第1章
导论与核心概念

本章前言

本章对微观计量经济学中的一些基本术语和核心概念进行定义和解释，阐述微观计量经济学建模所涉及的基本问题，并对后续章节所涉及的计量方法进行总览性介绍，对本章内容的掌握是学习后续内容的前提条件和重要基础。本章首先区分了变量间的因果性和相关性，接着在介绍被广泛应用于政策评估领域的随机控制实验、自然实验与准实验等因果识别策略基础上，引入了“反事实”的研究框架和处理效应的概念。最后，为了正确估计政策的处理效应，本章还考虑了选择性偏差问题。

1.1 节阐述了事件之间的两种基本关系：因果关系和相关关系。1.2 节对三类不同的实验方法进行介绍。1.3 节对处理效应的基本定义和形式进行解释，介绍了不同条件下如何估计平均处理效应。1.4 节介绍了非随机分配条件下基于可观测和不可观测因素的选择，以及这两类选择下适用的评估政策处理效应的计量方法。1.5 节对选择性偏差的定义、表现形式、分解与解决方法进行了阐述。1.6 节对本书介绍的政策评估计量方法进行总览性分类与评述。1.7 节为本章小结。1.8 节为经典案例分析。

本章关键词 ›››

1. 政策评估（policy evaluation）：在经济学研究领域中，政策评估也被称为项目评估（program evaluation）或影响评估（impact evaluation）。其关注于政策的绩效与效率，旨在通过对政策的评估深入了解政策的效果及其背后的作用机制。

2. 控制实验（controlled experiment）：控制实验是在实验中设置比较对象（控制组）的一种科学方法，是指除了实验要求的研究因素或操作处理外，其他因素保持一致的情况下对实验结果进行比较的实验，并以此阐明一定因素对一个对象的影响。

3. 随机控制实验（randomized controlled experiment）：随机控制实验的基本方法是将研究对象随机分组，对不同组实施不同的干预，以对照效果的不同。

4. 自然实验（natural experiment）：自然实验是一种自然发生的实验，而并非为了实验目的，即利用自然发生的事件（如气候变化、地理因素等）的冲击得到类似随机控制实验的随机分配效果的一种实验。

5. 准实验（quasi-experiment）：准实验设计是利用一定的标准（例如我国高考分数线的设置）在现有的观察数据上构造类似随机控制实验的随机分配，从而得到可比的处理组和控制组的一种研究方法。

6. 处理效应（treatment effect）：处理效应是指在经济学研究领域中进行项目效应评估时，对某一项目或者政策实施后产生的效果的量度。

7. “反事实”（counter-factual）：对于一个政策冲击而言，在两种潜在处理变量（受到政策处理与未受到政策处理）对应的两种潜在结果变量（受到政策处理时的结果与未受到政策处理时的结果）之间，仅能观察到一种潜在结果，而另一种无法观测到的潜在结果就被称为“反事实”。

8. 选择性偏差（selection bias）：研究过程中样本选择的非随机性导致的结论偏差，包括自选择偏差和样本选择偏差。

导入案例 ›››

教育政策的因果效应评估问题

2020—2035 年是实现我国“两个一百年”奋斗目标的关键阶段，各项改革均已进入深水区，教育发展面临新时代“培养什么样的人”和“怎样培养人”的重大课题。许多教育改革政策如雨后春笋般纷纷涌现，而每一项教育改革政策的出台都有其预期目标，

比如义务教育均衡发展的目标是促进全国省区间、县域间以及城乡间的义务教育均衡化，但政策实施后的真正效果仍需要一定的评估与检验。

经济学和其他社科领域学科在评估和检验政策效果时，最重要的就是做好因果识别。由于学科、研究问题和数据条件的不同，学者们也往往采用不同的因果推断方法。其中，常见的有随机控制实验、自然实验与准实验等，而随机控制实验因具有能明确实验的所有可能结果、相同条件下可重复进行等特点，成了最常见的因果识别方法。

一般而言，理想的随机控制实验包括两个阶段：第一阶段是完全随机抽样，但通常情况下难以从总体样本中完全随机抽取“被试”[①]对象。比如要评估独生子女政策对学生近视率的影响，通常的做法是在全国范围内选择几个地区，再从被选地区内选择几所学校，基于学校层面随机抽取学生。

第二阶段是令个体随机进入处理组或者控制组。有时这一阶段也难以实现。例如，探究小班教学对学生成绩的影响，父母对于子女是否进入小班学习持不同态度，就可能使小班教学对学生成绩的影响评估出现偏差。此外，进入小班学习的同学自身可能有着更好的学习能力与交际能力，但是这些能力往往难以准确测量，致使因果估计出现偏误。这在计量经济学中被称作遗漏变量偏误（omitted variable bias）问题[②]。基于不完全随机控制实验获得的样本进行政策评估，会造成遗漏变量偏误问题、逆向因果效应问题[③]，从而使评估政策或项目的因果效应时得出的结论有偏差。

由上述可知，随机控制实验在现实中难以实现。那么在不完全随机控制实验的情况下，如何通过技术手段降低估计偏差？有什么办法可以更好地进行因果关系的识别？为了回答以上问题，计量经济学家们开始致力于研究基于准实验数据和观察数据估测政策或项目效果的计量方法。双重差分法（DID）、工具变量法（IV）、匹配技术（matching）、断点回归（RD）等方法应运而生。这些方法在公共政策和项目评估领域的推广应用，提升了政策或项目干预效应评估研究的内在效度，我们将在本书的第 2 章到第 5 章对这些方法进行系统介绍和讲解，以帮助读者更好地学习和应用。

1.1　因果性与相关性

在社会科学领域，许多问题的核心和实质是政策评估：独生子女政策是否降低了学生视力？移民是否加剧了不平等现象？科技创新是否有助于缓解环境压力？建立国家层

① 被试：被试多用于生物或心理实验中，在经济学中的应用并不很广泛。此处指该随机试验中接受测试的对象。

② 遗漏变量偏误：在回归估计中，由于遗漏自变量而导致其他自变量的估计出现偏误。

③ 逆向因果效应：参与者认为接受干预对其有利而主动选择进入处理组所产生的效应。

面或者超国家的财政规则是否有助于控制政府债务规模？为了回答这些问题，政策评估者需要明确政策的真实因果效应并在此基础上探寻其背后的作用机制。

在识别因果关系前，首先需要了解因果性与相关性、因果关系与相关关系的内涵及相互关系。

1.1.1　因果性

因果性是指一个变量的变化完全由另一个变量的变化引起，其核心在于推断潜在的、看不见的“反事实”结果。社会科学研究中因果推论的应用最初是受到了一些自然科学现象和观点的启发，经济学家基于控制实验的思想，通过多种推断分析方法，试图像自然科学实验一样获得准确的结果，使经济学更加严谨、科学。

1.1.2　相关性

当人们在观察某个研究对象时，如果能够得到一个结论，即这个研究对象的变化总是与另一个对象的变化相随变动，那么我们就可以说这两者是相关的。

一般来讲，相关性本质可以被定义为：如果变量 A 的变化总是伴随着变量 B 的变化，那么可以说 A 和 B 是相关的。需要注意此处用的词是“伴随”，如果变量 A 的变化总是“引起”变量 B 的变化，那么它们之间不仅相关，很可能还存在因果关系。

1.1.3　因果关系与相关关系

1. 因果关系和相关关系的概念

所谓因果关系，是指某个因素的存在一定会导致某个特定结果的产生。简单地说，就是 $A\rightarrow B$，即事件 A 的发生导致事件 B 的发生。因果关系中最常见的形式是一因一果，另外还有一因多果、一果多因、多因多果等形式。

相关关系，是指两个事件的发生存在关联。如以 X 和 Y 来代表两个事件，用相关系数 ρ_{XY}（$|\rho_{XY}|\leqslant 1$）表示两个事件之间的关系。若 $\rho_{XY}\neq 0$，则称 X 与 Y 相关；若 $\rho_{XY}>0$，X 与 Y 正相关；若 $\rho_{XY}<0$，X 与 Y 负相关。需注意，当 $\rho_{XY}=1$ 时，X 与 Y 完全正相关；当 $\rho_{XY}=-1$ 时，X 与 Y 完全负相关；当 $\rho_{XY}=0$ 时，说明 X 与 Y 不存在线性相关，但两个事件之间仍有可能存在其他相关关系。

因果关系在统计学意义上是由概率体现出来的：如果一个事件 A 发生的概率对另一个事件 B 发生的概率有影响，并且这两个事件在发生时间上有先后顺序（例如 A 前 B 后），那么我们便可以说 A 是 B 的原因。但这种定义在研究上存在很大的缺陷，因为我们不

可能收集到全部信息，以至于无法严格证明 A 对 B 存在直接影响，只能定性说明。

2. 因果性和相关性的区别

理解因果关系和相关关系之间的区别是至关重要的。下面简要介绍一个案例：由于患有心血管疾病的高危群体往往倾向于服用阿司匹林，假设服用阿司匹林的人在 5 年内的死亡率高于不服用阿司匹林的人的死亡率，如果医生得知这一统计事实后，简单基于“服用阿司匹林的病人比不服用的病人死亡风险更大”的结果而决定不让病人服用阿司匹林，这位医生将会因玩忽职守而被起诉，因为他没有真正明白因果关系和相关关系之间的区别。服用阿司匹林与病人死亡之间只存在着相关关系而不存在因果关系，更高的死亡率是患者本身的心血管疾病导致的，只是因为这些高危群体更倾向于服用阿司匹林，使死亡与阿司匹林之间出现相随变动的现象。这个案例说明，相关关系并不一定反映因果关系，甚至可能还会导致相反的结论，我们需要在理论分析的基础上，选择恰当的实证模型来进一步识别因果关系。

下面我们以心脏移植为例进行讲解。假设病人患有心脏病需要进行心脏移植，进行心脏移植可能死亡也可能痊愈，我们用 A 表示处理变量，其中 $A=1$ 表示病人接受处理，进行了心脏移植，而 $A=0$ 表示病人未接受处理，未进行心脏移植；用 Y 表示结果变量，其中 $Y=1$ 表示病人死亡，$Y=0$ 表示病人痊愈，$\Pr(Y=1|A=1)=\frac{7}{13}$，$\Pr(Y=1|A=0)=\frac{3}{7}$。图 1-1 描述了因果关系与相关关系差异，所有的病人（用菱形表示）被分为一个白色区域（接受处理的）和一个较小的蓝色区域（未接受处理的）。

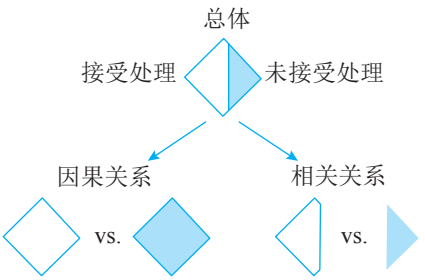

图 1-1　因果关系关联与相关关系关联的差异

因果关系的定义意味着整个白色菱形（所有个体都被处理过）和整个蓝色菱形（所有个体都未被处理过）之间的对比，而相关关系则意味着原始菱形的白色区域（被处理过）和蓝色区域（未处理过）之间的对比。也就是说，关于因果关系的推论关注的是“反事实”世界中的问题，比如“如果每个人都接受治疗，风险会是什么”以及“如果每个人都不接受治疗，会有什么风险”；而关于相关关系的推论则与现实世界中的问题有关，比如“被治疗者的风险是什么”以及“未经治疗的风险是什么”。这些截然不同的定义解释了“相关关系不是因果关系”这句话。在这个案例中存在着相关关系，因为接受治

疗的死亡率（7/13）大于未经治疗的死亡率（3/7）。但是，这个案例中没有因果关系，因为每个人接受治疗面临的概率（1/2）和未经治疗面临的概率（1/2）是相同的，即是否接受治疗的概率相等。

1.1.4　常见错误

在因果关系的确认中，常见的错误主要有以下几种形式：

1. 错误确定因果关系

（1）存在共同解释变量（混淆变量）：有相关关系而因果关系需进一步识别，在图 1-2 中也有解释。有个常见的误区是：“如果 B 紧跟着 A 发生，那么 A 一定导致 B。”在这里，或许 A 是 B 的原因，B 是 A 的结果，但也可能 A 和 B 相互之间并没有因果关系，这一现象的发生是受第三因素 C 的影响，在 C 的影响下 A 和 B 一起发生。

（2）存在共同影响变量（对撞变量）：无相关关系也无因果关系。很多情况下，我们只能确定两个变量之间具有相关性，是否具有因果关系仍是未知。因此，在明确推断变量之间是否存在因果关系之前，不能急于得出定论。比如某位深圳交警在处理交通违章事件中发现，天秤座、处女座、天蝎座的人违章概率更高，这时我们可能得出错误结论，认为交通违章是天秤座、处女座、天蝎座的人性格鲁莽导致的。但其实是因为 9—11 月为北半球低纬度地区生育高峰，天蝎座、天秤座、处女座的人口数占深圳总人口的比例较高，自然在被抓到的违章人数中占比也较大，造成了三个星座的性格是违章概率更高的原因的错误因果判断。

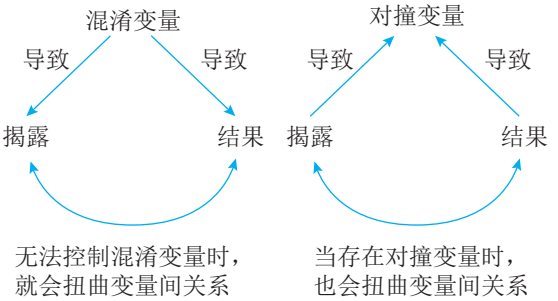

图 1-2　错误确定因果关系的表现形式

2. 小样本错误

小样本错误是一种数据“陷阱”。原因是样本数量太少，即使分析和推理过程是正确的，人们也有可能无法得出正确的结论。经典数学故事“苏格兰黑山羊”问题就属于这种类型。有三个农夫在苏格兰看到一群黑色山羊，第一个农夫说，“原来苏格兰的山羊都是黑色的”；第二个农夫说，“应该说苏格兰有黑山羊”；第三个人思维严谨、善

于思辨，他说，“这只能说明从这个角度看过去，这群山羊有一面是黑色的”。小样本的观察和推理并不具有普适性，更不能作为因果推断的基础。

1.2　随机分配与自然实验

实验方法是因果关系研究的有力工具，依据学科、研究问题和条件的不同，通常采用不同的实验方法。经济社会科学问题的研究依赖于政策或项目的因果效应，因此政策评估领域会广泛应用随机控制实验、自然实验与准实验等因果识别策略。

1.2.1　控制实验

一般是在理想的物理、化学等实验中，实验者有条件直接控制除自变量x以外的全部因素，使其保持不变，单独令x变化，观察因变量y变化的情况。

控制实验是依据研究目的人为地设置一个特定的非自然状态环境，按一定程序改变某些因素或控制条件，并保持其他因素不变，通过观察和分析两个以上变量的变化过程，以测试变量间相互关系和变化规律的研究方法。

例：在研究物体加速度与物体质量的关系时（$F = ma$），控制施加给物体的外力为F不变，改变物体的质量m，观察物体的加速度，借助于打点计时器计算物体的加速度，从而得出结论：物体的质量 m 与物体的加速度 a 成反比。

1.2.2　随机控制实验

随机控制实验是指基于理想随机化的实验模式，可以使实验对象随机地进入实验组或是控制组的实验。比如，通过抛硬币的方式来决定个体样本进入实验组还是控制组，他们能获得多大的实验“处理水平”都是随机决定的，完全独立于实验对象的个体特征，以及其他可能影响实验结果的因素。

这一步骤的目的是控制除了实验处理因素以外的其他因素，但这在实践中很难达成。例如，探究医学研究中某一种新药治疗效果的实验，接受处理的个体有着不同的生活方式，实验前的体质也有差异（即使用动物做实验，如小白鼠，它们之间也存在差异），很难达到控制实验的目的。当然，部分非实验处理因素也可以被控制，有时候，为了消除科研人员心理上的主观意愿造成的偏差，实验会采用“双盲实验”（double blind experiment）的方法，双盲实验是指具体实验个体是被分配到实验组还是控制组，同时对科研人员和实验对象保密。在医学研究中，一般是将实验对象随机地分成实验组和控

制组，其中实验组服用真药，而控制组服用安慰剂，所有实验对象都不知道自己被分在哪一组，以此避免心理作用的干扰。

1.2.3　自然实验与准实验

自然实验是指实验本身遵循随机定律，实验为完全随机的“自然事件”，而非指研究对象为自然现象。我国经济学界常用的自然实验概念，往往对应着准实验，但是自然实验和准实验之间有着细微却显著的区别：准实验无法在实验环境中完全随机地选择并对实验对象进行分组，其中包括人为干预的影响，如实验样本挑选与分组是人为进行的；而自然实验遵从随机定律，是完全随机的“自然事件”。

我国学者常常混淆自然实验和准实验，实际上自然实验与准实验在选取处理组和控制组时有明显差异：自然实验要求由自然发生的事件带来的外生冲击产生实验所需的处理组和控制组，被研究对象接受自然事件的冲击是随机的，其中部分人群不得不接受处理。然而，日常生活中很少出现诸如法律制度变更、自然灾害等外生冲击，也很难找到可视为自然实验的环境，在这种情况下准实验就应运而生了。准实验不是通过随机分配得到处理组和控制组，而是利用观测数据和统计学的方法，打造出类似于随机实验的状态。由于准实验利用了特定的统计方法，因而无法直接比较处理组和控制组，但自然实验和准实验设计本质上是相同的：都是得到可比的处理组和控制组，然后基于随机控制实验方法的思想进行比较，从而得到感兴趣的变量的因果效应。

1. 自然实验案例

自然实验是一种自然发生的实验，即利用自然发生的事件（如政策颁布、地理因素等）的冲击得到类似随机控制实验的随机分配的一种实验。根据自然事件的不同，我们通常可以将自然实验分为两类。

（1）由自然现象变化产生的自然实验。这类自然实验往往将自然的随机事件作为某些变量的工具变量，这些随机变化通常来自生物、气候机制，包括双胞胎、性别、出生日期以及天气事件等。这些由自然变化得到的自然实验往往具有较好的随机性。

（2）由于政策、法律法规等颁布实施使部分群体被迫随机地接受干预的自然实验。这类自然实验通常涉及政策评估，在数据可得的情况下可以用计量方法来估计某些政策的效应。在自然实验中，即使研究者既没有创造也没有指定处理方法，一些随机控制实验中的关键元素也会自己发生。下面是自然实验的一些例子。

例：选票上候选人的顺序是否会影响当选的机会？ 1978—2002 年间，美国加利福尼亚州在州级公职投票选举过程中采用随机分配的方式决定选票上候选人的顺序，这种随机化的分配方式依赖于抽签或摇骰子等方法，在这些方法中候选人抽到第一个或第二

个都是靠自己的“运气”，这其中的“运气”是研究者既无法创造也无法指定的处理方法，是完全随机的，因而是一个典型的自然实验。利用这一事实，何和今井（Ho 和 Imai，2008）试图寻找选票顺序对选举结果的影响，他们的研究结论是选票顺序对主要政党的候选人没有明显的影响，但确实影响了一些小党派的候选人。

例：人类遗传学提供了另一个自然实验的例子。人类遗传学中一项常见而重要的研究设计是测量生物学上的兄弟姐妹的基因和他们的疾病结果。除了基因的性别特征 X 和 Y 染色体，父亲和母亲各自拥有两个不同的遗传基因。父亲和母亲各自在自己的两个基因中遗传一种基因给孩子。母亲的两个基因中，哪一个遗传给孩子本质上是随机的，父亲的两个基因也是如此。因此，当父母生育孩子时，基因就存在一定程度的随机化，不受人为的控制。母亲两个基因中的一个和父亲两个基因中的一个随机结合，从而形成每一个孩子的一对基因，在这样一对生物学上的兄弟姐妹中，遗传标记和疾病之间的关联是某种遗传因素导致疾病的有力证据。

2. 准实验案例

准实验不随机分配实验对象，而是利用一定的标准，在现有的观察数据上构造类似随机控制实验的随机分配（比如，利用我国高考分数线的设置进行断点回归），从而得到可比的处理组和对照组。

周黎安和陈烨（2005）在对农村税费改革问题进行政策评估时，为保证农村税费改革实验的分组具备随机性进行了丰富的实验前测试。此问题研究是一次典型的准实验探究问题。首先，中央政府选择安徽省作为试点可能有其是农业大省的原因，在这个层面上只能近似为随机控制实验；再者，在安徽省选择 21 个试点县进行税费改革，并非通过抓阄或者掷硬币等随机方式选取，可能经过了省政府精心筛选。这个农村税费改革问题探讨是一次典型的准实验分析，因为其样本分组是实验者（政府）按照一定标准抽选的。

1.2.4　自然实验、准实验与随机控制实验：相互作用

自然实验、准实验设计这类非随机控制实验的因果识别策略与随机控制实验一起成为当下实证研究革命中最重要的部分，为致力于寻求现实情境中因果识别的政策评估学者提供了强大的工具。自然实验、准实验研究与随机控制实验研究看似是在两种完全不同的研究路径上，但从实际结果来看，它们之间联系紧密，并互相促进。这不仅体现在因果识别上非随机控制实验方法是以随机控制实验为基准，更体现在以下两方面：一方面，它们之间具有互补关系。沿着随机控制实验的因果识别框架，自然实验、准实验在一定条件下很好地解决了观察数据中由于非随机化而带来的因果识别困扰。另一方面，自然实验、准实验与随机控制实验联系紧密并互相促进。自然实验、准实验研究让随机

控制实验在识别方法上更加完善；随机控制实验使自然实验和准实验研究在研究设计和方法应用上更加成熟，更类似于随机控制实验。

1. 识别策略上更为成熟

随机控制实验的广泛开展使自然实验和准实验的因果识别策略更加成熟。在因果关系的识别方面，自然实验和准实验方法通常是以随机控制实验的研究为基准的。虽然在一定程度上，随机控制实验的有效性被某些学者怀疑，但是对于那些将随机控制实验看作一个特定的实证策略，或者说是一个明确定义的研究基准的非实验研究者来说，他们会更加努力地思考因果识别策略。因此，研究者在应用非随机化的数据和自然实验进行识别时，会更具有创造力。可以看到的是，如今，从研究选题、识别策略来看，非随机化控制实验研究的论文相较于之前，有了很大的提升。莫泽和沃纳（Moser 和 Voena，2012）将 1917 年实施的对敌贸易法案（Trading with the Enemy Act）作为自然实验，利用双重差分法（DID）考察了强制许可制度（compulsory licensing）对技术进步的影响。为了得到稳健的分析，作者进行了包括三重差分（DDD）、工具变量（IV）等在内的一系列稳健性检验。阿莱尼亚等（Alesina 等，2013）在探讨“性别角色的差异来源于传统农业实践方法”的假设时，匹配了 1256 个族群的历史数据和现代国家层面以及个人层面的数据，为了控制反向因果的影响，他们又使用了地理气候条件作为工具变量。可以看出，在自然实验和准实验研究的文献中，研究者选择因果识别这种更加标准的项目评估方法，使结论更为可靠。

2. 实验机制上日趋完善

随着自然实验、准实验设计在非随机控制实验的因果识别方法中的广泛应用，随机控制实验研究者意识到，当随机化方法存在不足的时候，将非随机化实验的因果识别方法应用到随机控制实验中，能够使随机控制实验的研究结果更加稳健。其中，最常用的便是当存在部分依从问题[①]时，将实际处理效应作为分配处理效应的工具变量［多比和弗赖尔（Dobbie 和 Fryer），2011；弗赖尔（Fryer），2013；戴明等（Deming 等），2014］。

此外，学者在随机控制实验研究中将随机控制实验方法与自然实验、准实验方法相结合，以得到更稳健的结果。多比和弗赖尔（Dobbie 和 Fryer，2011）在进行美国哈林儿童区[②]（Harlem Children’s Zone）特许学校（charter schools）与社区项目结合的模式对学生学业表现的影响评估时，利用工具变量对随机抽样的缺陷进行了改进。虽然，当

① 部分依从问题：指被试者没有完全遵从随机处理的要求，只有部分接受了干预或控制。比如实验对象可能由于个人偏好的差异而不接受分配，进而导致部分被分配到实验组的人拒绝接受干预，而控制组的人则可能会想办法得到干预，或者退出实验等。

② 哈林儿童区：针对哈林区贫困儿童和家庭的非营利组织，旨在改善他们的学业表现。

申请特许学校的学生超过特许学校名额时，利用抽签的方式来选择学生能够有效克服非随机的样本问题，但是多比和弗赖尔也意识到简单比较进入特许学校和没进入特许学校的学生的表现是有问题的：在随机抽样的样本中并不包含某些特定的群体（如第一年入学的群体），同时该特许学校在第一年并没有达到超额的情况，因此评估这类群体的影响是很困难的。因此，他们考虑利用工具变量法对随机评估进行补充，并将学生群体的入学年份（cohort year）与其是否住在儿童区的边界的虚拟变量作为工具变量，这是因为，虽然任何人都有资格进入特许学校，但只有住在儿童区内的学生才会被录取，此外，受开学时间以及年龄的限制，部分学生不具有进入特许学校的资格。研究利用这两个工具变量得出了相同的结论，即哈林儿童区特许学校能改善学生的学业表现。

比约克曼和斯文森（Björkman 和 Svensson，2009）在评估社区的监管模式有效性时，基于基准的实验回归模型，进一步利用实验前和实验后的数据通过双重差分模型（DID）来评估干预的影响，同时还使用似不相关回归（seemingly unrelated regression，SUR）模型[①]考察了干预对若干相关结果的影响。此外，杜弗洛等（Duflo 等，2011）在肯尼亚开展随机控制实验考察分班（tracking）的影响，利用断点回归（RD）设计考察了分班对处理组学校中学生的影响。由于在实验中，处理组学校的学生依据前一学年的成绩好坏分为两组，而这一分组恰好使成绩较好的组与较差的组在成绩的 50% 分位数上存在断点，因此可以利用断点回归来考察分班对成绩位于中位数附近的学生的影响。结果表明，分班对于这类学生的影响很小。

1.3　处理效应

1.3.1　处理、干预与处理效应

处理效应的英文 treatment effect 直译为“治疗效果”，该术语最早盛行于医学领域对药物疗效的评估中，但事实上，当处理效应应用于微观计量经济学中时，它与流行病学或者其他社会科学考虑的问题在本质上并无区别。具体来看，其主要应用于政策或者项目实施的效应评估，核心研究要点在于，在排除了所有潜在的易混淆因素后，研究特定的处理变量对结果变量的影响效果。故将处理效应定义为：在经济学的研究领域中进行项目效应评估时，对某一项目或者政策实施后产生效果的度量。在不同的学科背景下，处理变量和结果变量是各种各样的，表 1-1 展示了一些处理变量与结果变量的例子。

① 似不相关回归：各方程的变量之间没有联系，但各方程的扰动项之间存在相关性。如一个学生的不可观测因素同时对数学成绩和英语成绩造成影响，则包含这两个不同解释变量的方程的扰动项相关。如果进行联合估计，可以提高估计效率。

表 1-1　处理变量与结果变量的例子

处理变量	结果变量
锻炼	血压
职业培训	工资
大学教育	终身收入
药物	胆固醇水平
税收政策	工作时长

值得注意的是，在利用微观计量经济学的研究框架来对现实生活中的政策处理效应进行研究时，我们还需要考虑到处理效应中混杂的自我选择因素的干扰。比如在职业培训对工资影响的研究中，如果是否参加职业培训可以由参与研究的个体样本通过自我选择来确定，那么最终工资水平的结果将不仅仅只是受到了是否参加职业培训的影响，还会受到自我选择和样本其他特征的影响，例如，自我选择的因素中包含受教育水平，一般来说，受教育水平高的个体的收入高于受教育水平低的个体的收入。还比如在衡量运动量对胆固醇的影响时，一般来说运动量越大，胆固醇越低，二者为负相关关系，但由于样本中年龄分布的不同，很有可能出现运动量和胆固醇正相关的关系。当我们需要衡量政策的处理效应时，政策评估希望研究的“处理”应理解为“干预”，即一个外生的冲击。对于“干预”的理解，政府对产品价格的调控是一个很好的例子。价格是市场确定的，但政府可能会进行干预，要求实际价格高于或低于市场价格，以了解需求的变化情况。而一个反面的例子是，某人可以选择自愿服用而不是被强制注射某种药物，此时自愿服用该药物不能理解为“干预”，由于存在自我选择，因此不是我们在政策评估中希望研究的“处理”。

1.3.2　“反事实”

对于一个政策冲击而言，在两种潜在处理变量（受到政策处理与未受到政策处理）对应的两种潜在结果变量（受到政策处理时的结果与未受到政策处理时的结果）之间，仅能观察到一种潜在结果，而另一种无法观测到的潜在结果就被称为“反事实”结果（counter-factual），这是处理效应分析中的基本问题。例如在职业教育培训对工资收入影响的例子中，每个样本只有一个真实的工资收入（受到职业培训或未受到职业培训的收入），每个样本对应的另一种情况是现实中无法观测到的，例如，我们不能够通过直接观察知道受到职业培训的人如果没有接受培训收入会是多少，我们也无法知道未受到职业培训的人如果接受了职业培训收入又会是多少，这些不能够直接观测到的另一种结果就是“反事实”。

如前文所述，应用微观计量经济学来进行政策评估研究的目的不是单纯寻找两个现象之间的相关性，而是想要得到一个确切的因果关系。故在进行政策评估时，根据范子英（2018），使用“反事实”方法的基本逻辑是：在假定其他条件相同时，政策实施后与假设政策没有实施的结果之间的差异为政策的处理效应，这种处理效应可以体现出因果关系。例如，1993 年诺贝尔经济学奖得主福格尔（Fogel，1965）通过对比 1890 年真实的美国经济及其所构造的 1890 年没有铁路的美国经济，评估铁路对美国经济发展的影响。

因此，政策评估研究的重点难点在于如何构建“反事实”，比如说医学领域在研究处理效应时不是仅仅分为服药组和未服药组，而是分为服药组、服用安慰剂组和完全不处理组（什么都不服用），服用安慰剂的组相当于是服药组的“反事实”①。本书介绍的各种评估政策处理效应的方法，本质上是构建“反事实”的不同方法。在政策评估中，通常把受到政策处理的样本称为实验组，把未受到政策处理的样本称为控制组。

以一个旨在提高公司专利申请积极性的政策为例。图 1-3 中呈现的是在 t_0 和 t_1 期间实施的政策对公司 A 专利申请的影响，其中实线表示公司 A 被观测到的表现，虚线表示的是公司 A 在没有此项政策实施的情况下的表现，即“反事实”表现。可以看到，在 t_0（即政策实施前）时刻，该公司申请了 12 项专利，而在 t_1（即政策实施后）时刻，该公司申请了 8 项专利，比政策实施前更少。若我们只关注可观测到的直观差异，可能会认为该政策的实施并没能实现其目标。然而，一旦考虑到“反事实”的情况，我们就可以清楚地认识到，如果没有实施这项政策，t_1 时刻公司申请的专利数量将会更少（5 项）。虽然得到的结论似乎是反直觉的，但从“反事实”的角度来看该项政策是成功的，因为最终的效果是积极的（处理效应 α=8−5=3>0）。

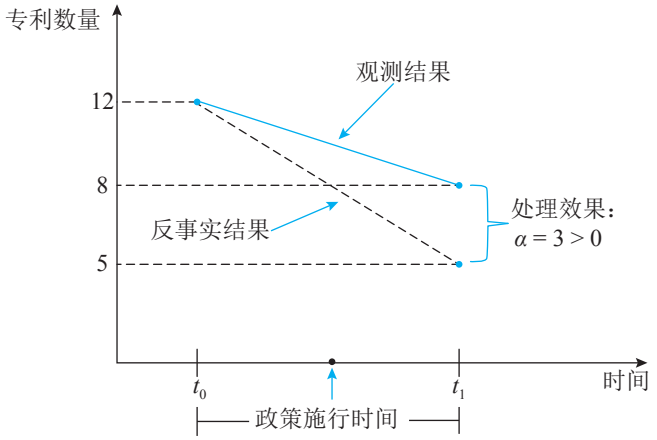

图 1-3　“反事实”因果关系

① 服用安慰剂组作“反事实”和完全不处理组（什么都不服用）作“反事实”相比能够多控制“服用”这个变量，处理效应会更加准确。

1.3.3　平均处理效应

1. 平均处理效应的定义

通过以上部分可知，政策评估的研究目标是衡量和估计处理效应。引入因果推理中的潜在结果模型，用二元处理变量 D_i 表示样本个体 i 是否得到了处理，D_i 取 1 值时表示样本个体 i 得到了处理，取 0 值表示样本个体 i 未被处理。假定我们感兴趣的可观测到的结果变量为 Y_i，对于样本个体 i 来说，其结果变量 Y_i 存在两种状态，即

$$Y_i = \begin{cases} Y_{0i}, D_i = 0 \\ Y_{1i}, D_i = 1 \end{cases} \tag{1-1}$$

其中，Y_{0i} 表示个体 i 未被处理时的潜在结果变量，Y_{1i} 表示个体 i 得到处理时的潜在结果变量，“潜在”的含义是这两个结果是个体 i 本身一直具备的两个结果，只不过我们并不总是能在现实中真实观测到。因为现实中样本个体 i 只能是得到处理或者是未得到处理其中的一种，故分析者只能观察到两个潜在结果变量当中的一个，如果个体 i 得到了处理，则 Y_{1i} 能被观测到，Y_{0i} 不能被观测到；反之，如果个体 i 未被处理，则 Y_{0i} 能被观测到，Y_{1i} 不能被观测到，即 Y_{0i} 和 Y_{1i} 不能被同时观测到。例如，当处理变量为补贴，结果变量为公司投资行为时，我们可以观察到某个获得补贴公司的投资情况，但我们无法观测到这家公司在没有得到补贴时的投资情况。

可以将等式（1-1）的分段函数改写为

$$Y_i = (1 - D_i)Y_{0i} + D_i Y_{1i} = Y_{0i} + D_i(Y_{1i} - Y_{0i}) \tag{1-2}$$

等式（1-2）将潜在结果和真实可观测结果联系了起来，被称为潜在结果模型（potential outcome model，POM）。

此时样本个体 i 得到处理的因果效应或处理效应（treatment effect，TE）为

$$\mathrm{TE}_i = Y_{1i} - Y_{0i} \tag{1-3}$$

TE_i 等于个体 i 被处理时的潜在结果变量 Y_{1i} 和个体 i 未被处理时的潜在结果变量 Y_{0i} 之间的差异。因此，为更好地衡量与估计处理效应，我们需要通过随机控制实验和准实验设计构建实验组和控制组，以控制组的情况来代替无法观测的“反事实”（实验组无法观测到的潜在结果），并以两组样本之间结果变量的差异来估计处理效应。

需要注意的是，考虑到上文中的 TE_i 仅仅涉及单一样本的对比，即实验组和控制组都只包含一个样本，若以此来估计处理效应，可能会由于个体差异导致估计结果有偏，无法评估政策的真实效果。因此在政策评估实践中，为了克服个体差异，我们一般需要扩大样本范围，使实验组和控制组均包含一定数量的样本，并以两个分组中的整体差异来衡量政策的真实效果。同时，引入“个体处理稳定性假设”（stable unit treatment value assumption，SUTVA），即样本个体 i 自身受到处理与否不会对其他个体产生影

响，对个体 i 的处理仅仅影响到个体 i 自身的结果，本书默认该假设一直成立。此时由于平均值具有一些很好的性质，如 $E(Y_{1i}-Y_{0i})=E(Y_{1i})-E(Y_{0i})$，同时平均值在样本的特征值中最能够代表其样本数据特征，因此我们引入平均处理效应（average treatment effect，ATE）的概念。ATE 的计算公式如下：

$$\text{ATE}=E(Y_{1i}-Y_{0i})=E(Y_{1i})-E(Y_{0i}) \tag{1-4}$$

ATE 等于个体被处理时潜在结果变量的均值与同一个体未被处理时潜在结果变量的均值之差，即不管个体是否得到处理，ATE 是对总体中随机抽取的个体求期望得到的结果。因为 ATE 是在总体上进行平均，其中会包括不适合接受处理的个体，有学者对此进行了批评，例如我们在估计职业培训项目的平均处理效应时，并不希望百万富翁——这些不适合接受处理的个体包含在内。也有学者认为此批评是不成立的，因为我们能够通过某种方法将永远不适合接受处理的个体排除在总体之外。在上例中，我们可以通过限定个体培训前收入水平将百万富翁严格地排除在外。在接下来的介绍中，为简便起见，除非特殊需要，将不会再使用下标来指明个体 i。

此外，在估计 ATE 过程中，还涉及两个被广泛关注的相关参数：参与者的平均处理效应（average treatment effect on the treated，ATET）和未参与者的平均处理效应（average treatment effect on the untreated，ATENT）。前者是以接受处理的样本计算的平均处理效应（$D=1$），后者是以未被处理的样本计算的平均处理效应（$D=0$），分别表示如下：

$$\text{ATET}=E(Y_1-Y_0\mid D=1)=E(Y_1\mid D=1)-E(Y_0\mid D=1) \tag{1-5}$$

$$\text{ATENT}=E(Y_1-Y_0\mid D=0)=E(Y_1\mid D=0)-E(Y_0\mid D=0) \tag{1-6}$$

一般情况下，ATET 与 ATENT 是不相等的，ATE 是 ATET 和 ATENT 的一个加权平均值，$\rho(D=1)$ 表示在总体中样本被处理的概率，$\rho(D=0)$ 表示在总体中样本未被处理的概率，根据期望迭代定律（law of iterated expectations，LIE）可得

$$\text{ATE}=\text{ATET}\bullet\rho(D=1)+\text{ATENT}\bullet\rho(D=0) \tag{1-7}$$

ATE、ATET 和 ATENT 这些不同的处理效应参数可以回答不同的问题。依旧是职业教育培训对工资收入影响的例子，假设此时职业培训对工资收入有积极影响，如果想知道职业教育培训对所有人的平均影响，需要估计的处理效应参数就是 ATE，它衡量的是如果所有人均受到职业培训相对于所有人均未受到职业培训的平均工资收入的增量；如果想知道的是职业培训给处理者（接受者）带来了多大程度工资收入的增加，需要估计的处理效应参数就是 ATET；如果想知道的仅是那些未受到职业培训的人如果受到了职业培训，他们的工资收入将会增长多少，需要估计的处理效应参数就是 ATENT。

在政策评估过程中，ATET 与 ATENT 的数值大小也具有重要意义。例如，在评估一项政策时，我们计算得出了 $\text{ATET}=50$，$\text{ATENT}=100$，初步来看，这项政策是成功的，因为 $\text{ATET}=50$，实验组个体平均得到了 50 的处理效应，即该项政策对实验组个

体产生了正向的影响。而 ATENT = 100 则表示，如果控制组个体被处理的话，控制组个体会平均得到 100 的处理效应。 ATENT > ATET 则意味着，如果当初选择这些控制组个体为实验组进行处理，政策的效应会更好，如果政策机构是以政策效应最大化为目标，那么此时政策的目标群体选择就存在一定问题，因为此时实验组个体并未使政策成效最大化。当然，在现实生活中，政策机构的目标很多时候并不是政策效应最大化，而是考虑对弱势群体的帮扶等社会福利方面，所以政策机构会将政策的目标群体定为表现较差的个体。

例如一个小额信贷项目，如果政策机构的目标是通过小额信贷缓解贫困状况，虽然贫困户通过此项目最终增加的收入很有可能是比那些条件更好的个体通过此项目能够增加的收入少，但该项目仍然选择贫困户作为支持的目标群体。该例子也说明，我们在实际进行政策评估时，一定要根据政策的直接目标来选择或者构建控制组（“反事实”组），否则，最终得到的政策评估结果很有可能是错误的。

此外，在政策评估过程中，研究者除了关注样本个体的结果变量 Y 和处理变量 D，还会关注一些协变量，比如样本个体的个人属性如年龄、性别和民族等，以及样本个体的经济社会特征如收入、受教育程度等，用协变量的行向量 $\boldsymbol{x}$ 来表示。这些协变量均不会受到处理变量 D 的影响，但是这些协变量很有可能影响政策处理效应的估计结果，应予以考虑。所以，在考虑行向量 $\boldsymbol{x}$ 的情形下，可以对前文的相关参数进行重新定义：

$$\mathrm{ATE}(\boldsymbol{x}) = E(Y_1 - Y_0 \mid \boldsymbol{x}) \tag{1-8}$$

$$\mathrm{ATET}(\boldsymbol{x}) = E(Y_1 - Y_0 \mid D = 1,\ \boldsymbol{x}) \tag{1-9}$$

$$\mathrm{ATENT}(\boldsymbol{x}) = E(Y_1 - Y_0 \mid D = 0,\ \boldsymbol{x}) \tag{1-10}$$

由于每个样本个体都有独自的 $\boldsymbol{x}$ 值，式（1-8）、式（1-9）、式（1-10）也可被称作“特定个体下的平均处理效应”，进一步可得：

$$\mathrm{ATE} = E_x\left\{\mathrm{ATE}(\boldsymbol{x})\right\} \tag{1-11}$$

$$\mathrm{ATET} = E_x\left\{\mathrm{ATET}(\boldsymbol{x})\right\} \tag{1-12}$$

$$\mathrm{ATENT} = E_x\left\{\mathrm{ATENT}(\boldsymbol{x})\right\} \tag{1-13}$$

式（1-11）、式（1-12）、式（1-13）表明，可以通过对 $\mathrm{ATE}(\boldsymbol{x})$、$\mathrm{ATET}(\boldsymbol{x})$ 和 $\mathrm{ATENT}(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 上求期望得到政策整体的平均处理效应、参与者的平均处理效应和未参与者的平均处理效应。

2. 平均处理效应的估计

1）随机分配下的 ATE 估计

基于上述分析，我们可以知道，ATE 估计的一大难点就是我们无法观测到个体的“反事实”结果。但是在随机分配条件下，即样本是随机抽取的，个体 i 是否受到处理由抛

硬币、电脑随机数分配等随机化方式决定。我们可以用处理样本结果变量的均值与未处理样本结果变量的均值之差（difference-in-means，DIM）来准确、一致地估计 ATE。因为在随机分配条件下，$(Y_1;\ Y_0)$ 独立于 D，即样本选择过程以及由此形成的 D 与每一种潜在结果都是不相关的，我们把这称之为独立性假设（independence assumption，IA），数学上表示为

$$(Y_1;\ Y_0) \perp D \tag{1-14}$$

其中，符号 $\perp$ 表示相互独立。同时，独立性假设的满足也意味着：

$$E(Y_0 \mid D=1) = E(Y_0 \mid D=0) = E(Y_0) \tag{1-15}$$

$$E(Y_1 \mid D=1) = E(Y_1 \mid D=0) = E(Y_1) \tag{1-16}$$

因此，在随机分配条件下，结合潜在结果模型（POM），即式（1-2），我们可得

$$\begin{aligned} \text{DIM} &= E(Y \mid D=1) - E(Y \mid D=0) = E(Y_1 \mid D=1) - E(Y_0 \mid D=0) \\ &= E(Y_1) - E(Y_0) = \text{ATE} \end{aligned} \tag{1-17}$$

也可知道在此种情况下：

$$\begin{aligned} \text{ATET} &= E(Y_1 - Y_0 \mid D=1) = E(Y_1 - Y_0 \mid D=0) = \text{ATENT} \\ &= E(Y_1 - Y_0) = \text{ATE} \end{aligned} \tag{1-18}$$

DIM 的样本估计量形式为

$$\widehat{\text{DIM}} = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{1,\ i} - \frac{1}{N_0}\sum_{i=1}^{N_0} Y_{0,\ i} \tag{1-19}$$

其中，N_1 是处理样本总量，N_0 是未处理样本总量。

2）非随机分配下的 ATE 估计

在实际的政策实施过程中，人们几乎不会随机选择处理个体，也不会随机选择非处理个体。一方面，个体会自我选择是否参与政策项目，例如企业在自我选择是否接受一项旨在提高固定投资的工业激励时，会将其接受此项政策后的收益与其参与项目需要付出的成本作比较，这个自我选择过程并不是随机的。另一方面，现实中的政策项目大多会根据其想要达到的目标来选择政策的处理个体，例如一项政策的目标是帮助受教育较少的人，那么处理的个体一定是受教育较少的人，即受益的个体并不是随机选择的。

当处理个体和非处理个体是非随机选择的，也就是非随机分配的时候，DIM 估计量不再能够准确估计 ATE。此时，独立性假设不再满足，结合式 (1-2) 和期望运算，可得：

$$\begin{aligned} \text{DIM} &= E(Y \mid D=1) - E(Y \mid D=0) = E(Y_1 \mid D=1) - E(Y_0 \mid D=0) + E(Y_0 \mid D=1) - E(Y_0 \mid D=1) \\ &= E(Y_0 \mid D=1) - E(Y_0 \mid D=0) + \text{ATET} \end{aligned} \tag{1-20}$$

$$\begin{aligned} \text{DIM} &= E(Y \mid D=1) - E(Y \mid D=0) = E(Y_1 \mid D=1) - E(Y_0 \mid D=0) + E(Y_1 \mid D=0) - E(Y_1 \mid D=0) \\ &= E(Y_1 \mid D=1) - E(Y_1 \mid D=0) + \text{ATENT} \end{aligned} \tag{1-21}$$

结合式（1-7），可得

$$\begin{aligned}&\mathrm{ATE}=\mathrm{ATET}\bullet\rho(D=1)+\mathrm{ATENT}\bullet\rho(D=0)\\&=\rho(D=1)\bullet\left[\mathrm{DIM}-E(Y_0\mid D=1)+E(Y_0\mid D=0)\right]\\&+\rho(D=0)\bullet\left[\mathrm{DIM}-E(Y_1\mid D=1)+E(Y_1\mid D=0)\right]\\&=\mathrm{DIM}-\begin{bmatrix}\rho(D=1)\bullet(E(Y_0\mid D=1)-E(Y_0\mid D=0))\\+\rho(D=0)\bullet(E(Y_1\mid D=1)-E(Y_1\mid D=0))\end{bmatrix}\end{aligned}\tag{1-22}$$

显然，我们无法从现实观测中得到式（1-22）中的 $E(Y_0\mid D=1)$ 和 $E(Y_1\mid D=0)$，此时使用根据观测数据计算得到的 DIM 估计 ATE 是不准确的。从式（1-22）也可以看出，只有在满足 $E(Y_0\mid D=1)=E(Y_0\mid D=0)$ 和 $E(Y_1\mid D=1)=E(Y_1\mid D=0)$ 这两个条件的情况下，即 Y_0 和 Y_1 不依赖于 D 的情况下，DIM 才可以准确估计 ATE。因此，在非随机分配条件下，需要一些附加条件来估算 ATE、ATET 和 ATENT，具体详见 1.4 节。

1.3.4　Excel专栏：平均处理效应ATE的估计

Excel 演示文件

本部分基于 Excel 操作，展示在穿透视角和真实视角下平均处理效应 ATE 的计算方法。本案例研究技能培训项目对工资收入的影响，我们假设决定收入（Y）的因素主要有智商（X_1）、家庭背景（X_2）以及是否接受技能培训（D）。本例使用的数据是人工生成的，共 200 个样本，令前 100 个样本作为实验组接受培训，后 100 个样本作为控制组不接受培训[①]，并根据穿透视角下设定的关系式生成数据，具体见表 1-2。表 1-3 展示了真实视角下的相关数据。

表 1-2　关于培训与收入的研究案例——穿透视角

序号	智商（X_1）	家庭背景（X_2）	培训（D）	Y_0	Y_1	Y
90	2	1	1	26.09	39.31	39.31
91	1	1	1	25.46	33.56	33.56
92	1	3	1	31.31	38.97	38.97
93	2	2	1	33.98	43.16	43.16
94	4	4	1	50.58	63.21	63.21
95	3	3	1	41.78	53.75	53.75
96	2	2	1	35.80	44.46	44.46
97	3	1	1	31.76	44.17	44.17
98	2	3	1	37.42	41.18	41.18
99	4	4	1	48.90	61.27	61.27

① 注意，此时为非随机分配，个体是否进入实验组并不是通过随机分配的方式进行的。

续表

序号	智商（X_1）	家庭背景（X_2）	培训（D）	Y_0	Y_1	Y
100	1	1	1	22.53	26.76	26.76
101	1	2	0	34.23	35.25	34.23
102	1	2	0	21.78	33.16	21.78
103	3	1	0	34.90	42.06	34.90
104	2	1	0	32.04	49.26	32.04
105	2	4	0	35.75	48.71	35.75
106	4	3	0	46.41	55.67	46.41
107	2	1	0	30.33	37.84	30.33
108	2	2	0	33.58	43.21	33.58
109	3	1	0	35.26	46.46	35.26
110	2	3	0	39.98	46.10	39.98

表 1-3　关于培训与收入的研究案例——真实视角

序号	智商（X_1）	家庭背景（X_2）	培训（D）	Y_0	Y_1	Y
90	2	1	1		39.31	39.31
91	1	1	1		33.56	33.56
92	1	3	1		38.97	38.97
93	2	2	1		43.16	43.16
94	4	4	1		63.21	63.21
95	3	3	1		53.75	53.75
96	2	2	1		44.46	44.46
97	3	1	1		44.17	44.17
98	2	3	1		41.18	41.18
99	4	4	1		61.27	61.27
100	1	1	1		26.76	26.76
101	1	2	0	34.23		34.23
102	1	2	0	21.78		21.78
103	3	1	0	34.90		34.90
104	2	1	0	32.04		32.04
105	2	4	0	35.75		35.75
106	4	3	0	46.41		46.41
107	2	1	0	30.33		30.33
108	2	2	0	33.58		33.58
109	3	1	0	35.26		35.26
110	2	3	0	39.98		39.98

1）穿透视角下平均处理效应 ATE 的计算

在穿透视角下，我们已知上述收入 Y 与 X_1、X_2 和 D 各变量之间的准确影响关系，我们将这个潜在关系假定为：

$$Y = 10 + 10 \bullet D + 6 \bullet X_1 + 5 \bullet X_2 + \varepsilon$$

其中，ε 为随机误差项，预设的 ATE 值为 10。因而结合 $D=0$，$Y=Y_0$ 和 $D=1$，$Y=Y_1$ 两等式，通过此关系式生成了 200 个样本的 Y_0 与 Y_1 数值，其中 4 ～ 103 行为前 100 数值，104 ～ 203 行为后 100 数值。数据生成过程以 Y_1 为例，$Y_1 = 10 + 10 * \$E\$2 + 6 * C4 + 5 * D4 + \text{NORMINV(RAND(), 0, 4)}$，其中 E 列为是否接受培训，C 列为智商，D 列为家庭背景，G 列为 Y_1 的数值，F 列为 Y_0 的数值。运用 NORMINV(RAND()，0，4)[①] 生成误差项。部分具体数据情况详见表 1-2，其中 $D=0$ 时 Y_1 取值和 $D=1$ 时 Y_0 的取值为结果变量的“反事实”，是真实视角中无法观测到的数据。

以表 1-2 的数据为基础，在穿透视角下我们能够完全看到实验组以及控制组中各个样本的“反事实”情况，从而能够通过前文所述公式得出 ATET、ATENT 以及 ATE 的准确数值，利用 Excel 中的 AVERAGE 函数，我们能够经过简单计算得出在该数据样本中：

$$\begin{aligned}\text{ATET} &= E(Y_1 - Y_0 \mid D=1) \\ &= \text{AVERAGE}(G4:G103) - \text{AVERAGE}(F4:F103) \\ &= 10.36\end{aligned}$$

$$\begin{aligned}\text{ATENT} &= E(Y_1 - Y_0 \mid D=0) \\ &= \text{AVERAGE}(G104:G203) - \text{AVERAGE}(F104:F203) \\ &= 10.04\end{aligned}$$

$$\begin{aligned}\text{ATE} &= \text{ATET} \bullet \rho(D=1) + \text{ATENT} \bullet \rho(D=0) \\ &= 0.5 \times \text{ATET} + 0.5 \times \text{ATENT} \\ &= 10.20\end{aligned}$$

通过上述计算结果我们能够看到，在穿透视角下计算得出的 ATE = 10.20，与本例中预先设定的 10 存在差别，该差别来源于数据生成时，同一个体 Y_1 和 Y_0 包含的随机误差项 ε 数值不同，即随机误差项的存在。

此时，运用 AVERAGE 函数简单计算 DIM 为

$$\begin{aligned}\text{DIM} &= E(Y \mid D=1) - E(Y \mid D=0) \\ &= \text{AVERAGE}(G4:G103) - \text{AVERAGE}(F104:F203) \\ &= 10.33\end{aligned}$$

DIM 估计量与 ATE 相差较大，表明此时 DIM 估计量无法准确估计 ATE 。

2）真实视角下平均处理效应 ATE 的估计

以表 1-3 关于培训与收入的研究案例——真实视角为基础，在真实视角下，我们并

① NORMINV(RAND()，0，4) 返回的随机数为均值为 0，标准差为 4 的正态分布中 RAND() 的分位数，RAND() 为 [0，1] 的随机数。

不知道 Y 与 X_1、X_2 和 D 各变量之间的准确影响关系，此时假设我们在现实中获得的关于个体智商（X_1）、家庭背景（X_2）、收入（Y）、是否参加培训（D）的数据与我们在“穿透视角”下生成的数据一致。

一方面，我们无法获得实验组以及控制组中各个样本的“反事实”情况，也就不能通过上述公式得出 ATET、ATENT 以及 ATE 的准确数值。另一方面，根据前文计算的数据可知，DIM 与 ATE 不相等，这表明能够在真实视角下计算得到的 DIM 也无法准确估计 ATE。因此，我们需要利用各类微观计量方法来近似得出实验组以及控制组中各个样本的“反事实”情况，在此基础上估计得出 ATE。

1.4　可观测选择与不可观测选择

1.4.1　可观测选择与不可观测选择定义

一般来说，影响个体非随机分配的因素会具有可观测的特性或不可观测的特性，由此引出可观测选择和不可观测选择两个定义。可观测选择是指样本选择完全是基于可观测因素进行的，即研究者知晓并能够精准地观察到影响个体自我选择或者政策机构选择的因素，此时只要控制可观测因素便能够识别出政策的真实效果。不可观测选择是指样本选择还受到与潜在结果相关的不可观测或者难以观测因素的影响：一种不可观测因素来自信息缺失，即数据不可得；另一种不可观测因素则是在信息充足的情况下也难以观测的变量，例如企业道德风险、抗风险能力等。

1.4.2　可观测选择

基于可观测因素进行选择时，个体 i 对 D 的选择完全取决于可观测的向量 $\boldsymbol{x}$，即使个体是非随机分配的，但对影响非随机分配的可观测因素 $\boldsymbol{x}$ 进行控制，ATE 也能被识别出来。

此时引入罗森鲍姆和鲁宾（Rosenbaum 和 Rubin，1983）提出的条件独立假设（conditional independence assumption，CIA），即在已知 $\boldsymbol{x}$ 的条件下，$(Y_0;\ Y_1)$ 依概率独立于 D，记为

$$(Y_0;\ Y_1) \perp D \mid \boldsymbol{x} \tag{1-23}$$

该假设意味着给定 $\boldsymbol{x}$，潜在结果和处置状态无关，$(Y_0;\ Y_1)$ 在实验组与控制组的分布完全一样，记为

$$F(Y_0, Y_1 \mid \boldsymbol{x}, D=1) = F(Y_0, Y_1 \mid \boldsymbol{x}, D=0) \tag{1-24}$$

其中，$F(\bullet)$ 为分布函数，该假设也意味着一旦影响样本选择的因素被控制住，就可以认为是满足随机条件了。但在考察平均处理效应时，CIA 假设过于严格，一般只需更弱的条件均值独立假设（conditional mean independence，CMI），表述为

$$E(Y_0 \mid \boldsymbol{x}, D=1) = E(Y_0 \mid \boldsymbol{x}, D=0) = E(Y_0 \mid \boldsymbol{x}) \tag{1-25}$$

$$E(Y_1 \mid \boldsymbol{x}, D=1) = E(Y_1 \mid \boldsymbol{x}, D=0) = E(Y_1 \mid \boldsymbol{x}) \tag{1-26}$$

该假设意味着，在给定 $\boldsymbol{x}$ 的情况下，Y_0 与 Y_1 的均值都独立于 D，即实验组和控制组的平均潜在结果相同。CMI 条件是对 ATE 、 ATET 、 ATENT 进行一致估计的基础。

因而，结合潜在结果模型（POM），即式（1-2）可得

$$\begin{aligned} E(Y \mid \boldsymbol{x}, D) &= E(Y_0 \mid \boldsymbol{x}, D) + D\left[E(Y_1 \mid \boldsymbol{x}, D) - E(Y_0 \mid \boldsymbol{x}, D)\right] \\ &= E(Y_0 \mid \boldsymbol{x}) + D\left[E(Y_1 \mid \boldsymbol{x}) - E(Y_0 \mid \boldsymbol{x})\right] \end{aligned} \tag{1-27}$$

式（1-27）在 $D=1$ 和 $D=0$ 情形可表述为：

$$\text{当} D=1\text{：}\ E(Y \mid \boldsymbol{x}, D=1) = E(Y_1 \mid \boldsymbol{x}) \tag{1-28}$$

$$\text{当} D=0\text{：}\ E(Y \mid \boldsymbol{x}, D=0) = E(Y_0 \mid \boldsymbol{x}) \tag{1-29}$$

式（1-28）和式（1-29）相减可得

$$E(Y \mid \boldsymbol{x}, D=1) - E(Y \mid \boldsymbol{x}, D=0) = E(Y_1 \mid \boldsymbol{x}) - E(Y_0 \mid \boldsymbol{x}) = \text{ATE}(\boldsymbol{x}) \tag{1-30}$$

整理可得：

$$\text{ATE}(\boldsymbol{x}) = E(Y \mid \boldsymbol{x}, D=1) - E(Y \mid \boldsymbol{x}, D=0) \tag{1-31}$$

式（1-31）右边是所有可观测变量，表明 $\text{ATE}(\boldsymbol{x})$ 能够被识别出来且无偏误。

根据式（1-11）、式（1-12）和式（1-13）可得

$$\text{ATE} = E_x\left\{\text{ATE}(\boldsymbol{x})\right\} = E_x\left\{E(Y \mid \boldsymbol{x}, D=1) - E(Y \mid \boldsymbol{x}, D=0)\right\} \tag{1-32}$$

$$\text{ATET} = E_x\left\{\text{ATE}(\boldsymbol{x}) \mid D=1\right\} \tag{1-33}$$

$$\text{ATENT} = E_x\left\{\text{ATE}(\boldsymbol{x}) \mid D=0\right\} \tag{1-34}$$

式（1-32）、式（1-33）和式（1-34）表明平均处理效应 ATE 、参与者的平均处理效应 ATET 和未参与者的平均处理效应 ATENT 能够被识别出来。

在可观测选择下，本书介绍的进行处理效应估计的计量经济学方法主要包括以下两种：

（1）匹配（matching），详见本书第 2 章。匹配的核心思想是基于 CIA 假设和重叠假设，认为可观测因素相同或者近似的样本对政策的反应是相同或近似的，因此可以通过实验组和控制组可观测因素特征的匹配，为每一个实验组个体或者控制组个体构建“反事实”。

（2）清晰断点回归（sharp RD），详见本书第 5 章。因为在临界值的邻域内是近似满足随机分配的自然实验，故能够通过比较临界值左右个体结果变量的期望得到政策

处理效应。清晰断点回归指在变量可观测前提下，个体是否接受处理由关键变量 x 是否超过临界值来决定，在断点处，断点完全决定干预分配的状态，即当个体的关键变量 x 的值大于临界值时，个体得到政策处理；反之小于临界值时，个体未得到处理，小于临界值的个体可以作为一个很好的控制组反映实验组个体没有得到政策处理时的状况。

1.4.3　不可观测选择

当个体选择不仅受可观测因素影响，还受到与潜在结果相关的不可观测因素影响时，CIA 假设或者 CMI 假设不再被满足，DIM 不再能够识别出 ATE。在此种情况下：

$$E(Y_0 \mid \boldsymbol{x},\ D) \neq E(Y_0 \mid \boldsymbol{x}) \tag{1-35}$$

$$E(Y_1 \mid \boldsymbol{x},\ D) \neq E(Y_1 \mid \boldsymbol{x}) \tag{1-36}$$

因而结合潜在结果模型（POM），即式（1-2）可得：

$$\begin{aligned} E(Y \mid \boldsymbol{x},\ D=1) - E(Y \mid \boldsymbol{x},\ D=0) &= E(Y_1 \mid \boldsymbol{x},\ D=1) - E(Y_0 \mid \boldsymbol{x},\ D=0) \\ &+ \left[E(Y_0 \mid \boldsymbol{x},\ D=1) - E(Y_0 \mid \boldsymbol{x},\ D=1)\right] \\ &= \left[E(Y_0 \mid \boldsymbol{x},\ D=1) - E(Y_0 \mid \boldsymbol{x},\ D=0)\right] + \mathrm{ATET}(\boldsymbol{x}) \end{aligned} \tag{1-37}$$

式（1-37）表明，当不可观测因素显著地影响非随机选择时，仅依赖可观测因素 $\boldsymbol{x}$ 估计 ATE 的方法将不再可行，因为即使是在能被 $\boldsymbol{x}$ 识别的个体中，$E(Y_0 \mid \boldsymbol{x},\ D=1)$ 也是不可观测的，运用 DIM 估计得到的处理效应是有偏误的。

简单来讲，在回归方程中，不可观测因素 a 属于误差项的一部分，不可观测因素 a 与处理变量 D 相关，导致处理变量 D 与误差项相关，此时二元处理变量 D 为内生变量，导致 OLS 回归产生有偏估计量。所以在不可观测选择下，本书中介绍的处理效应 ATE 的估计方法主要包括以下四种：

（1）工具变量法（IV），详见本书第 4 章。运用工具变量法的前提是至少找到一个与处理变量 D 相关，同时与结果变量 Y 不相关的工具变量，在政策评估问题中，找到满足条件的工具变量存在一定难度。此外，如果个体对于政策的反应不同，即个体的处理效应具有异质性，只有该异质性对个体是否得到处理并无影响时，工具变量才能够一致识别 ATE，但在实际操作中，此要求过于严格，许多研究者会选择忽略个体处理效应的异质性。反之，当个体处理效应的异质性会影响到其是否得到处理时，工具变量最终识别出来的是顺从者的平均处理效应，即局部平均处理效应（local average treatment effect，LATE）。

（2）双重差分法（DID），详见本书第 3 章。双重差分法是除工具变量法之外的一种被广泛使用的解决内生选择问题的方法，使用双重差分法的前提是实验组个体和控制组个体处理前后的观测值均可得且不可观测因素不随时间变化。

（3）倾向得分匹配－双重差分法（propensity score matching difference in difference, PSM-DID），详见本书第 6 章。PSM-DID 方法是倾向得分匹配与双重差分两种方法的结合，使用的前提为 CMI 假设与重叠假设成立、不可观测因素不随时间改变且具有面板数据格式，此方法能够减小选择性偏差，增强结果的可信度。

（4）模糊断点回归（fuzzy RD），详见本书第 5 章。模糊断点回归是指由于存在不完全依从者，是否接受处理不完全是由关键变量 x 来决定。因为即使关键变量 x 大于临界值，个体也不一定得到处理，即处理变量 D 还受到一些不可观测因素的影响，此时由于在断点附近近似存在随机分组，所以仍旧可以一致地估计平均处理效应。

1.5　选择性偏差

1.5.1　选择性偏差的定义

“选择性偏差”（selection bias）指在研究过程中，样本选择的非随机性而导致的结论偏差，包括自选择偏差和样本选择偏差。自选择偏差（self-selection bias）是实验组的非随机分配导致的，即是否进入实验组是个体选择的结果，个体选择的过程会使处理效应的估计结果产生偏差，例如探究职业培训对工资收入的影响时，个体是否接受职业培训会受到个体受教育程度的影响，同时，受教育程度高的人工资收入一般也比较高，最终导致处理效应的估计有偏。样本选择偏差（sample-selection bias）是指非随机选择的样本不能够反映总体的某些特征，从而使结果估计量产生偏差。例如探究农村教育对农村居民收入的影响时，能力突出的农村居民在城市化进程中通过升学、参军等途径进入城市体系，由于无法观测到这些已经成为城市居民的原农村居民的相关数据，农村居民样本将无法包含这些原农村居民。所以，农村居民的样本就是一个选择性的样本，用此样本估计得到的结果会低估农村教育对农村居民收入的影响。本书中介绍的大部分评估政策处理效应的方法的应用目的就是尽可能地降低选择性偏差，从而阐释处理变量 D 对结果变量 Y 的“纯”影响。

1.5.2　选择性偏差的刻画

根据计量知识和统计理论可知[①]，式（1-19）中的 DIM 样本估计量就是通过对式（1-38）进行 OLS 估计得到的系数 α：

① 具体推导参考赵西亮（2017）。

$$Y=\mu+\alpha D+\varepsilon \tag{1-38}$$

其中，μ为截距项，ε为随机误差项。在随机分配条件下，根据式（1-38）可得

$$E(Y\mid D=1)=\mu+\alpha \tag{1-39}$$

$$E(Y\mid D=0)=\mu \tag{1-40}$$

$$\alpha=E(Y\mid D=1)-E(Y\mid D=0)=\mathrm{DIM} \tag{1-41}$$

而在非随机分配下，若是否进行处理的选择是由因素 x 决定的，式（1-38）可变形为

$$Y=\mu_a+\alpha_a D+\beta x+\varepsilon_a \tag{1-42}$$

其中，为与式（1-38）中的 μ、α 和 ε 相区别，式（1-42）添加了脚标。

式（1-42）也可写成

$$Y'=\mu_a+\alpha_a D+\varepsilon_a \tag{1-43}$$

其中，$Y'=Y-\beta x$，式（1-42）与式（1-38）的回归形式相同，则可以得到

$$\alpha_a=E(Y'\mid D=1)-E(Y'\mid D=0) \tag{1-44}$$

进一步可以得到：

$$\begin{aligned}\alpha_a&=E(Y-\beta x\mid D=1)-E(Y-\beta x\mid D=0)\\&=\left[E(Y\mid D=1)-E(Y\mid D=0)\right]-\beta\left[E(x\mid D=1)-E(x\mid D=0)\right]\end{aligned} \tag{1-45}$$

式（1-45）也可以写成

$$\alpha_a=\mathrm{DIM}-\mathrm{BIAS} \tag{1-46}$$

其中，$\mathrm{DIM}=E(Y\mid D=1)-E(Y\mid D=0)=\alpha$，$\mathrm{BIAS}=\beta[E(x\mid D=1)-E(x\mid D=0)]$。

结合以上内容可知，在探究 D 对 Y 的影响时，选择因素 x 导致非随机分配下的结果与随机分配下的结果不一致，而选择性偏差就是这两种情况下 D 对 Y 影响的差：

$$\mathrm{BIAS}=\alpha-\alpha_a \tag{1-47}$$

因而，在一个存在非随机分配的政策效果评估中，如果认为其是随机分配的，最终估计出来的政策影响就会存在上述偏差 BIAS，在此处就是

$$\mathrm{BIAS}=\beta\left[E(x\mid D=1)-E(x\mid D=0)\right] \tag{1-48}$$

其中，β 用于衡量 x 对结果 Y 产生的影响，而 $\left[E(x\mid D=1)-E(x\mid D=0)\right]$ 用于衡量实验组和控制组个体在 x 上是否平衡，当 β 和 $E(x\mid D=1)-E(x\mid D=0)$ 同时不等于 0 时，会出现选择偏差。

在一项政策研究中，如果处理组和控制组在 x 上严重不平衡，但是研究者并不予以控制，使用式（1-41）而不是式（1-45）来估计政策效果，最终会得出具有偏误的政策评价结果。根据式（1-48）也可以看出，在 $\beta<0$，$E(x\mid D=1)<E(x\mid D=0)$ 的情况下，

选择性偏差 BIAS < 0，从而 DIM 会高估处理效应。下面运用一个简单的例子说明。

对一个旨在提高个人理解能力的培训项目进行评估，我们想要知道该培训项目对个人理解一篇论文的影响。有两组人群分别作为实验组（组 1）和控制组（组 0），只有实验组接受培训。由于选拔机制，组 1 都是年轻人，组 0 都是老人，两组平均年龄相差 40 岁。我们通过考试的形式监测培训结果，最终发现组 1 考试的平均分为 80 分，而组 0 考试的平均分只有 30 分，因此在本例子中，$\text{DIM} = 80 - 30 = 50$，表明该培训项目是能够提升人们理解能力的，但是这个结论明显是错误的。因为本例中，实验组与控制组在年龄上是严重不平衡的，而年龄显然会对考试得分产生显著负向影响，假设 $\beta = -1.5$，可计算得到

$$\alpha_a = \text{DIM} - \beta\left[E(x \mid D=1) - E(x \mid D=0)\right] = (80-30) - (-1.5) \times (-40) = -10$$

这表明，一旦处理组和控制组在年龄上是平衡的，该项目的真实效果很有可能是负向的，即该培训项目很有可能并不能够提升人们的理解能力。此例中，选择偏差 $\text{BIAS} = 60$。

1.5.3　选择性偏差的表现形式

选择性偏差的第一种表示方法 B_1：

$$\underbrace{E(Y_1 \mid D=1) - E(Y_0 \mid D=0)}_{\text{参与者与未参与者的平均差异DIM}} = \underbrace{E(Y_1 \mid D=1) - E(Y_0 \mid D=1)}_{\text{参与者平均处理效应ATET}} + \underbrace{E(Y_0 \mid D=1) - E(Y_0 \mid D=0)}_{\text{选择偏差}B_1} \tag{1-49}$$

此处 $B_1 = E(Y_0 \mid D=1) - E(Y_0 \mid D=0)$ 就是选择性偏差。本质上来讲，因为 ATET 的估计式中 $E(Y_0 \mid D=1)$ 是不可观测的，所以我们用可观测的 $E(Y_0 \mid D=0)$ 代替它，也就是使用 DIM 估计 ATET 时，如果 $E(Y_0 \mid D=1) > E(Y_0 \mid D=0)$，DIM 会高估 ATET，如果 $E(Y_0 \mid D=1) < E(Y_0 \mid D=0)$，DIM 会低估 ATET，只有当 $E(Y_0 \mid D=1) = E(Y_0 \mid D=0)$ 时，即处理组 Y_0 的均值与控制组 Y_0 的均值相等时，DIM 才是 ATET 的准确估计。

选择性偏差的第二种表示方法 B_2：

$$\underbrace{E(Y_1 \mid D=1) - E(Y_0 \mid D=0)}_{\text{参与者与未参与者的平均差异DIM}} = \underbrace{E(Y_1 \mid D=1) - E(Y_1 \mid D=0)}_{\text{选择偏差}B_2} + \underbrace{E(Y_1 \mid D=0) - E(Y_0 \mid D=0)}_{\text{未参与者平均处理效应ATENT}} \tag{1-50}$$

此处 $B_2 = E(Y_1 \mid D=1) - E(Y_1 \mid D=0)$ 也是选择性偏差。同样，因为 ATENT 的估计式中 $E(Y_1 \mid D=0)$ 是不可观测的，所以当我们用可观测的 $E(Y_1 \mid D=1)$ 代替它时，也就是使用 DIM 估计 ATENT 时，如果 $E(Y_1 \mid D=1) > E(Y_1 \mid D=0)$，DIM 会高估 ATENT，如果 $E(Y_1 \mid D=1) < E(Y_1 \mid D=0)$，DIM 会低估 ATENT，只有当 $E(Y_1 \mid D=1) = E(Y_1 \mid D=0)$ 时，即

处理组 Y_1 的均值与控制组 Y_1 的均值相等时，DIM 才是 ATENT 的准确估计。

结合式（1-49）和式（1-50），进一步可得

$$\mathrm{DIM}=\mathrm{ATET}+B_1 \tag{1-51}$$

$$\mathrm{DIM}=\mathrm{ATENT}+B_2 \tag{1-52}$$

$$\mathrm{DIM}=\frac{1}{2}(\mathrm{ATET}+\mathrm{ATENT})+\frac{1}{2}(B_1+B_2) \tag{1-53}$$

此时，$B_1+B_2=B$ 为整体偏差。

1.5.4　选择性偏差的分解

赫克曼等（Heckman 等，1998）提出的选择性偏差的分解方式指出，选择性偏差 B_1（或者 B_2）可以被分解成以下三种可解释的偏差：

$$B_1=B_A+B_B+B_C \tag{1-54}$$

首先要提供一些定义：（1）定义 $B_1(\boldsymbol{x})$ 为 B_1 在 $\boldsymbol{x}$ 条件下的偏差；（2）定义 $S_{1x}=\{\boldsymbol{x}:f(\boldsymbol{x}\mid D=1)>0\}$ 和 $S_{0x}=\{\boldsymbol{x}:f(\boldsymbol{x}\mid D=0)>0\}$ 为 $\boldsymbol{x}$ 在 $D=1$ 和 $D=0$ 条件下的子集，其中 $f(\boldsymbol{x}\mid D)$ 指 $\boldsymbol{x}$ 在 D 上的条件密度函数；（3）定义 $S_x=S_{1x}\cap S_{0x}$ 为重叠域；（4）定义 $\bar{y}_{00}=E(Y_0\mid \boldsymbol{x},\ D=0)$ 和 $\bar{y}_{01}=E(Y_0\mid \boldsymbol{x},\ D=1)$。结合这些定义，以 B_1 为例，B_1 可进一步分解为

$$\begin{aligned}
B_1&=E(Y_0\mid D=1)-E(Y_0\mid D=0)\\
&=\int_{S_{1x}}\bar{y}_{01}f(\boldsymbol{x},\ D=1)\mathrm{d}\boldsymbol{x}-\int_{S_{0x}}\bar{y}_{00}f(\boldsymbol{x},\ D=0)\mathrm{d}\boldsymbol{x}\\
&=\int_{S_{1x}}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_{0x}}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=0)\\
&=\int_{S_{1x}-S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_{0x}-S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=0)+\\
&\quad\int_{S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=0)+\\
&\quad\left[\int_{S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=1)\right]\\
&=\underbrace{\int_{S_{1x}-S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_{0x}-S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=0)+}_{B_A}\\
&\quad\underbrace{\int_{S_x}\bar{y}_{00}\left[\mathrm{d}F(\boldsymbol{x},\ D=1)-\mathrm{d}F(\boldsymbol{x},\ D=0)\right]+}_{B_B}
\end{aligned}$$

$$\underbrace{\int_{S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=1)}_{B_C} \tag{1-55}$$

分解后的三种偏差分别表述为

$$B_A=\int_{S_{1x}-S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_{0x}-S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=0) \tag{1-56}$$

$$B_B=\int_{S_x}\bar{y}_{00}\left[\mathrm{d}F(\boldsymbol{x},\ D=1)-\mathrm{d}F(\boldsymbol{x},\ D=0)\right] \tag{1-57}$$

$$B_C=\int_{S_x}\bar{y}_{01}\mathrm{d}F(\boldsymbol{x},\ D=1)-\int_{S_x}\bar{y}_{00}\mathrm{d}F(\boldsymbol{x},\ D=1) \tag{1-58}$$

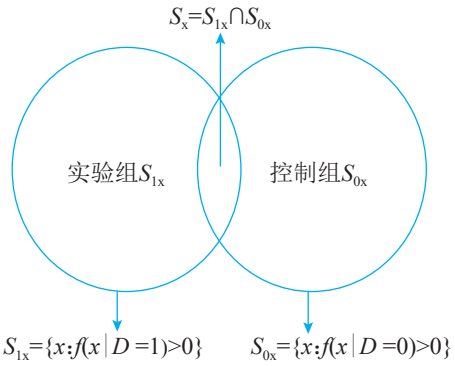

图 1-4　协变量的行向量 $\boldsymbol{x}$ 的分布情况

B_A：偏差应归于弱重叠。如图 1-4 所示，只要 $\{S_{1x}-S_x\}$ 和 $\{S_{0x}-S_x\}$ 非空，偏差 B_A 就存在，这意味着实验组和控制组的一些 $\boldsymbol{x}$ 的子集没有交集，即存在某种具有特殊特征 $\boldsymbol{x}$ 的个体全部都在实验组或全部都在控制组，这些个体是不能匹配的。如果 $\boldsymbol{x}$ 集合的范围不是完全重合的，就会存在弱重叠问题，只有 $S_{1x}=S_{0x}=S_x$，才不存在弱重叠问题。例如在研究素食主义对寿命的影响时，如果处理组样本只包括女性，控制组样本包括女性和男性，此时就会出现弱重叠问题，因为 $\{S_{0x}-S_x\}$ 非空。

B_B：偏差应归于弱平衡。尽管两群人的集合 $\boldsymbol{x}$ 重合了，但当 $F(\boldsymbol{x},\ D=1)\neq F(\boldsymbol{x},\ D=0)$ 时，B_B 偏差就会出现。在随机分布中，若是实验组和控制组中的协变量 $\boldsymbol{x}$ 来自于不同的分布，偏差就归结于这种不平衡的现象，并且可以被 B_B 衡量。如图 1-5 所示，实验组与控制组的一个协变量 $\boldsymbol{x}$ 的范围是强重叠的，但二者呈现的分布不同，因而出现弱平衡问题。例如在研究素食主义对寿命的影响的问题时，若吃素的人当中女性更多，吃肉的人当中男性更多，就会存在弱平衡问题。

B_C：这种偏差归结于不可观测的选择的存在。即使控制了包含在 x 集合中的可观测混杂因素，这种偏差仍然存在。只有当 $E(Y_0\mid\boldsymbol{x},\ D=1)=E(Y_0\mid\boldsymbol{x},\ D=0)$，即 $\bar{y}_{01}=\bar{y}_{00}$（条件均值独立假设，即 CMI 隐含的条件）时，这种偏差为 0，详见公式（1-25）。当

CMI 假设不被满足，政策项目中的选择受不可观测的混杂因素影响时，偏差 B_C 就会出现。这也意味着，在控制 $\boldsymbol{x}$ 的情况下，D 仍然是非随机分配的，例如在研究素食主义对寿命的影响时，女性是否吃素并不是随机分配的，很有可能受到宗教信仰的影响（假设宗教数据是不可得的），也就是说宗教信仰这个不可观测因素的存在导致了 $\bar{y}_{01}=\bar{y}_{00}$ 不成立，即处理组吃素女性如果不吃素时的期望寿命与控制组不吃素女性的期望寿命不等，此时会存在偏差 B_C。

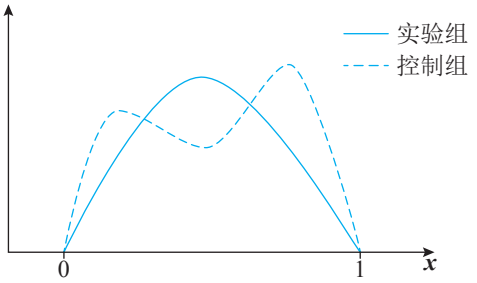

图 1-5　强重叠与弱平衡情况下协变量 $\boldsymbol{x}$ 的分布

运用计量方法并不能够同时消除这三种偏差，一般只能消除一部分。例如，匹配（matching）能够有效减少 B_A 和 B_B，但不能减少 B_C；工具变量法（IV）则只能减少 B_C，不能减少 B_A 和 B_B。

1.5.5　Excel专栏：选择性偏差分解

Excel 演示文件

本专栏第一部分是基于Excel操作展示穿透视角下选择性偏差的分解，本案例仍旧是研究技能培训项目对工资收入的影响。我们假设决定劳动者收入（Y）的因素主要有智商（X_1）、家庭背景（X_2）以及是否接受培训（D），并假定家庭背景属于不可观测因素，即在真实视角中 X_2 数据不可得。在本例中，我们总共构造了 200 个样本，令前 100 个样本作为实验组接受培训，后 100 个样本则为控制组不接受培训，并根据穿透视角下设定的 $Y=10+10\cdot D+6\cdot X_1+5\cdot X_2+\varepsilon$ 关系式生成 Y_0 与 Y_1 数值，其中 ε 为随机误差项。

本例中的部分具体数据详见表 1-4，本例的协变量包括智商（可观测）、家庭背景（不可观测）。在“穿透视角”下，我们已知培训对收入影响为 +10，也知道真实视角下无法观测到的个体结果变量的“反事实”和真实视角下无法观测到的 X_2，其中，“反事实”指的是处理组的 Y_0 和控制组的 Y_1。

表 1-4　关于培训与收入的研究案例——“穿透视角”

序号	智商（X_1）	家庭背景（X_2）	培训（D）	Y_0	Y_1	Y
90	4	1	1	38.93	49.25	49.25
91	4	1	1	43.17	48.33	48.33
92	4	3	1	45.87	56.93	56.93
93	4	2	1	37.13	58.17	58.17
94	4	4	1	57.84	69.58	69.58
95	4	3	1	48.64	62.13	62.13
96	4	2	1	44.49	60.07	60.07
97	4	1	1	34.95	51.92	51.92
98	4	3	1	48.17	57.35	57.35
99	4	4	1	55.66	66.12	66.12
100	4	1	1	37.45	51.05	51.05
101	1	2	0	26.95	31.13	26.95
102	1	2	0	30.82	38.37	30.82
103	1	1	0	16.60	33.20	16.60
104	1	1	0	23.73	26.80	23.73
105	1	4	0	39.39	44.50	39.39
106	1	3	0	33.97	37.80	33.97
107	1	1	0	25.20	28.22	25.20
108	1	2	0	24.13	38.55	24.13
109	1	1	0	20.34	32.08	20.34
110	1	3	0	31.01	43.50	31.01

基于表 1-4，在“穿透视角”下，由于已知个体结果变量的“反事实”，所以通过前文所述公式可得出 ATET、ATENT 以及 ATE 的准确数值，利用 Excel 中的 AVERAGE 函数经过简单计算得出，在该数据样本中：

$$\begin{aligned}\text{ATET} &= E(Y_1 - Y_0 \mid D=1)\\ &= \text{AVERAGE}(Q4:Q103) - \text{AVERAGE}(P4:P103)\\ &= 9.54\end{aligned}$$

$$\begin{aligned}\text{ATENT} &= E(Y_1 - Y_0 \mid D=0)\\ &= \text{AVERAGE}(Q104:Q203) - \text{AVERAGE}(P104:P203)\\ &= 10.60\end{aligned}$$

$$\begin{aligned}\text{ATE} &= \text{ATET} \bullet \rho(D=1) + \text{ATENT} \bullet \rho(D=0)\\ &= 0.5 \times \text{ATET} + 0.5 \times \text{ATENT}\\ &= 10.07\end{aligned}$$

$$\text{DIM} = E(Y \mid D=1) - E(Y \mid D=0)$$

$$= \text{AVERAGE}(Q4:Q103) - \text{AVERAGE}(P104:P203)$$

$$= 13.46$$

$$B_1 = \text{DIM} - \text{ATET} = 13.46 - 9.54 = 3.92$$

根据图 1-6 和图 1-7 可知，本例既存在弱重叠，也存在弱平衡，同时我们假设家庭因素是不可观测的，则在此例中，B_A、B_B 和 B_C 同时存在。

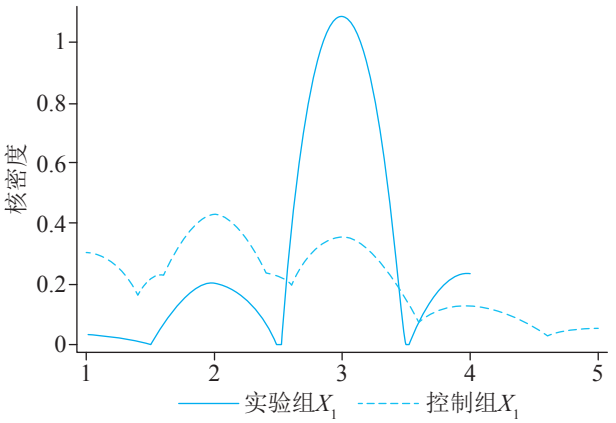

图 1-6　智商（X_1）在实验组和控制组的分布

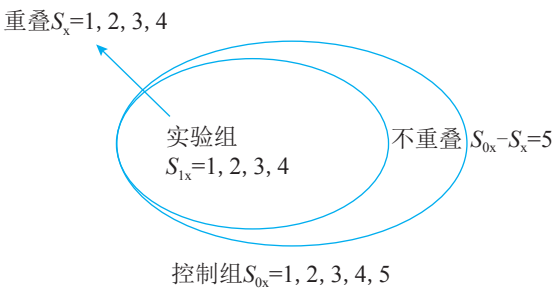

图 1-7　智商（X_1）在控制组和实验组的子集

根据选择偏差分解的公式，可计算得到

$$B_A = \int_{S_{1x}-S_x} \bar{y}_{01}\mathrm{d}F(x,\ D=1) - \int_{S_{0x}-S_x} \bar{y}_{00}\mathrm{d}F(x,\ D=0)$$

$$= \sum_{x=\phi} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - \sum_{x=5} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=0)$$

$$= 0 - E(Y_0 \mid x=5,\ D=0) \bullet P(x=5 \mid D=0)$$

$$= 0 - 51.10 \times \frac{4}{100}$$

$$= -2.04$$

$$B_B = \int_{S_x} \bar{y}_{00}\left[\mathrm{d}F(x,\ D=1)\right] - \int_{S_x} \bar{y}_{00}\left[\mathrm{d}F(x,\ D=0)\right]$$

$$= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=1) - \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=0)$$

$$= 40.77 - 34.26$$

$$= 6.51$$

$$B_C = \int_{S_x} \bar{y}_{01} \mathrm{d}F(x,\ D=1) - \int_{S_x} \bar{y}_{00} \mathrm{d}F(x,\ D=1)$$

$$= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=1)$$

$$= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - 40.77$$

$$= 40.22 - 40.77$$

$$= -0.55$$

$$B_A + B_B + B_C = B_1 = \mathrm{DIM} - \mathrm{ATET} = 3.92$$

因此在本例中，选择偏差 B_1 可分解为 B_A、B_B 和 B_C。

第二部分是基于第一部分的 Excel 操作展示组内差异对选择性偏差大小的影响。在此部分，我们在第一部分生成的 200 个样本数据的基础之上改变了四个处理组样本的 X_1 数值，如表 1-5 所示，此四个样本的原始 X_1 数值为 2，我们将其改为 1，并基于公式重新生成 Y_0 和 Y_1 值，也就表明第一部分中的处理组与此部分中的处理组存在差异。

表 1-5　原始数据与修改数据对比表

序号	智商（X_1）	家庭背景（X_2）	培训（D）	Y_0	Y_1	Y
原始数据						
3	2	2	1	30.89	43.91	43.91
4	2	1	1	29.94	30.29	30.29
5	2	4	1	34.40	55.14	55.14
6	2	1	1	33.47	36.54	36.54
修改后数据						
3	1	2	1	24.89	37.91	37.91
4	1	1	1	23.94	24.29	24.29
5	1	4	1	28.40	49.14	49.14
6	1	1	1	27.47	30.54	30.54

此时，我们基于新的 200 个样本的数据重新做选择偏差的分解。在此数据样本中，由于修改 X_1 数值的四个样本依旧是按照原始公式生成的 Y_0 和 Y_1 值，所以 ATET、ATENT 以及 ATE 的值并不会发生变化，依旧为 $\mathrm{ATET} = 9.54$，$\mathrm{ATENT} = 10.60$，$\mathrm{ATE} = 10.07$，但 DIM 和选择偏差 B_1 的值会发生变化，可计算得到：

$$\mathrm{DIM} = E(Y \mid D=1) - E(Y \mid D=0)$$

$$= 13.22$$

$$B_1 = \text{DIM} - \text{ATET}$$
$$= 3.68$$

可以发现，我们人为构造的处理组组内差异导致选择性偏差 B_1 的减小。重复第一部分中选择偏差分解的计算过程：

$$\begin{aligned}
B_A &= \int_{S_{1x}-S_x} \bar{y}_{01} \mathrm{d}F(x,\ D=1) - \int_{S_{0x}-S_x} \bar{y}_{00} \mathrm{d}F(x,\ D=0) \\
&= \sum_{x=\phi} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - \sum_{x=5} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=0) \\
&= 0 - E(Y_0 \mid x=5,\ D=0) \bullet P(x=5 \mid D=0) \\
&= 0 - 51.10 \times \frac{4}{100} \\
&= -2.04 \\
B_B &= \int_{S_x} \bar{y}_{00} \left[\mathrm{d}F(x,\ D=1)\right] - \int_{S_x} \bar{y}_{00} \left[\mathrm{d}F(x,\ D=0)\right] \\
&= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=1) - \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=0) \\
&= 40.56 - 34.26 \\
&= 6.30 \\
B_C &= \int_{S_x} \bar{y}_{01} \mathrm{d}F(x,\ D=1) - \int_{S_x} \bar{y}_{00} \mathrm{d}F(x,\ D=1) \\
&= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=0) \bullet P(x \mid D=1) \\
&= \sum_{x=1,2,3,4} E(Y_0 \mid x,\ D=1) \bullet P(x \mid D=1) - 40.56 \\
&= 39.98 - 40.56 \\
&= -0.58
\end{aligned}$$

$$B_A + B_B + B_C = B_1 = \text{DIM} - \text{ATET} = 3.68$$

与第一部分中的对应数值相比较可知，B_A 的值没有发生改变，B_B 与 B_C 的值均减小，这说明修改后的样本数据在 X_1 上更加平衡，受不可观测因素的影响也更小。

1.5.6　选择性偏差的解决方法

此部分选择性偏差的解决方法与 1.3.3 节平均处理效应估计的内容具有对应性。解决选择性偏差的方法之一是随机化，第 1.3.3 节证明了随机分配下有 DIM = ATE = ATET = ATENT，此时选择性偏差为 0，随机化实验是指通过一定的随机机制，比如投硬币的方式决定哪些个体进入实验组，哪些个体进入控制组。随机化的关键作用是使实验组和控制组个体除处理变量之外的因素（包括可观测因素和不可观测因

素）分布平衡或者相似，从而使实验组与控制组结果变量的差距就是政策处理效应的一致估计。回到教育培训对工资收入影响的例子，假设可以做随机化实验，考察教育培训 D 对工资收入 Y 的影响。对于一个特定总体，对他们是否受到培训进行随机化分配，通过扔硬币的方式，正面不让他受到培训，反面让他受到培训，直接比较实验组和控制组的工资收入就可以得到教育培训对工资收入的平均处理效应，即使这两组个体在智力、家庭背景等因素方面存在一定差距。正如前文所述，现实中个体是否得到处理往往是个人选择的结果，例如潜在培训收益率高的个体会选择接受培训，潜在培训收益率低的个体则不会选择接受培训，所以通过随机化实验得到的政策处理效应很有可能与现实中的政策处理效应不同。此外，随机化实验的实施时间长、实施成本高，故而在大多数情况下，会使用以下两类方法降低选择性偏差：第一类方法就是基于 1.4.2 节中的可观测选择假设的计量方法，例如匹配和清晰断点回归；第二类方法就是基于 1.4.3 节中的不可观测选择假设的计量方法，例如工具变量法、模糊断点回归和双重差分等，具体介绍详见对应章节。

1.5.7　辛普森悖论

根据上文假定，当协变量 x 不平衡时会出现选择偏差问题。如果选择偏差和实际的 ATET 符号是相反的，甚至会出现 DIM 和分组的计算结果相反的极端情况，辛普森悖论就是一个很好的例子。辛普森悖论是指当人们尝试探究两种变量是否具有相关性的时候，会分别对之进行分组研究。然而，在分组比较中都占优势的一方，在总评中有时反而是劣势的一方，或者在总评比较中占优势的一方，在分组中有时反而是劣势的一方，即：

$$E(Y\mid D=1)-E(Y\mid D=0)>0\text{但是}E(Y\mid x,\ D=1)-E(Y\mid x,\ D=0)<0\forall x \tag{1-59}$$

$$E(Y\mid D=1)-E(Y\mid D=0)<0\text{但是}E(Y\mid x,\ D=1)-E(Y\mid x,\ D=0)>0\forall x \tag{1-60}$$

1.5.8　Excel专栏：辛普森悖论

德根等（Dehghan 等，2018）在德国慕尼黑举行的欧洲心血管疾病大会上公开的研究结果表明，吃红肉（非加工）和奶制品多的人，早死的概率比全体人口减少了 25%，该学者发现肉类的饱和脂肪对于心脏有保护作用。本部分结合上述背景，以素食主义对寿命的影响作为案例，对本节中的辛普森悖论予以展示。

Excel 演示文件

下面基于“穿透视角”，利用 Excel 操作展示本案例中的辛普森悖论。本案例研究的是素食主义对寿命的影响，在该例中我们假设决定寿命（Y）的因素为是否是女性（X）和是否吃素（D），本例共包括 200 个样本，我们让前 100 个样本作为实验组，即 $D=1$，后 100 个样本则为控制组，即 $D=0$，并根据“穿透视角”下设定的 $Y=65-5\cdot D+15\cdot X+\varepsilon$ 关系式生成 Y_0 与 Y_1 数值，其中 ε 为随机误差项，即我们预设平均处理效应为 −5，素食主义对寿命存在负面影响。部分具体数据情况详见表 1-6，处理组的 Y_0 和控制组的 Y_1 均为“反事实”，是在真实视角中无法观测到的数据。

表 1-6　素食主义对寿命的影响——“穿透视角”

序号	是否女性（X）	是否吃素（D）	Y_0	Y_1	Y
90	0	1	64.85	57.33	57.33
91	1	1	77.07	75.14	75.14
92	0	1	69.05	60.45	60.45
93	1	1	81.62	75.38	75.38
94	1	1	80.76	70.33	70.33
95	0	1	65.77	57.72	57.72
96	1	1	82.21	79.74	79.74
97	1	1	72.37	77.85	77.85
98	1	1	73.42	74.94	74.94
99	0	1	62.04	57.38	57.38
100	1	1	75.45	71.95	71.95
101	1	0	78.34	70.30	78.34
102	0	0	65.16	55.72	65.16
103	1	0	74.58	78.60	74.58
104	0	0	63.86	61.85	63.86
105	1	0	84.50	77.07	84.50
106	0	0	67.35	63.73	67.35
107	0	0	70.68	58.94	70.68
108	0	0	71.64	63.26	71.64
109	1	0	80.74	78.06	80.74
110	0	0	60.20	61.67	60.20

在本例中，肉食者中 20% 为女性，素食者中 70% 为女性。$D=1$ 代表素食主义，$X=1$ 代表女性，Y 代表寿命。可知：$P(X=1|D=0)=0.2$，$P(X=1|D=1)=0.7$，即 D 与 X 是不独立的。

则分组计算结果为

$$E(Y|X=0,\ D=0)=64.67,\quad E(Y|X=0,\ D=1)=59.79$$

$$E(Y|X=0,\ D=1)-E(Y|X=0,\ D=0)=-4.88$$

$$E(Y|X=1,\ D=0)=79.76,\quad E(Y|X=1,\ D=1)=74.79$$

$$E(Y \mid X=1,\ D=1)-E(Y \mid X=1,\ D=0)=-4.97$$

$$E(Y \mid X,\ D=1)-E(Y \mid X,\ D=0)<0$$

由上面结果可知，男性和女性的素食效应均为 -5，即男性组和女性组吃素都会缩短寿命，分组计算的结果表明素食主义会缩短寿命，与我们预设的值方向相同，大小相近。

但是，若直接合并所有个体计算 DIM，则得到：

$$\begin{aligned}E(Y \mid D=0)&=E(Y \mid X=0,\ D=0)P(X=0 \mid D=0)+\\&\quad E(Y \mid X=1,\ D=0)P(X=1 \mid D=0)\\&=64.67\times 0.8+79.76\times 0.2=67.69\end{aligned}$$

$$\begin{aligned}E(Y \mid D=1)&=E(Y \mid X=0,\ D=1)P(X=0 \mid D=1)+\\&\quad E(Y \mid X=1,\ D=1)P(X=1 \mid D=1)\\&=59.79\times 0.3+74.79\times 0.7=70.29\end{aligned}$$

$$\mathrm{DIM}=E(Y \mid D=1)-E(Y \mid D=0)=2.60$$

此时出现辛普森悖论：

$$E(Y \mid D=1)-E(Y \mid D=0)>0$$

但是

$$E(Y \mid X,\ D=1)-E(Y \mid X,\ D=0)<0 \forall X$$

根据合并所有个体计算得到的 DIM 值会得出错误结论：素食主义能够延长寿命。导致错误结论的原因是，尽管男性和女性都有负的素食效应，但由于女性寿命比男性长得多（肉食者为 +15.27，素食者为 +15），因此积极的女性效应主导着消极的素食效应，最终导致整体估计出现偏差。该案例中，因果关系如图 1-8 所示，其中性别是混淆变量（confounding variable），同时影响自变量和因变量。

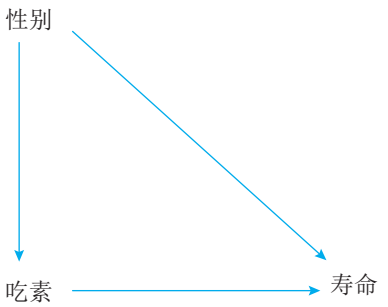

图 1-8　因果关系

合并数据可能会对真实情况产生干扰，因此要控制性别变量消除女性效应，应在原公式中将 $P(X \mid D)$ 替换为 $P(X)$，避免 D 对 X 的干扰，从而使两组的性别变量得以平衡。因为男性与女性在此案例中的比例为 11 ： 9，所以 $P(X=0)=0.55$，$P(X=1)=0.45$，可得：

$$E(Y \mid D=0)=E(Y \mid X=0,\ D=0)P(X=0)+E(Y \mid X=1,\ D=0)P(X=1)$$
$$=64.67\times 0.55+79.76\times 0.45=71.46$$
$$E(Y \mid D=1)=E(Y \mid X=0,\ D=1)P(X=0)+E(Y \mid X=1,\ D=1)P(X=1)$$
$$=59.79\times 0.55+74.79\times 0.45=66.54$$
$$\mathrm{DIM}=E(Y \mid D=1)-E(Y \mid D=0)=-4.92$$

由上式结果可知，我们最终得到了与分组计算结果和预设值相近的值。

1.6　政策评估计量方法

此部分对本书中介绍的政策评估计量方法进行简单的总览性分类与评述，具体见表 1-7 和表 1-8。表 1-7 是根据选择机制和数据类型对政策评估计量方法进行的大概分类，其中 PSM-DID、合成控制法与三重差分法均是双重差分法的延伸与拓展。表 1-8 是政策评估计量方法的应用优势与局限性的简要介绍，具体的分析将在后续章节中展开，表 1-8 的内容也说明不存在十全十美的评估政策处理效应的计量方法，每一种方法都有自己的局限性，所以在应用过程中，要根据实际数据情况与分析情况选择合适的计量方法。

表 1-7　根据选择机制和数据类型对政策评估计量方法的分类

评估方法	选择机制		数据类型	
	适用处理可观测选择	适用处理不可观测选择	截面数据	面板、重复截面数据
匹配	√		√	
双重差分	√	√		√
工具变量	√	√	√	
断点回归	√（清晰 RD）	√（模糊 RD）	√	
PSM-DID	√	√		√
合成控制法	√	√		√
三重差分	√	√		√

表 1-8　政策评估计量方法应用优势与局限性

评估方法	应用优势	应用局限性
匹配	适用于基于可观测变量的选择 非参数估计 不需要对分布进行假设	较强的假设前提：CIA 假设和重叠假设 不允许不可观测因素的存在，不能解决内生性问题 数据量要求大
双重差分	允许不随时间变化的不可观测因素的存在 不需要对分布进行假设	面板数据要求更为严苛 很强的识别假设：共同趋势假设 参数估计

续表

评估方法	应 用 优 势	应用局限性
工具变量	适用处理可观测选择和不可观测选择	工具变量的选择具有难度 参数估计 处理效应异质性框架下只能估计出 LATE
断点回归	断点附近近似于完全随机化实验 不对分布进行假设 内部有效性强	断点和带宽的选择 外部有效性较弱，断点处的因果效应难以推广到其他样本值
PSM-DID	允许不随时间变化的不可观测因素的存在能大幅降低选择偏差，共同趋势假设更容易满足	满足重叠假设
合成控制法	允许随着时间变化的不可观测因素的存在 适用只有一个实验对象的政策评估 非参数估计	要求政策实施前后具有较长的时间跨度 实验组个体特征远远大于或远远小于控制组个体特征时，无法找到合适权重拟合实验组个体
三重差分	允许随着时间变化的不可观测因素的存在	要求时间趋势差异满足共同趋势假设

1.7 本章总结

在政策评估研究中，随机控制实验、自然实验以及准实验为得到政策真实的因果效应（处理效应）提供了强大的工具支撑。而在经济学实证研究中，以随机控制实验、自然实验以及准实验为主的因果识别方法，正推动着微观计量经济学的可信性革命发展。

因果识别最重要的就是要在“反事实”的框架中消除选择性偏差。随机控制实验通过研究样本的随机化分组解决选择性偏差问题，但相较于观察性研究来说，随机控制实验研究通常面临着高昂的成本以及无法回避的伦理问题，因此更多的学者往往寻求通过观察数据实现有效的因果识别策略并进行近似随机化的处理。由非实验经济学者发展起来的基于自然实验、准实验的工具变量（IV）、双重差分法（DID）、断点回归（RD）设计和倾向得分匹配法（PSM）在满足一定的假设条件下，能够使非实验数据近似得到随机化处理，进而得到精准的因果识别，很好地弥补了随机控制实验无法在非实验数据中进行因果识别的缺陷，已经成为致力于因果识别的研究者在微观实证研究中的“宝典”。

从发展进程来看，随机控制实验研究和非随机控制实验研究之间联系紧密，并互相影响。对随机控制实验方法和非随机控制实验方法的对比研究明晰了非随机控制实验方法在真实项目评估中的有效性，并在一定程度上推动了非随机控制实验研究方法的完善与进一步发展。随机控制实验研究作为因果识别的基准，使非随机控制实验研究在识别策略上更加完善、解释问题上更加谨慎。而实验研究者对工具变量解决部分依从问题上有限性的认识，促使了更加复杂、完善的识别方法和研究设计形成。同时，非随机控制

实验方法由于在解决部分不完全随机化问题上具有的优势，也开始被用于随机控制实验研究中，这使随机控制实验研究的结果更加稳健。

可以看到，随机控制实验研究和非随机控制实验研究之间的界限并不是非常清晰，实验无论是基于随机控制实验、自然实验展开，还是基于人工选择的准实验研究，其本质都是在揭示不同变量之间的因果关系，进而探寻经济现象和社会现象的内在规律及背后真理，因此需要在合适的条件下为合适的问题选择合适的研究工具。

1.8　经典案例分析

随机控制实验中的因果推断——积极疗法是否比保守疗法更有效?

文献来源：保罗 • 罗森鲍姆 . 观察与实验：因果推理导论 [M]. 美国剑桥城：哈佛大学出版社，2017.

1. 研究思路

该研究通过随机控制实验的方法，对患有感染性休克的病人采用不同治疗方法产生的效果进行了分析，通过对费舍尔无效果假设（H_0：积极疗法与保守疗法对死亡率的影响无差异）进行检验，试图识别患者死亡率的差异究竟是来源于个体差异还是治疗方法的优劣，从而确认治疗方案是否有效。

2. 研究设计

该案例利用 ProCESS 实验[①] 数据研究积极疗法和保守疗法对死亡率的影响。首先，对样本中处理组和对照组的协变量（例如导致败血症的原因、收缩压等指标）进行描述性统计发现（如表 1-9 所示），协变量间无显著差异。进一步地，对两组患者在院 60 天的存活状况进行统计后发现，处理组（积极疗法）比对照组（保守疗法）的死亡率高 2.8%，如表 1-10 所示，因而提出研究需要解决的问题：处理组和对照组死亡率的差异来源于不同的治疗方法还是仅仅偶然发生？

表 1-9　ProCESS 实验中的协变量[②]

处 理 组	积 极 疗 法	保 守 疗 法
样本容量（病人数量）	439	446
年龄（年份，均值）	60	61
男性占比（%）	53	56
从护士之家来的占比（%）	15	16

① 该实验是美国 31 个急诊部门的研究人员合作完成的针对早期感染性休克护理的调查。
② 表 1-9 内容摘自罗森鲍姆（Rosenbaum，2017）中的表 1.1。

续表

处 理 组	积 极 疗 法	保 守 疗 法
肺炎	32	34
泌尿感染	23	20
腹腔感染	16	13
健康评估分数（均值）	20.8	20.6
收缩压（均值）	100	102
血清乳酸（均值）	4.8	5

表 1-10　ProCESS 实验中患者在医院死亡的结果 ①

住院 60 天的存活与否				
处理组	在院死亡	其他	总计	死亡率（%）
积极	92	347	439	21.0
保守	81	365	446	18.2
总计	173	712	885	

为了解答这一问题，该研究构建无治疗效果假设 H_0：患有感染性休克的病人的治疗效果（存活或死亡）取决于感染的性质和病人自身的健康状况，而与采用哪种治疗方案（积极疗法或保守疗法）无关，即积极疗法和保守疗法的效果无差异。对于所有的病人 i，在积极治疗方案下的生存率 r_{Ti} 和保守治疗方案下的生存率 r_{Ci} 是相同的，即 $r_{Ti}=r_{Ci}$（$i=1,\cdots, I$；$I=885$），其中 i 为第 i 个病人。

在对全样本 $I=885$ 名病人进行 ProCESS 实验之前，先假设一个处于极端情况下的小型随机控制实验。假设在一个随机临床实验中，有 $I=8$ 个病人，选取其中 $m=4$ 个病人接受积极疗法，$I-m=4$ 个病人接受保守疗法。如果病人 i 接受了积极疗法则记为 $Z_i=1$，若接受保守疗法则记为 $Z_i=0$，其中 $4=m=Z_1+\cdots+Z_8$。因此，共有 70 种从 8 名患者中选择 4 名患者进行积极治疗的组合方式，每一种组合情况都有 1/70 的概率发生。使用随机数选择出其中一种方法，然后按随机数指定的方式对患者进行治疗，并记录死亡人数 R_i。假设得到了一种（极端的）结果：所有接受积极治疗的人都死亡，而所有接受保守治疗的人都幸存，那么，这种结果是偶然发生的还是由于治疗方法不同所导致的？

在小型随机控制实验中，为了检验无效果假设，首先假定该假设是正确的，那么通过“反事实”的情况可以得知，在此条件下随机选择一种分配治疗方案，可以得到确定的结果（死亡人数），根据随机控制实验中零值假设下 T 的分布情况，可以得到相应的概率分布表，如表 1-11 所示。其中处理组死亡人数 T 为 4 的概率是 $\Pr(T=4)=1/70=0.0143$，即随机分配的治疗方案将所有死亡病人都放在积极治疗组

① 表 1-10 内容摘自罗森鲍姆（Rosenbaum，2017）中的表 1.2。

中，在 100 次实验中可能会出现 1 ～ 2 次，这是一种小概率事件，根据小概率不可能发生原理[①]，该研究认为在小型随机控制实验中拒绝原假设，即死亡率的差异主要是治疗方案的不同导致的。

表 1-11　70 种可能的处理分配表[②、③]

组别编号	概　率	被处理患者	处理指示器							
			Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
1	1/70	1，2，3，4	1	1	1	1	0	0	0	0
2	1/70	1，2，3，5	1	1	1	0	1	0	0	0
⋮										
44	1/70	2，3，6，8	0	1	1	0	0	1	0	1
⋮										
70	1/70	5，6，7，8	0	0	0	0	1	1	1	1

对于全样本的 ProCESS 实验，依据上述方法由小型随机控制实验拓展得到的 ProCESS 实验的表格将有 885 列（885 个病人），6.7×10^{24} 行（6.7×10^{24} 种分配方式）。除了表格大小之外，逻辑上与小型实验没有任何不同。因此，可用与小型实验中相同的检验步骤来解决全样本 ProCESS 实验中的问题。同理可得，全样本 ProCESS 实验检验无效果的概率为 0.1676。如果积极治疗与保守治疗相比效果没有区别，那么在 6.7×10^{24} 种可能的分配方式中，约 17% 的方式可能导致死亡率差异大于或等于全样本处理组与控制组死亡率的差异（据上文可知该差异为 2.8%）。由于事先不知道效应的方向，通常将双侧 P 值报告为该单侧 P 值的两倍：0.3352=2×0.1676。显然，该结果不能拒绝原假设。

3. 研究结果

通过计算产生差异结果的偶然发生概率，该研究发现，并没有理由拒绝积极治疗和保守治疗的方案效果无差异这一假设。当积极治疗和保守治疗之间没有区别时，处理组与控制组死亡率差异为 2.8% 是一件司空见惯的事情，大约有 33.52% 的概率仅凭偶然就可以产生这种差异。

---

① 小概率不可能发生原理：假设小概率事件在一次实验中是几乎不可能发生的，如果在一次实验中小概率事件 A 发生了，那么我们就有理由怀疑原假设的真实性，因而拒绝原假设。本案例中小概率事件（将所有死亡病人都放在积极治疗组中）发生的概率很低，即结果并非偶然发生，因此我们拒绝治疗方案效果无差异的原假设。

② 表 1-11 内容摘自罗森鲍姆（Rosenbaum，2017）中的表 3.2。

③ 从 I= 8 个病人中选出 m=4 个病人有 70 种方法。比如第一行选择第 1、2、3、4 个病人进行治疗——这是从 8 个病人中选择 4 个的一种方法。或者可以像第二行一样选择第 1、2、3、5 个病人进行治疗——这是第二种方法，以此类推直到把 70 行都填满。

本章习题

1. 简述自然实验与准实验的联系与区别，并说明它们与随机控制实验的区别与联系。

参考答案：自然实验与准实验都是因为某些外部突发事件的发生恰好使一部分个体近似于被随机地分到实验组（或控制组），区别是自然实验通过法律制度变更、自然灾害等自然产生的外生冲击形成处理组和对照组，但准实验要经过一定的统计方法来模拟随机状态（如高考分数线造成的划分），生成处理组和对照组。准实验与随机控制实验的不同之处在于：没有人为地选择控制方式，但又产生了近乎随机控制实验的控制效果。自然实验和准实验的联系在于：一方面，三者具有互补关系，自然实验、准实验用随机控制实验的因果识别框架，在一定条件下解决了观察数据中非随机化而带来的因果识别困扰；另一方面，自然实验、准实验与随机控制实验联系紧密并互相促进。自然实验、准实验让随机控制实验在识别方法上更加完善；随机控制实验使自然实验和准实验在研究设计和方法应用上更加成熟，更类似于随机控制实验。

2. 在随机控制实验的第一阶段和第二阶段，即随机样本选择阶段和参与者进入处理组或控制组阶段，可能存在哪些问题阻碍评估目标的实现？

参考答案：第一阶段，一般而言，由于时间、成本等因素的限制，在现实中很难实现完美的随机控制实验，样本的选取可能会受限于地域，例如教育类的政策评估，抽样范围可能仅仅涉及局部地区，无法实现全国层面的样本获取。在第二阶段，参与者（所选样本对象）可能会根据实验的内容，对是否参与进行主观评价，进而对进入处理组持反对态度或干预处理过程。另一方面，即使进入不同组别的过程完全随机，也存在着处理组或控制组的某些特征变量（如性别、受教育程度等）的不同分布影响实验结果的情形。其中，部分变量还具有不可观测的特点，如学生的交往能力等，这些在计量经济学中被称作遗漏变量。此时自选择问题、遗漏变量问题等将造成评估结果的偏差。

3. 探究音乐培训对儿童认知能力的影响，需要考虑哪些因素呢？请简要设计方案。

参考答案：首先，确定被试对象，随机选择一定数量的儿童（如 100 人），观察他们的年龄、性别、教育背景、是否接受过音乐培训等，对他们进行认知能力测试并记录分数。其次，对这部分儿童进行 10 周的音乐课程培训，再次进行测试。最后，通过两次测试的分数变化情况得出相应的分析结果。由于随着时间的推移，儿童认知能力会自然增长，实验中也可能出现如个人成长经历、家庭学习环境等不可观测变量，可以考虑

使用 PSM-DID 的方法。对于可能存在的内生性问题，可以引入家庭到音乐培训地点的距离作为工具变量进行研究。方案还可以进一步完善细化，留予同学们思考。

4. 简要说明选择性偏差解决的思路和方法。

参考答案：选择性偏差是指在研究过程中因样本选择的非随机性而出现的结论偏差，因此，计量经济学家们开始致力于研究基于准实验数据和观察数据来估测政策或项目效果的计量方法，双重差分法（DID）、工具变量方法（IV）、匹配技术（matching）、断点回归（RD）等应运而生。最理想的解决方法是使研究的样本符合随机分配假设：一是随机选择样本，消除样本选择偏差；二是对选择的样本进行随机化实验，使样本随机进入实验组和控制组，消除自选择偏差。

在满足理想随机控制实验的情况下，可使用计量方法消除一部分选择性偏差。例如用匹配法可以消除由于弱重叠和弱平衡导致的选择性偏差，但不能消除不可观测因素导致的偏差；用工具变量法、双重差分法可以消除不可观测因素导致的偏差，但不能消除弱重叠和弱平衡导致的偏差。为了更好地消除偏差，还可将多种方法结合起来，如倾向得分匹配和双重差分相结合，既可减小弱重叠和弱平衡导致的偏差，又能减小不可观测因素带来的偏差。

5. 尝试对生活中的一个例子做出解释：假设两名同学需要为班级团建寻找一个活动场所，为了满足班里男生和女生的个性化需求，他们分性别统计了网站上的用户评价结果。在男性对 *A* 餐厅和 *B* 餐厅评价的调查中，统计结果显示 *B* 餐厅的活动体验感更好。女性对于 *A* 餐厅和 *B* 餐厅评价的调查也显示 *B* 餐厅优于 *A* 餐厅。最后他们去掉“性别”这一筛选条件，使用相同的数据集，发现就整体用户而言，对 *A* 餐厅的评价更好。如何解释这种现象呢？

参考答案：事实上，两名同学在不觉中进入了辛普森悖论的世界。在辛普森悖论里，当原本分离的数据被组合起来，之前出现的统计现象可能会发生逆转，导致在总评中某一方不再占优。例如此题中，调查的同学忽略了各个餐厅参与评价的人的性别组成。如果对 *B* 餐厅评价的人中男性整体大于女性，而 *A* 餐厅中女性参评人更多，即男女人数在 *A* 餐厅和 *B* 餐厅两组的权重不一致时，组合数据后统计结果可能与分组的评价产生差异。

这个问题在学术研究中也会出现。由于处理组和对照组中协变量（如性别）的分布并不平衡，如果没有排除协变量的影响就对总处理效应进行估计，会使估计的处理效应中混淆着协变量对结果变量的影响。如果协变量分布不平衡对结果变量的效应与真实处理效应相反，抵消了真实处理效应，就有可能使估计出的处理效应和真实处理效应相反。

参考文献 ›››

[1] 范子英 . 如何科学评估经济政策的效应 ?[J]. 财经智库，2018，3（03）：42-64+142-143.

[2] 乔万尼 • 赛鲁利 . 社会经济政策的计量经济学评估：理论与应用 [M]. 邸俊鹏译 . 上海：格致出版社，2020：1-41.

[3] 赵西亮 . 基本有用的计量经济学 [M]. 北京：北京大学出版社，2017.

[4] 周黎安，陈烨 . 中国农村税费改革的政策效果：基于双重差分模型的估计 [J]. 经济研究，2005（8）：44-53.

[5] Alesina A, Giuliano P, Nunn N. On the origins of gender roles: women and the plough[J]. The Quarterly Journal of Economics, 2013, 128（2）：469-530.

[6] Björkman M, Svensson J. Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda[J]. The Quarterly Journal of Economics, 2009, 124（2）： 735-769.

[7] Dehghan M, Mente A, Yusuf S. Associations of fats and carbohydrates with cardiovascular disease and mortality—PURE and simple? –Authors'reply[J]. Lancet(London, England), 2018, 391(10131)： 1681.

[8] Deming D J, Hastings J S, Kane T J, et al. School choice, school quality, and postsecondary attainment[J]. American Economic Review, 2014, 104（3）：991-1013.

[9] Dobbie W, Fryer R G Jr. Are high-quality schools enough to increase achievement among the poor? Evidence from the Harlem Children's Zone[J]. American Economic Journal: Applied Economics, 2011, 3（3）：158-187.

[10] Duflo E, Dupas P, Kremer M. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya[J]. American Economic Review, 2011, 101（5）：1739-1774.

[11] Fogel R W. Railroads and American economic growth: essays in econometric history[J]. Canadian Journal of Economics and Political Science, 1965, 31（4）：611-612.

[12] Fryer R G. Teacher incentives and student achievement: evidence from New York City public schools[J]. Journal of Labor Economics, 2013, 31（2）：373-407.

[13] Heckman J J, Ichimura H, Smith J A, et al. Characterizing selection bias using experimental data[J]. Econometrica, 1998, 66（5）：1017-1098.

[14] Ho D E, Imai K. Estimating causal effects of ballot order from a randomized natural experiment: the California alphabet lottery, 1978–2002[J]. Public Opinion Quarterly, 2008, 72（2）：216-240.

[15] Moser P, Voena A. Compulsory licensing: evidence from the trading with the enemy act[J]. American Economic Review, 2012, 102（1）：396-427.

[16] Rosenbaum P R. Observation and experiment: an introduction to causal inference[M]. Cambridge, Massachusetts: Harvard University Press, 2017.

[17] Rosenbaum P R, Rubin D B. The central role of the propensity score in observational studies for causal effects[J]. Biometrika, 1983, 70（1）：41-55.