七、虚拟变量与随机解释变量

目录

虚拟解释变量

- 虚拟解释变量
  - 含义：反映定性 (或属性) 因素变化，取值为 0 和 1 的人工变量，也称哑变量 D
  - 作用：作为定性 (或属性) 因素的代表，描述和测定其影响；
    反映经济变量之间的相互关系，提高模型的精度；
    便于处理异常数据
  - 设置规则：一个定性因素有 m 个互斥类型，引入 $m-1$ 个虚拟变量；
    m 个定性因素，每个因素有 m_i 个不同属性类型，引入 $\sum(m_i-1)$ 个虚拟变量；
    应从分析问题的目的出发；
    在单一方程中，虚拟变量可以作为解释变量或因变量

- 虚拟解释变量
  - 引入方式：

    加法类型：$y_t = b_0 + b_1 x_t + \alpha D_t + u_t$，取 1 或 0 影响方程的截距；

    乘法类型：$y_t = b_0 + b_1 x_t + \alpha D_t x_t + u_t$，取 1 或 0 影响方程的斜率；

    一般方式：根据散点图或经济分析，大致判断类型，选择加法或乘法模型，也可交叉使用 $y_t = b_0 + b_1 x_t + \alpha_1 D_t x_t + \alpha_2 D_t + u_t$
  - 特殊应用：

    季节调整模型：$D_i = \begin{cases} 1, & \text{第}i\text{季度} \\ 0, & \text{其他} \end{cases} (i = 2,3,4)$，加法模型；常数项 b_0 对应第一季度的系数，若 D_i 的回归系数 α_{i-1} 显著不为 0，表示第 i 季度对最终数值有显著影响，b_i 表示第 i 季度与第一季度的差值

- 虚拟解释变量
  - 特殊应用：

    模型结构稳定性检验：同一总体两个样本的回归模型为

    $y_t = b_0 + b_1 x_t + u_t = a_0 + a_1 x_t + u_t$，令 $D = \begin{cases} 1, & \text{样本2} \\ 0, & \text{样本1} \end{cases}$，针对模型

    $y_t = b_0 + (a_0 - b_0) D_t + b_1 x_t + (a_1 - b_1) X D_t + u_t$ 进行回归分析；

    $a_1 = b_1, a_0 = b_0$，两者没有显著差异，“重合回归”，稳定；

    $a_1 = b_1, a_0 \neq b_0$，差异只体现在截距上，“平行回归”，不稳定；

    $a_1 \neq b_1, a_0 = b_0$，差异只体现在斜率上，“汇合回归”，不稳定；

    $a_1 \neq b_1, a_0 \neq b_0$，两者完全不同，“相异回归”，不稳定；

- 虚拟解释变量
  - 特殊应用：

分段回归：$y_t = \begin{cases} a_0 + a_1 x_t + u_t, & x_{\min} < x < x_1 \\ b_0 + b_1 x_t + u_t, & x_1 \le x < x_2 \\ c_0 + c_1 x_t + u_t, & x_2 \le x < x_{\max} \end{cases}$，令

$D_1 = \begin{cases} 0, & x_{\min} < x < x_1 \\ 1, & x_1 \le x < x_{\max} \end{cases}, D_2 = \begin{cases} 0, & x_{\min} < x < x_2 \\ 1, & x_2 \le x < x_{\max} \end{cases}$，模型为

$y = b_0 + b_1 x + b_2 (x - x_1) D_1 + b_3 (x - x_2) D_2 + u$;

虚拟解释变量

- 虚拟解释变量
  - 特殊应用：

    混合回归：使用时序数据和截面数据；

    异常值问题：一元线性回归模型，若 $\left|\frac{e_{t_0}}{\hat{\sigma}}\right| > 2$，模型在 t_0 处很可能存在异常值问题；

    因为 $E(u_t) = \begin{cases} 0, & t \neq t_0 \\ C, & t = t_0 \end{cases}$，令 $D_t = \begin{cases} 0, & t \neq t_0 \\ 1, & t = t_0 \end{cases}$，模型为 $y_t = b_0 + b_1 x_t + C \cdot D_t + v_t$，且 $v_t = u_t - C \cdot D_t$

  - 对 OLS 估计量的影响：

    加法模型：参数将无法估计，易产生完全共线性

虚拟被解释变量

- 线性概率模型 (LPM)
  - 线性概率模型的回归形式：$y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} + u_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B} + u_i$，
    $E(y_i) = p_i = P(y_i = 1) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B}$，属于内生变量
  - 估计：
    $E(y_i) = p_i \in \mathbf{R}$；$u_t$ 不服从正态分布，$u_i = \begin{cases} 1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B}, & y_i = 1 \\ -\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B}, & y_u = 0 \end{cases}$；
    $\mathrm{Var}(u_i) = (1 - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B}) \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{B} \neq constant$；
    $\text{Count } R^2 = \dfrac{\text{正确预测的次数}}{\text{预测的总次数}}$；$\dfrac{\partial E(y_i)}{\partial x_{ji}} = b_j$；
    $\dfrac{p_i}{1 - p_i}$ 是机会比率，$L_i = \ln \dfrac{p_i}{1 - p_i}$ 是对数单位

虚拟被解释变量

- 非线性概率模型
  - 特征：p 随 x 的变化而变化，但 $p \in [0,1]$；

    $x \to -\infty, p \to 0$，$x \to +\infty, p \to 1$
  - Probit 模型：

    $$P(y_i=1)=\Phi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})=\int_{-\infty}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B}}\frac{1}{\sqrt{2\pi}\exp\left(-\frac{t^2}{2}\right)}\mathrm{d}t\,;$$

    边际效应分析：$\dfrac{\partial P(y_i=1)}{\partial x_{ji}}=\phi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})b_j$；

    比例因子：$\phi(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})$ 或 $\bar{\phi}=\dfrac{1}{n}\sum\phi(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})$；

    平均边际效应：$\phi(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})b_j$

虚拟被解释变量

- 非线性概率模型
  - Logit 模型：

    $$P(y_i=1)=\Lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})=\frac{1}{1+\exp(-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})};$$

    边际效应分析：$\dfrac{\partial P(y_i=1)}{\partial x_{ji}}=\Lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})[1-\Lambda(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})]b_j$;

    比例因子：$\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})[1-\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})]$ 或

    $$\bar{\Lambda}(1-\bar{\Lambda})=\left[\frac{1}{n}\sum\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})\right]\left[1-\frac{1}{n}\sum\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})\right];$$

    平均边际效应：$\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})[1-\Lambda(\bar{\boldsymbol{x}}_i^{\mathrm{T}}\hat{\boldsymbol{B}})]b_j$

  - 估计：$\hat{\boldsymbol{B}}_{\mathrm{Logit}}\approx 4\hat{\boldsymbol{B}}_{\mathrm{OLS}}$，$\hat{\boldsymbol{B}}_{\mathrm{Probit}}\approx 2.5\hat{\boldsymbol{B}}_{\mathrm{OLS}}$，$\hat{\boldsymbol{B}}_{\mathrm{Logit}}\approx 1.6\hat{\boldsymbol{B}}_{\mathrm{Probit}}$

虚拟被解释变量

- 非线性概率模型
  - 极大似然估计：

    似然函数：$L=\prod[F(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})]^{y_i}[1-F(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})]^{1-y_i}$
  - 模型检验：

    拟合优度检验：Mcfadden $R^2=1-\dfrac{\ln L}{\ln L_0}$，越接近 1，拟合效果越好；

    总体显著性检验：$H_0:b_0=b_1=\cdots=c_k=0, H_1:b_j$ 不全为 0；

    似然比统计量：$\mathrm{LR}=2(\ln L-\ln L_0)\to\chi^2(k)$；$\mathrm{LR}>\chi^2_\alpha(k)$，拒绝 H_0，认为总体显著；$\mathrm{LR}<\chi^2_\alpha(k)$，接受 H_0，认为总体不显著

随机解释变量

- 估计量的渐进统计性质
  - 渐进无偏性 (不是无偏性)：$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$
  - 一致性：$p(\lim_{n\to\infty} \hat{\theta}_n = \theta) = 1$；$\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_n) = 0$
  - 大样本下，具有一致性；小样本下，一致性不起作用
- 随机解释变量及其产生的原因
  - 含义：$\mathrm{Cov}(x_{jt}, u_t) \neq 0$，解释变量中某些变量为随机变量，模型存在随机解释变量
  - 原因：省略的解释变量；

    经济变量取值一般难以确定；

    被解释变量往往受到前若干期值的影响

随机解释变量

- 一元线性回归模型下随机解释变量的影响
  - 若 $\text{Cov}(x_t, u_t) = 0$，OLS 估计量 $\hat{b}_0, \hat{b}_1$ 是 b_0, b_1 的无偏估计量
  - 若在小样本下 $\text{Cov}(x_t, u_t) \neq 0$，大样本下 $p\lim\limits_{n\to\infty} \frac{\dot{x}_t \dot{u}_t}{n} = 0$，OLS 估计量 $\hat{b}_0, \hat{b}_1$ 小样本下无偏，大样本下一致
  - 若 x_t, u_t 高度相关，且 $p\lim\limits_{n\to\infty} \frac{\dot{x}_t \dot{u}_t}{n} \neq 0$，OLS 估计量 $\hat{b}_0, \hat{b}_1$ 有偏、不一致
  - 若 x, u 相互独立，OLS 估计量无偏、一致
  - 若 x, u 同期不相关、异期相关，OLS 估计量小样本下有偏，大样本下一致
  - 若 x, u 同期相关，OLS 估计量有偏、非一致

随机解释变量

- 工具变量法 (IV)
  - 一元线性回归模型：$y_t = b_0 + b_1 x_t + u_t$ 的离差形式 $\dot{y}_t = b_1 \dot{x}_t + \dot{u}_t$；
    方程 $\sum \dot{z}_t \dot{y}_t = b_1 \sum \dot{z}_t \dot{x}_t + \sum \dot{z}_t \dot{u}_t$，其中 z_t 是工具变量，解得
    $$\begin{cases} \hat{b}_1 = \frac{\sum (z_t - \bar{z})(y_t - \bar{y})}{\sum (z_t - \bar{z})(x_t - \bar{x})}, \\ \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \end{cases}$$
  - 多元线性回归模型：$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U}$，将解释变量矩阵 $\boldsymbol{X} = [\mathbf{1}_{n\times 1}, x_{1i}, x_{2i}, \cdots, x_{ki}]$ 中的 x_{1t}, x_{kt} 换位工具变量 z_{1t}, z_{kt}，得到工具变量矩阵 $\boldsymbol{Z} = [\mathbf{1}_{n\times 1}, z_{1i}, x_{2i}, \cdots, z_{ki}]$，解得 $\hat{\boldsymbol{B}}_{\mathrm{IV}} = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}$，$\hat{\boldsymbol{B}}_{\mathrm{IV}}$ 是 $\boldsymbol{B}$ 的有偏、一致估计量

- 工具变量法 (IV)
  - 两阶段最小二乘法 (2SLS):

    一元：用 OLS 法对 $\hat{x}_t = \hat{a}_0 + \hat{a}_1 z_t$ 进行回归；

    以第一步得到的 $\hat{x}_t$ 进行 OLS 法回归，得 $y_t = b_0 + b_1 \hat{x}_t + u_t$；

    二元：$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + u_t$，其中 x_{1t} 为内生变量，x_{2t} 为外生变量，工具变量 z_{1t}, z_{2t}；

    两次 OLS 法回归，方程分别为 $\hat{x}_{1t} = \hat{a}_0 + \hat{a}_1 x_{2t} + \hat{a}_2 z_{1t} + \hat{a}_3 x_{2t}$，

    $y_t = b_0 + b_1 \hat{x}_{1t} + b_2 x_{2t} + u_t$

- 工具变量法 (IV)
  - 豪斯曼检验 (Hausman 检验)：$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + u_t$，其中 x_{1t} 为随机解释变量 (怀疑有内生性)，x_{2t} 为外生变量，工具变量 z_t；
    用 OLS 法对 $x_{1t} = a_0 + a_1 x_{2t} + a_2 z_t + v_t$ 进行回归，得到残差项 $\hat{v}_t$；
    再用 OLS 法对 $y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \delta \hat{v}_t + \varepsilon_t$ 进行回归；
    δ 显著为 0，认为 x_{1t} 不具有内生性；δ 不显著为 0，认为 x_{1t} 具有内生性

代码输出结果分析

- 回归分析结果

同第二章：

常数和解释变量	参数估计值	参数标准误差	t统计量	双侧概率
C(b_0)	331.5264	57.16954	5.799003	0.0000
PI(b_1)	0.692812	0.006279	110.3337	0.0000
决定系数	0.997297	被解释变量均值		4662.514
调整的决定系数	0.997215	被解释变量标准差		4659.100
回归标准误差	245.8925	赤池信息准则		13.90311
残差平方和	1995283.	施瓦兹信息准则		13.99199
对数似然函数	−241.3044	汉南准则		13.93379
F统计量	12173.53	DW统计量		0.180221
F统计量的概率	0.000000			

- 各种检验的输出结果分析

同第四章:

英文	含义	英文	含义
`Heterpskedasticity Test`	检验方法	`F-statistic`	回归模型的 F 统计量
`Obs*R-squared`	F 检验统计量	`Prob. Chi-Square(2)`	F 统计量对应的 p 值
`Prob. F(a, b)`	自由度为 a,b 的 F 分布临界值	`Scaled explained SS`	LM 统计量