

基于刀切岭估计的线性回归参数的最小体积置信集

郭童格, 胡宏昌

(湖北师范大学数学与统计学院, 湖北 黄石 435002)

摘要: 本文基于岭估计研究了正态线性回归模型中未知参数的最小体积置信集问题. 利用 Jackknife 方法, 获得了在不同情况下未知参数的最小体积置信集. 最后, 与经典置信集进行比较, 在最小体积意义下我们所得到的置信集是最佳的.

关键词: 刀切岭估计; 线性回归; 置信集; 岭估计

MR(2010) 主题分类号: 62J05; 62J07

中图分类号: O212.4

文献标识码: A

文章编号: 0255-7797(2023)06-0529-08

1 引言

本文考虑以下线性回归模型:

$$y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n), \quad (1.1)$$

其中 y 为 $n \times 1$ 维观测向量, $X = (x_1, x_2, \dots, x_n)'$ 为 $n \times p$ 阶设计矩阵, $\text{rank}(X) = p$, β 为 $p \times 1$ 维未知参数向量, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ 为 $n \times 1$ 维随机误差向量. 相应的最小二乘估计和岭估计分别为 $\hat{\beta} = (X'X)^{-1}X'y$, $\hat{\beta}_k = (X'X + kI)^{-1}X'y$, 其中岭参数 $k > 0$, I 为单位矩阵.

在回归分析中, 自变量之间存在多重共线性是一个常见的问题, 这对分析产生了严重的不良影响. 它对普通最小二乘法 (OLS) 的一个主要后果是, 估计产生了巨大的采样方差, 这可能导致模型中排除了重要系数. 为了处理这种不稳定性, 学者们提出了许多方法, 其中最著名的是 Hoerl 和 Kennard^[1] 提出的岭估计, 它是基于在 $X'X$ 对角线上添加少量正值, 这使得岭估计有偏差, 但确保了比 OLS 更小的均方误差 (MSE). Farebrother^[2] 利用广义逆讨论岭回归估计均方误差的进一步结果; Firinguetti^[3] 通过大量的仿真实验研究了共线性和自相关干扰对几种岭回归估计的影响. 这些基于岭估计的文献由于其计算可行性和一些最优性质受到了相当大的关注, 但它们可能具有严重的偏差. 此后学者们为了减小偏差, 同时又尽可能地保留岭估计的优良性质, Quenouille^[4] 提出将刀切方法应用于有偏估计以减小偏差. 刀切方法提供的估计不仅具有小的偏差, 而且具有所有理想的大样本特性. 文献 [5-11] 研究了利用刀切法减小岭估计的偏差以及刀切岭估计的性质. 另外, Chaubey^[13] 指出刀切岭估计在处理更复杂的推理问题 (如获得置信区间) 时可提供精细的解决方案.

*收稿日期: 2023-06-01 接收日期: 2023-08-25

基金项目: 湖北省自然科学基金-黄石联合资助项目 (2022CFD042), 湖北师范大学"研究生创新科研" 立项建设项目 (2023Z079).

作者简介: 郭童格 (1998-), 女, 河南三门峡, 硕士, 主要研究方向: 统计模型的统计推断及其应用.

通讯作者: 胡宏昌 (1971-), 男, 湖北黄冈, 教授, 主要研究方向: 统计模型的统计推断及其应用.

置信区间在应用统计领域已经非常成熟,它是指由样本统计量所构造的总体参数的估计区间,置信集则是置信区间推广到多维的形式. 先前已有学者对线性回归模型中参数的置信集做出大量研究,例如: Vinod^[12] 研究了通过 Bootstrap 和 Stein 方法构造岭回归的参数置信区间; Chaubey 等人^[13] 利用 Bootstrap 和 Jackknife 方法在岭估计基础上提出和比较回归模型中回归系数的不同置信区间; Firinguetti 和 Bobadilla^[14] 在基于岭估计和 Edgeworth 展开建立回归系数的渐近置信区间; 张金^[15] 提出了基于最小二乘估计的线性回归模型参数的最小体积置信集. 尽管学者们对于基于最小二乘估计的参数最小体积置信集的研究成果十分丰富,但基于刀切岭估计的参数最小体积置信集的研究还未展开,因此本文对此进行了研究.

为了下文的方便,我们引入模型 (1.1) 的典则形式

$$y = Z\gamma + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n), \quad (1.2)$$

其中 $Z = XP, \gamma = P'\beta, Z'Z = P'X'XP = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p), \lambda_i \geq 0, i = 1, 2, \dots, p$. γ 的最小二乘和岭估计分别为 $\hat{\gamma} = (Z'Z)^{-1}Z'y, \hat{\gamma}_k = (Z'Z + kI)^{-1}Z'y$. 另外, Hinkly 在文献 [11] 中定义 (加权) 刀切岭估计为

$$\tilde{\gamma}_k = [I + k(Z'Z + kI)^{-1}] \hat{\gamma}_k = [I - k^2(Z'Z + kI)^{-2}] \hat{\gamma}. \quad (1.3)$$

于是原模型 (1.1) 的刀切岭估计可表示为

$$\tilde{\beta}_k = P\tilde{\gamma}_k = [P - Pk^2(\Lambda + kI)^{-2}] \hat{\gamma} = [I - Pk^2(\Lambda + kI)^{-2}P'] \hat{\beta}.$$

2 基于刀切岭估计的最小体积置信集

为了得到基于刀切岭估计的最小体积置信集,我们先给出如下引理.

引理 2.1^[15] 假设

- (1) $s \in S(Z)$ 是 θ 的充分统计量, 其概率密度函数 (pdf) 为 $f(s, \theta)$, 其中 $S(Z) = \Theta$;
- (2) 对于函数 $p(\theta) > 0, \tilde{f}(S, \theta) = f(S, \theta)/p(\theta)$ 为一个枢轴量;
- (3) 对任意 $\theta \in \Theta$, 置信集 $C_m(S)$ 定义为 $C_m(S) = \{\theta : \tilde{f}(S, \theta) \geq m\}$, 其中 $m > 0$ 是由 $P(\theta \in C_m(S)) = 1 - \alpha, \forall \alpha \in (0, 1)$ 所定义临界值;
- (4) 对函数 $p(\theta) > 0$, 满足 $|C_m^*(\theta)| = q(\theta)|C_m(\theta)|$, 以及对任意 $C(S)$ 和 $\theta \in \Theta$ 都有 $|C^*(\theta)| \leq q(\theta)|C(\theta)|$.

若满足以上条件, 则 $C_m(S)$ 是 θ 的置信水平为 $1 - \alpha$ 的最小体积置信集.

定理 2.1 当 σ^2 已知, 对任意置信集 $C(S)$ 都有 $|C^*(\beta)| \leq |C(\beta)|$, 则 β 的置信水平 $1 - \alpha$ 的最小体积置信集为

$$\left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P'\tilde{\beta}_k - X\beta \right\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha). \quad (2.1)$$

证 在模型的典则形式 (1.2) 式中, 令 $S = \hat{\gamma} = (Z'Z)^{-1}Z'y$, 则它是 γ 的充分统计量, S 的 pdf 为

$$f(s, \gamma) \propto \exp\left(-1/2\sigma^2 \|Zs - Z\gamma\|^2\right).$$

由引理 2.1 得 $\tilde{f}(S, \gamma) = f(S, \gamma)$ 并且

$$C_m(S) = \{\gamma : \tilde{f}(S, \gamma) \geq m\} = \left\{ \gamma : \|\hat{y} - Z\gamma\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha) \right\},$$

$$|C_m^*(\gamma)| = \int_{\|\hat{y} - Z\gamma\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha)} ds = \frac{\pi^{p/2} [\sigma^2 \chi_p^2(1 - \alpha)]^{p/2}}{|Z'Z|^{1/2} \Gamma(p/2 + 1)} = |C_m(\gamma)|.$$

于是在 $|C^*(\gamma)| \leq |C(\gamma)|$ 的限制条件下, β 的置信水平 $1 - \alpha$ 的最小体积置信集为

$$\|\hat{y} - Z\gamma\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha).$$

因为 $\left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 = \|\hat{y} - Z\gamma\|^2$, 所以结论成立.

定理 2.2 当 β 已知, 对任意置信集 $C(S)$ 都有 $|C^*(\sigma^2)| \leq |C(\sigma^2)|$, 则 σ^2 的置信水平 $1 - \alpha$ 的最小体积置信区间为

$$\left[\frac{\|y - X\beta\|^2}{\chi_n^2(1 - \alpha')}, \frac{\|y - X\beta\|^2}{\chi_n^2(\alpha - \alpha')} \right]. \quad (2.2)$$

证 因为 $S = 1/\|y - Z\gamma\|^2$ 是 σ^2 的充分统计量, 其 pdf 为 $\tilde{g}(s, \sigma^2) = g_n(1/\sigma^2 s)(1/\sigma^2 s)^2, t > 0$, 其中 g_n 为 χ_n^2 的 pdf. 由引理 2.1 得

$$C_m(S) = \{\sigma^2 : \tilde{g}(S, \sigma^2) \geq m\} = \{\sigma^2 : m_1 \leq 1/(\sigma^2 S) \leq m_2\},$$

其中 $m_1 = \chi_n^2(\alpha - \alpha'), m_2 = \chi_n^2(1 - \alpha)$. 令 $\alpha'_1 = \arg \min_{0 < \alpha_1 < \alpha} \{m_1^{-1} - m_2^{-1}\}$, 即

$$(d/d\alpha_1)(m_1^{-1} - m_2^{-1}) = m_1^{-2} g_n^{-1}(m_1) - m_2^{-2} g_n^{-1}(m_2) = 0, m_1^2 g_n(m_1) = m_2^2 g_n(m_2) \neq 0.$$

$$|C_m^*(\sigma^2)| = \frac{1}{\sigma^2} \left(\frac{1}{\chi_n^2(\alpha - \alpha'_1)} - \frac{1}{\chi_n^2(1 - \alpha'_1)} \right) = |C_m(\sigma^2)|.$$

因此 $|C^*(\sigma^2)| \leq |C(\sigma^2)|$ 的限制条件下, σ^2 的置信水平为 $1 - \alpha$ 的最小体积置信区间为

$$C_m(S) = \left[\|y - Z\gamma\|^2 / \chi_n^2(1 - \alpha'), \|y - Z\gamma\|^2 / \chi_n^2(\alpha - \alpha'_1) \right].$$

由于 $\|y - Z\gamma\|^2 = \|y - X\beta\|^2$, 故结论成立.

定理 2.3 当 β, σ^2 未知, 对任意置信集 $C(S)$ 都有 $|C^*(\beta, \sigma^2)| \leq \sigma^p |C(\beta, \sigma^2)|$, 则 (β, σ^2) 的置信水平 $1 - \alpha$ 的最小体积置信集为

$$C_m(S_1, S_2) := \left\{ (\beta, \sigma^2) : \begin{aligned} & g \left(\frac{1}{n\sigma^2} \left\| y - XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k \right\|^2 \right) \\ & + \frac{1}{n\sigma^2} \left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 \leq m_\alpha \end{aligned} \right\},$$

等价形式

$$C_m(S_1, S_2) := \left\{ \begin{aligned} & \left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 \\ & \leq n\sigma^2 \left[m_\alpha - g \left(\frac{1}{n\sigma^2} \left\| y - XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k \right\|^2 \right) \right], \\ & \frac{1}{nm_2} \left\| y - XP(I - k^2(\Lambda + K)^{-2})^{-1} P' \tilde{\beta}_k \right\|^2 \leq \sigma^2 \\ & \leq \frac{1}{nm_1} \left\| y - XP(I - k^2(\Lambda + K)^{-2})^{-1} P' \tilde{\beta}_k \right\|^2, \end{aligned} \right.$$

其中 m_α, m_1, m_2 见下文 (2.5) 式.

证 由因子分解定理, $S = (\hat{\gamma}/\|\hat{\varepsilon}\|, 1/\|\hat{\varepsilon}\|^2)$ 是 (γ, σ^2) 的充分统计量, 其中 $\|\hat{\varepsilon}\|^2 = \|y\|^2 - \|Z\hat{\gamma}\|^2$. 令 $(N_1, N_2) = (\hat{\gamma}, \|\hat{\varepsilon}\|^2)$, 则其 pdf 为

$$|Z'Z/(2\pi\sigma^2)|^{p/2} \exp\left\{-1/2\sigma^2\|Zn_1 - Z\gamma\|^2\right\} f_{n-p}(n_2/\sigma^2) 1/\sigma^2.$$

于是 $S = (S_1, S_2)$ 的 pdf 为

$$f(s_1, s_2, \gamma, \sigma^2) = \frac{|Z'Z|^{1/2}}{(2\pi\sigma^2)^{p/2}} \exp\left\{-1/2\sigma^2\|Zs_1/\sqrt{s_2} - Z\gamma\|^2\right\} f_{n-p}(1/\sigma^2 s_2) \frac{1}{\sigma^2} \left|\frac{\partial(n_1, n_2)}{\partial(s_1, s_2)}\right|, s_2 > 0,$$

其中 $(N_1, N_2) = (S_1/\sqrt{S_2}, 1/S_2)$, 雅可比行列式 $|\partial(n_1, n_2)/\partial(s_1, s_2)| = 1/s_2^{2+p/2}$. 由引理 2.1 得

$$\tilde{f}(s_1, s_2, \gamma, \sigma^2) = \exp\left\{-1/2\sigma^2\|Zs_1/\sqrt{s_2} - Z\gamma\|^2\right\} f_{n-p}(1/\sigma^2 s_2) (1/\sigma^2 s_2)^{2+p/2}, s_2 > 0,$$

$$C_m(s_1, s_2) = \{(\gamma, \sigma^2) : \tilde{f}(s_1, s_2, \gamma, \sigma^2) \geq m\},$$

$$|C_m^*(\gamma, \sigma^2)| = \iint_{\tilde{f}(s_1, s_2, \gamma, \sigma^2) \geq m} ds_1 ds_2 = \iint_{\tilde{f}(\gamma, \sigma^2, \tilde{s}_1, \tilde{s}_2) \geq m} \sigma^p d\tilde{s}_1 d\tilde{s}_2 = \sigma^p |C_m(\gamma, \sigma^2)|.$$

其中 \tilde{s}_1, \tilde{s}_2 由 $s_1 = \sigma\tilde{s}_1 + \gamma\sqrt{\tilde{s}_2} - \gamma, s_2 = \tilde{s}_2$ 决定.

因此, 在 $|C^*(\gamma, \sigma^2)| \leq \sigma^p |C(\gamma, \sigma^2)|$ 的限制条件下, (β, σ^2) 的置信水平 $1 - \alpha$ 的最小体积置信集为

$$C_m(S_1, S_2) = \left\{(\gamma, \sigma^2) : g\left(\frac{\|\hat{\varepsilon}\|^2}{n\sigma^2}\right) + \frac{\|\hat{y} - Z\gamma\|^2}{n\sigma^2} \leq m_\alpha\right\}, \quad (2.3)$$

其中 $g(x) = x - (1 + 2/n)\log x$ 是凸函数和 m_α 是被确定的临界值. 因为

$$\|y - XP(I - k^2(\Lambda + kI)^{-2})^{-1}P'\tilde{\beta}_k\|^2 = \|\hat{\varepsilon}\|^2,$$

$$\|XP(I - k^2(\Lambda + kI)^{-2})^{-1}P'\tilde{\beta}_k - X\beta\|^2 = \|\hat{y} - Z\gamma\|^2,$$

所以结论成立.

又因为

$$C_m(S_1, S_2) := \begin{cases} \|\hat{y} - Z\gamma\|^2 \leq n\sigma^2 \left[m_\alpha - g\left(\|\hat{\varepsilon}\|^2/n\sigma^2\right) \right], \\ g\left(\|\hat{\varepsilon}\|^2/n\sigma^2\right) \leq m_\alpha. \end{cases}$$

并且, $g\left(\|\hat{\varepsilon}\|^2/n\sigma^2\right) \leq m_\alpha$ 等价于 $m_1 \leq \|\hat{\varepsilon}\|^2/n\sigma^2 \leq m_2, g(m_1) = g(m_2) = m_\alpha$.

因此, (2.3) 式的最小体积置信集的等价形式为

$$C_m(S_1, S_2) := \begin{cases} \|\hat{y} - Z\gamma\|^2 \leq n\sigma^2 \left[m_\alpha - g\left(\|\hat{\varepsilon}\|^2/n\sigma^2\right) \right], \\ \|\hat{\varepsilon}\|^2/nm_2 \leq \sigma^2 \leq \|\hat{\varepsilon}\|^2/nm_1, \end{cases} \quad (2.4)$$

其中 m_α 取决于

$$1 - \alpha = n \int_{m_1}^{m_2} f_{n-p}(nx) F_p(nm_\alpha - ng(x)) dx, \quad (2.5)$$

F_{n-p} 为 χ_{n-p}^2 的分布函数.

3 置信集的比较

首先, 参数 β, σ^2 的经典置信集如下, 分以下三种情况

情况 1: 当 σ^2 已知时, β 的置信水平为 $1 - \alpha$ 置信集为

$$\left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha), \quad (3.1)$$

情况 2: 当 β 已知时, σ^2 的置信水平为 $1 - \alpha$ 置信区间为

$$\left[\|y - X\beta\|^2 / \chi_n^2(1 - \alpha_1) \|y - X\beta\|^2 / \chi_n^2(\alpha_2) \right], \quad (3.2)$$

其中 $0 < \alpha < 1$, 且 $\alpha_1 + \alpha_2 = \alpha$.

情况 3: 当 β 和 σ^2 未知时, (β, σ^2) 的置信水平为 $1 - \alpha$ 的经典置信集为

$$\begin{aligned} \left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 &\leq \sigma^2 \chi_p^2(1 - \alpha), \\ \frac{\|y - XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k\|^2}{\chi_{n-p}^2(1 - \alpha_1'')} &\leq \sigma^2 \leq \frac{\|y - XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k\|^2}{\chi_{n-p}^2(\alpha_2'')} \end{aligned} \quad (3.3)$$

其中 $(1 - \alpha')(1 - \alpha'') = 1 - \alpha, \alpha_1'' + \alpha_2'' = \alpha''$. 上式结果由以下两式可得

$$\begin{aligned} P \left[\left\| XP(I - k^2(\Lambda + kI)^{-2})^{-1} P' \tilde{\beta}_k - X\beta \right\|^2 \leq \sigma^2 \chi_p^2(1 - \alpha) \right] &= 1 - \alpha', \\ P \left[\chi_{n-p}^2(\alpha_2'') \leq \|\hat{\varepsilon}\|^2 / \sigma^2 \leq \chi_{n-p}^2(1 - \alpha_1'') \right] &= 1 - \alpha'', \end{aligned}$$

其中 $Z\hat{\gamma}$ 与 $\hat{\varepsilon}$ 相互独立.

下面将以上经典置信集和基于刀切岭估计的最小体积置信集进行比较.

情况 1: 当 σ^2 已知, β 未知时. 在 $|C^*(\gamma)| \leq |C(\gamma)|$ 的限制条件下, 置信水平为 $1 - \alpha$ 的 β 经典置信集 (3.1) 式和最小体积置信集 (2.1) 式相同.

情况 2: 当 β 已知, σ^2 未知时. 在置信水平为 $1 - \alpha$ 中 σ^2 的经典置信区间 (3.2) 式, 其常用区间“等尾”($\alpha_1 = \alpha_2 = \alpha/2$), 在 $|C^*(\sigma^2)| \leq |C(\sigma^2)|$ 的限制条件下, 最优经典置信区间与最小体积置信区间 (2.2) 式相同, 是所有置信区间中最优.

情况 3: 当 β, σ^2 未知时. (3.3) 式中常用的经典置信集是“等尾”($\alpha_1'' = \alpha_2'' = \alpha''/2$)“等概率” $1 - \alpha' = 1 - \alpha'' = \sqrt{1 - \alpha}$. (3.3) 式的经典置信集有体积

$$\begin{aligned} |C(S)| &= \int_{\|\hat{\varepsilon}\|^2 / \chi_{n-p}^2(\alpha_2'')}^{\|\hat{\varepsilon}\|^2 / \chi_{n-p}^2(1 - \alpha_1'')} \frac{\pi^{p/2} [\sigma^2 \chi_p^2(1 - \alpha')]^{p/2}}{|Z'Z|^{1/2} \Gamma(p/2 + 1)} d\sigma^2 \\ &= \frac{\pi^{p/2} \|\hat{\varepsilon}\|^{p+2}}{|Z'Z|^{1/2} \Gamma(p/2 + 2)} \left[\frac{\chi_p^2(1 - \alpha')^{p/2}}{\chi_{n-p}^2(\alpha_2'')^{p/2+1}} - \frac{\chi_p^2(1 - \alpha')^{p/2}}{\chi_{n-p}^2(1 - \alpha_1'')^{p/2+1}} \right], \end{aligned}$$

(2.3) 式或 (2.4) 式最小体积置信集的体积为

$$\begin{aligned} |C_m(S)| &= \iint_{g(\|\hat{\varepsilon}\|^2/n\sigma^2) + \|\hat{y} - Z\gamma\|^2/n\sigma^2 \leq m_\alpha} d\sigma^2 d\gamma = \iint_{g(\|\hat{\varepsilon}\|^2/n\sigma^2) + \|(Z'Z)^{1/2}(\hat{\gamma} - \gamma)\|^2/n\sigma^2 \leq m_\alpha} d\sigma^2 d\gamma \\ &= \frac{\|\hat{\varepsilon}\|^{p+2}}{n|Z'Z|^{1/2}} \iint_{g(x) + \|y\|^2 \leq m_\alpha} x^{-p/2-2} dx dy = \frac{\pi^{p/2} \|\hat{\varepsilon}\|^{p+2}}{n|Z'Z|^{1/2} \Gamma(p/2+1)} \int_{m_1}^{m_2} \frac{[m_\alpha - g(x)]^{p/2}}{x^{p/2+2}} dx. \end{aligned}$$

假设 $C(S)$ 是 (3.3) 式的置信集, 其中 $S = (S_1, S_2) = (\hat{\gamma}/\|\hat{\varepsilon}\|, 1/\|\hat{\varepsilon}\|^2)$ 则 $C(S)$ 满足 $|C^*(\theta)| = \sigma^p |C(\theta)|, \theta \in \Theta$. 事实上, 注意到

$$\begin{aligned} |C(S)| &= |C(\hat{\gamma}/\|\hat{\varepsilon}\|, 1/\|\hat{\varepsilon}\|^2)| = \frac{\pi^{p/2} \|\hat{\varepsilon}\|^{p+2}}{|Z'Z|^{1/2} \Gamma(p/2+2)} \left[\frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(\alpha_2'')^{p/2+1}} - \frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(1-\alpha_1'')^{p/2+1}} \right], \\ |C(\theta)| &= |C(\gamma, \sigma^2)| = \frac{\pi^{p/2} \sigma^{-p-2}}{|Z'Z|^{1/2} \Gamma(p/2+2)} \left[\frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(\alpha_2'')^{p/2+1}} - \frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(1-\alpha_1'')^{p/2+1}} \right], \end{aligned}$$

于是

$$\begin{aligned} |C^*(\theta)| &= \int_{[\sigma^2 \chi_{n-p}^2(1-\alpha_1'')]^{-1}}^{[\sigma^2 \chi_{n-p}^2(\alpha_2'')]^{-1}} \frac{\pi^{p/2} [\sigma^2 \chi_p^2(1-\alpha') s_2]^{p/2}}{|Z'Z|^{1/2} \Gamma(p/2+1)} ds_2, \\ |C^*(\theta)| &= \frac{\pi^{p/2} \sigma^{-2}}{|Z'Z|^{1/2} \Gamma(p/2+2)} \left[\frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(\alpha_2'')^{p/2+1}} - \frac{\chi_p^2(1-\alpha')^{p/2}}{\chi_{n-p}^2(1-\alpha_1'')^{p/2+1}} \right] = \sigma^p |C(\theta)|. \end{aligned}$$

因此由引理 2.1 得 $|C_m(S)| \leq |C(S)|$.

4 应用

在本节中, 我们利用文献 [16] 给出的数据来说明上文置信集的可靠性. 该数据集来自对不同组成的波兰水泥在凝固和硬化过程中产生的热量 y 与生产水泥的四种化学成分 x_1, x_2, x_3, x_4 含量之间的关系的实验研究. 该数据样本量包括 $n = 13$ 个观测值, 一个响应变量和四个解释变量.

首先, 通过文献 [16] 可得到 $X'X$ 的特征值为 $\lambda_1 = 105.406, \lambda_2 = 809.952, \lambda_3 = 5965.339, \lambda_4 = 44663.303$, 其条件数为

$$k = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{44663.303}{105.406} = 423.726,$$

该信息表明回归向量之间存在严重的多重共线性. 我们得到了 β 和 σ^2 的最小二乘估计

$$\hat{\beta} = (2.193, 1.1533, 0.7582, 0.4863)', \hat{\sigma}_{OLS}^2 = 5.8455.$$

然后, 考虑随机约束 $r = R\beta + e, e \sim N(0, \sigma^2)$, 其中 $R = (1, -1, 1, 0), r = 0$. 由文献 [16] 可知, 随机约束 Liu- 型估计 (SRLTE) 为

$$\hat{\beta}_{SRLTE}(k, d) = (X'X + kI)^{-1} (X'X - dI) (X'X + R'R)^{-1} X'y, k > 0, -\infty < d < +\infty.$$

令 $k = 0.05, d = -0.5$, 则 $\hat{\beta}_{SRLTE}(0.05, -0.5) = (2.17985, 1.15688, 0.7466, 0.48857)'$, 在 $k = 0.05, d = -0.5$ 时, $MSE(\hat{\beta}_{SRLTE}) = 0.0641$.

经计算得 β 的刀切岭估计及其均方误差分别为

$$\tilde{\beta}_k = (2.1930457, 1.153326, 0.7585, 0.48632)'$$

$$MSE(\tilde{\beta}_k) = \sigma^2 \sum_{i=1}^4 \frac{\lambda_i(\lambda_i + 2k)^2}{(\lambda_i + k)^4} + k^4 \sum_{i=1}^4 \frac{\gamma_i^2}{(\lambda_i + k)^4} = 0.0637,$$

其中 σ^2 取 $\hat{\sigma}_{OLS}^2$, γ_i 取 $\hat{\gamma}_i = P'\hat{\beta}$.

因此, 当 $k = 0.05, d = -0.5$ 时, $MSE(\tilde{\beta}_k) < MSE(\hat{\beta}_{SRLTE})$, 此时我们提出的刀切岭估计 $\tilde{\beta}_k$ 更优势.

此外, 置信水平为 95% (β, σ^2) 的最小体积置信集为

$$\frac{\|y - XP(I - k^2(\Lambda + kI)^{-2})^{-1}P'\tilde{\beta}_k\|^2}{13\sigma^2} - 1.1538 \log \frac{\|y - XP(I - k^2(\Lambda + kI)^{-2})^{-1}P'\tilde{\beta}_k\|^2}{13\sigma^2} +$$

$$\frac{\|(\tilde{\beta}_k - \beta)' [XP(I - k^2(\Lambda + kI)^{-2})^{-1}P' - X]' [XP(I - k^2(\Lambda + kI)^{-2})^{-1}P' - X] (\tilde{\beta}_k - \beta)\|^2}{13\sigma^2}$$

$\leq m_{0.05}$,

其中 $m_{0.05} = 2.249$, P 为 $X'X$ 的特征向量, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, λ_i 为 $X'X$ 的特征根.

令上述最小体积置信集中的 β 为 $\hat{\beta}_{SRLTE}(0.05, -0.5)$, 通过 Matlab 计算, 得到左边等式为 1.1166. 由于 $1.1166 < 2.249$, 说明 $\hat{\beta}_{SRLTE}$ 包含在基于刀切岭估计的最小体积置信集中.

参 考 文 献

- [1] Hoerl A E, Kennard R W. Ridge regression: biased estimation for nonorthogonal problems[J]. *Technometrics*, 1970, 12(1): 55-67.
- [2] Farebrother R W. Further results on the mean square error of ridge regression[J]. *Journal of the Royal Statistical Society: Series B*, 1976, 38(3): 248-250.
- [3] Firinguetti L L. A simulation study of ridge regression estimators with autocorrelated errors[J]. *Communication in Statistics-Simulation and Computation*, 1989, 18(2): 673-702.
- [4] Quenouille M H. Notes on bias in estimation[J]. *Biometrika*, 1956, 43(3-4): 353-360.
- [5] Singh B, Chaubey Y P, Dwivedi T D. An almost unbiased ridge estimator[J]. *Sankhy ā: The Indian Journal of Statistics, Series B*, 1986, 48(3): 342-346.
- [6] Kadiyala K. A class almost unbiased and efficient estimators of regression coefficients[J]. *Economics Letters*, 1984, 16(3-4): 293-296.
- [7] Ohtani K. On small sample properties of the almost unbiased generalized ridge estimator[J]. *Communications in Statistics-Theory and Methods*, 1986, 15(5): 1571-1578.
- [8] Singh B, Chaubey Y P. On some improved ridge estimators[J]. *Statistische Hefte*, 1987, 28(1): 53-67.
- [9] Nomura M. On the almost unbiased ridge regression estimator[J]. *Communications in Statistics-Simulation and Computation*, 1988, 17(3): 729-743.
- [10] Gruber M. Improving efficiency by shrinkage: the james-stein and ridge regression estimators[M]. New York: Routledge, 1998.
- [11] Hinkley D V. Jackknifing in unbalanced situations[J]. *Technometrics*, 1977, 19(3): 285-292.

- [12] Vinod H D. Confidence intervals for ridge regression parameters[J]. Springer Netherlands, 1987, 36(3): 279–300.
- [13] Chaubey Y P, Khurana M, Chandra S. Confidence intervals based on resampling methods using ridge estimator in linear regression model[J]. *New Trends in Mathematical Sciences*, 2018, 4(6): 77–86.
- [14] Firinguetti L, Bobadilla G. Asymptotic confidence intervals in ridge regression based on the edge-worth expansion[J]. *Statistical Papers*, 2011, 52(2): 287–307.
- [15] Zhang Jin. Minimum-volume confidence sets for normal linear regression models[J]. *Statistics*, 2018, 52(4): 874–884.
- [16] Nilg ü n Y. A new stochastic restricted Liu-type estimator in linear regression model[J]. *Communications in Statistics-Simulation and Computation*, 2019, 48(1): 91–108.

MINIMUM-VOLUME CONFIDENCE SETS OF PARAMETERS BASED ON JACKKNIFED RIDGE ESTIMATOR FOR LINEAR REGRESSION MODELS

GUO Tong-ge, HU Hong-chang

(*Department of Mathematics and Statistics, Hubei Normal University, Huangshi 435002, China*)

Abstract: In this paper, based on the ridge estimator, we investigate the minimum volume confidence sets of unknown parameters in the normal linear regression models. By using the Jackknife method, the minimum volume confidence sets of unknown parameters in different cases are obtained. Finally, the minimum volume confidence sets are compared with the classical confidence sets, and our confidence sets are the best in the sense of minimum volume.

Keywords: Jackknife ridge estimation; linear regression; confidence set; ridge estimator

2010 MR Subject Classification: 62J05; 62J07