第八章 面板数据分析方法 第一节至第四节

汇报人：陈露滢

2020/11/26

Overview

- 8.1 什么是面板数据
- 8.2 面板数据的信息来源
- 8.3 面板数据因果关系分析的直观理解
- 8.4 面板数据分析的3种常见模型

8.1 什么是面板数据

1.1 面板数据的结构

1.2 面板数据的分类

1.1 面板数据的结构

- 面板数据是包含多个个体，并且同一个体由一系列不同时间观测点的数据
  - 包含了横截面和时间序列两个维度的数据：个体维度(i = 1,2,...,N)和时间维度(t = 1,2,...,T)

ID	YEAR	INC	EDU	AGE	GENDER
1	2017	800	3	23	1
1	2018	1000	4	24	1
1	2019	1200	5	25	1
2	2017	1200	5	30	0
2	2018	1250	6	31	0
2	2019	1300	7	32	0

1.1 面板数据的结构

- 并不是所有包含个体和时间两个维度的数据都是面板数据
- 合并横截面数据：没有跟踪记录同一个个体，观测点属于不同个体
  - 可以看作是横截面数据的简单合并

YEAR	INC	EDU	AGE	GENDER
2017	800	3	23	1
2018	1000	4	24	1
2019	1200	5	25	1
2017	1200	5	30	0
2018	1250	6	31	0
2019	1300	7	32	0

1.2 面板数据的分类

- 短面板和长面板
  - 短面板是指个体维度N较大，时间维度T较小
  - 长面板是指数据的N较小，T较大
- 平衡面板与非平衡面板
  - 平衡面板：每个个体都有相同时间T的观测点
  - 非平衡面板：有部分个体没有相同时间T的观测点
  - 若非平衡面数据由随机原因造成的，那么处理方法和平衡面板一样，但是如果数据缺失由非随机原因造成的，则必须考虑缺失的原因：如样本选择偏差

8.2 面板数据的信息来源

2 面板数据的信息来源

- 两个维度的信息
  - 不同个体间的差异和同一个个体在不同时间上的差异
- 总方差可以分解为个体间方差和个体内方差

总方差（total variation）=个体间方差（between variation）+个体内方差（within variation）

2 面板数据的信息来源

$$s_0^2 = \frac{1}{NT-1}\sum_{t=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \bar{\bar{X}}\right)^2$$

$$s_B^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{X}_i - \bar{\bar{X}}\right)^2$$

$$s_w^2 = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \bar{X}_i)^2$$

$$s_0^2 \approx s_B^2 + s_W^2$$

X_{it}是个体i在时间t的值，$\bar{\bar{X}}$是X在数据里总的平均值，$\bar{X}_i$是X在个体i中的平均值。计算样本的方差，做了N-1和NT-1的调整

8.3 面板数据因果关系分析的直观理解

3 面板数据因果关系分析的直观理解

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \theta TALENT_i + \varphi LUCK_{it}$$

- 上式中i表示个体，t表示时间
- 收入INC_{it}、受教育程度EDU_{it}的值可观测并随时间变化，称为**可观测的随时间变化的变量**
- 性别$GENDER_i$可观测到，但他的值不随时间变化，称为**可观测的不随时间变化的变量**
- 个人天赋$TALENT_i$观测不到，且不随时间变化，称为**不可观测且不随时间变化的变量**
- 个人运气$LUCK_{it}$不可观测且随时间变化的，称为**不可观测且随时间变化的**

3 面板数据因果关系分析的直观理解

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \mathrm{e}_{\mathrm{it}}$$

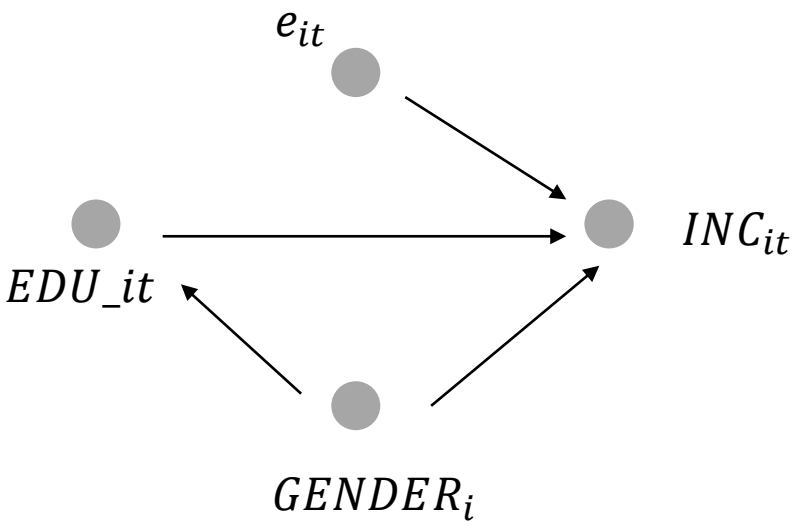
- 上式把所有不可观测的因素，包括TALENT和LUCK都归于干扰项e，那么要得到 β的正确估计 ，**需要EDU和e不相关，即EDU与天赋和运气都不相关**

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \alpha_i + u_{it}$$

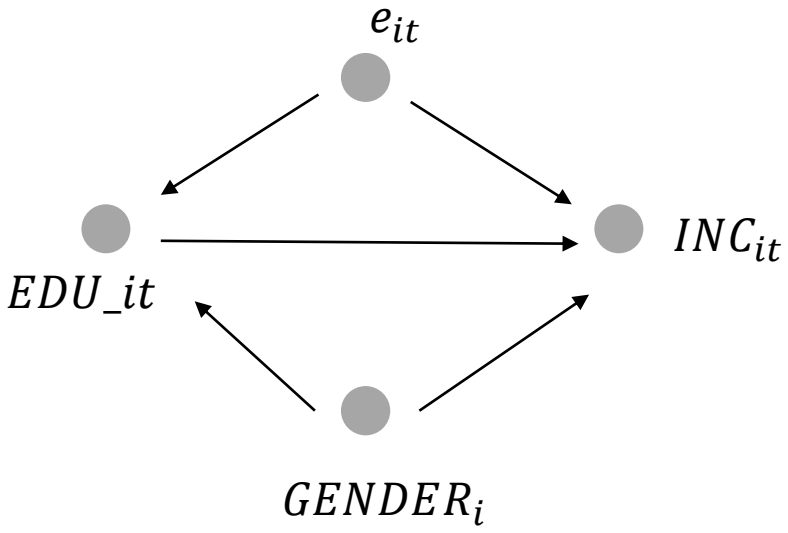
- 将干扰项e分解为α_i和u，α_i(个体效应)是个体不可观测且不随时间变化的因素$\theta TALENT_i$ ，u是个体不可观测且随时间变化的因素$\varphi LUCK_{it}$
- 此时α_i控制了天赋因素，**要正确估计$\boldsymbol{\beta}$只需要满足EDU与LUCK不相关**

3 面板数据因果关系分析的直观理解

- $INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \mathrm{e_{it}}$（简单回归模型）



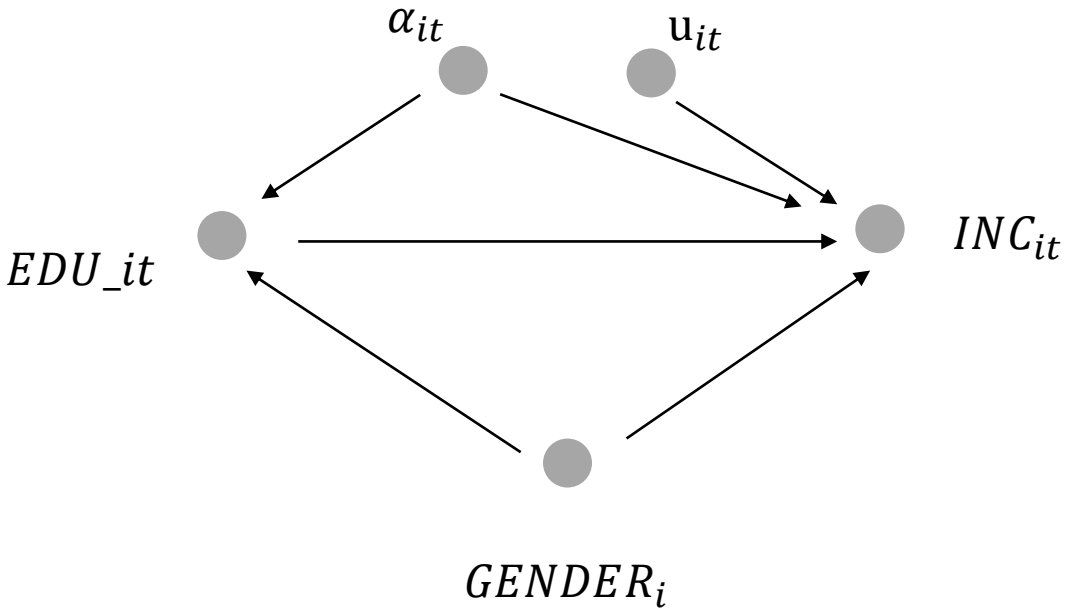
不存在混淆路径



存在混淆路径

3 面板数据因果关系分析的直观理解

- $INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \alpha_i + u_{it}$（固定效应模型）



面板数据的变量路径图：通过控制不可观测且不随时间变化的变量**截断混淆路径**

8.4 面板数据分析的3种常见模型

4 模型基本假设

- $INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \alpha_i + u_{it}$
- 需要假设 不可观测且随时间变化的变量u_{it}与可观测变量不相关
  - $\mathrm{E}(u_{it} | EDU_{it}, GENDER_{it}) = 0$
  - 在本例中，即为$\mathrm{E}(\mathrm{LUCK}_{it} | EDU_{it}, GENDER_{it}) = 0$
- 三个模型的关键差别在于：对个体不可观测且不随时间变化的变量α_i的假设

4.1 合并横截面模型

- 假设α_i不存在，即不存在会影响收入的不可观测且不随时间变化的因素，本例中$\theta TALENT_i$假设为零

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + u_{it}$$

- 假设此时也满足$\mathrm{E}(u_{it}|\, EDU_{it}, GENDER_{it}\,) = 0$
- 这只是简单的横截面数据在时间上的叠加
- 若α_i此时存在并且与可观测变量相关，就会导致缺失变量问题

4.2 随机效应模型

- 假设α_i存在，但α_i与可观测变量不相关，即$\mathrm{E}(\alpha_{it}|\, EDU_{it}, GENDER_{it}) = 0$，此时将$\alpha_i$放进干扰项不会造成估计误差

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \mathrm{e_{it}}$$

$$e_{it} = \alpha_i + u_{it}$$

- 由于同一个个体的干扰项e_{it}在不同时间包含了相同的α_i，即干扰项在同一个个体内是相关的，其相关系数为:(其中σ_α^2是α_i的方差，σ_u^2是u_{it}的方差)

$$Corr(e_{it}, e_{it-s}) = Corr(\alpha_i + u_{it}, \alpha_i + u_{it-s})$$

$$= \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_u^2)$$

4.2 随机效应模型

- 当干扰项已知时，采用GLS估计，现将模型转换为同方差：

$$INC_{it}^{*} = \alpha^{*} + \beta EDU_{\mathrm{it}}^{*} + \gamma GENDER_{i}^{*} + \mathrm{e}_{\mathrm{it}}^{*}$$

$$\mathrm{INC}_{it}^{*} = \mathrm{INC}_{\mathrm{it}} - \theta\overline{INC_{i}}$$

$$\theta = 1 - \frac{\sigma_{\alpha}}{\sqrt{\sigma_{\alpha}^{2} + T\sigma_{u}^{2}}}$$

$$e_{it}^{*} = e_{it} - \theta\overline{e_{i}}$$

- 对转换之后的模型使用OLS，估计出来的$\widehat{\beta^{RE}}$，在α_i与可观测值无关的情况下是无偏、一致且有效的估计量

4.3 固定效应模型

- 假设α_i存在，且α_i与可观测变量相关，即$\mathrm{E}(\alpha_{it}|\, EDU_{it}, GENDER_{it}) \neq 0$，需要把$\alpha_i$看作解释变量处理
- 若没有把α_i看作解释变量，而是作为干扰项的一部分处理，即$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \mathrm{e}_{\mathrm{it}}$
- 对上述模型求条件期望值：

4.3 固定效应模型

$$E(INC_{it}\,|EDU_{it}, GENDER_i)$$

$$= \alpha + \beta EDU_{it} + \gamma GENDER_i + E(e_i|EDU_{it}, GENDER_i)$$

$$= \alpha + \beta EDU_{it} + \gamma GENDER_i + E(\alpha_i + u_{it}|EDU_{it}, GENDER_i)$$

$$= \alpha + \beta EDU_{it} + \gamma GENDER_i + E(\theta TALENT_i + \varphi LUCK\_it|EDU_{it}, GENDER_i)$$

$$= \alpha + \beta EDU_{it} + \gamma GENDER_i + \theta E(TALENT_i|EDU_{it}, GENDER_i)$$

- 假设TALENT$_i$与可观测变量EDU_{it}、$GENDER_i$的相关关系表示如下：

$$E(TALENT_i|EDU_{it}, GENDER_i) = \phi_o + \phi_1 EDU_{it} + \phi_2 GENDER_i$$

4.3 固定效应模型

- 将相关关系代入原式得

$$E(INC_{it}\,|EDU_{it}, GENDER_i) = (\alpha + \theta\phi_0) + (\beta + \theta\phi_1)EDU_{it} + (\gamma + \theta\phi_2)GENDER_i$$

- 可以看出，若将α_i看作干扰项，估计出来的系数是$\beta + \theta\phi_1$，而不是β，会有缺失变量$TALENT_i$对INC_{it}及与EDU_{it}相关的部分
- 存在缺失变量误差
- 因此要把α_i作为解释变量来使用模型

4.3 固定效应模型

$$INC_{it} = \alpha + \beta EDU_{\mathrm{it}} + \gamma GENDER_i + \alpha_i + u_{it}$$

解释变量　　干扰项

- 固定效应模型：认为每个个体对应一个与其他解释变量相关的固定个体效应α_i
- 随机效应模型：认为个体效应是从某个分布中随机抽取，与其他解释变量无关（不太合理）