机器学习在社会科学实证研究中的应用*

以中国教育追踪调查数据为例

宋骁†

2020 年 5 月 22 日

摘要

作为社会科学定量研究者使用的主流方法，广义线性模型存在很多局限性和应用误区。比如，模型的不确定性使得研究者难以得到稳健的研究结果；基于 P 值和假设检验的决策方式容易导致学术不规范问题；模型驱动的研究范式将社会事实过度抽象，它难以刻划现实世界的复杂特性，同时无法很好地对新数据进行预测。相比于传统定量模型，机器学习算法能够更好地拟合数据、处理高维变量并挖掘出数据潜在的信息。常常被忽略的是，机器学习算法具有优良的可解释性，它们同样能够揭示变量间复杂的作用关系。这使得机器学习算法可以成为社会科学研究的重要工具。作者以中国教育追踪调查基线数据为例，使用传统方法进行回归分析，并通过重抽样法展示广义线性模型的不确定性。其次，通过比较逐步回归法和 LASSO 回归，展示 LASSO 如何通过系数的收缩来进行变量选择。随后，作者使用 Xgboost 集成学习算法计算变量重要性系数，展示机器学习算法如何解释变量机制。最后，统一使用 RMSE 和 R^2 评价每一种模型的表现，发现机器学习算法的泛化能力高于广义线性模型。在结论中，作者对机器学习在社会科学中的应用做出了未来展望，并认为社会科学研究者应当增强 “可重复性研究” 的共识。

关键词：机器学习，统计显著性，定量研究，社会科学实证研究

†个人主页: https://xsong.ltd ， Email: xsong@stu.ecnu.edu.cn。

Abstract

As a mainstream method used by quantitative researchers in social sciences, generalized linear models have many limitations and misunderstandings. For example, the uncertainty of the model makes it difficult for researchers to obtain robust research results; decision-making methods based on P-values and hypothesis testing are likely to lead to academic irregularities; model-driven research paradigms overly abstract social facts, and it is difficult to characterize the real world The complex characteristics of the data, and cannot predict the new data well. Compared with traditional quantitative models, machine learning algorithms can better fit the data, process high-dimensional variables, and mine the potential information of the data. It is often overlooked that machine learning algorithms have excellent interpretability, and they can also reveal complex interactions among variables. This makes machine learning algorithms an important tool for social science research. The author takes the baseline data of the China Education Panel Survey as an example, uses traditional methods for regression analysis, and demonstrates the uncertainty of the generalized linear model through resampling. Secondly, by comparing the stepwise regression method with LASSO regression, it shows how LASSO uses the contraction of coefficients to select variables. Subsequently, the author uses the Xgboost integrated learning algorithm to calculate the variable importance coefficients, showing how machine learning algorithms can explain variable mechanisms. Finally, using RMSE and R^2 to evaluate the performance of each model, it is found that the generalization ability of the machine learning algorithm is higher than that of the generalized linear model. In the conclusion, the author has made a future outlook on the application of machine learning in social sciences, and believes that social science researchers should enhance the consensus of reproducible research.

Keywords: Machine Learning, Statistical Significance, Quantitative Research, Empirical Research

目录

1 研究背景

社会本身的复杂性导致研究者们将面对高维[1]、非线性和带有交互作用的数据，为了定量地刻划这些复杂的关系，社会科学自发源起就将统计学作为主要研究工具之一。尽管定量方法带有一定局限性 (即无法得出 “放之四海而皆准” 的规律)，统计学方法仍然能够通过组间差异和组内差异的统计信息描述复杂社会的变异性 (谢宇2012)，为研究者提供可靠的经验证据。

单就社会学而论，统计工具在不同的时代有不同的流行趋势。1960 年之前，交互表分析与对数线性模型是定量分析的主流；这之后，基于广义线性模型的分析大行其道，在它的基础上又出现了生存分析、结构方程模型、因果推断方法 (如工具变量法) 的应用 (Raftery 2001)。目前，这些模型在定量研究中占据着主流地位。统计方法的应用并非一成不变，它应该跟随技术和时代发展。在计算机软件和编程语言工具的快速发展背景下，社会科学应当顺延潮流，主动更新自己所采用的研究工具。

随着广义线性模型的大规模使用，它的缺陷也逐渐暴露出来。模型的不确定性就是其中一例。其次，广义线性模型将问题进行了大幅抽象和简化，这导致相当多的数据信息被统计模型所抛弃，使得研究者难以从数据中获得有价值的未知信息。此外，基于 P 值的假设检验模式也开始为人所诟病。随着 P 值和假设检验成为实证研究的 “金标准”，一些统计学家认为相当多的学科存在 “P 值滥用” 的情况 (Nuzzo 2014)。社会科学无法避免过度使用 P 值和假设检验范式的指责，这给社会科学定量研究的可靠性带来了前所未有的挑战。

在这样的背景下，本文提出机器学习算法[2]能够有效地结合传统定量统计工具。机器学习模型能够处理高维与非线性特征的数据，并且能够更好地拟合数据。本文还对机器学习在社会科学中的应用做出了文献回顾，并对其核心思想进行解释。作者认为，机器学习能够成为传统定量方法的补充。Susan Athey(2018) 和陈硕 (2018) 从不同角度分别讨论了机器学习在计量经济学的应用。本文在它们的基础上进一步讨论传统定量方法的不足。本文基于中国教育追踪调查数据，分别使用传统定量方法和机器学习进行分析，并

[1]本研究所称 “高维统计问题” 指的当数据包含较多自变量 (>50) 时所带来的问题。但在严格意义上，高维数据指变量数多于样本数的数据。读者应注意区分二者的不同。

[2]本文 “机器学习” 主要指监督学习 (Supervised Learning)，即对有标注 (因变量) 数据进行分类或回归。

比较这些方法的优劣。

2 传统定量方法与机器学习方法

社会科学研究大量应用了统计学方法，在社会学、经济学、心理学和教育学研究中，统计学与具体学科的结合成为所谓的“定量研究”、“计量方法”等分支领域。传统社会科学定量研究通常使用广义线性模型 (Generalize Linear Model, GLM) 作为主要研究工具[1]。GLM 因为优良的可解释性与基于 P 值的假设检验决策，成为社会学研究的常用方法。由于社会学问题的复杂性，研究者也会使用其他模型对其进行改进，如基于面板数据的模型、多层线性模型等。它们都隶属于“广义线性模型”的范畴。但是这些方法存在一定的问题，包括模型的不确定性，对高维数据 (变量数 >50) 的处理能力不足等。

然而，随着计算机技术的发展，从 20 世纪 80 年代开始，一系列新兴的方法涌现，由于它们能够根据固定的程序改进自己的性能，因此被称为“机器学习”方法。它们对训练样例的归纳能力十分出色，在学术界和工业界得到了广泛应用。但社会科学定量研究仍然处在发展初期，并未及时吸纳和应用机器学习方法。在相当多的定量研究中，假设的验证往往停留在基于 P 值的统计显著性推断层面，而统计有效性问题常被忽略，下面将具体论证社会科学定量研究的误区，以及机器学习方法何以成为传统研究工具的补充。

2.1 广义线性模型的不稳定性

社会科学家非常重视定量研究中的“内生性问题”，并且将刻划社会科学的因果关系、解释因果机制为己任。经典计量经济学教材作者 Angrist 和 Pischke(2008) 非常重视自然实验与反事实框架，认为社会科学的首要目的是分析因果关系。同样，因果分析的必要性在于它能够更好地对未来事件进行预测。本质上，因果推断问题也是预测问题，只不过它涉及到对不可观测的反事实情况进行预测[2]。

[1]本文说称广义线性模型包含线性回归和使用联结函数对线性回归的泛化，如 Logistic 回归；同时包括处理生存数据的 Cox 回归、处理嵌套数据的混合效应模型等等。

[2]如在是否上大学对收入的因果关系的经典问题中，感兴趣的因果关系通常为大学生群体的收入减去他不上大学的收入。而不可能观测到大学生不上大学的反事实情况，因此因果推断实际上是在预测反事实情况下的因变量。

GLM 的可解释性与估计的无偏性非常适用于分析因果效应。然而，传统的 GLM 具有不确定性。在相同拟合程度度量下，被纳入的变量集容易产生震荡；数据的轻微扰动容易造成 GLM 结果迥异；不同的变量组合导致估计系数变化很大甚至方向相反。Leo Breiman(1996, 2001) 将这种现象称为 “罗生门效应”，他使用 GLM 模型从 30 个变量的数据集中选取 5 个最佳变量。当测试误差取相同水平时，有 3 种包含 5 个变量的组合可供选择。而这些变量的系数不同，每一种组合都反映着不同的变量关系机制。Draper(1995) 也指出统计模型存在 2 种不确定性:

- 参数的不确定性
- 模型形式的不确定性

前者指模型估计出的参数往往会在一定的区间内变动，后者指研究者根据不同的理论和个人的判断往往会提出很多备选模型。社会科学研究者往往通过理论甚至直观的方式进行选择，这给予研究者较大的随意性。研究者容易倾向于 “模型套理论”、“模型套假设”，选择更易解释、更符合研究假设的模型。这样做的严重后果是研究结论的不可靠与不可重复。而在机器学习术语中，这被称作 “特征选择”(Feature Selection) 问题。机器学习发展出了很多方法对候选进行选择并且根据模型有效性进行决策，这显然要优于随意性较强的传统定量方法。

而从预测准确的角度来说，基于 GLM 的因果推断方法并不能保证对因变量的准确拟合。虽然 GLM 声称能够计算干预效应的无偏估计量，但它仍然缺乏统计有效性。因为大部分 GLM 并没有经过交叉验证和测试集的检验。社会科学研究者 (如胡安宁 (2016)) 同样注意到了 GLM 的这一特性并指出了一些可能的解决方案 (如倾向值方法)。但作者认为机器学习方法能够更好地解决模型的不确定性问题。事实上，机器学习正是为准确分类和预测的需求所设计，大量的实践也证明机器学习算法的预测能力比传统统计模型具有更优秀的预测性能。后文中给出了 GLM 与机器学习有效性的比较。

GLM 假设变量的线性关系，如以下形式的线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

它将 y 设置为多个 x 乘以对应的系数并相加的模式，如果真实的 x 和 y 的关系不是线性的，模型的可靠性就会大打折扣。因为它假定数据服从于特定统计分布而且变量之间的关系是线性的，这些假定往往在现实世界中很难满足。当模型不能很好地拟合数据时，那么通过线性模型得到的结论并不可靠。

谢宇 (2013) 的教材中指出社会学定量模型应该遵守 “奥卡姆剃刀” 原则，即如果模型的解释能力一致，除非有很强的证据，优先选择最简单的模型。但 Leo Breiman(2001) 曾指出，数据模型的简约性与准确性相悖，即使 GLM 回归能够清晰地解释 y 与 x 之间的关系，由于模型的不稳定性和较差的泛化能力，GLM 的有效性与可靠性仍然是一个问题。为社会科学定量学者所重视的无偏性估计和因果效应如果没有有效性的保证，则成为一句空谈。

应该相信数据、模型还是理论？这是困扰定量社会学研究者的问题。谢宇 (2012) 认为社会科学研究者并不应当关注模型是否真实，而更应该关注模型是否符合社会学理论与人们对社会现象的经验理解。这意味着研究者应当通过理论与理解选择模型，而不是传统 GLM 的模型驱动范式。一般而言，社会科学研究者拒绝模型驱动的研究范式，认为模型驱动研究方法应该受到具体领域知识的领导，不应该盲目崇拜模型 (Angrist and Pischke 2008; 谢宇2012)。在这一点，数据驱动方法和社会科学研究者们是一致的。因为和模型驱动方法不同，数据驱动方法从数据本身出发，不对数据做先验假定；而是尽可能利用数据中的信息，挖掘数据中的信息，最终把解释的权利留给研究者。但数据科学知识普及的缺乏导致数据驱动范式较少被研究者所使用，它们无法放弃长期以来使用 GLM 的传统，社会科学定量研究观念的改变仍然是一个漫长而艰巨的任务。

2.2 P 值与假设检验的误用

P 值已经成了最被广泛使用的统计工具之一。不仅仅在社会学领域，P 值与假设检验方法在经济学、心理学乃至自然科学都有着极为广泛的应用。P 值的定义是假设零假设正确时，利用观测数据得到和零假设结果更为极端的概率。当这个值越小时，说明零假设越有可能是错误的。而 P 值最初并没有包含零假设、显著性水平、两类错误的概念，它最初并未用来进行假设检验 (Nuzzo 2014; 于淼2017a)。它经常的使用误区在于：

- 置信水平和 P 值这两个概念经常被混淆
- 假设检验只能拒绝掉一个假设而不能证明一个假设成立
- P<0.05 常常被当做统计显著的 “金标准”
- 假设检验使得统计结果成为了非此即彼的 “是-否决策”

研究中，置信水平往往是人为规定的，P 值则通过计算得出。其次，研究者在得出结果时往常常用词不当，比如，“无法拒绝某假设” 被说成 “接受某假设”。此外，0.05 这个阈值设置具有非常大的主观性。这些原因导致研究者，直接用这套工具决策某种现象是否存在，并依此得到研究结果。统计过程不应成为一种 “是-否决策”。

Nuzzo(2014) 认为我们应该更关注干预效应的具体剂量值。因为即便一个系数统计上显著，它的实际效应也可能非常微弱。于淼 (2017b) 指出随着测量技术的提高和测量精度的上升，随着样本量的增大，干预效应几乎总会被识别，但同时它的差异可能微不足道。事实上，讨论两组样本值的差异是无意义的，在更高的精确度下，差异总是存在的 (Tukey 1991)。在有理论支持的情况下，这种分析方法并不能带来新知识[1]。此外，在现有的假设检验范式中，研究者混淆了零假设条件 H_0 下出现观测数据 D 的条件概率和观测值下零假设的概率。也就是假设 $p(D|H_0) = p(H_0|D)$。这导致假设检验方法无法告诉研究者真正感兴趣的信息 (Cohen 2016)。

假设检验范式同时为科研工作者带来 P 值作弊的问题，即尝试大量的模型直到通过显著性检验为止。同时，学术刊物也存在一定的发表歧视，即倾向于发表显著的结果[2] 。但这样的研究成果是不可靠并且不可重复的，因为研究者可能刻意追求能够造成显著的变量组合与模型。即使没有改变原有数据，也使得统计结果无法重复。置信水平和 P 值常常被研究者们混淆，置信水平 α 值往往是人为规定的，而 P 值则是计算得出。

P 值衡量的是在某假设下观测值出现的概率，它忽略了备择假设为真时它被接受的概率，即统计功效 $(1-\beta)$。在实际的研究中，如果零假设没有被拒绝 (P 值不够小)，很少

[1]作者常常听到人们对社会学定量研究 “总在证明常识” 的诟病，这应被归咎于对假设检验的滥用。因为人们根据 “常识” 提出的假设在精确的测量和特定的显著性水平下几乎总能被 “证明”。

[2]Brodeur(2016) 等人收集了经济学顶刊 AER，JPE 和 QJE 文章中的 49297 个 P 值，并发现 P 值残差呈现 “双峰分布”，他们认为除了期刊更倾向于发表 “统计显著” 和 “非常不显著” 的文章，研究者还会选择性地报告更为显著的结果。这份数据集连同其他几份研究数据集也显示，不仅是经济学，相当多的学科在 P=0.05 处存在密度集中。这些数据都被收录于 R 语言 `tidypvals` 包，详细说明参见：`https://github.com/jtleek/tidypvals`.

有研究者讨论零假设是错误的，却未被拒绝的情况。很多不显著的研究结果都可能存在这种可能并且值得被发表。统计功效分析在社会科学定量研究中几乎是一个空白。

P 值和假设检验方法存在很多的缺陷，目前并没有一种成熟的替代方案。作者认为机器学习方法能够很好地拓展并补充传统的 P 值决策。因为基于准确率和统计有效性的评价方式比 P 值更加直观易理解，而且避免了上文所述的缺陷，下面详述机器学习方法的特点。

2.3 机器学习算法的特点

传统社会科学定量方法面对的任务通常为因果推断 (以计量经济学为代表)，趋势分析 (以人口学为代表) 为代表。虽然大量实证研究者以研究因果关系为己任，但作者认为广义线性回归的传统方法限制了社会科学研究者的视野，并非所有社会科学研究都需要进行严谨的因果推断。至少探索性研究没有此需求，因为它们缺少足够的理论基础，而一个设计合理的探索性研究却能够为理论建设提供启发。机器学习算法隶属于数据驱动 (Data Driven) 的研究范式，能够摆脱研究假设的限制，刻划数据中隐藏的变量关系。

在 “奥卡姆剃刀” 的理念驱使下，定量研究总是尽量避免纳入高维变量，简洁的模型意味着较少的变量维度。当加入过多变量时，GLM 容易产生多重共线性问题，这时系数的标准误升高，判定系数 R^2 也会增加，体现为过拟合 (Wooldridge 2016)。一般而言，模型过拟合表现在因变量能够被完美地预测，一旦运用到新数据中，模型的表现会迅速下降。机器学习算法能够较好地处理高维变量问题，它能够在学习到数据集中大量变量信息的前提下，确保模型的泛化能力，避免过拟合问题 (Breiman and others 2001; 吴喜之2019)。

使用机器学习算法，研究者同样不用担心降维的问题。由于量纲的不同，线性回归的回归系数默认不可比较大小，除非研究者计算变量的标准化回归系数。但这种方法无法实现对多变量的降维，导致模型不够简洁。在进行效应可比性研究时，一个可行的解决办法是 Heise(1972) 的 “系数集束化” 方法，具体是在拟合线性回归之后，估计一系列自变量 x 的效应，使得多个自变量的效应合成少数几个 “主效应”，从而达到降维和相互比较的目的。机器学习算法 (如决策树、随机森林、Xgboost) 通过计算变量的重要性系数进行效应的比较，如果某 2 个效应对因变量作用类似，算法会自动降低其中一个变量的重

要性，因此机器学习能够自动完成降维的效果。本文第 4.2节给出了基于 CEPS 数据的示范。

如前所述，不仅仅是教育研究，社会科学中大多数都是高维统计问题。面对 50 个以上的变量，为了避免前文所述的高维度问题，需要进行变量选择。以基于线性模型的逐步回归法为例，它基于 P 值删除不显著的变量。而显著性检验的缺点前文已经论证。机器学习的另一种降维和变量选择的方式是纳入惩罚项。以 LASSO 回归为例，它使用了惩罚系数 λ 对不重要的变量进行惩罚，使得回归系数在原有线性回归基础上向 0 收缩。完全收缩为 0 的变量相当于被模型删除，在这个过程中 LASSO 回归完成了变量的选择。而 LASSO 变量选择都是通过算法一步步迭代优化完成的，所以能够更好地以数据驱动的方式得到结论。

相比于使用具体理论来挑选合适的模型，机器学习算法通过训练、调参和交叉验证的方法来得到最优的模型。本质上，算法训练和调参是一个模型选择的过程。因此这种数据分析范式可以被称作 “数据驱动”。算法训练常用的方法有 k 折交叉验证。交叉验证法对预测准确性的评价更为客观，因为它总是在用于拟合算法之外的数据子集中进行评价。本文接下来将给出具体例子进行阐述。

3 数据分析：学业成绩的影响因素

对于学习成绩的研究是社会科学领域的经典课题。和大多数社会学问题一样，学业成绩成因是高度复杂且非线性的 (谢军2018)。相当多的社会学研究发现，家庭背景对于学业成就有着非常大的影响 (方超，黄斌2018)，它甚至超过了学校层次的作用。但是，家庭在教育的不同阶段有着不同的重要性。例如：随着教育阶段的上升，从小学到大学，家庭因素在进入下一级教育阶段的作用逐渐减弱。家庭社会地位对升学的影响逐渐被学校等级所取代 (唐俊超2015)。

谢军 (2018) 发现，在小学阶段，学生的学习成绩和家庭因素极为相关，母亲的教育水平是学习成绩的第一解释因素。在初中阶段，校级差异对学生学习成绩的方差解释力度更大；约 22.66% 的学生成绩变异可被校际差异所解释。此外，和贫寒家庭相比，经济条件更好的家庭也更可能为子女提供更好的入学条件，上更好的学校 (李忠路，邱泽奇2016)。

程诚 (2017) 发现同伴社会资本对个人的学习成绩能够产生较大影响，无论是直接的还是间接的。大学生的人力资本积累能够受到同伴的学业能力的显著提升。网络资源能够对学生的学业成就产生影响。这种影响多发生在家庭阶层背景相同的同伴中间，而来自不同家庭背景的学生能够产生的影响非常小。因此，我们同时考虑学校和班级层面变量的对学习成绩的影响。

综上所述，在考虑教育成就问题时，必须同时考虑家庭背景、学校差异和朋辈影响。本质上，学业成绩是一个典型的高维、非线性、交互性统计问题。高维是指影响学业成绩的因素成千上万，没有单一因素能够决定；非线性是指影响因素和学业成绩的关系可能是多次方的；交互性是指一个因素对学业成绩的影响可能依赖于另一个因素的取值。正因为模型与方法的限制，以往研究无法比较家庭、学校、班级、个人禀赋等各因素之间的重要性。究竟哪些因素影响初中阶段的学习成绩？这些因素中哪些更为重要？为了突显机器学习算法的特点，本文将从预测效果的层面 (而不仅仅是显著性水平) 评估模型。

表 1: 数据字段解释

变量字段	含义	类型	变量字段	含义	类型
stdmat	数学标准分	连续	huko	孩子户口类型	分类
grade9	9 年级学生	二分类	eduy	最高教育年限	连续
sex	性别	二分类	dangy	党员	二分类
onechi	是否是独生子女	二分类	houspro	住房生产经营用	二分类
drunk	爸爸是否经常酗酒	二分类	classtm	总课时	连续
qurel	父母经常吵架	二分类	clpre	备课时间	连续
relation	父母之间关系很好	二分类	revitm	批改时间	连续
desk	家中有独立书桌	二分类	know	认识家长数	连续
net	家里有电脑网络	分类	subject	教授本班科目	分类
maedu	母亲教育水平	连续	drsmok	有抽烟喝酒学生	二分类
faedu	父亲教育水平	连续	commhr	与学生交流时间	连续
eduexp	父母教育期望	连续	schcsrm	教室数量	连续
dialect	家庭交流方言	分类	comno	学校电脑数	连续
chmwk	父母检查功课	有序因子	buget	生均财政拨款	连续
chcos	父母指导作业	有序因子	eduqua	教师资格证人数	连续
qianzi	检查作业频率	有序因子	teainc	高级教师年收入	连续
futcfd	对孩子未来的信心	有序因子	schtype	学校性质	二分类
dial	孩子交流方言	分类	fight	打架斗殴	二分类
chidia	和孩子交流的方言	分类	brkpb	破坏公物	二分类
eduyexp	对孩子的教育期望	连续	smok	吸烟	二分类
			drink	饮酒	二分类

3.1 数据介绍

本文使用中国教育追踪调查[1](China Education Panel Survey, CEPS)2013-2014 学年基线数据进行对比分析。CEPS 采用两阶段分层的概率比例抽样方法，从全国抽取 28 个县级行政区进行调查。CEPS 在每个抽中的县级行政区内分别抽取 4 所初中学校。并在每所入

[1]项目主页：https://ceps.ruc.edu.cn/

样学校中分别抽取包括 2 个初一班级和 2 个初三班级，共 4 个班级。CEPS 以学校为单位进行调查，共抽取 112 所学校和 438 个班级，所有被抽中班级的全体学生进入样本。约 20000 名学生在基线调查中被抽中作为入样样本。

本研究使用了学生回答数据、家长、班主任、校领导回答数据进行匹配，得到了丰富的信息。字段解释如表 1 所示。

CEPS 基线数据仅提供了期中考试语文、数学、英语三门成绩的原始分和标准分 (重构为均值为 70、标准差为 10 的标准成绩)。而 2014-2015 追访数据 (此时受访者均已升至八年级) 则提供了总分成绩。从追访数据来看。共有总分 100 分制、120 分制、130 分制、150 分制 4 种情况。通过追访数据无法前推基线数据的总分。作者从基线数据中探查发现，存在原始分相同的标准分不同。可能是 CEPS 数据提供者考虑总分差异问题。故本文采用标准成绩作为因变量[1]。从图 1 中也可以发现，标准化过后的成绩的学校间差异降低。学校间的数学成绩近似同一个统计分布。

从表 1 可以发现数据集中含有一些相似特征，如学生本人填答的 “家庭交流方言” 和家长填答的 “孩子交流方言”、“和孩子交流的方言”；学生填答的 “父亲、母亲教育年限” 和家长填答的 “父母最高教育年限”；学生填答的 “父母教育期望” 和家长填答的 “对孩子的教育期望” 等。这些变量由于填答方认知的不同导致不严格的变量共线性 (Colinearity)，它们可能会为数据集带来噪音。

然而，本文使用的 LASSO、随机森林、Xgboost 算法都能够进行特征选择，处理高维数据和多重共线性问题。共线性变量不会影响算法的预测效力。此外，很多有价值的信息可能会隐藏在这些变量的组合中。因此，本研究不会对数据集变量作出显式的特征选择，而将特征选择步骤交给模型自己处理。

图 2 展示了对数据的探索性分析，直观来看，父母对孩子的信心评级 (`futcfd`) 与数学成绩有强相关关系，得分为 3、4 的组有较多离群值。父母的教育期望 (`eduexp`) 也和子女的数学成绩有着正相关关系。更深层次的信息有待机器学习算法的进一步挖掘。

[1]为了处理缺失值问题，本文使用 R 语言的 `mice`(Buuren and Groothuis-Oudshoorn 2011) 包进行缺失值插补。`mice` 提供了对连续变量进行均值预测、对二分变量进行 Logistic 回归、对多分类变量和有序因子进行多分类 Logistic 回归的方法。最终的数据集已没有缺失值。

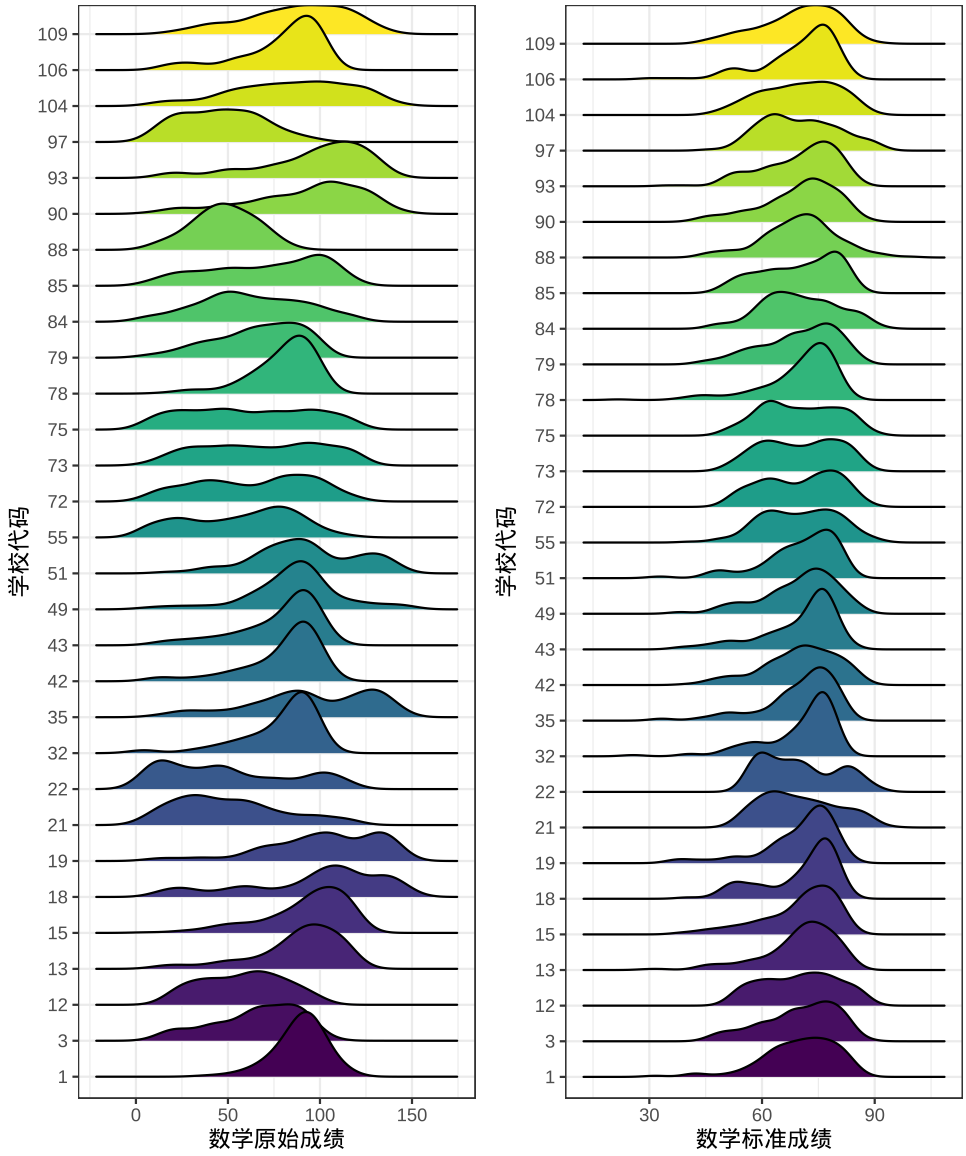

图 1: 随机抽样 30 所学校的数学成绩差异：纵轴为被抽中的学校代码，山丘形状为数学成绩的密度分布。左图是数学原始分布，右图是标准化分布。标准化后学校间差异变小。

3.2 算法介绍

为了对比传统定量方法与机器学习算法，本文使用基于 OLS 模型的逐步回归法来筛选变量。机器学习与之对应的方法是 LASSO 回归 (Tibshirani 1996)。LASSO 回归是对线性回归的补充，LASSO 回归通过牺牲参数估计的无偏性来换取有效性 (James et al. 2013)。它使用 L1 范数对线性回归系数进行 “惩罚”。被惩罚的变量将会向 0 值收缩，完全收缩为 0 值变量相当于被删除。变量系数收缩的幅度越大，说明它对于预测越不重要。通过这种方式，LASSO 实现了变量选择。在这个意义上，LASSO 也是一种逐步回归。LASSO 回归的优点是预测准确性更好，相比于逐步回归删除整个变量，LASSO 模型对于变量系数的收缩是连续的而不是一刀切的，它能够更好地保留变量信息。而机器学习中的 LASSO 回归则使用了惩罚项的方法对不重要的变量进行惩罚。LASSO 使用交叉验证的方法选择 λ 参数，促使 LASSO 能够在新的数据集上表现出更好的性能。

为了对教育成就的交互影响因素进行机制解释，本文将使用集成学习方法中的 Xgboost 算法。Xgboost 算法全名为极端梯度提升 (Extreme Gradient Boosting)，它是一种经过改进的 Booting 算法。Xgboost 算法组合 N 个基学习器 (通常是决策树) 提升数据拟合的效果；不同的基学习器通过有顺序的方式运行，即每一棵决策树学习上一棵树的残差。这样的话，每一轮训练都会更新一次目标函数，为了简化优化目标，Xgboost 通过二阶泰勒展开来近似目标函数。同时，Xgboost 通过计算每个叶节点的权重来得出每棵树的结构分数，它用来衡量树对于预测的改善作用。如果一棵决策树对预测的贡献不大，就禁止它继续分裂下去。此外，Xgboost 使用直方图算法近似分割特征，加快了运行速度，同时使得特征层面能够并行。Xgboost 还纳入了正则项防止算法对数据的过拟合。值得注意的是，它能够计算某一特征在每一棵决策树中的变量重要性的平均值，它衡量了各变量对预测的贡献程度，这是我们进行机制解释的重要方法。此外，本研究加入随机森林算法 (Breiman 2001) 进行更多的方法间比较。

对于不同的算法，需要一个统一的指标进行筛选和评价。评价回归问题的度量指标通常为误差均方根 (Root Mean Square Error, RMSE) 与 R^2，公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2}$$

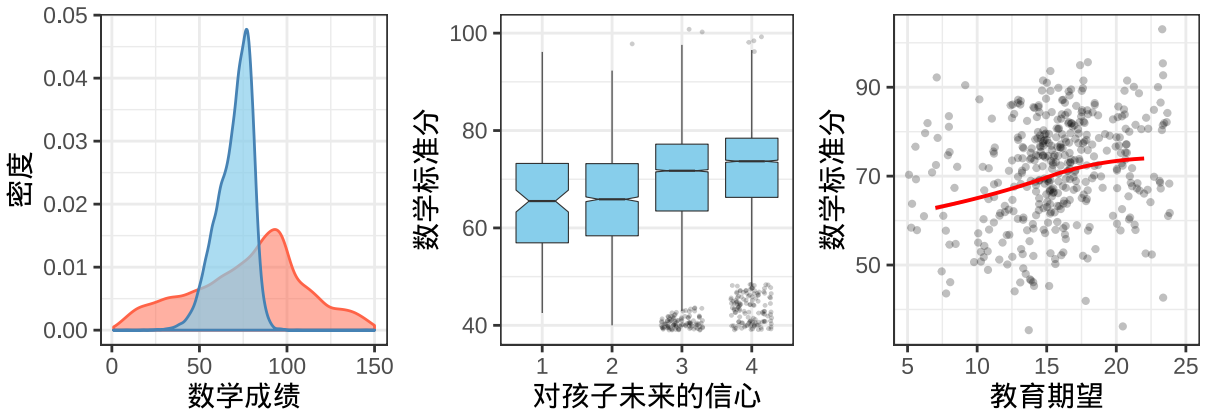

图 2: 探索性数据分析展示。左图：数学成绩的原始分和标准分。红色为原始分；蓝色为标准分。在之后的分析中离群点被重编码。大于 100 的标准分被重编码为 100；小于 40 的标准分被重编码为 40。中图：父母对孩子未来信心与数学成绩的关系。右图：父母对孩子教育期望 (年) 与数学成绩的关系。为防止图形堆叠，散点经过再抽样和抖动处理。

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

为了更客观地计算这些指标，本研究将使用 k 折交叉验证，它将数据集划分成 k 个相同大小的子样本，每次取出其中 1 份作为验证集，另外的 $k-1$ 份作为训练集，循环计算 RMSE 和 R^2，最终取 k 次计算的均值。本文表 3 展示了各算法的交叉验证和测试集的 RMSE 与 R^2。

4 分析结果

4.1 传统定量方法：广义线性模型

表 2 展示了传统定量研究范式的回归结果。其中，模型 1 和模型 2 均是线性回归结果，模型 3 和模型 4 为多层线性模型。一般而言，研究者会根据社会科学理论选择变量，将变量加入模型。模型 1 报告了未纳入学校哑变量的线性回归模型，模型 2 则在它的基础上加入了学校编码的哑变量 (系数省略)。多层线性模型则更好地刻划了数据的嵌套结构，如学生嵌套于班级，班级嵌套于学校。由于社会关系的濡染效应，相同层次内部的组具有相似性，高层次变量能够与低层次变量互相影响。因此在模型 3 中，作者设置了

随机截距效应，使各协变量的影响在不同学校中的斜率相同，截距不同。

在模型 1 中，可以看出母亲教育年限每增加 1 年，学生的标准数学成绩会下降约 0.017 分。而在加入了学校哑变量的模型 2 中，此系数又变为了 0.312，即母亲多受 1 年教育，学生的数据成绩将上升 0.312(P<0.01)。但在模型 3 和模型 4 中，母亲教育水平的系数变为了小于 1 的正数 (不显著)。为了进一步展示 GLM 的不稳定性，作者进行了一个简单的统计模拟，使用有放回抽样 (Bootstrap) 从 53 个自变量中抽取 10 个作为 “母亲教育年限” 的协变量，并且重复 25 次，每一次都得到一个回归模型[1]，计算每一个 “母亲教育年限” 的系数，图 3 展示了这 200 个系数值的大小，作者进行排序后发现，不同的变量组合造成了较大的系数变化，最大的系数达到了 1.0 附近，甚至有 13 个模型系数小于 0。这幅图直观地展示了社会科学定量研究中广泛存在的 “罗生门效应”。

[1]请注意在 25 个模型中均包含 “母亲教育年限” 变量

表 2: GLM 回归结果，因变量为数学标准成绩

	线性回归		多层线性模型	
	模型 1	模型 2	模型 3	模型 4
女性	−0.915***	−0.927***	−0.915***	−0.918***
	(0.135)	(0.136)	(0.135)	(0.135)
9 年级	0.626***	0.716***	0.666***	0.677***
	(0.136)	(0.137)	(0.137)	(0.136)
母亲教育年限	−0.017	0.312**	0.095	0.126
	(0.091)	(0.099)	(0.094)	(0.095)
父亲教育年限	0.396***	0.631***	0.469***	0.500***
	(0.093)	(0.096)	(0.094)	(0.095)
父母教育期望	2.757***	2.946***	2.834***	2.846***
	(0.070)	(0.072)	(0.071)	(0.071)
非农业户口	−0.881***	−0.513**	−0.780***	−0.724***
	(0.153)	(0.168)	(0.159)	(0.160)
其他户口	−1.548	−1.613	−1.616	−1.544
	(1.059)	(1.065)	(1.060)	(1.059)
党员家庭	0.168	0.347	0.251	0.262
	(0.221)	(0.225)	(0.222)	(0.223)
生均财政拨款				−0.143
				(0.100)
生均财政拨款 ×				−0.062
母亲教育水平				(0.073)
截距项	70.541***	68.336***	70.495***	70.512***
	(0.138)	(0.882)	(0.153)	(0.155)
随机效应 (标准差)				
学校 id			0.639	0.607
生均财政拨款				0.345
残差			9.403	9.399
样本量	19,487	19,487	19,487	19,487
学校数			112	112
校际差异解释力 (ICC)			0.063	0.059
R^2	0.086	0.096		
AIC			142,740.3	142,740.7
BIC			142,826.9	142,858.9

*p < .05; **p < .01; ***p < .001

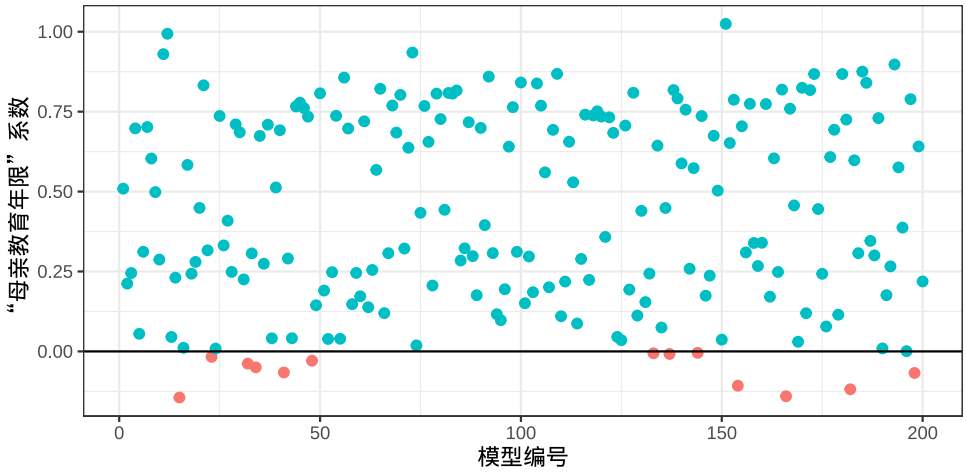

图 3: 使用有放回抽样随机抽取 10 个协变量后，通过这种方式抽取 200 次生成了 200 种不同的变量组合，使用这些自变量分别进行线性回归，散点展示了“母亲教育年限”系数的大小，小于 0 的系数用红点标出。

4.2 机器学习：LASSO 回归

LASSO 算法使用 K 折交叉验证的方法，找出 K 个验证集最小的 RMSE 对应的 λ 值。这时，算法包含的变量组合最为稳健，最能准确预测未知的新数据集。图 4 展示了 LASSO 回归的训练过程，系数估计值是 λ 的函数[1]。随着 λ 参数 (在图 4中取对数处理) 的不断增大，各个变量 x 数开始向 0 不断收缩，直到 λ 足够大时，所有系数均收缩为 0。那么，系数消失越晚说明变量的效应越强。因此，从图4中可以直观地看到，eduyexp(对孩子的教育期望) 收缩最晚，说明它对数学成绩的影响较大[2]。

图 5 展示了经过训练后 LASSO 回归分析的结果，红点对应的横坐标为各变量 LASSO 回归的系数。从结果中看，对于数学成绩有正向作用的变量为父母对未来的期望，父母指导作业和父母检查作业；对于数学成绩有负向作用的是孩子交流方言、家庭交流方言和班级内有抽烟喝酒的学生。具体而言，家长对孩子未来的期望值越高，孩子的学习成绩越好。有趣的是，在控制其他协变量后，父母每天检查作业反而会使孩子的数学成绩

[1]数据中的多分类变量在建模前被转换为哑变量。因此，图 4，5，6和表1 的变量名有所出入，是因为分类变量转换为哑变量后变为了变量名 + 类别的形式，如 smok 变量的 yes 类别转化为 smokyes。

[2]关于变量重要性和 LASSO 系数的理解，可参见作者在统计之都论坛的讨论：https://d.cosx.org/d/421033-lasso/7。

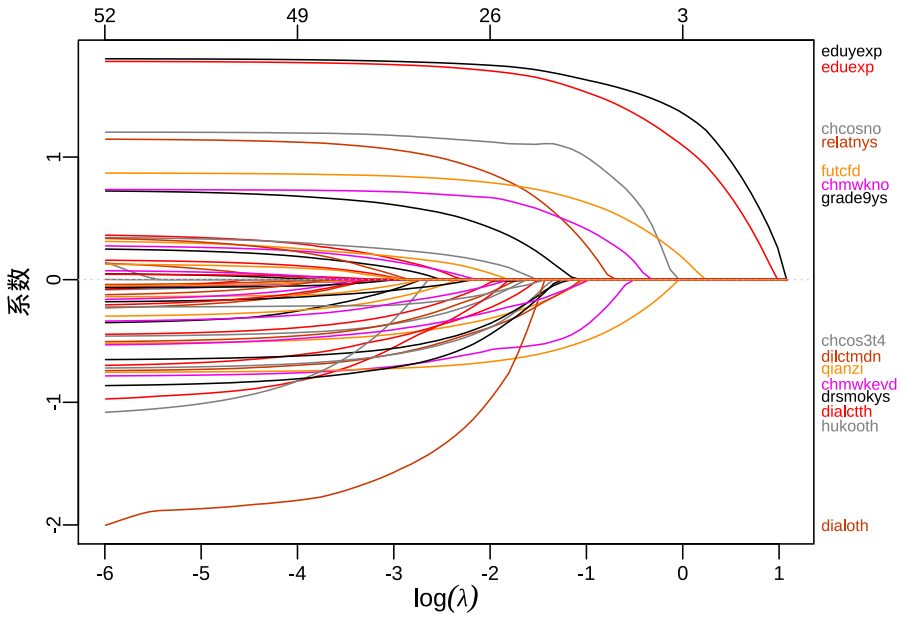

图 4: LASSO 回归系数收缩图。右侧单词为变量名标签；上方坐标轴为模型自由度，曲线颜色与变量对应。随着惩罚项 $\log(\lambda)$ 值不断增大，回归系数向 0 不断收缩。

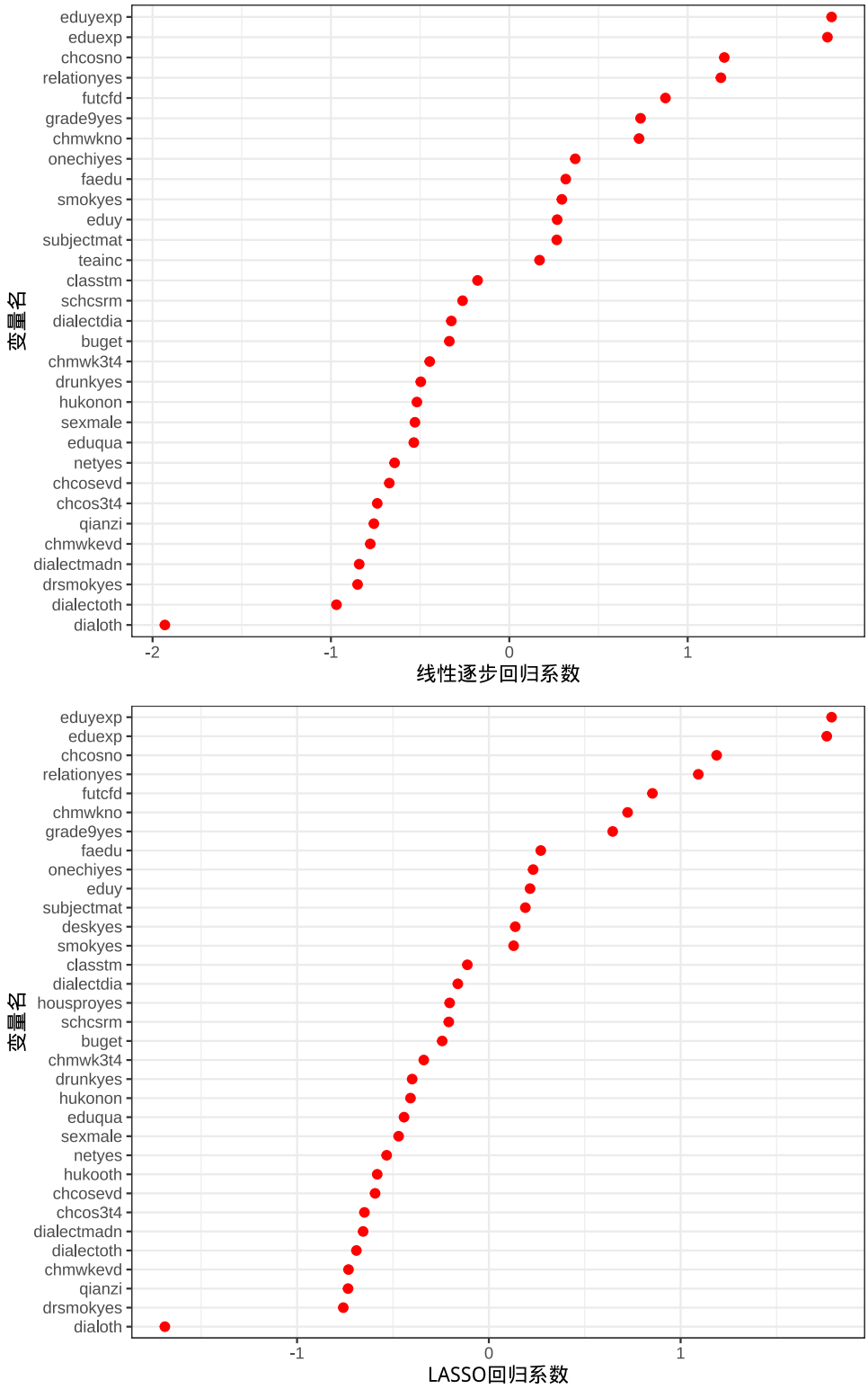

图 5: 上图：线性逐步回归系数。下图：LASSO 回归系数。过小的回归系数未在图中展示。二者对变量的选择略有差别。

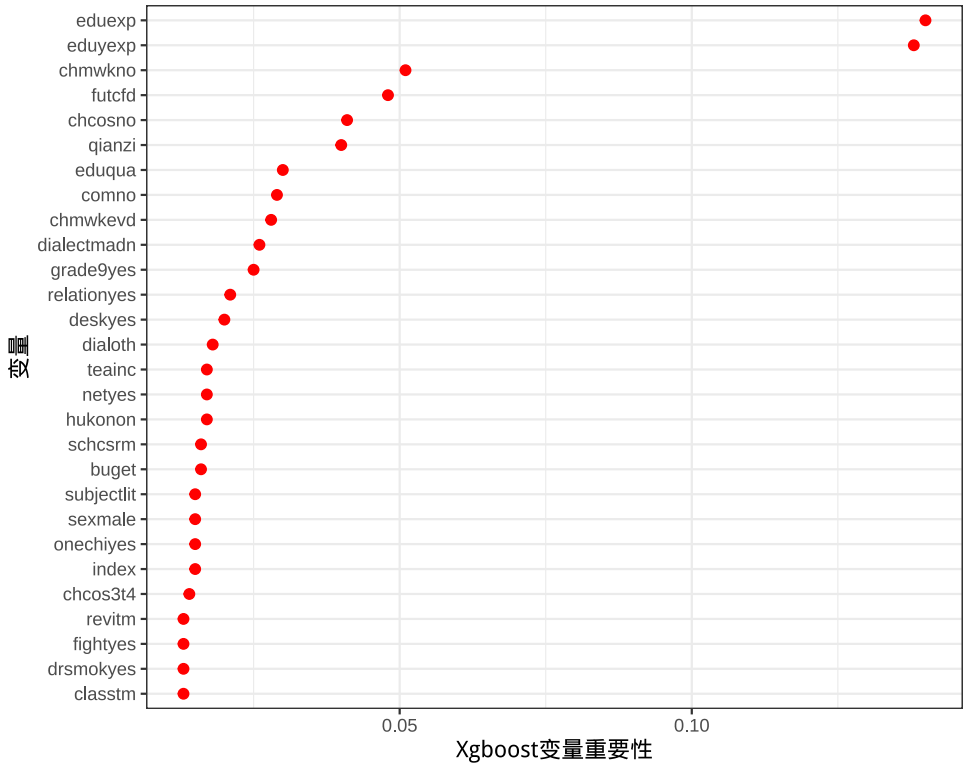

图 6: Xgboost 算法的变量重要性排序

下降，这可能因为孩子的学习习惯在初中阶段前就已建立。学习习惯较好的学生无需每天检查作业，学习习惯较差的学生促使家长不得不频繁检查作业，由于基础较差他们同时也无法得到较好的成绩。此外，在家中说其他语言 (非普通话也非方言) 的学生的数学成绩最差。如果学生所在的班级内有抽烟喝酒的同学，学生的数学成绩会受到影响。

4.3 机器学习：Xgboost 算法

Xgboost 使用特征分裂节点时的纯度 (取所有树的平均值) 来计算变量的重要性系数，它衡量每个自变量对于预测因变量的贡献程度。图 6 展示了使用 Xgboost 算法估计的变量重要性系数，和 LASSO、逐步回归的结果相近，父母的教育期望 (`eduexp`，`eduyexp`，`futctd`)、检查作业和辅导行为 (`chmwkno`，`chcos`，`qianzi`) 等系列变量的预测能力较强；其次重要的变量分别是教师资格证人数 (`eduqua`)、学校电脑数 (`comno`)。通过变量重要性系数，我们可以定量化地描述每个因素对数学成绩的预测能力，这是 GLM 分析方法所不具备的。

然而，图 6展示的统计结果无法体现严谨的因果关系。比如可能较好的成绩导致了更高的教育期望，而非本文结果所呈现的那样，教育期望直接导致更好的学习成绩。变量重要性系数并非无偏估计量，在 “相关关系不等于因果关系” 的认知下，本文乃至所有机器学习算法的结果必定会遭到质疑。但作者要指出的是，变量重要性的衡量的并不是 “相关关系” 而是 “预测关系”，即纳入某变量后多大程度上有助于预测因变量。不稳定、不可靠的 “因果关系” 和震荡程度较小的 “预测关系”，哪个更有价值、更有助于新知识、新的问题意识的发现和积累？哪个更有助于学术研究？研究者和社会大众需要在二者中做出抉择。作者认为这个问题并不是有答案的，而是需要根据问题的特殊性作出决策。

表 3 中的 10 折交叉验证法将训练集划分为 10 份，循环使用每 1 份作为验证集，剩下 9 份作为训练集进行评估，对每次的 R^2 和 RMSE 计算平均值得到上面的表格。从表 3 的准确性评估中我们可以看出。线性回归的交叉验证 R^2 高于测试集 R^2 ，说明线性回归在新数据中的预测能力不足。尽管 LASSO 回归在训练集交叉验证的表现不如线性回归，它在新数据集中的表现并未下滑，LASSO 算法迁移到测试集后 RMSE 的上升幅度也较低。以随机森林为代表的集成学习方法具有更好的预测能力，和线性回归、LASSO 相比，随机森林的各项指标均超越了前两者。而 Xgboost 的预测能力更是超过了随机森林。

表 3: 所有算法的准确性评估结果

算法	交叉验证 RMSE	测试集 RMSE	交叉验证 R^2	测试集 R^2
线性回归 (逐步回归法筛选)	8.991	9.143	0.158	0.153
LASSO	9.041	9.135	0.153	0.155
随机森林	9.026	8.905	0.157	0.197
Xgboost	8.868	8.769	0.185	0.221

5 结论与展望

社会科学中的因果机制常常是非线性、高维并且交互的。然而传统社会科学定量研究方法过度重视解释各变量相互影响的机制，忽视了模型不确定性、预测能力弱等问题。此外，P 值与假设检验导致研究的不规范和不透明。由于缺少交叉验证的方法，评价这

些模型的标准并不客观。这些原因使得很多社会科学定量研究失去了可靠性与可重复性。本文使用以预测精度为衡量指标机器学习模型研究数学成绩的影响因素，并根据精度最高的模型进行变量重要性排序，从而找到对预测数学成绩最重要的变量并加以比较。

本文展示了如何基于社会科学数据集使用机器学习算法：通过分析中国教育追踪调查 2014 年基线数据 (CEPS)，本研究使用三种经典的机器学习算法：LASSO 回归、随机森林和 Xgboost 对学生的数学成绩进行分析预测。由于影响学习成绩的因素往往非常复杂，也就是说，单一因素无法决定学习成绩，传统工具并不能刻划教育影响因素的复杂机制。本文通过数据驱动的方法发现父母的教育期望、作业检查行为对于数学成绩的影响最为重要。具体是使用交叉验证的方法训练 LASSO 算法，通过绘制变量收缩图判断不同变量对于因变量的预测能力，本文还对数据集进一步拟合 Xgboost 算法，通过计算变量重要性系数做出机制解释，比较不同因素的影响强度。此外，本文还给出了 GLM 方法的分析，以进行方法优劣性的比较。

机器学习算法之于社会科学并非完全陌生的新领域。事实上，一些简单的机器学习算法已经在社会科学研究中得到了广泛应用。如非监督学习中的主成分分析、聚类分析。如非参数方法中的 LOWESS 回归、GAM 回归等[1]。但更为成熟和复杂的机器学习算法 (尤其是监督学习算法) 并没有被社会科学研究者们所熟知。本文希望通过真实数据的例子，促进机器学习在社会科学中的传播。

然而，机器学习算法的一些特点导致它不能被简单移植到现有社会科学问题中。它最为人所诟病的一点是可解释性较差，因此甚至被称作 “黑箱模型”。除了决策树等能输出明确决策规则的算法之外，很多机器学习算法很难给出让人类能够直观理解的推断机制。比方说，神经网络类算法将大量神经元的输出当做下一层神经元的输入，层层叠加，原始特征被进行了多次转换，因此很难建立原始的 x 与 y 之间的直观联系。支持向量机算法将原始样本映射到更高维的特征空间从而寻找到一个超平面来划分不同类别的样本，这同样将原本的信号进行了大幅度转化。这些特质导致了很多机器学习算法的决策边界[2]十分复杂。这个特点阻碍了机器学习技术在社会科学研究中的广泛应用。使机器学

[1]本文图 2 中最右图使用 LOWESS 回归展示双变量的分布趋势。

[2]机器学习算法的决策边界是一个多维空间中的曲面，它将不同类别样本点划分开来。机器学习算法的任务是找到一个决策边界使得错误划分样本尽可能少。

习算法更具有可解释性是一个前沿的研究方向，包括提炼出简单的决策边界，刻划交互效应这种复杂机制等。Molnar 的著作 (2020) 总结了这个领域的研究成果。

尽管如此，机器学习仍然无法完全取代 GLM。目前已有研究者尝试二者结合的方式，使用机器学习模型回答因果推断等问题。在 GLM 类方法所有需要准确预测数据的任务中都可以使用机器学习算法，如二阶段最小二乘法的第一阶段回归。Susan Athey(2015) 使用机器学习中的交叉验证方法进行因果推断工作，她还提出了基于随机森林的因果推断方法 (Wager and Athey 2018)。一些研究者开发出了使用决策树代替线性混合效应模型中的固定效应部分的模型 (Sela and Simonoff 2012)。在计量经济学领域中，Bellon 和 Chernozhukov(2014; 2016) 等人引入机器学习方法来辅助估计因果推断中的干预效应。

现有的统计工具容易使社会科学定量研究的结果难以重复，这不利于学术交流和学科的发展。陈云松等 (2012) 讨论了社会科学定量研究的可重复性问题，他认为社会科学定量研究应当拥抱开源和可复制性，研究者应当公开数据分析的源代码和研究过程中的技术细节。目前，定量写作的数据分析与论文撰写是完全割裂的，人们习惯于将 A 软件中的导入到 B 软件中，将 B 软件的结果粘贴到 C 软件中。这种手工操作不可避免会产生细微的错误，还给了统计造假行为以可乘之机。相比之下，“可重复性动态报告” 工具能够更好地在技术层面支持可重复性研究，它将数据分析源代码与文档相结合，只要得到程序源码，可以将整篇文档复制出来[1]。

在使用更先进的统计工具的前提下，社会科学定量研究者应当响应哈佛大学统计系孟晓犁 (2019) 教授的号召[2]，在学术研究中，应当：

- 讨论数据的收集、预处理、质量、限制以及这些因素的影响
- 阐明、评估并讨论数据分析和建模的假设及其结果
- 进行调查，并且表现出对选择偏差、混淆因素、何时/是否可以得到因果关系的结论的深刻理解

[1]本文是一个可重复性研究例子，本文写作均基于动态文档程序包 knitr(Xie 2017) 和标记语言 RMarkdown(Baumer et al. 2014)。本文呈现的统计图形在源文档中是 R 代码，运行文档后，原有代码的位置被输出的统计图形自动替换。关于可重复性报告可以参见 knitr 的中文介绍：https://bookdown.org/yihui/r-ninja/auto-report.html

[2]文章中文版可见统计之都网站文章《从统计地显著到显著地统计》，https://cosx.org/2019/08/significantly-statistical/

- 对多元关系及分布表现出一致的概论意义上的思考和处理
- 合理运用统计学方法并且承认它们的不足之处
- 对不确定性进行适当的传播分析、量化及表示
- 表现出对统计学原理的深入理解，比如统计推断与偏差–方差权衡

作为一种多学科交叉的工具，机器学习算法的应用已不仅仅局限于对表格数据的分析和推断。它同样能够在“非结构化数据”中大展拳脚，如语音数据、文本数据和图像数据。机器学习中的语音识别技术，可以辅助田野研究者进行访谈文本转化。根据自然语言处理 (Natural Language Processing, NLP) 知识，研究者可以从网页获取大量的文本数据并进行分析。情感分析能够对文本的情感倾向进行评分。文本分类能够根据带标签的训练集对新数据进行分类[1] 。LDA 主题模型可以分析文档作者感兴趣的主题分布，以及每篇文档所涵盖的主题比例等。

如上文所述，机器学习算法不仅是解决以往社会科学研究误区的可行路径，更能开拓社会科学研究的视野，为其带来更广泛的问题域。借助这些工具，研究者们能够为社会科学的定量研究带来新鲜的血液。未来的社会科学研究者们应当积极拥抱数据科学，展开跨学科合作，建立更为透明的数据驱动研究范式。

[1]文本分类的具体应用可见作者编写的垃圾信息判别网页程序，源代码和程序详见：https://github.com/songxxiao/txtnb。

参考文献

Angrist, J. D., and Pischke, J.-S. (2008), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Athey, S. (2018), "The impact of machine learning on economics," in *The economics of artificial intelligence: An agenda*, University of Chicago Press, pp. 507–547.

Athey, S., and Imbens, G. W. (2015), "Machine learning methods for estimating heterogeneous causal effects," *stat*, 1050, 1–26.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014), "R markdown: Integrating a reproducible analysis tool into introductory statistics," *arXiv preprint arXiv:1402.1894*.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, Oxford University Press, 81, 608–650.

Breiman, L. (2001), "Random forests," *Machine Learning*, 45, 5–32.

Breiman, L., and others (1996), "Heuristics of instability and stabilization in model selection," *The annals of statistics*, Institute of Mathematical Statistics, 24, 2350–2383.

Breiman, L., and others (2001), "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, Institute of Mathematical Statistics, 16, 199–231.

Brodeur, A., Lé, M., Sangnier, M., and Zylberberg, Y. (2016), "Star wars: The empirics strike back," *American Economic Journal: Applied Economics*, 8, 1–32. https://doi.org/10.1257/app.20150044.

Buuren, S. van, and Groothuis-Oudshoorn, K. (2011), "Mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software, Articles*, 45, 1–67.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016), "Double/debiased machine learning for treatment and causal parameters," *arXiv preprint arXiv:1608.00060*.

Cohen, J. (2016), "The earth is round (p<. 05)," in *What if there were no significance tests?*,

Routledge, pp. 69–82.

Draper, D. (1995), "Assessment and propagation of model uncertainty," *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, 57, 45–70.

Heise, D. R. (1972), "Employing nominal variables, induced variables, and block variables in path analyses," *Sociological Methods & Research*, Sage Publications Sage CA: Thousand Oaks, CA, 1, 147–173.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An introduction to statistical learning: With applications in r*, Springer.

Molnar, C. (2020), *Interpretable machine learning: A guide for making black box models explainable*.

Nuzzo, R. (2014), "Scientific method: Statistical errors," *Nature News*, 506, 150.

Raftery, A. E. (2001), "Statistics in sociology, 1950–2000: A selective review," *Sociological Methodology*, Wiley Online Library, 31, 1–45.

Sela, R. J., and Simonoff, J. S. (2012), "RE-em trees: A data mining approach for longitudinal and clustered data," *Machine learning*, Springer, 86, 169–207.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, 58, 267–288.

Tukey, J. W. (1991), "The philosophy of multiple comparisons," *Statistical science*, JSTOR, 100–116.

Wager, S., and Athey, S. (2018), "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, Taylor & Francis, 113, 1228–1242.

Wooldridge, J. M. (2016), *Introductory econometrics: A modern approach*, Nelson Education.

Xiao Li, M. (2019), "From statistically significant to significantly statistical," *Institute of Mathematical Statistics President President's Column*.

Xie, Y. (2017), *Dynamic documents with r and knitr*, Chapman; Hall/CRC.

于淼 (2017a), “估计、p 值与科学决策,” 未发表.

于淼 (2017b), “假设检验的乌云,” 未发表.

吴喜之 (2019), “从模型驱动的集体推断到数据驱动的个体预测,” 第十二届中国 *R* 会议 *(*北京*)*.

唐俊超 (2015), “输在起跑线———再议中国社会的教育不平等 (1978-2008),” 社会学研究, 3, 123–145.

方超，黄斌 (2018), “家庭人力资本投资对儿童学业成绩的影响——基于 ceps 追踪数据的多层线性模型分析,” 安徽师范大学学报 *(*人文社会科学版*)*, 18.

李忠路，邱泽奇 (2016), “家庭背景如何影响儿童学业成就?———义务教育阶段家庭社会经济地位影响差异分析,” 社会学研究, 4, 121–144.

程诚 (2017), “同伴社会资本与学业成就——基于随机分配自然实验的案例分析,” 社会学研究, 141–164.

胡安宁 (2016), “统计模型的 ‘不确定性’ 问题与倾向值方法,” 社会, 37, 186–210.

谢军 (2018), “教育大数据与现代教育学,” 第十一届中国 *R* 会议 *(*上海*)*.

谢宇 (2012), 社会学方法与定量研究, 社会科学文献出版社.

谢宇 (2013), 回归分析, 社会科学文献出版社.

陈云松，吴晓刚 (2012), “走向开源的社会学定量分析中的复制性研究,” 社会, 32, 1–23.

陈硕，王宣艺 (2018), “机器学习在社会科学中的应用：回顾及展望,” 复旦大学工作论文.

附录

部分关键 R、Python 代码

```
### 图3 Bootstrap特征抽样并画图
`%>%` = magrittr::`%>%`
library(ggplot2)
set.seed(2020)
get_coef = function(x, replace = F,n = 10){ # 输出母亲教育年限系数
  features = subset(ceps1,
      select = -c(stdmat,maedu)) # 删除因变量和母亲教育年限
  samped = sample(colnames(features), n,
      replace = replace) # Bootstrap列抽样
  fmla = as.formula(paste('stdmat ~','maedu + ',
      paste(samped, collapse= '+'))) # 生成公式
  lmift = lm(fmla,ceps1) # 拟合线性回归
  coef = lmift[['coefficients']][['maedu']] # 提取母亲教育年限系数
  return(coef)
}

lst0 = sapply(1:200, get_coef, replace = T) # 重复200次
lst1 = data.frame(number=1:length(lst0),coef = lst0) # 生成数据框
lst1 %>%  # ggplot2画图
ggplot() +
  geom_point(aes(number,coef,col = factor(coef>0))
             ,size=2) +
  geom_hline(yintercept = 0,col='black',size=0.5) +
  labs(x='模型编号',y='“母亲教育年限”系数') +
  theme_bw() +
  theme(legend.position = 'none')

### LASSO计算交叉验证RMSE, R2
library(glmnet)
cvfit = cv.glmnet(x_train,y_train,type.measure='mse',alpha = 1)
#summary(cvfit)  # 打印结果
best = cvfit$lambda.min # 得到最优参数
best_mod = glmnet(x_train,y_train,alpha = 1,lambda=best)  # 隐式交叉验证
library(caret)
folds = createFolds(y = y_train, k = 10) # 划分数据集

cross_val = function(i,square = T){ # 编写交叉验证函数
tra_x = x_train[-folds[[i]],] # 训练集x
```

```
tra_y = y_train[-folds[[i]],] # 训练集y
valid_x = x_train[folds[[i]],] # 验证集x
valid_y = y_train[folds[[i]],] # 验证集y
mid = glmnet(tra_x, tra_y,alpha = 1, lambda=best) # 拟合LASSO
pred_y = predict(mid, s = best, newx= valid_x) #求出验证集预测值
r2 = R2(pred_y,valid_y) # 计算R方
rmses = RMSE(pred_y,valid_y) # 计算RMSE
if (square){return(round(r2,3))
  }else{return(round(rmses,3))}
}

sapply(1:10, cross_val,square = T) # 计算10次
#>  [1] 0.154 0.143 0.167 0.160 0.176 0.149 0.133 0.160 0.145 0.143
sapply(1:10, cross_val,square = F)
#>  [1] 9.219 9.074 9.118 8.884 8.974 9.082 9.090 8.968 9.024 9.006

sapply(1:10, cross_val,square = T) %>% # 10折交叉验证均值
   mean()
#> [1] 0.153
sapply(1:10, cross_val,square = F) %>%
   mean()
#> [1] 9.0439
# 测试集
y_pred = predict(best_mod,s = best,newx = x_test)

R2(y_pred,y_test) %>%
  as.numeric() %>%
  round(.,3)
#> [1] 0.155

RMSE(y_pred,y_test) %>%
  as.numeric() %>%
  round(.,3)
#> [1] 9.135
```

```
### Xgboost 交叉验证
from xgboost import XGBRegressor
from sklearn.model_selection import cross_val_score
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
import numpy as np
def print_r2(model,x,y,cv): # 模型, x, y, 交叉验证数
    '''交叉验证R2打印'''
    cvscore = cross_val_score(model, x, y, cv = cv,scoring='r2')
```

```
    print(cv,'折交叉验证R2为',round(cvscore.mean(),3))

def print_rmse(model,x,y,cv): # 模型, x, y, 交叉验证数
    '''交叉验证rmse打印'''
    cvscore = cross_val_score(model, x, y,
                              cv = cv,
                              scoring = 'neg_mean_squared_error')
    nmse = cvscore.mean()
    rmse0 = np.sqrt(-nmse)
    print(cv,'折交叉验证RMSE为',round(rmse0,3))

model = XGBRegressor(n_estimators=100,objective='reg:squarederror')
print_r2(model,x_train, y_train,10)
#> 10 折交叉验证R2为 0.185
print_rmse(model,x_train, y_train,10)
#> 10 折交叉验证RMSE为 8.868
def print_reg_metric(model,x_train,y_train,x_test,y_test):
    rgf = model.fit(x_train,y_train)
    ypred = rgf.predict(x_test)
    rmse = np.sqrt(mean_squared_error(y_test, ypred))
    r2 = r2_score(y_test, ypred)
    print('测试集R2为%.3f, 测试集RMSE为%.3f'%(r2,rmse))

print_reg_metric(model, x_train,y_train,x_test,y_test) #打印结果
#> 测试集R2为0.242, 测试集RMSE为8.552
```