线性回归运用篇——固定解释变量和随机解释变量

李非凡

2020.10.29

内容框架

- 1.固定解释变量和随机解释变量
- 2.固定解释变量下的回归模型假设
- 3.随机解释变量下的回归模型假设
- 4.小结

1.1 固定解释变量和随机解释变量

- （1）固定解释变量
- 解释变量在重复抽样中，值固定
- 常出现在控制实验中，实验条件可人为固定
- 例：解释变量X：水量和施肥量(可控)

  被解释变量Y：产量
- 若解释变量X固定，则被解释变量Y随机变化就不是由解释变量X造成的，而完全是由干扰项造成
- （2）随机解释变量
- 解释变量在重复抽样中，值随机

1.2 回归模型假设差异

- 通过实验或抽样得到N个观测点，每个观测点i的解释变量、被解释变量和干扰项满足以下线性关系：
- $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + e_i$
- k:第k个解释变量
- i:第i个观测点

- 上式可简化为矩阵：$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$

1.2 回归模型假设差异

固定解释变量	随机解释变量
1.$\boldsymbol{X}$是固定的	1.$\boldsymbol{X}$是随机的或随机和固定混合的
2.$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$	2.$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$
3.$\mathrm{E}[\boldsymbol{e}] = 0$	3.$\mathrm{E}[\boldsymbol{e}\|\boldsymbol{X}] = 0$
4.$\mathrm{E}(\boldsymbol{ee}') = \sigma^2\boldsymbol{I}$	4.$\mathrm{E}(\boldsymbol{ee}'\|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
$\mathrm{Var}(e_i) = \sigma^2$	$\mathrm{Var}(e_i\|\boldsymbol{X}) = \sigma^2$
$\mathrm{Cov}(e_i, e_j) = 0$	$\mathrm{Cov}(e_i, e_j\|\boldsymbol{X}) = 0$
5.$\boldsymbol{e}\sim\mathrm{N}(0, \sigma^2\boldsymbol{I})$	5.$\boldsymbol{e}\|\boldsymbol{X}\sim\mathrm{N}(0, \sigma^2\boldsymbol{I})$
6.解释变量不存在共线性并且观测点多于解释变量数	6.解释变量不存在共线性并且观测点多于解释变量数

2. 固定解释变量下的回归模型假设（6个）

- 假设1：解释变量X是固定的
- 对于N个观测点，每个观测点的解释变量X_i都是固定值

- 一个例子：温度对细菌生长的影响
- $X=$温度,$\mathrm{Y}=$细菌生长速度
- 第一次实验：$(30,$细菌生长速度$_1),(60,$细菌生长速度$_2)$
- 第二次实验：$(30,$细菌生长速度$'_1),(60,$细菌生长速度$'_2)$
- 固定解释变量：温度$(30°$和$60°)$
- 细菌生长速度随机原因：除温度外，还受其他观测不到的随机干扰因素影响（e）

2. 固定解释变量下的回归模型假设（6个）

- 假设2：$\boldsymbol{Y = X\beta + e}$
- 被解释变量是解释变量的线性函数加上一个干扰项

- 线性关系：被解释变量系数
- $Y = \beta_0 + \beta_1 X + \mu$
- $Y = \beta_0 + \beta_1 \sin(X) + \nu$
- $Y = \beta_0 + \beta_1 \log(x) + \omega$
- 非线性:$Y = \beta_0 + \beta_1 X^{\beta_2} + \varepsilon$
- 转换：$Y = \mathrm{A}X^{\beta} \exp(e) \Longrightarrow \log(Y) = \log(A) + \beta \log(X) + e$

2. 固定解释变量下的回归模型假设（6个）

- 假设3： $\mathrm{E}[\boldsymbol{e}] = 0$
- 每个观测点i的干扰项的均值为0
- 作用：解释变量固定，通过重复实验去除干扰项的影响$\Longrightarrow$得到解释变量和被解释变量的因果关系
- 例子：
- 单次实验：
- AB两点的虚线斜率

2. 固定解释变量下的回归模型假设（6个）

- 例子：
- 单次实验：AB两点的虚线斜率$\widehat{\beta_1}$

≡单次实验估计细菌生长速率和温度的线性关系系数

$=[(\text{细菌生长速度}_2)-(\text{细菌生长速度}_1)]/(\text{温度}_2-\text{温度}_1)$

$=[(\alpha+\beta_1 60+e_2)-(\alpha+\beta_1 30+e_1)]/(60-30)$

$=\beta_1+\underbrace{(e_2-e_1)/(60-30)}_{\text{单次实验干扰造成的误差}\neq 0}$

2. 固定解释变量下的回归模型假设（6个）

- 重复实验：CD两点的实线斜率$\mathrm{E}(\widehat{\beta_1})$

≡重复实验估计细菌生长速率和温度的线性关系系数

$=[\mathrm{E}(\text{细菌生长速度}_2)-\mathrm{E}(\text{细菌生长速度}_1)]/(\text{温度}_2-\text{温度}_1)$

$=[\mathrm{E}(\alpha+\beta_1 60+e_2)-\mathrm{E}(\alpha+\beta_1 30+e_1)]/(60-30)$

$=\beta_1+\underbrace{[\mathrm{E}(e_2)-\mathrm{E}(e_1)]/(60-30)}_{\text{重复实验干扰造成的误差=0}}$

$=\beta_1$

结论：重复实验得到的系数是单次实验所得系数$\widehat{\beta_1}$的均值，且这个均值等于真实值

2. 固定解释变量下的回归模型假设（6个）

- 不存在缺失解释变量：干扰项均值为0，包含除解释变量外所有其他不可观测但会影响被解释变量的因素
- 存在缺失解释变量：干扰项均值不为0，并且不同观测点干扰项的均值不同，影响因果关系的估计
- 例：假设湿度也会影响细菌生长，温度和适度正相关

那么，重复实验中，

观测点2（60）的平均湿度>观测点1（30）的平均湿度

湿度包含在干扰项里，因此有$\mathrm{E}(e_2) = a_2 > \mathrm{E}(e_1) = a_1$

2. 固定解释变量下的回归模型假设（6个）

- 重复实验得到估计系数的均值为：$\mathrm{E}(\widehat{\beta_1})$

≡重复实验估计细菌生长速率和温度的线性关系系数

$=[\mathrm{E}(\text{细菌生长速度}_2)-\mathrm{E}(\text{细菌生长速度}_1)]/(\text{温度}_2-\text{温度}_1)$

$=[\mathrm{E}(\alpha+\beta_1 60+e_2)-\mathrm{E}(\alpha+\beta_1 30+e_1)]/(60-30)$

$=\beta_1+[\mathrm{E}(e_2)-\mathrm{E}(e_1)]/(60-30)$

$=\beta_1+\underbrace{[a_2-a_1]/(60-30)}_{\text{重复实验干扰项造成的误差}\neq 0}$

$\neq\beta_1$

2. 固定解释变量下的回归模型假设（6个）

- $\mathrm{E}(\widehat{\beta_1}) \neq \beta_1$，原因：干扰项造成的误差均值不等于0
- 结论：当干扰项均值会随着解释变量值发生变化时，无法通过重复抽样消除干扰项的影响，解释变量和被解释变量的因果关系估计将受到随解释变量变化的干扰项的混淆
- 解释变量固定时，识别解释变量和被解释变量因果关系的关键：干扰项均值为0
- 解释变量固定时，不存在反向因果问题和解释变量侧脸发误差的内生性问题，此时$\mathrm{E}[\boldsymbol{e}] = 0$的假设可认为是“不存在缺失解释变量的假设”

2. 固定解释变量下的回归模型假设（6个）

- 假设4： $\mathrm{E}(\boldsymbol{ee}') = \sigma^2 \boldsymbol{I}$

- 展开理解： $\mathrm{E}(\boldsymbol{ee}') = \mathrm{E}\left(\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \cdots \\ e_N \end{bmatrix} \times [e_1 \quad e_2 \quad e_3 \quad \cdots \quad e_N] \right)$

$$= \begin{bmatrix} \mathrm{E}(e_1e_1) & \mathrm{E}(e_1e_2) & \mathrm{E}(e_1e_3) & \ldots & \mathrm{E}(e_1e_N) \\ \mathrm{E}(e_2e_1) & \mathrm{E}(e_2e_2) & \mathrm{E}(e_2e_3) & \cdots & \mathrm{E}(e_2e_N) \\ \mathrm{E}(e_3e_1) & \mathrm{E}(e_3e_2) & \mathrm{E}(e_3e_3) & \vdots & \mathrm{E}(e_3e_N) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{E}(e_Ne_1) & \mathrm{E}(e_Ne_2) & \mathrm{E}(e_Ne_3) & \cdots & \mathrm{E}(e_Ne_N) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \ldots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2N} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots & \sigma_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{N1} & \sigma_{N2} & \sigma_{N3} & \cdots & \sigma_N^2 \end{bmatrix}$$

2. 固定解释变量下的回归模型假设（6个）

- 若$\mathrm{E}(\boldsymbol{ee}') = \sigma^2\boldsymbol{I}$，则$\sigma_{ij} = 0(i \neq j)$

- $$\mathrm{E}(\boldsymbol{ee}') = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \ldots \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \cdots \sigma_{2N} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \cdots \sigma_{3N} \\ \cdots & \cdots & \cdots \cdots \cdots \\ \sigma_{N1} & \sigma_{N2} & \sigma_{N3} \cdots \sigma_N^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \cdots 0 \\ 0 & \sigma_2^2 & 0 \cdots 0 \\ 0 & 0 & \sigma_3^2 \cdots 0 \\ \cdots & \cdots & \cdots \cdots \cdots \\ 0 & 0 & 0 \cdots \sigma_N^2 \end{bmatrix} = \sigma^2\boldsymbol{I}$$

2. 固定解释变量下的回归模型假设（6个）

- 包含两个子假设：
- （1）同方差： $\mathrm{Var}(e_i) \equiv \sigma_i^2 = \sigma^2$
- 每个观测点干扰项的方差相同:干扰项的变化范围
- （2）不相关： $\mathrm{Cov}(e_i, e_j) \equiv \sigma_{ij} = 0$
- 不同观测点干扰项的协方差为0：一个观测点干扰项的信息不包含预测另一个观测点干扰项的信息
- 然而，这个假设在实际情况中通常不成立，不同观测点的干扰因素方差可能违背该假设会导致异方差和自相关问题

2. 固定解释变量下的线性回归假设（6个）

- 假设5： $\boldsymbol{e} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I})$
- 每个观测点的干扰项满足正态分布
- 假设3+4+5：每个干扰项满足均值为0、方差为σ^2的正态分布$e_i \sim \mathrm{N}(0, \sigma^2)$
- 正态分布的假设并不是必需的，只是在小样本里进行系数检验时需要

2. 固定解释变量下的回归模型假设（6个）

- 假设6：解释变量之间不存在共线性，并且观测点数量大于被解释变量数
- 这是能够通过N个观测点得到k个解释变量系数的必要条件：
- (1)如果解释变量间存在共线性，就无法分辨出所有解释变量系数的值。例：
- 投资 $= \alpha + \beta_1$经营性收入 $+\beta_2$ 非经营性收入 $+ \beta_3$ 总收入 $+ e$
- (2)如果观测点数少于解释变量数k，则意味着方程数量小于k，就无法求解k个变量的系数

3. 随机解释变量下的回归模型假设（6个）

- 假设1：解释变量$\boldsymbol{X}$是随机的，或随机和固定混合的
- 每个观测点i的解释变量值$\boldsymbol{X}_i$是从某个整体分布中随机抽取的
- 例：解释变量 温度
- 观测点1可能为30°或60°，观测点2也可能为30°或60°
- 与之前观测点1温度固定为30°、观测点2温度固定为60°不同

3. 随机解释变量下的回归模型假设（6个）

- 假设2：$\boldsymbol{Y = X\beta + e}$
- 线性关系的假设和固定变量下的假设没有本质区别，只是解释变量不再是固定的

3. 随机解释变量下的回归模型假设（6个）

- 假设3： $\mathrm{E}[\boldsymbol{e}|\boldsymbol{X}]=0$
- 对于任何给定值的解释变量$\boldsymbol{X}=\boldsymbol{x}$，干扰项$\boldsymbol{e}$的期望值为$0$
- 干扰项均值独立于解释解释变量：每个观测点的干扰项的均值不会随着解释变量的均值的变化而变化，即对任意一个观测点i，干扰项满足：

$$\mathrm{E}(e_i|X_1,\cdots,X_i,\cdots,X_N)=0$$

- 上式的意思是，对于给定所有观测点的解释变量值$(X_1=x_1\ \cdots\ X_i=x_i\ \cdots X_N=x_N)$，任何一个观测点$i$的干扰项$e_i$的均值为$0$

3. 随机解释变量下的回归模型假设（6个）

- $\mathrm{E}(e_i|\boldsymbol{X}) = 0$意味着$\mathrm{Cov}(e_i, \boldsymbol{X}) = 0$，即:
- $\mathrm{Cov}(e_i, X_1) = \cdots = \mathrm{Cov}(e_i, X_i) = \mathrm{Cov}(e_i, X_N) = 0$
- 故这个假设可简单理解为：观测点i的干扰项e_i与其他观测点的解释变量都不相关。换言之，任何观测点的解释变量都不包含干扰项的信息
- 这个假设也称为强外生假设

3. 随机解释变量下的回归模型假设（6个）

- 例子：
- 温度与细菌生长的例子中，有两个观测点，即$N = 2$
- $\mathrm{E}(e|\boldsymbol{X}) = 0$包含两个条件：
- $\mathrm{E}(e_1|温度_1, 温度_2) = 0$，$\mathrm{E}(e_2|温度_1, 温度_2) = 0$
- 即：观测点1的干扰因素与两个观测点的温度不相关，观测点2的干扰因素与两个观测点的温度不相关
- 这个假设的意义在于：在重复抽样里，无论两个观察点的温度是多少，干扰项均值都为0，这样就保证了每个观测点干扰项的均值不会随其他观测点温度的变化而变化，从而能够通过温度变化和细菌生长变化的相关性得到二者的因果关系

3. 随机解释变量下的回归模型假设（6个）

- 具体来看这个假设对因果关系的估计作用：
- 一次抽样得到的温度和细菌生长速度的系数估计值为：
- $$\widehat{\beta_1}=\frac{(\text{细菌生长速度}_2)-(\text{细菌生长速度}_1)}{\text{温度}_2-\text{温度}_1}$$

$$=\frac{(\alpha+\beta_1\text{温度}_2+e_2)-(\alpha+\beta_1\text{温度}_1+e_1)}{\text{温度}_2-\text{温度}_1}$$

$$=\beta_1+\underbrace{\frac{(e_2-e_1)}{\text{温度}_2-\text{温度}_1}}_{\text{单次实验干扰造成的误差}\neq 0}$$

- 一次抽样估计值和真实值β_1之间存在误差项

3. 随机解释变量下的回归模型假设（6个）

- 重复抽样后，系数估计值的均值为：

- $$\mathrm{E}(\widehat{\beta_1})=\beta_1+\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}\right)$$

$$=\beta_1+\mathrm{E}_{(温度_2,温度_1)}\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}\middle|温度_2,温度_1\right)$$

$$=\beta_1+\mathrm{E}_{(温度_2,温度_1)}\left(\frac{\mathrm{E}\left(e_2\middle|温度_1,温度_2\right)-\mathrm{E}\left(e_1\middle|温度_1,温度_2\right)}{温度_2-温度_1}\right)$$

$$=\beta_1+\mathrm{E}_{(温度_2,温度_1)}(0)$$

$$=\beta_1$$

3. 随机解释变量下的回归模型假设（6个）

- 这个证明用到了期望迭代法则$\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}\right)=\mathrm{E}_{(温度_2,温度_1)}\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}|温度_2,温度_1\right)$
- 计算过程：

(1)先给定温度，求其条件均值

$$\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}|温度_2,温度_1\right)=0$$

(2)再根据温度的分布，求条件均值的均值

$$\mathrm{E}_{(温度_2,温度_1)}\mathrm{E}\left(\frac{(e_2-e_1)}{温度_2-温度_1}|温度_2,温度_1\right)=0$$

3. 随机解释变量下的回归模型假设（6个）

- $\mathrm{E}[\boldsymbol{e}|\boldsymbol{X}]=0$保证了我们可以通过重复变量，使得误差项均值为0，从而估计系数的均值等于真实值，即系数估计时无偏的估计
- 这个条件对于独立抽样的样本可以进行简化：

  独立抽样⇒不同观测点信息不相关⇒ $\mathrm{E}[e_i|\boldsymbol{X}_j]=0, i\neq j$

  则只需要假设同一观测点的观测点和解释变量不相关，即 $\mathrm{E}[e_i|\boldsymbol{X}_i]=0$
- 在有时间序列的数据里，不同期的解释变量可能包含当期干扰项的信息，要避免干扰项对因果关系的估计，会要求强外生假设。例如：企业上一季的企业规模可能和当季的干扰项相关

3. 随机解释变量下的回归模型假设（6个）

- 假设4：$\mathrm{E}(\boldsymbol{ee}'|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
- 对于任何给定值的解释变量$\boldsymbol{X}$的值，所有干扰项具有相同的条件方差且相互条件独立
- 假设4意味着干扰项的方差和相关性不随着解释变量值的改变而改变，便于判断样本估计系数的准确度
- 实际情况中，该假设一般不成立，违背该假设会导致异方差和自相关问题

3. 随机解释变量下的回归模型假设（6个）

- $\mathrm{E}(\boldsymbol{ee}'|\boldsymbol{X})$

$$= \begin{bmatrix} \mathrm{E}(e_1e_1|\boldsymbol{X}) & \mathrm{E}(e_1e_2|\boldsymbol{X}) & \mathrm{E}(e_1e_3|\boldsymbol{X}) & \dots & \mathrm{E}(e_1e_N|\boldsymbol{X}) \\ \mathrm{E}(e_2e_1|\boldsymbol{X}) & \mathrm{E}(e_2e_2|\boldsymbol{X}) & \mathrm{E}(e_2e_3|\boldsymbol{X}) & \cdots & \mathrm{E}(e_2e_N|\boldsymbol{X}) \\ \mathrm{E}(e_3e_1|\boldsymbol{X}) & \mathrm{E}(e_3e_2|\boldsymbol{X}) & \mathrm{E}(e_3e_3|\boldsymbol{X}) & \vdots & \mathrm{E}(e_3e_N|\boldsymbol{X}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathrm{E}(e_Ne_1|\boldsymbol{X}) & \mathrm{E}(e_Ne_2|\boldsymbol{X}) & \mathrm{E}(e_Ne_3|\boldsymbol{X}) & \cdots & \mathrm{E}(e_Ne_N|\boldsymbol{X}) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\boldsymbol{I}$$

3. 随机解释变量下的回归模型假设（6个）

- 假设5： $\boldsymbol{e}|\boldsymbol{X}\sim \mathrm{N}(0,\sigma^2\boldsymbol{I})$
- 对于任何给定值的解释变量$\boldsymbol{X}$的值，干扰项的条件分布服从正态分布$\boldsymbol{e}|\boldsymbol{X}\sim \mathrm{N}(0,\sigma^2\boldsymbol{I})$
- 假设5意味着干扰项的条件分布不随解释变量值的改变而改变
- 这个假设也不是必需的，只是在小样本里进行系数检验时需要

3. 随机解释变量下的回归模型假设（6个）

- 假设6：解释变量之间不存在共线性，并且样本数量不少于被解释变量数
- 这个假设是为了保证能够得到 k 个解释变量的系数解，和固定变量下的要求没有区别

4.小结

- 线性回归模型的6个假设
- 固定解释变量和随机解释变量下线性回归模型假设的最大区别：对干扰项的假设不再是无条件的，以给定解释变量的某个值为条件
- 固定解释变量和随机解释变量假设条件的本质无区别
- 求解样本估计量$\hat{\beta}$的性质：
- 1.固定解释变量下：把固定解释变量当作常数
- 2.随机解释变量下：分两步
- (1)先得到给定X值情况下的结果
- (2)对不同X值的结果按X值的概率分布取平均值