第 32 卷 第 1 期
2006 年 1 月 1～6 页

作 物 学 报
ACTA AGRONOMICA SINICA

Vol 32, No 1
pp 1－6 Jan., 2006

回归分析中的病态矩阵及其改进

莫惠栋*

(扬州大学数量遗传研究室, 江苏扬州 225009)

摘 要: 在回归分析中, 信息矩阵 $X'X$ 的行列式值 det($X'X$)如果近于 0, 就会造成其逆阵($X'X$)$^{-1}$的极度膨胀, 进而大大增加回归系数的误差均方, 影响回归配合的稳健性和精确度。因而 det($X'X$ 近于 0 的 $X'X$ 被称为"病态矩阵"。本文提出以 X 变数的相关矩阵 R 的行列式值为综合指标, 当 det(R)在区间[－0.01, 0.01] 和[－0.0001, 0.0001] 但非 0 时, 可分别认为其对应的 $X'X$ 是"病态的"和"严重病态的"。$X'X$ 的病态源于 X 矩阵的高度列依赖, 可用简单相关系数、多重决定系数和状态指数度量其列依赖程度。为了改进或消除 $X'X$ 的病态, 建议选用(1)简化原回归模型, (2)增加新的资料, (3)对回归系数添加限制条件, (4)采用诸如脊回归、广义逆 M^- 回归等非常规回归程序。简要讨论了病态诊断的重要性和病态改进的评价。

关键词: 回归分析; 病态矩阵; 病态的诊断和改进

中图分类号: Q332; S11$^+$4

Ill-conditioned Matrix and Its Improvement in Regression

MO Hui-Dong

(Laboratory of Quantitative Genetics, Yangzhou University, Yangzhou 225009, Jiangsu, China)

Abstract: In regression analysis, the information matrix $X'X$ is an important factor because of $b=(X'X)^{-1}X'Y$. If the determinant value of $X'X$, det($X'X$), is close to zero, the inverse of the $X'X$, ($X'X$)$^{-1}$, will extremely inflate, the error mean square for regression coefficient will largely increase, and in consequence the regression fitting will be poor robustness and low precision. Thus the matrix $X'X$ of det($X'X$)$\approx$0 is called "ill-conditioned matrix". In this paper the determinant value of correlative matrix R of X variables, det(R), is used as a synthetic index for ill-conditioning, i.e. if the det(R) lies in the intervals[－0.01, 0.01] and[－0.0001, 0.0001] but nonzero, the corresponding matrix $X'X$ can be regarded as ill-conditioned and seriously ill-conditioned, respectively. The ill-conditioned $X'X$ results from the linear dependency among columns in X matrix. Three diagnostic criteria, including linear correlation coefficient, multiple determination coefficient and condition index, can measure the degree of the column dependency. In order to improve or eliminate the ill-conditioning of $X'X$, four methods, i.e. (1) to reduce the original regression model, (2) to collect the new data, (3) to add the restrictive condition for regression coefficients and (4) to adopt the non-customary regression procedure such as the ridge regression and the generalized inverse M^- regression, are suggested. The importance of diagnosing the ill-conditioning and the evaluation for improved ill-conditioning are also discussed briefly.

Key words: Regression analysis; Ill-conditioned matrix; Diagnosis and improvement of ill-conditioning

1 奇异矩阵和病态矩阵

线性回归分析的正规方程组可写成

$$X'Xb = X'Y \quad (1)$$

其最小平方解则为

$$b = (X'X)^{-1}X'Y \quad (2)$$

式(1)和(2)中的 X 为自变数的 $n\times m$(表示 n 行 m 列, 下同)矩阵; X' 为 X 的转置矩阵; $X'X$ 为对称的 $m\times m$ 方阵; ($X'X$)$^{-1}$为 $X'X$ 的逆阵; Y 为依变数的 $n\times 1$ 向量; b 为待解元的 $m\times 1$ 向量。这里的 n 为观察值组数, m 为待估计的回归系数数。在试验统计和数量遗传学科, 往往特称以上的 X 为模型矩阵或设计矩阵, $X'X$ 为信息矩阵[1～3]。

*基金项目: 国家自然科学基金(39670391)资助。

作者简介: 莫惠栋(1934－), 男, 浙江温岭人, 教授, 博士生导师, 研究方向: 生物统计学和数量遗传学。E-mail: mhd28993@yahoo.com.cn

Received(收稿日期): 2005-02-02; Accepted(接受日期): 2005-05-18.

需要指出的是, 式(2)并不是普遍成立的。一种例外情形是 $\boldsymbol{X}$ 的列间存在完全的线性依赖, 即它的某一或某些列元素正好是另一或另一些列元素的线性函数。这称为共线性或多重共线性(collinearity or multicollinearity)。$\boldsymbol{X}$ 的共线性必然导致 $\boldsymbol{X}'\boldsymbol{X}$ 的列间和行间存在共线性, 并使 $\boldsymbol{X}'\boldsymbol{X}$ 为奇异(singular), 即其行列式值为 0[4]

$$\det(\boldsymbol{X}'\boldsymbol{X}) = 0 \qquad (3)$$

由于在计算待解元 $\boldsymbol{b}$ 时, 都要用到以 $\det(\boldsymbol{X}'\boldsymbol{X})$为除数, 故当式(3)成立时, $\boldsymbol{X}'\boldsymbol{X}$ 将无逆, $\boldsymbol{b}$ 将无解, 或者说无确定解。

以上结论在数学上可能是早已明确的, 但应用上仍常被忽视。例如近年一些学者建议的复杂遗传模型(即 $\boldsymbol{X}$ 矩阵)就存在明显的共线性。这类模型, 如果不添加限制条件, 就不可能进行常规的回归分析[5]。

在回归分析中, 还存在近似于但不同于式(3)的另一类情况, 即虽然 $\det(\boldsymbol{X}'\boldsymbol{X})\neq 0$, 但近于 0

$$\det(\boldsymbol{X}'\boldsymbol{X}) \approx 0 \qquad (4)$$

符合式(4)的 $\boldsymbol{X}'\boldsymbol{X}$, 通常称为病态矩阵或近奇异矩阵[5,6]。病态矩阵是由 $\boldsymbol{X}$ 的列间存在高度的线性依赖引起的, 它对回归分析的影响尚缺少研究。本文试图从应用统计学角度, 研讨病态矩阵的问题以及如何发现 $\boldsymbol{X}'\boldsymbol{X}$ 为病态而做出相应改进的方法, 供应用回归分析的研究者参考。

由于回归资料数量级的千差万别, "近于 0"的数量界限至今尚不明确。在数学专业文献中, 亦只是将方程组"参数的小改变会引起解的大改变"定义为"病态矩阵"[6], 并未涉及"大"和"小"的具体界限。为便于分析研究, 本文提出的建议标准为, 当 $\boldsymbol{X}'\boldsymbol{X}$ 中的元素以标准化变量一线性相关系数 r 表示时(即将 $\boldsymbol{X}'\boldsymbol{X}$ 变换为相关矩阵 $\boldsymbol{R}$)[7], 若

$$\det(\boldsymbol{R}) = 0 \pm 0.01 \qquad (5)$$

即在 −0.01 ~ 0.01 区间内但非 0, 可认为"近于 0", 对应于该 $\boldsymbol{R}$ 的 $\boldsymbol{X}'\boldsymbol{X}$ 为病态; 若

$$\det(\boldsymbol{R}) = 0 \pm 0.0001 \qquad (6)$$

即在 −0.0001 ~ 0.0001 区间内但非 0, 则认为"非常近于 0", 对应于该 $\boldsymbol{R}$ 的 $\boldsymbol{X}'\boldsymbol{X}$ 为严重病态。$\det(\boldsymbol{R}) > |0.01|$ 的 $\boldsymbol{X}'\boldsymbol{X}$ 则视为"良态"。

2 病态矩阵的问题

病态矩阵只是近于奇异, 故仍能进行回归分析, 但结果不可靠。这主要由于 $\det(\boldsymbol{X}'\boldsymbol{X})$为小值, 而对 $\boldsymbol{X}'\boldsymbol{X}$ 求逆时必须用到该小值为除数, 因而造成 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 中元素取值的极度"膨胀"。由之产生以下直接危害。

2.1 不能精确估计回归参数

回归系数的误差均方为[7]

$$V(b_i) = s^2 c_{ii}$$

上式的 s^2 为离回归均方, c_{ii} 为 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 中的主对角线(第 i 行 i 列)元素。故 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 的膨胀必使回归系数具有很大的误差, 导致本来可能较好配合的回归模型却配合失败。

例 **1** 设有以下资料:

X_1	4	4	7	7	7.1	7.1
X_2	16	16	49	49	50.41	51.41
Y	19	20	37	39	36	38

用二元回归模型 $E(Y)=\beta_0+\beta_1 X_1+\beta_2 X_2$ 进行配合[此例 $X_2=X_1^2$, 故也可以说用多项式回归模型 $E(Y)=\beta_0+\beta_1 X_1+\beta_2 X_1^2$ 作配合, 结果相同]。这里的

$$\boldsymbol{X}=\begin{bmatrix}1 & 4 & 16\\ 1 & 4 & 16\\ 1 & 7 & 49\\ 1 & 7 & 49\\ 1 & 7.1 & 50.41\\ 1 & 7.1 & 50.41\end{bmatrix} \qquad \boldsymbol{Y}=\begin{bmatrix}19\\ 20\\ 37\\ 39\\ 36\\ 38\end{bmatrix}$$

由之得正规方程组和解为

$$\begin{bmatrix}6 & 36.2 & 230.82\\ 36.2 & 230.82 & 1529.822\\ 230.83 & 1529.822 & 10396.3362\end{bmatrix}\begin{bmatrix}b_0\\ b_1\\ b_2\end{bmatrix}=\begin{bmatrix}189.0\\ 1213.4\\ 8078.34\end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X} \qquad \boldsymbol{b} = \qquad \boldsymbol{X}'\boldsymbol{Y}$$

$$\begin{bmatrix}b_0\\ b_1\\ b_2\end{bmatrix}=\begin{bmatrix}8574.1602 & -3357.8456 & 303.7434\\ -3357.8456 & 1315.1877 & -118.9789\\ 303.7434 & -118.9789 & 10.7641\end{bmatrix}\begin{bmatrix}189.0\\ 1213.4\\ 8078.34\end{bmatrix}$$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad \boldsymbol{X}'\boldsymbol{Y}$$

$$= \begin{bmatrix}-151.1863\\ 63.5315\\ -5.2150\end{bmatrix}$$

即有二元回归方程 $\hat{Y}=-151.1863+65.5315X_1-5.2150X_2$。其有关标准误为 $s_{b0}=113.4078$, $s_{b1}=44.4162$, $s_{b2}=4.0182$; 二元决定系数为 $R^2=0.9897$。

上述结果表明, 本例 X_1 和 X_2 的变异已能说明 Y 变异的 98.97%, 但 b_0、b_1 和 b_2 都与 0 无显著差异, 即上述二元回归无显著意义。进一步分析可知, 上述结果正是源于 $\boldsymbol{X}$ 中 X_1 和 X_2 的高度线性依赖, 其 $r=0.99996937$, 导致 $\det(\boldsymbol{R})=0.6125\times 10^{-4}$, $\boldsymbol{X}'\boldsymbol{X}$ 具有严重病态。

2.2　约数误差可能左右分析结果

约数误差(roundoff error)是指统计运算过程中因中间数字的有效位数不足而造成背离应有意义的结果。例如计算 $10^6x=(a/b)-(c/d)$, 设 $a=10000$, $b=0.03$, $c=16666.6663$, $d=0.05$, 中间数字均保持 8 位, 则 $(a/b)=333333.33$, $(c/d)=333333.33$, $x=0$; 但较精确结果却是 $10^6x=(ad-bc)/bd=(500-499.999989)/0.0015=0.007333$, $x=7333$! 在回归分析中, 当作为除数的 $\det(\boldsymbol{X}'\boldsymbol{X})\approx 0$ 时, 极易发生类似以上的约数误差。所以 Freund 在检查了大量回归资料后曾警告说, 许多合理的结论有时完全是由变化无常的约数误差造成[8]。

在实践上, 人们常用"双精度算法"以减少约数误差的干扰。研究认为, 双精度使计算机工作的数字密度比通常加倍, 如作为标准技术将浪费时间, 而且也不是必须的谨慎; 只要 $\boldsymbol{X}'\boldsymbol{X}$ 存在病态, 约数误差仍会常常发生[5]。所以, 关键还是在于发现矩阵病态和改进病态矩阵。

3　矩阵病态的诊断

前述 $\det(\boldsymbol{R})$ 是度量 $\boldsymbol{X}'\boldsymbol{X}$ 矩阵全体共线性程度的一个综合指标。$\boldsymbol{X}'\boldsymbol{X}$ 的病态源于 $\boldsymbol{X}$ 中的高度列依赖, 必须具体检查 $\boldsymbol{X}$ 矩阵, 才能发现不同列间的线性依赖程度。这称为 $\boldsymbol{X}$ 矩阵的病态诊断, 基本方法如下。

3.1　相关系数法

计算 $\boldsymbol{X}$ 矩阵的任一 X_i 列和 X_j 列 $(i\neq j)$ 的线性相关系数 r 或决定系数 r^2。X_i 和 X_j 列的 $r=\pm 1$ 为完全线性依赖, $r=0$ 为完全独立。作者认为, 如 $|r|>0.99$ 应视为两列间有高度线性依赖, 必导致 $\boldsymbol{X}'\boldsymbol{X}$ 呈现病态。此方法最简单, 但不能提供若干列间复杂依赖的信息, 即不能发现多重的共线性; 而小的 $|r|$ 值也不一定表示不存在共线性。

3.2　多元决定系数法

若定义 R_i^2 是 $\boldsymbol{X}$ 矩阵的 X_i 列依其他 $(m-1)$ 列 X_j $(j\neq i)$ 的 $(m-1)$ 元决定系数, 则当 $\boldsymbol{X}'\boldsymbol{X}$ 可逆时可以证明[7]:

$$R_i^2=1-1/c'_{ii} \tag{7}$$

式(7)的 c'_{ii} 为 $\boldsymbol{X}$ 的相关矩阵 $\boldsymbol{R}$ 逆阵 $\boldsymbol{R}^{-1}$ 的主对角线元素。当 X_i 列独立于所有 X_j 列时 $c'_{ii}=1$, 并随着 X_i 列对所有 X_j 列线性依赖程度的增加而增大, 直至完全依赖时 $c'_{ii}\to\infty$。

例 **2**　Mather 和 Jinks 的 6 世代加性-显性-上位性遗传模型的设计矩阵 $\boldsymbol{X}$[9] 及其相关矩阵 $\boldsymbol{R}$ 为

$$\boldsymbol{X}=\begin{bmatrix}1&1&0&1&0&0\\1&-1&0&1&0&0\\1&0&1&0&0&1\\1&0&1/2&0&0&1/4\\1&1/2&1/2&1/4&1/4&1/4\\1&-1/2&1/2&-1/4&1/4&1/4\end{bmatrix}$$

$$\boldsymbol{R}=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0.447214&0\\0&0&1&-0.903738&0&0.932568\\0&0&-0.903738&1&0&-0.705431\\0&0.447214&0&0&1&0\\0&0&0.932568&-0.932568&0&1\end{bmatrix}$$

由 $\boldsymbol{R}$ 可求得其行列式值 $\det(\boldsymbol{R})=0.004009$, 表明该模型的信息矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 是病态的。但 $\boldsymbol{Z}$ 的任何两列间的 $|r|$ 都不大于 0.94, 不足以直接导致 $\boldsymbol{X}'\boldsymbol{X}$ 呈病态。由 $\boldsymbol{R}$ 可得其逆阵 $\boldsymbol{R}^{-1}$ 为

$$\boldsymbol{R}=\begin{bmatrix}1&0&0&0&0&0\\0&1.25&0&0&-0.559018&0\\0&0&100.229968&49.055972&0&-58.865651\\0&0&49.055972&26.000242&0&-27.406641\\0&-0.559018&0&0&1.25&0\\0&0&-58.865651&-27.406641&0&36.562729\end{bmatrix}$$

故根据式(7)进而得 $R_1^2=1-1/1=0$, $R_2^2=1-1/1.25=0.2$, $R_3^2=1-1/100.229968=0.990023$, $R_4^2=0.961539$, $R_5^2=0.2$ 和 $R_6^2=0.972650$。这表明 $\boldsymbol{X}$ 中的第 1 列完全独立于其余 5 列, 第 2 列变异则有

20%可为其余 5 列的变异所说明, 第 3 列变异则有 99.0023%可为其余 5 列的变异所说明, 等等。以第 3 列对其余 5 列的线性依赖度最高, 其次为第 6 列。

多元决定系数法对于评价 $\boldsymbol{X}$ 的某一列对其余 $(m-1)$列的线性依赖程度很有效, 但不能反映多列与多列间的线性依赖度。

3.3 状态指数法

此法由 Belsley 最先提出[10], 被认为是评价多列间线性依赖度的最有效方法[5]。它包括计算 $\boldsymbol{X}$ 中各列的状态指数 η_j 和分解回归系数方差的构成 q_{ij} 两个部分。

3.3.1 计算状态指数 η_j 以列平衡的(columnequ-ilibrate)$\boldsymbol{X}_{n\times m}$ 矩阵(即 $\boldsymbol{X}$ 的每列元素均除以该列平方和的根值或标准差)为基础, 作奇异值分解, 得到

$$\boldsymbol{X}_{n\times m} = \boldsymbol{U}_{n\times m}\boldsymbol{D}_{m\times m}\boldsymbol{V}'_{m\times m} \tag{8}$$

式(8)中的 $\boldsymbol{D}_{m\times m}$ 为对角阵, 即

$$\boldsymbol{D}_{m\times m} = \mathrm{diag}(\mu_1, \mu_2, \cdots, \mu_m) \tag{9}$$

其中 $\mu_j(j=1, 2, \cdots, m)$为 $\boldsymbol{X}$ 第 j 列的奇异值, 且非负。由之可得 m 个状态指数

$$\eta_j = \mu_{\max}/\mu_j \tag{10}$$

式(10)中的 $\mu_{\max}$为 μ_j 中的最大值。μ_j 愈近于 0, η_j 将愈大, 表示 X_j 列间的线性依赖度愈高。

3.3.2 分解回归系数方差 $V(\boldsymbol{b})$的构成 当 $\boldsymbol{X}'\boldsymbol{X}$ 可逆时,

$$\sigma^{-2}V(\boldsymbol{b}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}' \tag{11}$$

对于第 i 个回归系数 b_i 则

$$\sigma^{-2}V(b_i) = \frac{v_{i1}^2}{\mu_1^2} + \frac{v_{i2}^2}{\mu_2^2} + \cdots + \frac{v_{im}^2}{\mu_m^2} = (q_{i1} + q_{i2} + \cdots + q_{im})\sum_{j=1}^{m}\frac{v_{ij}^2}{\mu_j^2} \tag{12}$$

以上 v_{ij} 为 $\boldsymbol{V}_{m\times m}$ 中的第 i 行 j 列元素, q_{ij} 为 $V(b_i)$属于状态指数 η_j 的成数(比率), 具有 $q_{ij}\geqslant 0$ 和 $\sum_{j=1}^{m}q_{ij}=1$。$\boldsymbol{X}$ 列间的线性依赖由 η_j 和 q_{ij} 推断, 大的 η_j 表示高线性依赖, 而该 η_j 行的大 q_{ij} 则表示高线性依赖的列(回归系数)。

例 **3** 设有经过列平衡处理的 $\boldsymbol{X}$ 矩阵[10]

$$10^3\boldsymbol{X} = \begin{bmatrix} -733 & 553 & 430 & -3 & -3 \\ 139 & -477 & 501 & 3 & 3 \\ 654 & -498 & -119 & 47 & 47 \\ -119 & 456 & -716 & 252 & 252 \\ 30 & 55 & -167 & -816 & -816 \\ 40 & -83 & 95 & 518 & 518 \end{bmatrix}$$

对之作状态指数分析得表 1 结果(具体过程见文献[10])。在表 1 中, $\eta_5=5799$ 时的 $q_{45}=q_{55}=1$, 清楚地表明 $\boldsymbol{X}$ 中的 X_4 和 X_5(第 4 和第 5 列)为完全线性依赖; 而 $\eta_4=16$ 的 $q_{14}=q_{24}=0.994$ 和 $q_{34}=0.953$ 则表明 X_1、X_2 和 X_3 的高线性依赖。可以验证此结果: 用决定系数法求得上述 $10^3\boldsymbol{X}$ 中 X_4 和 X_5 的 $r^2=1$, X_1 依 X_2 和 X_3 的 $R^2=0.982$, 表明 X_4 和 X_5 可相互说明 100%的变异, 而 X_1 的变异则有 98.2%可被 X_2 和 X_3 的变异所说明。这与表 1 结果相符。

表 **1** 应用状态指数法评价 $\boldsymbol{X}$ 矩阵列间的线性依赖度

Table 1 Evaluating the linear dependence degree among columns of $\boldsymbol{X}$ matrix by condition index

状态指数* Condition index* (η_j)	$V(b_i)$的比率 q_{ij} Proportions q_{ij} of $V(b_j)$				
	$V(b_1)$	$V(b_2)$	$V(b_3)$	$V(b_4)$	$V(b_5)$
$\eta_1=1$	0.000	0.000	0.000	0.000	0.000
$\eta_2=1$	0.005	0.005	0.000	0.000	0.000
$\eta_3=1$	0.001	0.001	0.047	0.000	0.000
$\eta_4=16$	0.994	0.994	0.953	0.000	0.000
$\eta_5=5799$	0.000	0.000	0.000	1.000	1.000
Sum	1.000	1.000	1.000	1.000	1.000

注: *取约整数。

Note: * Rounded to the nearest integer.

4 病态矩阵的改进

改进或消除病态矩阵的病态, 可能有多种方法, 较为普适者如下。

4.1 简化原来的回归模型

当发现 $\boldsymbol{X}$ 任两列的$|r|\geqslant 0.99$ 或第 i 列依其余各列的 $R_i^2\geqslant 0.99$ 时, 表明原回归模型存在过参数(overparameterization)情形, 应毫不犹豫地删除第 i 列及与之关联的 b_i。这时, $\boldsymbol{X}$ 对 $\boldsymbol{Y}$ 的总回归决定度几乎不变, 但 $\boldsymbol{X}'\boldsymbol{X}$ 的病态却可消除。这是改进病态最直截了当的方法。

例 **4** 已知例 1 资料的 X_1 和 X_2 列的 $r>0.99$。若删去 X_2 列, 即改用回归模型 $E(Y)=\beta_0+\beta_1X_1$ 配合, 则

$$\boldsymbol{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 7 & 7 & 7.1 & 7.1 \end{bmatrix}$$

$$\boldsymbol{Y}' = [19 \quad 20 \quad 37 \quad 39 \quad 36 \quad 38]$$

由以上 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 可得回归方程 $\hat{Y}=-4.0293+5.8888X_1$ 和 $s_{b0}=2.3332$, $s_{b1}=0.3762$, Y 依 X 的线性回归为极显著。应注意, 本例的 Y 依 X 的线性决定系数 $r^2=0.9839$, 与例 1 的二元决定系数 $R^2=0.9897$ 仅相差 0.0058, 表明删除 X_2 对回归预测的

准确性并无明显影响; 但本例 $\det(\boldsymbol{R})=1$ 又表明, 例 1 发生的高度线性依赖已完全消除。本例还表明 $\boldsymbol{X}'\boldsymbol{X}$ 的是否病态与所配模型直接关联: 同一资料, 对模型 $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$, $\boldsymbol{X}'\boldsymbol{X}$ 为病态; 而对模型 $E(Y)=\beta_0+\beta_1X_1$, $\boldsymbol{X}'\boldsymbol{X}$ 是良态。

例 **5**　例 2 资料 $\boldsymbol{X}$ 中的第 3 列对其余各列的线性依赖度最高($R_3^2=99.0\%$), 若删除第 3 列就是不估计显性效应, 损失较大。求其次, 可牺牲第 6 列($R_6^2=97.3\%$), 即不估计显性× 显性互作。这时

$$\boldsymbol{R}=\begin{bmatrix}1 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.447214\\ 0 & 0 & 1 & -0.903738 & 0\\ 0 & 0 & -0.903738 & 1 & 0\\ 0 & 0.447214 & 0 & 0 & 1\end{bmatrix}$$

$$\boldsymbol{R}^{-1}=\begin{bmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 0.25 & 0 & 0 & -0.559018\\ 0 & 0 & 5.456798 & 4.931516 & 0\\ 0 & 0 & 4.931516 & 5.456798 & 0\\ 0 & -0.559018 & 0 & 0 & 1.25\end{bmatrix}$$

进而可得 $\det(\boldsymbol{R})=0.146606$ 和 $R_1^2=0$, $R_2^2=0.2$, $R_3^2=0.816742$, $R_4^2=0.816742$, $R_5^2=0.2$。说明删除例 2 中 $\boldsymbol{X}$ 的第 6 列, $\boldsymbol{X}'\boldsymbol{X}$ 即成为良态, 列 3 和 4 对其余各列的线性依赖度也变小。

4.2　增加新的资料

对病态的 $\boldsymbol{X}'\boldsymbol{X}$, 如要保持原回归模型, 收集、补充适当的新资料有时也是改进病态的一种有效方法。

例 **6**　例 1 资料的病态主要是由 X_1 的 7 和 7.1 近似于相等数值所引起。如果新增一组观察值 $X_1=10$, $X_2=100$ 和 $Y=35$, 仍配合模型 $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2$, 则有

$$\boldsymbol{X}=\begin{bmatrix}1 & 4 & 16\\ 1 & 4 & 16\\ 1 & 7 & 49\\ 1 & 7 & 49\\ 1 & 7.1 & 50.41\\ 1 & 7.1 & 50.41\\ 1 & 10 & 100\end{bmatrix}\quad \boldsymbol{Y}=\begin{bmatrix}19\\ 20\\ 37\\ 39\\ 36\\ 38\\ 35\end{bmatrix}$$

$$\begin{bmatrix}b_0\\ b_1\\ b_2\end{bmatrix}=\begin{bmatrix}14.907171 & -4.521055 & 0.318973\\ -4.521055 & 1.429840 & -0.104018\\ 0.318973 & -0.104018 & 0.007777\end{bmatrix}\begin{bmatrix}224.0\\ 1563.4\\ 11578.34\end{bmatrix}$$

$$\boldsymbol{b}=\qquad (\boldsymbol{X}'\boldsymbol{X})^{-1}\qquad \boldsymbol{X}'\boldsymbol{Y}$$

$$=\begin{bmatrix}-35.8369\\ 18.3402\\ -1.1259\end{bmatrix}$$

得到二元回归方程 $\hat{Y}=-35.8369+18.3402X_1-1.1259X_2$ 和 $s_{b0}=4.7502$, $s_{b1}=1.4711$, $s_{b2}=0.1085$, 二元决定系数 $R^2=0.9865$。这说明 X 的变异可决定 Y 变异的 98.65%, 且 3 个回归系数都极显著, 配合二元线性方程非常适合。进一步计算可得 $\det(\boldsymbol{R})=0.0270$, 表明扩大 $\boldsymbol{X}$ 的观察范围(仅增 1 个样本点), 例 1 中 $\boldsymbol{X}'\boldsymbol{X}$ 的病态即已转变成良态。

4.3　添加限制条件

如果能对回归系数的取值给予合理的线性限制, 可将该限制条件直接加入 $\boldsymbol{X}'\boldsymbol{X}$。这时, 原 $\boldsymbol{X}'\boldsymbol{X}$ 的病态、甚至奇异, 都可能得到矫正而成为良态。

例 **7**　设以 A、B、C 3 种抗菌液喷洒某品种柑橘树各 2 株, 观察指标为各树的病情指数 Y。当应用回归模型 $E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3$ 分析资料(如果 A, $X_1=1$; 如果 B, $X_2=1$; 如果 C, $X_3=1$)时,

$$\begin{matrix}X_0 & X_1 & X_2 & X_3\end{matrix}$$

$$\boldsymbol{X}=\begin{bmatrix}1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 1 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 1\end{bmatrix}\quad \boldsymbol{Y}=\begin{bmatrix}Y_{11}\\ Y_{12}\\ Y_{21}\\ Y_{22}\\ Y_{31}\\ Y_{32}\end{bmatrix}$$

其正规方程组是

$$\begin{bmatrix}6 & 2 & 2 & 2\\ 2 & 2 & 0 & 0\\ 2 & 0 & 2 & 0\\ 2 & 0 & 0 & 2\end{bmatrix}\begin{bmatrix}b_0\\ b_1\\ b_2\\ b_3\end{bmatrix}=\begin{bmatrix}\Sigma Y_{..}\\ \Sigma Y_{1.}\\ \Sigma Y_{2.}\\ \Sigma Y_{3.}\end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X}\qquad b\quad =\quad \boldsymbol{X}'\boldsymbol{Y}$$

注意上述的 $\det(\boldsymbol{X}'\boldsymbol{X})=0$, 为奇异(由 $\boldsymbol{X}$ 中的 $X_0=X_1+X_2+X_3$ 引起), 不能做出回归分析。解决此问题的一个简便方法是删去上述正规方程中的任一方程, 加入一个对回归系数的限制方程, 例如此处可以采用 $0b_0+b_1+b_2+b_3=0$。这样, 改进后的正规方程组可以是

$$\begin{bmatrix}0 & 1 & 1 & 1\\ 2 & 2 & 0 & 0\\ 2 & 0 & 2 & 0\\ 2 & 0 & 0 & 2\end{bmatrix}\begin{bmatrix}b_0\\ b_1\\ b_2\\ b_3\end{bmatrix}=\begin{bmatrix}0\\ \Sigma Y_{1.}\\ \Sigma Y_{2.}\\ \Sigma Y_{3.}\end{bmatrix}$$

或 $\begin{bmatrix}6 & 2 & 2 & 2\\ 2 & 2 & 0 & 0\\ 2 & 0 & 2 & 0\\ 0 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}b_0\\ b_1\\ b_2\\ b_3\end{bmatrix}=\begin{bmatrix}\Sigma Y_{..}\\ \Sigma Y_{1.}\\ \Sigma Y_{2.}\\ 0\end{bmatrix}$, …… 等。

上述的“$X'X$”均已成为良态，其解均为 $b_0=\bar{y}$，$b_1=(\overline{y_1}-\bar{y})$，$b_2=(\overline{y_2}-\bar{y})$，$b_3=(\overline{y_3}-\bar{y})$，与该资料的方差分析结果等同。

上述方法可推广到具有不等观察值数目的试验：设第 i 处理的观察值为 n_i 个，则在改进奇异或病态的 $X'X$ 时可添加的线性限制为

$$\sum n_i b_i = 0 \qquad (13)$$

在实践上，应注意式(13)限制的合理性，因为限制不同将直接导致解的不同。本例添加的限制使 b_1、b_2、b_3 与 A、B、C 3 处理的效应相当，故是充分合理的。

4.4 采用非常规回归方法

在模型和资料不变情况下，对于病态的 $X'X$，还可采用非常规回归方法，主要有广义逆 M^- 回归[11]和脊回归[12]。前者实质上是简化原设的回归模型；后者则是保持原模型但直接干预对应于 $X'X$ 的主对角线元素，使逆阵元素的取值变小，回归系数的误差减小。其详细程序可见文献[11]和[12]。

5 讨论

5.1 病态诊断的重要性

回归方法在农学、生物学领域有着相当广泛的应用。但人们通常只关注自变数 X 和依变数 Y 的数量关系，并不注意 X 的结构特征。本研究表明，如果 X 中存在高度的线性列依赖，就会产生病态的 $X'X$，使 $\det(X'X)$近于 0；由于对 $X'X$ 求逆时必须用到以 $\det(X'X)$为除数，故近于 0 的 $\det(X'X)$又导致 $(X'X)^{-1}$ 中元素取值的极度“膨胀”，回归系数的误差均方 $V(b_i)$猛增。其结果是回归配合的稳健性和精确度皆严重丧失，甚至完全失败。但是，如果能够及时发现 $X'X$ 的病态，则可能应用多种方法使之趋于良态或成为良态。这是应用统计学中尚需充分研究的问题。

据笔者的实践，多元回归分析中，$X'X$ 的病态是较为普遍的现象，而且随着自变数个数(X 的列数)的增多，病态有着更普遍、更严重的趋势。因此病态诊断应作为回归分析的必要准备，特别在分析结果出现异常时(如例 1)。

5.2 病态改进效应的度量

病态 $X'X$ 的回归分析与改进 $X'X$ 的回归分析，在回归系数、回归系数的误差方差和回归系数的个数上均可能不同。评估改进病态的效应，需综合考虑上述因素。

设由病态的 $X'X$ 得到的回归系数为 b_i，其误差方差为 c_{ii}，共 m 个，而由改进的 $X'X$ 得到的相对应值为 b_i'、c_{ii}' 和 m'，则病态时回归系数的平均误差变异系数

$$C = \sum_i \frac{\sqrt{c_{ii}}}{b_i} \Big/ m \qquad (14)$$

而改进后的相应值为

$$C' = \sum_i \frac{\sqrt{c_{ii}'}}{b_i'} \Big/ m' \qquad (15)$$

因此，改进的效应(improvement power，IP)可用改进后误差变异系数减少的成数来度量，即

$$IP = 1 - C'/C \qquad (16)$$

例 **8** 例 1 资料的 $X'X$ 为病态，由式(14)可得

$$C = \left(\frac{\sqrt{8574.1602}}{151.1863} + \frac{\sqrt{1315.1877}}{63.5315} + \frac{\sqrt{10.7641}}{5.2150} \right) / 3 = 0.6041$$

在例 6 经改进，由式(15)可得

$$C' = \left(\frac{\sqrt{14.907171}}{35.8369} + \frac{\sqrt{1.429840}}{18.3402} + \frac{\sqrt{0.007777}}{1.1259} \right) / 3 = 0.0837$$

故

$$IP = 1 - 0.0837/0.6041 = 0.8614$$

这说明例 6 的改进使回归系数的误差变异系数比原资料平均减少了 86.14%，应是非常有效的。

References

[1] Box G E P, Draper N R. Empirical Model Building and Response Surface. New York: John Wiley and Sons Inc, 1987

[2] Khuri A I, Cornell J A. Response Surface: Designs and Analysis. New York: Marcel Dekker Inc, 1987

[3] Mather K, Jinks J L. Biometrical Genetics. 3rd ed. London: Chapman and Hall, 1982

[4] Jiang E-X(蒋尔雄), Gao K-M(高坤敏), Wu J-K(吴景琨). Linear Algebra(线性代数). Shanghai: People's Education Press, 1978 (in Chinese)

[5] Draper N R, Smith H. Applied Regression Analysis. New York: John Wiley and Sons Inc, 1998

[6] Chen J-L(陈景良), Chen X-H(陈向晖). Special Matrixes(特殊矩阵). Beijing: Tsinghua Univ Press, 2001. pp 154-162 (in Chinese)

[7] Mo H-D(莫惠栋). Agricultural Experimentation(农业试验统计), 2nd ed. Shanghai: Shanghai Sci &Tech Press, 1992(in Chinese)

[8] Freund R J. A warning of round-off errors in regression. *Am Statistician*, 1963, Der 17, 13-15

[9] Kearsey M J, Pooni H S. The Genetical Analysis of Quantitative Traits. London: Chapman and Hall, 1996. p 232

[10] Belsley D A. Conditioning Diagnostics, Collinearity and Weak Data in Regression. New York: John Wiley and Sons Inc, 1991. pp 28-29

[11] Mo H-D(莫惠栋). Regression by generalized inverses M^- and its application. *J Yangzhou Univ* (Agric and Life Sci)(扬州大学学报·农业与生命科学版), 2002, 23(1): 35-39(in Chinese with English abstract)

[12] Mo H-D(莫惠栋). Ridge regression procedure and its application. *Acta Agron Sin*(作物学报), 2002, 28(4): 433-438 (in Chinese with English abstract)