第八章 虚拟变量回归

引子1　影响房地产价格的复杂因素

很多研究认为，影响商品房价格的因素有多个方面，例如：

1.成本费用因素：包括土地、建筑物建造成本、其他费用；

2.房地产供求因素：包括住房需求量、房地产开发量等；

3.经济因素：包括宏观经济状况、物价状况、居民收入状况等；

4. 人口因素：包括人口密度、家庭结构等；

5.社会因素：包括社会治安、城市化水平、消费心理等；

6.行政(政策)因素：包括土地与住房制度、房地产价格政策等；

7.区域因素：包括所处地段的市政基础设施、交通状况等；

8.个别因素：包括朝向、结构、材料、功能设计、施工质量等；

9.房地产投机因素：投机者在房地产市场中的投机活动；

10.自然因素：包括自然环境、地质、地形、地势及气候等。

在影响房地产价格的众多因素中，有定量的因素：

成本因素、房地产供求因素、经济因素、人口因素等；

也有定性的因素：

社会因素、行政因素、区位因素、个别因素、投机因素、 自然因素等。

在研究房地产价格影响机理时，需要分析那些不易量化的定性因素对房地产价格是否真的有显著影响。

能否把定性的因素也引入计量经济模型中呢？怎样才能在模型中有效地表示这些定性因素的作用呢？

引子2　男女大学生的消费真的有差异吗?

当代大学生在消费结构呈现出多元化趋势。大学生除了日常生活费开支以外，还有人际交往、网络通讯、书报、衣着、化妆品、电脑、旅游、食品、学习用品、各种考证等消费。不同性别大学生的消费结构有所不同，专科生、本科生、研究生的消费结构更有差异。不同年级之间，男女同学之间，消费水平、消费结构、消费方式上都存在着差异。

（注：来源于新华网等：共青团中央、全国学联共同发布的《2004中国大学生消费与生活形态研究报告》）

为了研究男女大学生、不同层次大学生、不同年级大学生的消费结构是否有差异，需要将这些定性的因素引入计量模型，怎样才能在模型中有效地表示这类定性因素的作用呢?

第一节　虚拟变量

一、什么是虚拟变量

数量变量与属性变量

可用数量表现的连续变量 ⟷ 只表明属性的不连续变量

属性变量：不能精确计量的说明某种属性或状态的定性变量，如性别、民族、战争、政治事件

◆本身是定性的二分类变量(非此即彼)

◆本来是连续变量也可转换为定性变量(如上线/不上线)

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

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 伊拉克战争、“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
- 比较两个回归模型的差异
- 虚拟被解释变量模型：
  被解释变量本身是定性变量

二、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

作用：假定其他因素都不变，只研究某种定性因素在某定量变量上是否表现出显著差异

2、 解释变量中既含定量变量，又含虚拟变量

作用：研究定量变量和虚拟变量同时对被解释变量的影响

3、 虚拟被解释变量模型：被解释变量本身取值为0或1

作用：对某社会经济现象进行“是”与“否”判断研究

（离散选择模型）

三、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：D=0 如果是女性（基础类型）

D=1 如果是男性（比较类型）

D=0 为“911事件”以前（基础类型）

D=1 为“911事件”以后（比较类型）

D=0 不是大学毕业生（基础类型）

D=1 是大学毕业生（比较类型）

虚拟变量的设置规则

又如，研究东、中、西部地区收入X与消费支出Y的关系：

$$Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta X_i + u_i$$

$\begin{cases} D_1 = 1 & \text{为东部地区} \\ D_1 = 0 & \text{为其他} \end{cases}$　　$\begin{cases} D_2 = 1 & \text{为中部地区} \\ D_2 = 0 & \text{为其他} \end{cases}$

D_1 和 D_2 取值均为0的类型—基础类型：$Y_i = \alpha_0 + \beta X_i + u_i$

是比较的基准，α_0 代表了基准组（西部地区）的截距

D_1 或 D_2分别取值为1的类型—是与基础类型比较的类型

$D_1 = 1$时　$Y_i = \alpha_0 + \alpha_1 + \beta X_i + u_i$

$D_2 = 1$时　$Y_i = \alpha_0 + \alpha_2 + \beta X_i + u_i$

α_1 和 α_2 为差异截距系数

虚拟变量的设置原则

虚拟变量的个数须按以下原则确定：

每一定性变量所需的虚拟变量个数要比该定性变量的类别数少1，即如果有m个属性类别，只在模型中引入m-1个虚拟变量。

例子：已知冷饮的销售量Y除受k种定量变量X_k的影响外，还受春、夏、秋、冬四季变化的影响，要考察该四季的影响，只需引入三个虚拟变量即可：

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

则冷饮销售量的模型为：

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

在上述模型中，若再引入第四个虚拟变量

$$D_{4t} = \begin{cases} 1 & \text{冬季} \\ 0 & \text{其他} \end{cases}$$

则冷饮销售模型变为：

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

其矩阵形式为：

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

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，则式中的：

$$(\mathbf{X},\mathbf{D})=\begin{pmatrix}1 & X_{11} & \cdots & X_{k1} & 1 & 0 & 0 & 0\\ 1 & X_{12} & \cdots & X_{k2} & 0 & 1 & 0 & 0\\ 1 & X_{13} & \cdots & X_{k3} & 0 & 0 & 1 & 0\\ 1 & X_{14} & \cdots & X_{k4} & 0 & 0 & 0 & 1\\ 1 & X_{15} & \cdots & X_{k5} & 0 & 1 & 0 & 0\\ 1 & X_{16} & \cdots & X_{k6} & 1 & 0 & 0 & 0\end{pmatrix} \quad \beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{pmatrix} \quad \alpha=\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \alpha_4\end{pmatrix}$$

显然，(X,D)中的第1列可表示成后4列的线性组合，从而(X,D)不满秩，参数无法唯一求出。

这就是所谓的“虚拟变量陷阱”，应避免。

使用虚拟变量需注意的问题

- 虚拟变量陷阱：若定性变量有m个类别，则引入m个虚拟变量将会产生完全多重共线性问题，避免方法：
  - 只引入（m-1）个虚拟变量
  - 引入m个虚拟变量但去掉截距项
- 哪种方法更好：包含截距项更方便，可以很容易地检验某个组与基准组之间是否存在显著差异以及差异程度。

2、避免落入“ 虚拟变量陷阱”

(1)在有截距的模型中

如果模型中每个定性因素有m个相互排斥的类型，模型中只能引入m-1个虚拟变量，否则会出现完全多重共线性

例如：一个定性因素有三种类型，若设三个虚拟变量

$$Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i$$

若 $D_1 = 1$则$D_2 = 0, D_3 = 0$ ； 若 $D_2 = 1$则$D_1 = 0, D_3 = 0$ 等等。

显然此时 $D_1 + D_2 + D_3 = 1$ ，而截距α_0对应的变量为1，$D_1 + D_2 + D_3 = 1$ 再次生成了截距项，则导致了完全的多重共线性

(2)若模型中无截距项

模型为 $Y_i=\alpha_1 D_{1i}+\alpha_2 D_{2i}+\alpha_3 D_{3i}+\beta X_i+u_i$

此时虽然有 $D_1=1$则$D_2=0,D_3=0$ ，若$D_2=1$则$D_1=0,D_3=0$ ，

若 $D_3=1$则$D_1=0,D_2=0$ 且 $D_1+D_2+D_3=1$ ，但因为没有截距项，不会出现完全的多重共线性。

注意： 此时 $\alpha_1,\alpha_2,\alpha_3$ 等参数不再是差异截距系数，而分别是相应类型的截距。

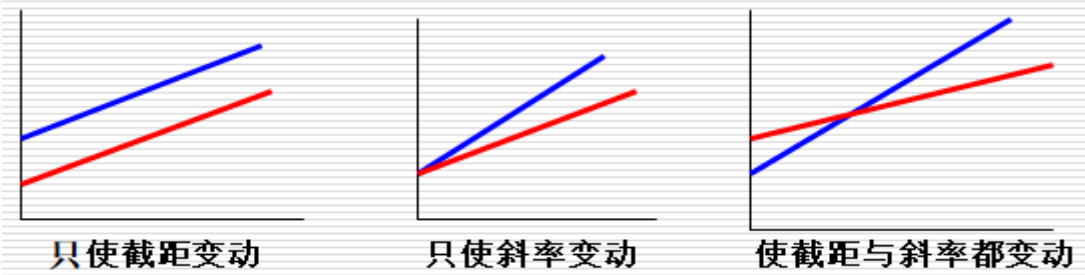
$D_1=1,D_2=0,D_3=0$时 $Y_i=\alpha_1+\beta X_i+u_i$

$D_2=1,D_1=0,D_3=0$时 $Y_i=\alpha_2+\beta X_i+u_i$

$D_3=1,D_1=0,D_2=0$时 $Y_i=\alpha_3+\beta X_i+u_i$

第二节　虚拟解释变量回归

定性变量作为解释变量，可以影响模型的截距，也可以影响模型的斜率，还可以同时影响截距和斜率



一、用虚拟变量表示不同截矩的回归——加法类型

虚拟变量以加法方式引入模型的作用：改变模型中截距，可分为各种情况去设置虚拟变量

虚拟变量的引入

1、加法方式

企业职工薪金模型中性别虚拟变量的引入采取了加法方式。

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

在该模型中，如果仍假定$E(\mu_i)=0$，则

企业女职工的平均薪金为：

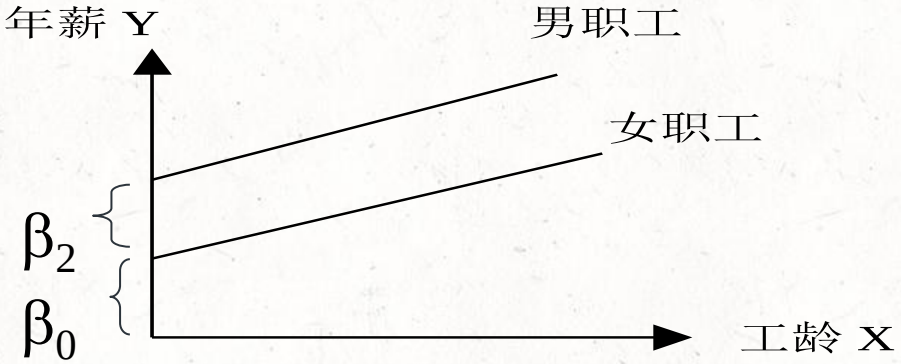
$$E(Y_i \mid X_i, D_i = 0) = \beta_0 + \beta_1 X_i$$

企业男职工的平均薪金为：

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

几何意义：

- 假定$\beta_2>0$，则两个函数有相同的斜率，但有不同的截距。意即，男女职工平均薪金对教龄的变化率是一样的，但两者的平均薪金水平相差β_2。
- 可以通过传统的回归检验，对β_2的统计显著性进行检验，以判断企业男女职工的平均薪金水平是否有显著差异。



又例：在横截面数据基础上，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：高中以下， 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：

$$Y_i=\beta_0+\beta_1X_i+\beta_2D_1+\beta_3D_2+\mu_i$$

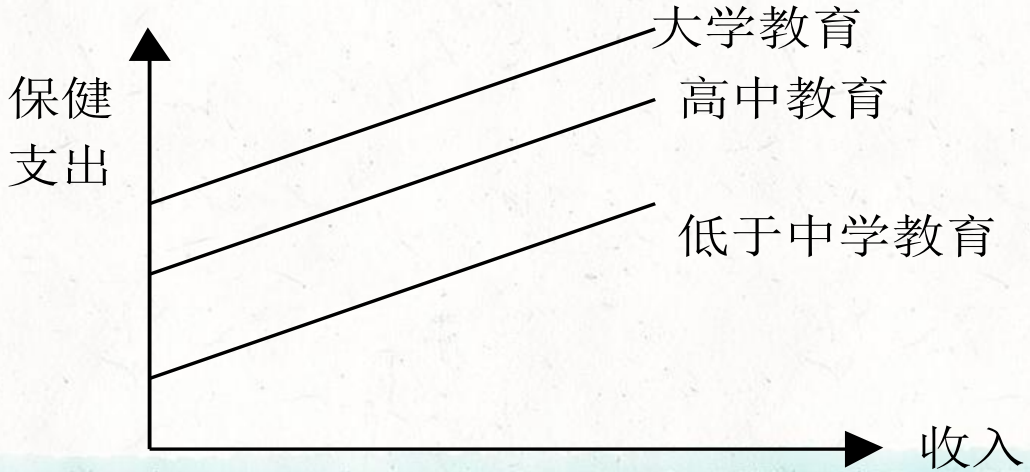
在$E(\mu_i)=0$的初始假定下，高中以下、高中、大学及其以上教育水平下个人保健支出的函数：

高中以下：$E(Y_i \mid X_i, D_1=0, D_2=0)=\beta_0+\beta_1 X_i$

高中：$E(Y_i \mid X_i, D_1=1, D_2=0)=(\beta_0+\beta_2)+\beta_1 X_i$

大学及其以上：$E(Y_i \mid X_i, D_1=0, D_2=1)=(\beta_0+\beta_3)+\beta_1 X_i$

假定$\beta_3>\beta_2$，其几何意义：



还可将多个虚拟变量引入模型中以考察多种“定性”因素的影响。

如在上述职工薪金的例中，再引入代表学历的虚拟变量D_2：

$$D_2=\begin{cases}1 & \text{本科及以上学历}\\ 0 & \text{本科以下学历}\end{cases}$$

职工薪金的回归模型可设计为：

$$Y_i=\beta_0+\beta_1X_i+\beta_2D_1+\beta_3D_2+\mu_i$$

于是，不同性别、不同学历职工的平均薪金分别为：

- 女职工本科以下学历的平均薪金：

$$E(Y_i \mid X_i, D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X_i$$

- 男职工本科以下学历的平均薪金：

$$E(Y_i \mid X_i, D_1 = 1, D_2 = 0) = (\beta_0 + \beta_2) + \beta_1 X_i$$

- 女职工本科以上学历的平均薪金：

$$E(Y_i \mid X_i, D_1 = 0, D_2 = 1) = (\beta_0 + \beta_3) + \beta_1 X_i$$

- 男职工本科以上学历的平均薪金：

$$E(Y_i \mid X_i, D_1 = 1, D_2 = 1) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 X_i$$

1. 解释变量只有一个分为两种类型的定性变量无定量变量的回归

这种模型又称方差分析模型 $Y_i = \alpha + \beta D_i + u_i$

其中：Y为公立学校教师工资，D=0为农村学校；D=1为城镇学校

分析条件期望：

基础类型：$E(Y_i | D = 0) = \alpha$

比较类型：$E(Y_i | D = 1) = \alpha + \beta$

β 为差异截距系数，通过对系数β的 t 检验：可检验

在其他因素不变的条件下，城乡教师的工资是否有显著差别

2、解释变量包含一个定量变量和一个分为两种类型的定性变量的回归

$$Y_i = \alpha_0 + \alpha_1 D_i + \beta X_i + u_i$$

例如：Y为服装消费　　X为收入，

D＝0为男性　　D=1为女性

分析条件期望：

基础类型：$E(Y_i | X_i, D=0) = \alpha_0 + \beta X_i$

比较类型：$E(Y_i | X_i, D=1) = (\alpha_0 + \alpha_1) + \beta X_i$

α_1 为差异截距系数

对系数α_1的 t 检验：可检验定性因素对截距是否有显著影响

注意：

- u_i 应服从基本假定
- 这里一个定性变量具有两种类型，只使用了一个虚拟变量（为什么？）

3、解释变量包含一个定量变量和一个两种以上类型的定性变量的回归

类型：高中以下、高中毕业、大学毕业及以上——三种类型

模型 $Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta X_i + u_i$

例如 Y_i ——年工资　　X_i——工龄

$\begin{cases} D_1 = 1 & \text{只是高中毕业} \\ D_1 = 0 & \text{其他} \end{cases}$　　$\begin{cases} D_2 = 1 & \text{大学毕业及以上} \\ D_2 = 0 & \text{其他} \end{cases}$

基础类型：$E(Y_i | X_i, D_1 = 0, D_2 = 0) = \alpha_0 + \beta X_i$　（高中以下）

比较类型：$E(Y_i | X_i, D_1 = 1, D_2 = 0) = (\alpha_0 + \alpha_1) + \beta X_i$　（高中）

$E(Y_i | X_i, D_1 = 0, D_2 = 1) = (\alpha_0 + \alpha_2) + \beta X_i$　（大学及以上）

差异截距系数为 α_1 和 α_2

问题：如果还要区分“专科”“本科”、“硕士”、“博士”应怎么办？

注意：

- u_i 应服从基本假定
- 一个定性变量有三种类型，使用了两个虚拟变量，

  D_1 和D_2代表的是同一个定性变量的两种不同类型
- 两个差异截距系数α_1和α_2表示的都是与基础类型的差异
- 一个定性变量有多种类型时，虚拟变量可同时取值为0，但不能同时取值为1，因同一定性变量的各类型间“非此即彼”

4、解释变量包含一个定量变量和两个定性变量

模型 $$Y_i=\alpha_0+\alpha_1 D_{1i}+\alpha_2 D_{2i}+\beta X_i+u_i$$

这里的D_{1i}和D_{2i}代表的是两个不同的定性变量,各分为两种类型

例如：Y为文化支出，X为收入

$$\begin{cases} D_{1i}=0 & \text{农村居民} \\ D_{1i}=1 & \text{城镇居民} \end{cases} \qquad \begin{cases} D_{2i}=0 & \text{高中以下文化程度} \\ D_{2i}=1 & \text{高中及以上文化程度} \end{cases}$$

基础类型：$$E(Y_i|X_i, D_1=0, D_2=0)=\alpha_0+\beta_1 X_i$$

对比类型：$$E(Y_i|X_i, D_1=1, D_2=0)=(\alpha_0+\alpha_1)+\beta_1 X_i$$

$$E(Y_i|X_i, D_1=0, D_2=1)=(\alpha_0+\alpha_2)+\beta_1 X_i$$

$$E(Y_i|X_i, D_1=1, D_2=1)=(\alpha_0+\alpha_1+\alpha_2)+\beta_1 X_i$$

用t检验分别检验 α_1 和 α_2 的统计显著性：验证两个定性变量对截距是否有显著影响

注意：

- u_i应服从基本假定
- 两个定性变量分别有两种类型，用了两个虚拟变量（为什么？）

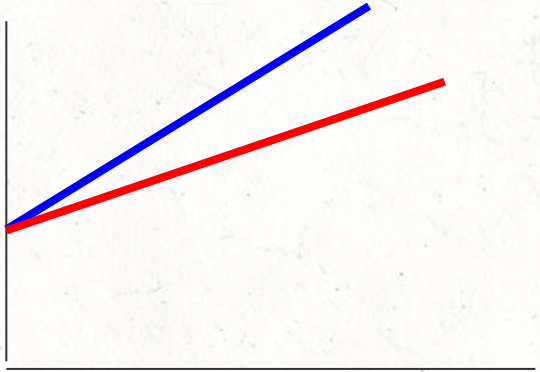
两个定性变量和一个定性变量三种类型都用了两个虚拟变量，但其性质是不同的

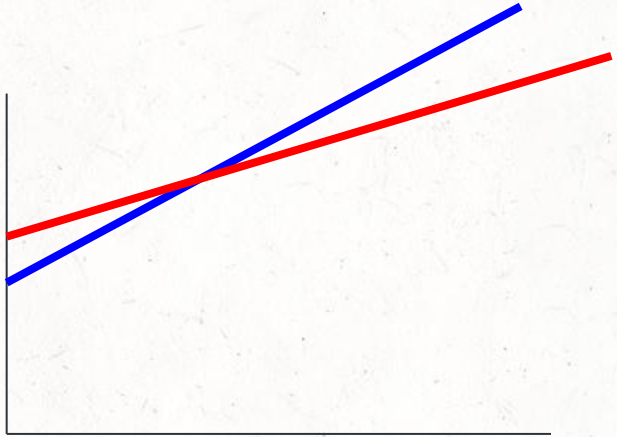
- K个定性变量可选用K个虚拟变量去表示，这不会出现“虚拟变量陷阱”
- 代表不同定性变量的虚拟变量，可以同时为0，也可同时为1，因为不同定性变量间没有“非此即彼”的关系。

二、用虚拟变量表示不同斜率的回归

——乘法类型

模型中斜率系数的差异，可用以乘法形式引入的虚拟变量去表示。





乘法方式

- 加法方式引入虚拟变量，考察：截距的不同，
- 许多情况下：往往是斜率就有变化，或斜率、截距同时发生变化。
- 斜率的变化可通过以乘法的方式引入虚拟变量来测度。

例：根据消费理论，消费水平C主要取决于收入水平Y，但在一个较长的时期，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通过在收入的系数中引入虚拟变量来考察。

如设

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

消费模型可建立如下：

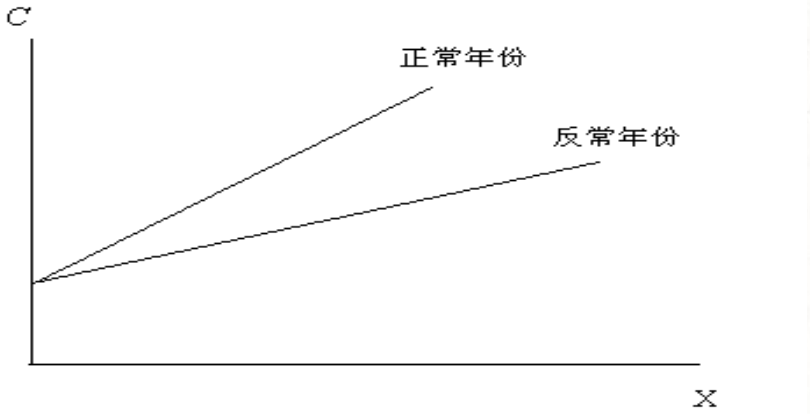
$$C_t = \beta_0 + \beta_1 X_t + \beta_2 D_t X_t + \mu_t$$

- 这里，虚拟变量D以与X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。

- 假定$E(\mu_i)=0$，上述模型所表示的函数可化为：

正常年份：$$E(C_t \mid X_t, D_t = 1) = \beta_0 + (\beta_1 + \beta_2)X_t$$

反常年份：$$E(C_t \mid X_t, D_t = 0) = \beta_0 + \beta_1 X_t$$



当截距与斜率发生变化时，则需要同时引入加法与乘法形式的虚拟变量。

例：考察1990年前后的中国居民的总储蓄-收入关系是否已发生变化。下表给出了中国1979~2001年以城乡储蓄存款余额代表的居民储蓄以及以GNP代表的居民收入的数据。

表 5.1.1　　1979~2001 年中国居民储蓄与收入数据（亿元）

90年前	储蓄	GNP	90年后	储蓄	GNP
1979	281	4038.2	1991	9107	21662.5
1980	399.5	4517.8	1992	11545.4	26651.9
1981	523.7	4860.3	1993	14762.4	34560.5
1982	675.4	5301.8	1994	21518.8	46670.0
1983	892.5	5957.4	1995	29662.3	57494.9
1984	1214.7	7206.7	1996	38520.8	66850.5
1985	1622.6	8989.1	1997	46279.8	73142.7
1986	2237.6	10201.4	1998	53407.5	76967.2
1987	3073.3	11954.5	1999	59621.8	80579.4
1988	3801.5	14922.3	2000	64332.4	88228.1
1989	5146.9	16917.8	2001	73762.4	94346.4
1990	7034.2	18598.4			

以Y为储蓄，X为收入，可令：

1990年前：$Y_i=\alpha_1+\alpha_2X_i+\mu_{1i}$　　$i=1,2\ldots,n_1$

1990年后：$Y_i=\beta_1+\beta_2X_i+\mu_{2i}$　　$i=1,2\ldots,n_2$

则有可能出现下述四种情况中的一种：

1. $\alpha_1=\beta_1$，且$\alpha_2=\beta_2$，称为**重合回归。**

2. $\alpha_1\neq\beta_1$,但$\alpha_2=\beta_2$，差异仅在其截距，称为**平行回归。**

3. $\alpha_1=\beta_1$，但$\alpha_2\neq\beta_2$，差异仅在其斜率，称为**同截距回归**

4. $\alpha_1\neq\beta_1$，且$\alpha_2\neq\beta_2$，两个回归完全不同，称为**非相似回归。**

虚拟变量模型的应用

- 虚拟变量是一个能处理一系列有趣问题的灵活工具。虚拟变量模型的应用包括：
  - 结构变化的检验
  - 虚拟变量的交互效应
  - 分段线性回归
  - 时间序列数据中的季节调整

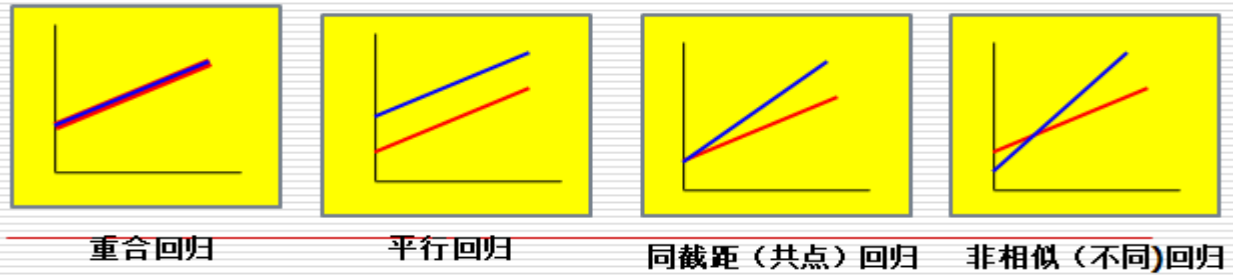
1. 回归模型比较——结构变化的检验

回顾：邹氏参数稳定性检验可以检验模型结构是否发生了变化：结构无变化 $\beta = \alpha$ 作受约束模型；结构变化 $\beta \neq \alpha$ 作无约束模型

$$F = \frac{(RSS_R - RSS_U)/k}{RSS_U/(n_1 + n_2 - 2k)} \square F[k, (n_1 + n_2 - 2k)]$$

邹氏检验只能检验模型结构是否发生变化，不能说明具体变化了多少，也不能说明究竟是截距变化还是斜率变化。

例如：怎样说明以下变化呢?



结构变化的检验

模型 $Y_i = \alpha_0 + \alpha_1 D_i + \beta_1 X_i + \beta_2 (D_i X_i) + u_i$

基础类型： $E(Y_i | X_i, D = 0) = \alpha_0 + \beta_1 X_i$

对比类型： $E(Y_i | X_i, D = 1) = (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2) X_i$

可看出：以加法引入虚拟变量D的系数是截距的差异系数，

以乘法引入虚拟变量D的系数是斜率的差异系数

用t检验分别检验 α_1 和 β_2 的显著性：可检验此定性变量对截距和斜率是否有显著影响，即检验两个回归的结构是否有差异

优点：

- 用一个回归替代了多个回归，简化了分析过程
- 可方便地检验各种假设
- 合并回归增加了自由度，提高参数估计的精确性

注意：

- 所比较的方程应是同方差，否则会出现异方差
- u_i 应服从基本假定

2. 交互效应分析

基本思想：分析两个定性变量对被解释变量影响的虚拟变量模型，暗含着假定：两个定性变量是分别独立影响被解释变量的。但在实际经济活动中，两个定性变量对被解释变量的影响可能存在交互作用。为描述这种交互作用，可把代表两个定性因素的虚拟变量的乘积以加法形式引入模型。

模型：$Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 (D_{1i} D_{2i}) + \beta X_i + u_i$

其中：D_{1i} ——代表第一个定性变量的虚拟变量

D_{2i} ——代表第二个定性变量的虚拟变量

（$D_{1i} D_{2i}$）——描述二者交互效应的虚拟变量

因为 $E(Y_i | X_i, D_1 = 1, D_2 = 1) = (\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3) + \beta X_i$

α_3是交互效应的截距差异系数，可以通过对 α_3 的显著性的检验，判断是否存在交互效应

例如 $Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 (D_{1i} D_{2i}) + \beta X_i + u_i$

其中：Y_i——种油菜籽和养蜂的收入　X_i ——投入资金

D_{1i}——代表是否种油菜籽的虚拟变量

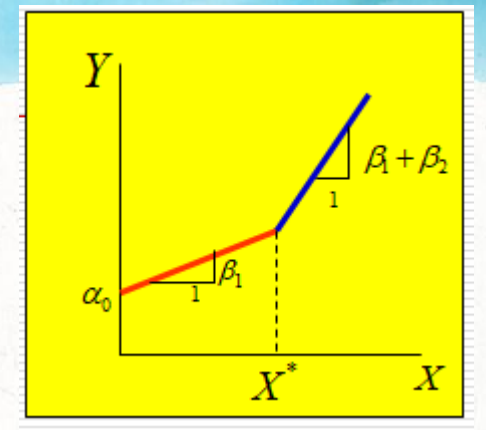
$D_{1i}=1$ 种油菜籽　　$D_{1i}=0$ 不种油菜籽

D_{2i}——代表是否养蜂的虚拟变量

$D_{2i}=1$ 养蜂　　$D_{2i}=0$ 不养蜂

（$D_{1i} D_{2i}$）——描述种油菜籽与养蜂的交互效应

3.分段线性回归



基本思想：

有的社会经济现象的变动，会在解释变量达到某个临界值时发生突变，为了区分不同阶段的截距和斜率可利用虚拟变量进行分段回归

第一段回归，当 $X_i < X^*$ 时（ X^* 是临界值）$Y_i = \alpha_0 + \beta_1 X_i + u_i^{(1)}$

第二段回归，当 $X_i \geq X^*$ 时

$$Y_i = (\alpha_0 + \beta_1 X^*) + (\beta_1 + \beta_2)(X_i - X^*) + u_i^{(2)}$$

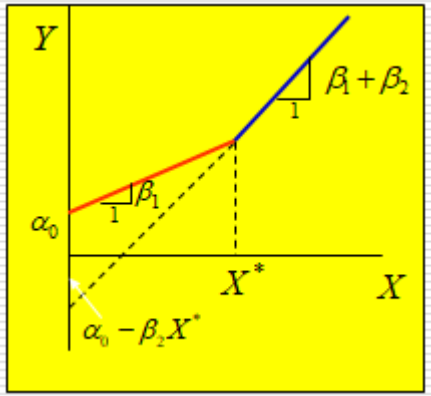
整理得 $Y_i = \alpha_0 - \beta_2 X^* + (\beta_1 + \beta_2) X_i + u_i^{(2)}$

例如:不同销售业绩的奖励方式不同

具体作法：

模型形式 $Y_I = \alpha_0 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$

其中：$D = \begin{cases} 1 \text{ 若 } X_i \geq X^* \\ 0 \text{ 若 } X_i < X^* \end{cases}$



第一段回归

$$E(Y_i \mid X_i, X^*, D_i = 0) = \alpha_0 + \beta_1 X_i$$

第二段回归

$$E(Y_i \mid X_i, X^*, D_i = 1) = \alpha_0 - \beta_2 X^* + (\beta_1 + \beta_2) X_i$$

注意：●第一、二段回归不仅截距不同，而且斜率也不同

●分为两段回归时用了一个虚拟变量

推理：分为K段回归时，可用K—1个虚拟变量

4.季节变动分析中的应用

思想：时间序列数据可分解为四个因素：

长期趋势； 季节变动； 循环变动； 随机（不规则）变动

为消除季节变动影响，常用修匀方法。为预测某季度变量又需加入季节因素。也可利用虚拟变量方法区分季节因素。

方法：例如某商品销售量Q与价格P有关，可能还与季节有关

(1)引入四个季度影响因素

$$Q_t = \alpha_0 + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \beta P_t + u_t$$

其中：Q_t —销售量 P_t —价格

$\begin{cases} D_1 = 1 \text{ 为二季度} \\ D_1 = 0 \text{ 为其它} \end{cases}$ $\begin{cases} D_2 = 1 \text{ 为三季度} \\ D_2 = 0 \text{ 为其它} \end{cases}$ $\begin{cases} D_3 = 1 \text{为四季度} \\ D_3 = 0 \text{为其它} \end{cases}$

(2)显著性检验

对 $\alpha_1,\alpha_2,\alpha_3$ 作 t 检验，若显著不为0，表明该季度有季节变化影响；若显著为0，表明不存在季节变动影响

(3)重建季节变动模型

如只是二季度有明显季节性变动，可省略D_2,D_3 重建模型

$$Q_t=\alpha_0+\alpha_1 D_{1t}+\beta P_t+u_t$$

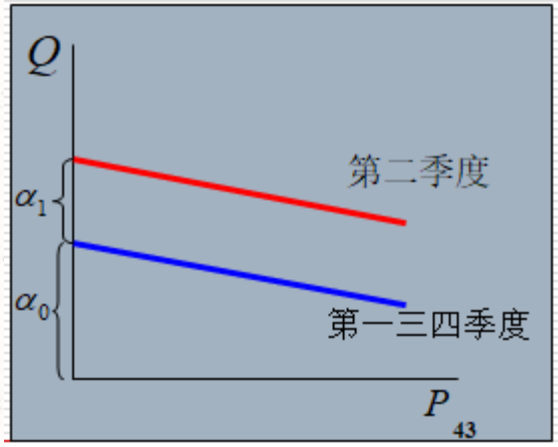
D=1为二季度； D=0为一、三、四季度

在一、三、四季度时

$$E(Q_t|P_t,D_1=0)=\alpha_0+\beta P_t$$

在二季度时

$$E(Q_t|P_t,D_1=1)=(\alpha_0+\alpha_1)+\beta P_t$$



第三节 虚拟被解释变量

有时所研究的经济现象本身可能是定性变量。

例如：是否购买住房？　是否购买汽车？

是否参加保险？　是否按期归还贷款？

定性的被研究对象作为被解释变量，也可用虚拟变量0或1表示，其取值可能受多种因素影响。

虚拟被解释变量模型的估计和检验会产生一些特殊的问题。将在中级计量经济学“离散选择模型”中讨论。

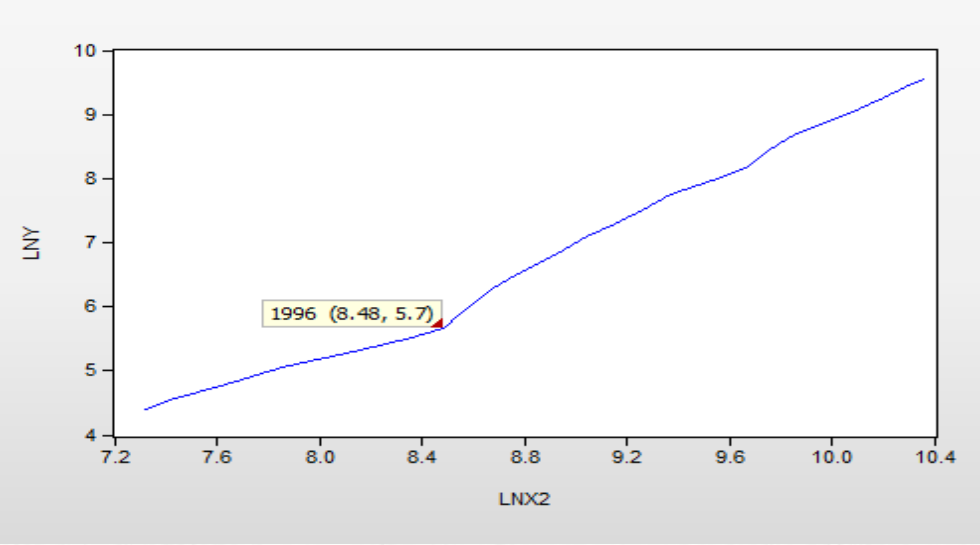
第四节　案例分析

一、问题提出：汽车工业向来有“火车头工业”之称，作为我国的支柱产业之一，很大程度上改变了中国人的生活方式。随着经济的快速发展，存在着很多影响私人汽车拥有量的因素，私人汽车拥有量正在以每年一千多万辆的速度持续增长。

二、数据：表8.1选取了1990-2015年中国私人汽车拥有量、城镇居民人均可支配收入、公路里程数、公路运营汽车拥有量、原油产量以及一年期贷款利率等相关数据，对中国私人汽车拥有量的主要影响因素进行分析（见p203-204）。

三、建立模型

私人汽车作为现代家庭的消费品，首先要受到居民可支配收入的较大影响。因此，先研究城镇居民可支配收入对私人汽车拥有量的弹性，为此对私人汽车拥有量Y和城镇居民可支配收入分别取对数，其相关关系如图所示：



从图中可以发现，城镇居民可支配收入的对数lnX2与私人汽车拥有量的对数lnY存在较明显的折线关系，转折点出现在1996年，转折点前后分别近似线性形式。

三、建立模型

为了分析私人汽车拥有量在1990—2015年不同时期的数量关系，以1996年度的转折点作为依据，引入虚拟变量D1，1996年度所对应的lnX2为8.484。据此，我们设定了如下以分段回归形式引入虚拟变量的模型：

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 (\ln X_{2t} - 8.484) D_{1t} + u_t$$

式中，

$$D_{1t} = \begin{cases} 1 & (t = 1996\text{年以后}) \\ 0 & (t = 1996\text{年及以前}) \end{cases}$$

注：虚拟变量D1需要在工作文件中命名一个新序列，即在命令窗中输入“series d1”，然后在d1中人工输入0和1的数值，1996之后为1，1996年及以前为0。

四、回归结果

命令窗中输入命令”**ls lny c lnx2 (lnx2-8.484)*d1**”，即可得到回归结果：

Dependent Variable: LNY
Method: Least Squares
Date: 12/17/19 Time: 09:45
Sample: 1990 2015
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.776594	0.488808	-9.771915	0.0000
LNX2	1.251881	0.059887	20.90402	0.0000
(LNX2-8.484)*D1	0.769620	0.079961	9.624889	0.0000

R-squared	0.997031	Mean dependent var	7.023911
Adjusted R-squared	0.996773	S.D. dependent var	1.611117
S.E. of regression	0.091527	Akaike info criterion	-1.836195
Sum squared resid	0.192676	Schwarz criterion	-1.691030
Log likelihood	26.87054	Hannan-Quinn criter.	-1.794393
F-statistic	3861.650	Durbin-Watson stat	0.376155
Prob(F-statistic)	0.000000		

四、回归结果

由于 $\ln X_{2t}$ 和 $(\ln X_{2t}-8.484)D_{1t}$ 参数估计量的 t 值均大于临界值 $t_{0.05}(26-3)=2.069$，表明各解释变量的斜率系数在显著性水平 $\alpha=0.05$ 下显著地不等于 0，私人汽车拥有量的回归模型分别为：

$$\widehat{\ln Y_t}=\begin{cases}-4.78+1.25\ln X_{2t} & t\le 1996\\ -11.31+2.02\ln X_{2t} & t>1996\end{cases}$$

这表明 1996 年前后私人汽车拥有量的回归方程在统计意义上确实有所不同。1996 年以前城镇居民可支配收入每增加 1%，平均说来私人汽车拥有量的增长率为 1.25%。1996 年以后随着城镇居民可支配收入的增加， 私人汽车拥有量的增速较 1996 年之前有所增加。上述模型同城镇居民可支配收入的对数与私人汽车拥有量的对数之间的散布图吻合，与这一时段中国的实际经济运行状况也是相符的。

五、进一步分析

需要注意的是，在上述建模过程中，为了方便通过图形特征建立模型，没有考虑公路里程数等其它影响私人汽车拥有量的因素。

如果要将其它因素考虑进来建立多元回归模型，它们之间的折线关系就不一定成立了，也就不能直接利用前面的折线结构去建立模型。

可是，在多元回归中，要绘制控制其它因素不变时两个变量的散布图比较困难，因此通常可利用经济背景来设定虚拟变量以改善模型的拟合效果。

首先用私人汽车拥有量的对数lnY对城镇居民可支配收入的对数lnX2以及公路里程等其它变量进行回归，结果见下页：

五、进一步分析

Dependent Variable: LNY
Method: Least Squares
Date: 12/17/19 Time: 10:11
Sample: 1990 2015
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.144258	0.287957	-14.39193	0.0000
LNX2	0.978048	0.060236	16.23691	0.0000
X3	0.001886	0.000291	6.472782	0.0000
X4	0.000464	5.72E-05	8.112988	0.0000
X5	9.74E-05	3.26E-05	2.987080	0.0073
X6	-0.016363	0.007454	-2.195105	0.0401

R-squared	0.999310	Mean dependent var	7.023911
Adjusted R-squared	0.999138	S.D. dependent var	1.611117
S.E. of regression	0.047315	Akaike info criterion	-3.064788
Sum squared resid	0.044775	Schwarz criterion	-2.774458
Log likelihood	45.84225	Hannan-Quinn criter.	-2.981184
F-statistic	5793.217	Durbin-Watson stat	1.675216
Prob(F-statistic)	0.000000		

从图中t检验的p值看出，各解释变量在显著性水平0.05下均显著。但DW统计量1.6752，显示模型不能排除存在自相关，这可能是由于模型忽略了某些重要变量造成的。

五、进一步分析

2001 年 12 月 11 日，世贸组织正式宣布中国成为世贸组织的一员，这意味着中国市场能与国际市场相连接，极大地促进了中国经济的发展和人民的生活水平的提高。这一变化同样也可能会影响到中国的私人汽车拥有量。尝试以 2001 年底加入 WTO 为转折点，在私人汽车拥有量的对数 lnY 对城镇居民可支配收入的对数 $\ln X_2$ 以及公路里程等其它变量进行回归的基础上，引入虚拟变量，建立以下多元回归模型：

$$\ln Y_t = \beta_1 + \beta_2 \ln X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + \beta_7 D_{2t} + u_t$$

式中

$$D_{2t} = \begin{cases} 1 & t = 2001\text{年以后} \\ 0 & t = 2001\text{年及以前} \end{cases}$$

注：引入虚拟变量d2的方式和输入回归命令的方式，均与前面类似。

五、进一步分析

Dependent Variable: LNY
Method: Least Squares
Date: 12/17/19 Time: 10:26
Sample: 1990 2015
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.191814	0.266712	-15.71661	0.0000
LNX2	0.959423	0.056286	17.04555	0.0000
X3	0.001664	0.000289	5.766804	0.0000
X4	0.000428	5.54E-05	7.725793	0.0000
X5	0.000111	3.08E-05	3.608367	0.0019
X6	-0.014315	0.006947	-2.060462	0.0533
D2	0.078505	0.037091	2.116532	0.0477

R-squared	0.999442	Mean dependent var	7.023911
Adjusted R-squared	0.999265	S.D. dependent var	1.611117
S.E. of regression	0.043669	Akaike info criterion	-3.199563
Sum squared resid	0.036232	Schwarz criterion	-2.860845
Log likelihood	48.59432	Hannan-Quinn criter.	-3.102024
F-statistic	5668.373	Durbin-Watson stat	1.995419
Prob(F-statistic)	0.000000		

五、进一步分析

引入加入WTO的虚拟变量以后，模型的所有解释变量在显著性水平0.1下均显著，并且DW统计量显示模型不存在自相关。因此相比于没有引入虚拟变量的模型，模型设定更为合理，模型结果也与实际经济运行相一致。

在解释此模型虚拟变量D2的经济意义时，需要注意被解释变量为私人汽车拥有量的对数，因此虚拟变量D2的参数0.0785表示的是，固定其它因素不变加入WTO以后平均来说私有汽车拥有量是加入WTO之前的1.08倍（$e^{0.0785}$倍）。或者说，加入WTO以后平均来说私有汽车拥有量比加入WTO之前多8%。

需要指出的是，在上述建模过程中，主要是从教学的目的出发，说明运用虚拟变量的规则和方法，没有考虑更多建模的可能性。在实证分析中，还可以进一步考虑分段回归更多的分段点、多元回归中更多的时间分割点以及虚拟变量对回归系数的影响等。