# 第九章　虚拟变量模型

在前面几章中，主要简介了经典线性回归模型及其在若干基本假定下的估计问题，并分析了一种或多种假定不满足时所产生的后果及其可能的改善措施。然而上述措施还不能处理经济生活中遇到的全部问题。

例如：

怎样考察某一突发事件、性别、季节、受教育程度等对经济行为带来的影响？？

本章将主要简介经典单方程计量经济学模型中引入虚拟变量并在此基础上对建立单方程计量经济学模型的措施论进行简朴的总结与讨论。

# 第九章　虚拟变量模型

## ◆ 学习目的

了解虚拟变量、虚拟变量模型的概念，掌握虚拟变量设置的原则和引入模型的措施。

## ◆ 基本要求

1)认识到虚拟变量是建立计量经济学模型经常会遇到的问题；

2)了解虚拟变量、虚拟变量模型的概念；

3)掌握虚拟变量设置的原则、虚拟变量模型的建模措施及应用。

# 第八章　虚拟变量模型

## 第一节　　虚拟变量

- ◆ 虚拟变量
  - ◆虚拟变量模型
    - ◆虚拟变量的引入
      - ◆虚拟变量的设置原则

# 一、虚拟变量

为何要引入“虚拟变量”？？

许多经济变量是能够定量度量的或者说是能够直接观察的

如商品需求量、价格、收入、产量等

但是也有某些影响经济变量的原因无法定量度量或者说无法直接观察

如职业、性别对收入的影响，战争、自然灾害对GDP的影响，季节对某些产品(如冷饮)销售的影响等。

为了能够在模型中反应这些原因的影响，并提升模型的精度，需要将它们人为地“量化”，这种“量化”一般是经过引入“虚拟变量”来完毕的。

这种用两个相异数字来表达对被解释变量有主要影响而本身又没有观察数值的一类变量，称为**虚拟变量(dummy variables)**。

虚拟变量也称为哑变量或定性变量。

## 虚拟变量的特点是：

1．虚拟变量是对经济变化有主要影响的不可测变量。

2．虚拟变量是赋值变量，一般根据这些原因的属性类型，构造只取“0”或“1”的人工变量，一般称为虚拟变量，记为D。这是为了便于计算而把定性原因这么数量化的，所以虚拟变量的数值只表达变量的性质而不表示变量的数值。

一般地，在虚拟变量的设置中，

基础类型和肯定类型取值为**1**；

比较类型和否定类型取值为**0**。

## 例如：

1）表达性别的虚拟变量可取为

$$D_1=\begin{cases}1 & \text{男性}\\0 & \text{女性}\end{cases}$$

2）表达文化程度的虚拟变量可取为

$$D_2=\begin{cases}1 & \text{本科及以上学历}\\0 & \text{本科下列学历}\end{cases}$$

3）表达地域的虚拟变量可取为

$$D_3=\begin{cases}1 & \text{城市}\\0 & \text{农村}\end{cases}$$

4）表达消费神理的虚拟变量可取为

$$D_4=\begin{cases}1 & \text{喜欢某种商品}\\0 & \text{不喜欢某种商品}\end{cases}$$

5）表达天气变化的虚拟变量可取为

$$D_5=\begin{cases}1 & \text{晴天}\\0 & \text{雨天}\end{cases}$$

## 二、虚拟变量模型

同步具有一般解释变量与虚拟变量的模型称为虚拟变量模型。

在模型中，虚拟变量可作为解释变量，也可作为被解释变量，但主要是用作解释变量。

**例如：** 一种以性别为虚拟变量来考察职员薪金的模型如下：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i \qquad (8\text{-}1)$$

其中

$Y_i$ ——为职员的薪金； $D_i$ =1 ——代表男性

$X_i$ ——为职员工龄； $D_i$ =0 ——代表女性

## 三、虚拟变量的引入

虚拟变量作为解释变量引入模型有两种基本方式：加法方式和乘法方式。

### 1. 加法方式

上述职员薪金模型（8-1）中性别虚拟变量的引入就采用了加法方式，

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$$

在该模型中，假如仍假定 $E(\mu_i)$ =0，则

女职员的平均薪金为：

$$E(Y_i | X_i, D_i = 0) = \beta_0 + \beta_1 X_i$$

男职员的平均薪金为：

$$E(Y_i | X_i, D_i = 1) = (\beta_0 + \beta_2) + \beta_1 X_i$$

从几何意义上看(图8-1)，

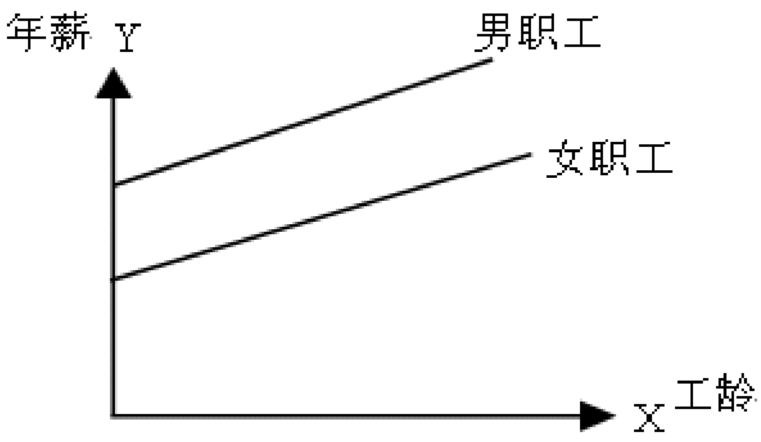


图8-1 男女职员平均薪金示意图

假定$\beta_2$ >0，

则两个函数有相同的斜率，但有不同的截距。

这意味着，男女职员平均薪金对工龄的变化率是一样的，但两者的平均薪金水平相差$\beta_2$。

能够经过老式的回归检验，对$\beta_2$的统计明显性进行检验，以判断男女职员的平均薪金水平是否有明显差别。

## 例如：

在截面数据基础上，考虑个人保健支出对个人收入和教育水平的回归。

教育水平考虑三个层次：高中下列，高中，大学及其以上

这时需要引入两个虚拟变量：

$$D_1=\begin{cases}1 & \text{高中}\\0 & \text{其他}\end{cases}\qquad D_2=\begin{cases}1 & \text{大学及其以上}\\0 & \text{其他}\end{cases}$$

模型可设定如下：

$$Y_i=\beta_0+\beta_1X_i+\beta_2D_{1i}+\beta_3D_{2i}+\mu_i \qquad (8\text{-}2)$$

在$E(\mu_i)$ =0的初始假定下，轻易得到高中下列、高中、大学及其以上

教育水平个人平均保健支出的函数：

高中下列：　$E(Y_i|X_i, D_{1i}=0, D_{2i}=0)=\beta_0+\beta_1 X_i$

高中：　$E(Y_i|X_i, D_{1i}=1, D_{2i}=0)=(\beta_0+\beta_2)+\beta_1 X_i$

大学及其以上：　$E(Y_i|X_i, D_{1i}=0, D_{2i}=1)=(\beta_0+\beta_3)+\beta_1 X_i$

假定$\beta_3 > \beta_2 > 0$，且 $\beta_0 > 0$ ，则其几何意义如图8-2所示。

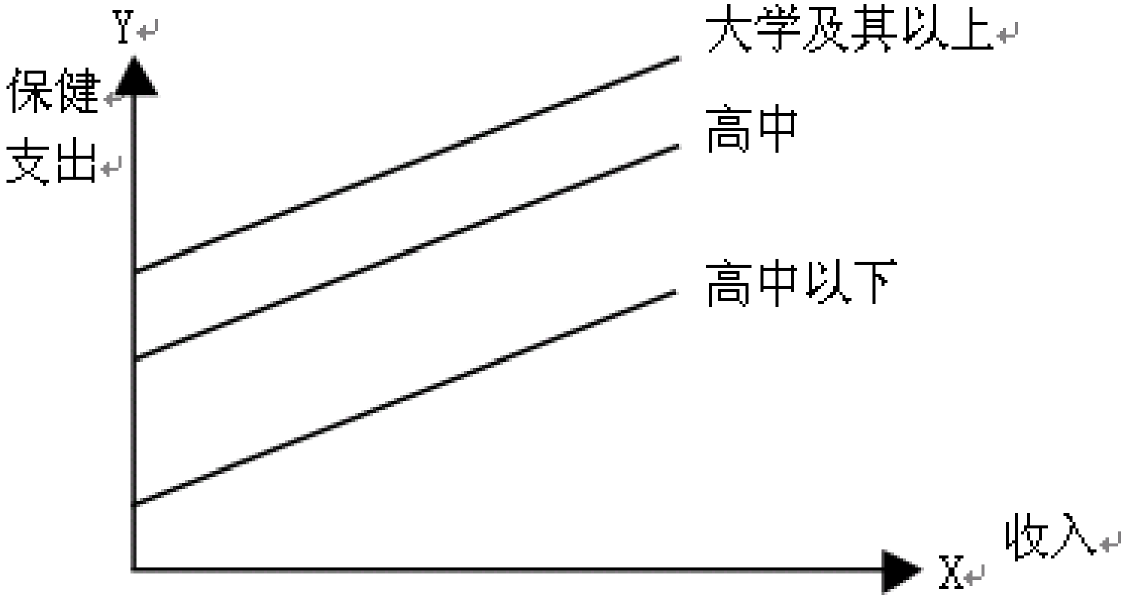


图8-2  不同教育程度人员保健支出示意图

还可将多种虚拟变量引入模型中以考察多种“定性”原因的影响。

例如：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$$

在职员薪金模型（8-1）的例子中，再引入学历的虚拟变量

$$D_2 = \begin{cases} 1 & \text{本科及以上学历} \\ 0 & \text{本科下列学历} \end{cases}$$

则职员薪金的回归模型可设计如下：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_{2i} + \mu_i \qquad (8\text{-}3)$$

于是，不同性别、不同学历职员的平均薪金分别由下面各式给出：

女职员本科下列学历的平均薪金：

$$E(Y_i|X_i，D_{1i}=0，D_{2i}=0)=\beta_0+\beta_1 X_i$$

男职员本科下列学历的平均薪金：

$$E(Y_i|X_i，D_{1i}=1，D_{2i}=0)=(\beta_0+\beta_2)+\beta_1 X_i$$

女职员本科以上学历的平均薪金：

$$E(Y_i|X_i，D_{1i}=0，D_{2i}=1)=(\beta_0+\beta_3)+\beta_1 X_i$$

男职员本科以上学历的平均薪金：

$$E(Y_i|X_i，D_{1i}=1，D_{2i}=1)=(\beta_0+\beta_2+\beta_3)+\beta_1 X_i$$

## 2. 乘法方式 ——斜率的变化

# 例如：

根据消费理论，消费水平C主要取决于收入水平X。但在一种较长的时期，人们的消费倾向会发生变化，尤其是在自然灾害、战争等反常年份，消费倾向往往出现变化。这种消费倾向的变化可经过在收入的系数中引入虚拟变量来考察。

设

$$D_t=\begin{cases}1 & \text{正常年份}\\ 0 & \text{反常年份}\end{cases}$$

则消费模型可建立如下：

$$C_t=\beta_0+\beta_1 X_t+\beta_2 D_t X_t+\mu_t \qquad \text{(8-4)}$$

这里，虚拟变量 $D_t$ 以与 $X_t$ 相乘的方式引入了模型中，从而可用来考察消费倾向的变化。

在E(μt)=0的假定下，上述模型所表达的函数可化为:

**正常年份：**

$$E(C_t|X_t, D_t = 1) = \beta_0 + (\beta_1 + \beta_2)X_t$$

**反常年份：**

$$E(C_t|X_t, D_t = 0) = \beta_0 + \beta_1 X_t$$

假定 $\beta_2 > 0$，

则其几何图形如图8-3所示。

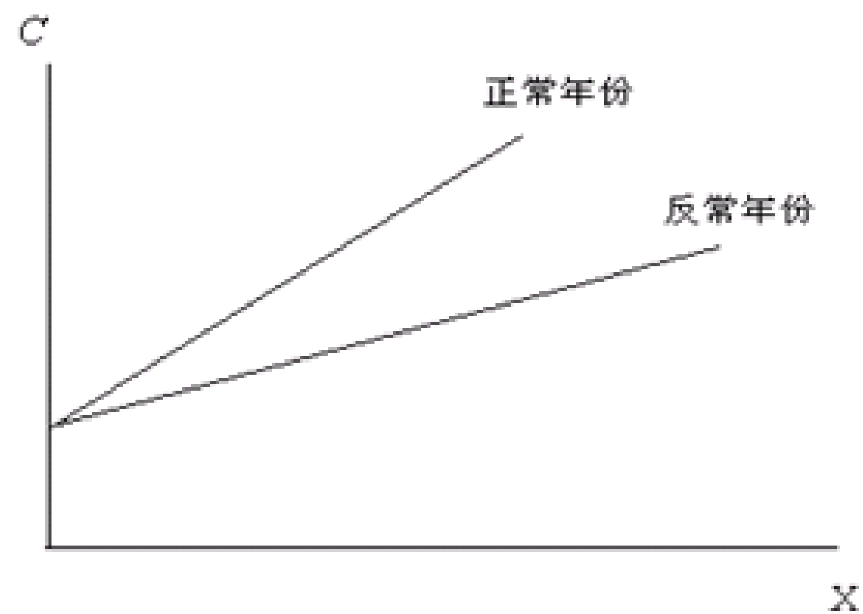


图8-3 不同年份消费倾向示意图

假如在模型中同步使用加法和乘法两种方式引入虚拟变量，则回归线的截距和斜率都会变化。

**例如：**

对于改革开放前后储蓄-收入模型，可设定为

$$Y_t = \alpha_0 + \alpha_1 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + \mu_t \quad （8-5）$$

其中，$Y$为储蓄，$X$为收入，$D_t$为虚拟变量

$$D_t = \begin{cases} 1 & \text{改革开放后来} \\ 0 & \text{改革开放此前} \end{cases}$$

显然在式（8-5）中，同步使用加法和乘法两种方式引入了虚拟变量。

在E(μt)=0的假定下，上述模型所表达的函数可化为:

**改革开放此前：**

$$E(Y_t|X_t,\ D_t=0)=\alpha_0+\beta_1 X_t$$

**改革开放后来：**

$$E(Y_t|X_t,\ D_t=1)=(\alpha_0+\alpha_1)+(\beta_1-\beta_2)X_t$$

假定 $\alpha_1>0$ 且 $\beta_2>0$，

则其几何图形如图8-4所示。

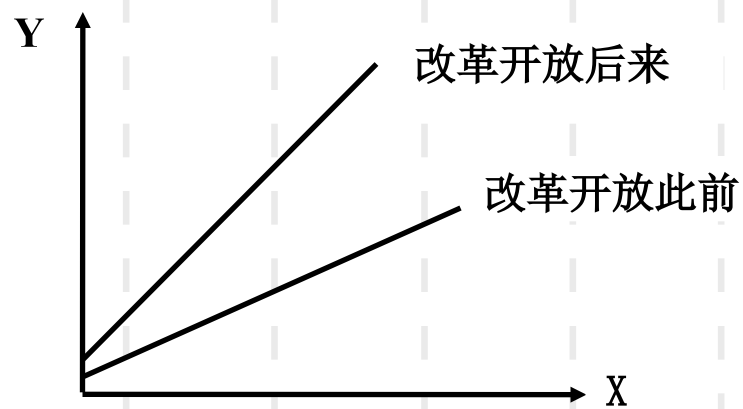


**图8-4 改革开放前后储蓄函数示意图**

## 3．临界指标的虚拟变量的引入

在经济发生转折时，可经过建立临界指标的虚拟变量模型来反应。

**例如：**

进口消费品数量*Y*主要取决于国民收入*X*的多少，中国在改革开放前后，*Y*对*X*的回归关系明显不同。

这时，能够$t^*=1979$为转折期，以1979年的国民收入$X_{t^*}$为临界值，设如下虚拟变量：

$$D_t=\begin{cases}1 & t\geq t^* \\ 0 & t<t^*\end{cases}$$

则进口消费品的回归模型可建立如下：

$$Y_t=\beta_0+\beta_1 X_t+\beta_2(X_t-X_t^*)D_t+\mu_t \qquad (8\text{-}6)$$

假如用OLS法得到该模型的回归方程为

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t + \hat{\beta}_2 (X_t - X_t^*) D_t \qquad (8\text{-}7)$$

则两个时期进口消费品函数分别为

当$t < t^* = 1979$时

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$$

当$t \geq t^* = 1979$时

$$\hat{Y}_t = (\hat{\beta}_0 - \hat{\beta}_2 X_t^*) + (\hat{\beta}_1 + \hat{\beta}_2) X_t$$

几何图形如图8-5所示

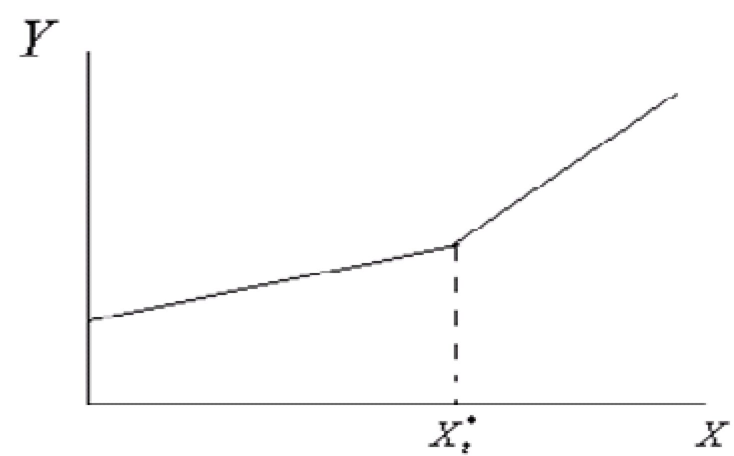


图8-5 转折期回归示意图

## 4．数值变量作为虚拟变量引入

有些变量虽然是数量变量，即能够取得实际观察值，但在某些特定情况下把它选用为虚拟变量则是以便的，以虚变量引入计量经济学模型愈加合理。

譬如年龄原因虽然能够用数字计量，但假如将年龄作为资料分组的特征，则可将年龄选作虚拟变量。

## 例如：

家庭教育经费支出不但取决于其收入，而且与年龄原因有关。

按年龄划分为三个年龄组：6—18岁年龄组（中小学教育）；19—22岁年龄组（大学教育）；其他年龄组。于是设定虚拟变量

$D_1 = \begin{cases} 1 & \text{6-18岁年龄组} \\ 0 & \text{其他} \end{cases}$　　　$D_2 = \begin{cases} 1 & \text{19-23年龄组} \\ 0 & \text{其他} \end{cases}$

则家庭教育经费支出模型可设定为

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \mu_i \qquad (8\text{-}8)$$

其中，$Y_i$是第$i$个家庭的教育经费支出；$X_i$是第$i$个家庭的收入；虚拟变量$D_{1i}$、$D_{2i}$分别表达第$i$家庭中是否有6—18岁和19—22岁的组员。

## 5. 虚拟变量交互效应分析

当分析解释变量对变量的影响时，大多数情形只是分析了解释变量本身变动对被解释变量的影响作用，而没有进一步分析解释变量间的相互作用对被解释变量影响。

前面讨论的分析两个定性变量对被解释变量影响的虚拟变量模型中，

暗含着一种假定：

**两个定性变量是分别独立地影响被解释变量的**

但是在实际经济活动中，两个定性变量对被解释变量的影响可能存在一定的交互作用，即一种解释变量的边际效应有时可能要依赖于另一种解释变量。

为描述这种交互作用，能够把两个虚拟变量的乘积以加法形式引入模型。

**例如：**

考虑下列模型

$$Y_i=\alpha_0+\alpha_1 D_{1i}+\alpha_2 D_{2i}+\beta X_i+\mu_i \qquad (8\text{-}9)$$

其中，$Y_i$为农副产品生产总收益，$X_i$为农副产品生产投入，$D_{1i}$为油菜籽生产虚拟变量，$D_{2i}$为养蜂生产虚拟变量。这里

$$D_{1i}=\begin{cases}1 & \text{发展油菜籽生产} \\ 0 & \text{其他}\end{cases} \qquad D_{2i}=\begin{cases}1 & \text{发展养蜂生产} \\ 0 & \text{其他}\end{cases}$$

显然，(8-9)式描述了是否发展油菜籽生产与是否发展养蜂生产的差别对农副产品总收益的影响。

虚拟解释变量$D_{1i}$和$D_{2i}$是以加法形式引入的，那么暗含着假定：

油菜籽生产和养蜂生产是分别独立地影响农副产品生产总收益。

但是，在发展油菜籽生产时，同步也发展养蜂生产，所取得的农副产品生产总收益可能会高于不发展养蜂生产的情况。即在是否发展油菜籽生产与养蜂生产的虚拟变量$D_{1i}$和$D_{2i}$之间，很可能存在着一定的交互作用，且这种交互影响对被解释变量—农副产品生产总收益会有影响。

为描述虚拟变量交互作用对被解释变量的效应，在(8-9)式中以加法形式引入两个虚拟解释变量的乘积，即

$$Y_i=\alpha_0+\alpha_1 D_{1i}+\alpha_2 D_{2i}+\alpha_3（D_{1i}D_{2i}）+\beta X_i+\mu_i \qquad (8\text{-}10)$$

（1）基础类型：不发展油菜籽生产，也不发展养蜂生产时农副产品生产平均总收益

$$E(Y_i|X_i,D_1=0,D_2=0)=\alpha_0+\beta X_i \qquad (8\text{-}11)$$

（2）比较类型：同步发展油菜籽生产和养蜂生产时，农副产品生产平均总收益

$$E(Y_i|X_i,D_1=1,D_2=1)=\alpha_0+\alpha_1+\alpha_2+\alpha_3+\beta X_i \qquad (8\text{-}12)$$

其中

$\alpha_1$为是否发展油菜籽生产对农副产品生产总收益的截距差别系数；
$\alpha_2$为是否发展养蜂生产对农副产品生产总收益的截距差别系数；
$\alpha_3$为同步发展油菜籽生产和养蜂生产时对农副产品生产总收益的交互效应系数。
$\alpha_0\sim\alpha_3$构成截距水平。