病态线性回归模型系数的主成分—岭估计

◎熊幼林　(湖北师范学院　435002)

本文针对岭估计和主成分估计的不足,从模型病态的根本原因出发,将模型分解成两个线性回归模型,对参数的两部分分别采用 LS 估计和岭估计,从而定义了一个新的估计,即主成分—岭估计. 通过研究该估计的性质,证明了在均方误差意义下,主成分—岭估计优于岭估计、$0-c$ 型岭估计和 $0-K$ 型广义岭估计,从而为病态线性回归模型系数的估计提供了一种改进的技术途径.

1. 引　言

考虑线性回归模型: $y = X\beta + e, E(e) = 0, \mathrm{Cov}(e) = \sigma^2 I_n$. (1.1)

其中,为 $n\times 1$ 随机观测向量,X 为 $n\times p$ 的设计矩阵且已中心化和标准化,$\mathrm{rank}(X) = p$,β 为 $p\times 1$ 的未知参数向量,e 为 $n\times 1$ 随机误差向量,I_n 为 n 阶单位矩阵. 存在 $p\times p$ 正交矩阵 Φ 使得 $X'X = \Phi\Lambda\Phi'$,其中 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$,$\Phi = (\varphi_1, \varphi_2, \cdots, \varphi_p)$,这里 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ 为 $X'X$ 的特征值,$\varphi_1, \varphi_2, \cdots, \varphi_p$ 为对应的标准正交化特征向量. 令: $Z = X\Phi$, $\alpha = \Phi'\beta$,则得到线性回归模型(1.1)的典则形式:

$$y = Z\alpha + e, E(e) = 0, \mathrm{Cov}(e) = \sigma^2 I_n. \quad (1.2)$$

α 的 LS 估计为: $\hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y$,从而原始参数 β 的 LS 估计为 $\hat{\beta} = \Phi\hat{\alpha}$. 当设计矩阵 X 的列向量之间出现复共线关系时,称模型(1.1)为病态线性回归模型. 由文献可知,当且仅当 $X'X$ 存在很小特征值时模型出现病态. 不妨假设 $\lambda_{r+1}, \cdots, \lambda_p \approx 0$,此时,$\mathrm{MSE}(\hat{\beta}) = \dfrac{\sigma^2 \sum_{i=1}^{p} 1}{\lambda_i}$ 非常大. 因而在均方误差意义下 LS 估计不再是一个好的估计. 为了解决这个问题,统计学家们做了大量工作. 目前应用最为广泛的有两种方法: 一是 Hoerl 和 Kennard 于 1970 年提出的岭估计. 对模型(1.1),回归系数 β 的岭估计定义为:

$$\hat{\beta}(k) = (X'X + kI_p)^{-1}X'y, k > 0. \quad (1.3)$$

二是主成分估计. 岭估计是以牺牲无偏性换取方差部分的大幅度减小,达到最终降低其均方误差的目的. 但从上述分析可知,真正使得 LS 估计变坏的原因在于 $\lambda_{r+1}, \cdots, \lambda_p$ 很小,因而增大 $\lambda_{r+1}, \cdots, \lambda_p$ 是有必要的. 对于主成分估计,虽然后面 $p-r$ 个主成分对因变量影响较小,但毕竟是影响 y 的一些因素,若轻易剔除,显然有失真之弊. 为了弥补这些不足,本文提出主成分 — 岭估计的设想.

2. 主成分 — 岭估计的定义

在模型(1.2)中,不妨假设 $\lambda_{r+1}, \cdots, \lambda_p \approx 0$,令 $\Phi = (\Phi_1 \vdots \Phi_2)$,其中 Φ_1 为 $p\times r$ 矩阵,Φ_2 为 $p\times(p-r)$ 矩阵,则模型变为: $y = X(\Phi_1, \Phi_2)(\Phi_1', \Phi_2')^T\beta + e = Z_1\alpha_1 + Z_2\alpha_2 + e$. (2.1)

其中 $Z_1 = X\Phi_1, Z_2 = X\Phi_2, \alpha_1 = \Phi_1'\beta, \alpha_2 = \Phi_2'\beta$. 记 $c = \dfrac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$,称为前 r 个主成分的贡献率.

将模型(2.1)变为: $cy = Z_1\alpha_1 + \dfrac{e}{2}$,

$$(1-c)y = Z_2\alpha_2 + \frac{e}{2}. \quad (2.2)$$

其中 $E(e) = 0, \mathrm{Cov}(e) = \sigma^2 I_n$,$c$ 的取值可根据实际需要预先取定. 易见,以上几个模型本质是相同的.

定义: 在模型(2.2)中,回归系数 $\alpha = (\alpha_1, \alpha_2)^T$ 的主成分 — 岭估计估计定义为:

$\alpha^* \triangleq (\alpha_1^*, \alpha_2^*(k))^T = (c(Z_1'Z_1)^{-1}Z_1'y, (1-c)(Z_2'Z_2 + kI_{p-r})^{-1}Z_2'y)^T$,相应地,原回归系数 $\beta = \Phi\alpha$ 的主成分 — 岭估计定义为: $\beta^* = \Phi(c\Lambda_1^{-1}Z_1'y, (1-c)(\Lambda_2 + kI_{p-r})^{-1}Z_2'y)^T$. (2.3)

3. 主成分 — 岭估计的基本性质

引理 1 α_1, α_2 的估计 $\alpha_1^*, \alpha_2^*(k)$ 具有下列性质: (1) $c'\alpha_1^*$ 是 $c'\alpha_1$ 的最佳线性无偏估计; (2) $\alpha_2^*(k)$ 是 α_2 的一个有偏估计; (3) $\mathrm{Cov}(\alpha_1^*) = \dfrac{\sigma^2\Lambda_1^{-1}}{2}$, $\mathrm{Cov}(\alpha_2^*(k)) = \dfrac{\sigma^2(\Lambda_2 + kI_{p-r})^{-1}\Lambda_2(\Lambda_2 + kI_{p-r})^{-1}}{2}$. (2.4)

引理 2 β^* 具有以下基本性质: (1) β^* 是最小二乘估计 $\hat{\beta}$ 向原点的一种压缩,且存在 $k > 0$,使得 β^* 是岭估计 $\hat{\beta}(k)$ 向原点的一种压缩; (2) β^* 比岭估计 $\hat{\beta}(k)$ 具有更小的偏差.

4. 主成分 — 岭估计的均方误差

定理 1 β^* 的均方误差为: $\mathrm{MSE}(\beta^*) = \dfrac{1}{2}\sigma^2\sum_{i=1}^{r}\dfrac{1}{\lambda i} + \dfrac{1}{2}\sigma^2\sum_{i=r+1}^{p}\dfrac{\lambda_i}{(\lambda_i + k)^2} + k^2\sum_{i=r+1}^{p}\dfrac{a_i^2}{(\lambda_i + k)^2}$. (4.1)

证明 $\because \mathrm{Cov}(\beta^*) = \mathrm{Cov}(\Phi\alpha^*) = \Phi\mathrm{Cov}(\alpha^*)\Phi'$, $\mathrm{Cov}(\alpha^*) = E[(\alpha_1^*, \alpha_2^*(k))^T - E(\alpha_1^*, \alpha_2^*(k))^T] \cdot [(\alpha_1^*, \alpha_2^*(k)) - E(\alpha_1^*, \alpha_2^*(k))] = \begin{pmatrix} \mathrm{Cov}(\alpha_1^*) & \mathrm{Cov}(\alpha_1^*, \alpha_2^*(k)) \\ \mathrm{Cov}(\alpha_2^*(k), \alpha_1^*) & \mathrm{Cov}(\alpha_1^*(k)) \end{pmatrix}$,

$\therefore \mathrm{tr}(\mathrm{Cov}(\beta^*)) = \mathrm{tr}(\Phi\mathrm{Cov}(\alpha^*)\Phi') = \mathrm{tr}(\mathrm{Cov}(\alpha^*)) = \mathrm{tr}(\mathrm{Cov}(\alpha_1^*)) + \mathrm{tr}(\mathrm{Cov}(\alpha_2^*(k)))$.

由(2.4)式知

$$\mathrm{tr}(\mathrm{Cov}(\beta^*)) = \frac{1}{2}\sigma^2\sum_{i=1}^{r}\frac{1}{\lambda i} + \frac{1}{2}\sigma^2\sum_{i=r+1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2}. \quad (4.2)$$

故 $\mathrm{MSE}(\beta^*) = \mathrm{tr}(\mathrm{Cov}(\beta^*)) + \|E(\beta^*) - \beta\|^2 = \dfrac{1}{2}\sigma^2\sum_{i=1}^{r}\dfrac{1}{\lambda_i} + \dfrac{1}{2}\sigma^2\sum_{i=r+1}^{p}\dfrac{\lambda_i}{(\lambda_i + k)^2} + k^2\sum_{i=r+1}^{p}\dfrac{a_i^2}{(\lambda_i + k)^2}$.

证明 令 $g(k) = \mathrm{MSE}(\hat{\beta}(k)) - \mathrm{MSE}(\beta^*)$.

$\because g(0) = \dfrac{1}{2}\sigma^2\sum_{i=1}^{p}\dfrac{1}{\lambda i} > 0$,而 $g(k)$ 在 $k \geqslant 0$ 时连续,

$\therefore \exists k^* > 0$,当 $k \in (0, k^*)$ 时有 $g(k) > 0$,从而就有 $\mathrm{MSE}(\beta^*) < \mathrm{MSE}(\hat{\beta}(k))$.