第四章　经典单方程计量经济学模型：放宽基本假定的模型

前述计量经济学模型的回归分析，是在对线性回归模型提出若干基本假定的条件下，应用普通最小二乘法得到了无偏、有效且一致的参数估计量。但是，在实际的计量经济学问题中，完全满足这些基本假定的情况并不多见。不满足基本假定的情况，称为**基本假定违背**。对截面数据模型来说，违背基本假定的情形主要包括：

（1）解释变量之间存在严重的多重共线性；

（2）随机干扰项序列存在异方差性；

（3）解释变量具有内生性；

（4）模型设定有偏误。

在进行计量经济学模型的回归分析时，必须对所研究对象是否满足普通最小二乘法下的基本假定进行检验，即检验是否存在一种或多种违背基本假定的情况，这种检验称为**计量经济学检验**。经过计量经济学检验发现出现一种或多种基本假定违背时，则不能直接使用普通最小二乘法进行参数估计，而必须采取补救措施或发展新的估计方法。

§4.1　多重共线性

在本节首先讨论模型的解释变量之间存在多重共线性这一违背基本假设的问题。

一、多重共线性的含义

对于模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i \tag{4.1.1}$$

其基本假设之一是解释变量 $X_1, X_2, \cdots, X_k$ 是相互独立的。**如果某两个或多个解释变量之间出现了相关性，则称为存在多重共线性**(multicollinearity)。

如果存在

$$c_1 X_{i1} + c_2 X_{i2} + \cdots + c_k X_{ik} = 0 \tag{4.1.2}$$

其中，c_i 不全为 0，即某一个解释变量可以用其他解释变量的线性组合表示，则称为解释变量间存在**完全共线性**(perfect multicollinearity)。如果存在

$$c_1 X_{i1} + c_2 X_{i2} + \cdots + c_k X_{ik} + v_i = 0 \tag{4.1.3}$$

其中，c_i 不全为 0，v_i 为随机干扰项，则称为**近似共线性**(approximate multico llinearity)或**交互相关**(intercorrelated)。

在矩阵表示的线性回归模型

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\mu}$$

中，完全共线性指秩 $R(\boldsymbol{X})<k+1$，即矩阵

$$X=\begin{pmatrix}1 & X_{11} & X_{12} & \cdots & X_{1k}\\ 1 & X_{21} & X_{22} & \cdots & X_{2k}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk}\end{pmatrix}$$

中，至少有一个列向量可由其他列向量(不包括第一列)线性表出。例如，$X_2=\lambda X_1$，这时 X_1 与 X_2 的相关系数为 1，解释变量 X_2 对被解释变量的作用完全可由 X_1 代替。

完全共线性的情况并不多见，一般出现的是在一定程度上的共线性，即近似共线性。

二、实际经济问题中的多重共线性

一般地，产生多重共线性的主要原因有以下三个方面：

1. 经济变量相关的共同趋势

样本数据中发生多重共线性的主要原因在于许多经济变量存在相关的共同趋势。例如，以某一行业的企业为样本建立企业生产函数模型，以产出量为被解释变量，选择资本、劳动力、技术等投入要素为解释变量。这些投入要素的数量往往与产出量成正比，产出量高的企业，投入的各种要素都比较多，这就使得投入要素之间出现线性相关性。如果以简单线性关系作为模型的数学形式，那么多重共线性是难以避免的。

2. 模型设定不谨慎

在计量模型设定中，往往由于不谨慎而导致模型解释变量间出现严重多重共线性。例如，为估计一个常弹性消费函数的扩展形式，将模型设定为

$$\ln C_i=\beta_0+\beta_1\ln Y_i+\beta_2\ln Y_i^2+\mu_i$$

其中，C 为家庭人均消费、Y 为家庭人均收入。显然，模型中引入的家庭人均收入的对数项与人均收入平方的对数项之间有着完全的线性相关性。

又例如，在考察学校支出对学生平均成绩的影响时，将学校的总支出 X_0 分解为对教职员工的工资性支出 X_1 以及其他支出 X_2，并设定如下模型

$$Y_i=\beta_0+\beta_1X_{i0}+\beta_2X_{i1}+\beta_3X_{i2}+\mu_i$$

其中，Y 代表学校的平均成绩。显然，由于 $X_1+X_2=X_0$，模型的解释变量间存在完全共线性。

3. 样本资料的限制

由于完全符合理论模型所要求的样本数据较难收集，在现有数据条件下，特定样本

可能存在某种程度的多重共线性。例如，将家庭孩子的考试分数 Y 与教育支出 X_1 和家庭人均收入 X_2 相关联而设定如下模型

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

一般经验告诉我们，教育支出 X_1 与家庭人均收入 X_2 之间存在一定程度的相关性，如果由于样本收集的原因，恰好导致它们之间显示出很强的相关性，则会出现严重的多重共线性。

三、多重共线性的后果

计量经济学模型一旦出现多重共线性，如果仍采用普通最小二乘法估计模型参数，会产生下列不良后果。

1. 完全共线性下参数估计量不存在

多元线性回归模型

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mu}$$

的普通最小二乘参数估计量为

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

如果出现完全共线性，则 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在，无法得到参数的估计量。

例如，对二元线性回归模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu \tag{4.1.4}$$

如果两个解释变量完全相关，如 $X_2 = \lambda X_1$，该二元线性回归模型退化为一元线性回归模型

$$Y = \beta_0 + (\beta_1 + \lambda\beta_2) X_1 + \mu$$

这时，只能确定综合参数 $\beta_1 + \lambda\beta_2$ 的估计值

$$\widehat{\beta_1 + \lambda\beta_2} = \frac{\sum x_{i1} y_i}{\sum x_{i1}^2}$$

却无法确定 β_1, β_2 各自的估计值。

2. 近似共线性下普通最小二乘法参数估计量的方差变大

在近似共线性下，虽然可以得到普通最小二乘参数估计量，但是由参数估计量方差的表达式

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

可见，由于此时 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$，引起 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 主对角线元素较大，使得参数估计量的方差增大，从而不能对总体参数作出准确推断。

仍以二元线性回归模型(4.1.4)式为例。离差形式下容易推出 $\hat{\beta}_1$ 的方差为（参见《计量经济学学习指南与练习（第二版）》，潘文卿，李子奈编著，高等教育出版社，2015）

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1)&=\frac{\sigma^2\sum x_{i2}^2}{\sum x_{i1}^2\sum x_{i2}^2-(\sum x_{i1}x_{i2})^2}\\&=\frac{\dfrac{\sigma^2}{\sum x_{i1}^2}}{1-\dfrac{(\sum x_{i1}x_{i2})^2}{\sum x_{i1}^2\sum x_{i2}^2}}=\frac{\sigma^2}{\sum x_{i1}^2}\cdot\frac{1}{1-r^2}\end{aligned}\tag{4.1.5}$$

其中，$\dfrac{(\sum x_{i1}x_{i2})^2}{\sum x_{i1}^2\sum x_{i2}^2}$ 恰为 X_1 与 X_2 的线性相关系数的平方 r^2，由于 $r^2\leqslant 1$，故 $\dfrac{1}{1-r^2}\geqslant 1$。

当完全不共线性时，

$$r^2=0,\qquad \mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_{i1}^2}$$

当近似共线性时，

$$0<r^2<1,\qquad \mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_{i1}^2}\cdot\frac{1}{1-r^2}>\frac{\sigma^2}{\sum x_{i1}^2}$$

即多重共线性使参数估计量的方差增大，**方差膨胀因子**(variance inflation factor,VIF)为

$$\mathrm{VIF}(\hat{\beta}_1)=\frac{1}{1-r^2}\tag{4.1.6}$$

其增大趋势如表 4.1.1 所示。

当完全共线性时，

$$r^2=1,\qquad \mathrm{Var}(\hat{\beta}_1)=+\infty$$

表 4.1.1 方差膨胀因子表

相关系数平方	0	0.5	0.8	0.9	0.95	0.96	0.97	0.98	0.99	0.999
方差膨胀因子	1	2	5	10	20	25	33	50	100	1 000

3. 参数估计量经济意义不合理

如果模型中两个解释变量具有线性相关性，如 X_1 和 X_2，那么它们中的一个变量可以由另一个变量表征。这时，X_1 和 X_2 前的参数并不反映各自与被解释变量之间的结构关系，而是反映它们对被解释变量的共同影响，所以各自的参数已经失去了应有的经济含义，于是经常表现出似乎反常的现象，例如估计结果本来应该是正的，结果却是负的。**经验告诉我们，在多元线性回归模型的估计中，如果出现参数估计值的经济意义明显不合理的情况，应该首先怀疑是否存在多重共线性。**

4. 变量的显著性检验和模型的预测功能失去意义

存在多重共线性时，参数估计值的方差与标准差变大，从而容易使通过样本计算的 t 值小于临界值，误导做出参数为零的推断，可能将重要的解释变量排除在模型之外。

变大的方差容易使预测值区间预测的“区间”变大，使预测失去意义。

四、多重共线性的检验

由于多重共线性表现为解释变量之间具有相关关系，所以用于多重共线性的检验方法主要是统计方法，如判定系数检验法、逐步回归检验法等。多重共线性检验的任务是：(1) 检验多重共线性是否存在；(2) 判明存在多重共线性的范围。

1. 检验多重共线性是否存在

（1）对两个解释变量的模型，采用简单相关系数法。求出 X_1 与 X_2 的简单相关系数 r，若 $|r|$ 接近 1，则说明两变量存在较强的多重共线性。

（2）对多个解释变量的模型，采用综合统计检验法。若在普通最小二乘法下，模型的 R^2 与 F 值较大，但各参数估计的 t 检验值较小，说明各解释变量对 Y 的联合线性作用显著，但各解释变量间存在共线性而使得它们对 Y 的独立作用不能分辨，故 t 检验不显著。

2. 判明存在多重共线性的范围

如果存在多重共线性，需进一步确定多重共线性究竟由哪些变量引起。

（1）判定系数检验法。使模型中每个解释变量分别以其余解释变量为解释变量进行回归，并计算相应的拟合优度，也称为判定系数。如果在某一种形式中判定系数较大，则说明在该形式中作为被解释变量的 X_j 可以用其他解释变量的线性组合代替，即 X_j 与其他解释变量间存在共线性。

可进一步对上述出现较大判定系数的回归方程作 F 检验：

$$F_j = \frac{R_{j\cdot}^2/(k-1)}{(1-R_{j\cdot}^2)/(n-k)} \sim F(k-1, n-k) \tag{4.1.7}$$

其中 $R_{j\cdot}^2$ 为第 j 个解释变量对其他解释变量的回归方程的决定系数。若存在较强的共线性，则 $R_{j\cdot}^2$ 较大且接近于 1，这时 $1-R_{j\cdot}^2$ 较小，从而 F_j 的值较大。因此，可以给定显著性水平 α，通过计算的 F 值与相应的临界值的比较来进行检验。此时，原假设为 X_j 与其他解释变量间不存在显著的线性关系。

另一等价的检验是：在模型中排除某个解释变量 X_j，估计模型，如果拟合优度与包含 X_j 时十分接近，则说明 X_j 与其他解释变量之间存在共线性。

（2）逐步回归法。以 Y 为被解释变量，逐个引入解释变量，构成回归模型，进行模型估计。根据拟合优度的变化决定新引入的变量是否可以用其他变量的线性组合代替，而不是作为独立的解释变量。如果拟合优度变化显著，则说明新引入的变量是一个独立解释变量；如果拟合优度变化很不显著，则说明新引入的变量不是一个独立解释变量，它可以用其他变量的线性组合代替，也就是说它与其他变量之间存在共线性的关系。

五、克服多重共线性的方法

如果模型被证明存在多重共线性，则需要发展新的方法估计模型，最常用的方法有两类。

1. 第一类方法：排除引起共线性的变量

找出引起多重共线性的解释变量，将它排除出去，是最有效克服多重共线性问题的方法，所以逐步回归法得到了最为广泛的应用。但是，需要特别注意的是，当排除了某个或某些变量后，保留在模型中的变量的系数的经济意义将发生变化，其估计值也将发生变化。例如，在对数线性生产函数模型中，当包含资本、劳动、技术等投入要素时，资本的系数表示资本的产出弹性；但是，当资本和劳动存在共线性因而排除劳动时，资本的系数所表示的经济意义就不是资本的产出弹性，其估计值也将大于资本的产出 弹性。

*2. 第二类方法：减小参数估计量的方差

多重共线性的主要后果是参数估计量具有较大的方差。若采取适当方法减小参数估计量的方差，虽然没有消除模型中的多重共线性，却能消除多重共线性造成的后果。例如，增加样本容量，可使参数估计量的方差减小。

20 世纪 70 年代发展的岭回归法(ridge regression)，以引入偏误为代价减小参数估计量的方差。具体方法是：引入矩阵 $\boldsymbol{D}$，使参数估计量为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{4.1.8}$$

矩阵 $\boldsymbol{D}$ 一般选择主对角矩阵，即

$$\boldsymbol{D} = l\boldsymbol{I} \tag{4.1.9}$$

其中 l 为大于 0 的常数。显然，与普通最小二乘估计量相比，(4.1.8)式的估计量有较小的方差。

如何选择 l 是一个复杂的问题，何瑞尔(Hoerl)和肯纳德(Kennard)于 1975 年提出一种估计方法。首先对原模型的解释变量与被解释变量的离差形式进行标准化处理：

$$x_{ik}^{*} = \frac{x_{ik}}{\sqrt{\sum x_{ik}^2}}, \qquad y_{ik}^{*} = \frac{y_{ik}}{\sqrt{\sum y_{ik}^2}}$$

得到下列模型：

$$y_i^{*} = \beta_1^{*}x_{i1}^{*} + \beta_2^{*}x_{i2}^{*} + \cdots + \beta_k^{*}x_{ik}^{*} + \mu_i^{*}, \quad i = 1,2,\cdots,n$$

用普通最小二乘法估计该模型，得到参数与随机干扰项方差的估计值 $\hat{\beta}_1^{*},\hat{\beta}_2^{*},\cdots,\hat{\beta}_k^{*}$ 和 $\hat{\sigma}^2$。选择

$$\hat{l}=\frac{(k-1)\hat{\sigma}^2}{\sum_{j=1}^{k}(\hat{\beta}_j^*)^2}$$

作为(4.1.9)式中 l 的估计值。

最后需要指出的是，多重共线性是一种样本现象。同一个模型在一个样本下可能表现出多重共线性，而在另一个样本下可能就不存在多重共线性，因此增加样本容量就有可能消除多重共线性。

另外，多重共线性的主要问题在于使参数估计量的方差变大，而从(4.1.5)式知，随机干扰项的方差、变量的变异程度与方差膨胀因子一起决定着参数估计量的方差。如果存在多重共线性，但随机干扰项的方差很小，或变量的变异程度很大，都可能得到较小的参数估计量的方差。这时，即使有较严重的多重共线性，也不会带来不良后果。因此，只要回归方程估计的参数标准差较小，t 统计值较大，就没有必要过于关心是否存在多重共线性的问题。

六、案例

例 4.1.1

根据理论和经验分析，影响粮食生产(Y)的主要因素有：粮食播种面积(X_1)、有效灌溉面积(X_2)、化肥施用量(X_3)、大型拖拉机(X_4)、小型拖拉机(X_5)、农用排灌柴油机(X_6)。表 4.1.2 列出了中国 31 个省、市、自治区粮食生产的相关数据，拟建立 2013 年中国粮食生产函数模型。

表 4.1.2 中国粮食生产与相关投入资料

	粮食产量 Y (万吨)	粮食播种面积 X_1 (千公顷)	有效灌溉面积 X_2 (千公顷)	化肥施用量 X_3 (万吨)	大型拖拉机 X_4 (千台)	小型拖拉机 X_5 (千台)	农用排灌柴油机 X_6 (千台)
北　京	96.1	158.9	153.0	12.8	6.5	2.4	37.7
天　津	174.7	332.8	308.9	24.3	15.6	9.2	63.1
河　北	3 365.0	6 315.9	4 349.0	331.0	234.3	1424.2	1 523.9
山　西	1 312.8	3 274.3	1 382.8	121.0	107.2	347.4	144.2
内蒙古	2 773.0	5 617.3	2 957.8	202.4	623.4	428.2	180.5
辽　宁	2 195.6	3 226.4	1 407.8	151.8	208.0	322.5	809.9
吉　林	3 551.0	4 789.9	1 510.1	216.8	440.4	670.8	197.6
黑龙江	6 004.1	11 564.4	5 342.1	245.0	873.3	645.3	131.2
上　海	114.2	168.5	184.1	10.8	6.7	3.6	13.5
江　苏	3 423.0	5 360.8	3 785.3	326.8	131.3	925.4	415.9
浙　江	734.0	1 253.7	1 409.4	92.4	11.7	139.3	863.3
安　徽	3 279.6	6 625.3	4 305.5	338.4	179.9	2249.7	1 174.2

续表

	粮食产量 Y (万吨)	粮食播种面积 X_1 (千公顷)	有效灌溉面积 X_2 (千公顷)	化肥施用量 X_3 (万吨)	大型拖拉机 X_4 (千台)	小型拖拉机 X_5 (千台)	农用排灌柴油机 X_6 (千台)
福 建	664.4	1 202.1	1 122.4	120.6	3.1	104.5	65.1
江 西	2 116.1	3 690.9	1 995.6	141.6	10.2	289.8	221.5
山 东	4 528.2	7 294.6	4 729.0	472.7	500.7	1997	1 259.8
河 南	5 713.7	10 081.8	4 969.1	696.4	357.8	3 513.2	1 100.5
湖 北	2 501.3	4 258.4	2 791.4	351.9	149.4	1 141.2	698.1
湖 南	2 925.7	4 936.6	3 084.3	248.2	106.6	227.5	1 067.8
广 东	1 315.9	2 507.6	1 770.8	243.9	23.9	329.2	349.7
广 西	1 521.8	3 076.0	1 586.4	255.7	34.2	456.8	271.6
海 南	190.9	421.8	260.9	47.6	44.5	52.7	38.0
重 庆	1 148.1	2 253.9	675.2	96.6	3.8	7.8	759.5
四 川	3 387.1	6 469.9	2 616.5	251.1	121.8	119.0	307.3
贵 州	1 030.0	3 118.4	926.9	97.4	41.9	85.8	225.0
云 南	1 824.0	4 499.4	1 660.3	219.0	287.0	377	121.6
西 藏	96.2	175.9	239.3	5.7	66.4	138.3	0.9
陕 西	1 215.8	3 105.1	1 209.9	241.7	99.3	198.7	322.6
甘 肃	1 138.9	2 858.7	1 284.1	94.7	130.4	575.6	130.7
青 海	102.4	280.0	186.9	9.8	11.1	243.9	2.5
宁 夏	373.4	801.6	498.6	40.4	42.6	179.8	27.0
新 疆	1 377.0	2 234.8	4 769.9	203.2	397.2	316.9	69.8

资料来源：《中国统计年鉴》(2014)。

设粮食生产函数为

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \beta_4 \ln X_4 + \beta_5 \ln X_5 + \beta_6 \ln X_6 + \mu$$

1. 用普通最小二乘法估计模型

$$\ln \hat{Y} = -1.100 + 0.757 \ln X_1 + 0.246 \ln X_2 + 0.000\ 2 \ln X_3$$

$$(-2.24) \qquad (8.20) \qquad (2.53) \qquad (0.002)$$

$$+0.030 \ln X_4 - 0.032 \ln X_5 + 0.051 \ln X_6$$

$$(0.92) \qquad (-0.96) \qquad (1.22)$$

$$R^2 = 0.985\ 0 \qquad \overline{R}^2 = 0.981\ 2 \qquad F = 262.32$$

由于 R^2 较大且接近于 1，而且 $F=262.32 > F_{0.05}(6,24)=2.51$，故认为粮食生产与上述解释变量间总体线性关系显著。但由于其中 X_3、X_4、X_5、X_6 前参数估计值未能通过 t 检验，而且 X_5 的参数符号的经济意义也不合理，故认为解释变量间存在多重共线性。

2. 检验简单相关系数

$\ln X_1$，$\ln X_2$，$\ln X_3$，$\ln X_4$，$\ln X_5$，$\ln X_6$ 的相关系数如表 4.1.3 所示。表中数据显示 $\ln X_1$、$\ln X_2$、$\ln X_3$ 间存在高度相关性，同时，$\ln X_3$ 与 $\ln X_6$ 间的相关性也较高。

表 4.1.3 相关系数表

	$\ln X_1$	$\ln X_2$	$\ln X_3$	$\ln X_4$	$\ln X_5$	$\ln X_6$
$\ln X_1$	1.0000	0.9345	0.9453	0.6736	0.7509	0.7908
$\ln X_2$	0.9345	1.0000	0.9285	0.6847	0.7838	0.7496
$\ln X_3$	0.9453	0.9285	1.0000	0.5946	0.7182	0.8579
$\ln X_4$	0.6736	0.6847	0.5946	1.0000	0.7260	0.3342
$\ln X_5$	0.7509	0.7838	0.7182	0.7260	1.0000	0.4400
$\ln X_6$	0.7908	0.7496	0.8579	0.3342	0.4400	1.0000

3. 找出最简单的回归形式

分别作 $\ln Y$ 关于 $\ln X_1$，$\ln X_2$，$\ln X_4$，$\ln X_5$，$\ln X_6$ 的回归，发现 $\ln Y$ 关于 $\ln X_1$ 的回归具有最大的可决系数：

$$\ln \hat{Y} = -0.684 + 1.004 \ln X_1$$

$$(-3.08) \quad (35.14)$$

$$R^2 = 0.9771 \quad \overline{R}^2 = 0.9763$$

可见，粮食生产受粮食播种面积的影响最大，与经验相符合，因此选该一元回归模型为初始的回归模型。

4. 逐步回归

将其他解释变量分别导入上述初始回归模型，寻找最佳回归方程(表 4.1.4)。

表 4.1.4 逐 步 回 归

	C	$\ln X_1$	$\ln X_2$	$\ln X_3$	$\ln X_4$	$\ln X_5$	$\ln X_5$	$\overline{R}^2$
$Y=f(X_1)$	−0.684	1.004						0.9763
t 值	(−3.08)	(35.14)						
$Y=f(X_1, X_2)$	−0.915	0.812	0.238					0.9810
t 值	(−4.26)	(11.3)	(2.87)					
$Y=f(X_1, X_2, X_3)$	−0.722	0.769	0.209	0.071				0.9808
t 值	(−2.25)	(8.62)	(2.31)	(0.81)				
$Y=f(X_1, X_2, X_4)$	−0.90	0.813	0.241		−0.005			0.9803
t 值	(−3.65)	(11.03)	(7.79)		(−0.18)			
$Y=f(X_1, X_2, X_5)$	−0.789	0.820	0.281			−0.041		0.9817
t 值	(−3.46)	(11.60)	(3.24)			(−1.43)		
$Y=f(X_1, X_2, X_6)$	−1.081	0.761	0.231				0.050	0.9823
t 值	(−4.75)	(10.15)	(2.89)				(1.76)	

讨论：

第一步，在初始模型中引入 X_2，模型的 $\overline{R}^2$ 提高，且参数符号合理，变量也通过了显著性水平为 5%的 t 检验；

第二步，引入 X_3，模型的 $\overline{R}^2$ 有所下降，虽然参数符号合理，但变量甚至未通过显著性水平为 10%的 t 检验；

第三步，去掉 X_3，引入 X_4，模型的 $\overline{R}^2$ 仍没有只有 X_1、X_2 时高，同时，X_4 的参数未能通过 10%显著性水平下的 t 检验，且参数符号与经济意义不符；

第四步，去掉 X_4，引入 X_6，模型的 $\overline{R}^2$ 虽有所提高，但 X_5 的参数未能通过 10%显著性水平下的 t 检验，且参数符号与经济意义不符。

第五步，去掉 X_5，引入 X_6，模型的 $\overline{R}^2$ 比只有 X_1、X_2 时有所提高，且 X_5 的参数符号与经济意义相符，并通过了 10%显著性水平下的 t 检验。

在第五步所得模型的基础上，再尝试引入单个的 X_3、X_4、X_5，或者引入它们的任意线性组合，均达不到以 X_1、X_2、X_6 为解释变量的回归效果。因此，最终的粮食生产函数应以 $Y=f(X_1,X_2,X_6)$ 为最优，拟合结果如下：

$$\ln\hat{Y}=-1.081+0.761\ln X_1+0.231\ln X_2+0.050\ln X_6$$

§4.2 异 方 差 性

对于模型

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_kX_{ik}+\mu_i,\qquad i=1,2,\cdots,n \tag{4.2.1}$$

同方差性假设为

$$\mathrm{Var}(\mu_i\mid X_1,X_2,\cdots,X_k)=\sigma^2,\qquad i=1,2,\cdots,n$$

如果出现

$$\mathrm{Var}(\mu_i\mid X_{i1},X_{i2},\cdots,X_{ik})=\sigma_i^2,\qquad i=1,2,\cdots,n$$

即**对于不同的样本点，随机干扰项的方差不再是常数，而是互不相同，则认为出现了异方差性**(heteroscedasticity)。

一、异方差的类型

同方差性假定的意义是指，每个 μ_i 围绕其零平均值的方差并不随解释变量 $\boldsymbol{X}_i$ 的变化而变化，不论解释变量是大还是小，每个 μ_i 的方差保持相同，即

$$\sigma_i^2=\text{常数}\neq f(\boldsymbol{X}_i)$$

在异方差的情况下，σ_i^2 已不是常数，它随 $\boldsymbol{X}_i$ 的变化而变化，即

$$\sigma_i^2=f(\boldsymbol{X}_i)$$

异方差一般可归结为三种类型(图 4.2.1)：

（1）单调递增型：σ_i^2 随 X 的增大而增大；

（2）单调递减型：σ_i^2 随 X 的增大而减小；

（3）复杂型：σ_i^2 与 X 的变化呈复杂形式。

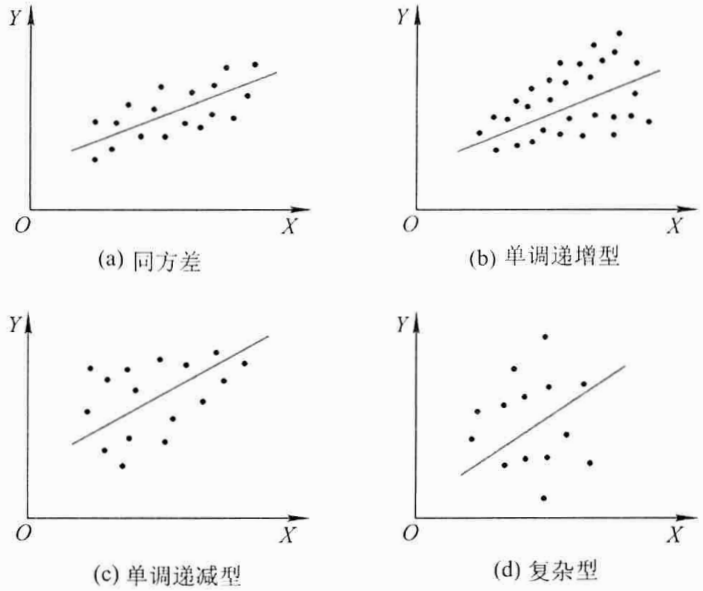


图 4.2.1 异方差的类型

二、实际经济问题中的异方差性

在实际经济问题中，哪些情况容易出现异方差性？下面以三个例子加以说明。

例 4.2.1

以截面数据为样本研究居民家庭的储蓄行为：

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

其中，Y_i 为第 i 个家庭的储蓄额，X_i 为第 i 个家庭的可支配收入。在该模型中，μ_i 项的方差为常数这一假定往往不符合实际情况。对高收入家庭来说，储蓄的差异较大；低收入家庭的储蓄则更有规律性(如为某一特定目的而储蓄)，差异较小。因此 μ_i 的方差往往随 X_i 的增加而增加，呈单调递增型变化。

例 4.2.2

以绝对收入假设为理论假设，以截面数据为样本建立居民消费函数(C)：

$$C_i = \beta_0 + \beta_1 Y_i + \mu_i$$

将居民按照收入 Y 等距离分成 n 组，取每组平均数为样本观测值。我们知道，一般情况下居民收入服从正态分布，所以处于每个收入组中的人数是不等的，处于中等收入组中的人数最多，处于两端收入组中的人数最少。人数多的组的平均数的误差小，人数少的组的平均数的误差大。所以样本观测值的观测误差随着解释变量观测值的不同而不同。如果样本观测值的观测误差构成随机干扰项的主要部分，那么对于不同的样本点，随机干扰项的方差互不相同，出现了异方差性。更进一步分析，在这个例子中，随机干扰项的方差随着解释变量 Y 的观测值的增大而呈 U 形变化，是复杂型的一种。

例 4.2.3

以某一行业的企业为样本建立企业生产函数模型

$$Y_i = \beta_0 A_i^{\beta_1} K_i^{\beta_2} L_i^{\beta_3} e^{\mu_i}$$

产出量(Y)为被解释变量，选择资本(K)、劳动(L)、技术(A)等投入要素为解释变量，那么每个企业所处的外部环境对产出量的影响被包含在随机干扰项中。由于每个企业所处的外部环境对产出量的影响程度不同，造成了随机干扰项的异方差性。这时，随机干扰项的方差并不随某一个解释变量观测值的变化而呈规律性变化，为复杂型的一种。

一般经验告诉我们，对于采用截面数据作样本的计量经济学问题，由于在不同样本点上解释变量以外的其他因素的差异较大，所以往往存在异方差性。

三、异方差性的后果

计量经济学模型一旦出现异方差性，如果仍采用普通最小二乘法估计模型参数，会产生一系列不良的后果。

1. 参数估计量非有效

根据 §3.2 中关于参数估计量的无偏性和有效性的证明过程，可以看出，当计量经济学模型出现异方差性时，其普通最小二乘法参数估计量仍然具有线性性、无偏性，但不具有有效性。因为在有效性证明中利用了

$$E(\boldsymbol{\mu}\boldsymbol{\mu}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$$

而且，在大样本情况下，尽管参数估计量具有一致性，但仍然不具有渐近有效性。

2. 变量的显著性检验失去意义

在 §3.3 关于变量的显著性检验中，构造了 t 统计量，它是建立在随机干扰项共同的方差 σ^2 不变而正确估计了参数方差 $S_{\hat{\beta}_j}$ 的基础之上的。如果出现了异方差性，估计的 $S_{\hat{\beta}_j}$ 出现偏误(偏大或偏小)，t 检验失去意义。其他检验也是如此。

如对一元回归模型

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

的普通最小二乘估计有

$$\hat{\beta}_1 = \beta_1 + \sum k_i \mu_i = \beta_1 + \frac{\sum x_i \mu_i}{\sum x_i^2}$$

可以证明，存在异方差的情况下正确的 $\hat{\beta}_1$ 的方差应为

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2} \tag{4.2.2}$$

而普通最小二乘法仍按下式给出 $\hat{\beta}_1$ 的方差估计

$$\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_i^2} \tag{4.2.3}$$

显然，只有同方差性满足时，(4.2.2)式与(4.2.3)式才会相同，否则普通最小二乘法给出的估计结果就会出现偏误，在有偏误的方差基础上构造的 t 统计量不再服从真实的 t 分布，相应的 t 检验也就失去了意义。

3. 模型的预测失效

一方面，由于上述后果，使得模型不具有良好的统计性质；另一方面，在预测值的置信区间中也包含有参数方差的估计量 $S_{\hat{\beta}_j}$。所以，当模型出现异方差性时，仍然使用普通最小二乘估计量，将导致预测区间偏大或偏小，预测功能失效。

四、异方差性的检验

异方差性的检验方法是计量经济学中一个重要的课题。在一些计量经济学教科书和文献中，可以见到十多种检验方法，如图示检验法、等级相关系数法、戈里瑟(Gleiser)检验、巴特列特检验、G-Q 检验等，很难说哪种方法是最好的。这些方法尽管不同，但存在一个共同的思路。正如上面所指出的，**异方差性，即相对于不同的样本点，也就是相对于不同的解释变量观测值，随机干扰项具有不同的方差，那么检验异方差性，也就是检验随机干扰项的方差与解释变量观测值之间的相关性**。各种检验方法就是在这个思路下发展起来的。

问题在于用什么来表示随机干扰项的方差。一般的处理方法是首先采用普通最小二乘法估计模型，以求得随机干扰项的估计量

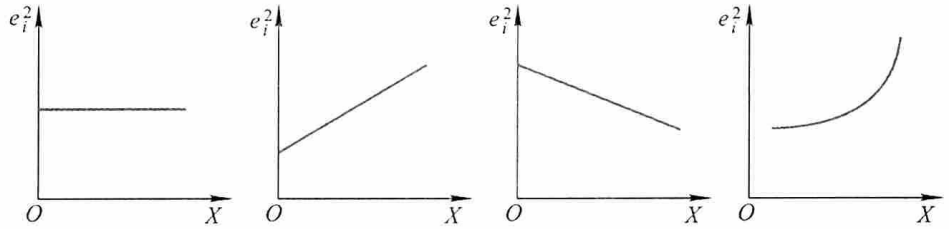
$$e_i=Y_i-(\hat{Y}_i)_{\mathrm{OLS}} \tag{4.2.4}$$

再用 e_i^2 来表示随机干扰项的方差。

下面有选择地介绍三种异方差性的检验方法。

1. 图示检验法

既可用 Y-X 的散点图进行判断，也可用某一个 e_i^2-X 的散点图进行判断。对前者看是否存在明显的散点扩大、缩小或复杂型趋势(即不在一个固定的带形域中)，如图 4.2.1 所示；对后者看是否形成一条斜率为零的直线，如图 4.2.2 所示。



(a)同方差　(b) 单调递增型异方差　(c) 单调递减型异方差　(d) 复杂型异方差

图 4.2.2　不同异方差类型

图示检验法只能进行大概的判断，其他的统计检验方法则更为严格。

2. 布罗施-帕甘（Breusch-Pagan）检验

布罗施-帕甘检验（B-P 检验）是一种较现代的最为常用的异方差检验方法，它具备将所有检验都放在同一框架之中的好处。

对线性模型

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

同方差性意味着

$$\mathrm{Var}(\mu_i \mid X_{i1}, X_{i2}, \cdots, X_{ik}) = \mathrm{Var}(\mu_i \mid \boldsymbol{X}_i) = \sigma^2$$

在随机干扰项具有零条件均值的基本假设下，同方差性也就意味着

$$\mathrm{E}(\mu_i^2 \mid \boldsymbol{X}_i) = \mathrm{E}(\mu_i^2) = \sigma^2$$

即随机干扰项的平方 μ^2 与一个或多个解释变量不相关。异方差的存在就意味着 μ^2 是部分或全部解释变量的某种函数。一个简单的方法就是假定该函数为线性函数：

$$\mu_i^2 = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \cdots + \delta_k X_{ik} + \varepsilon_i$$

则检验同方差性就是检验如下联合假设：

$$\mathrm{H}_0 : \delta_0 = \delta_1 = \delta_2 = \cdots = \delta_k = 0 \tag{4.2.5}$$

由于观测不到真实的 μ_i^2，可用它的 OLS 估计 e_i^2 近似替代，则对原模型随机干扰项同方差性的检验，就是针对辅助回归

$$e_i^2 = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \cdots + \delta_k X_{ik} + \varepsilon_i \tag{4.2.6}$$

检验联合假设（4.2.5）式。这可通过以（4.2.5）式为约束条件的受约束 F 检验或拉格朗日乘数（LM）检验来进行：

$$F = \frac{R_{e^2}^2 / k}{(1 - R_{e^2}^2)/(n-k-1)} \tag{4.2.7}$$

$$LM = n \cdot R_{e^2}^2 \tag{4.2.8}$$

其中，$R_{e^2}^2$ 为辅助回归（4.2.6）式的可决系数。可以证明，（4.2.7）式与（4.2.8）式所构造的 F 统计量与 LM 统计量在大样本下分别渐近地服从 $F(k, n-k-1)$ 分布与 $\chi^2(k)$ 分布（证明超出本教材的范围）。如果计算的 F 值或 LM 值大于给定显著性水平下的临界值，则拒绝 H_0，表明存在异方差性。

3. 怀特（White）检验

怀特检验可以看成是对布罗施-帕甘检验的一种拓展。既然随机干扰项的同方差性意味着 μ^2 与一个或多个解释变量不相关，而异方差性又意味着 μ^2 是部分或全部解释变量的某种函数，因此这种函数可以是非线性的，即可以包含解释变量的平方项以及不同解释变量间的交叉项。下面以两个解释变量的回归模型为例说明怀特检验的基本思想与步骤。

假设回归模型为

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

可先对该模型作 OLS 回归，并得到残差项的平方 e_i^2，然后做如下辅助回归：

$$e_i^2 = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \delta_3 X_{i1}^2 + \delta_4 X_{i2}^2 + \delta_5 X_{i1} X_{i2} + \varepsilon_i$$

要检验的同方差性假设为 $H_0 : \delta_1 = \delta_2 = \cdots = \delta_5 = 0$。

类似于布罗施-帕甘检验，对上述同方差性假设的检验可通过(4.2.7)式的 F 检验或(4.2.8)式的 LM 检验来进行。同样可以证明，在同方差假设下，(4.2.7)式的 F 统计量渐近地服从 F 分布，(4.2.8)式的 LM 统计量渐近地服从 χ^2 分布。如对这里包含两个解释变量及其平方项、交叉项做辅助回归，当得到可决系数 $R_{e^2}^2$ 后，用于检验异方差的 LM 统计量为

$$LM = n \cdot R_{e^2}^2 \sim \chi^2(5)$$

需要注意的是，怀特检验采用的辅助回归，仍是检验 e_i^2 与解释变量可能的组合的显著性，因此，辅助回归方程中还可引入解释变量的更高次方。如果存在异方差性，则表明 e_i^2 确与解释变量的某种组合有显著的相关性，这时往往显示出有较大的可决系数 $R_{e^2}^2$，并且某一参数的 t 检验值较大。当然，在多元回归中，由于辅助回归方程中可能有太多解释变量，从而使自由度减少，有时可去掉交叉项或（和）平方项。

五、异方差的修正

1. 加权最小二乘法(WLS)

如果模型被证明存在异方差性，则需要发展新的方法估计模型，最常用的方法是**加权最小二乘法**(Weighted Least Squares, WLS)。

加权最小二乘法是对原模型加权，使之变成一个新的不存在异方差性的模型，然后采用普通最小二乘法估计其参数。加权的基本思想是：在采用普通最小二乘法时，对较小的残差平方 e_i^2 赋予较大的权数，对较大的 e_i^2 赋予较小的权数，以对残差提供的信息的重要程度作一番校正，提高参数估计的精度。

加权最小二乘法就是对加了权重的残差平方和实施普通最小二乘法：

$$\sum w_i e_i^2 = \sum w_i [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k)]^2 \tag{4.2.9}$$

其中，w_i 为权数。

例如，如果在检验过程中已经知道

$$\mathrm{Var}(\mu_i) = \mathrm{E}(\mu_i^2) = \sigma_i^2 = f(X_{ji})\sigma^2$$

即随机干扰项的方差与解释变量 X_j 之间存在相关性，那么可以用 $\sqrt{f(X_j)}$ 去除原模型，使之变成如下形式的新模型：

$$\frac{1}{\sqrt{f(X_{ji})}} Y_i = \beta_0 \frac{1}{\sqrt{f(X_{ji})}} + \beta_1 \frac{1}{\sqrt{f(X_{ji})}} X_{1i} + \beta_2 \frac{1}{\sqrt{f(X_{ji})}} X_{2i} + \cdots +$$

$$\beta_k \frac{1}{\sqrt{f(X_{ji})}} X_{ki} + \frac{1}{\sqrt{f(X_{ji})}} \mu_i$$

在该模型中，存在

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{\sqrt{f(X_{ji})}}\mu_i\right] &= \left[\frac{1}{\sqrt{f(X_{ji})}}\right]^2 \mathrm{Var}(\mu_i) \\ &= \frac{1}{f(X_{ji})} f(X_{ji})\sigma^2 = \sigma^2\end{aligned}$$

即满足同方差性，于是可以用普通最小二乘法估计其参数，得到关于参数 $\beta_0,\beta_1,\cdots,\beta_k$ 的无偏的、有效的估计量。这就是加权最小二乘法，在这里权数就是 $\frac{1}{\sqrt{f(X_{ji})}}$。

加权最小二乘法具有比普通最小二乘法更普遍的意义，或者说普通最小二乘法只是加权最小二乘法中权恒取 1 时的一种特殊情况。从此意义看，加权最小二乘法也称为**广义最小二乘法**(Generalized Least Squares, GLS)。

实施加权最小二乘法的关键是寻找适当的“权”，或者说是寻找模型中随机干扰项 μ 的方差与解释变量间适当的函数形式。如果发现

$$\mathrm{Var}(\mu_i \mid X_{i1},X_{i2},\cdots,X_{ik}) = \sigma^2 f(X_{i1},X_{i2},\cdots,X_{ik})$$

则加权最小二乘法中的权即为 $1/\sqrt{f(X_{i1},X_{i2},\cdots,X_{ik})}$。但如何寻找 μ 的方差与各 X 间的函数关系呢？帕克检验指出可以进行各种形式的尝试，但下面给出一种相对灵活、有着广泛应用的方法。

假设 μ 的方差具有如下指数函数的形式：

$$\mathrm{Var}(\mu_i \mid X_{i1},\cdots,X_{ik}) = \sigma^2 \exp(\alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_k X_{ik}) \tag{4.2.10}$$

则可等价地写出

$$\mu_i^2 = \sigma^2 \exp(\alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_k X_{ik})\varepsilon_i$$

其中，ε_i 可看成是条件均值为 1 的随机项。如果假设 ε_i 与各 X 独立，进一步有

$$\ln(\mu_i^2) = \delta_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_k X_{ik} + v_i \tag{4.2.11}$$

其中，v_i 为独立于各 X，且条件均值为 0 的随机项。由于(4.2.11)式满足普通最小二乘法的基本假设，当用可观测的值 e_i 代替不可观测的 μ_i 时，用普通最小二乘法估计

$$\ln(e_i^2) = \delta_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_k X_{ik} + v_i \tag{4.2.12}$$

即可得到各 α_j 的无偏、一致且有效的估计 $\hat{\alpha}_j$ $(j=1,2,\cdots,k)$。于是得到 μ 的方差估计：

$$\hat{\sigma}_i^2 = \hat{\mu}_i^2 = \hat{f}_i = \exp(\hat{\delta}_0 + \hat{\alpha}_1 X_{i1} + \hat{\alpha}_2 X_{i2} + \cdots + \hat{\alpha}_k X_{ik}) \tag{4.2.13}$$

从而，估计的权为

$$\hat{w}_i = 1/\hat{\sigma}_i = 1/\sqrt{\hat{f}_i} = 1/\sqrt{\exp(\hat{\delta}_0 + \hat{\alpha}_1 X_{i1} + \hat{\alpha}_2 X_{i2} + \cdots + \hat{\alpha}_k X_{ik})} \tag{4.2.14}$$

最后需指出，(4.2.10)式的指数函数中只列出了各解释变量 X 的水平项，可根据估计

的显著性，对各 X 进行取舍；此外，还可根据需要加入适当的 X 的高次方项。

由于加权最小二乘法中的权，或者说原模型中 μ 的方差与各 X 间适当的函数关系是估计出来的，因此这一广义最小二乘法也称为**可行的广义最小二乘法**(feasible GLS, FGLS)，由广义最小二乘法得到的原模型中的估计量称为**可行的广义最小二乘估计量**，广义最小二乘估计量具有 BLUE 的特征。

2．异方差稳健标准误法

加权最小二乘法的关键是寻找模型中随机扰动项 μ 的方差与解释变量间的适当的函数形式，而这并非一件易事。在有些情况下很难得到正确的 μ 的方差与解释变量的函数关系式，这时，可采用下面介绍的异方差稳健标准误法来消除异方差的存在带来的不良后果。

由于回归模型随机干扰项出现异方差时，普通最小二乘法只是影响到了参数估计量方差或标准差的正确估计，从而无法保证普通最小二乘估计量的有效性，但并不影响估计量的无偏性与一致性。因此，另一种针对异方差的修正的估计方法是：仍采用普通最小二乘估计量，但修正相应的方差。

如何修正普通最小二乘估计量相应的方差呢？怀特 1980 年提出的方法是，用普通最小二乘法估计的残差的平方 e_i^2 作为相应 σ_i^2 的代表。如在一元线性回归中，估计的斜率 $\hat{\beta}_1$ 正确的方差应为

$$\mathrm{Var}(\hat{\beta}_1)=\frac{\sum x_i^2\sigma_i^2}{\left(\sum x_i^2\right)^2} \tag{4.2.2}$$

于是用普通最小二乘法估计的残差的平方 e_i^2 作为相应 σ_i^2 的代表，即用下式作为 $\mathrm{Var}(\hat{\beta}_1)$ 的估计：

$$\frac{\sum x_i^2 e_i^2}{\left(\sum x_i^2\right)^2} \tag{4.2.15}$$

怀特证明了大样本下，(4.2.15)式是(4.2.2)式的一致估计。(4.2.15)式的平方根称为 $\hat{\beta}_1$ 的**异方差稳健标准误**（heteroscedasticity-robust standard error），这种估计方法也称为异方差稳健标准误法。

在存在异方差时，异方差稳健标准误法虽不能得到有效的估计量，但由于可以得到普通最小二乘估计量正确的方差估计，从而使得以估计量方差为基础的各种统计检验不再失效、建立的预测区间也更加可信，因此异方差稳健标准误法就成为在不能较好地实施加权最小二乘法时，消除异方差性不良后果的主要手段。多元回归模型中进行怀特的异方差稳健标准误处理的算法较为复杂，已超出本教材的范围，但任何一款应用软件都有标准的处理程序可直接使用。

六、案例

例 4.2.4

中国农村居民人均消费支出主要由人均纯收入来决定。农村人均纯收入除从事农业经营的收入外，还包括从事其他产业的经营性收入以及工资性收入、财产收入和转移支付收入等。在改革开放的早期，农村居民从事农业经营的收入占到了其纯收入的一个不小的部分，但其他来源的收入可能会在不同的地区差异较大。为了考察从事农业经营的收入和其他收入对中国农村居民消费支出增长的影响，可使用如下双对数模型：

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

其中，Y 表示农村家庭人均消费支出，X_1 表示从事农业经营的纯收入，X_2 表示其他来源的纯收入。表 4.2.1 列出了 2001 年中国内地各地区农村居民家庭人均纯收入及消费支出的相关数据。

表 4.2.1 2001 年中国各地区农村居民家庭人均纯收入与消费支出 单位：元

地区	人均消费支出 Y	从事农业经营的纯收入 X_1	其他来源纯收入 X_2	地区	人均消费支出 Y	从事农业经营的纯收入 X_1	其他来源纯收入 X_2
北 京	3 552.1	579.1	4 446.4	湖 北	1 649.2	1352.0	1 000.1
天 津	2 050.9	1 314.6	2 633.1	湖 南	1 990.3	908.2	1 391.3
河 北	1 429.8	928.8	1 674.8	广 东	2 703.36	1 242.9	2 526.9
山 西	1 221.6	609.8	1 346.2	广 西	1 550.62	1 068.8	875.6
内蒙古	1 554.6	1 492.8	480.5	海 南	1 357.43	1 386.7	839.8
辽 宁	1 786.3	1 254.3	1 303.6	重 庆	1 475.16	883.2	1 088.0
吉 林	1 661.7	1 634.6	547.6	四 川	1 497.52	919.3	1 067.7
黑龙江	1 604.5	1 684.1	596.2	贵 州	1 098.39	764.0	647.8
上 海	4 753.2	652.5	5 218.4	云 南	1 336.25	889.4	644.3
江 苏	2 374.7	1 177.6	2 607.2	西 藏	1 123.71	589.6	814.4
浙 江	3 479.2	985.8	3 596.6	陕 西	1 331.03	614.8	876.0
安 徽	1 412.4	1 013.1	1 006.9	甘 肃	1 127.37	621.6	887.0
福 建	2 503.1	1 053.0	2 327.7	青 海	1 330.45	803.8	753.5
江 西	1 720.0	1 027.8	1 203.8	宁 夏	1 388.79	859.6	963.4
山 东	1 905.0	1 293.0	1 511.6	新 疆	1 350.23	1 300.1	410.3
河 南	1 375.6	1 083.8	1 014.1				

注：从事农业经营的纯收入由从事第一产业的经营总收入与从事第一产业的经营支出之差计算，其他来源的纯收入由总纯收入减去从事农业经营的纯收入后得到。

资料来源：《中国农村住户调查年鉴（2002）》、《中国统计年鉴（2002）》。

OLS 法的估计结果如下：

$$\ln \hat{Y} = 1.655 + 0.317 \ln X_1 + 0.508 \ln X_2$$

$$(1.87) \qquad (3.02) \qquad (10.04)$$

$$R^2 = 0.783\ 1 \qquad \bar{R}^2 = 0.767\ 6 \qquad F = 50.53 \qquad RSS = 0.823\ 1$$

估计结果显示，即使在 1%的显著性水平下，都拒绝了从事农业经营的纯收入与其他来源的收入对农村居民人均消费支出无影响的假设。当然，从参数估计值的大小看，居民消费支出关于其他来源纯收入的弹性更大，意味着即使与从事农业经营有着相同百分比的增长，其他来源的收入对农户人均消费支出的增长有更大的刺激作用。下面对该模型进行异方差性检验。

可以认为不同地区农村人均消费支出的差别主要来源于非农经营收入及工资收入、财产收入等其他收入的差别上，因此，如果存在异方差性，则可能是 X_2 引起的。模型普通最小二乘回归得到的残差平方项 e_i^2 与 $\ln X_2$ 的散点图表明（图 4.2.3），可能存在着递增型异方差性。

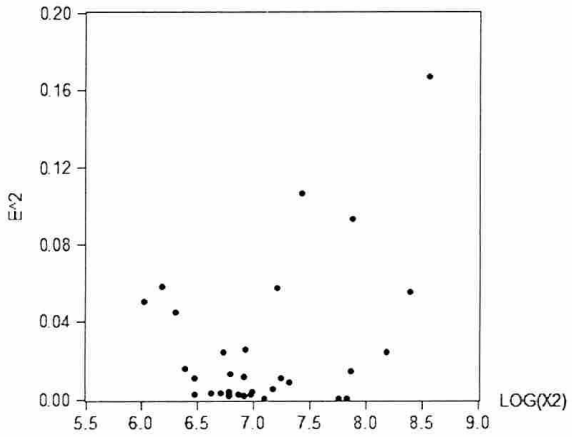


图 4.2.3 异方差性检验图

再进行进一步的统计检验。首先采用布罗施-帕甘（B-P）检验。

将原模型普通最小二乘估计的残差项 e 平方后关于 $\ln X_1$、$\ln X_2$ 做回归：

$$\hat{e}^2 = -0.141 + 0.023\ 6 \ln X_2$$

$$(-1.96) \qquad (2.33)$$

$$R_{e^2}^2 = 0.158\ 0 \qquad F = 5.44$$

由（4.2.7）式与（4.2.8）式计算的 F 统计量与 LM 统计量的值分别为

$$F = \frac{R_{e^2}^2 / k}{(1 - R_{e^2}^2)/(n-k-1)} = \frac{0.158\ 0/1}{(1-0.158\ 0)/29} = 5.44$$

$$LM = n \cdot R_{e^2}^2 = 31 \times 0.158\ 0 = 4.90$$

在5%的显著性水平下，自由度为(1,29)的 F 分布的临界值为 $F_{0.05}$=4.18,自由度为1的 χ^2 分布的临界值为 $\chi^2_{0.05}$=3.84。因此，5%显著性水平下拒绝原模型随机干扰项方差相同的假设。

再采用怀特检验。记 e_i^2 为对原始模型进行普通最小二乘回归得到的残差平方项，将其与 $\ln X_1$、$\ln X_2$ 及其平方项与交叉项作辅助回归，得

$$\hat{e}^2 = -0.17 + 0.10\ln X_1 - 0.06\ln X_2 + 0.01(\ln X_1)^2 + 0.03(\ln X_2)^2 - 0.04\ln X_1 \ln X_2$$

（-1.04）(0.10)　　(-0.12)　　(0.21)　　(1.47)　　(-1.11)

$$R^2_{e^2} = 0.4638 \qquad F = 4.32$$

对应的 F 统计量与 LM 统计量的值分别为

$$F = \frac{0.463\ 8/5}{(1-0.463\ 8)/25} = 4.32 \quad LM = 31\times 0.463\ 8 = 14.38$$

5%显著性水平下，F 分布的临界值为 $F_{0.05}(5,25)$=2.60，χ^2 分布的临界值 $\chi^2_{0.05}(5)$=11.07。因此，拒绝同方差的原假设。

去掉交叉项后的辅助回归结果为

$$\hat{e}^2 = 3.843 - 0.570\ln X_1 - 0.540\ln X_2 + 0.041(\ln X_1)^2 + 0.038(\ln X_2)^2$$

(1.37)(-0.64)　　(-2.76)　　(0.64)　　(2.90)

$$R^2_{e^2} = 0.437\ 4$$

显然，其他收入 $\ln X_2$ 项与它的平方项的参数的 t 检验是显著的，且 F 统计量与 LM 统计量的值分别为 F=5.05，LM=13.56，因此，在5%的显著性水平下，仍是拒绝同方差这一原假设。

下面采用加权最小二乘法对原模型进行回归。

经试算，发现原模型普通最小二乘回归残差平方项的对数 $\ln e_i^2$ 与 $\ln X_2$ 及其平方有显著的回归关系：

$$\ln e^2 = 94.22 - 27.652\ln X_2 + 1.915(\ln X_2)^2$$

(2.56)　(-2.73)　　(2.75)

$$R^2 = 0.2188$$

于是，用 $w_i = 1/\sqrt{\hat{f}_i} = 1/\sqrt{\exp(94.22 - 27.652\ln X_{i2} + 1.915(\ln X_{i2})^2)}$ 作为适当的权，对原模型进行加权最小二乘估计（WLS）得到

$$\ln \hat{Y} = 1.835 + 0.371\ln X_1 + 0.422\ln X_2$$

(2.82)　(4.46)　　(7.55)

$$R^2 = 0.7578 \qquad F = 43.81$$

可以看出，$\ln X_1$ 参数的 t 统计量的值有了显著的改进。同时，$\ln X_1$ 前的参数估计值比普通最小二乘估计略有增加，而 $\ln X_2$ 前的参数估计值比普通最小二乘估计略有减小，

但总体说来，变化不大。这一定程度地表明，原模型的设定是正确的，而且满足了随机干扰项条件零均值的基本假设。

下面我们检验是否经加权的回归模型已不存在异方差性。记经 w_i 加权的回归模型为

$$w\ln Y = \beta_0 w + \beta w\ln X_1 + \beta_2 w\ln X_2 + \mu_*$$

该模型的普通最小二乘回归结果为

$$w\ln \hat{Y} = 1.835w + 0.371w\ln X_1 + 0.421w\ln X_2$$

记该回归模型的残差估计的平方为 $\tilde{e}^2$，将其与 w、$w\ln X_1$、$w\ln X_2$ 作辅助回归，得

$$\tilde{e}^2 = 1.62 - 0.632w - 0.165w\ln X_1 + 0.257w\ln X_2$$

$$R^2 = 0.219\ 4$$

B-P 检验的 LM 统计量 $LM=n\,R^2=31*0.219\ 4=6.80$，该值小于 5%显著性水平下、自由度为 3 的 χ^2 分布的临界值 $\chi^2_{0.05}=7.81$，因此，不拒绝同方差的原假设。

最后，给出异方差稳健标准误法修正的结果：

$$\ln \hat{Y} = 1.655 + 0.317\ln X_1 + 0.508\ln X_2$$

$$(2.18) \qquad (3.03) \qquad (7.43)$$

$$R^2 = 0.783\ 1 \qquad \bar{R}^2 = 0.767\ 6 \qquad F = 50.53 \qquad RSS = 0.823\ 1$$

可以看出，估计的参数与普通最小二乘法的结果相同，只是由于参数的标准差得到了修正，从而使得 t 检验值与普通最小二乘法的结果不同。当然，这里异方差稳健标准误法得到的结论与普通最小二乘法及加权最小二乘法的结论基本一致：2001年，从事农业经营的收入与其他来源的收入都对农村居民人均消费支出有影响，但后者的影响力度更大一些。

§4.3 内生解释变量问题

线性计量经济学模型中有一个重要的假设是随机干扰项的条件零均值假设，如果该假设成立，则称解释变量是外生解释变量或具有严格外生性，否则称为内生解释变量或称解释变量具有内生性。解释变量的严格外生性假设要求任何观测点处的解释变量与任何观测点处的随机干扰项不相关。违背这一基本假设的问题称为内生解释变量问题。

一、内生解释变量问题的提出

对于模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i \tag{4.3.1}$$

其基本假设之一是解释变量 $X_1, X_2, \cdots, X_k$ 是严格外生变量。如果存在一个或多个随机变量是内生解释变量，则称原模型存在**内生解释变量问题**。为讨论方便，假设(4.3.1)式中 X_2 为内生解释变量。对于内生解释变量问题，又分两种不同情况：

1. 内生解释变量与随机干扰项同期无关但异期相关

即

$$\mathrm{Cov}(X_{i2}, \mu_i) = \mathrm{E}(X_{i2}\mu_i) = 0 \tag{4.3.2}$$

$$\mathrm{Cov}(X_{i2}, \mu_{i-s}) = \mathrm{E}(X_{i2}\mu_{i-s}) \neq 0, \qquad s \neq 0 \tag{4.3.3}$$

2. 内生解释变量与随机干扰项同期相关

即

$$\mathrm{Cov}(X_{i2}, \mu_i) = \mathrm{E}(X_{i2}\mu_i) \neq 0 \tag{4.3.4}$$

需要说明的是，对于截面数据模型，第 1 种情况几乎不存在。因此截面数据模型中的内生解释变量问题就主要表现在内生解释变量与随机干扰项的同期相关性上，这时称内生变量为同期内生变量。

二、实际经济问题中的内生解释变量问题

在实际经济问题中，同期内生变量问题往往出现在下面三种情形之中：一是被解释变量与解释变量具有联立因果关系（simultaneous causality）；二是模型设定时遗漏了重要的解释变量，而所遗漏的变量与模型中的一个或多个解释变量具有同期相关性（omitting relevant variables）；三是解释变量存在测量误差（errors-in-variables）。下面通过几个例子对前两种情形予以简单说明。第三种情形留作练习供读者尝试考证。

例 4.3.1

为了考察企业引进外资是否真正提高了企业的效益，以企业资金利润率 LR 为被解释变量，以企业资产中外资所占比例 WR 和其他外生变量 $\boldsymbol{X}$ 为解释变量，建立如下模型

$$LR_i = \alpha_0 + \alpha_1 WR_i + \boldsymbol{\beta}\boldsymbol{X}_i + \mu_i \quad i = 1, 2, \cdots, n \tag{4.3.5}$$

通过对企业引进外资情况的实际考察，不难发现，凡是效益好的企业，比较容易引进外资，凡是效益差的企业，引进外资就很困难。那么，在模型(4.3.5)式中，解释变量 WR 既影响被解释变量 LR，同时它也受被解释变量的影响，而 LR_i 与 μ_i 具有同期相关性，从而导致 WR_i 与 μ_i 具有同期相关性。这就是上述的第一种情形。

例 4.3.2

劳动经济学领域中，劳动者的工资 *wage* 主要由劳动者的受教育程度 *educ*、工作经验 *exper*、个人能力 *abil* 等诸多因素决定：

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp er_i + \beta_3 abil_i + \mu_i$$

但在具体估计该模型时，由于劳动者个人能力的大小很难测度，因此该解释变量无法真正地引入到模型中，于是它只能进入到随机干扰项 μ_i 之中，即实际用于回归的模型为

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp er_i + \mu_i \tag{4.3.6}$$

而个人能力的大小往往与其所受教育程度有着较为密切的联系，这就导致了实际用于回归的模型(4.3.6)中的劳动者个人的受教育程度变量 $educ_i$ 与随机干扰项 μ_i 间出现同期相关性。这就是上述的第二种情形。

例 4.3.3

联立方程模型中的每个结构方程一般存在内生解释变量问题，可以用下面的例子说明。

从全国来看，某类消费品的需求量 Q 主要由该类商品的价格 P 以及居民的收入水平 Y 决定。某些消费品，由于存在流通障碍，不同地区的价格也是不同的，那么，以地区作为样本，一个简单的商品需求函数可表示为

$$Q_i = \beta_0 + \beta_1 P_i + \beta_2 Y_i + \mu_i \quad i = 1,2,\cdots n \tag{4.3.7}$$

其中，Q_i, P_i, Y_i 表示各个地区的需求量、价格和居民收入。经济学基本理论指出，商品价格是由供给与需求的均衡关系决定的，因此商品的需求量是影响价格的重要因素。为了方便推理，可设定如下简单的价格函数：

$$P_i = \alpha_0 + \alpha_1 Q_i + \varepsilon_i \tag{4.3.8}$$

由于商品的需求与均衡价格是在市场上被联合决定的，(4.3.7)式与(4.3.8)式也被称为**联立方程模型**（simultaneous equations model）。因此，即使我们单独估计(4.3.7)式，由于 Q_i 同时影响 P_i，而 Q_i 与 μ_i 具有同期相关性，从而导致 P_i 与 μ_i 具有同期相关性。事实上，通过**联立方程模型**(4.3.7)式与(4.3.8)式，可推导出价格 P_i 的如下式子：

$$P_i = \frac{\alpha_0 + \alpha_1\beta_0}{1-\alpha_1\beta_1} + \frac{\alpha_1\beta_2}{1-\alpha_1\beta_1} Y_i + \frac{\alpha_1\mu_i + \varepsilon_i}{1-\alpha_1\beta_1} \tag{4.3.9}$$

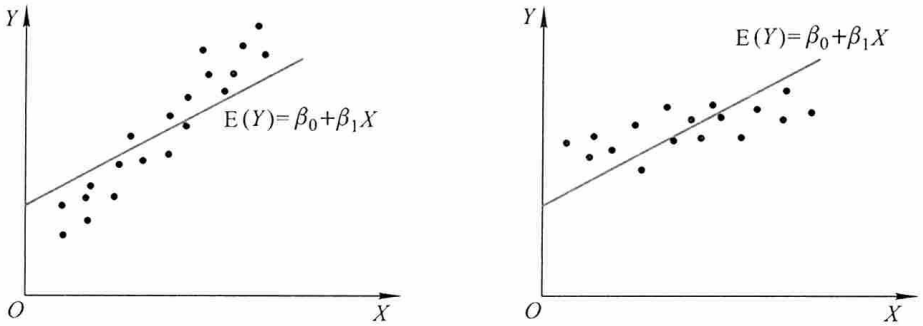
(4.3.9)式显示出 P_i 与 μ_i 具有同期相关性。可见，商品的需求函数(4.3.7)式中价格 P 与需求量 Q 的双向影响关系，决定了 P 是内生解释变量。商品的需求函数(4.3.7)式中的居民收入水平 Y 直接影响着居民对该类商品的购买量，而购买量不直接影响居民的收

入水平，因此 Y 是需求函数(4.3.7)式的外生解释变量。对由(4.3.7)式与(4.3.8)式构成的联立方程模型来说，商品需求 Q 与价格 P 由联立方程模型来决定，称为模型的**内生变量**(endogenous variables)，居民收入水平 Y 不由联立方程模型决定，称为模型的**外生变量**(exogenous variables)。而用外生变量与随机干扰项表示的价格 P 的方程（4.3.9）式称为 P 的**简化式方程**(reduced form equation)。

三、内生解释变量的后果

计量经济学模型一旦出现同期内生解释变量，即与随机干扰项同期相关，如果仍采用普通最小二乘法估计模型参数，将导致产生不良的后果。下面以一元线性回归模型为例进行说明。

从图形上看(图 4.3.1)，如果内生解释变量与随机干扰项正相关，则在抽取样本时，容易出现 X 值较小的点在总体回归线下方，而 X 值较大的点在总体回归线上方的情况，因此，拟合的样本回归线则可能低估(underestimate)截距项，而高估(overestimate)斜率项。反之，如果随机解释变量与随机干扰项负相关，则往往导致拟合的样本回归线高估截距项而低估斜率项。



(a) 正相关　　(b) 负相关

图 4.3.1　随机解释变量与随机干扰项相关图

对一元线性回归模型

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

在第二章曾得到如下最小二乘估计量：

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \beta_1 + \frac{\sum x_i \mu_i}{\sum x_i^2} \tag{4.3.10}$$

如果 X_i 与 μ_i 相关，则容易由（4.3.10）式得到

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{E}(\sum \frac{x_i}{\sum x_i^2} \mu_i) = \beta_1 + \sum \mathrm{E}(k_i \mu_i) \neq \beta_1$$

$$\underset{n\to\infty}{\mathrm{P\lim}}\left(\beta_1+\frac{\sum x_i\mu_i}{\sum x_i^2}\right)=\beta_1+\frac{\mathrm{P\lim}\left(\frac{1}{n}\sum x_i\mu_i\right)}{\mathrm{P\lim}\left(\frac{1}{n}\sum x_i^2\right)}$$
$$=\beta_1+\frac{\mathrm{Cov}(X_i,\mu_i)}{\mathrm{Var}(X_i)}\neq\beta_1$$

即参数估计量是有偏的，同时也是不一致的。

四、工具变量法

模型中出现内生解释变量并且与随机干扰项同期相关时，普通最小二乘估计量是有偏且不一致的。这时，为了得到大样本下的一致估计量，最常用的估计方法是**工具变量**(instrument variable)**法**。

1. 工具变量的选取

工具变量，顾名思义是在模型估计过程中被作为工具使用，以替代与随机干扰项相关的内生解释变量。如果选 Z 作为内生解释变量 X_j 的工具变量，Z 必须满足以下条件：

（1）与所替代的随机解释变量高度相关：$\mathrm{Cov}(Z, X_j)\neq 0$；

（2）与随机干扰项不相关：$\mathrm{Cov}(Z, \mu)=0$；

（3）与模型中其他解释变量不高度相关，以避免出现严重的多重共线性。

2. 工具变量的应用

工具变量法是克服解释变量与随机干扰项同期相关影响的一种参数估计方法，它是矩估计的一种形式。下面仍以一元回归模型为例说明。

记一元线性回归模型如下：

$$Y_i=\beta_0+\beta_1X_i+\mu_i \tag{4.3.11}$$

矩估计是在两个重要的特征 $\mathrm{E}(\mu_i)=0$ 与 $\mathrm{E}(X_i\mu_i)=0$ 下，以之作为总体矩条件，并写出相应的样本矩条件

$$\frac{1}{n}\sum\left(Y_i-\hat\beta_0-\hat\beta_1X_i\right)=0,\quad \frac{1}{n}\sum X_i\left(Y_i-\hat\beta_0-\hat\beta_1X_i\right)=0$$

后得到一个关于参数估计量的正规方程组：

$$\begin{cases}\sum Y_i=n\hat\beta_0+\hat\beta_1\sum X_i\\ \sum X_iY_i=\hat\beta_0\sum X_i+\hat\beta_1\sum X_i^2\end{cases} \tag{4.3.12}$$

求解该正规方程组，得到

$$\hat\beta_1=\frac{\sum x_iy_i}{\sum x_i^2},\qquad \hat\beta_0=\overline{Y}-\hat\beta_1\overline{X}$$

然而，如果 X_i 与 μ_i 相关，则无法得到(4.3.12)式。

如果按照工具变量的选择条件选择 Z 为 X 的**工具变量**，则有总体矩条件

$$\mathrm{E}(\mu_i)=0\,,\qquad \mathrm{Cov}(Z_i,\mu_i)=\mathrm{E}(Z_i\mu_i)=0$$

于是，在一组容量为 n 的样本下，可写出相应的样本矩条件

$$\frac{1}{n}\sum\left(Y_i-\tilde{\beta}_0-\tilde{\beta}_1X_i\right)=0\,,\qquad \frac{1}{n}\sum Z_i\left(Y_i-\tilde{\beta}_0-\tilde{\beta}_1X_i\right)=0$$

并由此得到一个关于参数估计量的正规方程组：

$$\begin{aligned}\sum Y_i&=n\tilde{\beta}_0+\tilde{\beta}_1\sum X_i\\ \sum Z_iY_i&=\tilde{\beta}_0\sum Z_i+\tilde{\beta}_1\sum Z_iX_i\end{aligned}\tag{4.3.13}$$

于是得到

$$\tilde{\beta}_1=\frac{\sum z_iy_i}{\sum z_ix_i}\,,\qquad \tilde{\beta}_0=\overline{Y}-\tilde{\beta}_1\overline{X}\tag{4.3.14}$$

这种求模型参数估计量的方法称为**工具变量法**，$\tilde{\beta}_0,\tilde{\beta}_1$ 称为**工具变量法估计量**(instrumental variable estimator)。

对于多元线性回归模型，其矩阵形式为

$$\boldsymbol{Y}_i=\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{\mu}_i\qquad i=1,2,\cdots,n$$

其中，$\boldsymbol{X}_i=(1,\boldsymbol{X}_{i1},\boldsymbol{X}_{i2},\cdots,\boldsymbol{X}_{ik})$，$\boldsymbol{\beta}=(\beta_0,\beta_1,\cdots,\beta_k)'$。假设 X_2 与随机干扰项相关，用工具变量 Z 替代，于是得到工具变量矩阵

$$\boldsymbol{Z}=\begin{bmatrix}1 & X_{11} & Z_1 & \cdots & X_{1k}\\ 1 & X_{21} & Z_2 & \cdots & X_{2k}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & X_{n1} & Z_n & \cdots & X_{nk}\end{bmatrix}$$

记 $\boldsymbol{Z}_i=(1,X_{i1},Z_i,\cdots,X_{ik})$，则在 $\boldsymbol{Z}_i$ 满足总体矩条件 $\mathrm{E}(\boldsymbol{Z}_i'\mu_i)=\boldsymbol{0}$ 时，对容量为 n 的一组样本，相应的样本矩条件

$$\frac{1}{n}\sum\boldsymbol{Z}_i'(Y_i-\boldsymbol{X}_i\tilde{\boldsymbol{\beta}})=\boldsymbol{0}\tag{4.3.15}$$

或

$$\frac{1}{n}\boldsymbol{Z}'(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}})=\boldsymbol{0}\tag{4.3.16}$$

于是，参数估计量为

$$\tilde{\boldsymbol{\beta}}=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}\tag{4.3.17}$$

需要注意，通常情况下，工具变量矩阵 $\boldsymbol{Z}$ 由工具变量及原模型中的外生解释变量组成。这时，对于没有选择另外的变量作为工具变量的解释变量，可以认为用自身作为工具变量。

3. 工具变量法估计量是一致估计量

一元回归中，用工具变量法所求的参数估计量 $\tilde{\beta}_1$ 与总体参数真值 β_1 之间的关系为

$$\tilde{\beta}_1=\frac{\sum z_i y_i}{\sum z_i x_i}=\frac{\sum z_i Y_i}{\sum z_i x_i}=\frac{\sum z_i(\beta_0+\beta_1 X_i+\mu_i)}{\sum z_i x_i}$$
$$=\frac{\beta_1\sum z_i x_i}{\sum z_i x_i}+\frac{\sum z_i\mu_i}{\sum z_i x_i}=\beta_1+\frac{\sum z_i\mu_i}{\sum z_i x_i}$$

两边取概率极限得

$$\operatorname{Plim}(\tilde{\beta}_1)=\beta_1+\frac{\operatorname{Plim}\left(\frac{1}{n}\sum z_i\mu_i\right)}{\operatorname{Plim}\left(\frac{1}{n}\sum z_i x_i\right)}$$

如果工具变量 Z 选取恰当，即有

$$\operatorname{Plim}\left(\frac{1}{n}\sum z_i\mu_i\right)=\operatorname{Cov}(Z_i,\mu_i)=0$$
$$\operatorname{Plim}\left(\frac{1}{n}\sum z_i x_i\right)=\operatorname{Cov}(Z_i,X_i)\neq 0$$

因此，
$$\operatorname{Plim}(\tilde{\beta}_1)=\beta_1$$

尽管工具变量法估计量在大样本下具有一致性，但容易验证在小样本下，由于

$$\mathrm{E}\left(\frac{1}{\sum z_i x_i}\sum z_i\mu_i\right)\neq\mathrm{E}\left(\frac{1}{\sum z_i x_i}\right)\mathrm{E}\left(\sum z_i\mu_i\right)=0$$

工具变量法估计量仍是有偏的。

4. 两阶段最小二乘法：多个工具变量的情形

对工具变量法，经常产生一种误解，以为采用工具变量法是将原模型中的随机解释变量换成工具变量，即改变了原来的模型。实际上，从上面一元线性回归模型的例子中可以看出，**工具变量法并没有改变原模型，只是在原模型的参数估计过程中用工具变量的信息“替代”了内生解释变量的信息**。或者说，上述工具变量法估计过程可等价地分解成下面两个阶段的普通最小二乘回归：

第一阶段，用普通最小二乘法进行 X 关于工具变量 Z 的回归：

$$\hat{X}_i=\hat{\alpha}_0+\hat{\alpha}_1 Z_i$$

第二阶段，以第一步得到的 $\hat{X}_i$ 为解释变量，进行如下普通最小二乘回归：

$$Y_i=\beta_0+\beta_1\hat{X}_i+\mu_i \tag{4.3.18}$$

容易验证，(4.3.18)式中的参数 $\tilde{\beta}_1$ 与(4.3.14)式相同（留作练习）。(4.3.18)式表明，工具变量法仍是 Y 对 X 的回归，而不是对 Z 的回归。这里采用两个阶段的普通最小二乘法来估计模型参数，也称为**两阶段最小二乘法**(two stage least squares, 2SLS)。

当对一个内生解释变量寻找到 1 个工具变量时，工具变量法，或上述两阶段最小二乘法（2SLS）可以得到参数的一致估计量。而当对一个内生解释变量寻找到多个工具变量，且不想损失这些工具变量提供的信息时，仍然可以采用两阶段最小二乘法（2SLS）

来得到参数的一致估计。在多元线性回归中，其基本做法与上述一元回归两个阶段的OLS法相同，只不过第一阶段是将内生变量关于所有工具变量以及模型中已有的外生变量进行OLS回归。下面以二元模型为例进行说明。

对于二元线性回归模型：

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

式中，假设X为同期内生变量，Z为外生变量。如果对内生变量X寻找到了两个工具变量Z_1、Z_2，则两阶段最小二乘估计过程为：

第一阶段，做内生变量X关于工具变量Z_1、Z_2及模型中的外生变量Z的OLS回归，并记录X的拟合值：

$$\hat{X}_i = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{i1} + \hat{\alpha}_2 Z_{i2} + \hat{\alpha}_3 Z_i$$

第二阶段，以第一阶段得到的$\hat{X}_i$替代原模型中的X_i，进行如下OLS回归：

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 Z_i + \mu_i$$

如果一个内生解释变量可以找到多个互相独立的工具变量，人们希望充分利用这些工具变量的信息，就形成了**广义矩方法**(Generalized Method of Moments, GMM)。在GMM中，矩条件大于待估参数的数量，于是如何求解成为它的核心问题。GMM是近20年计量经济学理论方法发展的重要方向之一。两阶段最小二乘法（2SLS）是GMM的一种特殊的估计方法，而当一个内生变量只寻找到了一个工具变量时所采用的工具变量法（IV），则是两阶段最小二乘法的一个特例。同样地，如果所有解释变量都是外生变量，则OLS法也可看成是工具变量法的特例。

最后需要说明的是，两阶段最小二乘法可以完成模型的工具变量估计，但不能简单地以第二阶段的OLS回归得到相关解释变量的标准差以及可决系数R^2。正确的计算过程较为复杂，已超过本教材的范围，但任何一个计量经济学软件都设计了计算程序，因此，直接通过软件可以得到解释变量正确的标准差及R^2。同样地，如果在两阶段最小二乘估计中怀疑或检验出存在着异方差性，还可以通过White的方法获得异方差稳健标准误。虽然手工操作相当复杂，但大多数计量经济学软件都拥有此功能。

五、内生性检验与过度识别约束检验

1. 解释变量的内生性检验

回归模型的基本假设要求随机解释变量与模型的随机干扰项不存在同期相关性，即随机解释变量至少是同期外生变量。那么如何判断所设定的模型中各解释变量是同期外生变量呢？经济学的相关知识能帮助我们做出一些基本的判断，如由于消费惯性的存在，可以认为前期的消费支出对当期的消费支出有着一定的影响，但不能反过来说当期的消费支出对前期的消费支出有影响。除此之外，豪斯曼(Hausman)从计量技术上给出了一个检验随机解释变量是否是同期外生变量的方法。

假设有如下设定的二元线性回归模型

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{i1} + \mu_i \tag{4.3.19}$$

其中，X 与 Z_1 是随机解释变量，而且明确知道 Z_1 是外生变量，但怀疑 X 是同期内生变量。如何检验 X 是否具有内生性呢？豪斯曼提出的检验的基本思想是：如果 X 是内生变量，则需寻找一外生变量 Z_2 作为工具变量并对(4.3.19)式进行工具变量法估计，将工具变量法的估计结果与对(4.3.19)式直接进行普通最小二乘法的估计结果对比，看差异是否显著。如果两者有显著的差异，则表明 X 是内生变量。由于工具变量法等价于两阶段最小二乘法，因此该检验法可具体如下进行。

第一步，将怀疑是内生变量的 X 关于外生变量 Z_1，Z_2 作普通最小二乘估计：

$$X_i = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + v_i \tag{4.3.20}$$

得到残差项 $\hat{v}$。这里假设随机干扰项 v 满足所有线性回归基本假设。该普通最小二乘回归的目的是为了得到残差项 $\hat{v}$，因此可认为是辅助回归。

第二步，将第一步得到的残差项 $\hat{v}$ 加入到原模型后，再进行普通最小二乘估计：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{i1} + \delta \hat{v}_i + \varepsilon_i \tag{4.3.21}$$

仍假设随机干扰项 ε 满足所有线性回归基本假设，并与 v 不同期相关。如果 $\hat{v}$ 前的参数 δ 显著为零，则表明(4.3.20)式的随机干扰项 v 与 Y 同期无关，进而与原模型(4.3.19)式的随机干扰项 μ 同期无关，而 Z_1，Z_2 是外生变量，它们肯定与 μ 同期无关，由(4.3.20)式知 X 与 μ 同期无关。因此，(4.3.21)式的普通最小二乘回归不拒绝 $\delta=0$ 的假设，则可判断原模型(4.3.19)式中的解释变量 X 是同期外生变量，否则判断 X 是同期内生变量。

最后，有三点需要说明。

第一，由(4.3.20)式知，判断 X 与 μ 是否同期相关，等价于判断 v 与 μ 是否同期相关；而对(4.3.21)式的普通最小二乘回归，等价于对下式进行普通最小二乘回归：

$$\mu_i = \delta v_i + \varepsilon_i$$

第二，如果一个被怀疑的内生解释变量有多个工具变量，则在第一步中需将该解释变量关于所有的工具变量及原模型中已有的外生变量进行 OLS 回归。

第三，如果原回归模型有多个随机解释变量被怀疑与随机干扰项同期相关，则需寻找多个外生变量，并将每个所怀疑的解释变量与所有外生变量(包括原模型中已有的外生变量)作普通最小二乘回归，取得各自的残差项，并将它们全部引入到原模型中再进行普通最小二乘估计，通过 t 检验或多种情形的受约束 F 检验，可判断哪些解释变量确实是内生变量。

2. 过度识别约束检验

工具变量法的核心是要寻找到适当的工具变量，它应与原模型的随机干扰项不同期相关。当一个内生解释变量有多于一个的工具变量时，就可以对该组工具变量的外生性进行检验，这就是所谓的**过度识别约束检验**(overidentifying restrictions test)。

过度识别约束检验的基本思路是：如果寻找到的工具变量具有外生性，则它们应与原模型中的随机干扰项不同期相关。因此，只需对原模型进行两阶段最小二乘回归（2SLS），将记录的残差项再关于所有工具变量及原模型中的外生变量进行 OLS 回归，并对该回归中的所有工具变量前的参数都为零的假设进行联合性 F 检验。可以证明，在所有工具变量都是外生的假设下，当样本容量趋于无穷大时，F 统计量逐渐接近于精确的 F 分布，同时，样本容量与该回归的可决系数的乘积 nR^2 的渐近分布为 χ^2 分布(自由度为额外工具变量的个数)。于是，可以通过 F 统计量的值或 nR^2 的值与相关分布临界值的比较，来判断是否拒绝所有工具变量都具有外生性的假设。由于 nR^2 的计算较为方便，人们大多通过 χ^2 统计量来进行过度识别约束检验，nR^2 也称为 ***J* 统计量**（J-statistic）。下面以二元线性回归模型为例进行简单总结：

对有一个内生变量 X 与一个外生变量 Z 的二元线性回归模型：

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

如果对内生变量 X 寻找到了两个工具变量 Z_1、Z_2，记两阶段最小二乘回归（2SLS）的参数估计为 $\tilde{\beta}_0$、$\tilde{\beta}_1$、$\tilde{\beta}_2$，残差为 $\tilde{\mu}_i$：

$$\tilde{\mu}_i = Y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 X_i + \tilde{\beta}_2 Z_i)$$

将 $\tilde{\mu}_i$ 关于所有工具变量 Z_1、Z_2 及原模型中的外生变量 Z 做如下辅助回归：

$$\tilde{\mu}_i = \delta_0 + \delta_1 Z_{i1} + \delta_2 Z_{i2} + \delta_3 Z_i + \varepsilon_i$$

记该辅助回归的可决系数为 R^2，则在所有工具变量为外生变量的假设下（大样本下）

$$J = nR^2 \sim \chi^2(1)$$

这里，一个内生变量对应着 2 个工具变量，额外的工具变量个数为 1。因此，当 nR^2 的值大于给定显著性水平下自由度为 1 的 χ^2 分布的临界值时，拒绝 Z_1、Z_2 同时为外生变量的假设，意味着它们中至少有一个不是外生的。

需要指出的是，对工具变量的外生性检验就是对总体矩条件 $\mathrm{E}(\boldsymbol{Z}_i'\mu_i)=\mathbf{0}$ 的检验，当工具变量的个数恰好等于内生变量的个数时，无论总体矩条件是否成立，对应的样本矩

$$\frac{1}{n}\sum \boldsymbol{Z}_i'(Y_i - \boldsymbol{X}_i\tilde{\boldsymbol{\beta}}) = \mathbf{0}$$

总有唯一的解。这意味着当工具变量的个数恰好等于内生变量的个数时，工具变量的外生性是无法检验的。

六、案例

例 4.3.4

利用美国各州的数据为样本观测值，建立香烟需求模型。根据商品需求函数理论，对香烟的人均消费需求 Q 与居民的收入水平 Y 及香烟的销售价格 P 有关，即香烟的需求模型可写为

$$\ln Q = \beta_0 + \beta_1 \ln Y + \beta_2 \ln P + \mu \tag{4.3.22}$$

然而，如果考虑到在市场均衡时香烟的销售价格也同时受香烟的需求量的影响，则 Q 与 P 之间存在着双向因果关系。因此，由于 P 的内生性将导致对（4.3.22）式的 OLS 回归带来有偏且不一致的估计，这时需要寻找适当的工具变量来对（4.3.22）式进行工具变量或两阶段最小二乘估计。考虑到香烟价格的组成部分更多地是政府对烟草的课税，而香烟的人均消费量本身不会直接影响政府对香烟的课税政策，因此香烟的消费税可能是一个适当的工具变量。表 4.3.1 给出了 1995 年美国 48 个州的人均香烟消费量 Q、香烟的平均价格 P、每个州的人均收入水平 Y 以及香烟的平均消费税 Tax，同时大多数州还对香烟征收了特别消费税 $Taxs$，表中也同时将其列出。表中的香烟平均价格、税以及人均收入都经过了居民消费价格指数的调整。

表 4.3.1 1995 年美国 48 个州人均香烟消费、收入与对香烟的课税

地区	人均香烟消费量 Q（盒）	人均收入 Y（千美元）	香烟平均销售价格 P（美分/盒）	香烟平均消费税 Tax（美分/盒）	香烟平均特别消费税 $Taxs$（美分/盒）
AL	101.09	12.92	103.92	26.57	0.92
AR	111.04	12.17	115.19	36.42	5.49
AZ	71.95	13.54	130.32	42.87	6.21
CA	56.86	16.07	138.13	40.03	9.04
CO	82.58	16.32	109.81	28.87	0.00
CT	79.47	20.96	143.23	48.56	8.11
DE	124.47	16.66	108.66	31.50	0.00
FL	93.07	15.43	123.17	37.99	6.97
GA	97.47	14.59	102.74	23.62	0.94
IA	92.40	13.90	125.26	39.37	5.96
ID	74.85	12.88	117.87	34.12	5.61
IL	83.27	16.83	130.23	44.62	7.37
IN	134.26	14.33	101.40	25.92	4.83
KS	88.75	14.36	114.97	31.50	5.47
KY	172.65	12.61	95.79	17.72	5.42
LA	105.18	12.82	110.10	28.87	4.23
MA	76.62	18.41	142.46	49.21	6.78
MD	77.47	17.65	122.07	39.37	5.81
ME	102.47	13.28	129.42	40.03	7.33
MI	81.39	15.73	158.04	64.96	8.95
MN	82.95	16.13	144.59	47.24	9.46
MO	122.45	14.50	103.17	26.90	0.91
MS	105.58	11.28	111.04	27.56	7.26
MT	87.16	12.31	102.50	27.56	0.00
NC	121.54	14.40	98.42	19.03	3.79
ND	79.81	12.52	126.15	44.62	7.14
NE	87.27	14.56	119.54	38.06	5.69
NH	156.34	16.41	109.34	32.15	0.00

续表

地区	人均香烟消费量 Q（盒）	人均收入 Y（千美元）	香烟平均销售价格 P（美分/盒）	香烟平均消费税 Tax（美分/盒）	香烟平均特别消费税 $Taxs$（美分/盒）
NJ	80.37	19.21	133.26	41.99	7.54
NM	64.67	12.37	115.58	29.53	5.50
NV	93.53	16.93	135.56	38.71	8.87
NY	70.82	18.19	145.58	52.49	5.60
OH	111.38	15.02	108.85	31.50	5.18
OK	108.68	12.73	111.64	30.84	5.32
OR	92.16	14.87	124.87	40.68	0.00
PA	95.64	15.58	115.59	36.09	6.54
RI	92.60	15.78	147.28	52.49	9.64
SC	108.08	12.78	100.27	20.34	4.77
SD	97.22	13.02	110.26	30.84	4.24
TN	122.32	14.30	109.62	24.28	8.12
TX	73.08	14.12	130.05	42.65	7.36
UT	49.27	12.37	118.75	33.14	5.65
VA	105.39	16.05	109.36	17.39	5.21
VT	122.33	14.02	115.25	28.87	5.49
WA	65.53	15.67	156.90	52.82	10.26
WI	92.47	14.81	132.14	40.68	6.29
WV	115.57	11.75	109.26	26.90	6.18
WY	112.24	14.12	104.03	23.62	0.00

资料来源：根据 Introduction to Econometrics (2nd edition)整理。

首先，对（4.3.22）式进行普通最小二乘回归如下：

$$\ln\hat{Q} = 10.341 + 0.344\ln Y - 1.406\ln P$$

$$(10.11) \quad (1.46) \qquad (-5.60)$$

$$R^2 = 0.432\ 8 \qquad \overline{R}^2 = 0.407\ 5 \qquad F = 17.17$$

可见，价格确实是影响人均香烟消费的重要因素。但正是因为价格与消费需求可能存在的双向因果关系，使得模型中 P 具有内生性，从而普通最小二乘估计有偏且不一致。

其次，用香烟消费税 Tax 为工具变量，重新对（4.3.22）式进行工具变量法回归，得到如下结果：

$$\ln\hat{Q} = 10.023 + 0.299\ln Y - 1.315\ln P$$

$$(9.27) \quad (1.24) \qquad (-4.85)$$

$$R^2 = 0.431\ 1 \qquad \overline{R}^2 = 0.405\ 8 \qquad F = 13.27$$

可以看出，工具变量法估计得到的人均香烟消费关于价格的弹性要低于普通最小二乘估计的结果。

由于许多州都对香烟有额外的特别消费税 Taxs，它也可以作为价格 P 的一个工具变量。用 *Tax* 及 *Taxs* 两个工具变量进行的两阶段最小二乘估计结果如下：

$$\ln\hat{Q} = 9.894 + 0.281\ln Y - 1.277\ln P$$

$$(9.35) \quad (1.18) \quad (-4.85)$$

$$R^2 = 0.429\ 4 \qquad \overline{R}^2 = 0.404\ 1 \qquad F = 13.28$$

可见，估计的人均香烟消费关于价格的弹性进一步下降了。

下面再进行过度识别约束检验，以检验 *Tax*、*Taxs* 是否是外生变量。

根据过度识别约束检验的过程，用 *Tax* 及 *Taxs* 两个工具变量对原模型进行两阶段最小二乘估计后，记录残差项 $\tilde{\mu}$：

$$\tilde{\mu} = \ln Q - (9.894 + 0.281\ln Y - 1.277\ln P)$$

再做 $\tilde{\mu}$ 关于工具变量 *Tax*、*Taxs* 以及原模型外生变量 $\ln Y$ 的普通最小二乘回归，该辅助回归的结果为：

$$\hat{\tilde{\mu}} = -0.066 - 0.002Tax + 0.006Taxs + 0.032\ln Y$$

$$(-0.11) \ (-0.41) \quad (0.54) \qquad (0.14)$$

$$R^2 = 0.007\ 0 \qquad F = 0.103$$

首先，从 F 统计量看，即使在 10%的显著性水平下（这时临界值 $F_{0.1}(3,44) = 2.21$），也不拒绝 *Tax*、*Taxs* 及 $\ln Y$ 前参数都为零的假设；其次，由于

$$nR^2 = 48 \times 0.0070 = 0.336$$

在有一个额外工具变量的情况下，5%显著性水平下、自由度为 1 的 χ^2 分布的临界值为 $\chi^2(1) = 3.84$，可见 χ^2 检验也不拒绝 *Tax*、*Taxs* 作为工具变量的外生性假设。

下面再用豪斯曼检验来判定香烟价格 P 是否确实是内生变量。

可选择香烟消费税 *Tax* 与香烟特别消费税 *Taxs* 作为 $\ln P$ 的工具变量，将 $\ln P$ 关于 $\ln Y$、*Tax*、*Taxs* 进行 OLS 估计得：

$$\ln\hat{P} = 4.103 + 0.108\ln Y + 0.009Tax + 0.011Taxs$$

记录残差序列 $\hat{v}$，并将其加入原模型后进行普通最小二乘估计得：

$$\ln\hat{Q} = 9.894 + 0.281\ln Y - 1.277\ln P - 1.565\hat{v}$$

$$(9.59) \quad (1.21) \qquad (-4.98) \quad (-1.75)$$

在10%与5%的显著性水平下，t 分布的临界值分别为 $t_{0.05}(44) = 1.68$、$t_{0.025}(44) = 2.02$，因此，在10%的显著性水平下，拒绝 $\hat{v}$ 前的参数为0的假设，可判断香烟价格是内生变量；但在5%的显著性水平下，不拒绝 $\hat{v}$ 前的参数为0的假设，判断香烟价格不是内生变量。如果模型中不存在内生解释变量，则应采用普通最小二乘法进行模型估计，因为普通最小二乘估计量是最佳线性无偏估计量。

§4.4 模型设定偏误问题

到目前为止，经典计量经济学模型的回归分析，都集中在对模型的估计和对解释变量、随机干扰项经典假设的检验方面，而较少关注模型的具体设定形式。在线性回归模型的经典假设中，还有一个重要的假设就是模型设定是正确的。然而，如果我们设定了一个“错误的”或者说是“有偏误的”模型，即使其他经典假设都满足，得到的估计结果也会与“实际”有偏误，这种偏误称为模型设定偏误。只有模型设定“正确”，并且通过了所有经典假设检验，才能认为得到了一个较为“满意”的模型估计结果，从而可以进一步用于经济分析及预测。

一、模型设定偏误的类型

模型设定偏误主要有两大类：一类是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；另一类是关于模型函数形式选取的偏误。

1．相关变量的遗漏(omitting relevant variable)

在建立模型时，由于人们认识上的偏差、理论分析的缺陷，或者是有关统计数据的限制，可能有意或无意地忽略了某些重要变量。例如，如果“正确”的模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu \tag{4.4.1}$$

而我们将模型设定为

$$Y = \alpha_0 + \alpha_1 X_1 + v \tag{4.4.2}$$

也就是说，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**。

2．无关变量的误选(including irrelevant variable)

无关变量的误选是指在设定模型时，包括了无关的解释变量。例如，如果

(4.4.1)式仍为“真”，但将模型设定为

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + v \tag{4.4.3}$$

也就是说，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)

错误的函数形式是指在设定模型时，选取了不正确的函数形式。最常见的就是当“真实”的函数形式为非线性时，却选取了线性的函数形式。例如，如果“真实”的回归函数为

$$Y = AX_1^{\beta_1} X_2^{\beta_2} e^{\mu} \tag{4.4.4}$$

但却将模型设定为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \tag{4.4.5}$$

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。这种偏差的性质和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(4.4.1)式，而我们却对(4.4.2)式进行回归，X_1的参数估计为

$$\hat{\alpha}_1 = \frac{\sum x_{i1} y_i}{\sum x_{i1}^2} \tag{4.4.6}$$

将正确模型(4.4.1)式的离差形式

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \mu_i - \bar{\mu}$$

代入(4.4.6)式得

$$\begin{aligned}\hat{\alpha}_1 &= \frac{\sum x_{i1} y_i}{\sum x_{i1}^2} = \frac{\sum x_{i1}(\beta_1 x_{i1} + \beta_2 x_{i2} + \mu_i - \bar{\mu})}{\sum x_{i1}^2} \\ &= \beta_1 + \beta_2 \frac{\sum x_{i1} x_{i2}}{\sum x_{i1}^2} + \frac{\sum x_{i1}(\mu_i - \bar{\mu})}{\sum x_{i1}^2} \\ &= \beta_1 + \beta_2 \frac{\sum x_{i1} x_{i2}}{\sum x_{i1}^2} + \sum \frac{x_{i1}}{\sum x_{i1}^2} \mu_i \end{aligned} \tag{4.4.7}$$

（1）如果漏掉的X_2与X_1相关，则(4.4.7)式中的第二项在小样本下求条件期望与大样本下求概率极限都不会为零（第三项的条件期望与概率极限都等于零），从而使得普通最小二乘估计量在小样本下是有偏的，在大样本下也是非一致的。

事实上，在正确模型为(4.4.1)式的情况下对(4.4.2)式进行回归，则(4.4.2)式的随机干扰项就包括了X_2，即$v = \beta_2 X_2 + \mu$，从而X_1与v是同期相关的。因此，如果$\beta_2 > 0$，且X_2与X_1正相关，则X_1与v正相关，导致X_1的参数被高估，而常数项被低估。事实上，§4.3 节已讨论到了遗漏变量而导致模型中该解释变量的内生性问题。

（2）如果X_2与X_1给定的样本下不相关（正交），则由(4.4.7)式易知α_1的估计满足无偏性与一致性，但这时α_0的估计却是有偏且非一致的（留作练习）。

（3）随机干扰项的方差估计$\hat{\sigma}^2$也是有偏的。在同样的样本下，(4.4.2)式给出的样本残差与(4.4.1)式给出的样本残差也不相同，因此，由两组样本残差估计的随机干扰项的方差也会不同。如果(4.4.1)式是正确的估计，(4.4.2)式的估计则是有偏误的。

（4）$\hat{\alpha}_1$的方差是正确估计量$\hat{\beta}_1$的方差的有偏估计。由(4.4.2)式与(4.4.1)式估计的X_1的参数的方差分别为

$$\mathrm{Var}(\hat{\alpha}_1)=\frac{\sigma^2}{\sum x_{i1}^2} \tag{4.4.8}$$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1)&=\sigma^2\frac{\sum x_{i2}^2}{\sum x_{i1}^2\sum x_{i2}^2-(\sum x_{i1}x_{i2})^2}\\&=\frac{\sigma^2}{\sum x_{i1}^2(1-r_{x_1x_2}^2)}\end{aligned} \tag{4.4.9}$$

其中，$r_{x_1x_2}^2$ 为 X_1 与 X_2 的相关系数的平方。如果 X_2 与 X_1 相关，显然有 $\mathrm{Var}(\hat{\alpha}_1)\neq\mathrm{Var}(\hat{\beta}_1)$，即使 X_2 与 X_1 不相关，由于由(4.4.2)式与(4.4.1)式估计的随机干扰项的方差不同，估计的 X_1 的参数的方差也会不同。

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)。

设正确的模型为(4.4.2)式，而我们却对(4.4.1)式进行估计。对于(4.4.1)式，如果 $\beta_2=0$，则与(4.4.2)式相同，因此，可将(4.4.1)式视为以 $\beta_2=0$ 为约束的正确模型(4.4.2)式的特殊形式。由于所有的经典假设都满足，因此对(4.4.1)式进行普通最小二乘估计，可得到无偏且一致的估计量。由于 $\beta_2=0$，因此，$\mathrm{E}(\hat{\beta}_2)=0$。

尽管在包含无关变量的情况下，普通最小二乘估计量是无偏的，但却不具有最小方差性。事实上，对 X_1 前的参数的方差而言，正确模型(4.4.2)式与错误模型(4.4.1)式估计的方差分别由(4.4.8)式与(4.4.9)式给出。显然，当 X_1 与 X_2 完全线性无关时，两模型参数估计的方差相同，否则，包含无关变量的模型参数的方差大于正确模型参数估计的方差，即 $\mathrm{Var}(\hat{\beta}_1)>\mathrm{Var}(\hat{\alpha}_1)$。

由此可见，在多选无关解释变量的情形下，普通最小二乘估计量仍是无偏且一致的，随机干扰项的方差 σ^2 也能被正确估计，但普通最小二乘估计量却往往是无效的。也就是说，包含无关变量的偏误主要表现为“错误”模型的普通最小二乘估计量的方差一般会大于“正确”模型相应参数估计量的方差。

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(4.4.4)式给出的幂函数的形式，而在模型估计时设定的模型却为(4.4.5)式所示的线性形式。显然，模型(4.4.4)式中的参数 β_1 为弹性，而按(4.4.5)式估计出的 $\hat{\beta}_1$ 却是对一个单位 X_1 变化带来的 Y 相应变化的测量。两者具有完全不同的经济意义，估计结果一般也是不相同的。

三、模型设定偏误的检验

一旦模型设定有偏误，普通最小二乘估计可能带来不良后果。因此，对模型的设定

偏误进行检验就显得非常重要。

1．检验是否含有无关变量

对于无关变量的误选检验比较简单，可用统计检验中的 t 检验与 F 检验完成。检验的基本思想是，如果模型中误选了无关变量，则其系数的真值应为零。因此，只需对无关变量系数的显著性进行检验即可。例如，对所选定的一个 k 元回归模型

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

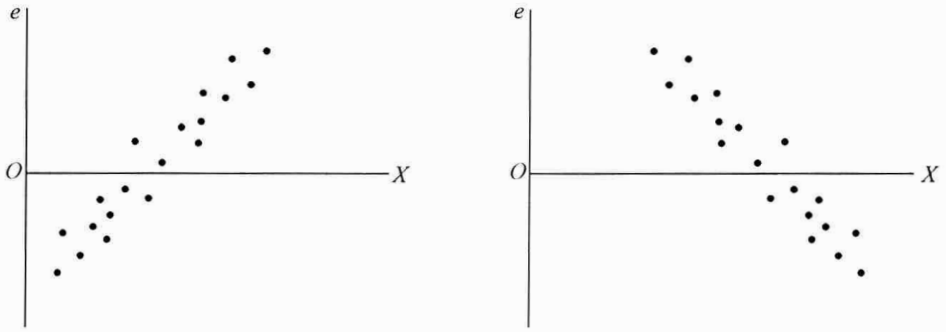
如果我们怀疑其中第 j 个变量是与 Y 无关的变量，只需用通常的 t 检验去检验 β_j 的显著性即可。而要检验 X_2 与 X_3 是否同时应包括在模型中来，只需检验联合假设 H_0：$\beta_2=\beta_3=0$ 即可，第三章已介绍了适用的 F 检验。

2．检验是否有相关变量的遗漏或函数形式设定偏误

在上面所列出的三种模型设定偏误中，遗漏相关变量与设定错误的函数形式的后果比多选不相关变量的情形要严重得多。不仅估计量有偏且不一致，而且随机干扰项的方差也往往被高估，从而使通常的推断程序变得无效，甚至参数的经济意义也可能不合理。而在多选不相关变量的情形下，后果仅是效率的损失。下面，我们着重介绍遗漏相关变量与设定错误的函数形式这两种模型设定偏误的检验。

（1）残差图示法。对所设定的模型进行普通最小二乘回归，得到估计的残差序列 e_t，做出 e_t 与某解释变量 X 的散点图，从图形考察估计的残差序列 e_t 是否有规律地变动，来判断模型设定时是否遗漏了重要的解释变量或函数形式选取有偏误。

图 4.4.1 给出了残差序列随某解释变量持续上升与持续下降的两类图形。前者预示着模型设定时可能遗漏了与某解释变量正向关的变量；后者则表明模型设定时可能遗漏了与某解释变量负相关的变量。



（a）正相关　　（b）负相关

图 4.4.1　残差序列变化图

当模型函数形式出现偏误时，残差序列也往往表现出某种有规律的变化特征。图 4.4.2 给出了一元回归模型中，真实模型呈幂函数形式，但却选取了线性函数进行回归的

情形。在这种情形下，容易知道残差序列呈现先正、后负、再正的变化特征。

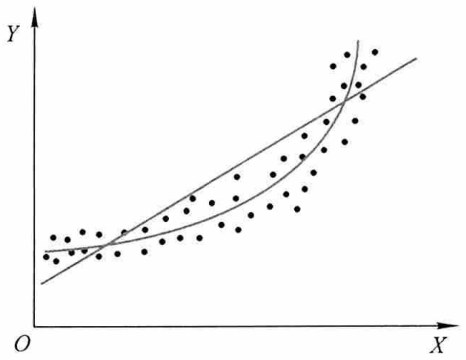


图 4.4.2 模型函数形式设定偏误时残差序列呈现正负交替变化

（2）一般性设定偏误检验。残差图示法能够帮助我们初步判定在模型设定时是否遗漏了重要的解释变量，或者是否设定了有偏误的函数形式。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)。

我们仍假设正确模型为(4.4.1)式，却对(4.4.2)式进行估计。如果我们明确知道遗漏了一个相关变量 X_2，则问题变得相对简单。只需再估计(4.4.1)式，并检验变量 X_2 前的参数是否显著不为零即可。如果是显著的，就能判定(4.4.2)式的模型设定有误。

但问题是我们事先并不知道哪个变量被漏掉了，即无法确定 X_2 是什么。当然，如果能有一个替代变量 Z 来替代 X_2，我们就能进行上述检验。在拉姆齐的 RESET 检验中，采用(4.4.2)式中被解释变量 Y 的估计值 $\hat{Y}$ 的若干次幂来充当该“替代”变量。即先估计(4.4.2)式，得

$$\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1$$

再用通过残差项 e 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，进行普通最小二乘估计。如 e 与 $\hat{Y}$ 的图形呈现曲线形变化时，回归模型可选为

$$Y = \beta_0 + \beta_1 X_1 + \gamma_1 \hat{Y}^2 + \gamma_2 \hat{Y}^3 + \mu \tag{4.4.10}$$

再根据§3.7 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。当然，若仅增加一个“替代”变量，也可通过 t 检验来判断。

如果检验结果表明一个或若干个“替代”变量能够引入到模型中去，则说明模型设定时遗漏了相关变量。由于在进行(4.4.10)式的回归时，可引入若干个“替代”变量来判明是否有多于一个的变量被漏掉，因此，该方法被称为**一般性设定偏误检验**(test for general mis-specification)。

RESET 检验也可用来检验函数形式设定偏误的问题。例如，在一元回归模型中，假设真实的函数形式是非线性的，可用泰勒定理将其近似地表示为如下多项式的形式：

$$Y=\beta_0+\beta_1X_1+\beta_2X_1^2+\beta_3X_1^3+\cdots+\mu \tag{4.4.11}$$

因此，如果我们用线性函数设定了模型，即意味着遗漏了相关变量 X_1^2, X_1^3，等等。所以，在一元回归中可以通过检验(4.4.11)式中 X_1 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。

对多元回归模型，非线性函数可能是关于若干个或全部解释变量的非线性，这时上述一元回归检验的程序已不适用，因为，模型中包含太多解释变量的高次幂及交叉项，容易导致自由度的损失以及出现多重共线性。这时仍可以按上面介绍的遗漏变量的程序进行检验。例如，我们对(4.4.1)式的二元线性模型进行估计，但却怀疑真实的函数形式是非线性的。这时，只需以(4.4.1)式估计出的 $\hat{Y}$ 的若干次幂为“替代”变量，进行类似于如下模型的估计：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\gamma_1\hat{Y}^2+\gamma_2\hat{Y}^3+\mu \tag{4.4.12}$$

再判断各“替代”变量的参数是否显著地不为零即可。这里采用 $\hat{Y}$ 的高次幂，既避免了自由度的损失与多重共线性的问题，同时它已包含了解释变量的高次幂及交叉项提供的信息。

例 4.4.1

§ 4.3 中的例 4.3.4 考察了 1995 年美国香烟的人均消费问题，设定的香烟消费函数如下

$$\ln Q=\beta_0+\beta_1\ln Y+\beta_2\ln P$$

在对香烟价格 P 的内生性检验中，在 10%的显著性水平下拒绝了它为外生变量的假设，但在 5%的显著性水平下不拒绝这一假设。在不拒绝该假设的情况下，原模型的普通最小二乘估计更适合。下面对该式进行模型设定的 RESET 检验。

首先，用原模型的普通最小二乘估计式

$$\ln\hat{Q}=10.341+0.344\ln Y-1.406\ln P \tag{4.4.13}$$

$$(10.11)\quad(1.46)\qquad(-5.60)$$

$$R^2=0.4328\qquad\overline{R}^2=0.4075\qquad F=17.17$$

估计出对香烟的人均消费的对数序列 $\ln\hat{Q}$。

其次，在原回归模型中加入新的解释变量 $(\ln Q)^2$ 后重新进行估计，得：

$$\ln\tilde{Q}=-122.47-5.364\ln Y+21.718\ln P+1.820(\ln\hat{Q})^2$$

$$(-1.82)\quad(-1.85)\quad(1.86)\qquad(1.98)$$

$$R^2=0.479\ 1$$

原回归模型的可决系数为 $R^2=0.432\ 8$，由(3.7.17)式计算 F 统计量：

$$F=\frac{(0.479\ 1-0.432\ 8)/1}{(1-0.479\ 1)/(48-4)}=3.91$$

该值小于5%显著性水平下、自由度为（1，44）的 F 分布的临界值4.06，因此不拒绝 $(\ln\hat{Q})^2$ 的参数显著为零的假设，表明原模型不存在遗漏相关变量的设定偏误。继续引入 $(\ln\hat{Q})^3$ 项仍可验证原双对数模型不存在设定偏误问题。

如果将原双对数模型设定为简单的线性模型

$$Q=\beta_0+\beta_1 Y+\beta_2 P$$

其OLS估计结果为：

$$\hat{Q}=198.46+1.882Y-1.080P \tag{4.4.14}$$

$$(8.63)\quad(1.22)\quad(-5.32)$$

$$R^2=0.415\ 0\qquad \overline{R}^2=0.389\ 0\qquad F=15.96$$

那么该模型的设定是否正确呢？仍进行 RESET 检验。由该模型计算的人均香烟消费序列记为 $\hat{Q}$，并将它的平方项作为解释变量加入到模型中进行普通最小二乘估计得：

$$\tilde{Q}=-509.55-7.010Y+3.910P+0.025\hat{Q}^2$$

$$(-1.99)\quad(-1.99)\quad(2.16)\quad(2.77)$$

$$R^2=0.502\ 1$$

原模型(4.4.14)式的可决系数为 $R^2=0.415\ 0$，由(3.7.17)式计算 F 统计量：

$$F=\frac{(0.502\ 1-0.415\ 0)/1}{(1-0.502\ 1)/(48-4)}=7.70$$

该值大于5%显著性水平下、自由度为（1，44）的 F 分布的临界值4.06，表明简单的线性模型存在着设定偏误问题。因此，双对数线性模型的设定要优于简单线性模型的设定。

本章练习题

1. 对一元回归模型

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

（1）假如其他基本假设全部满足，但 $\mathrm{Var}(\mu_i)=\sigma_i^2\neq\sigma^2$，试证明估计的斜率项仍是无偏的，但方差变为

$$\mathrm{Var}(\tilde{\beta}_1)=\frac{\sum x_i^2\sigma_i^2}{(\sum x_i^2)^2}$$

（2）如果 $\mathrm{Var}(\mu_i)=\sigma^2K_i$，试证明上述方差的表达式为

$$\mathrm{Var}(\tilde{\beta}_1)=\frac{\sigma^2}{\sum x_i^2}\cdot\frac{\sum x_i^2K_i}{\sum x_i^2}$$

该表达式与同方差假定下的方差 $\mathrm{Var}(\hat{\beta}_1)$ 之间有何关系？分 K_i 大于1与小于1两种情况讨论。

2. 对习题1中的一元线性回归模型，如果已知 $\mathrm{Var}(\mu_i)=\sigma_i^2$，则可对原模型以权 $\frac{1}{\sigma_i}$ 相乘后变换成

如下的二元模型：

$$\frac{Y_i}{\sigma_i}=\beta_0\frac{1}{\sigma_i}+\beta_1\frac{X_i}{\sigma_i}+\frac{\mu_i}{\sigma_i}$$

对该模型进行普通最小二乘估计就是加权最小二乘法。试证明该模型的随机干扰项是同方差的，并求出 β_1 的上述加权最小二乘估计量。

3. 试证明：二元线性回归模型

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\mu_i$$

中变量 X_1 与 X_2 的参数的普通最小二乘估计可以写成

$$\hat{\beta}_1=\frac{(\sum y_ix_{i1})(\sum x_{i2}^2)-(\sum y_ix_{i2})(\sum x_{i1}x_{i2})}{\sum x_{i1}^2\sum x_{i2}^2(1-r^2)}$$

$$\hat{\beta}_2=\frac{(\sum y_ix_{i2})(\sum x_{i1}^2)-(\sum y_ix_{i1})(\sum x_{i1}x_{i2})}{\sum x_{i1}^2\sum x_{i2}^2(1-r^2)}$$

其中，r 为 X_1 与 X_2 的相关系数。讨论 r 等于或接近于 1 时，该模型的估计问题。

4. 在教材中已指出，对截面数据，引起解释变量内生性的原因主要有三种情形，其中第三种情形是解释变量存在着测量误差。对于一元回归模型

$$Y_i=\beta_0+\beta_1X_i^*+\mu_i$$

假设解释变量 X_i^* 的实测值 X_i 与之有偏误：$X_i=X_i^*+e_i$，其中 e_i 是具有零均值、不序列相关，且与 X_i^* 及 μ_i 不相关的随机变量。试问：能否将 $X_i^*=X_i-e_i$ 代入原模型，使之变换成 $Y_i=\beta_0+\beta_1X_i+v_i$ 后进行估计？其中，v_i 为变换后模型的随机干扰项。

5. 试证明 § 4.3 中工具变量法部分给出的（4.3.15）式 $\frac{1}{n}\sum Z_i'(Y_i-X_i\tilde{\beta})=\mathbf{0}$ 与（4.3.16）式 $\frac{1}{n}Z'(Y-X\tilde{\beta})=\mathbf{0}$ 是等价的。

6. 验证教材（4.3.18）式中的参数估计与（4.3.14）式相同。

7. 产生模型设定偏误的主要原因是什么？模型设定偏误的后果以及检验方法有哪些？

8. 如果 X_2 与 X_1 在给定的样本下不相关（正交），则通过教材（4.4.7）式证明：α_1 的估计满足无偏性与一致性；但 α_0 的估计却是有偏且非一致的。

9. 如果真实的模型是 $Y_i=\beta_1X_i+\mu_i$，但你却拟合了一个带截距项的模型 $Y_i=\alpha_0+\alpha_1X_i+v_i$，试评述这一设定误差的后果。

10. 在习题 9 中，假设真实的模型是带截距项的模型，而你却对过原点的模型进行了普通最小二乘回归。请评述这一模型误设的后果。

11. 下表列出了某年中国部分省市城镇居民每个家庭平均全年可支配收入 X 与消费性支出 Y 的统计数据。

单位：元

地区	可支配收入 X	消费性支出 Y	地区	可支配收入 X	消费性支出 Y
北　京	10 349.69	8 493.49	河　北	5 661.16	4 348.47
天　津	8 140.50	6 121.04	山　西	4 724.11	3 941.87
内蒙古	5 129.05	3 927.75	河　南	4 766.26	3 830.71
辽　宁	5 357.79	4 356.06	湖　北	5 524.54	4 644.50
吉　林	4 810.00	4 020.87	湖　南	6 218.73	5 218.79
黑龙江	4 912.88	3 824.44	广　东	9 761.57	8 016.91

续表

地区	可支配收入 X	消费性支出 Y	地区	可支配收入 X	消费性支出 Y
上　海	11 718.01	8 868.19	陕　西	5 124.24	4 276.67
江　苏	6 800.23	5 323.18	甘　肃	4 916.25	4 126.47
浙　江	9 279.16	7 020.22	青　海	5 169.96	4 185.73
山　东	6 489.97	5 022.00	新　疆	5 644.86	4 422.93

(1) 试用普通最小二乘法建立居民人均消费支出与可支配收入的线性模型；

(2) 检验模型是否存在异方差性；

(3) 如果存在异方差性，试采用适当的方法估计模型参数。

12. 经济理论指出，家庭消费支出 Y 不仅取决于可支配收入 X_1，还取决于个人财富 X_2，即可设定如下回归模型：

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

试根据下表的资料进行回归分析，并说明估计的模型是否可靠，给出你的分析。

单位：元

编号	Y	X_1	X_2	编号	Y	X_1	X_2
1	700	800	8 100	6	1 150	1 800	18 760
2	650	1 000	10 090	7	1 200	2 000	20 520
3	900	1 200	12 730	8	1 400	2 200	22 010
4	950	1 400	14 250	9	1 550	2 400	24 350
5	1 100	1 600	16 930	10	1 500	2 600	26 860

13. 下表列出了 2006 年中国城镇居民人均消费支出(Y)、人均可支配收入(X_1)以及 2005 年人均消费支出(X_2)、人均可支配收入(Z_1)以及人均政府消费支出(Z_2)的相关数据。

单位：元

地区	2006 年人均消费支出 Y	2006 年人均可支配收入 X_1	2005 年人均消费支出 X_2	2005 年人均可支配收入 Z_1	2005 年人均政府消费支出 Z_2
北　京	14 825	19 978	13 244	17 653	10 058
天　津	10 548	14 283	9 653	12 639	6 728
河　北	7 343	10 305	6 700	9 107	5 313
山　西	7 171	10 028	6 343	8 914	3 964
内蒙古	7 667	10 358	6 929	9 137	5 432
辽　宁	7 987	10 370	7 369	9 108	4 449
吉　林	7 353	9 775	6 795	8 691	3 603
黑龙江	6 655	9 182	6 178	8 273	4 042
上　海	14 762	20 668	13 773	18 645	7 243
江　苏	9 629	14 084	8 622	12 319	5 876
浙　江	13 349	18 265	12 254	16 294	6 081
安　徽	7 295	9 771	6 368	8 471	2 812
福　建	9 808	13 753	8 794	12 321	5 400
江　西	6 646	9 551	6 109	8 620	2 982
山　东	8 468	12 192	7 457	10 745	6 058

续表

地区	2006 年人均消费支出 Y	2006 年人均可支配收入 X_1	2005 年人均消费支出 X_2	2005 年人均可支配收入 Z_1	2005 年人均政府消费支出 Z_2
河　南	6 685	9 810	6 038	8 668	5 348
湖　北	7 397	9 803	6 737	8 786	3 490
湖　南	8 169	10 505	7 505	9 524	4 194
广　东	12 432	16 016	11 810	14 770	4 564
广　西	6 792	9 899	7 033	9 287	4 186
海　南	7 127	9 395	5 929	8 124	3 418
重　庆	9 399	11 570	8 623	10 243	3 690
四　川	7 525	9 350	6 891	8 386	3 659
贵　州	6 848	9 117	6 159	8 151	3 968
云　南	7 380	10 070	6 997	9 266	5 037
西　藏	6 193	8 941	8 617	9 431	13 716
陕　西	7 553	9 268	6 656	8 272	2 526
甘　肃	6 974	8 921	6 529	8 087	4 171
青　海	6 530	9 000	6 245	8 058	7 066
宁　夏	7 206	9 177	6 404	8 094	5 146
新　疆	6 730	8 871	6 208	7 990	6 732

如果设定 2006 年中国城镇居民人均消费函数的计量模型如下：

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

试回答以下问题：

（1）对上述模型进行 OLS 估计；试问该模型存在内生解释变量问题吗？

（2）如果认定人均可支配收入 X_1 是内生变量，选择 2005 年人均可支配收入 Z_1 以及人均政府消费支出 Z_2 为工具变量，对上述模型进行两阶段最小二乘估计（2SLS）。

（3）请问能选取 Z_1 以及 Z_2 作为 X_1 的工具变量吗？你如何检验它们的外生性？

（4）检验 X_1 的内生变性。

14. 教材例 4.4.1 中，通过引入 $\ln\hat{Q}$ 的平方项后检验了双对数模型

$$\ln Q = \beta_0 + \beta_1 \ln Y + \beta_2 \ln P$$

式不存在设定偏误，请进一步引入 $\ln\hat{Q}$ 的立方项检验该模型是否仍不存在设定偏误问题。

15. 在第三章习题 17 中，假设有人认为原幂函数模型不是正确设定的模型，而下面的线性形式是正确设定的模型 $Y_i = \beta_0 + \beta_1 K_i + \beta_2 L_i + \mu_i$。你将如何检验哪一个模型设定更正确？