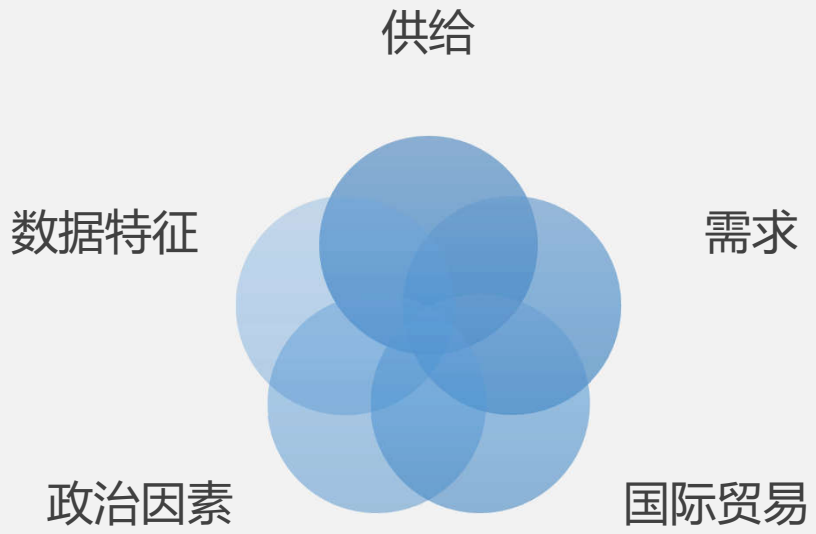
3-1

虚拟变量的概念和设置规则

主讲人：陈娟

- 我国是石油生产大国，更是石油消费大国。

- 影响石油进口的因素有很多：



一、什么是虚拟变量

●回顾：定量因素与定性因素

定量因素：可直接测度的数值型因素，如国内生产总值、居民收入和消费等。

定性因素：不能精确计量的说明某种属性或状态存在与否的非数值型因素，如性别、户口、学历、民族等。

虚拟变量：人工构造的取值为0和1的作为属性变量代表的变量称虚拟变量，一般常用D (dummy) 表示

$D=0$　表示某种属性或状态不出现或不存在
$D=1$　表示某种属性或状态出现或存在

二、虚拟变量的作用

- 作为属性因素的代表
  如：性别
- 作为某些非精确计量的数量因素的代表
  如：受教育程度(高中及以下、专科、本科及以上)
- 作为某些偶然因素或政策因素的代表
  如：“911事件”、四川汶川大地震
- 时间序列分析中作为季节（月份）的代表
- 分段回归——研究斜率、截距的变动，或比较两个回归模型的差异

三、虚拟变量模型

虚拟变量模型：包含有虚拟变量的模型称虚拟变量模型

三种类型：

1、解释变量中只包含虚拟变量√

2、解释变量中既含定量变量，又含虚拟变量√

3、虚拟被解释变量模型：被解释变量本身取值为0或1

四、虚拟变量的设置规则

1、虚拟变量取值

- 虚拟变量D取值为0，还是取值为1，要根据研究的目的去决定。

D取值为0的类型—**基础类型**，作为比较的基准

D取值为1的类型—**与基础类型相比较的类型**

●例如：为比较政府税收政策变动对居民收入X与消费支出Y关系的影响，模型可以设定为

$$Y_t = \alpha_0 + \alpha_1 D_t + \beta X_t + u_t$$

其中：$D_t = \begin{cases} 0, & \text{基础类型(政府税收政策不变)} \\ 1, & \text{比较类型(政府税收政策变动)} \end{cases}$

当$D_t = 0$时，$Y_t = \alpha_0 + \beta X_t + u_t$

当$D_t = 1$时，$Y_t = \alpha_0 + \alpha_1 + \beta X_t + u_t$

●思考：

●依据虚拟变量的取值原则，一个定性变量有m种类型，可以对每个类型都设置一个虚拟变量，一共设置了m个虚拟变量。

●在一个有截距项的模型中引入虚拟变量，如果把m个虚拟变量都引入模型中，会出现什么问题呢？

●例：已知冷饮的销售量Y除受k个定量因素X_1，X_2,...,X_k的影响外，还受春、夏、秋、冬四季变化的影响。

●冷饮销售模型设定为：

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots \beta_k X_{kt} + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + \mu_t$$

$$D_{1t} = \begin{cases} 1 & \text{春季} \\ 0 & \text{其他} \end{cases} \qquad D_{2t} = \begin{cases} 1 & \text{夏季} \\ 0 & \text{其他} \end{cases} \qquad D_{3t} = \begin{cases} 1 & \text{秋季} \\ 0 & \text{其他} \end{cases} \qquad D_{4t} = \begin{cases} 1 & \text{冬季} \\ 0 & \text{其他} \end{cases}$$

其矩阵形式为：

$$\mathbf{Y} = (\mathbf{X}, \mathbf{D}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} + \boldsymbol{\mu}$$

●如果只取六个观测值，其中春季与夏季取了两次，秋、冬各取到一次观测值，则解释变量矩阵可表示为：

$$(\mathbf{X},\ \mathbf{D}) = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} & 1 & 0 & 0 & 0 \\ 1 & X_{12} & \cdots & X_{k2} & 0 & 1 & 0 & 0 \\ 1 & X_{13} & \cdots & X_{k3} & 0 & 0 & 1 & 0 \\ 1 & X_{14} & \cdots & X_{k4} & 0 & 0 & 0 & 1 \\ 1 & X_{15} & \cdots & X_{k5} & 0 & 1 & 0 & 0 \\ 1 & X_{16} & \cdots & X_{k6} & 1 & 0 & 0 & 0 \end{pmatrix}$$

●显然，解释变量矩阵中的第1列可表示成后4列的线性组合。

● “虚拟变量陷阱”

●虚拟变量的设置原则

●如果1个定性因素有m个相互排斥的类型，按照模型设定中有无截距项，虚拟变量个数的设置规则分为两种情况：

1、在有截距项的模型中，只能引入m - 1个虚拟变量，否则会落入“虚拟变量陷阱”；

2、在无截距项的模型中，可以引入m个虚拟变量。

●评价

●这两种设置方法是等价的；

●但是模型中包含截距项更方便，可以很容易地检验某组与基准组之间是否存在显著差异以及差异程度。