2025年第1期 中山大学学报（社会科学版） No. 1 2025
第65卷 JOURNAL OF SUN YAT-SEN UNIVERSITY Vol. 65
（总313期） （SOCIAL SCIENCE EDITION） General No. 313

# 计量经济学中的因果推断：过去、现在与未来*

陈　强

**摘　要**：自从可信度革命以来，因果推断在计量经济学中的地位日益重要，因果推断方法也层出不穷，甚至令实证研究者眼花缭乱。本文系统梳理了因果推断在计量经济学中的历史渊源与发展脉络，涵盖随机实验、自然实验、断点回归、匹配估计、双稳健估计、工具变量法、双重差分法、合成控制法、回归控制法、分位数控制法等主流因果推断方法。与已有的文献综述相比，本文更注重因果推断方法的历史起源，以帮助实证研究者更好地把握其精神实质，同时将所涉文献更新至2024年的最新前沿（尤其是工具变量法与双重差分法的最新进展），并展望未来的发展方向。

**关键词**：计量经济学；因果推断；可信度革命
DOI：10.13471/j.cnki.jsysusse.2025.01.008

## 引　言

1930年12月，由耶鲁大学Irving Fisher、奥斯陆大学Ragnar Frisch与康奈尔大学Charles Roos等发起的计量经济学会（the Econometric Society）在美国俄亥俄州克利夫兰市成立。1932年，从事投资咨询的美国商人Alfred Cowles在科罗拉多州建立考尔斯委员会（the Cowles Commission），给予计量经济学会急需的财力与人力资助，最初预算为每年1.2万美元①。1933年1月，计量经济学会会刊*Econometrica*正式出版，并在发刊词首创新词“econometrics”（Frisch，1933），宣告计量经济学诞生。

然而，当时世界范围内仍鲜有从事计量理论与应用的研究者。在计量经济学草创之初，既无锣鼓开道，也无鲜花喝彩，反而不乏重量级的质疑者。1939年，荷兰经济学家丁伯根应用多元回归方法，研究经济周期中的投资决定（Tinbergen，1939）②。同年，当时如日中天的英国经济学家凯恩斯在其主编的*Economic Journal*发表评论文章（Keynes，1939），指出丁伯根实证研究的诸多缺陷，由此引发激烈辩论（Qin，1993，p. 20）。针对丁伯根的宏观计量实证研究，凯恩斯的质疑包括：（1）线性回归模型的误设问题；（2）遗漏变量问题；（3）结构变动问题（经济环境的稳定性问题）；（4）双向因果问题（伪相关问题）；（5）变量的度量误差问题；（6）时间滞后阶数的选择问题；（7）时间趋势项的选择问题。更重要的，凯恩斯认为丁伯根仅满足于对统计关系的描述，而实证研究的终极目的是“归纳泛化”（inductive generalization），即现在常说的因果推断（causal inference）。虽然凯恩斯并非计量经济学家，但其批判性评论无疑切中了计量经济学的要害。与其说这是凯恩斯对于初生的计量经济学的怀疑，倒不如说是指出了计量经济学面临的挑战与发展蓝图。唯有克服这些问题，计量经济学才能成为可信且令人尊敬的学科。

* **收稿日期**：2024—06—11
**作者简介**：陈强，山东大学经济学院（济南 250100）。

① 有关计量经济学会与考尔斯委员会的详细起源，可参见Christ（1952）。

② Jan Tinbergen与Ragnar Frisch于1969年因为对于计量经济学的贡献而获得首届诺贝尔经济学奖。

初创的计量经济学有太多的基础理论问题需要解决,而考尔斯委员会在此阶段发挥了核心作用。1939年9月,考尔斯委员会从科罗拉多州迁往芝加哥大学,吸引了更多优秀计量经济学家加盟[①]。在这一阶段与考尔斯委员会关系密切而后来获得诺贝尔经济学奖的计量经济学家包括Tryve Haavelmo,Leonid Hurwicz,Lawrence Klein,Tjalling Koopmans与James Tobin(Christ, 1952)。此时考尔斯委员会主要使用时间序列进行宏观经济学的建模,而理论研究则集中于"联立方程组"(simultaneous equations)的识别、估计与推断[②]。计量经济学中常见的"内生变量"(endogenous variable)与"外生变量"(exogenous variable)的区分即源于此。到了20世纪50年代中期,考尔斯委员会的计量方法已成为当时的主流计量理论。

然而,康奈尔大学华裔经济学家刘大中(Liu, 1960)指出,为了识别联立方程组而施加的排他性约束常常是"人为的"(artificial),而经济理论则一般要求在回归方程中放入更多变量。换言之,这些联立方程组很可能"不可识别"(underidentified),而相应的估计存在偏差,并不能揭示真正的因果关系。2000年诺贝尔经济学奖获得者、芝加哥大学经济学家Heckman(2000)指出,到了20世纪60年代中期,考尔斯委员会的计量方法被广泛地视为"理论成功,但实证失败"(an intellectual success but an empirical failure)。到了20世纪80年代,这种状况仍未根本改观。例如,加州大学洛杉矶分校经济学家Edward Leamer(1983, p. 37)不无讽刺地指出,"几乎无人把数据分析当真。或许更准确地说,几乎无人把别人的数据分析当真"[③]。值得一提的是,1980年6月,宾夕法尼亚大学经济学家克莱因(Lawrence Klein)率领7位美国经济学家,在颐和园举行了为期7周的计量经济学讲习班,首次将计量经济学传入中国。

幸运的是,大约从1990年开始,计量经济学的实证研究越来越重视通过"好的研究设计"(a good research design),寻找数据中合适的"外生变动"(exogenous variation)来识别因果关系。这些外生变动的来源,可能来自随机实验、自然实验或准实验(例如断点回归)、工具变量,或处理组与控制组的某种相似性(例如双重差分法的平行趋势假定)。Heckman(2000)将这种趋势称为"自然实验运动"(natural experiment movement),而Angrist & Pischke(2010)则更乐观地称之为"可信度革命"(a credibility revolution),因为实证研究的可信度有了质的飞跃,而Leamer(1983)的抱怨不再成立。

显然,因果推断的方法在可信度革命中起到了核心作用。为此,本文其余部分将系统梳理计量经济学中主流因果推断方法的历史渊源与发展前沿,涵盖随机实验、自然实验、断点回归、匹配估计、双稳健估计、工具变量法、双重差分法、合成控制法、回归控制法、分位数控制法等主流因果推断方法。与已有的文献综述相比(例如Imbens & Wooldridge, 2010; Abadie & Cattaneo, 2018),本文更注重因果推断方法的历史起源,以帮助实证研究者更好地把握其精神实质,同时将所涉文献更新至2024年的最新前沿,并展望未来发展方向。

## 一、随机实验

1919年,一位毕业于剑桥大学、年仅29岁的统计学家罗纳德·费雪(Ronald Fisher)携家带口来到位于伦敦北部的罗森斯特农业实验站(Rothamsted Agricultural Experimental Station),他的工作职责是分析

① 1955年7月考尔斯委员会迁往耶鲁大学,并延续至今,参见https://cowles.yale.edu/about-us/move-yale。截至2018年,与考尔斯委员会有关联的诺贝尔经济学奖得主达到27名,参见https://cowles.yale.edu/archives/nobel-laureates。

② 考尔斯委员会的联立方程计量理论有时称为"结构式"(structural form)或"结构计量经济学"(structural econometrics)。

③ 原文为"Hardly anyone takes data analysis seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analysis seriously."

该农业站自1843年创立以来积累的农业试验数据。在费雪到来之前,农业站的试验方法可简化如下。譬如想知道某化肥对农作物产量的作用,则将一大块农地一分为二,其中一块施肥,而另一块不施肥,然后比较二者的产量差异①。然而,费雪很快沮丧地发现,试图从农业站70多年积累的"大数据"得到科学结论,无异于"在粪堆中搜索"(raking over the muck heap)②,因为化肥对产量的作用存在很多混杂因素(confounders),例如不同地块的肥力不同,而土壤肥力无法精确度量。无奈之下,费雪只得另起炉灶,并找到了一个天才的解决方案,即"随机化"(randomization)。这意味着,将一大块地分成许多较小的方块,然后随机决定每个小方块是否施肥。随机化切断了混杂因素的干扰,使得施肥的小方块(构成处理组)与不施肥的小方块(构成控制组)不再有任何系统差别,而二者平均产量的差异就是施肥的平均处理效应,参见图1。

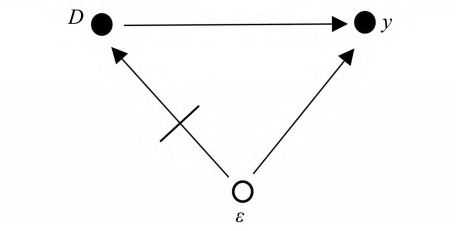


图1 通过随机化切断混淆变量的影响

在图1中,由于存在不可观测(以空心圆表示)的混淆变量$\varepsilon$(比如土壤肥力等),同时影响可观测(以实心圆表示)的处理变量$D$(是否施肥)与结果变量$y$(农作物产量),故即使化肥对农作物产量无因果效应,二者也依然存在相关关系。混淆变量$\varepsilon$究竟是什么以及如何度量,研究者通常无从知晓。破解此遗漏变量偏差的巧妙方法是对于处理变量$D$的取值进行随机分配(randomly assigned),从而切断任何混淆因素对于处理变量$D$的可能影响。在理论上,处理变量$D$的取值只与随机分组的细节有关(比如随机数的种子),而与宇宙中的任何其他事物无关。如果使用线性回归方程表示,则有:

$$y_i = \alpha + \beta D_i + \varepsilon_i \tag{1}$$

其中,下标$i$表示个体,比如第$i$个地块。由于处理变量$D_i$的取值随机分配,故与扰动项$\varepsilon_i$不相关,使用普通最小二乘法(OLS)即可得到对因果效应$\beta$的一致估计。可以证明,OLS估计量可写为处理组与控制组平均结果之差:

$$\hat{\beta}_{OLS} = \bar{y}_{treat} - \bar{y}_{control} \tag{2}$$

这就是所谓"差分估计量"(the difference estimator)。1925年,费雪首次在其著作《研究工作者的统计方法》中解释了随机实验(randomized experiment或randomized control trial)的原理(Fisher, 1925)。1935年,费雪在其经典著作《实验的设计》中(Fisher, 1935),全面深入地阐述了随机实验的理论与方法。费雪首倡的随机实验方法很快在农学领域得到普及,并进而推广到药学、化学、工业质量控制、经济学等领域,被誉为因果推断的"黄金标准"(gold standard)。

从20世纪80年代中期开始,哈佛大学经济学家Michael Kremer与其合作者在非洲肯尼亚进行了一系列有关教育与健康的随机实验。当时肯尼亚乡村小学普遍教材奇缺,而世界银行正计划给肯尼亚贷款,以增加教材供给。但提供教材真能改进教育质量吗?为此,研究者从100所肯尼亚小学中随机抽取25所(构成处理组),由政府提供免费教材;而其余75所小学则构成控制组。随机实验的结果显示,提供教材可显著提升优秀学生(以实验前成绩衡量)的成绩,但对于其他学生则影响甚微(Glewwe et al., 2009)。可能原因包括,教材为英文教材,而英语只是多数学生不熟悉的第三语言,且肯尼亚小学课程体系偏向于优秀学生。Michael Kremer与其合作者在肯尼亚进行的其他随机实验包括为学校提供科学挂图(Glewwe et al., 2004)、为学生除蠕虫(Miguel & Kremer, 2004)以及提供免费早餐(Vermeersch &

① 农业站实际进行的试验更为复杂。例如,根据化肥的量(一份或两份化肥)以及施肥时间(二月、三月或五月),使用6个不同的地块,再加上一个不施肥的地块;参见Yates (1964)。

② 参见Salsburg (2001),第33页。

Kremer, 2005),结果发现这些干预均无法显著提升学生成绩,尽管除蠕虫与免费早餐可显著减少旷课。

在肯尼亚进行的上述田野实验(field experiments)表明,单纯地增加学校的资源投入未必能提高教学质量。基于这些发现,麻省理工学院经济学家Abhijit Banerjee、Esther Duflo与合作者在印度开展随机实验,考察为差生增设补习项目的效应。其中一个项目从社区雇佣年轻女士为差生提供专门的语文与数学补习,而另一项目则通过计算机提供数学游戏。结果发现,这两个项目均显著提升了学生成绩,尽管其政策效应在项目结束一年后即出现了衰减(Banerjee et al., 2007)。至此,随机实验在发展经济学中的应用已呈爆炸趋势(Banerjee & Duflo, 2009),而有些学者对于实验方法的推崇几近宗教狂热(Heckman, 2020)。2019年,Abhijit Banerjee,Esther Duflo与Michael Kremer因使用实验方法研究如何降低全球贫困而获得诺贝尔经济学奖①。

然而,将起源于农学的随机实验推广至社会经济领域,必然遇到新的挑战(Banerjee & Duflo, 2009; Heckman, 1992、2020)。例如,无论如何将地块随机分组,都不会改变土壤、化肥的性质;但将个人或组织进行随机分组,则可能改变其行为模式,即所谓"霍桑效应"(Hawthorne effect)。另外,尽管理想的随机实验具有很强的"内部有效性"(internal validity),但社会经济实验的结果还依赖于其所处的制度环境,故"外部有效性"(external validity)可能存在局限。事实上,社会经济现实中的随机实验通常并不完美,实验参与者未必遵从实验设计(non-compliance)或中途退出(attrition),导致其内部有效性也受到威胁②。

随机实验对于因果推断的深远影响还体现在,它催生了"潜在结果"(potential outcomes)的反事实框架(counterfactual framework)。1923年,年仅29岁的波兰裔统计学家内曼(Jersy Neyman)考察在$m$个地块播种$v$类种子的随机实验,并引入了"潜在产出"(potential yield)的概念,以$U_{ik}$表示地块$k$($k = 1, \cdots, m$)播种第$i$类种子($i = 1, \cdots, v$)的产出(Neyman, 1923[1990])。显然,每个地块只能播种一类种子,其相应的$U_{ik}$为观测结果,但其余的$U_{ik}$则为不可观测的潜在结果。然而,Neyman (1923)仅在随机实验的框架下探讨潜在结果,并未推广到一般的观测数据(Imbens & Rubin, 2015, p. 23)。直到半个世纪后,美国统计学家唐纳德·鲁宾(Donald Rubin)才完成这步关键的飞跃(Rubin, 1974)。彼时,鲁宾仅三十出头,不久前刚从哈佛大学获得统计学博士学位。Rubin (1974)将潜在结果放在了因果推断的核心位置,并适用于更一般的观测数据。记个体$i$的处理变量为$D_i \in \{0, 1\}$,其中1表示受政策冲击,0表示未受处理;而结果变量为$y_i$。进一步,以潜在结果$y_{0i}$表示个体$i$未受处理的潜在结果,而$y_{1i}$表示个体$i$受处理的潜在结果,则个体$i$的处理效应可定义为:

$$\tau_i \equiv y_{1i} - y_{0i} \tag{3}$$

显然,由于个体只能处于一种状态,故只能观测到$y_{0i}$或$y_{1i}$,而无法同时观测二者。这实际上是一种数据缺失(missing data)问题,也正是"因果推断的基本问题"(Holland, 1986)。具体而言,若个体$i$受处理,则$y_{1i}$可观测,而$y_{0i}$为不可观测的反事实结果。不同因果推断方法的区别主要在于,如何在一定的合理假定下估计$\hat{y}_{0i}$,从而得到因果效应的估计值$(y_{1i} - \hat{y}_{0i})$。时至今日,基于反事实框架的因果推断方法,早已成为经济学中因果推断的主流方法,并统称为"鲁宾因果模型"(Rubin causal model)。

① 诺贝尔奖官网给出的获奖理由为"for their experimental approach to alleviating global poverty",参见https://www.nobelprize.org/prizes/economic-sciences/2019/summary。

② 对于前者(未遵从实验设计),可使用工具变量法进行估计,参见第五节。对于后者,如果处理组与控制组的退出率(attrition rate)显著不同,可能导致样本中个体的处理状态不再为随机分配。

## 二、自然实验

随机实验的最大局限性或许在于,有许多重要的社会经济问题无法通过随机实验得到回答,比如宏观经济的运行规律,或道德成本高昂的微观问题(比如吸烟对肺癌的作用)。为此,有些研究者开始关注"自然实验"(natural experiment)或"准实验"(quasi experiment)(Meyer, 1995),即虽然研究者只有观测数据(observational data),但由于某种自然发生的外部突发事件(例如政府政策变化或自然现象),使得处理组与控制组仿佛被随机分配(as if randomly assigned)。20世纪90年代,加州大学伯克利分校经济学家David Card使用自然实验进行了一系列颠覆性研究,包括移民以及最低工资立法对劳动力市场的影响。由于1980年古巴政经危机,约12.5万古巴人从马里矣尔港乘船涌入美国,即所谓"马里矣尔船运"(Mariel boatlift);其中半数定居于迈阿密,使得迈阿密的劳动力增加7%。Card (1990)以1980年古巴移民潮对迈阿密劳动力市场的外生冲击作为自然实验,并以四个与迈阿密相似但未受移民冲击的美国城市构成控制组。通过比较案例分析(comparative case study)发现,马里矣尔船运对于迈阿密劳动力市场的工资与失业率几乎没有影响。

1992年,美国新泽西州通过法律将最低工资(minimum wage)从每小时$4.25提高到$5.05,但在相邻的宾夕法尼亚州最低工资却保持不变。Card & Krueger (1994)慧眼识珠,将其视为最低工资影响低技能劳动力需求的自然实验。在新泽西州实施新法前,由于新泽西州与相邻的宾夕法尼亚州东部的经济状况相似,且经贸关系密切,故可将这两个地区的快餐店视为来自同一总体。在新泽西州实施新法后,这两个地区的快餐店则仿佛被随机分配到处理组(新泽西州)与控制组(宾夕法尼亚州东部)。双重差分法的估计结果显示,提高最低工资对于雇佣人数的处理效应并不显著,不支持传统理论所预期的负效应。2021年,David Card因为使用自然实验研究劳动力市场而获得诺贝尔经济学奖的一半奖金[①]。

事实上,自然实验的思想方法已被应用于经济学的不同领域,包括宏观经济学(Fuchs-Schündeln & Hassan, 2016)与经济史(Cantoni & Yuchtman, 2021)。当然,自然实验毕竟不是随机实验,其处理状态也并非由研究者所随机分配,故处理组与控制组未必具有可比性(Sekhon & Titiunik, 2012)。回到Card & Krueger (1994)的最低工资案例,在新泽西州实施新法前,新泽西州与宾夕法尼亚州东部的快餐店特征也不完全相同;例如,新泽西州快餐店"全餐"(full meal)的平均价格显著高于宾夕法尼亚州东部的快餐店(Card & Krueger, 1994, p. 775)。因此,自然实验的可信度一般弱于随机实验。这也提示我们,如果自然实验将个体分为两组,则研究者可通过考察协变量在处理组与控制组的分布是否相同,来判断自然实验偏离随机实验的程度。另外,在经济学文献中,有时也将任何外生冲击均视为自然实验(Rosenzweig & Wolpin, 2000),而不要求存在明显的处理组与控制组(即处理变量未必是虚拟变量)。

## 三、局部随机实验与断点回归

1960年,美国心理学家Donald Thistlethwaite与Donald Campbell发表论文,提出一种新的准实验方法,即"断点回归"(regression discontinuity),并以此研究奖学金对于学业的影响。Thistlewaite & Campbell (1960)的样本由1957年参加美国奖学金竞赛的高中生组成,其中5,126名获优秀奖,而另外2,848名仅收到鼓励信。学生究竟获得优秀奖或鼓励信,完全取决于其CEEB SQT的考试成绩[②]。因此,成绩刚好达到获奖标准与差点达到的学生具有可比性,可作为对方的反事实结果。事实上,由于学生无法精确控制其成绩,故在断点(即获奖分数线)附近的学生可视为随机分组,构成在断点附近的"局部随机实

① 2021年诺贝尔经济学奖的另一半奖金由Joshua Angrist与Guido Imbens所分享,参见第五节。

② 该考试的全称为"College Entrance Examination Board Scholarship Qualifying Test"。

验”(local randomized experiments)。结果发现,优秀奖获得者更可能得到其他渠道的奖学金,但对于其对学术的态度及职业生涯计划则无显著影响。但Thistlewaite & Campbell (1960)开创的断点回归在很长时间里并未引起学界的广泛关注,仅在心理学、教育学、统计学等领域缓慢发展,部分原因是断点回归在这些领域的应用场景较少(Cook, 2008)。自20世纪90年代末开始,经济学家注意到断点回归在经济学领域的巨大应用价值,才使得断点回归重获新生,并趋于成熟。

断点回归设计包括三大要素,即驱动变量(running variable)、断点(cutoff)与不连续的处理配置规则(discontinuous treatment assignment rule)。假设数据为横截面的随机样本$\{y_i, D_i, x_i\}_{i=1}^{n}$,其中$y_i$,$D_i$与$x_i$分别为结果变量、处理变量与驱动变量。断点回归的处理配置规则为[①]:

$$D_i = \mathbf{1}(x_i \geqslant c) \tag{4}$$

其中,$c$为某已知断点,而$\mathbf{1}(\cdot)$为示性函数(若括弧中表达式成立,则返回值为1;反之,则为0)。断点回归之所以在经济学大放光芒,根本原因在于许多经济现象的分组规则均由貌似随意的断点所决定,例如班级规模超过40名学生须分成两个班(Angrist & Lavy, 1999),学区的地理分界线(Black, 1999),获得奖学金的分数线(Van Der Klaauw, 2002),议会选举得票超过最强竞争对手即可当选议员(Lee, 2008),1992年农村人均纯收入不足700元的贫困县即为扶贫对象(Meng, 2013),在济南市区购置建筑面积90平米以上商品住宅即可获得济南户籍(Chen et al., 2019)。

早期的断点回归一般在断点两侧分别进行多项式回归,但可能因函数误设而导致偏差。Hahn et al. (2001)在潜在结果的框架下,首次证明了断点回归的非参数识别(nonparametric identification)条件。记个体$i$的两个潜在结果分别为$y_{0i}$与$y_{1i}$,则在断点$c$处的局部平均处理效应(Local Average Treatment Effect,简记LATE)可定义为:

$$\tau \equiv \mathrm{E}[y_{1i} - y_{0i} | x_i = c] \tag{5}$$

假定驱动变量在断点附近的密度为正,且潜在结果的条件期望函数$\mathrm{E}[y_{0i}|x_i]$与$\mathrm{E}[y_{1i}|x_i]$在断点$x_i = c$处连续,Hahn et al. (2001)证明:

$$\tau = \lim_{x_i \downarrow c} \mathrm{E}(y_i|x_i) - \lim_{x_i \uparrow c} \mathrm{E}(y_i|x_i) \tag{6}$$

其中,$\lim_{x_i \downarrow c}$与$\lim_{x_i \uparrow c}$分别表示在$x_i = c$处的右极限与左极限,参见图2。Hahn et al. (2001)的核心假设是潜在结果的条件期望在断点处连续,故称为“基于连续性的分析框架”(continuity-based framework),是目前断点回归的主流方法。

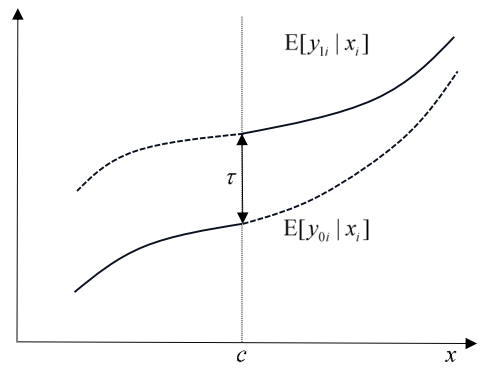



图2 基于连续性的分析框架

为了估计方程(6)中的左极限与右极限,Hahn et al. (2001)建议使用非参数的局部线性回归(local linear regression)。Imbens & Kalyanaraman (2012)与Calonico et al. (2014)提出了使估计量均方误差(MSE)最小化的最优带宽(optimal bandwidth),而后者为前者的改进。然而,由于非参数估计使用了断点附近的观测值,故一般存在偏差。尽管此偏差将在大样本下消失(故不影响一致性),但仍会影响估计量的渐近分布。为此,Calonico et al. (2014)提出了“偏差校正”(bias correction)的方法,即先使用MSE最优带宽得到

① 限于篇幅,本文仅关注“精确断点回归”(sharp regression discontinuity)。在断点处,若个体受处理的概率从$a$跃升为$b$,其中$0 < a < b < 1$,则为“模糊断点回归”(fuzzy regression discontinuity)。

点估计$\hat{\tau}$,然后估计其偏差$\hat{B}$,再将$(\hat{\tau}-\hat{B})$作为偏差校正的估计量。由于对点估计的偏差校正本身也会带来不确定性,故偏差校正估计量的方差还须进行调整,由此可构建偏差校正稳健(robust bias-corrected)的置信区间。

从方程(6)及图2可知,断点回归(尤其是基于连续性的分析框架)仅识别在断点处的局部平均处理效应,故尽管其内部有效性较强,但外部有效性较弱。另外,为了保证断点附近两侧个体的可比性,还须排除"内生分组"(endogenous sorting)的可能性。为此,McCrary (2008)提出使用非参数方法检验驱动变量$x_i$的密度函数在断点处是否连续,即所谓"密度检验"(density test)。如果个体可通过自身努力而完全操控驱动变量(complete manipulation),则可自行选择进入处理组或控制组,导致内生性。Cattaneo et al. (2020)使用局部多项式密度估计量(local polynomial density estimator),得到了更有效率的密度检验,也称为"操纵检验"(manipulation test)。

尽管连续性框架是目前断点回归的主流方法,但实证研究者常非正式地将断点回归视为局部随机实验。为此,Cattaneo et al. (2015, 2017)提出"局部随机化的框架"(local randomization framework),通过一系列协变量平衡检验来选择断点附近的小窗口,使得在此小窗口内个体的驱动变量及处理状态可视为近乎随机分配(as-if randomly assigned),然后使用分析随机实验的方法进行估计与推断。由于局部随机化框架出现较晚,目前仍主要作为断点回归的替补方法或稳健性检验。刘冲等(2022a)与陈强等(2024a)详细比较了断点回归两大分析框架的优缺点。局部随机化框架的优势在于所选窗口一般更窄,故驱动变量的外生性条件更易满足,且适用于离散驱动变量的情形;但在更窄的窗口内有效样本容量可能较小,不便于统计推断。

## 四、条件随机实验与匹配估计

Rubin (1974)提出因果推断的反事实框架,也适用于观测数据,但具体如何应用仍有待落实。假设数据为横截面的随机样本$\{y_i, D_i, \mathbf{x}_i\}_{i=1}^{n}$,其中$\mathbf{x}_i$为处理前的一些协变量(pretreatment covariates),通常为个体$i$的某些特征。在一定条件下,如果个体$i$与个体$j$的特征相同或很接近,即$\mathbf{x}_i \approx \mathbf{x}_j$,则二者就有可比性,可作为对方的反事实。更严格地,Rosenbaum & Rubin (1983)引入"可忽略性"(ignorability)假定,即在给定协变量$\mathbf{x}_i$的条件下,处理变量$D_i$独立于潜在结果$(y_{0i}, y_{1i})$:

$$D_i \perp (y_{0i}, y_{1i}) | \mathbf{x}_i \tag{7}$$

其中,"⊥"表示相互独立。此假定也被称为"非混杂性"(unconfoundedness),"条件独立假定"(conditional independence assumption)或"依可测变量选择"(selection on observables)。由于个体存在自我选择,故$D_i$一般与$(y_{0i}, y_{1i})$相关。但在可忽略性假定之下,一旦给定个体特征$\mathbf{x}_i$,则$D_i$与$(y_{0i}, y_{1i})$不再相关。这意味着,给定$\mathbf{x}_i$,则$D_i$的取值可视为随机分配(as good as randomly assigned conditional on $\mathbf{x}_i$),故相当于"条件随机实验"(conditionally randomized trial)或"分层随机实验"(stratified randomized experiment)。直观上,对于$\mathbf{x}_i$完全相同的个体,其进入处理组或控制组完全随机地决定(例如抛硬币或抽签),构成随机分组。基于此条件随机实验,估计处理效应的一个自然方法为"匹配估计量"(matching estimator)。具体而言,假设个体$i$属于处理组,寻找控制组的某个体$j$,使得$\mathbf{x}_i \approx \mathbf{x}_j$,则可将$y_j$作为$y_{0i}$的估计量,即$\hat{y}_{0i} = y_j$,由此得到对个体$i$处理效应的估计$(y_i - \hat{y}_{0i}) = y_i - y_j$。将处理组每位个体的处理效应估计值进行平均,即可得到"处理组平均处理效应"(average treatment effect on the treated,简记ATT)。类似地,对控制组每位个体也可进行匹配,并计算相应的处理效应。若对样本中每位个体的处理效应估计进行平均,则可得到"平均处理效应"(average treatment effect,简记ATE)。

为了保证$\mathbf{x}_i \approx \mathbf{x}_j$,可使用马氏距离(Mahalanobis distance)度量二者的距离,即$\sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}(\mathbf{x}_i - \mathbf{x}_j)}$,其中$\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}$为协变量的样本协方差矩阵之逆矩阵。马氏距离是欧氏距离(Euclidean distance)的推广,也称为"统计距离"(statistical distance)。使用马氏距离进行匹配也称为"近邻匹配"(nearest neighbors matching)。具体匹配方法则包括一对一匹配(最近邻匹配)、一对多匹配($K$近邻匹配)、卡尺匹配(也称半径匹配)及卡尺内$K$近邻匹配等。一对一匹配仅使用最近的邻居,故偏差最小,但方差较大;而一对多匹配使用了更远的邻居,故偏差较大,但方差更小。为了降低均方误差,需考虑偏差与方差之间的权衡(比如一对四匹配)。$K$近邻匹配仅限制匹配的个数,而最近的邻居也可能相去甚远,故卡尺匹配限制邻居的绝对距离,而卡尺内$K$近邻匹配则同时限制邻居的距离与个数。此外,在匹配时还可选择"有放回"(with replacement)或"无放回"(without replacement);一般认为前者更有效率。由于近邻匹配存在偏差,还可进行偏差校正(Abadie & Imbens, 2011)。

为满足可忽略性假定,$\mathbf{x}_i$通常需包括较多变量。故若直接以$\mathbf{x}_i$进行匹配,可能遇到数据稀疏的问题,这是"维度诅咒"(curse of dimensionality)的一种表现。为此,Rosenbaum & Rubin (1983)提出使用"倾向得分"(propensity score)进行匹配,称为"倾向得分匹配"(propensity score matching)。其中,个体$i$的倾向得分为,在给定$\mathbf{x}_i$的情况下,个体$i$进入处理组的条件概率,即$p(\mathbf{x}_i) \equiv \mathrm{P}(D_i = 1|\mathbf{x} = \mathbf{x}_i)$,可使用形式灵活的Probit或Logit进行估计。显然,倾向得分$p(\mathbf{x}_i)$将多维(甚至高维)的$\mathbf{x}_i$压缩到一维的[0, 1]区间。Rosenbaum & Rubin (1983)证明了重要的倾向得分定理,即如果可忽略性假定成立,则只要给定$p(\mathbf{x}_i)$,则$D_i$就独立于$(y_{0i}, y_{1i})$:

$$D_i \perp (y_{0i}, y_{1i}) | p(\mathbf{x}_i) \tag{8}$$

上式意味着,只要倾向得分相同的个体就具有可比性,而无须要求$\mathbf{x}_i \approx \mathbf{x}_j$。这使得倾向得分匹配的适用范围扩大,在实践中广为流行。另外,倾向得分还具有"平衡性质"(balancing property),即对于倾向得分相同的个体而言,其协变量在处理组与控制组的分布相同。因此,在进行倾向得分匹配后,可据此进行"数据平衡检验"(data balancing test),考察匹配样本中(matched sample),处理组与控制组的协变量分布特征(例如均值与方差)是否接近。

匹配估计量并非是基于可忽略性的唯一估计方法。早在20世纪50年代,两位美国统计学家即提出通过加权的方式,将原本不平衡的处理组与控制组人为地变得平衡(Horvitz & Thompson, 1952)。若个体$i$的倾向得分为$p(\mathbf{x}_i) = 0.2$,则具备特征$\mathbf{x}_i$的那些个体大约只有五分之一进入处理组,而其余约五分之四则进入控制组,导致协变量$\mathbf{x}_i$在处理组与控制组之间的分布不平衡。此时,可将倾向得分的倒数作为权重,即所谓"逆概加权"(inverse probability weighting,简记IPW)来构造平衡的"伪总体"(pseudo-population)。例如,对于处理组中具备特征$\mathbf{x}_i$的稀有个体,给予权重1/0.2;而对于控制组中具备特征$\mathbf{x}_i$的常见个体,则给予权重1/0.8。在可忽略性假定下,可以证明,平均处理效应$\mathrm{ATE} \equiv \mathrm{E}(y_{i1} - y_{i0}) = \mathrm{E}(w_i y_i)$,其中个体$i$的权重$w_i$为:

$$w_i = \begin{cases} \dfrac{1}{p(\mathbf{x}_i)} & 若D_i = 1 \\ -\dfrac{1}{1 - p(\mathbf{x}_i)} & 若D_i = 0 \end{cases} \tag{9}$$

在上式中代入$p(\mathbf{x}_i)$的估计值$\hat{p}(\mathbf{x}_i)$,即可得权重估计值$\hat{w}_i$,以及相应的逆概加权估计量$\widehat{\mathrm{ATE}}_{ipw} = \dfrac{1}{n}\sum_{i=1}^{n}\hat{w}_i y_i$。类似地,也可得到对于处理组平均处理效应(ATT)的逆概加权估计。显然,如果倾向得分$p(\mathbf{x}_i)$的函数形式设定有误,则逆概加权估计量将存在偏差。另外,若权重分母中的$p(\mathbf{x}_i)$很接近于0或1,可能导致逆概加权估计量的不稳定。

如果对于倾向得分的函数形式没把握,还可使用“回归调整”(regression adjustment,简记RA)估计量。在可忽略性假定下,根据迭代期望定律(law of iterated expectation),可将平均处理效应写为:

$$\begin{aligned}\text{ATE} \equiv \text{E}(y_{1i}) - \text{E}(y_{0i}) &= \text{E}_{\mathbf{x}_i}\left[\text{E}(y_{1i}|\mathbf{x}_i)\right] - \text{E}_{\mathbf{x}_i}\left[\text{E}(y_{0i}|\mathbf{x}_i)\right] \\ &= \text{E}_{\mathbf{x}_i}\left[\text{E}(y_i|D_i = 1, \mathbf{x}_i)\right] - \text{E}_{\mathbf{x}_i}\left[\text{E}(y_i|D_i = 0, \mathbf{x}_i)\right]\end{aligned} \tag{10}$$

其中,$\mu_1(\mathbf{x}_i) \equiv \text{E}(y_i|D_i = 1, \mathbf{x}_i)$为处理组的结果回归(outcome regression),即使用处理组数据,将结果变量$y_i$对协变量$\mathbf{x}_i$进行回归,记所得估计结果为$\hat{\mu}_1(\mathbf{x}_i)$。类似地,$\mu_0(\mathbf{x}_i) \equiv \text{E}(y_i|D_i = 1, \mathbf{x}_i)$为控制组的结果回归,记所得估计结果为$\hat{\mu}_0(\mathbf{x}_i)$。在方程(10)中以样本均值替代期望,即可得回归调整估计量$\widehat{\text{ATE}}_{RA} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(\mathbf{x}_i) - \hat{\mu}_0(\mathbf{x}_i)\right]$;类似地,可计算$\widehat{\text{ATT}}_{RA}$。显然,若结果回归的函数设定有误,则回归调整估计量将出现偏差。

20世纪90年代,哈佛大学生物统计学家James Robins与其合作者巧妙地将IPW与RA估计量相结合,提出了具有“双稳健性”(doubly robustness)的估计量(Robins et al. , 1994):

$$\widehat{\text{ATE}}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i y_i}{\hat{p}(\mathbf{x}_i)} - \frac{D_i - \hat{p}(\mathbf{x}_i)}{\hat{p}(\mathbf{x}_i)}\hat{\mu}_1(\mathbf{x}_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - D_i)y_i}{1 - \hat{p}(\mathbf{x}_i)} + \frac{D_i - \hat{p}(\mathbf{x}_i)}{1 - \hat{p}(\mathbf{x}_i)}\hat{\mu}_0(\mathbf{x}_i)\right] \tag{11}$$

双稳健性意味着,只要倾向得分函数$p(\mathbf{x}_i)$或结果回归(即$\mu_1(\mathbf{x}_i)$与$\mu_0(\mathbf{x}_i)$)二者之一设定正确,则该估计量就是一致估计。与IPW及RA估计量相比,双稳健估计量相当于“双保险”,有更多机会得到一致估计。当然,如果倾向得分函数与结果回归均设定错误,则双稳健估计量依然不一致。逆概加权估计、回归调整估计,特别是双稳健估计,在最近兴起的交叠DID模型中有着重要应用,参见第六节。

## 五、工具变量法

如果观测数据无法被视为自然实验、局部随机实验(断点回归)或条件随机实验(满足可忽略性假定),则工具变量法就是解决内生性的一种通用方法。20世纪20年代末,美国经济学家Philip Wright出版了专著《动物与植物油的关税》(Wright, 1928)。该书附录指出,仅使用价格与销量的数据,无法估计需求或供给函数(因为存在双向因果,即联立方程偏差),并首次提出使用工具变量回归(instrumental variable regression,简记IV)来解决内生性[①]。对于需求方程,Wright (1928)使用替代品的价格(price of a substitute)作为价格的工具变量;而对于供给方程,则使用每英亩产出(yield per acre)作为价格的工具变量。

但Wright (1928)仅考虑了恰好识别的情形(即工具变量个数等于内生变量个数),其方法并不适用于过度识别的情形(即工具变量个数大于内生变量个数)。到了20世纪50年代,荷兰经济学家Henri Theil提出“二阶段最小二乘法”(Two Stage Least Squares,简记2SLS)(Theil, 1953)。考虑结果变量$y_i$对处理变量$D_i$的一元线性回归(为论述方便,忽略其他协变量),也称为“结构方程”(structural equation):

$$y_i = \alpha + \beta D_i + \varepsilon_i, \quad \text{cov}(D_i, \varepsilon_i) \neq 0 \tag{12}$$

其中,由于$D_i$与扰动项$\varepsilon_i$相关,故OLS无法得到一致估计。假设存在工具变量$z_i$,满足如下两个条件,即相关性($\text{cov}(D_i, z_i) \neq 0$)与外生性($\text{cov}(z_i, \varepsilon_i) = 0$)。在2SLS的第一阶段回归中(first stage regression),将内生变量对工具变量进行OLS回归[②]:

① 由于此附录与该书正文内容迥异,有些学者怀疑此附录为Philip Wright的儿子Sewall Wright所著,后者是一位著名的生物统计学家。然而,Stock & Trebbi (2003)使用“文体学”(stylometrics)的方法,通过考察二人用词的风格差异,发现Philip Wright是该附录的作者。

② 若存在多个工具变量,则将所有工具变量都放入2SLS的第一阶段回归中。另外,若存在外生的解释变量,也应放入第一阶段回归,即将外生的解释变量作为自己的工具变量。

$$D_i = \gamma + \pi z_i + u_i \tag{13}$$

在第二阶段回归中,将第一阶段回归(13)的拟合值$\hat{D}_i$替代结构方程(12)的内生变量$D_i$进行OLS回归即可。直观上,第一阶段回归的作用是从内生变量$D_i$中分离出外生的部分$\hat{D}_i$(因为$\hat{D}_i$是工具变量$z_i$的线性函数),以保证第二阶段的OLS回归为一致估计。在过度识别的情况下,若扰动项同方差且无自相关,则2SLS是最有效率的IV估计。然而,若扰动项存在异方差或自相关,则2SLS并非最有效率,因为它没有考虑不同矩条件所包含的信息量差异(例如,方差较大的矩条件所含信息量较小)。芝加哥大学经济学家Lars Hansen提出"广义矩估计"(Generalized Method of Moments,简记GMM)(Hansen, 1982),通过选择最优权重矩阵(optimal weighting matrix),达到在过度识别情况下最有效率的IV估计。Lars Hansen因为这项重要工作而分享了2013年诺贝尔经济学奖。

**1. 工具变量的外生性**

有效的工具变量必须满足相关性与外生性。但扰动项不可观测,如何才能保证工具变量与扰动项不相关呢?显然,最可靠的工具变量来自随机实验。例如,Angrist (1990)研究越战期间是否服兵役($D_i$)如何影响退役后的长期收入($y_i$)。越战期间,美国对全国年轻男子以生日抽签的方式进行征兵,但是否参军还取决于体检,且有些人得到豁免,而另一些人未抽中却自愿参军,故$D_i$仍存在内生性。为此,Angrist (1990)使用抽签结果($z_i$)作为$D_i$的工具变量。IV估计结果显示,服兵役会减少白人的长期收入,但不影响非白人的长期收入。

如果随机实验不可得,退而其次,还可使用自然实验作为工具变量。例如,在研究教育投资的回报率时,Angrist & Krueger (1991)将出生季度(quarter of birth)作为教育年限的工具变量。美国多数州要求青少年在满16岁生日之前须在校上学,而儿童在入学那年1月1日须满6周岁。这使得年初出生的学生在其受教育过程中,相比年末出生的学生,更早达到法定退学年龄,故1季度出生者所受教育平均而言低于4季度出生者。显然,出生季度并非由人为操控的随机实验所决定,但可视为大自然带来的准实验。

如果工具变量并非源自随机实验或自然实验,在恰好识别的情况下,一般须通过"排他性约束"(exclusion restriction)的定性讨论来考察工具变量的外生性。直观上,由于IV估计已经利用了工具变量的外生性(即$\mathrm{cov}(z_i, \varepsilon_i) = 0$),故无法再通过数据来检验此外生性(比如考察$z_i$与残差的相关性)。排他性约束的基本逻辑是,由于扰动项$\varepsilon_i$包含了除处理变量$D_i$以外,影响结果变量$y_i$的所有因素,故工具变量的外生性要求,工具变量仅通过处理变量$D_i$这唯一渠道影响结果变量,而排除所有其他渠道,参见图3。

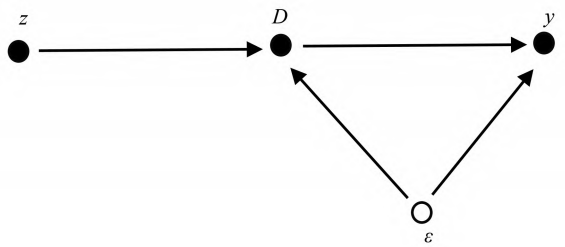



图3 工具变量的排他性约束

然而,实践中的排他性约束讨论常常并不令人信服,有时甚至缺失,这无疑降低了工具变量法的可信度。当然,在过度识别的情况下(即工具变量个数大于内生变量个数),可以通过"过度识别检验"(overidentification test)考察工具变量的外生性(即2SLS的Sargan检验或GMM的Hansen检验)。但过度识别检验依然依赖于恰好识别的大前提,故排他性约束的讨论仍然是必要的。

**2. 弱工具变量**

即使研究者对于工具变量的外生性很有信心,也仍需担心"弱工具变量"(weak instrument)的问题,即虽然$\mathrm{cov}(D_i, z_i) \neq 0$,但二者的相关性很弱。弱工具变量问题一直未引起重视,直到Bound et al.

(1995)发现,如果将 Angrist & Krueger (1991)的工具变量"出生季度"随机打乱,使用2SLS也能得到相似的教育投资回报率与统计显著性!这说明在弱工具变量的情况下,IV估计结果并不可信。直观上,弱工具变量问题相当于有效样本容量太小,因为弱工具变量仅能分离出内生变量微小的一部分信息。事实上,弱工具变量的后果可能更为严重,因为 Angrist & Krueger (1991)所用样本容量高达20多万或30多万。将第一阶段回归方程(13)代入原方程(12)可得"简化式回归"(reduced form regression):

$$y_i = (\alpha + \beta\gamma) + \delta z_i + (\beta u_i + \varepsilon_i), \tag{14}$$

其中,$\delta = \beta\pi$。由此可知,$\beta = \dfrac{\delta}{\pi}$,故IV估计量事实上是一个比值估计量(ratio estimator),即 $\hat{\beta}_{IV} = \dfrac{\hat{\delta}}{\hat{\pi}}$,其中 $\hat{\delta}$ 与 $\hat{\pi}$ 分别为简化式回归与第一阶段回归的OLS估计量。显然,如果 $D_i$ 与 $z_i$ 的相关性很弱(即 $\pi$ 很接近于0),则若此式的分母 $\hat{\pi} \approx 0$,将使得 $\left|\hat{\beta}_{IV}\right|$ 变大且不稳定。例如,Jiang (2017)发现,在金融学三大国际顶刊发表的255篇论文中,超过80%的IV估计值大于OLS估计值,而前者的平均值是后者的9倍之多。Lal et al. (2024)考察了政治学三大国际顶刊使用IV估计的67篇论文,也得到类似的结论。进一步,在弱工具变量的情况下,尽管 $\hat{\delta}$ 与 $\hat{\pi}$ 均为渐近正态,但作为二者的比值,$\hat{\beta}_{IV}$ 却不再服从渐近正态分布,致使常规的统计推断失效。更糟糕的是,若存在弱工具变量与内生工具变量的"并发症",则使用内生工具变量带来的偏差将被放大。由于 $\hat{\beta}_{IV}$ 为比值估计量,故 $\hat{\beta}_{IV}$ 的偏差在形式上也是一个比值:

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_{IV} = \beta + \frac{\operatorname{Cov}(D_i, \varepsilon_i)}{\operatorname{Cov}(D_i, z_i)} \tag{15}$$

因此,在大样本中,IV估计量 $\hat{\beta}_{IV}$ 的偏差为 $\dfrac{\operatorname{Cov}(D_i, \varepsilon_i)}{\operatorname{Cov}(D_i, z_i)}$。显然,若 $\operatorname{Cov}(D_i, \varepsilon_i) \neq 0$,而 $\operatorname{Cov}(D_i, z_i) \approx 0$,则此偏差将进一步被放大,可谓"雪上加霜"。一个重要问题是如何识别弱工具变量。Staiger & Stock (1997)提出检验弱工具变量的一个"经验规则"(rule of thumb),即在进行第一阶段回归时,计算检验所有工具变量联合显著性的 $F$ 统计量;若此 $F$ 统计量大于10,则认为不存在弱工具变量。Stock & Yogo (2005)进一步给出了此检验的具体临界值(依赖于工具变量的个数),但实践中仍主要流行"$F > 10$"的经验规则。

但 Staiger & Stock (1997)与 Stock & Yogo (2005)的弱工具变量检验均假设扰动项同方差且无自相关(故其 $F$ 统计量也使用非稳健的普通标准误),在实践中未必成立。另一方面,实证研究者有时使用异方差或聚类稳健标准误来计算第一阶段回归的 $F$ 统计量,并将其与10相比;尽管这样做并无理论上的依据。为此,Montiel Olea & Pflueger (2013)[①]提出了第一阶段回归的"有效 $F$ 统计量"(effective $F$ statistic),在异方差、自相关与聚类数据的情况下依然稳健。有效 $F$ 统计量是传统的非稳健 $F$ 统计量的"缩放版本"(a scaled version),二者表达式的分子相同,仅分母不同。若加上同方差且无自相关的假定,则有效 $F$ 统计量还原为非稳健的 $F$ 统计量。另外,在恰好识别情况下,若存在异方差或聚类数据,则有效 $F$ 统计量在数值上等价于异方差或聚类稳健的 $F$ 统计量;但在过度识别的情况下,二者并不相等。总之,Montiel Olea & Pflueger (2013)将弱工具变量检验推广到存在异方差、自相关或聚类数据的情形,并为 Andrews et al. (2019)所推荐。但有效 $F$ 统计量目前仅适用于单一内生变量的情形。

针对一个内生变量与一个工具变量的恰好识别情形,Lee et al. (2022)提出了在弱工具变量情况下依然稳健的统计推断方法,称为"$tF$ 方法"($tF$ procedure)。Lee et al. (2022)指出,即使第一阶段回归的 $F$ 统计量大于10,$\hat{\beta}_{IV}$ 的标准误也可能被低估,导致相应的 $t$ 统计量被高估;而相应的95%置信区间覆盖真实 $\beta$ 的概率也可能远离0.95的名义置信度。$tF$ 方法包括以下步骤。首先,使用异方差或聚类稳健的标

① 该论文第一作者的姓氏并非"Olea",而是"Montiel Olea"(复姓)。

准误,计算与$\hat{\beta}_{IV}$相应的$t$统计量。其次,根据第一阶段$F$统计量的大小(反映工具变量的强弱),计算此$t$检验的临界值(而非常规的固定临界值1.96)。此临界值是第一阶段$F$统计量的函数,称为"$tF$临界值函数"($tF$ critical value function),由Lee et al.(2022)以列表的形式提供[①]。最后,若要计算置信区间,可使用经过校正的"$tF$标准误"($tF$ standard error):

$$\mathrm{SE}_{tF} = \mathrm{SE} \times \left( \frac{tF\text{临界值}}{1.96} \right) \tag{16}$$

其中,SE为常规的异方差或聚类稳健标准误,而$\left( \frac{tF\text{临界值}}{1.96} \right)$为校正的倍数。由此可得置信区间$[\hat{\beta} - 1.96 \times \mathrm{SE}_{tF},\ \hat{\beta} + 1.96 \times \mathrm{SE}_{tF}]$。将$tF$方法应用于发表在《美国经济评论》的61篇论文,Lee et al.(2022)发现在5%的显著性水平上,四分之一论文的校正标准误至少比传统2SLS标准误大49%。

事实上,早在20世纪40年代末,美国统计学家Ted Anderson与Herman Rubin即提出了在弱工具变量下依然适用的"安德森—鲁宾检验"(Anderson-Rubin test,简记AR)(Anderson & Rubin, 1949)。考虑一个内生变量与一个工具变量的情形[②],并检验原假设$H_0$: $\beta = \beta_0$。根据简化式方程(14),$\delta = \beta\pi$。因此,在原假设下,$\delta - \beta_0\pi = 0$。相应的AR检验统计量为:

$$AR(\beta_0) = \hat{\delta} - \beta_0\hat{\pi} \xrightarrow{d} N\left(0, \sigma^2_{AR}(\beta_0)\right) \tag{17}$$

其中,由于$\hat{\delta}$与$\hat{\pi}$均为渐近正态,故二者的线性组合$(\hat{\delta} - \beta_0\hat{\pi})$也服从渐近正态分布,而$\sigma^2_{AR}(\beta_0) = \mathrm{Var}(\hat{\delta}) + \beta_0^2\,\mathrm{Var}(\hat{\pi}) - 2\beta_0\,\mathrm{Cov}(\hat{\delta}, \hat{\pi})$为其渐近方差。由于AR统计量并非比值的形式,故此结果对于任意的$\pi$取值都成立(即使$\pi = 0$)。由此可对原假设"$H_0$: $\delta - \beta_0\pi = 0$"进行$t$检验(在过度识别情况下可进行$F$检验或$\chi^2$检验)。若担心存在异方差或聚类数据,可使用异方差或聚类稳健的标准误。AR检验的最大优势在于,它即使在弱工具变量的情况下依然成立。

进一步,还可使用"检验求逆"(test inversion)的方法,通过AR检验构造95%的置信区间。具体而言,给定$\beta_0$,对于$H_0$: $\beta = \beta_0$在5%的显著性水平上进行AR检验。若接受此原假设,则将$\beta_0$放入95%的置信集中。所有不被拒绝的$\beta_0$取值集合,即为95%的置信集。如果工具变量很弱,则AR置信集可能是整个实数轴,意味着无法识别$\beta$。在某些情况下,AR置信集也可能不是一个区间。另外,在过度识别的情况下,AR置信集还可能是空集,这或许意味着过度识别约束并不成立。由于AR检验无论工具变量强弱均适用,且在恰好识别情形下是有效率的检验,故Andrews et al.(2019)与Keane & Neal(2023)均推荐使用AR检验,尽管目前实践中仍较少应用。

**3. 异质性工具变量法**

以上有关工具变量法的论述均假设"同质性的处理效应"(homogeneous treatment effects),即所有个体的处理效应均为$\beta$。然而,"异质性的处理效应"(heterogeneous treatment effects)显然是更现实的假定。例如,个体$i$的处理效应为$\beta_i$,且不尽相同。假定处理变量$D_i$与工具变量$z_i$均为虚拟变量。例如,$D_i$为实际的处理状态,$z_i$为实验设计者随机分配的处理状态,但由于个体未必遵从实验设计,故二者可能不一致。根据潜在结果的框架,若$z_i = 0$,记相应的"潜在处理"(potential treatment)为$D_{0i}$;反之,若$z_i = 1$,记相应的潜在处理为$D_{1i}$。根据潜在处理$(D_{0i}, D_{1i})$的不同取值,美国经济学家Guido Imbens与Joshua Angrist将所有个体分为四类(Imbens & Angrist, 1994),参见图4。

在图4中,原点$(D_{0i}, D_{1i}) = (0, 0)$意味着,无论$z_i$如何取值,个体$i$都不会参与项目(受到处理),故称

① 在此表中,第一阶段$F$统计量的最小值为4,故若第一阶段$F$统计量小于4(工具变量太弱),则无法使用$tF$方法。若第一阶段$F$统计量的取值未列于表中,可使用线性插值的方法进行近似。

② 安德森—鲁宾检验也适用于过度识别的情形,在此从略。

为“恒拒者”(never taker)。另一方面,若$(D_{0i}, D_{1i}) = (1, 1)$,则个体总会参与项目,故称为“恒受者”(always taker)。显然,工具变量$z_i$的外生变动并不影响恒拒者或恒受者的处理状态,故IV估计无法识别这两类个体的处理效应。

进一步,若$(D_{0i}, D_{1i}) = (0, 1)$,则当$z_i = 0$时,个体不参与项目$(D_i = 0)$;而当$z_i = 1$时,个体参与项目$(D_i = 1)$,故称为“顺从者”(complier)。反之,若$(D_{0i}, D_{1i}) = (1, 0)$,则当$z_i = 0$时,个体却参与项目$(D_i = 1)$;而当$z_i = 1$时,个体反而不参与项目$(D_i = 0)$,故称为“逆反者”(defier)。一个基本结论是,如果样本中存在逆反者,则无法识别处理效应。不妨假设所有个体的处理效应为正。当工具变量$z_i$从0变为1时,顺从者的结果变量将增加,但逆反者的结果变量反而减少,二者部分地相互抵消,使得平均处理效应的估计出现偏差。因此,在异质性处理效应的情况下,为了识别平均处理效应,除了常规的工具变量相关性与外生性假定外,还须施加“单调性”(monotonicity)假定$D_{1i} \geqslant D_{0i}$,以排除逆反者的存在(对于逆反者,$D_{1i} < D_{0i}$)。Imbens & Angrist (1994)证明,在这些假定下,IV估计可以识别“顺从者的平均处理效应”$\mathrm{E}(y_{1i} - y_{0i} | D_{1i} > D_{0i})$,也称为“局部平处理效应”(Local Average Treatment Effect,简记LATE)。但究竟哪些个体为顺从者,则无从判断。而且,由于顺从者根据工具变量而定义,故使用不同的工具变量,其相应的顺从者群体也可能不同。在算法层面,基于异质性处理效应的IV估计与传统工具变量法并无二致,但二者的解释大相径庭。另外,针对上文提及的IV估计常大于OLS估计的现象,异质性处理效应的LATE理论也提供了新的解释,即IV估计所度量的顺从者平均处理效应,可能与全样本的平均处理效应有明显差别。2021年,Guido Imbens与Joshua Angrist由于异质性工具变量法的开创性研究而分享了诺贝尔经济学奖的一半奖金①。

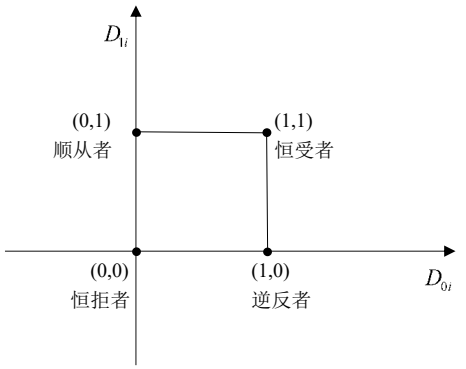


图4 异质性处理效应下的四类个体

## 六、双重差分法

上文主要探讨如何使用横截面数据进行因果推断。从本节开始,将考察怎样在面板数据中做因果推断。政策效应研究的一种常见情形是,样本中个体可分为两组,即处理组(treatment group)与控制组(control group);而时间可分为两段,即处理前(pretreatment)与处理后(posttreatment)。在处理前,所有个体均未受到政策冲击;而在处理后,只有处理组个体受到政策冲击,而控制组个体始终未受处理。这是双重差分法(difference-in-differences,简记DID)的经典适用场景。近年来,双重差分法蓬勃发展,已衍生出不同类型,包括两期DID、多期DID(含标准DID与交叠DID)、一般DID等,下面分别阐述。

**1. 两期DID**

1849年,伦敦因霍乱而死亡万余人。关于霍乱的原因,当时有两种解释,即空气污染(bad air)与水污染(bad water)。当时伦敦供水主要来自两家公司,即Lambeth Waterworks(简记LW)与Southwark and Vauxhall Water(简记SV),二者水源均来自伦敦市区受污染的泰晤士河。英国流行病学家John Snow认为,伦敦霍乱因水污染而起(Snow, 1855[1936])。1852年,LW公司将其水源地移到了未受污染的泰晤士河上游,但SV公司仍从伦敦市区的泰晤士河取水。1853—1854年,伦敦再次爆发霍乱。John Snow收

① 2021年诺贝尔经济学奖的另一半奖金颁发给了David Card,参见第二节。

集数据后发现,LW公司客户的霍乱死亡率明显低于SV公司客户。1855年,SV公司也将水源地移到了泰晤士河上游。这项研究被称为John Snow的"伟大实验"(grand experiment),因为超过30万的伦敦人参与了这项自然实验。这或许就是双重差分法的最早起源。LW公司客户构成处理组,而SV公司客户构成控制组;1849年为处理前,而1853—1854为处理后。由于处理组与控制组在处理前的死亡率可视为无差别,故二者在处理后死亡率的差异即为双重差分估计。

假设面板数据只有两期,其中$t=1$表示处理前,而$t=2$表示处理后。为了评估处理组的政策效应,一种天真的做法是比较处理组均值的前后差异,即$\Delta\bar{y}_{\text{treat}} \equiv \bar{y}_{\text{treat},2} - \bar{y}_{\text{treat},1}$,其中$\bar{y}_{\text{treat},2}$与$\bar{y}_{\text{treat},1}$分别为处理组在处理后与处理前的样本均值。但$\Delta\bar{y}_{\text{treat}}$显然混杂了时间效应。若控制组的时间趋势与处理组相同,则可使用控制组均值的前后差异,即$\Delta\bar{y}_{\text{control}} \equiv \bar{y}_{\text{control},2} - \bar{y}_{\text{control},1}$,来估计此时间效应。为了剔除处理组的时间效应,综合以上两个差分,即可得到双重差分估计:

$$\Delta\bar{y}_{\text{treat}} - \Delta\bar{y}_{\text{control}} = (\bar{y}_{\text{treat},2} - \bar{y}_{\text{treat},1}) - (\bar{y}_{\text{control},2} - \bar{y}_{\text{control},1}) \tag{18}$$

由此可知,DID模型的基本前提是,处理组若未受政策干预,其时间趋势应与控制组相同,即所谓"平行趋势假定"(parallel trends assumption)或"共同趋势假定"(common trends assumption),参见图5。

记$y_{it}(0)$与$y_{it}(1)$分别为个体$i$在第$t$期若未受处理及受处理的潜在结果,而处理组虚拟变量为$treat_i$,则平行趋势假定可写为:

$$\mathrm{E}\left[y_{i2}(0) - y_{i1}(0) | treat_i = 1\right] = \mathrm{E}\left[y_{i2}(0) - y_{i1}(0) | treat_i = 0\right] \tag{19}$$

其中,"$treat_i = 1$"表示处理组,而"$treat_i = 0$"表示控制组。上式意味着,若处理组与控制组均未受政策干预,则二者的时间趋势完全相同。将上式移项,可得到处理组在第2期的反事实结果:

$$\mathrm{E}\left[y_{i2}(0) | treat_i = 1\right] = \mathrm{E}\left[y_{i1}(0) | treat_i = 1\right] + \mathrm{E}\left[y_{i2}(0) - y_{i1}(0) | treat_i = 0\right] \tag{20}$$

基于此反事实结果,即可识别处理组的平均处理效应(ATT):

$$\begin{aligned}\tau_{\text{ATT}} &\equiv \mathrm{E}\left[y_{i2}(1) | treat_i = 1\right] - \mathrm{E}\left[y_{i2}(0) | treat_i = 1\right] \\ &= \mathrm{E}\left[y_{i2}(1) | treat_i = 1\right] - \left\{\mathrm{E}\left[y_{i1}(0) | treat_i = 1\right] + \mathrm{E}\left[y_{i2}(0) - y_{i1}(0) | treat_i = 0\right]\right\} \\ &= \left\{\mathrm{E}\left[y_{i2}(1) | treat_i = 1\right] - \mathrm{E}\left[y_{i1}(0) | treat_i = 1\right]\right\} - \left\{\mathrm{E}\left[y_{i2}(0) | treat_i = 0\right] - \mathrm{E}\left[y_{i1}(0) | treat_i = 0\right]\right\}\end{aligned} \tag{21}$$

上式为总体期望的双重差分,而相应的样本估计值正是DID估计量(18)。在实践中,双重差分估计一般写为等价的双向固定效应模型(two-way fixed effects,简记TWFE),更便于统计推断:

$$y_{it} = \alpha + \beta\, treat_i \times post_t + u_i + \lambda_t + \varepsilon_{it} \quad (i = 1, \cdots, n;\ t = 1, 2) \tag{22}$$

其中,$post_t$为处理期虚拟变量(处理后取值为1,而处理前取值为0)。交互项$treat_i \times post_t$是处理变量,表示个体$i$在第$t$期是否受处理。$u_i$为个体固定效应,$\lambda_t$为时间固定效应,而扰动项$\varepsilon_{it}$为"个殊性冲击"(idiosyncratic shock)。在实践中,一般对方程(22)进行OLS估计,并使用以个体为聚类的聚类标准误。两期DID的一个经典案例是Card & Krueger (1994),参见第二节。

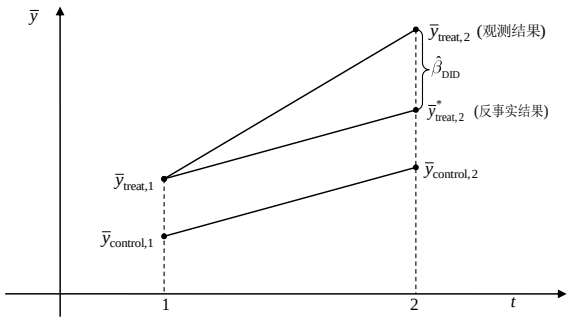



图5 双重差分法的平行趋势假定

对于两期DID,只要将其作一阶差分,即可得到横截面数据,而此截面数据的结果变量为原结果变量的前后变化。此时,若存在协变量$\mathbf{x}_i$(通常为处理前的个体特征),则可使用估计截面数据处理效应的系列方法(参见第四节),例如倾向得分匹配(即PSM-DID,Heckman et al.,1997, 1998),逆概加权法(Abadie, 2005),回归调整法(即结果回归法),以及双稳健估计(Sant'Anna & Zhao, 2020)。这些方法的基本前提是,

对于个体特征 $\mathbf{x}_i$ 相同的处理组与控制组个体,二者的时间趋势平行:

$$\mathrm{E}\left[y_{i2}(0)-y_{i1}(0)|\mathbf{x}_i,\ treat_i=1\right]=\mathrm{E}\left[y_{i2}(0)-y_{i1}(0)|\mathbf{x}_i,\ treat_i=0\right] \tag{23}$$

上式称为"条件平行趋势假定"(conditional parallel trends assumption)。显然,条件平行趋势假定(23)比通常的(无条件)平行趋势假定(19)更弱,在实践中更易满足。基于条件平行趋势假定的方法在交叠DID中有着重要应用,详见下文。

2. **标准**DID

对于两期DID,政策冲击只可能发生于第2期,故处理组所有个体的受处理时间必然同步。对于多期DID,如果处理组所有个体受处理时点均相同(单一处理时点),则称为"标准DID"(standard DID)。反之,若处理组个体的政策冲击时点不尽相同(多个处理时点),则称为"交叠DID"(staggered DID)。对于标准DID,由于政策冲击同步,故仍可使用交互项 $treat_i \times post_t$ 作为处理变量,其回归方程可写为:

$$y_{it}=\alpha+\beta\, treat_i\times post_t+\boldsymbol{\gamma}'\mathbf{w}_{it}+u_i+\lambda_t+\varepsilon_{it}(i=1,\ \cdots,\ n;\ t=1,\cdots,T) \tag{24}$$

其中,$\mathbf{w}_{it}$ 为随时间而变的一些协变量。这依然是双向固定效应模型(TWFE)。多期DID的优势在于(假定处理前不止一期),可使用"事件分析法"(event study)进行平行趋势检验[①]:

$$y_{it}=\alpha+\sum_{s=2}^{T}\beta_s\, treat_i\times\mathbf{1}(s=t)+\boldsymbol{\gamma}'\mathbf{w}_{it}+u_i+\lambda_t+\varepsilon_{it}(i=1,\ \cdots,\ n;\ t=1,\cdots,T) \tag{25}$$

其中,$\mathbf{1}(s=t)$ 为第 $t$ 期的时间虚拟变量(第 $t$ 期取值为1,反之取值为0)。平行趋势假定要求,所有处理前的 $\beta_s$ 均不显著,或联合不显著;而处理后的 $\beta_s$ 则为"动态处理效应"(dynamic treatment effects)。美国经济学家Ashenfelter(1978)使用多期DID估计就业培训项目对于收入的处理效应,这或许是经济学中最早使用DID的案例。然而,Ashenfelter(1978)发现,在就业培训前,处理组的平均收入不仅相对于控制组下降,而且绝对下降,即所谓"阿氏沉降"(Ashenfelter's dip),故存在个体自我选择,导致平行趋势假定不满足。

3. **交叠**DID

对于多期DID,有时每位个体受政策冲击的初始时间并不一致;比如,某试点政策在不同城市分批推出,称为"交叠处理"(staggered adoption)。更严格地,交叠DID需满足两个要素,即个体开始受处理的时间不尽相同,且个体一旦受处理,则不能退出。反之,若存在政策退出或逆转,则称为"一般DID"(general DID)。对于交叠DID,虽然无法将处理变量写为交互项的形式,但仍可定义处理变量 $D_{it}$,表示个体 $i$ 在时期 $t$ 是否受处理:

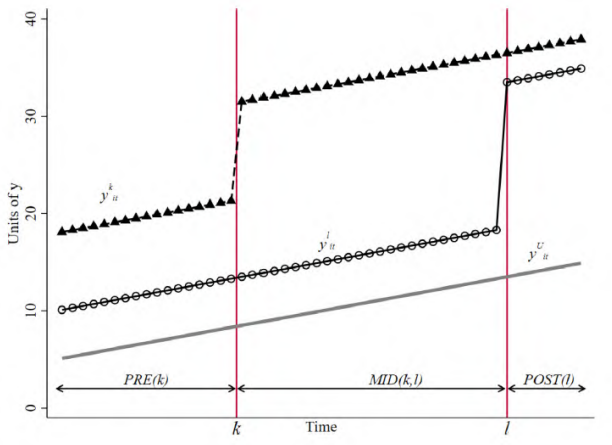
$$y_{it}=\alpha+\beta D_{it}+\boldsymbol{\gamma}'\mathbf{w}_{it}+u_i+\lambda_t+\varepsilon_{it}(i=1,\ \cdots,\ n;\ t=1,\cdots,T) \tag{26}$$

在形式上,上式依然是双向固定效应模型,传统上使用OLS估计;并可通过事件分析法检验平行趋势假定。但近年研究表明,除非处理效应同质(例如,处理效应不随时间而变),使用TWFE估计交叠DID一般会有偏差。理解此偏差的方便工具是"培根分解"(Bacon decomposition)。根据培根分解定理(Goodman-Bacon, 2021)[②],使用TWFE估计交叠DID模型,所得估计系数为使用该样本所有可能的标准DID估计系数的加权平均,其中所有权重均非负,且权重之和为1。例如,假定样本中有一个未处理组(untreated group),其结果变量记为 $y_{it}^{U}$;一个早处理组(early treatment group),其结果变量记为 $y_{it}^{k}$;以及一个迟处理组(late treatment group),其结果变量记为 $y_{it}^{l}$;参见图6。

此样本共包括四个标准DID模型,即"早处理组vs未处理组","迟处理组vs未处理组","早处理组vs晚处理组"(去掉晚处理组受处理后的样本),以及"晚处理组vs早处理组"(去掉早处理组受处理前的样本)。其中,"晚处理组vs早处理组"的标准DID模型将已受处理的早处理组作为控制组进行TWFE估

① 有关事件分析法及其在DID模型的应用,可参见张子尧和黄炜(2023)的文献综述。

② "Goodman-Bacon"为复姓,但其分解定理一般简称为"Bacon decomposition theorem"。




注:此图来自Goodman-Bacon (2021), Fig. 1。

图6 培根分解的示意图

计。显然,若早处理组的处理效应随时间而变,则相应的标准DID估计将出现偏差,从而污染交叠DID的最终估计结果。这意味着,针对交叠DID模型的TWFE估计仅在处理效应为常数(constant treatment effect)的情况下才是一致估计,但处理效应为常数的假定在实践一般难以成立。为此,文献中涌现出在异质性处理效应情况下依然稳健(heterogeneity-robust)的一系列交叠DID估计方法。所谓异质性处理效应(heterogeneous treatment effects),即允许处理效应可随个体或时间而变。

记个体$i$的初始处理时间(initial treatment timing)为$G_i$。根据$G_i$的不同取值,可将样本分为若干"组群"(cohort);例如,上文例子中的早处理组、晚处理组与未处理组。在交叠DID模型中,除了通常的日历时间(calendar time),还可定义相对于政策冲击的"相对时间"(relative time)或"事件时间"(event time)$l_{it} = t - G_i$,以及相对时间的虚拟变量$D_{it}^{l} = \mathbf{1}\{t - G_i = l\}$(表示是否距政策冲击过了$l$期)。针对交叠DID模型,我们感兴趣的估计目标一般为组群$g$在相对时间$l$的"组群平均处理效应"(cohort average treatment effect,简记CATT):

$$CATT_{g,l} = \mathrm{E}\left[y_{i,g+l} - y_{i,g+l}(0) | G_i = g\right] \tag{27}$$

其中,$y_{i,g+l}(0)$表示个体$i$若从未受处理在第$(g+l)$期的潜在结果。CATT也称为"Group-time average treatment effect"。Sun & Abraham (2021)提出以下"交互加权估计量"(interaction-weighted estimator,简记IW-DID)来估计CATT:

$$y_{it} = \sum_{g \notin C} \sum_{l \neq -1} \delta_{g,l}\left(\mathbf{1}\{G_i = g\} \cdot D_{it}^{l}\right) + u_i + \lambda_t + \varepsilon_{it} \tag{28}$$

其中,$C$为"从未处理组"(never treated),作为参照系。若没有从未处理组,可将最后处理组(last treated)设为从未处理组,同时去掉该组受处理后的样本数据。在形式上,上式依然是TWFE模型,可用OLS进行估计。在平行趋势假定下,可以证明,$\hat{\delta}_{g,l}$是$CATT_{g,l}$的一致估计。IW-DID虽然易懂且操作方便,但仅以从未处理组作为控制组,而未考虑将"尚未处理组"(not-yet treated group)作为控制组。另外,IW-DID依赖于无条件平行趋势假定,无法适用于给定处理前协变量的条件平行趋势假定。

Callaway & Sant'Anna (2021)提出CSDID方法,通过"长差分"(long difference)来估计CATT,或许是目前最为流行的交叠DID方法。假设以从未处理组作为控制组,为了估计组群$g$在相对时间$l(l \geqslant 0)$的$CATT_{g,l}$,CSDID的基本步骤如下。

(1)保留时期$(g-1)$与$(g+l)$的数据,去掉其他各期的数据。

(2)保留组群$g$与组群$C$(从未处理组)的数据,去掉其他组群的数据。

(3)针对所得的两期DID,如果无协变量,则使用TWFE计算$\widehat{CATT_{g,l}}$。

(4)针对所得的两期DID,如果有协变量$\mathbf{x}_i$,则可使用逆概加权法、回归调整法或双稳健估计来计算$\widehat{CATT_{g,l}}$(参见第四节)。若协变量$\mathbf{w}_{it}$随时间而变,可令$\mathbf{x}_i \equiv \mathbf{w}_{i1}$,即第1期的协变量取值。

在此算法中,若以尚未处理组(not-yet treated group)作为控制组,可将上述第(2)步中的"从未处理组"更改为"尚未处理组"(至第$(g+l)$期尚未受处理的个体)即可。进一步,对于所得的不同$\widehat{CATT_{g,l}}$,可以观测值的比重作为权重,进行加权平均,以计算加总的组群平均处理效应。例如,固定相对时间$l$,对

不同组群的$\widehat{CATT}_{g,l}$进行加权平均;固定组群$g$,对该组群处理后各期的$\widehat{CATT}_{g,l}$进行加权平均;以及各组群与各期$\widehat{CATT}_{g,l}$的总平均(grand mean)等。

异质性稳健的交叠DID估计方法还包括堆叠回归(stacked regression)(Cengiz et al., 2019),二阶段DID法(two-stage difference-in-differences)(Gardner, 2021),扩展TWFE法(extended TWFE)(Wooldridge, 2021),插补法(imputation)(Borusyak et al., 2024),以及反事实估计量(counterfactual estimators)(Liu et al., 2024)等,限于篇幅从略。需要指出,这些方法有些尚未正式发表(例如二阶段DID法与扩展TWFE法),而有些虽已发表却缺乏足够理论证明(例如堆叠回归、反事实估计量),故仍需更多时间才能尘埃落定。有关交叠DID的文献综述可参见Baker et al.(2022),Roth et al.(2023),Wing et al.(2024),刘冲等(2022b)以及许文立(2023)。另外,对于标准DID与交叠DID,还可通过安慰剂检验(placebo test)进行证伪(falsification),包括时间、空间与混合安慰剂检验,在实践中日益流行;参见陈强等(2025)的文献综述。

4. **一般DID**

标准DID与交叠DID均假定政策不可逆,故处理变量只能从0变到1,在样本期间不可能从1变成0。更一般地,政策可逆(reversible treatment)的DID模型称为"一般DID"(general DID)。针对一般DID,de Chaisemartin和D'Haultfoeuille(2020)考虑估计处理状态切换者(switcher)的即时处理效应,即处理状态切换当期的平均处理效应(instantaneous ATE for all switchers):

$$\delta^S = \mathrm{E}\left\{\frac{1}{N_S}\sum_{t\geqslant 2,\ D_{it}\neq D_{i,t-1}}\left[y_{it}(1)-y_{it}(0)\right]\right\} \tag{29}$$

其中,$N_S=\sum_{t\geqslant 2}\mathbf{1}\left\{D_{it}\neq D_{i,t-1}\right\}$为切换者的总数。切换者还可进一步分为两类,即"切入者"(switch in),满足$D_{i,t-1}=0$, $D_{it}=1$;以及"切出者"(switch out),满足$D_{i,t-1}=1$, $D_{it}=0$。对于第$t$期的切入者,可使用第$(t-1)$期与第$t$期的两期DID,估计其平均处理效应,记为$DID_{+,t}$。类似地,可估计第$t$期切出者的平均处理效应,记为$DID_{-,t}$。将各期的$DID_{+,t}$与$DID_{-,t}$进行加权平均(以观测值为权重),则可得到所有切换者在第$t$期的即时平均处理效应。但即时平均处理效应仅估计当期的处理效应,而政策冲击可能存在滞后效应。为此,de Chaisemartin和D'Haultfoeuille(2022)将即时处理效应推广到"跨期处理效应"(inter-temporal treatment effects)。de Chaisemartin和D'Haultfoeuille(2020, 2022)方法的一个优点是,并不要求处理变量为虚拟变量,也适用于处理变量为排序数据的情形(non-binary ordered treatment)。然而,若将此法应用于标准DID或交叠DID,则可能是低效的,因为处理状态切换者在样本中所占比重或许较小。

5. **平行趋势假定不满足怎么办**

DID模型的基本前提为平行趋势假定。如果平行趋势假定不满足,其他备选因果推断方法包括三重差分法与合成双重差分法。三重差分法(difference-in-differences-in-differences,简记DDD)由美国麻省理工学院经济学家Jonathan Gruber(1994)提出。若平行趋势假定不成立,则DID估计有偏差。DDD的基本思想是,如果能找到另一对"准处理组"与"准控制组",其DID估计也有同样的偏差,则这两个DID估计之差就是无偏的。在具体操作上,DDD一般通过在回归方程中引入一个"三重交互项"(triple interaction)来实现。事实上,DDD仍依赖于某种形式更复杂的平行趋势假定,以保证两个DID的估计偏差正好抵消(Oden & Moen, 2022)。

使用DDD的重要前提是,能找到合适的"准处理组"与"准控制组",但这在实践中未必可行。若平行趋势假定不成立,一个更为通用的解决方法为美国斯坦福大学团队Arkhangelsky et al.(2021)所提出的"合成双重差分法"(synthetic difference-in-differences,简记SDID)。SDID借鉴了合成控制法的权重思想(参见第七节),希望通过选择最优权重,使得加权之后的控制组时间趋势与处理组平行。具体而

言,对于个体$i$第$t$期的观测值,其最优权重为$\hat{w}_{it}=\hat{\omega}_i\times\hat{\lambda}_t$,其中$\hat{\omega}_i$为个体$i$的最优权重,而$\hat{\lambda}_t$为时期$t$的最优权重。然后,SDID以$\hat{w}_{it}$为权重,使用加权最小二乘法(weighted least squares)进行TWFE估计,即求解如下最小化问题:

$$\min_{\alpha,\beta}\sum_{i=1}^{n}\sum_{t=1}^{T}w_{it}(y_{it}-\alpha-\beta D_{it}-u_i-\lambda_t)^2 \tag{30}$$

另外,即使平行趋势检验通过,平行趋势假定也仍可能不成立,因为犯第II类错误的概率依然为正(甚至较大)。此时,可使用Rambachan & Roth (2023)的"诚实DID"方法(honest DID)进行敏感性分析(sensitivity analysis)。首先,仍以常规DID方法进行估计(比如标准DID或交叠DID的方法)。其次,考察需要平行趋势假定被违背到何种程度,估计结果的显著性才会消失。如果平行趋势假定需要被严重违背,才会影响到估计结果的显著性,则认为DID估计结果是稳健的。

## 七、合成控制法

经济学实证研究有时会遇到处理组只有一位个体的情形。在社会科学领域,针对此情形的传统方法是"比较案例分析"(comparative case study)。例如,为了研究马里矣尔船运(Mariel boatlift)引发的古巴移民对迈阿密劳动力市场的影响,Card (1990)选择了四个与迈阿密相似但未受移民冲击的美国城市构成控制组(参见第二节)。但文中并未说明为何选择这四个城市,基本由研究者主观决定。另外,这四个城市与迈阿密的相似度显然不同,是否应给予不同的权重,而非简单算术平均?

为了克服比较案例分析的局限性,美国麻省理工学院经济学家Alberto Abadie与合作者提出"合成控制法"(synthetic control method,简记SCM)(Abadie & Gardeazabal, 2003),并用来研究西班牙巴斯克地区(Basque country)恐怖活动的经济成本。SCM的基本思想是,使用控制组中各地区的线性组合作为"合成控制"(synthetic control),以估计处理地区若未受政策冲击的反事实结果。SCM的优点是,可通过"数据驱动"(data-driven)的方式选择合成控制的最优权重,从而避免研究者手工选择控制组的主观随意性。

假设共有$(N+1)$个地区,其中第1个地区受到政策冲击[①],而其余$N$个地区未受处理,构成潜在的控制组,称为"捐献池"(donor pool)。将合成控制地区的权重记为向量$\mathbf{w}=(w_2\cdots w_{N+1})'$,其中$w_i$为地区$i$的权重,并要求所有权重非负,且权重之和为1。记结果变量为$y$,而$K$维向量$\mathbf{x}_i$为地区$i$在处理前的协变量均值。SCM的目标是,寻找最优权重$\mathbf{w}$,使得控制地区的协变量$(\mathbf{x}_2\cdots\mathbf{x}_{N+1})$,经过线性组合后,尽可能接近处理地区的协变量$\mathbf{x}_1$:

$$\mathbf{X}_0\mathbf{w}=(\mathbf{x}_2\cdots\mathbf{x}_{N+1})\begin{pmatrix}w_2\\ \vdots\\ w_{N+1}\end{pmatrix}=w_2\mathbf{x}_2+\cdots+w_{N+1}\mathbf{x}_{N+1}\approx\mathbf{x}_1 \tag{31}$$

其中,控制组协变量的矩阵$\mathbf{X}_0\equiv(\mathbf{x}_2\cdots\mathbf{x}_{N+1})$。为此,求解如下有约束的"加权平方和"最小化问题:

$$\begin{aligned}&\min_{\mathbf{w}}(\mathbf{x}_1-\mathbf{X}_0\mathbf{w})'\mathbf{V}(\mathbf{x}_1-\mathbf{X}_0\mathbf{w})\\ &s.t.\ w_i\geqslant 0\ \ (i=2,\cdots,N+1);\ \sum_{i=2}^{N+1}w_i=1\end{aligned} \tag{32}$$

其中,$\mathbf{V}$为$K\times K$对角矩阵,其主对角线元素$(v_1,\cdots,v_K)$均非负,表示每个协变量对于预测结果变量的重要性(故这些协变量也称"预测变量")。由于存在非线性约束,故此最小化问题一般须进行数值求解。显然,最优解依赖于$\mathbf{V}$,记为$\mathbf{w}^*(\mathbf{V})$。进一步,可选择最优的$\mathbf{V}^*$,使得在处理前,合成控制地区的结果变

① 若有多个地区受处理,可分别使用SCM进行实证分析。

量尽量接近处理地区。得到最优权重 $\mathbf{w}^* = \mathbf{w}^*(\mathbf{V}^*)$ 后,即可预测处理地区若未受政策冲击的反事实结果,并计算相应的处理效应。

Abadie et al. (2010)探讨了SCM的统计性质,并以SCM研究加州控烟法对加州香烟销量的影响。Abadie et al. (2010)证明,若最优合成控制 $\mathbf{w}^*$ 能完美再现处理地区的协变量特征与处理前结果变量,则当处理前时期数 $T_0$ 趋向无穷大时,SCM估计量是渐近无偏的(asymptotically unbiased)。在统计推断方面,由于实践中的控制地区个数 $N$ 与处理前时期数 $T_0$ 通常不大,故Abadie et al. (2010)提出使用"空间安慰剂检验"(in-space placebo test),将控制地区作为"伪处理单位"(fake treatment units)进行SCM估计,从而得到安慰剂效应的分布进行统计推断。进一步,Abadie et al. (2015)提出"时间安慰剂检验"(in-time placebo test),即将政策冲击时间后移至处理前的某个时期,作为"伪处理时间"(fake treatment time)进行证伪检验。但时间安慰剂检验本身无法提供检验的 $p$ 值。为此,Chen & Yan (2023)提出"混合安慰剂检验"(mixed placebo test),同时使用伪处理时间与伪处理单位,即可获得时间安慰剂检验的 $p$ 值。

合成控制法的最优解常常是稀疏的。例如,在西班牙巴斯克地区恐怖活动的案例中(Abadie & Gardeazabal, 2003),16个控制地区中仅有两个地区的最优权重为正,而其余控制地区的权重均为0。在一篇有关SCM的文献综述中,Abadie (2021)从预测空间(predictor space)的角度出发,在几何上将SCM的最优权重解释为处理地区的协变量 $\mathbf{x}_1$ 向控制组协变量 $\mathbf{X}_0 = (\mathbf{x}_2 \cdots \mathbf{x}_{N+1})$ 构成的"凸包"(convex hull)所做的投影,以此解释SCM的稀疏性。Chen & Li (2024)进一步证明,如果SCM存在唯一解,则SCM最优权重为正的个数不会超过协变量的个数。这意味着,如果实证研究中SCM正权重数目超过协变量个数,则SCM不存在唯一解,其最优权重可能不稳定。另外,Chen & Li (2024)还从参数空间(parameter space)的角度,给出了SCM稀疏性的新解释,在性质上类似于拉索估计量(Lasso)的稀疏性。

## 八、回归控制法与分位数控制法

针对处理组只有一位个体的情形,合成控制法是流行的因果推断方法。但SCM要求合成控制须很好地复现处理个体在处理前的协变量与结果变量,否则无法使用。另一方面,SCM要求所有权重非负,限制了拟合效果。为此,Hsiao et al. (2012)另辟蹊径,提出"政策评估的面板数据方法"(a panel data approach to program evaluation)。由于此法利用面板数据中横截面单位之间的相关性(cross-sectional correlation),直接使用回归的方法构造反事实的控制个体,故也称为"回归控制法"(regression control method,简记RCM)。

假设观测到面板数据 $\{y_{it}\}(i = 1, \cdots, N+1;\ t = 1, \cdots, T)$,记相应的潜在结果分别为 $y_{it}^0$(若未受处理)与 $y_{it}^1$(若受处理)。假定只有第1个地区受到政策干预,而其余 $N$ 个地区未受冲击。记 $T_0$ 为处理前的时期数。我们感兴趣第1个地区的处理效应 $\tau_{1t} = y_{1t} - y_{1t}^0\ (t = T_0 + 1, \cdots, T)$。假设反事实结果 $y_{it}^0$ 由以下线性因子模型(linear factor model)决定:

$$y_{it}^0 = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + \varepsilon_{it} \tag{33}$$

其中,$\alpha_i$ 为个体固定效应,$\mathbf{f}_t = (f_{1t} \cdots f_{rt})'$ 为 $r \times 1$ 维不可观测的共同因子(common factors)或共同冲击(common shocks),$\mathbf{b}_i = (b_{i1} \cdots b_{ir})'$ 为 $r \times 1$ 维不可观测的因子载荷(factor loadings),$\mathbf{b}_i'\mathbf{f}_t = b_{i1}f_{1t} + \cdots + b_{ir}f_{rt}$ 为交互固定效应(interactive fixed effects),而 $\varepsilon_{it}$ 为个殊性冲击(idiosyncratic shock)。

由于共同因子 $\mathbf{f}_t$ 的存在,故不同个体的结果变量存在同期截面相关(cross-sectional correlation)。利用此截面相关,Hsiao et al. (2012)将共同因子 $\mathbf{f}_t$ 消去,并把处理地区的 $y_{1t}^0$ 表示为同期控制地区 $y_{2t}^0, \cdots, y_{N+1,t}^0$ 的函数。记第 $t$ 期控制地区的结果变量为 $\tilde{\mathbf{y}}_t = (y_{2t} \cdots y_{N+1,t})'$,可使用处理前的样本数据,进

行如下线性回归:

$$y_{1t} = \delta_1 + \boldsymbol{\delta}'\tilde{\mathbf{y}}_t + v_{1t}\ (t = 1, \cdots, T_0) \tag{34}$$

然后,以 $\hat{y}_{1t}^0 = \hat{\delta}_1 + \widehat{\boldsymbol{\delta}}'\tilde{\mathbf{y}}_t\ (t = T_0 + 1, \cdots, T)$ 估计处理地区在处理后的反事实结果,以及处理效应 $\hat{\tau}_{1t} = y_{1t} - \hat{y}_{1t}^0$。然而,在回归方程(34)中,如果 $N \geqslant T_0$(解释变量个数大于样本容量),则为高维数据,无法进行OLS回归。另一方面,即使 $N < T_0$,也可能发生"过拟合"(overfitting),即样本内拟合很好,但外推预测的效果较差。为此,Hsiao et al.(2012)推荐使用"校正AIC准则"(corrected AIC)来选择最优模型。由于传统AIC准则对于模型复杂程度的惩罚不够严厉,故校正AIC准则对此进行了校正(增加了额外的惩罚项)。作为应用演示,Hsiao et al.(2012)使用RCM评估了1997年香港回归,以及2004年香港与内地经济整合对于香港经济增长的影响。

Li & Bell(2017)与Carvalho et al.(2018)提出使用拉索估计量(Lasso)进行模型选择,计算更为简便,在高维情况下依然适用。相对于合成控制法,RCM的一个优势是无须协变量也能使用。Hsiao & Zhou(2019)则将协变量引入RCM,以进一步提高反事实预测的效率。在统计推断方面,Fujiki & Hsiao(2015)在独立同分布(iid)的较强假定下,推导了RCM估计量的置信区间。事实上,也可使用安慰剂检验进行RCM的统计推断,包括空间、时间与混合安慰剂检验;其原理类似于SCM的安慰剂检验(Yan & Chen, 2022)。

然而,使用安慰剂检验进行统计推断,要求较强的假定(比如随机分组),在SCM与RCM的实践中一般难以满足。为此,Chen et al.(2024)提出使用分位数回归(quantile regression),对方程(34)进行2.5%与97.5%的分位数回归,由此得到反事实结果 $y_{1t}^0$ 的2.5%与97.5%分位数,以构造每期处理效应的95%置信区间。但传统的线性分位数回归收敛速度较慢,且可能存在函数误设的偏差。Chen et al.(2024)提出使用"分位数随机森林"(quantile random forest)的机器学习方法进行分位数回归,以构造处理效应的置信区间,并称此方法为"分位数控制法"(quantile control method,简记QCM)。

作为一种非参数的集成学习方法(例如随机森林为1000棵决策树的平均),QCM在异方差、自相关或函数误设的情况下依然稳健,且收敛速度更快,在小样本中表现优良。例如,在控制组个体数 $N = 30$,处理前时期数 $T_0 = 30$ 的情况下,Chen et al.(2024)考察了13种不同的数据生成过程,模拟结果显示使用QCM构造的置信区间,其覆盖真实处理效应的概率均已超过90%,接近95%的名义置信度,且优于文献中的其他方法,包括Fujiki & Hsiao(2015),Bai & Ng(2021),Cattaneo et al.(2021)以及Chernozhukov et al.(2021)所提出的方法。

## 结论与展望

自20世纪30年代计量经济学草创迄今,计量经济学日渐成熟,而因果推断方法也不断推陈出新,演变为20世纪90年代以来的可信度革命,至今方兴未艾。事实上,目前有关因果推断的文献已十分庞大。作为一篇综述,本文只是将计量经济学中的因果推断方法刻画了素描梗概,难免挂一漏万。例如,本文未涉及因果图(causal graph)(Pearl, 2009)、中介效应分析(尤其是近年兴起的因果中介分析)[①],以及机器学习在因果推断中的诸多应用。即使着重阐述的因果推断方法,因篇幅限制,也有很多变种或细节未包括;譬如,在第三节介绍断点回归时,未涉及"模糊断点回归"(fuzzy regression discontinuity)与"拐点回归"(regression kink design)。尽管如此,本文仍试图将因果推断的发展脉络与精神实质用最生动通俗的语言呈现给读者。

展望未来,不难看出计量经济学中因果推断方法的两大趋势。首先,因果推断方法发展迅猛,创新

① 关于中介效应分析的最新进展,可参考陈强等(2024b)的文献综述。

不断。例如,针对处理组只有一位个体的面板数据,合成控制法(Abadie & Gardeazabal, 2003; Abadie et al., 2010)、回归控制法(Hsiao et al., 2012)、分位数控制法(Chen et al., 2024)出现的历史不过二十年或更短,且已在实践中大量应用。另一方面,一些貌似古老的因果推断方法,例如工具变量法与双重差分法,近年来再次焕发了青春,尤其在弱工具变量、交叠DID等领域。其次,机器学习方法有望在因果推断中发挥日益重要的作用。传统的因果推断方法一般依赖于参数模型,所作的线性假设很可能误设,但传统的非参数统计方法则受制于"维度诅咒"(curse of dimensionality),不适用于高维数据。起源于人工智能领域的机器学习方法则天然地适用于高维的非线性模型。另外,由于因果推断的核心在于预测反事实结果,而机器学习尤其擅长预测,故机器学习在因果推断领域大有用武之地。例如,将Lasso应用于工具变量法(Belloni et al. 2012),因果树(causal tree)(Athey & Imbens, 2016),因果森林(causal forest)(Wager & Athey, 2018),双重机器学习(double machine learning)(Chernozhukov et al., 2018),以及将分位数随机森林应用于分位数控制法(Chen et al., 2024)。总之,我们有理由相信,计量经济学的可信度革命不仅将延续,且会不断壮大,使得未来的因果推断方法更加稳健而可信。毫无疑问,我们正处在因果推断的一个黄金时代。

## [参 考 文 献]


陈强, 齐霁, 颜冠鹏. 断点回归的两大分析框架:我们究竟该用哪一个? 经济学动态, 2024a(11):128—144.

陈强, 齐霁, 颜冠鹏. 双重差分法的安慰剂检验:一个实践的指南. 管理世界, 2025,即将发表.

陈强, 齐霁, 颜冠鹏. 经济学实证研究中的中介效应分析. 工作论文, 2024b.

刘冲, 沙学康, 张妍. 交错双重差分:处理效应异质性与估计方法选择. 数量经济技术经济研究, 2022b(9): 177—204.

刘冲, 诸宇灵, 李皓宇. 断点回归设计:理论前沿进展与新应用场景. 经济学报, 2022a(3): 325—366.

许文立. 双重差分法的最新理论进展与经验研究新趋势. 广东社会科学, 2023(5): 51—62.

张子尧,黄炜. 事件研究法的实现、问题和拓展. 数量经济技术经济研究, 2023(9): 71—92.

Abadie, A. Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, 2005, 72(1): 1-19.

Abadie, A. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 2021, 59(2): 391-425.

Abadie, A., Diamond A., & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American statistical Association*, 2010, 105(490): 493-505.

Abadie, A., Diamond A., & Hainmueller, J. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 2015, 59(2): 495-510.

Abadie, A., & Gardeazabal, J. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 2003, 93(1): 113-132.

Abadie, A., Imbens, G. W. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics*, 2011, 29(1): 1-11.

Anderson, T. W., & Rubin, H. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics*, 1949, 20 (1): 46-63.

Andrews, I., Stock, J. H., & Sun, L. Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, 2019, 11 (1): 727-753.

Angrist, J. D. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administration Records. *American Economic Review*, 1990, 80(3): 313-336.

Angrist, J. D., & Krueger, A. B. Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*, 1991, 106(4): 979-1014.

Angrist, J. D., & Lavy, V. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, 1999, 144(2): 533-576.

Angrist, J. D., & Pischke, J. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the

Con out of Econometrics. *Journal of Economic Perspective*, 2010, 24(2): 3-30.
Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. Synthetic Difference-in-Differences. *American Economic Review*, 2021, 111(12): 4088-4118.
Ashenfelter, O. Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics*, 1978, 60(1): 47-57.
Athey, S., & Imbens G. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of National Academy of Sciences USA*, 2016, 113(27): 7353-7360.
Bai, J., & Ng, S. Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data. *Journal of the American Statistical Association*, 2021, 116(536): 1746-1763.
Baker, A. C., Larcker, D. F., & Wang, C. C. Y. How Much Should We Trust Staggered Difference-in-Differences Estimates? *Journal of Financial Economics*, 2022, 144(2), 370-395.
Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 2007, 122(3): 1235-1264.
Banerjee, A. V., & Duflo, E. The Experimental Approach to Development Economics. *Annual Review of Economics*, 2009, 1: 151-178.
Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 2012, 80(6): 2369-2429.
Black, S. E. Do Better Schools Matter? Parental Valuation of Elementary Education. *Quarterly Journal of Economics*, 1999, 114(2): 577-599.

备注:因空间限制,完整参考文献参见作者个人网页(www. econometrics-stata. com)。

## Causal Inference in Econometrics: The Past, the Present and the Future

Chen Qiang

**Abstract**: Since the Credibility Revolution, causal inference has played an increasingly important role in econometrics. New methods of causal inference emerge constantly, which may bewilder empirical researchers. This paper systematically reviews the historical origins and development of causal inference in econometrics, including mainstream methods for causal inference such as randomized experiment, natural experiment, regression discontinuity, matching estimation, doubly robust estimation, instrumental variable regression, difference in differences, synthetic control method, regression control method, and quantile control method. Compared with existing literature reviews, this paper focuses more on the historical origins of causal inference in order to help empirical researchers understand the essence, while updating the literature to the frontier of 2024 (especially the latest advances in instrument variable regression and difference in differences) and looking forward to the directions of future development.

**Keywords**: econometrics; causal inference; credibility revolution

【责任编辑:周吉梅;责任校对:周吉梅,赵洪艳】