第 25 卷第 6 期
2025 年 11 月

经 济 学（季 刊）
China Economic Quarterly

Vol. 25, No. 6
November, 2025

工具变量法最新理论发展与应用展望

王 也 李海风 杨汝岱 易君健*

摘 要：现有因果推断方法主要基于三种思路解决由不可观测的异质性导致的内生性问题，其中基于排他性约束的工具变量法（IV）是最主要的方法之一。当存在异质性处理效应时，局部平均处理效应（LATE）框架为 IV 估计量提供了一个清晰合理的解释。但是，当使用多个工具变量或内生变量为非二元变量时，受限于对个体偏好结构和选择结构的简单假设，经典的 LATE 框架不能为 IV 估计量提供一个合理的解释。同时，经典的 LATE 框架不能估计非依从者的平均处理效应。关于这些问题的讨论推动了工具变量法的不断发展，本文从三方面对其进行总结和阐述。首先，本文讨论了最新文献在多个工具变量或内生变量为非二元变量时对 LATE 框架的补充和拓展；其次，梳理了边际处理效应（MTE）与 LATE 的关系以及如何基于 MTE 将 LATE 外推得到非依从者的平均处理效应；最后，本文还归纳了 Bartik 工具变量这一特殊的连续型工具变量的最新理论发展。

关键词：局部平均处理效应；边际处理效应；Bartik 工具变量
DOI：10.13821/j.cnki.ceq.2025.06.01

一、引 言

因果推断方法需要解决的问题是：不可观测的异质性（unobserved heterogeneity）会影响样本是否接受处理，从而导致样本选择偏误（selection bias），此时得到的变量之间的关系不能被解释为因果关系。现有因果推断方法主要基于三种思路解决该问题。第一种思路是通过随机（randomization）的方法使得控制组和处理组不可观测的异质性保持平衡（balanced）。比如随机对照试验（randomized controlled trial, RCT）和精确断点回归法（sharp regression discontinuity, sharp RD）①。第二种思路是直接对不可观测的异质性的结构进行具体假设。比如固定效应模型（fixed effects model）假设不可观测的异质性不会同时随时间和个体的变化而变化，匹配法（matching）假设所有的异质性都可以用可观测变量表示出来，即不存在不可观测的异质性。第三种思路是使用排他性约束（exclusion restriction），通过引入与内生变量相关而与不可观测的异质性不相关的工具变量，从内生变量中分离出与不可观测的异质性不相关的部分。比如工具变量法（in-

* 王也、杨汝岱，北京大学经济学院；李海风，福州大学经济与管理学院；易君健，北京大学国家发展研究院。通信作者及地址：李海风，福建省福州市福州大学城乌龙江北大道 2 号福州大学经济与管理学院，350108；电话：15321596109；E-mail：lhf_0730@163.com。

① 精确断点回归法也可以被理解为对不可观测的异质性作连续性假设。

strumental variable，IV)和模糊断点回归法(fuzzy RD)。这些方法各有其特定的适用情境和条件。其中工具变量法是最主要的方法之一，在因果推断的实证研究中具有重要地位。从 2000 年到 2022 年，五本经济学重要中文期刊中使用常见因果推断方法的论文共有 1 174 篇，其中使用工具变量法的论文有 721 篇，占比为 61.41%。① 从图 1 可以看出，在中外经济学重要期刊中，与工具变量法相关的论文占比呈现递增趋势。②

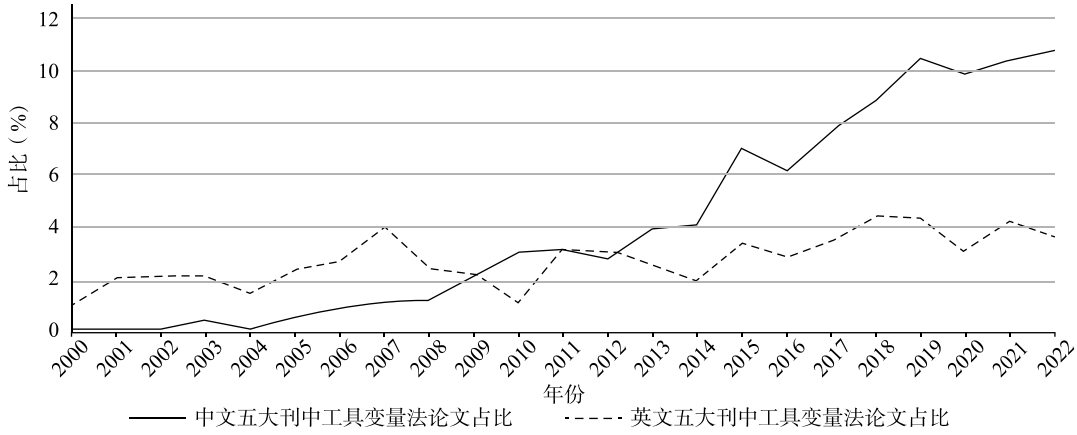

图 1 经济学重要期刊中与工具变量法相关的论文占比

从工具变量法的发展历程来看，这种基于排他性约束处理内生性问题的思路历史悠久，最早可以追溯到 20 世纪 20 年代(Stock and Trebbi，2003)。传统的工具变量法假设所有个体的处理效应都是相同的，所以个体的处理效应和总体的平均处理效应相等，只要满足排他性和相关性假设，IV 估计量就可以被解释为总体的平均处理效应。但是当存在异质性处理效应(heterogeneous treatment effect)且是否接受处理受到异质性处理效应的影响时，IV 估计量如何解释一直未有定论。直到 Imbens and Angrist(1994)提出了局部平均处理效应(local average treatment effect，LATE)理论，在简洁的非参设定下为 IV 估计量提供了一个清晰有力的解释。在 LATE 框架下，工具变量和内生变量均为二元变量，当满足独立性、排他性、相关性、单调性四个假设时，IV 估计量可以被解释为依从者(compliers)的平均处理效应。该框架允许异质性处理效应存在，不进行任何分布和方程形式的假设。在过去的三十年里，LATE 理论极大地推动了经济学在因果推断研究上的发展。

但是经典的 LATE 理论也存在一些问题和局限性。第一，当使用多个工具变量或内生变量为非二元变量时，经典的 LATE 框架对个体的偏好结构和选择结构的简单假设使

① 常见的因果推断方法包括工具变量法、双重差分法、断点回归法、匹配法等。五本经济学重要中文期刊指《经济研究》、《管理世界》、《世界经济》、《经济学》(季刊)和《中国工业经济》。参考黄炜等(2022)，我们在知网中检索了上述五本期刊中主题、篇名、摘要或关键词出现“工具变量”的论文以及主题、篇名、摘要或关键词出现“工具变量”“双重差分”“倍差”“断点回归”或“倾向得分匹配”的论文，进而求出两者之间的比例。

② 中文五大经济学期刊同上，我们在知网中检索了这五本期刊中主题、篇名、摘要或关键词出现“工具变量”的论文比例。英文五大经济学期刊指 *American Economic Review*、*Quarterly Journal of Economics*、*Econometrica*、*Journal of Political Economy* 和 *Review of Economic Studies*，我们在 Web of Science 中检索了这五本期刊中主题、标题、摘要或索引中出现“instrument”“instrumental variable”或“instrumental variables”的论文比例。

其失去对 IV 估计量的合理解释力。第二，LATE 框架只能估计依从者的平均处理效应，而无法估计非依从者的平均处理效应，例如始终接受者(always takers)、从不接受者(never takers)和政策制定者关心的目标政策受众群体的平均处理效应。第三，在工具变量或内生变量为连续变量的情况下，LATE 理论无法对异质性处理效应下的 IV 估计量进行解释。这些问题和局限性引发了理论方法层面的一系列研究讨论和创新突破，推动了工具变量法的不断发展，本文旨在对这些最新进展进行总结和阐述。首先，本文第二部分到第四部分从 LATE 和边际处理效应(marginal treatment effect，MTE)两方面综述了异质性处理效应下工具变量法的最新进展。第二部分主要介绍当使用多个工具变量或内生变量为非二元变量时，前沿文献对 LATE 框架下单调性假设提出的挑战与补充拓展。第三部分介绍最新文献如何基于选择模型(selection model)和边际处理效应将依从者的平均处理效应外推得到始终接受者和从不接受者的平均处理效应。第四部分讨论在异质性处理效应下 LATE 和 MTE 估计量的经济学基础，并与结构估计法进行对比，介绍三者在政策评估应用中的差异。其次，本文在第五部分介绍一种特殊的工具变量，即 Bartik 工具变量的最新文献发展。最后，本文基于工具变量法在实际应用中的优势与挑战，对未来的应用实践提出建议与展望。

本文认为，研究方法的拓展都是希望能更好地应对研究过程中新的理论和实践问题，研究方法的适应性同样有一般性和特殊性的区分。工具变量法作为因果推断的一种基本的一般性方法，在推动经济学和其他相关学科发展过程中起到了重要的作用。但随着研究的不断深入，随着对研究结论准确性的要求越来越高，对研究方法提出了更高的要求。在这个背景下，本文从三个方面对工具变量法的最新前沿进展做了较为详细的综述性讨论，这可以帮助研究者了解这些前沿方法的优点和局限性。尤其需要注意的是，特定维度的方法进展有着特定的适应性，主要是针对特定的理论问题和现实问题，我们更希望研究者通过本文的研究能够去真正深入思考方法拓展的内在逻辑，思考方法为什么会沿着这一方向拓展。只有这样，在现实的研究过程中，才能真正做到方法的选择是服务于理论问题研究，真正做到理论、方法、现实问题的协调统一，真正提高研究水平。

二、局部平均处理效应(LATE)的最新发展

当使用多个工具变量或内生变量为非二元变量时，经典 LATE 理论的简单假设难以有效刻画异质性个体的偏好结构和选择结构，从而无法合理地解释该情形下的 IV 估计量。为此，最新文献主要针对 LATE 框架下异质性偏好的相关假设展开讨论和拓展。在本部分中，首先回顾 LATE 定理的基本假设和结论，并在此框架下分别讨论当多个工具变量或内生变量有多个取值时，LATE 定理对异质性偏好刻画的局限性及其最新理论发展。

(一) LATE 定理的回顾

Imbens and Angrist(1994)提出，在异质性处理效应的框架下，当满足独立性、排他性、工具变量相关性、单调性四个假设时，工具变量法的估计结果应该被解释为依从者的

平均处理效应，这种效应被称为局部平均处理效应。具体地，我们假设 y 和 d 为可观测到的结果变量和处理变量，y^0、y^1 为代表了处理变量 d 取值为0和1时 y 的潜在结果；同理，d^0、d^1 为代表了工具变量 z 取值为0和1时 d 的潜在结果。此外，定义 $y^{d,z}$ 为在工具变量 z、处理变量 d 下 y 的潜在结果，记 β^{IV} 为使用工具变量法得到的平均处理效应。LATE定理的正式表达如下：

LATE定理（Imbens and Angrist，1994）。假设：

（1）假设1（独立性，independence）：$\{y^1, y^0, d^1, d^0\} \perp z$；

（2）假设2（排他性，exclusion）：$y^{d,z} = y^d$，对于 $d = 0, 1$；

（3）假设3（工具变量相关性，instrument relevance）：$E[d^1 - d^0] \neq 0$；

（4）假设4（单调性，monotonicity）：对于所有个体，要么 $d^1 \geqslant d^0$ 成立，要么 $d^0 \geqslant d^1$ 成立。

若满足以上四个假设，则

$$\beta^{IV} = \frac{E[y \mid z=1] - E[y \mid z=0]}{E[d \mid z=1] - E[d \mid z=0]} = E[y^1 - y^0 \mid d^1 - d^0 = 1].$$ ①

当内生处理变量和工具变量都只有两个取值时，所有的人被分为始终接受者（always takers）、依从者（compliers）、从不接受者（never takers）和违背者（defiers）四个类型（Angrist et al.，1996）。以估计大学的教育回报率为例，假设 y 为个体的收入，处理变量 d 为个体是否上大学（$d=1$ 为上大学，$d=0$ 为未上大学），采用个体与大学的居住距离远近作为工具变量 z（$z=1$ 为距离近，$z=0$ 为距离远）。不失一般性，假定距离近（$z=1$）的个体更有激励上大学（$d=1$）。在此框架下，所有人可以被分为以下四个类型。第一，始终接受者。无论距离远近，始终接受者都会接受大学教育。第二，依从者。这类人群对工具变量的变化做出完全的反应：当大学距离近时，他们接受大学教育；当大学距离远时，他们不接受大学教育。第三，从不接受者。无论距离远近，这类人群都不会去上大学。第四，违背者。违背者与依从者完全相反：当大学距离近时，他们不愿意上大学；当大学距离远时，他们反而更愿意上大学。

在Imbens and Angrist（1994）的框架下，单调性假设要求依从者和违背者不能同时存在。当假设不存在违背者时，工具变量法的估计结果被解释为依从者的平均处理效应，即会因为距离远近而改变上大学决策的人群（表1中第（2）类）的平均处理效应。

表1 二元处理变量和二元工具变量下的四类群体

		d^0	d^1
（1）	始终接受者（always takers，*at*）	1	1
（2）	依从者（compliers，*c*）	0	1
（3）	从不接受者（never takers，*nt*）	0	0
（4）	违背者（defiers，*d*）	1	0

① 为简化公式，除个别必要情况外，本文在异质性处理效应下省略所有变量的下标 i。并且，LATE定理中的四个假设对所有个体都成立。

近年来，学界进一步讨论了 LATE 理论中的异质性问题，主要集中在对单调性假设的补充和拓展。首先，一些学者澄清了单调性假设的概念。Vytlacil(2002)、Heckman and Vytlacil(2005)等认为，由 Imbens and Angrist(1994)所定义的单调性假设，应该称其为一致性假设(uniformity assumption)更为合理，它本质上定义的是人们之间行为的一致性，而非意味着某个特定的个体对工具变量的反应必须是一个单调的函数。其次，拓展了单调性假设的适用范围。研究发现，当使用多个工具变量或内生变量为非二元变量时，单调性假设不是一个完备的假设。此时，以上四类群体的假设并不足以穷尽所有人的类型，无法完全刻画出所有人的理性选择。在下文中，我们将分别论述当使用多个工具变量或内生变量有多个取值时，工具变量方法的最新理论进展。

（二）多个工具变量的情况

1. 单调性假设的局限性

当有多个工具变量时，单调性假设限制了个体选择的异质性(Mogstad et al., 2021)。仍以上文研究大学的教育回报率为例，除了个体与大学的居住距离，引入大学学费作为第二个工具变量。此时，LATE 框架下的单调性假设要求所有人的偏好具有一致性，具体表现为，对于所有人而言，学费对个体是否去上大学的影响都应该大于距离对个体是否去上大学的影响，或者反过来成立。

为了更为直观地说明，引入一个随机效应模型(random utility model)。假设个体 i 选择 d 的间接效用函数为 $V_i(d,z)$，那么当且仅当 $V_i(1,z)\geqslant V_i(0,z)$ 时，个体选择 $d^z=1$，即：

$$d^z=\underset{d\in\{0,1\}}{\operatorname{argmax}}\,V_i(d,z)\equiv\begin{cases}1, & \text{if } V_i(z)\geqslant 0\\ 0, & \text{if } V_i(z)<0\end{cases},\tag{1}$$

其中，$V_i(z)\equiv V_i(1,z)-V_i(0,z)$ 为净效用函数。

在本例中，工具变量 $z=(z_1,z_2)$，其中，z_1 代表距离远近($z_1=0$ 为远距离，$z_1=1$ 为近距离)，z_2 代表学费高低($z_2=0$ 为高学费，$z_2=1$ 为低学费)。假设任何一个工具变量的取值越大，越能激励个体接受大学教育。图 2 分别绘制了个体 k 和个体 j 的两条无差异曲线，在无差异曲线上，个体上大学与不上大学所带来的效用没有差异。横轴代表距离，纵轴代表学费。由于两个工具变量都是二元变量，因此 $z=(z_1,z_2)$ 有四种可能取值：$A(0,0)$、$B(0,1)$、$C(1,0)$、$D(1,1)$。

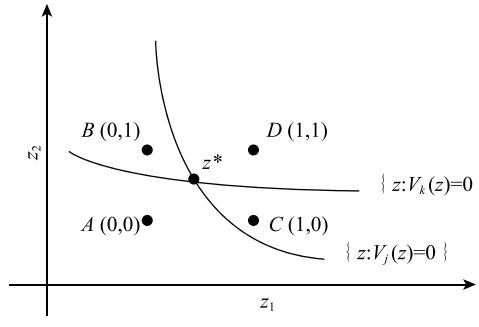

图 2　多个工具变量情况下的单调性假设

注：横轴箭头方向代表近距离，纵轴箭头方向代表低学费。

图2展示了单调性假设对个体选择所施加的限制。对于个体 j 而言，无论学费高低，他只会在与大学距离近的时候（$z_1=1$）接受大学教育；反之，对于个体 k 而言，无论距离远近，他只会在大学学费低的时候（$z_2=1$）接受大学教育。可以证明，这种情形不满足LATE理论下的单调性假设。单调性假设要求，所有人对低学费但远距离（对应点 B）和近距离但高学费（对应点 C）这两种组合的偏好是一致的，即单调性假设不允许以下情况存在：在考虑是否接受大学教育的决策时，有些人（例如个体 j）只在乎距离而不在乎学费，但另一些人（例如个体 k）只在乎学费而不在乎距离。因此，单调性假设对人们偏好的同质性施加了过强的约束，不能准确地刻画现实情况。

当工具变量连续且净间接效用函数可微时，单调性假设本质上限制了任意两个个体在上述无差异曲线的交点上的边际替代率相同。Mogstad et al.(2021)提出并证明了该性质：

性质1 假设 d^z 由式(1)决定。令 z^* 为集合 $\mathbb{Z}$ 内的一个点，$I(z^*)\equiv\{i\in I:V_i(z^*)=0\}$ 表示在 z^* 处对是否接受处理无差异的所有个体。假设对所有 $i\in I$，$V_i(z)$ 在 z^* 的邻域上都是连续可微的函数。那么，单调性假设意味着：

$$\partial_1 V_j(z^*)\partial_2 V_k(z^*)=\partial_1 V_k(z^*)\partial_2 V_j(z^*)，对所有 j,k\in I(z^*)，$$

其中，$\partial_l V_i(z)\equiv\partial V_i(z)/\partial z_l$，对于 $l=1,2$。

上述性质对偏好施加了很强的同质性限制，与现实不完全一致。例如，在上述教育回报率的例子中，假设个体 i 的净效用函数 $V_i(z)$ 由下式给出

$$V_i(z)=B_{i,0}+B_{i,1}z_1+z_2 \text{ 且 } d^z=\mathbf{1}[B_{i,0}+B_{i,1}z_1+z_2\geqslant 0]，$$

其中，$B_{i,1}$ 代表个体对距离相对学费的偏好程度。由性质1，单调性假设意味着 $B_{i,1}$ 无法随着个体 i 的变化而变化①。因此，对于所有人而言，学费对个体是否去上大学的影响都应该大于距离对个体是否去上大学的影响（$B_{i,1}\geqslant 1$），或者反过来（$B_{i,1}<1$）。

2. 部分单调性假设

从上文的分析可以看到，单调性假设下的同质性偏好在现实中很难得到满足。为了解决这一问题，Mogstad et al.(2021)提出了部分单调性假设，允许个体在多个工具变量时的异质性偏好。具体而言，部分单调性假设要求，在给定其他工具变量取值时，人们只需要对某一个工具变量的偏好具有一致性。为了说明这一假设，将工具变量 $z\in\mathbb{Z}$ 划分为第 l 个分量 z_l 和所有其他（$L-1$）个分量 z_{-l}，即 $z=(z_l,z_{-l})$。基于此设定，部分单调性假设（partial monotonicity，PM）的表述如下：任取 $l=1,\cdots,L$，使得 $z=(z_l,z_{-l})$ 和 $z'=(z'_l,z_{-l})$ 是 $\mathbb{Z}$ 中的两点。那么对于所有的个体 i，要么 $d^z\geqslant d^{z'}$ 成立，要么 $d^z\leqslant d^{z'}$ 成立。

相比于单调性假设，部分单调性假设是其“降维”后的更弱的假设。可以看到，当只有一个工具变量时，部分单调性假设等同于单调性假设。在多个工具变量的情况下，部分单调性则拓展了单调性假设的适用范围。单调性假设的局限性在于它需要对不同工具变量进行直接比较，例如，图2中 $B(0,1)$ 和 $C(1,0)$ 的比较。为了避免不同工具变量

① 这是由于 $B_{i,1}=\left(\frac{\partial V_i(z)}{\partial z_1}\right)/\left(\frac{\partial V_i(z)}{\partial z_2}\right)$。

之间比较的困难，部分单调性采用“降维”的思路，在给定其他工具变量取值的条件下，仅考虑某一工具变量取值的比较，例如，图 2 中 $A(0,0)$ 和 $B(0,1)$ 的比较。在估计大学的收益率的例子中，部分单调性假设只要求：对所有个体，当距离 z_1 取值相同的时候，学费低都使其更愿意上大学；或者对所有个体，当学费 z_2 取值相同的时候，距离近都使其更愿意上大学。因此，部分单调性假设允许不同个体存在异质性的偏好，更为合理。

那么，如何在部分单调性假设下解释工具变量法的估计结果呢？沿用前文符号，y^0、y^1 分别代表了 d 取值为 0 和 1 时 y 的潜在结果，可观测的结果 $y=dy^1+(1-d)y^0=y(d)$。类似地，可观测的处理变量可以写作 $d=\sum_{z\in\mathbb{Z}}\mathbf{1}[z=z]\cdot d^z=d(z)$，简单起见，考虑有两个工具变量 $z=(z_1,z_2)$，其中，$z_1\in\{0,1\}$，$z_2\in\{0,1\}$，其支撑集为 $\mathbb{Z}=\{0,1\}^2$。此时，工具变量一共有四个取值：$z_A=(0,0)$，$z_B=(0,1)$，$z_C=(1,0)$，$z_D=(1,1)$，分别对应图 2 中 A,B,C,D 四个点。

在部分单调性假设下，可以定义以下偏好：

$$d^{(0,0)}\leqslant d^{(0,1)}\leqslant d^{(1,1)}\quad 且\quad d^{(0,0)}\leqslant d^{(1,0)}\leqslant d^{(1,1)}. \tag{2}$$

在式(2)的偏好排序下，所有的人被分为以下六种类型(表 2)：

表 2　偏好排序式(2)下的六类群体

		z_A	z_B	z_C	z_D
		$d^{(0,0)}$	$d^{(0,1)}$	$d^{(1,0)}$	$d^{(1,1)}$
(1)	始终接受者(always takers,*at*)	1	1	1	1
(2)	急切依从者(eager compliers,*ec*)	0	1	1	1
(3)	勉强依从者(reluctant compliers,*rc*)	0	0	0	1
(4)	从不接受者(never takers,*nt*)	0	0	0	0
(5)	z_1 依从者(z_1 compliers,1*c*)	0	0	1	1
(6)	z_2 依从者(z_2 compliers,2*c*)	0	1	0	1

表 2 沿用并拓展了 Angrist et al.(1996)的术语。同表 1，始终接受者和从不接受者都不会因为任何工具变量的取值的变化而改变选择。无论两个工具变量取值是多少，始终接受者都会接受处理；反之，从不接受者都不会接受处理。此外，根据个体对两个工具变量的反应情况，依从者的概念被拓展为四类人群。第一，急切依从者。只要有任意一个工具变量的值从 0 变为 1，这类个体就会接受处理。第二，勉强依从者。只有两个工具变量的值都从 0 变为 1，这类人群才会接受处理。第三，z_1 依从者。无论 z_2 的值为多少，只要 z_1 的值从 0 变为 1，z_1 依从者便会接受处理。第四，z_2 依从者。类似地，无论 z_1 取值多少，z_2 依从者仅对 z_2 的取值变化做出反应。为了方便，本文将这四类会因一个或多个工具变量的取值变化而改变处理状态的个体统一称为“依从者们”。

观察可知，该偏好满足部分单调性假设，但不满足单调性假设。考虑工具变量 z 的取值从(0,1)变到(1,0)，此时，z_1 依从者会由不接受处理变为接受处理，z_2 依从者则会由接受处理变为不接受处理，而 z_1 依从者和 z_2 依从者这两类人同时存在，就违背了单调

性假设。而部分单调性假设对个体在(0,1)和(1,0)这两个点上的偏好不作限制，从而允许了偏好的异质性。需要提醒的是，在部分单调性假设下，由式(2)所定义的偏好并不是必要的，可以被放松。

在部分单调性假设下，工具变量法的估计结果可以被解释为“依从者们”的加权平均处理效应。设个体 i 所在组别为 G_i，$G_i \in \{at, ec, rc, nt, 1c, 2c\}$ 为表2所定义的六个组别之一。定义 $\pi_g \equiv \Pr[G_i = g]$ 和 $\Delta_g \equiv E[y^1 - y^0 \mid G_i = g]$，分别代表 g 组在总体中所占比例及其平均处理效应。Mogstad et al.(2021)提出且证明了以下性质：

性质2 假设工具变量 z 的支撑集为 $Z = \{0,1\}^2$，而且在部分单调性假设下，偏好由式(2)所定义。如果满足独立性假设①，且IV的估计结果存在，则：

$$\beta^{2sls} = \sum_{g \in \{1c, 2c, ec, rc\}} \omega_g \Delta_g,$$

其中，权重 ω_g 的总和为1②。ω_{ec} 和 ω_{rc} 总是非负的。如果 $\pi_{1c} \geqslant \pi_{2c}$，那么 ω_{1c} 为非负，而 ω_{2c} 的正负性由下式决定：

$$\mathrm{sgn}(\omega_{2c}) = \mathbf{1}[\pi_{2c} > 0] \times \mathrm{sgn}(\Pr[d = 1 \mid z_2 = 1] - \Pr[d = 1 \mid z_2 = 0]);$$

反之，如果 $\pi_{1c} < \pi_{2c}$，那么 ω_{2c} 为非负，而 ω_{1c} 的正负性则由下式决定：

$$\mathrm{sgn}(\omega_{1c}) = \mathbf{1}[\pi_{1c} > 0] \times \mathrm{sgn}(\Pr[d = 1 \mid z_1 = 1] - \Pr[d = 1 \mid z_1 = 0])。$$

性质2表明，在部分单调性假设下，IV的估计结果是四组“依从者们”的平均处理效应的线性组合。值得注意的是，我们无法保证这些权重的取值范围在0到1之间。当存在负的权重(negative weights)时，IV的估计结果便不能被解释为一个正的加权平均处理效应。具体而言，勉强依从者和急切依从者的权重总是非负的，但 z_1 依从者和 z_2 依从者的权重却不一定为正。如果 z_1 依从者的比例大于 z_2 依从者，那么 z_1 依从者的权重一定非负，但 z_2 依从者的权重可能是正的，也可能是负的。

这背后的直觉是，当工具变量 z 的取值从(0,1)变到(1,0)时，z_1 依从者会更多地接受处理，而 z_2 依从者则会退出处理，表现得像 z_1 的“违背者”。如果 z_1 依从者的比例大于 z_2 依从者，工具变量 z 的这一变化仍会使得更多个体接受处理。此时，由于 z_2 的值由1变为0，z_2 依从者则会退出处理，表现为工具变量 z_1 的“违背者”，因此赋予负的权重。然而，赋予 z_2 依从者的最终权重是正是负，还取决于其作为工具变量 z_1 的“违背者”时对处理变量产生的负效应，是否能被其作为工具变量 z_2 的依从者时发挥的正效应所抵消。这种正效应指的是，工具变量从(0,0)变为(0,1)，以及从(1,0)变为(1,1)时，z_2 依从者对处理变量产生的正向影响。

以上分析表明，在多个工具变量下传统的单调性假设限制了个体选择的异质性，因此，Mogstad et al.(2021)提出一个“降维”的部分单调性假设，从而允许了偏好的异质性。此外，他们还证明了在所有“依从者们”权重为正的条件下，两阶段最小二乘的结果仍然可以被解释为“依从者们”的局部平均处理效应的正加权平均(positively weighted average of local average treatment effects)。但是，需要注意的是，当存在负的权重时，这

① 此时，独立性假设为 $(y^1, y^0, d^z{}_{z \in Z}) \perp z$。

② ω_g 的具体表达式详见Mogstad et al.(2021)的附录部分。

一最新理论发展仍然无法在 LATE 框架下给予 IV 估计量一个合理的解释。

（三）内生变量有多个取值的情况

当内生变量是多元时，传统的 LATE 理论对偏好结构和选择结构的刻画也存在局限性，因而影响其对 IV 估计结果的解释力（Heckman et.al，2006；Heckman et.al，2008）。接下来将介绍在内生处理变量是多元的情况下，工具变量法的最新理论进展。

在多元选择模型（multiple choice models）中，个体面临的不再是“非此即彼”的二元选择，此时单调性假设并非一个完备的假设，无法刻画出所有依从者的选择路径。在二元选择模型中，工具变量的变化所引致的依从者的选择路径是单一的：当工具变量 z 从 0 变为 1 时，依从者的选择变量 d 将相应地由 0 变为 1。但当选择变量是多元变量时，例如 $d\in\{0,1,2\}$，此时依从者的选择路径很有可能不是单一的。当工具变量 z 从 0 变为 1 时，单调性假设只能保证依从者的选择变量 d 变为 1，却无法刻画出在工具变量 $z=0$ 时依从者选择 0 还是 2。当不同的引致路径存在时，就不能简单地将工具变量法的估计结果解释为个体的某一选择相较于另一选择的处理效应。

为了说明这一问题，首先将问题简化为内生选择变量有三个取值，工具变量有两个取值的情况。假设个体在经济学（E）、法学（L）、数学（M）三个专业中进行选择，重点关注个体选择经济学专业相对于数学专业的教育回报率。此时，y 为可观测到的收入，y^E、y^L、y^M 分别表示个体选择以上三个专业的潜在收入。工具变量 $z_E\in\{0,1\}$ 为个体是否被随机分配到经济学专业，潜在选择结果 $d^{z_E}\in\{M,E,L\}$ 表示在工具变量 z_E 下个体选择的专业类型，可观测的选择变量 $d=\mathbf{1}[z_E=0]\cdot d^0+\mathbf{1}[z_E=1]\cdot d^1=d(z_E)$，根据 z_E 和 d^{z_E} 的不同取值，所有的人被分为以下九个类型（表 3）：

表 3　多元内生变量下的九类群体

		d^0	d^1
(1)	M 类从不接受者（M-type never takers）	M	M
(2)	M 类依从者（M-type compliers）	M	E
(3)	始终接受者（always takers）	E	E
(4)	L 类依从者（L-type compliers）	L	E
(5)	L 类从不接受者（L-type never takers）	L	L
(6)	不满足单调性	M	L
(7)	不满足单调性	E	M
(8)	不满足单调性	E	L
(9)	不满足单调性	L	M

依据 LATE 理论，可以将单调性拓展为：

$$d^1\neq d^0\Rightarrow d^1=E. \tag{3}$$

式(3)意味着，如果个体由于 z_E 的变化（收到经济学专业的录取通知）而改变了决策，那么这个决策必然是选择了经济学专业（$d^1=E$）。或者说，z_E 从 0 变到 1，不可能促

使个体去选择其他专业(法学、数学)。但该单调性的假设并不排除始终愿意选择法学、数学的这类人群的存在。根据单调性假设的限制,可以排除表3中的(6)、(7)、(8)、(9)类人群。

表3依然沿用了前文的术语习惯。始终接受者和从不接受者仍然不会因为工具变量 z_E 取值的变化而改变选择。不同的是,在多元选择模型下,从不接受者有两种类型:M类从不接受者(M-type never takers)和L类从不接受者(L-type never takers),他们分别代表了数学专业和法律专业的狂热者,无论有没有收到经济学专业录取通知,他们都将坚定地选择数学专业或法律专业。根据不同的引致路径,依从者也有两类:M类依从者(M-type compliers)和L类依从者(L-type compliers)。M类依从者和L类依从者分别以数学、法律作为其次优选择的专业(next-best alternatives),当未收到经济学录取通知($z_E=0$)时,M类依从者会选择数学专业,L类依从者会选择法律专业,当 z_E 的值从0转变为1时,他们都将从各自的次优选择转为选择经济学专业。同样地,将这两类拥有不同次优选择、但都会因为工具变量 z_E 从0变到1而选择经济学专业的个体称为"依从者们"。

在多元选择模型下,工具变量法的估计结果是各类"依从者们"的局部平均处理效应的加权平均(Kirkeboen et al., 2016;Kline and Walters,2016)。如式(4)所示,基于拓展的单调性假设式(3),工具变量法得到的估计结果是混合"依从者们"选择经济学专业的平均效应:

$$\frac{E[y|z_E=1]-E[y|z_E=0]}{E[\mathbf{1}\{d=E\}|z_E=1]-E[\mathbf{1}\{d=E\}|z_E=0]}$$
$$=E[y^E-y^{d^0}|d^1=E,d^0\neq E]\equiv LATE_E. \tag{4}$$

进一步地,可以将这一平均处理效应分解为各类型"依从者们"的子平均处理效应(subLATEs)的加权平均:

$$LATE_E=S_L LATE_{LE}+(1-S_L)LATE_{ME},$$

其中,$LATE_{aE}\equiv E[y^E-y^a|d^1=E,d^0=a]$ 为 a 类依从者平均处理效应,$a\in\{L,M\}$。权重 S_L 为L类依从者在两类依从者们中的占比:$S_L\equiv\frac{P(d^1=E,d^0=L)}{P(d^1=E,d^0\neq E)}$。

在多元选择模型下,工具变量法所得到的是各个子平均处理效应的加权平均。此时,一开始想要估计的 $LATE_E$ 不再是选择经济学专业相对于选择数学专业的教育回报率,而是由工具变量 z_E 的变化所引致的两类依从者们选择经济学专业的加权平均处理效应。通常情况下,若没有额外的假设,工具变量法并不能识别出这些"依从者们"各自的平均处理效应,也就无法识别出某类个体或群体选择一个专业相对于选择另一个专业的收入回报。

Kirkeboen et al.(2016)巧妙利用挪威的大学录取数据解决了这一问题。在挪威,高考数据详细记录了个人分数、志愿学校和志愿专业排序、录取情况和各大学专业的录取分数线。一方面,详尽的大学录取数据为每个专业选择变量提供一个相应的工具变量(个人成绩是否过该专业录取线),极大程度地减少了内生性问题;另一方面,个人志愿学校和专业的偏好排序提供了关键的信息:个体的次优选择。在此特殊的数据结构下,他们

估计出了选择不同专业相对于选择次优专业的劳动力市场回报。

具体地，考虑一个更为一般的大学专业选择情境。假设每个人可以在三个专业中进行选择，$d\in\{0,1,2\}$，个体收入由下式决定：

$$y=\beta_0+\beta_1 d_1+\beta_2 d_2+\varepsilon,$$

其中，y 为可观测到的收入；处理变量 $d_j\in\{0,1\}$ 表示个体是否选择 j 专业，$j\in\{0,1,2\}$；工具变量为个体被随机分配到的专业 $z\in\{0,1,2\}$，具体地，设 $z_j\in\{0,1\}$ 为个体是否被随机分配到 j 专业，$j\in\{0,1,2\}$。①

Kirkeboen et al.(2016)证明，基于不相关性假设(irrelevance)和次优选择(next-best alternatives)的信息，可以识别多元选择模型中的某类个体或群体的某个选择相对于另一个选择的平均处理效应。性质 3 更为具体地概括了这一过程：

性质 3　假设 LATE 定理中的假设(1)至假设(4)都得到满足，且 $d_1^1=d_1^0=0\Rightarrow d_2^1=d_2^0$，$d_2^2=d_2^0=0\Rightarrow d_1^2=d_1^0$，且条件于 $d_1^0=d_2^0=0$，那么

$$\beta_1=E[y^1-y^0|d_1^1-d_1^0=1,d_2^0=0],$$

$$\beta_2=E[y^2-y^0|d_2^2-d_2^0=1,d_1^0=0],$$

其中，$d_j^z\in\{0,1\}$ 代表当个体被随机分配到 z 专业时，他/她是否会选择 j 专业。在性质 3 中，不相关性假设指的是，如果工具变量 z 从 0 变为 1(2)不会使得个体选择专业 1(专业 2)，那么这一变化也不可能引致个体去选择专业 2(专业 1)。给定专业 0 为次优选择，β_1 测度了选择专业 1 以替代专业 0 的平均收入回报，β_2 测度了选择专业 2 以替代专业 0 的平均收入回报。通过比较 β_1 和 β_2 的大小，可以得到，对于以专业 0 为次优选择的个体，他们选择专业 2 而非专业 0 的收益是否大于选择专业 1 而非专业 0 的收益，这一方法有利于比较大学各个专业的收入回报。

直觉上，这一做法也采用了“降维”的思路。在多元选择模型中，IV 估计的困难在于无法明确“依从者们”是从哪个次优选择的状态引致而来，从而得到了一个“杂糅”的处理效应。然而，一旦给定了偏好排序(ranking)，就能将“依从者们”限制为从某一特定次优选择引致而来的某一类依从者，此时，IV 估计就又“降维”为一个二元选择的问题。

但是，大多时候，我们缺少关于次优选择的完整信息。Kline and Walters(2016)提出可以利用模型刻画人们的偏好选择，进行结构方程估计。评估美国学前早教 Head Start 项目时，由于竞争项目(competing preschool programs)的存在，工具变量法估计的局部平均处理效应是两类依从者的子处理效应的加权平均。为了解决潜在竞争项目所带来的替代效应的影响，Kline and Walters(2016)基于结构估计的方式，通过建立选择模型(selection model)，刻画个人的偏好和选择过程，估计出不同类型的依从者的处理效应，以此来评估美国的学前早教 Head Start 项目的成本收益。他们也强调了在评估公共项目时，必须考虑到替代品的存在对政策评估结果的影响。

总结来说，在多元选择模型中，由于工具变量变化所导致的选择路径不再是单一的，因此 LATE 框架下的单调性假设是不完备的。此时，IV 的估计结果是几类拥有不同的次优选择的“依从者们”的平均处理效应的加权平均。借助次优选择的信息，可以估计出

① 在实际应用中，Kirkeboen et al.(2016) 采用个人成绩是否过该专业录取线(cutoff) 作为工具变量，在不可预测的专业录取线附近，个体是否过该专业录取线可被视为一种随机分配的过程。

每类依从者的平均处理效应。若没有次优选择的信息，还可以通过模型刻画个人选择偏好，利用结构方程估计出每类依从者的平均处理效应。

上述理论拓展了LATE框架的适用范围，然而LATE框架仍然存在无法估计非依从者的平均处理效应的局限性。以下介绍边际处理效应的最新发展，这一理论框架可以为上述问题提供解决方案。

三、边际处理效应(MTE)最新发展

在LATE理论发展的同时，有一支与其紧密相关但又相对独立的文献也在迅速发展，即MTE理论。Heckman and Vytlacil(1999，2005，2007)、Heckman et al.(2006)、Carneiro et al.(2011)在Björklund and Moffitt(1987)的基础上发展了MTE的概念，这支文献的目标是估计处理效应的整体分布，在研究中的运用愈加广泛(Nybom，2017；Black et al.，2024；Borghesan and Vasey，2024)。

本部分对MTE的最新进展进行详细介绍。首先，基于选择模型(selection model)介绍MTE的具体概念，并概述选择模型与LATE模型的等价关系；其次，阐述如何基于选择模型和MTE将依从者的平均处理效应外推得到始终接受者和从不接受者的平均处理效应；最后，梳理工具变量取值连续和离散时MTE的估计方法。

(一) 选择模型与边际处理效应

接下来参考Brinch et al.(2017)对MTE的概念进行具体介绍。考虑广义罗伊模型(generalized Roy model)这一最为常见的选择模型，d 表示是否接受处理($d=1$ 表示接受处理，$d=0$ 表示不接受处理)，接受处理的潜在结果为 y^1，不接受处理的潜在结果为 y^0，潜在结果由可观测变量 X 和随机项决定，潜在结果方程为：

$$y^j=\mu^j(X)+U^j,\ j=0,1,$$

其中，U^j 为随机项，$\mu^j(\cdot)$ 为任意函数，假设 $E(U^j\mid X)=0$。实际观测结果 y 为：

$$y=dy^1+(1-d)y^0.$$

是否接受处理由可观测变量 X、工具变量 z 和不可观测的个体异质性 U_D 决定，U_D 为一个连续的随机变量，选择方程为：

$$I_D=\mu_D(Z)-U_D,$$

其中，$Z=(X,z)$，$\mu_D(\cdot)$ 为任意函数。当且仅当 $I_D>0$ 时，$d=1$。U_D 可以被标准化为服从 (0,1) 均匀分布，即 $U_D\sim U(0,1)$。[①]此时，当 Z 被给定时，接受处理的概率 $p(Z)\equiv \mathrm{Prob}(d=1\mid Z)=\mathrm{Prob}(I_D>0\mid Z)=\mathrm{Prob}(U_D<\mu_D(Z)\mid Z)=\mu_D(Z)$，因此 $\mu_D(Z)$ 可

① 假设 U_D 服从任意严格递增的分布 F_{U_D}，则可以通过如下数学变形将 U_D 标准化为服从 (0,1) 均匀分布：定义 $\tilde{\mu}_D(Z)=F_{U_D}(\mu_D(Z))$ 和 $\tilde{U}_D=F_{U_D}(U_D)$，则 $d=\mathbf{1}(\mu_D(Z)-U_D>0)=\mathbf{1}\{F_{U_D}(\mu_D(Z))>F_{U_D}(U_D)\}=\mathbf{1}(\tilde{\mu}_D(Z)-\tilde{U}_D>0)$。$\tilde{U}_D$ 的分布函数 $F_{\tilde{U}_D}(u)=\mathrm{Prob}(F_{U_D}(U_D)\leqslant u)=\mathrm{Prob}(U_D\leqslant F_{U_D}{}^{-1}(u))=F_{U_D}(F_{U_D}{}^{-1}(u))=u$，即 $\tilde{U}_D$ 服从 (0,1) 均匀分布。

以被称作倾向分数(propensity score)。①

假设给定 X 时，(U^0,U^1,U_D) 和 Z 相互独立，则工具变量 z 只会通过影响接受处理的概率 p 影响期望结果。此时 MTE 被定义为接受处理 $(d=1)$ 和不接受处理 $(d=0)$ 无差异(即 $I_D=0$，也即 $U_D=\mu_D(Z)=p$)的样本的平均处理效应：

$$\begin{aligned}MTE(x,p)&=E(y^1-y^0|X=x,U_D=p)\\&=\mu^1(x)-\mu^0(x)+E(U^1|X=x,U_D=p)-E(U^0|X=x,U_D=p)\\&\equiv\mu^1(x)-\mu^0(x)+k^1(x,p)-k^0(x,p)\\&\equiv\mu^1(x)-\mu^0(x)+k(x,p),\end{aligned}$$

其中，$k^j(x,p)\equiv E(U^j|X=x,U_D=p)$，$j=0,1$；$k(x,p)\equiv k^1(x,p)-k^0(x,p)$。

假设工具变量 z 为二元变量，$p^0(x)=\text{Prob}(d=1|X=x,z=0)$，$p^1(x)=\text{Prob}(d=1|X=x,z=1)$，则结合 LATE 和 MTE 的定义可以推导出：

$$\begin{aligned}LATE(x)&=\frac{E(y|z=1,X=x)-E(y|z=0,X=x)}{E(d|z=1,X=x)-E(d|z=0,X=x)}\\&=\frac{1}{p^1(x)-p^0(x)}\int_{p^0(x)}^{p^1(x)}MTE(x,p)dp,\end{aligned}$$

即 LATE 是对工具变量做出反应的依从者的 MTE 的均值。可见，MTE 的优点在于独立于工具变量的选取，并且能够将不同工具变量得到的不同 LATE 用一个统一的框架进行解释(Heckman and Vytlacil，2005)。

(二) 选择模型与 LATE 模型的等价关系

前文已经推导出由 LATE 框架得到的 LATE 估计量与由选择模型得到的 MTE 估计量之间的关系。本小节则主要介绍 LATE 框架和选择模型之间的等价关系。

由于处理的分配机制(treatment assignment mechanism)是理解选择偏误的关键，因此选择模型的思路是把选择是否接受处理的过程用潜在指数模型(latent index model)刻画出来。在这个过程中，需要对个体不可观测异质性 U_D 的分布进行假设。而 LATE 框架下只要求满足独立性、排他性、相关性、单调性四个假设，不进行任何分布和方程形式的假设。因此，选择模型看上去比 LATE 框架有更多的限制和约束。但是，实际上 LATE 框架和选择模型是等价的(Vytlacil，2002；Kline and Walters，2019)。接下来对上述选择模型进行简化。假设工具变量 z 为二元变量，X 只包含常数项，μ_D 是 z 的一次函数，即 $y^j=\mu^j+U^j$，$\mu_D(Z)=\varphi_0+\varphi_1 z$，并在该简化的设定下证明 LATE 框架和选择模型的等价关系。

要证明 LATE 框架和选择模型的等价性，首先需要证明选择模型满足 LATE 框架的四个假设。由于 $y^1=\mu^1+U^1$，$y^0=\mu^0+U^0$，$d^1=\mathbf{1}[\varphi_0+\varphi_1-U_D>0]$，$d^0=\mathbf{1}[\varphi_0-U_D>0]$ 且 (U^0,U^1,U_D) 和 z 相互独立，因此独立性假设和排他性假设满足。

① 以上讨论较为抽象，接下来仍以估计大学的教育回报率为例对模型进行具体说明。d 表示个体是否上大学($d=1$ 表示上大学，$d=0$ 表示未上大学)，y^1、y^0 分别表示个体上大学和未上大学的潜在收入。U^j 表示影响个体潜在收入的不可观测的异质性，U_D 表示影响个体选择是否上大学的不可观测的异质性，即个体对上大学的偏好。如果个体在选择是否上大学时将潜在收入因素考虑在内，则 U^j 和 U_D 之间有相关性，此时存在内生性问题。

假设 $\varphi_1 > 0$，则对于所有个体，$d^1 \geqslant d^0$ 成立，因此单调性条件满足。此时，$\text{Prob}(d^1 > d^0) = \text{Prob}(\varphi_0 + \varphi_1 > U_D \geqslant \varphi_0) > 0$，则 $E[d^1 - d^0] \neq 0$，因此相关性条件满足。

要证明 LATE 框架和选择模型的等价性，还需要证明，只要满足 LATE 框架，就能构建一个与之对应的选择模型。首先需要说明的是，对 U_D 的分布进行参数假设本质上并不是一种约束，因为 U_D 可以服从任意分布。假设 U_D 的分布函数为 F，对于任何严格递增的分布函数 G，

$$\begin{aligned} d &= \mathbf{1}(\varphi_0 + \varphi_1 z > U_D) = \mathbf{1}\{F(\varphi_0 + \varphi_1 z) > F(U_D)\} \\ &= \mathbf{1}\{G^{-1}(F(\varphi_0 + \varphi_1 z)) > G^{-1}(F(U_D))\} = \mathbf{1}(\tilde{\varphi}_0 + \tilde{\varphi}_1 z > \tilde{U}_D), \end{aligned}$$

其中，G^{-1} 表示 G 的反函数，$\tilde{\varphi}_0 = G^{-1}(F(\varphi_0))$，$\tilde{\varphi}_1 = G^{-1}(F(\varphi_0 + \varphi_1)) - G^{-1}(F(\varphi_0))$，$\tilde{U}_D = G^{-1}(F(U_D))$。经过上述变形之后，个体异质性 $\tilde{U}_D$ 的分布函数变为 G，但是该选择模型并未发生本质变化。从上述变形可以看出，在同一个选择模型中，个体异质性服从的分布可以是任意的。因此，想要证明在满足 LATE 框架时能构建一个与之对应的选择模型，只需要证明 LATE 模型可以被一个 U_D 服从某个特殊分布的选择模型表示即可。

接下来证明 LATE 模型可以被一个特殊的选择模型表示。假设 U_D 服从以下分布：

$$U_D = \begin{cases} u \times p^0, & d^0 = 1 \\ p^0 + u \times (p^1 - p^0), & d^1 > d^0, \\ p^1 + u \times (1 - p^1), & d^1 = 0 \end{cases}$$

其中，$u \sim U(0,1)$ 且 u 和 z 相互独立。这意味着始终接受者的 U_D 服从 $(0, p^0)$ 均匀分布，依从者的 U_D 服从 (p^0, p^1) 均匀分布，从不接受者的 U_D 服从 $(p^1, 1)$ 均匀分布。此时，可以基于这一特殊分布构建一个特殊的潜在指数模型 $d = \mathbf{1}[p^0 + (p^1 - p^0)z > U_D]$，使之能够表示和 LATE 模型相同的 d 的潜在结果，从而证明了 LATE 框架和选择模型的等价性。Vytlacil(2002)、Kline and Walters(2019)对更加一般的情形下 LATE 模型和选择模型的等价关系进行了详细阐述。

（三）LATE 的外推

MTE 还可以帮助我们将 LATE 外推得到 ATE。对于始终接受者(always takers, at)，$U_D < p^0(x)$，则 $E(U_D \mid at) = \dfrac{p^0(x)}{2}$；对于依从者(compliers, c)，$p^0(x) \leqslant U_D < p^1(x)$，则 $E(U_D \mid c) = \dfrac{p^0(x) + p^1(x)}{2}$；对于从不接受者(never takers, nt)，$U_D \geqslant p^1(x)$，则 $E(U_D \mid nt) = \dfrac{p^1(x) + 1}{2}$。根据下面的表达式：

$$E(y^1 \mid at) = E(y \mid d = 1, z = 0),$$

$$E(y^0 \mid nt) = E(y \mid d = 0, z = 1),$$

$$\begin{aligned} E(y \mid d = 1, z = 1) = {} & \frac{\text{Prob}(d = 1 \mid z = 0)}{\text{Prob}(d = 1 \mid z = 1)} E(y^1 \mid at) + \\ & \frac{\text{Prob}(d = 1 \mid z = 1) - \text{Prob}(d = 1 \mid z = 0)}{\text{Prob}(d = 1 \mid z = 1)} E(y^1 \mid c), \end{aligned}$$

$$E(y|d=0,z=0)=\frac{\text{Prob}(d=0|z=1)}{\text{Prob}(d=0|z=0)}E(y^0|nt)+\frac{\text{Prob}(d=0|z=0)-\text{Prob}(d=0|z=1)}{\text{Prob}(d=0|z=0)}E(y^0|c),$$

可以求出 $E(y^1|at)$、$E(y^0|nt)$、$E(y^1|c)$和 $E(y^0|c)$。这样只能得到依从者的平均处理效应，即 $E(y^1|c)-E(y^0|c)$，而不能得到始终接受者和从不接受者的平均处理效应。此时如果想由 LATE 外推得到 ATE，就需要对 $E(y^j|U_D)$的方程形式进行假设（Kline and Walters，2019）。图 3 可以为理解外推提供直觉，比如假设 $E(y^j|U_D)$为U_D的线性函数，则如图 3(a)所示，此时 A 点和 B 点都在 $E(y^1|U_D)$线上，C 点和 D 点都在 $E(y^0|U_D)$线上，可以通过将 AB 外推至 E 点得到 $E(y^1|nt)$，通过将 CD 外推至 F 点得到 $E(y^0|at)$，从而可以得到从不接受者和始终接受者的平均处理效应。当假设 $E(y^j|U_D)$为U_D的非线性函数时，如图 3(b)所示，根据詹森不等式（Jensen's inequality），A 点、B 点、C 点、D 点不一定在 $E(y^j|U_D)$线上，这区别于图 3(a)中的线性情况。但是仍然可以利用 $E(y^1|at)$和 $E(y^1|c)$求出 $E(y^1|U_D)$的表达式，利用 $E(y^0|c)$和 $E(y^0|nt)$求出 $E(y^0|U_D)$的表达式，并进而求出 $E(y^1|nt)$和 $E(y^0|at)$。更一般地，$E(y^1|U_D)$线和 $E(y^0|U_D)$线之间的距离即为 MTE，Heckman and Vytlacil(2005)指出，平均处理效应（ATE）、处理组的平均处理效应（ATT）、未处理组的平均处理效应（TUT）等都可以表示成 MTE 的加权平均数，因此 MTE 还可以帮助我们推算这些处理效应。

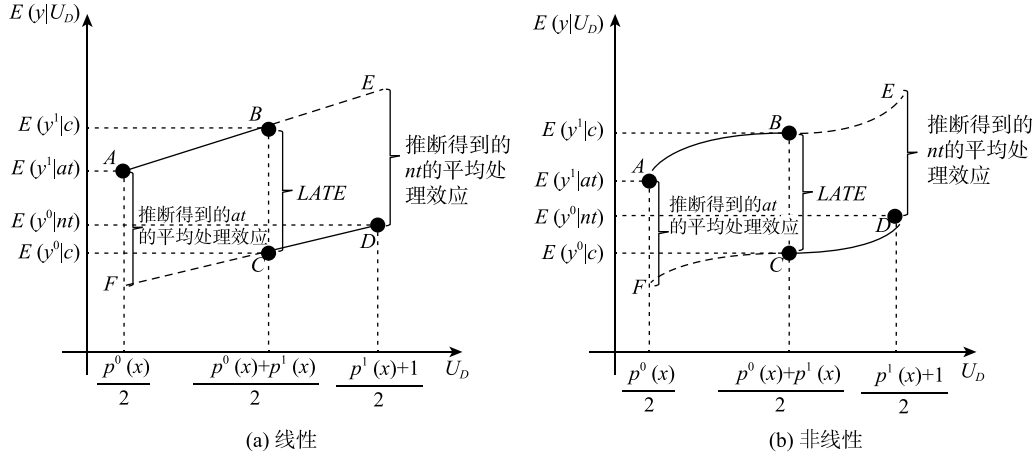

图 3　LATE 的外推

（四）MTE 的估计方法

以上主要在工具变量 z 为二元变量的情况下讨论 MTE 和 LATE 之间的关系，接下来则在更加一般的情形下介绍如何估计 MTE。

1. 工具变量取值连续时

当工具变量为连续变量时，估计 MTE 主要有两种方法。第一种是 Heckman and Vytlacil(1999，2005)、Heckman(2010)使用的局部工具变量法（local instrumental varia-

ble, LIV)。给定 X 和 $p(Z)$ 时，可以证明得到①：

$$MTE\ (x,p)=\frac{\partial E(y\mid P(Z)=p,X=x)}{\partial p},$$

因此 MTE 可以通过求 $E(y\mid P(Z)=p,X=x)$ 关于 p 的导数得到。

第二种是 Heckman and Vytlacil(2007)使用的分离估计法(separate estimation)。给定 X 和 $p(Z)$时，可以证明得到②：

$$MTE(x,p)=p\ \frac{\partial E(y^1\mid P(Z)=p,X=x,U_D\leqslant p)}{\partial p}+E(y^1\mid P(Z)=p,X=x,U_D\leqslant p)+$$

$$(1-p)\frac{\partial E(y^0\mid P(Z)=p,X=x,U_D>p)}{\partial p}-E(y^0\mid P(Z)=p,X=x,U_D>p),$$

因此 MTE 可以通过求 $E(y^1\mid P(Z)=p,X=x,U_D\leqslant p)$、$E(y^0\mid P(Z)=p,X=x,U_D>p)$以及二者各自关于 p 的导数得到。

当使用的工具变量使 p 在 $[0,1]$ 有连续取值时，局部工具变量法和分离估计法都可以通过非参数方法估计出 MTE，两者并无重大区别。

2. 工具变量取值离散时

上面讨论的主要是工具变量为连续取值的情形，但在实际应用中，很多工具变量都是离散的。比如用抽签结果作为是否服兵役的工具变量(Angrist, 1990)，用出生季度作为受教育年限的工具变量(Angrist and Krueger, 1991)，用第一个孩子和第二个孩子性别是否相同作为是否生第三个孩子的工具变量(Black et al., 2005)等。此时不能使用非参数方法估计 MTE，而需要进行额外的假设。主要有以下两种思路：第一种是在 X 和 U_D 不可分离时对 k 的方程形式进行额外假设(French and Song, 2014; Moffitt, 2008)；第二种是进一步假设 $E(y^j\mid X=x,U_D)=\mu^j(x)+E(U^j\mid U_D)$，此时 MTE 关于 X 和 U_D 加性可分(additively separable)(Brinch et al., 2017; Carneiro and Lee, 2009)。

(1) X 和 U_D 不可分离时

在沿用“给定 X 时，(U^0,U^1,U_D)和 Z 相互独立”的假设的基础上，对 k 的方程形式进行假设。首先来看一个特殊的例子，给定 X，假设 k^j 为 p 的一次函数且工具变量为二元变量③，则有：

$$E(U^0\mid X=x,U_D=p)=k^0(x,p)=\alpha^0(x)p-\frac{1}{2}\alpha^0(x),$$

$$E(U^1\mid X=x,U_D=p)=k^1(x,p)=\alpha^1(x)p-\frac{1}{2}\alpha^1(x).$$

可以计算得到④：

$$E(y^0\mid P(Z)=p,X=x,U_D>p)=\mu^0(x)+\frac{1}{2}\alpha^0(x)p, \tag{5}$$

① 证明请见附录Ⅰ(一)。限于篇幅，附录未在正文列示，感兴趣的读者可在《经济学》(季刊)官网(https://ceq.ccer.pku.edu.cn)下载。

② 证明请见附录Ⅰ(二)。

③ k^0和k^1一次函数的常数项是根据 $E(U^0\mid X=x)=0$ 和 $E(U^1\mid X=x)=0$ 得到的。

④ 笔者发现 Brinch et al.(2017)一文中第 994 页第 5 行 MTE 的表达式有印刷错误，经与作者确认，本文在此进行订正处理。

$$E(y^1|P(Z)=p,X=x,U_D\leqslant p)=\mu^1(x)+\frac{1}{2}\alpha^1(x)(p-1),\tag{6}$$

$$E(y|P(Z)=p,X=x)=\mu^0(x)+(\mu^1(x)-\mu^0(x))p+\frac{1}{2}(\alpha^1(x)-\alpha^0(x))p(p-1),\tag{7}$$

$$MTE(x,p)=\mu^1(x)-\mu^0(x)-\frac{1}{2}(\alpha^1(x)-\alpha^0(x))+(\alpha^1(x)-\alpha^0(x))p.\tag{8}$$

如果使用局部工具变量法，则是对式(7)进行估计。由于工具变量为二元变量，只能得到$(p^0,E(y|P(Z)=p^0,X=x))$和$(p^1,E(y|P(Z)=p^1,X=x))$两个点，而$E(y|P(Z)=p,X=x)$是关于p的二次函数，因此无法估计出$E(y|P(Z)=p,X=x)$，更无法进一步估计出 MTE。

如果使用分离估计法，则是对式(5)和式(6)进行估计。由于$E(y^0|P(Z)=p,X=x,U_D>p)$和$E(y^1|P(Z)=p,X=x,U_D\leqslant p)$都是关于$p$的一次函数，因此可以利用$(p^0,E(y^0|P(Z)=p^0,X=x,U_D>p^0))$和$(p^1,E(y^0|P(Z)=p^1,X=x,U_D>p^1))$两个点估计出$E(y^0|P(Z)=p,X=x,U_D>p)$；利用$(p^0,E(y^1|P(Z)=p^0,X=x,U_D\leqslant p^0))$和$(p^1,E(y^1|P(Z)=p^1,X=x,U_D\leqslant p^1))$两个点估计出$E(y^1|P(Z)=p,X=x,U_D\leqslant p)$，从而估计出 MTE。

更一般地，当k^j为L次函数且p有N个取值时，$E(y|P(Z)=p,X=x)$是关于p的$L+1$次函数，$E(y^0|P(Z)=p,X=x,U_D>p)$和$E(y^1|P(Z)=p,X=x,U_D\leqslant p)$是关于$p$的$L$次函数。上述三个函数已知点的个数都为$N$，因此当且仅当$N\geqslant L+2$时，可以用局部工具变量法估计出 MTE；当且仅当$N\geqslant L+1$时，可以用分离估计法估计出 MTE(Brinch et al., 2017)。①

(2) X和U_D可分离时

根据前文可知，当N较小时，k^j的函数形式受到较大约束。比如当工具变量为二元变量时，如果k^j为p的二次函数、三次函数或更高次函数，局部工具变量法和分离估计法都无法估计出 MTE。为了放松对k的函数形式的约束，可以进一步假设$E(y^j|X=x,U_D)=\mu^j(x)+E(U^j|U_D)$，此时$E(y^0|P(Z)=p,X=x,U_D>p)$、$E(y^1|P(Z)=p,X=x,U_D\leqslant p)$、$E(y|P(Z)=p,X=x)$、$MTE(x,p)$都关于$X$和$U_D$加性可分。

此时，当X取不同值时，k^j的函数形式保持不变，比如在上述k^j为p的一次函数的例子中α^0、α^1的取值与x不再相关。可以通过增加X的取值，在保持k^j的未知参数个数不变的情况下增加已知方程的数量，从而可以估计出k^j中更多的未知参数，因而可以估计出形态更为丰富的$MTE(x,p)$(Brinch et al., 2017)。

(五) 工具变量有效性检验

以上，本文介绍了异质性处理效应下 LATE 理论和 MTE 理论的最新文献进展。需

① 请注意此处参数L的含义与 Brinch et al.(2017)命题 1 中的参数L不同，该结论与 Brinch et al.(2017)的命题 1 仍然保持一致。

要注意的是，在LATE理论和MTE理论中，工具变量的有效性要求工具变量满足独立性、排他性和单调性三个假设。那么，在异质性处理效应下，应当如何进行工具变量有效性检验？Kitagawa(2015)最早将工具变量有效性检验拓展到异质性处理效应的框架，提出可以利用方差加权Kolmogorov-Smirnov统计量(variance-weighted Kolmogorov-Smirnov test statistic)对工具变量有效性进行检验，并扩展到工具变量有多个取值和有协变量的情况。Sun(2023)将其拓展为更一般的框架，允许多值的处理变量是有序或无序的多值变量，并允许无界的结果变量。针对Mogstad et al.(2021)提出的部分单调性假设，Jiang and Sun(2023)基于Kitagawa(2015)和Sun(2023)的工作，提供了部分单调性假设的非参检验方法。此外，Carr and Kitagawa(2023)、Mao and Sant'Anna(2020)基于半参数模型，将异质性处理效应下的工具变量有效性检验推广到MTE的应用中。其中，Carr and Kitagawa(2023)能够处理协变量维度较高的情况，从而可以对MTE和LATE中的工具变量进行有效性检验。

四、异质性处理效应下工具变量法的经济学基础与政策评估应用

作为最重要的因果推断方法之一，工具变量法被广泛应用于政策评估。与此同时，近年来结构估计法(structural estimation)在政策分析中也呈现出明显的“复兴”趋势。本部分将对比结构估计法，探讨异质性处理效应下工具变量法在政策评估中的经济学基础及应用。

应用因果推断方法进行政策分析，旨在达成三个层次的目标。第一，评估已实施的政策在既定环境中的影响；第二，评估一个已实施的政策在其他环境中的潜在影响；第三，评估一个尚未实施的政策在目标环境中的潜在影响(Heckman and Pinto, 2024)。围绕这三个目标，我们比较上述两种工具变量法(LATE理论、边际处理效应[①])和结构估计法的经济学基础。

首先是LATE理论。作为一种非参估计方法，LATE理论及其最新进展有利于学者进行“事后”的因果效应识别，实现政策分析的第一个目标。经典的LATE将IV估计量合理地解释为政策“依从者”的平均处理效应，其优势在于简洁的非参设定。在此基础上，最新理论进展进一步拓展了其适用范围，为学者研究多元选择问题，或应用多个工具变量进行政策评估提供了分析的“利器”。这一进展也提醒了学者需谨慎考虑可能遗漏的其他替代项目对政策评估所产生的影响。以评估普通高中教育的收入回报为例。在初中毕业后，个体事实上面临着上普通高中、上中职学校或直接就业这三种选择，若研究者忽略了上中职学校这一选择，则会使政策估计结果失去合理的解释。

然而，LATE理论对政策结果的解释高度受限于工具变量，因而无法将其外推以实现目标二和目标三。作为一种非参估计方法，LATE理论无需对个体的偏好和行为做出假设。这一估计并非建立在经济学理论的基础上，而是将其结果定义在工具变量之上，

① 在异质性处理效应下，如何对Bartik IV这类连续型IV的估计结果进行合理的经济学解释，目前为止仍然没有定论。因此，本部分不对Bartik IV进行讨论。

即将估计的政策效果解释为工具变量“依从者”的平均处理效应。因此，LATE 理论可以对已实施的某个政策在具体的环境下进行“事后”的成本收益分析。但是当工具变量发生变化，或者当政策“依从者”人群发生变化时，研究者便无法把“事后”的政策评估结果外推到不同环境（目标二）或不同政策（目标三）。例如，想要估计大学教育的收入回报，并假设以外生的学费减免政策作为工具变量。对于穷国和富国，这两个国家的个体对于学费减免可能有不同的敏感程度（即大学教育的价格弹性），因此工具变量所对应的“依从者”不同。此时，无法将穷国的大学教育收入回报外推为富国的大学教育潜在收入回报，即无法实现目标二。同时，对于一个尚未实施的政策（如税收减免政策），我们无法将学费减免这一已实施的政策效果外推为尚未实施政策的潜在效果，从而无法实现目标三。

其次是边际处理效应（MTE）理论。边际处理效应对选择结构进行刻画，具有一定的经济学基础，可以实现目标一和目标二。当研究教育、医疗等领域可能对不同群体产生异质性影响的政策时，基于 MTE 的估计结果可以得到 LATE，并外推得到始终接受者、从不接受者和目标政策受众群体的平均处理效应，因此 MTE 可以实现目标一中的因果效应识别。同时，在估计出一个已实施政策的 MTE 后，只要知道该政策在其他环境中受众群体的个体异质性 U_D 的取值范围，即可实现目标二中的因果效应识别。但是 MTE 在实现目标一和目标二时仍然存在缺陷。当工具变量取值非连续时，MTE 的估计依赖于对 k 的方程形式进行参数假设。此外，作为一种半参数估计方法（semi-parametric estimation），边际处理效应对异质性偏好结构的刻画也是有限的。在 MTE 的设定下，个体的异质性只体现在选择结构的一个维度上（即 U_D），而没有对偏好和选择结构进行更加“底层”的假设，进而无法根据一个已实施政策的 MTE 外推得到另一个未实施政策的 MTE，不能实现目标三。

最后是结构估计法。结构估计法基于经济学理论，直接设定了效用函数和生产技术等模型形式，进而刻画一个更为完整的异质性偏好和选择结构，可以实现上述三个目标。尽管如此，结构方程法对偏好结构的假定也并非没有代价，其准确性十分依赖模型假定和参数估计的正确性（Heckman and Urzua，2010）。针对这些挑战，最新文献在如何提高模型设定、分布假设以及参数估计的准确性和稳健性方面取得了显著进展（Andrews et al.，2017，2020）。

以上讨论主要聚焦于三种方法在因果识别上的差异。在福利分析上，这三种方法同样也存在较大差异。首先，由于 LATE 没有对个体选择行为进行建模，无法进行福利分析。其次，尽管 MTE 基于间接效用函数对个体选择结构进行了模型设定，但只有当选择模型中包含与货币相关的变量时，才能计算个体在不同选择下的支付意愿，进而实现福利分析。最后，结构方程直接定义了效用函数，通常能据此为个体效用变化提供一种货币度量，从而进行福利分析。

综上，政策评估的各种方法不是截然对立的，而应互为补充，更好地服务于中国经济学研究的具体实践。近年来，中国在经济、社会、环境、外交等多个领域实施了广泛而深入的改革政策，如何对这些政策进行科学的评估显得尤为重要。为此，我们不仅可以基

于LATE框架评估这些已实施的政策，还可以进一步基于MTE和结构估计，模拟未实施政策的反事实结果，进而探索最优政策设计。

五、Bartik工具变量法最新发展

最后，本文介绍一种特殊的工具变量，即Bartik工具变量的最新文献发展。首先，介绍Bartik工具变量的定义和应用场景；其次，依次阐述Bartik工具变量能够得到一致估计量的三种方法：份额(share)外生法、变动(shift)外生法、直接控制法。最后，总结上述三种方法的适用条件和使用注意点。

（一）背景介绍

Bartik工具变量也被称为变动份额(shift-share)工具变量，是最常见的工具变量之一(Cunningham，2021)。以经典的劳动供给弹性问题为例，假设经济体中有L个地区，K个行业，考虑以下等式：

$$Y_l = \beta X_l + \gamma W_l + \varepsilon_l, \tag{9}$$

其中，Y_l表示地区l的工资增长率，X_l表示地区l的劳动力增长率，W_l表示地区l的控制变量(含常数项)，β为劳动供给弹性的倒数。如果直接进行最小二乘回归，有可能存在遗漏变量等内生性问题，比如不可观测的技术冲击会同时影响工资和劳动力的变化，从而导致β的估计结果存在偏误。Bartik(1991)基于以下思路构建了一个工具变量。首先，地区l的劳动力增长率X_l可以分解成地区l行业k的劳动力增长率g_{lk}和地区l行业k的期初劳动力占比z_{lk}的乘积之和，即$X_l = \sum_{k=1}^{K} z_{lk} g_{lk}$。其次，地区$l$行业$k$的劳动力增长率$g_{lk}$又可以进一步分解成全国层面行业$k$的劳动力增长率$g_k$和地区异质性$\tilde{g}_{lk}$两部分，即$g_{lk} = g_k + \tilde{g}_{lk}$。因此，$X_l$最终可以被分解成$\sum_{k=1}^{K} z_{lk} g_k + \sum_{k=1}^{K} z_{lk} \tilde{g}_{lk}$，其中第一项表示由各行业的全国平均劳动力增长率引致的地区劳动力增长率，第二项表示由各行业的地区异质性劳动力增长率引致的地区劳动力增长率。Bartik(1991)认为各行业的全国平均劳动力增长率g_k和地区层面不可观测的扰动项ε_l之间几乎不存在相关性，因此$\sum_{k=1}^{K} z_{lk} g_k$是$X_l$的一个有效工具变量，即Bartik工具变量(记为$B_l$，也被称为shift-share工具变量)，$g_k$为变动(shift)，$z_{lk}$为份额(share)。

实际上，Bartik工具变量并不是由Bartik(1991)首次提出，而是由Bartik(1991)首次对该工具变量背后的分解和逻辑进行详细阐述，所以最终用Bartik来命名(Cunningham，2021)。从上述的分析可以看出，只要存在类似$X_l = \sum_{k=1}^{K} z_{lk} g_{lk}$和$g_{lk} = g_k + \tilde{g}_{lk}$的两个分解式，就可以“无中生有”地构造Bartik工具变量，即Bartik工具变量相对而言比较容易构建，且通过这种分解的方式构建出来的工具变量和内生变量之间的相关性往往较高，因此Bartik工具变量被广泛应用于人口流动、国际贸易等研究领域(Autor et al.，2013；Card，2009；Olney and Pozzoli，2021；Imbert et al.，2022；Berman et al.，2015；Hummels et al.，2014；Peri et al.，2015；Stuen et al.，2012；Xu，2022)。尽管

Bartik(1991)对于 Bartik 工具变量的有效性进行了一些阐述，但是实际上，由于 Bartik 工具变量同时受到多个变动和份额的影响，其识别假设(identification assumption)一直以来都像是“黑箱”(Goldsmith-Pinkham et al.，2020)。比如 Card(2009)指出，当期初劳动力占比和扰动项之间存在相关性时，Bartik 工具变量可能不能很好地解决内生性问题。因此，最近一些文献尝试打开这个“黑箱”，对 Bartik 工具变量的识别假设进行进一步探究(Goldsmith-Pinkham et al.，2020；Borusyak et al.，2022；Borusyak and Hull，2023；Jaeger et al.，2018；Adão et al.，2019)。

（二）识别假设

设使用 Bartik 工具变量得到的两阶段最小二乘(2SLS)估计量为 $\hat{\beta}^B$，则有①：

$$\hat{\beta}^B-\beta=\frac{L^{-1}\sum_l B_l\varepsilon_l^{\perp}}{L^{-1}\sum_l B_l X_l^{\perp}}=\frac{L^{-1}\sum_l\sum_k z_{lk}g_k\varepsilon_l^{\perp}}{L^{-1}\sum_l\sum_k z_{lk}g_k X_l^{\perp}}.$$

要证明 $\hat{\beta}^B$ 是 β 的一致估计量，即证明 $\operatorname{plim}\hat{\beta}^B-\beta=0$，也就是证明

$$\operatorname{plim} L^{-1}\sum_l\sum_k z_{lk}g_k\varepsilon_l^{\perp}=0\text{(外生性)和 }\operatorname{plim} L^{-1}\sum_l\sum_k z_{lk}g_k X_l^{\perp}\neq 0\text{(相关性)}.$$

由于 Bartik 工具变量往往通过对内生变量进行分解得到，相关性条件较容易满足。因此本小节核心关注外生性条件，在假设相关性条件满足的前提下讨论 $\hat{\beta}^B$ 为一致估计量的三种情形：份额外生法、变动外生法、直接控制法。

1. 份额外生法

(1) 估计量一致性证明

Goldsmith-Pinkham et al.(2020)发现当 GMM 权重矩阵满足一定条件时，$\hat{\beta}^B$ 和直接用 $z_{lk}(k=1,2,\ldots,K)$ 作工具变量得到的 GMM 估计量在数值上是相等的，并从份额 z_{lk} 外生的角度给出 $\hat{\beta}^B$ 满足一致性的充分条件为 $E(z_{lk}\varepsilon_l\mid W_l)=0$。证明如下：当变量在地区 l 层面满足独立同分布(i.i.d.)时，根据大数定律，$\operatorname{plim}L^{-1}\sum_l\sum_k z_{lk}g_k\varepsilon_l^{\perp}=\sum_k g_k E(z_{lk}\varepsilon_l^{\perp})=0$。

可以类比双重差分法来理解上述份额外生法的合理性。考虑式(9)表示的劳动供给弹性问题，份额外生并不要求份额 z_{lk} 与工资水平无关，只要求份额与工资增长率无关，即要求在变动 g_k 发生之前，份额 z_{lk} 不同的地区的工资满足平行趋势。以最简单的两个地区(A 和 B)、两个行业(1 和 2)的情况为例。假设地区 A 只有行业 1($z_{A1}=1$)，地区 B 只有行业 2($z_{B2}=1$)。如图 4 所示，在 $t=1$ 时，即劳动力需求冲击 g_k 发生之前，份额 z_{lk} 并不影响劳动力增长率和工资增长率；在 $t=2$ 时，行业 1 发生劳动力需求冲击，份额 z_{lk} 通过影响地区是否接受冲击(或接受冲击的大小)进而影响劳动力增长率和工资增长率。

① 变量 $V^{\perp}$ 表示变量 V 对控制变量 W 回归之后得到的残差，证明参考 Frisch-Waugh-Lovell 定理。

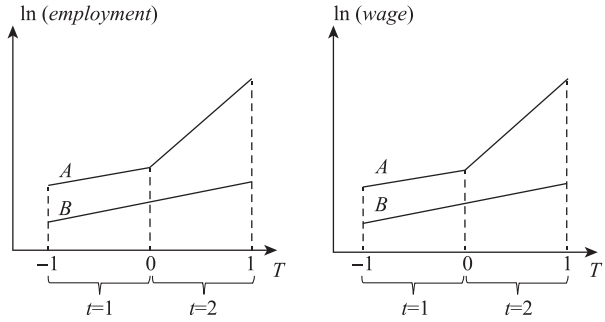

图 4 两地区两行业情况下的份额外生法

（2）工具变量有效性检验

由于每一个行业 k 都要求满足 $E(z_{lk}\varepsilon_l \mid W_l)=0$，当行业数量较多时，检验有效性的工作量较大，因此希望从中找到一些关键行业的 z_{lk} 进行检验。根据 Goldsmith-Pinkham et al.(2020)，可以将由 Bartik 工具变量得到的 $\widehat{\beta^B}=\dfrac{\sum_l B_l Y_l^{\perp}}{\sum_l B_l X_l^{\perp}}$ 进行如下行业层面的分解：$\widehat{\beta^B}=\sum_k \widehat{\alpha_k}\widehat{\beta_k}$。其中，$\widehat{\beta_k}=\dfrac{\sum_l z_{lk} Y_l^{\perp}}{\sum_l z_{lk} X_l^{\perp}}$ 表示用 z_{lk} 作为工具变量时得到的 2SLS 估计量，$\widehat{\alpha_k}=\dfrac{g_k \sum_l z_{lk} X_l^{\perp}}{\sum_{k'} g_{k'} \sum_l z_{lk'} X_l^{\perp}}$ 表示权重，$\sum_k \widehat{\alpha_k}=1$。进一步可以推出，$\widehat{\beta^B}-\beta=\sum_k \widehat{\alpha_k}(\widehat{\beta_k}-\beta)$，因此 $\widehat{\alpha_k}$ 可以衡量由 z_{lk} 造成的偏误会在多大程度上影响 $\widehat{\beta^B}$ 的一致性。这一分解的优点在于，当行业数量较多时，只需要重点关注 $\widehat{\alpha_k}$ 较大的几个行业的 z_{lk} 是否满足 $E(z_{lk}\varepsilon_l \mid W_l)=0$ 即可，因为这几个行业的 z_{lk} 造成的偏误会更严重地影响 Bartik 工具变量得到的 $\widehat{\beta^B}$ 的一致性。

由于 ε_l 不可观测，因此 $E(z_{lk}\varepsilon_l \mid W_l)=0$ 难以直接证明，但可以从相关变量、平行趋势、过度识别检验三个方面进行间接检验(Goldsmith-Pinkham et al., 2020)。第一，使用期初 z_{lk} 对期初地区特征变量进行回归。如果某些地区特征变量与 z_{lk} 相关，且同时能够影响 Y_l，则需要将其作为控制变量，以免对估计结果造成偏误。第二，类比双重差分法的思路，如果 z_{lk} 只通过影响接受冲击的程度来影响 Y_l，则在冲击发生之前，z_{lk} 不会影响 Y_l，即与平行趋势假设一致。第三，由于使用 Bartik 工具变量得到的 2SLS 估计量 $\widehat{\beta^B}$ 在数值上等于 GMM 权重矩阵满足一定条件时直接用 $z_{lk}(k=1,2,\ldots,K)$ 作为工具变量得到的 GMM 估计量，因此可以基于后者进行过度识别检验。

（3）份额外生法中变动的作用

在份额外生法中，冲击 g_k 的选取也非常重要，因为估计量的经济学含义取决于冲击的性质。仍然以两个地区(A 和 B)、两个行业(1 和 2)的情况为例，假设 $z_{A1}=1$，$z_{B2}=1$。

在第一种情况中，用劳动力需求冲击来估计劳动供给弹性。图 5(a)和图 5(b)表示，行业 1 的劳动力需求冲击来临之前($t=1$)，两个地区的劳动力增长率和工资增长率相

同；当劳动力需求冲击来临之后（$t=2$），地区 1 接受冲击，劳动力增长率提高，工资增长率提高，此时回归式(9)估计出来的 $\hat{\beta}$ 是劳动供给弹性的倒数。原理类似于图 5(c)中通过劳动力需求曲线的移动得到劳动力供给曲线上的不同点，从而求出劳动力供给曲线。

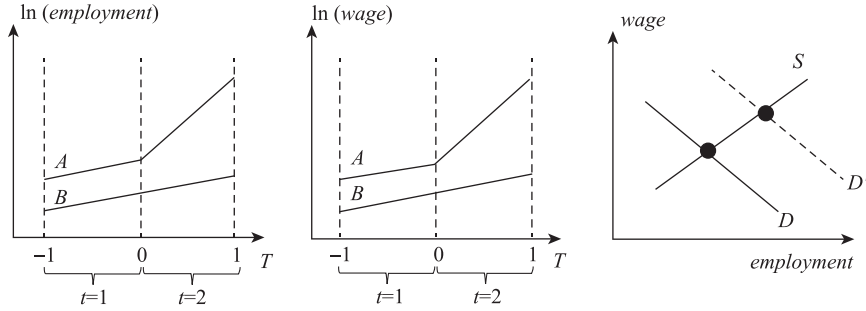

图 5　劳动力需求冲击的影响

在第二种情况中，用劳动力供给冲击来估计劳动需求弹性。图 6(a)和图 6(b)表示，行业 1 的劳动力供给冲击来临之前（$t=1$），两个地区的劳动力增长率和工资增长率相同；当劳动力供给冲击来临之后（$t=2$），地区 1 接受冲击，劳动力增长率提高，工资增长率降低，此时回归式(9)估计出来的 $\hat{\beta}$ 是劳动需求弹性的倒数。原理类似于图 6(c)中通过劳动力供给曲线的移动得到劳动力需求曲线上的不同点，从而求出劳动力需求曲线。

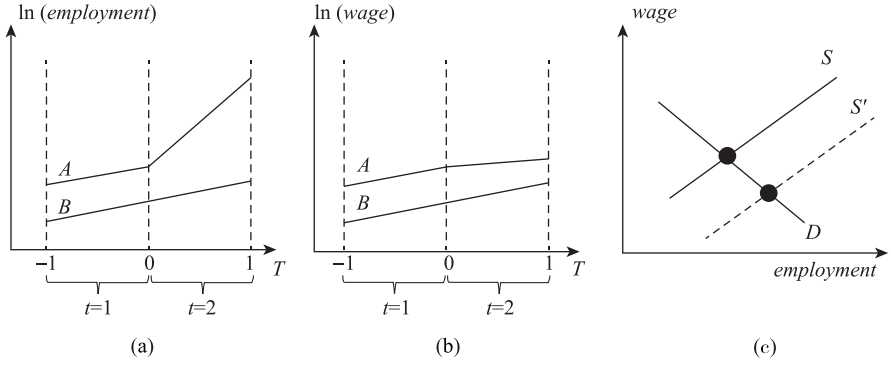

图 6　劳动力供给冲击的影响

但是现实中观察到的冲击 g_k 往往是劳动力供给冲击和需求冲击的混合。图 7(a)和图 7(b)表示，行业 1 的劳动力冲击来临之前（$t=1$），两个地区的劳动力增长率和工资增长率相同；当劳动力冲击来临之后（$t=2$），地区 1 接受冲击，劳动力增长率提高，但是工资增长率可能提高、可能降低、也可能保持不变，此时回归式(9)估计出来的 $\hat{\beta}$ 既不是劳动需求弹性的倒数，也不是劳动供给弹性的倒数，而是两者的混合。原理类似于图 7(c)中当劳动力供给曲线和需求曲线同时发生移动时，由新的均衡点和原来的均衡点得到的线既不是劳动力需求曲线，也不是劳动力供给曲线。

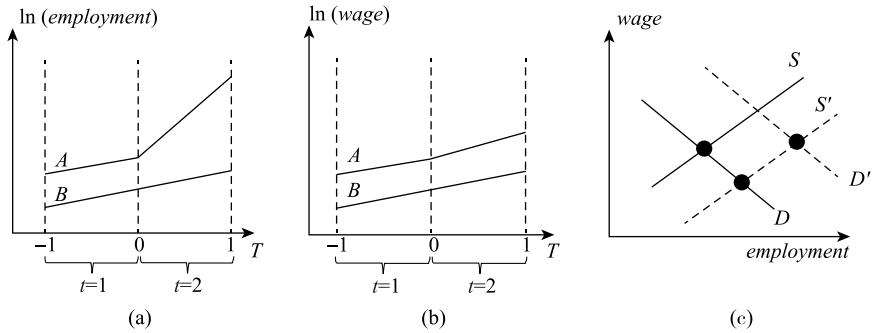

图 7 劳动力需求冲击和供给冲击的影响

从上述例子可以看出，在份额外生法中，无论冲击是否外生，都能得到一个满足一致性的估计量，但是冲击 g_k 决定了这个估计量的经济学含义，因此要根据想要识别的参数选取 g_k。而实际上，我们想要识别的参数往往需要通过一个外生的 g_k 才能实现。比如当研究的问题需要对需求侧和供给侧进行区分时，往往需要寻找一个外生的需求侧冲击或供给侧冲击。

2. 变动外生法

(1) 估计量一致性证明

Borusyak et al.(2022)指出拥有相似份额 z_{lk} 的地区的 X_l、B_l、ε_l 可能相似，因此变量在地区 l 层面不满足独立同分布(i.i.d.)，不能在 l 层面使用大数定律。因此 Borusyak et al.(2022)将地区 l 层面的统计量转化到行业 k 层面，从变动 g_k 外生的角度给出 $\widehat{\beta}^B$ 满足一致性的充分条件，具体为：①变动 g_k 随机，$E(g_k|z_{lk},\varepsilon_l)=\mu$；②各行业的变动互不相关，$\mathrm{cov}(g_k,g_{k'}|z_{lk},\varepsilon_l,k\neq k')=0$；③行业数量较多且在全国层面较为分散。证明思路[①]是首先根据 $\sum_l B_l\varepsilon_l^{\perp}=\sum_l\sum_k z_{lk}g_k\varepsilon_l^{\perp}=\sum_k g_k\sum_l z_{lk}\varepsilon_l^{\perp}=\sum_k g_k\overline{\varepsilon}_k^{\perp}$ 将统计量从 l 层面转化为 k 层面[②]，再基于条件②和条件③在 k 层面使用大数定律，最后使用条件①证明外生性[③]。

(2) 统计推断问题解决

在使用变动外生法时，不要求份额外生，因此份额 z_{lk} 在不同地区之间可能存在相关性。由于扰动项可能是变动份额的形式，因此扰动项在拥有相似份额的地区之间可能存在相关性[④]。如果不考虑这种相关性，就会造成标准误被低估，置信区间偏小，从而导致过度拒绝的问题(over-rejection problem)。考虑以 Bartik 工具变量直接作为解释变量的简约式回归(reduced-form regression)：

$$Y_l=\beta^{RF}B_l+\varepsilon_l=\beta^{RF}\sum_{k=1}^{K}z_{lk}g_k+\varepsilon_l, \tag{10}$$

① 具体证明请见附录Ⅰ(三)。

② 变量 $\overline{V}_k^{\perp}\equiv\sum_l z_{lk}V_l^{\perp}$。

③ 本文只讨论 $\sum_k z_{lk}=1$ 的情况，Borusyak et al.(2022)对 $\sum_k z_{lk}\neq 1$ 的情况进行了详细讨论。

④ 这一由变动份额形式导致的扰动项的相关性在 Borusyak et al.(2022)中被称为“exposure-based clustering”。

其中，被解释变量为 Y_l，解释变量为 B_l，扰动项 ε_l 由变动 v_k 和份额 z_{lk} 构成：

$$\varepsilon_l = \sum_{k=1}^{K} z_{lk} v_k, \tag{11}$$

则估计量的表达式为：

$$\hat{\beta}^{RF} = \frac{\sum_l B_l Y_l}{\sum_l B_l^2} = \beta^{RF} + \frac{\sum_l B_l \varepsilon_l}{\sum_l B_l^2}. \tag{12}$$

参考 Adão et al.(2019)，考虑如下反复抽样过程：条件于 z_{lk} 和 ε_l ①，每次独立随机地抽取 g_k，g_k 满足 $E(g_k|z_{lk},\varepsilon_l)=0$。在此基础上对 $\hat{\beta}^{RF}$ 进行统计推断，可以得到 $\hat{\beta}^{RF}$ 的渐近方差为：

$$V(\hat{\beta}^{RF}|z,\varepsilon) = \frac{\sum_k \mathrm{var}(g_k|z,\varepsilon)\left(\sum_l z_{lk}\varepsilon_l\right)^2}{\left(\sum_l B_l^2\right)^2},$$

令 $c(i)$、$c(j)$ 表示地区 i、j 所属的聚类②，则聚类稳健方差为：

$$V_{CL}(\hat{\beta}^{RF}|z,\varepsilon) = \frac{\sum_k \mathrm{var}(g_k|z,\varepsilon)\sum_i\sum_j \mathbf{1}\{c(i)=c(j)\}\, z_{ik}z_{jk}\varepsilon_i\varepsilon_j}{\left(\sum_l B_l^2\right)^2},$$

则两者的差值的期望为：

$$\begin{aligned} &E_\varepsilon[V(\hat{\beta}^{RF}|z,\varepsilon) - V_{CL}(\hat{\beta}^{RF}|z,\varepsilon)|z] \\ &= \frac{\sum_k \mathrm{var}(g_k|z,\varepsilon)\sum_i\sum_j \mathbf{1}\{c(i)\neq c(j)\}\, z_{ik}z_{jk}E(\varepsilon_i\varepsilon_j|z)}{\left(\sum_l B_l^2\right)^2}. \end{aligned}$$

由于扰动项 ε_l 为变动份额形式，假设各行业的 v_k 互不相关且条件期望为 0，则 $E[\varepsilon_i\varepsilon_j|z]=\sum_k E(v_k^2|z)z_{ik}z_{jk}\geqslant 0$，则 $E_\varepsilon[V(\hat{\beta}^{RF}|z,\varepsilon)|z]\geqslant E_\varepsilon[V_{CL}(\hat{\beta}^{RF}|z,\varepsilon)|z]$，从而证明了在使用变动外生法的情况下，如果扰动项中也包含变动份额部分，则扰动项在份额相似的地区之间会存在较大的相关性，导致通常使用的聚类稳健标准误偏低，由此产生过度拒绝的问题。

Adão et al.(2019)在 Autor et al.(2013)的基础之上进行安慰剂检验，对上述过度拒绝问题进行说明。被解释变量为 2000—2007 年就业率变化和工资变化，解释变量为 Bartik 工具变量，其中份额 z_{lk} 为 1990 年的就业结构，变动 g_k 从均值为 0、方差为 5 的正态分布中独立抽出。由于随机抽取的变动 g_k 独立于被解释变量和份额 z_{lk}，β^{RF} 真值为 0。但是在 30 000 次模拟中，如果使用稳健标准误或者地区聚类标准误，原假设(即 β^{RF} 为 0)在 5%的显著性水平下被拒绝的比例最高达到 56%左右。

① 这一设定允许 ε_l 之间存在各种形式的相关性(Adão et al., 2019)。

② $c(i)$ 指现有文献常用的聚类方式，比如在城市层面聚类、在省份层面聚类等，非按照份额进行聚类。

为了解决过度拒绝的问题，Adão et al.(2019)将份额引致的扰动项在地区之间的相关性考虑在内，构建了一个更加稳健的标准误，并给出了Stata、Matlab和R中的相应代码①。但是该标准误的构建方法要求地区数量大于行业数量(Adão et al., 2019)，当行业划分较细时这一条件较难满足。

Borusyak et al.(2022)则给出另一种更加一般化的解决方法。由于 $\widehat{\beta^B}=\frac{\sum_l B_l Y_l^{\perp}}{\sum_l B_l X_l^{\perp}}=\frac{\sum_l \sum_k z_{lk} g_k Y_l^{\perp}}{\sum_l \sum_k z_{lk} g_k X_l^{\perp}}=\frac{\sum_k g_k \sum_l z_{lk} Y_l^{\perp}}{\sum_k g_k \sum_l z_{lk} X_l^{\perp}}=\frac{\sum_k g_k \bar{Y}_k^{\perp}}{\sum_k g_k \bar{X}_k^{\perp}}$，Borusyak et al.(2022)指出可以将地区层面的回归式(9)转化为行业层面的回归：

$$\bar{Y}_k^{\perp}=\beta \bar{X}_k^{\perp}+\bar{\varepsilon}_k^{\perp}. \tag{13}$$

具体地，$\widehat{\beta^B}$ 可以通过以 g_k 为工具变量对 $\bar{Y}_k^{\perp}$ 和 $\bar{X}_k^{\perp}$ 进行两阶段最小二乘回归得到。从统计推断来看，在行业层面进行回归可以规避地区层面由于份额相似导致的扰动项相关的问题，因此能够直接使用上述两阶段最小二乘回归的稳健标准误。该方法既能借助Stata命令ssaggregate简便地实现，也不要求地区数量大于行业数量(Borusyak et al., 2022)。

(3) 工具变量有效性检验

根据Borusyak et al.(2022)，可以从安慰剂检验、有效样本量和过度识别检验三个方面进行检验。第一，由于不能直接检验 $E(B_l \varepsilon_l \mid W_l)=0$，因此可以寻找 ε_l 的代理变量 r_l (比如变动 g_k 发生之前的地区特征)，用 r_l 对 B_l 进行回归，并将地区层面的回归转化为行业层面的回归进行统计推断。② 第二，尽管基于式(13)以 g_k 为工具变量在行业层面进行两阶段最小二乘回归得到的稳健标准误是渐近有效的，但是当行业数量较少时，基于该标准误进行的统计推断则不一定准确。因此Borusyak et al.(2022)在实际应用中汇报了有效样本量，即 $1/\sum_k z_k^2$，其中 z_k 表示全国层面行业 k 的期初劳动力占比。③ Borusyak et al.(2022)利用蒙特卡罗方法证明了即使有效样本量仅为20，直接基于行业层面回归进行统计推断仍然较为准确。第三，Hahn et al.(2024)构建了一个一般性的框架来对使用Bartik工具变量的两阶段最小二乘估计进行过度识别检验，可以参考Hahn et al.(2024)进行检验。

---

① 在Stata中的相应代码详见https://github.com/zhangxiang0822/ShiftShareSEStata，在Matlab中的相应代码详见https://github.com/kolesarm/ShiftShareSEMatlab，在R中的相应代码详见https://github.com/kolesarm/ShiftShareSE。

② 在利用一阶段(first stage) X_l 对 B_l 的回归进行相关性检验时，也需要将地区层面的回归转化为行业层面的回归进行统计推断，注意由于 $\widehat{\beta^{FS}}=\frac{\sum_l B_l Y_l^{\perp}}{\sum_l B_l B_l^{\perp}}=\frac{\sum_l \sum_k z_{lk} g_k Y_l^{\perp}}{\sum_l \sum_k z_{lk} g_k B_l^{\perp}}=\frac{\sum_k g_k \sum_l z_{lk} Y_l^{\perp}}{\sum_k g_k \sum_l z_{lk} B_l^{\perp}}=\frac{\sum_k g_k \bar{Y}_k^{\perp}}{\sum_k g_k \bar{B}_k^{\perp}}$，转化之后行业层面的回归为用 g_k 作为工具变量对 $\bar{Y}_k^{\perp}$ 和 $\bar{B}_k^{\perp}$ 进行两阶段最小二乘回归。

③ 当行业平均发展($z_k=1/K$)时，有效样本量即为行业数量，行业数量越多则有效样本量越大。

3. 直接控制法

(1) 估计量一致性证明

变动外生法中讨论的 Bartik 工具变量为变动 g_k 和份额 z_{lk} 的线性组合，接下来将其拓展到更为一般化的形式：外生的变动和非随机的份额的任意形式组合，即 $B_l^* = f_l(g, z)$，其中 g 表示外生冲击，z 表示非随机的受冲击程度，f 表示函数关系。比如在研究交通基础设施发展的影响时，可以参考 Donaldson and Hornbeck(2016)构建市场可达性(MA_l)增长率指标作为代理变量。由于 $MA_l = \sum_{l' \neq l} \tau_{ll'}^{-\theta} N_{l'}$（其中 τ 表示与交通密切相关的贸易成本，θ 表示贸易弹性参数，N 表示人口数量），因此市场可达性增长率是外生的交通基础设施发展和内生的人口数量的非线性组合。

Borusyak and Hull(2023)指出即使在 g 为外生的条件下，B_l^* 也可能存在内生性问题。具体地，假设 g 的条件分布函数 $G(g|z)$ 已知，则可以求出 B_l^* 的条件期望 $\mu_l = E[f_l(g, z)|z]$，进一步结合 g 的外生性和期望迭代定理可以证明得到 $E[(1/N)\sum_l B_l^* \varepsilon_l^{\perp}] = E[(1/N)\sum_l \mu_l \varepsilon_l^{\perp}]$。[①]由于 μ_l 由非随机的 z_l 决定，因此 $E[(1/N)\sum_l \mu_l \varepsilon_l^{\perp}]$ 在一般情况下不等于 0，即 B_l^* 在一般情况下存在内生性问题。而在变动外生法中，Bartik 工具变量为变动和份额的线性组合的特殊情况下，由于 $\mu_l = E[\sum_k z_{lk} g_k|z] = \sum_k z_{lk} E(g_k|z) = \sum_k z_{lk}\mu = \mu$，才能推出 $E[(1/N)\sum_l \mu_l \varepsilon_l^{\perp}] = E[(1/N)\sum_l \mu \varepsilon_l^{\perp}] = E[(1/N)\mu \sum_l \varepsilon_l^{\perp}] = 0$，因此在这一特殊的 f 形式下 B_l^* 不存在内生性问题。

在此基础上，Borusyak and Hull(2023)提出了两种处理内生性问题的思路。第一，将 B_l^* 中内生决定的部分 μ_l 剔除，构建一个重新调整的工具变量 $\widetilde{B}_l^* = B_l^* - \mu_l$。由于 $E[(1/N)\sum_l \widetilde{B}_l^* \varepsilon_l^{\perp}] = 0$，因此 $\widetilde{B}_l^*$ 满足工具变量的外生性条件。第二，在控制变量中加入 μ_l，由 $\varepsilon_l^{\perp}$ 的构造可以推出 $\sum_l \mu_l \varepsilon_l^{\perp} = 0$，则 $E[(1/N)\sum_l B_l^* \varepsilon_l^{\perp}] = 0$，此时用 B_l^* 作为工具变量满足外生性条件。

(2) 统计推断问题解决与工具变量有效性检验

需要注意的是，这一设定同样存在变动外生法中由份额导致的聚类问题(exposure-based clustering)。Borusyak and Hull(2023)指出可以使用随机推断(randomization inference)的方法进行统计推断。在进行工具变量有效性检验时，除了可以和变动外生法中类似地对 ε_l 的代理变量 r_l 和 $\widetilde{B}_l^*$ 进行回归来间接检验 $E[(1/N)\sum_l \widetilde{B}_l^* \varepsilon_l^{\perp}]$ 是否为 0 之外，还可以对 μ_l 和 $\widetilde{B}_l^*$ 进行回归来检验两者是否相关，若 μ_l 能包含 B_l^* 中所有内生决定的部分，则 μ_l 和 $\widetilde{B}_l^*$ 应该不相关。

(三) 方法评价与总结

上述三种使用 Bartik 工具变量进行识别的方法有其各自的使用注意点。从份额外

① 具体证明过程如下：$E[(1/N)\sum_l B_l^* \varepsilon_l^{\perp}] = E[(1/N)\sum_l f_l(g, z)\varepsilon_l^{\perp}] = E[(1/N)\sum_l E[f_l(g, z)\varepsilon_l^{\perp}|z]] = E[(1/N)\sum_l E[f_l(g, z)|z]E[\varepsilon_l^{\perp}|z]] = E[(1/N)\sum_l \mu_l E[\varepsilon_l^{\perp}|z]] = E[(1/N)\sum_l \mu_l \varepsilon_l^{\perp}]$。

生法来看，Borusyak et al.(2022)指出份额的外生性事先(ex ante)很难满足。如果一些不可观测的冲击 v_k 也发生在行业层面(比如高技术移民冲击)，并通过 z_{lk} 作用于地区层面，即 $\varepsilon_l=\sum_k z_{lk}v_k$，则 $E(z_{lk}\varepsilon_l|W_l)=0$ 天然不可能满足。变动外生法面临的挑战是当行业数量很少或者只强调冲击对于几个特定行业的影响时，变动外生法不再适用(Goldsmith-Pinkham et al., 2020；Borusyak et al., 2022)。直接控制法的难点在于，由于 μ_l 的计算依赖于 g 的条件分布，而实际中通常只能观察到 g 的一次实现，因此只有基于特定的研究问题或者对 g 进行一些特殊假设，才能得到反事实的 g 及其分布。

综上，对 Bartik 工具变量的三种方法进行总结。当份额 z_{lk} 外生时，可以使用份额外生法得到无偏一致的估计量(Goldsmith-Pinkham et al., 2020)，份额 z_{lk} 越特殊、越为研究问题的内生变量“量身定制”，就越难通过其他冲击影响被解释变量，越有可能外生(Borusyak et al., 2022)。当变动 g_k 外生时，如果行业数量较大且变动和份额之间为线性组合，可以使用变动外生法得到无偏一致的估计量，并通过将地区层面的回归转化为行业层面的回归进行正确的统计推断(Borusyak et al., 2022)。如果变动和份额之间的组合较为复杂，则可以参考 Borusyak and Hull(2023)，基于冲击 g 的条件分布计算 Bartik 工具变量的条件期望值 μ，由于 μ 吸收了所有内生决定的部分，因此通过将 μ 作为控制变量或者将其从 Bartik 工具变量中剔除，可以得到无偏一致的估计量。需要指出的是，在使用份额外生法时，变动 g_k 决定了估计量的经济学含义，变动 g_k 非外生可能会导致识别出的参数并非目标参数。①

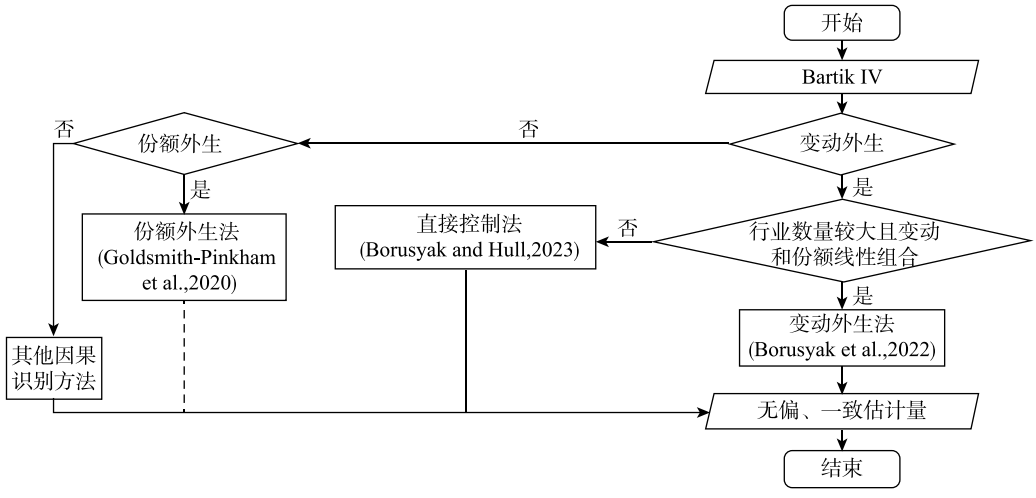

图 8 Bartik 工具变量使用方法总结

六、结论与展望

本文通过梳理工具变量法的最新进展，以期为学者更广泛、更准确地应用工具变量法研究具体经济学问题提供参考。本文主要对工具变量三个方面的前沿进展进行了讨

① 为了强调这一点，图 8 中“份额外生法”与“无偏、一致估计量”之间用虚线连接。

论。当研究多元选择问题(例如:专业选择、就业选择),或当研究中使用了多个工具变量时,LATE 理论的最新进展有助于研究者更科学、准确地解释工具变量法的估计结果。特别地,在评估公共项目时,学者需要考虑其他替代项目作为一种潜在选择所产生的影响。当研究教育、医疗等领域可能对不同群体产生异质性影响的政策时,边际处理效应理论的最新研究为学者将 LATE 外推得到 ATE 提供了行之有效的分析工具。当研究交通基础设施建设、人口流动或国际贸易等问题时,Bartik IV 方法的前沿理论为研究者如何使用 Bartik IV 得到一致估计量提供了参考。

工具变量法作为重要的因果推断方法之一,能与其他经济学方法互为补充,更好地服务于经济学研究的具体实践。一方面,在其他因果推断方法存在局限性时,研究者可以采用工具变量法作为有效补充。比如,由于"非完全"依从者的存在,直接采用随机对照试验方法进行估计的结果是有偏的,此时,工具变量法可以被视为一种有效补充。此外,文献中常利用双重差分法构造工具变量,此时,双重差分法的设定被用于工具变量法的第一阶段回归(Waldinger, 2010;马超等,2023)。

另一方面,工具变量法依然面临许多挑战。例如,当有多个工具变量时,尽管 Mogstad et al.(2021)提出的部分单调性假设赋予了 LATE 理论在该情形下"新的生命力"。但若存在负的权重,LATE 理论便又"黯然失色",失去其对估计结果的解释能力。同时,对于连续型的内生变量或工具变量,LATE 理论至今仍无法对其估计结果进行合理的解释。不仅如此,在估计始终接受者和从不接受者的平均处理效应时,LATE 理论也是"无能的"。在这些情形下,边际处理效应体现出在估计异质性处理效应时的优越性。但作为一种半参数估计方法,MTE 对个体异质性的刻画只体现在选择结构的一个维度上,而对定义在其他维度的异质性则无能为力。对于这样的情况,研究者可以利用结构方程法,基于一个更为完整的异质性偏好和选择结构进行估计。尽管如此,利用结构方程法进行估计也并非没有缺点,其准确性在很大程度上取决于模型结构和参数假设的合理性。

在经济学研究的具体实践中,学者应当"取长补短",结合具体研究问题灵活运用各种研究方法。以政策评估为例,由于 LATE 缺乏相关理论基础,无法将其估计结果外推到其他反事实的场景中。因此,在研究实践中,首先可以将 LATE 框架下的估计结果作为特征事实,并以此为研究参考,进一步建立理论模型。然而,是否需要基于 LATE 拓展到 MTE,或进一步拓展到结构估计,以及如何进行拓展,则需要研究者结合具体的研究问题和已有的数据进行判断,进而恰当地设计和丰富研究内容。

参 考 文 献

[1] Adão R., M. Kolesár, and E. Morales, "Shift-share Designs: Theory and Inference", *The Quarterly Journal of Economics*, 2019, 134(4), 1949-2010.

[2] Andrews, I., M. Gentzkow, and J. M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments", *The Quarterly Journal of Economics*, 2017, 132(4), 1553-1592.

[3] Andrews, I., M. Gentzkow, and J. M. Shapiro, "On the Informativeness of Descriptive Statistics for Structural Estimates", *Econometrica*, 2020, 88(6), 2231-2258.

[4] Angrist, J. D., "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Adminis-

trative Records", *The American Economic Review*, 1990, 80(3), 313-336.

[5] Angrist, J. D., G. W. Imbens, and D. B. Rubin, "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, 1996, 91(434), 444-455.

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

[7] Autor, D. H., D. Dorn, and G. H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States", *The American Economic Review*, 2013, 103(6), 2121-2168.

[8] Bartik, T. J., *Who Benefits from State and Local Economic Development Policies*? W. E. Upjohn Institute for Employment Research, 1991.

[9] Berman, N., A. Berthou, and J. Héricourt, "Export Dynamics and Sales at Home", *Journal of International Economics*, 2015, 96(2), 298-310.

[10] Björklund, A., and R. Moffitt, "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models", *The Review of Economics and Statistics*, 1987, 69(1), 42-49.

[11] Black, B., E. French, J. McCauley, and J. Song, "The Effect of Disability Insurance Receipt on Mortality", *Journal of Public Economics*, 2024, 229, 105033.

[12] Black, S. E., P. J. Devereux, and K. G. Salvanes, "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education", *The Quarterly Journal of Economics*, 2005, 120(2), 669-700.

[13] Borghesan, E., and G. Vasey, "The Marginal Returns to Distance Education: Evidence from Mexico's Telesecundarias", *American Economic Journal: Applied Economics*, 2024, 16(1), 253-285.

[14] Borusyak, K., and P. Hull, "Non-random Exposure to Exogenous Shocks", *Econometrica*, 2023, 91(6), 2155-2185.

[15] Borusyak, K., P. Hull, and X. Jaravel, "Quasi-experimental Shift-Share Research Designs", *The Review of Economic Studies*, 2022, 89(1), 181-213.

[16] Brinch, C. N., M. Mogstad, and M. Wiswall, "Beyond LATE with a Discrete Instrument", *Journal of Political Economy*, 2017, 125(4), 985-1039.

[17] Card, D., "Immigration and Inequality", *The American Economic Review*, 2009, 99(2), 1-21.

[18] Carneiro, P., J. J. Heckman, and E. J. Vytlacil, "Estimating Marginal Returns to Education", *The American Economic Review*, 2011, 101(6), 2754-2781.

[19] Carneiro, P., and S. Lee, "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality", *Journal of Econometrics*, 2009, 149(2), 191-208.

[20] Carr, T., and T. Kitagawa, "Testing Instrument Validity with Covariates", *Working Paper*, 2023.

[21] Cunningham, S., *Causal Inference: The Mixtape*. Yale university press, 2021.

[22] Donaldson, D., and R. Hornbeck, "Railroads and American Economic Growth: A 'Market Access' Approach", *The Quarterly Journal of Economic*, 2016, 131(2), 799-858.

[23] French, E., and J. Song, "The Effect of Disability Insurance Receipt on Labor Supply", *American Economic Journal: Economic Policy*, 2014, 6(2), 291-337.

[24] Goldsmith-Pinkham, P., I. Sorkin, and H. Swift, "Bartik Instruments: What, When, Why, and How", *American Economic Review*, 2020, 110(8), 2586-2624.

[25] Hahn, J., G. Kuersteiner, A. Santos, and W. Willigrod, "Overidentification in Shift-Share Designs", *Working Paper*, 2024.

[26] Heckman, J. J., "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy", *Journal of Economic Literature*, 2010, 48(2), 356-398.

[27] Heckman, J. J., and R. Pinto, "Econometric Causality: The Central Role of Thought Experiments", *Journal of Econometrics*, 2024, 105719.

[28] Heckman, J. J., and S. Urzua, "Comparing IV with Structural Models: What Simple IV Can and Cannot Identi-

fy", *Journal of Econometrics*, 2010, 156(1), 27-37.

[29] Heckman, J. J., S. Urzua, and E. Vytlacil, "Understanding Instrumental Variables in Models with Essential Heterogeneity", *The Review of Economics and Statistics*, 2006, 88(3), 389-432.

[30] Heckman, J. J., S. Urzua, and E. Vytlacil, "Instrumental Variables in Models with Multiple Outcomes: The General Unordered Case", *Annales d'Economie et de Statistique*, 2008, 151-174.

[31] Heckman, J. J., and E. J. Vytlacil, "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects", *Proceedings of the National Academy of Sciences of the United States of America*, 1999, 96(8), 4730-4734.

[32] Heckman, J. J., and E. Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation", *Econometrica*, 2005, 73(3), 669-738.

[33] Heckman, J. J. and E. J. Vytlacil, "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments", *Handbook of Econometrics*, 2007, 6, 4875-5143.

[34] 黄炜、张子尧、刘安然,"从双重差分法到事件研究法",《产业经济评论》,2022 年第 2 期,第 17—36 页。

[35] Hummels, D., R. Jørgensen, J. Munch, and C. Xiang, "The Wage Effects of Offshoring: Evidence from Danish Matched Worker-Firm Data", *The American Economic Review*, 2014, 104(6), 1597-1629.

[36] Imbens, G. W., and J. D. Angrist, "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 1994, 62(2), 467-475.

[37] Imbert, C., M. Seror, Y. Zhang, and Y. Zylberberg, "Migrants and Firms: Evidence from China", *American Economic Review*, 2022, 112(6), 1885-1914.

[38] Jaeger, D. A., J. Ruist, and J. Stuhler, "Shift-share Instruments and the Impact of Immigration", *NBER Working Paper*, 2018, (No. W24285).

[39] Jiang, H., and Z. Sun, "Testing Partial Instrument Monotonicity", *Economics Letters*, 2023, 233, 111400.

[40] Kirkeboen, L. J., E. Leuven, and M. Mogstad, "Field of Study, Earnings, and Self-selection", *The Quarterly Journal of Economics*, 2016, 131(3), 1057-1112.

[41] Kitagawa, T., "A Test for Instrument Validity", *Econometrica*, 2015, 83(5), 2043-2063.

[42] Kline, P., and C. R. Walters, "The Case of Head Start", *The Quarterly Journal of Economics*, 2016, 131(4), 1795-1848.

[43] Kline, P., and C. R. Walters, "On Heckits, LATE, and Numerical Equivalence", *Econometrica*, 2019, 87(2), 677-696.

[44] 马超、赵双雨、唐润宇,"上医治未病：免费体检计划对老年人医疗服务与健康福利的影响",《管理世界》,2023 年第 12 期,第 144—166 页。

[45] Mao, M., and P. H. C. Sant'Anna, "Testing Instrument Validity in Marginal Treatment Effects Models", *unpublished manuscript*, 2020.

[46] Moffitt, R., "Estimating Marginal Treatment Effects in Heterogeneous Populations", *Annales d'Economie et de Statistique*, 2008, (91/92), 239-261.

[47] Mogstad, M., A. Torgovitsky, and C. R. Walters, "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables", *American Economic Review*, 2021, 111(11), 3663-3698.

[48] Nybom, M., "The Distribution of Lifetime Earnings Returns to College", *Journal of Labor Economics*, 2017, 35(4), 903-952.

[49] Olney, W. W., and D. Pozzoli, "The Impact of Immigration on Firm-level Offshoring", *Review of Economics and Statistics*, 2021, 103(1), 177-195.

[50] Peri, G., K. Shih, and C. Sparber, "STEM Workers, H-1B Visas, and Productivity in US Cities", *Journal of Labor Economics*, 2015, 33(S1), S225-S255.

[51] Stock, J. H., and F. Trebbi, "Retrospectives: Who Invented Instrumental Variable Regression?", *Journal of Economic Perspectives*, 2003, 17(3), 177-194.

[52] Stuen, E. T., A. M. Mobarak, and K. E. Maskus, "Skilled Immigration and Innovation: Evidence from Enrolment Fluctuations in US Doctoral Programmes", *The Economic Journal*, 2012, 122(565), 1143-1176.

[53] Sun, Z., "Instrument Validity for Heterogeneous Causal Effects", *Journal of Econometrics*, 2023, 237(2), 105523.

[54] Vytlacil, E., "Independence, Monotonicity, and Latent Index Models: An Equivalence Result", *Econometrica*, 2002, 70(1), 331-341.

[55] Waldinger, F., "Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany", *Journal of Political Economy*, 2010, 118(4), 787-831.

[56] Xu, C., "Reshaping Global Trade: The Immediate and Long-run Effects of Bank Failures", *The Quarterly Journal of Economics*, 2022, 137(4), 2107-2161.

The Recent Development in Instrumental Variable Estimator and Its Applications

WANG Ye
(Peking University)
LI Haifeng*
(Fuzhou University)
YANG Rudai YI Junjian
(Peking University)

Abstract: The literature in applied microeconometrics primarily employs three approaches to establish causal inferences, particularly in the presence of treatment effect heterogeneity. Among these approaches, the instrumental variable (IV) estimator, founded on the assumption of exclusion restriction, emerges as one of the most prevalent. While the local average treatment effect (LATE) framework offers a clear interpretation for the IV estimator, its interpretation is unclear with multiple instrumental variables or multiple arms of treatment. In addition, the treatment effect for non-compliers in this framework is unable to be estimated. We present a comprehensive review of recent advancements in the IV estimator, which extends the classic LATE framework to accommodate multiple instrumental variables or multiple arms of treatment. Additionally, we summarize the relationship between the marginal treatment effect (MTE) and the LATE and how to obtain the average treatment effect for non-compliers based on MTE. Finally, we delve into recent developments in a specialized type of IV estimator—the Bartik IV estimator.

Keywords: LATE; MTE; Bartik IV

JEL Classification: C13, C26, C36

* Corresponding Author: LI Haifeng, The School of Economics and Management, Fuzhou University, No.2 Xue Yuan Road, University Town, Fuzhou, Fujian 350108, China; Tel: 86-15321596109; E-mail: lhf_0730@163.com.