doi: 10.12012/CJoE2021-0027

更精确的因果效应识别: 基于机器学习的视角

钱浩祺 [1,2], 龚嫣然 [3], 吴力波 [4,5]

(1. 复旦大学全球公共政策研究院, 上海 200433; 2. 复旦 - LSE 全球公共政策研究中心, 上海 200433; 3. 复旦大学大数据学院, 上海 200433; 4. 复旦大学经济学院, 上海 200433; 5. 复旦大学大数据研究院, 上海 200433)

摘 要 在传统计量经济方法学中引入机器学习方法已逐渐成为一个不可忽视的重要发展趋势, 本文从因果推断的两大主流分析框架出发, 分析了其各自的特征与内在关联, 在此基础上, 提出了机器学习方法可以从样本匹配与反事实预测两个方面对现有因果效应识别研究进行改进. 本文认为, 机器学习能够通过样本直接匹配以及提升倾向得分估计准确度来实现样本的精准匹配, 使研究样本更具备 "随机化" 实验的特征, 此外, 机器学习方法能够利用复杂关系建模、交叉验证以及正则化等方法来提升样本反事实预测的准确性. 本文接着从匹配法、断点回归法、双重差分法以及合成控制法这四个具体的方法出发, 详细阐述了机器学习在提升因果效应识别方面的理论基础, 同时在每一个方法部分都给出了若干实际应用案例, 以供应用计量研究者借鉴和参考.

关键词 机器学习; 因果推断; 因果效应; 政策评估; 匹配法; 断点回归法; 双重差分法; 合成控制法; 工具变量; 大数据

More Accurate Causal Inference: A Perspective of Machine Learning

QIAN Haoqi[1,2], GONG Yanran[3], WU Libo[4,5]

(1. Institute for Global Public Policy, Fudan University, Shanghai, 200433, China; 2. LSE-Fudan Research Centre for Global Public Policy, Fudan University, Shanghai 200433, China; 3. School of Data Science, Fudan University, Shanghai 200433, China; 4. School of Economics, Fudan University, Shanghai 200433, China; 5. Institute for Big Data, Fudan University, Shanghai 200433, China)

收稿日期: 2021-04-07

基金项目: 国家社会科学基金重大项目 (15ZDB148); 国家杰出青年科学基金 (71925010); 国家自然科学基金 (71703027); 上海市 "科技创新行动计划" 社会发展科技攻关项目 (20DZ1200600)

Supported by Key Program of National Social Science Foundation of China (15ZDB148); National Natural Science Funds for Distinguished Young Scholars (71925010); National Natural Science Foundation of China (71703027); Science and Technology Commission of Shanghai Municipality Grant (20DZ1200600)

作者简介: 通信作者: 钱浩祺, 副研究员, 博士, 研究方向: 政策分析与评估、政策仿真建模、能源与环境经济学、能源与经济大数据, E-mail: qianhaoqi@fudan.edu.cn; 龚嫣然, 博士研究生, 研究方向: 因果推断、机器学习、大数据与政策评估, E-mail: yrgong19@fudan.edu.cn; 吴力波, 复旦大学经济学院教授, 博士, 复旦大学能源经济与战略研究中心主任、国家杰出青年基金项目获得者、教育部青年长江学者, 研究方向: 能源环境气候经济学、政策建模与仿真、大数据应用统计分析, E-mail: wulibo@fudan.edu.cn.

Abstract There is an increasing trend towards combining machine learning methods with traditional econometric methodologies. Starting from comparing features and internal relations of two mainstream causal inference frameworks, this paper proposes that causal inference can be significantly improved with the introducing of machine learning methods in two ways, one is sample matching and one is counterfactual prediction. Firstly, machine learning techniques can enhance matching qualities by pairing samples directly or improving the accuracies of propensity score predictions. This can make the matched samples more similar to samples collected from randomized controlled trials. Secondly, machine learning techniques can improve the accuracies of counterfactual predictions by modeling complex relations, using cross-validation, and adopting regularization. This paper then introduces the theoretical foundations of combining machine learning techniques and causal inferences by reviewing four specific methods: Matching, regression discontinuity, difference-in-difference, and synthetic control method. At the meantime, several application cases are provided in each method section for researchers in applied econometrics as references.

Keywords machine learning; causal inference; causal effect; policy evaluation; matching; regression discontinuity; difference-in-differences; synthetic control method; instrumental variable; Big data

1 引言

定量实证分析已经是社会科学研究中不可或缺的组成部分, 近年来基于 “实验” 思路的实证分析, 更是成为了定量研究领域的前沿热门方向. 无论是基于研究人员有针对性设计的 “随机控制实验” 还是由政策变动或外生事件所产生的 “准自然实验” 所开展的研究, 其共同目标都是对政府政策或干预措施所产生的效果进行可靠评估, 这是一类典型的因果推断 (casual inference) 问题, 其所测算的效应也被称为因果效应 (casual effect) 或处理效应 (treatment effect). 目前在计量经济学领域, 主要使用四种主流的因果推断方法来估计因果效应, 分别是匹配法 (matching)、断点回归法 (regression discontinuity, RD)、双重差分法 (difference-in-differences, DID)、以及合成控制法 (synthetic control method), 这四类方法被广泛应用于财政学 (王玺和刘萌 (2020), 饶茜等 (2020), 高正斌等 (2020))、环境经济学 (齐绍洲等 (2018), 徐佳和崔静波 (2020), 吴力波等 (2021))、劳动经济学 (高玉娟等 (2018), 王妍等 (2019), 王乙杰和孙文凯 (2020)) 以及发展经济学 (毛其淋和许家云 (2016), 张俊 (2017), 俞秀梅和王敏 (2020)) 等各个经济学分支领域的实证研究.

虽然学界采用因果推断方法涌现了大量经济学实证研究, 但是诸如不可观测因素、共同支撑假定、共同趋势假定以及样本分配随机性等问题, 很大程度上限制了现有方法的进一步拓展和应用 (胡咏梅和唐一鹏 (2018)). 其根本原因在于, 因果推断研究所估计的因果效应, 主要由观测样本的实际观测值与其所假想或构造的反事实结果的差异而得到, 而由于前述影响因素的存在, 会不同程度地影响反事实结果的准确性, 并由此造成因果效应估计值的潜在偏误 (陈林和伍海军 (2015)). 在过去的很长一段时间中, 经济学家和统计学家也试图从引入局部核估计 (Heckman et

al. (1998))、逆概率加权估计 (Hirano et al. (2003))、大样本性质 (Abadie and Imbens (2006)) 以及非参数估计 (Abadie (2005), Branson et al. (2019), Lu et al. (2019)) 等多种途径来提升因果效应估计的可靠性, 并且也已经取得了一定的成效.

自 2009 年 "计算社会科学" (computational social science) 理念被首次大规模提出后 (Lazer et al. (2009)), 大数据 (Big data) 与机器学习 (machine learning) 也开始逐渐融入经济学实证研究中, 这一趋势于 2015 年左右呈现出爆炸式增长. 图 1 展示了通过搜索 Web of Science 核心数据库中商业与经济学领域研究中包含 "machine learning" 关键词的研究总数趋势, 从图 1 中可以看到, 相关论文数量从 2009 年起突破 100 篇, 于 2015 年突破 400 篇后迅速增加, 2019 年达到 2676 篇. 与传统经济学实证研究有所区别的是, 结合机器学习的计量经济分析弱化了对方程式中系数估计的关注, 而是更专注于对因变量进行估计, 即对研究对象进行更精准的 "预测" (Mullainathan and Spiess (2017)). 机器学习方法所展示出的强大精准预测能力, 使得经济学家们发现了一个新的提升因果效应估计可靠性的途径, 即利用机器学习方法来提升反事实结果估计的准确性 (Varian (2016), Athey (2017)). 通过对 25 本国际权威经济学期刊过去 10 年的发表论文进行统计, 图 2 的结果表明, 机器学习方法已经得到国际主流经济学界的认可, 与之相关的发表论文数自 2016 年后迅速增加, 并且将其运用于因果推断研究的论文数也大幅增加[1].

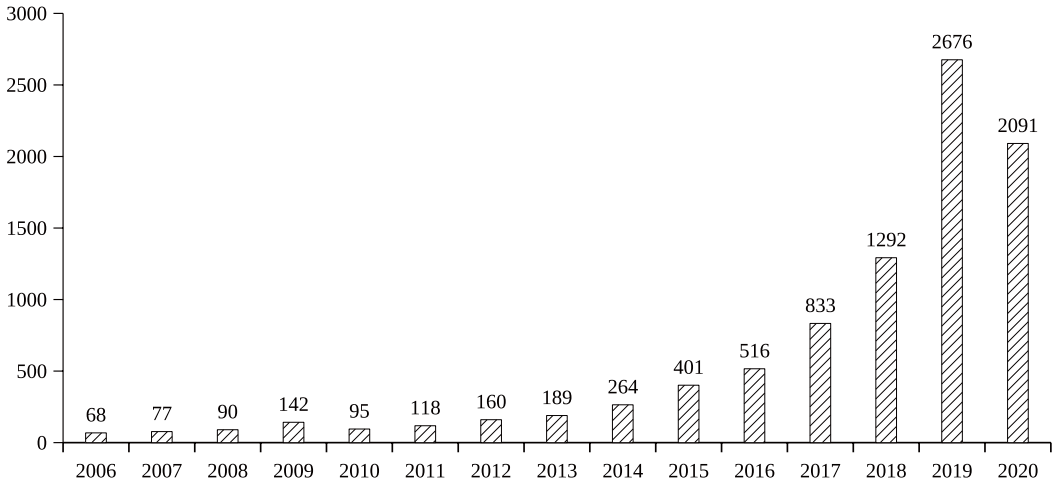

图 1　2006 年至 2020 年商业与经济学领域与机器学习相关研究数量

[1]根据经济学领域的主流分类方法, 其中评级为 A 的期刊共 5 本, 分别是: American Economic Review、Journal of Political Economy、Quarterly Journal of Economics、Econometrica、Review of Economic Studies; 评级为 A– 的期刊共 20 本, 分别是: Journal of Economic Literature、Review of Financial Studies、Review of Economics and Statistics、Journal of Economic Theory、Journal of Monetary Economics、American Economic Journal: Macroeconomics、American Economic Journal: Microeconomics、American Economic Journal: Applied Economics、American Economic Journal: Economic Policy、Journal of Finance、Journal of Financial Economics、Journal of the European Economic Association、Economic Journal、Rand Journal of Economics、Journal of Public Economics、International Economic Review、Journal of International Economics、Journal of Labor Economics、Journal of Econometrics、Games and Economic Behavior.

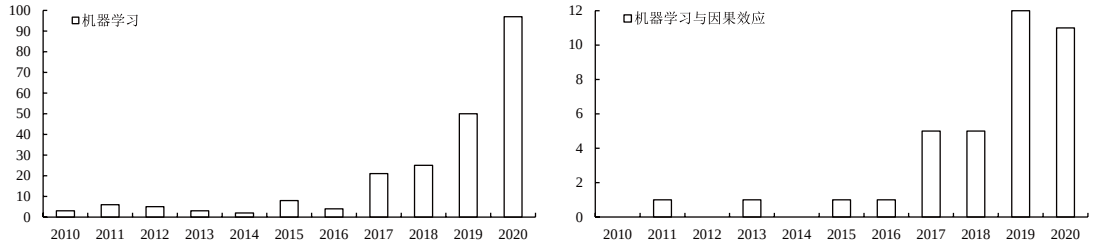

图 2 经济学权威期刊机器学习相关论文数量统计

针对经济学实证研究中的前沿发展领域, 现有学者从不同角度进行了一定的梳理. 例如, 有学者从更宽泛的领域讨论了机器学习方法在社会科学研究中的应用及其影响 (黄乃静和于明哲 (2018), Athey (2019), 陈云松等 (2020), 蔡宗武 (2021), 洪永淼 (2021)), 或从大数据的视角探讨了其与经济学之间的联系 (Athey and Imbens (2019), 洪永淼和汪寿阳 (2020), 萧政 (2021)), 也有学者更有针对性地讨论了机器学习方法在政策评估或因果效应估计领域的应用 (Kleinberg et al. (2015), Athey and Imbens (2017), 胡咏梅和唐一鹏 (2018)). 但是, 现有研究或聚焦于广义的经济学研究讨论, 或从机器学习不同方法视角切入开展研究, 对于机器学习方法究竟如何能够提升因果效应识别尚缺乏系统性的完整讨论. 基于此, 本文聚焦理论层面对机器学习方法在因果效应识别领域的改进原理进行研究, 并结合现有文献中的实际应用对其效果和进一步提升空间进行分析.

本文的结构安排如下: 第二部分从因果推断分析的核心概念出发, 对因果推断分析框架进行概述, 并对其潜在的关键改进方面进行分析; 第三部分详细分析机器学习方法在四种典型因果效应识别方法中的理论改进与具体应用; 第四部分对全文进行了总结, 并对计量经济学与机器学习方向进行交叉结合的潜在突破方向和需进一步解决的问题进行了展望. 通过本文的研究和分析, 期望给计量经济学理论研究者和实证分析研究者提供较为全面的前沿研究梳理, 能够有助于我国研究者在该领域能够进一步拓展和提升.

2 因果推断分析框架

2.1 结构因果模型框架与潜在结果框架

2.1.1 结构因果模型框架

因果推断研究有多种不同的分析框架, 但是目前主流研究主要集中在其中两种. 第一种因果推断框架为 “结构因果模型” (structural casual model, SCM) 框架, 这一概念可追溯至 Wright (1921) 提出的路径分析 (path analysis), 在这一框架下, 因果关系由以下一系列非线性和非参数方程所组成的结构方程组 (structural equation models, SEMs) 来刻画:

$$x_i = f_i(\mathrm{pa}_i, u_i), i = 1, \cdots, n, \tag{1}$$

其中, x_i 为我们所关注的对象变量, pa_i 为直接影响对象变量的父类变量, u_i 则为无法观测的扰动因素. 在经济学研究中, 纯理论建模通常会在一系列前提假设的情况下构建由式 (1) 所示的 SEMs

并进行均衡求解, 由此提出相应的理论假说, 随后在实证分析中, 通常会将式 (1) 的 SEMs 进一步展开为以下线性形式的简约式方程来进行参数估计和参数检验:

$$x_i = \sum_{k \neq 1} \alpha_{ik} + u_i, i = 1, \cdots, n, \tag{2}$$

其中 $\alpha_i \neq 0$ 所对应的变量, 即对应为式 (1) 中的父类变量. 上述 SCM 框架后续被 Pearl (1995, 2009) 使用 "因果图" (causal diagram) 和 "有向无环图" (directed acyclic graph, DAG) 的概念进行了发展推广. 图 3 展示了一个用 DAG 方法表述的经典经济学研究问题, 其中 Q 表示需求量, 其受到收入 I 和不可观测因素 U_1 影响, 而 P 表示价格水平, 其受到工资水平 W 和不可观测因素 U_2 影响, 同时需求量和价格水平之间也存在相互影响, d_1、d_2、b_1 和 b_2 分别为对应 SEMs 形式中的系数及其影响的大小.

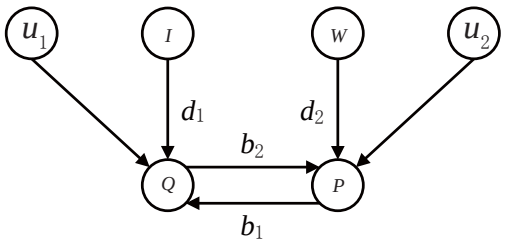

图 3　经济学研究典型问题的 DAG 表述方法

2.1.2　潜在结果框架

第二种因果推断框架则被称为 "潜在结果" (potential outcome, PO) 框架, 其可追溯到 Neyman (1923) 所提出的随机化实验研究, 在这一框架下, 因果效应被视作是实验中样本的实际结果与受到随机处理后所产生的潜在结果之间的差异, 即在一个 N 个样本的随机实验中, 若有 n_0 个未被处理的样本和 n_1 个被随机处理的样本, 则该实验的无偏平均因果效应 (average treatment effect, ATE) 被定义为:

$$\text{ATE} = \sum_{i=1}^{N} \frac{Y_i^1 - Y_i^0}{N}, \tag{3}$$

其中 Y_i^1 和 Y_i^0 为观测样本的潜在结果, 上标表示是否受到处理, 我们仅能观测到两者中的一个. 此时, ATE 的无偏估计量为:

$$\widehat{\text{ATE}} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} Y_j. \tag{4}$$

该框架后续则被 Rubin (1974, 2005) 拓展至非随机实验领域, 从而形成了著名的 "Neyman-Rubin 模型". 假设用哑变量 D 表示是否受到政策干预, $D = 1$ 表示受到政策干预, 其样本集合被称为处理组 (treatment group), $D = 0$ 则表示没有受到政策干预, 其样本集合被称为控制组 (control group). 处理组和控制组的潜在观测结果分为 Y_1 和 Y_0, 则实际观测结果可表示为 $Y = DY_1 + (1 - D)Y_0$, 此时 ATE 估计量为:

$$\text{ATE} = E(Y|D = 1) - E(Y|D = 0), \tag{5}$$

其中 $E(\cdot|\cdot)$ 为条件期望算子. 若研究者更关注参与者平均因果效应 (average treatment effect on the treated, ATT) 或非参与者平均因果效应 (average treatment effect on the untreated, ATU), 则可以通过以下两式分别得到 ATT 和 ATU 的估计量:

$$\text{ATT} = E[Y_1 - Y_0|D = 1], \tag{6}$$

$$\text{ATU} = E[Y_1 - Y_0|D = 0]. \tag{7}$$

2.1.3 两种框架的比较

从理论模型表达方式上看, SCM 框架和 PO 框架存在着较大的差异, 但是 Pearl (2009) 证明在非递归条件下, 这两个框架在理论上是等价的, 但这一等价性并不意味着在概念化或实践方面也是等价的. 例如, 由于 SCM 框架更专注刻画变量间的复杂关系, 因此常被用于经济学理论研究 (Heckman (2000)); SCM 框架还可以利用因果图工具对各类变量之间的关系进行直观的分析和解释, 因此在计算机、人工智能等领域得到了较广泛的应用 (Karimi et al. (2020), Löwe et al. (2020)). 此外, 在与因果推断主题相关的教材中, 基本也会同时结合两种框架来阐述相关概念 (Lee (2016)). 但是正如 Imbens (2020) 指出, PO 框架相对于 SCM 框架, 其在前提假设、经济理论关联性、模型简化性、异质性假设以及实验设计这五个领域相对于 SCM 框架存在优势, 从而促使 PO 框架成为了经济学实证领域中的主流框架, 也正是基于这一考量, 本文将主要探讨在 PO 框架下应用机器学习来提升因果效应识别准确性的问题.

2.2 机器学习与因果推断框架

PO 因果推断框架在实践运用中往往会进行一定程度的简化, 例如采用线性化假定、设定通用函数形式等做法, 以便于构建实证模型来进行因果效应估计. 但是这些做法可能在很多情况下会引入偏误, 导致估计结果产生偏差, 由于这些简化做法存在固有缺陷, 使得简单的方法优化与修正难以完全克服估计结果的偏差. 上述原因为通过引入机器学习方法来改进因果效应识别提供了可能性, 而改进路径主要体现在两个方面: 第一, 加强样本匹配的 “随机性”, 即提升控制组样本与处理组样本之间匹配过程的随机性; 第二, 提升反事实预测的准确度, 即对控制组样本和处理组样本的不可观测效应进行准确预测.

2.2.1 样本匹配随机性

在社会科学研究中, 往往很难去进行一个真正 “随机化” 的社会实验, 因此需要充分利用实际观测数据来尽可能地创造 “随机化” 条件, 即在 “Neyman-Rubin 模型” 的 PO 框架下对因果效应进行准确识别. 在处理 “非随机” 观测数据时, Rosenbaum and Rubin (1983) 在其经典论文中证明并强调了倾向得分 (propensity score) 在因果推断研究中所处的核心地位. 从理论上看, 倾向得分被定义为在给定样本特征 X 时不同样本接受处理的条件概率, 即 $P(D = 1|X)$. 倾向得分之所以重要, 因为它是能够将 “非随机化” 观测数据尽量模拟成 “随机化” 实验的重要桥梁, 通过选取倾向得分相近的处理组和控制组样本, 使得分组样本更接近于从总样本中随机挑选而出的随机样本, 样本之间的可比性大大增强, 从而提升因果效应估计准确性.

鉴于倾向得分是一个抽象表达式, 在将这一理论概念应用于实证分析时, 就必须首先对倾向得分的具体数值进行估计, 再基于倾向得分估计值对样本进行匹配后估计因果效应, 即为因果效

应识别的 “两步法”. 在第一步中, Logit 模型是目前经济学实证领域中最为广泛应用的模型, 和其他众多模型一样, Logit 模型利用观测数据拟合回归模型, 并用该回归模型对所有观测值计算出一个范围为 0 至 1 之间的拟合值, 作为各观测样本的倾向得分估计值. 上述估计步骤的本质是对倾向得分进行预测, 因此机器学习能够借助其强大的样本内与样本外预测能力, 来提升对倾向得分的估计准确性 (Pirracchio et al. (2015), Cannas and Arpino (2019)). 而且在实际研究中, 经常会在大数据环境下使用高维协变量来匹配或者控制个体与群体之间的特征差异 (黄乃静等 (2018)), 因此可以利用机器学习中的降维方法来进行协变量筛选, 以避免错误选择协变量而导致有偏的匹配结果 (Varian (2014)). 此外, 随着机器学习方法的快速发展, 逐渐出现利用回归树 (regression tree) 及其拓展方法等前沿算法绕开传统倾向得分估计步骤的理论框架与应用 (Athey and Imbens (2016)), 此时倾向得分匹配这一步骤被隐含在了机器学习算法的过程之中, 而样本间匹配的过程则被转化为了机器学习问题中的聚类问题. 上述这些发展, 均表明机器学习方法在提升样本匹配随机性方面所具有的巨大潜力.

2.2.2 反事实预测

在 PO 框架中, 式 (5) 中等号右侧要求同时得到样本的实际观测值与反事实观测值, 但是在实际情况中, 等号右侧的两项中只能获得其中一项的观测数据, 即被处理的结果 $E(Y|D=1)$ 或未被处理的结果 $E(Y|D=0)$. 这一问题在现有文献中也被视作是一个典型的 “缺失数据” 问题 (Swanson et al. (2018)), 因此如何合理的估计另一项, 是因果效应识别中的核心难点. 除了前文所述的提升样本匹配随机性之外, 可以利用已有观测数据对未知项进行反事实预测, 通过计算实际观测值与反事实预测值两者之间的差异来估算因果效应. 在 PO 框架下, 因果效应的识别问题被转化为了纯粹的精准预测问题, 因而机器学习方法同样能够凭借其样本内与样本外的精准预测能力, 来提升因果效应识别的准确性 (钱浩祺 (2020)).

与简单线性模型相比, 机器学习模型允许在因变量和自变量之间存在更灵活的关系, 因而能够更好地抓住变量间的复杂非线性关系, 也能够更好地利用非数值变量中所蕴含的信息, 来提升预测准确性 (Varian (2014)). 因此, 机器学习方法能够减少传统框架下所面临的前提假设束缚来进行因果效应识别, 例如, 当双重差分模型不满足平行趋势假定时, 可以直接使用机器学习方法来进行反事实预测并估计因果效应 (Cicala (2017)).

此外, 机器学习方法中也充分探讨了模型集成方法对提升预测准确度的贡献, 即集成多种预测方法来降低单一方法可能存在的预测结果偏误. 例如, 因果森林预测方法通过集成大量因果树模型, 来降低单一因果树的预测偏误, 并且大量的实验都证明了因果森林方法的预测结果要优于单一因果树模型 (Wager and Athey (2018)), Athey, Bayati and Imbens et al. (2019) 则发现在面板数据中使用集成方法能比使用传统方法得到更为精准的预测结果.

目前在经济金融政策、公共政策和健康政策等领域, 人们越来越关注个体或者子群体对某一项政策或措施的响应 (Heckman and García (2017), Shalit et al. (2017), Yoon et al. (2018), Jesson et al. (2020)). 借助于机器学习方法的预测优越性, 能够提升因果效应识别的准确性, 并被应用于上述异质性效应的估计, 这将是一个极具发展前景的理论研究和实践应用方向.

3 基于机器学习的因果效应识别提升方法

3.1 匹配法

由前文所述, 样本匹配是因果效应识别中极其重要的一环, 在社会科学研究中, 双胞胎匹配往往被看作是最理想的配对样本, 可以通过控制不可观测的个体因素而得到较为准确的因果效应 (McGue et al. (2010)). 但是对于绝大多数研究问题, 几乎无法找到类似 “双胞胎匹配” 这样的配对样本, 因此需要充分利用已获取的观测数据, 来尽可能提升样本之间的可比性. 传统的样本匹配方式主要分为两种: 一种是通过协变量进行直接匹配, 第二种是通过倾向得分进行匹配 (Steiner and Cook (2013)), 机器学习方法在上述两种方法中均能够被应用, 并为匹配效果带来一定的提升.

3.1.1 直接匹配

现有研究中, 使用协变量进行样本匹配的核心思想, 是构造一个基于多维协变量的标量 “距离” 值, 通过设定卡尺 (caliper) 来筛选距离处理组样本较近的控制组样本来进行配对匹配, 以此计算以下配对匹配估计量 (pair matching estimator) 来作为因果效应的估计值:

$$\mathrm{ATE} = \frac{1}{N_t}\sum_{t\in T}(Y_t - Y_{c(t)}), \tag{8}$$

其中 $t \in T$ 表示样本 t 属于处理组, $c(t)$ 表示匹配上的控制组样本, N_t 表示处理组样本数量. 由于距离函数的形式多样, 因此选取不同的距离函数会得到不同的匹配估计量, 并且也会因数据集的特点而发生改变.

在大数据背景下, 许多可供研究使用的数据集的样本数量快速增加, 并且样本的协变量特征也不断丰富. 在这种情况下, 基于机器学习方法的样本匹配将带来更优的匹配效果, 除了可以估计群体的平均因果效应之外, 也能够进一步地估计子样本分组因果效应、个体因果效应与分位数因果效应. 例如, Schwab et al. (2018) 在最近邻匹配的思想上结合神经网络提出了完美匹配的概念 (perfect match) 来应对不同的情景与数据集, Imai and Ratkovic (2013) 将协变量与处理变量的交互项加入回归模型, 并利用 LASSO 回归方法来估计异质性因果效应, Diamond and Sekhon (2013) 引入进化搜索算法, 用于改善样本匹配后的协变量平衡问题, 而 Louizos et al. (2017) 则提出了一种基于有限混合模型进行样本分类的因果效应识别方法, 该方法能够在处理变量为隐变量的情况下对样本进行分类并对因果效应进行识别.

在这些引入机器学习方法的研究中, 回归树模型是最为常见的一种应用. 回归树模型的本质是根据样本的协变量特征来对其进行分组, 在构造 “树” 的过程中, 顺序依次将协变量与阈值进行比较, 并将样本不断划分进入不同的分组中, 通过最小化各组残差平方和来确定回归树的最终形状, 最后对处于同一个分组内的控制组样本与处理组样本计算其对应的因果效应. Su et al. (2009) 提出了交互树 (interaction tree) 的概念, 通过引入交互树的随机森林直接识别影响因果效应异质性的变量重要程度. Chipman et al. (2010) 则在随机森林模型的基础上加入贝叶斯算法, 提出贝叶斯累加回归树 (bayesian additive regression trees, BART) 方法, 用以自动识别变量非线性关系以及估计异质性因果效应 (Green and Kern (2012)), 虽然 BART 方法有着较为不错的实证表现, 但是其大样本统计性质还尚未得到深入的研究. Athey and Imbens (2016) 则提出了

一个更具影响力的因果树 (causal tree) 概念, 其中, 全体样本被分为两组: 一组专门用以构建决策树, 而另一组则专门用以估计子样本因果效应, 这种被称为诚实估计 (honest estimation) 的估计算法可以将样本划分成具有不同程度因果效应的子样本, 并得到这些子样本的无偏异质性因果效应. 在此基础上, Wager and Athey (2018) 结合因果树和随机森林提出了因果森林 (causal forest) 算法, 该方法不仅能比传统的近邻匹配方法取得更好的样本匹配结果, 而且在异质性因果效应估计方面也具有逐点一致性和渐进正态性. 上述方法目前被 Athey, Tibshirani and Wager (2019) 提出的广义随机森林 (generalized random forest) 方法所统一. 更加准确的异质性因果效应估计, 能够帮助决策者制定更加有针对性的政策, Davis and Heller (2020) 选择对美国青年参加暑期工读项目 (summer youth employment programs, SYEP) 所受的影响来进行随机实验研究, 并利用因果森林方法对青年人样本进行精准分组, 并估计项目所造成的异质性影响. 研究结果清晰揭示了 SYEP 所针对的 "失联青年" 并不是受益者, 反而是那些更年轻、更有学习意愿、西班牙裔以及过往表现更好的青年人得到了更大的收益, 这一结论挑战了 SYEP 会通过增加犯罪机会成本而降低青年犯罪率的传统理论, 体现了异质性因果效应在政策设计领域中的重要性. 更为重要的是, 该例子展现了机器学习方法能够有效避免传统分层法 (stratification) 和传统基于交互项方法在面临高维变量时可能遇到的维数灾难问题, 由此获得更加优异的异质性因果效应估计效果. 这一方法也被 Handel and Kolstad (2017) 用来研究可穿戴设备对用户的异质性影响以及胡安宁等 (2021) 用来研究中国精英大学教育回报所产生的异质性影响等.

实践中受制于数据等因素, 我们不一定能够得到子样本组或个体的精确因果效应, 但是对于特定决策领域, 其关注点更聚焦于整体样本的分布特征, 因此只要能够得到一项政策对于样本群体在整个分布不同分位点上的异质性影响, 就足以实现决策效能的提升. 例如在劳动经济学领域, 实施政策对人群收入分布的影响是一个重要的关注领域 (朱平芳和邸俊鹏 (2017)), 又如在企业决策中, 由于极端影响可能会对用户留下更强烈的长期负面印象 (Baumeister et al. (2001)), 因此需要考察一项新措施对用户所产生影响的具体分布情况, 以尽可能减少尾部分布对用户产生的极端影响. 上述逻辑同样适用于在金融市场中用来防范极端负面影响, 因此也需要对分位数或分布影响进行估计. 在上述情况下, 相应的因果效应被称为分位数因果效应 (quantile treatment effect, QTE), 其定义为在分位数 τ 上, 两个逆累积分布函数 $F_{Y_1}^{-1}(\cdot)$ 的差值 (Doksum (1974), Lehman and D'Abrera (1975)):

$$\mathrm{QTE} = F_{Y_1}^{-1}(\tau) - F_{Y_0}^{-1}(\tau). \tag{9}$$

在分位数因果效应的估计中, 涉及到的是一类特殊的匹配, 即对控制组和处理组因变量的逆累积分布函数进行分位点匹配, 匹配结果的好坏将直接影响分位数因果效应估计的准确性. Meinshausen and Ridgeway (2006) 指出在一般的分位数回归问题中, 回归森林树方法能够得到因变量的完整条件分布估计, 未来具有用于改进分位数因果效应估计的潜力. Kallus et al. (2019) 提出了一种基于机器学习来估计局部分位数效应及其条件在险值的方法, 称为 LDML (localized debiased machine learning) 方法, 并以美国 401K 养老计划为例, 将 LDML 方法与传统分位数因果效应分析方法进行对比, 结果发现 LDML 方法在分位数因果效应分析方面能够较好地提升估计过程的整体效率. 总体而言, 将机器学习应用于分位数因果效应的研究, 目前尚处于初级起步

阶段.

3.1.2 倾向得分匹配

由前文所述, 倾向得分是在缺乏进行随机化条件下利用已有观测和统计数据进行因果效应识别的核心要素. 但是在实际应用中, 估计一个相对准确的倾向得分往往具有较大的难度, 主要体现在三个方面: 1) 影响样本是否被纳入处理组的因素复杂, 难以用简化的线性模型进行准确估计; 2) 由于不可观测因素的存在造成估计上的偏误; 3) 当样本数量较大或协变量维度较高时, 无法有效利用数据中蕴含的信息、或者由于数据稀疏问题而造成预测偏误. 以上问题, 都能够利用机器学习方法在一定程度上进行改进和完善.

首先, 机器学习方法可以通过构建表达式更加复杂的非线性模型, 来捕捉倾向得分中的复杂变量关系. Setoguchi et al. (2008) 比较了传统的 Logistic 回归、决策树和神经网络 (neural networks, NN) 等方法分别应用在倾向得分计算环节的表现并进行了模拟研究, 结果显示 NN 方法对于倾向得分的预测效果最优, Lee et al. (2010) 则在此基础上考察了更多机器学习模型, 如分类树、回归树、决策树、随机森林和提升树等模型在预测倾向得分上的表现优劣, 并且发现使用集成方法加强的树算法具有更小的偏误和更加一致的置信区间.

其次, 机器学习方法能够通过交叉验证 (cross validation) 方法来提升对倾向得分的样本外预测准确度, 这体现了机器学习方法以数据驱动为指导思想来提升预测能力的特点 (Athey and Imbens (2017)). 通过将样本分成不同的子样本, 每次将其中一份子样本作为验证集计算子样本因果效应, 剩下的样本用于机器学习训练模型, 如此往复循环, 当所有子样本都经过验证后, 取其所有子样本的因果效应平均值作为最终因果效应的估计值, 便能够极大地提升因果效应估计的准确性 (Chernozhukov et al. (2018)).

第三, 机器学习虽然具有很强的预测能力, 但是模型训练中也极易造成过拟合问题, 从而反过来降低其样本外预测能力. 为了克服过拟合问题, 可以在训练倾向得分拟合模型时, 通过在目标函数中加入惩罚项来解决, 这被称为正则化方法. 但是 Chernozhukov et al. (2017, 2018) 也指出使用正则化方法也有可能因正则偏差效应而造成因果效应估计量的一致性遭到破坏, 从而提出了双重机器学习 (double/debiased machine learning, DML) 方法来消除这一估计偏误, DML 方法运用两次机器学习方法对因果效应进行估计, 第一步中使用机器学习方法对传统回归方程进行训练, 机器学习引入的正则化方法在这个过程中会导致参数估计结果无法满足无偏性, 因此在第二步中, 将上一步经过正交化的残差序列视为工具变量来进行拟合, 得到最终的因果效应估计值, 上述过程可以通过交叉验证法来提升模型的样本外预测能力. 在实际应用中, Dube et al. (2020) 在研究劳动力市场中的垄断问题时, 利用双重机器学习方法帮助分离了报酬中不可观测的外生变化.

在因果效应估计方法中被长期广泛使用的工具变量法 (instrumental variable, IV) 与倾向匹配得分之间具有紧密的关联性, 但这一特性往往被人所忽略. 工具变量法通常被用于解决因果效应识别过程中的内生性问题 (Clarke and Windmeijer (2012)), 其基本原理是通过寻找一个工具变量, 使其与解释变量相关而与其他混杂因素无关, 从而来构建工具变量估计量, 实践中往往采用两阶段最小二乘估计 (two-stage least squares, 2SLS) 来实现这一目标. 目前已有学者从一般意

义上讨论了机器学习对工具变量法的贡献 (王芳等 (2020)), 但是从理论上看, 工具变量法与 PO 框架存在着十分紧密的联系 (Angrist et al. (1996)), 即当解释变量被替换为二元处理变量 D 时, PO 框架便成为了一种特殊形式的工具变量法. 在这种形式的 2SLS 估计中, 第一阶段估计工具变量对处理变量的关系, 第二阶段则估计处理变量对因变量的因果效应, 因此在 PO 框架下, 工具变量法的本质就是对倾向得分进行预测, 因而上述机器学习方法的优势也都是适用的, 大量研究者也在工具变量法的实践应用中引入机器学习. 例如, 现有大量研究尝试利用机器学习的正则化方法和非线性建模方法等来优化工具变量的构造过程 (Belloni et al. (2017), Hartford et al. (2017), Singh et al. (2020)), 此外, 研究者也在工具变量法中运用诸如正则化或平滑参数等机器学习方法来更精准的预测倾向得分 (Belloni et al. (2012), Carrasco (2012)) 、使用机器学习算法针对高维度数据进行估计 (Singh and Sun (2019)) 以及利用贝叶斯机器学习算法来估计工具变量情景下的异质性因果效应 (Bargagli-Stoffi et al. (2019)) 等.

3.2 断点回归法

断点回归法是一种较为特殊的因果效应识别方法, 相比其他方法, 断点回归方法在研究设计之初已经完成了控制组与处理组的样本匹配过程. 例如, 在不同的研究问题中, 断点可以为时间 (Hausman and Rapson (2018), 吴力波等 (2020))、考试成绩 (Jepsen et al. (2017))、甚至是地理位置分界线 (徐志伟和刘晨诗 (2020), Jia et al. (2021)) 等, 当研究的样本对象分别位于断点附近的位置但却受到了不同的政策对待时, 其被天然地分为了控制组和处理组. 当样本在政策前没有受到特殊的区分对待时, 我们认为两个分组中的样本在各个协变量特征方面是无限接近的, 样本之间存在高度可比性. 正因为断点回归在研究设计中已经包含了样本匹配的过程, 因此其最终目的是对断点两侧的因变量真实值进行估计, 从而计算局部平均因果效应 (local average treatment effect, LATE). 断点回归估计主要分为两类, 第一类是精确断点回归 (sharp regression discontinuity, SRD), 第二类是模糊断点回归 (fuzzy regression discontinuity, FRD).

在精确断点回归中, 假定存在一个驱动变量 R (running variable), 使得当该变量处于断点 r 两侧时, 因变量 Y 可能存在一个明显的跳跃, 而两侧对应样本的被处理概率分别为 0 和 1, 相应的 LATE 则由下式估计得到:

$$\text{LATE} = E[Y_{1i} - Y_{0i}|R = r] = \lim_{R\to r^+} E[Y_i|R] - \lim_{R\to r^-} E[Y_i|R]. \tag{10}$$

式 (9) 中的 LATE 也可通过估计式 (10) 中的 δ 系数所得到:

$$Y_i = f(R_i) + \delta D_i + \epsilon_i, \tag{11}$$

其中 $f(R_i)$ 表示驱动变量在断点附近对因变量的影响, 其函数形式一般而言是未知的, 在实践中会选择线性表达式、多项式表达式以及非参数表达式等多种形式来进行拟合, 但是控制全局高阶多项式的拟合方法在噪声估计、阶数敏感性等方面存在着较为明显的缺陷, 因此被研究者所诟病 (Gelman and Imbens (2019)). 此时, 利用机器学习方法可以更好地捕捉驱动变量与因变量在断点附近的关系, 从而得到更加精确的断点处预测值. 并且机器学习方法也能够更好地捕捉断点两侧可能存在的门槛效应, 以此避免传统方法可能造成的估计偏误. 例如, Imbens and Wager (2019) 提出了一种基于数据驱动 (data-driven) 的方法对断点回归进行了改进, 通过借鉴类似机器学习中的正则化思路, 利用凸优化方法直接计算得到断点回归中的效应估计值, 并且能够为无论是连

续型还是离散型驱动变量提供置信区间的估计值, 作者利用该方法研究了调整最低离校年龄政策对未来收入影响以及参加暑期学校对未来成绩影响等问题. 在断点处的样本权重设定方面, 传统的局部线性回归方法往往会为靠近断点的样本设置更大的权重, 对此 Branson et al. (2019) 提出了一种基于高斯过程的回归方法, 使得在估计因果效应时可以更灵活的确定权重, 从而提升因果效应估计结果.

断点回归对于模型设定和参数设定也存在较强的依赖性. 例如, 断点回归对于断点两侧的带宽选择也较为敏感, 即使是使用机器学习方法, 若训练样本中包含了太多无效信息, 也会在一定程度上影响在断点处的预测准确性. 为此, Imbens and Kalyanaraman (2012) 基于数据驱动的思想, 利用交叉验证法提出了一种最优的带宽选择方法, 并且该方法在实际数据集上取得了较为出色的表现. 此外, 在断点回归中, 通常会通过添加处理前的协变量来提升因果效应估计的准确性, 但加入额外的变量反而可能会对估计结果和标准误产生较大影响, 这一点在小样本研究中尤为明显. 鉴于上述情况, Anastasopoulos (2019) 提出可以通过引入机器学习中的 LASSO 算法, 使其自动选择最优模型, 来提升整体的估计效果. 在实际应用中, Chen et al. (2021) 等研究了中国税率降低对企业研发投资的影响, 由于当研发投资占比超过一定阈值的企业能够享受税率优惠, 因此该阈值就成了一个能够估计政策效果的断点, 交叉验证法被用来证明断点回归估计结果的稳健性. 最后, 在断点回归中, 如果样本是否受到政策处理并不仅仅由驱动变量是否在断点两侧而决定, 而是还受到其他因素影响的情况下, 这时候研究问题便成了模糊断点回归问题, 此时问题将变得更加复杂, 虽然 Imbens and Wager (2019) 提出的基于数据驱动的方法仍然能够适用于 FRD 情况下的因果效应估计, 但是诸如估计量的渐进性质以及其结论的外推性质仍有待未来的进一步研究 (Bertanha and Imbens (2020)).

Narayanan and Kalyanam (2020) 的研究则给了我们在断点回归中如何更好地应用机器学习以较大的启发. 在目前数据驱动型商业模式不断发展的背景下, 互联网公司尤其是平台型公司, 他们业务中很重要的一部分, 就是使用前沿的机器学习算法并基于多维特征属性来给用户进行 “打分”, 企业进而根据得分情况划分用户类型并实行针对性的经营策略. 此时, 由于绝大多数机器学习算法计算出的用户得分都是连续的, 因此就在不同用户分类得分阈值处产生了断点, 而且这类断点无法被用户本身所观测到. 在上述情况下, 机器学习算法便能够在断点识别和用户匹配方面发挥极大的用武之地, 实现 “以己之矛, 攻己之盾” 的效果.

3.3 双重差分法

双重差分法和断点回归方法不同, 控制组与处理组之间的样本因变量差异是 “第一重差分”, 两个样本组的因变量差异在两个时期之间的差异则是 “第二重差分”, 利用上述双重差分的步骤得到的因变量差异就是估计所得的因果效应, 因此时间因素是双重差分法中的重要组成元素. 目前众多的经济学因果效应实证研究中, 双重差分法是最为广泛使用的一种方法, 原因除了实际操作方法简单易用外, 经济变量往往以面板数据形式为主, 因而符合双重差分法的数据前提条件. 此外, 各类经济社会政策不仅数量较多, 而且这些政策实施能够被视为是一种 “准自然实验”, 较适合运用双重差分法来开展政策效果评估研究.

双重差分法作为一种应用于潜在因果框架的估计方法, 自然也需要考虑样本随机性因素. 如

果实证研究基于随机控制实验 (randomized controlled trial, RCT), 其样本随机性能够得到满足, 则可以直接使用双重差分法的公式计算对应的平均因果效应. 但是, 由于大多数经济社会政策的实施是一种 “准自然实验”, 因此在绝大多数情况下数据样本并不满足随机性要求, 在这种情况下, 现有研究者往往会首先对样本进行直接匹配或者通过倾向得分进行匹配, 再使用双重差分法来估计因果效应, 即 Matching-DID 和 PSM-DID 等各类改进方法 (唐为和王媛 (2015), 张琦等 (2019), Zhu et al. (2019), 陈浩等 (2020), 王班班等 (2020)). 在这种情况下, 可以利用前文所述的基于机器学习的匹配方法来进一步改进第一步的匹配过程, 从而实现对双重差分法进行提升的目的.

由于双重差分法中涉及到了时间这一重要元素, 因此一个满足良好样本随机性的匹配结果, 应当能够使得经过匹配的控制组和处理组样本的因变量呈现出 “平行趋势” (parallel trend), 即当没有政策干预存在的情况下, 两个样本组的因变量应该以相同的趋势发生变化 (Lee (2016)), 其数学含义由式 (12) 所表示:

$$E[Y_{T_2}^0 - Y_{T_1}^0 | D = 1] = E[Y_{T_2}^0 - Y_{T_1}^0 | D = 0]. \tag{12}$$

当满足平行趋势假设前提时, 反事实结果 $E[Y_{T_2}^0 | D = 1]$ 是可以被准确估计的, 从而可以通过双重差分法求出参与者平均因果效应 $E[Y_{T_2}^1 - Y_{T_2}^0 | D = 1]$, 如式 (13) 所示:

$$E[Y_{T_2}^1 - Y_{T_2}^0 | D = 1] = (E[Y_{T_2} | D = 1] - E[Y_{T_1} | D = 1]) - (E[Y_{T_2} | D = 0] - E[Y_{T_1} | D = 0]). \tag{13}$$

但是, 当平行趋势假设无法得到满足时, 通过式 (13) 所估计得到的因果效应就会存在偏误, 双重差分法的结果可靠性会受到较大影响, 并且随着样本期长度的增加, 平行趋势假设能够得到满足的可能性也不断下降, 因果效应估计结果有偏差, 这也是长期以来双重差分法所面临的挑战之一. 虽然也有学者指出平行趋势的假设过于严格, 在适当的情况下可以放宽该假设, 即不同样本组因变量的变动趋势处于一个合理区间内即可 (Manski and Pepper (2018)), 但是尽量满足平行趋势假设依然是目前实证研究中的主流做法. 造成样本数据违背平行趋势假设的原因, 往往在于实验选择和因变量受到多重混杂因素的影响, 其中又可以分成两个方面: 第一个方面, 遗漏了部分重要混杂因素, 导致因变量的变化不再严格外生, 因此造成遗漏变量偏误. 第二个方面, 虽然充分考虑了混杂因素的影响, 但是其影响形式采用了线性模型来进行简单刻画, 遗漏了混杂因素可能造成的非线性影响, 因此造成模型设定偏误. 在传统实证领域中已有不少研究试图找到这个问题的答案, 如 Heckman et al. (1997) 考虑了当存在由不可观测因素导致的实验选择性偏误时, 需要改善匹配方法并提出了经过半参数回归调整的广义 DID 估计量, Abadie (2005) 在此基础上进一步发展了半参数双重差分法, 使得其能够更好地估计异质性因果效应, Athey and Imbens (2006) 则提出了一个广义非线性双重差分模型, 该模型能够估计出结果变量的反事实分布情况, 用于更好地估计分位数因果效应和异质性因果效应等, 还给出了估计量的统计推断性质.

机器学习方法在解决上述造成平行趋势不满足的问题中也能够发挥其积极作用, 但是直接将机器学习运用于改进双重差分方法的应用目前并不多见. 机器学习能够借助前沿算法来深入挖掘潜在混杂因素与因变量之间的复杂关系, 从而增加样本满足平行趋势的可能性. Knittel and Stolper (2021) 使用双重差分法研究了美国加利福尼亚州圣马科斯的家庭能源报告措施对家庭用电量的影响, 虽然该措施基于随机控制实验所开展, 但是作者在使用因果森林方法进行因果效应

估计时, 依然使用了 DML 方法中的正交残差序列来获得尽可能无偏的估计效果. Ning et al. (2020) 则基于 Abadie (2005) 半参数 DID 估计量, 提出了在 DML 方法的基础上, 通过控制高维协变量来将平行趋势假设放宽为条件平行趋势 (conditional parallel trend) 假设, 并以此来估计异质性因果效应. Chang (2020) 同样提出了基于机器学习方法的半参数 DID 估计方法 (DML-DID), 能够利用高维变量使得数据满足条件平行趋势, 为实践研究者在利用多样化的机器学习方法和大数据集来估计双重差分法下的因果效应提供了指导.

除了上述对复杂混杂因素的改进处理使数据满足平行趋势假设, 另一种改进双重差分法的路径, 是通过机器学习方法来对反事实潜在结果进行更加精确的预测, 从而得到准确的因果效应估计结果. Athey et al. (2021) 将机器学习领域的矩阵填补 (matrix completion) 思想应用到面板数据的因果效应识别问题, 由于反事实潜在结果在因果效应研究中是不可观测数据, 所以可以被视作为样本数据矩阵中缺失了相应的结果变量. 作者从理论上证明了双重差分方法与矩阵填补问题是等价的, 因此可以通过对矩阵缺失值的填补来完成对反事实潜在结果的预测估计, 进而得到因果效应估计值. Cicala (2017) 在研究美国电网从国家计划发电转变为市场自动调整发电这一举措所带来的收益时, 同样发现数据无法满足传统双重差分方法所要求的平行趋势假设, 于是直接利用随机森林算法对每个地区当没有市场介入时的发电量反事实结果进行了预测, 并在此基础上利用双重差分法估计了政策措施的因果效应.

目前还有一些针对双重差分法的改进方法与本文即将讨论的合成控制法所结合在一起, 相关内容我们将在下一部分进行详细展开和阐述.

3.4 合成控制法

使用合成控制法来估计因果效应的核心思想与双重差分法是一致的, 即考察控制组和处理组的结果变量在时间维度上的变动差异, 将其作为因果效应的估计值. 如前文所述, 除了随机控制实验, 一般需要在控制组中找到和处理组目标个体最相似的个体进行匹配后, 进行因果效应估计. 但是在双重差分法的实际应用中, 尤其是宏观层面的研究中, 往往存在样本量过小的情形, 这就会导致根本就不存在与处理组样本非常相似的控制组样本, 无法得到准确的因果效应估计值. 此时, 合成控制法将会通过样本构造的方法, 从现有样本中利用线性组合方法构造出与处理组样本高度相似的 “合成样本”, 随后对因果效应进行估计 (Abadie and Gardeazabal (2003), Abadie et al. (2010), Abadie et al. (2015), Sergio and Vitor (2018), Abadie (2021)), 其中 Abadie et al. (2010) 的方法在文献研究中被称作为 ADH 方法. 从数学表达式上看, 假设共有 J 个样本, 样本 1 是唯一受到政策措施干预的处理组样本, 剩余 $J-1$ 个样本均为控制组样本, 合成控制法即试图找出最优的非负权重向量 $\boldsymbol{w^*}=(w_2^*,\cdots,w_J^*)^{\mathrm{T}}$ 且满足 $\sum_{j=2}^{J} w_j^*=1$, 控制组样本在经过最优权重向量加权后形成合成样本 $\sum_{j=2}^{J} w_j^* Y_{jt}$, 该合成样本与样本 1 在受到干预措施前具有非常相似的性质, 对应的合成控制估计量 $\hat{\alpha}_{1t}$ 可以表示为:

$$\hat{\alpha}_{1t}=Y_{1t}-\sum_{j=2}^{J} w_j^* Y_{jt}, \tag{14}$$

其中, 下标 t 表示时间, Y 表示结果变量. 从式 (14) 看, 双重差分法的匹配过程可以看作是合成控制法的一个特例, 即控制组中取一个样本的权重为 1, 其余样本的权重为 0. 此外, 两者的联系还

在于双重差分法假设潜在干扰因子的影响是不随时间变化的, 但是合成控制法则允许通过重新加权来表示这些影响随时间的动态变化. 一般而言, 经过样本合成所构建的合成样本, 能够与处理组样本之间存在更好的平行趋势, 而且合成控制法不仅在小样本上具有较好的表现, 在面对大样本以及高维变量时, 合成控制法相比双重差分法往往具有更好的因果效应估计的性质 (Gobillon and Magnac (2016), Kinn (2018)).

从式 (14) 可以看到, 合成控制法的关键在于如何求得最优权重向量 $\boldsymbol{w}^* = (w_2^*, \cdots, w_J^*)^{\mathrm{T}}$, 一个理想的合成样本需要具有两方面特征: 第一个特征是具有良好的拟合特征, 即合成样本与处理组样本之间拥有较好的处理前平行趋势; 第二个特征是具有良好的样本外预测能力, 即合成样本能够准确地反映出处理组样本在受政策干预后的反事实潜在结果. 从这两方面看, 机器学习方法的优势与上述需求能够完全契合, 从拟合性能角度来看, 机器学习能够挖掘出变量间的非线性特征, 并且部分算法能够充分利用非数值变量的信息, 使得合成样本能够在政策干预前尽可能地与处理组样本保持一致, 平行趋势也就自然得到了满足. Xu (2017) 将合成控制法与交互固定效应模型相结合, 提出广义合成控制法 (generalized synthetic control method), 双重差分法依然是这个方法的特例, 广义合成控制法通过为处理组中每一个样本构建一个合成样本, 以此来达到估计个体因果效应的目的, 同时, 由于每一个处理样本都对应一个合成样本, 因此处理组样本与控制组样本之间的总体平行趋势能够得到更好地满足. Amjad et al. (2018) 的解决思路, 则是使用机器学习中的奇异值分解 (singular value decomposition, SVD) 算法去除数据中的噪音干扰后再使用合成控制法, 这样做的好处主要在于能够在随机缺失数据和协变量信息不足的情况下依然得到较好的合成样本拟合权重以及较稳定的因果效应估计结果, 作者通过引入贝叶斯估计框架验证了这一方法是足够稳健的.

在合成控制法中引入机器学习不仅仅是为了合成样本具有更好的政策干预前拟合效果, 其在政策干预后的样本外预测能力同样重要, 即要防止在训练合成样本时出现过拟合的情况. 正则化是最常用的用来克服这一问题的方法, 在合成控制法中, 以最小二乘法拟合为例, 通过加入惩罚项, 其目标函数就变为如下式所示:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{N} \left[\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} \|\beta_j\|_m \right] \right\}, \tag{15}$$

其中 λ 为外生给定的惩罚参数, 惩罚项 $\|\beta\|_m$ 随 m 取值的不同表示不同的范数, 例如 m 取 1 和 2 时分别代表 LASSO 回归和岭回归 (Belloni et al. (2014), Chernozhukov et al. (2015), Bloniarz et al. (2016)), 并且式 (15) 中的惩罚参数 λ 和训练模型都可以通过交叉验证法来确定, 以进一步提升模型的样本外预测能力 (Dube and Zipperer (2015)). 正则化除了能够防止过拟合外, 也能够起到筛选变量的作用, 这对于面临高维变量时可能遇到的维数灾难具有帮助. Doudchenko and Imbens (2016) 在传统的权重限制上, 允许权重为负且不再限制权重之和, 并结合正则化对合成控制法进行了改进, 使其能够在大样本的情况下表现更好. Ben-Michae et al. (2021) 发现当控制组样本无法实现政策干预前的完美拟合匹配时, 合成控制法的估计结果会出现偏误, 并由此提出了增强合成控制方法, 该方法使用岭回归作为结果模型来估计合成样本拟合过程中的偏误, 并基于该偏误大小来调整合成控制法的估计结果, 即偏误越大, 则增强合成控制法与传统合成控制法的差别越大. 作者通过模拟实验和 2012 年美国堪萨斯州减税政策对经济增长影响研究验证了该方

法的有效性. Cole et al. (2020) 使用两步法研究了新冠疫情封锁对武汉市四种空气污染物浓度的影响. 第一步中, 使用机器学习方法消除天气条件混杂因素对污染物浓度的影响, 第二步中, 使用了增强合成控制方法, 以未处于封锁状态的城市为控制组, 估计了封锁对空气污染物浓度的影响. Carvalho (2018) 则提出了一种称为人工反事实方法 (artificial counterfactual, ArCo) 来改进合成控制法, 除了使用了正则化方法外, ArCo 方法的优势在于能够同时检验政策干预对多个因变量的影响, 并且能够在政策干预时间未知的情况下进行因果效应估计, 大大提高了合成控制法的实际应用领域. 最后, 鉴于合成控制法与双重差分法之间的紧密关联性, Arkhangelsky et al. (2019) 将两种方法进行了结合, 提出了合成双重差分法 (synthetic difference in differences, SDID), 并从理论上证明了 SDID 比单独两种方法具有更好的统计性质.

4 研究结论与启示

4.1 研究结论

在传统计量经济方法学中引入机器学习方法已逐渐成为一个不可忽视重要发展趋势, 本文从目前实证研究中非常重要的因果推断框架作为切入点, 首先简要论述了因果推断两大主流分析框架各自的特征与内在关联, 在此基础上, 提出了机器学习方法可以从样本匹配与反事实预测两个方面对现有的因果效应识别研究提供有效的改进途径.

在因果推断框架中最重要的样本匹配方面, 机器学习能够利用精准预测的特性将 “非随机化” 的观测样本尽可能向 “随机化” 实验靠拢, 并且在面对大量协变量时展现出优异的变量选择能力. 此外, 无论是对于处理组样本还是控制组样本, 机器学习方法能够对其反事实结果进行一定程度的精准预测, 从而提升了因果效应识别能力. 此外, 机器学习方法能够提升样本的反事实预测的准确性, 从而提升因果效应识别效果.

本文接着从匹配法、断点回归法、双重差分法以及合成控制法四个具体的方法出发, 详细阐述了机器学习在提升因果效应识别结果方面的理论基础及其实际应用. 深入分析了机器学习是如何在未改变传统估计量的情况下, 通过复杂关系建模、交叉验证以及正则化等方法来使得模型在更 “理想化” 的情景下实现更精准的因果效应估计.

4.2 研究启示

虽然本文在因果效应分析框架之下, 着重讨论与分析了引入机器学习方法能够为提升因果效应识别效果带来的一定的优势, 但是机器学习方法也不是万能的, 其本身也存在着诸多缺陷, 使得在因果效应识别理论和应用中, 仍有大量的困难和未知等待着研究者们去攻克与探索 (汪寿阳等 (2019)). 首先, 机器学习方法中由于引入了大量的参数以及复杂的非线性变量关系, 使得过拟合问题较为严重, 这将会导致机器学习方法所引以为傲的样本外预测能力被大打折扣, 从而影响对因果效应估计的准确性. 因此, 现有机器学习方法中发展出了大量诸如正则化、样本分割等技术, 以期能够避免这一困境. 目前还没有绝对完美的方法能够解决这一难题, 因此在因果效应中使用机器学习方法时, 研究者们必须将这一点牢记于心. 其次, 机器学习方法虽然具有较好的预测能力, 但是在部分实践应用场景中, 其在微观层面的预测结果并不稳定, 较容易受到数据噪音、数据缺失、超参数取值等因素的影响, 在很多情况下, 这些问题需要长期从事机器学习领域的专业研

究者利用一定的专门技巧来加以缓解, 因此, 经济学以及其他社会科学领域的研究者在应用机器学习方法时, 需要对预测结果进行谨慎的考察, 必要时可以使用多种机器学习方法进行结果的交叉验证. 第三, 部分机器学习方法由于建模方法复杂, 导致其预测结果的可解释性较差, 因此很难弄清楚这些预测结果的直观机理, 这对于很多政策决策领域的应用来说, 就必须持相对谨慎的态度来看待相应的预测结果, 在很多情况下, 需要通过收集更多类型、更多区域和更长时间的样本来进行充分的稳健性检验, 才能够将因果效应的研究结果最终运用于决策支撑.

此外, 计量经济学的本质是统计学, 而机器学习的本质也是统计学, 由此也造就了两者的交叉应用. 除了在经济学的因果效应识别中应用机器学习外, 我们发现在其他众多研究领域, 尤其是计算机领域中的计算机视觉 (computer vision, CV) (Shen et al. (2018), Qi et al. (2020))、自然语言处理 (natural language processing, NLP) (Veitch et al. (2019), Zhang et al. (2020)) 等, 近年来都涌现了大量有关于因果推理和反事实预测相关的研究和应用, 未来的研究中需要进行更加深入的学科间交叉研究和融合, 来更好地提升因果推断研究的进一步提升和完善.

除了理论研究正经历着不断的突破, 各类基于机器学习方法的因果效应估计算法包也在不断涌现, 例如微软研究实验室 (Microsoft Research Lab) 开发了一套基于 Python 语言的异质性因果效应算法包 EconML, 包含了本文已经介绍过的成熟机器学习因果效应估计算法, 如双重机器学习法、因果森林法等 (Battocchi et al. (2019)), 而在另一个被业界和学界广泛使用的统计程序 R 语言中, 也能够找到双重机器学习法 (算法包名 DoubleML) 和因果森林 (算法包名 grf) 等算法包. 相信随着各类算法包的不断完善和提升, 能够极大地帮助应用经济学研究者们对更多的棘手现实问题进行深入剖析, 服务于我们经济社会决策的各类实际需求.

参　考　文　献

蔡宗武, (2021). 基于面板数据的处置效应估计的计量方法最新进展 [J]. 计量经济学报, 1(2): 233–249.
Cai Z W, (2021). Recent Developments in Estimating Treatment Effects for Panel Data[J]. China Journal of Econometrics, 1(2): 233–249.

陈浩, 冯艳, 魏文栋, (2020). 环境污染信息公开是否提升了城市技术创新?[J]. 环境经济研究, 5(3): 56–75.
Chen H, Feng Y, Wei W D, (2020). Does the Disclosure of Environmental Pollution Information Improve Urban Technological Innovation?[J]. Journal of Environmental Economics, 5(3): 56–75.

陈林, 伍海军, (2015). 国内双重差分法的研究现状与潜在问题 [J]. 数量经济技术经济研究, 32(7): 133–148.
Chen L, Wu H J, (2015). Research Status and Potential Problems of Differences-in-Differences Method in China[J]. The Journal of Quantitative and Technical Economics, 32(7): 133–148.

陈云松, 吴晓刚, 胡安宁, 贺光烨, 句国栋, (2020). 社会预测: 基于机器学习的研究新范式 [J]. 社会学研究, 35(3): 94–117.
Chen Y S, Wu X G, Hu A N, He G Y, Ju G D, (2020). Social Prediction: A New Research Paradigm Based on Machine[J]. Sociological Studies, 35(3): 94–117.

高玉娟, 白钰, 马跃, 史耀疆, (2018). 正负效应的先来后到: 父母外出对留守儿童学业表现的影响研究 [J]. 劳动经济研究, 6(3): 97–113.
Gao Y J, Bai Y, Ma Y, Shi Y J, (2018). Arrival Order for Positive and Negative Effects of Parental Migration on the Academic Performance of Left-behind Children in Rural China[J]. Studies in Labor Economics, 6(3): 97–113.

高正斌, 张开志, 倪志良, (2020). 减税能促进企业创新吗?—— 基于所得税分享改革的准自然实验 [J]. 财政研究, 4(8): 86–100.
Gao Z B, Zhang K Z, Ni Z L, (2020). Can Tax Reduction Promote Firm Innovation? A Quasi-natural Experiment Based on the Income Tax Sharing Reform[J]. Public Finance Research, 4(8): 86–100.

洪永淼, (2021). 理解现代计量经济学 [J]. 计量经济学报, 1(2): 266–284.
Hong Y M, (2021). Understanding Modern Econometrics[J]. China Journal of Econometrics, 1(2): 266–284.

洪永淼, 汪寿阳, (2020). 数学、模型与经济思想 [J]. 管理世界, 36(10): 15–27.
Hong Y M, Wang S Y, (2020). Mathematical, Model and Economic Thought[J]. Management World, 36(10): 15–27.

胡安宁, 吴晓刚, 陈云松, (2021). 处理效应异质性分析 —— 机器学习方法带来的机遇与挑战 [J]. 社会学研究, 36(1): 91–114.
Hu A N, Wu X G, Chen Y S, (2021). Analysis of Heterogeneous Treatment Effect: New Opportunities and Challenges with Machine Learning Techniques[J]. Sociological Studies, 36(1): 91–114.

胡咏梅, 唐一鹏, (2018). 公共政策或项目的因果效应评估方法及其应用 [J]. 华中师范大学学报 (人文社会科学版), 57(3): 168–181.
Hu Y M, Tang Y P, (2018). The Causality Evaluation Method on Public Policies and Programs and Its Application[J]. Journal of Central China Normal University (Humanities and Social Sciences), 57(3): 168–181.

黄乃静, 于明哲, (2018). 机器学习对经济学研究的影响研究进展 [J]. 经济学动态, 4(7): 115–129.
Huang N J, Yu M Z, (2018). Research Progress on the Influence of Machine Learning on Economic Research[J]. Economic Perspectives, 4(7): 115–129.

毛其淋, 许家云, (2016). 政府补贴、异质性与企业风险承担 [J]. 经济学 (季刊), 15(4): 1533–1562.
Mao Q L, Xu J Y, (2016). Government Subsidy, Heterogeneity and Corporate Risk-taking[J]. China Economic Quarterly, 15(4): 1533–1562.

齐绍洲, 林屾, 崔静波, (2018). 环境权益交易市场能否诱发绿色创新? —— 基于我国上市公司绿色专利数据的证据 [J]. 经济研究, 53(12): 129–143.
Qi S Z, Lin S, Cui J B, (2018). Do Environmental Rights Trading Schemes Induce Green Innovation? Evidence from Listed Firms in China[J]. Economic Research Journal, 53(12): 129–143.

钱浩祺, (2020). 环境大数据应用的最新进展与趋势 [J]. 环境经济研究, 5(4): 152–180.
Qian H Q, (2020). The Perspective of Application of Environmental Big Data[J]. Journal of Environmental Economics, 5(4): 152–180.

饶茜, 杨雨虹, 郭世俊, 向丹, (2020). 增值税进项加计抵减对企业价值的影响 [J]. 财政研究, 4(10): 102–114.
Rao Q, Yang Y H, Guo S J, Xiang D, (2020). The Impact of the Input VAT Additional Deduction on Firm Value[J]. Public Finance Research, 4(10): 102–114.

唐为, 王媛, (2015). 行政区划调整与人口城市化: 来自撤县设区的经验证据 [J]. 经济研究, 50(9): 72–85.
Tang W, Wang Y, (2015). Administrative Boundary Adjustment and Urbanization of Population: Evidence from City-county Merger in China[J]. Economic Research Journal, 50(9): 72–85.

汪寿阳, 洪永淼, 霍红, 方颖, 陈海强, (2019). 大数据时代下计量经济学若干重要发展方向 [J]. 中国科学基金, 33(4): 386–393.
Wang S Y, Hong Y M, Huo H, Fang Y, Chen H Q, (2019). Several Influential Research Directions of Econometrics in Big Data Era[J]. Bulletin of National Natural Science Foundation of China, 33(4): 386–393.

王班班, 莫琼辉, 钱浩祺, (2020). 地方环境政策创新的扩散模式与实施效果 —— 基于河长制政策扩散的微观实证 [J]. 中国工业经济, (8): 99–117.
Wang B B, Mo Q H, Qian H Q, (2020). The Diffusion Models and Effects of the Local Environmental Policy Innovation — A Micro-econometric Evidence from the Diffusion of River Chief Policy[J]. China Industrial Economics, (8): 99–117.

王芳, 王宣艺, 陈硕, (2020). 经济学研究中的机器学习: 回顾与展望 [J]. 数量经济技术经济研究, 37(4): 146–164.
Wang F, Wang X Y, Chen S, (2020). Machine Learning Economics Research: Review and Prospective[J]. The Journal of Quantitative and Technical Economics, 37(4): 146–164.

王玺, 刘萌, (2020). 研发费用加计扣除政策对企业绩效的影响研究 —— 基于我国上市公司的实证分析 [J]. 财政研究, 4(11): 101–114.
Wang X, Liu M, (2020). Research on the Impact of R&D Expense Additional Deduction Policy on Firm Performance — An Empirical Analysis Based on Listed Companies in China[J]. Public Finance Research, 4(11): 101–114.

王妍, 白钰, 刘承芳, 史耀疆, (2019). 父母返乡对留守儿童学业表现的影响 —— 基于西北贫困农村 130 所学校的研究 [J]. 劳动经济研究, 7(1): 78–98.
Wang Y, Bai Y, Liu C F, Shi Y J, (2019). The Effect of Parental Returning on Academic Performance of Left-behind Children: Evidence from 130 Schools in Rural Northwest China[J]. Studies in Labor Economics, 7(1): 78–98.

王乙杰, 孙文凯, (2020). 户口改变对流动人口家庭消费的影响 —— 来自微观追踪数据的证据 [J]. 劳动经济研究, 8(2): 68–100.
Wang Y J, Sun W K, (2020). Effect of Hukou Change on Migrant Household Consumption: Evidence from Panel Data[J]. Studies in Labor Economics, 8(2): 68–100.

吴力波, 任飞州, 徐少丹, (2020). 新冠肺炎疫情防控一级响应对城市空气污染物减排的影响 [J]. 环境经济研究, 5(3): 1–20.
Wu L B, Ren F Z, Xu S D, (2020). Impact of First-level Response to COVID-19 on the Reduction of Urban Air Pollutants in China[J]. Journal of Environmental Economics, 5(3): 1–20.

吴力波, 任飞州, 徐少丹, (2021). 环境规制执行对企业绿色创新的影响 [J]. 中国人口 · 资源与环境, 31(1): 90–99.
Wu L B, Ren F Z, Xu S D, (2021). Influence of Environmental Regulation Enforcement on Enterprises' Green Innovation[J]. China Population Resources and Environment, 31(1): 90–99.

萧政, (2021). 大数据时代关于预测的几点思考 [J]. 计量经济学报, 1(1): 1–16.
Hsiao C, (2021). Some Thoughts on Prediction in the Presence of Big Data[J]. China Journal of Econometrics, 1(1): 1–16.

徐佳, 崔静波, (2020). 低碳城市和企业绿色技术创新 [J]. 中国工业经济, (12): 178–196.
Xu J, Cui J B, (2020). Low-carbon Cities and Firms' Green Technological Innovation[J]. China Industrial Economics, (12): 178–196.

徐志伟, 刘晨诗, (2020). 环境规制的 "灰边" 效应 [J]. 财贸经济, 41(1): 145–160.
Xu Z W, Liu C S, (2020). Grey-edge Effect of Environmental Regulation[J]. Finance and Trade Economics, 41(1): 145–160.

俞秀梅, 王敏, (2020). 阶梯电价改革对我国居民电力消费的影响 —— 基于固定电表月度面板数据的研究 [J]. 经济学 (季刊), 19(2): 731–756.
Yu X M, Wang M, (2020). The Impact of Tiered Pricing Reform on China's Residential Electricity Consumption[J]. China Economic Quarterly, 19(2): 731–756.

张俊, (2017). 高铁建设与县域经济发展 —— 基于卫星灯光数据的研究 [J]. 经济学 (季刊), 16(4): 1533–1562.
Zhang J, (2017). High-speed Rail Construction and County Economic Development: The Research of Satellite Light Data[J]. China Economic Quarterly, 16(4): 1533–1562.

张琦, 郑瑶, 孔东民, (2019). 地区环境治理压力、高管经历与企业环保投资 —— 一项基于《环境空气质量标准 (2012)》的准自然实验 [J]. 经济研究, 54(6): 183–198.
Zhang Q, Zheng Y, Kong D M, (2019). Local Environmental Governance Pressure, Executive's Working Experience and Enterprise Investment in Environmental Protection: A Quasi-natural Experiment — Based on China's "Ambient Air Quality Standards 2012"[J]. Economic Research Journal, 54(6): 183–198.

朱平芳, 邸俊鹏, (2017). 无条件分位数处理效应方法及其应用 [J]. 数量经济技术经济研究, 34(2): 139–155.
Zhu P F, Di J P, (2017). Unconditional Quantile Treatment Effects and Application in Policy Evaluation[J]. The Journal of Quantitative and Technical Economics, 34(2): 139–155.

Abadie A, (2005). Semiparametric Difference-in-Differences Estimators[J]. The Review of Economic Studies, 72(1): 1–19.

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

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

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

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

Abadie A, Imbens G W, (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects[J]. Econometrica, 74(1): 235–267.

Amjad M, Shah D, Shen D, (2018). Robust Synthetic Control[J]. The Journal of Machine Learning Research, 19(1): 802–852.

Anastasopoulos J, (2019). Principled Estimation of Regression Discontinuity Designs with Covariates: A Machine Learning Approach[J]. https://arxiv.org/abs/1910.06381.

Angrist J D, Imbens G W, Rubin D B, (1996). Identification of Causal Effects Using Instrumental Variables[J]. Journal of the American Statistical Association, 91(434): 444–455.

Arkhangelsky D, Athey S, Hirshberg D A, Imbens G W, Wager S, (2019). Synthetic Difference in Differences[R]. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w25532/w25532.pdf.

Athey S, (2017). Beyond Prediction: Using Big Data for Policy Problems[J]. Science, 355(6324): 483–485.

Athey S, (2019). The Impact of Machine Learning on Economics[M]// Agrawal A, Gans J, Goldfarb A. The Economics of Artificial Intelligence: An Agenda. Chicago: University of Chicago Press: 507–547.

Athey S, Bayati M, Doudchenko N, Imbens G, Khosravi K, (2021). Matrix Completion Methods for Causal Panel Data Models[J]. Journal of the American Statistical Association: 1–15.

Athey S, Bayati M, Imbens G, Qu Z, (2019). Ensemble Methods for Causal Effects in Panel Data Settings[J]. AEA Papers and Proceedings, 109: 65–70.

Athey S, Imbens G W, (2006). Identification and Inference in Nonlinear Difference-in-Differences Models[J]. Econometrica, 74(2): 431–497.

Athey S, Imbens G W, (2016). Recursive Partitioning for Heterogeneous Causal Effects[J]. Proceedings of the National Academy of Sciences, 113(27): 7353–7360.

Athey S, Imbens G W, (2017). The State of Applied Econometrics: Causality and Policy Evaluation[J]. Journal of Economic Perspectives, 31(2): 3–32.

Athey S, Imbens G W, (2019). Machine Learning Methods Economists Should Know About[J]. https://arxiv.org/abs/1903.10075.

Athey S, Tibshirani J, Wager S, (2019). Generalized Random Forests[J]. The Annals of Statistics, 47(2): 1148–1178.

Bargagli-Stoffi F J, De-Witte K, Gnecco G, (2019). Heterogeneous Causal Effects with Imperfect Compliance: A Novel Bayesian Machine Learning Approach[J]. https://arxiv.org/abs/1905.12707.

Baumeister R F, Bratslavsky E, Finkenauer C, Vohs K D, (2001). Bad is Stronger than Good[J]. Review of General Psychology, 5(4): 323–370.

Belloni A, Chen D, Chernozhukov V, Hansen C, (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain[J]. Econometrica, 80(6): 2369–2429.

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C, (2017). Program Evaluation and Causal Inference with High-dimensional Data[J]. Econometrica, 85(1): 233–298.

Belloni A, Chernozhukov V, Hansen C, (2014). High-dimensional Methods and Inference on Structural and Treatment Effects[J]. Journal of Economic Perspectives, 28(2): 29–50.

Ben-Michael E, Feller A, Rothstein J, (2021). The Augmented Synthetic Control Method[J]. Journal of the American Statistical Association: 1–34.

Bertanha M, Imbens G W, (2020). External Validity in Fuzzy Regression Discontinuity Designs[J]. Journal of Business and Economic Statistics, 38(3): 593–612.

Bloniarz A, Liu H, Zhang C H, Sekhon J S, Yu B, (2016). Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments[J]. Proceedings of the National Academy of Sciences, 113(27): 7383–7390.

Branson Z, Rischard M, Bornn L, Miratrix L W, (2019). A Nonparametric Bayesian Methodology for Regression Discontinuity Designs[J]. Journal of Statistical Planning and Inference, 202: 14–30.

Cannas M, Arpino B, (2019). A Comparison of Machine Learning Algorithms and Covariate Balance Measures for Propensity Score Matching and Weighting[J]. Biometrical Journal, 61(4): 1049–1072.

Carrasco M, (2012). A Regularization Approach to the Many Instruments Problem[J]. Journal of Econometrics, 170(2): 383–398.

Carvalho C, Masini R, Medeiros M C, (2018). ArCo: An Artificial Counterfactual Approach for High-dimensional Panel Time-series Data[J]. Journal of Econometrics, 207(2): 352–380.

Chang N C, (2020). Double/debiased Machine Learning for Difference-in-Differences Models[J]. The Econometrics Journal, 23(2): 177–191.

Chen Z, Liu Z, Suárez Serrato J C, Xu D Y, (2021). Notching R&D Investment with Corporate Income Tax Cuts in China[J]. American Economic Review, 111(7): 2065–2100.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al. (2017). Double/debiased/neyman Machine Learning of Treatment Effects[J]. American Economic Review, 107(5): 261–265.

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, et al. (2018). Double/debiased Machine Learning for Treatment and Structural Parameters[J]. The Econometrics Journal, 21(1): C1–C68.

Chernozhukov V, Hansen C, Spindler M, (2015). Valid Post-selection and Post-regularization Inference: An Elementary, General Approach[J]. Annual Review of Economics, 7(1): 649–688.

Chipman H A, George E I, McCulloch R E, (2010). BART: Bayesian Additive Regression Trees[J]. The Annals of Applied Statistics, 4(1): 266–298.

Cicala S, (2017). Imperfect Markets Versus Imperfect Regulation in US Electricity Generation[R]. National Bureau of Economic Research. http://www.nber.org/papers/w23053.

Clarke P S, Windmeijer F, (2012). Instrumental Variable Estimators for Binary Outcomes[J]. Journal of the American Statistical Association, 107(500): 1638–1652.

Cole M A, Elliott R J R, Liu B, (2020). The Impact of the Wuhan COVID-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach[J]. Environmental and Resource Economics, 76(4): 553–580.

Davis J M V, Heller S B, (2020). Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs[J]. Review of Economics and Statistics, 102(4): 664–677.

Diamond A, Sekhon J S, (2013). Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies[J]. Review of Economics and Statistics, 95(3): 932–945.

Doksum K, (1974). Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-sample Case[J]. The Annals of Statistics: 267–277.

Doudchenko N, Imbens G W, (2016). Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis[R]. National Bureau of Economic Research. https://www.nber.org/papers/w22791.

Dube A, Jacobs J, Naidu S, Suri S, (2020). Monopsony in Online Labor Markets[J]. American Economic Review: Insights, 2(1): 33–46.

Dube A, Zipperer B, (2015). Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies[J]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2589786.

Gelman A, Imbens G, (2019). Why High-order Polynomials Should Not Be Used in Regression Discontinuity Designs[J]. Journal of Business and Economic Statistics, 37(3): 447–456.

Gobillon L, Magnac T, (2016). Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls[J]. Review of Economics and Statistics, 98(3): 535–551.

Green D P, Kern H L, (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees[J]. Public Opinion Quarterly, 76(3): 491–511.

Handel B, Kolstad J, (2017). Wearable Technologies and Health Behaviors: New Data and New Methods to Understand Population Health[J]. American Economic Review, 107(5): 481–485.

Hartford J, Lewis G, Leyton-Brown K, Taddy M, (2017). Deep IV: A Flexible Approach for Counterfactual Prediction[C]// International Conference on Machine Learning. PMLR: 1414–1423.

Hausman C, Rapson D S, (2018). Regression Discontinuity in Time: Considerations for Empirical Applications[J]. Annual Review of Resource Economics, 10: 533–552.

Heckman J J, (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective[J]. The Quarterly Journal of Economics, 115(1): 45–97.

Heckman J J, García J L, (2017). Social Policy: Targeting Programmes Effectively[J]. Nature Human Behaviour, 1(1): 1–2.

Heckman J J, Ichimura H, Todd P E, (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme[J]. The Review of Economic Studies, 64(4): 605–654.

Heckman J J, Ichimura H, Todd P, (1998). Matching as an Econometric Evaluation Estimator[J]. The Review of Economic Studies, 65(2): 261–294.

Hirano K, Imbens G W, Ridder G, (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score[J]. Econometrica, 71(4): 1161–1189.

Imai K, Ratkovic M, (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation[J]. The Annals of Applied Statistics, 7(1): 443–470.

Imbens G W, (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics[J]. Journal of Economic Literature, 58(4): 1129–1179.

Imbens G W, Kalyanaraman K, (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator[J]. The Review of Economic Studies, 79(3): 933–959.

Imbens G W, Wager S, (2019). Optimized Regression Discontinuity Designs[J]. Review of Economics and Statistics, 101(2): 264–278.

Jepsen C, Mueser P, Troske K, (2017). Second Chance for High School Dropouts? A Regression Discontinuity Analysis of Postsecondary Educational Returns to the GED[J]. Journal of Labor Economics, 35(S1): S273–S304.

Jesson A, Mindermann S, Shalit U, Gal Y, (2020). Identifying Causal-effect Inference Failure with Uncertainty-aware Models[J]. Advances in Neural Information Processing Systems, 33: 11637–11649.

Jia J, Liang X, Ma G, (2021). Political Hierarchy and Regional Economic Development: Evidence from a Spatial Discontinuity in China[J]. Journal of Public Economics, 194: 104352.

Kallus N, Mao X, Uehara M, (2019). Localized Debiased Machine Learning: Efficient Inference on Quantile Treatment Effects and Beyond[J]. https://arxiv.org/abs/1912.12945v3.

Karimi A H, Von Kügelgen J, Schölkopf B, Valera I, (2020). Algorithmic Recourse Under Imperfect Causal Knowledge: A Probabilistic Approach[J]. https://arxiv.org/abs/2006.06831.

Kinn D, (2018). Synthetic Control Methods and Big Data[J]. https://arxiv.org/abs/1803.00096.

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z, (2015). Prediction Policy Problems[J]. American Economic Review, 105(5): 491–95.

Knittel C R, Stolper S, (2021). Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use[J]. AEA Papers and Proceedings, 111: 440–444.

Lazer D, Pentland A, Adamic L, Aral S, Barabasi A L, et al. (2009). Computational Social Science[J]. Science (New York), 323(5915): 721–723.

Lee B K, Lessler J, Stuart E A, (2010). Improving Propensity Score Weighting Using Machine Learning[J]. Statistics in Medicine, 29(3): 337–346.

Lee M, (2016). Matching, Regression Discontinuity, Difference in Differences, and Beyond[M]. Oxford: Oxford University Press.

Lehmann E L, D'Abrera H J, (1975). Nonparametrics: Statistical Methods Based on Ranks[M]. San Francisco: Holden-day.

Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, et al. (2017). Causal Effect Inference with Deep Latent-variable Models[J]. https://arxiv.org/abs/1705.08821.

Löwe S, Madras D, Zemel R, Welling M, (2020). Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-series Data[J]. https://arxiv.org/abs/2006.10833.

Lu C, Nie X, Wager S, (2019). Robust Nonparametric Difference-in-Differences Estimation[J]. https://arxiv.org/abs/1905.11622.

Manski C F, Pepper J V, (2018). How Do Right-to-carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-variation Assumptions[J]. Review of Economics and Statistics, 100(2): 232–244.

McGue M, Osler M, Christensen K, (2010). Causal Inference and Observational Research: The Utility of Twins[J]. Perspectives on Psychological Science, 5(5): 546–556.

Meinshausen N, Ridgeway G, (2006). Quantile Regression Forests[J]. Journal of Machine Learning Research, 7(35): 983–999.

Microsoft Research, (2019). EconML: A Python Package for ML — Based Heterogeneous Treatment Effects Estimation[J]. https://github.com/microsoft/EconML.

Mullainathan S, Spiess J, (2017). Machine Learning: An Applied Econometric Approach[J]. Journal of Economic Perspectives, 31(2): 87–106.

Narayanan S, Kalyanam K, (2020). Behavioral Targeting, Machine Learning and Regression Discontinuity Designs[J]. https://gsb-faculty.stanford.edu/sridhar-narayanan/files/2021/01/Working-Paper_RDML_Dec2020.pdf.

Neyman J S, (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Translated and Edited by Dabrowska D M and Speed T P, (1990). Statistical Science, (5): 465–480)[J]. Annals of Agricultural Sciences, 10: 1–51.

Ning Y, Peng S, Tao J, (2020). Doubly Robust Semiparametric Difference-in-Differences Estimators with High-dimensional Data[J]. https://arxiv.org/abs/2009.03151.

Pearl J, (1995). Causal Diagrams for Empirical Research[J]. Biometrika, 82(4): 669–688.

Pearl J, (2009). Causal Inference in Statistics: An Overview[J]. Statistics Surveys, 3: 96–146.

Pirracchio R, Petersen M L, Van Der Laan M. (2015). Improving Propensity Score Estimators' Robustness to Model Misspecification Using Super Learner[J]. American Journal of Epidemiology, 181(2): 108–119.

Qi J, Niu Y, Huang J, Zhang H, (2020). Two Causal Principles for Improving Visual Dialog[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 10860–10869.

Rosenbaum P R, Rubin D B, (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects[J]. Biometrika, 70(1): 41–55.

Rubin D B, (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies[J]. Journal of Educational Psychology, 66(5): 688–701.

Rubin D B, (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions[J]. Journal of the American Statistical Association, 100(469): 322–331.

Schwab P, Linhardt L, Karlen W, (2018). Perfect Match: A Simple Method for Learning Representations for Counterfactual Inference with Neural Networks[J]. https://arxiv.org/abs/1810.00656.

Setoguchi S, Schneeweiss S, Brookhart M A, Glynn R J, Cook E F, (2008). Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study[J]. Pharmacoepidemiology and Drug Safety, 17(6): 546–555.

Shalit U, Johansson F D, Sontag D, (2017). Estimating Individual Treatment Effect: Generalization Bounds and Algorithms[C]// International Conference on Machine Learning. PMLR: 3076–3085.

Shen Z, Cui P, Kuang K, Li B, Chen P, (2018). Causally Regularized Learning with Agnostic Data Selection Bias[C]// Proceedings of the 26th ACM International Conference on Multimedia: 411–419.

Singh A, Hosanagar K, Gandhi A, (2020). Machine Learning Instrument Variables for Causal Inference[C]// Proceedings of the 21st ACM Conference on Economics and Computation: 835–836.

Singh R, Sun L, (2019). De-biased Machine Learning for Compliers[C]// Advances in Neural Information Processing Systems Workshop on Causal Machine Learning.

Steiner P M, Cook D, (2013). Matching and Propensity Scores[J]. The Oxford Handbook of Quantitative Methods, 1: 237–259.

Su X, Tsai C L, Wang H, Nickerson D M, Li B, (2009). Subgroup Analysis via Recursive Partitioning[J]. Journal of Machine Learning Research, 10(2): 114–158.

Swanson S A, Hernán M A, Miller M, Robins J M, Richardson T S, (2018). Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes[J]. Journal of the American Statistical Association, 113(522): 933–947.

Varian H R, (2014). Big Data: New Tricks for Econometrics[J]. Journal of Economic Perspectives, 28(2): 3–28.

Varian H R, (2016). Causal Inference in Economics and Marketing[J]. Proceedings of the National Academy of Sciences, 113(27): 7310–7315.

Veitch V, Sridhar D, Blei D M, (2019). Using Text Embeddings for Causal Inference[J]. https://openreview.net/forum?id=YiMfKTUm3jN.

Wager S, Athey S, (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests[J]. Journal of the American Statistical Association, 113(523): 1228–1242.

Wright S, (1921). Correlation and Causation[J]. https://naldc.nal.usda.gov/catalog/IND43966364.

Xu Y, (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models[J]. Political Analysis, 25(1): 57–76.

Yoon J, Jordon J, Van Der Schaar M, (2018). GANITE: Estimation of Individualized Treatment Effects Using Generative Adversarial Nets[C]// International Conference on Learning Representations.

Zhang D, Zhang H, Tang J, Hua X, Sun Q, (2020). Causal Intervention for Weakly-supervised Semantic Segmentation[J]. https://arxiv.org/abs/2009.12547.

Zhu J, Fan Y, Deng X, Xue L, (2019). Low-carbon Innovation Induced by Emissions Trading in China[J]. Nature Communications, 10(1): 1–8.