第三章

经典单方程计量经济学模型

多元线性回归模型

第三章 多元线性回归模型

3.1 多元线性回归模型概述

3.2 多元线性回归模型的参数估计

3.3 多元线性回归模型的统计检验

3.4 非线性回归模型的线性化

3.5 含有虚拟变量的多元线性回归模型

3.1 多元线性回归模型概述

一、多元线性回归模型的形式

二、多元线性回归模型的基本假设

一、多元线性回归模型的形式

定义：模型中含有多个解释变量的回归模型。

一般形式

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu$$

给出观察值的情况下

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$$

其中： k——解释变量数目；

β_j（$j = 1,2,\ldots,k$）——回归系数

总体回归函数的确定性形式：

$$E(Y|X_{i1}, X_{i2}, \cdots, X_{ik}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

总体回归函数表示：在给定解释变量X下，Y的平均响应。即各解释变量X值固定时Y的平均响应。

- β_j 又称为偏回归系数，表示在其他解释变量保持不变的情况下，X_j每变化1个单位时，Y的均值$E(Y|X)$的变化。
- 或者说，β_j给出了X_j的单位变化对Y均值的“直接”影响或“净”影响（不含其他变量的影响）。

总体回归函数的确定性形式：

$$E(Y|X_{i1}, X_{i2}, \cdots, X_{ik}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

总体回归函数的随机形式：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$$

样本回归函数的确定性形式：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}$$

样本回归函数的随机形式：

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} + e_i$$

$e_i = Y_i - \hat{Y}_i$，e_i为残差项，可以看作总体回归函数中μ_i的近似替代。

多元线性回归模型的矩阵表达式

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i \qquad i = 1,2, \ldots n$$

$$\begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_k X_{1k} + \mu_1 \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_k X_{2k} + \mu_2 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_k X_{nk} + \mu_n \end{cases}$$ ，即

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n*1} = \begin{pmatrix} 1 & X_{11} & X_{12} \cdots & X_{1k} \\ 1 & X_{21} & X_{22} \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \ldots & X_{nk} \end{pmatrix}_{n*(k+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1)*1} + \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}_{n*1}$$

即 $\boldsymbol{Y = X\beta + \mu}$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}$$

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} \cdots & X_{1k} \\ 1 & X_{21} & X_{22} \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \cdots & X_{nk} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \longrightarrow \boldsymbol{\hat{Y} = X\hat{\beta}}$$

$$\boldsymbol{e_i = Y_i - \hat{Y}_i}$$

$$\boldsymbol{e} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} \qquad \boldsymbol{e = Y - \hat{Y}} \longrightarrow \boldsymbol{Y = X\hat{\beta} + e}$$

二、多元线性回归模型的基本假设

假设1：回归模型是正确设定的。

假设2：解释变量$X_1, X_2, \ldots, X_k$在所抽取的样本中具有变异性，且$X_j(j = 1, 2, \ldots, k)$之间不存在严格的线性相关性（无完全多重共线性）。

〈多元模型特有〉

$R(X) = k + 1 \rightarrow R(X'X) = k + 1$(方阵) $\rightarrow$ 即X′X可逆

假设3：随扰动项具有条件零均值。

$$E(\mu_i|X_1\ ,\ X_2\ ,\ \cdots\ ,\ X_k)=0$$

重要推论：$E(X_i\mu_i)=0$，即X与μ同期不相关。

假设4：随机干扰项具有条件同方差和不序列相关性。

$$Cov(\mu_i,\mu_j)=E[(\mu_i-E\mu_i)(\mu_j-E\mu_j)]=E(\ \mu_i\mu_j\ )=\begin{cases}\sigma^2 & i=j\text{时}\\ 0 & i\neq j\text{时}\end{cases}$$

假设5：随机干扰项满足正态分布。

$$\mu_i|X_1\ ,\ X_2\ ,\ \cdots\ ,\ X_k\sim N(0,\sigma^2)$$

- **随机干扰项条件零均值、同方差和不序列相关的矩阵表示**

$$E(\mu)=E\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}=\begin{pmatrix}E\mu_1\\ E\mu_2\\ \vdots\\ E\mu_n\end{pmatrix}=0$$

$$Var(\mu|X)=E(\mu\mu'|X)=E\left[\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}(\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n)|X\right]$$

$$=E\begin{pmatrix}{\mu_1}^2 & \cdots & \mu_1\mu_n\\ \vdots & \ddots & \vdots\\ \mu_n\mu_1 & \cdots & {\mu_n}^2\end{pmatrix}=\begin{pmatrix}Var(\mu_1) & \cdots & Cov(\mu_1,\mu_n)\\ \vdots & \ddots & \vdots\\ Cov(\mu_n,\mu_1) & \cdots & Var(\mu_n)\end{pmatrix}$$

$$=\begin{pmatrix}\sigma^2 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma^2\end{pmatrix}=\sigma^2 I$$

μ的方差-协方差矩阵

第三章 多元线性回归模型

3.1 多元线性回归模型概述

3.2 多元线性回归模型的参数估计

3.3 多元线性回归模型的统计检验

3.4 非线性回归模型的线性化

3.5 含有虚拟变量的多元线性回归模型

3.2 多元线性回归模型的参数估计

估计任务
- 模型的结构参数——β_0，β_1，…，β_k
- 模型分布参数（随机扰动项的方差）——σ^2

一、模型的OLS估计

目标：多元线性回归模型中，要估计的是一个平面或超平面——拟合平面（通常称为拟合直线）。

原理：最优“拟合直线”——点到拟合直线的纵向距离最小。即拟合值尽可能逼近真值，使残差平方和最小。

$$min\sum e_i^2 = min\sum(Y-\hat{Y})^2$$

估计过程（利用解析式估计）：

$$Q=\sum e_i{}^2=\sum(Y_i-\hat{Y}_i)^2=\sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_kX_{ik})]^2$$

求偏导，得到正规方程组

$$\begin{cases}\sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_kX_{ik})]=0 & => \quad \sum e_i=0\\ \sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_kX_{ik})]X_{i1}=0 & => \quad \sum X_{i1}e_i=0\\ \qquad\vdots & \qquad\vdots\\ \sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_{i1}+\hat{\beta}_2X_{i2}+\cdots+\hat{\beta}_kX_{ik})]X_{ik}=0 & => \quad \sum X_{ik}e_i=0\end{cases}$$

其矩阵表达为：

$$\begin{bmatrix} \sum e_i \\ \sum X_{i1} e_i \\ \vdots \\ \sum X_{ik} e_i \end{bmatrix} = 0 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ \vdots & \vdots & & \vdots \\ X_{1k} & X_{2k} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{X'e}$$

对样本回归函数 $\boldsymbol{Y = X\hat{\beta} + e}$

将两边左乘$\boldsymbol{X'}$ $\boldsymbol{X'Y = X'X\hat{\beta} + X'e}$

得到 $\boldsymbol{X'Y = X'X\hat{\beta}}$ ， 即$\boldsymbol{\hat{\beta} = (X'X)^{-1}X'Y}$

二、随机扰动项的OLS估计

多元线性回归模型：$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1} = \frac{e'e}{n-k-1}$

一元线性回归模型：$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1-1} = \frac{\sum e_i^2}{n-2}$

三、参数估计量的统计性质

1.线性性

由于

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{XX}')^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{CY}$$

其中，$\boldsymbol{C} = (\boldsymbol{XX}')^{-1}\boldsymbol{X}'$仅与固定的$\boldsymbol{X}$有关，可见，参数估计量是被解释变量$\boldsymbol{Y}$的线性组合。

2.无偏性

在解释变量样本值给定的条件下，参数估计量$\hat{\beta}$具有无偏性，证明如下：

$$
\begin{aligned}
E(\hat{\beta}|X) &= E((X'X)^{-1}X'Y|X) \\
&= E((X'X)^{-1}X'(X\beta+\mu)|X) \\
&= \beta+(X'X)^{-1}X'E(\mu|X) \\
&= \beta
\end{aligned}
$$

3.有效性

参数β的方差-协方差矩阵可进行如下转变：

$$Var(\hat{\beta}|X) = E[\left(\hat{\beta} - E(\hat{\beta})\right)(\hat{\beta} - E(\hat{\beta}))'|X]$$

$$\hat{\beta} = (XX')^{-1}X'Y$$
$$= (XX')^{-1}X'(X\beta + \mu)$$
$$= \beta + (XX')^{-1}X'\mu$$

$$E(\mu\mu'|X) = \sigma^2 I_n$$

$$= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X]$$
$$= E[(X'X)^{-1}X'\mu\mu'X(X'X)^{-1}|X]$$
$$= (X'X)^{-1}X'E(\mu\mu'|X)X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}$$

根据高斯-马尔可夫定理，该方差在所有无偏估计量的方差中是最小的，所有该参数估计量具有有效性。

四、样本容量问题

模型依赖于实际样本，样本容量越大越好，然而获取样本需要成本，因而如何选择样本容量，使其既能满足建模需要，又能减轻收集数据的困难，就成为一个重要的实际问题。

最小样本容量：$n \geq k+1$

（欲得到估计量，必须满足$(\mathbf{X}'\mathbf{X})^{-1}$存在，$\mathbf{X}'\mathbf{X}$应为$k+1$阶满秩矩阵，即$\boldsymbol{R}(\mathbf{X}) \geq k+1$，$\mathbf{X}$最大秩为$k+1$）

满足基本要求的样本容量：$n \geq 30$，Z检验成立，

$n-k \geq 8$，t分布稳定，检验有效。

第三章 多元线性回归模型

3.1 多元线性回归模型概述

3.2 多元线性回归模型的参数估计

3.3 多元线性回归模型的统计检验

3.4 非线性回归模型的线性化

3.5 含有虚拟变量的多元线性回归模型

3.3 多元线性回归模型的统计检验

- 统计检验不涉及模型的经济内涵，旨在检验模型是否满足数学理论与方法上的要求——统计差异显著性。
- 统计检验的结果表明模型是否能拟合样本数据，或者说观察到的经济事实是否支持理论模型。

1、拟合优度检验（R^2检验）

2、方程的显著性检验（F检验）

3、变量的显著性检验（t检验）

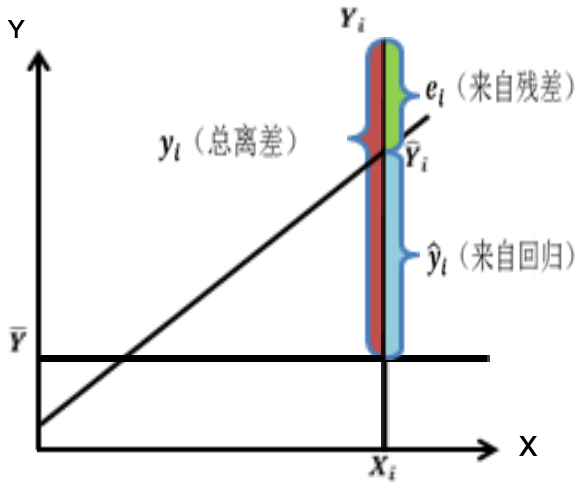
1、拟合优度检验（R^2检验）

（一）可决系数R^2

$TSS = \sum(Y_i - \bar{Y})^2$ —— 总离差平方和

$ESS = \sum(\hat{Y}_i - \bar{Y})^2$ —— 回归平方和

$RSS = \sum(Y_i - \hat{Y}_i)^2$ —— 残差平方和



$$
\begin{aligned}
TSS &= \sum\left(Y_i - \bar{Y}\right)^2 \\
&= \sum\left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right]^2 \\
&= \sum\left(Y_i - \hat{Y}_i\right)^2 + 2\sum\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) + \sum\left(\hat{Y}_i - \bar{Y}\right)^2 \\
&= \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{Y}\right)^2 \\
&= RSS + ESS
\end{aligned}
$$

可决系数 $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

$R^2 \in [0,1]$，R^2越接近1，说明实际观测值离样本回归线越近，拟合效果越好。

【这里产生了一个问题：】

- 实际应用中发现，如果在模型中增加一个解释变量，R^2往往会增大。这是因为残差平方和往往随着解释变量个数增加而减少，至少不会增加。
- 容易产生一种错觉，要模型拟合的好，只需要增加解释变量数目即可。
- 但实际上，通过增加解释变量引起的R^2的增大与拟合好坏无关。所以，R^2需调整 。

（二）调整的可决系数——$\bar{R}^2$

引入$\bar{R}^2$的目的：解释变量的增加会使残差平方和减小，增大R^2，R^2的这一增加与拟合好坏无关，所以，引入$\bar{R}^2$对多元模型进行比较。

调整$\bar{R}^2$的思路：在样本容量一定的情况下，增加解释变量必定使得自由度减少，所以调整的思路是将残差平方和与总离差平方和分别除以各自的自由度，以剔除变量个数对拟合优度的影响。

自由度：对模型复杂程度的度量。

自由度=可自由变动的变量个数—约束条件的个数

统计量	自由度
$TSS = \sum (Y_i - \bar{Y})^2$	$n-1$
$RSS = \sum (Y_i - \hat{Y}_i)^2$	$n-k-1$
$ESS = \sum (\hat{Y}_i - \bar{Y})^2$	k

将$R^2 = \frac{ESS}{TSS}$转化为$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$

或 $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$

2、方程的显著性检验（F检验）

目的：对模型中被解释变量与解释变量之间的线性关系在总体上是否显著成立做出判断。

即检验模型$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$ 中的参数$\beta_1, \ldots, \beta_k$是否显著不为0。

原假设与备择假设：

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \beta_j \text{不全为0}$$

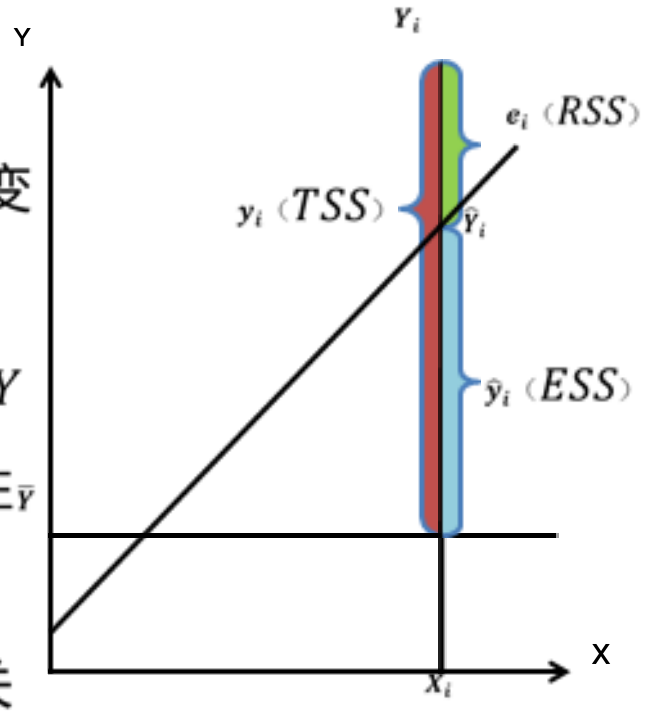
逻辑意义：如果H_0成立，说明解释变量X对被解释变量Y不起解释作用，Y由随机误差项μ决定。

【注意】若至少有一个X对Y有解释作用，则拒绝H_0，接受H_1。

F统计量： $F = \dfrac{ESS/k}{RSS/(n-k-1)}$

基本思想：

- 来自于总离差平方和的分解（$TSS = ESS + RSS$）
- 回归平方和ESS是解释变量X对被解释变量Y的线性作用的结果
- 如果ESS/RSS值较大，则X的联合体对Y的解释程度高，可以认为总体存在线性关系
- 反之，二者不存在总体上的线性相关关系。



- F检验与拟合优度检验的关系

区别

拟合优度检验：从已估计的模型出发，检验其对样本观测值的拟合程度

F检验：从样本观测值出发，检验模型总体线性关系的显著性

联系：模型对样本观测值的拟合程度高，模型总体线性关系的显著性就强。

拟合优度检验和F检验的数量关系：

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1+kF} = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}$$

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{ESS/k}{RSS/(n-k-1)}$$

- F和R^2同向变化
- $R^2 = 0$时，$F = 1$；$R^2 = 1$时，F无穷大，R^2越大，F值越大。

- **F检验的步骤**

<1> 提出假设$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \beta_j (j = 1, 2, \ldots, k) \text{不全为零} \end{cases}$

<2> 计算H_0成立时的F 统计量， $F = \frac{ESS/K}{RSS/(n-k-1)} \sim F(k, n-k-1)$。

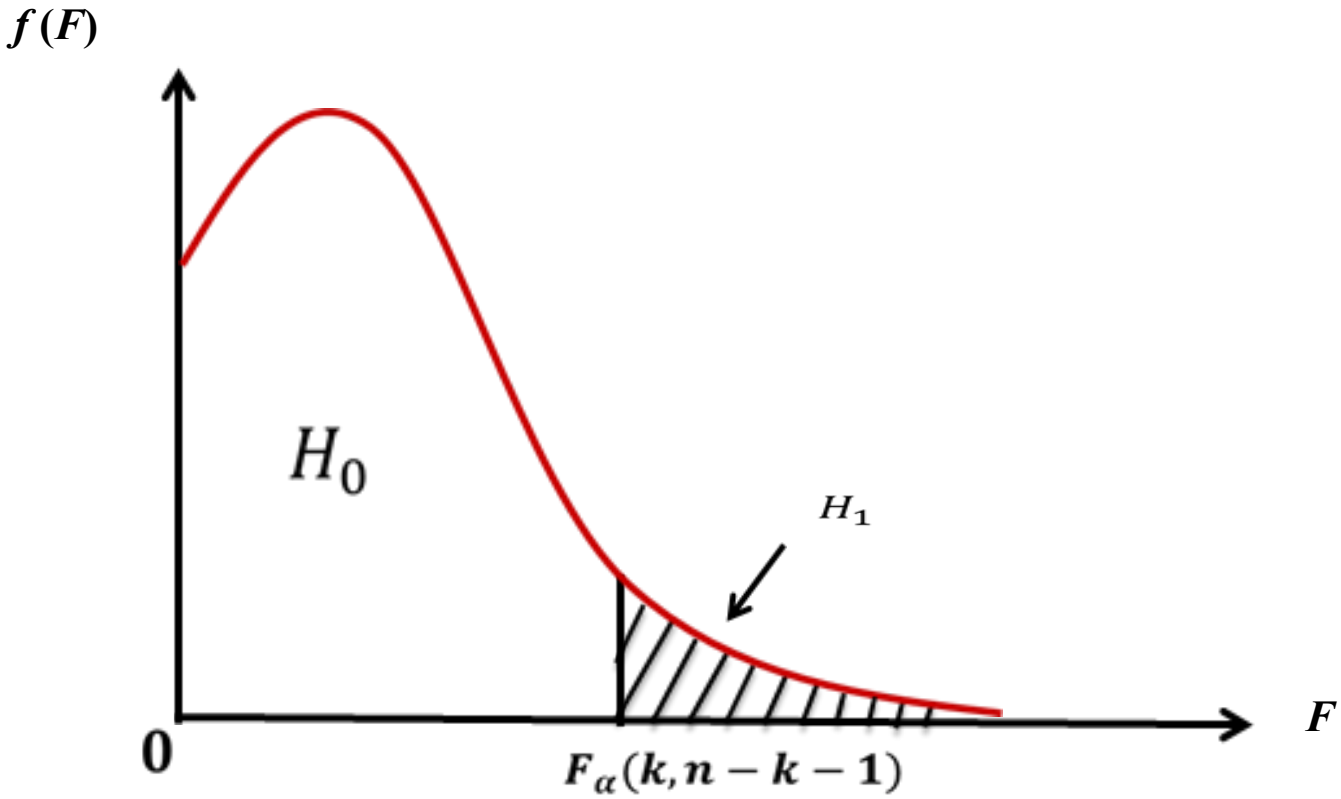
<3> 给定显著性水平α，查表得到临界值$F_\alpha(k, n-k-1)$。

<4> 比较，判断

若$F > F_\alpha(k, n-k-1)$，拒绝H_0。模型在总体上存在显著的线性相关关系。

若$F \leq F_\alpha(k, n-k-1)$，不拒绝$H_0$。模型在总体上线性相关关系不显著。

F检原假设的拒绝域



【注意】F检验只把模型作为一个整体，对总体线性关系进行检验。除此之外还应对模型中的各个变量进行显著性检验，以决定它们是否应当保留在模型之中。

3、变量的显著性检验（t检验）

目的：剔除模型中回归系数与0差异不显著的解释变量，使模型更简洁实用。

要求：对模型进行整体检验，一次只能剔除一个最不显著的解释变量。

检验步骤：

<1> 提出假设 $\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases} \quad (i = 1, 2, \dots, k)$

<2> 计算t统计量 $t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}}$

<3> 查表，得临界值 $t_{\alpha/2}(n-k-1)$

<4> 比较，判断

若 $|t| > t_{\frac{\alpha}{2}}(n-k-1)$，拒绝$H_0$。

若 $|t| \leq t_{\frac{\alpha}{2}}(n-k-1)$，不拒绝$H_0$。系数显著为0。

第三章 多元线性回归模型

3.1 多元线性回归模型概述

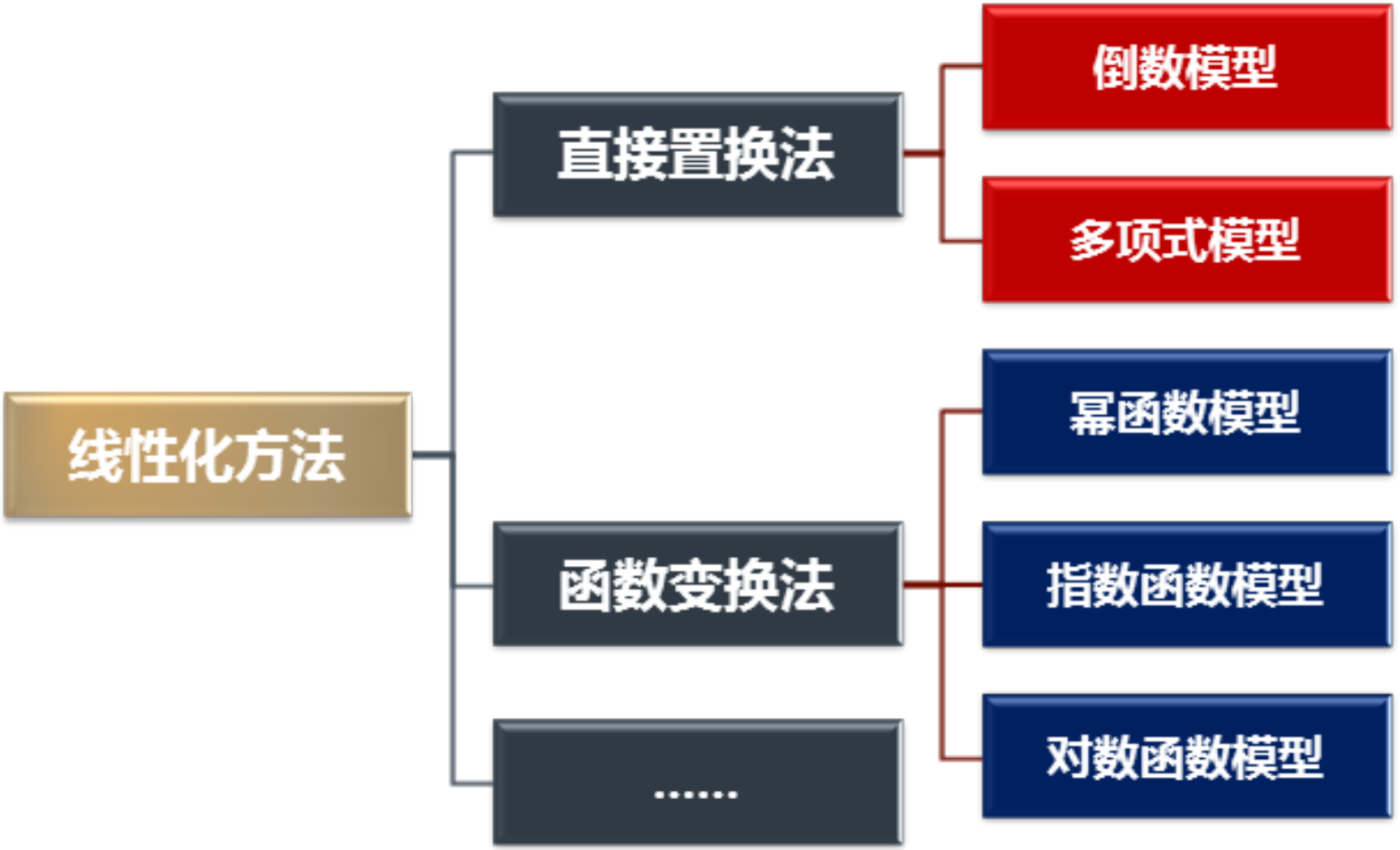
3.2 多元线性回归模型的参数估计

3.3 多元线性回归模型的统计检验

3.4 非线性回归模型的线性化

3.5 含有虚拟变量的多元线性回归模型

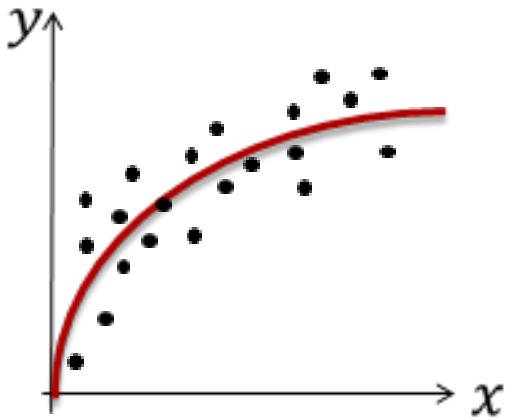
3.4 非线性回归模型的线性化



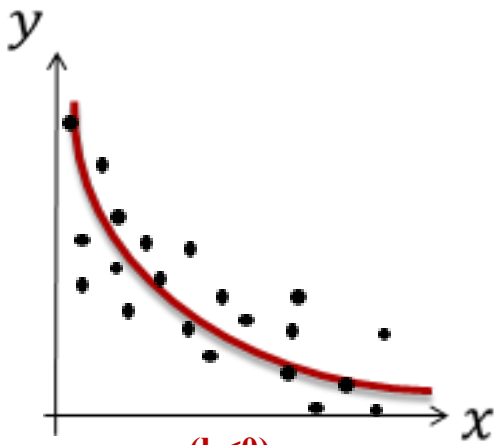
一、直接置换法

1.倒数函数模型（双曲函数模型）

$$\frac{1}{y} = a + \frac{b}{x} + \mu$$



(b>0)



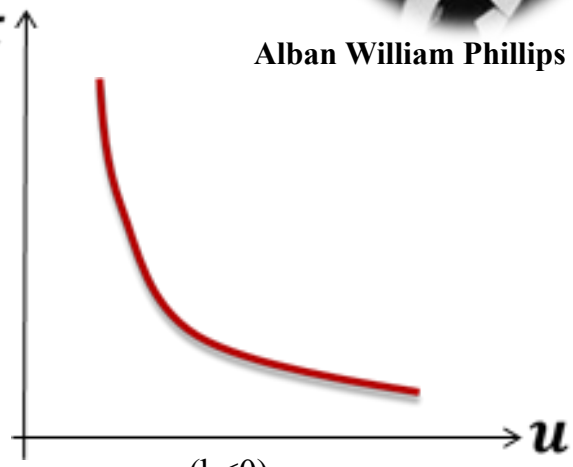
(b<0)

【例】菲利普斯曲线 $\pi = a + b\frac{1}{u} + \mu$

其中：π—通胀率、u—失业率

Alban William Phillips

菲利普斯曲线表明——失业与通货膨胀存在一种交替关系的曲线，通货膨胀率高时，失业率低；通货膨胀率低时，失业率高。

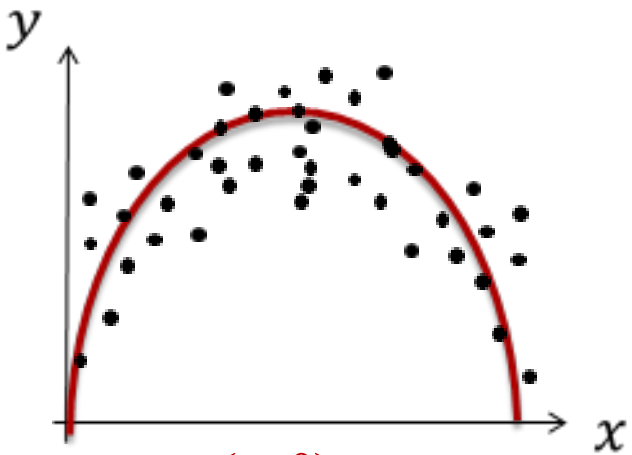


(b<0)

线性变换方法：令 $\frac{1}{u} = U^*$，得到：$\pi = a + bU^* + \mu$

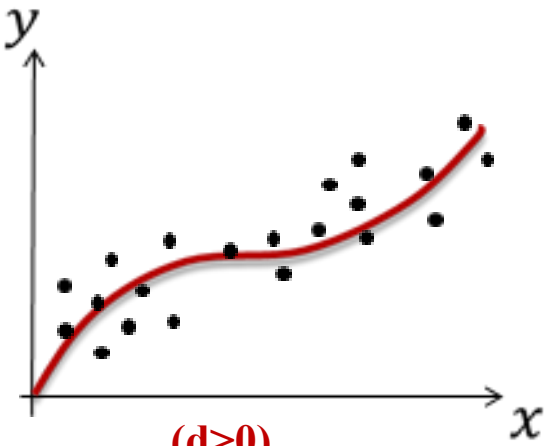
2.多项式函数模型

$$y = a + bx + cx^2 + \mu$$



(c<0)

$$y = a + bx + cx^2 + dx^3 + \mu$$



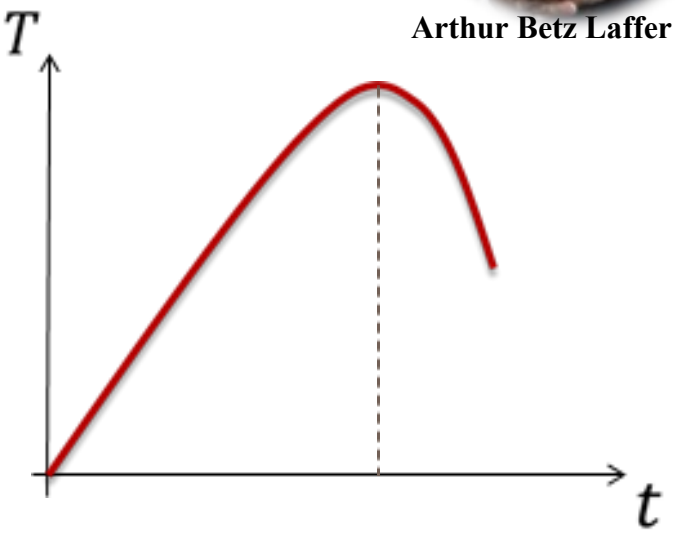
(d>0)

【例】拉弗曲线 $T = a + bt + ct^2 + \mu$

其中：T—税收收入，t—税率

Arthur Betz Laffer

拉弗曲线描绘了政府的税收收入与税率之间的关系，当税率在一定的限度以下时，提高税率能增加政府税收收入，但超过这一限度时，较高的税率将抑制经济的增长，使税基减小，再提高税率反而导致政府税收收入减少。

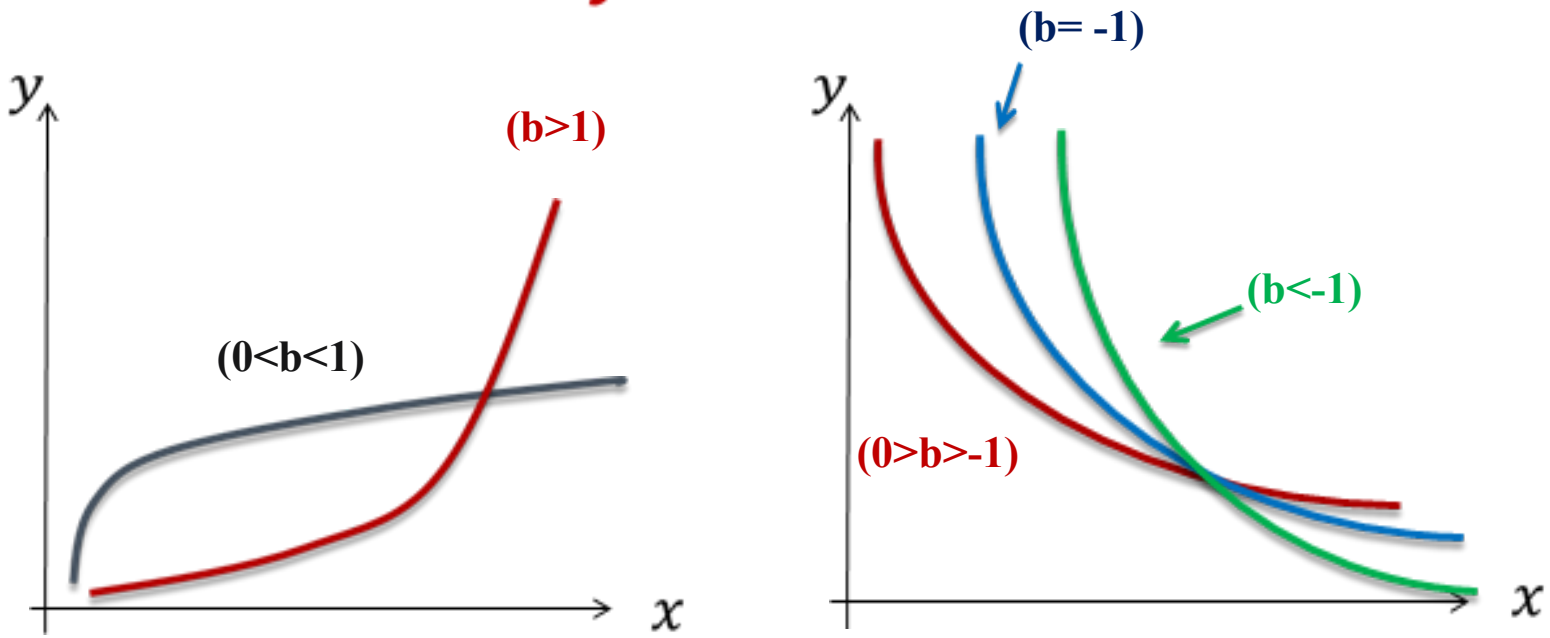


线性变换方法：令$t^2 = t^*$，得到：$T = a + bt + ct^* + \mu$

二、函数变换法

1.幂函数模型

$$y = ax^b e^{\mu}$$



【例】柯布道格拉斯生产函数 $Q = AK^{\alpha}L^{\beta}e^{\mu}$

其中：Q—产出量， A—效率系数， K—资本，

L—劳动

线性变换方法：

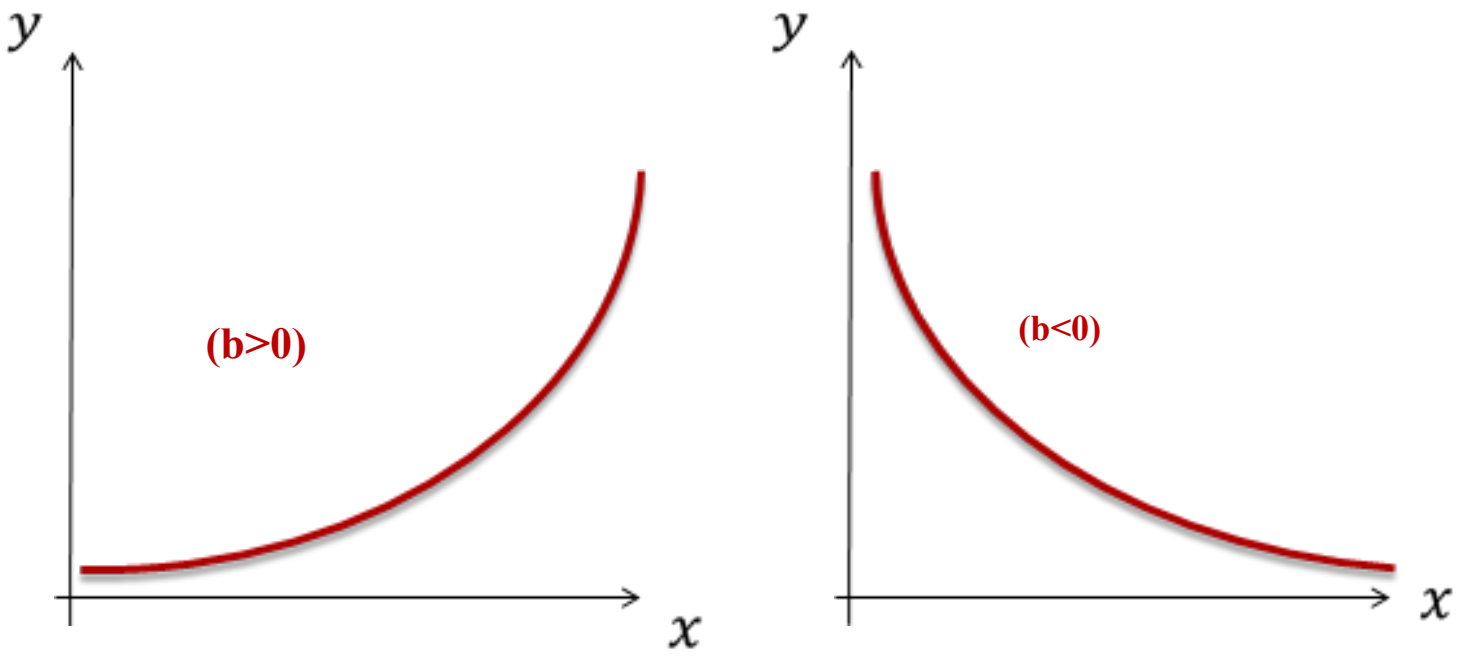
方程两边同时取对数

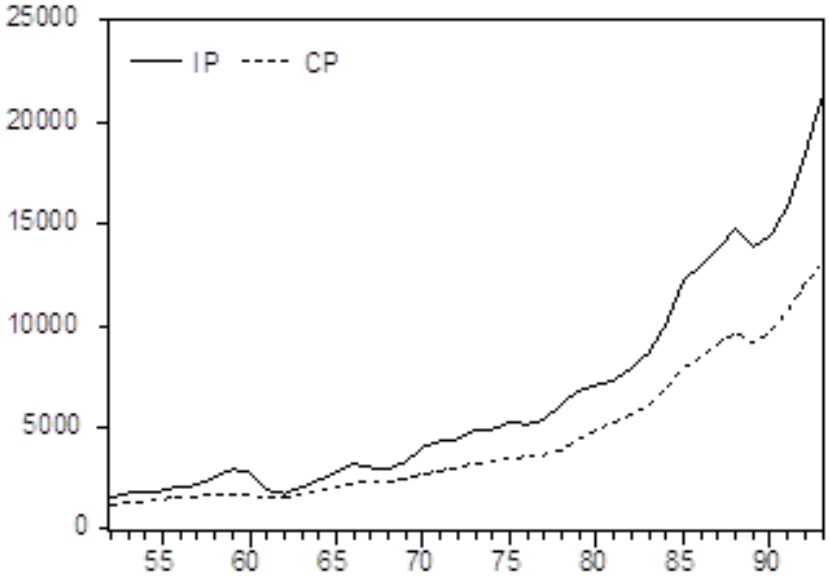
令$lnQ = Q^{*}$，$lnA = A^{*}$，$lnK = K^{*}$，$lnL = L^{*}$，

得到 $Q^{*} = A^{*} + \alpha K^{*} + \beta L^{*} + \mu$

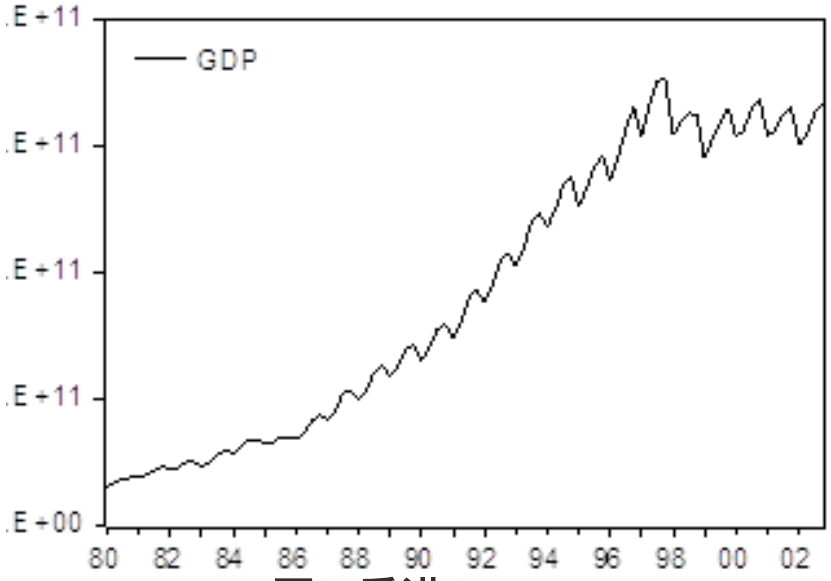
2.指数函数模型

$$y = ae^{bx+\mu}$$





图a 我国国民收入和消费



图b 香港GDP

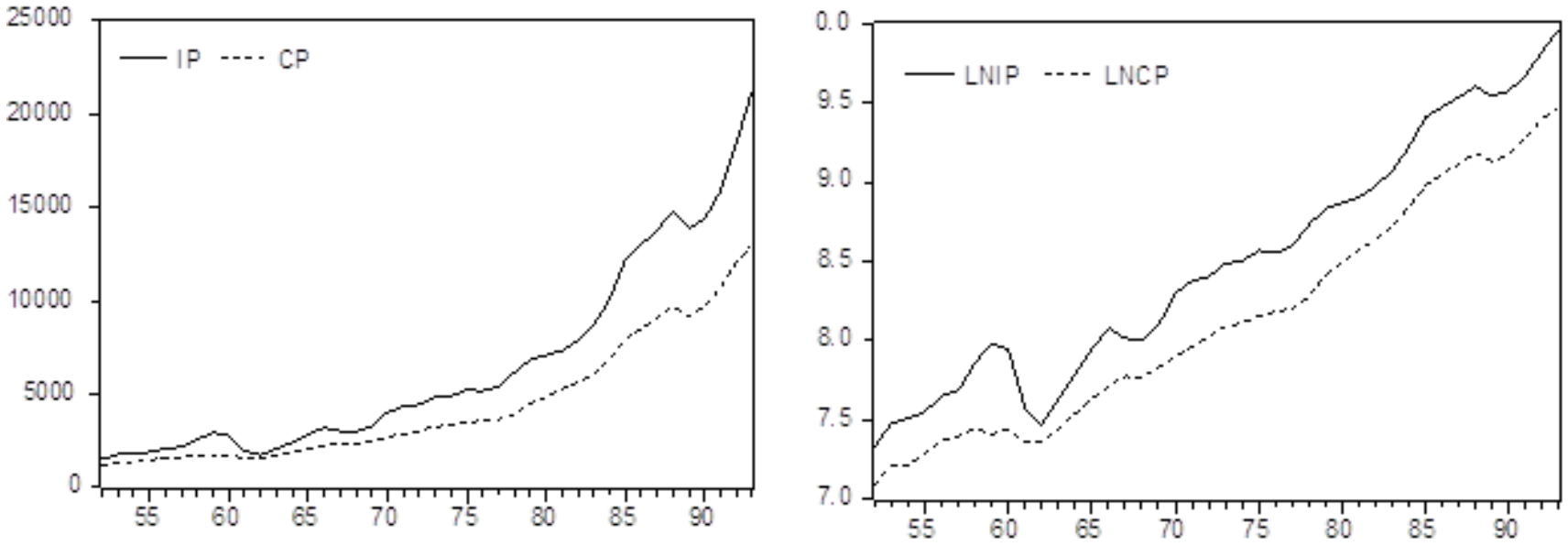
【例】

增长模型 $y = ae^{bt+\mu}$，其中：y—经济变量，t—时间

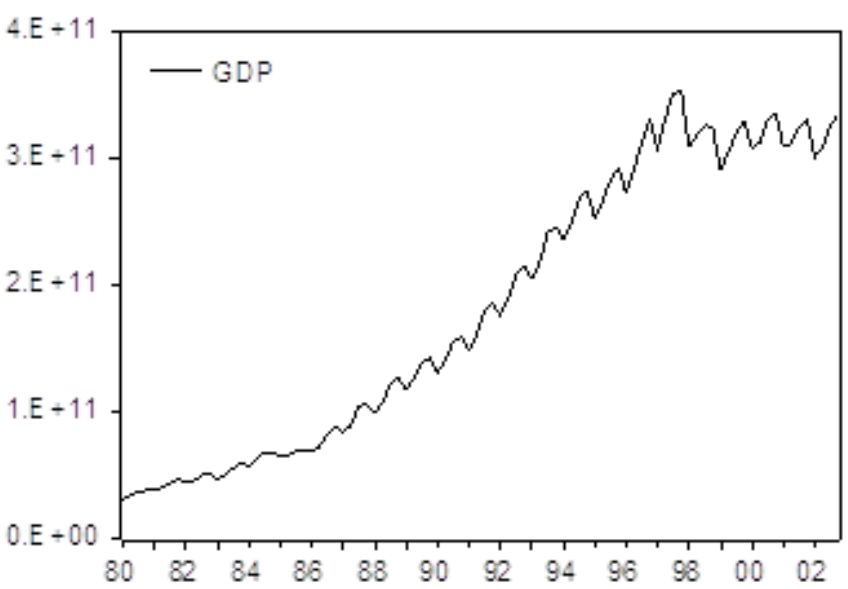
线性变换方法：

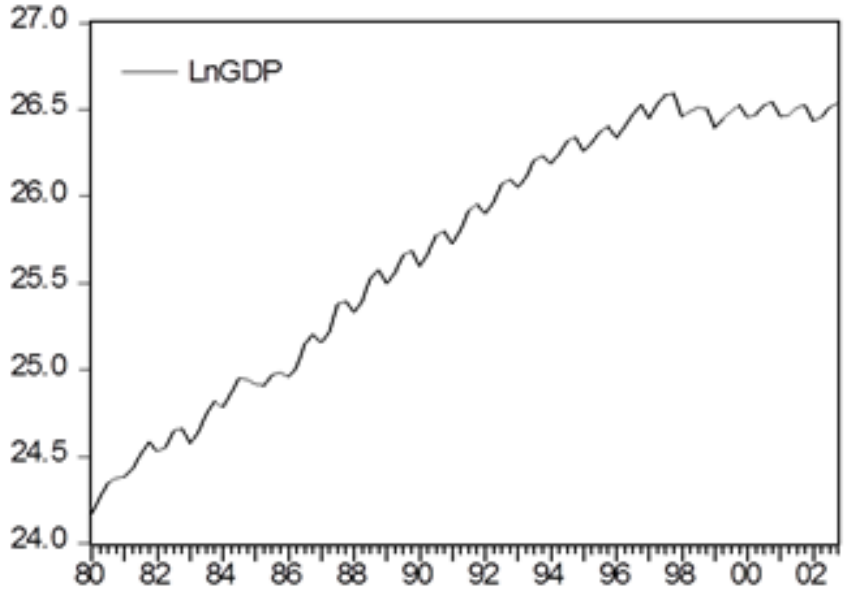
方程两边同时取对数，令$lny = y^*$，$lna = a^*$

得到 $y^* = a^* + bt + \mu$



图c 我国国民收入和消费与其对数序列

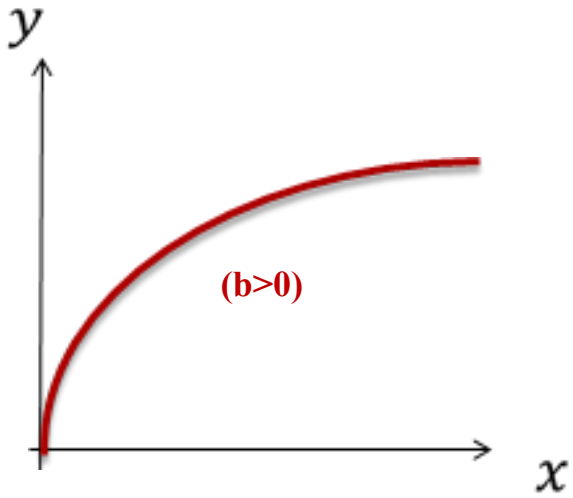


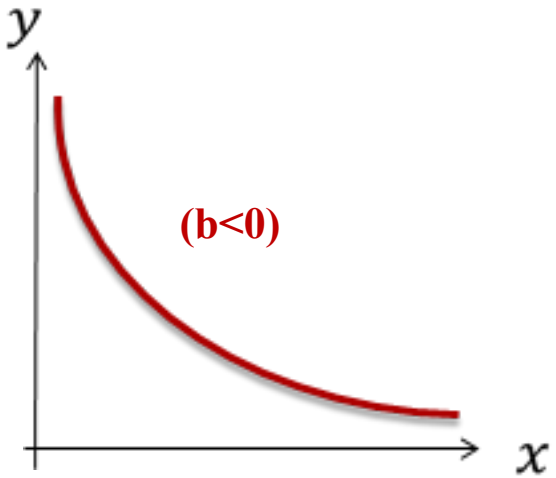


图d 香港GDP与其对数序列

3.对数函数模型

$$y = a + b\ln x + \mu$$





线性变换方法：令$\ln x = x^*$，得到 $y = a + bx^* + \mu$

第三章 多元线性回归模型

3.1 多元线性回归模型概述

3.2 多元线性回归模型的参数估计

3.3 多元线性回归模型的统计检验

3.4 非线性回归模型的线性化

3.5 含有虚拟变量的多元线性回归模型

3.5 含有虚拟变量的多元线性回归模型

引入虚拟变量的原因：将无法量化的变量进行“量化”。

一、含有虚拟变量的模型

二、虚拟变量的引入

三、虚拟变量的设置原则

一、含有虚拟变量的模型

定义：为了将无法量化的因素引入模型，提高模型精度构造只取“0”和“1”的人工变量，通常称为虚拟变量，记为“D”。

【例】反映性别的虚拟变量

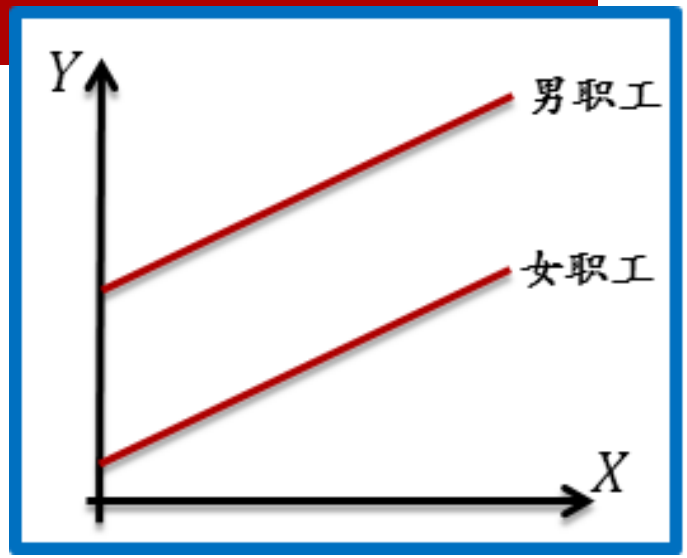
$$D=\begin{cases}1, \text{男性}\\ 0, \text{女性}\end{cases}$$

虚拟变量中，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。得到以性别为虚拟变量来考察员工薪资的模型如下

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

其中：Y_i－员工薪金，X_i－工龄，$D_i=1$代表男性，$D_i=0$代表女性。

二、虚拟变量的引入



两种引入方式：加法方式和乘法方式。

（一）加法方式

定义：将虚拟变量以相加方式引入模型之中。

【例】对上例模型$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$，假定$E(\mu_i) = 0$，则男、女职工平均薪金分别为

$$\begin{cases} \text{女职工平均薪金：} & E(Y_i|X, D=0) = \beta_0 + \beta_1 X_i \\ \text{男职工平均薪金：} & E(Y_i|X, D=1) = (\beta_0 + \beta_2) + \beta_1 X_i \end{cases}$$

若$\beta_2 > 0$，则两个函数斜率相同，截距不同。可以通过变量显著性检验对β_2进行显著性检验，以判断男女职工的平均薪资是否存在显著差异。

● 两个类型的虚拟变量引入

对上例，继续引入学历因素：

$$D_1=\begin{cases}1，男\\0，女\end{cases}\quad，\quad D_2=\begin{cases}1，本科及以上学历\\0，本科以下学历\end{cases}$$

得到：$\boldsymbol{Y_i=\beta_0+\beta_1X_i+\beta_2D_1+\beta_3D_2+\mu_i}$

即，女员工平均薪金为

$$\begin{cases}\text{女员工本科以下学历：}\ \boldsymbol{E(Y_i|X,D_1=0,\ D_2=0)=\beta_0+\beta_1X_i}\\ \text{女员工本科以上学历：}\ \boldsymbol{E(Y_i|X,D_1=0,\ D_2=1)=\beta_0+\beta_1X_i+\beta_3}\end{cases}$$

【例】考虑个人保健支出对个人收入和教育水平的回归，教育水平分为三个层次：高中以下、高中和大学及其以上。引入两个虚拟变量：

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

模型设定为：$Y_i=\beta_0+\beta_1X_i+\beta_2D_1+\beta_3D_2+\mu_i$

其中：Y_i—保健支出，X_i—个人收入

可以得到个人保健支出函数：

$$\begin{cases}\text{高中以下：}E(Y_i|X,D_1=0,\ D_2=0)=\beta_0+\beta_1X_i\\ \text{高\quad 中：}E(Y_i|X,D_1=1,\ D_2=0)=(\beta_0+\beta_2)+\beta_1X_i\\ \text{大学以上：}E(Y_i|X,D_1=0,\ D_2=1)=(\beta_0+\beta_3)+\beta_1X_i\end{cases}$$

（二）乘法方式

采用乘法方式引入虚拟变量的原因：如果虚拟变量引起了斜率变化，或斜率和截距同时变化，这需要用乘法方式引入。

【例】对中国城镇和农村居民边际消费倾向不同的考察

$$D_i = \begin{cases} 1, & \text{农村居民} \\ 0, & \text{城镇居民} \end{cases}$$

则全体居民消费模型为：

$$C_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + \mu_i = \beta_0 + (\beta_1 + \beta_2 D_i) X_i + \mu_i$$

其中： C-人均消费，X-人均收入。

$$\begin{cases} \text{农村居民}: E(C_i|X, D=1) = \beta_0 + (\beta_1 + \beta_2) X_i \\ \text{城镇居民}: E(C_i|X, D=0) = \beta_0 + \beta_1 X_i \end{cases}$$

如果 β_2 显著异于0，就可判定城镇和农村居民边际消费倾向不同。

三、虚拟变量的设置原则

原则：每一定型变量所需的虚拟变量个数要比该定性变量的类别数少1。

【例】冷饮销量（Y）除了受k种定性变量（X_k）影响，还受到四季因素影响，需要引入3个虚拟变量：

$$D_{i1}=\begin{cases}1, & \text{春季}\\ 0, & \text{其他}\end{cases},\quad D_{i2}=\begin{cases}1, & \text{夏季}\\ 0, & \text{其他}\end{cases},\quad D_{i3}=\begin{cases}1, & \text{秋季}\\ 0, & \text{其他}\end{cases}$$

则冷饮销售模型为：

$$Y_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}+\alpha_1D_{i1}+\alpha_2D_{i2}+\alpha_3D_{i3}+\mu_i$$

如果引入第4个虚拟变量，虚拟变量组成变为：

$$D_{i1}=\begin{cases}1, & \text{春季}\\ 0, & \text{其他}\end{cases},\quad D_{i2}=\begin{cases}1, & \text{夏季}\\ 0, & \text{其他}\end{cases}$$

$$D_{i3}=\begin{cases}1, & \text{秋季}\\ 0, & \text{其他}\end{cases},\quad D_{i4}=\begin{cases}1, & \text{冬季}\\ 0, & \text{其他}\end{cases}$$

则冷饮销售模型变为：

$$Y_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}+\alpha_1D_{i1}+\alpha_2D_{i2}+\alpha_3D_{i3}+\alpha_4D_{i4}+\mu_i$$

其矩阵形式为：$Y = (X\ D)\begin{pmatrix}\beta\\ \alpha\end{pmatrix} + \mu$

如果只取6个观测值，其中春、夏取两次，秋冬取一次，则其中

$$(X\ D) = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} & 1 & 0 & 0 & 0 \\ 1 & X_{21} & \cdots & X_{2k} & 0 & 1 & 0 & 0 \\ 1 & X_{31} & \cdots & X_{3k} & 0 & 0 & 1 & 0 \\ 1 & X_{41} & \cdots & X_{4k} & 0 & 0 & 0 & 1 \\ 1 & X_{51} & \cdots & X_{5k} & 0 & 1 & 0 & 0 \\ 1 & X_{61} & \cdots & X_{6k} & 1 & 0 & 0 & 0 \end{pmatrix}$$

如果取4个虚拟变量，$(X\ D)$矩阵将不满秩，参数无法唯一求出。

这就是所谓的“虚拟变量陷阱”。

第三章 · 小结