§5.3 模型设定偏误问题

一、模型设定偏误的类型

二、模型设定偏误的后果

三、模型设定偏误的检验

回顾　多元线性回归模型的基本假定

假设1，解释变量是非随机的或固定的，且各X之间互不相关（无多重共线性）。

假设2，随机误差项具有零均值、同方差及不序列相关性

$$E(\mu_i) = 0$$

$$Var(\mu_i) = E(\mu_i^2) = \sigma^2 \qquad i \neq j \quad i, j = 1,2,\cdots,n$$

$$Cov(\mu_i, \mu_j) = E(\mu_i \mu_j) = 0$$

假设3，解释变量与随机项不相关

$$Cov(X_{ji}, \mu_i) = 0 \qquad j = 1,2\cdots,k$$

假设4，随机项满足正态分布

$$\mu_i \sim N(0, \sigma^2)$$

假设5，样本容量趋于无穷时，各解释变量的方差趋于有界常数

假设6，回归模型的设定是正确的。

一、模型设定偏误的类型

- 模型设定偏误主要有两大类:

(1)关于解释变量选取的偏误，主要包括漏选相关变量和多选无关变量，

(2)关于模型函数形式选取的偏误。

1、相关变量的遗漏
（omitting relevant variables）

- 例如，如果“正确”的模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu$$

而我们将模型设定为

$$Y = \alpha_0 + \alpha_1 X_1 + v$$

即设定模型时漏掉了一个相关的解释变量。

这类错误称为遗漏相关变量。

- 动态设定偏误（dynamic mis-specification）：遗漏相关变量表现为对Y或X滞后项的遗漏 。

2、无关变量的误选
(including irrevelant variables)

- 例如，如果

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$$

仍为“真”，但我们将模型设定为

$$Y=\alpha_0+\alpha_1X_1+\alpha_2X_2+\alpha_3X_3+\mu$$

即设定模型时，多选了一个无关解释变量。

3、错误的函数形式 (wrong functional form)

- 例如，如果“真实”的回归函数为

$$Y = AX_1^{\beta_1} X_2^{\beta_2} e^{\mu}$$

但却将模型设定为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

二、模型设定偏误的后果

- 当模型设定出现偏误时，模型估计结果也会与“实际”有偏差。这种偏差的性质与程度与模型设定偏误的类型密切相关。

1、遗漏相关变量偏误

采用遗漏相关变量的模型进行估计而带来的偏误称为遗漏相关变量偏误（omitting relevant variable bias）。

设正确的模型为

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$$

却对

$$Y=\alpha_0+\alpha_1X_1+v$$

进行回归，得

$$\hat{\alpha}_1=\frac{\sum x_{1i}y_i}{\sum x_{1i}^2}$$

将正确模型 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$ 的离差形式

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i - \bar{\mu}$$

代入 $\hat{\alpha}_1 = \dfrac{\sum x_{1i} y_i}{\sum x_{1i}^2}$ 得

$$\hat{\alpha}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2} = \frac{\sum x_{1i}(\beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i - \bar{\mu})}{\sum x_{1i}^2}$$

$$= \beta_1 + \beta_2 \frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i}(\mu_i - \bar{\mu})}{\sum x_{1i}^2}$$

(1)如果漏掉的X_2与X_1相关，则上式中的第二项在小样本下求期望与大样本下求概率极限都不会为零，从而使得OLS估计量在小样本下有偏，在大样本下非一致。

(2)如果X_2与X_1不相关，则α_1的估计满足无偏性与一致性；但这时α_0的估计却是有偏的。

(3) 随机扰动项μ的方差估计$\hat{\sigma}^2$也是有偏的。

(4)$\hat{\alpha}_1$的方差是真实估计量$\hat{\beta}_1$的方差的有偏估计。

由 $Y=\alpha_0+\alpha_1X_1+v$ 得

$$Var(\hat{\alpha}_1)=\frac{\sigma^2}{\sum x_{1i}^2}$$

由 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$ 得

$$Var(\hat{\beta}_1)=\sigma^2\frac{\sum x_{2i}^2}{\sum x_{1i}^2\sum x_{2i}^2-(\sum x_{1i}x_{2i})^2}=\frac{\sigma^2}{\sum x_{1i}^2(1-r_{x_1x_2}^2)}$$

如果X_2与X_1相关，显然有 $Var(\hat{\alpha}_1)\neq Var(\hat{\beta}_1)$

如果X_2与X_1不相关，也有 $Var(\hat{\alpha}_1)\neq Var(\hat{\beta}_1)$

2、包含无关变量偏误

采用包含无关解释变量的模型进行估计带来的偏误，称为包含无关变量偏误（including irrelevant variable bias）。

设 $Y=\alpha_0+\alpha_1X_1+v$ (*)

为正确模型，但却估计了

$Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$ (**)

如果$\beta_2=0$，则(**)与(*)相同，因此，可将(**)式视为以$\beta_2=0$为约束的(*)式的特殊形式。

由于所有的经典假设都满足，因此对

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu \qquad (**)$$

式进行OLS估计，可得到无偏且一致的估计量。

注意： 由于$\beta_2=0$，因此，$E(\hat{\beta}_2)=0$。

但是，OLS估计量却不具有最小方差性。

$Y=\alpha_0+\alpha_1X_1+v$ 中X_1的方差：$Var(\hat{\alpha}_1)=\dfrac{\sigma^2}{\sum x_{1i}^2}$

$Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$ 中X_1的方差：$Var(\hat{\beta}_1)=\dfrac{\sigma^2}{\sum x_{1i}^2(1-r_{x_1x_2}^2)}$

当X_1与X_2完全线性无关时：$Var(\hat{\alpha}_1)=Var(\hat{\beta}_1)$

否则：$Var(\hat{\beta}_1)>Var(\hat{\alpha}_1)$

总结：

- ✓ 在多选无关解释变量的情形下，OLS估计是无偏且一致的估计量
- ✓ 随机干扰项的方差也能被正确估计
- ✓ 但是OLS估计往往是无效的

3、错误函数形式的偏误

当选取了错误函数形式并对其进行估计时，带来的偏误称错误函数形式偏误（wrong functional form bias）。

容易判断，这种偏误是全方位的。

例如，如果“真实”的回归函数为

$$Y = AX_1^{\beta_1} X_2^{\beta_2} e^{\mu}$$

却估计线性式

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v$$

显然，两者的参数具有完全不同的经济含义，且估计结果一般也是不相同的。

三、模型设定偏误的检验

1、检验是否含有无关变量

可用t 检验与F检验完成。

检验的基本思想：如果模型中误选了无关变量，则其系数的真值应为零。因此，只须对无关变量系数的显著性进行检验。

t检验：检验某1个变量是否应包括在模型中；

F检验：检验若干个变量是否应同时包括在模型中

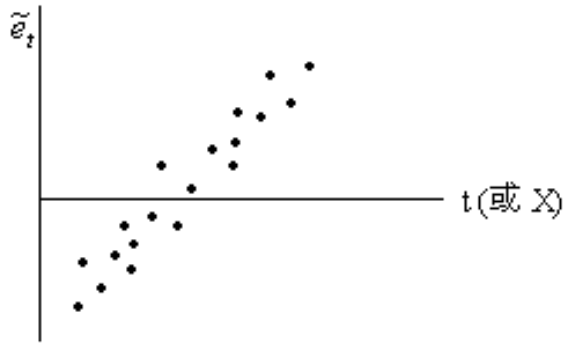
2、检验是否有相关变量的遗漏或函数形式设定偏误

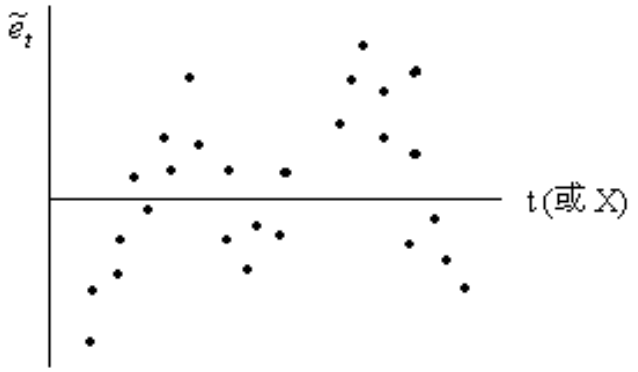
（1）残差图示法

对所设定的模型进行OLS回归，得到估计的残差序列$\tilde{e}_t$；

做出$\tilde{e}_t$与时间t或某解释变量X的散点图，考察$\tilde{e}_t$是否有规律地在变动，以判断是否遗漏了重要的解释变量或选取了错误的函数形式。

- 残差序列变化图

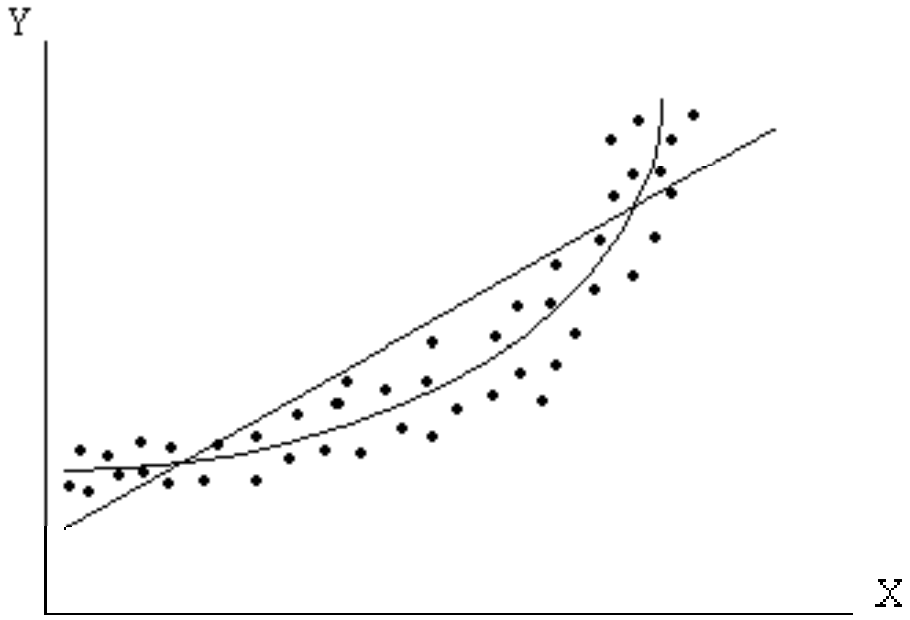




（a）趋势变化：
模型设定时可能遗漏了一随着时间的推移而持续上升的变量

（b）循环变化：
模型设定时可能遗漏了一随着时间的推移而呈现循环变化的变量

- 模型函数形式设定偏误时残差序列呈现正负交替变化



图示：一元回归模型中，真实模型呈幂函数形式，但却选取了线性函数进行回归。

（2）一般性设定偏误检验

但更准确更常用的判定方法是拉姆齐(Ramsey)于1969年提出的所谓**RESET 检验**（regression error specification test）。

基本思想：

如果事先知道遗漏了哪个变量，只需将此变量引入模型，估计并检验其参数是否显著不为零即可；

问题是不知道遗漏了哪个变量，需寻找一个替代变量Z，来进行上述检验。

RESET检验中，采用所设定模型中被解释变量Y的估计值$\hat{Y}$的若干次幂来充当该“替代”变量。

例如，先估计 $Y=\alpha_0+\alpha_1X_1+v$ 得

$$\hat{Y}=\hat{\alpha}_0+\hat{\alpha}_1X_1$$

再用通过残差项$\tilde{e}_t$与估计的$\hat{Y}$的图形判断引入$\hat{Y}$的若干次幂充当“替代”变量。

如$\tilde{e}_t$与Y的图形呈现曲线形变化时，回归模型可选为：

$$Y=\beta_0+\beta_1X_1+\gamma_1\hat{Y}^2+\gamma_2\hat{Y}^3+\mu$$

再根据第三章第五节介绍的增加解释变量的F检验来判断是否增加这些“替代”变量。

若仅增加一个“替代”变量，也可通过t检验来判断。

RESET检验也可用来检验函数形式设定偏误的问题。

例如，在一元回归中，假设真实的函数形式是非线性的，用泰勒定理将其近似地表示为多项式：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \mu \qquad (*)$$

因此，如果设定了线性模型，就意味着遗漏了相关变量X_1^2、 X_1^3，等等。

因此，在一元回归中，可通过检验(*)式中的各高次幂参数的显著性来判断是否将非线性模型误设成了线性模型。

对多元回归，非线性函数可能是关于若干个或全部解释变量的非线性，这时可按遗漏变量的程序进行检验。

例如，估计 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\mu$

但却怀疑真实的函数形式是非线性的。

这时，只需以估计出的$\hat{Y}$的若干次幂为“替代”变量，进行类似于如下模型的估计

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\gamma_1\hat{Y}^2+\gamma_2\hat{Y}^3+\mu$$

再判断各“替代”变量的参数是否显著地不为零即可。

案例：中国商品进口模型

我们主要研究中国商品进口与国内生产总值的关系。（下表）。

表 4.2.1　**1978~2001 年中国商品进口与国内生产总值**

	国内生产总值 GDP （亿元）	商品进口 M （亿美元）		国内生产总值 GDP （亿元）	商品进口 M （亿美元）
1978	3624.1	108.9	1990	18547.9	533.5
1979	4038.2	156.7	1991	21617.8	637.9
1980	4517.8	200.2	1992	26638.1	805.9
1981	4862.4	220.2	1993	34634.4	1039.6
1982	5294.7	192.9	1994	46759.4	1156.1
1983	5934.5	213.9	1995	58478.1	1320.8
1984	7171.0	274.1	1996	67884.6	1388.3
1985	8964.4	422.5	1997	74462.6	1423.7
1986	10202.2	429.1	1998	78345.2	1402.4
1987	11962.5	432.1	1999	82067.46	1657
1988	14928.3	552.7	2000	89442.2	2250.9
1989	16909.2	591.4	2001	95933.3	2436.1

资料来源：《中国统计年鉴》（1995、2000、2002）。

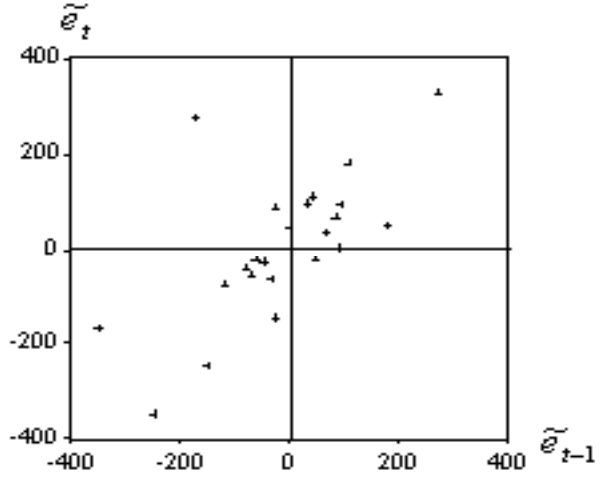
1. 通过OLS法建立如下中国商品进口方程：

$$\hat{M}_t = 152.91 + 0.02GDP_t$$

（2.32）　（20.12）

R^2=0.948　$\overline{R^2}$ =0.946　SE=154.9　DW=0.628

2. 进行序列相关性检验。



上例估计了中国商品进口M与GDP的关系，并发现具有强烈的一阶自相关性。

然而，由于仅用GDP来解释商品进口的变化，明显地遗漏了诸如商品进口价格、汇率等其他影响因素。因此，序列相关性的主要原因可能就是建模时遗漏了重要的相关变量造成的。

下面进行RESET检验。

用原回归模型估计出商品进口序列

$$\hat{M}_t = 152.91 + 0.020 GDP_t$$

$$R^2 = 0.9484$$

在原回归模型中加入$\hat{M}_t^2$、$\hat{M}_t^3$后重新进行估计，得：

$$\tilde{M}_t = -3.860 + 0.072GDP - 0.0028\hat{M}_t^2 + 8.59E-07\hat{M}_t^3$$

（-0.085）　（8.274）　（-6.457）　　（6.692）

$R^2=0.9842$

$$F = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-(k+q+1))} = \frac{(0.984-0.948)/2}{(1-0.984)/(24-4)} = 22.5$$

在$\alpha=5\%$下，查得临界值$F_{0.05}(2, 20)=3.49$

判断：拒绝原模型与引入新变量的模型可决系数无显著差异的假设，表明原模型确实存在遗漏相关变量的设定偏误。

例1　　在回归模型中由于自变量的_____，进行最小二乘法估计引起的后果有：(1)估计量是有偏的；(2)估计量的方差变小。由于自变量的_____，进行最小二乘法估计引起的后果有：(1)估计量是无偏的；(2)估计量的方差变大。

例 2、一个估计某行业 ECO 薪水的回归模型如下

$$\ln(salary) = \beta_0 + \beta_1 \ln(sales) + \beta_2 \ln(mktval) + \beta_3 profm\,\text{arg} + \beta_4 ceoten + \beta_5 comten + \mu$$

其中，salary 为年薪 sales 为公司的销售收入，mktval 为公司的市值，profmarg 为利润占销售额的百分比，ceoten 为其就任当前公司 CEO 的年数，comten 为其在该公司的年数。一个有 177 个样本数据集的估计得到 R^2=0.353。若添加 $ceoten^2$ 和 $comten^2$ 后，R^2=0.375。问：此模型中是否有函数设定的偏误？

例3　假设真实模型是

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$$

但你估计了

$$Y_i = \alpha_1 + \alpha_2 X_i + v_i$$

如果你利用Y在$X=-3$、-2、-1、0、1、2、3处的观测并估计了“不正确”的模型，这估计值将出现什么偏误？③

例 4、假设真实模型是：

$$Y_i = \beta_1 X_i + u_i$$

但没有拟和这个过原点回归，却例行拟和了通常带有截距项的模型：

$$Y_i = \alpha_0 + \alpha_1 X_i + v_i$$ 。

评述这一设定误差的后果。

例 5、假设真实模型是：

$$Y_i = \alpha_0 + \alpha_1 X_i + v_i$$

但拟和的模型为：

$$Y_i = \beta_1 X_i + u_i$$ 。

评述这一设定误差的后果。

例 6　考虑如下“真实”（柯布-道格拉斯）生产函数：

$$\ln Y_i = \alpha_0 + \alpha_1 \ln L_{1i} + \alpha_2 \ln L_{2i} + \alpha_3 \ln K_i + u_i$$

其中Y＝产出；

L_1＝生产性劳动；

L_2＝非生产性劳动；

K＝资本。

但若在经验研究中实际用的回归是：

$$\ln Y_i = \beta_0 + \beta_1 \ln L_{1i} + \beta_2 \ln K_i + u_i$$

假定你拥有有关变量的横截面数据，

a. 我们会得到 $E(\hat{\beta}_1) = \alpha_1$ 和 $E(\hat{\beta}_2) = \alpha_3$ 吗?

b. 如果知道 L_2 是生产函数中的一个无关变量，（a）中的答案能否成立