第六章 多重共线性的情形及其处理

6 .1 多重共线性产生的背景和原因
6 .2 多重共线性对回归模型的影响
6 .3 多重共线性的诊断
6 .4 消除多重共线性的方法
6 .5 主成分回归
6 .6 本章小结与评注

第六章 多重共线性的情形及其处理

如果存在不全为0的$p+1$个数$c_0,c_1,c_2,\ldots,c_p$,使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}=0\ ,\ i=1,2,\ldots,n \qquad (6.1)$$

则称自变量$x_1,x_2,\ldots,x_p$之间存在着完全多重共线性。

在实际经济问题中完全的多重共线性并不多见，常见的是（6.1）式近似成立的情况，即存在不全为0的$p+1$个数$c_0,c_1,c_2,\ldots,c_p$,使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}\approx 0\ ,\ i=1,2,\ldots,n \qquad (6.2)$$

称自变量$x_1,x_2,\ldots,x_p$之间存在着多重共线性(Multi-collinearity),也称为复共线性。

§6.1多重共线性产生的经济背景和原因

当我们所研究的经济问题涉及到时间序列资料时,由于经济变量随时间往往存在共同的变化趋势,使得它们之间就容易出现共线性。

例如, 我们要研究我国居民消费状况,影响居民消费的因素很多,一般有职工平均工资、农民平均收入、银行利率、全国零售物价指数、国债利率、货币发行量、储蓄额、前期消费额等,这些因素显然既对居民消费产生重要影响,它们之间又有着很强的相关性。

§6.1多重共线性产生的经济背景和原因

许多利用截面数据建立回归方程的问题常常也存在自变量高度相关的情形。

例如,我们以企业的截面数据为样本估计生产函数,由于投入要素资本K,劳动力投入L,科技投入S,能源供应E等都与企业的生产规模有关,所以它们之间存在较强的相关性。

§6.2 多重共线性对回归模型的影响

设回归模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\varepsilon$$

存在完全的多重共线性,即对设计矩阵**X**的列向量存在不全为零的一组数$c_0,c_1,c_2,\ldots,c_p$,使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}=0\ ,\ i=1,2,\ldots,n$$

设计矩阵**X**的秩rank（**X**）＜p+1，此时|**x′x**|=0，正规方程组的解不唯一，（**x′x**）$^{-1}$不存在，回归参数的最小二乘估计表达式 $\hat{\boldsymbol{\beta}}=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ 不成立。

§6.2 多重共线性对回归模型的影响

对非完全共线性, 存在不全为零的一组数$c_0,c_1,c_2,\ldots,c_p$,使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\ldots+c_px_{ip}\approx 0\ ,\quad i=1,2,\ldots,n$$

此时设计矩阵 $\mathbf{X}$ 的秩 rank（$\mathbf{X}$）=p+1 虽然成立，但是此时$|\mathbf{x}'\ \mathbf{x}|\approx 0$，

$(\mathbf{x}'\ \mathbf{x})^{-1}$ 的对角线元素很大，$\hat{\boldsymbol{\beta}}$的方差阵 D（$\hat{\boldsymbol{\beta}}$)=$\sigma^2(\mathbf{X}'\ \mathbf{X})^{-1}$ 的

对角线元素很大，而 D（$\hat{\boldsymbol{\beta}}$)的对角线元素即为 $\mathrm{var}(\hat{\beta}_0)$，$\mathrm{var}(\hat{\beta}_1),\cdots,\mathrm{var}(\hat{\beta}_p)$

因而 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计精度很低。这样，虽然用 OLSE 还能得到 β 的无偏估计,但估计量$\hat{\boldsymbol{\beta}}$的变差很大,不能正确判断解释变量对被解释变量的影响程度,甚至出现估计量的经济意义无法解释。

§6.2 多重共线性对回归模型的影响

我们做y对两个自变量x_1,x_2的线性回归,假定y与x_1,x_2都已经中心化，此时回归常数项为零，回归方程为

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

记$L_{11} = \sum_{i=1}^{n} x_{i1}^2$，$L_{12} = \sum_{i=1}^{n} x_{i1} x_{i2}$，$L_{22} = \sum_{i=1}^{n} x_{i2}^2$，

则x_1与x_2之间的相关系数为

$$r_{12} = \frac{L_{12}}{\sqrt{L_{11} L_{22}}}$$

§6.2 多重共线性对回归模型的影响

$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ 的协方差阵为

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

$$X'X = \begin{pmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{|X'X|}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix} = \frac{1}{L_{11}L_{22} - L_{12}^2}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

$$= \frac{1}{L_{11}L_{22}(1 - r_{12}^2)}\begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

§6.2 多重共线性对回归模型的影响

由此可得

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{(1-r_{12}^2)L_{11}} \qquad (6.3)$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)L_{22}} \qquad (6.4)$$

可知,随着自变量 x_1 与 x_2 的相关性增强,$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差将逐渐增大。

当 x_1 与 x_2 完全相关时,r=1,方差将变为无穷大。

§6.2 多重共线性对回归模型的影响

当给不同的r_{12}值时,由表6.1可看出方差增大的速度。

为了方便，我们假设$\sigma^2/L_{11}=1$,相关系数从0.5变为0.9时,回归系数的方差增加了295%,相关系数从0.5变为0.95时,回归系数的方差增加了670%。

表6.1

r_{12}	0.0	0.2	0.50	0.70	0.80	0.90	0.95	0.99	1.00
$\mathrm{var}(\hat{\beta}_1)$	1.0	1.04	1.33	1.96	2.78	5.26	10.26	50.25	∞

§6.2 多重共线性对回归模型的影响

在例3.3中,我们建立的中国民航客运量回归方程为:

$$\hat{y}=450.9+0.354x_1-0.561x_2-0.0073x_3+21.578x_4+0.435x_5$$

其中：y—民航客运量(万人),

x_1—国民收入(亿元), x_2—消费额(亿元),

x_3—铁路客运量(万人), x_4—民航航线里程(万公里),

x_5—来华旅游入境人数(万人)。

5个自变量都通过了t检验，但是x_2的回归系数是负值，x_2是消费额，从经济学的定性分析看，消费额与民航客运量应该是正相关，负的回归系数无法解释。问题出在哪里？这正是由于自变量之间的复共线性造成的。

§6.3 多重共线性的诊断

一、方差扩大因子法

对自变量做中心标准化，则$\mathbf{X}^{*\prime}\mathbf{X}^{*}=(r_{ij})$为自变量的相关阵。记

$$\mathbf{C}=(c_{ij})=(\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1} \qquad (6.5)$$

称其主对角线元素$VIF_j=c_{jj}$为自变量x_j的方差扩大因子(Variance Inflation Factor,简记为VIF)。根据（3.31）式可知，

$$\mathrm{var}(\hat{\beta}_j)=c_{jj}\sigma^2/L_{jj},\qquad j=1,\cdots,p$$

其中L_{jj}是x_j的离差平方和，由（6.6）式可知用c_{jj}做为衡量自变量x_j的方差扩大程度的因子是恰如其分的。

§6.3 多重共线性的诊断

记 R_j^2 为自变量 x_j 对其余 $p-1$ 个自变量的复判定系数，

可以证明

$$c_{jj} = \frac{1}{1 - R_j^2} \qquad (6.7)$$

（6.7）式同样也可以作为为方差扩大因子 VIF_j 的定义，由此可知 $VIF_j \geqslant 1$。

§6.3 多重共线性的诊断

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	450.909	178.078		2.532	.030		
X1	.354	.085	2.447	4.152	.002	.001	1963
X2	-.561	.125	-2.485	-4.478	.001	.001	1741
X3	-7.E-03	.002	-.083	-3.510	.006	.315	3.171
X4	21.578	4.030	.531	5.354	.000	.018	55.5
X5	.435	.052	.564	8.440	.000	.040	25.2

a. Dependent Variable: Y

§6.3 多重共线性的诊断

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	x5, x3, x4, [a]x2	.	Enter

a. All requested variables entered.

b. Dependent Variable: x1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.9997452991[a]	.999	.999	175.08601

a. Predictors: (Constant), x5, x3, x4, x2

§6.3 多重共线性的诊断

经验表明,当$VIF_j \geq 10$时,就说明自变量x_j与其余自变量之间有严重的多重共线性,且这种多重共线性可能会过度地影响最小二乘估计值。

还可用p个自变量所对应的方差扩大因子的平均数来度量多重共线性。当

$$\overline{VIF} = \frac{1}{p}\sum_{j=1}^{p} VIF_j$$

远远大于1时就表示存在严重的多重共线性问题。

§6.3 多重共线性的诊断

当某自变量 x_j 对其余 p-1 个自变量的复判定系数 R_j^2 超过一定界限时，SPSS 软件将拒绝这个自变量 x_j 进入回归模型。

称 $Tol_j=1-R_j^2$ 为自变量 x_j 的容忍度（Tolerance），SPSS 软件的默认容忍度为 0.0001。也就是说，当 $R_j^2>0.9999$ 时，自变量 x_j 将被自动拒绝在回归方程之外，除非我们修改容忍度的默认值。

§6.3 多重共线性的诊断

以下用SPSS软件诊断例3.2中国民航客运量一例中的多重共线性问题。

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	450.909	178.078		2.532	.030		
X1	.354	.085	2.447	4.152	.002	.001	1963
X2	-.561	.125	-2.485	-4.478	.001	.001	1741
X3	-7.E-03	.002	-.083	-3.510	.006	.315	3.171
X4	21.578	4.030	.531	5.354	.000	.018	55.5
X5	.435	.052	.564	8.440	.000	.040	25.2

a. Dependent Variable: Y

§6.3 多重共线性的诊断

二、特征根判定法

（一）特征根分析

根据矩阵行列式的性质，矩阵的行列式等于其特征根的连乘积。因而，当行列式$|X'X|\approx 0$时，矩阵$X'X$至少有一个特征根近似为零。反之可以证明，当矩阵$X'X$至少有一个特征根近似为零时，X 的列向量间必存在复共线性，证明如下：

§6.3 多重共线性的诊断

记$X = (X_0, X_1, \ldots, X_p)$，其中

X_i为X 的列向量，

$X_0 = (1, 1, \ldots, 1)'$ 是元素全为1的n维列向量。

λ 是矩阵$X'X$的一个近似为零的特征根，$\lambda \approx 0$

$c=(c_0, c_1, \ldots, c_p)'$ 是对应于特征根 λ 的单位特征向量，则

$$X'Xc = \lambda c \approx 0$$

§6.3 多重共线性的诊断

上式两边左乘c′，得　$c' X' X c \approx 0$

从而有　$X c \approx 0$

即　$c_0 X_0 + c_1 X_1 + \ldots + c_p X_p \approx 0$

写成分量形式即为

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \ldots + c_p x_{ip} \approx 0 \quad , \quad i = 1, 2, \ldots, n$$

这正是（6.2）式定义的多重共线性关系。

§6.3 多重共线性的诊断

（二）条件数

特征根分析表明，当矩阵$\mathbf{X}'\mathbf{X}$有一个特征根近似为零时，设计矩阵$\mathbf{X}$ 的列向量间必存在复共线性。那么特征根近似为零的标准如何确定哪？这可以用下面介绍的条件数确定。记$\mathbf{X}'\mathbf{X}$的最大特征根为λ_m，称

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}, \quad \mathrm{i} = 0,1,2,\cdots,\mathrm{p}$$

为特征根λ_i的条件数（Condition Index）。

§6.3 多重共线性的诊断

用条件数判断多重共线性的准则

$0<k<10$时,设计矩阵$\mathbf{X}$没有多重共线性;

$10\leq k<100$时,认为$\mathbf{X}$存在较强的多重共线性;

当$k\geq 100$时,则认为存在严重的多重共线性。

§6.3 多重共线性的诊断

对例3.2中国民航客运量的例子，用SPSS软件计算出特征根与条件数如下：

Collinearity Diagnostics[a]

Dimension	Eigenvalue	Condition Index	Variance Proportions					
			(Constant)	X1	X2	X3	X4	X5
1	5.578	1.000	.00	.00	.00	.00	.00	.00
2	.378	3.842	.00	.00	.00	.00	.00	.00
3	3.745E-02	12.205	.01	.00	.00	.00	.03	.19
4	4.203E-03	36.431	.17	.00	.01	.09	.50	.04
5	1.939E-03	53.643	.72	.00	.01	.66	.15	.71
6	8.080E-05	262.762	.10	.99	.99	.25	.31	.06

a. Dependent Variable: Y

§6.3 多重共线性的诊断

方差比例是用于判断哪几个自变量之间存在共线性的。实际上共线性关系可以根据（6.9）式直接从特征向量看出来，只是SPSS软件在线性回归模块中没有输出特征向量阵。

把特征向量按照特征值由大到小排成行向量，每个数值平方后再除以特征值，然后再把每列数据除以列数据之和，使得每列数据之和为1，这样就得到了输出结果6.2的方差比。

再次强调的是线性回归分析共线性诊断中设计阵***X***包含代表常数项的一列1，而因子分析模块中给出的特征向量是对标准化的设计阵给出的，两者之间有一些差异。

§6.3 多重共线性的诊断

（三）直观判定法

1.当增加或剔除一个自变量,或者改变一个观测值时,回归系数的估计值发生较大变化。

2.从定性分析认为,一些重要的自变量在回归方程中没有通过显著性检验。

3.有些自变量的回归系数所带正负号与定性分析结果违背。

4.自变量的相关矩阵中,自变量间的相关系数较大。

5.一些重要的自变量的回归系数的标准误差较大。

§6.4 消除多重共线性的方法

一、剔除一些不重要的解释变量

在剔除自变量时,可以将回归系数的显著性检验、方差扩大因子VIF以及自变量的经济含义结合起来考虑，以引进或剔除变量。

§6.4 消除多重共线性的方法

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Tolerance	VIF
(Constant)	450.909	178.078		2.532	.030		
X1	.354	.085	2.447	4.152	.002	.001	1963
X2	-.561	.125	-2.485	-4.478	.001	.001	1741
X3	-7.E-03	.002	-.083	-3.510	.006	.315	3.171
X4	21.578	4.030	.531	5.354	.000	.018	55.5
X5	.435	.052	.564	8.440	.000	.040	25.2

a. Dependent Variable: Y

§6.4 消除多重共线性的方法

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	695.039	264.525		2.627	.024		
X2	-5.257E-02	.042	-.233	-1.262	.233	.013	77.546
X3	-1.170E-02	.003	-.134	-4.207	.001	.431	2.319
X4	32.037	4.951	.788	6.471	.000	.030	33.812
X5	.399	.080	.517	4.988	.000	.041	24.469

§6.4 消除多重共线性的方法

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	591.876	257.730		2.296	.040		
X3	-1.037E-02	.003	-.119	-3.934	.002	.504	1.984
X4	26.436	2.249	.650	11.754	.000	.150	6.650
X5	.317	.048	.411	6.568	.000	.117	8.514

§6.4 消除多重共线性的方法

二、增大样本容量

例如, 由（6.3）式和（6.4）式

$$\mathrm{var}(\hat{\beta}_1)=\frac{\sigma^2}{(1-r_{12}^2)L_{11}} \qquad \mathrm{var}(\hat{\beta}_2)=\frac{\sigma^2}{(1-r_{12}^2)L_{22}}$$

可以看到,在r_{12}固定不变时，当样本容量n增大时,L_{11}和L_{22}都会增大，两个方差均可减小，从而减弱了多重共线性对回归方程的影响。

§6.4 消除多重共线性的方法

三、回归系数的有偏估计

消除多重共线性对回归模型的影响是近30年来统计学家们关注的热点课题之一,除以上方法被人们应用外,统计学家还致力于改进古典的最小二乘法,提出以采用有偏估计为代价来提高估计量稳定性的方法,如:

岭回归法

主成分回归法

偏最小二乘法等。

§6.5　主成分回归

主成分分析（Principal Components Analysis, 简记为PCA）是多元统计分析的一个基本方法，是对数据做一个正交旋转变换，也就是对原有变量做一些线性变换，变换后的变量是正交的。为了避免变量的量纲不同所产生的影响，要求先把数据做中心标准化，中心标准化后的自变量样本观测数据矩阵（即设计阵）就是n行p列的矩阵，$\mathbf{r} = (\mathbf{X}^*)'\mathbf{X}'$ 就是相关阵。

§6.5 主成分回归

以例3.3民航客运量的数据为例

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.991	79.826	79.826
2	.932	18.641	98.468
3	.065	1.303	99.771
4	.011	.224	99.995
5	.000	.005	100.000

§6.5 主成分回归

Factor1	Factor2	Factor3	Factor4	Factor5
-1.2894	-1.481	-0.27458	0.42456	1.35596
-1.15466	-1.05109	-0.27713	0.28188	1.04786
-1.0025	-0.57432	0.2561	-0.08416	0.00132
-0.8846	-0.31539	0.27878	-0.61988	-0.88559
-0.79664	0.06441	0.82047	-0.24731	-1.14311
-0.67695	0.60404	0.99481	0.38279	-0.23507
-0.47273	0.94985	0.92127	0.05603	0.64251
-0.23379	1.07489	0.10751	0.19928	0.56835
-0.05087	0.72689	-0.92133	-1.38813	-1.02769
0.23403	0.98323	-0.8562	-2.1089	0.06758
0.59455	1.83545	-1.11974	0.83275	1.61644

§6.5　主成分回归

现在用y对前两个主成分Factor1和Factor2做普通最小二乘回归，得主成分回归回归方程：

$$\hat{y}=1159.125+936.781Factor1-185.876Factor2$$

不过以上回归方程的自变量是用两个主成分Factor1和Factor2表示的，应该转换回到用原始自变量表示的回归方程。

§6.5 主成分回归

分别用两个主成分Factor1和Factor2做因变量，以5个原始自变量做自变量做线性回归，所得的回归系数就是所需要的线性组合的系数。得到

$$Factor1 = -2.464 + 0.000\,037\,14x_1 + 0.000\,058\,31x_2 + 0.000\,009\,394x_3 + 0.010\,22x_4 + 0.000\,195\,7x_5$$

$$Factor2 = -8.426 - 0.000\,026\,72x_1 - 0.000\,033\,32x_2 + 0.000\,088\,51x_3 - 0.009\,708x_4 + 0.000\,110\,5x_5$$

§6.5　主成分回归

还原后的主成分回归方程为：

$$\hat{y} = 416.8 + 0.039\,76x_1 + 0.060\,82x_2 - 0.007\,652x_3 + 11.37x_4 + 0.162\,8x_5$$

每个回归系数的解释也都合理。

§6.5 主成分回归

载荷矩阵

Component Matrix(a)

	Component				
	1	2	3	4	5
x1	.985	-.165	.018	.047	.012
x2	.990	-.132	-.001	.055	-.011
x3	.413	.908	.066	.007	.000
x4	.963	-.214	.150	-.064	-.001
x5	.972	.128	-.195	-.043	.000

Extraction Method: Principal Component Analysis.

a 5 components extracted.

§6.6 本章小结与评注

当解释变量之间的简单相关系数很大时,可以断定自变量间存在着严重的多重共线性；但是一个回归方程存在严重的多重共线性时,解释变量之间的简单相关系数不一定很大。例如假定3个自变量之间有完全确定的关系

$$x_1 = x_2 + x_3$$

再假定x_2与x_3的简单相关系数r_{23}=-0.5，x_2与x_3的离差平方和L_{22}=L_{33}=1，此时

$$L_{23} = r_{23}\sqrt{L_{22}L_{33}} = -0.5$$

§6.6 本章小结与评注

$$L_{11}=\sum(x_1-\bar{x}_1)^2$$
$$=\sum(x_2+x_3-(\bar{x}_2+\bar{x}_3))^2=\sum((x_2-\bar{x}_2)+(x_3-\bar{x}_3))^2$$
$$=\sum(x_2-\bar{x}_2)^2+\sum(x_3-\bar{x}_3)^2+2\sum(x_2-\bar{x}_2)(x_3-\bar{x}_3)=1+1+2(-0.5)=1$$

$$L_{12}=\sum(x_1-x_1)(x_2-\bar{x}_2)$$
$$=\sum(x_2+x_3-(\bar{x}_2+\bar{x}_3))(x_2-\bar{x}_2)=\sum((x_2-\bar{x}_2)+(x_3-\bar{x}_3))(x_2-\bar{x}_2)$$
$$=L_{22}+L_{23}=1-0.5=0.5$$

$$r_{12}=L_{12}/\sqrt{L_{11}L_{22}}=0.5$$

同理 r_{13}=0.5