经济学类各专业核心课程

计量经济学

计量经济学

第八章

虚拟变量回归

引子：男女大学生消费真有差异吗?

在校生的消费行为呈现多元化的结构：人际交往消费、手机类消费、衣着类消费、化妆品类消费、电脑类消费、旅游类消费占有较大的比例；而食品类消费、学习用品类消费不突显。

男女生在消费上存在差异。为了了解男、女生的消费支出结构差异，应当如何建立模型？

面临的问题：如何把男女生这样的非数量变量引入方程？

问题的一般性描述

在实际建模中，许多经济变量是可以定量度量，如：商品需求量、价格、收入、产量等

也有一些影响经济变量的因素无法定量度量，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）销售的影响等等。

如何对非定量因素进行回归分析？

采用“虚拟变量”对定性变量进行量化。

第八章　虚拟变量回归

本章主要讨论：

- 虚拟变量
- 虚拟解释变量的回归
- 虚拟被解释变量的回归 (选讲, 不包括)

第一节　虚拟变量

本节基本内容：

- 基本概念
- 虚拟变量设置规则

一、基本概念

定量因素：可直接测度、数值性的因素。

定性因素：属性因素，表征某种属性存在与否的非数值性的因素。

基本思想：将定性因素定量化。

这种“量化”通常是通过引入“虚拟变量”来完成的。根据这些因素的属性类型，构造只取“0”或“1”的人工变量，通常称为**虚拟变量（dummy variables）**，记为D。

- 例如，反映文程度的虚拟变量可取为：

$$D=\begin{cases}1, & \text{本科学历}\\ 0, & \text{非本科学历}\end{cases}$$

一般地，在虚拟变量的设置中：

- **基础类型、肯定类型取值为1；**
- **比较类型，否定类型取值为0。**

概念：

同时含有一般解释变量与虚拟变量的模型称为虚拟变量模型或者方差分析（analysis-of variance: ANOVA）模型。

一个以性别为虚拟变量考察企业职工薪金的模型：

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

其中：Y_i为企业职工的薪金，X_i为工龄，

D_i=1，若是男性，D_i=0，若是女性。

二、虚拟变量设置规则

虚拟变量的设置规则涉及三个方面：

1.“0”和“1”选取原则

2.属性（状态、水平）因素与设置虚拟变量数量的关系

3.虚拟变量在回归分析中的角色以及作用等方面的问题

1. “0”和“1”选取原则

- 虚拟变量取“1”或“0”的原则，应从分析问题的目的出发予以界定。
- 从理论上讲，虚拟变量取“0”值通常代表比较的基础类型；而虚拟变量取“1”值通常代表被比较的类型。

“0”代表基期（比较的基础，参照物）；

“1”代表报告期（被比较的效应）。

例如，比较收入时考察性别的作用。当研究男性收入是否高于女性时，是将女性作为比较的基础（参照物），故有男性为“1”，女性为“0”。

例1

（1） $D=\begin{cases}1 & \text{男}\\ 0 & \text{女}\end{cases}$

（2） $D=\begin{cases}1 & \text{改革开放以后}\\ 0 & \text{改革开放以前}\end{cases}$

（3） $D_1=\begin{cases}1 & \text{天气阴}\\ 0 & \text{其他}\end{cases}$

（4） $D_2=\begin{cases}1 & \text{天气雨}\\ 0 & \text{其他}\end{cases}$

问题：

为何只选0、1，选2、3、4行吗？为什么？

2. 属性的状态（水平）数与虚拟变量数量的关系

定性因素的属性既可能为两种状态，也可能为多种状态。例如，性别（男、女两种）、季节（4种状态），地理位置（东、中、西部），行业归属，所有制，收入的分组等。

$$\text{如：}\ (D_1, D_2) = \begin{cases} (1,0) & \text{天气阴} \\ (0,1) & \text{天气雨} \\ (0,0) & \text{其　他} \end{cases}$$

虚拟变量数量的设置规则

1. 若定性因素具有m（m>2)个相互排斥属性(或几个水平)，当回归模型有截距项时，只能引入(m-1)个虚拟变量；

2. 当回归模型无截距项时，则可引入m个虚拟变量；否则，就会陷入“虚拟变量陷阱”。（为什么？）

3. 虚拟变量在回归模型中的角色

虚拟变量既可作为被解释变量，也可作为解释变量，分别称其为虚拟被解释变量和虚拟解释变量。

虚拟被解释变量的研究是当前计量经济学研究的前沿领域，如MacFadden、Heckmen等人的微观计量经济学研究，大量涉及到虚拟被解释变量的分析。

本课程只是讨论虚拟解释变量的问题

第二节　虚拟解释变量的回归

本节基本内容：

- 加法类型
- 乘法类型
- 虚拟解释变量综合应用

一、虚拟变量的引入

虚拟变量做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

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。

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：

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

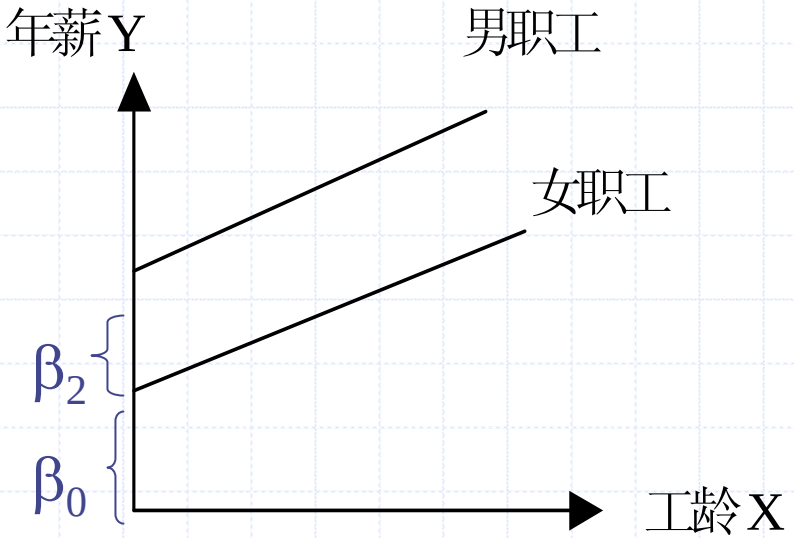
企业男职工的平均薪金为：

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

例子

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：

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

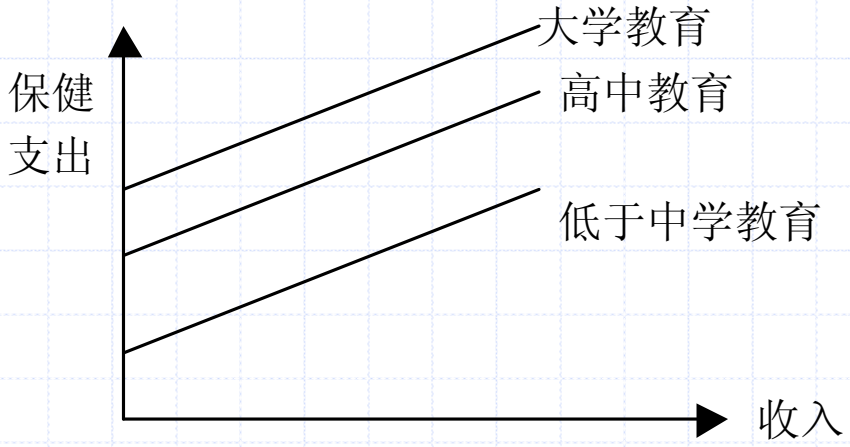
模型可设定如下：

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

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：

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

2、乘法方式

- 加法方式引入虚拟变量，考察：截距的不变
- 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。

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：

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

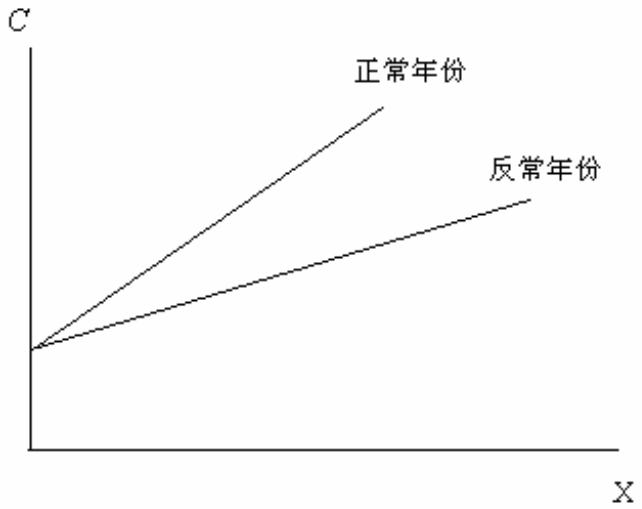
- 这里，虚拟变量**D**以与**X**相乘的方式引入了模型中，从而可用来考察消费倾向的变化。
- 假定**$E(\mu_i) = 0$**，上述模型所表示的函数可化为：

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：

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



截距和斜率均发生变化

模型形式：

$$Y_i = f\left(X_t, D_t, D_t X_t\right) \Rightarrow \alpha = \alpha_0 + \alpha_1 D, \beta = \beta_1 + \beta_2 D$$

例，同样研究消费支出 Y 、收入 X 、年份状况 D 间的影响关系。

$$Y_t = \alpha_0 + \beta_1 X_t + \alpha_1 D_t + \beta_2 (D_t X_t) + \mu_t$$

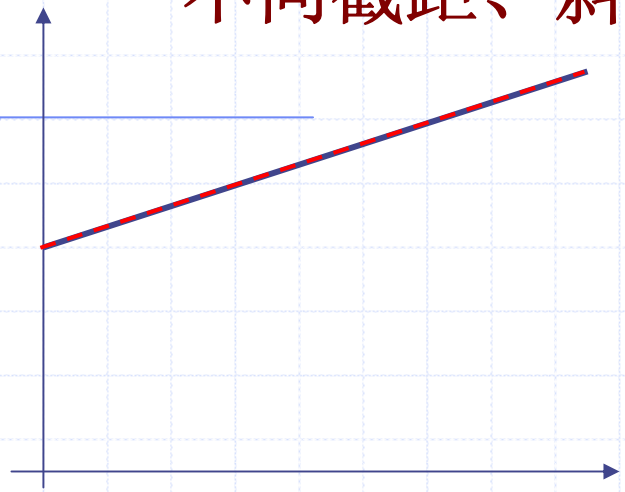
其中：Y – 消费支出；X – 收入；$D_t = \begin{cases} 1 & \text{反常年份} \\ 0 & \text{正常年份} \end{cases}$

反常年份 $\mathrm{E}\left(Y_t \mid X_t, D_t = 1\right) = \left(\alpha_0 + \alpha_1\right) + (\beta_1 + \beta_2) X_t$

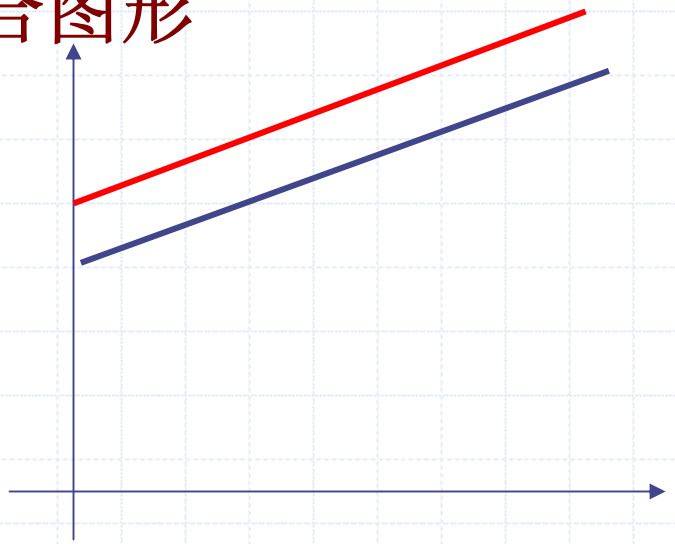
正常年份 $\mathrm{E}\left(Y_t \mid X_t, D_t = 0\right) = \alpha + \beta_1 X_t$

在正常年份基础上比较，截距和斜率系数都改变，为什么？

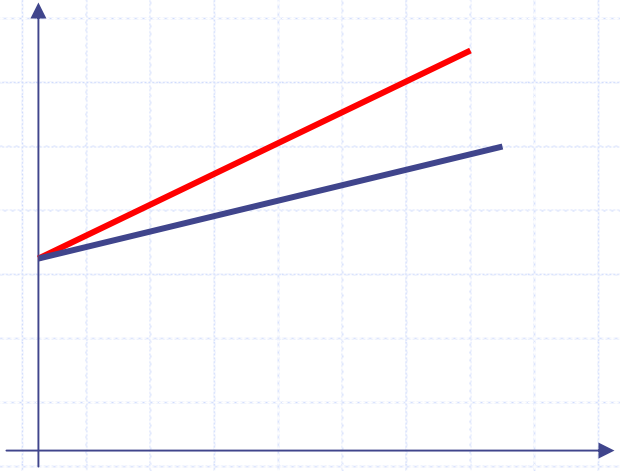
不同截距、斜率的组合图形



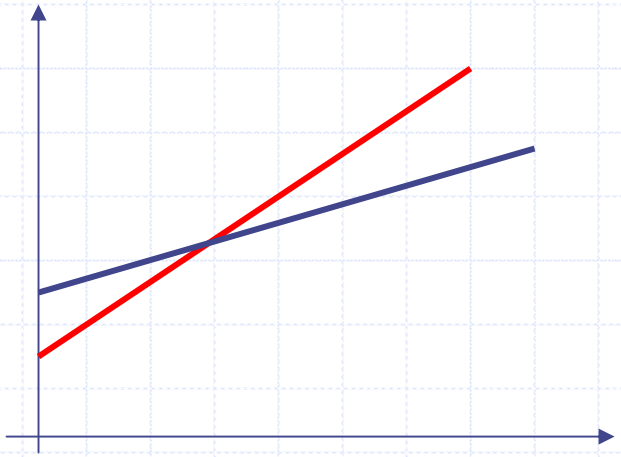
重合回归：截距斜率均相同



平行回归：截距不同斜率相同



共点回归：截距相同斜率不同



交叉（不同）回归：截距斜率均不同

二、虚拟解释变量综合应用

所谓综合应用是指将引入虚拟解释变量的加法方式、乘法方式进行综合使用。

基本分析方式仍然是条件期望分析。

本节主要讨论

（1）结构变化分析；

（2）交互效应分析；

（3）分段回归分析

（1）结构变化分析

结构变化的实质是检验所设定的模型在样本期内是否为同一模型。

平行回归模型的假定是斜率保持不变；

共点回归模型的假定是截距保持不变；

不同回归模型的假定是截距、斜率均变动。

例：比较改革开放前、后我国居民（平均）“储蓄—收入”总量关系是否发生了变化？

模型的设定形式为 ：

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t \quad (1)$$

其中： Y_t为储蓄总额， X_t为收入总额。

$$D = \begin{cases} 1 & \text{改革开放后} \\ 0 & \text{改革开放前} \end{cases}$$

回归方程：

改革开放后 $\mathrm{E}(Y_t \mid X_t, D=1)=(\alpha_1+\alpha_2)+(\beta_1+\beta_2)\ X_t \quad (2)$

改革开放前 $\mathrm{E}(Y_t \mid X_t, D=0)=\alpha_1+\beta_1 X_t \quad (3)$

显然，只要 α_2 β_2 不同时为零，上述模型就能刻画改革开放前后我国居民储蓄收入模型结构是否发生变化。

问题：

1.本例中，平行、共点回归、不同的回归三模型的经济学背景解释是什么？

2.如何进行结构变化判断？

3.是否可对(2)、(3)分别进行 OLS 估计？为什么？

4.若分别对(2)、(3)进行 OLS 估计应注意什么？

（2）交互效应分析

交互作用：

一个解释变量的边际效应有时可能要依赖于另一个解释变量。为此，**Klein**和**Morgen**(1951)提出了有关收入和财产在决定消费模式上相互作用的假设。他们认为消费的边际倾向不仅依赖于收入，而且也依赖于财产的多少——较富有的人可能会有不同的消费倾向。

为了捕获该影响，设 $C = \alpha + \beta Y + u$ 。假设边际消费倾向 β 依赖于财产 Z。一个简单的表示方法就是 $\beta = \beta_1 + \beta_2 Z$ 。代入消费函数，有：

$$C = \alpha + \beta_1 Y + \beta_2 YZ + u$$

由于YZ捕获了收入和财产之间的相互作用而被称为交互作用项。

显然，刻画交互作用的方法，在变量为数量(定量)变量时，是以乘法方式引入虚拟变量的。

例：是否发展油菜籽生产与是否发展养蜂生产的差异对农副产品总收益的影响研究。

模型设定为：

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

其中：Y_i（农副产品收益）；X_i（农副产品投入）

$$D_2 = \begin{cases} 1 & \text{发展油菜籽生产} \\ 0 & \text{其他} \end{cases}; \quad D_3 = \begin{cases} 1 & \text{发展养蜂生产} \\ 0 & \text{其他} \end{cases}$$

（1）式中，以加法形式引入虚拟变量暗含何假设?

（1）式以加法形式引入，暗含的假设为：

菜籽生产和养蜂生产是分别独立地影响农副品生产总收益。是否存在着一定的交互作用，且这种交互影响对被解释变量会有影响？

为了反映交互效应，将（1）变为：

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{2i} D_{3i} + \beta X_i + u_i$$

同时发展油菜籽和养蜂生产：

$$Y_i = (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) + \beta X_i + u_i$$

发展油菜籽生产：

$$Y_i = (\alpha_1 + \alpha_2) + \beta X_i + u_i$$

发展养蜂生产：

$$Y_i = (\alpha_1 + \alpha_3) + \beta X_i + u_i$$

基础类型：

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

如何检验交互效应是否存在？

看$(D_{2i}D_{3i})$系数α_4对应的t值：

即检验：$\begin{cases} \mathrm{H}_0: \ \alpha_4 = 0 \\ \mathrm{H}_1: \ \alpha_4 \neq 0 \end{cases}$

若拒绝原假设，即交互效应对Y产生了影响（应该引入模型）。

（3）分段回归分析

作用：提高模型的描述精度。

虚拟变量也可以用来代表数量因素的不同阶段。

分段线性回归就是类似情形中常见的一种。

一个例子： 研究不同时段我国居民的消费行为。

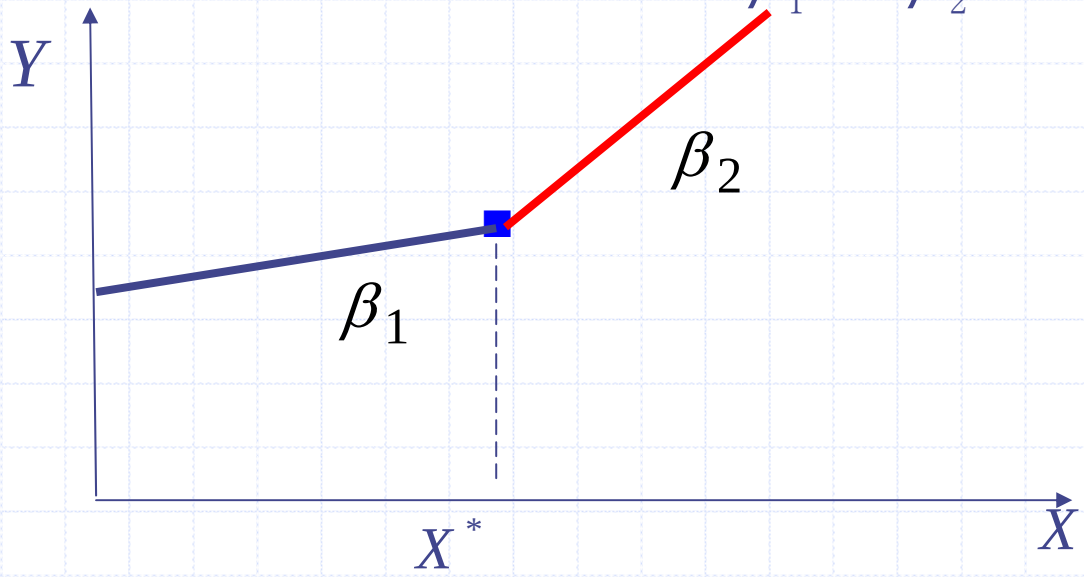
实际数据表明，1979年以前，我国居民的消费支出呈缓慢上升的趋势；从1979年开始，居民消费支出为快速上升趋势。

如何刻画我国居民在不同时段的消费行为？

分析

1979年之前，回归模型的斜率为 β_1；

1979年之前，回归模型的斜率为 $\beta_1 + \beta_2$；



若统计检验表明，β_2显著不为零，则我国居民的消费行为在**1979**年前后发生了明显改变。

采用如下描述我国居民在不同时段消费行为模型：

$$Y_t = \beta_0 + \beta_1 t + \beta_2 (t - X^*) D + u_t$$

其中：$D = \begin{cases} 1 & t \geq X_t^* \\ 0 & t < X_t^* \end{cases}$ (t=1955，1956，...，2004)

居民消费趋势方程：

1979年以前：$Y_t = \beta_0 + \beta_1 t + u_t$

1979年以后：$Y_t = \beta_0 - \beta_2 X^* + (\beta_1 + \beta_2) t + u_t$

第三节　案例分析

为了考察改革开放以来中国居民的储蓄存款与收入的关系是否已发生变化，以城乡居民人民币储蓄存款年底余额代表居民储蓄（Y），以国民总收入**GNI**代表城乡居民收入，分析居民收入对储蓄存款影响的数量关系，并建立相应的计量经济学模型。

表8.1　　国民总收入与居民储蓄存款　　单位：亿元

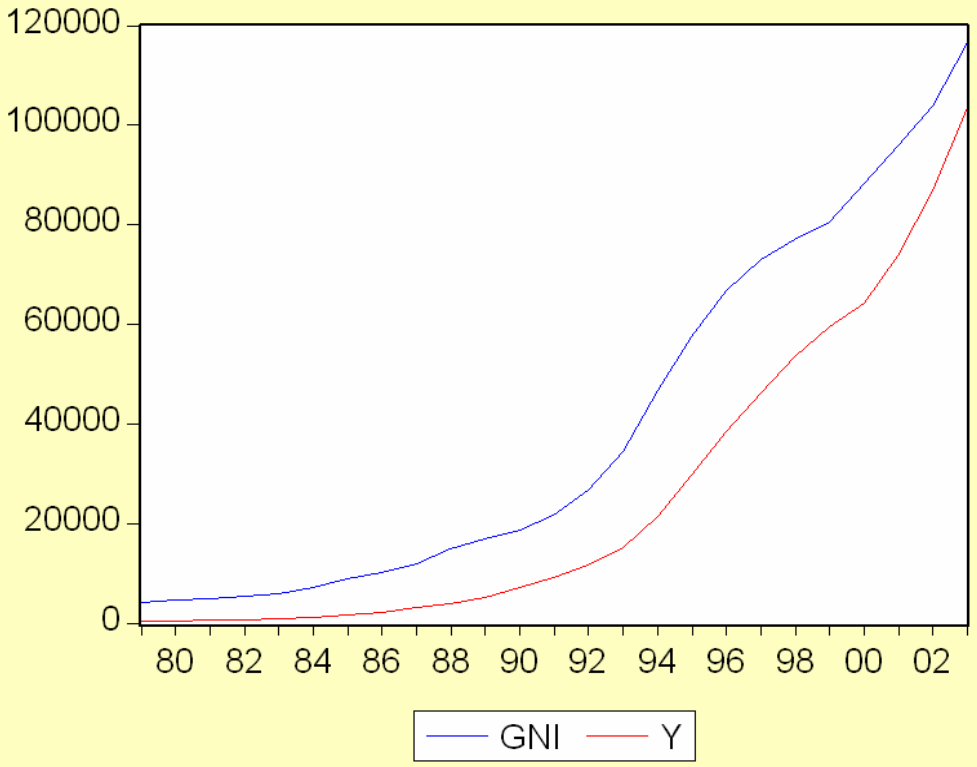
年份	国民总收入（GNI）	城乡居民人民币储蓄存款年底余额（Y）	城乡居民人民币储蓄存款增加额（YY）	年份	国民总收入（GNI）	城乡居民人民币储蓄存款年底余额（Y）	城乡居民人民币储蓄存款增额（YY）
1978	3624.1	210.6	NA	1991	21662.5	9241.6	2121.8
1979	4038.2	281	70.4	1992	26651.9	11759.4	2517.8
1980	4517.8	399.5	118.5	1993	34560.5	15203.5	3444.1
1981	4860.3	532.7	124.2	1994	46670	21518.8	6315.3
1982	5301.8	675.4	151.7	1995	57494.9	29662.3	8143.5
1983	5957.4	892.5	217.1	1996	66850.5	38520.8	8858.5

数据来源：《中国统计年鉴2004》，中国统计出版社。表中“城乡居民人民币储蓄存款年增加额”为年鉴数值，与用年底余额计算的数值有差异。

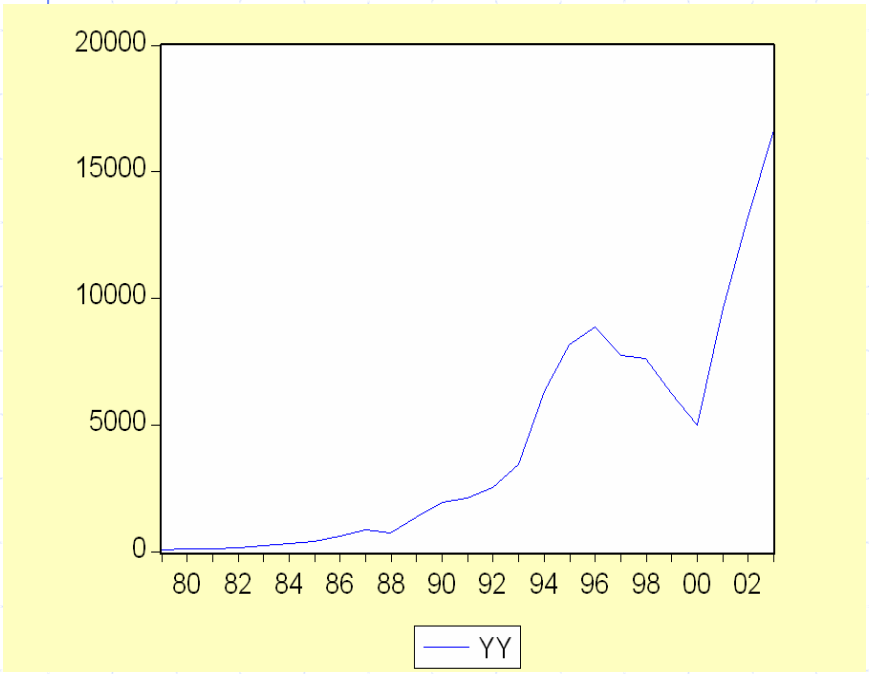
表8.1　国民总收入与居民储蓄存款（续）　　单位：亿元

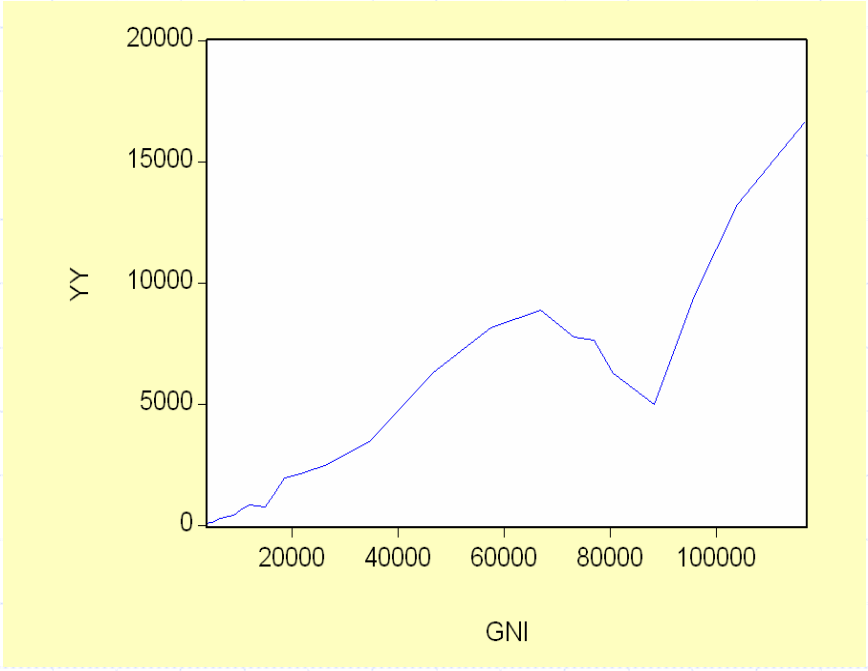
年份	国民总收入（GNI）	城乡居民人民币储蓄存款年底余额（Y）	城乡居民人民币储蓄存款增加额（YY）	年份	国民总收入（GNI）	城乡居民人民币储蓄存款年底余额（Y）	城乡居民人民币储蓄存款增加额（YY）
1984	**7206.7**	**1214.7**	**322.2**	1997	**73142.7**	**46279.8**	**7759**
1985	**8989.1**	**1622.6**	**407.9**	1998	**76967.2**	**53407.5**	**7615.4**
1986	**10201.4**	**2237.6**	**615**	1999	**80579.4**	**59621.8**	**6253**
1987	**11954.5**	**3073.3**	**835.7**	2000	**88254**	**64332.4**	**4976.7**
1988	**14922.3**	**3801.5**	**728.2**	2001	**95727.9**	**73762.4**	**9457.6**
1989	**16917.8**	**5146.9**	**1374.2**	2002	**103935.3**	**86910.6**	**13233.2**
1990	**18598.4**	**7119.8**	**1923.4**	2003	**116603.2**	**103617.7**	**16631.9**

为了研究**1978—2003**年期间城乡居民储蓄存款随收入的变化规律是否有变化，考证城乡居民储蓄存款、国民总收入随时间的变化情况，如下图所示：



从上图中，尚无法得到居民的储蓄行为发生明显改变的详尽信息。若取居民储蓄的增量（YY），并作时序图（见左下图）：





从居民储蓄增量图(上页左图)可以看出，城乡居民的储蓄行为表现出了明显的阶段特征：在1996年和2000年有两个明显的转折点。再从城乡居民储蓄存款增量与国民总收入之间关系的散布图看（见上页右图），也呈现出了相同的阶段性特征。

为了分析居民储蓄行为在1996年前后和2000年前后三个阶段的数量关系，引入虚拟变量 D_1 和 D_2。

D_1 和 D_2 的选择，是以1996、2000年两个转折点作为依据，并设定了如下以加法和乘法两种方式同时引入虚拟变量模型：

$$YY_t = \beta_1 + \beta_2 GNI_t + \beta_3 (GNI_t - 66850.50) D_{1t} + \beta_4 (GNI_t - 88254.00) D_{2t} + u_t$$

其中：

$$D_{1t} = \begin{cases} 1 & t = 1996\text{年以后} \\ 0 & t = 1996\text{年及以前} \end{cases} \qquad D_{2t} = \begin{cases} 1 & t = 2000\text{年以后} \\ 0 & t = 2000\text{年及以前} \end{cases}$$

对上式进行回归后，有：

EViews - [Equation: EQ01 Workfile: CHAPTER 8-CASE

File Edit Object View Proc Quick Options Window Help

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: YY
Method: Least Squares
Date: 08/02/05 Time: 21:54
Sample (adjusted): 1979 2003
Included observations: 25 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-830.4045	172.1626	-4.823374	0.0001
GNI	0.144486	0.005740	25.17001	0.0000
(GNI-66850.50)*DUM1	-0.291371	0.027182	-10.71920	0.0000
(GNI-88254.00)*DUM2	0.560219	0.040136	13.95810	0.0000

R-squared	0.989498	Mean dependent var	4168.652
Adjusted R-squared	0.987998	S.D. dependent var	4581.447
S.E. of regression	501.9182	Akaike info criterion	15.42040
Sum squared resid	5290359.	Schwarz criterion	15.61542
Log likelihood	-188.7550	F-statistic	659.5450
Durbin-Watson stat	1.677712	Prob(F-statistic)	0.000000

即有：

$$YY_t = -830.4045 + 0.1445GNI_t - 0.2914(GNI_t - 66850.50)D_{1t}$$

$$se = \quad 172.1626 \qquad 0.0057 \qquad 0.0272$$

$$t = \quad -4.8234 \qquad 25.1700 \qquad -10.7192$$

$$+ 0.5602(GNI_t - 88254.00)D_{2t}$$

$$se = \quad 0.0401$$

$$t = \quad 13.9581$$

由于各个系数的t检验均大于2，表明各解释变量的系数显著地不等于0，居民人民币储蓄存款年增加额的回归模型分别为：

$$YY_t = \begin{cases} YY_t = -830.4045 + 0.1445GNI_t + \varepsilon_{1t} & t \le 1996 \\ YY_t = 18649.8312 - 0.1469GNI_t + \varepsilon_{2t} & 1996 < t \le 2000 \\ YY_t = -30790.0596 + 0.4133GNI_t + \varepsilon_{3t} & t > 2000 \end{cases}$$

这表明三个时期居民储蓄增加额的回归方程在统计意义上确实是不相同的。**1996**年以前收入每增加**1**亿元，居民储蓄存款的平均增加**0.1445**亿元；在**2000**年以后，则为**0.4133**亿元，已发生了很大变化。

本章内容结束！