因果推断初探

李德山

[1] 西南科技大学 经济管理学院

[2]School of Economics and Management
Southwest University of Science and Technology

2021 年 3 月 15 日

主要内容

主要内容

1. 辛普森悖论
2. 变量关系路径图

主要内容

1. 辛普森悖论
2. 变量关系路径图
3. 因果关系估计偏差来源

主要内容

1. 辛普森悖论
2. 变量关系路径图
3. 因果关系估计偏差来源
4. 常用因果关系估计方法

辛普森悖论

- 因果关系必然导致相关关系，但相关关系未必一定反映因果关系。

- **辛普森悖论**（Simpson's Paradox）：两个变量 X 和 Y 在每个分组中的关系是正（负），但在总体（所有组加总）中关系会发生逆转变成负（正）。

表：辛普森悖论数据总结

		未服药	服药	健康状况差异
		(1)	(2)	(3) = (2) - (1)
30岁组	平均身体健康指数（人数）	80（6）	90（2）	10
40岁组	平均身体健康指数（人数）	60（3）	65（5）	5
所有人	平均身体健康指数（人数）	73.3（9）	72.1（7）	-1.2

辛普森悖论

- 我们用健康指数（Health）对服药与否（如果服药，Treat=1；如果未服药，Treat=0）进行回归分析：

$$Health = 73.3 - 1.2 \times Treat$$

如果控制年龄因素，则回归结果为：

$$Health = 146.9 + 7.2 \times Treat - 2.2 \times Age$$

- 由于服药个体中大多数年龄比较大，如果没有控制年龄因素，服药与否与健康状况的相关系就包含了个体年龄对健康状况的负作用，因此得到了负的治疗效果。
- 在剔除了年龄的影响后，假设不存在其他混淆因素的话，我们就可以将服药与否同健康指数的正相关关系归结于服药对健康有正向的因果效应。

- 相关关系不一定反映因果关系。在某些情况下，通过相关关系去推导因果关系还会自相矛盾。

变量关系路径图

- 变量关系路径图，也称为**有向无环图**（Directed Acyclic Graphy）。
  - **路径图**是由节点和单向前头组成，其中每个节点表示一个变量。我们用实心圆点表示观测得到的变量，空心圆点表示观测不到的变量。
  - 路径图是一个有向无环图：“有向”是指以单向箭头表示变量之间的因果关系；“无环”是指无法从某一个节点出发经过若干路径回到原节点。

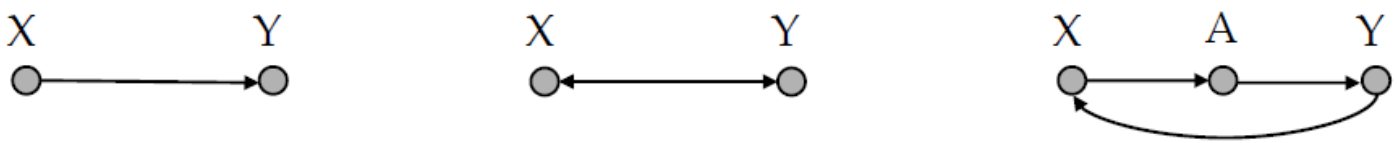

Figure: 路径图基本要素

变量关系路径图

- **路径**是指连接两个变量的一系列箭头线，这些箭头线并不一定要同方向，但只能通过一个变量一次。
- 路径可分为三类：因果路径（也称为链状路径 $A \to B \to C$）；混淆路径（也称为叉状路径 $A \leftarrow B \to C$）；对撞路径（也称为反叉状路径 $A \to B \leftarrow C$）。
- **因果路径**是从解释变量指向被解释变量的单向路径，其特点是所有箭头指向同一方向。
- 两个变量之间如果存在因果关系，他们就存在相关关系，所以因果路径为开放路径。

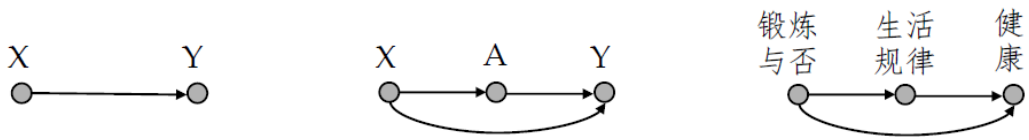

Figure: 因果路径

变量关系路径图

- **混淆路径**是指在解释变量 X 与被解释变量 Y 之间存在混淆变量的路径。混淆变量是同时影响 X 和 Y 的变量。
  - 混淆路径的存在也会造成两个变量的相关性，因此混淆路径也是开放路径。

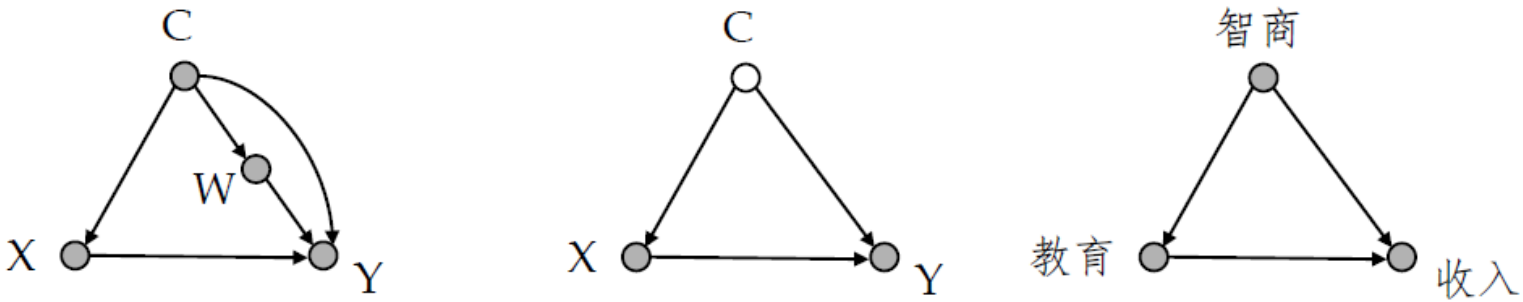

Figure: 混淆路径

变量关系路径图

- **对撞路径**是指包含对撞变量的路径。对撞变量是被两个变量共同影响的变量。
  - 对撞路径并不会造成两个变量的相关性，因此对撞路径是死路径。

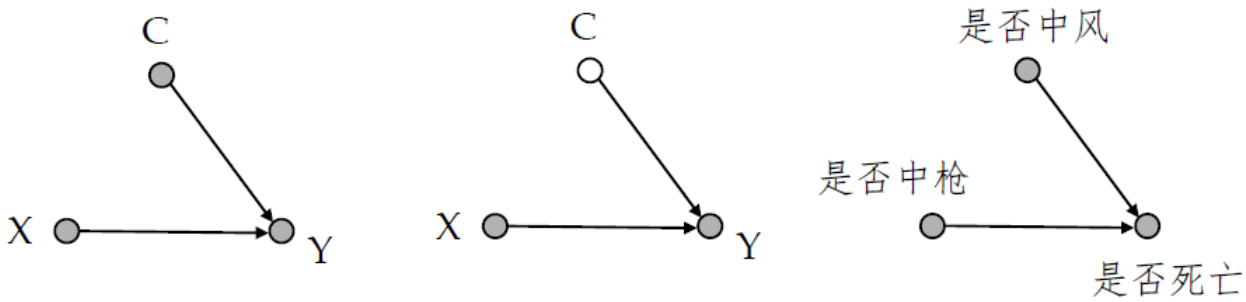

Figure: 对撞路径

变量关系路径

- 估计变量之间的因果关系的本质是找到二者之间所有的因果路径，同时去除二者之间的非因果关系路径。

- 在实际研究中如果没能正确处理变量间的路径关系，就会造成各种偏差。这些偏差主要归纳为三类：

- 1. 混淆偏差
- 2. 过度控制偏差
- 3. 内生选择偏差

变量关系路径

- **混淆偏差**是指在解释变量和被解释变量之间存在未截断的混淆路径，造成解释变量和被解释变量的相关系不仅包含因果关系，还包含非因果关系。
- 截断混淆路径是通过给定混淆变量，从而排除混淆变量的干扰。给定混淆变量可以简单地理解为固定混淆变量的值。在关系图中，我们给变量加个方框表示这个变量是给定的。
- 当混淆变量给定时，解释变量和被解释变量的相关性就与混淆变量无关，二者的相关性就反映了因果关系。在回归分析里，给定某个变量也称为控制某个变量。

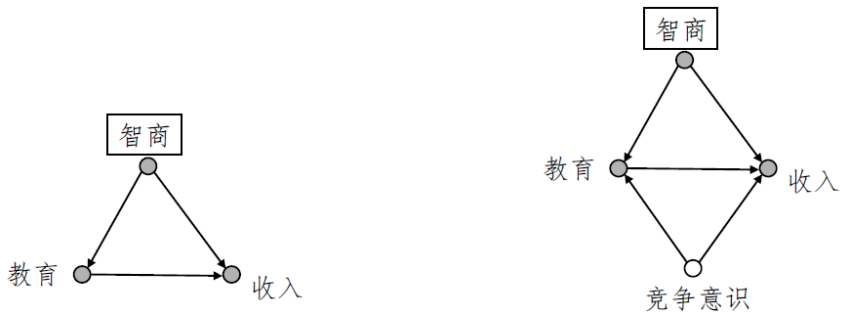

Figure: 混淆偏差例子

变量关系路径

- **过度控制偏差**是指控制了因果路径上的变量造成的偏差。
  - 在研究中，我们要避免控制受解释变量影响并会影响被解释变量的中介变量，负责会造成过度控制偏差。

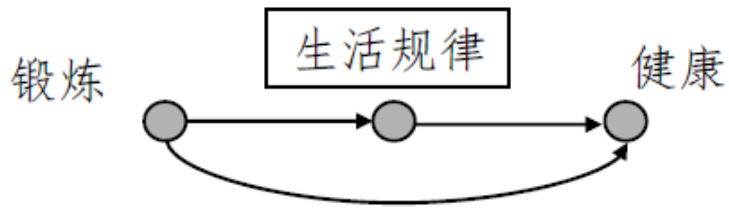

Figure: 过度控制偏差例子

变量关系路径

- **内生选择偏差**也称为对撞偏差。如果给定了两个变量的对撞变量，会造成两个本来不相关的变量之间产生相关关系，这个错误的相关性称为对撞路径。

- 当给定两个变量共同的被解释变量（对撞变量）时，两个变量之间会产生一个衍生路径。衍生路径会造成两个原本不相关的变量变为相关，或造成两本原本相关变量的相关性发生改变。

- 另外，给定对撞变量的被解释变量也会造成衍生路径和内生选择偏差。

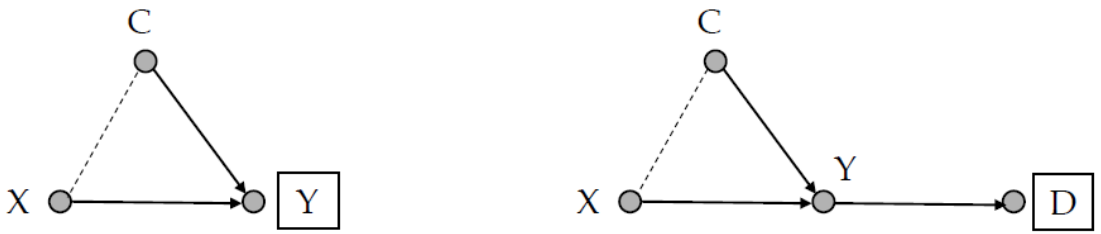

Figure: 内生选择偏差

变量关系路径

- 内生选择偏差的一个例子。

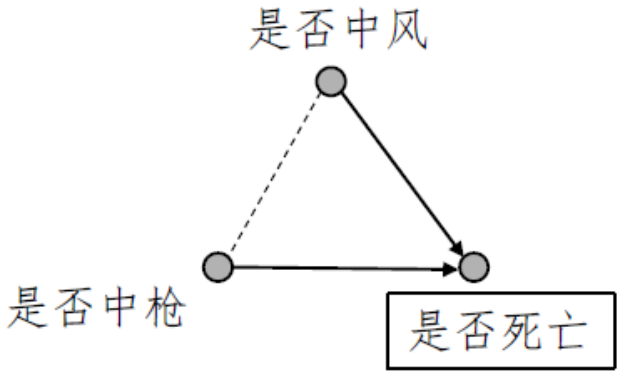

Figure: 内生选择偏差的例子

表：内生选择偏差的例子

是否死亡	是否中风	是否中枪
否	否	否
是	否	是

变量关系路径

- 由于因果关系通常无法直接观测到，我们只能通过变量间的相关性去推测因果关系，因此从路径的角度上来看，因果关系分析的本质就是发现因果路径，截断混淆路径，避免对撞路径产生的衍生路径。

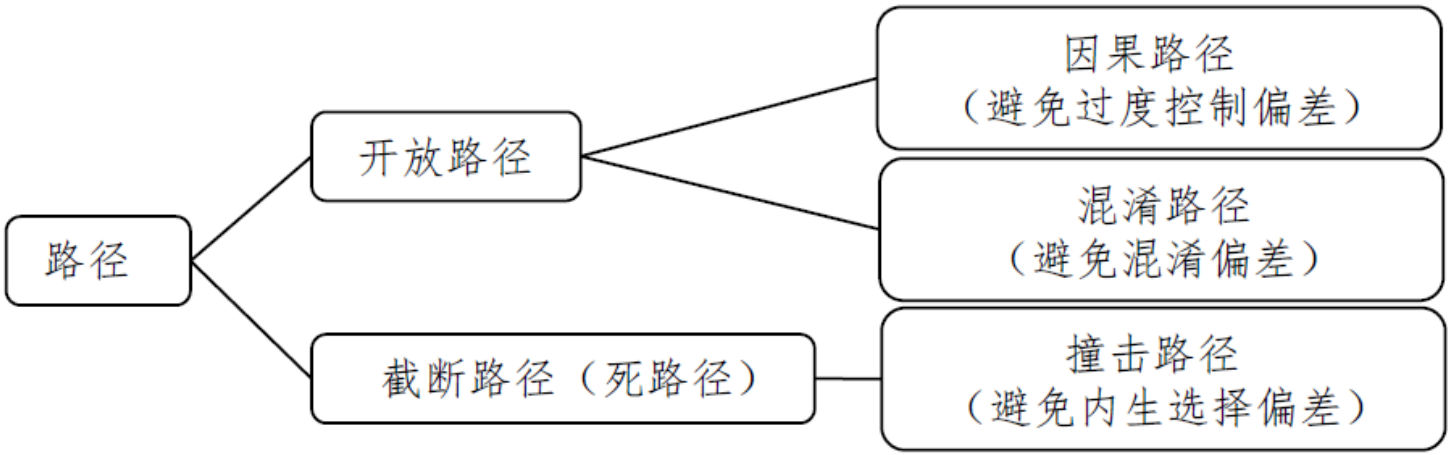

Figure: 路径和偏差种类总结

常用因果关系估计方法

- 因果关系可以直接定义为解释变量 X 的变化（因）导致被解释变量 Y 的变化（果），也可以通过潜在结果模型定义为处置效应。（后面章节会详细介绍）
- 控制实验：解释变量与任何其他可能的混淆变量都不相关。断点回归方法。例子：大学高考分数线。
- 准自然实验：事件的发生与否并不是个体自己能选择的。双重差分法。例子：一个省的税法改革。与控制实验的不同之处在于，准自然实验的干预行为不是随机分配的。
- 在实际研究中，我们通常面对的是观测数据，这类数据的特点是数据产生不具备随机安排并且是个体自行选择产生的。面板数据固定效应模型、工具变量法、匹配方法。例子：服药和健康。
- 在实证研究中，若存在内生选择偏差，它不是因为解释变量和不可观测因素在总体里存在相关性造成的，而是由于用来估计的样本不是从总体里随机抽取，导致样本里解释变量和不可观测因素存在相关性。Heckman 样本自选择模型。例子：受教育程度与收入。

常用因果关系估计方法

表：常见实证方法解决的估计偏差

方法	解决的因果关系中的偏差
简单回归、匹配法	可观测因素造成的混淆偏差
面板数据分析法	可观测因素+不随时间变化的不可观测因素造成的混淆偏差
工具变量法、双重差分法、断点回归法	可观测因素+不可观测因素造成的混淆偏差
样本自选择模型	包含不可观测因素造成的内生选择性偏差

参考文献：

Angrist and Pischke, Mostly Harmless Econometrics, 2009.

赵西亮，《基本有用的计量经济学》，2017.