多重共线性的情形及其处理

马学俊 (主讲) 晁越 (助教)

苏州大学

数学科学学院

https://xuejunma.github.io/

1. 多重共线性产生的经济背景和原因
2. 多重共线性对回归模型的影响
3. 多重共线性的诊断
4. 消除多重共线性的方法
5. 评注

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 $c_0, c_1, c_2, \cdots, c_p$，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} = 0, i = 1, 2, \cdots, n \quad (6.1)$$

则称自变量 $x_1, x_2, \cdots, x_p$ 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 $c_0, c_1, c_2, \cdots, c_p$，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0, i = 1, 2, \cdots, n \quad (6.1)$$

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

全模型和选模型

- 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。
- 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。

多重共线性对回归模型的影响

设回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

存在完全的多重共线性，即对设计矩阵 $\boldsymbol{X}$ 的列向量存在不全为 0 的一组数 $c_0, c_1, c_2, \cdots, c_p$，使得

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

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

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 $c_0, c_1, c_2, \cdots, c_p$，使得

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

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

- $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ 的估计精度很低
  - $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的对角元素很大，$\hat{\boldsymbol{\beta}}$ 的方差矩阵 $D(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的对角元素很大，而 $D(\hat{\boldsymbol{\beta}})$ 的对角元素即 $var(\hat{\beta}_0), var(\hat{\beta}_1), \cdots, var(\hat{\beta}_p)$
- 虽然用普通最小二乘估计能得到 $\boldsymbol{\beta}$ 的无偏估计，但估计量 $\hat{\boldsymbol{\beta}}$ 的方差很大
- 不能正确判断解释变量对被解释变量的影响程度
- 甚至会导致估计量的经济意义无法解释。

多重共线性对回归模型的影响

对于二元回归模型，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 相关系数为

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

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

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

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

多重共线性对回归模型的影响

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

由此可得

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

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

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。

多重共线性对回归模型的影响

- 当给不同的 r_{12} 值时，由下表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 671%。

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	∞

多重共线性的诊断：方差扩大因子法

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

$$\boldsymbol{C}=(c_{ij})=(\boldsymbol{X}^{*'}\boldsymbol{X}^{*})^{-1}$$

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)。根据书 (3.31) 式可知，

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

其中 L_{jj} 是 x_j 的离差平方和，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。

记 R_j^2 为自变量 x_j 对其余 $p-1$ 个自变量的复决定系数，可以证明

$$c_{jj}=\frac{1}{1-R_j^2}$$

也可以作为放长扩大因子 VIF_j 的定义，由此式可知，$VIF_j\geq 1$

R_j^2 度量了自变量 x_j 与其余 $p-1$ 个自变量的线性相关程度，这种相关程度越强，说明自变量之间的多重共线性越严重，R_j^2 越接近于 1，VIF_j 就越大。

方差扩大因子法

- 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

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

远远大于 1 时就表示存在严重的多重共线性问题。

多重共线性的诊断：特征根判定法

- 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 $|\boldsymbol{X}'\boldsymbol{X}| \approx 0$ 时，矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零。反之可以证明，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 至少有一个特征根近似为零时，$\boldsymbol{X}$ 的列向量间必存在复共线性，

证明:

记 $\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_p)$，其中 $\boldsymbol{X}_i, i = 0, 1, \cdots, p$ 为 $\boldsymbol{X}$ 的列向量，$\boldsymbol{X}_0 = (1, 1, \cdots, q)'$ 是元素全为 1 的 n 维列向量。λ 是矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的一个近似为零的特征根，$\lambda \approx 0$，$\boldsymbol{c} = (c_0, c_1.c_2, \cdots, c_p)'$ 是对应于特征根 λ 的单位特征向量，则

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} = \lambda\boldsymbol{c} \approx \boldsymbol{0}$$

上式两边左乘 $\boldsymbol{c}'$ 得 $\boldsymbol{c}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{c} \approx 0$

从而有 $\boldsymbol{X}\boldsymbol{c} \approx \boldsymbol{0}$ 即 $c_0\boldsymbol{X}_0 + c_1\boldsymbol{X}_1 + c_2\boldsymbol{X}_2 + \cdots + c_p\boldsymbol{X}_p \approx \boldsymbol{0}$

写成分量得形式

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0$$

这正是定义的多重共线性关系。

特征根判定法

- 如果矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有多个特征根近似为零，在上面的证明中，取每个特征根的特征向量为标准化正交向量
- 证明：$\boldsymbol{X}'\boldsymbol{X}$ 有多少个特征根接近于零
- 设计矩阵 $\boldsymbol{X}$ 就有多少个多重共线性关系，并且这些多重共线性关系的系数向量就等于接近于零的那些特征根对应的特征向量。

特征根判定法

- 特征根分析表明，当矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 有一个特征根近似为零时，设计矩阵 $\boldsymbol{X}$ 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？
- 记 $\boldsymbol{X}'\boldsymbol{X}$ 的最大特征根为 λ_m，称

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

为特征根 λ_i 的条件数（Condition Index）。

- 用条件数判断多重共线性的准则
  - $0 < k < 10$ 时，设计矩阵 $\boldsymbol{X}$ 没有多重共线性；
  - $10 \leq k < 100$ 时，认为 $\boldsymbol{X}$ 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。

多重共线性的诊断：直观判定法

1. 当增加或剔除一个自变量，其它自变量的系数估计值或显著性发生较大变化，则回归方程存在严重的多重共线性。
2. 当定性分析认为重要的一些自变量在回归方程中没有通过显著性检验时，可初步判断存在着严重的多重共线性。
3. 与因变量简单相关系数绝对值很大的自变量，在回归方程中没有通过显著性检验时，可初步判断存在着严重的多重共线性。
4. 有些自变量的回归系数的数值大小与预期相差很大，甚至正负号与定性分析结果相反时，存在严重多重共线性问题。
5. 自变量的相关矩阵中，自变量间的相关系数较大时，会出现多重共线性问题。
6. 一些重要的自变量的回归系数的标准误差较大时，我们认为可能存在多重共线性。

消除多重共线性的方法

- 剔除一些不重要的解释变量
  - 在选择回归模型时，可以将回归系数的显著性检验、方差扩大因子 VIF 的数值、以及自变量的经济含义结合起来考虑，以引进或剔除变量。
- 增大样本容量
  - 例如，我们的问题设计两个自变量 x_1 和 x_2，假设 x_1 和 x_2 都已经中心化。

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

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

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。

- 回归系数的有偏估计
  - 岭回归法、主成分回归法、偏最小二乘法等。

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

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

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

本章小结与评注

$$
\begin{aligned}
L_{12} &= \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \\
&= \sum (x_2 + x_3 - (\bar{x}_2 + \bar{x}_3))(x_2 - \bar{x}_2) \\
&= \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)(x_2 - \bar{x}_2) \\
&= L_{22} + L_{23} = 1 - 0.5 = 0.5
\end{aligned}
$$

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

同理 $r_{13} = 0.4$

由此看到，当回归方程中的自变量数目超过 2 时，并不能由自变量间的简单相关系数不高，就断定它们不存在多重共线性。