社会分层研究中的调节效应：一个整合性的分析框架

李适源　柳皑然①

摘要：本文旨在介绍社会分层研究中的调节效应。基于潜在结果与因果图，本文首先厘清了调节变量的概念界定，并据此指出好的调节变量应具备的基本特点；其后介绍针对调节效应的两种常见分析思路及其内在逻辑联系，在此基础上提供了一个整合性的分析框架；最后讨论如何利用观测数据来识别和估计调节效应，并着重介绍了基于潜在结果的实证分析策略。本文有助于完善和加深对调节效应分析的现有理解，同时也为未来的社会政策设计与评估提供了有用的分析工具。

关键词：社会分层　调节效应　潜在结果　模型误设　碰撞偏误

①作者简介：
李适源，北京大学光华管理学院社会研究中心博士候选人，研究方向：社会分层与流动、社会心态、青少年发展。
柳皑然（通讯作者），北京大学光华管理学院社会研究中心副教授、研究员。研究方向：儿童发展与教育、生命早期不平等、中国社会。

Moderating Effect in the Research on Social Stratification: A Comprehensive Analytical Framework

Shiyuan LI Airan LIU

ABSTRACT

By reviewing previous literature, this paper discusses the value and application of moderation analysis in social stratification research. First, we clearly define the concept of moderating variables using the notations of potential outcomes and causal diagrams, and we discuss generic characteristics shared by good moderating variables. Second, we propose a framework to apply moderation analysis in social policy evaluation under the existence of heterogenous treatment effects by using counterfactual analysis. Finally, we illustrated how to identify and estimate moderating effects based on observational data. We highlight the advantage of empirical strategy based on potential outcome framework, which provides direct answers to core questions in our analytical framework and is more flexible than the classical regression-based strategy. This paper enhances our understanding of moderation analysis and provides a reference manual for a better and more accurate assessment of moderation effects in policy evaluation.

KEY WORDS

Social stratification; Moderating effect; Potential outcome; Model misspecification; Collider bias

一、问题提出

在社会分层与流动的研究谱系中，描述社会分层的现实格局，探索不平等的系统性根源，寻找实现社会流动的可能渠道，始终是分层与流动研究[①]的核心关切（Hout and DiPrete, 2006; Xie, 2007; Grusky, 2019; Wu, 2019）。

（一）调节效应在社会分层研究中的兴起

近年来，以“潜在结果框架”(potential outcome framework; 见 Rubin, 2011) 和“因果图”(causal diagram; 见 Pearl, 2009) 为代表的因果推断框架蓬勃发展，并被广泛地应用于社会学研究当中（陈云松、范晓光，2010；彭玉生，2011；胡安宁，2012；胡安宁等，2021；句国栋、陈云松，2022；Gangl，2010；Morgan and Winship，2015）。在社会分层领域的实证研究当中，针对调节效应 (moderation effect) 的分析也日益得到重视。

考察调节效应的一种常见方式是“处理效应异质性”分析 (treatment effect heterogeneity)，它关注某一因果效应的方向和强度，如何随着社会群体特征（如阶层、族群、性别）的不同而展现出差异。社会分层研究中，一大关注点是社会劣势群体是否能够从某项政策或干预项目中“收益更多”，从而对他们原有劣势提供补偿[②]。一个典例是高等教育收入回报

[①]需要说明，社会分层与社会流动在概念定义、理论意涵方面均有区别，但本文的论述中将二者作为一组相互联系的概念整体，并未做出严格的区分。

[②]需要注意，实证研究中也涉及与此相反的情形，也即社会优势阶层从某项干预中收益更大，或社会劣势阶层在某项负面冲击中损失更多。本文聚焦于方法论本身的讨论，因此正文中只考虑了社会劣势阶层从干预中收益更大的“劣势补偿情形”。本文提供的分析框架，同样适用于其他的效应异质性模式。

的阶层异质性。布朗德与谢宇 (Brand and Xie, 2010) 基于美国的纵向调查数据发现，相比较高阶层出身的子女，那些较低阶层出身的子女由于接受高等教育而获得的收入回报更高[①]。

为更进一步理解“效应异质性”对于社会分层的具体意涵，社会分层研究者在考察效应异质性的基础之上，发展出了更细致而深入的分析思路 (Torche, 2011; Witteveen and Attewell, 2017; Karlson, 2019; Zhou, 2019)，本文将这种思路称为“反事实关联”分析。简单来说，“反事实关联”就是在“处理效应异质性”分析的基础上，进一步考察某项干预或处理（如，接受高等教育）如果在全人群中得到普遍推行，这是否能对社会不平等的“现状”带来改善，以及具体能够带来多大程度的改善[②]。

（二）“调节效应分析”的方法论挑战

目前，调节效应分析越来越多地在社会分层等领域得到了出色应用，对理解社会不平等提供了宝贵洞见。不过，调节效应分析仍存在三方面的方法论挑战，分别是：(1) 调节变量的概念界定尚不明确；(2) 调节效应对于改善社会不平等的政策意涵有待进一步阐明；(3) 针对调节效应的实证策略对因果识别与模型假定的认识尚且不足，易陷入“选择性偏误”或“模型误设”等常见陷阱，进而造成对调节分析结果的误

[①]需要说明，这一结论是基于“间接证据”而做出的。原文的直接证据是“上大学概率更低的子女从高等教育中获益更多”；与此同时，该研究发现，家庭社会经济地位的高低与上大学概率是正相关的。由此作者推测，家庭社会经济地位较低的子女，从高等教育中的获益更多。

[②]后文将对“反事实关联”的具体含义做详细介绍。

解与误读。

针对调节分析的现存挑战，本文回顾和梳理了社会分层研究中有关调节效应的代表性文献，基于因果推断范式中的“潜在结果”概念（并配合“因果图”），尝试为读者提供一个整合性的分析框架，以及相应的实证策略与操作指南。

本研究的结构安排如下。

首先，本文使用因果推断范式中的“潜在结果”概念对调节变量、调节效应做出正式界定，讨论调节变量与处理变量的核心区别，并探讨“好的调节变量”具备的常见特征。

第二，本文尝试提供一个在社会分层研究中考察调节效应的分析框架，构成这一框架的核心要素是“效应异质性”与“反事实关联”。在讨论这一分析框架的过程中，我们使用“潜在结果”作为概念工具，以更清晰表达“效应异质性”与“反事实关联”分析所关注的核心问题，并阐明这两种分析思路间的内在逻辑联系。本文提供的分析框架有两方面意义：一方面，它可以在认识论层面上帮助实证研究者更好地理解“处理效应异质性”对改善社会不平等现状的具体政策意涵；另一方面，它可以在方法论层面上为相应的实证策略提供指引，减少研究者对调节效应分析方法的误用以及对实证结果的误读。

第三，本文着重介绍“基于潜在结果的实证策略”。它基于潜在结果来定义因果效应，刻画“反事实”，能直接回答调节效应分析框架涉及到的核心问题。这种实证策略清晰地分离了“因果识别”与“模型估计”两个步骤，有助于更加明确地认识到调节效应分析的常见陷阱，增进分析结果的可靠性。

二、调节变量的概念界定

在较早的社会科学研究文献中，调节变量不具有明确的因果含义。社会心理学家巴隆与肯尼 (Baron and Kenny, 1986: 1174) 在一篇针对调节效应与中介效应的经典综述中指出，调节变量是指可影响到“因变量 Y 与某一自变量 X 关系强度和（或）方向”的变量[①]。给出定义后，作者紧接着使用“相关性分析”的术语来说明调节效应。他们并没有明确指出自变量与因变量的关系具有因果性。也即，他们所指的调节变量，调节的并不一定是对的“因果效应”。

基于传统定义，如果两个变量的关联性依赖于第三个变量的取值，就可以认为变量“调节了”这两个变量的关系。

本文无意判定传统定义是否正确。但是从实证研究角度看，由于传统定义中缺乏明确的因果概念，研究者往往难以区分调节变量与解释变量（处理变量），也难以区分调节效应与因果效应。这可能导致调节变量在研究中被误用或滥用。

据此，本文在“因果推断范式”中来定义调节变量。无论基于潜在结果框架 (Rubin, 2011)，或是基于因果图的逻辑 (Pearl, 2009)，只要能引入明确的因果效应（处理效应）[②]，就能在概念上澄清“调节变量的定义”[③]。

---

①原文为“a moderator is a qualitative (e.g., sex, race, class) or quantitative (e.g., level of reward) variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable.”

②本文讨论的因果效应都可以视作处理效应，本文将两者作为等价概念使用。

③为方便理解，本节用直观的因果图来说明调节变量的定义；在理论框架中再正式引入潜在结果的概念。

具体地，本文基于社会学家沃德克与阿尔米拉利 (Wodtke and Almirall, 2017) 在因果推断框架中对调节变量做出的定义：调节变量是指，“能够系统性地改变‘未来’处理变量对结果变量的效应形式、效应方向或效应强度”的变量[①]。

注意到，通过区分不同变量的生成时点，上述定义明确指出，调节变量应是“生成时点早于处理变量的前定变量”(pre-determined variable)。与此较类似的调节变量定义，也见于社会学家布兰德及其同事 (Brand et al., 2021）以及流行病学家范德韦尔 (VanderWeele, 2015) 的代表性研究或著作。

下面的因果图有助于进一步说明调节变量的定义。

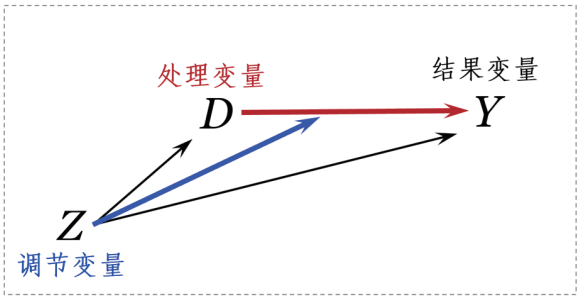

图 1 调节变量与调节效应的定义图示

图 1 当中，调节变量 的发生时点早于处理变量 D，而处理变量 D

[①]需要说明，尽管上述定义是在纵向研究情形 (longitudinal setting) 当中给出，但该定义对截面研究也具有适用性。在截面数据中，有许多“回溯型”变量都具有明确的先后时点。例如受访者 3 岁时的户口类型、受访者 13 岁时的父母职业类型，等等。

发生时点又早于结果变量 Y。其中，由处理变量指向结果变量的箭头 D → Y 定义为 D 对 Y 的因果效应（处理效应）；由调节变量指向处理效应 D → Y 的箭头表示 D 对于 Y 的因果效应存在“异质性”。因此，本文所定义的调节变量，是指可以带来“处理效应异质性”的变量。

具体地，在社会分层的研究中，“好的调节变量”往往具备以下几个特点。

首先，“调节变量”与“处理变量”在分析中的“角色定位”应有较明确的区分，这对后续的实证分析具有重要意义。在具体研究中，研究者往往面对两个在分析地位上几乎同等重要的自变量。例如，子女的“家庭阶层背景”以及“是否获得高等教育”。此时，研究者可以考虑自变量的取值“受到干预（或得到改变）的难度大小”，以判断这一自变量更适合视作调节变量，还是处理变量。据调节变量的定义，调节变量所要调节的是“自变量对因变量的因果效应”（即使实证研究中存在因果识别的经验难题，但至少应在理论和概念层面上具有因果意义）。

例如，若某自变量（高等教育获得）正好对应于某种社会政策与干预项目[①]，那么该变量就更适合当作处理变量，因为它与结果变量的关系（至少在概念上）更容易被理解为因果效应（处理效应）；相反，若某一自变量相对更难“受干预或被改变”（如家庭出身、族群身份等先赋特征），这类变量更宜被视作调节变量。而相应地，这类变量与结果变量的关系，更宜被理解为“经验（观测到）的关联性”(empirical

① 尽管是否选择读大学往往是个体选择的结果，但这并不影响在“概念上”将其设想为一种干预或处理。

association)，而不是因果关系。社会分层研究中，这种“经验关联性”刻画的，正是结果变量在不同社会群体间的分布差异。

其二，调节变量与处理变量的“测量时点”[①]往往具有先后顺序之分。对于给定的一段研究时期，调节变量总是“前定变量”。即，调节变量的取值“生成时点”，应该“早于”它所调节的处理变量的生成时点 (Wodtke and Almirall, 2017: 7)。调节变量的“前定性质”，在一定意义上可以保证调节变量不会受处理变量、结果变量的影响；反之，若调节变量发生在处理变量之后，将可能受到处理变量的影响，进而引发“后定变量偏误”，或言“碰撞偏误”(collider bias)、“样本选择偏误”(sample selection bias)，从而估计出“虚假的调节效应”[②]。

其三，调节变量的变量取值应该（至少在一段时期内）保持相对稳定。例如，家庭出身、性别与族群身份等“与社会分组有关”的变量。从方法层面看，具有社会分组性质的调节变量往往具有“前定性质”，在很大程度上能够避免“碰撞偏误”；从分层研究的实际意义看，如果调节变量对应某种特定的社会分组标准（如，阶层出身），那么调节效应分析的结果可以回应社会分层研究的核心关切。其中，调节变量与结果变量的关联性强度，就可以直接刻画（某一维度）的不平等程度；而基于调节变量分组得到的处理效应异质性，也可以反映出某种政策干预是否能对社会劣势群体带来回报上的补偿。

①更严格的说法应该是变量生成的时点。

②当然，在涉及多期处理变量、多期调节变量的纵向研究中，对于较晚时点的调节变量而言，或许不可避免地受到较早时点处理变量的影响。这时需要使用特殊的实证策略化解可能存在的效应偏误。详见(Wodtke and Almirall, 2017; Wodtke, 2020)。

三、调节效应的分析框架

（一）处理变量对结果变量的“效应异质性”

从本节开始，我们用“潜在结果”概念 (Rubin, 2011) 来正式定义因果效应。对于“二值处理变量（例如是否上大学）”D_i，个体 i 拥有两个潜在结果（例如收入的潜在结果）。其一是“如果接受了处理”对应的潜在结果 $Y_i(1)$（如果上了大学将会实现的收入水平）；其二是“如果没有接受处理”对应的潜在结果 $Y_i(0)$（如果没上大学将会实现的收入水平）。对于同一个体而言，只能观测到她真正实现的那一个潜在结果。比如，个体 i 事实上接受了大学教育，她的潜在结果 $Y_i(1)$ 就得以实现，成为了“观测结果”(observed outcome)；而她没有实现的潜在结果 $Y_i(0)$ 永远无法得到观测，成为了个体 i 的“反事实结果”(counterfactual outcome)。因此，个体层面的因果效应一般无法被识别，研究者主要考察“组群层面”(group-level) 或“全人群层面”(population-level) 的平均因果效应。

具体而言，社会分层研究中，研究者往往不只关注处理变量 对结果变量 在全人群当中的总体平均因果效应 (Average Treatment Effect, ATE)：

$$ATE = E[Y_i(1) - Y_i(0)]$$

而且关注特定的社会群体（或称社会分组，由调节变量 Z_i 表示）的“组群平均因果效应”（Group Average Treatment Effect, GATE; 详见 Knaus et al., 2021）。为了表述简便，下文将 GATE 统称为“组群因果效应”。针对特定社会群体 $\{Z_i = z\}$，组群因果效应定义为：

$$GATE(z) = E[Y_i(1) - Y_i(0)|Z_i = z]$$

在此基础上，我们想考察各社会群体的“组群因果效应”是否存在差异，这也是本文主要讨论的“效应异质性”(effect heterogeneity)。例如，对于两个社会群体 $\{Z_i = z\}$ 以及 $\{Z_i = z^*\}$ 而言，它们的“组群因果效应”之差为：

$$\Delta GATE(z^*, z) = GATE(z^*) - GATE(z)$$

概念上需要注意，本文讨论的“效应异质性”与方法论文献中提到的“条件平均因果效应”(Conditional Average Treatment Effect, CATE) 并不相同。本文所关注的“组群因果效应”(GATE) 基于单一的社会分组标准（如，出身阶层、族群、性别等），各社会群体的组群因果效应处于“适中的加总层次”(intermediate aggregate level)。相比之下，“条件平均因果效应”(CATE）的分组颗粒度更细致，它的分组标准主要有两种。其一，用整个控制变量集构成的“所有可能取值”来定义“最小群组”$\{X_i = x\}$，关注每个“最小群组”对应的因果效应；其二，用“接受处理的条件概率”，$\{P(D_i = 1|X_i = x) = p\}$，也即倾向得分来定义群组，进而考察倾向得分不同的群组在因果效应上表现出的差异 (Xie et al., 2012)。而本文所指的效应异质性 (GATE) 都基于单一分类的群组，而对于更细层面的效应异质性不再展开讨论[①]。

在本文中，我们将使用“高等教育的收入回报异质性”作为说明性例子[②]。考虑以下变量：其一，调节变量 Z 为二分类的家庭背景（1 高门

[①]有关 CATE 在社会学实证研究中的方法论介绍，读者可参考（Brand et al.，2021；胡安宁等，2021）。

[②]举例中涉及到多种对现实世界的假定，仅仅是为了帮助读者理解对应的形式化定义及相关结论。

子女，0 寒门子女）；其二，处理变量 D 表示个体是否受过高等教育（1 上过大学，0 没有上过大学）[①]；其三，结果变量 Y 表示个体在 30 岁时的收入水平。

就“处理效应异质性”分析而言，其主要目的是识别“针对不同社会分组的人群”（由调节变量 Z 的取值来定义），处理变量 D 对于结果变量 Y 的“组群因果效应”(GATE)，并比较各组的组群因果效应是否存在组间差异。具体到本例，这对应于高门子女读大学的收入回报 $GATE(Z_i = 1)$、寒门子女读大学的收入回报 $GATE(Z_i = 0)$，以及两组子女的效应差异 ΔGATE，它们的定义式如下：

$$GATE(Z_i = 1) \equiv E[Y_i(1) - Y_i(0)|Z_i = 1]$$
$$GATE(Z_i = 0) \equiv E[Y_i(1) - Y_i(0)|Z_i = 0]$$
$$\Delta GATE \equiv GATE(Z_i = 1) - GATE(Z_i = 0)$$

如果发现寒门子女接受高等教育的收入回报比高门子女更高，这就意味着“接受高等教育”（一项干预或者处理）可以对寒门子女原有的社会劣势提供补偿。研究者往往致力于探索“劣势补偿效应”是否存在，这可以视作是“效应异质性”分析对于推动社会分层研究、改善社会不平等的现实意义。

（二）深入理解“效应异质性”的现实意义

社会分层研究并不止步于对处理效应异质性的估计，而可能会进一

[①]严格来说，接受高等教育和完成高等教育并不等同，上大学也不只是高等教育的唯一形式。出于说明的简洁，本文在举例时并未严格区分高等教育获得与上大学之别。

步考虑“效应异质性”对于改善社会不平等现状的政策涵义。譬如，如果发现寒门子女上大学的教育回报比高门子女更高，那么推行“高等教育普及化”政策（也即让各阶层出身的子女都能接受高等教育），将能有效削减“不同阶层间现有的收入差距”。换句话说，假若高等教育在全人群中得到普及，阶层间现有的收入差距是否以及在多大程度上可以得到缩减，其实与高等教育的收入回报是否在阶层间存在异质性，以及效应异质性的方向、强度密切相关。

本文考虑，在特定处理状态（高等教育）普及到全人群的“反事实情形中”，结果变量（收入）与调节变量（阶层出身）的关联性强度，是否会相较现实情形得到改善。我们将这种分析思路称作“反事实关联”(counterfactual association) 分析，它可以视作是“处理效应异质性”分析的递进与深入。

我们继续以前文提到的“高等教育的收入回报异质性”为例来说明。其中，结果变量 Y 代表子女 30 岁时的收入，处理变量 D 代表子女是否接受了高等教育，调节变量 Z 是家庭背景（寒门子女取值为 0，高门子女取值为 1）[①]。处理效应异质性的存在，将会影响到某项“针对处理变量（高等教育）D 的社会政策”对于改变收入不平等“现状”的作用效果。

例如，在实施高等教育普及化政策的条件下，如果高等教育对于寒门子女的收入促进效应更大，那么这项普及化政策对削减“收入不平等现状”的改善作用，将会有别于“高等教育对寒门、高门子女收入影响

[①]父母社会经济地位等连续型变量，也可以是阶层出身、性别或族群身份等类别型变量，简化起见，本文将其定义为高门或寒门子女。

完全相同”的情况。为了具体评估在效应异质性存在与否的情况下，这类平等化政策对改善社会不平等的效果，可以运用“反事实关联”的分析思路。我们接下来展开一组“思想实验”，详细探讨“反事实关联”的分析思路与核心逻辑。

在下图 2 中，包含两条截距、斜率明显不同的直线：其中第一条线由 A 点指向 B 点，直线 AB 的斜率表示对于高门子女而言，读大学的收入回报；第二条线由 F 点指向 G 点，直线 FG 的斜率表示对寒门子女而言，读大学的收入回报。容易发现，直线 FG 的斜率明显比直线 AB 的斜率更大，这说明寒门子女读大学的收入回报相对更高。

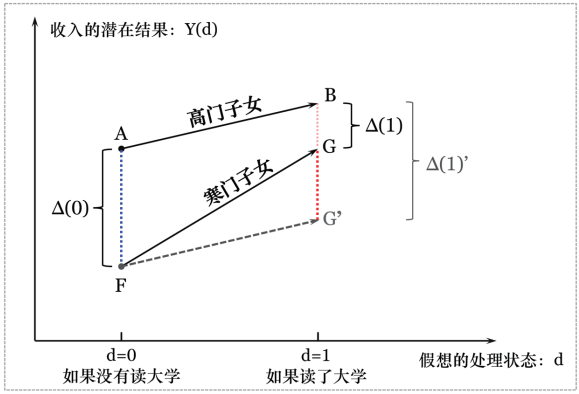

图 2“处理效应异质性”与“反事实关联”图示

现在设想两个针对处理状态（是否接受高等教育）的“反事实情形”，这两个情形恰好构成处理状态“全有”与“全无”的对照：其中，反事实情形 (1) 就是“高等教育普遍化”的情形。也即高等教育在全社会得

到了普及，所有子女都能获得高等教育；这一情形下，高门与寒门子女的收入差异为图 2 中 B 点与 G 点的纵坐标之差，也即 Δ(1)，我们称 Δ(1) 为“反事实关联 (1)”。

与此类似地，反事实情形 (2) 是“高等教育不存在”的情形。也即，全社会所有子女都没有接受高等教育。这一情形下，高门与寒门子女的收入差异就是图 2 中 A 点与 F 点的纵坐标之差，也即 Δ(0)，我们称 Δ(0) 为“反事实关联 (2)”。

最后介绍“事实情形”。在现实世界当中，无论高门还是寒门出身的子女，都各有一部分人接受了高等教育，也各有一部分人未接受高等教育。在事实情形下，高门与寒门这两组子女的收入差距就是“事实关联”。

为论证简便起见，不妨暂将“反事实情形 (2)”（所有子女都没接受高等教育）直接等同于“事实情形”。此时子女收入与其阶层出身的“事实关联”，就是 A 点与 F 点纵坐标之差 Δ(0)。

注意到，我们一旦知晓了“反事实关联”以及“事实关联”的具体取值（即不同情形下的阶层间收入差异），就能将这二者进行比较，从而衡量“高等教育如果得到全面普及”可以对收入不平等的现状带来多大程度的改善。具体说，图 2 左侧的 Δ(0) 与右侧的 Δ(1) 两者之差（即 G 点与 G’点的纵坐标之差），反映的就是高等教育普及化能够对现有阶层间收入差距带来的缩减幅度。

从形式上看，“反事实关联”分析其实可以视作一种较特殊的“双重差分”。第一重差分是指，分别在反事实情形 (1) 中（假设所有人都接受高等教育）、反事实情形 (2) 当中（假设所有人都不接受高等教育），高门与寒门子女的收入之差，也即上例中的“Δ(1)”与“Δ(0)”。而第

二重差分，则是用“反事实关联 Δ(1)”与“反事实关联 Δ(0)”做差。

需要注意的是，尽管形式相似，但在实质含义上，本文提供的“反事实关联”分析与经典双重差分存在着鲜明差异。在经典双重差分的设定中，横轴一般表示处理发生前与发生后，|AF|、|BG| 往往表示的是处理组与控制组在处理发生前、处理发生后，观测结果 (observed outcome) 的取值差异。也即，图 2 中的 A、F、B、G 四个点对应的取值都是某一组在某一期的“均值”，均可基于数据直接计算得出。若满足“平行趋势假定”，|BG| 与 |AF| 之差（经典双重差分的参数）可用于识别某项处理对处理组带来的因果效应。不同的是，本文“反事实关联”分析中，横轴表示的是两种“反事实情形”，|AF|、|BG| 表示由调节变量定义的两组人群（如高、低阶层），在不同的反事实情形中，相应的潜在结果 (potential outcome) 取值差异。换言之，A、F、B、G 四个点对应的取值并非可以直接观测到的均值，也不能直接基于数据计算得出[①]。而此处的双重差分 (|BG| 与 |AF| 之差) 并不是“高等教育对收入带来的因果效应”，它代表高等教育从“全无”到“普及”能对高低阶层间收入差距带来的缩小幅度。换言之，它评估的是“调节效应”而非“主效应”。

在本例中，“处理效应异质性”的存在（也即直线 AB 与直线 FG 的斜率具有差异），是“高等教育的普及化”能有效削减“收入不平等现状”的关键所在。

我们在图 2 中绘制了第三条直线 FG’来解释上述论断。注意到，直

[①]本文第四节将介绍在一系列假定条件下，可基于观测数据，通过回归拟合的方式间接计算出相应取值。

线 FG’与直线 AB 是平行的（两者斜率相同）。直线 FG’意味着寒门子女上大学的收入回报与高门子女完全相同。即，高等教育对收入的组群处理效应，对高、低阶层子女而言是完全同质的。此时，即使高等教育在全社会得到普及，也并不会改变高门与寒门子女之间的收入差异。反映在图中，当寒门与高门子女的收入回报相同，那么在“所有子女都接受了高等教育”的反事实情形下，高门与寒门子女的收入差距就是图中 B 点与 G’点的纵坐标之差 Δ(1)’。注意直线 FG’与直线 AB 是两条平行线，因此图右侧的 Δ(1)’与图左侧的 Δ(0) 是完全相等的。回忆在本例当中，Δ(0) 代表的是“事实关联”（即，高低阶层子女在现实世界中的收入差距），所以，当高等教育对收入的处理效应不存在阶层异质性时，针对高等教育的普及化政策，无法对阶层间的收入不平等现状带来改善。

从上述“思想实验”中发现，正是寒门子女接受高等教育的收入回报（相比高门子女的收入回报）更高，“高等教育的普及化政策”才能有效地削减高低阶层之间现存的收入差距。

需要指出，上文出于方便，直接将“所有子女都没上大学”视作“事实情形”，但现实社会中的真实情况是，有一定比例的寒门子女与高门子女接受了高等教育，且寒门子女中上大学的比例往往更低。但这一技术性假设不改变上述结论的本质意涵，也即“处理效应异质性”对于“改善社会不平等现状”的深刻意义[①]。

[①]现实社会中，高门、寒门子女的高等教育获得比例应分别处在图 2 线段 AB 内的一点 C、线段 FG 内的一点 H，而且高门子女对应的 H 点横坐标一般更大（高等教育获得比例更大）。此时，C 点与 H 点的纵坐标之差 |CH|，就是高门与寒门

（三）调节效应的分析框架及其社会学应用

如上所述，处理效应异质性不仅可以回答“谁获益（或受损）更多（或更少）”等社会分层研究的理论关切，而且也对学者评估一项（潜在的）普及化政策所能带来的对社会不平等现状的改善，具有重要意义。也即，是否存在处理效应异质性、效应异质性的具体模式与幅度，都将会影响到某项政策实施对改善不平等现状的效果。而为了评估“效应异质性”对于改善社会不平等现状具有的社会政策意涵，我们需要借助“反事实关联”这一实用的分析手段。

接下来，我们以“潜在结果”这一核心概念作为桥梁，结合图示和形式语言，进一步细致阐明“效应异质性”分析与“反事实关联”分析的内在逻辑联系，并将之整合为一个统一的分析框架。我们首先仍以高等教育的收入回报异质性为例，具体介绍这一分析框架当中的核心要素；然后说明该分析框架在社会分层研究中的广泛应用场景与现实意义。

子女在（接上页）现实世界中的事实收入差异。此时即使高低阶层子女不存在处理效应的异质性，高等教育普及化仍有可能会缩小收入不平等的现状。也即图 2 中的 Δ(1)’ 可能会小于 |CH|。但是当高等教育对收入存在“劣势补偿型”的处理效应异质性 (寒门子女的收入回报相对更高)，那么高等教育的普及化将会对阶层间收入差异的现状带来“更大程度的削减”。Δ(1) 将“更加地小于”|CH|。所以，严格意义上说，（特定模式）的处理效应异质性的存在，意味着“针对处理状态的普及化政策”可以对结果变量现存的不平等状况带来“额外的改善”。至于在“普及化政策”对不平等现状带来的改善当中，究竟有多大比例可归因于“效应异质性”的存在，借助本文提供的分析框架以及相应实证策略，可以较容易地得出答案。出于论证简单和方便读者理解的考虑，下文论证仍将“各阶层子女普遍不上大学”直接视作“事实情形”来考虑。

现在，我们对图2中各要素进行更详细和正式的标注与介绍，如图3所示。其中横轴是“假想的处理状态 d”（假设读了大学则 d = 1，假设没有读大学则 d = 0）；纵轴表示读大学、没读大学对应的潜在收入 Y(d)；直线 AB 的斜率表示高等教育对高门子女的组群处理效应 GATE(Z_i = 1)；直线 FG 的斜率表示表示高教育对寒门子女的组群处理效应 GATE(Z_i = 0)。注意到，高门与寒门子女的组群处理效应存在异质性，寒门子女的收入回报明显更高[①]。

$$GATE(Z_i = 1) = Slope(AB)$$
$$GATE(Z_i = 0) = Slope(FG)$$
$$\Delta GATE = Slope(AB) - Slope(FG)$$

在图 3 展示的说明性例子中，其实隐含了两点预设，它们分别是起点劣势与回报补偿。其一，起点劣势是指，如果都没上大学，社会劣势阶层（寒门子女）对应的潜在收入将会更低。图 3 中表现为直线 FG 的截距（相比 AB）更低。

$$起点劣势：E[Y_i(0)|Z_i = 0] < E[Y_i(0)|Z_i = 1]$$

其二，回报补偿是指，如果接受了相同处理或干预（干预前都不能上大学，接受干预后都能上大学），社会劣势阶层（寒门子女）从中获取的回报将相对更大（高等教育获得对寒门子女的收入促进效用更强），从而对其出身的劣势提供了补偿。

[①]本例中处理变量、调节变量都是二值变量，因此函数关系都可以由线性函数表示。实证研究当中的处理变量和调节变量可能都是连续变量，它们与结果变量的关系可能是非线性的。本文提供的分析框架仅仅考虑了最简单的情形，但这一分析框架蕴含的基本假定与结论，同样能为更复杂的情形提供有益启发。

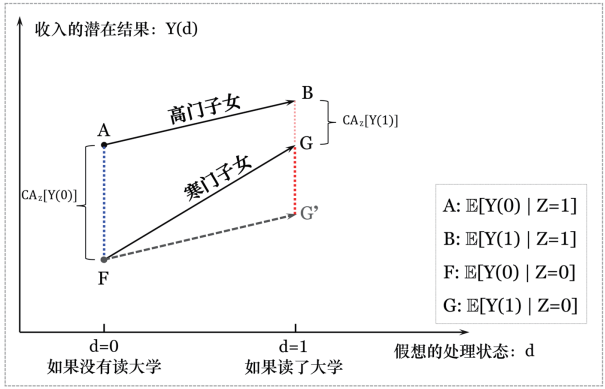

图 3 调节效应的分析框架图示

这在图 3 中对应于，直线 FG 相比直线 AB 更为“陡峭”。

$$\text{回报补偿：} GATE[Z_i = 0] > GATE[Z_i = 1]$$

现在，可以直接在图 3 中考察“反事实关联”分析涉及的核心问题。在图 3 中，反事实关联 (1) ——“所有子女普遍上大学的情形下”，高门子女与寒门子女的收入差异为 |BG|，即 B 点与 G 点纵坐标之差；反事实关联 (2) ——“所有子女都没有上大学的情形下”，高门与寒门子女的收入差异为 |AF|，即图中 A 点与 F 点的纵坐标之差。上述两个“反事实关联”的数学表达式如下：

$$|BG| = CA_Z[Y(1) = E[Y_i(1)|Z_i = 1] - E[Y_i(1)|Z_i = 0]$$
$$|AF| = CA_Z[Y(0) = E[Y_i(0)|Z_i = 1] - E[Y_i(0)|Z_i = 0]$$

再次说明，为了简单起见，我们将“反事实关联 (2)”直接等同于现

实世界中的高低阶层子女的收入差异（也即“事实关联”）。

如果上述例子中的两个预设（起点劣势、回报补偿）成立，那么可以得到一个具有重要现实意义的发现——在“高等教育得到普及化的情形下”，高低阶层间的收入差异，将会比“现实世界中的阶层间收入差异”更小。换言之，如果推行了高等教育普及化（各阶层子女都能获得高等教育）的社会政策，将会对阶层间的收入不平等“现状”带来有效改善。上述结论在图 3 当中对应于，竖直距离 |BG| 明显小于距离 |AF|。相反，如果不存在“回报补偿型”的效应异质性，寒门子女与高门子女的收入回报完全一样，那么“高等教育普及化”并不会对现有的阶层间收入差距带来任何改变。该结论就是说，竖直距离 |BG’| 与 |AF| 是相等的。

基于对上述实例的分析，可以概括出更具一般性的结论——“效应异质性”与“反事实关联”具有统一的内在逻辑联系。对于这一内在联系的清晰认识，有助于我们更深入地评估“效应异质性”存在与否、具体模式及其幅度大小，对于改善不平等“现状”所具有的社会政策意涵。在理论上，可以使用“潜在结果”的形式语言来说明这一内在联系：只要知晓了（由调节变量 Z 界定的）各社会群体的潜在结果均值 $E[Y(d)|Z_i = z]$，即可得到各社会群体对应的“组群因果效应”GATE(z)，以及效应异质性 ΔGATE；同时也能由此求出，在特定“反事实情形”当中，结果变量 Y 与调节变量 Z 的“关联性强度”（或言，Y 在各社会群体之间的组间均值差异），也即“反事实关联”$CA_Z[Y(d)]$。下列三个表达式是对上述论证的形式化概括[①]：

[①]出于论述简便，此处仅讨论调节变量 Z 为二值变量（取值为 0 或 1）的情形。事实上，调节变量 Z 的取值并不限于二值。后文将讨论更加复杂的情形，例如调节变量 Z 为多分类变量或连续型变量。

$$GATE(z) = E[Y_i(1) - Y_i(0)|Z_i = z]$$
$$\Delta GATE = GATE(Z_i = 1) - GATE(Z_i = 0)$$
$$CA_z = E[Y_i(d)|Z_i = 1] - E[Y_i(d)|Z_i = 0]$$

经验层面上，可以这样理解“效应异质性”与“反事实关联”的内在联系：结果变量 Y_i 与调节变量 Z_i 的关联性，之所以会在不同的反事实情形下存在差别；（或者说某项“将处理状态普及化”为导向的社会政策，之所以能对结果变量 Y_i 的不平等“现状”带来“有效改善”），这背后依靠的核心驱动力就是（特定模式的）“处理效应的组群异质性”。在我们的例子中，正是由于接受高等教育对寒门子女收入的正向影响相比高门子女更大，因此在“高等教育实现普及化”（所有人都接受高等教育）的反事实情形下，寒门子女与高门子女的组间收入差距（相比“所有人都没有接受高等教育”的情形）才会得到缩小。

这启示研究者，在完成效应异质性分析的基础上，可以进一步考察“反事实关联”，并与“事实关联”进行比较，这有助于理解“效应异质性”对改善社会不平等现状所具备的社会政策意涵。

在社会分层的经验研究中，上述分析框架具有广泛的应用场景。除了“高等教育的收入回报”这一典例外，调节效应的分析框架也能应用于当下的热点议题——与“双减”相关的社会分层研究。自 2021 年 7 月下旬开始推行的“双减”政策[①]，其核心目标之一是减轻学生的校外培训负担。在此背景下，有研究考察了“参加课外补习对学生学业表现、

[①]详见中华人民共和国教育部《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》，链接：http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202107/t20210724_546576.html

身心健康”的因果效应，以及学生的家庭阶层背景在其中所发挥的调节作用，以期为“双减”政策的落实与优化提供参考。具体而言，有研究发现课外补习参与将可能会导致学生的负向情绪增加；而且，补习对情绪健康的负面影响具有阶层异质性。较低阶层的子女为此付出的“情绪健康代价”或相对更大（李适源，刘爱玉，2022）。对此，我们可进一步考察上述“效应异质性模式”对健康分层的具体意涵。也即，相比(“双减”政策实施之前)的现实情形(各阶层子女都有一定比例参加了课外补习)，如若“课外补习被彻底取消”(各阶层子女都不再参加课外补习)[①]，那么，“子女情绪健康的阶层间差异”是否将会得到改善？具体可以得到多大程度的改善？可见，本研究提供的分析框架有助于更好地理解，“双减”政策对弥合不同阶层子女间心理健康差异的重要意义；此外，有研究者发现课外补习对于学业成绩的影响具有效应异质性（李昂然，2022），对此，我们可以进一步使用“反事实关联”的分析思路，考察在效应异质性存在的情况下，“如果取消课外补习”，是否能够、以及在多大程度上能缩小不同阶层出身的子女在学业表现上的差距。

（四）“反事实关联”分析：问题意识的来源与发展

在转向介绍针对调节效应的实证策略之前，还有必要对“反事实关联”关涉的理论脉络做出几点补充。实际上，“反事实关联”并不是一个全新概念视角，回顾社会分层与流动的经典议题，与之类似的分析思路并不少见。如，社会流动研究领域的学者们发现，由于父代与子代社会地位之间的关联强度在接受过高等教育的人群中更弱；因此，伴随着高等教育扩张和更多的人接受高等教育，整个社会的代际流动程度将会不断提

升 (Hout, 1988; Breen, 2010; Torche, 2011; Pfeffer and Hertel, 2015)。这种由人群的教育构成情况发生变化而引起的总人口中“代际流动性”增强，被称为（高等教育的）“构成效应”(compositional effect)。

注意到，本文介绍的“反事实关联”分析，与社会流动研究所关注的“高等教育 - 构成效应”具有紧密联系。本文讨论的调节效应分析框架中，若处理变量 为是否获得高等教育，调节变量 代表父母的社会地位，结果变量 代表子女的社会地位，那么此时的“反事实关联”分析要回答的问题就对应于在“高等教育覆盖率为 100%（各阶层子女普遍上大学）的假想情形下”，父母地位与子女地位的关联性（反事实关联）是否会比“现实中观测到的代际关联性（事实关联）”更为微弱。可见，此处的“反事实关联”分析，回答的问题恰好对应于“当高等教育扩张达到饱和状态时，（相比现实社会当中的高等教育分布状况），代际间地位关联性强度的下降幅度（即，代际社会流动性的改善程度）”。

尽管两者存在紧密联系，但本文讨论的“反事实关联”与社会流动领域中的“构成效应”不完全相同，它们存在两个主要区别：第一，分析层面与核心关注有所不同。社会流动研究中所指的“构成效应”，着眼于宏观层面的趋势分析。其核心关注在于“高等教育扩张的历史趋势”是否可以解释“代际关联性或社会流动性”长时期的变迁趋势（这类研究可见：石磊，2022；Hout, 1988; Breen, 2010; Pfeffer and Hertel, 2015）。相比之下，本文提出的“反事实关联”立足微观层面的因果推断。本质上关注的是“接受高等教育”对个人成就带来的处理效应是否在不同阶层出身的子女群体之间存在“效应异质性”。在这一基础上展开的“反事实关联”分析，目的是对“效应异质性”分析结果的深化与补充，

它尝试去衡量“效应异质性”对改善不平等现状的具体意义（此类研究可见：Torche 2011; Witteveen and Attewell, 2017; Karlson 2019; Zhou 2019）。第二，两者的适用场景不同。高等教育的“构成效应”主要用于解释较长历史时期、偏宏观取向的代际社会流动的变迁趋势；而本研究介绍的“反事实关联”作为调节效应分析的重要步骤，更适用于考察时间跨度更短、分析层次偏微观、具有明确因果含义的处理效应异质性对“社会不平等现状”带来的可能改善。

另外，由于研究者可以灵活选取处理变量、调节变量与结果变量，“反事实关联”分析不限于高等教育与社会流动这一经典议题，而且广泛适用于各类社会分层研究议题。例如，涉及儿童发展的议题，“反事实关联”分析可以用来评估“培育儿童的非认知能力”（处理变量）是否能削弱家庭背景（调节变量）与其学业表现、教育获得、成年期收入水平等结果变量的关联性 (Shanahan et al., 2014; Damian et al., 2015；Claro et al., 2016; Liu, 2019)。

四、调节效应的实证策略：因果识别与模型估计

在目前涉及调节效应的经验研究中，最常用的是“基于观测结果的实证策略”。例如研究者熟知的“交互项系数”或“分组回归系数比较”，大致都可以划归到此类实证策略类型。这类实证策略具有操作简便、结果易于理解等优势；但是，由于缺乏潜在结果与反事实的明确定义，在应用中可能面临两点问题。一是容易遭遇“模型设定偏误”；二是此类策略很容易受到“碰撞偏误”(Collider Bias) 的干扰，难以直接考察社会分层研究所关心的“反事实关联”。

为了更加清晰直观地回答调节效应分析框架关注的核心问题（“效应异质性”与“反事实关联”），同时为了更好地规避上述提到的两点常见问题，本文主张使用“基于潜在结果的实证策略”来进行调节效应分析。它直接使用潜在结果的概念来定义因果效应，刻画反事实情形，清晰分离了“因果识别”与“模型设定及参数估计”，有助于增进调节效应分析结果的可靠性。

（一）基于观测结果的实证策略

我们首先回顾“基于观测结果的实证策略”及其面临的常见问题。这类实证策略往往不涉及“潜在结果”的概念语言，也不会特意对“因果识别”与“模型设定”做出区分 (Imbens and Wooldridge, 2009)。它直接设定“总体模型”(population model) 或称“真实模型”(true model），假定在总体模型中，观测到的结果变量 Y_i 与自变量（包括处理变量 D_i 、调节变量 Z_i、其他控制变量 C_i）以及误差项 u_i 符合特定的函数关系。

1. 基于观测结果的“效应异质性分析”

针对“效应异质性”的模型设定，主要包括“设置交互项”与“分组回归系数比较”。下面分别介绍这两种模型设定及其可能面临的局限。

(1) 交互项模型

在基于观测结果 Y_i 的实证策略中，往往将结果变量 Y_i 对处理变量 D_i 的平均边际效应系数(Average Marginal Effect, AME) 理解为（总体平均）因果效应 (ATE)。以包含交互项的线性模型为例：

$$Y_i = \alpha + \beta_1 D_i + \beta_2 Z_i + \beta_3 (D_i Z_i) + \gamma C_i + u_i$$
$$E[u_i | D_i, Z_i, C_i] = 0$$

上式中，误差项 u_i 往往被假定“均值独立于”模型中的自变量，也即经典计量中的“外生性假定[①]”(Exogeneity Assumption)。若外生性假定得到满足，则称模型设定正确（correctly specified, 详见 Hong, 2020）。这一假定条件可视作基于观测结果的实证策略中最为核心的因果识别条件，它是总体模型各个参数能得到无偏估计的基本前提；据此，可导出结果变量 Y_i 关于所有自变量的边际效应系数：

$$\frac{\partial E[Y_i | D_i, Z_i = 0, C_i]}{\partial D_i} = \beta_1$$
$$\frac{\partial E[Y_i | D_i, Z_i = 1, C_i]}{\partial D_i} = \beta_1 + \beta_3$$

对于社会群体 $\{Z_i = 0\}$ 而言，结果变量关于处理变量的边际效应系数[②]就是 β_1；对于社会群体 $\{Z_i = 1\}$ 而言，边际效应系数就是 $\beta_1 + \beta_3$；而关键之处就在于此，基于观测结果的建模策略一般认为，总体模型中边际效应系数 β_1 以及 $\beta_1 + \beta_3$，恰好对应社会群体 $\{Z_i = 0\}$ 及 $\{Z_i = 1\}$ 的组群因果效应 $GATE[Z_i = z]$。而交互项系数 β_3，恰好就是我们感兴趣的“效应异质性”$\Delta GATE$。

[①]基于伍德里奇 (Wooldridge, 2015) 的表述，该假定意味着不存在：(1) 遗漏变量；(2) 自变量测量误差；(3) 函数形式误设；(4) 因变量对自变量的反向因果。

[②] 此处在概念上需要区分：“边际效应”是经典计量领域中的通常叫法，它虽然包含“效应”二字，但并不代表（基于潜在结果而定义的）因果效应。在因果推断框架中，如果缺乏因果识别条件，“边际效应”仅仅代表因变量与自变量的（偏）关联性或（偏）相关性。

(2) 分组回归模型

除了常见的“交互项模型”，当研究中的调节变量 Z_i 是分类变量（例如高门或寒门子女），研究者也常采用“分组回归”的方式来考察效应异质性。即，基于调节变量的取值 $\{Z_i = z\}$ 将总体划分为几个子群体，分别在每一个子群体当中，使用结果变量 Y_i 对处理变量 D_i，控制变量 C_i 做回归。如下式所示：

$$E[Y_i|D_i, C_i, Z_i = 0] = \alpha_{z=0} + \beta_{z=0}D_i + \gamma_{z=0}C_i$$
$$E[Y_i|D_i, C_i, Z_i = 1] = \alpha_{z=1} + \beta_{z=1}D_i + \gamma_{z=1}C_i$$

对于社会群体 $\{Z_i = 0\}$ 而言，结果变量关于处理变量的边际效应系数就是 $\beta_{z=0}$；对于社会群体 $\{Z_i = 1\}$ 而言，边际效应系数就是 $\beta_{z=1}$；而这两组的边际效应系数之差就是 $(\beta_{z=1} - \beta_{z=0})$，这也就是我们感兴趣的“效应异质性”ΔGATE。

相比“交互项模型”，“分组回归”主要放松了对控制变量组 C_i 施加的同质性假定。在分组回归模型中，控制变量 C_i 对应系数 γ_z 依赖于下标 z；也即允许对于不同的社会分组，控制变量 C_i 与结果变量 Y_i 的（偏）关联性不同。因此分组回归在一定程度上放松了对总体模型的函数形式假定。事实上，如果在交互项模型中，额外设定调节变量 Z_i 与其他所有控制变量 C_i 的交互项，那么交互项模型当中的 β_3 与分组回归模型中的 $(\beta_{z=1} - \beta_{z=0})$ 在数值上将完全相同 (Wooldridge, 2015)。

(3) 模型误设：调节效应的“陷阱”

基于观测结果来考察效应异质性，具有三大突出优势：操作步骤简洁、系数解读方便、结果易于理解。但是这种实证策略也可能面临一些

系统性风险：基于观测结果的实证策略，在考察“效应异质性”的过程中更可能将“模型设定错误（被遗漏的高次项或交互项）”误读为“效应异质性”。这一问题在新近的方法论研究中得到高度重视（胡安宁等，2021），被认为是调节效应分析中的常见“陷阱”（朱家祥、张文睿，2021；江艇，2022）。

我们使用因果图来直观地呈现上述“模型误设”问题。该问题可以理解为，研究者对控制变量组 的“统计控制不够充分”所致。以模型设定相对更灵活的分组回归为例。图 4 代表在子总体 {Z = z}（高门或寒门子女）当中的变量间关系。变量 C 代表的是“处理变量 D 和结果变量 Y 的共同原因”(common cause)，也被称为“混杂变量”(confounder)；变量 C 周围的虚线方框表示，尽管研究者观测到了所有的混杂变量，但却只在回归模型中设置了变量 C 的线性项，而并没有设置它的高次项，以及与其他自变量的交互项。这导致变量 C 当中仍可能保有“混杂因素的残余信息”，进而导致“非因果路径”D ← C → Y 未得到彻底阻断，最终仍然引发了“遗漏变量偏误”，由此不能正确识别 D → Y 对应的“组群因果效应”GATE(Z_i = z)。

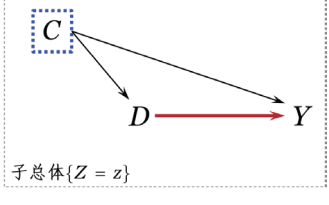

图 4 调节效应“陷阱”的因果图示

从表象上看，调节效应的陷阱意味着研究者在模型设定中出现了差

错，遗漏了自变量自身（或者自变量之间）的非线性项，进而“误读”了交互项系数的因果意义；从根源上看，这一问题其实反映了“观测结果模型”在认识论层面的局限。这类实证策略在模型设定过程中不外显地涉及到“潜在结果”概念，因此较难从因果效应的明确定义出发，来判断总体模型的特定参数是否等同于感兴趣的处理效应异质性①。

2. 基于观测结果的“反事实关联”分析

基于观测结果的实证策略，往往难以正确地识别“反事实关联”。这一识别难题的根源也在于，实证策略中并未明确涉及“潜在结果”与“反事实情形”的清晰定义，这致使我们难以判断，基于观测结果模型得到的“变量 Y_i 与变量 Z_i 的关联性”（如，收入水平与出身阶层的关联性）是否等同于本文感兴趣的“反事实关联”。回顾上一小节的交互项模型。结果变量 Y_i 代表收入，处理变量 D_i 代表是否接受了高等教育，调节变量 Z_i 代表阶层出身为高门还是寒门，并且假设唯一的控制变量 C_i 为高中时期的认知能力。

$$E[Y_i|D_i, Z_i, C_i] = \alpha + \beta_1 D_i + \beta_2 Z_i + \beta_3 (D_i Z_i) + \gamma C_i$$

需要注意，即便上述模型设定正确，总体回归方程中所有参数已知，也仍然无法直接推导出“反事实关联”。以“所有人都能上大学”对应的反事实关联为例。我们关注在这个反事实情形中，结果变量（收入）

①换言之，即使拥有了总体数据，也无法判断总体模型中的特定参数恰好就对应于因果效应及其异质性。

与调节变量（出身阶层）之间的关联性（也即高门子女与寒门子女的收入均值差异）。一个看似可行的思路是：将处理变量 D_i 的取值固定在“1”处，利用回归方程 $g(D_i, Z_i, C_i)$ 算出高门、寒门子女两组人的平均收入差异。这相当于是在计算，对于“事实上接受了高等教育的人群”而言，结果变量 Y 关于调节变量 Z 的边际效应系数 (Marginal Effect, ME)，记作 $ME_Z(D_i = 1)$，其具体表达式为：

$$\begin{aligned} & ME_Z(D_i = 1) \\ & \equiv E[Y_i \mid D_i = 1, Z_i = 1] - E[Y_i \mid D_i = 1, Z_i = 0] \\ & = (\beta_2 + \beta_3) + \gamma \cdot (E[C_i \mid D_i = 1, Z_i = 1] - E[C_i \mid D_i = 1, Z_i = 0] \end{aligned}$$

周翔（Zhou, 2019）将 $ME_Z(D_i = 1)$ 称为“条件关联性”(conditional association)。所谓“条件”，就是指研究者只在那些实际读了大学的子总体（而不是全人群）中去考察子女收入与阶层出身的关联性。

我们进一步用因果图来解释为何上述的边际效应系数（“条件关联性”）不是本文关心的“反事实关联”。注意到，在图 5 中，处理变量 D 恰好是“控制变量 C 和调节变量 Z 的共同结果 (common result)”，处理变量 D 构成了“变量 C 与 Z 的‘碰撞变量’”(collider)。此时，结果变量 Y 关于调节变量 Z 的边际效应系数 $ME_Z(D_i = 1)$ 可理解为，在“给定处理变量 D 的条件下”Y 与 Z 的关联性强度。

根据因果图的原理，如果给定碰撞变量，将引发“内生性选择偏误”[①]，或称“碰撞偏误”(colliding bias)。它将导致“碰撞变量”的两个“亲代

[①]注意此处的选择偏误类似于“样本选择偏误”，而不是“自选择偏误”。详见（句国栋，陈云松，2022）。

变量”——变量 C 和变量 Z 在总体中的真实关联性遭到“扭曲”，这种“关系的扭曲”体现为图 5 中左侧的红色弧形虚线。不仅如此，由于变量 C 是“结果变量 Y 和调节变量 Z 的‘共同原因’”，因此变量 C 与变量 Z 的扭曲关系将会进一步传导，导致结果变量 Y 与调节变量 Z 的真实关联性也同样会遭到“扭曲”①。

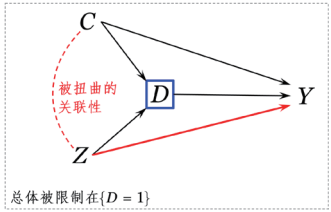

图 5 “碰撞偏误”的因果图示

具体来说，在“高等教育的收入回报”这一例中，变量 C（认知能力）、调节变量 Z（家庭阶层背景）两者都对 D（高等教育获得）有正向影响。研究者一旦给定 $\{D_i = 1\}$（也即，将分析对象限定在“事实上已接受了高等教育的人群”当中），高门与寒门子女在认知能力 C、阶层出身 Z 的组间差异（相比全人群）都会更小。这是因为，“谁能获得高等教育”具有高度的筛选性，筛选标准是基于认知能力和家庭背景。这种筛选性使得那些“有幸进入高等教育的子女”在（认知能力和阶层出身）两方面都具有更强的同质性 (homogeneity)。而且，C 与 Z 的关系扭曲还会进一步传导，

①需要注意，尽管在回归方程中已经控制了变量 C_i，然而基于边际效应系数得到的“条件关联性”仍会遭遇“碰撞偏误”。这一点从上文给出的边际效应系数表达式就可以看出。

导致结果变量 Y 与调节变量 的真实关联性（高门与寒门子女的收入差异）也会因此而下降。值得注意的是，变量 Y 与变量 Z 关联性强度的降低并不是因为“反事实情形”（高等教育普及化）对不平等的现状带来了改善，而是由于我们错误地控制了“碰撞变量”而引发了内生性选择偏差。

最后，从因果图中还可以发现一种特例：如果处理变量 D 不再具有选择性，那么基于观测结果模型得到的“条件关联性”就等同于“反事实关联”。当然，这对处理变量 D 要求非常高。例如，个体是否接受高等教育是完全随机分配的，不取决于个体特征和家庭背景。此时，“碰撞路径”C → D ← Z 就不复存在。

（二）基于潜在结果的实证策略

接下来介绍基于潜在结果的实证策略。这类策略直接使用潜在结果的概念来明确定义因果效应，刻画反事实情形，直接对应于调节效应分析框架涉及的核心问题，并且有助于化解调节效应分析中面临的“模型误设偏误”与“碰撞偏误”。

1. 因果识别条件

以潜在结果为基础的实证策略中，因果识别往往先于具体的模型设定与估计。因果识别的核心任务是提供“因果识别条件”，回答在何种条件下，研究者可以用已知的“观测结果”替代“潜在结果”，进而将基于潜在结果的因果效应定义式，转化为基于观测结果的表达式。

基于观测数据 (observational data) 的因果识别，最主要的两大识别条

件是“可忽略性假定”(Ignorability Assumption)及“共同支撑假定”(common support)[①]。为了方便在调节效应的方法框架中进行讨论，我们将“代表所有控制变量”的 X_i 拆分为调节变量 Z_i，以及“其余控制变量” C_i[②]。这两个假定的形式语言如下：

$$\text{可忽略性假定：} \{Y_i(1), Y_i(0)\} \perp D_i | C_i = c, Z_i = z$$
$$\text{共同支撑假定：} 0 < P(D_i = 1 | C_i = c, Z_i = z) < 1$$

对于由调节变量 Z_i 定义的任一组社会群体 $\{Z_i = z\}$（例如寒门子女），“可忽略性假定”是指，给定控制变量 C_i 的取值，潜在结果 $Y_i(1)$ 与 $Y_i(0)$ 的取值独立于“事实上的处理状态 D_i”。形象地说，若按照控制变量 C_i 的所有可能取值 $\{C_i = c\}$ 将总体划分成“格子”(cell)，在每个格子中，个体 i 的潜在收入 $\{Y_i(1), Y_i(0)\}$ 不再依赖于她事实上是否接受了高等教育。因此，高等教育带来的收入回报 $(Y_i(1) - Y_i(0))$ 也不再依赖于事实上的处理状态；而“共同支撑假定”是指每个格子 $\{C_i = c\}$ 中，既存在上了大学的个体，也存在没有上大学的个体。

“可忽略性假定”对于识别处理效应异质性具有重要意义。一方面，该假定条件意味着，研究者能够观测到所有可能的“混淆变量”，至少在理论上可排除遗漏变量偏误对因果效应识别的干扰；另一方面，（对

[①]有关因果推断基本假定的讨论并非本文重点，读者可以参考（Xie et al., 2012; 胡安宁 , 2012; 2020）。

[②]简洁起见，这里假定只有一个控制变量，也即“其余控制变量” C_i 为一个标量 (scalar) 而不是向量。但本节介绍的分析思路和相关结论，同样适用于“多个控制变量”的情形。

于社会群体 $\{Z_i = z\}$ 而言），如不满足“可忽略性假定”，则无法正确地识别“针对这一社会群体”的“组群因果效应” $GATE(Z_i = z)$。一般而言，研究者很难知晓各社会群体中（高门、寒门子女）被遗漏的“混杂变量”是否相同，也不易判定遗漏变量与处理变量、结果变量的关联模式是否相同，因此，处理效应异质性 ΔGATE 的识别往往也面临偏误[①]。此时可用敏感性分析 (sensitivity analysis) 来评估效应异质性的稳健程度[②]。

如果上述因果识别条件成立，则可以使用“总体数据”直接计算出我们关心的“平均潜在结果”和“平均因果效应”。注意到，对任一社会群体 $\{Z_i = z\}$ 而言（如寒门子女），给定控制变量组 C_i 取值，处理组的观测结果均值，等于控制组的反事实结果均值；同理，控制组的观测结果均值等于处理组的反事实结果均值。由此，可用观测结果的均值（下式左端）替换潜在结果的均值（下式右端）。

$$E[Y_i|D_i= 1, C_i = c, z] = E[Y_i(1)|D_i = 0, C_i = c, z] = E[Y_i(1)|C_i = c, z]$$
$$E[Y_i|D_i= 0, C_i = c, z] = E[Y_i(0)|D_i = 1, C_i = c, z] = E[Y_i(0)|C_i = c, z]$$

接下来使用“迭代期望法则”(Law of Iterated Expectation)，以去除条件期望函数中的控制变量，进而得到社会群体 $\{Z_i = z\}$ 的“平均潜在结果”（下式左端）。

[①]在极特殊的情形下，对于各社会群体（高门、寒门子女），遗漏变量对其“组群因果效应”带来的偏误是完全相同的，那么这种“同质性偏误”不会影响效应异质性的识别。

[②] 技术细节可参考 (Brand et al., 2021; Zhou, 2022)。

$$E[Y_i(1)|z\ ] = E_{C|z}\{E[Y_i(1)|C_i = c, z\ ]\} = \sum\nolimits_C P(C_i = c|z) \cdot E[Y_i(1)|C_i = c, z\ ]$$
$$E[Y_i(0)|z\ ] = E_{C|z}\{E[Y_i(0)|C_i = c, z\ ]\} = \sum\nolimits_C P(C_i = c|z) \cdot E[Y_i(0)|C_i = c, z\ ]$$

据定义，社会群体 $\{Z_i = z\}$ 的组群因果效应 GATE(z)，等于其“接受处理 (d = 1)”以及“不受处理 (d = 0)”分别对应的“平均潜在结果”之差：

$$\begin{aligned} GATE(z) &\equiv E[Y_i(1) - Y_i(0)|Z_i = z] \\ &= E[Y_i(1)|Z_i = z] - E[Y_i(0)|Z_i = z] \\ &= E_{C|z}(E[Y_i|D_i = 1, C_i = c, z\ ] - E[Y_i|D_i = 0, C_i = c, z\ ]) \\ &= \sum\nolimits_C P(C_i = c|Z_i = z) \cdot (E[Y_i|D_i = 1, c, z\ ] - E[Y_i|D_i = 0, c, z\ ]) \end{aligned}$$

至此，在平均意义上，我们已将所有感兴趣的潜在结果和因果效应全部转化为可观测的结果。上述推导的启示是，如果拥有全人群的观测结果（总体数据），且“因果识别条件”得到满足。那么可以直接基于“公式”计算出特定社会群体的潜在结果均值与组群因果效应。可见，因果识别并不涉及具体统计模型的设定。

2. 直接针对潜在结果进行建模

（1）通行思路

实证研究中，即便满足“因果识别条件”，由于样本量有限，往往缺乏足够的自由度，很难使用上述推导出的“理论公式”直接计算出平均潜在结果与组群因果效应。面对“有限样本”的难题，一个通行思路是，针对“潜在结果”进行模型设定、估计函数结构，预测潜在结果，计算平均潜在结果以及组群平均因果效应 (GATE)。这对应于以下五个步骤：

1) 设定函数结构：针对潜在结果进行建模，设定潜在结果关于控制

变量的条件期望的函数形式 (Functional Form)。

2) 估计函数结构：基于样本数据，估计“潜在结果模型的函数表达式”。

3) 预测潜在结果：将样本数据的相应变量信息带入“潜在结果的函数表达式”，预测出“个体层次的潜在结果”。

4)“均化”潜在结果：将上一步的预测值在“更高的群体层次”进行加总平均，得到“群体层次的平均潜在结果”。

5) 计算调节效应：基于“群体层次的平均潜在结果”，计算特定社会群体对应的“组群因果效应”（同时也可以计算出“反事实关联”）。

上述分析思路直接针对“潜在结果”进行模型设定、估计和预测，其最终的计算结果（效应异质性与反事实关联）与本文提出的理论分析框架完全对应[①]。它的突出优势是将“因果识别条件”与“模型设定与估计”做出了明确的分离。在因果识别条件得到满足的前提下，研究者无需局限在传统线性回归模型当中，而可以选择更为灵活的模型——例如，能对控制变量高次项进行筛选的“惩罚回归”(penalized regression)，能自动考虑自变量间交互效应的随机森林 (random forest) 等；这些相对灵活的模型有助于缓解“模型误设”对调节效应分析的误导。由于这些模型或算法的具体介绍已远超出本研究范畴[②]，因此我们仍以线性函数为例，对上述步骤的原理进行说明。

---

①在某些实证场景中，可能很难确定潜在结果与控制变量间的函数关系，而相对更加了解哪些特征的人群更有可能接受处理。此时可以转而针对处理变量进行建模，设定“处理变量关于控制变量的条件期望”为某种函数关系（也即倾向得分模型）；此外，分析者也可以同时设定“潜在结果模型”与“倾向得分模型”。囿于篇幅，详细介绍请参考（Jacob 2021；Knaus et al., 2021；胡安宁，2020：99-106）。

②有关机器学习应用于社会学研究的具体介绍，请参考陈云松等 (2020)。

（2）模型设定与估计的具体步骤

第一步，设定函数结构。这一步是指，将“潜在结果 $Y_i(d)$ 关于控制变量 C_i 的条件期望”设定为特定函数结构。具体地，针对特定社会群体 $\{Z_i = z\}$ 而言，设定这个群体的潜在结果 Y(1) 关于其控制变量的条件期望函数为 $g_1(C_i, z)$，设定该社会群体的潜在结果 Y(0) 关于其控制变量的条件期望函数为 $g_0(C_i, z)$；不妨将这两个条件期望的函数形式都设为常见的参数化线性函数，如下式所示：

$$E[Y_i(1)|C_i, Z_i = z] \Rightarrow g_1(C_i, z) = \alpha_{1,z} + \beta_{1,z}C_i$$
$$E[Y_i(0)|C_i, Z_i = z] \Rightarrow g_0(C_i, z) = \alpha_{0,z} + \beta_{0,z}C_i$$
$$E[Y_i(d)|C_i, Z_i = z] \Rightarrow g_d(C_i, z) = \alpha_{d,z} + \beta_{d,z}C_i$$

其中，$\alpha_{d,z}$ 是条件期望函数中的截距项，$\beta_{d,z}$ 是条件期望函数中的斜率项。第一个下标“d（取值为 0 或 1）[①]”代表该参数刻画的函数结构是针对潜在结果 Y(0) 还是 Y(1)；第二个下标“z”用以代表特定社会群体 $\{Z_i = z\}$。

设定了潜在结果（关于控制变量）的函数形式后，需要估计函数结构中的未知参数 $\{\alpha_{d,z}, \beta_{d,z}\}$ 的取值。一旦估计出这些参数的取值，$g_d(C_i, z)$ 的函数表达式就变成了已知信息，这就意味着，我们只需将相应变量的样本观测值带入该函数，就能“输出”感兴趣的结果。这就进入了第二个步骤。

[①]再次说明，本文中的小写字母 d 均代表“假想的处理状态”，与“潜在结果”相对应；大写字母 D_i 代表的是观测到的（实际的）处理状态；除此以外，其他的小写字母仅表示“随机变量的特定取值”。例如，z 代表调节变量 Z_i 的特定取值，c 代表其他控制变量 C_i 的特定取值。

第二步，估计函数结构。为了估计潜在结果的条件期望函数中的未知参数，需要施加“可忽略性假定”（因果识别条件）：在给定控制变量 C_i 的条件下，潜在结果可以等价转化为观测结果。由此，我们可以基于“实际接受处理 $\{Di = 1\}$ 的样本数据”，使用结果变量 Y 对控制变量 C 进行线性回归[①]，进而估计出潜在结果的条件期望函数 $g_1(C_i, z)$；同理，可以基于“实际没有接受处理 $\{D_i = 0\}$ 的样本数据”，估计出潜在结果的条件期望函数 $g_0(C_i, z)$。

第二步的操作如下所示：

⇒基于样本数据 $\{D_i = 1, Z_i = z\}$，用 Y_i 对 C_i 回归得：$\hat{g}_1(C_i, z) = \hat{\alpha}_{1,z} + \hat{\beta}_{1,z} C_i$

⇒基于样本数据 $\{D_i = 0, Z_i = z\}$，用 Y_i 对 C_i 回归得：$\hat{g}_0(C_i, z) = \hat{\alpha}_{0,z} + \hat{\beta}_{0,z} C_i$

第三步，预测潜在结果。在社会群体 $\{Z_i = z\}$ 这个子样本中，将每个受访者的控制变量观测值 $\{C_i = c\}$ 带入“已知的潜在结果函数表达式”，即可“预测”(Predict) 出每个受访者对应的潜在结果取值，记作 $\hat{g}_{1,i,z}$ 和 $\hat{g}_{0,i,z}$，定义式如下：

$$\hat{g}_{1,i,z} = \hat{g}_1(C_i = c, z) = \hat{\alpha}_{1,z} + \hat{\beta}_{1,z} c$$
$$\hat{g}_{0,i,z} = \hat{g}_0(C_i = c, z) = \hat{\alpha}_{0,z} + \hat{\beta}_{0,z} c$$

需要注意，$\hat{g}_{1,i,z}$ 与 $\hat{g}_{0,i,z}$ 在严格意义上并不代表每位受访者的“个体潜在结果”(individual potential outcome)，而更宜被理解为“个体化的平均潜在结果”(Individualized Average Potential Outcome，IAPO)[②]。其具体含义是：

[①] 由于设定的是线性模型，所以可直接用线性回归的方式估计出未知参数。

[②] 此处的定义参考了(Knaus, 2021)。

对于可观测特征同为 $\{C_i = c, Z_i = z\}$ 这组人的“平均潜在结果”。沿用“格子”(cell) 的比喻，它们就是指在样本数据当中，控制变量 C_i 和调节变量 Z_i 的取值恰好落在 $\{C_i = c, Z_i = z\}$ 这个格子中的所有受访者的平均潜在结果。$\hat{g}_{1,i,z}$与$\hat{g}_{0,i,z}$是颗粒度最细的“平均潜在结果”，它们是估计其他“更高层面”平均潜在结果的基础性要素 (building block)。

第四步，“均化”潜在结果。这一步是在“社会群体 $\{Z_i = z\}$ 的子样本”当中，分别将$\hat{g}_{1,i,z}$以及$\hat{g}_{0,i,z}$进行加总平均，得到颗粒度更粗、加总层面更高（不再包含控制变量 ）的平均潜在结果。若潜在结果与控制变量的函数关系设定正确，则对于社会群体 $\{Z_i = z\}$ 而言， $\hat{g}_{1,i,z}$与$\hat{g}_{0,i,z}$ 的样本均值将会依概率收敛到对应（子）总体的“平均潜在结果”。具体推导如下[①]：

$$\frac{1}{N_z}\sum_{i=1}^{N_z}(\hat{g}_{1,i,z}) \to^{p} E[g_1(C_i, z)|Z_i = z] = E[Yi(1)|Z_i = z]$$

$$\frac{1}{N_z}\sum_{i=1}^{N_z}(\hat{g}_{0,i,z}) \to^{p} E[g_0(C_i, z)|Z_i = z] = E[Yi(0)|Z_i = z]$$

第五步，计算“组群因果效应”、效应异质性以及“反事实关联”。在操作上，我们可以用各个子样本数据 $\{Z_i = z\}$（例如寒门子女 $\{Z_i = 0\}$、高门子女 $\{Z_i = 1\}$ 这两个子样本），逐一运行上述四个步骤，估计出各社会群体的“平均潜在结果” $E[Y_i(d)|Z_i = z]$。

回顾调节效应分析的理论框架，一旦得到各个社会群体的“平均潜在结果”，即可计算“组群因果效应”GATE(z)，效应异质性 ΔGATE，以及“反事实关联” $CA_z[Y(d)]$。它们对应的样本估计 (sample estimates)

[①]其中 N_z 代表的是社会群体 $\{Z_i = z\}$ 对应的子样本的样本量。

形式如下[①]：

$$\hat{G}ATE(z) = \frac{1}{N_z}\sum_{i=1}^{N_z}(\hat{g}_{1,i,z}) - \frac{1}{N_z}\sum_{i=1}^{N_z}(\hat{g}_{0,i,z})$$

$$\hat{\Delta}GATE = \hat{\Delta}GATE\,(Z_i = 1) - \hat{\Delta}GATE\,(Z_i = 0)$$

$$\hat{C}A_z[Y(d)] = \frac{1}{N_{z=1}}\sum_{i=1}^{N_{z=1}}(\hat{g}_{d,i,z=1}) - \frac{1}{N_{z=0}}\sum_{i=1}^{N_{z=0}}(\hat{g}_{d,i,z=0})$$

（3）拓展：变量类型复杂化

在稍加调整后，上述基于二值处理变量、二值调节变量的实证策略，也可以应对于变量类型更为复杂的实证场景，主要包括：(1) 处理变量或（与）调节变量为多分类变量；(2) 处理变量、调节变量各类别对应的（子）样本量相差较为悬殊；(3) 处理变量或（和）调节变量为连续型变量的场景。

首先讨论第一种场景，也即“处理变量与调节变量都为多类别型变量”。从理论上看，如果满足因果识别条件（“可忽略性假定”与“共同支撑假定”），且各社会群体对应的样本数、各社会群体内接受处理与未受控制的样本数都较充足[②]，那么就可以沿用上文介绍的模型估计思路。一般地，若处理变量有 L 个分类，调节变量有 K 个分类，理论上总能将全样本划分为 (L×K) 个子样本，再分别使用每一个子样本数据估

[①]在 Stata 软件当中，可以使用 teffects 命令族来估计处理效应，并且可在主命令后添加选项“pomeans”以获得“平均潜在结果”的估计值。具体操作细节参考 Stata 官方提供的参考手册，对应网页链接为：https://www.stata.com/manuals/causalteffectsintro.pdf

[②]不同研究场景中，控制变量数量、具体估计方法都可能有所不同，因此各组样本的样本数是否充足并没有明确标准。但研究者可以判断，不同组别的样本数相比较而言，各组样本数的差异是否过于悬殊。

计出“针对特定社会群体 z、特定处理状态 d”的平均潜在结果，进而得到各群体的组群因果效应，以及特定处理状态对应的“反事实关联”。

至于第二种场景（子样本的人数差异较为悬殊）和第三种场景（连续型变量），可一并讨论，因为它们面临类似挑战：由于某些子样本 $\{Z_i = z, D_i = d\}$ 缺乏足够多的观测数，此时若仍然采用“分样本估计的方式”，将很难准确估计出“特定社会群体、特定处理状态对应的平均潜在结果” $E[Y_i(d)|Z_i = z]$。

对此，两点应对方案供参考：其一，合并相似类别，然后再用分样本估计。面对多分类的调节变量，考虑将样本数较少的社会群体进行合并；面对连续型调节变量（例如家庭收入），考虑按“分位数排序”(percentile-rank)，转化为类别型变量（例如五分类的家庭收入；详见 Zhou，2022）。其二，放弃“分样本估计”的方式，转而使用全样本来估计潜在结果的函数结构。即，将处理变量、调节变量和控制变量全部当作自变量，使用更灵活的模型（机器学习算法）估计结果变量与这些自变量的函数关系。当实证场景中的处理变量或（和）调节变量从简单的二值变量转向多分类乃至连续变量，研究者需在潜在结果模型设定是否正确（偏差）与估计精度（方差）之间做出权衡取舍。在此意义上讲，上述参考方案各有其局限。我们建议研究者兼用不同的建模方式，并清晰认识到特定方式隐含的约束条件。

此外，在结果变量方面，“基于潜在结果”的估计策略主要适用于连续型或二值因变量[①]。对于不适合简单视作连续变量的定序因变量，以及

[①]对于二值因变量（例如是否发生事件），上述步骤中的“平均潜在结果”可理解为事件在某一组群中发生的概率。

不适合直接转为二值变量的多分类因变量，潜在结果并不必然具有明确定义，因此需更谨慎地使用本文提供的估计策略，以避免陷入新的调节效应陷阱。

（三）两种实证策略的关系

最后需要说明“基于观测结果”与“基于潜在结果”这两种实证策略的关系：其一，就“效应异质性”而言，基于观测结果得到的组群因果效应及效应异质性，是基于潜在结果实证策略的一个特例；其二，若要进一步考察“反事实关联”，那么只要处理变量存在一定程度的“选择性”（非完全随机分配），即使基于观测结果的回归模型设定正确，也难以直接估计出结果变量与调节变量的“反事实关联”。相反，基于潜在结果的实证策略可以直接给出相应回答[①]。

我们用基于观测结果的“交互项模型”来说明上述第 (1) 条结论：寒门子女 $\{Z_i = 0\}$ 的组群因果效应对应于交互项模型当中的 β_1，高门子女 $\{Z_i = 1\}$ 的组群因果效应对应于模型中的 $(\beta_1 + \beta_3)$；容易证明，当潜在结果 Y(d) 关于控制变量 C 的条件期望函数为线性，而且“潜在结果 Y(1) 与变量 C 的（偏）相关性”等于“Y(0) 与变量 C 的（偏）相关性”；那么，潜在结果的条件期望函数 $g_1(C_i, z)$ 与 $g_0(C_i, z)$ 的截距项之差 $(\alpha_{1,z} - \alpha_{0,z})$ 就等于“交互项模型”当中社会群体 $\{Z_i = z\}$ 对应的组群因果效应。如果 $(\alpha_{1,z} - \alpha_{0,z})$ 下标 z 为 1（高门子女），那就等于交互项模型中的 $(\beta_1 + \beta_3)$；如果 $(\alpha_{1,z} - \alpha_{0,z})$ 下标 z 为 0（寒门子女），那就等于交互项模型中的 β_1。

① 这一结论在“基于观测结果的反事实关联分析”一节已进行详细论证，此处不再赘述。

可见，只要对潜在结果的模型设定施加相应约束条件，基于潜在结果的实证策略得出的效应异质性，在数值上等同于“观测结果模型”的分析结果；综上，本研究认为基于潜在结果的实证策略具有相对更好的兼容性、灵活性和扩展性。

（四）实证应用举例

为了更具体地阐释本研究提供的调节效应分析框架及实证策略，本小节提供一个说明性例子——高等教育的心理健康回报及其阶层异质性。近年来，分层研究者关注到，原生家庭社会阶层与个人受教育程度分别作为先赋型和后致型两种资源，它们对于个体心理健康的影响可能存在着“资源补偿效应”。相比于较高阶层出身的人，出身于较低阶层的人能从受教育中获得更多的心理健康回报（Schaan，2014；Anderssonand and Vaughan，2017；常青松等，2024）。即，出身阶层“调节了”受教育程度对个体心理健康的促进效应。

我们使用2018年的中国综合社会调查数据(CGSS)来考察上述调节效应。在参考代表性研究的基础上，为了方便演示，我们对变量与模型设定进行了必要的简化。本例中的处理变量为“个体是否获得高等教育”（二值变量，学历为大专及以上＝1，高中及以下＝0）；结果变量为“心理健康的标准化得分”[①]（连续变量，取值越高说明心理健康水平越高）；调节变量为“原

[①] 问卷中对应问题为“过去四周内，您感到心情抑郁或沮丧的频繁程度”，五点计分，1＝总是，5＝从不。需说明，本文是为了方便演示调节效应分析的实证策略，将该变量直接视作连续变量，转化为均值为0方差为1的标准分(z-score)。更严谨的方式是将其视作定序变量进行建模，详见（常青松等，2024）。

生家庭阶层地位（二值变量，较高阶层 = 1，较低阶层 = 0）”[①]。控制变量选取了受访者的年龄、年龄平方、性别、14 岁时的居住地（农村 = 1，城镇 = 0）。我们将分析样本限定在已完成教育的成年受访者当中。

首先，我们使用分组 OLS 回归以及设置交互项的方式来考察高等教育获得对于心理健康的影响是否存在阶层异质性[②]。结果如表 1 所示。通过比较表 1 的前两列系数，容易发现，高等教育对于出身阶层较低者的心理健康发挥了显著且幅度可观的促进效应，而对于出身阶层较高者的心理健康的正效应非常微弱且并不显著。第三列是全交互模型，也即引入了调节变量（出身阶层）与所有自变量的交互项。此时，出身阶层与高等教育的交互项系数，等价于分组回归中较高、较低阶层的高等教育效应系数的组间差异。这一交互项系数显著为负 ($p < 0.05$)，说明高等教育对心理健康在较高、较低阶层之间的效应差异是统计显著的。

表 1 高等教育与心理健康，原生家庭阶层的调节效应（基于 OLS 估计）

	(1) 较低出身阶层	(2) 较高出身阶层	(3) 全样本
是否接受高等教育	0.127***	0.007	0.127***
	(0.043)	(0.035	(0.043)

①问卷中对应问题为“您认为在您 14 岁时，您的家庭处在哪个等级上”。变量取值为从 1 到 10 的整数。我们基于该变量的分布，将评分位于前 50%（对应于 1~4）的受访者划分为低出身阶层，将取值分布在后 50% 的受访者（对应于 5~10）的受访者划分为高出身阶层。仍然需要强调，这种划分仅是出于应用举例的方便起见，更为严谨的处理方式请参考（常青松等，2024）。

②为了对总效应和调节效应进行因果识别，本例需要施加“可忽略性假定”和“共同支撑假定”，详见第四节的第二小节。

年龄	-0.005	-0.008	-0.005
	(0.006)	(0.005)	(0.006)
年龄平方	0.000	0.000*	0.000
	(0.000)	(0.000)	(0.000)
性别 (女性 =1)	-0.124***	-0.036	-0.124***
	(0.031)	(0.029)	(0.031)
居住地 (农村 =1)	-0.122***	-0.113***	-0.122***
	(0.034)	(0.031)	(0.034)
原生家庭阶层 (较高阶层 =1)			0.100
			(0.201)
高等教育 × 原生家庭阶层			-0.120**
			(0.055)
年龄 × 原生家庭阶层			-0.003
			(0.008)
年龄平方 × 原生家庭阶层			0.088**
			(0.043)
居住地 × 原生家庭阶层			0.009
			(0.046)
常数项	0.303**	0.404***	0.303**
样本数	4,243	4,025	8,268

注：*$p < 0.1$, **$p < 0.05$, ***$p < 0.01$；括号内是异方差稳健标准误。

以上是现有研究中考察调节效应的常见策略。接下来，我们应用本文介绍的“基于潜在结果的实证策略”，展开“反事实关联”分析，以更深入地考察原生家庭阶层的调节效应对于改善心理健康不平等现状的政策意涵。

具体来说，根据上文介绍的针对潜在结果的“五步建模步骤”（详见本节的第二小节），将心理健康的潜在结果关于控制变量的函数形式

设定为简单的线性函数；然后基于样本数据，分别计算出两类群体（高阶层、低阶层），两种状态下（假设所有人都接受高等教育、所有人都不接受高等教育）的四个“平均潜在结果”①，如下表 2（前两行）所示。

表 2 在反事实情形、事实情形下，高低出身阶层间的心理健康差异

三类情形	较低出身阶层 (Z=0)	较高出身阶层 (Z=1)	反事实关联 (2)-(1)	事实关联
假设全都接受高等教育 (d=1)	$\hat{E}[Y_i(1)\|Z_i=0]$ = 0.249	$\hat{E}[Y_i(1)\|Z_i=1]$ = 0.202	$\hat{CA}_Z[Y(1)]$ = -0.047	
假设均未接受高等教育 (d=0)	$\hat{E}[Y_i(0)\|Z_i=0]$ = 0.013	$\hat{E}[Y_i(0)\|Z_i=1]$ = 0.190	$\hat{CA}_Z[Y(0)]$ = 0.177	
实际接受高等教育的比例	$\hat{E}[Y_i\|Z_i=1]$ = 0.027	$\hat{E}[Y_i\|Z_i=1]$ = 0.190		=0.163

基于表 2 的四个潜在结果，可以计算出两种反事实情形下，较高与较低出身阶层的心理健康差异，对应于反事实关联 (1)，也即在所有人都接受高等教育的情形下，较高和较低出身阶层者的心理健康差异为 -0.047。此时，较低出身阶层者的心理健康水平甚至优于较高出身阶层者；反事实关联 (2)，也即在所有人都没有接受高等教育的情形下，较高和较低出身阶层者的心理健康差异为 0.177。此时，较低出身阶层者的心理健康会比那些较高出身阶层者的心理健康状况更差，差距约为 0.18 个标准差。通过比较高等教育从“全无”到“全有”这两种情形下的反事实关联，可以看出接受高等教育在很大程度上缩小（甚至逆转）了高

①如上所述，在 Stata 中可使用 teffects 命令配合 pomeans 选项获得对应的平均潜在结果。

低出身阶层间的心理健康差距。

更进一步，表 2 的第三行提供了样本数据中较高、较低阶层的平均心理健康水平，据此可以得出“事实情形下”较高、较低出身阶层者之间存在的心理健康差距为 0.163，也即“事实关联”。不难看出，如果将高等教育普及到全人群，相比于保持现状（样本数据中的高等教育分布），不仅有助于弥合现存的出身阶层间心理健康的不平等，而且可能带来差异的逆转[①]，使出身阶层较低者的心理健康水平超过出身阶层较高者。此外，如果研究者关注上述差距以及差距缩小量是否具有统计显著性，可以通过自助抽样法来对上述估计结果进行统计检验。

本例也再次说明了处理效应异质性与反事实关联分析的关系：正是因为接受高等教育对于心理健康的促进效应存在鲜明的阶层异质性，且出身阶层较低者的心理健康从高等教育中获益更多，我们才能通过反事实分析发现，普及高等教育能有效改善现有的阶层间心理健康不平等。

五、结论与讨论

（一）结论

本文在因果推断范式下，系统介绍了社会分层研究中的调节效应分析方法。

[①]关于本例中高低阶层间差异出现“逆转”，很可能是因为本例出于演示简便起见，控制变量选取较少、存在遗漏变量偏误所致，也可能是因为直接使用线性模型设定来预测心理健康的潜在结果，带来了较大的预测误差所致。在实际研究中更常见的情形是现有差距被缩小，但差距的方向并未改变。

其一，本文借助“潜在结果”与因果图，明确厘清了调节变量的概念界定，并说明了“好的调节变量”具备的常见特点：前定于处理变量、取值相对稳定、与“社会分组”（例如阶层出身、性别或族群身份等）密切相关。

其二，本文介绍了针对调节效应的常见分析思路。首先，是研究者较为熟悉的“处理效应异质性”分析，它往往涉及到考察不同社会群体对应的“组群处理效应”是否存在差异。我们尤其关心社会劣势群体是否从某项政策干预当中收获更多，从而对他们原有的劣势提供补偿；在此基础上，如果发现“效应异质性”的存在，则可进一步考虑展开“反事实关联”分析，以深入评估“效应异质性”模式对于有效改善社会不平等“现状”所具有的实际政策意涵。

其三，本文详细阐释了“效应异质性”与“反事实关联”的内在逻辑联系，并将它们整合为一条连贯而完整的分析链条，为社会分层领域的调节效应分析提供了一个可供参考的框架。这一分析框架的核心启示是，在一定的条件下，某项“针对处理状态的普及化（平等化）政策”之所以能对（某一维度的）社会不平等的“现状”带来有效的改善，其背后依靠的核心驱力正是不同社会群体的“处理效应异质性”。此外，该分析框架具有广泛的社会学应用场景，有助于研究者更为深入地理解和刻画“效应异质性”对于改善社会分层现状的具体意义。

最后，本文回到实证研究的具体操作层面，讨论了如何使用观测数据来识别和估计“效应异质性”与“反事实关联”，并着重介绍了“基于潜在结果的实证策略”。本文主张的这类实证策略具有较强的扩展性和灵活性，目前研究者常见的“交互项模型”或“分组回归”均可视为

这类实证策略的特例。这类实证策略直接基于潜在结果定义因果效应，明确地刻画了“反事实情形”，并能直接回应理论框架当中关涉的核心问题。分析思路上，它将“因果识别”与“模型估计”两个步骤进行了清晰分离，这有助于研究者注意避免调节效应分析的常见陷阱。

（二）讨论

需要指出，本文未能深入探讨“效应异质性的产生原因”。也即，不同社会群体之间究竟为什么会表现出处理效应异质性？事实上，即使研究者估计得出了类似的效应异质性模式，也可能存在多种“竞争性解释”，而不同的解释又可能对应着不同的具体政策意涵。作为文末讨论，我们简要介绍三种对于效应异质性的常见解释以及相应的实证研究建议[①]。

其一，对于不同的社会群体，驱使其接受处理的诱因不同，进而使其受处理之后的生命轨迹也产生分异。仍以高等教育为例。对寒门子女而言，驱动其接受高等教育的主要诱因可能是“经济激励”(economic incentive)；对于高门子女而言，主要驱动因素可能是“文化期待”(cultural expectation)。由于驱使接受高等教育的诱因不同，寒门子女在选择专业时更可能会选择“与收入回报关联更强”的经济管理等专业；而高门子女更有可能选择“短期经济回报并不明显”的基础学科。这在一定程度上能解释高等教育的收入回报异质性。可见，上述解释关注不同社会群体从处理变量到结果变量的“因果路径”（渠道）的分异，这进而导致

[①]严格来说，以下三种可能情形并不一定是互斥关系，本文将之做并列介绍是为了方便读者理解。

不同群体的组群处理效应表现出差异。目前，基于因果路径而展开的异质性分析思路已发展出较完备的概念框架和实证方法 (Zhou, 2022)。

其二，不同社会群体对处理状态的“获取方式”或“接受处理的质量 / 类型”可能不同，而正是不同的方式或质量 / 类型，塑造了因果效应的差异。例如，由于家境劣势，寒门子女往往要更积极努力地主动争取上大学的机会；而相比之下，家庭背景占优的高门子女可能拥有更多资源和途径，进而更轻松地获得高等教育。而寒门子女收入回报较高，可能正是由于寒门与高门子女“获取大学教育的方式差异”导致的结果[①]。类似地，如果发现寒门子弟上大学的收入回报反而更低，另一种可能解释是寒门子弟与高门子弟接受的高等教育的质量 / 类型不同，例如重点与非重点院校，本科与专科之别。

这启发研究者，需要紧密结合社会分层理论与经验观察，对处理变量做出更细致的操作化定义，审慎地考虑二分类测量方案的合理性。

还有一种可能是“基于可忽略性假定的因果识别条件”并未得到满足。换言之，基于“可观测特征”（调节变量）展现出的效应异质性，其实反映的是“未能观测的特征”引发的效应异质性。例如，个体能力越强，上大学的收入回报相对更高；同时，成功进入大学的寒门子女往往能力也较强。如若控制变量中遗漏了个体能力，那么寒门子女上大学的高回报可能是由于“未观测到的能力”在寒门子女组内带来的“高估偏误”

[①]由处理状态的获取方式差异而带来的效应异质性在分层研究中并不少见。例如考察城市户口获得对收入的影响，研究者发现只有基于个人努力而实现的户口转化（如上大学、参军等），才对收入有促进作用；而通过征地与移民安置实现的政策性户口转化，对收入并无因果影响 (Wu and Zheng 2018)。

更大所致。就观测性研究而言，一般无法证明必然满足“可忽略性假定”，但研究者可针对估计出的效应异质性进行“敏感性分析”(Brand et al., 2021; Zhou, 2022)，考察“未观测到的混杂因素”对“效应异质性”带来多大程度的威胁；另外，若能找寻到有效的工具变量 (valid instrument)，则可考虑估计“边际处理效应”(Marginal Treatment Effect, MTE)，它可以同时展示处理效应与“可观测特征”和“未观测特征”的关系 (Zhou and Xie, 2019)。

在社会分层研究领域，调节效应分析具有可观的应用前景，调节分析的结论可能对推动社会平等、促进社会流动带来宝贵的政策意涵。当下，分层研究中的调节效应在概念界定、理论解释与实证策略方面，都还存在着巨大的发展潜力，希望本文能起到抛砖引玉之用，为更加成熟完善的调节效应研究提供启发。

参考文献

常青松、胡景梁、刘子曦 . (2024). 原生家庭社会阶层如何影响教育的精神健康回报：资源补偿还是强化？*社会 , 1,* 213-234.

陈云松、范晓光 . (2010). 社会学定量分析中的内生性问题：测估社会互动的因果效应研究综述 . *社会 , 4,* 91-117.

陈云松、吴晓刚、胡安宁、贺光烨、句国栋 . (2020). 社会预测 : 基于机器学习的研究新范式 . *社会学研究 , 3,* 94-117+244.

胡安宁 . (2012). 倾向值匹配与因果推论 : 方法论述评 . *社会学研究 , 1,* 221-242+246.

胡安宁 . (2020). *应用统计因果推论* . 上海 : 复旦大学出版社 .

胡安宁、吴晓刚、陈云松 . (2021). 处理效应异质性分析——机器学习方法带来的机遇与挑战 . *社会学研究 , 1,* 91-114+228.

江艇 . (2022). 因果推断经验研究中的中介效应与调节效应 . *中国工业经济 , 5,* 100-120.

句国栋、陈云松 . (2022). 图形的逻辑力量 : 因果图的概念及其应用 . *社会 , 3,* 195-221.

李昂然 . (2022). 中国教育资源市场化与个体选择 : 初中课外补习效应异质性探究 . *社会 , 2,* 94-125.

李适源、刘爱玉 . (2022). “忧郁的孩子们” : 课外补习会带来负向情绪吗？基于中国教育追踪调查 (CEPS) 两期数据的因果推断 . *社会 , 2,* 60-93.

彭玉生 . (2011). 社会科学中的因果分析 . *社会学研究 , 3,* 1-32+243.

石磊 . (2022). 中国代际社会流动的变迁——基于多重机制的分析 . *社会学研究 , 2,* 156-178+229.

朱家祥、张文睿 . (2021). 调节效应的陷阱 . *经济学季刊 , 5,* 1867-1876.

Andersson, M. A., & Vaughan, K. (2017). Adult health returns to education by key childhood social and economic indicators: Results from representative European data. *SSM-Population Health, 3,* 411-418.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Brand, J. E., & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75*(2), 273-302.

Brand, J. E., Xu, J., Koch, B., & Geraldo, P. (2021). Uncovering sociological effect heterogeneity using tree-based machine learning. *Sociological Methodology, 51*(2), 189-223.

Breen, R. (2010). Educational expansion and social mobility in the 20th century. *Social Forces, 89*(2), 365-388.

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences, 113*(31), 8664-8668.

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., & Roberts, B. W. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adulthood. *Journal of Personality and Social Psychology, 109,* 473-489.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology, 36*(1), 21-47.

Grusky, D. B. (Ed.). (2019). *Social stratification: Class, race, and gender in sociological perspective.* New York: Routledge.

Hong, Y. (2020). Foundations of modern econometrics: *A unified approach.* New Jersey: World Scientific.

Hout, M. (1988). More universalism, less structural mobility: The American occupational structure in the 1980s. *American Journal of Sociology, 93*(6), 1358-1400.

Hout, M., & DiPrete, T. A. (2006). What we have learned: RC28's contributions to knowledge about social stratification. *Research in Social Stratification and Mobility, 24*(1), 1-20.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5-86.

Jacob, D. (2021). CATE meets ML-Conditional average treatment effect and machine learning. *arXiv:2104.09935 [econ.EM].*

Karlson, K. B. (2019). College as equalizer? Testing the selectivity hypothesis. *Social Science Research, 80,* 216-229.

Knaus, M. C. (2021). Double machine learning based program evaluation under unconfoundedness. *arXiv:2003.03191 [econ].*

Knaus, M., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal, 24*(1), 134-161.

Liu, A. (2019). Can non-cognitive skills compensate for background disadvantage? The moderation of non-cognitive skills on family socioeconomic status and achievement during early childhood and early adolescence. *Social Science Research, 83,* 102306.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference.* Cambridge University Press.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Pfeffer, F. T., & Hertel, F. R. (2015). How has educational expansion shaped social mobility trends in the United States? *Social Forces, 94*(1), 143-180.

Rubin, D. B. (2011). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100*(469), 322-331.

Schaan, B. (2014). The interaction of family background and personal education on depressive symptoms in later life. *Social Science & Medicine, 102,* 94-102.

Shanahan, M. J., Bauldry, S., Roberts, B. W., Macmillan, R., & Russo, R. (2014). Personality and the reproduction of social class. *Social Forces, 93*(1), 209-240.

Torche, F. (2011). Is a college degree still the great equalizer? Intergenerational mobility across levels of schooling in the United States. *American Journal of Sociology, 117*(3), 763-807.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

Witteveen, D., & Attewell, P. (2017). Family background and earnings inequality among college graduates. *Social Forces, 95*(4), 1539-1576.

Wodtke, G. T. (2020). Regression-based adjustment for time-varying confounders. *Sociological Methods and Research, 49*(4), 906-946.

Wodtke, G. T., & Almirall, D. (2017). Estimating moderated causal effects with time-varying treatments and time-varying moderators: Structural nested mean models and regression with residuals. *Sociological