经济学研究中的机器学习：回顾与展望①

王 芳[1] 王宣艺[2] 陈 硕[2]

（1. 华东师范大学公共管理学院；2. 复旦大学经济学院）

研究目标：随着数据的可得和计算机的发展，机器学习技术在经济学领域的应用发展非常迅速。本文旨在系统介绍机器学习在经济学中的应用。**研究方法**：简单介绍机器学习的定义后，本文将从数据生成、预测以及因果识别（DID、RD 和 IV）三个方面详细介绍机器学习在经济学中的应用。**研究发现**：局限于经济学因果识别方法的成熟及样本大小限制，本文认为机器学习虽然拓展了研究的边界，但并不会颠覆社会科学研究范式。**研究创新**：将机器学习的最新应用进行综述。**研究价值**：对机器学习在经济学中的已有应用进行分类归纳，并对未来研究进行展望。此外，本文也从学界不平等及可复制性等方面讨论了该技术在应用过程中可能带来的问题。

关键词 机器学习 数据生成 预测 因果识别

中图分类号 F224 **文献标识码** A

DOI:10.13653/j.cnki.jqte.2020.04.008

引 言

机器学习（Machine Learning，ML）指的是从数据中识别出规律并以此完成预测、分类及聚类等任务的算法总称②。在操作中，机器学习方法可以根据被解释变量是否已知被分成监督学习（Supervised Learning）和非监督学习（Unsupervised Learning）。监督学习是指被解释变量已知的机器学习。下面用 x 表示解释变量，y 表示被解释变量，那么监督学习首先依据已有样本数据建立 $y=f(x)$ 的函数关系。随后，当要预测解释变量为 x' 时被解释变量 y 的取值时，只需将 x' 代入等式右边即可获得预测值。非监督学习是指被解释变量未知的学习方式。换句话说，算法只知道解释变量 x 的取值。在这种学习方式下，算法会

① 本文获得国家自然科学基金面上项目“环境污染与公共健康：影响估计、成本核算及治理对策研究”（71773021）、重点项目“中国国家治理与长期经济发展的历史数据构建及计量分析”（71933002）、上海市教育委员会科研创新计划（2017-01-07-00-07-E00002）的资助。作者感谢复旦大学－清华大学“机器学习在社会科学研究中的应用”联合工作坊参会学者们的建议。

② 和机器学习相关并经常被同时提起的另外两个概念是人工智能（Artificial Intelligence，AI）与大数据（Big Data）。严格意义上，机器学习应归属于人工智能研究范畴。人工智能还包括诸如机械伦理学、自然语言处理和计算机图像识别等领域，机器学习更像是实现人工智能的手段和算法基础。为了和现有文献保持一致，这里对这两个概念不做明确区分，都称为广义的“机器学习”（Athey，2019；Camerer，2019）。大数据则是机器学习必需的数据支持，机器学习算法是基于大量数据来估计变量间的相关性并进行预测。另一个被广泛引用的机器学习定位是：如果一个算法在某项任务上的表现随着经验的积累而提升，那么就称该算法为机器学习（Mitchell，1997）。

分析 x 的内部结构，然后根据相似性把数据聚类（Cluster）。以垃圾邮件分类为例，在监督学习下，学者告诉计算机样本中哪些是垃圾邮件哪些是正常邮件。基于该信息，算法会找出邮件内容与类别间的规律并完成对未来邮件的分类。在非监督学习下，算法并不知道某一封邮件是垃圾邮件还是正常邮件。此时，算法依照邮件文本结构与相似度等指标把邮件归类。在整个过程中，算法只完成归类的工作，定义哪一类是垃圾邮件依然依赖于人工。

在实际研究中，学者一般根据被解释变量是否可得来选择使用监督还是非监督学习。比方说某项研究想回答某项政策能否提高民众幸福感。如果研究者能定期走访调查并建立覆盖政策前后时期的幸福感指数的话，作为被解释变量的幸福感就是已知的。这种情况下就可以选择监督学习。当然在该数据不可得时，学者也可以从互联网上搜集网民留言并分析其幸福程度。采用人力对浩如烟海的留言进行打分显然不切实际，此时可以借助非监督学习技术：让机器自动将留言分成幸福和不幸福两类以供后续研究。

在经济学领域，机器学习的使用虽然起步较晚，但发展非常迅速。例如，五大经济学英文顶尖期刊中涉及机器学习技术的文章数量在 2014 年之后以每年 74.7%的速度递增，2017 年的数量达到 16 篇。中文经济学权威期刊也有类似的趋势[①]。本文旨在系统介绍机器学习技术在经济学中的应用[②]。我们主要从三方面展开：第一，数据生成（Data Generating Process）：机器学习可以帮助学者获得以前很难或无法获得的数据，进而对一些更具挑战性的假设进行检验。第二，预测（Prediction）：机器学习可以更有效地探索变量之间的相关性，进而做出较为精准的预测。这部分将用公式表达的方式详细比较机器学习技术和传统基于回归方法在预测方面的异同。第三，因果识别（Causal Inference）：社会科学特别是经济学实证研究的核心是因果识别。由于机器学习在预测方面的优势，它可以被用来预测反事实进而获得因果效应[③]。最后，本文认为机器学习技术在上述方面的优势使其可以和经济学现有分析工具结合，检验之前无法用传统方法检验的假设，最终会拓展现有经济学乃至社会科学研究的边界。同时，研究者也应该对其带来的问题保持清醒认识，这些问题包括研究可复制性、过分依赖大数据及可能加剧学界不平等。本文最后一部分将对这些问题展开初步讨论。

一、关于数据生成

传统经济学实证研究基于的数据大都来自官方数据、问卷调查、实地调查、田野或者实验室实验等方式。最新一些研究试图利用机器学习技术拓展数据可得性。通过机器学习获得数据的主要方式是文本挖掘及图像识别。

就文本信息来说，研究者关心的是文本主题。为了在海量文本数据中提取主题，学者一

① 中文经济学权威期刊中涉及机器学习技术的论文在 2014～2017 年翻了 5 倍。正文中提到的五大期刊指的是 *American Economic Review*、*Econometrica*、*Journal of Political Economy*、*Quarterly Journal of Economics* 和 *Review of Economic Studies*。经济学中文权威期刊是指《管理世界》《金融研究》《经济学（季刊）》《经济研究》《世界经济》。此外还有另外两个更能反映未来研究趋势的指标：NBER 工作论文在 2016～2017 年的每个季度平均有 8 篇新文章是关于机器学习的；在美国前 20 名高校经济系讲座中，2016～2017 年每个季度平均有 2～3 个讲座涉及机器学习。这两个指标反映出来的趋势也充分证实机器学习近年来在经济学领域发展迅速。

② 由于篇幅所限，此处没有综述机器学习在业界和自然科学研究中的应用。对这部分内容感兴趣的读者，可以向作者索取。

③ 值得注意的是，本文与 Athey（2019）的综述性文章并不相同，由于篇幅所限，此处没有展开。对这部分内容感兴趣的读者，可以向作者索取。

般使用 Latent Dirichlet Allocation（LDA）方法[①]。例如，Hansen 等（2018）就利用该方法探究透明度政策如何影响政府内决策过程。这篇论文的研究背景是美国联邦公开市场委员会（Federal Open Market Committee，FOMC）在 1993 年通过决议公开了内部会议的发言记录。作者将该项政策视作自然实验以观察委员会成员的发言内容在该年前后的变化。研究基于的文本信息包含 5 万多次发言，总计 500 多万个单词，人工检索几乎不可能。作者便利用上文提到的 LDA 模型，从这些海量文本中提取 40 个不同的主题。任意一个成员的每一条发言都可以对应到这些主题中的一个或几个上。每个成员发言中各个主题的占比及成员间发言的相似度等指标就可以被计算出来，作者便可以使用常规 OLS 检验透明度政策对这些被解释变量的影响了。

除了提取文本主题外，类似使用机器学习从文本中生成变量的研究还有很多。包括利用朴素贝叶斯算法（Naive Bayes）对文本进行分类（Antweiler 和 Frank，2004），以及通过自动非参数文本分析（Automated Nonparametric Content Analysis）和支持向量机（Support Vector Machine）来判断文本作者的身份等（King 等，2017；Qin 等，2017）。

除了文本，机器学习也可以从图像中提取变量。卫星图像就是一个被经济学家广泛研究的图像信息[②]。例如，Engstrom 等（2017）试图测量一个地区的综合社会福利水平。在发达国家，研究者可以直接依赖官方数据或者调查数据。但很多落后国家由于没有足够财政维持经济普查机构的运转，其官方经济统计数据并不可得。为此，作者使用卷积神经网络（Convolutional Neural Networks，CNN）来识别卫星图片中建筑物、车辆及道路等固定资产，以此评估这些地区的福利水平[③]。我们认为未来的研究除卫星遥感照片外，学者还可以使用街景照片（比如 Google Street View）及卫星探测大气污染物图像等。结合机器学习及这类图像可以帮助学者识别出城市的边界或定位污染源，有助于研究区域经济学和环境经济学的相关问题。

另一个被广泛研究的图像信息是人像，一些研究采用机器学习技术识别人像的性别或对人像进行颜值打分，用以识别市场中是否存在基于性别或相貌的歧视。比如 Edelman 等（2017）通过用机器学习技术判别 Airbnb 上的用户头像性别进而分析租房平台上是否存在性别歧视；Cao 和 Chen（2018）在研究恋爱配对市场中颜值和物质条件发挥作用时，使用机器学习技术对研究对象的面貌进行打分并和人工打分比较。

上述研究主要涉及变量的“绝对”值，机器学习还可以为研究者生成“相对”意义上的变量。比较不同文本的相似度是该领域的典型应用。比如 Iaria 等（2018）试图研究一战冲击是否会影响跨国学术交流合作。在这个研究中，解释变量是战争的爆发，被解释变量则是论文的相似程度。作者预期战争会降低论文相似程度，基于如下逻辑：如果两个国家的学者经常交流，那么大家的研究兴趣和方向就会比较相似，这会导致论文成果也具有相似性。战争的爆发使得国家之间进入敌对状态，跨国学术合作被迫中断，这将导致同盟国和协约国各自的论文标题相似度下降。该研究需要解决的关键问题是比较论文间的相似度：样本包含 4

① LDA 可以通过不同单词在一段文本中出现的位置、频率和上下文，推测出这段文本中含有几个主题。该技术的具体细节和算法见 Blei 等（2003）。

② 这方面的工作和 Henderson 等（2012）的研究不同，该研究是将夜间灯光的明暗转变为数据。而这里介绍的则是从高分辨率的卫星图片中识别出房屋、汽车等和地区财富相关的指标。

③ 另一篇相似的研究是 Gechter 和 Tsivanidis（2018），两位作者利用 CNN 通过卫星图像定位贫民窟，从而检验了政府的房屋建设能否对贫民窟产生溢出效应。此外，通过卫星照片来估计地区财富水平的研究还有 Jean 等（2016）。

万多篇论文。作者采用基于机器学习的语义分析（Latent Semantic Analysis）来比较两两论文标题间的相似程度，实现了人工不大可能完成的工作。语义分析的结果发现论文相似度在一战爆发后显著下降，证实了作者上述猜想。

其他利用文本相似度进行研究的文献包括 Bleakley 和 Ferrie（2016）、Hoberg 和 Phillips（2016）等。Bleakley 和 Ferrie（2016）试图研究财富增加能否增加对后代的教育投资。由于普查数据来自多个年份，造成了部分父辈与子辈无法匹配（比如女子婚后改变姓氏）。作者使用机器学习技术并结合其他个人信息预测来自两个样本的不同个体是不是父子或父女关系以解决该问题。Hoberg 和 Phillips（2016）则研究了“9·11”事件如何影响军火企业。由于美国传统行业划分是不随时变化的，这就导致了那些由于这一事件进入或退出军火行业的企业无法被识别出来。为解决该问题，作者同样采用机器学习技术：分析公司每年的产品描述文档并根据其相似度划分行业分类。结果发现“9·11”事件后进入军火行业的企业数目显著增多。

除了对海量文本进行归类和比较外，机器学习技术还可以测量文字背后的情感。比如，Hills 等（2019）试图研究历史上人们的主观幸福指数。学者现在可以依靠社会调查数据测量现代社会公众的幸福感，但该方法并不适用于古代。作者采用的策略是利用机器学习计算不同时期出版图书中的幸福感。研究数据来自谷歌图书（Google Books Corpora），该数据库收录了 1500 年以来将近 1000 万本书。作者首先利用语言学和心理学文献中已有的“幸福感词典”，定义出每一个词所代表的“幸福值”，然后用计算机计算出每一本书的幸福感指数。为了验证该方法可靠性，作者比较了 1970 年后分别利用该方法和 Eurobarometer 社会调查所建立起来的意大利公众幸福指数，这两个指数之间呈现了强相关性，证实了该方法是可靠的。

另一个被学者特别是政治学者广泛关注的是文本中体现的政治立场[①]。利用机器学习技术对文本进行立场分析的相关研究包括：从各个党派党章或宣言中推测党派政策立场，通过新闻报纸措辞来判断报纸党派倾向，通过国会发言来测定党派分歧等（Laver 等，2003；Gentzkow 和 Shapiro，2010；Gentzkow 等，2019）[②]。我们认为基于文本生成变量将对未来经济学研究提供大量数据来源。上文提及的新闻报道、政府文件以及书籍出版物可以为政治经济学和福利经济学的研究提供大量实时的微观数据。此外，如微信朋友圈和微博等社交媒体上的网民发言及互动信息对于研究社会结构的学者也有一些帮助。

二、关于预测

在使用机器学习之前，经济学研究者主要依赖最小二乘回归（OLS）进行预测。本小节首先用公式推导的方法比较该领域被广泛使用的 Ridge（岭回归）技术和 OLS 在预测上的差异，并简单评价这两种方法的优劣；其次介绍利用机器学习进行预测的最新文献。

1. OLS 与 Ridge 在预测上的差异。

预测的目的在于找出两个变量间的相关关系。假设这两个变量间的真实关系是 $y=f$

① 当然，文字能反映的感情显然不局限于幸福感和政治倾向，研究者可以根据研究所需生成想要的情感倾向变量。比如 Wu（Forthcoming）就利用机器学习来甄别出哪些词汇被经常用来刻画男性形象，哪些词汇更可能描述女性。

② 除了情感和政治立场分析之外，也有一些研究利用机器学习分析主流媒体的报道来预测军事冲突。Mueller 和 Rauh（2018）首先利用和正文中提到的 Hansen 等人研究中类似的 LDA 方法提取《纽约时报》《华盛顿邮报》《经济学人》上的新闻主题并计算每年占比，然后将占比作为解释变量以预测军事冲突的可能发生地点。

$(x)+\varepsilon$。此处函数关系 f 客观存在但不为研究者所知。无论依赖于机器学习还是计量经济学，研究者的目的都是找到一个与 f 尽可能接近的函数 g，使得该函数估计值 $\hat{y}=g(x)$ 能够非常好地吻合真实值 y。评价一种预测方法好坏最常用的标准是均方误差（Mean Squared Error，MSE），也就是残差平方的期望，可表述为（Hastie 等，2016）：

$$MSE = \mathrm{E}[(y-\hat{y})^2] \tag{1}$$

当解释变量取值为 x_0 时，预测值 $\hat{y}=g(x_0)$ 与被解释变量真实 $y=f(x_0)$ 间的差异可被写为：

$$\begin{aligned} Err &= \mathrm{E}[(y-\hat{y})^2 \mid x=x_0] \\ &= \underbrace{[\mathrm{E}(\hat{y})-y]^2}_{\text{偏差平方}} + \underbrace{\mathrm{E}\{[\hat{y}-\mathrm{E}(\hat{y})]^2\}}_{\text{估计值方差}} + \underbrace{\varepsilon^2}_{\text{扰动方差}} \end{aligned} \tag{2}$$

通过式（2），均方误差被分解为三部分：估计值与真实值间的偏差（Bias）、估计值方差（Variance）以及真实值的扰动方差（Noise）。其中，扰动方差完全来自随机扰动项 ε，该部分不会消除且也不会由于预测方法的不同而存在差异。因此，不同预测方法减小均方误差的途径就是在偏差和方差间进行取舍①。

下面将从偏差、方差以及最终均方误差三方面，比较 OLS 和 Ridge 在预测方面的差异。为了推导的简洁，假设 Y 与 X 的真实函数关系 $f(x)$ 为线性且解释变量 X 为正交矩阵②：

$$Y = X\beta+\varepsilon \qquad X^{\mathrm{T}}X = I \tag{3}$$

OLS 的预测函数 $g(x)$ 可表示为 $\hat{Y}=X\hat{\beta}$，对式中 $\hat{\beta}$ 的估计方法是最小化残差的平方和，表示为：

$$\widehat{\beta_{OLS}} = \operatorname{argmin}\sum_{i=1}^{n}(y_i-\sum_{j=1}^{m}\beta_j x_{ij})^2 = (X^TX)^{-1}X^TY \tag{4}$$

此时，$\widehat{\beta_{OLS}}$ 的偏差是：

$$Bias_{OLS} = \mathrm{E}[(X^TX)^{-1}X^TY]-\beta = \beta-\beta = 0 \tag{5}$$

方差是：

$$Var_{OLS} = \frac{\sigma^2}{X^TX} = \sigma^2 \tag{6}$$

可知 $\widehat{\beta_{OLS}}$ 是真实值 β 的无偏估计量③。

由以上可以看出利用 OLS 进行预测的优点在于估计系数偏差为 0，缺点是方差可能较大。换句话说，选择若干个随机样本进行多次回归，无偏性保证所获系数的均值接近于系数的真实值。方差较大则意味着单次回归系数偏离均值较远，可能会异常大或小。当解释变量间存在多重共线性时，这一问题尤为严重。

① 偏差和方差的权衡取舍（Bias-Variance Tradeoff）是机器学习预测中的一个重大问题，详细讨论见 Bishop（2006）、Murphy（2012）以及 Hastie 等（2016）。

② 对于一般的非正交矩阵情况，Ridge 的预测能力也是优于 OLS 的，严格数学证明见 Theobald（1974）。若读者对 Ridge 感兴趣，该技术的更多内容可以参考 van Wieringen（2015）。

③ 当 X 不是正交矩阵时，OLS 得到的系数估计也是无偏的，这容易从式（4）中看出。

针对此问题，Ridge 在最小化目标函数中引入估计系数平方作为惩罚项，表示为[①]：

$$\widehat{\beta_{Ridge}} = \operatorname{argmin}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_j x_{ij})^2 + \underbrace{\lambda\sum_{j=1}^{m}\beta_j^2}_{\text{惩罚项}} = (X^TX + \lambda I)^{-1}X^TY \tag{7}$$

式（7）在直觉上非常容易理解：OLS 的缺点在于方差大，也就是估计系数的上下波动很剧烈。为了防止这种情况，机器学习在最小化过程中通过加入估计系数的平方或绝对值来“抑制”系数大小。如此便可以减小估计系数的方差使得预测更加稳定。这种思路可以理解为对系数大小的一种惩罚：过大则赋予较小权重，过小则相反。另一种常见的带有惩罚项的方法是 LASSO，它和 Ridge 的不同就体现在惩罚系数的选取上：Ridge 惩罚项为系数的平方 $\sum_{j=1}^{m}\beta_j^2$，而 LASSO 则是系数的绝对值 $\sum_{j=1}^{m}|\beta_j|$[②]。引入惩罚项后，Ridge 最小化的目标函数较之 OLS 更为复杂，而 LASSO 甚至无法导出估计系数的解析表示，只能求得数值解。操作上，最小化问题往往借助机器学习技术实现。

以下将分别比较 OLS 和 Ridge 估计系数的偏差、方差和均方误差的大小。首先，根据 X 是正交阵假设，由式（7）可得系数 $\widehat{\beta_{Ridge}}$ 为：

$$\widehat{\beta_{Ridge}} = \frac{\beta}{1+\lambda} \tag{8}$$

此时估计系数的偏差是：

$$Bias_{Ridge} = \mathrm{E}\left[\frac{\beta}{1+\lambda}\right] - \beta = \frac{-\lambda\beta}{1+\lambda} \neq 0 \tag{9}$$

方差是：

$$Var_{Ridge} = \frac{\sigma^2}{(1+\lambda)^2} \tag{10}$$

式（9）表明 Ridge 系数估计是一个有偏估计量，而式（10）则表示其方差比 OLS 要更小。换一句话说，在方差和误差的权衡中，Ridge 以有偏为代价换取更小的方差。

在获得了 OLS 和 Ridge 估计系数的偏差和方差后，根据式（2）分别计算两者的均方误差：

$$Err_{OLS} = 0 + \sigma^2 = \sigma^2 \tag{11}$$

$$Err_{Ridge} = \left(\frac{\lambda\beta}{1+\lambda}\right)^2 + \frac{\sigma^2}{(1+\lambda)^2} \tag{12}$$

为了比较两种方法的预测能力，将式（11）和式（12）做差：

$$Err_{OLS} - Err_{Ridge} = (0+\sigma^2) - \left[\left(\frac{\lambda\beta}{1+\lambda}\right)^2 + \frac{\sigma^2}{(1+\lambda)^2}\right] \tag{13}$$

若式（13）取值为正，那么 OLS 预测误差更大，反之 Ridge 的误差更大。可以看到，

① 可以看到，λ 反映了“惩罚力度”，当 $\lambda=0$ 时意味着无惩罚，此时 Ridge 和 OLS 完全一样。

② 严格来说，LASSO 和 Ridge 的“惩罚项”具有不同的含义：LASSO 进行的是 L1 范数正则化，Ridge 则是 L2 范数正则化。关于正则化详细讨论见 Bishop（2006）、Murphy（2012）以及 Hastie 等（2016）。

式（13）实际是一个关于 λ 的函数，其正负性也依赖于 λ 的值。现在问题转变为：什么样的 λ 会使得式（13）取正值或者负值？回答这个问题可以先考察该函数极值，如果该函数极大值都小于零，OLS 的均方误差将恒小于 Ridge；反之，如果该函数的极大值大于零，意味着一定能找到 λ 使得 Ridge 的均方误差更小。为找到极值，令式（13）的导数为零，得到一阶条件 $\lambda=\frac{\sigma^2}{\beta^T\beta}$。将该一阶条件代入，此时 $Err_{OLS}-Err_{Ridge}=\frac{2(\beta^T\beta)(\sigma^T\sigma)+(\sigma^T\sigma)^2}{\beta^T\beta+\sigma^T\sigma}$。式中分子分母都为正数，因此式（13）大于零，这意味着 Ridge 预测能力优于 OLS。事实上，Theobald（1974）证明了将该条件放宽到 $\lambda<2\sigma^2(\beta^T\beta)^{-1}$ 时，Ridge 的均方误差都是小于或等于 OLS 的。

下面将从“无偏性”和“可解释性”两方面评价传统计量经济学方法和机器学习方法在预测方面的优劣。正如本节开头所说，任何预测方法都是在偏差和方差间进行权衡取舍。社会科学实证研究，特别是经济学研究，特别强调因果推论。基于这种考虑，计量经济学回归模型都致力于获得一致的估计系数。这意味着在这一方差－偏差权衡中，计量经济学方法宁愿付出方差较大的代价，也不能放弃无偏这一性质（Athey，2019）。比方说上面所提到的 OLS 的估计系数正体现这一思路。而机器学习的目的就是进行预测——它并不在乎用以做出预测的估计系数是否具有无偏性特点。这就意味着在无偏性上，机器学习做出了“让步”：选择用偏差来换取更小的方差以提高预测性能。“可解释性”指的是从模型估计出的结果能够容易地被解释。计量经济学的目的不仅是预测，更在于解释现实中的现象以找到背后规律。从这个意义上来说，用来预测的函数形式越简单越好。因为复杂模型需要廓清模型拟合好坏的原因及解释变量与被解释变量间的互动关系等诸多问题。机器学习则恰恰相反，只要这个函数能够很好地模拟现实，哪怕函数形式再复杂也无所谓[①]。在这一点上，机器学习不拘泥于“可解释性”，灵活地选择函数形式进行拟合数据，这使得其预测能力强过了计量经济学传统方法[②]。

2. 用机器学习预测在文献中的应用

本小节将从个体和社会两个方面对现有利用机器学习进行预测的文献进行简单梳理。在个体层面上，机器学习可以帮学者更好地预测个人信息、决策或未来行为。此类研究包括 Oster（2018）、Kleinberg 等（2017）、Goel 等（2016）、Chalfin 等（2016），这里以 Oster（2018）为例进行详细介绍。该文研究糖尿病患者在确诊后是否会改变饮食结构。糖尿病患者是否将饮食结构调整为合理对其健康状况有很大影响，因此测量患者改变饮食的愿意程度是非常重要的，但以往文献由于缺少数据而无法测度这种愿意程度。为解决该问题，Oster（2018）首先基于个体消费记录（Nielsen HomeScan Panel）中的“血糖仪”“试纸”等关键字，利用随机森林（Random Forest）预测某人是否患糖尿病。得到预测结果后，再分析该个体在患糖尿病前后的食品消费记录来推断其是否改变饮食习惯。该文发现患者因为患糖尿病而改变饮食习惯的幅度非常小。Goel 等（2016）同样采用随机森林方法预测哪些行人更有可能携带武器；Kleinberg 等（2017）则通过梯度提升决策树（Gradient Boosted Decision Trees）预测被保释犯人是否会出席庭审；Chalfin 等（2016）采用随机梯度提升（Stochas-

① 很多机器学习技术都欠缺可解释性。比如，神经网络技术甚至能被视为黑箱：研究者无法得知解释变量通过何种机制影响被解释变量。一些致力于机器学习的学者也注意到了这个问题，正着手提高复杂模型的可解释性。相关讨论见 Bau 等（2017）。

② 用机器学习预测除 Ridge 之外，还可以采用其他非线性、半参或非参预测方法。Mullainathan 和 Spiess（2017）对比了这些方法与 OLS 的差异，结果发现机器学习的预测能力普遍强于 OLS。

tic Gradient Boosting）决策树来预测警察的工作质量。作者通过警员入职申请中提及的社会经济状况、婚姻、是否服役等信息，预测其未来工作中是否偏向使用暴力。

国内采用机器学习在个体层面上进行预测的相关研究也已经开始发展，它们大多集中在金融领域，主要包括对个人或企业的贷款风险信用进行预测。针对个体违约风险的研究中起步较早的是方匡南等（2010），该文通过个体的人力资本指标以及过往金融信用对信用卡违约风险进行预测；针对企业风险预测的早期研究包括郭英见和吴冲（2009）、徐晓萍和马文杰（2011），他们分别利用神经网络和决策树模型，基于财务指标估计了企业的信用风险。此后吕劲松等（2016）、马晓君（2015）、钱争鸣等（2010）分别就预测指标和模型选择上进行了改进。苏治等（2017）对金融领域中机器学习的实证应用做了一个综述。限于篇幅，这里无法做过多展开，读者可参考原文。

在社会经济层面，机器学习能够帮助研究者预测经济指标。比如，Blumenstock 等（2015）试图研究发展中国家的财富分布情况。作者遇到的问题和本文提到的 Engstrom 等（2017）的研究类似：落后地区的官方数据质量较差。和 Engstrom 等（2017）依赖于卫星图像不同，Blumenstock 等（2015）认为手机元数据（Mobile Phone Metadata，如通话历史等信息）不光能用来推断手机使用者的财富状况，同时也具有覆盖范围广、质量高且廉价的优势。作者先收集 856 人的手机元数据及他们的经济状况，再利用弹性网络（Elastic Net）建立手机数据与经济状况间的函数关系。作者发现该函数关系可以很好地预测财富分布。类似文献还有 Glaeser 等（2017），作者试图测量基层（Granular Level）实时（Close to Real Time）经济状况。但政府统计报表公布频度太低，时效性差。作者基于网络平台 Yelp 数据并通过随机森林来预测微观经济活动。国内学者也尝试通过机器学习完成预测目标，比如刘涛雄和徐晓飞（2015）、孙毅等（2014）通过互联网搜索数据分别对 GDP 及通货膨胀率进行预测。

关于机器学习在预测方面的应用，以下两个方面或许值得进一步研究。第一是基于家庭户消费数据的预测。随着网购数据和信用卡消费数据的逐步可得，研究者可以采用机器学习技术基于此类大数据进行预测分析。这能够帮助我们完成以前难以实现的研究，包括消费者行为改变或网购与实体店交互影响的实证研究等。第二是结合互联网微观大数据对宏观经济指标进行预测和度量。最近，大众点评的消费数据以及互联网打车软件的数据也逐步开放获取，这些大数据都可以用来预测地区层面的经济增长。

三、关于因果识别

社会科学，尤其是经济学实证研究的核心目标是获得因果推论，以探究干预（Treatment）措施是否导致预期结果并廓清作用发生机制。本部分将讨论机器学习技术在这方面的应用：首先基于著名的 Neyman-Rubin 反事实框架（Neyman-Rubin Counterfactual Framework）给出“因果效应”的定义（Neyman，1923；Rubin，1974）；随后结合目前应用微观计量经济学广泛使用的三种因果推论方法：双重差分（Difference-in-Differences，DID）、断点回归（Regression Discontinuity，RD）以及工具变量（Instrumental Variable，IV）展示该技术在其中的应用。

1. 因果关系与反事实

这里依然沿用 Rubin（1974）中患者吃药的例子来探究药物能否“导致”疾病被治愈。在现实世界，能够被观察到的“药物效果”是那些头痛并吃了药的人的健康状况减去那些健

康且没吃药的人的健康状况，用公式表述为：

$$Observed\ Effect = \mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 0] \tag{14}$$

其中，Y 表示个体健康程度，第一个下标为 1 表示该个体实际吃了药，0 表示没有吃药；第二个下标 i 代表第 i 个观察对象。D_i 取值 1 或 0 分别表示该患者是否患有头痛。因此 $\mathrm{E}[Y_{1i} \mid D_i=1]$ 就表示那些头痛且的确吃了药的患者健康程度，$\mathrm{E}[Y_{0i} \mid D_i=0]$ 则表示健康且没有吃药的人的健康程度。

显而易见，式（14）并不代表药物的因果作用：那些没吃药的人相对较为健康，与那些因为头痛而吃药的患者不可比。在这种情况下研究者无法区分两个群体在吃药后身体状况的差异到底来自药物效果还是来自个体差异。真正的效果应该是除了是否吃药外，其他所有因素都一样。换句话说，药物作用应当是同样一群患者（D_i 均为 1），分别测量没有吃药和吃药后的健康程度的差别，公式表述为：

$$Average\ Treatment\ Effect = \mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 1] \tag{15}$$

式（15）可以理解为药物作用是吃药的患者健康程度（$\mathrm{E}[Y_{1i} \mid D_i=1]$）减去如果他没有吃药时的健康程度（$\mathrm{E}[Y_{0i} \mid D_i=1]$）。但遗憾的是，这两者在真实世界中永远无法同时被观察到。这种无法观察到的情况被定义为吃药患者健康程度的“反事实”。虽然式（15）在真实世界中没有可操作性，但这不妨用该公式来重新组织式（14）：

$$\begin{aligned} &Observed\ Effect = \mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 0] \\ &= \underbrace{\mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 1]}_{\text{平均因果效应}} + \underbrace{\mathrm{E}[Y_{0i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 0]}_{\text{选择偏差}} \end{aligned} \tag{16}$$

式（16）可由式（14）中减掉反事实再加上反事实后获得，这依然是观察到的“因果效应”。但此时式（16）由两部分组成：第一部分正是药物的真实作用（式（15）），第二部分是 $\mathrm{E}[Y_{0i} \mid D_i=1] - \mathrm{E}[Y_{0i} \mid D_i=0]$，这里将其定义为选择偏差（Selection Bias）。该部分字面意义是那些有病没吃药的患者的健康程度减去那些没病也没吃药的个体健康程度。显而易见，该选择偏差小于 0，这是因为有病但没吃药的患者的健康程度当然差于没病也没吃药的个体健康程度。自此，可以知道观察法获得的所谓“因果效应”等于真实因果效应加上一个小于 0 的选择偏差。换句话说，观察法获得的“因果效应”低估了真实因果效应[①]。

缺乏反事实使得观察法获得的因果效应并不等于真实因果效应。该问题被称为因果推论的根本问题（Fundamental Problem in Causal Inference）。而从以上分析可以发现，该问题的根源在于个体差异：健康个体和病患个体存在诸多差异，因而前者无法作为后者的“反事实”。传统计量经济学所发展出来的所有分析工具，不管采用何种研究设计，其最终目的都是构建出介入组（Treatment Group）的反事实。一般来说，这些方法通过寻找恰当的控制组（Control Group）并提供证据来论证或假定该组可以作为介入组的反事实。找到适当的控制组后，它在介入后的取值即可以作为介入组的反事实，二者间差异为介入效果。

这个过程就为机器学习的应用提供了机会，即与其直接计算介入组和控制组在介入后的差异，不如利用控制组中样本构建出某种函数（比如样本的加权平均），使得该函数的取值与介入组足够相似，从而便可将该函数在介入后的取值作为反事实。用公式表述为：

① 在大多数情况下，研究者并不清楚选择偏差是大于还是小于零。这意味着在大多数情况下，偏差方向是未知的。

$$\widehat{Y_T(post)} = f(Y_C(pre), Y_T(pre)) \tag{17}$$

其中，Y_C（pre）是控制组没有被介入时的取值，Y_T（pre）是介入组在介入前的取值。而$\widehat{Y_T(post)}$是反事实的预测值，表示介入组如果没有被介入时的取值。该函数作为介入组反事实的合理性可以通过其在介入前的值与介入组在介入前的值 Y_T（pre）的差值反映。如果两者差异不大，表示该函数和介入组足够相似，即用它作为反事实是可靠的。一般来说，这个差异可以用该函数的预测值和 Y_T（pre）的差值平方来评价，公式表示为：

$$Error = \sum_{i \in pre} [Y_T(i) - \widehat{Y_T(i)}]^2 \tag{18}$$

如果将式（18）与式（1）对比，可以发现两者最小化目标函数十分相似，式（18）是最小化残差平方和，而式（1）是最小化均方误差（MSE）。从这意义上来说，对反事实的估计可以被看作一种预测（Varian，2014、2016；Athey，2019）。以下将依次结合 DID 及 RD 具体展示机器学习如何实现对反事实的估计。

2. 双重差分方法

双重差分是当前应用微观计量经济学中最常用的政策评估方法之一。图 1（a）展示了该方法识别因果的策略。

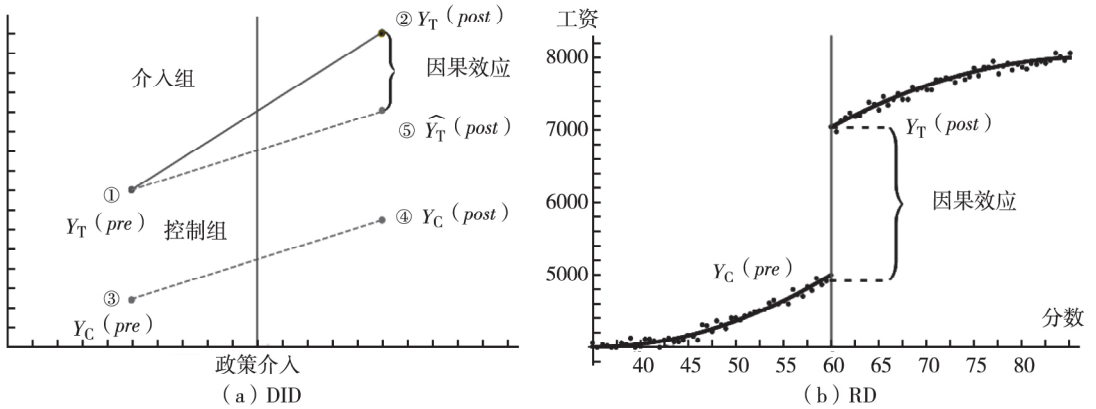

图 1 因果识别

注：（a）中—代表介入组（地区 T），……代表控制组（地区 C），竖线代表在地区 T 实施的某项政策。

假设某个地区（地区 T）被政策介入，该地区某项研究者感兴趣的指标在政策前（竖线左边）的取值为①，政策实施后（竖线右边）的取值为②。很明显，②－①并不是政策效果，因为研究者并不知道该地区如果没有政策的话，该指标取值是多少（反事实）。为了解决该问题，DID 的策略是寻找另外一个没有被政策介入的地区（地区 C），该指标在 C 地区政策前后的取值分别是③和④。①、②、③和④的取值分别表示为 Y_T（pre）、Y_T（$post$）、Y_C（pre）和 Y_C（$post$）。DID 假设⑤是 T 的反事实$\widehat{Y_T(post)}$。⑤的取值可以通过①、③和④点获得：⑤＝①＋（④－③），其中④－③为 C 地区的时间趋势，假定与 T 地区相同。因此，政策效果被表示为：②－[①＋(④－③)]，也可以表示为（②－①）－(④－③)，这也就是该方法被称为双重差分的原因。当然，出于展示的便利，图 1 中仅用 4 个点表示所有样本的四种情况。在实际研究中，这四个点背后会有很多观察值。此时，①＋（④－③）就被表述为：

$$\widehat{Y_T(post)} = \frac{1}{N}\sum_{i\in C} Y_{C,i}(post) + Constant \tag{19}$$

其中，Y 的第一个下标仍然代表控制组（地区 C），而第二个下标 i 则表示样本中的第 i 个观察值[①]。式子末尾的常数项为实验组与控制组在施加处理前的差异，也就是图 1 中①和③的垂直距离。因此，双重差分方法比较的是这四个点背后所有观察值的算数平均的差值。

从以上分析可得，DID 方法的有效性依赖于两个地区存在相似的时间趋势。如果研究者用样本算数平均数来构建反事实，上述平行趋势假设并不容易满足。此时，采用更加一般化的加权平均方法来构建反事实的效果可能更好。这是因为构建反事实的根本在于找出与介入组 T 尽可能同质的控制组 C，然而 C 中每一个观察对象与介入组的相似度可能各不相同。研究者自然会想到给那些与 T 相似的观察对象赋予更大的权重，而并非给所有对象赋予相同大小的权重。该思路被称之为合成控制法（Synthetic Control Method，SCM）（Abadie 等，2010）。在该方法下，反事实被表示为：

$$\widehat{Y_T(post)} = \sum_{i\in C}\omega_i Y_{C,i}(post) \qquad \sum_{i\in C}\omega_i = 1 \tag{20}$$

下一步需要解决的问题是每一个观察值被赋予的权重 ω_i，这可以通过最小化式（18）的残差平方和得到：

$$\begin{aligned} &\hat{\omega}_i = \arg\min_{\omega_i}\sum_{j\in pre}(Y_T(j) - \sum_{i\in pre}\omega_i Y_{C,i}(j))^2 \\ &\text{s. t.}\ \sum_{i=1}^{N}\omega_i = 1 \qquad \omega_i \geqslant 0 \qquad i = 1,2,\cdots,N \end{aligned} \tag{21}$$

受到 SCM 的启发，Doudchenko 和 Imbens（2016）将加权平均进一步放松为更加一般的线性组合函数来构建反事实，这也成为了机器学习在 DID 中应用的基本思路。此时该一般线性函数可被表达为：

$$\widehat{Y_T(post)} = \sum_{i\in C} k_i Y_{C,i}(post) + b \tag{22}$$

较之 SCM，此处并不要求 k_i 是权重，k_i 甚至可以取负值。同时，对 $\sum k_i$ 也不作任何要求。接下来的问题是如何找到参数 k_i 和 b 来最小化式（18）。两位学者使用了正则化回归（Regularised Regression），具体细节可以参照原文。至此，机器学习技术“改善”了 DID 方法中对反事实的估计，即利用控制组和介入组在政策实施前的信息建立线性函数并预测出反事实。

DID 与机器学习相结合的技术较为前沿，Cicala（2017）是目前少有的将 DID 与机器学习结合进行因果识别的研究[②]。该研究试图评估美国电网从国家计划发电到市场自动调整发电量这一措施所带来的收益。由于政府计划发电的废除是逐年逐地区推行的，这使得 Cicala（2017）可以用一个典型的 DID 框架来分析市场介入带来的因果效应。如上文所说，DID 的有效性依赖于样本在时间上具有平行趋势，但是这一假设在本文的设置下较难满足，这是因为影响发电的重要因素是燃料价格，即当前的燃料价格相对于历史价格的变动会导致发电成

① 为了展示的便利，此处假设介入组 Y_T 中仅有一个样本。因此，$\widehat{Y_T(post)}$ 没有 i 下标。该假设并不损失一般性（Doudchenko 和 Imbens，2016）。

② 在 Doudchenko 和 Imbens（2016）的研究中，两位作者首先用理论推导的方法表述机器学习如何与 DID 结合进行因果识别，然后以现有文献研究过的三个问题对其展示。

本改变，而不同地区的发电能力存在差异，因此这一成本变动在不同地区大小不同。最终相同的燃料价格变动会对不同地区的电量供给产生不同的变动，DID 无法控制这一异质性，导致构建的反事实距离真正的反事实有较大偏差。要解决这个问题，更准确地说，解决这个利用历史数据构建反事实的问题就是前文提到的预测问题。将燃料价格波动纳入考量，作者采用了随机森林的方法预测了每个地区倘若没有市场介入时发电量的反事实。

3. 断点回归方法

断点回归方法也是另外一种被广泛使用的因果识别方法。和双重差分方法用政策前后区分介入－控制样本组不同，样本是否被介入依赖于其中某一变量 X（Running Variable）相对于断点（Cut-off）的大小，可以表述为[①]：

$$\text{Treatment} = \begin{cases} 1 & X \geqslant cut\text{-}off \\ 0 & X < cut\text{-}off \end{cases} \tag{23}$$

该式子表示当 $X \geqslant cut\text{-}off$ 时，对应的样本被定义为介入组；而当 $X < cut\text{-}off$ 时，对应样本被定义为控制组。介入组和控制组都有相对应的被解释变量取值。假设被解释变量在相应控制组和介入组内部是连续变化的，如果该变量在断点左右出现跳跃，学者可以将其归咎于“基于断点”的介入所导致的因果效应（Imbens 和 Lemieux，2008）。

可以通过一个例子更加直观地展示上述研究设计的逻辑并在其中指出机器学习发挥作用的地方。最常见使用 RD 方法的研究问题是大学教育对工资的影响。研究者当然不能直接比较大学生和那些没上大学学生的平均工资差异，因为这两群人在能力上可能并不相同。这两群人在业界的工资差异除了反映大学经历的作用外，也提取了诸如能力等个体异质性因素。而在实际操作中，能力非常难以精确度量。断点回归的目的是通过比较两组“相似”的样本进而剔除掉这些异质性因素，以下是研究设计。

假设大学录取分数线是 60 分，那么总有学生因为 1 分之差和大学失之交臂。在这种情况下，研究者往往认为考 59 分的人运气太差，而非能力不足。换句话说，那些考了 59 分没能上大学的学生和恰好踩线 60 分得以进入大学的同学在能力上可能没什么太大差别，最终能否上大学完全是运气，而运气是随机的。这就使得学者在真实世界难能可贵地找到了用随机选择分配大学的机会。在该设定下，那些考 59 分的同学就无限接近于那些考了 60 分同学的反事实：考了 60 分但没读大学的同学。所以 59 分同学和 60 分同学在未来的工资差别可以视作为大学教育对收入的因果效应。

图 1（b）展示了上述研究设计思路，其中横坐标代表考生分数，纵坐标代表工资。图中的 60 分，也就是能否上大学的分界线，用竖线加以表示。可以看到，那些考上 60 分的学生们被划分到介入组 T，59 分的学生们则是控制组 C。在 DID 设置中，前者就是 Y_T（$post$）（介入组被介入后，②），后者就是 Y_C（pre）（控制组在介入前，③）。和 DID 的区别在于，RD 的研究设置中不存在 Y_T（pre）和 Y_C（$post$）——前者表示考上 60 分的学生（介入组）没有上大学的工资（①），后者表示 60 分以下的学生（控制组）上了大学的工资（④）。由于缺乏时间维度带来的这两个信息，RD 无法使用 DID 中采用的①、③及④的信息构建出反事实⑤。在这种情况下，RD 采用的策略是绕开①和④的信息，直接假设③等同于⑤。该假设在 RD 中之所以成立的原因在于上文所说的考 59 分的学生和考 60 分的学生高度同质。此

① 不妨假设大于断点的那部分样本被介入。

时，用公式表述大学的因果效应为：

$$\begin{aligned} Treatment\ Effect &= \lim_{\varepsilon \to 0^{+}} \mathrm{E}[Y_i \mid score_i = 60 + \varepsilon] - \lim_{\varepsilon \to 0^{-}} \mathrm{E}[Y_i \mid score_i = 60 + \varepsilon] \\ &\approx \mathrm{E}[Y_i \mid score_i = 60] - \mathrm{E}[Y_i \mid score_i = 59] \\ &= \widehat{Y_Y(post)} - \widehat{Y_C(pre)} \end{aligned} \tag{24}$$

其中，$score_i$ 代表第 i 个学生的分数，Y_i 表示其工资。第一行利用了左、右极限刻画了真实的因果效应：同样考 60 分的一群学生仅仅由于是否上大学而导致的工资差。随后依据上文假设，RD 近似地认为 59 分与 60 分的学生同质，从而得到了后两行的结论。

即使如此，挑战依然存在。考 59 分和 60 分的学生在能力上相同的假设可能太强，因为 1 分之差也是能力的差别。此时，RD 效果依然会提取能力作用进而导致大学对工资的影响被高估。机器学习可以帮助研究者消除这一差异（Varian，2016）。通过各种预测模型，机器学习技术能够通过以下步骤构建出考到 60 分但却没有读大学的同学的未来工资（⑤）。首先，机器学习可以利用③的信息（小于等于 59 分的样本）归纳出没有上大学人群中工资与分数间的关系：

$$Y_C(score) = f(score) \qquad score \in pre \tag{25}$$

然后扩大函数的“定义域”：将 60 分作为解释变量代入式（25）右端的函数中。此时函数的取值就是那些“倘若”考到 60 分但却没有接受大学教育的同学未来的工资，将此作为考到 60 分且读大学同学的反事实（⑤）：

$$\widehat{Y_T(60)} = \widehat{f(60)} \tag{26}$$

此时，比较机器学习得到的反事实 $\widehat{f(60)}$ 与传统 RD 方法得到的反事实 Y_C（59），可以发现机器学习已经将 59～60 分之间的能力剔除在外了，进而获得更加精确的因果效应。

现在问题转变为怎样的机器学习预测函数 f 能够达成推测因果效应的目的。第一，f 应当具有较好的预测性能，即尽可能减小均方误差。第二，f 给出的估计量应当具有良好的统计性质：包括在大样本下渐近一致性以及较窄置信区间等。针对这些要求，Imbens 和 Wager（2019）利用凸优化的数值方法（Numerical Convex Optimization Method）来进行断点回归的因果推断，超越了传统上用来进行 RD 识别的局部线性回归（Local Linear Regression）[①]。

4. 工具变量方法[②]

除了双重差分及断点回归方法之外，应用微观计量经济学者也经常使用工具变量方法（Instrumental Variable Approach）来识别因果关系。和以上两种方法依赖于寻找同质样本的思路不同，工具变量方法试图寻找外生变量来克服异质性与样本是否被介入间的关系。实际操作采用两阶段最小二乘法（Two-Stage Least Squares，2SLS）实现。在第一阶段通过 OLS 线性估计用外生工具变量“替代”内生解释变量（是否介入或者介入的程度），从而获

① 传统方法有两大不足：首先，传统方法的前提假设是变量的连续性，在上文例子中该假设意味着分数必须是连续的，然而很多实证研究里的变量都是离散的。其次，传统方法虽然具有一致性，但它并不是最小化均方误差的估计量。两位作者的研究利用了机器学习中常用的凸优化数值方法改进了这两个问题，具体细节请参考原文。局部线性回归的相关研究见 Armstrong 和 Kolesár（2018）、Kolesár 和 Rothe（2018）。

② 机器学习在工具变量方法中的应用本质上仍然属于预测部分，而并不涉及反事实的估计。处于内容的相关性，这部分被放在因果推论部分论述。

得内生解释变量的预测值。该预测值的方差都是由于外生工具变量所解释，与异质性之间的关系便不再存在。在第二阶段中，用解释变量预测值和被解释变量回归，获得解释变量的一致性估计系数。

在这里仍然采用教育对收入的作用来展示工具变量方法的操作方法。该例子来自 Angrist 和 Krueger（1991），作者试图估计教育时长对工资的作用，估计公式为：

$$wage_i = \beta_0 + \beta_1 edu_i + u_i \tag{27}$$

上文提到，不管是否上大学还是大学教育时间长短均和个体异质性有关系，这就意味着上述因果关系中存在内生性问题：Cov（edu，u）$\neq 0$。这导致研究者无法区分观察到的收入差异到底来自于教育还是个体异质性。为了应对该内生性问题，两位作者采用工具变量方法，他们认为出生时间 z 是一个很好的工具变量。对该变量的评价需要了解一下美国的义务教育制度：义务教育法律规定学童在年满 6 周岁时要入学读书，年满 16 周岁后才可以离开学校①。法律规定的“年满 6 周岁”指的是当年 1 月 1 日年满 6 周岁。该一刀切的规定会导致出生月份不同的学童实际接受教育时长存在差别。举一个极端例子，一个 12 月 31 日出生的学童，在 6 年后的 1 月 1 日时恰好 6 周岁多一天。按照法律规定，该学童符合入学条件。而另一个在 1 月 2 日出生的学童在入学日时却只有 5 周岁 364 天。虽然只有 1 天之差，但依然不能入学，必须等到下一年 1 月 1 日。那时他已经 6 周岁 364 天。由于离校都是 16 周岁的那天，这会导致 1 月 2 日出生的学童比 12 月 31 日出生的学童少接受 364 天教育。当然大部分学童受到的教育时长都小于该极端值。可以看出，上述制度设置所导致的教育时长差异是由于出生月份导致，如果假设能力和出生月份无关的话，那么该变量就是教育时长的有效工具变量。

工具变量方法的实施关键在于第一阶段，不光需要给出证据证明工具变量具有外生性，还要通过统计指标说明该工具变量和内生解释变量之间存在足够强的相关关系。在这篇研究中，作者给出一些证据比如 Z 值的显著性来说明出生季节的确和教育时长之间存在相关关系，但后续许多学者认为该相关关系并不强以至于影响最终的估计结果（Bound 等，1995；Staiger 和 Stock，1997；Card，1999）②。该问题本质上仍然是外生 z 对内生 edu 的预测能力，而这正是机器学习最擅长的地方（Varian，2016；Mullainathan 和 Spiess，2017）。因此，工具变量方法的第一阶段完全可以采用机器学习技术预测内生解释变量。这一领域已经积累起了较多的理论计量文献：有些学者采用正则化回归，比如 LASSO 和 Ridge 等方法来构建第一阶段的估计（Belloni 等，2012；Carrasco，2012；Hansen 和 Kozbur，2014）；另一些学者则采用神经网络等非线性方法来进行第一阶段的估计（Hartford 等，2017）。

四、展望及结论

经济学中机器学习的应用近几年获得了长足发展，无论在数据生成、预测还是因果识别等诸多方面都得到应用。但是本文认为机器学习技术尚未成为经济学研究的主流方法，目前机器学习的贡献主要在于提供更多变量生成以及预测工具，对整个社会科学因果识别的研究范式的冲击有限。就数据生成来说，机器学习一方面提高了数据搜集和整理的生产率，将以前需要通过庞大人力及大量时间才能生成的变量通过以机器学习算法辅助的方式自动生成；

① 有些州规定的离校最低年龄是 17 岁，这不会影响下面的讨论。

② 限于篇幅，这里无法给出弱 IV 为什么会使得估计结果产生偏差或产生不一致的系数估计。关于 IV 影响估计结果的讨论，见 Bound 等（1995）、Staiger 和 Stock（1997）、Chao 和 Swanson（2005）。

另一方面通过机器学习手段可以将图像与文本进行量化，增加了经济学研究的数据来源。目前国内社交媒体逐步开放大数据获取更进一步凸显了机器学习在变量生成上的优势。但这些由机器学习生成的数据依然以变量形式进入传统经济学研究框架内，本质上没有改变经济学的研究方法。就预测来说，机器学习相比计量经济学更加宽松的设置使得其预测性能远远超过传统计量手段，这使得其在学术界得到了比较多的运用。然而目前经济学在该领域的应用在很大程度上是对业界已有成果的复制，引领这一领域发展的驱动力依然是商业应用。就最有可能产生颠覆意义的因果识别来说，虽然利用机器学习的预测优势构建处理组的反事实在方法论上行得通，但目前并没有被研究者所广泛接受和使用。本文认为其原因有两个：第一，在很大程度上社会科学，特别是经济学在识别因果上已经发展出非常成熟的范式。除非能够带来颠覆性的边际贡献，研究没有理由抛弃传统因果识别方法。本文认为目前一个较为务实的做法是将机器学习识别因果的相关证据作为稳健性检验方式放入原有框架。第二，充分发挥机器学习的预测能力依赖于海量数据，当前经济学研究的样本量远没有达到能够让其获得精准预测的下限①。

机器学习使得研究者获得了以前通过人工投入无法获得的海量数据，检验了一些依靠传统方法无法有效论证的假设，这在一定程度上拓展了经济学研究的边界。本文相信未来几年会有越来越多的学者在研究中尝试机器学习技术。但学者也必须对该技术在应用过程中可能带来的问题有清醒的认识，这主要涉及学者间不平等及数据可复制性问题。机器学习依赖海量数据，这些数据的产生者主要来自业界和政府组织。可以想象，获得这些数据的主要方式并不是团队劳动投入，而是学者通过个人和组织的网络关系获得使用许可。这无疑给大部分学者设置了进入障碍，进而造成赢者通吃并可能加剧学界内部的不平等。机器学习带来的另外一个问题是研究的可复制性。学者通过公布数据及程序代码可以让其他学者和学生复制论文结论。但基于大数据的研究，学者虽然可以公布代码，但数据的公开必须获得数据提供方的许可。和一般数据相比，业界和政府可能更不情愿公布这些海量数据。这可能导致基于大数据研究的可复制性降低。本文对此的建议是，学者在获得数据的同时一并争取获得在未来公布其中的若干部分（比如数据量的万分之一）的权利：随机取样的子样本依然具有重复复制的价值。

参 考 文 献

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

[2] 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.

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

[4] Antweiler W.，Frank M. Z.，2004，*Is All That Talk Just Noise*？*The Information Content of Internet Stock Message Boards* [J]，Journal of Finance，59 (3)，1259～1294.

① 2017 年《美国经济学评论》（*American Economic Review*）和《经济研究》实证研究样本量的中位数分别是 16068 和 4557。

[5] Armstrong T. B., Kolesár M., 2018, *Optimal Inference in a Class of Regression Models* [J], Econometrica, 86 (2), 655～683.

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

[7] Athey S., 2019, *The Impact of Machine Learning on Economics* [A], In Agrawal A. K., Gans J., Goldfarb A. (eds.), *The Economics of Artificial Intelligence: An Agenda* [C], Chicago: University of Chicago Press.

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

[9] Bau D., Zhou B., Khosla A., Oliva A., Torralba A., 2017, *Network Dissection: Quantifying Interpretability of Deep Visual Representations* [A], The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C], 6541～6549.

[10] 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.

[11] Bishop C. M., 2006, *Pattern Recognition and Machine Learning* [M], New York: Springer.

[12] Bleakley H., Ferrie J., 2016, *Shocking Behavior: Random Wealth in antebellum Georgia and Human Capital across Generations* [J], Quarterly Journal of Economics, 131 (3), 1455～1495.

[13] Blei D. M., Ng A. Y., Jordan M. I., 2003, *Latent Dirichlet Allocation* [J], Journal of Machine Learning Research, 3 (4～5), 993～1022.

[14] Blum A. L., Langley P., 1997, *Selection of Relevant Features and Examples in Machine Learning* [J], Artificial Intelligence, 97 (1～2), 245～271.

[15] Blumenstock J., Cadamuro G., On R., 2015, *Predicting Poverty and Wealth from Mobile Phone Metadata* [J], Science, 350 (6264), 1073～1076.

[16] Bound J., Jaeger D. A., Baker R. M., 1995, *Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak* [J], Journal of the American Statistical Association, 90 (430), 443～450.

[17] Camerer C. F., 2019, *Artificial Intelligence and Behavioral Economics* [A], In Agrawal A. K., Gans J., Goldfarb A. (eds.), *The Economics of Artificial Intelligence: An Agenda* [C], Chicago: University of Chicago Press.

[18] Cao Y., Chen S., 2018, *Bad Romance* [R], Fudan University Working Paper.

[19] Card D., 1990, *The Impact of the Mariel Boatlift on the Miami Labor Market* [J], ILR Review, 43 (2), 245～257.

[20] Card D., 1999, *The Causal Effect of Education on Earnings* [A], In Card D., Ashenfelter O. (eds.), *Handbook of Labor Economics* [C], Amsterdam: North-Holland.

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

[22] Chalfin A., Danieli O., Hillis A., Jelveh Z., Luca M., Ludwig J., Mullainathan S., 2016, *Productivity and Selection of Human Capital with Machine Learning* [J], American Economic Review, 106 (5), 124～127.

[23] Chao J. C., Swanson N. R., 2005, *Consistent Estimation with a Large Number of Weak Instruments* [J], Econometrica, 73 (5), 1673～1692.

[24] Cicala S., 2017, *Imperfect Markets versus Imperfect Regulation in U. S. Electricity Generation* [R], NBER Working Paper, No. 23053.

[25] Dash M., Liu H., 1997, *Feature Selection for Classification* [J], Intelligent Data Analysis, 1 (1～4), 131～156.

[26] Doudchenko N., Imbens G. W., 2016, *Balancing, Regression, Difference-In-Differences and*

Synthetic Control Methods: *A Synthesis* [R], NBER Working Paper, No. 22791.

[27] Edelman B., Luca M., Svirsky D., 2017, *Racial Discrimination in the Sharing Economy*: *Evidence from a Field Experiment* [J], American Economic Journal: Applied Economics, 9 (2), 1～22.

[28] Engstrom R., Hersh J., Newhouse D., 2017, *Poverty from Space*: *Using High Resolution Satellite Imagery for Estimating Economic Well-Being and Geographic Targeting* [R], World Bank Policy Research Working Paper, No. 8284.

[29] Gechter M., Tsivanidis N., 2018, *The Welfare Consequences of Formalizing Developing Country Cities*: *Evidence from the Mumbai Mills Redevelopment* [R], Dartmouth College Working Paper,

[30] Gentzkow M., Shapiro J. M., 2010, *What Drives Media Slant*? *Evidence from U. S. Daily Newspapers* [J], Econometrica, 78 (1), 35～71.

[31] Gentzkow M., Shapiro J. M., Taddy M., 2019, *Measuring Group Differences in High-Dimensional Choices*: *Method and Application to Congressional Speech* [J], Econometrica, 87 (4), 1307～1340.

[32] Glaeser E. L., Kim H., Luca M., 2017, *Nowcasting the Local Economy*: *Using Yelp Data to Measure Economic Activity* [R], Harvard Business School NOM Unit Working Paper, No. 18-022.

[33] Goel S., Rao J. M., Shroff R., 2016, *Precinct or Prejudice*? *Understanding Racial Disparities in New York City's Stop-and-Frisk Policy* [J], Annals of Applied Statistics, 10 (1), 365～394.

[34] Goodfellow I., Bengio Y., Courville A., 2016, *Deep Learning* [M], Cambridge: MIT Press.

[35] Hansen C., Kozbur D., 2014, *Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE* [J], Journal of Econometrics, 182 (2), 290～308.

[36] Hansen S., McMahon M., Prat A., 2018, *Transparency and Deliberation Within the FOMC*: *A Computational Linguistics Approach* [J], Quarterly Journal of Economics, 133 (2), 801～870.

[37] Hartford J., Lewis G., Leyton-Brown K., Taddy M., 2017, *Deep IV*: *A Flexible Approach for Counterfactual Prediction* [R], Proceedings of the 34th International Conference on Machine Learning (PMLR).

[38] Hastie T., Tibshirani R., Friedman J., 2016, *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction* [M], 2nd Edition, New York: Springer.

[39] Henderson J. V., Storeygard A., Weil D. N., 2012, *Measuring Economic Growth from Outer Space* [J], American Economic Review, 102 (2), 994～1028.

[40] Hills T., Proto E., Sgroi D., Seresinhe C. I., 2019, *Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books* [J], Nature Human Behaviour, 3, 1271～1275.

[41] Hoberg G., Phillips G., 2016, *Text-Based Network Industries and Endogenous Product Differentiation* [J], Journal of Political Economy, 124 (5), 1423～1465.

[42] Huff C., Kertzer J. D., 2018, *How the Public Defines Terrorism* [J], American Journal of Political Science, 62 (1), 55～71.

[43] Iaria A., Schwarz C., Waldinger F., 2018, *Frontier Knowledge and Scientific Production*: *Evidence from the Collapse of International Science* [J], Quarterly Journal of Economics, 133 (2), 927～991.

[44] Imbens G. W., Lemieux T., 2008, *Regression Discontinuity Designs*: *A Guide to Practice* [J], Journal of Econometrics, 142 (2), 615～635.

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

[46] Jean N., Burke M., Xie M., Davis W. M., Lobell D. B., Ermon S., 2016, *Combining Satellite Imagery and Machine Learning to Predict Poverty* [J], Science, 353 (6301), 790～794.

[47] King G., Pan J., Roberts M. E., 2017, *How the Chinese Government Fabricates Social Media Posts for Strategic Distraction*, *Not Engaged Argument* [J], American Political Science Review, 111 (3), 484～501.

[48] Kleinberg J., Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S., 2017, *Human Decisions*

and Machine Predictions [J], Quarterly Journal of Economics, 133 (1), 237～293.

[49] Kohavi R., John G. H., 1997, *Wrappers for Feature Subset Selection* [J], Artificial Intelligence, 97 (1～2), 273～324.

[50] Kolesár M., Rothe C., 2018, *Inference in Regression Discontinuity Designs with a Discrete Running Variable* [J], American Economic Review, 108 (8), 2277～2304.

[51] Laver M., Benoit K., Garry J., 2003, *Extracting Policy Positions from Political Texts Using Words as Data* [J], American Political Science Review, 97 (2), 311～331.

[52] Mitchell T. M., 1997, *Machine Learning* [M], New York: McGraw Hill.

[53] Mueller H. F., Rauh C., 2018, *Reading between the Lines: Prediction of Political Violence Using Newspaper Text* [J], American Political Science Review, 112 (2), 358～375.

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

[55] Murphy K. P., 2012, *Machine Learning: A Probabilistic Perspective* [M], Cambridge: MIT Press.

[56] Neyman J., 1923, *Sur Les Applications de La Théorie Des Probabilités Aux Experiences Agricoles: Essai Des Principes* [J], Roczniki Nauk Rolniczych, 10, 1～51.

[57] Oster E., 2018, *Diabetes and Diet: Purchasing Behavior Change in Response to Health Information* [J], American Economic Journal: Applied Economics, 10 (4), 308～348.

[58] Peri G., Yasenov V., 2019, *The Labor Market Effects of a Refugee Wave: Synthetic Control Method Meets the Mariel Boatlift* [J], Journal of Human Resources, 54 (2), 267～309.

[59] Qin B., Strömberg D., Wu Y., 2017, *Why Does China Allow Freer Social Media? Protests versus Surveillance and Propaganda* [J], Journal of Economic Perspectives, 31 (1), 117～140.

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

[61] Staiger D., Stock J. H., 1997, *Instrumental Variables Regression with Weak Instruments* [J], Econometrica, 65 (3), 557～586.

[62] Theobald C. M., 1974, *Generalizations of Mean Square Error Applied to Ridge Regression* [J], Journal of the Royal Statistical Society. Series B (Methodological), 36 (1), 103～106.

[63] van Wieringen W. N., 2015. *Lecture Notes on Ridge Regression* [R], arXiv: 1509. 09169.

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

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

[66] Wu A., Forthcoming, *Gender Bias in Rumors Among Professionals: An Identity-based Interpretation* [J], Review of Economics and Statistics.

[67] Zheng S., Wang J., Sun C., Zhang X., Kahn M. E., 2019, *Air Pollution Lowers Chinese Urbanites' Expressed Happiness on Social Media* [J], Nature Human Behaviour, 3, 237～243.

[68] **方匡南、吴见彬、朱建平、谢邦昌：《信贷信息不对称下的信用卡信用风险研究》[J]，《经济研究》2010 年第 S1 期。**

[69] **方意：《主板与中小板、创业板市场之间的非线性研究："市场分割"抑或"危机传染"?》[J]，《经济学（季刊）》2016 年第 1 期。**

[70] **郭英见、吴冲：《基于信息融合的商业银行信用风险评估模型研究》[J]，《金融研究》2009 年第 1 期。**

[71] **李静：《文化创意产业与乡村旅游产业的融合发展研究》[J]，《管理世界》2017 年第 6 期。**

[72] **刘键、戴俭：《我国工业设计产业竞争力分析与发展对策研究》[J]，《管理世界》2017 年第 5 期。**

[73] **刘涛雄、徐晓飞：《互联网搜索行为能帮助我们预测宏观经济吗?》[J]，《经济研究》2015 年第 12 期。**

[74] **吕劲松、王志成、隋学深、徐权：《基于数据挖掘的商业银行对公信贷资产质量审计研究》[J]，**

《金融研究》2016 年第 7 期。

[75] 马晓君：《基于数据挖掘的新标准客户信用风险管理规则的构建——以央企中航国际钢铁贸易公司为例》[J]，《管理世界》2015 年第 3 期。

[76] 钱争鸣、李海波、于艳萍：《个人住房按揭贷款违约风险研究》[J]，《经济研究》2010 年第 S1 期。

[77] 沈艺峰、王夫乐、黄娟娟、纪荣嵘：《高管之“人”的先天特征在 IPO 市场中起作用吗？》[J]，《管理世界》2017 年第 9 期。

[78] 苏治、卢曼、李德轩：《深度学习的金融实证应用：动态、贡献与展望》[J]，《金融研究》2017 年第 5 期。

[79] 孙毅、吕本富、陈航、薛添：《大数据视角的通胀预期测度与应用研究》[J]，《管理世界》2014 年第 4 期。

[80] 徐晓萍、马文杰：《非上市中小企业贷款违约率的定量分析——基于判别分析法和决策树模型的分析》[J]，《金融研究》2011 年第 3 期。

[81] 于焕杰、杜子芳：《基于随机森林的企业监管方法研究》[J]，《管理世界》2017 年第 9 期。

[82] 张自然：《寿险公司的财务评估——基于主成分分析的 RIDIT 方法》[J]，《管理世界》2016 年第 3 期。

Machine Learning in Economics Research: Review and Prospective

Wang Fang[1]　Wang Xuanyi[2]　Chen Shuo[2]

(1. School of Public Administration, East China Normal University;
2. School of Economics, Fudan University)

Research Objectives: With the increasing availability of data and advances in computer technology, machine learning (ML) has been applied in economics studies and develops quickly. This article introduces the applications of ML in economics research systematically. **Research Methods**: Following a definition of ML, we introduce its applications in three aspects: data generating process, prediction, and causal inference (DID, RD and IV). **Research Findings**: We believe ML can expand the scope of social science research, but will not shift its paradigm. **Research Innovations**: This paper reviews the latest applications of ML in economics research. **Research Value**: We put applications of ML into three categories, and suggest some potential applications of ML in the future studies. In addition, the last section of the article discusses possible consequences of ML such as the inequality in academia and the research reproducibility.

Key Words: Machine Learning; Data Generating Process; Prediction; Causal Inference

JEL Classification: C19; C55; C80

（责任编辑：焦云霞）