计量经济学-作业6-虚拟变量

😀 [redacted]

P128 习题1

1 不查阅书本（或笔记），给出下列术语的定义，然后与书本上的相比较。

a. 双对数函数形式	b. 弹性	c. 交叉项	d. 截距虚拟变量
e. 滞后	f. 参数是线性的	g. 变量是线性的	h. 对数
i. 自然对数	j. 多项式函数形式	k. 半对数函数形式	l. 斜率虚拟变量

回答

a. 双对数函数形式：**所有**的变量（包括解释变量和被解释变量）是**以自然对数**的形式表示，其形式例如 $\ln Y = \beta_0 + \beta_1 \ \ln X_1 + \beta_2 \ \ln X_2 + \varepsilon$

b. 弹性 :保持方程中其他变量不变，**解释变量变化1%时，引起被解释变量变化的百分比**

c. 交叉项：回归方程中的一个解释变量，这个解释变量是由方程中**两个或两个以上的其他解释变量相乘**得到的。

d. 截距虚拟变量：描述解释变量和被解释变量之间的**截距**因虚拟变量的设定是否被满足而有所不同的情况

e. 滞后：并不是所有的解释变量和被解释变量的因果关系都是瞬时的。在很多情况下，解释变量的变化对被解释变量的影响**要延迟一段时间**。这段时间叫作滞后期。

f. 参数是线性的：方程的参数(β_s)都是以简单的线性形式呈现，方程中这些**参数**的幂都是1次的，没有和其他参数相乘或相除，它们自身也没有包含其他的函数形式（如对数或指数）。例如： $f(Y) = \beta_0 + \beta_1 f(X)$

g. 变量是线性的：方程的变量都是以简单的线性形式呈现，方程中这些**变量**的幂都是1次的，没有和其他变量相乘或相除，它们自身也没有包含其他的函数形式（如对数或指数）。根据X和Y的值描绘出来的函数图像是一条直线。例如： $Y = \beta_0 + \beta_1 X + \varepsilon$

h. 对数：如果a的b次方等于X,那么，b就是X的以a为底的对数值

> 即 $\log_a \ (X) = b$ 表示的是 $a^b = X$

i. 自然对数：自然对数是以常数e为底数的对数

> 即 $\ln \ (X) = b$ 表示的是 $e^b = X$

j. 多项式函数形式：被解释变量Y由**解释变量的函数形式**表示。如
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_2)^2 + \varepsilon$

k. 半对数函数形式：**只有一部分**的变量（包括解释变量和被解释变量）是以自然对数的形式表示，具体又分为左半对数函数形式和右半对数函数形式。

l. 斜率虚拟变量：描述解释变量和被解释变量之间的**斜率**因虚拟变量的设定是否被满足而有所不同的情况，其是一种特殊的交叉项，为一个虚拟变量和普通解释变量相乘得到

P128 习题3

3 仔细观察下面的方程，说出它们是变量是线性的方程，还是参数是线性的方程，或者都是，或者都不是：

a. $Y_i = \beta_0 + \beta_1 X_i^3 + \varepsilon_i$　　b. $Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$　　c. $\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i$

d. $Y_i = \beta_0 + \beta_1 X_i^{\beta_1} + \varepsilon_i$　　e. $Y_i^{\beta_1} = \beta_1 + \beta_2 X_i^2 + \varepsilon_i$

回答

a. 是**参数线性但不是变量线性**。因为变量 X 存在三次方

b. 是**参数线性但不是变量线性**。因为变量 X 包含了对数形式（$\ln\ X$）

c. 是**参数线性但不是变量线性**。因为变量 X、Y 包含了对数形式（$\ln\ X$、$\ln\ Y$）

d. **不是参数线性也不是变量线性**。因为变量 X 存在 β_2 次方，同时参数 β_2 也是变量的指数

e. **不是参数线性也不是变量线性**。因为变量 Y 存在 β_2 次方，变量 X 存在2次方，同时参数 β_2 也是变量的指数

P128 习题5

5 为了解释区域工资差异，你搜集了 7 338 名非技术工人工资收入的数据，将整个国家分成四个部分（东北部、南部、中西部和西部），并估计得出了以下的方程（括号内的数值为标准差）：

$$\hat{Y}_i = 4.78 - 0.038E_i - 0.041S_i - 0.048W_i$$
$$(0.019) \quad (0.010) \quad (0.012)$$
$$\overline{R}^2 = 0.49 \qquad N = 7338$$

式中，Y_i 代表第 i 位非技术工人每小时的工资（单位：美元）；E_i 代表第 i 位工人是否住在东北部，为虚拟变量，如果是为 1，否则为 0；S_i 代表第 i 位工人是否住在南部，为虚拟变量，如果是为 1，否则为 0；W_i 代表第 i 位工人是否住在西部，为虚拟变量，如果是为 1，否则为 0。

a. 方程中被省略的状态是什么？

b. 如果增加一个虚拟变量来代表被省略的状态，而保持 E_i，S_i 和 W_i 不变，会发生什么情况？

c. 如果增加一个虚拟变量来代表被省略的状态，去掉方程中的 E_i，新变量的估计参数的符号是什么？

d. 下面三种表述中哪种最准确？哪种最不准确？请说明理由。

ⅰ. 仅仅用区域变量，方程解释了 Y 围绕其均值变化的 49%，因此，工资差异有相当一部分由于是区域差异引起的。

ⅱ. 区域变量的参数实质上是相同的，因此，工资不存在较大的区域差异。

ⅲ. 区域变量的参数值与平均工资相比较小，因此，工资不存在较大的区域差异。

e. 如果要在方程中再增加一个变量，应该加什么？证实你的选择。

回答

a. 方程中被省略的状态是**居住于中西部**。当 E_i、S_i、W_i 三个虚拟变量均取值为0时候代表这个情况。

b. 增加一个虚拟变量会导致出现**完全共线性**，当 E_i、S_i、W_i 三个虚拟变量均取值为0时，这个虚拟变量的值一定为1，即存在 $E_i + S_i + W_i + M_i = 1$ (假设新变量为 M_i)

c. 增加一个虚拟变量并去掉方程中的 E_i ，这个**变量符号预计为正**。

d. 表述中**iii最准确，i最不准确**

> 虽然从假设检验的角度，存在显著性差异，但是——
>
> 参数值的数量级均为 10^{-2} ，远小于平均工资4.78，因此处于四个地区的实际工资差异并不大，差异数量级不超过 10^{-2} ，因此可以认为工资**不存在较大的区域差异**

e. 可以增加 `地区消费水平` ，往往一个地区消费水平越高，其工资相对较高。

P128 习题8

8 理查德·福尔斯（Richard Fowles）和彼得·勒布（Peter Loeb）研究了饮酒、海拔和交通死亡情况之间的关系。[⊖]作者假设酒后驾车死亡事故在高海拔的地方比在低海拔的地方更容易发生。这是因为高海拔地方的大脑供氧量较少，增强了酒精对人体的作用。为了检验这个假设，他们在方程中用到了海拔和啤酒消费的交叉项。估计得了如下的机车死亡率（根据美国的报告）的截面模型：

$$\hat{F}_i = -3.36 - 0.002B_i + 0.17S_i - 0.31D_i + 0.011B_iA_i \qquad (7\text{-}22)$$
$$(-0.08)\quad (1.85)\quad (-1.29)\quad (4.05)$$
$$N = 48 \qquad \overline{R}^2 = 0.499$$

式中，F_i 代表在第 i 个州机车行驶每单位英里发生的交通死亡次数；B_i 代表在第 i 个州人均消费的啤酒（麦芽饮料）；S_i 代表在第 i 个州高速公路的平均驾驶速度；D_i 代表第 i 州有没有机车安全监察程序，为虚拟变量，如果有为 1，否则为 0；A_i 代表第 i 个州的主要城市的平均海拔（单位：千米）。

a. 在 5%的显著水平下，对变量 B，S 和 D 的参数进行假设检验。检验结果说明了方程中的什么计量经济学问题？请说明理由。

b. 思考交叉项衡量的是什么？详细说明 $B \cdot A$ 的参数含义。

c. 在 5%的显著水平下，对交叉项进行假设检验。

d. 注意变量 A_i 只包含在交叉项中，但并没有作为一个单独的解释变量。那么，交叉项中的两个部分应该作为两个单独的解释变量吗？为什么？（提示：在斜率虚拟变量那部分内容中，我们强调了斜率虚拟变量和截距虚拟变量必须同时存在于方程中。）

e. 在方程中增加解释变量 A_i，结果如方程（7-23）所示。你认为哪个方程更合理？请说明理由。

$$\hat{F}_i = -2.33 - 0.024B_i + 0.14S_i - 0.24D_i - 0.35A_i + 0.023B_iA_i \qquad (7\text{-}23)$$
$$(-0.80)\quad (1.53)\quad (-0.96)\quad (-1.07)\quad (1.97)$$
$$N = 48 \qquad \overline{R}^2 = 0.501$$

回答

a. 分别对 B、S、D 三个参数进行假设检验。在显著性水平5%的条件，首先，样本总数为48，可以求得自由度为 $n-k-1=48-4-1=43$,查表得单侧 $t_c = 1.682$ 。三个参数预期的符号分别为正、正、负。

- B :做出假设 $H_0:\beta \leq 0$ $H_A:\beta > 0$,且预测参数为正 *（因为根据常识啤酒消费越高，越容易饮酒，死亡率越高）* 那么 $t_1 = -0.08$ ，可知 $|-0.08| < 1.682$ 即 $|t_1| < t_c$ ，因此**不能拒绝** H_0
- S :做出假设 $H_0:\beta \leq 0$ $H_A:\beta > 0$,且预测参数为正 *（因为根据常识驾驶速度越高，越容易发生交通事故，死亡率越高）* 那么 $t_2 = 1.85$ ，可知 $|1.85| > 1.682$ 即 $|t_2| > t_c$ ，且其预期符号和对立假设相同（均为正数），因此**可以拒绝** H_0
- D :做出假设 $H_0:\beta \geq 0$ $H_A:\beta < 0$,且预测参数为负 *（因为根据常识没有安全检测程序，越容易发生交通事故，死亡率越高）* 那么 $t_3 = -1.29$ ，可知 $|-1.29| < 1.682$ 即 $|t_3| < t_c$ ，因此**不能拒绝** H_0

检验结果的问题：

1. B 的斜率参数并不显著，不显著的原因可能是遗漏了某个变量 *（回归方程的符号和预期不一致）* ，也有可能是交叉项吸收了啤酒消费的对死亡率的影响。

2. D 的斜率参数也不显著，但是他是一个逻辑正确且合理的数值，我们无法将其视为无关变量。

b. 交差项是衡量 啤酒饮用对交通死亡人数是否随着城市海拔高度的上升而上升 的指标。

   参数意义为：**方程中所有其他自变量保持不变的情况下，B和A的乘积每增加一个单位，F就会增加0.011**。但是，系数的大小本身并没有真正的直观意义。

c. 做出假设 $H_0:\beta \leq 0$ $H_A:\beta > 0$,且预测参数为正 *（因为根据题意，高海拔地方的大脑供氧量较少，增强了酒精对人体的作用，死亡率越高）* 那么 $t_4 = 4.05$ ，可知 $|4.05| > 1.682$ 即 $|t_4| > t_c$ ，且其预期符号和对立假设相同（均为正数），因此**可以拒绝** H_0

d. 虽然这个不是斜率虚拟变量，但是参照斜率虚拟变量的规则，**应该作为两个单独的解释变量；**同时，一般来讲，**需要将交叉项中的两个部分应该作为两个单独的解释变量**。主要原因是为了避免 省略的交叉项的某个因子的影响 附加 到交叉项系数上使交叉项系数变得显著。

   > 这也是在第一小问提出的检验结果的其中一个问题

e. 我认为**方程(7-23)更合理**，原因如下：

   i. 首先，单独增加了 A 剔除了因为遗漏交叉项中 A 的影响 附加 到交叉项系数上使交叉项系数变得显著的可能性，即第4小问描述的问题

   ii. 将 A 加入变量中，使得 $\overline{R^2}$ 增大了，方程的拟合优度更好。更加稳健。