DOI:10.16382/j.cnki.1000-5560.2017.04.001

教育研究中的因果关系推断*

——相关方法原理与实例应用

黄斌 方超 汪栋

(南京财经大学公共管理学院/公共财政研究中心,南京 210023)

摘 要:近二十年来,因果关系推断方法快速发展成熟,并逐渐占据微观计量方法领域的主流地位。本文首先对因果关系推断方法兴起的背景进行了介绍;其次,探讨了判定因果关系需满足的三个条件,对在实验数据和非实验(观测)数据条件下进行因果判定的主要困难,以及观测数据研究中异质性残值的产生原因与构成进行了剖析;其三,借助小班化教学与"新机制"改革效果评价的实际案例,依次阐述了断点回归、工具变量、倾向得分结合倍差等准实验研究方法的基本原理与实现过程;最后,对准实验研究所面临的内部有效性质疑进行了回应,强调对选用方法背后隐含假设进行稳健性检验的重要性。

关键词:因果推断;教育研究;准实验;异质性残值;内部有效性

一、引言

当今社会,公共政策的科学制定越来越依赖因果数量证据的获取,因果推断方法在近二十年间快速发展成熟,并逐渐替代传统的相关分析方法(例如相关性检验与 OLS 回归),占据了微观计量方法领域的主流地位。因果方法之所以能在短时间取得如此成就,主要得益于世界各国政府制定公共政策由依靠主观个体经验模式向客观证据导向(evidence-based)模式的转变。首先,全球经济增长整体放缓,政府财政增收乏力,但民生性财政支出需求却居高不下,经济建设投入与民生投入之间、不同民生支出项目之间的预算竞争愈发激烈。为了在有限的财力中分得更大的份额,各利益相关部门不得不寻求数据证据的支持,以表明自身增支的合理性;其次,在应对更加多元、复杂与庞大的社会系统治理时,既往"摸着石头过河"的公共政策制定模式日渐暴露出其高"犯错"风险与高"试错"成本的弊端,致使决策者在制定政策时慎之又慎;其三,伴随着民主化进程,社会公众对于公共政策亦从原先被动的接受者转变为主动的参与者与质疑者,由此推动政府内部基于绩效问责的管理制度变革,政府需要通过对项目投入与产出之间的因果关系判定与绩效分析,提高公共资源分配与使用效率,并以此来回应外界对公共政策合法性与合理性的质疑。

从统计与计量学理上看,传统相关分析方法处于被淘汰地位亦存在其必然性。相关方法虽然能够揭示政策与现实结果之间的数量变动关系,却无法为现实结果的形成究竟是不是由某一政策促成的这一问题提供可靠的答案。政策 A 与结果 B 有相关关系,并不意味一定是 A 导致 B,因为如果存在其他

* 基金项目:国家社会科学基金教育学一般课题"2000 年后我国义务教育财政制度改革效果评价研究"(BFA140039)。

一个事件 C,它同时对 A 和 B 有影响,那么在未控制 C 的情况下,我们就无法得到 A 对 B 一定有因果效应(causal effect)的结论,由此就不能保证政策 A 一定是一种能通向结果 B 的有效的政策干预手段。若决策者错误地将相关结果视为具有因果意义的依据,极易形成无效的政策干预。例如,分析显示冰淇淋销量与儿童溺水死亡数之间存在数量上的正相关关系,但这并不意味着控制冰淇淋销量能够降低儿童溺水死亡数,因为这两个变量之间关系只是一种伪相关(spurious correlation),它们同时受到气温的影响。此外,相关分析通常无法提供变量间因果走向信息。例如,教育财政研究者常发现地方生均教育支出与财政转移支付之间有负相关关系,但对这一结果我们可以作两种截然不同的解读:一是受财政资助越多,地方的生均教育支出水平越低,这表明转移支付效果较差;二是那些生均支出水平较低的地方得到了更多的转移支付,这表明转移支付的分配符合公平性原则。究竟谁为因,谁为果,相关分析难以给出答案。相比之下,因果推断方法能够提供变量间可靠的因果关系信息,帮助决策者设计出能达成预期政策目标的有效的政策工具,从而成为证据导向型政策研究的主流方法选择。

近年来,因果推断方法被世界各国大量应用于公共政策制定中。美国是当前在公共教育政策领域中最重视因果推断法运用的国家。2002 年布什政府出台《不让一个孩子落后法》(No Child Left Behind Act),该法案在其 A 部分第 9101 节中指出,教育政策在对教育现实进行干预前需得到科学基准研究(Scientifically Based Research)的支持,而符合科学基准的研究应满足两个条件:其一,教育行为或项目的相关信息,必须经由严格、系统以及客观的程序获取;其二,研究设计需采用随机实验或准实验(quasi - experiment)的方法,将个体与组织、项目或行为分配至不同的条件下,运用合理的控制进行项目评价,在各类项目测评方法中优先承认随机实验的研究成果。该法案所提及的随机实验和准实验正是当前最重要的两种因果研究方法。一般认为,前一种方法达成的因果推断的内部有效性(internal validity)[①]要优于后一种方法,因而在上述法案中,随机实验结果被赋予了更高的优先权(Kaplan,2009)。[②]该法案的出台极大刺激了美国教育政策研究对因果推断方法的需求,吸引了众多统计学家和计量经济学家进入到教育政策研究领域。大量新方法的应用还推动了传统教育经济学在研究视角、研究对象与研究领域方面的转变。早先教育经济学研究主要关注教育作为一种数量级的投入要素(劳动力的受教育年限)对于促进个人收入和社会经济发展的工具性作用,而近年来,越来越多的学者开始关注教育作为人身固有权利的实现程度,探讨应如何制定公共教育政策才能切实提高学生学业成绩、改善学校的教学质量。有关教育政策或项目评价性的经济学文献数量不断增多,涉及教育市场化改革(例如宪章学校、教育券)、小班化政策、教师培训项目、教学手段革新等宏观和微观领域(例如,Angrist & Lavy,1999; Angrist et al. ,2006)。因果研究不再追求精细的模型推演与复杂的方法应用,而是强调通过科学的研究设计构造出一种随机实验或类似于随机实验的数据环境,以最小的统计假设为代价获得更加可靠的因果关系结论。常用的因果推断方法包括随机实验、自然实验与断点回归、工具变量法、倾向得分法、倍差法等,本文将循序渐进地对这几种因果推断方法的基本原理及其在教育研究中的实例应用进行介绍。为了降低教育研究者的阅读难度,增强文章的可阅性,我们将在论述中尽量采用非技术性语言,尽可能地避免数学化的表述。

二、因果关系的判定方法

(一)如何判定因果

根据 J. S. Mill(1851)在其著作《逻辑体系》(A System of Logic)中的界定,判定变量间因果关系需满足以下三个条件:第一,在时间顺序上,假设的“因”应该在“果”之前发生,即作为“因”的变量应为前定变量(predeterminant),它应该在“果”变量之前就已经发生了;第二,如果假设的“因”发生了系统性的变动,那么“果”也应当呈现相应的变动;第三,假设的“因”对“果”的影响已经考虑了其他所有可能

的解释，在考虑和控制了其他所有可能的解释后，假设的“因”对“果”依然具有相当的解释力。上述第一个条件明确了因果关系的走向，若两个变量存在因果关系，只可能是先发生的变量影响后发生的变量。第二条件为相关条件，两个变量如果有因果关系，它们首先应是相关的。第三个条件是最难满足的条件，因为若要确定“因”对“果”有影响，就必须考虑并控制所有可能对因果两变量同时具有影响的其他变量。由于存在无限种“其他变量”的可能性，实现完全控制的难度极大。

为了解决这一难题，统计学家先后研发出多种方法，大致可以分为两大类：一是经典的随机实验法。在随机实验下，每个个体被随机挑选接受处理（treatment），每个个体是否接受处理不受其他任何变量的影响，这就保证了再也没有“其他变量”会对因果两变量同时具有影响，在绝对意义上实现了完全的控制。二是准实验方法，此类方法适用于观测数据（即非随机实验数据）的因果推断。在观测数据中个体是否接受处理是非随机的，因此准实验方法就必须采用一定的手段，将可能混淆因果关系的其他变量的影响剔除干净，从而得到假定的“因”对果的净效应（net effect）。当然，要实现“剔除干净”并不容易，这使得准实验结果的内部有效性常遭到质疑，需小心应对。

（二）随机实验中的因果推断问题

在严格意义上，处理效应（treatment effect）应等于个体接受处理后的结果减去该个体如果没有接受处理的结果。前一个结果是我们可以观测到的，而后一个结果被称为反事实结果（counterfactual outcome），它是无法观测到的，因为时间不能倒流，我们永远不可能观测到同一个体在某一时间段处于不同境遇下的所有结果。运用随机试验可以解决这一问题。在随机实验中，所有观测对象被随机分配至处理组与控制组中，每个个体被挑中的概率是相同的，这使得这两组人具有的各项特征（譬如身高、体重等）不具有统计学意义上的显著差异。此时，控制组与处理组可以简单地视为是完全相同的两组人[③]，因此无论谁接受处理，其结果都一样，同理，无论谁接受控制，其结果也一样。只要这一假设得到满足，我们就可以将控制组的结果当作处理组如果未接受处理的结果，而处理效应就等于处理组接受处理的结果减去控制组不接受处理的结果，这两种结果都是可以观测到的。如果处理组与控制组的结果存在显著的差异，那么这一差异只可能是处理手段引起的，于是便可以判定处理手段对结果具有因果效应。

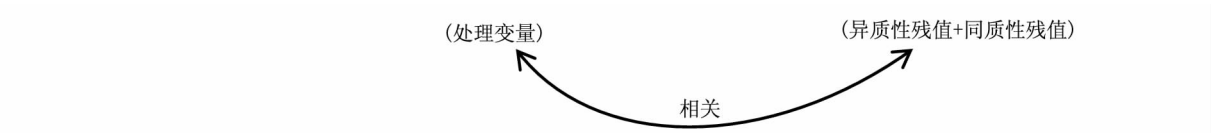
随机实验是自然科学研究进行因果关系分析的经典方法，但当我们将这一方法运用于人文社会科学研究时往往面临着许多困难。首先，随机实验成本较高。为获得相同数据的样本，随机实验通常要比非随机研究花费更多的钱。一项追踪性的人群随机试验费用少则十几万元，多则上百、上千万元，高不可及的实验费用将绝大多数的人文社会科学研究者阻隔在实验研究门槛之外。其次，随机实验运用于人类行为的研究时，有时会引发较大的道德争议，尤其是当处理手段对被试对象福利具有重大影响时，巨大的社会压力常使得随机分配难以得到彻底执行（Borman，2009）。例如对某一种新的教学法进行随机实验研究，如果家长们意识到这一教学法很可能对学生成绩产生提升作用，他们就会干预或破坏实验的随机进程。其三，随机实验通常需要对研究对象进行多期测量，在这一过程中，被试对象可能会对研究者的指令采取不遵从行为，如果被试对象采取不遵从行为的概率与其个人、家庭、学校或社区环境特征变量存在一定相关性，那么样本的随机性就会被“污染”。例如，想探究纠正视力能否提升儿童学业成绩，我们随机分配一部分近视儿童佩戴眼镜，另一部分近视儿童不佩戴眼镜。实验初期先对着两组儿童做一轮学业水平前测，过一段时间再后测，这时我们发现控制组中有部分儿童采取了不遵从行为，家长为他们配备了眼镜，导致这部分原本不应接受处理的儿童实际上接受了处理，并且这些选择不遵从策略（主动为孩子配备眼镜）父母的收入与受教育水平要显著高于其他选择遵从策略的父母。处理组也发生了不遵从行为，有部分儿童把眼镜丢失或损坏了，他们的家长未及时向课题组报告，导致这部分原本应接受处理的儿童实际上并未接受处理，并且这些选择不遵从策略（不报告）父母的收入与

受教育水平要显著低于其他选择遵从策略的父母。在前测时,被试对象是随机分配的,此时控制组和处理组儿童的父母社会经济背景无显著差别,但在后测时,不遵从行为发生了,并且发生的概率与父母社会经济背景特征呈相关性,有部分高社会经济背景的控制组儿童流向处理组,有部分低社会经济背景的处理组儿童流向控制组,由此导致实际接受控制与处理的两组儿童在父母社会经济背景特征上出现了显著差别。此时,若简单将这两组儿童的后测学业成绩之差视为是纠正视力对儿童学业成绩的因果效应,明显是有误的。在父母收入与受教育水平对子女学业成绩存在正向影响的条件下,这一结果高估了纠正视力的处理效应。

(三)准实验中的残值异质性问题

实施随机实验困难重重,逼迫研究者不断思考如何利用更易获得的观测(非实验)数据实现有效的因果关系推断,从而发展出一系列准实验方法。对于准实验研究来说,最棘手的技术性问题是如何妥善地解决残值的异质性(residual heterogeneity)问题,保证对因果效应的无偏估计(unbiased estimate)。在观测数据环境中,处理组与控制组的分配是非随机的,因此两组人群在某些特征变量上很可能存在着显著差异,处于非平衡(unbalance)状态。在这些存在差异的特征变量中,有些变量是在现有数据条件下可以获得的或可以被观测到的,对于这些变量我们在回归模型中直接予以控制即可。但还有些变量是无法获得或不可观测的,因而无法直接控制。如果这些不可测变量与处理变量无关,那么不控制这些变量也不会引起因果关系偏估;但如果与处理变量存在相关性,不控制这些变量就会引起因果关系偏估。

$$income_i = \beta_0 + \beta_1 \cdot treat_i + \beta_2 \cdot gender_i + \beta_3 \cdot ses_i + (ability_i + \varepsilon_i) \quad (1)$$



举一个例子,估计上大学对个人收入的因果效应。如方程(1),结果变量(*income*)表示个人的收入水平,处理变量(*treat*)取值 1 和 0,分别表示个人上没上过大学。个人上没上过大学不是随机分配的,是由高考成绩决定的,而高考成绩的高低又受制于许多因素,因此上过大学(处理组)和没上过大学(控制组)两组人必定在不少特征变量上存在差异。这些变量有些是容易获得或可观测的,如个人的性别(*gender*)与家庭社会经济背景(*ses*)等,而有些是难以观测到的,如个人的天生能力(*ability*)。天生能力对个人是否上大学和收入水平肯定有影响,如果该变量不能得到控制,就会进入到残值中,使得残值由 ε 变为了 $ability + \varepsilon$。原本残值 ε 与处理变量是无关的,而新残值($ability + \varepsilon$)由于包含了能力变量,就变得与处理变量相关了。当残值与处理变量有关时,残值就具有了异质性特征,会引发估计偏差,经济学家称其为“内生性”问题。之所以被称为“内生”,是因为如果模型中有变量需被控制但未能得到控制,那么处理变量就不再是外生给定的,而是由模型残值隐含的一些变量内生决定的。如本例中,是否上大学即为内生变量,它被能力变量内生决定。由于个人是否上过大学与能力正相关,因此在不控制能力变量的条件下,上大学对个人收入的因果效应(β_1)必定被高估。

既然异质性残值是导致因果推断失效的重要原因,那么分析异质性残值产生的原因并按一定的构成将其分解,便成为寻找解决办法的关键所在。导致残值异质性有三方面原因(Heckman,1979;Card,1999;Li & Luo,2004):一是遗漏变量(omitted variable)。如上例,上大学对个人收入的影响之所以被偏估就是因为模型遗漏了个人能力这一重要变量。二是样本选择(sample selection)。观测对象的自我选择(self – selection)行为是形成选择性偏差的主要原因。例如在对小班化的教学效果进行评估中,有地方官员或校长经常将实行小班教学学校的学生成绩与实行大班教学学校的学生成绩直接进行对比,以

此作为小班化教学效果的证据，这是不正确的。在中国，农村学校班额一般偏小，城市学校班额偏大，名校班额一般偏大，普校班额偏小，并且学生就读于何种类型学校或班额并不是随机分配的，它是学生在家庭社会经济背景、天生能力、努力程度等一系列个体特征变量影响下“自我”选择的结果④，由此导致现实中的大班与小班在学生构成、学校师资力量与办学条件等方面存在着一定差异。若不对这些差异予以控制，测评的结果必定是有偏的。三是测量误差（measurement error）。所有的数据都可能存在测量误差，如果测量误差是随机发生的，则不会引起估计偏差；但如果测量误差是非随机的，误差的发生概率与个体某些特征变量存在相关性，就会引起估计偏差。例如在进行入户调查时，我们发现被访者经常会误报自己的受教育年限，并且被访者的受教育水平越低，误报的可能性就越高（黄斌、钟晓琳，2012）。与遗漏变量导致偏估的原理相同，非随机的测量误差如果没能得到妥善的技术处理，就会进入到残值项中，形成异质性残值。

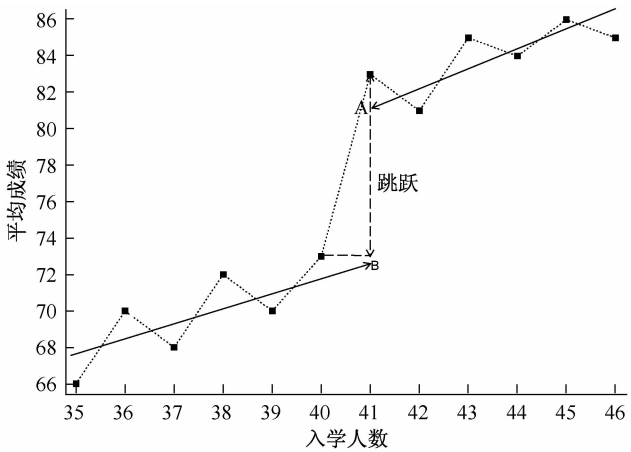
残值异质性还可以按其构成进行分解，一项观测研究可能会同时包含几种异质性残值。如前所述，之所以产生异质性残值是由于在非随机分配下处理组和控制组在一些特征变量上存在着显著差异。在这些存在差异的特征变量中，有些是可以被直接观测到的，但研究者却没有很好地控制它们，由此便会导致可观测异质性；还有些特征变量是在现有数据与测量技术条件下无法或很难观测得到，由此形成所谓的不可观测异质性。此外，按照异质性是否随时间变化的情况，我们又将不可观测异质性分为随时间变化和不随时间变量的异质性两种。前例中，天生能力变量即属于不会随时间变化的一类，而个人努力程度、自我效能（self-efficacy）、社会经济制度等则属于可能随时间变化的一类。从处理难度上来看，解决可观测的异质性最容易，其次是不随时间变化的异质性，最难解决的是随时间变化的异质性。当前流行的几种准实验方法在处理上述三种异质性构成上各具一定的偏向：倾向得分法（Propensity Score Method）只能用于解决可观测的异质性，倍差法（Difference in Difference）可用于消除一切不随时间变化的异质性，而断点回归（Regression Discontinuity）与工具变量法（Instrumental Variable）是将三种异质性构成作整体上的解决。也正是由于这一原因，从达成因果推断的效力上看，断点回归和工具变量法要优于倍差法和倾向得分法。有学者指出，在准实验方法中，断点回归是最具有因果说服力的方法，它也是美国教育部教育科学研究院（Institute of Education Science）唯一认可的具有因果判定效力的准实验方法（Shadish et al.，2002；Smith，2014）。以下，我们将就这些准实验方法一一进行介绍。

三、断点回归

断点回归方法最早由 Thistlethwaite & Campbell（1960）在一本教育心理学专业期刊上提出，之后沉寂多年，直到 1980 年后才逐渐发展成熟并得到大量应用。与随机实验相似，断点回归也是试图利用一种随机安排形成两组无显著差异的个体进行对照比较。不同在于，断点回归的随机安排并不是人为事先就设计好的，而是研究者事后利用在数据形成过程中自然发生的事件构造出来的。这些“自然事件”应是一些外生事件并对结果变量具有冲击作用⑤，例如自然灾害、爆发战争、实行已久的政策戛然而止、政策对象的随机指派，等等。在外来的冲击作用下，观测对象的某一特征变量会在取值上出现一个断点，观测对象被随机分配至断点两边，一方为接受处理的处理组，另一方为未接受处理的控制组。如果处理手段对于观测对象的行为结果具有因果效应，那么，在该断点上观测对象的结果变量取值必定会有一个明显的跳跃变化，而断点回归的核心任务就是侦测这一跳跃变化是否真的存在。Angrist & Lavy（1999）曾使用断点回归对以色列小班化的教学效果进行过研究，本节我们就以该研究为案例来呈现断点回归的一般设计思路。

1996 年以色列政府颁布班额政策，采用 12 世纪犹太教学者 Maimonides 有关集体研修圣经人数最多不能超过 40 个人的教义，规定所有中小学校班额不得超过 40 人，如果学校同年级入学人数超过 40

人就必须拆分为两个班级授课。如前所述,学生就读小班抑或大班受到诸多个人与家庭特征的影响,但在以色列的班额政策下,学生是否就读小班不再受制于这些因素,而是由同年级入学人数决定,这就为研究者实现断点设计创造了条件。如图 1,横坐标为各学校入学人数,该变量被称为“驱动变量”(forcing variable)或“流动变量”(running variable),它的取值应具有连续性(continuum)特质且断点正位于其连续取值中的某一点上。本例中,不同学校是否采用小班教学取决于入学人数是否达到 41 人,因此 41 人这一点即成为断点。该断点很清晰地将不同学校分成两类,断点左边的学校为控制组,采用大班教学;断点右边的学校为处理组,采用小班教学。从图上看,学生平均成绩在断点上存在着一个明显的垂直跳跃,这预示着小班教学对学生学业成绩具有一定的因果效应。那么,这一因果效应的程度有多大呢?



注:为简化讨论,我们对 Angrist & Lavy(1999)的原始数据进行了适当的修改

图 1 断点回归原理图解

先看位于断点附近的两个点,一个是入学人数达到 40 人的学校,样本中这样的学校共有 9 所,学生平均成绩为 73 分;另一个是入学人数达到 41 人的学校,样本中这样的学校共有 28 所,学生平均成绩为 83 分。由于两类学校的入学人数仅相差 1 人,学生对于自己就读学校的同年级同学是 40 人还是 41 人(决定了他们是接受大班还是小班教学)几乎没有自我选择的可能,因此可以看成是一种随机安排。于是,我们就可以如随机实验一般,把接受大班教学学生的平均成绩(73 分)当成是接受小班教学的学生如果没有接受小班教学而是接受大班教学的反事实结果。当然,我们也可以把接受小班教学学生的平时成绩(83 分)当成是接受大班教学的学生如果接受小班教学的反事实结果。无论从何种角度考虑,都可以得到小班教学能使学生平均成绩提高(83 - 73 =)10 分的结论。我们可以采用 t 检验对这两组学校学生的 10 分平均成绩差异是否足够明显(即是否显著)进行检验:如果结果显著,即可判定小班教学对学生学业成绩具有因果效应;若不显著,则证实两者不存在因果关系。在以上整个分析中,我们只采用了均值、单次差与 t 检验三种统计技术,它们都属于统计学的“入门级”方法。可见,良好的研究设计可以极大地降低估计参数所需技术的复杂性,只要构造出满足随机分配的数据环境,研究者只需采用最简单的统计技术便可得到可靠的因果结论。

在以上分析中,我们只采用断点附近两个数据点,纳入分析的学校样本数仅有 37 所,样本容量过小会降低研究的统计功效(statistical power)[⑥]与代表性。为了解决这一问题,放宽取样区间(bandwidth)是必然的办法。例如,我们可以将学校入学人数的取样区间由 40 - 41 人放宽至 35 - 46 人,这样我们就可以利用断点左右两边各六个数据点,运用 OLS 估计出两条拟合线,并据此预测出:当学校入学人数达到 41 人且采用小班教学时的学生平时成绩(即 A 点),以及当学校入学人数达到 41 人但依然采

用大班教学时的学生平时成绩(即 B 点)。B 点是 A 点的反事实结果,分别表示入学人数为 41 人学校的学生在小班和大班教学下的平均成绩,用 A 值减去 B 值就可得到小班教学的教学效果。以上估计过程可采用如下计量模型实现:

$$Achievement_i = \beta_0 + \beta_1(ENROL_i - 41) + \beta_2 SMALL + \varepsilon_i \quad (2)$$

上式中,因变量 *Achievement* 为设定的取样区间中各校学生的平均成绩,*ENROL* 表示各学校的入学人数,对该变量做有关断点 41 人的中心化处理⑦,*SMALL* 为小班教学虚拟变量,取值 1 表示学校入学人数在 41 人及以上,取值 0 表示入学人数 41 人以下。运用简单 OLS 回归,便可估计出 *SMALL* 变量的估计系数 β_2,它正是我们想得到的小班教学的处理效应。在 OLS 回归中,我们同样可以对系数 β_2 的显著性进行 t 检验,以判定因果关系是否存在。

断点回归方法可以说是与随机实验血缘关系最近的一种准实验方法(Lee & Lemieus,2010),在实际研究中有着广泛的应用,但也存在被"滥用"的倾向。为实现有效的断点研究设计,需注意以下几点:一是通常情况下,所选择的驱动变量应为连续(continuum)或有序(ordinal)变量,以往研究惯用的驱动变量包括学生学业成绩、学生人数、出生日期、地理位置、气温变化、交通拥挤程度,等等。驱动变量应有着清晰的概念界定,最好能够被直接观测或测量,尽量不要采用需复杂构建或存在测量争议的变量作为驱动变量。二是自然事件的发生不应受模型其他变量的影响,断点应保证外生特质,能够清晰地界定处理组和控制组,这样才能保证观测对象在断点左右的落点是随机的。传统的断点研究大多采用精确断点回归(Sharp Regression Discontinuity,SRD),要求所有观测对象都应严格遵从断点的指示,绝大部分归于处理组的个体最终都应接受处理,绝大部分归于控制组的个体最终都应接受控制,如果样本中采取不遵从行为的个体比例超过 5%,便很可能产生偏估的结果(Trochim,1984)。为了克服这一方法缺陷,统计学家又研发出模糊断点回归(Fuzzy Regression Discontinuity,FRD),允许处理组和控制组中有一定比例的个体采取不遵从行为。严格来说,Angrist & Lavy(1999)的研究就属于模糊断点一类,因为在他们数据中有一部分学校虽然入学人数不足 41 人但也采用了小班化教学。当然,模糊断点并不意味着对个体遵从比例没有要求,如果样本中采取不遵从行为的比例很高,断点设计也就丧失了形成有效因果推断的作用。运用工具变量法估计处理效应是处理模糊断点的常用手段(Jacob & Lefgren,2004),有关这一点我们在下一节再作说明。三是取值区间的选择是一把"双刃剑"。一方面,扩宽取值区间能增大分析数据的样本容量,使估计系数更容易通过显著性检验,增大了研究结果推及总体的可一般化(generalization)范围;另一方面,扩宽取值区间意味着样本中有更多的观测对象是来自离断点较远的区域,此时控制组与处理组分配是否还具有随机性特质便会受到质疑,如本例中,家庭对于子女就读学校的入学人数是 40 人还是 41 人不太存在自我选择的可能,但对于子女就读学校的入学人数是 30 人还是 50 人,家庭有选择的空间。为了让孩子有更高概率接受小班教学,有条件的家庭可能会特意搬迁到学龄儿童人数一贯偏少且教育质量较好的学区居住,由此导致了处理组和控制组个体在家庭社会经济背景上存在着一定的差异。因此,一个好的断点研究设计通常需要做有关估计区间的稳健性检验,取不同宽度的估计区间得到若干估计结果,并进行对比分析。若估计结果未随着估计区间的放宽而发生明显的变化,就说明这一结果是稳健可靠的。四是断点回归函数形式的选择对于估计结果有重要影响。如本例,我们仅采用简单线性函数形式,假定学生学习成绩与学校入学人数之间的变动关系为线性,但如果此二者真实的函数关系是非线性的,那么图 1 中 A 点与 B 点的估计位置就会发生变化。为避免出现这一问题,我们需先通过观察散点图来推测驱动变量与结果变量之间可能的函数关系,并采用多种函数形式进行回归分析,观测估计结果在不同函数关系下的稳健性表现。

断点设计也存在着一些难以克服的缺陷。首先,断点回归只具有有限的一般化能力,其估计结果只在断点两侧较窄的估计区间内有效。如本例,所估计出的小班化教学效果只对那些入学规模在 35 -

46 人之间的中等规模学校是有效的,这一结论能否推及其他小规模和大规模学校并不确定。其次,断点回归对样本容量有很高的要求,所需样本数通常是随机试验的 1.5 – 2 倍以上,尤其是当断点位于驱动变量分布的尾部或顶部时,所需数据量就会更大。(Cappelleri et al. ,1994)

四、工具变量法

与断点回归相同,工具变量法对残值异质性也是采用整体解决的方案,但在研究设计上,工具变量法则另辟蹊径。如前所述,之所以会出现因果关系估计偏差,是因为异质性残值与处理变量出现了相关,导致处理变量的内生化。在此基础上,研究者进一步思考,如果我们能够通过一定方法将处理变量的变异进行分解,将处理变量变异中与残值相关的那部分变异(即内生变异部分)剔除,仅保留与残值不相关的那部分变异(即外生变异部分),并只用这部分外生变异对结果变量进行回归的话,其估计必定是无偏的。

仍以小班教学为例,如图 2 所示,为估计班级规模对学生成绩的因果效应,我们需控制一切可能与班级规模相关并对学生成绩有影响的变量。在不可能实现完全控制的情况下,班级规模成为内生变量。为剥离该内生变量的异质性变异,我们需找到一个工具变量,它满足以下三个条件(Angrist & Steffen,2015):一是它对内生变量具有因果效应;二是它是外生的,与异质性残值不存在相关关系;三是该工具变量除通过内生变量对结果变量产生影响外,不再可能通过其他渠道对结果变量产生影响。

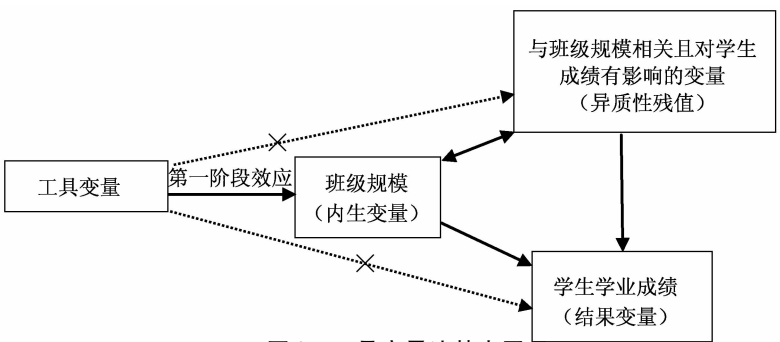


图 2 工具变量法基本原理

运用这三个条件,我们就构建了唯一的一条可由工具变量通往结果变量的因果关系链条,即工具变量→班级规模→学生学业成绩。如果我们能证实工具变量对学生学业成绩有因果效应,那么这一因果效应必定是通过班级规模变量实现的,由此便可证明班级规模必定对学生学业成绩具有因果效应。根据这一因果链条,可得到以下等式:

工具变量对结果变量的效应 = 工具变量对内生变量的效应 × 内生变量对结果变量的效应

上述等式左边第一个效应称为简化效应(reduced effect),右边的第一个效应称为第一阶段效应(first – stage effect),第二个效应即为我们关心的处理效应。对该等式进行移项,便可得 Imbens & Angrist(1994)提出的局部平均处理效应(Local Average Treatment Effect,LATE)估计式:

$$\text{处理效应} = \text{简化效应} \div \text{第一阶段效应} \tag{3}$$

根据公式(3),由于工具变量与异质性残值无关,因此在第一阶段效应中,工具变量所解释的处理变量变异部分必定是外生的(即与异质性残值无关),用简化效应除以第一阶段效应就相当于只用处理变量变异中的外生变异对结果变量变异进行解释,表示结果变量会随着处理变量的外生变动发生怎样的变化,由此就排除了所有可能引发内生性偏估的嫌疑,得到了处理变量对结果变量的真实效应。凡事有好的一面,也有坏的一面。工具变量分离处理变量变异的做法虽然确保了研究的内部有效性,但也造成研究外部有效性的减损,工具变量法的因果结论只对那些受工具变量影响的处理组个体有效,

无法推广至更广泛的群体，这也正是估计式（3）被称为“局部”效应的原因。事实上，对现有文献的阅读我们发现一个现象，一项研究设计越是精妙，内部有效性表现越是优异，其结论一般化的能力往往越弱。如何调和内部有效性与外部有效性之间此消彼长的矛盾关系，实现此二者在同一研究中的共同增进，是当前因果推断技术发展亟待突破的一大瓶颈问题。

工具变量法可通过两阶段回归或矩估计实现，其计算过程相对复杂，研究者可利用现代统计软件（例如 Stata 或 R）提供的命令快速运算出结果。工具变量法最大的难题在于难以寻找到严格满足条件的工具变量。找到一个工具变量与内生变量强相关相对容易，但同时还要保证第二和第三个条件则困难得多，因为异质性残值本身包含了一些无法得到测量的变量，是一个未知的量，我们如何能证明工具变量与未知量之间是否存在相关性呢？此外，影响结果变量的变量有很多，我们如何才能保证工具变量只能通过内生变量才会对结果变量产生影响呢？为保证第二和第三条件成立，最稳妥的办法就是找一个随机变量来作为工具变量。上一节介绍的模糊断点回归就是采用了这个思路：有部分入学人数未达到 41 人的学校也采取了小班教学，这些学校与严格遵守政策采取大班教学的学校在某些特征上可能存在着差异，在未对这些差异进行控制的情况下，小班教学变量就变成了一种非随机安排的内生变量，此时小班教学效果估计系数 β_2 就也不再是无偏的。小班教学安排非随机，但观测对象落在断点两侧是随机的，这就为我们提供了一个“天然”的工具变量：首先，学校入学人数的落点与学校最终是否采取小班教学之间具有很强的相关性，落在左边的学校大多采用大班教学，落在右边的学校大多采用小班教学（满足第一条件）；其次，学校入学人数的落点是随机的，那么就应该与其他变量不具有相关性（满足第二条件）；其三，学校入学人数的落点除了通过影响学校班额规模外，再无其他会对学生成绩产生影响的渠道（第三条件）。由此，我们就可以用学校入学人数是否达到 41 人[⑧]这一虚拟变量作为工具变量，以学校最终是否采取小班教学作为内生变量进行两阶段回归或矩估计，从而得到小班教学效果无偏的估计值。模糊断点的设计巧妙之处在于，它充分利用断点设计为内生性处理变量提供了一个无需任何统计验证的确凿有效的工具变量。

如果选择的工具变量不具有随机性，那么工具变量法的估计结果就可能受到方方面面的质疑。Hoxby（2000）曾以学区人口出生率作为工具变量，对美国康涅狄格州小学小班化教学效果进行过估计。受人口流动与分层居住的影响，学区人口出生率并不是一个由“上帝”决定的随机变量，一个具有高社会经济背景的优质学区的人口出生率与一个具有低社会经济背景的劣质学区的人口出生率可能有较大落差。为此，Hoxby 采用一定较复杂的技术，将学区人口出生率变化中具有自然特质的变异部分分离出来，并将这一部分自然变异作为工具变量进行回两阶段回归。Hoxby 的工具变量设计可谓精妙，但依然面临着内部有效性的质疑，因为很难证明她所提取出的那部分自然变异真的是完全外生的，不会受到任何个人、家庭、学校、社区特征变量的影响。不得以，Hoxby 在文末采用断点回归法对小班化效果再做了一次估计，将工具变量结果与断点回归结果对比，发现两者的结果一致，都显示小班化教学对学生成绩无显著影响。

如前所述，工具变量法在估计中只使用了处理变量的一部分变异，而处理变量变异的缩小会使得处理效应的标准误增大，显著性检验结果常不尽如人意。针对这个问题，可考虑：（1）在模型中增加控制变量，但要确保新加入的控制变量应是外生变量，否则又将产生新的内生性偏估问题；（2）采用多个工具变量，这样可以增大内生变量被分离部分的变异度，使处理效应标准误下降。此外，导致变量内生的原因有多种，例如在估计教育收益率时，个人受教育年限变量就可能同时面临两种内生性：一是模型遗漏个人能力变量导致的内生性，二是受教育年限测量误差导致的内生性。前者会高估教育收益率，后者会低估教育收益率。对于不同的内生性问题，我们可以采用不同的工具变量分别解决，对不同方向的偏估同时进行纠偏（Card，1999；Li & Luo，2004；黄斌，钟晓琳，2012）。

五、倾向得分法结合倍差法

在实际数据分析中,受制于观测数据的事后性,有效的断点与工具变量往往难以找寻,这严重掣肘着因果关系推断方法的推广应用。近年来,倾向得分法结合倍差法的研究设计逐渐盛行,该方法的基本思路是对异质性残值的不同构成分别运用不同的方法予以解决:运用倾向得分匹配法解决可观测异质性,运用倍差法解决不随时间变化的不可观测异质性,对于随时间变化的不可观测异质性则采用一些补救手段尽可能地予以消除。下面,我们将以农村义务教育经费保障新机制改革(下文简称"新机制"改革)的效果评价研究为例(黄斌,苗晶晶,金俊,2016),介绍如何结合运用这两种方法。

大家或许都知道,2006 年"新机制"改革是我国义务教育财政体制的一次重大改革。在"新机制"改革前后,中国农村义务教育财政现实状况发生了巨大的变化,这是不争的事实(黄斌,汪栋,2016)。然而,引起地方教育财政状况发生变化的因素有很多,改革与现实变化之间是否存在直接的因果关系?改革对于现实变化起到了多大的作用?为了准确回答这些问题,研究者需要就改革的因果效应进行科学的量化分析。"新机制"改革面向多个教育财政目标,为了简化讨论,我们在此仅关注改革对于农村中小学生均公用经费水平的因果效应。中国的重大制度改革多采用先试点后全国推广的渐进模式,"新机制"改革亦是如此,这为研究者运用倾向得分法与倍差法提供了机会。如果改革采用的是爆炸模式,所有地区同时进行改革,研究者就无法在某一历史时间段上同时寻找到控制组和处理组。"新机制"试点改革始于 2006 年春季,2007 年春季结束,试点仅持续一年,这逼迫我们只能采用 2005 - 2006 年数据对改革的短期效应进行估计。2001 年实行"以县为主"的体制后,县级政府取代乡、村成为农村义务教育的财政支出主体,因此我们有针对性地采用全国县级数据展开分析。采用县级数据的另一大好处是它能提供更大的样本数,全国省级单位仅 34 个,地级市单位不足 300 个,而县级单位(包括县、县级市、市辖区)则接近 3000 个。对于统计分析来说,拥有更大的样本数就意味着更高的统计功效、更精确的估计结果,以及能够采用更加精细的研究设计,获得更丰富的估计结果。

表 1 未发生改革时的数据平衡检验结果

变量	改革县与未改革县均值差		
	数据匹配前	数据匹配后	
		近邻配合半径匹配法	马氏距离匹配法
人均 GDP	-2133.5***	-127.27	-106.98
人均一般性财政转移支付	198.29***	-66.69***	2.48
人均专项财政转移支付	145.3***	33.60***	3.69
总人口数	-16.08***	4.30**	-1.13
人口密度	-0.02***	0.00	0.00
财政供养人口比例	0.01***	0.004**	0.000
农村人口占比	-0.02**	0.00	0.00
东部地区	-0.14***	0.06**	0.00
样本数	3754	1166	428

注:样本仅包括县、县级市,剔除了所有的市辖区。为保持行政区划一致,我们把 2005 - 2006 年间所有发生行政区划变动的县级单位剔除,最终形成了 2005 和 2006 年各 1877 个县级单位的面板数据。表中数字等于各变量的改革县均值减去未改革县均值,数字右上标星号表示各变量均值差的 t 检验结果,***0.01 水平上显著,**0.05 水平上显著,*0.1 水平上显著

"新机制"改革变量是一个典型的内生变量,因为改革试点县的挑选并不是随机的,这意味着每个县被挑选作为试点县的概率是不一样的,具有某些特定特征的县被挑选作为试点的概率要高于其他县,由此导致改革县与未改革县存在严重的数据非平衡问题。如表 1 第二列,在未发生改革的 2005 年,改革县与未改革县在经济、财政、人口等诸多特征变量上存在着系统性差异,试点改革县大多来自于中西部,经济相对欠发达,接受上级政府转移支付相对较多。这些经济、财政与人口变量通常对地方

教育财政支出水平具有影响，如果不予以控制，改革效应的估计值必定是有偏的。实施倾向得分法的目标就是要消除可观测特征变量在两组之间的差异性，实现数据平衡，而要实现这一点，通常需要经过三个步骤（Rosenbaum & Rubin，1985；Guo & Fraser，2010）：

第一步，估计倾向分值。利用2005年（此时所有县都未发生改革）数据，以各县是否发生改革作为因变量，以表1中呈现差异的种种特征变量作为自变量进行逻辑回归（logistical regression）[⑨]，并根据回归结果，预测出各县发生改革的概率值（即所谓倾向分值）。

第二步，实施数据匹配，为每一个改革县找到一个与其倾向分值最相近的未改革县相匹配，删除所有未实现匹配的个体，得到一个新的匹配样本（matching sample）。实施数据匹配有多种方法，可分为贪婪匹配（greedy matching）和最优匹配（optimal matching）两大类。贪婪匹配是实际研究中最常用的数据匹配方法，它包含了若干种具体的匹配法，例如基于倾向分值的近邻匹配（Nearest neighbor matching）与近邻结合半径匹配（Nearest neighbor matching within a caliper），以及基于矩阵运算的马氏距离匹配（Mahalanobis metric matching）。这些方法看似不同，实则原理相似，都是以距离最小化为基本原则来为处理组个体挑选匹配对象。以本研究为例，若采用近邻匹配法，就是将两个倾向分值差距最小的改革县与未改革县一对一地匹配起来。近邻匹配简单易行，但有明显的缺陷。两个距离最近的匹配未必是“最好”的匹配。譬如，样本中有一个改革县发生改革的概率为0.30，与其距离最近的未改革县发生改革的概率值仅为0.01，此二者倾向分值相差巨大，把它们强行匹配在一起明显不合适。为了克服这一缺陷，我们可以设定一个匹配半径[⑩]，为每一个处理组个体划定一个匹配范围。在这个半径范围内，如果没有任何可供匹配的控制组个体，就放弃匹配，将该处理组个体删除；如果存在可供匹配的对象，就采用近邻法挑选距离最近的作为匹配对象。马氏距离匹配的基本原理与近邻匹配相似，不同之处在于，近邻匹配是采用基于逻辑回归所得到的倾向分值来计算个体之间的距离，而马氏匹配是直接利用个体的特征变量向量来计算距离值。

第三步，对匹配后形成的样本再次进行数据平衡检验，检视匹配后留存的样本数量，对匹配策略进行优劣评价。一般来说，一个好的匹配策略应满足两个条件：一是它应该能消除控制组与处理组下所有特征变量的差异，这是首要条件；二是尽量让处理组与控制组倾向得分在分布上拥有广阔的重合区域。[⑪]重合区域越广，成功匹配的概率越高，匹配后剩余的样本数量就越多，研究的统计功效就会越强。如表1，马氏距离匹配后改革县与未改革县所有特征变量的差异都消失了，样本数量由原先的3 754个减少为428个。相比之下，近邻配合半径匹配虽然留存了更多的样本，但没有很好地实现数据平衡，改革县与未改革县依然在人均一般性转移支付、人均专项转移支付等特征变量上存在显著差异。

值得注意的是，倾向得分法的匹配结果对于研究者所采用的匹配策略具有很强的敏感性。估计倾向得分时采用不同的函数形式与估计法，实施匹配时采用不同的匹配方法[⑫]，都有可能导致匹配样本的数量、结构发生很大的变化。因此，有研究者（Murnane & Willett，2011）指出，实施倾向得分匹配最好同时采用多种数据匹配策略，观测不同匹配策略下估计结果的稳健性。倾向得分法的另一个重要缺陷在于，它只能对可观测变量做数据平衡处理，控制可观测异质性，对不可观测的异质性无能为力。如本例，除表1所列示的变量外，改革县和未改革县还可能在经济制度、财政体制安排、历史与文化传统、官员能力与执政偏好等方面存在着差异，这些变量应当在模型中得到控制，然而现有的县级数据并不提供这些变量，由此产生不可观测的异质性残值。对于这一问题，一种常用的解决方法是对数据匹配后形成的新样本再进行倍差法回归（Khandker et al.，2010）。

设定倍差法模型如以下形式：

$$（2005\text{年未实施试点改革时}）Exp_{i,05} = \alpha_{05} + \beta \cdot REFORM_{i,05} + (\eta_i + \varepsilon_{i,05}) \quad (4)$$

$$（2006\text{年已实施试点改革时}）Exp_{i,06} = \alpha_{06} + \beta \cdot REFORM_{i,06} + (\eta_i + \varepsilon_{i,06}) \quad (5)$$

在模型中，*REFORM* 表示“新机制”改革，其估计系数 β 就是我们想要的改革的处理效应。2005年所有县都未发生“新机制”改革，因此所有县的 *REFORM* 变量值都为零，2006年试点改革开始，此时改

革县的 *REFORM* 变量取值 1,未改革县取值为 0。模型的残值由两部分组成:一是 η,其下标不带有年代符号,表示不可观测且不随时间变化的异质性残值;二是 ε,这部分残值既包含了(不会产生内生性偏估)的同质性残值,也可能包含了随时间变化的异质性残值,因此其下标带有年代符号。

我们将试点前后两期相减,即用模型(5)减去(4),便可以将模型中不随时间变化的异质性残值 η 彻底消除,得到:

$$(Exp_{i,06} - Exp_{i,05}) = \Delta Exp_i = \alpha + \beta \cdot REFORM_{i,06} + \varepsilon_i \quad (6)$$

倍差法的实现过程相对简单,只需两次差分:第一次差分是用后一阶段的结果变量(因变量)和自变量减去前一阶段的结果变量与自变量,得到结果变量和自变量的两期差值。如本例,结果变量差值为:$\Delta EXP_i = EXP_{i,06} - EXP_{i,05}$;自变量差值为:$REFORM_{i,06} = REFORM_{i,06} - 0$;第二次差分是用第一次差分形成的因变量差值与自变量差值进行 OLS 回归,估计出改革的处理效应 β。如表 2,倾向得分匹配后再实施倍差法的估计结果显示,2006 年"新机制"试点改革使得农村小学和初中生均公用经费平均上升了 102.96 元和 115.07 元,该估计值要比 OLS 回归的估计值(分别为 66.42 和 79.77)高出不少。可见,传统 OLS 严重低估了"新机制"改革的效果,运用倾向得分结合倍差法可以在一定程度上纠正这一偏差。

表 2 "新机制"改革水平效应的估计结果

变量	小学		初中	
	OLS	马氏匹配 + 倍差法	OLS	马氏匹配 + 倍差法
截距	-204.96*** (19.35)	63.43*** (10.45)	-322.93*** (43.38)	106.56*** (20.61)
新机制改革	66.42*** (10.75)	102.96*** (13.59)	79.77*** (24.10)	115.07*** (26.35)
其他控制变量(略)	…	…	…	…

注:***0.01 水平上显著 **0.05 水平上显著,*0.1 水平上显著;小括号内为估计系数标准误;在回归中,我们还控制了其他一些变量,包括各县自有财力、上级各类转移支付、小学和初中在校生人数、人口数量与密度等

倾向得分结合倍差法最大的问题在于它无法解决随时间变化的异质性问题,在公式(6)残值 ε 中依然包含着未知的异质性残值。对于这一问题,可能的解决办法是:(1)在倍差法回归中尽量控制一些可反映观测对象随时间变化特征的变量,并尝试采用多种函数形式,观测估计结果在不同函数形式与控制环境下的稳健性;(2)减少数据的时间跨度,选取短时期数据进行分析,因为数据跨期越短,出现随时间变化异质性的风险就越低。当然这样做带来的负面后果是,我们只能估计改革的短期效应。从达成因果推断的效力看,倾向得分结合倍差法确实不如断点回归与工具变量法,甚至可以说,倾向得分结合倍差法只是在难以形成有效的断点或工具变量设计情况下的一种折衷的方法,但倾向得分结合倍差法依然有其科学性的一面,只要数据处理得当,研究设计合理,运用此种方法同样能产生令人信服的因果推断结论。

六、延伸讨论

在实验与准实验方法统治整个微观计量学界并盛行于社会科学各个研究领域的当下,我们还需反思这些因果推断方法是否存在被误用甚至被滥用的可能。产生这一可能的主要原因在于,当今的量化研究者过于关注估计参数的显著性表现,过于坚持估计结果与预期结果的一致性,而忽视了实现有效因果关系推断所需满足的基本条件。并不是所有冠之以"随机试验"或"准实验"的研究结论都一定具有因果推断的效力。任何一项意图达成因果推断结论的研究,都必须满足处理无关假设(the ignorable treatment assignment assumption),要求样本中所有观测对象,无论是处理组还是控制组个体,其潜在结果(potential outcome)与是否接受处理安排都应当是无关的(Rubin,1986;Heckman,2005)。

具体而言,断点回归要求所设计的断点应是外生的,它能够将断点两侧个体随机分配为处理组与控制组,以此保证结果变量取值在断点上所发生的垂直跳跃与断点之间存在着一一对应的因果关系。

工具变量法要求所构造的工具变量应满足与内生变量强相关、与异质性残值无关，以及只能通过内生变量方能对结果变量产生影响这三个假设。只有这三个条件得到满足，才能形成工具变量→内生变量→结果变量的因果关系链条。倾向得分结合倍差法只能用于解决可观测异质性与不随时间变化的不可观测异质性，其估计结果只有在不存在随时间变化异质性条件下才是无偏的。为了达成这一目的，研究者需采用多种策略组合对数据进行匹配，检视在不同匹配策略下处理组与控制组倾向得分的分布重合状况与数据平衡实现情况，对处理组在接受处理前后的变化趋势是否与控制组变化趋势保持平行变化态势进行检验分析[13]。

从当前社会科学对于人之行为及结果的研究成果来看，我们已知的远不如未知的多。一个计量模型所包含的特征变量对于结果变量变异的解释力达到40%便属于高解释力模型，而剩余60%所包含的种种未知特征变异都有可能成为形成异质性残值或内生性问题的诱因。可以说，几乎所有的准实验研究都或多或少面临着有关估计结果是否有偏的质疑，因此一个好的科学研究应当注重对方法背后所隐含假设的检验，重视估计结果在不同样本、不同估计方法、不同函数形式之下的稳健性或敏感性分析。在现有数据和技术条件下所做的最好的研究，就是最大程度地实现研究的内部有效性，准确界定研究结果所适用的人群范围，客观评价研究结果推及总体的一般化能力。研究者对研究结果的自我审查、质疑与辩伪要比获得一个显著的、与预期完全一致的估计结果具有更加重大的意义，它既是形成可靠知识的有效累积，又是推动教育研究科学化发展最为重要的原生动力。

参考文献

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long－term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *The American Economic Review*, 96(3), 847－862.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2), 533－575.

Angrist, J. D., & Steffen, P. J. (2015). *Mastering metrics: the path from cause to effect*. NJ: Princeton University Press.

Borman, G. D. (2009). The use of randomized trials to inform education policy. In Sykes, G., Schneider, B. & Plank, D. N. (Eds.). *Handbook of education policy research* (pp. 129－138). New York: Routledge.

Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power analysis of cutoff－based randomized clinical trials. *Evaluation Review*, 18(2), 141－152.

Card, D. (1999). The Causal Effect of Education on Earnings. In Ashenfelter, O. & Card, D. (Eds.). *Handbook of Labor Economics*, 3*A* (pp. 1801－1864). New YorK: Elsevier.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis*. Thousand Oaks: Sage.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153－161.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology*, 35(1), 1－97.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4), 1239－1285.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2), 467－475.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences: An introduction*. New York: Cambridge University Press.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression－discontinuity analysis. *Review of Economics and Statistics*, 86(1), 226－244.

Kaplan, D. (2009). Causal inference in non－experimental educational policy research. In Sykes, G., Schneider, B. & Plank, D. N. (Eds.). *Handbook of education policy research* (pp. 139－153). New York: Routledge.

Khandker, S. R., Koolwal, G. B., &Samad, H. A. (2010). *Handbook on impact evaluation: quantitative methods and practices*. Washington, D. C.: World Bank Publications.

Lee, D. S., &Lemieuxa, T. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281－355.

Li, H., & Luo, Y. (2004). Reporting errors, ability heterogeneity, and returns to schooling in China. *Pacific Economic Review*, 9(3), 191－207.

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford: Oxford

University Press.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33 – 38.

Rubin, D. B. (1986). Which ifs have causal inference. *Journal of the American Statistical Association*, 81, 961 – 962.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton, Mifflin and Company.

Smith, W. C. (2014). Estimating unbiased treatment effects in education using a regression discontinuity design. *Practical Assessment, Research & Evaluation*, 19(9), 2.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of educational Psychology*, 51(6), 309.

Trochim, W. M. K. (1984). *Research design for program evaluation: the regression-discontinuity approach*. CA: Sage.

黄斌,钟晓琳. (2012). 中国农村地区教育与个人收入——基于三省六县入户调查数据的实证研究. 教育研究, (3),18 – 26.

黄斌,汪栋. (2016). 中国义务教育财政投入的回顾与展望. 华中师范大学学报:人文社会科学版, 55(4),154 – 161.

黄斌,苗晶晶,金俊. (2016). "新机制"改革对农村中小学公用经费的因果效应分析——基于准实验研究设计(工作论文). 南京: 南京财经大学公共财政研究中心.

注 释:

①内部有效性是指研究结果能否真实地反映出样本中变量间的因果关系,外部有效性是指样本分析所得到的因果关系能否推广至总体。

②2015 年奥巴马政府颁布了新的法案《每一个孩子都成功法》(Every Student Succeeds Act)。新法案对老法案作出了一定的修订,包括放松联邦政府对州一级的绩效考核,改进原有严格基于学生学业成绩的学校问责与拨款制度等,但有关强调科学因果推断结论对于形成教育政策的重要性与优先地位的相关论述未变。

③这也就说是,我们运用随机分配实现了完全的控制,处理组和控制组只在是否接受处理上存在差异,其他完全相同。

④此处的自我选择可能是学生及家庭主动的选择,也可能是被动的选择。家庭背景好的学生就读的班额大小可能是自己主动寻求的结果,而家庭背景差的学生就读的班额大小可能是不得不接受、无从选择的结果。无论是自动还是被动选择,都表现为个体选择为非随机,受制于某些个体特征。

⑤也正是这个原因,断点回归又被称为"自然实验"(natural experiment)。

⑥所谓统计功效是指我们能够正确地拒绝一个错误假设,估计出真实处理效应的能力,一个研究的统计功效取决于多种因素,其中包括样本容量。

⑦之所以要采用(ENROL – 41)的中心化处理,是要保证公式(2)中的估计系数 β1 恰好等于图 1 中垂直跳跃的值(本例中该值即等于 10)。

⑧学校入学人数是否达到 41 人决定了学校在断点左右两侧的落点。

⑨注意此处必须使用改革之前的数据进行分析,以保证各自变量都发生在改革之前,都是改革的前定变量。

⑩通常以样本中倾向分值标准差的 1/4 为匹配半径。

⑪这一要求被称为共同支撑假设(common support assumption),我们可以绘制出处理组与控制倾向得分的分布图,通过观测这两个分布重复区域的大小以检测研究数据是否满足共同支撑假设。

⑫对于预测概率估计我们可以选择逻辑回归、probit 回归或更加复杂但更具有稳健性的广义自举回归(Generalized Boosted Modeling,GBM),函数形式可以选择线性或非线性形式,对于数据匹配我们可以选择近邻匹配、半径匹配、近邻结合半径匹配、马氏(Mahalanobis)距离匹配、核匹配(Kernel – based matching),等等。不同匹配策略各具优缺点,所形成的匹配样本数量亦常常有很大差别。具体讨论参见 Guo & Fraser(2010)与 Imbens& Rubin(2015)。

⑬如果不存在随时间变化异质性,那么所有不可观测异质性对于处理组和控制组结果变量的影响都不会随时间变化,于是我们就应观测到在接受处理前后处理组与控制组的结果变量应当具有相同的随时间变化趋势。该假设被称为平行趋势假设(parallel trend assumption)。

(责任编辑 童想文)