§2.10 模型中的特殊解释变量

10.1 随机解释变量（一般性了解）

10.2 工具变量

10.3 滞后变量（一般性了解）

10.4 虚拟变量（重点）（其中分段线性回归不讲）

10.5 时间变量

10.1随机解释变量

假定条件⑵规定解释变量是非随机的且与随机误差项相互独立，即 $\mathbf{E}(X'u)=0$.

（1）如果模型中的解释变量是随机的，但具有平稳性且与误差项相互独立，模型其他假定条件都成立，β的 OLS 估计量 $\hat{\beta}$ 仍具有无偏性，$E(\hat{\beta})=\beta$。

（2）如果模型中的解释变量 X 是随机的，与误差项 u 不独立，也不相关，模型其他假定条件都成立，β的 OLS 估计量具有一致性。 $\operatorname{plim}_{T\to\infty}\hat{\beta}=\beta$

（3）如果模型中的解释变量 X 是随机的，且与误差项 u 相关，$\mathrm{Cov}(X'u)\neq 0$，模型其他假定条件都成立，β的 OLS 估计量不具有无偏性，也不具有一致性。

10.2 工具变量法

工具变量法是解决随机解释变量X与误差项u相关时，β的OLS估计量不具有一致性的方法。

假定有变量Z与X高度相关，但与误差项u不相关，则用Z替换X，估计回归参数β，这种估计方法称作工具变量法，Z称作工具变量。β的工具变量法估计量具有一致性。

$$\operatorname*{p\lim}_{T\to\infty}\hat{\beta}_{IV}=\beta$$

10.2 工具变量法

例8.1 用最终消费**C1**对国内生产总值**Y**回归。假定**Y**与误差项***u***相关，但资本总额**K**与误差项***u***不相关，用**K**作**Y**的工具变量。

工具变量法的**EViews**操作：打开模型估计对话窗，选**TSLS**估计法。在方程设定区填入 **C1 C Y**

在工具变量列写区填入 **C K**，点击确定键。

Dependent Variable: C1
Method: Two-Stage Least Squares
Date: 02/12/07 Time: 15:27
Sample: 1978 1998
Included observations: 21
Instrument list: C K

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	628.2564	94.56630	6.643555	0.0000
Y	0.572700	0.002692	212.7480	0.0000

R-squared	0.999582	Mean dependent var	14984.05
Adjusted R-squared	0.999560	S.D. dependent var	14470.05
S.E. of regression	303.6112	Sum squared resid	1751416.
F-statistic	45261.70	Durbin-Watson stat	0.804629
Prob(F-statistic)	0.000000		

10.3 滞后变量（一般性了解）

滞后的原因。如：消费行为的滞后，央行上调银行存款准备金率，投资、项目研发周期长，一项政策的执行有滞后。

（1）分布滞后模型（权数法、阿尔蒙多项式法不讲）

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + u_t$$

可以用**OLS法估计参数**，但不具有有效性。容易引起多重共线性。最大滞后阶数由**AIC、SC**准则决定。

（2）自回归模型（柯依克变换不讲）

$$Y_t = \alpha + \beta_0 X_t + \gamma_1 Y_{t-1} + \ldots + \gamma_m Y_{t-m} + u_t$$

可以用**OLS法估计参数，为有偏、一致估计量**。最大滞后阶数由**AIC、SC**准则决定。

（3）自回归分布滞后模型

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + \gamma_1 Y_{t-1} + \ldots + \gamma_m Y_{t-m} + u_t$$

如消费模型：$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \gamma_1 Y_{t-1} + u_t$

10.4 虚拟变量（重点掌握）

在实际建模过程中，被解释变量不但受定量变量影响，同时还受定性变量影响。例如需要考虑性别、民族、不同历史时期、季节差异、企业所有制性质不同等因素的影响。

由于定性变量通常表示的是某种特征的有和无，所以量化方法可采用取值为**1**或**0**。这种变量称作虚拟变量（**dummy variable**），用***D***表示。虚拟变量应用于模型中，对其回归系数的估计与检验方法和定量变量相同。

10.4 虚拟变量

注意：**(1)** 当定性变量含有***m***个类别时，模型不能引入***m***个虚拟变量。最多只能引入***m* -1**个虚拟变量，否则当模型中存在截距项时就会产生完全多重共线性，无法估计回归参数。比如，对于季节数据引入**4**个虚拟变量，数据如下表，

t	x_t	D_1	D_2	D_3	D_4
1995.1	x_1	**1**	**0**	**0**	**0**
1995.2	x_2	**0**	**1**	**0**	**0**
1995.3	x_3	**0**	**0**	**1**	**0**
1995.4	x_4	**0**	**0**	**0**	**1**
1996.1	x_5	**1**	**0**	**0**	**0**
1996.2	x_6	**0**	**1**	**0**	**0**
1996.3	x_7	**0**	**0**	**1**	**0**
1996.4	x_8	**0**	**0**	**0**	**1**
1997.1	x_9	**1**	**0**	**0**	**0**

则必然会有，截距项对应的单位向量等于 $(D_1+ D_2+ D_3+ D_4)$ 。这意味着虚拟变量之间存在完全多重共线性。

10.4 虚拟变量

(2) 把虚拟变量取值为**0**所对应的类别称作基础类别。

(3) 当定性变量含有***m***个类别时，不能把虚拟变量的值设成如下形式。

$$D=\begin{cases} 0 & , \quad \text{第1个类别} \\ 1 & , \quad \text{第2个类别} \\ \cdots & , \quad \cdots \\ m-1 & , \quad \text{第}m\text{个类别} \end{cases}$$

这种赋值法在一般情形下与虚拟变量赋值是完全不同的两回事。

(4) 回归模型可以只用虚拟变量作解释变量，也可以用定量变量和虚拟变量一起做解释变量。

10.4 虚拟变量

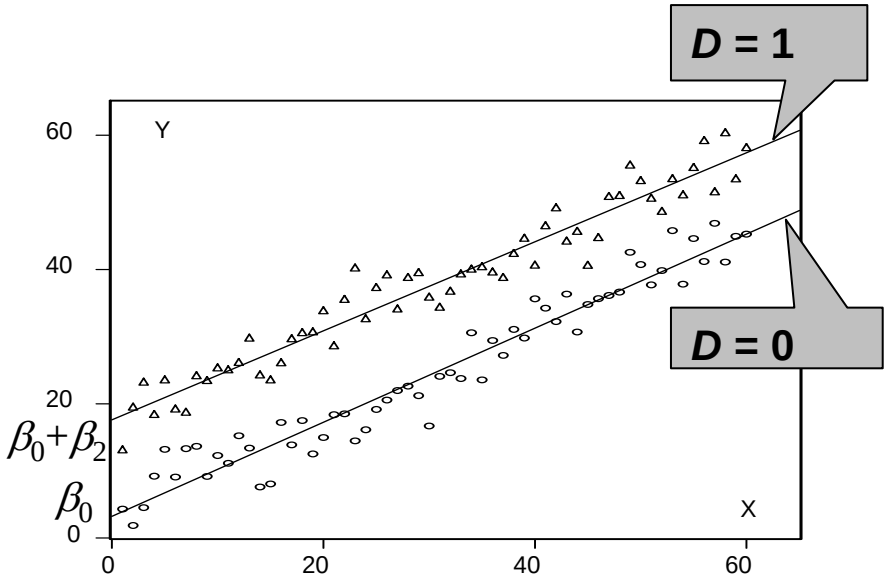
1. 用虚拟变量测量截距变动

设有模型，$y_t = \beta_0 + \beta_1 x_t + \beta_2 D + u_t$，

其中y_t，x_t为定量变量；D为定性变量。

当$D = 0$ 或1时，上述模型可表达为，

$$y_t = \begin{cases} \beta_0 + \beta_1 x_t + u_t & D = 0 \\ (\beta_0 + \beta_2) + \beta_1 x_t + u_t & D = 1 \end{cases}$$



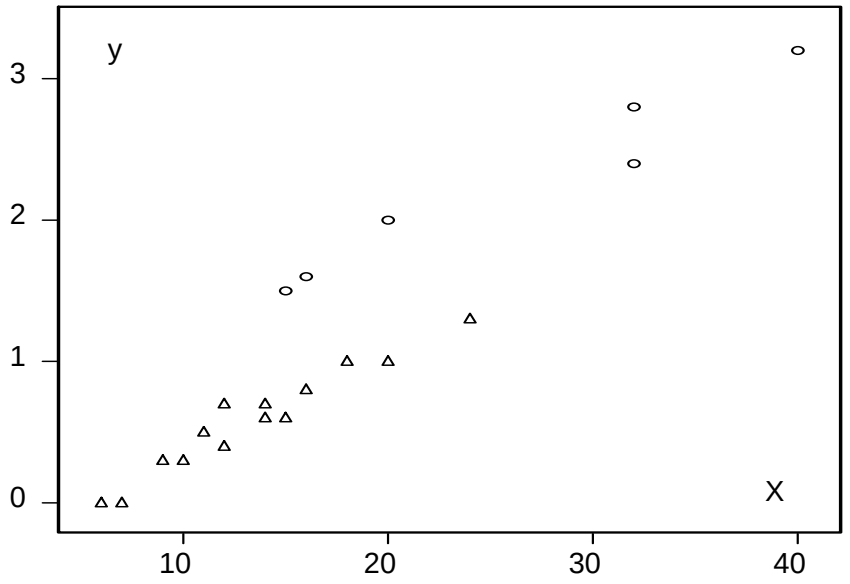
$D = 1$或0表示某种特征的有无。反映在数学上是截距不同的两个函数。

若β_2显著不为零，说明截距不同；若β_2为零，说明这种分类无显著性差异。

10.4 虚拟变量

例8.3 随机调查美国旧金山地区20个家庭的储蓄情况，拟建立年储蓄额Y_i(千美元) 对年收入X_i (千美元) 的回归模型。通过对样本点的分析发现，居于上部的6个点（用小圆圈表示）都是代表自己有房子的家庭；居于下部的14个点（用小三角表示）都是租房住的家庭。而这两类家庭所对应的观测点各自都表现出明显的线性关系。于是给模型加入一个定性变量“住房状况”，用D表示。定义如下：

$$D=\begin{cases}1, & \text{（有房户）}\\ 0, & \text{（租房户）}\end{cases}$$



10.4 虚拟变量

例8.3 建立回归模型

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

得估计结果如下，

$$\hat{Y}_i = -0.3204 + 0.0675 X_t + 0.8273 D_i$$

$$(-5.2) \qquad (16.9) \qquad (11.0) \quad R^2 = 0.99, \ DW = 2.27$$

由于回归系数**0.8273**显著地不为零，说明对住房状况不同的两类家庭来说，回归函数截距项确实明显不同。

当模型不引入虚拟变量“住房状况”时，得回归方程如下，

$$\hat{Y}_i = -0.5667 + 0.0963 X_i$$

$$(-3.5) \qquad (11.6) \quad R^2 = 0.88, \ DW = 1.85$$

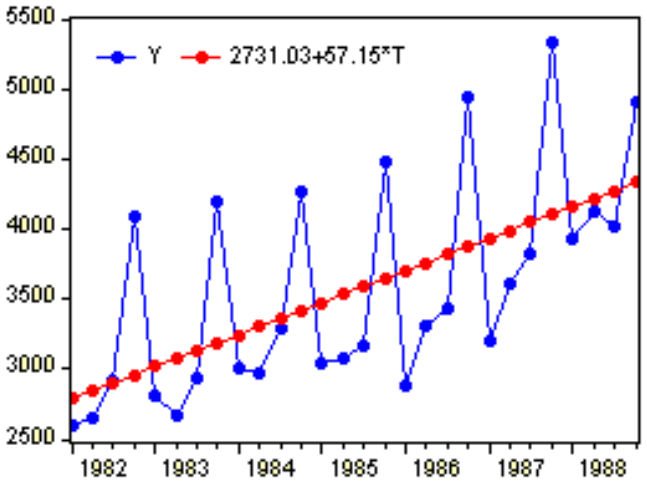
比较回归方程，前者的确定系数为**0.99**，后者的确定系数仅为**0.88**。说明该回归模型中引入虚拟变量非常必要。

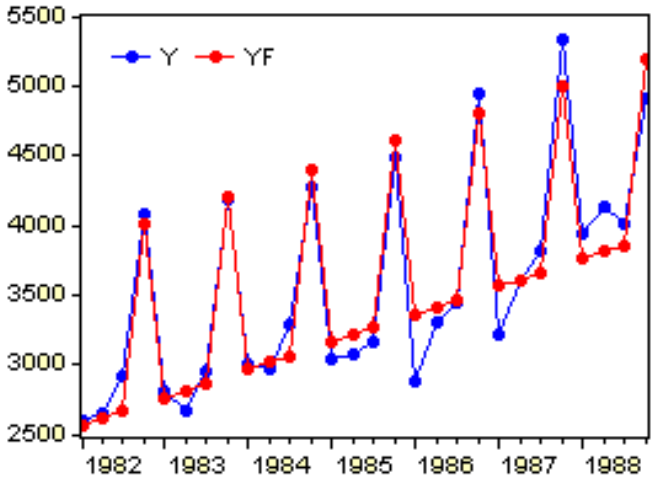
10.4 虚拟变量

“季节”是在研究经济问题中常常遇到的定性因素。比如，酒，肉的销量在冬季要超过其它季节，而饮料的销量又以夏季为最大。当建立这类问题的计量模型时，就要考虑把“季节”因素引入模型。由于一年有四个季节，所以这是一个含有四个类别的定性变量。应该向模型引入三个虚拟变量。

例8.4 市场用煤销售量模型。由于受取暖用煤的影响，每年第四季度的销售量大大高于其它季度。鉴于是季节数据可设三个季节变量如下：

$$D_1=\begin{cases}1, & \text{（第四季度）}\\ 0, & \text{（其他季度）}\end{cases}\qquad D_2=\begin{cases}1, & \text{（第三季度）}\\ 0, & \text{（其他季度）}\end{cases}\qquad D_3=\begin{cases}1, & \text{（第二季度）}\\ 0, & \text{（其他季度）}\end{cases}$$





10.4 虚拟变量例10.4

以时间 t 为解释变量（1982年1季度取$t = 1$）的煤销售量（Y_i）模型估计结果如下：

$$\hat{Y}_i = 2431.20 + 49.00\,t + 1388.09\,D_1 + 201.84\,D_2 + 85.00\,D_3$$

(26.04)　(10.81)　(13.43)　(1.96)　(0.83)

$R^2 = 0.95$, DW = 1.2, F=100.4, T=28, $t_{0.05}$ (28-5) = 2.07

由于D_2，D_3的系数没有显著性，说明第二、三季度可以归并入基础类别第一季度。于是只考虑加入一个虚拟变量D_1，把季节因素分为第四季度和第一、二、三季度两类。从上式中剔除虚拟变量D_2，D_3，得煤销售量（Y_i）模型如下：

$$\hat{Y}_i = 2515.86 + 49.73\,t + 1290.91\,D_1$$

(32.03　(10.63)　(14.79)

$R^2 = 0.94$, DW = 1.4, F = 184.9, T=28, $t_{0.05}$ (25) = 2.06

这里第一、二、三季度为基础类别。

10.4 虚拟变量

2. 测量斜率变动

以上介绍了用虚拟变量测量回归函数的截距变化。实际上，也可以用虚拟变量考察回归函数的斜率是否发生变化。方法是在模型中加入定量变量与虚拟变量的乘积项。设模型如下，

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i D_i) + u_i$$

按β_2，β_3是否为零，回归函数可有如下四种形式。

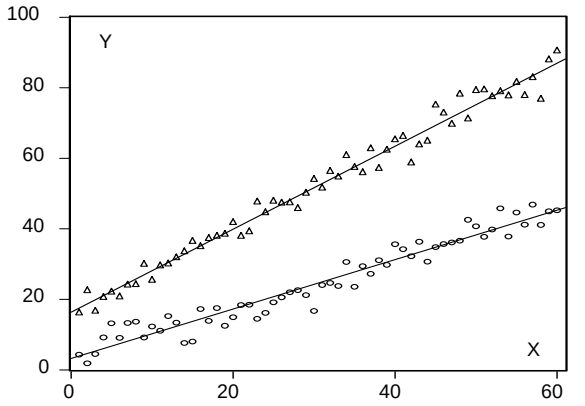
$$E(Y_i) = \beta_0 + \beta_1 X_i, \qquad (\text{当 } \beta_2 = \beta_3 = 0)$$

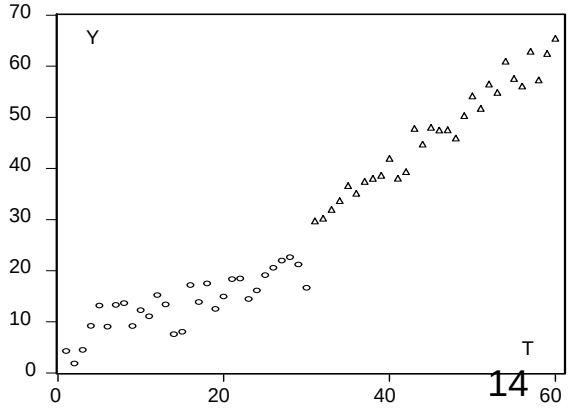
$$E(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i, \quad (\text{当 } \beta_2 \neq 0, \beta_3 \neq 0)$$

$$E(Y_i) = \beta_0 + (\beta_1 + \beta_3) X_i, \qquad (\text{当 } \beta_2 = 0, \beta_3 \neq 0)$$

$$E(Y_i) = (\beta_0 + \beta_2) + \beta_1 X_i, \qquad (\text{当 } \beta_2 \neq 0, \beta_3 = 0)$$

截距、斜率同时发生变化的两种情形见图。





3. 分段线性回归（不讲）

10.4 虚拟变量

例10.5 中国进出口贸易总额序列（**1950~1984**年）如图。试检验改革开放前后该时间序列的斜率是否发生变化。定义虚拟变量***D***如下，

$$D=\begin{cases}1, & (1950-1978)\\ 0, & (1979-1984)\end{cases}$$

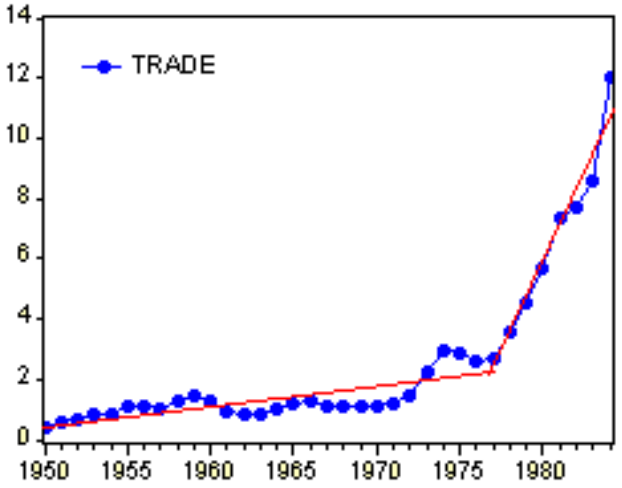
以时间***time***为解释变量，进出口贸易总额用***trade***表示，估计结果如下，

$$\hat{trade} = 0.2818 + 0.0746\ time - 35.8809D + 1.2559\ time\ D$$

(1.35)　(6.2)　　(-8.4)　　(9.6)

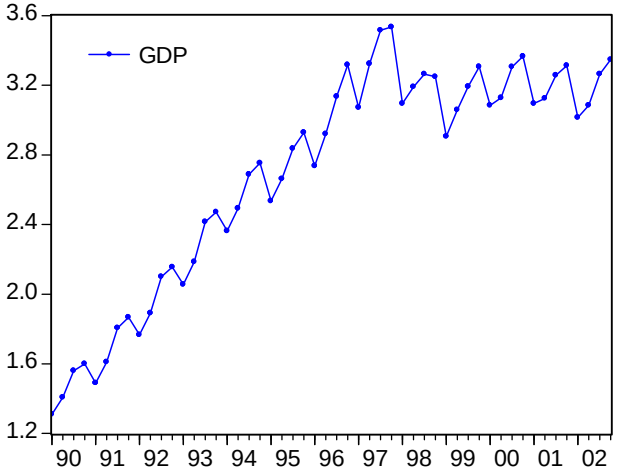
$$\hat{trade}=\begin{cases}0.2818+0.0746time, & (D=0,1950-1978)\\ -33.5991+1.3305time, & (D=1,1979-1984)\end{cases}$$

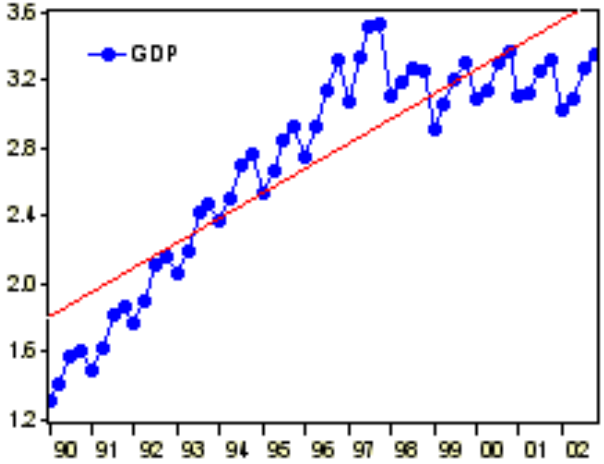
上式说明，改革开放前后相比无论截距和斜率都发生了变化。进出口贸易总额的年平均增长量扩大了近**17**倍。



10.4 虚拟变量

补充案例 ：香港季节GDP数据（千亿港元）的拟合（file:dummy6）





1990~1997年香港季度**GDP**呈线性增长。**1997**年由于遭受东南亚金融危机的影响，经济发展处于停滞状态，**1998~2002**年底**GDP**总量几乎没有增长（见上图）。对这样一种先增长后停滞，且含有季节性周期变化的过程简单地用一条直线去拟合显然是不恰当的。为区别不同季节，和不同时期，定义季节虚拟变量**D2**、**D3**、**D4**和区别不同时期的虚拟变量**DT**如下，

$$D2=\begin{cases}1, & \text{（第2季度）}\\ 0, & \text{（其他季度）}\end{cases}\quad D3=\begin{cases}1, & \text{（第3季度）}\\ 0, & \text{（其他季度）}\end{cases}\quad D4=\begin{cases}1, & \text{（第4季度）}\\ 0, & \text{（其他季度）}\end{cases}\quad DT=\begin{cases}0, & 1990m1\sim 1997m4\\ 1, & 1998\text{m1}\sim 2002\text{m4}\end{cases}$$

10.4 虚拟变量

例3：香港季节GDP数据（千亿港元）的拟合（file:dummy6）

得估计结果如下：

$$\hat{GDP}_t = 1.1573+0.0668t+0.0775D_2+0.2098D_3+0.2349D_4+1.8338DT- 0.0654DT\times t$$

(50.8) (64.6) (3.7) (9.9) (11.0) (19.9) (-28.0)

$R^2 = 0.99$, DW = 0.9, s.e. = 0.05, F=1198.4, T=52, $t_{0.05}$ (52-7) = 2.01

对于1990:1 ~1997:4

$$\hat{GDP}_t = 1.1573 + 0.0668\ t + 0.0775\ D_2 + 0.2098\ D_3 + 0.2349\ D_4$$

对于1998:1~2002:4

$$\hat{GDP}_t = 2.9911 + 0.0014\ t + 0.0775\ D_2 + 0.2098\ D_3 + 0.2349\ D_4$$

10.4 虚拟变量

补充案例：香港季节GDP数据的拟合

Dependent Variable: GDP
Method: Least Squares
Date: 04/04/04 Time: 15:54
Sample: 1990:1 2002:4
Included observations: 52

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.157300	0.022778	50.80814	0.0000
T	0.066843	0.001035	64.56095	0.0000
D2	0.077522	0.021139	3.667325	0.0006
D3	0.209829	0.021215	9.890799	0.0000
D4	0.234922	0.021341	11.00827	0.0000
DT	1.833786	0.092079	19.91526	0.0000
DT*T	-0.065419	0.002333	-28.03951	0.0000

R-squared	0.993780	Mean dependent var	2.695174
Adjusted R-squared	0.992951	S.D. dependent var	0.641144
S.E. of regression	0.053829	Akaike info criterion	-2.881376
Sum squared resid	0.130388	Schwarz criterion	-2.618708
Log likelihood	81.91576	F-statistic	1198.382
Durbin-Watson stat	0.910754	Prob(F-statistic)	0.000000

如果不采用虚拟变量拟合效果将很差。

$\hat{GDP}_t = 1.6952 + 0.0377\,t$

(20.6) (13.9) $R^2 = 0.80$, **DW = 0.3,** T**=52,** $t_{0.05}$ **(52-2) = 2.01**

10.5 时间变量

以时间变量t作解释变量。估计与检验方法与定量解释变量X_t相同。

$$Y_t = \alpha + \gamma t + \beta_0 X_t + u_t$$

$$Y_t = \alpha + \gamma t + u_t$$

如时间变量t在生产函数模型中代表技术进步。

$$Lny_t = \beta_0 + \gamma t + \beta_1 Lnx_{t1} + \beta_2 Lnx_{t2} + u_t$$