第一章：
因果推断常用计量方法图解与概览

展示：王健

CIRG，2020-10-22

内容框架

- 辛普森悖论
- 变量关系图
- 因果关系估计偏差来源
- 常用因果关系估计方法概览

1、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应因果关系

- 辛普森悖论：

  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）

1.1 一个例子

数据见书（P2）

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

1.1 一个例子

- 回归分析：

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

若考虑年龄因素：

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

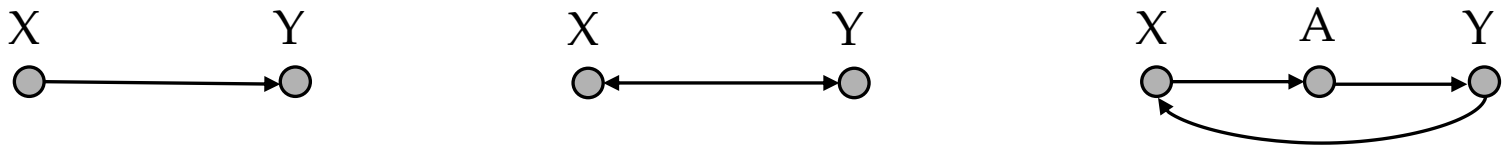
- 由于服药个题中大多数年龄比较大，如果没有控制年龄因素，服药与否与健康状况的相关性就包含了个体年龄对健康状况的副作用，因此得到了负的治疗效果。
- 而在剔除年龄的影响后，假设不存在其他混淆因素的话，我们就可以将服药与否同健康指数的正相关关系归结于服药对健康有正向的因果效应。

1.2 结论

- 相关关系不一定反映因果关系，甚至在某些情况下用相关关系推导因果关系还会自身矛盾。

2、变量关系图

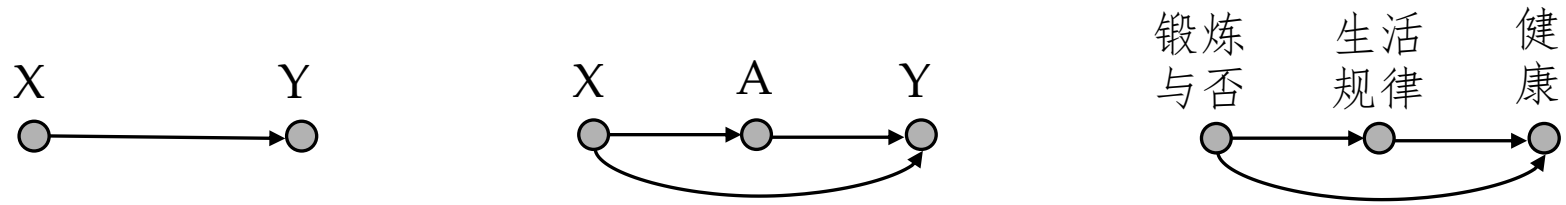
- 变量关系路径图，也成为无环图（Directed Acyclic Graph）
- 路径图是由节点和单项箭头组成的，其中每个节点表示一个变量。我们用实心圆点表示观测得到的变量，空心圆点表示观测不到的变量。
- 路径图是一个有向无环图。所以我们不能用无环有向图去描述互为因果（simultaneous causation）和反馈循环（feedback loops）。但是比较复杂的无环有向图可以描述以上情况，而实际问题中大多数问题可以通过无环有向图来表述。



图：路径基本要素

2.1 路径种类：因果路径

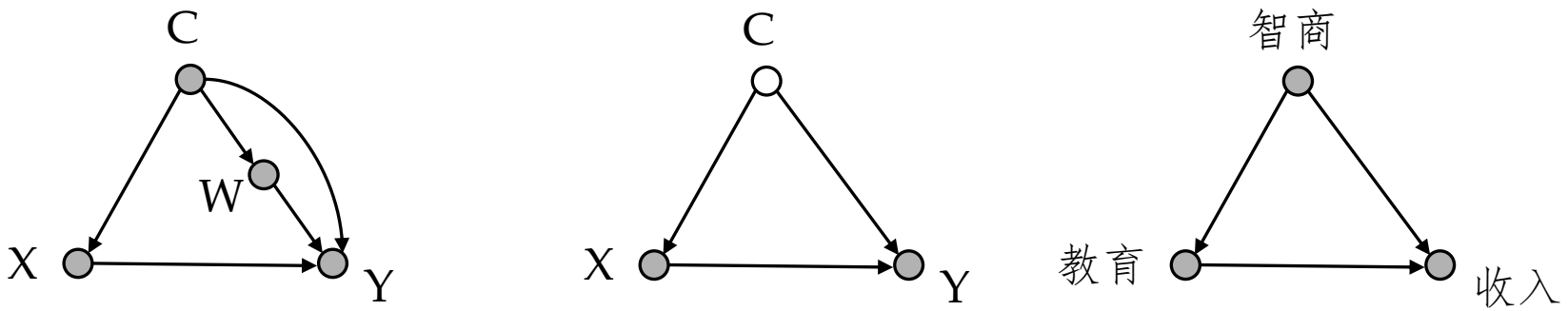
- 因果路径也称链状路径$A \to B \to C$，是解释变量指向被解释变量的单向路径，其特点是箭头指向同一方向。两个变量之间如果存在因果关系，它们就存在相关关系，所以因果路径为开放路径。



图：因果路径

2.1 路径种类：混淆路径

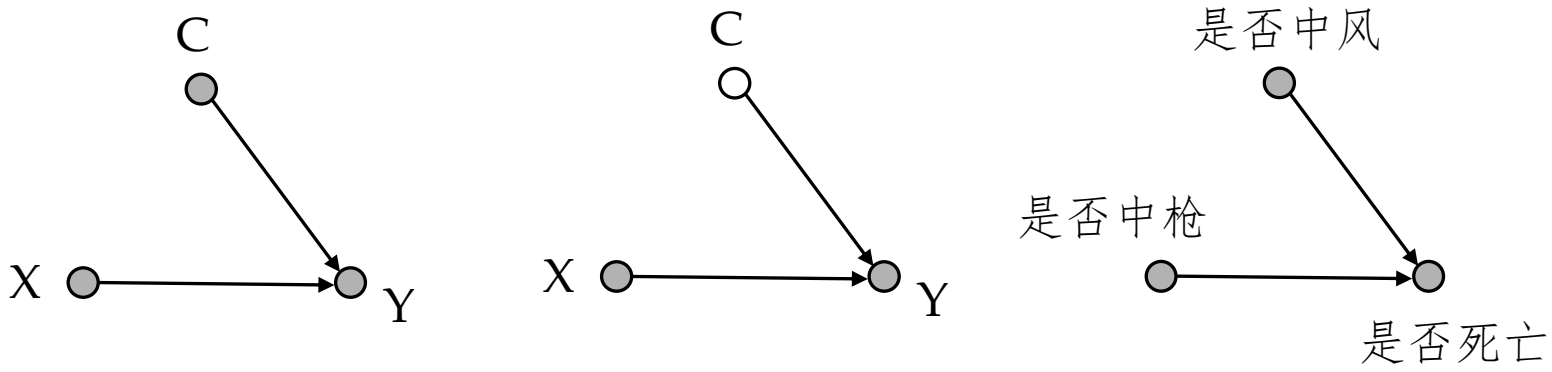
- 混淆路径也称叉状路径$A \leftarrow B \rightarrow C$，是指在解释变量X和Y之间存在混淆变量的路径，混淆变量同时影响X和Y的变量，而混淆变量的存在也会导致相关关系，所以混淆路径也是开放路径。



图：混淆路径

2.1 路径种类：对撞路径

- 对撞路径也称为反叉状路径$A \rightarrow B \leftarrow C$，是指包含对撞路径的变量，对撞变量是指两个变量共同影响的变量。对撞变量不会产生相关性，因此对撞变量是死路径。



图：对撞路径

3、因果关系估计偏差来源

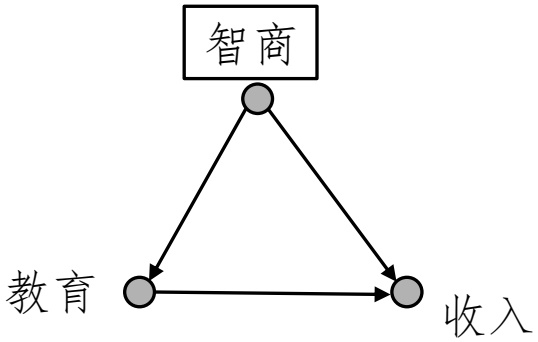
- 估计变量之间的因果关系的本质是找到二者间所有的因果路径，同时去除二者间的非因果关系路径。

- 在实际操作中会造成各种偏差，而偏差可以主要分为三类：
  - 混淆偏差
  - 过度控制偏差
  - 内生选择偏差

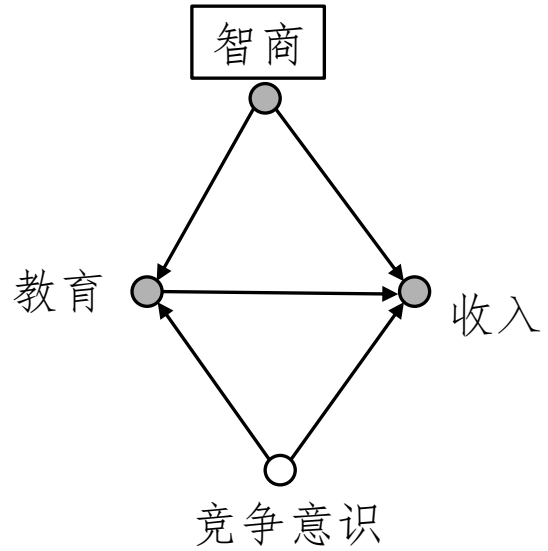
3.1 混淆偏差

- 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性不仅包含因果关系，还包含非因果关系。
- 截断混淆路径是通过给定混淆变量（conditional on confounding variable），从而排除混淆变量的干扰。给定混淆变量可以简单的理解为固定混淆变量的值，而在关系图中，我们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，二者的相关行就反映了因果关系。

3.1 混淆偏差：一个例子



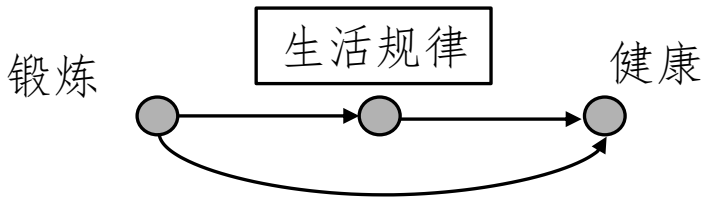
图：截断混淆路径



图：存在未截断的混淆路径

3.2 过度控制偏差

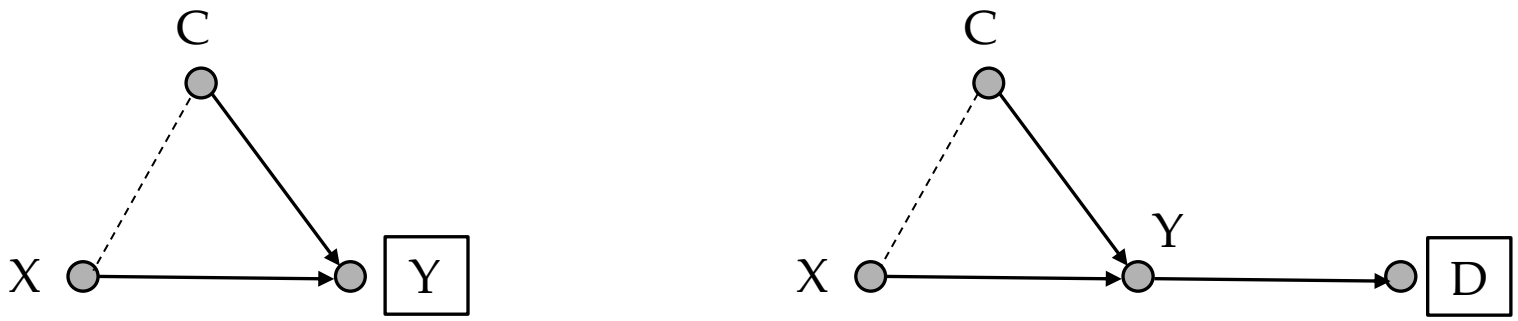
- 过度控制偏差是指控制了因果路径上的变量造成的偏差。
- 在研究中我们要避免控制受解释变量影响并会影响被解释变量的中介变量，否则会造成过度控制偏差。
- 一个例子：



图：过度控制偏差的例子

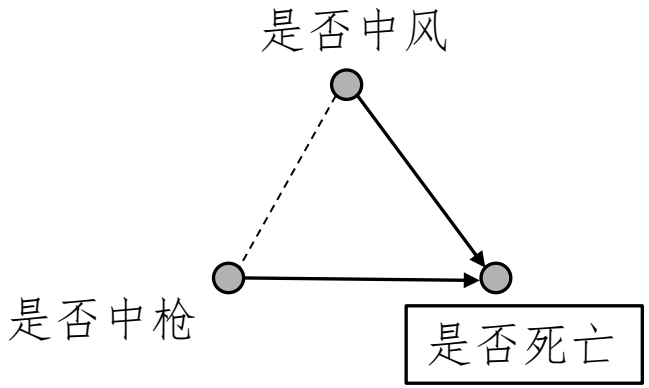
3.3 内生选择偏差

- 内生选择偏差可以理解为当给定两个变量共同的被解释变量（对撞变量）时（或者对撞变量的被解释变量），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，或造成两个原本相关的变量的相关性发生改变。



图：内生选择偏差

3.3 内生选择偏差：一个例子



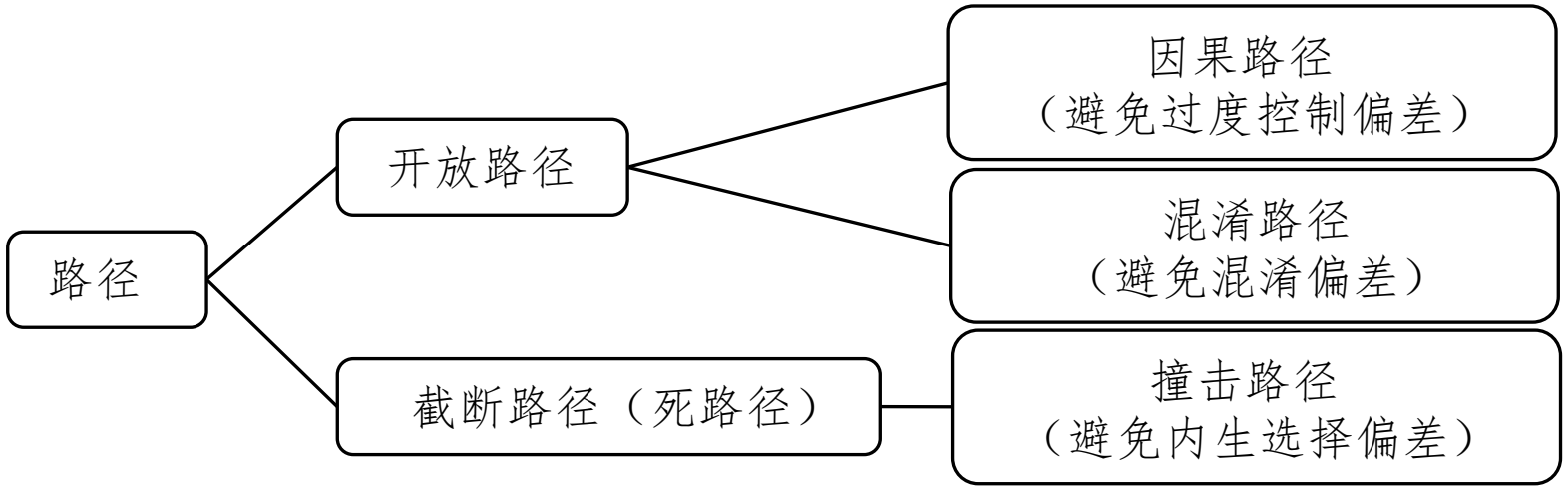
图：内生选择偏差的例子

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

3.4 小结

- 由于因果关系通常无法被直接观测到，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讲，分析因果关系的本质就是：
  - 发现因果关系
  - 截断混淆路径
  - 避免对撞路径产生的衍生路径



图：路径和偏差种类总结

4、常用的因果关系估计方法概览

- 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。
  - 处置效应模型引入新概念“潜在结果”和“处置效应”。
  - 潜在结果$=\begin{cases}Y_i(0), if\ D_i=0\\ Y_i(1), if\ D_i=1\end{cases}$，其中个体$i$接受了某种处置行为$D_i$则后果为$Y_i(1)$，反之为$Y_i(0)$。
  - D_i对个体的处置效应$=\gamma_i=Y_i(1)-Y_i(0)$，所以γ_i为D_i对Y_i的因果效用。

4.1 理想条件与实际研究

- 在理想状态下，我们可以用两种方法估计因果关系：
  - 控制实验（随机分配）：解释变量与任何其他可能的混淆变量都不相关。
  - 准自然实验：事件的发生并不是个体自己所能选择的。由于准实验法与控制实验不完全相似，因此对其数据有一定的要求，并且要使用相应的估计方法（双重差分）。

- 在实际研究中，由于我们常常面临观测数据，即数据产生不具备随机安排并且是个体自愿选择产生的。
  - 例如服药和身体健康的数据。

4.2 一个例子

- 假设我们探究教育程度（EDU）对收入（INC）的因果影响，其中性别（$GENDER$）、年龄（AGE）和其他不可观测的因素对收入都有因果影响。其中i表示个体，t表示时间。

4.2 一个例子

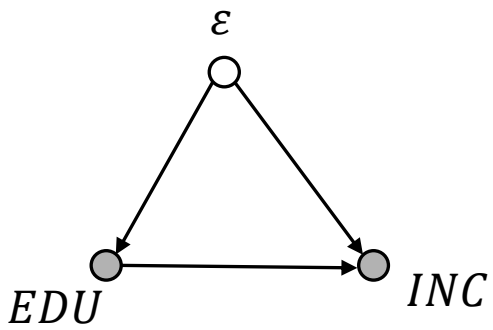
- 忽略性别和年龄的影响：

$$INC_{it} = \alpha + \beta_1 EDU_{it} + \varepsilon_{it}$$

  - 由于ε_{it}中含有年龄、性别和不可观测变量即：

$$\varepsilon_{it} = \beta_2 AGE_{it} + \beta_3 GENDER_i + e_{it}$$

ε_{it}与解释变量EDU_{it}的相关，导致了混淆偏差，所以无法识别因果影响系数β_1。



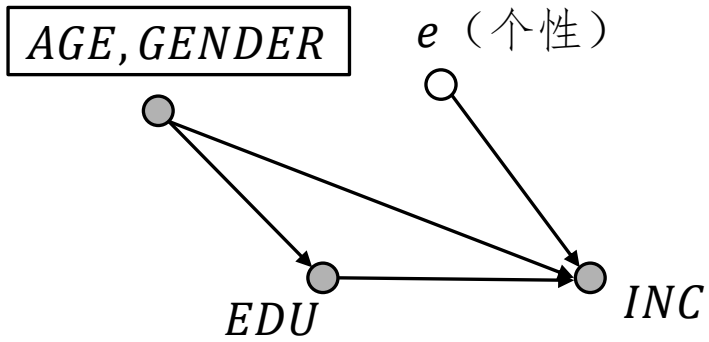
图：忽略年龄和性别的情况

4.2 一个例子

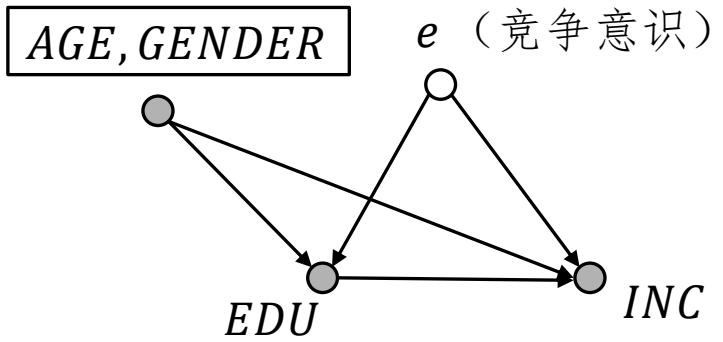
- 将残差项中的可观测变量分离进行控制

$$INC_{it} = \alpha + \beta_1 EDU_{it} + \beta_2 AGE_{it} + \beta_3 GENDER_i + e_{it}$$

  - 其中e为不可观测变量，如（个性、竞争意识）
  - EDU_{it}和AGE_{it}为可随时间变化的变量（observable time-variant variable）；$GENDER_i$为不随时间干扰的变量（observable time-invariant variable）



图：控制年龄和性别且e为无关变量的情况



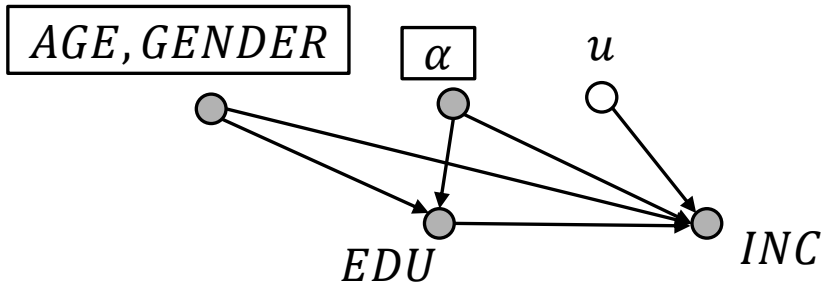
图：控制年龄和性别且e为混淆变量的情况

4.2 一个例子

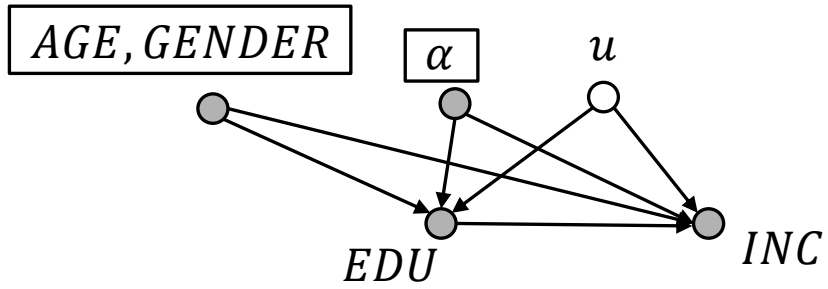
- 将残差项进一步分解为：不可观察的不随时间变化的变量α_i（unobservable time-invariant variable）和不可观测的随时间变化的变量u_{it}（ unobservable time-variant variable ），即$e_{it}= \alpha_i + u_{it}$。

$$INC_{it} = \alpha + \beta_1 EDU_{it} + \beta_2 AGE_{it} + \beta_3 GENDER_i + \alpha_i + u_{it}$$

  - 如果混淆路径是α_i造成的，我们希望控制α_i截断混淆路径。即采用面板数据分析法可以达到控制不可观测的不随时间变化的变量。



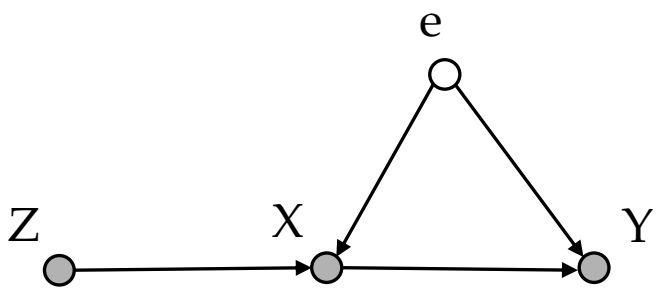
图：控制年龄、性别和α且u为无关变量的情况



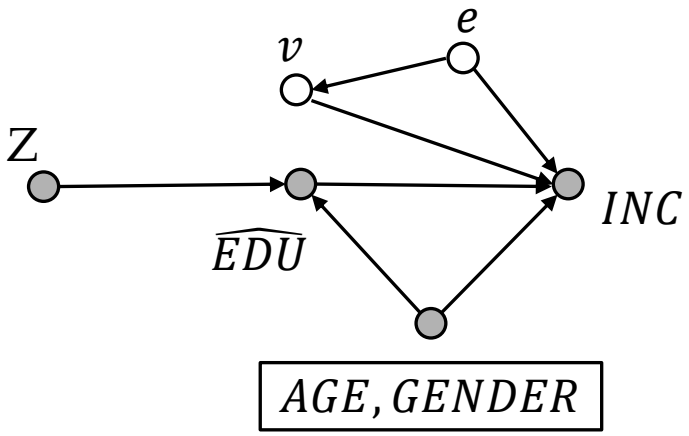
图：控制年龄、性别和α且u为混淆变量的情况

4.2 一个例子

- 引入工具变量Z_i分解出EDU_{it}变化中与e_{it}无关的部分，即$EDU_{it} = \widehat{EDU_{it}} + v_{it}$，其中$\widehat{EDU_{it}}$是$EDU_{it}$与$e_{it}$无关的部分。通过工具变量分解出解释变量中不被$e_{it}$混淆的信息来估计解释和被解释变量的因果关系。
  - 工具变量要符合两个条件：外生性和相关性
  - 这意味着Z对Y的作用 = Z对X的作用 × X对Y的作用



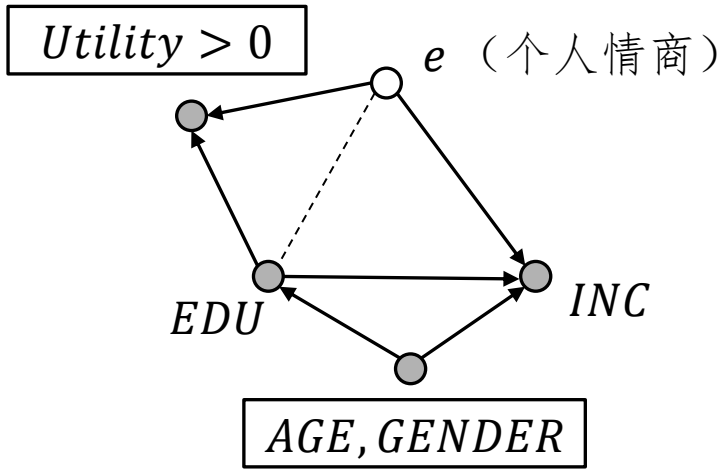
图：工具变量的相关性和外生性



图：引入工具变量的情况

4.2 一个例子

- 若存在内生选择偏差。内生选择偏差不是因为解释变量和不可观测因素e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的随机抽取，导致样本里解释变量和不可观测因素e存在相关性。
  - 由于样本中只包括了参加工作的个体，是否参加工作则有效用变量$Utility$表示。



图：产生内生选择偏差的情况

4.3 小结

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