评论 深度

2021诺贝尔经济学奖，为何是一场静悄悄的革命？

如何判断一场革命有没有成功？

诺贝尔经济学奖得主之一，美国麻省理工学院教授安格里斯特（Joshua D. Angrist)。图：网上图片

Ye Wang

Ye Wang | 2021-10-14

2021诺贝尔奖　　评论

“革命胜利了！”

在今年诺贝尔经济学奖得主公布之后，推特上一众社会科学学者异口同声地发出了感慨。

他们口中的革命，是发轫于统计学，并逐渐扩散到社会科学各个领域，由因果推断（causal inference）方法驱动，悄然间改变了实证研究基本面貌的“可置信性革命（credible revolution）”。

如今，当你翻开一篇社会科学中的实证论文，有很大概率会发现如下字眼的身影：“识别策略（identification strategy）”、“内生性（endogeneity）”、“准随机分配（quasi-random assignment）”，亦或“自然实验（natural experiment）”。虽然含义略有不同，但它们都体现了同样的思想：想要论证从X到Y的因果关系，我们必须要依赖X独立于Y发生的随机变动。这一变动可以来自研究者的人为干预，即真正的对照实验，也可以源于出乎意料的外生政策或事件冲击。在后一种情况中，研究者无法控制随机分配的过程，只能观测到最终的结果，就仿佛是在自然中恰好撞见了一场由第三方执行完毕的对照实验。因此这种情况得名“准实验”或“自然实验”。

一个经典的例子是今年诺奖得主Angrist于1990年发表于《美国经济评论》（American Economic Review）的论文。他感兴趣的问题是，服兵役会给个体未来的收入带来怎样的改变。显然，直接对比有无参军经历者当下的工资水平，得到的估计并不准确。因为具备某些特质（比如身体强壮或服从纪律）的个体参军意愿更高，而这些特质又会影响他们在劳动力市场上的表现。因此，我们很难知道，工资水平的差异究竟完全是由兵役导致，还是源自个体在其他方面的差异。这些会对因果识别产生干扰的差异，被统计学家们形象地称为“混淆变量（confounder）”。

如果我们可以开展一场实验，随机地决定每名被试需不需要参军，那自然就可以排除混淆变量的干扰。只不过，这样的实验若由学者执行，必然违反伦理，不具备可行性。Angrist独辟蹊径，考察了七十年代初美国政府在越战期间进行的军事动员。当时的美国国防部出于公平性的考虑，采用了抽签的方式来决定每名适龄男性是否要应征入伍。Angrist的分析发现，被抽中的越战老兵跟未参战的同龄人相比，在八十年代的收入要低15%。由于抽签的随机性，这一历史事件相当于是由政府实施的大规模实验，因此上述数字可以被视作对服兵役和收入水平之间因果关系的可信估计。

旧制度

因果识别需要随机变动，这是现代统计学的开山祖师，Ronald Fisher和Jerzy Neyman就已经清楚认识到的道理。他们二人分别发明的置换检验（permutation test）和置信区间（confidence interval），至今仍是实验分析中最为常用的工具。只不过由于时代的限制，他们并未试图将随机化的理念推广到非实验数据之中。更糟糕的是，二战前的统计学，有一段和种族主义互相纠葛的不光彩历史。著名统计学家，也

是回归分析的发明者高尔顿，曾试图用统计方法，在不同人种的颅骨尺寸和平均智力之间建立因果关系。就连Fisher自己，也因为是优生学的忠实拥趸而饱受争议。战后成长起来的一代统计学家对这段历史深为不齿，转而拥抱了另一位统计学巨匠，高尔顿门生皮尔逊所提倡的理念：相关非因果（Correlation does not imply causation）。一时间，统计学的潮流变成了寻找反例，来揭示因果推断中可能存在的各种谬误，以证明此类尝试的荒诞不经。

与此同时，在大战期间自然科学高速发展的带动下，社会科学迎来了各色理论蓬勃兴起的时代。经济学在萨缪尔森的带领下迅速完成了数学化，建立了植根于最优化理论的微观基础；而纳什和阿罗的开创性工作，让均衡的概念深入人心。在政治学和社会学中，帕森斯的结构功能主义占据了主导地位。这一时期的理论家们，试图去厘清给定封闭系统（一个经济体或者一个社会）内，均衡状态下各个变量之间的相互作用。在一定的模型假设下，这些作用可以被归结为一组联立的线性方程，即所谓的“结构方程模型（structural equation modeling）”。方程组中，有些变量同时出现在左右两端——既接受来自其他变量的作用，又作用于其他变量——因而是“内生”的；另一些变量只存在于右手端，是纯粹的输入，可以被视为“外生”的。利用观测到的数据，我们可以试图求解（“识别”）方程的各个系数。如果我们相信，手头的理论模型是对现实足够精确的近似，那么解出的模型就可以帮助我们预测，当某个外生变量（比如利率）取值改变时，感兴趣的内生变量（比如储蓄率）会如何响应。

随着统计技术的发展，学者们建立的模型也越来越复杂——不但形式愈发灵活，囊括的变量也更多。然而，这种“基于结果的视角（outcome-based perspective）”不得不面对一个根本性的问题：影响任意结果（比如个体收入）的因素，都可能有成百上千种，无法被一个模型一网打尽。应该增加哪些变量，删去哪些，最后往往变成了理念上的争执，没有经验证据作为支撑——如果我们认为，年龄是决定收入的重要因素，那年龄的平方项呢？立方项呢？跟其他变量的交互项呢？要怎么判断哪个模型是更好的选择？怎么确保研究者不会刻意去选择提供了有利证据的模型？

统计学家 Jerzy Neyman。图：网上图片

火种

当社会科学家们为这些问题争论不休时，他们并不知道，一场将要横扫各个社科领域的革命，已经在美国的东海岸播下了火种。哈佛毕业的统计学家Don Rubin当时供职于美国教育考试服务中心（ETS）。他对于心理学家同事们在对照实验中也要使用模型的做法感到疑惑不解，于是写了一篇短文加以反驳。在这篇短文中，Rubin提出了一种后来被称为“潜在结果（potential outcome）”的分析框架，以帮助人们理解实验数据。所谓潜在结果，看起来是个再简单不过的想法：比如在随机分配药物的对照实验中，我们或者观察到某个被试服用药物后的健康状况（潜在结果1），或者观察到其服用了安慰剂之后的健康状况（潜在结果2），但永远不可能同时观察到两者。未被观察到的那个潜在结果，即被称为实际结果的“反事实（counterfactual）”。Rubin指出，任何实验处理（treatment）所产生的因果效应，等于两个潜在结果之差在样本中的平均值，即“平均处理效应（average treatment effect）”。因果识别的根本挑战，就在于利用手头的信息，去推断未被观察到的反事实会是怎样（Holland，1986）。

这个看似简单的框架，跟随机化的理念相结合之后，爆发出了惊人的能量。Rubin证明，只要处理分配是随机的，那么处理组（treatment group）和控制组（control group）的平均结果，即为两个潜在结果平均值的可信估计。进而两组在平均结果上的差异，也就告诉了我们平均处理效应的大小。更重要的是，Rubin注意到了实验和自然实验之间的相似性，第一次将非实验数据放置于实验的框架下进行思考：如果我们想知道X和Y是否存在因果关系，那与其去关注哪些因素影响了Y，倒不如想想X的变动由哪些因素决定，再去寻找X相对于Y发生了随机变动的情境，并由此对反事实进行推断。换言之，我们不再在乎“结果

的原因”，而是去关心“原因的结果”。前者要求我们对事物的运作规律有完整的认识，而后者只需要自然的一次无心插柳。

反过来说，Rubin的框架也意味着，只有存在随机化的处理分配时，讨论因果性才有意义。因此，当研究者们想在非实验数据中找到因果关系时，他们必须要回答如下问题：在何种假设下，数据的生成过程可以被视为一次随机实验？如何论证假设的可信性？如果假设未被满足，分析结果会不会发生很大变化？本质上，Rubin要求实证研究者用实验设计的标准来评判非实验数据的生成过程。自变量被视为“处理状态（treatment status）”，因变量是“结果（outcome）”，通过随机分配识别因果关系的具体方法则是“识别策略”。这种范式后来被称为“基于设计的视角（design-based perspective）”。

滥觞

但是，对于自然实验的分析，很多时候并不如真实实验那么简单。由于处理的分配过程无法被控制，甚至不能被直接观测，研究者往往要依靠统计方法对数据进行调整，以期尽量分离出感兴趣的因果关系。而这，正是今年三位诺奖得主最为显著的贡献。比如在Angrist那个越战老兵的例子中，一个不可忽视的事实是，抽签结果跟实际入伍情况并不完全一致——有些未抽中的爱国青年会主动从军，而另一些抽中的人会想方设法逃避兵役。在实验设计里，这种现象被称为“不依从（non-compliance）”。如何在存在不依从的时候估计因果效应？在1995年的一篇经典论文中，Angrist、Imbens和Rubin一起解决了这个问题。他们指出，根据处理分配和实际的处理状态，我们可以把整个样本分成四类人：始终接受者（always-taker），无论抽中与否都会从军的人；从不接受者（never-taker），无论抽中与否都不会从军的人；依从者（complier），只有被抽中时才会从军的人；违逆者（defier），只有没抽中签才会从军的人。这其中，违逆者在现实中的比例应该可以忽略不计。那么，利用抽签的随机性，我们就能计算出其他三类人的比例——比如在抽中签的人里，没去服兵役的必定是从不接受者，而余下的则是依从者。因为只有依从者的选择跟抽签结果有关，所以抽签造成的收入差异，完全体现在他们身上。也就是说，越战征兵的自然实验，只能帮我们识别兵役对依从者这个群体的处理效应，这被三位作者称为“局部平均处理效应（local average treatment effect）”。

他们的文章，为因果推断中一类重要的方法——工具变量（instrumental variable），奠定了坚实的统计基础。所谓工具变量，即跟处理状态相关，但又不直接改变潜在结果的某个变量，比如例子中的中签与否：中签的人参军概率更高，但收入只跟参军挂钩，不直接由抽签决定。尽管这一方法早已在经济学中出现，但直到这篇文章，学者们才意识到了工具变量与不依从之间的深刻关系，其背后蕴含的假设，以及其估计值的真正含义。

运用工具变量法，Angrist和合作者Krueger又巧妙地回答了经济学中另一个经典问题：更长的教育年限是否能带来收入的增长？他们为教育年限找到的工具变量，是个体的出生季度。出生季度显然不会直接影响

收入，但出生在下半年的人，相比于出生在上半年的同班同学年龄更小，也更难满足辍学所需的年龄要求（在美国一般为16岁）。比如A和B分别出生在1990年的3月和10月，那么到了2006年9月新学年开始之前，A已经年满16岁，可以选择辍学，而B则不能，只好多接受一年中学教育。Angrist和Krueger估计，多出的这一年，可以在将来让B的收入获得7.5%的增长。

当显而易见的随机分配难以寻觅的时候，实证研究者常用的一种识别策略，是将特质相似的个体放在一组，分组估计处理效应再进行加总。这背后的思路，是让数据看起来尽可能像分组实验（blocking experiment）的产物——在根据固定特质划分的组别里，我们有更充分的理由相信，处理的变动源自随机分配。举例来说，我们想知道有女儿的美国法官会不会在判决时更为仁慈。在由全体法官构成的样本里，有没有女儿未必是随机的。很明显，年轻的法官更可能没有女儿，而判决标准则也许更为严苛。但是，在每个年龄相仿、履历类似、政治立场相近的法官子群体里，家里有没有女儿跟判决的潜在结果，看起来就更像是互相独立的事件。

在统计学中，这一策略被称为匹配（matching），最早也是由Rubin发明。现实世界中，完美的匹配经常难以实现，我们不得不依靠各种各样的近似，比如将最为相似的每五名法官分为一组，或者根据法官们有女儿的概率——所谓的“倾向评分（propensity score）”——进行分组。从2006年至今，Imbens和合作者Abadie发表了一系列论文，研究了不同近似法则对最终估计的影响，并推导了相应的统计分布。他们证明，从匹配得到的估计并不满足中心极限定理成立的条件，因此无法应用自助法（bootstrap）进行统计推断。Imbens还进一步探讨了如何利用倾向评分进行加权（weighting）估计，以及如何将匹配或者加权跟经典的回归分析相结合，以得到更加稳健的估计结果。他提出的这些方法，目前都已经成为了因果推断中最为常见的工具。

在一些情况下，即使分组也不能完全确保随机的处理分配。我们也许会担心，分配过程还受到不可观测的混淆变量干扰。Card和Krueger在研究最低工资水平和快餐店员工就业率之间的关系时，就遭遇了这一挑战。他们想利用的自然实验，是从1992年4月1日起，新泽西州将法定最低时薪从4.25美元上调到5.05美元，而接壤的宾夕法尼亚州则未有改变。那么，是不是比较一下两州边界上特质相似的快餐店雇员人数上的差异，就能断定最低工资的上涨有没有拉低员工就业率呢？并不一定。一种可能是，新泽西州的顾客更习惯于自助点餐，因而快餐店里本来员工人数就更少。Card和Krueger的调查数据显示，在1992年2月时薪调整之前，新泽西州快餐店的平均雇员人数是20.44人，确实少于宾州快餐店的23.33人。

为了更好地利用时薪调整中包含的随机性，Card和Krueger采用了一种名为双重差分法（difference-in-differences）的识别策略。他们假设，也许最低时薪的上涨不独立于快餐店的雇员人数本身，但至少独立于雇员人数的变化趋势。这意味着，混淆变量的影响在各个时期是恒定的。这样的话，我们可以先计算每家快餐店1992年4月1日前后的雇员人数变化（一重差分）以消除混淆变量，再对比两州平均雇员变化率上面的差异（双重差分）。双重差分的结果显示，在处理开始之后，新泽西快餐店的平均雇员人数跟宾州的相比，并未表现出更缓慢的增长。在另一项研究中，Card将双重差分用于衡量古巴移民的涌入对迈阿密

劳动力市场的冲击，发现效应也十分微弱。

Card、Angrist和Imbens的开创性研究，极大地动摇了经济学界对于传统方法的信心。根据经典理论，最低工资的上涨必然导致就业率降低，移民的涌入必然降低市场薪资水平。现在，实证结果跟理论预测相悖，那我们应该去质疑自然实验的效力，还是去相应地修正理论？三位诺奖得主不约而同地选择了后者。在他们看来，基于随机处理分配得到的估计值，才是检验因果关系的黄金标准，要远比抽象的逻辑推导更为贴近现实。两者出现的偏差，恰恰说明经典理论，以及由此而来的结构式估计（structural estimation），可置信度（credibility）远远没有我们想像得高。经济学想进一步发展，想对现实有更大的指导意义，需要一场可置信性革命，需要以实验和自然实验作为理论的试金石。

统计学家Donald Rubin。图：网上图片

燎原

革命永远激动人心，哪怕在经济学家中也是如此。追随着他们三人的脚步，经济学界对于自然实验的热情变得空前高涨。断点回归（regression discontinuity）、控制函数（control function）、合成控制法（synthetic control）……新的识别策略层出不穷，渗透的领域也日益广泛。从民主是否促进经济增长（Acemoglu et al., 2001），到加强警力能不能够遏制犯罪（Levitt, 2002），革命者们在一个又一个议题上向传统智慧发起了挑战。自然实验的人气也让对照实验变得更为流行。以19年诺贝尔经济学奖得主Banerjee、Duflo和Kremer为代表的学者，将众多发展中国家变成了自己的实验场。他们与政府或国际组织合作，单刀直入地在实地检验不同政策的效果，将发展经济学带入了崭新的时代。甚至在许多曾被认为无法用实验方法研究的领域，比如国际贸易和产业组织中，经济学家们也挖空心思地引入了随机干预，并取得了丰硕的成果。

随着革命进程的深入，其真正的意义也愈发清楚地展现在人们面前。首先，新方法可以直接帮助我们评估某项政策的收益，因此前所未有地拉近了经济学研究跟现实世界的距离。过去三十年间，Card和Angrist在劳动经济学和教育经济学上不断地将知识前沿向前推进：技能培训到底有没有用（Ashenfelter and Card, 1985）？小班教学的收益有多大（Angrist and Levy, 1999）？哪种择校模式更有效率（Abdulkadiroğlu et al., 2017）？他们的工作深刻地重塑了美国从联邦到地方的政策制定，也为各个国家的无数后来者所效仿。其次，出乎意料的实证结果破除了经济学家对经典理论的迷信，让行为经济学等一批“旁门左道”获得了更多注意力，也迫使理论研究者更注重从现实中汲取灵感。最后，也是最重要的是，新一代经济学家们意识到，你不用再学习复杂的动态一般均衡模型（DSGE），也能做出极具价值的研究。你需要的，只是一本《基本无害的计量经济学》（Angrist和Pischke合著的教材），一个好用的统计软件，以及一次还未被人注意到的自然实验。随着门槛的降低，经济学中的实证研究变得空前民主化和国际化了，哪怕你身处第三世界，并不了解经济理论的前沿进展，也能借由本国的案例为学科发展做出贡献。

没过多久，从事定量研究的政治学家和社会学家们也感受到了革命的召唤，而他们的皈依某种程度上甚至比经济学家们还要更加彻底。这两个学科本身没有结构式估计的传统，因此范式转换的成本更低，更愿意接受被经济学家视为异端的工具，比如联合选择实验（conjoint experiment）、贝叶斯推断（Bayesian inference）和中介分析（mediation analysis）。新方法的出现，打开了通往无限可能的大门，让不少子领域重焕生机。在美国政治中，向政治家寄出内容随机剪裁的信件以探知其回应性（responsiveness），或者借由票数接近的选举来考察党派跟政策之间的关系，已经成为了学者们的常规操作。

因果推断在社会科学中的成功，反过来又让更多统计学家和计量经济学家对相关问题产生了兴趣。他们的加入，促进了因果推断和统计学其他领域的交流，加速了新方法的生产和应用。Imbens近年的工作就主要集中于这一方面。他跟合作者一道将各类机器学习算法引入了因果推断（Athey and Imbens, 2019）。机器学习的灵活性，让我们获得了对误差更加稳健的识别策略，也得以更加具体地刻画处理效应在样本中的分布，设计更加有效率的处理分配。

如今，因果推断已经成为了一座各个学科汇聚一堂的大熔炉：统计学家、社会科学家、计算机科学家和医学工作者们，一边立足本学科的需求提出新的问题，一边从其他学科那里借鉴洞察和经验，时不时合作进行难题攻坚。计算机科学巨匠Judea Pearl提出的因果图模型（Pearl, 2009），被政治学家入江直树用于研究德国仇恨犯罪的蔓延（Egami, 2019）；经济学家Manski对部分识别（partial identification）的思考（Manski, 2000），启发了流行病学家对抽样方式的改进（Crawford et al., 2018）。大数据时代的到来，让因果推断有了更广阔的用武之地。大数据能承载更复杂的工具，导向更准确的估计，进而帮我们筛选出更有解释力的理论；理论再为下一步的研究设计指明方向。社会科学家关于人类行为的知识正在以空前的速度积累，也许很快就能给人类社会本身带来天翻地覆的变化。

2021年10月11日瑞典斯德哥尔摩新闻发布会上，瑞典皇家科学院秘书长宣布诺贝尔经济学奖得奖者。摄：Claudio Bresciani/Reuters/达志影像

退潮、和解与新希望

但就跟其他革命一样，可置信性革命的发展，也始终伴随着攻讦和质疑。一部分批评来自革命阵营内部。随着对因果推断的理解愈发深入，学者们逐渐发现，早年的那批开创性研究，确实存在不少值得商榷之处。比如以出生季度作为教育年限的工具变量，看似巧妙，实则漏洞颇多。一来出生季度跟教育年限之间的相关性实际上非常微弱，是一个弱工具变量，会给估计带来偏误；二来出生季度未必是随机分配，新生儿往往集中于特定月份出生，而且跟家庭背景有关。抨击双重差分方法的论文，则更是汗牛充栋。新方法固然更加透明可靠，但也只有在满足一定假设的前提下才能生效。门槛降低的副作用之一，就是研究者们对统计工具不加思索的滥用，面对非实验数据，不去思索背后的生成过程，就生硬地套上一个工具变量或者倾向评分匹配，然后声称自己得到了因果效应的估计。

但是，世间没有万灵药。在没有随机分配的情境中，强求因果识别，反而违背了“基于设计的视角”这一根本理念，只会降低人们对社会科学，乃至因果推断方法的期待。近年来，以工具变量作为识别策略的论文愈发少见，很大一个原因，正是之前发表的大批劣质研究，摧毁了学界对这种方法的信任。举例来说，研

愿及少见，很大一个原因，正是之前发表的大批劣质研究，摧毁了学界对这种方法的信任。举例来说，研究非洲的学者们曾经用降雨量作为农业收成的工具变量，去识别收成缩减如何加剧武装冲突。但2015年发表的一篇论文发现，在修建了水坝的地区，尽管降雨量不再能影响农业，却还是跟冲突发生的概率高度相关（Sarsons, 2015）。这说明降雨通过其他渠道发生了作用，并不符合工具变量的基本要求。防止此类情况再度出现，一要靠统计方法的不断进步，二要靠学术界内部的制度建设，用“事前注册（pre-register）”等方式，减少滥用工具的激励。

更加刺耳的声音，来自传统方法的拥护者们。他们中的一些人认为，所谓“局部平均处理效应”根本没有意义，因为每次实验的依从者群体可能都各不相同，无法作为政策制定的参考。还有一些人断言，缺乏理论的指导，单靠对因果效应的估计，我们还是很难真正理解世界的运作方式。就连诺奖得主们也不得不承认，这些意见不无道理。但他们也指出，传统方法并不能帮助我们做得更好。因果推断跟结构式估计，也从来不是非此即彼的互斥关系。相反，将两种方法结合起来，能够让我们对社会现象有更加深刻的认识。城市经济学家们已经尝试过，将柏林墙的建立和倒塌作为两次外生冲击，来识别城市地理模型中的深层参数，诸如个体对居住环境的偏好和生产力对区位的依赖（Ahlfeldt et al., 2015）。政治学家Svolik在考察美国人对民主的态度时，将调查实验（survey experiment）嵌入空间投票模型（spatial voting model）之中，让人们直观地感受到了美国人愿意为民主制度支付的代价（Graham and Svolik, 2020）。

如何判断一场革命有没有成功？领袖人物誉满天下，桂冠加持，自然是一个强烈的信号。但也许更为重要的是，革命带来的改变，已经不知不觉间成为了我们生活中的一部分，让人很容易就忘记了旧时代的模样。可置信性革命无疑符合这样的标准。各个社科院系的研究生们，有几个没听过《基本无害的计量经济学》，有几个对潜在结果和处理效应闻所未闻？也许总有一天，统计工具上的新旧之争，会被彻底遗忘。范式的选择将取决于具体问题，而非学者的师承或者无谓执念。当被问及在方法论上的偏好时，我们每个人都会像政治学家Adam Przeworski那样作答：“我是一个方法论上的机会主义者……我从来没有什么原则。”

参考文献：

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. econometrica, 74(1), 235-267.

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. Econometrica, 76(6), 1537-1557.

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2017). Research design meets market design: Using centralized assignment for impact evaluation. Econometrica, 85(5), 1373-1432.

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5), 1369-1401.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., & Wolf, N. (2015). The economics of density: Evidence

from the Berlin Wall. Econometrica, 83(6), 2127-2189.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. The American Economic Review, 313-336.

Angrist, J. D., & Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106(4), 979-1014.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. Journal of the American statistical Association, 91(434), 444-455.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly journal of economics, 114(2), 533-575.

Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics. Princeton university press.

Ashenfelter, O., & Card, D. (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. The Review of Economics and Statistics, 648-660.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. Annual Review of Economics, 11, 685-725.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. ILR Review, 43(2), 245-257.

Card, D., & Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania.

Crawford, F. W., Aronow, P. M., Zeng, L., & Li, J. (2018). Identification of homophily and preferential recruitment in respondent-driven sampling. American journal of epidemiology, 187(1), 153-160.

Egami, N. (2018). Identification of Causal Diffusion Effects Under Structural Stationarity. arXiv preprint arXiv:1810.07858.

Graham, M. H., & Svolik, M. W. (2020). Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. American Political Science Review, 114(2), 392-409.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4), 1161-1189.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American statistical Association, 81(396), 945-960.

Levitt, S. D. (2002). Using electoral cycles in police hiring to estimate the effects of police on crime: Reply. American Economic Review, 92(4), 1244-1250.

Manski, C. F. (2003). Partial identification of probability distributions. Springer Science & Business Media.

Pearl, J. (2009). Causality. Cambridge university press.

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics, 159-183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5), 688.

Sarsons, H. (2015). Rainfall and conflict: A cautionary tale. Journal of development Economics, 115, 62-72.