多元回归分析中多重共线性的处理

天津医科大学卫生统计教研室 孟庆和

多元回归分析是一种应用广泛的多变量分析方法。在应用时，其假设各自变量之间不存在直线相关关系，即多重共线性。如果该假设不满足，则很难得到正确的分析结果[1]。关于多重共线性的诊断，在许多的统计分析软件包中都有相应的分析方法[2]。但是，如何对多重共线的问题进行处理呢？常规的方法是将存在共线性的指标剔除，然后再进行回归分析，将所得到的回归方程进行比较，选择一个最佳的回归方程。此时，虽然消除了多重共线性的影响，但其可导致回归参数真值的估计产生严重的偏差。

假设反映自变量和因变量之间真实关系的多元回归模型为：

$$Y = X\beta + \epsilon = (X_1 X_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \epsilon$$
$$= X_1\beta_1 + X_2\beta_2 + \epsilon \qquad (1)$$

其中矩阵 X_2 由可能引起共线性的变量构成。当 X_2 从模型中被剔除以后，实际使用的模型为：

$$Y = X_1\beta_1 + \epsilon \qquad (2)$$

应用最小二乘法，估计回归参数向量 β_1 的值为：

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$$

由于公式(1)为真实模型，有：

$$E(\hat{\beta}_1) = E[(X_1'X_1)^{-1}X_1'Y]$$
$$= (X_1'X_1)^{-1}X_1'E(Y)$$
$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2)$$
$$= \beta_1 + A\beta_2$$

其中 $A = (X_1'X_1)^{-1}X_1'X_2$ 为偏误矩阵。

由此可见，从回归模型中剔除引起共线性的自变量，可引起回归参数的错误估计。

因此，在采用此方法时，应综合各方面的因素来分析，如果引起多重共线性的变量在专业上属于不重要的指标，则可将其剔除，否则，应将其留在模型中，应用其他的方法处理多重共线性问题。

在进行回归分析时，分析的变量往往是时间序列数据，由此可产生严重的多重共线性问题。本文下面介绍一种处理时间序列数据多重共线性的方法。

方 法

1. 收集具有共线性指标的横断面数据，即同一时点，不同地区的该指标资料和其相应的因变量资料，按最小二乘法估计该指标(自变量，X_{t1})单独对因变量的回归方程，其回归系数用 $\hat{\beta}_1$ 表示，将其代入回归方程得：

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1X_{t1} + \beta_2X_{t2} + \cdots + \beta_iX_{ti} + \epsilon \qquad (3)$$
$$Y_t - \hat{\beta}_1X_{t1} = \hat{\beta}_0 + \beta_2X_{t2} + \cdots + \beta_iX_{ti} + \epsilon$$

令：$Y_t^* = Y_t - \hat{\beta}_1X_{t1}$ (4)

则：$Y_t^* = \beta_0 + \beta_2X_{t2} + \cdots + \beta_iX_{ti} + \epsilon$ (5)

2. 用具有时间序列数据其他自变量估计对 Y_t^* 的回归方程，得：

$$Y_t^* = \hat{\beta}_0 + \hat{\beta}_2X_{t2} + \cdots + \hat{\beta}_iX_{ti} \qquad (6)$$

利用上述方法求出 $\hat{\beta}_0$、$\hat{\beta}_2$、…、$\hat{\beta}_i$ 以后，即可得到共线性减弱了的回归方程。

实例分析

分析某地区经济收入、卫生状况与某病发病率的关系，数据见表 1。

表 1　某地区 1985～1994 年某病发病率、经济收入、卫生状况

年份	某病发病率(‰) (Y_t)	经济收入(元/月) (X_1)	卫生状况(综合评分) (X_2)
1985	35.4	351	8.5
1986	42.5	326	7.4
1987	39.2	309	6.2
1988	36.7	348	8.3
1989	39.8	312	6.0
1990	42.6	328	7.1
1991	51.7	286	5.9
1992	48.5	291	6.1
1993	38.8	342	8.2
1994	37.4	349	8.4

应用 SAS 软件包进行多元回归分析, 结果见表 2。

表 2　多元回归分析结果

变量	回归系数	标准误	t 值	P 值
X_1	-0.4346	0.0911	-4.77	< 0.01
X_2	5.6173	2.0023	2.81	< 0.05

从模型的结果看, 两个变量在回归分析中均有较显著的意义。但是, 卫生状况对该病发病率的作用恰好相反, 很显然不合理。因此, 考虑是否存在共线性, 对其进行共线性诊断, 结果见表 3。

表 3　对 X_1、X_2 的共线性诊断结果

序号	特征值	条件数	方差比	
			X_1	X_2
1	1.9530	1.0000	0.0235	0.0235
2	0.0470	6.4490	0.9765	0.9765

从表 3 可见, 两者存在着共线关系。为了减弱经济收入和卫生状况的共线性, 可进行如下分析:

(1) 收集 X_2 的横断面数据, 也就是同一时间、不同地区的该病发病率和卫生状况资料, 并计算两者的回归系数, 结果见表 4。

表 4　不同地区发病率和卫生状况

地区	某病发病率(‰)	卫生状况(综合评分)
1	35.4	8.4
2	36.2	8.3
3	34.8	8.6
4	33.9	8.9
5	35.8	8.3
6	36.0	8.1

其回归系数为: -2.9492。然后, 按公式 (4) 计算 Y_t^*。

(2) 用时间序列数据 X_1 (经济收入) 估计对 Y_t^* 的回归方程, 得到回归方程为:

$$\hat{Y}_t^* = 123.3619 - 0.3188X_1$$

估计出 $\hat{\beta}_0$、$\hat{\beta}_1$ 以后, 即得到了共线性减弱的回归方程。

$$\hat{Y}_t = 123.3619 - 0.3188X_1 - 2.9492X_2$$

参 考 文 献

1. 李严洁. 多元回归中的多重共线性及其存在的后果. 中国卫生统计　1992; 9(1): 24
2. 柳丽. 回归分析中的多重共线性的诊断与处理. 中国卫生统计　1994; 11(1): 5
3. J. Tobin. A Statistical Demand Function for food in the U. S. A., Journal of Royal Statistical society, Ser. A, 113-141