第九讲 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

四、案例分析

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 $Var(\hat{\alpha}_1)\neq Var(\hat{\beta}_1)$ Why?

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 $$Y=\alpha_0+\alpha_1X_1+v \qquad (*)$$

为正确模型，但却估计了

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

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

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

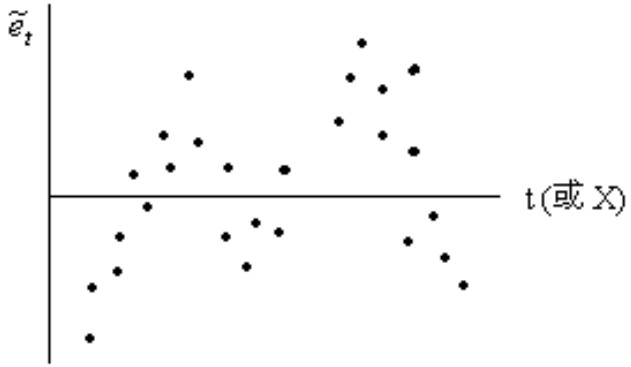
2、检验是否有相关变量的遗漏或函数形式设定偏误

（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。

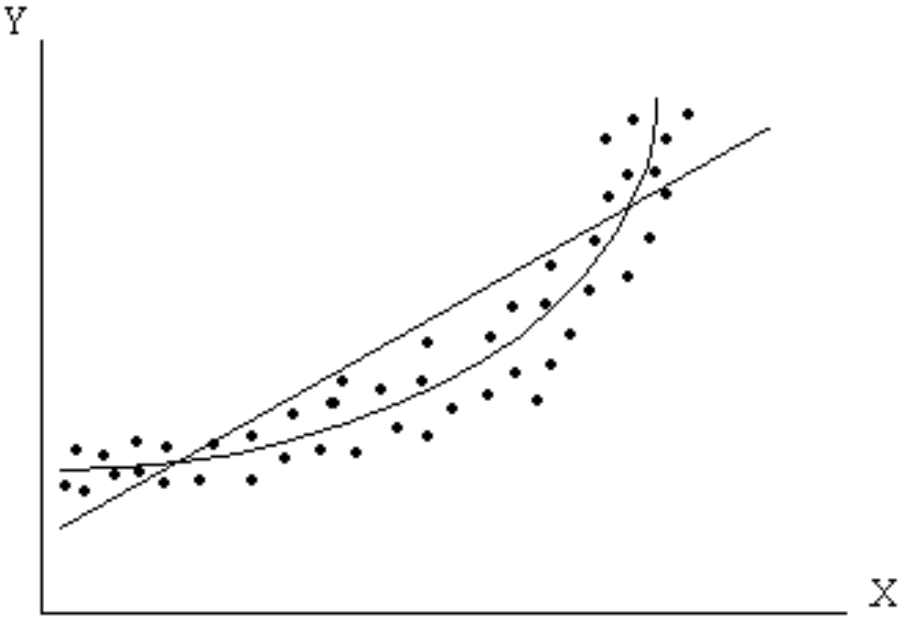
- 残差序列变化图



（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。

例5.3.1： 在§4.3商品进口的例中,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。

*（3）同期相关性的豪斯蔓（Hausman）检验

由于在遗漏相关变量的情况下，往往导致解释变量与随机扰动项出现同期相关性，从而使得OLS估计量有偏且非一致。

因此，对模型遗漏相关变量的检验可以用模型是否出现解释变量与随机扰动项同期相关性的检验来替代。这就是**豪斯蔓检验（1978）的主要思想。**

当解释变量与随机扰动项同期相关时，通过工具变量法可得到参数的一致估计量。

而当解释变量与随机扰动项同期无关时， OLS估计量就可得到参数的一致估计量。

因此，**只须检验IV估计量与OLS估计量是否有显著差异来检验解释变量与随机扰动项是否同期无关。**

对一元线性回归模型

$$Y=\beta_0+\beta_1 X+\mu$$

所检验的假设是 H_0：X与μ无同期相关。

设一元样本回归模型为

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

以Z为工具变量，则IV估计量为：

$$\tilde{\beta} = \frac{\sum z_i y_i}{\sum z_i x_i}$$

$$= \frac{\sum z_i (\hat{\beta}_1 x_i + e_i)}{\sum z_i x_i} = \hat{\beta}_1 + \frac{\sum z_i e_i}{\sum z_i x_i} \qquad (*)$$

其中，$\hat{\beta}_1$为OLS估计量。

(*)式表明，IV估计量与OLS估计量无差异当且仅当$\sum z_i e_i = 0$，即工具变量与OLS估计的残差项无关。

检验时，求Y关于X与Z的OLS回归式：

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\gamma}\, Z_i$$

如果$\hat{\gamma}$显著地异于零，就表明工具变量Z与$Y=\beta_0+\beta_1 X+\mu$式OLS估计的残差相关，因此，拒绝原假设，说明X与μ同期相关。

在实际检验中，豪斯蔓检验主要针对多元回归进行，而且也不是直接对工具变量回归，而是对以各工具变量为自变量、分别以各解释变量为因变量进行回归。

如对二元回归模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i \qquad (*)$$

如果选取了若干个工具变量$Z_1, Z_2, \ldots, Z_m$，分别以X_1与X_2为因变量关于所有工具变量做回归，求出数据序列$\hat{X}_1$与$\hat{X}_2$，再估计下面的方程：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma_1 \hat{X}_{1i} + \gamma_2 \hat{X}_{2i}$$

通过**增加解释变量的F检验**，检验联合假设：H_0：$\gamma_1=\gamma_2=0$ 。

拒绝原假设，就意味着（*）式中的解释变量与随机扰动项相关。

（4）线性模型与双对数线性模型的选择

无法通过判定系数的大小来辅助决策，因为在两类模型中被解释变量是不同的。

为了在两类模型中比较，可用Box-Cox变换:

第一步，计算Y的样本几何均值。

$$\tilde{Y} = (Y_1 Y_2 \cdots Y_n)^{1/n} = \exp(\frac{1}{n}\sum \ln Y_i)$$

第二步，用得到的样本几何均值去除原被解释变量Y，得到被解释变量的新序列Y*。

$$Y_i^* = Y_i / \tilde{Y}$$

第三步，用Y*替代Y，分别估计双对数线性模型与线性模型。并通过比较它们的残差平方和是否有显著差异来进行判断。

Zarembka（1968）提出的检验统计量为：

$$\frac{1}{2}n\ln\left(\frac{RSS_2}{RSS_1}\right)$$

其中，RSS_1与RSS_2分别为对应的较大的残差平方和与较小的残差平方和，n为样本容量。

可以证明：该统计量在两个回归的残差平方和无差异的假设下服从自由度为1 的χ^2分布。

因此，拒绝原假设时，就应选择RSS_2的模型。

例5.3.2 在§4.3中国商品进口的例中，

采用线性模型：$\mathbf{R^2=0.948}$;

采用双对数线性模型: $\mathbf{R^2=0.973}$，

但不能就此简单地判断双对数线性模型优于线性模型。下面进行Box-Cox变换。

计算原商品进口样本的几何平均值为：

$$\tilde{M} = \exp(\tfrac{1}{n}\sum \ln(M_t) = 583.12$$

计算出新的商品进口序列：

$$M_t^* = M_t / .\tilde{M}$$

以M_t^*替代M_t，分别进行双对数线性模型与线性模型的回归，得：

$$\ln(\hat{M}_t^*) = -1.3565 + 0.7836 \ln GDP_t \qquad RSS_1=0.5044$$

$$\hat{M}_t^* = 0.2622 + 0.000035 GDP_t \qquad RSS_2=1.5536$$

于是，$$\frac{1}{2} n \ln(\frac{RSS_2}{RSS_1}) = \frac{1}{2} \times 24 \ln(1.1249) = 13.49$$

在$\alpha=5\%$下，查得临界值$\chi^2_{0.05}(1)=3.841$

判断：拒绝原假设，表明**双对数线性模型确实“优于”线性模型。**

四、案例分析

1.数据来源

- 根据统计资料得到了美国工资的横截面数据，变量主要包括：wage=工资，educ=受教育年限，exper=工作经验年限，tenure=任职年限，lwage=工资的对数值。采用数据名为：“wage1.dta”工作文件。

$$\text{wage} = \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure}$$

- 利用wage1的数据，分别利用Link方法和Ramsey方法检验模型
- 是否遗漏了重要的解释变量。

- **2.实验操作指导**

- **（1）使用Link方法检验遗漏变量**
- Link方法进行检验的基本命令语句为：
- linktest [if] [in] [, cmd_options]
- 在这个命令语句中，linktest是进行Link检验的基本命令，if是表示条件的命令语句，in是范围语句，cmd_options表示Link检验的选项应该与所使用的估计方法的选项一致，例如检验之前使用的回归regress命令，则此处的选项应与regress的选项一致。

例如，利用wage1的数据，检验模型

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure$$

是否遗漏了重要的解释变量，应该输入以下命令：

use c:\data\wage1.dta,clear

reg lwage educ exper tenure

linktest

第一个命令表示打开数据文件wage1，第二个命令语句是对模型进行回归估计，第三个命令就是进行遗漏变量的Link检验，检验结果如图7.1所示。

从第二个表格中，可以看到hatsq项的p值为0.018，拒绝了hatsq系数为零的假设，即说明被解释变量lwage的拟合值的平方项具有解释能力，所以可以得出结论原模型可能遗漏了重要的解释变量。

为了进一步验证添加重要变量是否会改变Link检验的结果，我们生成受教育年限educ和工作经验年限exper的平方项，重新进行回归并进行检验，这时输入的命令如下：

```
gen educ2=educ^2
gen exper2=exper^2
reg lwage educ exper tenure educ2 exper2
linktest
```

第一个命令语句的作用是生成变量educ2，使其值为变量educ的平方；第二个命令语句的作用是生成变量exper2，使其值为变量exper的平方；第三个命令语句的作用是对进行回归估计;第四个命令就是进行遗漏变量的Link检验，检验结果如图7.2所示。

（2）使用Ramsey方法检验遗漏变量

Ramsey方法进行检验的基本命令语句为：

estat ovtest [, rhs]

在这个命令语句中，estat ovtest是进行Ramsey检验的命令语句，如果设定rhs，则在检验过程中使用解释变量，如果不设定rhs，则在检验中使用被解释变量的拟合值。

例如，利用wage1的数据，使用Ramsey方法检验模型：

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure$$

是否遗漏了重要的解释变量，应该输入以下命令：

use c:\data\wage1.dta,clear

reg lwage educ exper tenure

estat ovtest

- 在这组命令语句中，第一个命令的功能是打开数据文件，第二个命令是对模型进行回归估计，第三个命令就是进行遗漏变量的Ramsey检验，检验结果如图7.3所示。

- 在图7.3中，第一个图表仍然是回归结果，第二部分则是Ramsey检验的结果，不难发现Ramsey检验的原假设是模型不存在遗漏变量，检验的p值为0.0048，拒绝原假设，即认为原模型存在遗漏变量。

- 为了进一步验证添加重要变量是否会改变Ramsey检验的结果，我们采取Link检验中的方法，生成受教育年限educ和工作经验年限exper的平方项，重新进行回归并进行检验，这时输入的命令如下：
- gen educ2=educ^2
- gen exper2=exper^2
- reg lwage educ exper tenure educ2 exper2
- estat ovtest
- 这里不再赘述这些命令语句的含义，调整之后的检验结果如图7.4所示，可以发现此时检验的p值为0.5404，无法拒绝原假设，即认为模型不再存在遗漏变量。