2.线性回归-理解篇

相耐汀

本章框架

- 2.1：线性回归模型满足条件
- 2.2：介绍求解线性回归函数系数的最小二乘法
- 2.3：运用图形直观说明最小二乘法估计模型系数的机理
- 2.4介绍多元线性回归分解法

2.1线性回归模型，条件期望函数与因果推断

2.1.1解释变量、被解释变量与干扰项

- 假设我们研究受教育程度（EDU）对与收入水平（INC）的因果影响程度：

$$INC = \alpha + \beta_1 EDU + \beta_2 IQ + e$$

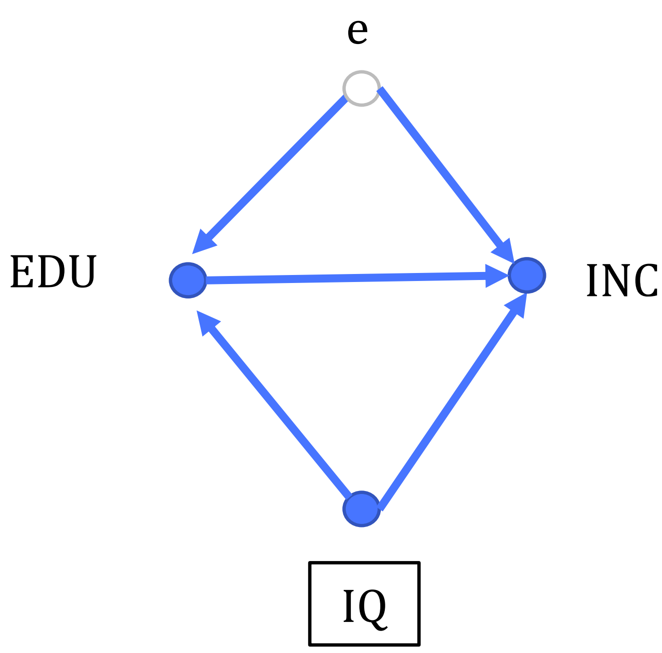
- 其中：
  - EDU处置变量
  - IQ控制变量
  - 其他无法观测到的但会影响INC的变量称为干扰项e

2.1.1解释变量、被解释变量与干扰项

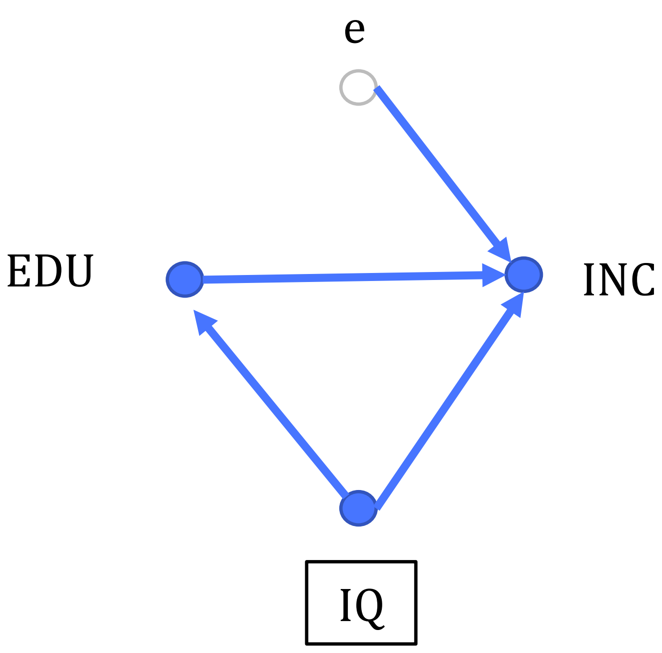
- 若干扰项e包含混淆变量，则无法识别EDU对INC的因果影响有多大，即β_1。
  - 人格特性：勤勉

2.1.1解释变量、被解释变量与干扰项

- 变量路径图



干扰项与解释变量相关



干扰项与解释变量不相关

2.1.2因果关系条件期望函数

- 若干扰项条件均值独立于解释变量，此时可识别出因果影响系数：

$$\mathbb{E}(e|EDU,IQ) = \mathbb{E}(e) = c$$

- 例：对于同样IQ=iq的人，EDU从10到11，干扰项e均值不变，INC均值变化为：

$$\begin{aligned}
&\mathbb{E}\left(INC|EDU = 11, IQ = iq\right) \\
&- \mathbb{E}(INC|EDU = 10, IQ = iq) \\
&= [\alpha + \beta_1 11 + \beta_2 iq + \mathbb{E}(e|EDU = 11, IQ = iq)] \\
&- [\alpha + \beta_1 10 + \beta_2 iq + \mathbb{E}(e|EDU = 10, IQ = iq)] = \beta_1
\end{aligned}$$

2.1.2因果关系条件期望函数

- 由于线性回归模型里包含常数项，所以我们可以把$\mathbb{E}(e|EDU,IQ)=\mathbb{E}(e)=c$的常数c并入常数项，使：

$$\mathbb{E}(e|EDU,IQ)=\mathbb{E}(e)=0$$

- 综上所述，线性回归模型要满足以下两个假设：
  - 线性关系假设：

$$INC=\alpha+\beta_1 EDU+\beta_2 IQ+e$$

  - 干扰项条件均值为0假设：

$$\mathbb{E}(e|EDU,IQ)=\mathbb{E}(e)=0$$

2.1.2因果关系条件期望函数

- 对线性函数两边取期望值，得到线性条件期望函数CEF：

$$
\begin{aligned}
&\mathbb{E}(INC|EDU,IQ)\\
&=\alpha+\beta_1 EDU+\beta_2 IQ+\mathbb{E}(e|EDU,IQ)\\
&=\alpha+\beta_1 EDU+\beta_2 IQ
\end{aligned}
$$

- 将条件期望函数对EDU求偏导：偏回归系数

$$\frac{d\mathbb{E}(INC|EDU,IQ)}{dEDU}=\beta_1$$

2.1.3相关关系条件期望函数

- 若干扰项均值独立于解释变量的假设不成立，干扰项的存在如何影响因果关系的估计？
- 假设只观测到了INC和EDU：

$$INC = \alpha + \beta_1 EDU + \varepsilon,\quad \varepsilon = \beta_2 IQ + e$$

- 此时：

$$\mathbb{E}(\varepsilon|EDU) = \mathbb{E}(\beta_2 IQ + \mathrm{e}|EDU) = \beta_2 \mathbb{E}(IQ|EDU) \neq 0$$

- 此时模型为：

$$INC = \alpha + \beta_1 EDU + \varepsilon, \mathbb{E}(\varepsilon|EDU) \neq 0$$

2.1.3相关关系条件期望函数

- 若我们误以为干扰项与解释变量不相关，我们就“扭曲”了干扰项,此时我们称u为伪干扰项：

$$INC = \gamma_0 + \gamma_1 EDU + u, \mathbb{E}(u|EDU = 0)$$

- 通过CEF来理解γ_1和β_1的关系：

$$\mathbb{E}(INC|EDU) = \gamma_0 + \gamma_1 EDU$$

  - 求偏导：

$$\frac{d\mathbb{E}(INC|EDU)}{dEDU} = \gamma_1$$

- γ_1反映了INC 的期望值如何随EDU变化，但并没有控制IQ不变。

2.1.3相关关系条件期望函数

- 计算γ_1和β_1的关系：
- $\mathbb{E}(INC|EDU) = \mathbb{E}(\alpha + \beta_1 EDU + \beta_2 IQ + \mathrm{e}|EDU) = \alpha + \beta_1 EDU + \beta_2 \mathbb{E}(IQ|EDU)$
  - 对EDU求导：

$$\frac{d\mathbb{E}(INC|EDU)}{dEDU} = \beta_1 + \beta_2 \frac{d\mathbb{E}(IQ|EDU)}{dEDU}$$

  - 即

$$\gamma_1 = \beta_1 + \beta_2 \frac{d\mathbb{E}(IQ|EDU)}{dEDU}$$

2.1.3相关关系条件期望函数

- 假设受教育程度于智商之间存在线性相关关系：

$$\mathbb{E}(IQ|EDU) = \phi_0 + \phi_1 EDU$$

  - 即

$$\frac{d\mathbb{E}(IQ|EDU)}{dEDU} = \phi_1$$

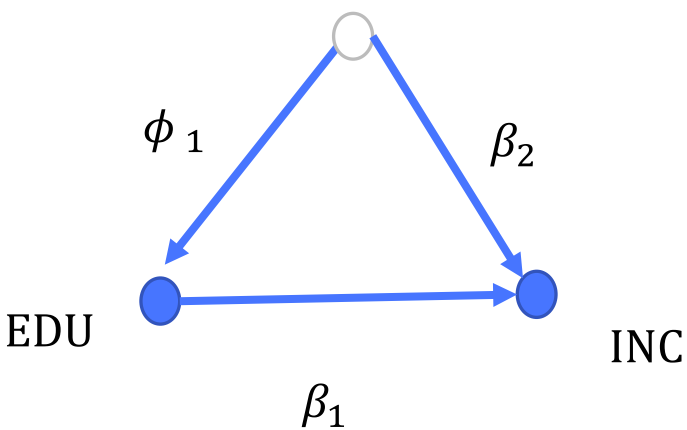
  - 则

$$\gamma_1 = \beta_1 + \beta_2 \phi_1$$

2.1.3相关关系条件期望函数

- 由此可见，γ_1反映了EDU与INC 的相关性，包含了：
  - EDU对INC的因果影响β_1
  - EDU与IQ的相关性ϕ_1乘以IQ对INC的因果影响β_2。

$$\varepsilon = \beta_2 IQ + \mathrm{e}$$



2.2最小二乘法

2.2.1总体最小二乘法

- 假设我们要估计的模型是：

$$Y = \boldsymbol{X}'\boldsymbol{\beta} + \epsilon, \mathrm{E}(\epsilon|X) = 0$$

- 展开：$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$
- 回归模型对应的CEF：

$$\mathbb{E}(Y|X) = \boldsymbol{X}'\boldsymbol{\beta}$$

2.2.1总体最小二乘法

- 利用最小二乘法求解系数$\hat{\beta}$

$$\widehat{\boldsymbol{\beta}} = \underset{b}{argmin}\, \mathbb{E}[(Y - \boldsymbol{X}'\boldsymbol{b})^2]$$

由一阶条件可得：

$$\mathbb{E}\left[\boldsymbol{X}(Y - \boldsymbol{X}'\widehat{\boldsymbol{\beta}})\right] = 0$$

此条件同等与：

$$\mathbb{E}\left[\boldsymbol{X}(Y - \boldsymbol{X}'\widehat{\boldsymbol{\beta}})\right] = \mathbb{E}[\boldsymbol{X}\hat{\epsilon}] = 0$$

- 由此可见，最小二乘法的本质是求解系数$\widehat{\boldsymbol{\beta}}$，使得解释变量$\mathbf{X}$与残差$\hat{\epsilon}$不相关

$$\widehat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}']^{-1}\mathbb{E}[\boldsymbol{X}Y]$$

2.2.1总体最小二乘法

- 将线性回归模型代入上式：

  $$\widehat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}']^{-1}\mathbb{E}[\boldsymbol{X}Y] = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}']^{-1}\mathbb{E}[\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\beta} + \epsilon)]$$
  $$= \beta + \mathbb{E}[\boldsymbol{X}\boldsymbol{X}']^{-1}\mathbb{E}[\boldsymbol{X}\epsilon]$$

- 其中：$\mathbb{E}(\epsilon|X) = 0$
  - 故$\mathbb{E}[\boldsymbol{X}\epsilon] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}(\boldsymbol{X}\epsilon|\boldsymbol{X})] = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\mathbb{E}(\epsilon|\boldsymbol{X})] = \boldsymbol{0}$
  - 故$\widehat{\boldsymbol{\beta}}$=$\boldsymbol{\beta}$

- 以上讨论说明，最小二乘法的解对应的系数$\widehat{\boldsymbol{\beta}}$是模型$Y = \boldsymbol{X}'\boldsymbol{\beta} + \epsilon, \mathbb{E}(\epsilon|X) = 0$，即条件期望函数$\mathbb{E}(Y|X) = \boldsymbol{X}'\boldsymbol{\beta}$的系数$\boldsymbol{\beta}$

2.2.2干扰项和残差

- 干扰项ϵ包含了除解释变量以外的其他影响被解释变量的因素，与解释变量是否相关无法检验
- 残差$\hat{\epsilon}$使用最小二乘法算出来的，总是与解释变量不相关。

$$\mathbb{E}\left[\boldsymbol{X}(Y-\boldsymbol{X}'\widehat{\boldsymbol{\beta}})\right]=\mathbb{E}[\boldsymbol{X}\hat{\epsilon}]=0$$

- 只有当干扰项与解释变量不相关时，残差才是干扰项的正确估计。