 考虑如下两种非线性函数关系 $f_{reg_a}(X)=210+27.4X_1+13.7(X_2+X_3+X_4^2)$，$f_{reg\_b}(X)=210+13.7(1+X_4)X_4$，第一种非线性模型中 X_1,X_2,X_3 是线性参与的，第二种非线性模型中只有 X_4 的线性项和二次项. 如果不考虑 X_4 的高阶项，第一种模型的误设程度相对较低. 相应的两种数据生成过程为

$$\begin{aligned}
&\text{DGP2a: } Y_0(0)=f_{reg_a}(X)+v_a(X,D)+\varepsilon_0,\\
&Y_1(d)=2f_{reg_a}(X)+v_a(X,D)+\varepsilon_1(d),\\
&d=0,1\\
&p(Z)=\frac{\exp(f_{ps}(Z))}{1+\exp(f_{ps}(Z))},\\
&D=1\{p(Z)\geqslant U\}\\
&\text{DGP2b: } Y_0(0)=f_{reg_b}(X)+v_b(X,D)+\varepsilon_0,\\
&Y_1(d)=2f_{reg_b}(X)+v_b(X,D)+\varepsilon_1(d),\\
&d=0,1\\
&p(Z)=\frac{\exp(f_{ps}(Z))}{1+\exp(f_{ps}(Z))},\\
&D=1\{p(Z)\geqslant U\}
\end{aligned}$$

其中 X 的生成过程，ε_0，$\varepsilon_1(d)$，$d=0,1$ 及 U 的分布与前一个数值模拟完全一致. $v_j(X,D)$ 是一个服从均值为 $D\times f_{reg_j}(X)(j=a,b)$，方差为 1 的独立正态随机变量. 观测到的数据集为 $\{Y_{0,i},Y_{1,i},D_i,X_i\}_{i=1}^n$，其中 $Y_0=Y_0(0)$，$Y_1=DY_1(1)+(1-D)Y_1(0)$. 在上述数据生成过程中，为方便分析相对误差大小，ATT 的真值设为 6，$v_j(X,D)$ 代表不随时间变化的不可观测异质性，在 $Y_1(d)$ 方程中 $f_{reg\_j}(X)$ 为时间趋势项.

由于本小节关注于筛选出的不平衡控制变量 X_4 简单加入到回归模型和被重点控制加入到回归模型中的情景，在倾向得分匹配过程中，本小节假设匹配仅基于 X_1,X_2,X_3 进行. 随后进行加权双重差分估计时，可获得简单线性控制非平衡变量的 PSM-IPW-DID 估计量 $\widehat{\tau}^{psm\_ipw\_s}$ 和重点非线性控制非平衡变量（如加入 X_4 和 X_4^2 等高阶项）的 PSM-IPW-DID 估计量 $\widehat{\tau}^{psm\_ipw\_c}$. 蒙特卡洛数值模拟的结果如表 6 中所示.

表 6　非平衡变量对 PSM-IPW-DID 估计值影响分析

Table 6 Analysis of the impact of unbalanced covariates on the PSM-IPW-DID estimator

$N=1\,000,T=2$	DGP2a：回归模型 a“正确”和倾向得分方程不正确设定			
	Bias	Ave. SE	5%Sign. rate	10%Sign. rate
$\hat{\tau}^{psm\_ipw\_s}$	-0.539	4.634	0.043	0.144
$\hat{\tau}^{psm\_ipw\_c}$	-0.539	3.158	0.386	0.540
$N=1\,000,T=2$	DGP2b：回归模型 b“正确”和倾向得分方程不正确设定			
	Bias	Ave. SE	5%Sign. rate	10%Sign. rate
$\hat{\tau}^{psm\_ipw\_s}$	1.429	4.032	0.408	0.671
$\hat{\tau}^{psm\_ipw\_c}$	1.429	2.063	0.949	0.976

注：模拟结果来自 1 000 次重复，每次重复中 ATT 的真值为 6. $\hat{\tau}^{psm\_ipw\_s}$ 是方程（23）中提出的 PSM-IPW-DID 估计量且简单线性控制非平衡控制变量，$\widehat{\tau}^{psm\_ipw\_c}$ 是方程（23）中提出的 PSM-IPW-DID 估计量且重点非线性控制非平衡控制变量. Bias 和 Ave. SE 分别指对应估计量的偏误和平均标准差；5%Sign. rate 和 10%Sign. rate 分别指 5%显著性水平下估计量显著的概率和 10%显著性水平下估计量显著次数的占比.

表 6 中数值模拟结果显示,控制非平衡变量的高阶项的估计值与只控制非平衡变量的线性项估计值相比,其大小相同且均存在一定偏误,但标准误更小,回归结果更显著. 更小的标准误使得了无论是在 5%还是 10%显著性水平上,控制非平衡变量的高阶项的估计值都更显著. 比较两种不同误设程度的回归模型 DGP2a 和 DGP2b,可以发现,误设程度较高的 DGP2b 模型的估计值偏误较大(|1.429|和|−0.539|),但是一旦控制非平衡变量的高阶项后,标准误就极大降低,而且显著性水平极大提高. 这些结论进一步验证了考虑控制非平衡变量的非线性函数关系的重要性,对实证分析也有重要指导意义.

4 实例分析: PSM-IPW-DID

本节将提出的 PSM-IPW-DID 方法应用于第二章所阐述的两个政策评估实例中,并与文献中常用的 PSM-DID 方法进行比较. 通过实例对比研究,加深对本文所提出方法的理解和应用推广.

4.1 “营改增”实例分析

首先,利用第三节中倾向得分匹配成功并依据共同支撑假设进行删减后的样本进行双重差分估计,结果如下表 7 中第(1)列、表 7 第(2)列所示. 表 7 第(1)列中是不包括控制变量的固定效应模型回归结果,表 7 第(2)列中是包含了控制变量的回归结果. 回归结果显示,“营改增”政策对主营构成第一名在营业收入占比有显著的负向作用,其大小为 3.4 到 3.7 个百分点. 在加入所有企业层面控制变量和企业层面的固定效应后,平均意义上来讲,“营改增”政策减少了试点地区上市制造业公司 3.7 个百分点的主营构成第一名在营业收入占比,这说明制造业上市公司有分散经营的动向.

表 7 “营改增”对营业收入占比的影响

Table 7 The impact of “replacement of business tax with value-added tax" policy on the proportion of operating income

变量	PSM-DID		PSM-IPW-DID	
	(1)	(2)	(3)	(4)
Treat × *Post*⑨	−0.034***	−0.037***	−0.041***	−0.043***
	(0.012)	(0.012)	(0.014)	(0.014)
Lnasset		−0.014**		−0.012
		(0.006)		(0.009)
Lev		0.040		−0.266*
		(0.052)		(0.150)
Inexr		−0.028		−0.016
		(0.030)		(0.034)
Profit		−0.005		−0.006
		(0.004)		(0.008)
Market		0.173***		0.020
		(0.034)		(0.062)
Intasst		−0.091**		0.129
		(0.046)		(0.128)
*Lev*2				0.126
				(0.167)
*Profit*2				−0.001***
				(0.000 3)
*Intasst*2				−0.106
				(0.126)
常数项	0.849***	1.145***	0.952 5***	1.258***
	(0.006)	(0.135)	(0.012 3)	(0.202)
企业固定效应	是	是	是	是
年份固定效应	是	是	是	是
观测值	3 415	3 414	3 414	3 414
R^2	0.004	0.021	0.795	0.800

注: *、** 与*** 分别表示 10%、5%与 1%的显著性水平, 回归结果均为稳健标准误估计.

⑨DID 模型中的 *Treat* 和 *Post* 被企业固定效应和年份固定效应吸收,因此表 7 中回归结果没有这两项.

在表 7 第(3)列和表 7 第(4)列为使用本文提出的 PSM-IPW-DID 方法估计的政策处理效应. 表 7 第(3)列中是不包括控制变量的固定效应模型回归结果,表 7 第(4)列中是包含了控制变量的回归结果. 具体地,表 7 第(4)列加入了所有控制变量以及表 2 中非平衡变量 *Lev*,*Profit*,*Intasst* 的二次项,即 *Lev*2,*Profit*2 和 *Intasst*2. 回归结果显示,“营改增”政策对主营构成第一名在营业收入占比有显著的负向作用,其大小为 4.1 到 4.3 个百分点. 在加入所有企业控制变量和企业层面的固定效应后,平均意义上来讲,“营改增”政策减少了试点地区上市制造业公司 4.3 个百分点的主营构成第一名在营业收入占比,这说明制造业上市公司经营更为多元化.

值得注意的是,表 7 中第(4)列是考虑了企业个体固定效应且加控制变量的 IPW 双重差分模型,与表 7 第(2)列相比,控制变量的显著性明显降低(表 7 第(2)列中有三个在 1%的显著性上显著,表 7 第(4)列中只有一个在 1%的显著性上显著). 控制变量的显著性大幅减少能有效缓解由于双重差分模型误设所导致的估计偏误. 这些都从实证上验证了本文所提出来的 PSM-IPW-DID 比 PSM-DID 更加稳健.

4.2　“智慧城市”实例分析

类似地,首先,利用倾向得分匹配成功的样本进行双重差分估计,结果如表 8 中第(1)列、表 8 第(2)列所示. 表 8 中第(1)列、表 8 第(2)列为使用固定效应方法进行面板数据回归的结果,其中表 8 第(1)列未控制城市层面控制变量,仅控制了年份固定效应和城市固定效应,表 8 第(2)列在其基础上加入了城市控制变量. 尽管表 8 第(1)列回归结果仍然显示该政策对城市 $PM_{2.5}$ 没有显著影响,但表 8 第(2)列为最为严格控制的模型,其回归结果显示相比非试点城市,试点城市的 $PM_{2.5}$ 显著降低 0.904,所有城市 $PM_{2.5}$ 的均值为 36.68,因此平均而言,城市的 $PM_{2.5}$ 显著降低 2.5%.

表 8　“智慧城市”对 $PM_{2.5}$ 的影响

Table 8 The impact of "smart cities" policy on $PM_{2.5}$

变量	PSM-DID		PSM-IPW-DID	
	(1)	(2)	(3)	(4)
Treat × *Post*⑩	−0.402 (0.417)	−0.904** (0.442)	−0.332 (0.562)	−0.678 (0.557)
Pop		−0.958* (0.500)		−1.205 (0.797)
Economic		−1.943*** (0.646)		−11.771 (13.535)
Finance		−1.629*** (0.516)		11.569* (6.248)
Urban		−0.036*** (0.009)		−0.019 (0.081)
Open		−0.023 (1.026)		−2.016 (1.637)
*Economic*2				0.336 (0.618)
*Finance*2				−0.458** (0.188)
*Urban*2				0.000 (0.000)
常数项	36.470*** (0.244)	86.722*** (9.155)	39.594*** (0.290)	74.448 (67.415)
城市固定效应	是	是	是	是
年份固定效应	是	是	是	是
观测值	3 067	2 837	2 847	2 835
R^2	0.231	0.239	0.962	0.960

注：*、** 与*** 分别表示 10%、5%与 1%的显著性水平，回归结果均为稳健标准误估计.

⑩　DID 模型中的 Treat 和 Post 被城市固定效应和年份固定效应吸收,因此表 8 中回归结果没有这两项.

随后,将本文提出的 PSM-IPW-DID 的方法应用在上述的“智慧城市”数据上,得到的结果如表 8 中第(3)列、表 8 第(4)列所示. 表 8 第(3)列未加入城市层面控制变量,仅控制了年份固定效应和城市固定效应,表 8 第(4)列在其基础上加入了城市控制变量,并加入表 2 中非平衡变量 *Economic*,*Finance*,*Urban* 的二次项,即 *Economic*2,*Finance*2 和 *Urban*2. 回归结果显示,采用改进后的 PSM-IPW-DID 方法后,智慧城市试点政策对城市 $PM_{2.5}$ 的排放量没有显著的影响,表明原 PSM-DID 方法简单平衡性测度错误判断了匹配前后样本的平衡性,从而导致了估计结果并非真实的因果效应,呈现出了虚假的统计显著性,得出了智慧城市对城市 $PM_{2.5}$ 有显著降低作用的错误结果.

值得注意的是,表 8 中第(4)列加入控制变量和非平衡变量二次项的 IPW 双重差分模型与表 8 第(2)列传统双重差分模型相比,控制变量的显著性减少了一半(粗略比较,表 8 第(2)列有 4 个显著,表 8 第(4)列只有 2 个显著,且前者显著性明显高于后者),与表 8 第(3)列不加控制变量的 IPW 双重差分模型相比 DID 系数更加接近(−0.678 和 −0.332). 控制变量的显著性大幅减少能有效缓解由于双重差分模型误设所导致的估计偏误. 因此,本案例在城市层面的分析也验证了本文提出的 PSM-IPW-DID 比 PSM-DID 更加稳健,PSM-DID 现有平衡性测度导致样本非均衡时可能呈现出虚假显著性的回归结果.

5 结束语

本文基于文献中常用的 PSM-DID 估计方法及其平衡性检验的不足,提出了多种平衡性测度方法,并在此基础上提出了一种更加稳健性的 PSM-IPW-DID 方法. 在具体分析中,本文基于文献中研究较多的“营改增”及“智慧城市”的政策评价分析进行探讨,以期能提供更多可操作性的指引和建议. 总结下来,本文建议:1)选择合适的倾向得分匹配方法,如 1∶4 近邻匹配、半径匹配或者核匹配,以确保匹配后不丢失过多样本;2)在进行倾向得分匹配后需要选用多种平衡性测度来检验控制变量的平衡性,简单的均值差异 t 检验是不全面的,本文推荐使用多种多维度平衡性测度;3)如果发现平衡性无法满足,建议采用本文所提出的 PSM-IPW-DID 方法进行估计,以期在不进一步损失样本情况下得到更加稳健的估计结果.

本文关注的平衡性测度问题是在倾向得分匹配之后进行的. 在未来的研究中,研究者可以关注于倾向得分匹配的具体匹配操作,如基于横截面估计的倾向得分在面板结构的数据如何进行精准匹配(Matching),最新的研究如谢申详等[37]就探讨了该类问题. 另外,倾向得分在固定效应模型(不仅仅是双重差分模型)中的应用也是一个值得深入探讨和研究的领域,如 Arkhangelsky 和 Imben[21]. 以上有益的尝试为未来倾向得分匹配方法的研究提供了新的视角和方向.

参 考 文 献:

[1]孙 亮, 刘 春, 陈 凡. 政府赋予型声誉有激励效应吗?[J]. 管理科学学报, 2022, 25(1): 39 – 63.
Sun Liang, Liu Chun, Chen Fan. Does government-granted reputation have incentive effects? [J]. Journal of Management Sciences in China, 2022, 25(1): 39 – 63. (in Chinese)

[2]黄 俊, 赵 宇, 胡丹奇, 等. 司法改善与企业投资——基于我国巡回法庭设立的经验研究[J]. 经济学(季刊), 2021, 21(5): 1521 – 1544.
Huang Jun, Zhao Yu, Hu Danqi, et al. Judicial improvement and corporate investment: Empirical analysis on the establishment of circuit court[J]. China Economic Quarterly, 2021, 21(5): 1521 – 1544. (in Chinese)

[3]吴要武. 产业转移的潜在收益估算——一个劳动力成本视角[J]. 经济学(季刊), 2013, (4): 373 – 398.
Wu Yaowu. Estimation of potential benefits of industrial transfer: A labor cost perspective[J]. China Economic Quarterly, 2013, (4): 373 – 398. (in Chinese)

[4]贾俊雪, 秦 聪. 农村基层治理, 专业协会与农户增收[J]. 经济研究, 2019, (9): 123 – 140.
Jia Junxue, Qin Cong. Rural grassroots governance, specialized cooperative organizationsand farmers' income growth[J]. Economic Research Journal, 2019, (9): 123 – 140. (in Chinese)

[5]万海远，李　实. 户籍歧视对城乡收入差距的影响[J]. 经济研究，2013，(9)：43 -55.
Wan Haiyuan，Li Shi. The effects of household registration system discrimination on urban-rural income inequality in China[J]. Economic Research Journal，2013，(9)：43 -55. (in Chinese)
[6]王　庶，岳希明. 退耕还林、非农就业与农民增收—基于 21 省面板数据的双重差分分析[J]. 经济研究，2017，52(4)：106 -119.
Wang Shu，Yue Ximing. The grain-for-green project，non-farm employment，and the growth of farmer income[J]. Economic Research Journal，2017，52(4)：106 -119. (in Chinese)
[7]龙　玉，赵海龙，张新德，等. 时空压缩下的风险投资—高铁通车与风险投资区域变化[J]. 经济研究，2017，52(4)：195 -208.
Long Yu，Zhao Hailong，Zhang Xinde，et al. High-speed railway and venture capital investment[J]. Economic Research Journal，2017，52(4)：195 -208. (in Chinese)
[8]孙文凯，王乙杰. 父母外出务工对留守儿童健康的影响—基于微观面板数据的再考察[J]. 经济学(季刊)，2016，15(2)：963 -988.
Sun Wenkai，Wang Yijie. The impact of parental migration on left-behind children's health in China[J]. China Economics Quarterly，2016，15(2)：963 -988. (in Chinese)
[9]张　杰，张　帆，陈志远. 出口与企业生产率关系的新检验：中国经验[J]. 世界经济，2016，(6)：54 -76.
Zhang Jie，Zhang Fan，Chen Zhiyuan. A re-examination of the relationship between export and productivity：New micro evidence from China[J]. World Economy，2016，(6)：54 -76. (in Chinese)
[10]王桂军，张　辉. "一带一路"与中国 OFDI 企业 TFP：对发达国家投资视角[J]. 世界经济，2020，43(5)：49 -72.
Wang Guijun，Zhang Hui. The Belt and Road initiative and the TFP of China's OFDI enterprises：A perspective on investing in developed countries[J]. World Economy，2020，43(5)：49 -72. (in Chinese)
[11]Imai K，Ratkovic M. Covariate balancing propensity score[J]. Journal of the Royal Statistical Society：Series B (Statistical Methodology)，2014，(76)：243 -263.
[12]Imbens G W，Rubin D B. Causal Inference for Statistics，Social，and Biomedical Sciences：An Introduction[M]. Cambridge：The Cambridge University Press，2015，ISBN 9781139025751.
[13]邰凌楠，王春雨，田茂再. 缺失数据下的逆概率多重加权分位回归估计及其应用[J]. 统计研究，2018，35(9)：115 -128.
Tai Lingnan，Wang Chunyu，Tian Maozai. Inverse probability multiple weighted quantile regression estimation and its application with missing data[J]. Statistical Research，2018，35(9)：115 -128. (in Chinese)
[14]Negi A. Doubly Weighted M-estimation for Nonrandom Assignment and Missing Outcomes[R]. Working Paper，arXiv：2011，11485.
[15]Sant'Anna P，Song X，Xu Q. Covariate distribution balance via propensity scores[J]. Journal of Applied Econometrics，2022，37(6)：1093 -1120.
[16]Frölich M. Finite-sample properties of propensity-score matching and weighting estimators[J]. The Review of Economics and Statistics，2004，(86)：77 -90.
[17]Huber M，Lechner M，Wunsch C. The performance of estimators based on the propensity score[J]. Journal of Econometrics，2013，(175)，1 -21.
[18]Busso M，DiNardo J，McCrary J. New evidence on the finite sample properties of propensity score reweighting and matching estimators[J]. The Review of Economics and Statistics，2014，96(5)：885 -897.
[19]King G，Nielsen R. Why propensity scores should not be used for matching[J]. Political Analysis，2019，27(4)：435 -454.
[20]Wooldridge J M. Econometric Analysis of Cross Section and Panel Data[M]. Cambridge：The MIT Press，2010，ISBN 9780262232586.
[21]Arkhangelsky D，ImbensG. The Role of the Propensity Score in Fixed Effect Models[R]. Working Paper，2022. Cambridge：National Bureau of Economic Research.
[22]Imbens G W. Matching Methods in practice：Three examples[J]. The Journal of Human Resources，2015，50(2)：373 -419.
[23]Athey S，Imbens G W. The state of applied econometrics：Causality and policy evaluation[J]. The Journal of Economic Perspectives，2017，31(2)：3 -32.
[24]林智平，徐　迪. 税制营改增下资金约束供应链的融资均衡[J]. 管理科学学报，2018，21(10)：14 -31.
Lin Zhiping，Xu Di. Equilibrium financing in capital-constrained supply chains under replacing business tax with value-added tax[J]. Journal of Management Sciences in China，2018，21(10)：14 -31. (in Chinese)
[25]孙晓华，张竣喃，郑　辉. "营改增"促进了制造业与服务业融合发展吗[J]. 中国工业经济，2020，(8)：5 -23.
Sun Xiaohua，Zhang Junnan，Zheng Hui. Will replacing BT with VAT promote the integrated development of manufacturing and services[J]. China Industrial Economics，2020，(8)：5 -23. (in Chinese)
[26]Abadie A，DrukkerD J，Herr L. Implementing matching estimators for average treatment effects in Stata[J]. Stata Journal，

2004, (3), 290 – 311.
[27] Austin P C. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity score matched samples[J]. Statistics in Medicine, 2009, 28(25): 3083 – 3107.
[28] 石大千, 丁 海, 卫 平, 等. 智慧城市建设能否降低环境污染[J]. 中国工业经济, 2018, (6): 117 – 135.
Shi Daqian, Ding Hai, Wei Ping, et al. Can smart city construction reduce environmental pollution[J]. China Industrial Economics, 2018, (6): 117 – 135. (in Chinese)
[29] Horvitz D G, Thompson D J. A generalization of sampling without replacement from a finite universe[J]. Journal of the American Statistical Association, 1952, (47): 663 – 685.
[30] Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects[J]. Econometrica, 1998, (66): 315 – 331.
[31] Hirano K, Imbens G W, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score [J]. Econometrica, 2003, (71): 1161 – 1189.
[32] Słoczyński T, Wooldridge J. A general double robustness result for eEstimating average treatment effects[J]. Econometric Theory, 2018, 34(1): 112 – 133.
[33] Abadie A. Semiparametric difference-in-differences estimators[J]. The Review of Economic Studies, 2005, 72(1): 1 – 19.
[34] Sant'Anna, Pedro H C, Zhao Jun. Doubly robust difference-in-differences estimators[J]. Journal of Econometrics, 2020, 219(1): 101 – 122.
[35] Słoczyński T, Uysal S D, Wooldridge J M. Doubly Robust Estimation of Local Average Treatment Effects Using Inverse Probability Weighted Regression Adjustment[J]. 2022. arXiv preprint arXiv: 2208.01300.
[36] Kang J D Y, Schafer J L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data[J]. Statist. Sci. 2007, 22(4): 569 – 573.
[37] 谢申祥, 范鹏飞, 宛圆渊. 传统 PSM-DID 模型的改进与应用[J]. 统计研究, 2021, 38(2): 146 – 160.
Xie Shenxiang, Fan Pengfei, Wan Yuanyuan. Improvement and application of classical PSM-DID model[J]. Statistical Research, 2021, 38(2): 146 – 160. (in Chinese)

PSM-DID in program evaluation: Current research and an improved method

CAI Jun[1], *YANG Lan*[2 *], *ZHOU Ya-hong*[3]

1. School of Management, Huazhong University of Science and Technology, Wuhan 430074, China;
2. School of Statistics, Southwest University of Finance and Economics, Chengdu 611130, China;
3. School of Economics & Dishui Lake Advanced Finance Institute, Shanghai University of Finance and Economics, Shanghai 200433, China

Abstract: The propensity score matching-differenced-differences model (PSM-DID) is one of the most popular methods for policy evaluation and causal inference. However, it faces the challenge of unbalanced control variables between the treatment group samples and the control group samples. The traditional balance test based on the mean difference t test is prone to one-sided and imprecise conclusions, which makes subsequent causal inferences biased and incredible. In order to overcome the above problems, this paper proposes the following improvements on the traditional balance test: First, it recommends a more comprehensive and multi-dimensional balance measurement index, which facilitates (after applying propensity score matching) the balance comparison between the treated group and the control group in a more rigorous method. Second, a new estimation method for unbalanced samples is proposed: Propensity score matching-inverse probability weighted-differences (PSM-IPW-DID). This method combines the merits of propensity score matching (PSM) in overcoming sample self-selection endogeneity and robustness to unbalanced samples, and the advantage of inverse probability weighting (IPW) in taking advantage of the full-sample information, to obtain a more robust difference-in-differences estimation method without further trimming samples. Furthermore, numerical simulation and real data applications show that the proposed new method can evaluate the effects of various macro and micro policies comprehensively and objectively, and yield credible causal inferences.

Key words: propensity score matching-difference in differences; covariate balance; inverse probability weighting; double robustness