DOI:10.19744/j.cnki.11-1235/f.2025.0136

研究方法

实证研究中的控制变量选择:原理与原则*

张子尧　黄　炜

摘要:控制变量的选择是决定因果推断类实证研究有效性的关键环节。本文系统阐释控制变量在观测性实证研究中的作用原理、选择标准、使用原则与实践建议。利用潜在结果框架和线性回归模型估计量分解,明确控制变量在因果识别和统计推断两个核心环节的重要作用。在因果识别环节,"好"控制变量通过合理分层让观测性数据在局部尽可能近似于随机化实验,而"坏"控制变量会引入选择性偏误。在统计推断环节,"好"控制变量有助于减少数据噪音干扰,提高估计精度,而"坏"控制变量则会放大估计误差、降低统计功效。控制变量的"好"与"坏"取决于其在因果结构中的位置,而研究者对因果结构的理解来源于社会科学理论和现实制度背景,故控制变量的选择应由理论驱动而非数据驱动。在此基础上归纳实证研究中若干类常见控制变量的分类判别方法,总结提炼控制变量使用的5个基本原则:基于因果结构选择控制变量、高度重视坏控制变量问题、关注控制变量的重叠性、在复杂情况下权衡控制变量"利""弊",以及避免过度解读控制变量系数,并基于上述原则提出具体的实践建议。本文为社会科学实证研究者优化研究设计提供了富有操作性的方法论框架,对提升实证研究的可信性、透明性和可复制性具有指导意义和参考价值。

关键词:控制变量　因果推断　观测性研究　选择性偏误

一、引言

严谨可信的实证研究是学术进步的基础和政策制定的重要依据。因果推断作为实证研究的核心目标,旨在从观测数据中识别和量化变量之间的因果关系,为理解复杂社会现象和制定有效的公共政策提供科学支撑。自费希尔(1935)提出并倡导使用随机化实验方法研究因果关系以来,随机化实验已经成为公认的因果推断"黄金标准"。但是,社会科学的许多重要问题往往不适合进行随机化实验①。与此同时,使用观测性数据(observational data)进行因果推断研究的可信性仍然存在许多担忧和质疑(坎贝尔、厄尔巴赫,1970;利默,1983;拉隆德,1986)。鲁宾(1974)开创性地提出通过精心合理的研究设计(research design),使用观测性数据同样可以进行可信的因果推断研究②。以此为起点,基于设计的研究范式(design-based approach)逐渐成为社会科学实证研究的主流研究范式(安格里斯特、皮施克,2010;赫尔等,2022;许琪,2024;陈强,2025)。

实证研究常用的因果推断诸方法中,控制变量无疑是使用最为广泛的基础性工具之一。通过合理地引入控制变量,研究者得以在观测性研究中最大程度地消除遗漏变量的干扰,更加准确地估计变量间的因果效应。然而,尽管控制变量在实证研究中扮演着举足轻重的角色,但长期以来,如何科学合理地选择控制变量却缺乏明确的标准和规范。这种选择标准上的模糊性导致了当前经济学和管理学等诸多社会科学领域的实证研究中控制变量选择的随意性和主观性问题日益突出,无依据地随意添加和盲目堆砌控制变量,按既有研究惯例而机械照搬控制变量(怀特德等,2022),甚至为了迎合预期研究结论而人为筛选控制变量等现象屡见不

收稿时间:2025-3-4;反馈外审意见时间:2025-7-9;拟录用时间:2025-8-26。

*本项研究得到国家社会科学基金重大专项"生育友好型社会背景下生育支持政策体系和激励机制研究"(项目编号:24ZDA091)、国家社会科学基金重大专项"完善收入分配制度的理论建构和制度优化研究"(项目编号:24ZDA087)、国家自然科学基金面上项目"人力资本外部性与经济高质量增长:现象、机制和影响"(项目编号:72373003)、国家自然科学基金青年科学基金项目"新发展阶段下市场结构变迁与资本劳动要素收入分配关系研究"(项目编号:72403253)的资助。感谢黄文佳出色的研究助理工作。黄炜为本文通讯作者。

鲜(布罗德等,2016;布罗德等,2020;米顿,2022)[③]。不规范的控制变量选择和使用行为严重削弱了实证研究结果的可信度和科学价值,阻碍了相关学科的知识累积和进步。20世纪80年代以来,以因果推断为核心的"可信性革命"深刻地推动了实证研究范式的革新(安格里斯特、皮施克,2010;王美今、林建浩,2012)。在现代因果推断方法的视角下,研究者们对控制变量的作用机理有了全新的理解,服务于因果推断的控制变量选择准则得到广泛认同(鲁宾,2008;因本斯,2015;奇内利等,2022;因本斯、徐,2024)。在此背景下,重新审视控制变量在实证研究中的作用机理和选择原则,总结出切实可行的控制变量实践规范,对于提高经济学、管理学等社会科学领域实证研究的规范性和可信性具有重要的理论意义和实践价值。

实证研究的基本目的是将理论命题转化为可检验的经验假设,并利用观测数据进行检验与评估(赫克曼,2000)。因果推断的目的就是通过观测到的样本数据识别和估计出总体中变量间的因果关系,以检验和回应理论命题。因此,完整的因果推断类实证研究必然要经过两个环节(见图1)。第一,从样本相关性过渡到总体相关性。具体来说,样本中呈现的变量间的相关关系在总体中是否仍然存在?这是统计推断(statistical inference)环节需要回答的问题。第二,从总体相关性过渡到总体因果性。具体而言,如果在总体中变量间存在相关关系,这种相关关系能否能被识别为因果关系?这是因果识别(causal identification)环节需要回答的问题。

本文借助潜在结果框架(因本斯、鲁宾,2015)和因果图(珀尔,2009)两个强有力的因果推断工具,在基于设计的研究范式下系统地分析了控制变量在因果识别和统计推断两个实证研究核心环节的重要作用和内在机理。本文的核心观点在于,控制变量的"好"与"坏"并非是绝对的,而是内生于特定的因果结构。在识别环节,好控制变量(good control)能够对观测性数据进行合理分组从而近似于分块随机化实验(blocked randomized experiment),以实现消除选择性偏误的核心功能。对撞变量(collider)、中介变量(mediator)等坏控制变量(bad control)则会破坏近似分块随机化实验的有效性,不但无法消除观测性数据原本所固有的选择性偏误,还会人为引入新的选择性偏误,进一步放大因果效应的估计偏误。在统计推断环节,本文通过理论分析和数值模拟方法,证明好控制变量有助于剔除其他因素等数据噪音的干扰,提高系数估计精度和推断效力,而坏控制变量则会放大估计误差、降低统计功效。综上所述,对于准确估计因果效应这一实证研究的核心目标而言,避免坏控制变量和选取好控制变量具有同等重要地位,研究者们需要像重视遗漏变量问题一样重视坏控制变量问题。

基于上述理论分析和实证模拟,本文进一步归纳和总结了实证研究中若干类常见控制变量的分类判别方法,并强调控制变量的选择应以理论驱动而非数据驱动。在此基础上,本文提炼出实证研究中控制变量选择的5个基本原则,包括:基于因果结构选择控制变量、高度重视坏控制变量问题、注意控制变量的重叠性、复杂情况下的控制变量"利""弊"权衡、避免过度解读控制变量系数。基于上述原则,本文进一步总结了实践中的控制变量使用建议。以上原则和建议旨在为社会科学实证研究者提供一套操作性强、易于掌握的指导性框架,能够帮助研究者在实证研究中更加科学合理地选择和使用控制变量,也可以用于分析特定控制变量的作用和合理性,最终提升研究设计的严谨性、研究过程的透明性和研究结果的可复现性,实现科学知识的有效积累。

本文的边际贡献主要体现在以下3个方面。第一,本文借助现代因果推断理论进展,遵循"让观测性研究逼近随机化实验"的核心精神(鲁宾,2008),系统阐释了在观测性研究中如何正确利用控制变量以实现模拟随机化实验。相较于奇内利等(2022)等已有文献侧重于理论分类与因果图分析,本文更关注控制变量在模拟随机化实验中的作用,系统阐述了控制变量如何通过合理的分组方式构造出可比的控制组,从而在局部上近似于随机化实验。这一视角

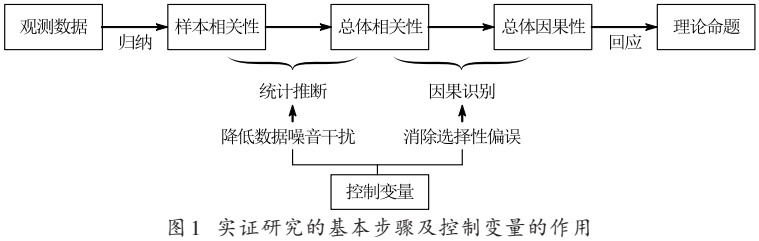


图1 实证研究的基本步骤及控制变量的作用

为研究者提供了更为直观的思维框架，能够更好地理解和体现控制变量如何服务于因果识别的根本目的，并为甄别和诊断研究设计中的潜在缺陷提供了清晰的基准。第二，不同于既有文献侧重于非参数因果推断方法下的识别问题，本文聚焦于线性回归模型这一最常用的实证工具，详细分析了控制变量对线性回归模型中的估计系数产生影响的途径，明确控制变量不仅会通过消除或引入选择性偏误影响估计结果，还会通过影响分块权重和改变有效样本范围两个经常被忽略的途径影响估计结果。区分3种途径的不同影响对于正确理解线性回归模型中控制变量的作用至关重要。第三，紧密结合当前我国经济学与管理学等社会科学领域的实证研究现状，归纳总结针对性的原则与建议。针对当前我国经济学、管理学等社会科学领域的实证研究中滥用、误用控制变量等问题日益严重的现状，本文立足于提升本土实证研究质量的根本目的，将实践中亟待强调的要点提炼为核心原则，旨在对当前研究中存在的机械照搬或盲目堆砌控制变量的现象起到警示作用，提醒研究者们应正确地使用和解读控制变量的作用。期望本文总结的原则与建议能够有助于提升实证研究的质量和可信度，增强社会科学实证研究结论的严肃性和参考价值，减少p值操纵等违反学术研究规范的不当行为，促进我国经济学、管理学等社会科学领域实证研究的健康发展。

本文余下部分的结构安排如下：第二部分阐释控制变量的理论基础；第三部分深入解析坏控制变量的作用机制；第四部分系统地提出控制变量的选择标准；第五部分总结实证研究中使用控制变量的基本原则和实践建议；第六部分总结全文并展望未来研究方向。

二、控制变量的理论基础

本节首先简要回顾基于模型的研究范式（model-based approach）中线性回归模型与控制变量作用，总结其面临的挑战。而后详细介绍基于设计的研究范式对线性回归模型中估计系数的理解方式，分析控制变量的作用原理，明确实证研究中控制变量在识别和推断两个核心步骤中发挥的重要作用。

（一）基于模型的研究范式中的控制变量作用原理

1.FWL定理与控制变量作用原理

假设结果变量Y_i的数据生成过程（data generating process，DGP）为线性结构模型：

$$Y_i=a+bD_i+cX_i+\varepsilon_i \tag{1}$$

影响结果变量Y的原因变量分为3类：一是研究者感兴趣的特定原因变量D（也被称为处理变量）；二是影响Y且与D存在相关性的变量X（即$\mathrm{cov}(D_i,X_i)\neq0$）；三是影响$Y_i$且与$D_i$无关的因素，它们被整体打包在扰动项$\varepsilon_i$之中（即$\mathrm{cov}(D_i,\varepsilon_i)=0$）。这里的“影响”一词描述的是单向的因果关系而非双向的统计相关关系。线性结构模型式中的参数称为结构参数（structural parameter）或因果参数（causal parameter），代表处理变量对结果变量的因果效应量度。在基于模型的结构式计量经济学视角下，DGP中的参数或其某种组合衡量了变量间的因果效应，估计因果效应等价于估计模型参数。

在实证研究中研究者通常会构建线性回归模型（linear regression model）来量化处理变量D_i和结果变量Y_i的相关性。例如，一元线性回归模型为：

$$Y_i=\alpha+\beta D_i+u_i \tag{2}$$

最常用的参数估计方法是最小二乘法（ordinary least squares，OLS），其中β的OLS估计量（OLS estimator）为：

$$\hat{\beta}^{ols}=\frac{\mathrm{cov}(Y_i,D_i)}{\mathrm{var}(D_i)} \tag{3}$$

若$\hat{\beta}^{ols}$的概率极限是因果参数b，则称OLS估计量为一致估计量。将线性结构模型代入上式，可以得到OLS估计量$\hat{\beta}^{ols}$和因果参数b之间的关系为：

$$\hat{\beta}^{ols}=\frac{\mathrm{cov}(Y_i,D_i)}{\mathrm{var}(D_i)}=\frac{\mathrm{cov}(a+bD_i+cX_i+\varepsilon_i,D_i)}{\mathrm{var}(D_i)}=b+c\times\underbrace{\frac{\mathrm{cov}(X_i,D_i)}{\mathrm{var}(D_i)}}_{\lambda} \tag{4}$$

上式即线性回归模型中OLS估计量的遗漏变量偏误(omitted variables bias,OVB)公式(安格里斯特、皮施克,2009)。该公式表明$\hat{\beta}^{ols}$等于因果参数b加上一个偏误项$c\times\lambda$,其中c为X_i对Y_i的因果效应,λ是X_i对D_i做回归的OLS估计系数。偏误项等于0有两种情形:一是X_i不影响Y_i(c=0);二是X_i与D_i不相关(λ=0)。OVB公式揭示了OLS估计量产生偏误的内在机制:若模型遗漏了Y_i与D_i的共同原因X_i,估计系数除了因果效应外,还包含共同原因X_i导致的非因果相关性。在这种情况下OLS估计量偏离真实的因果参数,不能被解释为因果效应。

当外生性假设$E[u_i|D_i]$=0成立时,此时不存在遗漏变量偏误。绝大多数基于观测性数据的实证研究中都不满足外生性假设。任何结果变量Y_i都会有大量的影响因素,而这些因素之间通常又有着很强的相关性,因此几乎所有的观测性研究都会存在或多或少的遗漏变量。为了在线性回归模型中控制遗漏变量的干扰,可以把这些变量加入到多元线性回归模型中:

$$Y_i=\alpha^l+\beta^l D_i+\gamma^l X_i+u^l_i \tag{5}$$

若满足以下条件外生性假设(conditional exogeneity assumption):

$$E[u^l_i|D_i,X_i]=E[u^l_i|X_i]=0 \tag{6}$$

可以证明式(5)中的估计系数$\hat{\beta}^l$是因果参数b的一致估计。借助FWL定理(Frisch-Waugh-Lovell Theorem)可以更为直观地理解多元线性回归模型的控制变量如何发挥消除估计偏误的作用。FWL定理表明,形如(5)式的多元线性回归模型中的估计系数$\hat{\beta}^l$可以分为两步估计:首先,分别将Y_i和D_i对控制变量X_i做回归,残差记为$\tilde{Y}_i$和$\tilde{D}_i$;而后,将残差$\tilde{Y}_i$对$\tilde{D}_i$进行回归:

$$\tilde{Y}_i=\alpha^s+\beta^s\tilde{D}_i+u^s_i \tag{7}$$

我们将形如(5)式的多元线性回归模型称为长回归(long regression),形如(7)式的一元线性回归模型称为短回归(short regression)。短回归中$\tilde{D}_i$的OLS估计系数为:

$$\hat{\beta}^s=\frac{\mathrm{cov}\left(\tilde{Y}_i,\tilde{D}_i\right)}{\mathrm{var}\left(\tilde{D}_i\right)} \tag{8}$$

FWL定理证明了$\hat{\beta}^s$与长回归方程中的估计系数$\hat{\beta}^l$相等。根据FWL定理,在回归方程中加入控制变量可以看作是一种"剔除"遗漏变量影响的方法。根据线性回归模型的基本原理,被解释变量的变动性可以分解为可解释部分(拟合值)和不可解释部分(残差),因此,Y_i和D_i对控制变量X_i做回归后得到的残差$\tilde{Y}_i$和$\tilde{D}_i$与控制变量X_i无关,这相当于从原始的Y_i和D_i中"剔除"掉了与控制变量X_i相关的部分,只保留下了与其无关的残差$\tilde{Y}_i$和$\tilde{D}_i$。第二步的估计系数$\hat{\beta}^s$衡量了$\tilde{Y}_i$和$\tilde{D}_i$的相关性,由于$\tilde{Y}_i$和$\tilde{D}_i$与X_i无关,那么这种相关性不可能是由于遗漏X_i所导致的伪相关关系,只可能是来自于D_i和Y_i之间所蕴含的因果关系。

2.传统实证研究范式面临的一些潜在挑战

自20世纪20年代计量经济学这门学科诞生起,直到80年代"可信性革命"(安格里斯特、皮施克,2010)之前,基于模型的研究范式一直是计量经济学最主流的研究范式。然而,对于计量经济学方法论的争论和质疑从凯恩斯(1939)一直发展延续至今。

首先,传统范式缺乏对因果关系的明确定义。内曼于1923年在他的博士论文中已经初步提出了潜在结果的概念(内曼,1923),哈维尔莫(1944)进一步引入干预性概率模型的思想实验形式,但这些方法并未在早期计量经济学体系中得到广泛重视。传统方法往往试图通过结构模型中的参数来定义因果效应,这种做法在本质上混淆了因果效应本身、估计目标与估计量之间的区别。从现代因果推断的角度来看,结构参数只是对因果效应的某种函数表示,而非其本体。当存在模型误设问题时,结构参数的含义变得模糊难解。例如,线性结构模型中的个体因果效应存在异质性,即真实数据生成过程为:

$$Y_i=a+b_iD_i+cX_i+\varepsilon_i \tag{9}$$

在这种情形下,如何解释线性回归模型的OLS估计系数β^l成为一个棘手的问题。换言之,在异质性因果效应的情形下,β^l的估计目标并不清楚,也就无法评判估计量是否存在偏误。总之,缺乏明确的因果关系定义方式可能导致研究者们对估计结果的解读产生了一些根本性混淆。

其次，该范式未能清晰地区分处理变量与控制变量的理论角色，这也是与本文最相关的一点。虽然研究者可以人为地将处理变量和控制变量进行区分，但是线性回归模型中处理变量和控制变量都是方程右侧的解释变量，没有体现出明显的角色区别。变量地位的模糊性使得许多研究者在选择控制变量时，主要是基于处理变量和备选控制变量的相关性关系而非因果性关系来判断是否应该对其进行控制。然而，基于相关性选择控制变量会面临一个悖论：如果研究者认为X是D的遗漏变量，那反过来也能够说D是X的遗漏变量。至少从计量模型中的角色和相关性关系来看两者是对称的，没有明显差异[④]。

可以说，上述问题仍未超出凯恩斯（1939）对计量经济学的批判。从某种意义上讲，计量经济学的发展正是伴随着对上述难题的不断探索而逐步取得的。直到20世纪80年代起，随着计量经济学领域经历了“可信性革命”的重要学术研究范式变革，现代因果推断方法开始成为实证研究的方法论基础，基于设计的研究范式成为新的实证研究主流范式。

（二）基于设计的研究范式中的控制变量作用原理

现代因果推断方法建立在潜在结果框架（potential outcomes framework）这一强有力的因果推断工具的基础上[⑤]。采用现代因果推断方法的实证研究范式强调良好的研究设计对于因果效应识别和实证研究结果可信性的核心作用（赫尔等，2022），因此该范式亦被称为基于设计的研究范式（design-based approach）。鲁宾（2008）指出基于设计的研究范式核心精神是“让观测性研究逼近随机化实验”，在这种精神的指导下，线性回归方法以及控制变量在其中的作用都呈现出了全新的理解。接下来本文首先简要介绍在潜在结果框架下的因果效应定义与识别假设，然后从研究设计的视角阐述在线性回归模型中添加控制变量的作用原理。

1.潜在结果框架下的因果效应定义与识别挑战

在潜在结果框架下，个体接受处理时的潜在结果为$Y_i(1)$，未接受处理时的潜在结果为$Y_i(0)$。处理变量D_i对结果变量Y_i的个体因果效应τ_i定义为两个潜在结果之差：

$$\tau_i \equiv Y_i(1)-Y_i(0) \tag{10}$$

相应的，群体的平均因果效应是群体内所有个体因果效应的平均值：

$$ATE \equiv E[\tau_i]=E[Y_i(1)-Y_i(0)] \tag{11}$$

类似的，也可以定义受处理群体的平均处理效应（average treatment effects of treated，ATT）和未受处理群体的平均处理效应（average treatment effects of untreated，ATU）。

使用潜在结果框架定义因果效应的优势有以下几点。第一，精确定义不同类型的因果效应，明确实证研究的估计目标。第二，定义不依赖于具体的结构模型函数形式，区分开了结构模型参数和因果效应之间的区别。第三，揭示了因果推断面临的核心挑战。因果效应定义为两个潜在结果之差，而研究者在现实中至多只能够观测到其中一个潜在结果，所以因果效应永远无法被直接观测到[⑥]。因此，因果推断问题从本质上讲是一个数据缺失问题（missing data problem），而因果推断的核心任务就是寻找恰当的方法，从已观测到的数据中挑选合适的对象为缺失的潜在结果作插值（因本斯、鲁宾，2015）。

估计因果效应最直观的估计方法是比较处理组与对照组的平均结果（Simple Difference in Means，SDM），根据潜在结果框架可得：

$$\begin{aligned} SDM &= \underbrace{E\left[Y_i(1)\middle|D_i=1\right]-E\left[Y_i(0)\middle|D_i=1\right]}_{ATT}+\underbrace{\left\{E\left[Y_i(0)\middle|D_i=1\right]-E\left[Y_i(0)\middle|D_i=0\right]\right\}}_{\text{选择性偏误}} \\ SDM &= \underbrace{E\left[Y_i(1)\middle|D_i=0\right]-E\left[Y_i(0)\middle|D_i=0\right]}_{ATU}+\underbrace{\left\{E\left[Y_i(1)\middle|D_i=1\right]-E\left[Y_i(1)\middle|D_i=0\right]\right\}}_{\text{选择性偏误}} \end{aligned} \tag{12}$$

上式在因果推断中的关键作用在于将可观测数据和研究者感兴趣的特定估计目标联系了起来，等式左侧为可观测的组间均值差异，等式右侧为研究者关心的估计目标和选择性偏误项。消除选择性偏误需要满足如下识别假设。

识别假设一：独立性假设（independence assumption）。

观测样本$\{Y_i(1),Y_i(0),D_i\}_{i=1}^n$互相独立且均服从同一概率分布$\{Y(1),Y(0),D\}$，该概率分布满足$\{Y(1),Y(0)\}\perp D$。

独立性假设意味着处理状态是随机分配的，故处理组和控制组在各项特征上均不存在显著差异。当独立性假设成立时，选择性偏误项被消除，处理组和控制组的简单均值差异即可反映处理变量对结果变量的因果效应。然而除了随机对照实验等特殊研究方法外，独立性假设很难得到满足。特别是在目前经济学、管理学等社会科学实证研究中占绝对主流的观测性研究，处理状态往往不是随机分配的，而是直接或间接地受到个体自选择行为的影响。因此，由于个人自选择行为所导致的处理状态非随机分配问题是观测性研究面临的最大挑战⑦。

2.控制变量与分块随机化实验

虽然独立性假设在观测性数据中成立条件过于苛刻，但这并不意味着使用观测性研究无法进行严谨的因果推断。其解决之道在于，将比较的范围从全样本缩小到具有可比性的子样本中。为此需要引入新的识别假设。

识别假设二：条件独立性假设（conditional independence assumption，CIA）。

观测样本$\{Y_i(1),Y_i(0),X_i,D_i\}_{i=1}^n$互相独立且均服从同一概率分布$\{Y(1),Y(0),X,D\}$，该概率分布满足$\{Y(1),Y(0)\}\perp D|X$。

条件独立性假设在一些文献中也被称为可忽略性假设（ignorability）或无混杂性假设（unconfoundedness）。条件独立性假设成立意味着，对于控制变量X取值相等的个体，它们受到处理的概率是相同的。罗森鲍姆和鲁宾（1983）将个体接受处理的概率表示为控制变量X的函数，定义$e(X)\equiv\Pr\{D=1|X\}$为倾向得分（propensity score）。如果将控制变量取值相同的个体划分到一个子样本内，不妨把这个子样本称为一个分块（block），在每个分块内部所有个体接受处理的概率是相同的，分块内处理状态近似于随机分配。此时有：

$$\begin{aligned}E\left[Y_i(0)|D_i=1,X\right]&=E\left[Y_i(0)|D_i=0,X\right]=E\left[Y_i(0)|X\right]\\E\left[Y_i(1)|D_i=1,X\right]&=E\left[Y_i(1)|D_i=0,X\right]=E\left[Y_i(1)|X\right]\end{aligned}\tag{13}$$

上式意味着在每个分块内，处理组和控制组具有相同的潜在结果，故可以使用可观测结果对不可观测的潜在结果作插值，从而估计出分块内部的条件平均处理效应：

$$\tau(X)\equiv E[Y_i(1)-Y_i(0)|X]=E[Y_i(1)|X]-E[Y_i(0)|X]=E[Y_i(1)|D_i=1,X]-E[Y_i(0)|D_i=0,X]\tag{14}$$

使用条件独立性假设来识别因果效应时，还需要另一个重要的但常常被忽略的重要假设“重叠性假设”。

识别假设三：重叠性假设（overlap assumption）。

对控制变量X的任何可能取值$X=x$，倾向得分$e(X)$均满足$0<e(X)<1$。

在一些文献中重叠性假设也被称为共同支撑假设（common support assumption）。重叠性假设意味着，无论控制变量X的取值为何，个体既可能被分配到处理组，也有可能被分配到对照组。重叠性假设确保了对于任意处理组个体都能够找到特征相似的控制组作为插值对象。若倾向得分$e(X)$等于0或1，即$X=x$的分块内只存在处理组或控制组个体，由于无法找到合适的比较对象，$\tau(X)$无法被识别，故该部分样本被排除在因果效应的估计过程外。

在识别出条件平均处理效应$\tau(X)$后，根据每个分块占总样本权重$\Pr(X=x)$进行加权即可得到平均处理效应ATE的估计值：

$$ATE=\sum\nolimits_{x:0<e_x<1}\frac{\Pr(X=x)}{\sum_{x:0<e_x<1}\Pr(X=x)}\tau(X)\tag{15}$$

上述估计方法即是基于控制变量X的精确匹配估计量（exact matching estimator）。事实上这种估计方法对应着特定的研究设计——分块随机化实验。通过将全样本按照控制变量取值划分为多个分块，在CIA假设成立的前提下，每一个分块内部都近似于一个随机化实验。这样，通过良好的研究设计，整个观测性研究被分解成了多个小型随机化实验，研究者只需估计出每个小型随机化实验的因果效应，再依据估计目标选择不同的加权方式，即可得到所需要的因果效应估计结果。通过上述研究设计过程，研究者们成功地使得观测性研究

逼近了随机化实验。

3. 线性回归模型中的控制变量与研究设计

现代因果推断方法是一种非参数方法，对因果效应的定义和估计不依赖于具体的结构模型形式，那么如何理解线性回归模型和OLS估计量的内涵，以及控制变量在其中的作用？接下来本文将说明，在基于设计的研究范式下，线性回归模型中的控制变量作用与精确匹配估计类似，都是对应着分块随机化实验的研究设计，而OLS估计量则是一种特殊加权的匹配估计量。

首先讨论一类特殊的线性回归模型——饱和回归模型（saturated regression model，后文简称饱和回归模型）。饱和回归模型是指对离散型控制变量的每一个可能取值都使用一个特定的虚拟变量进行控制的线性回归模型。假设控制变量X的取值范围是$\{x_1, x_2, \cdots, x_G\}$，饱和回归模型形式如下：

$$Y_i = \beta^{sat} D_i + \sum_{g=1}^{G} \gamma^g \mathbb{I}\left(X_i = x_g\right) + u_i \tag{16}$$

上式中的$\mathbb{I}(\cdot)$为示性函数（indicator function），当括号内条件成立时取1，否则取0。当控制变量X_i存在G个可能取值时，相应添加G个虚拟变量。可以证明，饱和回归模型的估计系数是一系列条件平均处理效应的加权平均（安格里斯特、皮施克，2009；丁，2024）：

$$\hat{\beta}^{sat} = \sum_{g=1}^{G} w_g \hat{\tau}\left(X_i = x_g\right) = \sum_{g=1}^{G} \frac{\Pr\left(X_i = x_g\right)\mathrm{var}\left(D_i \middle| X_i = x_g\right)}{\sum_{g=1}^{G} \Pr\left(X_i = x_g\right)\mathrm{var}\left(D_i \middle| X_i = x_g\right)} \hat{\tau}\left(X_i = x_g\right) \tag{17}$$

其中，$\hat{\tau}(X_i=x_g) \equiv E[Y_i|D_i=1, X_i=x_g] - E[Y_i|D_i=0, X_i=x_g]$是$X_i=x_g$时的条件平均处理效应。根据上式，可以这样理解饱和回归模型估计量的计算过程：首先根据控制变量的取值将观测样本分为G个分块（group）；而后估计每个分块内的条件平均处理效应$\hat{\tau}(X_i=x_g)$；最后按照特定的权重w_g把全部分块的条件平均处理效应加总为平均处理效应。

基于设计的研究范式下，饱和回归模型估计量的计算过程是一种特殊的匹配估计量。与上一节的精确匹配估计量类似，饱和回归模型事实上相当于近似实行了一次以控制变量的取值为分组依据的分块随机化实验，在满足CIA假设的前提下，通过将处理组和控制组的差异比较限制在分块内部，实现了近似随机化实验，从而消除了遗漏变量对因果推断的影响。总的来说，饱和回归模型和精确匹配估计量所对应的研究设计完全一致，二者的差异主要体现在加权方式上。饱和回归模型权重w_g取决于两个因素：一是样本占总体的比重$\Pr(X_i=x_g)$，占比越大的分块权重越大；二是处理状态的方差$\mathrm{var}(D_i|X_i=x_g)$，衡量分块内部个体处理状态的变异性，分块内的处理组和控制组分布得越为均匀，权重也就越大。精确匹配估计量的权重则只考虑样本占总体比重$\Pr(X_i=x_g)$。

在多数实证研究中，控制变量往往为连续型，或同时包含连续型与离散型变量。在此情形下，直接使用饱和回归模型容易遭遇维数诅咒（curse of dimensionality），进而引发不可识别问题。因此，研究者通常更倾向于使用如下形式的多元线性回归模型：

$$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon_i \tag{18}$$

在基于设计的研究范式下，可将上述线性回归模型中的估计系数视为一种特殊的匹配估计量。为说明这一点，首先引入线性投影算子$L(\cdot|X_i)$，该算子代表向量在X张开的线性空间上的投影，例如$L(D_i|X_i)$表示D_i对X_i做线性回归后的拟合值$\hat{D}_i$。在这种设定下，可以将$\hat{D}_i$理解为使用线性概率模型（Linear Probability Model，LPM）所估得的倾向得分，简记作$L(X_i)$。斯沃琴斯基（2022）证明多元线性回归模型中处理变量的估计系数β与以下线性回归模型的估计系数相等：

$$Y_i = \alpha + \beta D_i + \varphi L(X_i) + \varepsilon_i \tag{19}$$

上述结果表明，在多元线性回归模型中添加控制变量X等价于控制使用LPM估计的倾向得分$L(X_i)$。进一步地，若将倾向得分$L(X_i)$做离散化（如通过再分组（subclassification））并进行饱和模型回归，该研究设计便等价于一种基于倾向得分的分层匹配方法。对于线性倾向得分$L(X_i)$是连续变量的一般化情形，多元线性回归模型实际上可以看做是一种使用LPM估计倾向得分、使用倾向得分条件方差进行加权调整的特殊倾向得分模糊匹配估计量。

综上所述，无论控制变量是离散型还是连续型，无论采用饱和回归模型还是一般形式的多元线性回归模型，线性回归模型的估计系数都对应着某种匹配估计方法，而匹配估计方法的本质是模拟分块随机化实验。因此，在基于设计的研究范式下，线性回归模型不再是对于结果变量的一种结构建模方法，而是一种让观测数据逼近随机化实验的研究设计方式。

控制变量在分块随机化实验中发挥着至关重要的作用，通过控制变量进行合理分组，保证每个分块内的个体都有相同的倾向得分，此时处理状态近似于随机分配。换言之，基于设计的研究范式下，控制变量是否合理既不取决于它是否影响结果变量，也不取决于它是否和处理变量存在相关性，而是取决于它在处理分配机制中的角色，即对处理变量是否存在因果性关系。为进一步说明控制变量在这一框架中的作用，可以证明一般型控制变量的线性回归模型估计系数β^{ols}与离散型控制变量的饱和回归模型估计系数β^{sat}存在如下关系⑧：

$$\beta^{ols}=\beta^{sat}+\frac{\mathrm{var}(e(X_i)-L(X_i))}{\mathrm{var}(D_i-L(X_i))}\left\{\beta^{sat}-\frac{\mathrm{cov}\left[(e(X_i)-L(X_i))Y_i\right]}{\mathrm{var}(e(X_i)-L(X_i))}\right\} \quad (20)$$

借助上式，可以从基于设计的研究范式重新理解线性回归模型的模型误设问题。该范式强调，研究者关注的核心是控制变量与处理变量之间的关系（即倾向得分函数）是否建模合理。当倾向得分函数为线性函数时，饱和回归模型中的分组虚拟变量拟合出的分块倾向得分$e(X_i)$与线性回归模型拟合得到的线性倾向得分$L(X_i)$相同，此时两种模型估计的处理效应将完全一致。然而，若倾向得分函数并非线性，使用线性模型估算的倾向得分将存在模型误设偏误（misspecification），线性倾向得分$L(X_i)$不能准确估计出正确的倾向得分$e(X_i)$，此时若根据$L(X_i)$进行匹配，可能会因为线性外推（linear extrapolation）引起估计偏误。

4. 线性回归模型中的控制变量作用示例

本节使用一个模拟示例来展示线性回归模型中的控制变量作用，以及线性回归模型所对应的研究设计⑨。表1展示了一个大学教育与个人收入的模拟数据集，其中包含个体是否上大学（D）、能力（X）、收入（Y）的观测结果和潜在结果。

假设研究者的研究目的是估计上大学对收入的平均处理效应。根据表1的数据，组间均值差异SDM估计量为56.67。可以看到，由于能力更强的人更容易上大学，同时能力更强的人的潜在收入水平也较高，因此上大学和未上大学人群的潜在收入水平不具有可比性，SDM不能正确识别因果效应。然而，如果将比较范围限制在能力相同的人群中，就能够消除选择性偏误。例如，分块2中的3个人能力相同，他们的潜在收入水平$Y(0)$也相同，此时组间均值差异可以正确识别条件平均处理效应，计算可知$\hat{\tau}(X=4)=50$。类似的，可以计算出分块3和分块4的条件平均处理效应。然而，由于分块1和分块5内部只有处理组或控制组，它们的条件平均处理效应无法被识别。计算出每个块的条件平均处理效应后，可将所有条件平均处理效应加权求和，最终得到的平均处理效应估计结果为35。

表1 模拟数据：大学教育与个人收入

编号	按能力分块	上大学（D=1）	未上大学（D=0）	能力（X）	Y	$Y(1)$	$Y(0)$	个体因果效应	$CATE$	样本占比	var(D_i)
1	1	1	0	10	120	120	100	20	20*	1/6	0
2		1	0	10	120	120	100	20			
3	2	1	0	4	130	130	80	50	50	1/4	2/9
4		1	0	4	130	130	80	50			
5		0	1	4	80	130	80	50			
6	3	1	0	3	100	100	70	30	30	1/4	2/9
7		0	1	3	70	100	70	30			
8		0	1	3	70	100	70	30			
9	4	1	0	2	70	70	50	20	20	1/6	1/4
10		0	1	2	50	70	50	20			
11	5	0	1	1	30	40	30	10	10*	1/6	0
12		0	1	1	30	40	30	10			

说明：CATE是条件平均处理效应，为分块内个体处理效应的平均值。由于分块1只有处理组，分块5只有控制组，这两组的CATE无法被估计，故以*号注明。

接下来使用饱和回归模型估计因果效应。表2的第（3）列报告了使用全样本估计的饱和回归模型估计结果为34.545，与精确匹配估计的结果基本一致，两者间的细微差异是由于分块权重的计算方式不同所引起的⑩。表2的第（4）列将样本范围限制在满足重叠性假

表2 不同估计方法的结果比较

	（1）	（2）	（3）	（4）	（5）	（6）	（7）
	简单均值差异	精确匹配估计	饱和回归	饱和回归	饱和回归+调整权重	多元线性回归	多元线性回归
ATT估计值	56.667	35.000	34.545	34.545	35.000	42.137	34.079
是否控制能力	不控制	控制	控制	控制	控制	控制	控制
样本范围	全样本	重叠样本	全样本	重叠样本	重叠样本	全样本	重叠样本
样本数	12	8	12	8	8	12	8

设的样本中，饱和回归模型估计结果与全样本完全一致，表明对于饱和回归模型而言，不满足重叠性假设的样本不会进入到因果效应的估计过程中，不会对估计结果产生任何影响。

既然饱和回归模型和精确匹配估计量本质上是相同的研究设计，二者的差异反映在加权权重上，那么理论上可以通过再加权（reweight）的方式让两者结果完全等价。吉本斯等（2019）给出了修正权重的表达式：

$$\hat{w}_i=\left[\widehat{\mathrm{Var}}(\tilde{D}_i|X=x_g)\right]^{-1/2} \quad (21)$$

上式中的$\tilde{D}_i=D_i-\hat{e}(X_i)$。研究者可以首先计算出个体的经验倾向得分$\hat{e}(X_i)$，而后计算出个体权重$\hat{w}_i$并以其为权重估计加权线性回归模型。表2的第（5）列使用上述方法进行加权回归，可以看到经过权重调整后的饱和回归模型与精确匹配估计量的估计结果完全一致。上述结果有力地证明了饱和回归模型本质上是一种遵循特殊加权方式的精确匹配估计量，二者仅是加权权重有所不同。

表2的（6）、（7）两列分别展示了使用全样本和满足重叠性假设的子样本时多元线性回归模型的估计结果。可以发现，对于多元线性回归模型，不满足重叠性假设的样本同样会影响估计结果。这是因为多元线性回归模型是用线性函数拟合倾向得分，对于不满足重叠性假设的样本使用拟合结果做线性外推，这部分样本仍然会参与因果效应的估计。然而，线性外推带来的一个潜在不良后果是可能会存在不可忽视的外推偏误，特别是对于一些控制变量取值较为极端的样本。例如，线性倾向得分$L(X_i)$在极端情况下甚至会超过1，这显然是不合理的。此外，由于线性回归估计量存在着权重逆反（reversed weights）的特性（斯沃琴斯基，2022），较为稀少的极端值样本往往会被赋予较大的权重，这会使得整个估计系数受到极端值的显著干扰。可以看到，第（6）列使用全样本的估计系数与真实因果效应存在较为明显的差异，这一估计偏误就是由于线性外推所导致的。如果将线性回归模型限制在满足重叠性假设的子样本，可以有效减少极端值样本对估计系数的干扰。可以看到，第（7）列多元线性模型在排除极端值样本后的估计系数与真实值非常接近。综上所述，重叠性假设对线性回归模型估计系数的稳定性具有至关重要的作用。在实践中，研究者们应该注意检查控制变量的重叠性假设是否满足，同时要高度关注极端值对估计系数的重要影响，通过各类稳健性检验（例如排除部分极端值样本）来确保估计结果的稳健性和可信性（因本斯、徐，2024）。

（三）两种研究视角下的控制变量作用比较

基于模型的研究范式与基于设计的研究范式代表了两种实证研究的主导性方法论。基于模型的传统研究范式强调对结果变量的数据生成过程（DGP）进行建模拟合，从而揭示变量之间的因果关系，其重点在于通过对结果变量的精确建模和模型参数估计。线性回归模型被视为一个用于刻画DGP的参数化模型，并通过控制变量来减少遗漏变量偏误，以提升模型的拟合程度与参数估计的一致性。然而，如前所述，基于模型的分析方法面临若干理论挑战，例如对因果关系的界定不清、无法完全避免模型误设问题以及依赖相关性而非因果性选择控制变量等。基于设计的研究范式以潜在结果框架为分析基础，以随机化实验作为估计因果关系的黄金标准。在这一范式下，研究者的重点从预测结果变量转向了理解处理分配机制，在此基础上让观测性研究尽可能逼近随机化实验。借助于潜在结果框架这一描述因果关系的强大工具能够明确界定因果关系，并自然地容纳个体层面的异质性因果效应。估计目标也可以明确为研究者所感兴趣特定群体的平均因果效应，这就能够很好地解决异质性因果效应情形下的线性回归模型系数定义问题。研究重点的转变从根本上规避了对结果变量DGP进行建模时所无法完全克服的模型误设问题。

两种研究范式最核心区别在于二者的重点不同。基于模型的研究范式聚焦于对结果变量的数据生成过程进行建模，本质上是一个数据建模和预测问题。因此，该范式强调控制变量如何改进统计模型中的参数估计一致性和估计效率。基于设计的研究范式重点则在于理解处理分配机制的基础上消除观测性数据中的选择性偏误，使观测数据在某些条件下能够近似于随机化实验。因此，基于设计的研究范式强调控制变量的核心作用在于理解处理分配机制，进而通过合理分组以在局部逼近随机化实验，而非提升模型的预测能力。对于处理分配机制的关注是否处于实证研究的核心地位是两种研究范式最为核心的区别。

上述比较分析对于实证研究者具有重要的启示。首先，在研究设计的视角下，虽然线性回归模型仍然保

留了统计模型的外形，但其内核已经被替换为近似随机化实验识别思路，其实现的是合理分组、组内估计和跨组加权平均的一整套因果效应识别和估计程序。从这一视角看，以提升拟合优度等模型预测表现作为选择控制变量的标准并不恰当，因为这种选择标准并没有体现如何使得观测性数据得以逼近随机化实验的作用。其次，基于设计的研究范式下，控制变量的选择思路不再是基于变量间的相关性，而是基于因果效应的可识别性，这种思路转变带来了一个非常重要的发现：控制变量并不一定都能够消除选择性偏误，还可能存在引起选择性偏误的坏控制变量。这一发现深刻地改变了研究者对于控制变量的选择和判断标准。最后，只要确定了控制变量的选择，线性回归模型和匹配方法二者所需要的识别假设完全相同，二者在识别效力上是等价的，区别在于倾向得分估计方法、匹配策略和组别权重选择。因此，因果推断实证研究的首要核心应该是如何选择合适的控制变量以消除选择性偏误，至于选择何种估计方法则是一个次要问题。

（四）控制变量与统计推断

控制变量不仅会影响因果效应的识别，还会影响估计效率。效率更高的估计量能够在给定样本量不变的前提下获得更为精确的估计值，提高统计推断效力。对于线性回归模型，误差项满足同方差假设时OLS估计系数$\hat{\beta}^s$的标准误为：

$$se\left(\hat{\beta}^s\right)\approx\sqrt{\frac{1}{N}\times\frac{\mathrm{var}\left(u_i^s\right)}{\mathrm{var}\left(\tilde{D}_i\right)}} \tag{22}$$

上式中，N为样本量，误差项u_i^s是原始误差项u_i^l剔除了可以被控制变量X解释部分后的剩余部分，同理，$\tilde{D}_i$是D_i剔除了可以被控制变量X解释部分后的剩余部分[⑪]。误差项的方差$\mathrm{var}(u_i^s)$越大意味着数据中其他因素越多、噪音越大，导致估计结果更为粗糙，标准误更大。处理变量的方差$\mathrm{var}(\tilde{D}_i)$越大表示处理组和控制组分布越均匀，越有利于平均掉误差项的扰动，估计效率越高。

借助(22)式，我们可以考察添加控制变量对OLS估计系数标准误的影响。如果控制变量对误差项的解释力度很强，添加控制变量后能够大大吸收误差项的变动性，使得误差项的方差$\mathrm{var}(u_i^s)$大幅降低，此时估计系数的标准误会下降，估计效率提升。但是，如果控制变量对处理变量的解释力度很强，添加控制变量后会很大程度上吸收处理变量的变动性，处理变量的方差$\mathrm{var}(\tilde{D}_i)$大幅降低会导致估计系数的标准误显著增加，估计效率降低。因此，从改善估计效率、增加统计推断效力的角度，好控制变量应该对结果变量具有较高的解释力，同时尽可能与处理变量无关；反之，坏控制变量与结果变量相关性很小，但与处理变量高度相关。

在统计推断方面发挥作用的控制变量与前述因果识别方面的控制变量具有显著不同。两者最大的差异在于添加控制变量的目标不同：因果识别方面的控制变量是为了消除选择性偏误，使得估计系数能够反映真实的因果效应；统计推断方面的控制变量旨在减少其他因素对因果效应估计的干扰，提高估计效率和推断效力。研究者们在使用控制变量时应当清楚地意识到不同方面的控制变量所针对的特定问题和目的，确保正确发挥控制变量的作用。

三、实证研究中的坏控制变量问题

实证研究中，控制变量的作用在于剥离遗漏变量的影响，从而准确估计变量间的因果关系。因此，控制变量的选取是研究设计的核心环节。然而，并非所有控制变量的引入都能优化研究设计——某些变量看似合理，实则会导致估计偏误。坏控制变量是研究设计中一个典型误区，它通过引入选择性偏误或过度控制间接效应，使得因果效应估计失真甚至方向错误。本节从因果推断视角系统解析坏控制变量如何导致因果推断失效。

（一）坏控制变量与选择性偏误

第一类坏控制变量是研究者关注的处理变量D与影响结果变量的其他因素u的共同结果变量W，亦称为对撞变量(collider)。如图2所示，假设处理变量D是随机分配的，如果研究者不控制对撞变量W，处理变量D与结果变量Y之间不存在因果关系之外的其他相关关系，此时观测性数据中D和Y的相关关系可以被识别为因果关系。然而，如果研究者控制了对撞变量

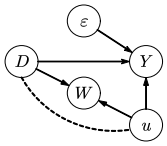


图2 坏控制变量与选择性偏误

W，就会使得W的共同原因D和u相关，由于此时u同时与D和Y相关，u成为了会导致选择性偏误的遗漏变量（安格里斯特、皮施克，2009；奇内利等，2022）。在这个例子中，不控制W不会引起因果效应估计偏误，而控制了W反而会使得因果效应的估计中混杂入选择性偏误。正因为如此，研究者们将W这类变量称为坏控制变量。

表3 坏控制变量的模拟数据：大学教育与个人收入

编号	按职业分块	上大学（D=1）	未上大学（D=0）	能力（X）	职业（W）	W^1	W^0	Y	Y^1	Y^0	个体因果效应
1	1	1	0	h	1	1	1	130	130	100	30
2	1	1	0	h	1	1	1	130	130	100	30
3	1	1	0	l	1	1	0	100	100	80	20
4	1	1	0	l	1	1	0	100	100	80	20
5	1	0	1	h	1	1	1	100	130	100	30
6	1	0	1	h	1	1	1	100	130	100	30
7	2	0	1	l	0	1	0	80	100	80	20
8	2	0	1	l	0	1	0	80	100	80	20

表3继续沿用大学教育与个人收入的例子来进一步解释坏控制变量引起选择性偏误的内在机制。首先，个体是否能够上大学是随机分配的，与能力无关。从表3中可以看到，高能力（X=h）和低能力（X=l）群体上大学的可能性都是50%。其次，个人的职业选择取决于能力与学历，能力较强、有大学学历或二者兼有的个体能够进入高薪职业（W=1），能力较低且没有上大学的个体只能进入低薪职业（W=0）。在该数据中，能力X扮演了类似于图2的变量u的角色。由于是否上大学D是随机分配的，与能力无关，因此并不存在第二节中所述的选择性偏误，直接比较上大学和未上大学人群的平均收入差异即可得到上大学对收入的因果效应。然而，由于能力和学历同时决定职业，在职业相同的群体内部，能力与个人是否上大学存在显著的相关性。从表3中可以看到，在进入高薪职业的群体中（分块1），高能力个体上大学的概率只有50%，而低能力个体上大学的概率是100%，能力与学历两者存在负相关关系。因此，根据职业分块后，每个块内的组间均值差异无法准确估计条件平均处理效应。

表4 坏控制变量对因果效应估计的影响

	（1）	（2）	（3）
	真实因果效应	简单均值差异	添加坏控制变量
ATT估计值	25	25	15
样本数	8	8	8

表4展示了添加坏控制变量如何干扰正确的因果效应估计。可以看到，当不添加坏控制变量时，线性回归模型等价于计算简单均值差异SDM，由于不存在选择性偏误，SDM可以正确估计真实因果效应。正如本文上一节所阐述的，在线性回归中添加控制变量，本质上就是按照控制变量进行分块估计条件平均处理效应，而后再计算加权平均的总体平均处理效应。添加坏控制变量使得第一步估计条件平均处理效应失效，自然也无法得到总体平均处理效应的正确估计结果。

借助潜在结果框架可以更为清晰地理解添加坏控制变量引发的选择性偏误问题。使用$W_i(1)$和$W_i(0)$分别表示个体在上大学或未上大学状态下的职业选择。如果控制了职业W，条件平均处理效应估计结果可以表示为：

$$\begin{aligned}
&E\left[Y_i|D_i=1,W_i=1\right]-E\left[Y_i|D_i=0,W_i=1\right]=E\left[Y_i(1)|W_i(1)=1\right]-E\left[Y_i(0)|W_i(0)=1\right]\\
&=E\left[Y_i(1)|W_i(1)=1\right]-E\left[Y_i(0)|W_i(1)=1\right]+E\left[Y_i(0)|W_i(1)=1\right]-E\left[Y_i(0)|W_i(0)=1\right]\\
&=\underbrace{E\left[Y_i(1)-Y_i(0)|W_i(1)=1\right]}_{\text{条件平均因果效应}}+\underbrace{E\left[Y_i(0)|W_i(1)=1\right]-E\left[Y_i(0)|W_i(0)=1\right]}_{\text{选择性偏误}}
\end{aligned}\tag{23}$$

上式清晰地说明了坏控制变量引发选择性偏误的机制在于将上大学后才能获得高薪职业的群体（即W_i(1)=1）和不上大学也可以获得高薪职业的群体（即W_i(0)=1）放在一起进行比较，而这两个群体在能力方面存在着天然的差异，不是恰当的比较对象。所以，坏控制变量相当于人为选择将“苹果和橘子”进行错误比较，导致产生选择性偏误。因此，坏控制变量事实上是对研究设计的一种破坏，不但没有使得观测性数据逼近随机化实验，反而使得选择性偏误愈发严重，甚至人为引入了原本不存在的选择性偏误，极大地破坏了研究设计的合理性和可靠性。

（二）中介变量、间接效应与过度控制

第二类坏控制变量是处理变量对结果变量产生因果效应的具体机制变量，也被称为中介变量（mediator）。对于识别总体因果效应而言，控制中介变量会存在两重负面影响：一是过度控制，即由于控制间接效应造成对总因果效应的错误估计；二是可能像第一类坏控制变量那样，引入新的选择性偏误。

首先解释过度控制问题。假设数据生成过程如图3a所示，处理变量D对结果变量Y的因果效应由两条路

径构成，一是 D 对 Y 的直接效应 $D\rightarrow Y$，二是 D 通过影响 W 进而影响 Y 的间接效应 $D\rightarrow W\rightarrow Y$。在这种情况下，$D$ 对 Y 的总因果效应是直接效应和间接效应之和。若控制住 W，那么间接效应的因果路径被阻断，此时观测性数据中的 D 和 Y 之间的相关关系只能反映直接效应的影响，导致因果效应估计偏误。更重要的是，这种偏误程度的大小无法评估，甚至在极端情况下能够直接改变估计系数的符号。例如，假设直接效应为 2，间接效应为 −4，总因果效应为两者之和等于 −2。但是，如果将中介变量作为控制变量，研究者将得到大小为 2 的因果效应估计值，从而错误地认为处理变量对结果变量存在正向因果影响。

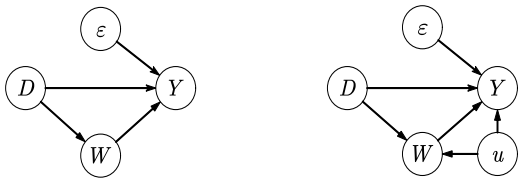
上述分析一方面解释了为什么控制中介变量会导致因果效应估计出现错误，但另一方面似乎也隐含着一种可能性：研究者可以通过比较未控制中介变量的因果效应估计值（包含直接效应和间接效应）与控制中介变量的因果效应估计值（只包含直接效应）两者之间的差异来量化间接效应的大小。这种做法正是所谓的中介效应分析三步法（巴伦、肯尼，1986）。然而，现实中观测性数据几乎一定会存在某些不可观测因素 u 是中介变量 W 和结果变量 Y 的共同原因（见图 3b），此时中介变量 W 同时也是一个对撞变量。在这种情况下将 W 作为控制变量，一方面会剥离掉间接效应，另一方面还会引入新的选择性偏误，对于因果效应估计值的影响将会更加复杂和难以判断。正因为如此，近年来许多学者都对实践中使用中介效应分析三步法提出了担忧，建议研究者们要慎重采用这种方法（江艇，2022）[12]。

总而言之，在观测性研究的绝大多数研究情景中，控制中介变量都很有可能引入新的选择性偏误。即使不存在导致中介变量成为对撞变量的因素，控制中介变量仍然存在过度控制问题，导致总体因果效应的估计偏误。从谨慎和保守的科学研究精神考虑，研究者应该尽可能避免将中介变量作为控制变量。

（三）统计推断方面的坏控制变量

还有一类特殊的控制变量，控制它们不会导致选择性偏误，但是会降低估计效率，表现为增大估计系数的标准误，使得置信区间变宽、统计功效下降。理解这类变量的作用机制，对优化研究设计、平衡偏差与方差至关重要。

前文已经说明，如果一类变量与误差项高度相关的同时与处理变量不相关，在线性回归模型中控制这类变量会起到吸收误差项的变动性、不影响处理变量的变动性的效果，从而降低估计系数标准误、提高统计功效。反之，如果控制变量与处理变量高度相关、与误差项不相关，这类控制变量则会增大估计系数的标准误、降低统计功效。图 4 通过数据模拟直观地展示了上述结论。分别对无控制变量、统计推断的好控制变量（与 u 相关且与 D 无关）和坏控制变量（与 u 无关且与 D 相关）3 个线性回归模型进行蒙特卡洛模拟，可以看到 3 个线性回归模型均能一致估计参数，但估计效率存在很大差异。与不添加任何控制变量的基准估计结果相比，控制与误差项 u 相关的控制变量后估计系数的分布更加集中，表明添加此类控制变量能够有效减少噪音对系数估计的影响，起到降低估计系数标准误的作用。然而，如果控制与处理变量 D 相关的控制变量，反而会使得估计系数的分布变得更加分散，导致估计系数的标准误大大提高，降低统计功效。因此，选择合适的控制变量



(a) 过度控制间接效应　　(b) 引入选择性偏误

图 3　中介变量与坏控制变量

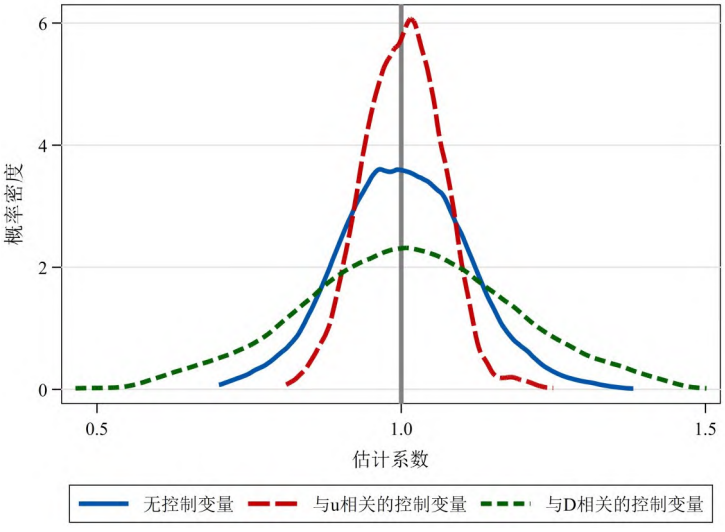


图 4　控制变量对统计推断的影响

注：分别对 3 类模型进行了 1000 次估计，图中是估计系数的分布。真实系数等于 1。

对于实证研究的统计推断环节至关重要，良好的控制变量有助于研究者得到更为精确的因果效应估计结果，而不恰当的控制变量则会放大估计误差，使得研究者无法获得因果效应的精确估计值。

综上所述，本文总结了3类坏控制变量，其中前两类控制变量是因果识别方面的坏控制变量，第三类是统计推断方面的坏控制变量。因果识别方面的坏控制变量会引入新的选择性偏误，导致因果效应估计失效，换言之，会破坏估计系数的一致性。而统计推断方面的坏控制变量则会过度吸收处理变量的变异性，导致估计系数的标准误增大、统计功效降低。两方面的坏控制变量作用有本质性区别，研究者在添加控制变量时要注意区分，避免相互混淆。

四、基于设计的研究范式下的控制变量选择标准

（一）控制变量类型的简明分类

明确控制变量的判断标准对于实证研究工作者而言至关重要。本文以图5为对象，介绍实证研究中常见的一些控制变量的分类[19]。

从因果识别的角度来看，好的控制变量应该能够消除选择性偏误，剥离因果关系中的遗漏变量影响。图5中因果识别的好控制变量用方框标出。首先，处理变量D和结果变量Y的共同原因变量X是选择性偏误的来源之一，故它是一个好控制变量。其次，对于不可观测的共同原因变量ε，虽然无法直接控制它，但可以通过控制ε对D产生影响的机制变量G来达到剥离D和ε相关性的目的，因此G作为ε的代理变量也是一个好控制变量[19]。

因果识别方面的坏控制变量作用恰恰相反，会造成新的选择性偏误。图5中因果识别的坏控制变量用圆形标出。首先，中介变量M是一个典型的坏控制变量，控制它会导致不可观测因素u同时与D和Y产生相关性，造成选择性偏误。即使不存在u这类变量，控制M也会剥离间接效应，造成因果效应低估甚至方向错误。其次，D和Y的共同结果变量Q以及Y的结果变量R也是坏控制变量。控制Q会导致D和Y产生非因果的相关性关系，干扰因果效应估计，控制R则会导致将潜在结果不具有可比性的群体放在一起进行比较，产生新的选择性偏误。这些坏控制变量都会在原有的选择性偏误之外引发新的选择性偏误，在研究中需要高度警惕，避免错误控制。

从统计推断的角度来看，控制某些变量可以提高估计的精确性。变量P是Y的原因变量且与D不相关，控制它有助于减少噪声，提高估计的稳健性。变量W是D的原因变量，与误差项不相关但与D高度相关，控制W会大大降低处理变量的变异性，导致估计误差增大、统计效力下降。

需要说明的是，图5是一个非常简化的因果模型，现实中有大量复杂情形，比如，一个变量可能既是因果识别方面的好控制变量，又是统计推断方面的坏控制变量，X就属于这一类变量，控制X一方面能够消除选择性偏误，但另一方面也会增大估计误差，这种偏差—方差权衡（bias-variance trade-off）贯穿了因果推断类实证研究的方方面面。当面临这类权衡时，没有一个公认的选择准则。是一致但粗糙的估计结果更符合科学研究的要求，还是一个有偏但精确的估计结果更有实际意义？研究者们需要仔细地思考控制变量的利弊，根据研究目标来权衡和选择。

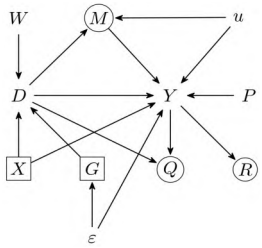


图5 常见控制变量的分类与判断准则

注：图中方框内的变量为因果识别的好控制变量，圆圈内为坏控制变量。

（二）复杂情形下的控制变量权衡

现实中的实证研究情景往往要比图5复杂得多，甚至会出现有些变量可能既是好控制变量，也是坏控制变量。图6是一个例子。在图6的左图中，u是D和Y的共同原因，是一个典型的遗漏变量。对u和Y的因果路径上的机制变量W加以控制，可以阻断u和Y的相关性，克服遗漏变量问题。然而，W同时也是u和ε的共同结果，控制W会使得u和ε产生相关性，进而使得D和Y产生因果关系之外的相关性，引起新的选择性偏误。同理，在图6的右图中，不控制W会导致ε

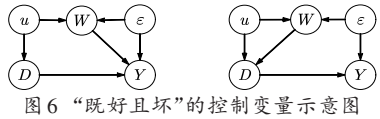


图6 “既好且坏”的控制变量示意图

成为一个遗漏变量，而控制了W后又会使得u成为遗漏变量。在这两种情形中，控制或不控制W都无法完全消除选择性偏误。

在实证研究中，“既好且坏”的控制变量并非罕见。以高等教育回报率的研究为例，众多文献试图估计我国高等教育对个人收入的因果效应[15]。在我国义务教育阶段入学制度和高等学校录取制度的政策大背景下，户籍一方面决定了义务教育阶段的教育质量，另一方面决定了个人在何处参加高考和录取，因此户籍是决定个人能否接受高等教育的核心因素之一。同时，由于户籍制度的限制和城乡经济差距等客观现实，非农户口的城镇居民能够获得更多高质量的就业机会，所以户籍也是个人收入水平的重要影响因素。综合上述两点，在研究高等教育回报率问题时，显然户籍是一个遗漏变量，应该加以控制。然而，个人户籍并非是不可改变的，在相当长的一段时间里，通过接受高等教育并进入非农部门工作是个人改变户籍的主要途径之一。因此，户籍又是高等教育的结果变量之一，并且是高等教育影响个人收入的中介变量。根据图7，如果要克服户籍导致的选择性偏误，理想的做法应该是控制个人在出生时的户籍信息（前定变量）。但是相关研究使用的人口普查数据、城镇住户调查数据等微观个体数据都只包含了个人当前户籍信息（后定变量）。在这种情形下，不控制当前户籍会存在遗漏变量问题，而控制当前户籍又会过度控制间接效应，低估高等教育对个人收入的真实因果效应。如果还存在其他同时影响当前户籍状态和个人收入的因素（如家庭社会资本），控制当前户籍状态还会进一步引入新的选择性偏误。因此，当前户籍是一个“既好且坏”的控制变量。

上述示例生动地展示了观测性研究中控制变量方法的局限性。事实上，即使研究者能够完整地获得关于原因和结果的因果结构知识，也无法确保一定可以通过控制遗漏变量的方式实现可靠的因果推断。究其原因在于，在一个高度复杂的因果结构网络中，可能不存在一个同时剔除全部选择性偏误的控制策略。特别是当研究者所关注的处理变量和结果变量两者都存在非常多的影响因素时，整个因果结构会变得极其复杂，很容易出现加与不加控制变量都存在偏误的两难情形。一旦出现这种情况，单纯依靠控制变量方法无法完全克服选择性偏误，从而无法准确估计因果效应。当研究者遇到此类情形时，应当清晰地意识到控制变量在消除选择性偏误作用的局限性，转而寻找新的研究设计，例如寻找准自然实验来获得处理变量的外生变动（exogenous variation）并使用双重差分法、断点回归法等识别策略[16]。若研究对象既无法寻找到合适的控制策略，又无法寻找到处理变量的外生变动，那么这类问题从根本上很难使用观测性数据实现可信的因果推断。此时研究者可以选择随机对照实验等实验研究方法开展因果推断研究，或退而求其次使用观测性数据进行相关性研究，提供具有参考价值的建议性证据（suggestive evidence）。

五、实证研究中控制变量选取和使用原则与建议

至此本文已经详细阐述了研究设计中的控制变量选择的重要性。可以说，如何挑选出好控制变量、避免坏控制变量是高质量实证研究的核心要求。本文基于控制变量的基本原理，结合当前我国社会科学实证研究现状，归纳总结出以下几点控制变量选取和使用的原则与建议。

（一）控制变量的基本原则

原则一：明确因果模型，基于因果结构选择控制变量。

一个控制变量到底能够消除还是引入选择性偏误，取决于控制变量在因果模型中的位置。可以说，控制变量的分类判别是否合理完全依赖于研究者对因果模型的认识，如果研究者对于研究对象的因果模型没有清晰的理解，就不可能准确判别控制变量所发挥的作用，更不可能挑选出那些必须控制的好控制变量，也无法分辨出必须加以避免的坏控制变量。因此，在实证研究中研究者首先要基于对现有理论知识以及对制度背景的深入了解，形成对研究对象的因果模型的先验认识，然后才可能有理有据地挑选出恰当的控制变量。从这个意义上讲，好控制变量应该是也必然是社会科学理论驱动的产物，而非数据挖掘的结果。唯有坚持理论优先、因果结构优先的控制变量选择逻辑，才能确保控制变量的选择和使用真正服务于因果推

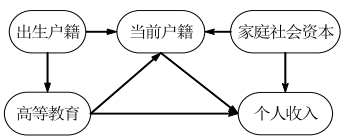


图7 高等教育与个人收入的因果图

断[17]。

原则一对于提高实证研究的可信性、透明性和可复现性具有重要意义。明确选取控制变量应建立在理解因果模型的基础上,有助于减少当前一些实证研究中出现的控制变量"误用""滥用"现象。在当前我国社会科学领域的实证研究中,忽视因果结构而机械添加控制变量的现象仍较普遍。例如,有的研究者仍根据变量间的相关性而非因果性作为控制变量的选取标准,认为所有和处理变量相关的变量都应该被纳入控制变量,或认为控制变量有益无害、多多益善,忽略了存在坏控制变量的可能性。有的研究者通过逐步回归、机器学习等方法,依据统计显著性或模型拟合度筛选控制变量,却没有进一步思考控制变量在因果模型中的角色。有的研究者机械地照搬既有文献中的控制变量,却没有认真推敲既有控制变量在全新的研究情景和制度背景下是否仍然适用。更有极少数研究者为了追求显著性结果随意更换和挑选控制变量,通过反复尝试不同变量组合直至获得预期结果。这些错误的做法不仅没有正确发挥出控制变量在因果推断中的核心作用,反而对实证研究的研究设计严谨性和研究结论可信性造成了危害。总而言之,以因果模型作为基本出发点应是实证研究者们选择控制变量时的首要原则。

原则二:高度重视坏控制变量问题。

实证研究的最终目标是准确估计变量间的因果效应,因此,实证研究的首要问题在于通过好的研究设计来尽可能消除选择性偏误。遗漏变量问题和坏控制变量问题都会产生选择性偏误。因此,选取合适的好控制变量以克服遗漏变量偏误,与避免坏控制变量所导致的选择性偏误,两者在实证研究中应该具有相同的重要性。然而,一些研究者在论证内生性问题和控制遗漏变量偏误方面花费了极大的心血和功夫,但对坏控制变量问题则不够重视,无意中控制了一些不恰当的坏控制变量。在此情形下,即使研究者能够非常理想地消除所有遗漏变量偏误问题,但坏控制变量问题仍然会使得因果效应估计存在不可预期的偏误,导致研究结论可信性大大降低,使得前期在研究设计阶段的努力前功尽弃,这不得不说是一种遗憾。因此,在研究设计和数据分析阶段,研究者们需要高度重视坏控制变量问题,提高研究设计的合理性和稳健性,以增强实证研究结论的可信性。

原则三:注意检验控制变量的重叠性假设。

使用线性回归模型估计因果效应时,控制变量在处理组和控制组分布是否满足重叠性假设是一个重要的但常被忽略的假设。线性回归模型的估计系数实际上是一系列条件平均处理效应的加权平均,如果控制变量的重叠性假设不满足,那么某些样本可能被赋予非常不合理的权重[18]。因本斯(2015)指出当处理组和控制组的控制变量分布存在系统性差异时,线性回归估计系数对模型设定形式会非常敏感。因本斯和徐(2024)发现在重叠性假设满足的时候,线性回归和匹配等各类估计方法得到的因果效应估计值非常接近。因此,重叠性假设对于因果效应估计的稳健性至关重要。

重叠性假设本质上是为了保证在数据中能够为每一个分块的处理组都能找到相似的控制组进行比较。本文将满足重叠性假设的样本称之为有效样本。在很多情形下控制变量是通过改变有效样本的范围来影响估计系数,而不是消除选择性偏误。下面提供一个假想示例。假如研究者试图研究中彩票对劳动供给的影响,研究者收集了10个城市的居民调查数据,其中只有1个城市的居民曾中过彩票(城市内居民的处理状态方差大于0),剩下9个城市的居民全都未中奖(城市内居民的处理状态方差为0)。由于中奖是完全随机的,直接将劳动供给Y对是否中奖D进行线性回归即可得到因果效应的一致估计。然而,如果研究者们进一步控制了城市固定效应,那么事实上的有效样本只包含中过奖的1个城市的居民,剩余9个城市的居民数据即使保留在回归模型中,也不会对估计结果产生任何影响。此时研究者得到的估计系数实际上只能够反映有效样本内的1个城市居民的因果效应。如果考虑到不同城市居民的因果效应可能存在异质性,添加城市固定效应后的估计系数只在有效样本范围内具有内部有效性(internal validity),不能简单地推广到全部样本。如果研究者没有认识到这一点,很可能会错误地解读估计系数的经济含义和适用范围。

明确重叠性假设的作用对于实证研究中控制变量的选取判断具有重要意义。研究者如果添加了一个与

处理变量无关但重叠性较差的控制变量，这不会影响选择性偏误，但会引起加权权重和有效样本范围的显著变化，进而对估计系数产生巨大影响。如果研究者未能洞察其背后的机制，很可能会将这种由权重和样本范围变化导致的系数变动，错误地解读为该变量成功控制了某种遗漏变量偏误[19]。综上所述，实证研究者必须对控制变量在处理组和控制组的分布状况是否满足重叠性假设给予更多的关注。当添加控制变量使得估计系数发生变化时，研究者要厘清系数变化到底是因为消除选择性偏误还是改变有效样本范围。只有搞清楚这些问题，研究者才能够准确理解控制变量的内在机制，合理使用和发挥控制变量的作用。

原则四：在复杂情形下权衡控制变量的利弊。

在实际研究过程中，同一个控制变量在复杂情形下可能同时存在多重作用，研究者们需要进行权衡取舍。例如，一个因果识别方面的好控制变量也一定是统计推断方面的坏控制变量，在消除遗漏变量导致的选择性偏误同时也会减少处理变量的变异性、降低估计效率。控制变量越精细，越有助于消除遗漏变量的影响，但同时也越容易降低统计功效。在极端情况下，过于精细的控制变量甚至会导致有效样本范围的缩小，削弱研究结论的内部有效性和外部有效性[20]。此时研究者必须进行权衡：是选择“系数一致但估计不精确且适用范围小”的研究方法，还是选择“系数可能不完全一致但估计精确且适用范围大”的研究方法，这需要研究者基于具体的研究情景和论证目的作出最恰当的选择。

当面临控制变量既“好”且“坏”的情形时，不存在可以完全消除选择性偏误的控制策略，此时研究者可以转换控制思路，讨论两种偏误的相对重要性。如果遗漏变量导致的选择性偏误要大于坏控制变量导致的选择性偏误，那么添加控制变量所消除的选择性偏误要大于其造成的选择性偏误，综合来看仍然能够降低总的选择性偏误。此时添加控制变量有助于获得更为准确的因果效应估计。以我国高等教育回报率的实证研究为例，目前学界基本上还是认为户籍作为一个遗漏变量所产生的选择性偏误更为严重一些，所以大多数的研究者还是会选择控制户籍（刘生龙、胡鞍钢，2018；鄢伟波，2022）。即便如此，研究者们也应该清醒的认识到，这种控制并非是理想的做法，而是一种“两害相权取其轻”的权衡之举。

在现实中研究者可能遇到各种各样的复杂情形，控制变量所发挥的作用不是那么的清晰和明确。对于这种情况，研究者们应该首先尝试理解研究对象的因果结构，而后借助因果模型来判断特定的控制变量对估计结果的各种可能影响渠道，形成对控制变量作用的完整理解，明确控制变量的“利”与“弊”。最后，研究者们就可以根据特定的研究目的来权衡“利”“弊”，选择最为恰当的做法扬长避短，充分发挥控制变量在因果推断研究中的作用。

原则五：避免过度解读控制变量的估计系数。

基于设计的研究范式的现代因果推断方法最重要的特征之一是将所有的解释变量明确区分为处理变量和控制变量两类，并且明确估计目标是处理变量对于结果变量的因果效应。控制变量的作用是为了让处理变量实现某种程度的随机化分配的效果以更好地消除选择性偏误，至于控制变量本身是否具有选择性偏误则是次要的问题。所以，良好的研究设计虽然能够保证处理变量的估计系数能够反映因果效应，但若没有对控制变量的选择性偏误做针对性处理，无法保证控制变量的估计系数也具有类似的因果含义。一些研究者会通过论证控制变量的估计系数方向与理论预期相符的方式来尝试说明控制变量的可靠性。然而，除非控制变量原本就满足外生性假设，或研究者尝试克服了控制变量的内生性问题，否则控制变量的估计系数并没有因果含义，自然无法与理论预期进行对应和比较。总之，控制变量在实证研究中的核心作用是消除处理变量存在的选择性偏误问题，其本身的估计系数在绝大多数实践情景中不具有明确的因果含义。研究者们应避免过度解读控制变量的估计系数，特别要注意，在缺乏对控制变量内生性问题做出讨论和特殊应对的情形下，不要强行赋予其因果解释。

（二）实践中的控制变量使用建议

依据上述基本原则，本文对于实践中的控制变量使用提出以下几点具体建议。

建议一：选择控制变量需以社会科学理论为基本指导，并充分结合现实制度背景。

控制变量的核心作用是让观测性数据在局部逼近随机化实验，深入理解处理分配机制、明确其中的核心因素，进而选择相应的控制变量加以针对性处理，是良好研究设计的关键所在。在研究实践中，研究者应从社会科学理论和现实制度背景入手，对处理分配机制展开研究分析。例如，在教育回报率的研究中，个体受教育年限的决定因素构成了潜在的遗漏变量集合，研究者需要依据家庭教育行为决策方面的理论知识来寻找出可能的遗漏变量并加以控制。社会科学理论落实到具体研究情景时还需进一步结合现实制度背景。例如，对于受教育机会这同一个变量而言，在中国情景下户籍是一个核心因素（孙等，2025），而在美国情境下种族则可能更为关键（切蒂等，2020）。综上所述，研究者在选择控制变量时首先要借助社会科学相关理论并结合现实制度背景，深入分析具体研究问题可能存在的遗漏变量。在此基础上，充分详实地描述所选取的控制变量或代理变量的作用，明确其能够控制哪一个特定的遗漏变量，做到每一个控制变量的选取都要有理有据。

在实证研究中，研究者需要谨慎对待以下几种做法。第一，直接照搬相关文献的控制变量。对于同一个结果变量，不同的处理变量所需控制的遗漏变量完全不同。即使是相同的结果变量和处理变量，在不同时期、不同制度背景下，需要控制的遗漏变量也会发生变化。因此，直接照搬其他相关文献的控制变量的做法不可取。第二，不加说明地大量堆砌控制变量。若研究者不对控制变量的具体作用加以说明，既不利于读者理解文章研究设计的精巧之处，又可能引起文章存在人为筛选合意结果的怀疑，增添了p值操纵的风险。第三，单纯依靠统计显著性来选取控制变量。一些文献以回归模型中的控制变量是否显著作为选取标准，但是统计显著性并不能作为选择控制变量的充分理由，甚至也不是必要理由，有些不显著的控制变量依然有助于减少遗漏变量偏误[①]。从根本上讲，利用回归结果来判断变量是否控制，相当于给定结果去寻找原因，极易落入主观挑选合意结果的陷阱中。正因为如此，鲁宾（2008）明确指出，为了保证因果识别的客观性，研究者使用观测性数据进行实证研究时应基于处理分配机制来选择控制变量，不应使用任何涉及结果变量的信息（包括回归结果）[②]。因此，根据回归结果显著性等涉及结果变量信息的方法来控制变量是不正确的。

建议二：避免错误或不必要的控制变量，特别是要避免坏控制变量问题。

对于坏控制变量的判断，最根本的方法是依据社会科学理论和具体制度背景。现实中研究者们可能也会遇到一些实践方面的困难，例如在研究初期尚缺乏对因果结构的全面理解，无法判定某个特定变量在因果模型中的具体位置。在这种情况下，有一些基本的经验法则可供参考。首先，选取前定变量作为控制变量。好控制变量在因果模型中处于处理变量的上游，是先于处理变量确定的前定变量，而坏控制变量一般是受到处理变量直接或间接影响的后定变量。虽然在极个别的特殊情形中前定变量也可能是坏控制变量（如M-bias），但总体上看，以前定变量作为控制变量选择标准可以极大程度地避免坏控制变量。其次，对于双重差分法等动态分析模型，若研究者们认为某个变量是遗漏变量，可以选择政策发生前一年的值或前几年的均值作为政策前定特征，将其与特定形式的时间趋势项的交互项作为控制变量[③]。

坏控制变量问题在当前我国经济学和管理学实证研究中并非是偶发个例，而是一种普遍存在的现象。例如，许多政策评估类文献使用双重差分法研究政策冲击对企业行为的影响，这些研究普遍在回归模型中添加了企业营收、资产规模、资产回报率等时变控制变量。然而，政策发生后的企业特征必然或多或少、或直接或间接地受到政策影响，此种情形属于典型的坏控制变量。由于坏控制变量问题导致的估计偏误方向和大小都不确定，极端情况下甚至会使得估计系数的符号发生反转，对研究结论的可靠性产生较大危害。研究者在使用双重差分法时应着力避免这一问题。

建议三：使用线性回归模型时，需要对控制变量的重叠性做更多稳健性检验。

线性回归模型会对不满足重叠性假设的样本做线性外推，较为容易引起外推偏误，使得个别极端值样本变化可能对估计结果产生较大影响，导致研究结论不够稳健。本文提出以下几点建议以供参考：第一，在数据预处理阶段仔细观察变量的分布，删去极端值样本，具体操作时可将缩尾后仍偏离变量均值3倍或5倍标准差作为极端值判断标准；第二，在文中分组报告处理组和控制组的控制变量分布区间，明确其中满足重叠性假设

的区间范围;第三,在稳健性检验部分需要报告剔除不满足重叠性假设样本后的估计结果。第四,在线性回归模型之外,使用双重稳健估计量(double-robust estimator)等更为稳健的估计方法进行交叉验证。

建议四:添加固定效应时要特别关注有效样本范围的变化。

固定效应本质上是一种分组方式,固定效应层级越精细意味着分块数量越多,每个分块内的样本数量就会越少,越可能不满足重叠性假设。因此,固定效应层级精细度与有效样本范围之间往往存在着权衡关系,更精细的固定效应能够提高识别效力,但同时也会缩小结果适用范围。对此本文提出以下3点建议。第一,使用固定效应模型时要明确有效样本范围,将不影响估计结果的无效样本加以剔除。例如非平衡面板数据中只出现一期的个体样本,在添加个体固定效应后不再满足重叠性假设,应该予以剔除[29]。第二,避免控制过度精细的固定效应。固定效应不是免费的午餐,控制更为精细的固定效应是以降低统计功效和缩小有效样本范围为代价的。因此,研究者应选择与不可观测的遗漏变量相同层级的固定效应加以控制,而非不加思考直接控制更精细的固定效应。第三,不得控制与处理变量的变动性层级完全相同的固定效应。若固定效应与处理变量的层级完全相同,根据固定效应分组后,每一个分块内部都变得只有处理组或控制组,所有样本都会被排除在有效样本范围之外,此时研究设计完全失效,无法估计出相应系数,这一现象近年来在国际和国内期刊上均有发生,例如:使用中国宏观经济不确定性指数作为处理变量时(仅在年份层面有变异性)控制年份固定效应;使用时间断点回归方法时(仅在时间层面有变异性)控制日期固定效应;使用双重差分法研究城市层面政策冲击(如"宽带中国"政策,仅在城市—年份层面有变异性)的政策效应时控制城市—年份联合固定效应等等。如果研究者能够从根本上理解固定效应等价于分组比较的研究设计本质,相信能够极大程度地避免发生此类错误。

建议五:正确理解和使用机器学习等新兴控制变量选择方法。

近年来机器学习方法和人工智能与计量经济学的结合日趋紧密(郭峰、陶旭辉,2023)。使用机器学习方法帮助研究者在高维控制变量的情形下选择合适的控制变量越来越常见,代表性方法包括事后双重选择方法(贝洛尼等,2014)和双重机器学习(切尔诺朱科夫等,2018)等。机器学习方法主要用于在高维控制变量中挑选最具有解释力的控制变量,或是用非参数方法等手段捕捉控制变量的非线性影响。需要明确的是,机器学习方法在选择控制变量的标准主要是基于模型预测能力而非因果结构知识,因此不能够用于判断控制变量的"好"与"坏"。机器学习方法的适用情境是研究者已经筛选出了备选的好控制变量集合,此时可以使用机器学习方法帮助进一步筛选控制哪些变量、使用什么形式控制。在控制变量性质未知的情形下,机器学习方法不能帮助研究者区分出哪些是好控制变量,哪些是坏控制变量。一些研究者简单照搬文献中的做法,不加判断地将全部备选变量放入到机器学习模型中加以筛选,这种做法不可避免地会导致坏控制变量问题,最终得到有偏的估计结果。综上所述,本文建议研究者应正确理解机器学习等新兴方法的基本原理,明确其适用场景与应对的具体问题,做到方法的正确使用。在实践中,本文建议研究者应优先依据社会科学理论和现实制度背景挑选出可能的好控制变量集合,而后再使用双重机器学习的新兴方法做进一步筛选。

六、结论与展望

在以因果推断为核心的现代实证研究范式下,控制变量的选择已成为决定实证研究成败与否的核心环节。然而,近年来国内经济学、管理学等社会科学领域的实证研究使用控制变量时存在标准不清、挑选随意等现象,机械照搬既有文献、盲目堆砌控制变量,甚至为追求显著性结果而随意筛选控制变量的现象普遍存在,这些做法严重削弱了研究结论的可信度与学术价值。本文通过系统阐释控制变量在因果识别与统计推断中的作用机理,旨在纠正上述常见误区,为本土实证研究质量提升提供方法论上的保障。本文立足于鲁宾(2008)所强调的"让观测性研究逼近随机化实验"的实证研究核心精神,指出控制变量的核心作用在于其通过合理分层,在观测数据中构造出局部可比的处理组和控制组,从而逼近随机化实验的效果,实现因果推断

的核心目标。进一步研究揭示控制变量的双重机制：在因果识别层面，好控制变量能够消除选择性偏误，使处理分配在局部上近似随机；相对地，对撞变量与中介变量等坏控制变量不仅不能消除原有的偏误，反而会引入新的选择性偏误，从根本上破坏研究设计的有效性。在统计推断层面，与结果变量相关但与处理变量无关的控制变量能够有效降低方差、提升估计功效，而与处理变量高度相关的控制变量则会显著削弱估计精度。基于上述理论分析并结合当前我国实证研究的现状，本文提炼出选取和使用控制变量的5项基本原则，包括基于因果结构选择控制变量、高度重视坏控制变量问题、关注控制变量的重叠性、在复杂情况下权衡控制变量“利”“弊”，以及避免过度解读控制变量系数，并基于上述原则提出具有操作性的实践建议，从而为研究者提供一套能够提升研究设计严谨性、研究过程透明性与结果可复制性的控制变量选择框架。期望本文能够推动社会科学各领域实证研究的规范化、透明化与可信化，以实现学术知识的有效积累，进一步推动各学科发展与进步。

本文是对实证研究中控制变量问题的一次尝试性综述，受限于文章篇幅和笔者自身阅历及知识储备所限，仍然存在诸多不足和局限。第一，虽然本文以线性回归模型为例进行了深入剖析，但研究设计的视角本身具有更广泛的适用性，不应被线性回归模型所局限。例如在广义线性模型等非线性模型中，控制变量的选择与应用同样可以从研究设计的角度重新进行审视。第二，在基于设计的研究范式下，双重差分法、工具变量法、断点回归法等准自然实验方法的控制变量选择标准值得进一步深入探究。本文对控制变量选取原则的讨论是基于条件独立性识别假设所展开的。当识别假设发生变化时，控制变量的选择和使用原则理应随之改变。例如，双重差分法依赖于平行趋势假设而非条件独立性假设，在该识别假设下研究者们应该如何选择控制变量[25]？这些问题虽然文中有一定提及，但受限于篇幅，未能展开深入讨论。第三，本文主要分析了静态因果关系下的控制变量选择问题，但现实经济社会现象具有典型的动态特征，因果效应可能存在滞后、累积或动态反馈等不同形式。在动态因果关系框架下，控制变量的选择问题更为复杂，如何选择和使用各类控制变量以准确识别和估计动态因果效应，是未来研究的重要方向。第四，本文虽强调避免过度解读控制变量系数的重要性，但对于如何在实证研究中更加规范和透明地报告控制变量的使用情况，仍有待进一步深入探讨以达成学界共识。

控制变量是实证研究者开展因果推断的利器，同时也是一把双刃剑：正确选择可消除混杂偏误或提高估计精度，错误纳入则可能引入新偏误和损害估计效率。通过严谨的理论框架与数据模拟，本文为社会科学研究者提供系统、实用且前瞻性的控制变量选择方法论框架，希望能够为提升社会科学各领域实证研究的规范性、可信性与可重复性作出一些贡献。正如统计学家乔治·博克斯最广为人知的名言所说的：“所有模型都是错的，但有些是有用的”。唯有在理解控制变量的作用与局限性的基础上，才能正确发挥其作用，让其真正服务于社会科学诸领域的科学探索工作[26]。

（作者单位：张子尧，中南财经政法大学财政税务学院、中南财经政法大学收入分配与现代财政学科创新引智基地；黄炜，北京大学中国经济研究中心、北京大学国家发展研究院）

注释

①囿于伦理道德与高昂的研究成本限制，许多重要的经济社会问题不适合开展随机化实验研究。此外，卡特赖特（2007）、迪顿和卡特赖特（2018）等指出随机化实验的结论往往是“黑箱”式的，缺乏对内在因果机制的充分理解，同时也存在缺乏外部有效性（external validity）、研究结论难以推广等问题。

②德赫贾和瓦赫巴（1999，2002）使用拉隆德（1986）相同的数据，通过精巧地选择控制变量和估计方法，证明使用观测性数据能够得到与实验性数据相似的研究结论。

③布罗德等（2016）分析了50000多个发表在经济学期刊上的实证研究统计量，发现显著性水平在1%、5%和10%附近存在异常集中的现象。这暗示着研究者可能通过人为筛选控制变量等方式操纵实证研究结果，这种行为被统称为“p值操纵”（p-hacking）。布罗德等（2020）对发表在25本权威经济学期刊上的21000余个假设检验进行了统计分析，工具变量法（IV）表现出较高的p-hacking程度，而随机对照实验（RCT）和断点回归设计（RDD）则相对较低。米顿（2022）统计了公司金融实证研究的方法差异性，发现更换变量（包括选择不同的变量与变换变量函数形式）对估计系数的显著性具有明显影响。

④现代因果推断方法是基于因果结构而非相关关系来挑选控制变量，从根源上解决了该悖论。

⑤现在一般认为，内曼在1920~1930年代的工作是潜在结果框架的雏形。鲁宾在1970~1980年代的一系列重要工作系统化地发

展了潜在结果框架，成为目前因果推断方法的基础性工具。因此潜在结果框架也被称为鲁宾因果模型（Rubin Causal Model，RCM）。

⑥霍兰（1986）将之称为因果推断的基本问题（fundamental problem of causal inference）。

⑦在实证研究中，选择性偏误外和样本选择偏误既有区别又有联系。选择性偏误所导致的问题在于，可观测数据中观测到的变量相关关系无法被干净地识别为因果关系；样本选择偏误导致的问题在于，样本中观测到的变量相关关系无法代表总体相关关系。两者的联系在于，当研究者关心的是总体中的因果关系时，选择性偏误和样本选择偏误均会导致实证研究结论偏离真实情况。两者的区别在于，选择性偏误来源于处理状态分配过程，而样本选择偏误来源于数据抽样过程，两种问题的应对思路和处理方法完全不同，不可混淆。

⑧详细的数学证明过程见《管理世界》网络发行版附录2。

⑨这个例子的灵感来自于安格里斯特和皮施克（2014）的表2.1。

⑩详细证明见《管理世界》网络发行版附录3。

⑪严格来说，该表达式只是在控制变量个数远小于样本数时近似成立，精确的标准误表达式还需要根据控制变量个数和样本数来调整自由度。

⑫今井等（2010）提出和发展了一种因果中介分析方法试图更好地探究和量化中介变量相关的间接效应大小。赵西亮（2025）对该方法进行了详尽的综述。该方法需要序贯可忽略性假设（sequential ignorability assumption）成立，隐含着中介变量和结果变量之间不存在遗漏变量的要求。但正如前文所述，在观测性研究中几乎一定会存在中介变量和结果变量的共同原因，因此序贯可忽略性假设在实践中几乎不可能成立，大大限制了该方法的使用情景。

⑬图5是一个较为简化的因果模型，仅包含了实证研究中最常见的一些情形。对于复杂情况下的控制变量分类判断，读者可参考埃尔韦特和温希普（2014）与奇内利等（2022）等文献。

⑭这里是较为理想状态下的结论，若代理变量存在测量误差时该结论不一定成立。弗罗斯特（1979）指出当代理变量和遗漏变量的相关性非常微弱时，控制代理变量反而可能会放大选择性偏误。不过，在实践中这种极端情况非常少见，绝大多数情况下控制不完美的代理变量引起的偏误要远远小于其消除的选择性偏误，总体来看仍然有助于降低总偏误（伍德里奇，2019）。

⑮关于我国高等教育回报率的研究汗牛充栋，其中代表性研究包括邢春冰（2014）、马光荣等（2017）、刘生龙和胡鞍钢（2018）、贾和李（2021）等。

⑯伟大的统计学家乔治·博克斯曾有一句非常著名的名言“能分组的就分组，不能分组的就随机化”（Block what you can and randomize what you cannot.）。在笔者看来，这句话一语道尽了社会科学实证研究设计的精髓：分组控制用于处理可观测的遗漏变量，随机化用于处理不可观测的遗漏变量，结合二者能够更好地实现因果推断。

⑰当然，这并不意味着基于理论的先验知识是不可更改、不可动摇的。随着研究逐步深入，研究者可以基于已有的数据分析结果不断地更新对因果模型的认识，而后可以根据新的因果模型重新选择更为合适的控制变量。这一过程正是科学研究不断“扬弃”的体现。

⑱因本斯（2015）发现个别样本的控制变量存在极端值时，这些样本甚至可能被赋予负权重。张征宇和吴路遥（2025）发现双重差分法中不满足重叠性假设的极端控制变量有可能导致负权重问题。

⑲这个例子也从另一个角度说明，单纯地通过逐步回归或机器学习方法，根据估计系数的变化与显著性来判断控制变量的“好”与“坏”是不恰当的。

⑳例如本文第二章的模拟示例，使用饱和回归模型控制能力因素时，会使得有效样本范围缩小1/3，并且此时得到的因果效应估计不适用于能力极高和极低的群体。

㉑《管理世界》网络发行版附录4提供了一个具体示例。

㉒研究者可以想象自己正在做一项随机化实验，根据事前预设好的处理分配机制来确定个体处理状态，而在实验结束前无法观测到结果变量的信息。因此，基于处理分配机制来选择控制变量并不需要任何结果变量的信息。

㉓时间趋势项可以取一次项线性趋势或多次项非线性趋势，也可以用最为灵活的时间固定效应。

㉔Stata估计固定效应模型常用的reghdfe命令会自动排除此类单独样本（该命令将其称为singletons）。如果研究者手动添加了keepsingletons选项，Stata会保留这部分样本，此时估计系数保持不变，但由于样本数错误增加，可能会导致低估标准误、错误拒绝原假设，得到假阳性（false positive）的误导性结论。

㉕卡埃塔诺和卡拉韦（2024）、卡里姆和韦伯（2024）以及张征宇等（2024）讨论了双重差分法设定下的控制变量选择和使用方式，感兴趣的读者可进一步阅读。

㉖中外文人名（机构名）对照：费希尔（Fisher）；坎贝尔（Campbell）；厄尔巴赫（Erlebacher）；利默（Leamer）；拉隆德（Lalonde）；鲁宾（Rubin）；安格里斯特（Angrist）；皮施克（Pischke）；赫尔（Hull）；怀特德（Whited）；布罗德（Brodeur）；米顿（Mitton）；因本斯（Imbens）；奇内利（Cinelli）；徐（Xu）；赫克曼（Heckman）；珀尔（Pearl）；凯恩斯（Keynes）；哈维尔莫（Haavelmo）；罗森鲍姆（Rosenbaum）；丁（Ding）；吉本斯（Gibbons）；斯沃琴斯基（Słoczyński）；巴伦（Baron）；肯尼（Kenny）；孙（Sun）；切蒂（Chetty）；贝洛尼（Belloni）；切尔诺朱科夫（Chernozhukov）；卡特赖特（Cartwright）；迪顿（Deaton）；德赫贾（Dehejia）；瓦赫巴（Wahba）；内曼（Neyman）；霍兰（Holland）；今井（Imai）；埃尔韦特（Elwert）；温希普（Winship）；弗罗斯特（Frost）；伍德里奇（Wooldridge）；贾（Jia）；李（Li）；卡埃塔诺（Caetano）；卡拉韦（Callaway）；卡里姆（Karim）；韦伯（Webb）。

参考文献

（1）陈强：《计量经济学中的因果推断：过去、现在与未来》，《中山大学学报（社会科学版）》，2025年第1期。

（2）郭峰、陶旭辉：《机器学习与社会科学中的因果关系：一个文献综述》，《经济学（季刊）》，2023年第1期。

（3）江艇：《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》，2022年第5期。

（4）刘生龙、胡鞍钢：《大学教育回报：基于大学扩招的自然实验》，《劳动经济研究》，2018年第4期。

（5）马光荣、纪洋、徐建炜：《大学扩招如何影响高等教育溢价？》，《管理世界》，2017年第8期。
（6）王美今、林建浩：《计量经济学应用研究的可信性革命》，《经济研究》，2012年第2期。
（7）邢春冰：《教育扩展、迁移与城乡教育差距——以大学扩招为例》，《经济学（季刊）》，2014年第1期。
（8）许琪：《因果推断五十年：成就、挑战与应对》，《学术月刊》，2024年第11期。
（9）鄢伟波：《高等教育溢价变动的新趋势与解释——以流动人口为例》，《劳动经济研究》，2022年第5期。
（10）张征宇、林丽花、曹思力、周亚虹：《双重差分设计下固定效应估计量何时可信？——若干有用的建议》，《管理世界》，2024年第1期。
（11）张征宇、吴路遥：《双重差分法下固定效应估计量的负权重问题——新的机制与解决方案》，《数量经济技术经济研究》，2025年第2期。
（12）赵西亮：《因果中介分析的理论进展及其应用》，《数量经济技术经济研究》，2025年第2期。
（13）Angrist, J. D. and Pischke, J. S., 2009, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.
（14）Angrist, J. D. and Pischke, J. S., 2010, "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics", *Journal of Economic Perspectives*, vol.24(2), pp.3~30.
（15）Angrist, J. D. and Pischke, J. S., 2014, *Mastering 'Metrics: The Path from Cause to Effect*, Princeton: Princeton University Press.
（16）Baron, R. M. and Kenny, D. A., 1986, "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations", *Journal of Personality and Social Psychology*, vol.51(6), pp.1173~1182.
（17）Belloni, A., Chernozhukov, V. and Hansen, C., 2014, "Inference on Treatment Effects after Selection among High-Dimensional Controls", *The Review of Economic Studies*, vol.81(2), pp.608~650.
（18）Brodeur, A., Cook, N. and Heyes, A., 2020, "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics", *American Economic Review*, vol.110(11), pp.3634~3660.
（19）Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y., 2016, "Star Wars: The Empirics Strike Back", *American Economic Journal: Applied Economics*, vol.8(1), pp.1~32.
（20）Caetano, C. and Callaway, B., 2024, "Difference-in-Differences When Parallel Trends Holds Conditional on Covariates", Working Paper.
（21）Campbell, D. T. and Erlebacher, A., 1970, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful", in Hellmuth, J., eds: *Compensatory Education: A National Debate*, New York: Brunner/Mazel.
（22）Cartwright, N., 2007, "Are RCTs the Gold Standard?", *BioSocieties*, vol.2(1), pp.11~20.
（23）Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E. and Hansen, C., 2018, "Double/Debiased Machine Learning for Treatment and Structural Parameters", *The Econometrics Journal*, vol.21(1), pp.1~68.
（24）Chetty, R., Hendren, N., Jones, M. and Porter, S., 2020, "Race and Economic Opportunity in the United States: An Intergenerational Perspective", *The Quarterly Journal of Economics*, vol.135(2), pp.711~783.
（25）Cinelli, C., Forney, A. and Pearl, J., 2022, "A Crash Course in Good and Bad Controls", *Sociological Methods & Research*, vol.53(3), pp.1071~1104.
（26）Deaton, A. and Cartwright, N., 2018, "Understanding and Misunderstanding Randomized Controlled Trials", *Social Science & Medicine*, vol.210, pp.2~21.
（27）Dehejia, R. H. and Wahba, S., 1999, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, vol.94(448), pp.1053~1062.
（28）Dehejia, R. H. and Wahba, S., 2002, "Propensity Score-Matching Methods for Nonexperimental Causal Studies", *Review of Economics and Statistics*, vol.84(1), pp.151~161.
（29）Ding, P., 2024, *A First Course in Causal Inference*, New York: CRC Press.
（30）Elwert, F. and Winship, C., 2014, "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable", *Annual Review of Sociology*, vol.40, pp.31~53.
（31）Fisher, R. A., 1935, *The Design of Experiments*, London: Oliver and Boyd.
（32）Frost, P. A., 1979, "Proxy Variables and Specification Bias", *The Review of Economics and Statistics*, vol.61(2), pp.323~325.
（33）Gibbons, C. E., Suárez Serrato, J. C. and Urbancic, M. B., 2019, "Broken or Fixed Effects?", *Journal of Econometric Methods*, vol.8(1), Article 20170002.
（34）Haavelmo, T., 1944, "The Probability Approach in Econometrics", *Econometrica*, vol.12, pp.iii~115.
（35）Heckman, J. J., 2000, "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *The Quarterly Journal of Economics*, vol.115(1), pp.45~97.
（36）Holland, P. W., 1986, "Statistics and Causal Inference", *Journal of the American Statistical Association*, vol.81(396), pp.945~960.
（37）Hull, P., Kolesár, M. and Walters, C., 2022, "Labour by Design: Contributions of David Card, Joshua Angrist, and Guido Imbens", *The Scandinavian Journal of Economics*, vol.124(3), pp.603~645.
（38）Imai, K., Keele, L. and Tingley, D., 2010, "A General Approach to Causal Mediation Analysis", *Psychological Methods*, vol.15(4), pp.309~334.
（39）Imbens, G. W., 2015, "Matching Methods in Practice: Three Examples", *Journal of Human Resources*, vol.50(2), pp.373~419.

(40) Imbens, G. W. and Rubin, D. B., 2015, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press.

(41) Imbens, G. W. and Xu, Y., 2024, "LaLonde (1986) after Nearly Four Decades: Lessons Learned", Working Paper.

(42) Jia, R. and Li, H., 2021, "Just Above the Exam Cutoff Score: Elite College Admission and Wages in China", *Journal of Public Economics*, vol.196, Article 104371.

(43) Karim, S. and Webb, M. D., 2024, "Good Controls Gone Bad: Difference-in-Differences with Covariates", Working Paper.

(44) Keynes, J. M., 1939, "Professor Tinbergen's Method", *The Economic Journal*, vol.49(195), pp.558~577.

(45) LaLonde, R. J., 1986, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *The American Economic Review*, vol.76(4), pp.604~620.

(46) Leamer, E. E., 1983, "Let's Take the Con Out of Econometrics", *The American Economic Review*, vol.73(1), pp.31~43.

(47) Mitton, T., 2022, "Methodological Variation in Empirical Corporate Finance", *The Review of Financial Studies*, vol.35(2), pp.527~575.

(48) Neyman, J., 1923, "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with Discussion). Section 9.", *Statistical Science*, pp.465~472.

(49) Pearl, J., 2009, *Causality*, Cambridge: Cambridge University Press.

(50) Rosenbaum, P. R. and Rubin, D. B., 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, vol.70(1), pp.41~55.

(51) Rubin, D. B., 1974, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, vol.66(5), pp.688~701.

(52) Rubin, D. B., 2008, "For Objective Causal Inference, Design Trumps Analysis", *The Annals of Applied Statistics*, vol.2(3), pp.808~840.

(53) Słoczyński, T., 2022, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights", *The Review of Economics and Statistics*, vol.104(3), pp.501~509.

(54) Sun, Y., Zhao, L. Q. and Zhao, Z., 2025, "Hukou Status and Children's Education in China", *Economic Development and Cultural Change*, vol.73(2), pp.979~1021.

(55) Whited, R. L., Swanquist, Q. T. and Shipman, J. E., 2022, "Out of Control: The (Over) Use of Controls in Accounting Research", *The Accounting Review*, vol.97(3), pp.395~413.

(56) Wooldridge, J., 2019, *Introductory Econometrics: A Modern Approach*, Boston: Cengage Learning.

Control Variable Selection in Empirical Research: Theory and Principles

Zhang Ziyao[a,b] *and Huang Wei*[c,d]

(a. School of Public Finance and Taxation, Zhongnan University of Ecomics and Law; b. Innovation and Talent Base for Income Distribution and Public Finance, Zhongnan University of Economics and Law; c. China Center for Economic Research, Peking University; d. National School of Development, Peking University)

Abstract: The selection of control variables is a critical determinant of the validity of causal inference in empirical research. This paper develops a systematic framework that clarifies the conceptual foundations, selection criteria, and practical guidelines for their use. Leveraging the potential outcomes framework and a decomposition of linear regression estimators, we distinguish two essential functions of controls: causal identification and statistical inference. For identification, good controls enable observational studies to approximate randomized experiments by conditioning on relevant strata, whereas bad controls may generate selection bias. For inference, suitable controls attenuate noise and enhance precision, while poor choices inflate estimation error and weaken statistical power. Whether a control is "good" or "bad" depends on its position within the causal structure, which should be grounded in theory and institutional context rather than data-driven selection. Building on this insight, we classify common types of controls and articulate five principles for applied research: grounding control selection in causal structure, rigorously addressing bad controls, ensuring overlap, weighing trade-offs in complex settings, and avoiding misinterpretation of coefficients. Practical recommendations are provided accordingly. This framework offers an operational methodology to improve research design and strengthens the credibility, transparency, and replicability of empirical work in the social sciences.

Keywords: control variables; causal inference; observational studies; selection bias

Control Variable Selection in Empirical Research: Theory and Principles

Zhang Ziyao[a,b] *and Huang Wei*[c,d]

(a. School of Public Finance and Taxation, Zhongnan University of Ecomics and Law; b. Innovation and Talent Base for Income Distribution and Public Finance, Zhongnan University of Economics and Law; c. China Center for Economic Research, Peking University; d. National School of Development, Peking University)

Summary: The selection of control variables is a central determinant of the credibility and validity of causal inference in observational studies. Despite their widespread use in applied microeconomic research, control variables are often selected in an ad hoc or data-driven manner, or selected mechanically from prior studies. Such practices risk introducing selection bias, inflating estimation variance, and undermining the interpretability and replicability of empirical findings. Motivated by the growing emphasis on design-based inference following the "credibility revolution", this paper proposes a theory-driven framework for the proper selection and use of control variables.

The paper begins by contrasting two major empirical paradigms: the model-based approach, which focuses on specifying the data-generating process (DGP) of the outcome variable, and the design-based approach, which emphasizes the treatment assignment mechanism and seeks to approximate randomized experiment. Within the design-based framework, both saturated and standard linear regressions are shown to function as special cases of matching estimators, with control variables enabling local comparisons across units with similar covariates. This reinterpretation positions regression analysis not as a structural model per se, but as a method for implementing quasi-experimental designs, thereby reframing the role of control variables from statistical adjustment to identification strategy.

We highlight the dual role of control variables. For identification, valid controls eliminate selection bias by conditioning on pre-treatment covariates that jointly affect treatment and outcome. In contrast, bad controls—such as colliders or mediators—can introduce new sources of bias or block causal pathways, distorting the estimated effect. For inference, control variables that reduce the residual variance of the outcome can improve estimation precision, while those highly correlated with the treatment can inflate standard errors and reduce power.

Based on this framework, the paper offers a set of concise principles to guide empirical applications. Researchers should select control variables based on a clearly defined causal model, rather than statistical significance or predictive performance. Variables that are post-treatment or lie on the causal pathway should generally be excluded to avoid bias from bad controls. Ensuring sufficient covariate overlap between treated and untreated units is critical, and researchers should use diagnostics such as propensity score distributions to assess common support. When overlap is weak, trimming or reweighting can mitigate extrapolation risks. Finally, the inclusion of control variables should be justified with explicit reference to their identification role, and empirical strategies should distinguish between variables included for bias reduction and those added for improving precision.

The contributions of the paper are threefold. First, it provides a systematic synthesis of control variable logic across model-based and design-based paradigms, clarifying how control variables function not only to address omitted variable bias but also to simulate blocked randomized designs. Second, it builds a formal connection between linear regression estimators and matching estimators, including under misspecified propensity score models, thereby offering new interpretive clarity on how control variables shape weighting, extrapolation, and identification. Third, the paper addresses pressing challenges in applied research, particularly within the Chinese context, where mechanical control variable selection, overcontrol, and insufficient attention to causal structure remain common. The proposed framework offers concrete, applicable guidance for improving research design quality, enhancing empirical transparency, and strengthening the credibility of causal claims.

Keywords: control variables; causal inference; observational studies; selection bias

JEL Classification: C01

《实证研究中的控制变量选择：原理与原则》附录

1. 饱和回归模型估计量的因果解释推导过程

根据FWL定理，可以使用两步法估计饱和回归模型的估计系数。第一步将处理状态D_i对控制变量X_i的虚拟变量集合回归：

$D_i=\sum_{g=1}^{G}\gamma^g\mathbb{I}(X_i=x_g)+u_i^d$

估计系数γ^g等于$X_i=x_g$组内的处理状态变量均值$\overline{D}_{ig}$（即处理组个体比重），也可以理解为$X_i=x_g$时的倾向得分$e_g=\Pr(D_i=1|X_i=x_g)$。

第二步将结果变量Y_i对第一步回归的残差$u^d_i\equiv D_i-\overline{D}_{ig}$进行回归：

$Y_i=\beta^{sat}(D_i-\overline{D}_{ig})+u^Y_i$

饱和回归模型估计量的表达式为如下形式：

$$\hat{\beta}^{sat}=\frac{\frac{1}{N}\sum_i^N(D_i-\bar{D}_{ig})Y_i}{\frac{1}{N}\sum_i^N(D_i-\bar{D}_{ig})^2}$$

分子可以重新表述为如下形式：

$$\frac{1}{N}\sum_i^N(D_i-\bar{D}_{ig})Y_i=\frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(D_i-e_g)Y_i=\frac{1}{N}\sum_{g=1}^{G}\left[\sum_{i=1}^{n_g^t}(1-e_g)Y_i-\sum_{j=1}^{n_g^c}e_gY_j\right]=\frac{1}{N}\sum_{g=1}^{G}n_g\left[(1-e_g)\frac{n_g^t}{n_g}\bar{Y}_g^t-e_g\frac{n_g^c}{n_g}\bar{Y}_g^c\right]$$

$$=\sum_{g=1}^{G}\frac{n_g}{N}(1-e_g)e_g(\bar{Y}_g^t-\bar{Y}_g^c)=\sum_{g=1}^{G}\Pr(X_i=x_g)\mathrm{var}(D_i|X_i=x_g)\hat{\tau}(X_i=x_g)$$

分母可以重新表述为如下形式：

$$\frac{1}{N}\sum_i^N(D_i-\bar{D}_{ig})^2=\frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(D_i-e_g)^2=\frac{1}{N}\sum_{g=1}^{G}\left[\sum_{i=1}^{n_g^t}(1-e_g)^2+\sum_{j=1}^{n_g^c}e_g^2\right]=\sum_{g=1}^{G}\frac{n_g}{N}\left[\frac{n_g^t}{n_g}(1-e_g)^2+\frac{n_g^c}{n_g}e_g^2\right]$$

$$=\sum_{g=1}^{G}\frac{n_g}{N}\left[e_g(1-e_g)^2+(1-e_g)e_g^2\right]=\sum_{g=1}^{G}\frac{n_g}{N}e_g(1-e_g)=\sum_{g=1}^{G}\Pr(X_i=x_g)\mathrm{var}(D_i|X_i=x_g)$$

综合上述结果，饱和回归模型估计量等于：

$$\hat{\beta}^{sat}=\sum_{g=1}^{G}\frac{\Pr(X_i=x_g)\mathrm{var}(D_i|X_i=x_g)}{\sum_{g=1}^{G}\Pr(X_i=x_g)\mathrm{var}(D_i|X_i=x_g)}\hat{\tau}(X_i=x_g)=\sum_{g=1}^{G}w_g\hat{\tau}(X_i=x_g)$$

2. 一般型控制变量情形下的多元线性回归模型OLS估计量推导过程

证明：

$$\hat{\beta}^{ols}=\frac{\frac{1}{N}\sum_{i=1}^{N}\tilde{D}_iY_i}{\frac{1}{N}\sum_{i=1}^{N}\tilde{D}_i^2}$$

分子等于

$$\frac{1}{N}\sum_{i=1}^{N}\tilde{D}_iY_i=\frac{1}{N}\sum_{i=1}^{N}(D_i-L(X_i))Y_i=\frac{1}{N}\sum_{x_i=x_1}^{x_G}\sum_{i=1}^{n_g}(D_i-L(X_i=x_g))Y_i$$

$$=\frac{1}{N}\sum_{x_1}^{x_G}\left\{\sum_{i\in t}^{n_g^t}(1-L(x_g))Y_i-\sum_{i\in c}^{n_g^c}L(x_g)Y_i\right\}=\frac{1}{N}\sum_{x_1}^{x_G}\left\{n_g^t(1-L_g)\frac{1}{n_g^t}\sum_{i\in t}^{n_g^t}Y_i-L_gn_g^c\frac{1}{n_g^c}\sum_{i\in c}^{n_g^c}Y_i\right\}$$

$$=\sum_{x_1}^{x_G}\frac{n_g}{N}\left\{\frac{n_g^t}{n_g}(1-L_g)\bar{Y}_g^t-\frac{n_g^c}{n_g}L_g\bar{Y}_g^c\right\}=\sum_{x_1}^{x_G}\frac{n_g}{N}\left\{e_g(1-L_g)\bar{Y}_g^t-(1-e_g)L_g\bar{Y}_g^c\right\}$$

$$=\sum_{x_1}^{x_G}\frac{n_g}{N}\left\{\left[e_g(1-e_g)-e_g(L_g-e_g)\right]\bar{Y}_g^t-\left[(1-e_g)e_g+(1-e_g)(L_g-e_g)\right]\bar{Y}_g^c\right\}$$

$$=\sum_{x_1}^{x_G}\frac{n_g}{N}\left\{\left[e_g(1-e_g)\right](\bar{Y}_g^t-\bar{Y}_g^c)\right\}-\sum_{x_1}^{x_G}\frac{n_g}{N}(L_g-e_g)\left[e_g\bar{Y}_g^t+(1-e_g)\bar{Y}_g^c\right]$$

当使用饱和回归模型估计时有$e_g=FE_g$，故OLS估计量可以表示为

$$\hat{\beta}^{ols}=\frac{\frac{1}{N}\sum_{i=1}^{N}(D_i-e_g)^2}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}\frac{\sum_{x_1}^{x_G}\frac{n_g}{N}e_g(1-e_g)(\bar{Y}_g^t-\bar{Y}_g^c)}{\frac{1}{N}\sum_{i=1}^{N}(D_i-e_g)^2}-\frac{\sum_{x_1}^{x_G}\frac{n_g}{N}(L_g-e_g)\left[e_g\bar{Y}_g^t+(1-e_g)\bar{Y}_g^c\right]}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}$$

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}\left[(D_i-L_g)+(L_g-e_g)\right]^2}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}\hat{\beta}^{sat}-\frac{\sum_{x_1}^{x_G}\frac{n_g}{N}(L_g-e_g)\left[e_g\bar{Y}_g^t+(1-e_g)\bar{Y}_g^c\right]}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}$$

$$=\frac{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2-\frac{1}{N}\sum_{i=1}^{N}(L_g-e_g)^2}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}\hat{\beta}^{sat}-\frac{\sum_{x_1}^{x_G}\frac{n_g}{N}(L_g-e_g)\left[e_g\bar{Y}_g^t+(1-e_g)\bar{Y}_g^c\right]}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L_g)^2}$$

$$=\left(1-\frac{\frac{1}{N}\sum_{i=1}^{N}(L(X_i)-e(X_i))^2}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L(X_i))^2}\right)\hat{\beta}^{sat}-\frac{\frac{1}{N}\sum_{i=1}^{N}(L(X_i)-e(X_i))Y_i}{\frac{1}{N}\sum_{i=1}^{N}(D_i-L(X_i))^2}$$

$$p\lim\rightarrow\left(1-\frac{\mathrm{var}(L(X_i)-e(X_i))}{\mathrm{var}(D_i-L(X_i))}\right)\beta^{sat}-\frac{\mathrm{cov}\left[(L(X_i)-e(X_i))Y_i\right]}{\mathrm{var}(D_i-L(X_i))}$$

第二项为

$$\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(D_i-L(X_i)\right)}=\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(L(X_i)-e(X_i)\right)}\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}$$

综合起来有

$$\beta^{sat}-\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\beta^{sat}-\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(L(X_i)-e(X_i)\right)}\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}=\beta^{sat}-\left\{\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\beta^{sat}+\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(D_i-L(X_i)\right)}\right\}$$

$$\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\beta^{sat}+\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(D_i-L(X_i)\right)}=\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\beta^{sat}+\frac{\operatorname{var}\left(L(X_i)-e(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\frac{\operatorname{cov}\left[\left(L(X_i)-e(X_i)\right)Y_i\right]}{\operatorname{var}\left(L(X_i)-e(X_i)\right)}$$

$$=\frac{\operatorname{var}\left(e(X_i)-L(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\left\{\beta^{sat}-\frac{\operatorname{cov}\left[\left(e(X_i)-L(X_i)\right)Y_i\right]}{\operatorname{var}\left(e(X_i)-L(X_i)\right)}\right\}$$

进一步运算：

$$\beta^{ols}\to\beta^{sat}+\frac{\operatorname{var}\left(e(X_i)-L(X_i)\right)}{\operatorname{var}\left(D_i-L(X_i)\right)}\left\{\beta^{sat}-\frac{\operatorname{cov}\left[\left(e(X_i)-L(X_i)\right)Y_i\right]}{\operatorname{var}\left(e(X_i)-L(X_i)\right)}\right\}$$

当倾向得分模型是线性模型时有 $e_g=L_g$，此时 $\hat{\beta}^{ols}=\hat{\beta}^{sat}$。

饱和回归模型与线性回归的另一个区别在于，不满足重叠性假设的样本不影响饱和回归模型，但会影响线性回归模型。

3. 精确匹配估计量和饱和回归模型OLS估计量的对比

精确匹配估计结果为：

$$\hat{\beta}^{match}=\sum_{g\in\{0<e_g<1\}}\frac{\Pr\left(X_i=x_g\right)}{\sum_{g\in\{0<e_g<1\}}\Pr\left(X_i=x_g\right)}\hat{\tau}\left(X_i=x_g\right)=\frac{1/4}{1/4+1/4+1/6}\times 50+\frac{1/4}{1/4+1/4+1/6}\times 30+\frac{1/4}{1/4+1/4+1/6}\times 20=35$$

饱和回归模型的OLS估计量为：

$$\hat{\beta}^{sat}=\sum_{g=1}^{G}\frac{\Pr\left(X_i=x_g\right)\operatorname{var}\left(D_i\middle|X_i=x_g\right)}{\sum_{g=1}^{G}\Pr\left(X_i=x_g\right)\operatorname{var}\left(D_i\middle|X_i=x_g\right)}\hat{\tau}\left(X_i=x_g\right)=\frac{\frac{1}{4}\times\frac{2}{9}}{\frac{1}{4}\times\frac{2}{9}+\frac{1}{4}\times\frac{2}{9}+\frac{1}{6}\times\frac{1}{4}}\times 50+\frac{\frac{1}{4}\times\frac{2}{9}}{\frac{1}{4}\times\frac{2}{9}+\frac{1}{4}\times\frac{2}{9}+\frac{1}{6}\times\frac{1}{4}}\times 30+\frac{\frac{1}{6}\times\frac{1}{4}}{\frac{1}{4}\times\frac{2}{9}+\frac{1}{4}\times\frac{2}{9}+\frac{1}{6}\times\frac{1}{4}}\times 20$$

≈ 35.545

可以看到，两种估计方法的条件平均处理效应估计结果完全相同，只有加权权重存在部分差异。

4. 不显著的控制变量与遗漏变量偏误

本节提供一个具体示例，用于说明控制变量不显著时仍然需要加以控制。

设数据生成过程如下：

$Y=X+P-1.65Q+\varepsilon^{Y}$

$X=P+\varepsilon^{X}$

$P=Q+\varepsilon^{P}$

其中 $Q\sim N(0,3)$，$\varepsilon^{Y},\varepsilon^{X}\sim N(0,1)$，$\varepsilon^{P}\sim N(0,2)$。

使用该数据生成过程进行模拟，结果报告在附表1中。P 是 Y 和 X 的共同原因，故其是一个遗漏变量。P 对 Y 的真实系数为1，然而由于 Q 的存在，第(2)列中 P 的估计系数因选择性偏误偏离了真实系数并且不具有统计显著性。对比第(1)列和第(2)列的结果，可以看到添加控制变量 P 后，X 的估计系数更加接近真实系数1。因此，即使控制变量 P 不显著，对其进行控制仍然有助于消除选择性偏误。

附表1　不显著的控制变量与估计结果

	(1)	(2)
X	0.871***	0.982***
	(0.0247)	(0.0920)
P		−0.118
		(0.0945)
观测值	1000	1000