第四讲 多重共线性

Multi-Collinearity

- 一、多重共线性的概念
- 二、实际经济问题中的多重共线性
- 三、多重共线性的后果
- 四、多重共线性的检验
- 五、克服多重共线性的方法
- 六、案例
- *七、分部回归与多重共线性

一、多重共线性的概念

对于模型

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}+\mu_i$$

$$i=1,2,\ldots,n$$

其基本假设之一是解释变量是互相独立的。

如果某两个或多个解释变量之间出现了相关性，则称为多重共线性(Multicollinearity)。

如果存在

$$c_1X_{1i}+c_2X_{2i}+\ldots+c_kX_{ki}=0 \qquad i=1,2,\ldots,n$$

其中：c_i不全为0，**则称为解释变量间存在完全共线性（perfect multicollinearity）**。

如果存在

$$c_1X_{1i}+c_2X_{2i}+\ldots+c_kX_{ki}+v_i=0 \qquad i=1,2,\ldots,n$$

其中c_i不全为0，v_i为随机误差项，则称为 **近似共线性（approximate multicollinearity）**或**交互相关(intercorrelated)**。

在矩阵表示的线性回归模型

$$Y=X\beta+\mu$$

中，**完全共线性指：秩(X)<*k*+1**，即

$$X=\begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix}$$

中，至少有一列向量可由其他列向量（不包括第一列）线性表出。

如：$X_2=\lambda X_1$，则X_2对Y的作用可由X_1代替。

注意：

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

时间序列样本：经济繁荣时期，各基本经济变量（收入、消费、投资、价格）都趋于增长；衰退时期，又同时趋于下降。

横截面数据：生产函数中，资本投入与劳动力投入往往出现高度相关情况，大企业二者都大，小企业都小。

（2）滞后变量的引入

在经济计量模型中，往往需要引入滞后经济变量来反映真实的经济关系。

例如，消费=f(当期收入, 前期收入）

显然，两期收入间有较强的线性相关性。

（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，特定样本可能存在某种程度的多重共线性。

一般经验：

时间序列数据样本：简单线性模型，往往存在多重共线性。

截面数据样本：问题不那么严重，但多重共线性仍然是存在的。

二、多重共线性的后果

1、完全共线性下参数估计量不存在

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

的OLS估计量为：

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

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

例：对离差形式的二元回归模型

$$y = \beta_1 x_1 + \beta_2 x_2 + \mu$$

如果两个解释变量完全相关，如$x_2 = \lambda x_1$，则

$$y = (\beta_1 + \lambda\beta_2)x_1 + \mu$$

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

$$\widehat{\beta_1 + \lambda\beta_2} = \sum x_{1i} y_i \Big/ \sum x_{1i}^2$$

2、近似共线性下OLS估计量非有效

近似共线性下，可以得到OLS参数估计量，但参数估计量方差的表达式为

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

由于|X'X|≈0，引起$(X'X)^{-1}$主对角线元素较大，使参数估计值的方差增大，**OLS参数估计量非有效。**

仍以二元线性模型 $y=\beta_1 x_1+\beta_2 x_2+\mu$ 为例:

$$\mathrm{var}(\hat{\beta}_1)=\sigma^2(X'X)^{-1}_{11}=\frac{\sigma^2\sum x_{2i}^2}{\sum x_{1i}^2\sum x_{2i}^2-(\sum x_{1i}x_{2i})^2}=\frac{\sigma^2/\sum x_{1i}^2}{1-(\sum x_{1i}x_{2i})^2/\sum x_{1i}^2\sum x_{2i}^2}$$

$$=\frac{\sigma^2}{\sum x_{1i}^2}\cdot\frac{1}{1-r^2}$$

$\dfrac{(\sum x_{1i}x_{2i})^2}{\sum x_{1i}^2\sum x_{2i}^2}$ 恰为X_1与X_2的线性相关系数的平方r^2

由于 $r^2\leq 1$，故 $1/(1-r^2)\geq 1$

当完全不共线时，$r^2=0$　　$\mathrm{var}(\hat{\beta}_1)=\sigma^2/\sum x_{1i}^2$

当近似共线时，$0<r^2<1$　　$\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum x_{1i}^2}\cdot\frac{1}{1-r^2}>\frac{\sigma^2}{\sum x_{1i}^2}$

多重共线性使参数估计值的方差增大，$1/(1-r^2)$为**方差膨胀因子(Variance Inflation Factor, VIF)**

表 4.3.1　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	1000

当完全共线时，　$r^2=1$，　　$\mathrm{var}(\hat{\beta}_1)=\infty$

3、参数估计量经济含义不合理

如果模型中两个解释变量具有线性相关性，例如 $X_2 = \lambda X_1$，

这时，X_1和X_2前的参数β_1、β_2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。

β_1、β_2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β_1本来应该是正的，结果恰是负的。

4、变量的显著性检验失去意义

存在多重共线性时

↓

参数估计值的方差与标准差变大

↓

容易使通过样本计算的**t**值小于临界值，
误导作出参数为**0**的推断

↓

可能将重要的解释变量排除在模型之外

5、模型的预测功能失效

变大的方差容易使区间预测的“区间”变大，使预测失去意义。

注意：

除非是完全共线性，多重共线性并不意味着任何基本假设的违背；

因此，即使出现较高程度的多重共线性，OLS估计量仍具有线性性等良好的统计性质。

问题在于，即使OLS法仍是最好的估计方法，它却不是“完美的”，尤其是在统计推断上无法给出真正有用的信息。

三、多重共线性的检验

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）估计多重共线性的范围，即判断哪些变量之间存在共线性。

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OLS法下：R^2与F值较大，但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，需进一步确定究竟由哪些变量引起。

(1) 判定系数检验法

使模型中每一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。

如果某一种回归

$$X_{ji}=\alpha_1 X_{1i}+\alpha_2 X_{2i}+\ldots\alpha_L X_{Li}$$

的**判定系数**较大，说明X_j与其他X间存在**共线性**。

具体可进一步对上述回归方程作F检验：

构造如下F统计量

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

式中：$R_{j\cdot}^2$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，给定显著性水平α，计算F值，并与相应的临界值比较，来判定是否存在相关性。

另一等价的检验是:

在模型中排除某一个解释变量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与其它解释变量之间存在共线性。

方差膨胀因子法：

$$VIF_j = \frac{1}{(1-R_j^2)}$$

(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独立。

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与其它变量之间存在共线性关系。

四、克服多重共线性的方法

如果模型被检验证明存在多重共线性，则需要发展新的方法估计模型，最常用的方法有三类。

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。

以**逐步回归法**得到最广泛的应用。

- **注意：**

这时，剩余解释变量参数的经济含义和数值都发生了变化。

*2、第二类方法：差分法

时间序列数据、线性模型：将原模型变换为差分模型：

$$\Delta Y_i=\beta_1 \Delta X_{1i}+\beta_2 \Delta X_{2i}+\ldots+\beta_k \Delta X_{ki}+ \Delta \mu_i$$

可以有效地消除原模型中的多重共线性。

一般讲，增量之间的线性关系远比总量之间的线性关系弱得多。

例如：

表 4.3.2　　中国 GDP 与居民消费 C 的总量与增量数据（亿元）

年份	C	Y	C/Y	△C	△Y	△C/△Y
1978	1759.1	3605.6	0.488			
1979	2005.4	4074.0	0.492	246.3	468.4	0.526
1980	2317.1	4551.3	0.509	311.7	477.3	0.653
1981	2604.1	4901.4	0.531	287.0	350.1	0.820
1982	2867.9	5489.2	0.522	263.8	587.8	0.449
1983	3182.5	6076.3	0.524	314.6	587.1	0.536
1984	3674.5	7164.4	0.513	492.0	1088.1	0.452
1985	4589.0	8792.1	0.522	914.5	1627.7	0.562
1986	5175.0	10132.8	0.511	586.0	1340.7	0.437
1987	5961.2	11784.7	0.506	786.2	1651.9	0.476
1988	7633.1	14704.0	0.519	1671.9	2919.3	0.573
1989	8523.5	16466.0	0.518	890.4	1762.0	0.505
1990	9113.2	18319.5	0.497	589.7	1853.5	0.318
1991	10315.9	21280.4	0.485	1202.7	2960.9	0.406
1992	12459.8	25863.7	0.482	2143.9	4583.3	0.468
1993	15682.4	34500.7	0.455	3222.6	8637.0	0.373
1994	20809.8	46690.7	0.446	5127.4	12190.0	0.421
1995	26944.5	58510.5	0.461	6134.7	11819.8	0.519
1996	32152.3	68330.4	0.471	5207.8	9819.9	0.530
1997	34854.6	74894.2	0.465	2702.3	6563.8	0.412
1998	36921.1	79003.3	0.467	2066.5	4109.1	0.503
1999	39334.4	82673.1	0.476	2413.3	3669.8	0.658
2000	42911.9	89112.5	0.482	3577.5	6439.4	0.556

由表中的比值可以直观地看到，**增量的线性关系弱于总量之间的线性关系**。

进一步分析：

Y与C(-1)之间的判定系数为0.9988，

△Y与△C(-1)之间的判定系数为0.9567

3、第三类方法：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，所以

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，但确能消除多重共线性造成的后果。

例如：

①**增加样本容量，可使参数估计量的方差减小。**

*②岭回归法（Ridge Regression）

70年代发展的岭回归法，**以引入偏误为代价减小参数估计量的方差**，受到人们的重视。

具体方法是：引入矩阵**D**，使参数估计量为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{D})^{-1}\mathbf{X}'\mathbf{Y} \qquad (*)$$

其中矩阵**D**一般选择为主对角阵，即

$$\mathbf{D} = a\mathbf{I}$$

a为大于0的常数。

显然，与未含D的参数B的估计量相比，(*)式的估计量有较小的方差。

六、案例——中国粮食生产函数

根据理论和经验分析，影响粮食生产（Y）的主要因素有：

农业化肥施用量（X_1）；粮食播种面积(X_2)

成灾面积(X_3)；　　农业机械总动力(X_4)；

农业劳动力(X_5)

已知中国粮食生产的相关数据，建立中国粮食生产函数：

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

表 4.3.3　中国粮食生产与相关投入资料

年份	粮食产量 Y (万吨)	农业化肥施用量X_1（万公斤）	粮食播种面积X_2（千公顷）	受灾面积 X_3（公顷）	农业机械总动力X_4（万千瓦）	农业劳动力X_5（万人）
1983	38728	1659.8	114047	16209.3	18022	31645.1
1984	40731	1739.8	112884	15264.0	19497	31685.0
1985	37911	1775.8	108845	22705.3	20913	30351.5
1986	39151	1930.6	110933	23656.0	22950	30467.0
1987	40208	1999.3	111268	20392.7	24836	30870.0
1988	39408	2141.5	110123	23944.7	26575	31455.7
1989	40755	2357.1	112205	24448.7	28067	32440.5
1990	44624	2590.3	113466	17819.3	28708	33330.4
1991	43529	2806.1	112314	27814.0	29389	34186.3
1992	44264	2930.2	110560	25894.7	30308	34037.0
1993	45649	3151.9	110509	23133.0	31817	33258.2
1994	44510	3317.9	109544	31383.0	33802	32690.3
1995	46662	3593.7	110060	22267.0	36118	32334.5
1996	50454	3827.9	112548	21233.0	38547	32260.4
1997	49417	3980.7	112912	30309.0	42016	32434.9
1998	51230	4083.7	113787	25181.0	45208	32626.4
1999	50839	4124.3	113161	26731.0	48996	32911.8
2000	46218	4146.4	108463	34374.0	52574	32797.5

1、用OLS法估计上述模型：

$$\hat{Y} = -12816.44 + 6.213X_1 + 0.421X_2 - 0.166X_3 - 0.098X_4 - 0.028X_5$$

(-0.91)　(8.39)　(3.32)　(-2.81)　(-1.45)　(-0.14)

R^2=0.9828　$\overline{R^2}$ =0.9756　F=137.11　DW=1.81

R^2接近于1；

给定α=5%，得F临界值 $F_{0.05}(5,12)$=3.11

F=638.4 > 15.19，

故认上述粮食生产的总体线性关系显著成立。

但X_4、X_5的参数未通过t检验，且符号不正确，故**解释变量间可能存在多重共线性**。

2、检验简单相关系数

列出X_1，X_2，X_3，X_4，X_5的相关系数矩阵：

	X1	X2	X3	X4	X5
X1	1.00	0.01	0.64	0.96	0.55
X2	0.01	1.00	-0.45	-0.04	0.18
X3	0.64	-0.45	1.00	0.69	0.36
X4	0.96	-0.04	0.69	1.00	0.45
X5	0.55	0.18	0.36	0.45	1.00

- **发现：** X_1与X_4间存在高度相关性。

3、找出最简单的回归形式

分别作Y与X_1，X_2，X_4，X_5间的回归：

$$\hat{Y} = 30867.64 + 4.576X_1$$
(25.58)　(11.49)
R^2=0.8919　F=132.1　DW=1.56

$$\hat{Y} = -33821.18 + 0.699X_2$$
(-0.49)　(1.14)
R^2=0.075　F=1.30　DW=0.12

$$\hat{Y} = 31919.0 + 0.380X_4$$
(17.45)　(6.68)
R^2=0.7527　F=48.7　DW=1.11

$$\hat{Y} = -28259.19 + 2.240X_5$$
(-1.04)　(2.66)
R^2=0.3064　F=7.07　DW=0.36

- 可见，应选第1个式子为初始的回归模型。

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。

	C	X1	X2	X3	X4	X5	$\overline{R}^2$	DW
Y=f(X1)	30868	4.23					0.8852	1.56
t 值	25.58	11.49						
Y=f(X1,X2)	-43871	4.65	0.67				0.9558	2.01
t 值	-3.02	18.47	5.16					
Y=f(X1,X2,X3)	-11978	5.26	0.41	-0.19			0.9752	1.53
t 值	0.85	19.6	3.35	-3.57				
Y=f(X1,X2,X3,X4)	-13056	6.17	0.42	-0.17	-0.09		0.9775	1.80
t 值	-0.97	9.61	3.57	-3.09	-1.55			
Y=f(X1,X3,X4,X5)	-12690	5.22	0.40	-0.20		0.07	0.9798	1.55
t 值	-0.87	17.85	3.02	-3.47		0.37		

5、结论

回归方程以$Y=f(X_1，X_2，X_3)$为最优：

$$Y = -11978 + 5.26X_1 + 0.41X_2 - 0.19X_3$$

*七、分部回归与多重共线性

1、分部回归法(Partitioned Regression)

对于模型

$$\mathbf{Y} = \mathbf{XB} + \mathrm{N}$$

将解释变量分为两部分，对应的参数也分为两部分：

$$\mathbf{Y} = \mathbf{X}_1\mathrm{B}_1 + \mathbf{X}_2\mathrm{B}_2 + \mathrm{N}$$

在满足解释变量与随机误差项不相关的情况下，可以写出关于参数估计量的方程组：

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_2'\mathbf{x}_1 & \mathbf{x}_2'\mathbf{x}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathrm{B}}_1 \\ \hat{\mathrm{B}}_2 \end{bmatrix}$$

$$\hat{\mathbf{B}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\mathbf{B}}_2$$

$$= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{Y} - \mathbf{X}_2\hat{\mathbf{B}}_2)$$

如果存在 $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$

则有 $\hat{\mathbf{B}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$

这就是仅以X_1作为解释变量时的参数估计量

同样有 $\hat{\mathbf{B}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}$

这就是仅以X_2作为解释变量时的参数估计量。

2、由分部回归法导出

- 如果一个多元线性模型的解释变量之间完全正交，可以将该多元模型分为多个一元模型、二元模型、...进行估计，参数估计结果不变；

- 实际模型由于存在或轻或重的共线性，如果将它们分为多个一元模型、二元模型、...进行估计，参数估计结果将发生变化；

- 当模型存在共线性，将某个共线性变量去掉，剩余变量的参数估计结果将发生变化，而且经济含义有发生变化；

- 严格地说，实际模型由于总存在一定程度的共线性，所以每个参数估计量并不真正反映对应变量与被解释变量之间的结构关系。