

政治学研究中的因果推断策略及其应用

顾燕峰*

一、导 论

科学是由方法论驱动的。政治科学在长期的发展中形成了三大方法论传统,即描述性传统、解释性-预测性传统和实践性传统。^①受培根经验哲学和牛顿-伽利略自然科学方法的影响,传统社会科学在两百年前逐步形成了科学主义研究范式,以期在对社会现象进行描述的基础上,致力于追求对社会现象的因果解释,最终实现社会预测功能。政治学的解释性-预测性传统正是在这一发展过程中逐步形成的。然而,由于政治现象的复杂性,政治学离实现预测功能还相去甚远,但政治学在通过引入行之有效的因果推断方法改善识别效果方面已有长期的努力。

社会科学因果推断的基础是鲁宾因果模型(也称潜在因果模型)。该模型包括两个必不可少的部分。第一部分是通过潜在因果定义因果效应。稳定单位处理值假设(SUTVA)是这个部分的重要假设,它确保任意单位的潜在结果与其他单位是否接受处理不相关,并且对每个单位,每个处理的水平在形式上没有差异。第二部分是随机分配机制。随机分配机制要求研究单位被随机分配进处理组和对照组,这可以确保处理组和对照组在统计学意义上拥有相同的特征。这构成了因果推断的基础。

* 顾燕峰,博士,复旦大学社会科学高等研究院专职研究员、副教授。

① Johann Lauer, "Methodology and Political Science: The Discipline Needs Three Fundamentally Different Methodological Traditions", *SN Social Sciences*, Vol.1, no.43(2021), pp.1—29.

随着社会科学逐步强调探求社会现象间的因果关系,实验法逐步被引入社会科学(尤其是经济学),并成为社会科学中最常用的检验因果关系的手段之一。随机实验是因果推断的黄金标准。^①政治学利用实验(尤其是随机实验)进行因果推断已有很长的历史。然而从20世纪20年代到70年代,实验在政治学中仍然属于边缘性方法。但这期间的积累蕴含了80年代突破的种子。首先,政治学家越来越追求对政治现象的因果关系解释,定量研究已成为政治学研究的主流方法。在20世纪70年代当统计学家正努力解决通过观察数据进行因果推断的一系列关键技术问题时,许多社会科学家在实证研究中已认识到定量分析的这一缺陷。精心设计的实验在识别因果关系方面往往是有效的,因此,在统计学和计量经济学发展出一套成熟的因果推断技术并引入政治研究前,实验注定将被越来越多的政治学家采用。其次,实验在其他社会科学领域不断增长的应用促进了其在政治学领域的应用。经济学领域许多重要的实验成果在20世纪50年代就已经出现,1962年实验经济学的诞生,70年代对实验经济学研究基金支持的增加,最终促成了80年代实验经济学爆发性的增长。这无疑将促进政治学研究中实验的应用。最终,政治学中的实验在90年代开始出现某种突破。不过,至少到2002年,实验仍未成为政治学中的主要研究方法。但经过21世纪初期的持续增长后,实验在政治科学研究中的应用终于在2010年以后出现了爆发式的增长。

政治学的许多问题无法通过实验来回答。因此,政治学家必须尝试用观察数据来解决因果推断问题。自然实验是基于实验思想,利用观察数据进行实验推断的典型方法。由于自然实验使用来自真实世界中的观察数据,避免了研究人员对实验环境的干预,从而可以具有较好的外部有效性。因此,自然实验的独特优势吸引了许多社会科学家,尤其是来自经济学领域的学者。而这最终也促成了政治学家利用该方法进行因果推断。但自然实验也有其缺点,即实验者不能挑选实验干预的具体内容,实验者也不能挑选干预的地点和时间。实际上,这可能导致自然实验违反鲁宾因果模型中的单位稳定处理值假设和随机假

^① Donald B. Rubin, "For Objective Causal Inference, Design Trumps Analysis", *The Annals of Applied Statistics*, Vol.2, no.3(2008), pp.808—840.

设。为了确保自然实验符合随机分配机制,学者采用匹配法来确保处理组和对照组的可比性。这项技术也被越来越多的政治学家所采用。

基于实验思想发展出来的断点回归、双重差分模型和满足鲁宾因果模型框架的工具变量方法是社会科学进行因果推断的重要工具。这些技术被引入政治学研究的历史虽有差异,但一个共同的特征是随着政治学对因果推断的关注,这些技术在政治学中的应用在 21 世纪,尤其是 2010 年以后呈现爆发式增长。

机器学习是在统计学和大数据基础上发展起来的一项技术。机器学习的一项特征是强调对被解释变量的拟合值,而非回归系数的无偏一致性。这项特征被用于因果推断的主要优势在于因果推断的过程中需要使用预测程序时,机器学习有助于提升预测的准确性,从而帮助研究者改进因果推断的准确性。机器学习在因果推断中的一个典型的应用是,在工具变量估计第一阶段回归中,利用机器学习方法以改善预测质量;另一个典型应用是用于基于观测数据的匹配法。这些方法在统计学和计量经济学中已经得到长足的发展,但在政治学中几乎未有涉及。

二、因果推断的基础:鲁宾因果模型

假定 Y_i^1 接受处理的单位(unit)的潜在结果, i 指单位 i , 1 表示接受处理;反之 Y_i^0 是指没有接受处理的单位的潜在结果,0 表示没有接受处理。因此,对 i 来说,处理效应是:

$$\delta_i = Y_i^1 - Y_i^0 \quad (1)$$

如果有许多单位,则其平均处理效应(average treatment effect, ATE)为:

$$ATE = E(\delta_i) = E(Y_i^1) - E(Y_i^0) \quad (2)$$

但问题是我们无法同时观测 i 的潜在结果 Y_i^1 和 Y_i^0 。显然, δ_i 不能直接测量,因此 ATE 也无法直接计算。

现在假定有 N 个单位($N > 1$),我们将之分为处理组和对照组,并对处理组施加实验处理,并将处理组和对照组的潜在结果表示为:

$$Y_i = \begin{cases} Y_i^1 & \text{if } T_i = 1 \\ Y_i^0 & \text{if } T_i = 0 \end{cases} \quad (3)$$

$T_i = 1$ 表示处理组的单位, $T_i = 0$ 表示对照组的单位。由此, 我们可以计算平均处理效应:

$$ATE = E(\delta_i) = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 0) \quad (4)$$

δ_i 的值可能因单位而异。因此实验处理的平均处理效应在总体与处理组 (treatment group) 中可能不一致。处理组的平均处理效应 (average treatment effect on treated, ATT), 更正式的表述应为:

$$ATT = E(\delta_i | T_i = 1) = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 1) \quad (5)$$

显然, ATT 才是我们期望计算的平均处理效应, 然而事实上却不可测; 而我们却可通过处理组和对照组的区分来计算 ATE。

利用可观测的实验数据计算的 ATE, 和理论上 ATT 之间的差异为:

$$D = E(Y_i^0 | T_i = 1) - E(Y_i^0 | T_i = 0) \quad (6)$$

从公式(6)可知, 如果能确保

$$E(Y_i^0 | T_i = 1) = E(Y_i^0 | T_i = 0) \quad (7)$$

我们就能确保 ATT 等于 ATE。显然, 我们无法直接验证公式(7)。

鲁宾因果模型有两个必要的部分。一是通过潜在结果定义了因果效应。稳定单位处理值假设 (SUTVA) 是这个部分的重要假设。该假设要求任意单位的潜在结果与其他单位是否接受处理不相关; 且对每个单位, 每个处理的水平在形式上没有差异。^① SUTVA 是一个排除性约束 (exclusion restriction), 它排除了单位间互相扰动和处理变异的可能, 从而确保每个单位只有两个潜在结果 (y_i^0 和 y_i^1)。^② 二是随机分配机制。随机分配机制要求研究单位被随机分配进处

^① Guido W. Imbens and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press, 2015.

^② SUTVA 并不要求所有单位所有处理的每个水平的形式都是相同的, 但必须明确定义每个处理水平的潜在结果。Guido W. Imbens and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press, 2015.

理组和对照组,这可以确保处理组和对照组在统计学意义上拥有相同的特征。满足这两个条件即可确保公式(7)成立,即可以估计 ATT。

这实质上构成现代社会科学因果推断的理论基础。尽管从社会科学因果探索历史上看,最初从实验数据和观察数据进行因果推断探索几乎是齐头并进,并分别建立起各自的方法,但最终都归结到鲁宾因果模型。一个明显的例子是工具变量,正如本文下面将讨论到的,工具变量最初应用于观察数据检验因果关系,通常是从联立方程的语言进行讨论的,但最终学者发现其原来可以用鲁宾因果模型解释。^①

三、实验法及其在政治科学研究中的应用

(一) 实验法在政治学中的引入

政治现象纷繁复杂,单纯依靠观察难以触及政治现象的本质。因此,政治学家开始尝试利用定量方法来探究其背后抽象的逻辑。早在 20 世纪 20 年代《美国政治学评论》就已发表了第一篇使用定量分析的论文。^②早期的定量政治学家主要依赖直接经验观察,这无疑在方法论上是有缺陷的。因此,20 世纪 20 年代方法论学者开始转向系统的数据收集。^③这一方面导致政治学研究议题的扩展,也促成了多元统计技术(如相关分析、回归分析和因子分析)在政治科学中的应用。^④但统计技术在政治学中并未因此进入快速扩张期。根据加里·金(Gary King)的研究,《美国政治学评论》上定量分析论文的数量在 1960 年之前一直保持在一个非常低的比例,直至 60 年代定量论文的比例才从约 1/4 增长至

① Joshua D. Angrist, Guido W. Imbens & Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, Vol.91(1996), pp.434, 444—455.

② Harold F. Gosnell, "An Experiment in the Stimulation of Voting", *American Political Science Review*, Vol.20(1926), pp.869—874.

③ Gary King, "On Political Methodology", *Political Analysis*, Vol.2(1990), pp.1—29.

④ Gary King, "On Political Methodology", *Political Analysis*, Vol.2(1990), pp.1—29. David J. Gow, "Quantification and Statistics in the Early Years of American Political Science, 1880—1922", *Political Methodology*, Vol.11, no.1/2(1985), pp.1—18.

一半。^①至此,定量分析在政治学中已成为主流分析技术。整个 70 年代,定量论文的比例在该杂志亦未能形成突破。

尽管至少在 20 世纪 60 年代后期,定量分析已成为政治学研究主流方法。虽然此时统计学领域已经建立以实验数据为基础的因果推断的理论,却未能完成利用观察数据进行因果推断的理论准备。反事实框架和潜在结果模型是因果推断的理论基础。以 18 世纪的休谟和 19 世纪的密尔为代表的哲学家首先从反事实框架讨论了因果关系,并且密尔还初步萌发了潜在结果的思想。但真正将反事实框架和潜在结果思想应用于因果推断的是 20 世纪的统计学家。1923 年内曼(Neyman)首次将潜在结果的数学公式引入统计学。^②内曼的重要贡献是解决了完全随机实验最重要的理论问题。费希尔(Fisher)的贡献在于引入了随机实验。^③在 1925 年费希尔引入随机实验之后,内曼的潜在结果模型很快就成为随机实验中定义平均因果关系的标准。^④不过,如何将潜在结果模型应用于观察数据,即在潜在结果方面获得对因果效应的一般正式定义以及分配机制的正式定义问题,却要到 20 世纪下半叶才能解决。^⑤只有解决了这一问题后,统计学才有可能建立起比较完善的因果推断技术,从而可以不仅利用实验数据,还可能利用观察数据进行因果推断。^⑥既然当时统计学未能解决如何利用观察数据进行因果推断的一些关键问题,因此政治学家对政治现象的因果推断主要依赖于实验。

实验最初起源于自然科学。随后由心理学家将实验法引入对人的研究。早在 18 世纪中叶开始,他们把自然科学的研究方法引入心理学,利用某些类似物理实验和生理实验方法研究心理学问题。至 19 世纪中叶已经有了一系列成果。因此,1862 年德国著名心理学家威廉·冯特(Wilhelm Wundt)提出了“实验心理学”。1879 年,冯特在莱比锡大学建立了世界上第一个正式的心理实验

① Gary King, “On Political Methodology”, *Political Analysis*, Vol.2(1990), pp.1—29.

② Donald B. Rubin, “[On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9], Comment: Neyman(1923) and Causal Inference in Experiments and Observational Studies”, *Statistical Science*, Vol.5, no.4(1990), pp.465—480.

③ R. A. Fisher, *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, 1925.

④⑤ Guido W. Imbens and Donald B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge: Cambridge University Press, 2015.

⑥ 今天常用的如双重差分、断点回归、倾向值匹配等因果推断技术就是在解决了这一问题后发展起来的。

室。20 世纪初在美国兴起的行为主义学派主张心理学的研究对象为人的行为,并开始将实验用于对人的行为的研究。但实验心理学研究的议题与政治学关注的问题少有重合,因此,从实验的角度而言,心理学对政治学影响主要体现在实验方法上。

随着社会科学逐步强调探求社会现象间的因果关系,实验法逐步被引入社会科学(尤其是经济学),并成为社会科学中最常用的检验因果关系的手段之一。实验在经济学领域的发展尤为引人注目。20 世纪 20 年代经济学逐步被塑造成一门以理性的、自利的行为者为基本假设的“真正的”科学,40 年代“理性人”假设已成为经济学的正统观念。因此经济学逐步形成以关注个人选择、预期效用和博弈论为重点的行为经济学。显然这些议题都是实验(尤其是实验室实验)研究的理想对象。因此至 20 世纪 50 年代初期,经济学家已经开始集中使用实验来检验经济理论(尤其是市场理论和博弈论)。此后实验在经济学研究中的地位持续上升,到 20 世纪 70 年代,美国出现了首个国家级的大型基金(national science foundation)支持利用实验研究经济行为。^①

然而,实验在政治学中的应用未能如在经济学中般顺利。其实,实验在政治学中的应用至少可追溯至 1924 年戈斯内尔在芝加哥对无党派人士进行邮件拉票对投票影响的实验。^②但至少在 20 世纪 80 年代前的发展极为缓慢。以政治学的顶级杂志《美国政治学评论》为例,该杂志 20 世纪 80 年代以前刊发的实验论文寥寥无几(见图 1)。1956 年埃尔德斯韦尔德在该杂志发表了一篇有关投票行为的实验论文。^③尽管到 60 年代末期定量研究已经成为政治学的主流研究方法,但实验却未能在政治学中实现突破。^④整个 60 年代,《美国政治学评论》

① Rose McDermott, “Experimental Methodology in Political Science”, *Political Analysis*, Vol.10, no.4(2002), pp.325—342.

② 在这项实验中,实验者将 6 000 多选民分成实验组和对照组,并向实验组的选民邮寄个人非党派呼吁书,以检验实验组与对照组之间投票行为的差异。Harold F. Gosnell, “An Experiment in the Stimulation of Voting”, *American Political Science Review*, Vol.20(1926), pp.869—874.

③ Samuel J. Eldersveld, “Experimental Propaganda Techniques and Voting Behavior”, *American Political Science Review*, Vol.50, no.1(1956), pp.154—165.

④ 一项统计发现,2009—2017 年间《美国政治学评论》发表的定量研究的比重超过 92%,同期《美国政治学杂志》完全未发表过任何纯定性论文。参见郑冰岛:《社会科学研究中的倾向分值匹配方法》,载《复旦政治学评论》2019 年第 1 期。

仅发表了三篇实验论文。^①此后随着实验在社会科学领域的逐步兴起,70年代《美国政治学杂志》实验论文的数量略有增长,共计有7篇实验论文。^②总体上,这一时期政治学对实验还未引起足够的兴趣,许多与政治现象相关的实验论文被政治学期刊不公平地拒绝了。^③因此最终发表于经济学或心理学杂志上。^④1971年创刊的《政治学实验研究》(*The Experimental Study of Politics*)起初推动了实验在政治学中的应用,但最终未能在更广泛的领域引起政治学家足够的兴趣,因此仅维持了四年就停刊了。^⑤

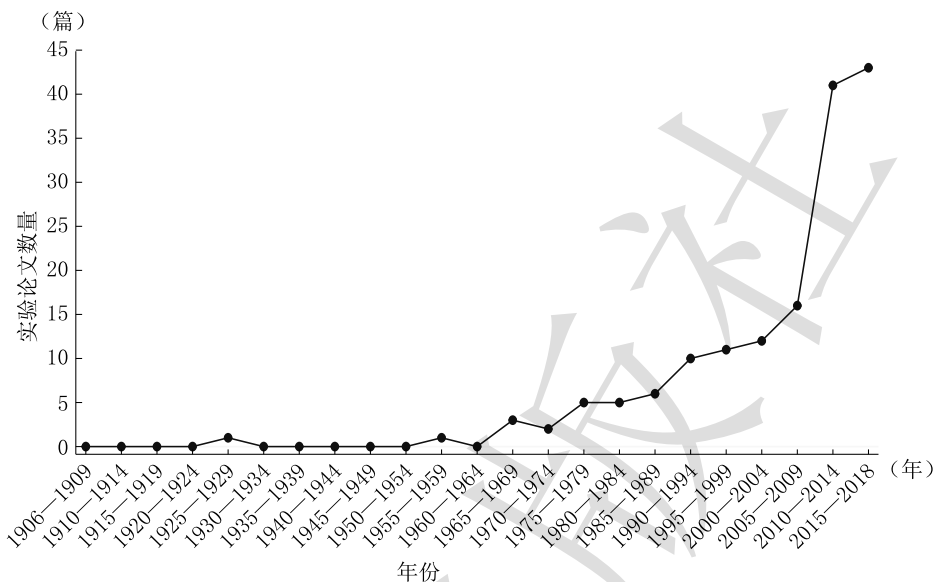
① 这三篇文章分别是 J. A. Laponce, “An Experimental Method to Measure the Tendency to Equibalance in a Political System”, *American Political Science Review*, Vol.60, no.4(1966), pp.982—993; James A. Robinson, Lee F. Anderson, Margaret G. Hermann and Richard C. Snyder, “Teaching with Inter-Nation Simulation and Case Studies”, *American Political Science Review*, Vol. 60, no.1(1966), pp.53—65; William H. Riker, “Bargaining in a Three-Person Game”, *American Political Science Review*, Vol.61, no.3(1967), pp.642—656。

② 这7篇文章分别是 William H. Riker and William J. Zavanon, “Rational Behavior in Politics: Evidence from a Three Person Game”, *American Political Science Review*, Vol.64(1970), pp.48—60; Daniel W. Fleitas, “Bandwagon and Underdog Effects in Minimal-Information Elections”, *American Political Science Review*, Vol.65, no.2(1971), pp.434—438; David Collier and Richard E. Messick, “Prerequisites Versus Diffusion: Testing Alternative Explanations of Social Security Adoption”, *American Political Science Review*, Vol.69, no.4(1975), pp.1299—1315; Michael J. Robinson, “Public Affairs Television and the Growth of Political Malaise: The Case of ‘The Selling of the Pentagon’”, *American Political Science Review*, Vol.70, no.2(1976), pp.409—432; Steven R. Brown, “Political Literature and the Response of the Reader: Experimental Studies of Interpretation, Imagery, and Criticism”, *American Political Science Review*, Vol.71, no.2(1977), pp.567—584; Barry S. Rundquist, Gerald S. Strom and John G. Peters, “Corrupt Politicians and Their Electoral Support: Some Experimental Observations”, *American Political Science Review*, Vol.71, no.3(1977), pp.954—963; Morris P. Fiorina and Charles R. Plott, “Committee Decisions under Majority Rule: An Experimental Study”, *American Political Science Review*, Vol.72, no.2(1978), pp.575—598。此外,还有学者统计了政治学三本顶级杂志——《美国政治学评论》《美国政治学杂志》(*American Journal of Political Science*)和《政治学杂志》(*Journal of Politics*)——上实验论文的发文数量,发现20世纪50年代仅有1篇,60年代和70年代分别有4篇和14篇。Rebecca B. Morton and Kenneth C. Williams, “Experimentation in Political Science”, in Jane M. Box-Steffensmeier et al. (eds.), *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press, 2008, pp.339—356。

③ J. McConahay and Jeanne N. Knutson, “Experimental Research”, in Jeanne N. Knutson (ed.), *Handbook of Political Psychology*, San Francisco: Jossey-Bass Publishers, 1973, pp.356—382。

④ 例如 Charles Plott, “A Notion of Equilibrium and Its Possibility Under Majority Rule”, *American Economic Review*, Vol.57(1967), pp.787—806。

⑤ Rose McDermott, “Experimental Methodology in Political Science”, *Political Analysis*, Vol.10, no.4(2002), pp.325—342。



资料来源:1906—2009 年数据来自 James N. Druckman, Donald P. Greene, James H. Kuklinski and Arthur Lupia (eds.), *Cambridge Handbook of Experimental Political Science*, Cambridge: Cambridge University Press, 2011。作者补充了 1926 年的数据,并重新收集并检查了 2005—2009 年的数据;2010—2018 年的数据由作者自己收集。

图 1 《美国政治学评论》实验论文数量

(二) 实验法在政治学中的发展

尽管 20 世纪 70 年代实验在政治学研究中并未形成爆发式增长,但却蕴含了突破的种子。一方面,政治学家越来越追求对政治现象的因果关系解释,定量研究已成为政治学研究的主流方法。然而统计学自诞生之初就深受因果推断问题的困扰。如何区别相关关系和因果关系是统计学发展过程中重要的脉络之一。20 世纪 70 年代当统计学家正努力解决通过观察数据进行因果推断的一系列关键技术问题时,许多社会科学家在实证研究中已认识到定量分析的这一缺陷。例如杰克逊对通过引入内生性变量的概念尝试提出了新的模型来研究选民的立场和对政党的评价如何影响政党认同和投票行为,从而开启了对政党认同理论的修正之路。^①此后,越来越多的政党和投票行为研究都绕不开因果

^① John E. Jackson, "Issues, Party Choices, and Presidential Votes", *American Journal of Political Science*, (1975), pp.161—185.

推断中的内生性问题。^①从因果推断技术角度来看,这样的尝试难以让政治学家满意。内生性、自我选择和遗漏变量偏差等限制了因果推断,并引起许多争论。然而,精心设计的实验在识别因果关系方面往往是有效的。因此,在统计学和计量经济学发展出一套成熟的因果推断技术并引入政治研究前,实验被越来越多的政治学家所采用几乎是注定的。

另一方面,其他社会科学(尤其是在研究议题上与政治学有许多重合的经济学)领域的应用已获得长足的发展,这必然会促进实验在政治学中的应用。1952年,法国经济学家莫里斯·阿莱基于一系列可控实验提出了“阿莱悖论”。弗农·史密斯1962年在《政治经济学杂志》发表《竞争市场行为的实验研究》,标志着实验经济学的诞生。经过20世纪60年代实验经济学的稳定增长、70年代研究基金支持带来的进一步增长后,80年代和90年代早期实验经济学迎来了爆发式增长。伴随着这种增长,实验经济学出现了成为主流学科的迹象。^②1988年阿莱凭借“阿莱悖论”获得诺贝尔经济学奖意味着实验经济学的真正到来。^③可以说,“阿莱悖论”彰显了实验的魅力。2002年被称为“实验经济学之父”的弗农·史密斯因其为创立实验经济学研究领域奠定了基础而获得诺贝尔经济学奖。实验经济学和政治学在许多研究议题上有重叠。^④因此,实验经济学的发展无疑会促进政治学研究中实验的应用。

图1显示,相比于20世纪六七十年代,80年代实验论文在《美国政治学评论》上的发文量呈现稳定增长,10年间实验论文的发表数量达到11篇,相当于该杂志自创刊到70年代末实验论文发表数量的总和。90年代开始实现某种突破,10年间发表的实验论文数量达到21篇,大约相当于80年代发表的实验论

① Philip E. Converse and Gregory B. Markus, “Plus Ça Change...: The New CPS Election Study Panel”, *American Political Science Review*, Vol.73, no.1(1979), pp.32—49. Benjamin I. Page and Calvin C. Jones, “Reciprocal Effects of Policy Preferences, Party Loyalties and the Vote”, *American Political Science Review*, Vol.73, no.4(1979), pp.1071—1089. Morris P. Fiorina, *Retrospective Voting in American National Elections*, New Haven: Yale University Press, 1981.

② Alvin E. Roth, “Introduction to Experimental Economics”, in John H. Kagel and Alvin E. Roth(eds.), *Handbook of Experimental Economics*, Princeton, NJ: Princeton University Press, 1995, pp.1—109.

③④ Rose McDermott, “Experimental Methodology in Political Science”, *Political Analysis*, Vol.10, no.4(2002), pp.325—342.

文数量的2倍。^①不过,至少到2002年实验仍未成为政治学中的主要研究方法。^②在经过21世纪初期的持续增长后,实验在政治科学研究中的应用终于在2010年以后出现了爆发式的增长。从2010—2018年,仅《美国政治学评论》的发文数量就高达84篇,已超过该杂志此前100多年实验论文发表数量的总和。另一本政治学顶级杂志——《美国政治学杂志》——2010—2018年发表的实验论文数量高达133篇,而2003—2009年仅发表了33篇实验论文。^③以2015—2018年实验论文的平均发表数量计算,《美国政治学评论》和《美国政治学杂志》实验论文的发表数量约占其总发表数量的15%和25%。^④可以说,实验已成为21世纪的第二个10年政治学研究的主流方法。

一项研究方法要成为学科主流,必然要求这种方法被广泛用于研究学科内的各项议题。麦克德莫特(McDermott)曾统计了自1926年到2002年主流政治学杂志发表的105篇实验论文,发现80%的论文研究了美国投票行为。这种情况在21世纪发生显著的变化。表1展示了2006—2018年间《美国政治学评论》和《美国政治学杂志》上实验论文研究议题的分布。表1不仅清晰地显示了与投票行为相关的实验论文比重的持续下降,更清晰地展示了这十多年间实验论

① 杜鲁克曼等学者在统计1906—2004年《美国政治学评论》上传统定义上实验论文数量时发现,超过半数的实验论文发表于1992年以后,参见James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia, "The Growth and Development of Experimental Research in Political Science", *American Political Science Review*, Vol.100, no.4(2006), pp.627—635。根据莫顿等人的统计,实验论文在政治学三大顶级期刊的发表数量在20世纪80年代达到25篇,已经超过此前实验论文发表数量的总和;90年代则进一步增加至42篇。Rebecca B. Morton and Kenneth C. Williams, "Experimentation in Political Science", in Jane M. Box-Steffensmeier et al. (eds.), *The Oxford Handbook of Political Methodology*, Oxford: Oxford University Press, 2008, pp.339—356。莫顿和威廉进一步统计了2000—2007年三大顶级期刊实验论文的数量,发现这段时间内实验论文发表数量持续增长。Rebecca Morton and Kenneth Williams, "The Advent of Experimental Political Science", in Rebecca Morton and Kenneth Williams, *Experimental Political Science and the Study of Causality: From Nature to the Lab*, Cambridge: Cambridge University Press, 2010, pp.3—28。

② Rose McDermott, "Experimental Methodology in Political Science", *Political Analysis*, Vol.10, no.4(2002), pp.325—342。James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia, "The Growth and Development of Experimental Research in Political Science", *American Political Science Review*, Vol.100, no.4(2006), pp.627—635。

③ 顾燕峰:《探求因果关系:实验在政治学中的应用》,载郭苏建主编:《转型中国的社会科学理论、范式和方法问题研究》,上海:格致出版社2019年版,第233—263页。

④ 如果算上自然实验、准实验论文及实验方法论文,则远超上述比率。参见上文。

文研究议题的扩展。首先,与投票行为相关的实验论文从 2006—2009 年的 27.03%下降至 2015—2018 年的 14.29%,投票行为不再是实验政治学的第一大研究议题,取而代之的是有关态度、观念与偏好议题(约占实验论文的18.37%)。其次,实验论文研究议题的扩展至代表性、性别、民主、人均关系、政党与政体、政治交流、腐败等议题。

表 1 实验论文研究主题分布

	2006—2009 (N=37)	2010—2014 (N=82)	2015—2018 (N=98)	总计 (N=217)
选民、选举与投票	27.03%	21.95%	14.29%	19.35%
态度、观念与偏好	18.92%	12.20%	18.37%	16.13%
族群与认同	16.22%	8.54%	12.24%	11.52%
认知、信息与决策	13.51%	7.32%	6.12%	7.83%
党派	2.70%	9.76%	7.14%	7.37%
组织、制度与行为	5.41%	7.32%	6.12%	6.45%
精英行为	2.70%	2.44%	6.12%	4.15%
政策	0.00%	3.66%	4.08%	3.23%
争端、暴力、恐怖主义与战争	2.70%	2.44%	3.06%	2.76%
政治参与	2.70%	1.22%	4.08%	2.76%
媒体	8.11%		2.04%	2.30%
代表性		1.22%	3.06%	1.84%
性别		3.66%	1.02%	1.84%
民主		3.66%	1.02%	1.84%
人际关系		3.66%		1.38%
政党与政体		3.66%		1.38%
政治交流		2.44%		0.92%
腐败			2.04%	0.92%
其他		4.88%	9.18%	5.99%
总计	100.00%	100.00%	100.00%	100.00%

资料来源:数据来自《美国政治学杂志》和《美国政治学评论》,2006—2018 年。表格原始数据来自顾燕峰:《探求因果关系:实验在政治学中的应用》,载郭苏建主编:《转型中国的社会科学理论、范式和方法问题研究》,上海:格致出版社 2019 年版,第 233—263 页。

一项研究方法要成为学科的主流,就必然需要引起同行的认同与关注。对学术研究而言,论文的引用率是一项重要的指标。德鲁克曼等人对发表于《美国政治学评论》上的 57 篇实验论文和与之研究主题相近的 193 篇非实验论文引

用率的比较发现,实验性文章($n=57$)平均被引用 40.1 次($SD=40.5$),而同期的非实验性文章($n=193$)平均被引用 30.5 次($SD=37.7$)。此外,同期的定量文章平均引用率是 29.7 次($n=89$, $SD=37.0$),内容相近的同期文章($n=114$)的平均引用率是 32.8 次($SD=41.3$)。^①显然,由于实验在检验因果关系上独特的优势,容易引起同行的关注与认可。

早期的实验法在社会科学中一直饱受实验的外部有效性的批评。批评主要集中在两个方面。首先,人为操控的实验环境有别于真实社会环境,因此实验的结论能否推广至现实世界还有待检验。其次,实验对象往往缺乏随机抽样,因此样本的代表性存疑。社会科学针对这两个批评进一步改进实验方法,越来越多的研究采用调查实验和现场实验替代实验室实验。调查实验结合了调查的优势,可以实现真正的随机性。随着网络技术的普及,调查实验被越来越多的社会科学家(包括政治学家)所采用。而现场实验是在实际环境中进行的,具备较强的外部有效性,因此也越来越多地被社会科学家(包括政治学家)所采用。通过比较 2006—2018 年间《美国政治学评论》和《美国政治学杂志》上的实验论文,可以发现现场实验论文和调查实验论文的比率分别从 2006—2009 年 21.1% 和 42.1% 增长至 2015—2018 年的 29.6% 和 52%,而实验室实验论文在实验论文中的比例从 2006—2009 年的 36.8% 下降至 2015—2018 年的 10.2%。从年均发表数量看,现场实验论文和调查实验论文的年均发表数量分别从 2006—2009 年的 2 篇和 4 篇增长至 2015—2018 年的 7.25 篇和 12.75 篇,而实验室实验论文的年均发表数量已从 2006—2009 年的 3.5 篇下降至 2015—2018 年的 2.5 篇。^②

四、实验法的扩展:自然实验

理论上,随机对照实验被认为是获得对干预的真实影响的无偏估计的唯一

^① James N. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia, "The Growth and Development of Experimental Research in Political Science", *American Political Science Review*, Vol.100, no.4(2006), pp.627—635.

^② 原始数据见顾燕峰:《探求因果关系:实验在政治学中的应用》,载郭苏建主编:《转型中国的社会科学理论、范式和方法问题研究》,上海:格致出版社 2019 年版,第 233—263 页。

手段。然而实际上,社会科学实验经常受到外部有效性、实验伦理等方面的质疑,也承受着实验成本的困扰。更重要的是,社会科学中的许多问题无法进行实验验证,例如战争、侵略、灾害、疾病和死亡等。因此研究者将目光转向自然实验。对自然实验的一个普遍共识是,当实施了特定干预的环境不受研究人员控制时,自然实验就会发生。^①与随机对照实验使用实验数据不同,自然实验使用来自真实世界中的观察数据,这避免了研究人员对实验环境和干预,从而可以具有较好的外部有效性,这也是自然实验最具吸引力之处。同时,研究者在进行自然实验时,近乎随机地将非实验对象分配到治疗组和对照组。只要能够证明分配是随机的,就可以证明研究者能够通过这种自然实验获得对“干预”的因果效应的无偏估计。^②

正因为自然实验的独特优势,这种方法被越来越多的社会科学家所采用,尤其是经济学和流行病学。^③政治学家使用自然实验研究政治现象也有较长的历史,并且在2010年后呈现爆发式增长。通过谷歌学术检索《美国政治学评论》上文字含有“自然实验”字段的论文,发现最早的两篇论文出现在1963年和1983年。^④此后整个20世纪90年代共有8篇论文提及“自然实验”,21世纪的前10年这一数字增加至12篇,而到了2010—2018年暴增到32篇。《美国政治

① Peter Craig et al., “Using Natural Experiments to Evaluate Population Health Interventions: New Medical Research Council Guidance”, *Journal of Epidemiology and Community Health*, Vol.66, no.12(2012), pp.1182—1186. Peter Craig et al., *Using Natural Experiments to Evaluate Population Health Interventions: New Medical Research Council Guidance*, London, UK: Medical Research Council, 2011.

② 当然,这对自然实验设计提出了更高的要求。实验者必须证明分配机制是随机的,才能说明实验的内部有效性。

③ 邓宁通过谷歌学术搜索“自然实验”发现点击率超过100万。经济学和流行病学对该词条的使用率最高,见Thad Dunning, “Improving Causal Inference: Strengths and Limitations of Natural Experiments”, *Political Research Quarterly*, Vol.61, no.2(2008), pp.282—293。在经济学领域,从1968年到2000年,《经济学文献杂志》(*Journal of Economic Literature*)就出现了68篇在题目或摘要中含有“自然实验”术语的论文,见Mark R. Rosenzweig and Kenneth I. Wolpin, “Natural Experiments”, *Journal of Economic Literature*, Vol.38, no.4(2000), pp.827—874。

④ J. David Singer, “Inter-Nation Influence: a Formal Model”, *American Political Science Review*, Vol.57, no.2(1963), pp.420—430. Robert D. Putnam, Robert Leonardi, Raffaella Y. Nanetti and Franco Pavnello, “Explaining Institutional Success: The Case of Italian Regional Government”, *American Political Science Review*, Vol.77, no.1(1983), pp.55—74.

学杂志》的情况与之类似。该杂志最早出现含有“自然实验”文字的论文是1980年。^①2000年前共有7篇论文提及“自然实验”,2000—2009年这一数字增加至16篇,2010—2018年迅速扩张至28篇。

自然实验的主要缺点来自其性质本身,即实验者不能挑选实验干预的具体内容,实验者也不能挑选干预的地点和时间。^②前者表明由于自然实验并非按照实验者的意图给予实验干预,因此实验干预可能无法与实验者的意图完全一致,这就导致自然实验无法完整地测量到实验者想要东西,因此限制了自然实验的使用场景。后者表明自然实验中实验者对实验干预缺乏控制,因此自然实验能否满足随机分配机制是存疑的。而随机分配机制可以确保处理组和对照组的可比性,这恰恰是实验核心的要求之一。^③为了克服这一缺陷,社会科学家在使用自然实验的时候,经常以匹配法(matching)确保处理组和对照组的可比性。^④本质上,匹配法是通过获得具有相似协变量分布的处理组和对照组来近似地复制随机实验,从而减少由协变量引起的偏差,其核心在于通过寻找一系列足够完备的个体特征来预测个体进入处理组和对照组的倾向性,从而使除了这些被用于预测的个体特征之外,再没有其他任何因素能够影响个体经历某一事件^⑤,即在控制了一系列个体特征后,个体进入处理组或对照组是一个随机过程。匹配法只是一种统计技术,不需要任何识别假设,其主要吸引力在于它是

① Lutz Erbring, Edie N. Goldenberg and Arthur H. Miller, “Front-Page News and Real-World Cues: A New Look at Agenda-Setting by the Media”, *American Journal of Political Science*, Vol.24, no.1(1980), pp.16—49.

② Glenn W. Harrison and John A. List, “Field Experiments”, *Journal of Economic Literature*, Vol.42, no.4(2004), pp.1009—1055.

③ 鲁宾因果模型(Rubin causal model)将潜在结果模型从实验数据推广到观察数据,为因果推断提供了正式的数学框架。该模型实际上包含了两个基本部分和一个可选部分。前两个部分分别是:(1)使用“潜在结果”来定义所有情形下的因果效应;(2)分配机制,即将“干预”分配给“单位”的明确的概率模型作为所有可观测的数量(包括所有潜在结果)的函数。可选部分是分配机制中条件数量的可选分布,包括潜在结果,从而允许基于模型的贝叶斯“后验预测”因果推理。见 Guido W. Imbens and Donald B. Rubin, “Rubin causal model”, in Steven N. Durlauf and Lawrence E. Blume(eds.), *Microeconometrics*, London: Palgrave Macmillan, 2010, pp.229—241.

④ 匹配法和断点回归不仅是自然实验的辅助手段,其本身也是基于鲁宾因果模型发展起来的计量模型。本文将在后面进一步说明。基尔对匹配法作了很好的总结,参见 Luke Keele, “The Statistics of Causal Inference: a View from Political Methodology”, *Political Analysis*, Vol.23, no.3(2015), pp.313—335.

⑤ 郑冰岛:《社会科学研究中的倾向分值匹配方法》,载《复旦政治学评论》2019年第1期。

一种完全非参数化并可以完全定制的调整形式,并且只需要在关键变量上进行非常接近或精确的匹配,而在不太重要的协变量上进行较宽松的限制。^①但这种方法是在建立在强假定基础上的,即在观察到的协变量的条件下,处理组和对照组之间不存在未观察到的差异。^②为了满足这一假设,匹配过程中必须包含所有已知和实验刺激分配与结果相关的变量。^③政治学家在近十年来越来越认识到自然实验的这一缺陷,因此也开始频繁在自然实验中使用匹配法和断点回归。根据作者在谷歌学术上检索发现,1960—2018年《美国政治学评论》上同时含有“自然实验”和“匹配”样的论文共有14篇,其中13篇是2010年及以后发表的。

五、实验的进一步扩展:基于计量模型的方法

(一) 断点回归模型

断点回归(regression discontinuity)通常被视为自然实验的一种。断点回归的核心思想是实验刺激(或处理)的分配完全或部分由位于固定阈值两侧的协变量(X_i)的值决定的,即个体进入处理组或对照组的概率是一个已知连续性协变量的函数。由于该协变量在临界点两侧是连续的,因此个体针对 X 的取值落入该临界点任意一侧是随机的,也就是说,在临界值附近的样本是随机分配进处

① Luke Keele, “The Statistics of Causal Inference: a View from Political Methodology”, *Political Analysis*, Vol.23, no.3(2015), pp.313—335.

② 但是这个假定无法用观测值来验证,因而无法证伪。Charles F. Manski, “Nonparametric Bounds on Treatment Effects”, *The American Economic Review*, Vol.80, no.2(1990), pp.319—323.

③ Donald B. Rubin and Neal Thomas, “Matching Using Estimated Propensity Scores, Relating Theory to Practice”, *Biometrics*, Vol.52, no.1(1996), pp.249—264. James J. Heckman, Hidehiko Ichimura and Petra Todd, “Matching as an Econometric Evaluation Estimator”, *Review of Economic Studies*, Vol.65, no.2(1998). Steven Glazerman, Dan M. Levy and David Myers, “Nonexperimental versus Experimental Estimates of Earnings Impacts”, *The Annals of the American Academy of Political and Social Science*, Vol.589, no.1(2003), pp.63—93. Jennifer Hill, Jerome Reiter and Elaine Zanutto, “A Comparison of Experimental and Observational Data Analyses”, in Andrew Gelman and Meng Xiao-Li(eds.), *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective*, Chichester: John Wiley & Sons Ltd., 2004.但如果仅用少量预测变量(如人口统计数据)时,匹配的效率高通常较差。William R. Shadish, M. Clark and Peter M. Steiner, “Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments”, *Journal of the American Statistical Association*, Vol.103, no.484(2008), pp.1334—1344.

理组和对照组的,因而构成了一个准自然实验。断点回归就是利用靠近断点附近的样本值来估计处理效应。这样做的困难是,要选择合适的带宽(bandwidth selection)。特别是,当使用非参数估计时,需要对 X_i 临界值两侧的平均值做出准确估计的话,带宽过大会导致有偏估计;但如果带宽太小,则样本可能太少。通常可以用局部线性回归或局部多项式回归来解决这个问题。但在经验研究中,一般仍然采用参数估计,并对接近临界值的观测值赋予更大的权重。^①

在一般的回归框架中,当可忽略性(ignorability assumption)假设(即与处理分配有关的协变量都被控制或者没有遗漏变量)的情况下,对实验平均处理效应的估计是无偏的。这是个强假定,因而在实践中经常不被认为是一个特别可行的假设。但在断点回归设计中,由于协变量(X_i)的取值决定了处理虚拟变量(the treatment dummy)的取值(即可以确定研究单位是否接受实验刺激),显然这个假设在断点回归设计中很容易被满足。因此,断点回归可以被视为“可观察因素的选择”模型(“selection on observables” model)的一个特例。^②

断点回归首先被应用于心理学研究。由西斯尔思韦特和坎贝尔在1960年研究优秀奖对未来学生学术成果的影响时首次使用。^③到20世纪90年代后期,经济学也开始应用断点回归。^④而政治学领域使用这项技术的时间更晚。几乎

① Joshua D. Angrist and Jörn S. Pischke, *Mostly Harmless Econometrics*, Princeton: Princeton University Press, 2008.

② David S. Lee and Thomas Lemieux, “Regression Discontinuity Designs in Economics”, *Journal of Economic Literature*, Vol.48, no.2(2010), pp.281—355.

③ Donald L. Thistlethwaite and Donald T. Campbell, “Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment”, *Journal of Educational Psychology*, Vol. 51, no. 6 (1960), pp.309—317.

④ 例如 Sandra Black, “Do Better Schools Matter? Parental Valuation of Elementary Education”, *Quarterly Journal of Economics*, Vol. 114, no. 2 (1999), pp. 577—599; Wilbert Van Der Klaauw, “Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach”, *International Economic Review*, Vol.43, no.4 (2002), pp.1249—1287; Joshua D. Angrist and Victor Lavy, “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement”, *Quarterly Journal of Economics*, Vol.114, no.2 (1999), pp.533—575,等等。因本斯(Imbens)和勒米厄(Lemieux)在他们2008年的研究中,以及李(Lee)和勒米厄在他们2010年的文章中分别对断点回归作了很好的总结,见 Guido W. Imbens and Thomas Lemieux, “Regression Discontinuity Designs: A Guide to Practice”, *Journal of Econometrics*, Vol. 142, no. 2 (2008), pp.615—635; David S. Lee and Thomas Lemieux, “Regression Discontinuity Designs in Economics”, *Journal of Economic Literature*, Vol.48, no.2(2010), pp.281—355。

可以肯定,这项技术在2010年以后才被政治学家广泛使用。根据谷歌搜索显示,1960—2018年《美国政治学评论》上含有“断点回归”字样的论文共计30篇,其中最早的记录是2009年。^①而且,同时含有“自然实验”和“断点回归”字样的论文有14篇,并且全部是2010年及以后发表的,占同期使用断点回归设计的论文的一半。

(二) 双重差分模型

双重差分模型的基本思想是,将所有样本分成处理组和对照组,并引入实验前测和后测,然后测量两组的后测平均结果差异与前测平均结果差异之差异。假定 $Y(i, t)$ 是感兴趣的结果, i 代表个体(研究单位), t 代表时间, $t \in \{0, 1\}$,其中 $t=0$ 代表处理前, $t=1$ 代表处理后。二分类变量 $D(i, t)$ 代表是否接受处理, $d \in \{0, 1\}$,其中 $d=0$ 代表未接受处理, $d=1$ 代表接受处理。由于只有在 $t=1$ 时个体才接受处理,因此,平均处理效应是:

$$\begin{aligned} ATE = & \{E[Y(i, 1) | D(i, 1) = 1] - E[Y(i, 1) | D(i, 1) = 0]\} \\ & - \{E[Y(i, 0) | D(i, 1) = 1] - E[Y(i, 0) | D(i, 1) = 0]\} \end{aligned} \quad (8)$$

其中前半部分代表处理组和对照组在处理后的(即 $t=1$)的差异,后半部分代表两者在处理前($t=0$)的差异。

双重差分模型的一个优势在于通过引入实验前测,可以放松可忽略性假设,即可以通过固定效应模型或者滞后被解释变量模型排除在时间上不变的(time-invariant)遗漏变量的影响。因此,研究者只要专注于考虑个体在未受干预时表现出的趋势特征,并在处理组和对照组中控制趋势特征,从而比较两者在差异。^②双重差分模型必须满足几个重要的假定:首先必须满足因果推断中的标准假设,即稳定单位处理值假设和外生性假设。前者假定,条件变量不受实

^① Andrew C. Eggers and Jens Hainmueller, “MPs for Sale? Returns to Office in Postwar British Politics”, *American Political Science Review*, Vol. 103, no. 4 (2009), pp. 513—533. Rikhil R. Bhavnani, “Do Electoral Quotas Work after They are Withdrawn? Evidence from a Natural Experiment in India”, *American Political Science Review*, Vol. 103, no. 1 (2009), pp. 23—35.

^② 对此详细的讨论,请参考 Joshua D. Angrist and Jörn S. Pischke, *Mostly Harmless Econometrics*, Princeton: Princeton University Press, 2008。

验处理的影响。其次,双重差分模型关键性的假设是平行性假设(或者共同趋势假设),即假定在处理前,处理对待处理的人群没有影响。这实际上排除了处理组在处理前的行为的变化,例如由于期待未来参与实验处理,因而在处理前就改变了行为模式。利用处理前多个时期重复观测的数据,研究者可以很容易地检验平行性假设。

双重差分思想由斯诺(Snow)在1855年研究19世纪中期伦敦霍乱问题就首次使用,并且在20世纪在心理学、经济学等领域获得广泛的关注。^①但政治学领域使用双重差分模型起步较晚。依据谷歌学术搜索,1906—2018年《美国政治学评论》上含有“双重差分”(difference-in-difference)的论文仅有17篇,均发表于2010—2018年间;与此类似,同期《美国政治学杂志》上含有上述文字的论文共有15篇,均发表于2011—2017年。这表明至少在政治学顶级期刊上,双重差分模型的使用起步很晚,但近期已经引起政治学家的较多关注。

(三) 工具变量估计

实验并不适合所有社会科学问题,因此在探求社会科学因果关系的过程中,社会科学家针对观察数据提出了不同于实验设计的因果推断方法。例如计量经济学家提出了联立方程模型,社会学家提出了路径分析等。^②工具变量法最初可以追溯到1928年赖特(Wright)对农产品市场的研究。^③工具变量法的基本想法是通过使用与因变量相关但不直接影响因变量的一个变量(即工具变量)来估计回归中与误差项相关的内生变量的影响。工具变量法最初被用于估计联立方程模型系数,但安格里斯特、因本斯和鲁宾在1996年提出工具变量估计可以嵌入鲁宾因果模型,用以解决观察数据的遗漏变量和测量误差问题。^④工具

① Michael Lechner, “The Estimation of Causal Effects by Difference-in-Difference Methods”, *Foundations and Trends in Econometrics*, Vol.4, no.3(2011), pp.165—224.

② 这些模型更一般的称谓是结构方程模型。

③ Phillip G. Wright, *The Tariff on Animal and Vegetable Oils*, New York: Macmillan, 1928.可参考斯托克和特莱比对估计变量起源的回顾,见 James H. Stock and Francesco Trebbi, “Retrospectives: Who Invented Instrumental Variable Regression?” *Journal of Economic Perspectives*, Vol.17, no.3(2003), pp.177—194.

④ Joshua D. Angrist, Guido W. Imbens and Donald B. Rubin, “Identification of Causal Effects Using Instrumental Variables”, *Journal of the American Statistical Association*, Vol.91(1996), pp.434, 444—455.

变量估计的优势是在使用观察数据进行因果分析时,可以避免在模型中加入过多控制变量,并且可以解决遗漏变量和测量误差问题。工具变量估计的难点在于满足所有条件的工具变量很难找。^①这可能限制了工具变量在因果推断中的应用。

政治学使用工具变量已有较长的历史。索韦和格林对1985—2008年《美国政治学评论》《美国政治学杂志》和《世界政治》上使用工具变量的论文进行统计分析表明政治学对估计变量的使用不仅在增加,而且对工具变量的使用也更加规范、成熟,但总体上依然还有许多提升的空间,例如仅有少数论文报告了第一阶段的结果、评估弱工具变量和过度识别约束等技术细节。^②

六、因果推断的新进展:机器学习技术的发展

机器学习是近年来在统计学和大数据基础上发展起来的一项技术,使社会科学可以处理更为复杂的数据形式并建构更为灵活的模型。与传统因果推断技术相比,机器学习技术有几项独特的特征。一是机器学习关注的重点是被解释变量的拟合值 $\hat{y}$,而非回归系数的无偏一致性;二是机器学习不关注理论相关性。因此,在因果推断框架下,机器学习本质上并没有改变社会科学研究的 basic 范式。^③但恰恰是这两个特征却确实给社会科学因果解释模型提供了修正的可能,并且被越来越多的社会科学家采用。

当研究者必须通过预测程序获得因果关系识别(即想要获得对 $\hat{\beta}$ 的无偏估计)时,机器学习可以通过提升预测的准确性,从而帮助研究者改进因果推断的准确性。例如在利用工具变量进行因果推断时,需要利用二阶段最小二乘法估

① 估计变量需要满足五个假设,即稳定单位处理值假设、随机分配假设、排他性约束假设、Z对D非零平均处理效应假设、单调性假设。详见Joshua D. Angrist, Guido W. Imbens and Donald B. Rubin, "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, Vol.91(1996), pp.434, 444—455。

② Allison J. Sovey and Donald P. Green, "Instrumental Variables Estimation in Political Science: A Readers' Guide", *American Journal of Political Science*, Vol.55, no.1(2011), pp.188—200.

③ 顾燕峰:《大数据与实证社会科学研究范式的延续与变革》,载《东南学术》2021年第1期。

计。第一阶段首先利用 $x = \gamma'Z + \delta$, 从而获得拟合值 $\hat{x}$; 然后利用拟合值 $\hat{x}$ 来估计第二阶段方程 $y = \beta'\hat{x} + \varepsilon$, 从而获得 β' 。显然, 要想获得 β 的无偏估计 β' , 取决于拟合值 $\hat{x}$ 在多大程度上捕获了 $\gamma'Z$ 的信号。如果在第一阶段回归方程中对 $\hat{x}$ 的过度拟合, 意味着拟合值 $\hat{x}$ 不仅拾取了 $\gamma'Z$ 的信号, 捕获了 ε 的信号, 这时, $\hat{x}$ 趋向于等于 x , 那么在第二阶段回归方程中 β' 会趋向于最小二乘法 ($y = \beta''x + \varepsilon$) 中的系数 β'' 。由于 x 是个内生变量, 这意味着 β' 是有偏的, 因此, 从模型估计的角度看, 工具变量估计的质量取决于拟合值 $\hat{x}$ 的质量。 $\hat{x}$ 的过度拟合将导致第二阶段回归中 β' 的估计是有偏的。在小样本或弱工具变量的情况下, 这个问题尤为严重。^①机器学习将样本分为训练集、验证集和测试集, 并利用训练集拟合模型, 再利用验证集优化调整模型的参数, 最后用测试集数据交叉验证。这可以改善对一个函数的预测质量。应用这些技术会产生分割样本工具变量和“刀切法”工具变量。^②最近有许多文章探讨如何将机器学习中的随机森林^③、套索^④、岭回归^⑤、神经网络^⑥等技术用于工具变量估计。

机器学习在因果推断中另一个典型的应用是用于基于观测数据的匹配法。鲁宾因果模型的随机分配假设要求研究单位被随机分配给处理组和对照组, 从而

① Sendhil Mullainathan and Jann Spiess, “Machine Learning: An Applied Econometric Approach”, *Journal of Economic Perspectives*, Vol.31, no.2(2017), pp.87—106.

② Joshua D. Angrist and Alan B. Krueger, “Split-Sample Instrumental Variables Estimates of the Return to Schooling”, *Journal of Business and Economic Statistics*, Vol.13, no.2(1995), pp.225—235. Joshua D. Angrist, Guido W. Imbens and Alan B. Krueger, “Jackknife Instrumental Variables Estimation”, *Journal of Applied Econometrics*, Vol.14, no.1(1999), pp.57—67. 参考 Sendhil Mullainathan and Jann Spiess, “Machine Learning: An Applied Econometric Approach”, *Journal of Economic Perspectives*, Vol.31, no.2(2017), pp.87—106。

③ Chen Jau-Er and Chen-Wei Hsiang, “Causal Random Forests Model Using Instrumental Variable Quantile Regression”, *Econometrics*, Vol.7, no.4(2019), p.49.

④ Alexandre Belloni, Daniel Chen, Victor Chernozhukov and Christian Hansen, “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain”, *Econometrica*, Vol.80, no.6(2012), pp.2369—2429.

⑤ Christian Hansen and Damian Kozbur, “Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE”, *Journal of Econometrics*, Vol.182, no.2(2014), pp.290—308. Marine Carrasco, “A Regularization Approach to the Many Instruments Problem”, *Journal of Econometrics*, Vol.170, no.2(2012), pp.383—398.

⑥ Jason Hartford, Greg Lewis, Kevin Leyton Brown and Matt Taddy, “Counterfactual Prediction with Deep Instrumental Variables Networks”, arXiv:1612.09596(2016).

确保处理组和对照组在统计学上无差异。在随机实验中,很容易实现受试者分配的随机性。然而在利用观察数据进行因果推断时,随机分配假设经常受到挑战。如前所述,研究者可以通过匹配法获得具有相似协变量分布的处理组和对照组来近似地复制随机实验,从而减少由于协变量引起的偏差。匹配法只是一种统计技术,不需要任何识别假设,其主要吸引力在于它是一种完全非参数化并可以完全定制的调整形式。以常用的倾向值匹配为例,在传统计量模型框架下,研究者往往依赖一个 Logistic 回归模型,利用一系列可以观测到的协变量预测个体进入处理组的概率,即估计倾向分值。应用倾向分值匹配方法进行因果推断的效果,很大程度上取决于对倾向值预测的准确性。^①如果我们把对倾向分值匹配方法的估计看作一个预测问题,那么机器学习将很有助于我们改进冗余模型的设定。^②近期的前沿研究发展出一套去偏差机器学习的方法,它的核心是使用机器学习技术同时改进两个预测问题,即对干预本身和对结果变量的预测。^③这些算法结合对样本的拆分以及交叉拟合,能够很好地改进匹配模型的因果效应估计。

通过机器学习改善因果推断是一项前沿课题,目前统计学和经济学对该方法的理论和应用有较多的讨论。对政治学而言,机器学习是一个全新的议题,但近年来引起了越来越多的政治学家的兴趣。例如瓦曼等人尝试利用机器学习进行数据编码。^④米勒和劳利用机器学习技术分析大量报纸文本资料来预测冲突。^⑤

① Daniel Westreich, Justin Lessler and Michele J. Funk, "Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta-Classifiers as Alternatives to Logistic Regression", *Journal of Clinical Epidemiology*, Vol.63, no.8(2010), pp.826—833.

② 具体的改进算法比较多,可参考 Brian K. Lee, Justin Lessler and Elizabeth A. Stuart, "Improving Propensity Score Weighting Using Machine Learning", *Statistics in Medicine*, Vol.29, no.3(2010), pp.337—346; Richard Wyss, Alan R. Ellis, M. Alan Brookhart, Cynthia J. Girman, Michele Jonsson Funk, Robert LoCasale and Til Stürmer, "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score", *American Journal of Epidemiology*, Vol.180, no.6(2014), pp.645—655.

③ Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Dufflo, Christian Hansen, Whitney Newey and James Robins, "Double/Debiased Machine Learning for Treatment and Structural Parameters", *The Econometrics Journal*, Vol.21, no.1(2018).

④ Michael Wahman, Nikolaos Frantzeskakis and Tevfik M. Yildirim, "From Thin to Thick Representation: How a Female President Shapes Female Parliamentary Behavior", *American Political Science Review*, Vol.115, no.2(2021), pp.360—378.

⑤ Hannes Mueller and Christopher Rauh, "Reading between the lines: Prediction of Political Violence Using Newspaper Text", *American Political Science Review*, Vol.112, no.2(2018), pp.358—375.

雅各布斯等人利用机器学习技术对新闻进行分类,进而发现美国的经济新闻的基调强烈且不成比例地追踪最富有家庭的财富,而对非富人的收入变化几乎不敏感。^①不过值得注意的是,虽然政治学家对使用机器学习研究政治现象的兴趣与日俱增,但到目前为止,政治学领域应用机器学习的案例大多数仅利用机器学习来对数据进行整理、分类和编码,而将其应用于改善因果效应识别的案例非常鲜见。

七、结 论

为政治现象提供因果解释是政治科学的基本任务。因此,如何利用因果识别工具对政治现象提供因果解释是每一位政治科学家要考虑的问题。本文从因果推断技术发展史的视角,回顾了从20世纪20年代至今基于实验的因果识别技术在政治学中的应用历史和现状。自然实验是因果推断的黄金标准,本文首先回顾了自然实验在政治学中的应用发展历程,发现经历了20世纪20—70年代缓慢的增长后,自然实验在政治学中的应用终于在80年代开始大量增长,并在2010年后出现爆发性增长。但同时,实验的外部有效性却饱受批评。自然实验使用来自真实世界中的观察数据,这避免了研究人员对实验环境和干预,从而可以具有较好的外部有效性。因此越来越多的研究者将目标转向了自然实验,2000年以后自然实验论文在政治学顶级期刊的发表数量越来越多。本文还考虑了三种基于实验思想的因果推断技术——断点回归模型、双重差分模型和工具变量估计——在政治学中的应用,发现尽管三种技术被政治学家采用的起始时间有差异,但2010年以后三种技术在政治学中越来越多地被采用,但仍有待普及。最后,本文还考察了机器学习技术在因果推断中的应用前景,并指出机器学习技术在政治学中目前的主要应用场景是数据清洗、编码等基础性工作,而真正将机器学习应用于因果识别的研究在政治学中仍然非常鲜见。

通过回溯实验政治科学中的引入和发展历程,本文一方面揭示了政治科学

^① Alan M. Jacobs, J. Scott Matthews, Timothy Hicks and Eric Merkley, "Whose News? Class-Biased Economic Reporting in the United States", *American Political Science Review*, Vol. 115, no. 3 (2021), pp. 1016—1033.

和统计学、经济学等学科的交流与相互借鉴的过程,另一方面展示了政治科学作为社会科学的一部分追求因果解释的不懈追求的历史。完全有理由相信政治科学还将继续吸收来自统计学、计量经济学乃至大数据科学的最新成果来进一步丰富政治科学研究的基础。近年来,机器学习技术在政治科学中的初步应用正好证明了这一点。实验技术和这些方法的结合,可以进一步扩宽实验,尤其是田野实验和自然实验的应用场景,甚至允许政治科学家放松某些实验条件,从而在更多的情境中应用实验来探明政治现象的因果关系。

此外,通过对实验在政治科学中的应用历史的回溯,本文也清晰地展示了与经济学相比,政治科学在对实验的应用始终处于跟随的地位。这与两个学科的发展程度密切相关。一方面,经济学在长期的发展历程中已形成独有的概念框架和理论体系,最终发展出一套较为严密的数学论证体系,这为量化研究提供了基础。另一方面,计量经济学不断吸收数学和统计学的研究成果,最终发展出一套较为严密的因果识别方法,这为实验法在经济学中的广泛应用奠定了基础。反观政治科学,无论从概念框架和理论体系,还是从方法论体系上,始终未能形成一套在学科意义上完整的理论体系。这注定了政治科学始终处于追随者的地位。然而,幸运的是,近十年来,政治科学明显加快了追随的脚步,并取得了显著的进展。