文章编号:1000－2367(2019)02－0001－07 **DOI**:10.16366/j.cnki.1000－2367.2019.02.001

协方差阵扰动对 Stein 岭型主成分估计的影响分析

朱宁[a,b],黄荣臻[a],张茂军[a],邓超海[a]

(桂林电子科技大学 a 数学与计算科学学院;b 信息科技学院,广西 桂林 541004)

摘 要:针对线性回归模型中协方差阵扰动对 Stein 岭型主成分估计 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_G$ 的影响问题进行研究.证明了 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_G$ 的某种极限是数据删除模型的 Stein 岭型主成分估计;建立了 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_G$ 与 G-M 模型的 Stein 岭型主成分估计 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})$ 之间的关系;定义了度量扰动影响的距离测度 D_G,并给出了 D_G 的多种计算式;最后通过实例验证其有效性.

关键词:Stein 岭型主成分估计;协方差阵扰动模型;数据删除模型;影响分析;Cook 距离

中图分类号:O212.1 **文献标志码**:A

考察 Gauss-Markov 模型

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},E(\boldsymbol{\varepsilon})=0,\mathrm{Cov}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}_n, \tag{1}$$

其中,$\boldsymbol{Y}$ 是 $n\times 1$ 观测向量,$\boldsymbol{X}$ 是 $n\times p$ 列满秩设计矩阵,$\boldsymbol{\beta}$ 是 $p\times 1$ 未知参数向量,$\boldsymbol{\varepsilon}$ 是 $n\times 1$ 随机误差向量,σ^2 是未知参数,$\boldsymbol{I}_n$ 是 $n\times n$ 单位矩阵.

对于模型(1),在 $\boldsymbol{\beta}$ 的所有线性无偏估计中,最小二乘估计 $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ 是唯一最小方差的估计[1].但当设计阵 $\boldsymbol{X}$ 含有多重共线关系时,$\boldsymbol{X}'\boldsymbol{X}$ 接近奇异,此时利用最小二乘估计得到的参数将严重偏离实际值.为了改进最小二乘估计,提出了一系列的有偏估计.同时,在研究客观过程中,当对模型(1)做任何微小扰动时,都会引起统计推断的改变.因此,很有必要探讨模型的扰动方式,给出度量扰动对统计推断影响大小的标准.

文献[2－8]分别研究了协方差阵扰动对不同有偏估计的影响分析;文献[9]提出了一种新的有偏岭-压缩组合估计,称之为 Stein 岭型主成分估计(SRPCE);文献[10]利用几乎无偏的思想对 SRPCE 进行优化,并证明了其在均方误差准则下的优良性;文献[11－12]分别研究了 SRPCE 下单个或多个数据删除模型的影响分析.

本文探讨了模型(1)在协方差阵存在扰动时,对文献[9]提出的有偏估计——Stein 岭型主成分估计(SRPCE)的影响分析.证明了协方差阵扰动模型下 SRPCE 的某种极限是数据删除模型的 SRPCE,建立了协方差阵扰动模型下 SRPCE 与模型(1)SRPCE 之间的关系,讨论了一组或多组协方差阵扰动对 SRPCE 的影响,求得 SRPCE 的基于有偏估计的 Cook 距离.

本文讨论的协方差阵扰动模型为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},E(\boldsymbol{\varepsilon})=0,\mathrm{Cov}(\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{G}^{-1}, \tag{2}$$

其中 $\boldsymbol{G}$ 为正定阵.

本文讨论的数据删除模型为

$$\boldsymbol{Y}(\boldsymbol{J})=\boldsymbol{X}(\boldsymbol{J})\boldsymbol{\beta}+\boldsymbol{\varepsilon}(\boldsymbol{J}),E(\boldsymbol{\varepsilon}(\boldsymbol{J}))=0,\mathrm{Cov}(\boldsymbol{\varepsilon}(\boldsymbol{J}))=\sigma^2\boldsymbol{I}_{n-m}, \tag{3}$$

其中,$\boldsymbol{J}=\{j_1,j_2,\cdots,j_m\},1\le j_1\le j_2\le\cdots\le j_m\le n$,$\boldsymbol{Y}(\boldsymbol{J})$,$\boldsymbol{X}(\boldsymbol{J})$,$\varepsilon(\boldsymbol{J})$ 分别为 $\boldsymbol{Y}$,$\boldsymbol{X}$,$\boldsymbol{\varepsilon}$ 删除 $\boldsymbol{J}$ 中各行后得到的向量或矩阵[12].

当 $\boldsymbol{J}=i$ 时,有

收稿日期:2018-03-19;**修回日期**:2018-05-12.

基金项目:国家自然科学基金(71461005);广西研究生教育创新计划资助项目(YCSW2017143).

作者简介:朱宁(1957－),男,湖南宁乡人,桂林电子科技大学教授,研究方向为线性统计模型,E-mail:znqx@guet.edu.cn.

通信作者:黄荣臻(1992－),男,广西南宁人,研究方向为应用统计,E-mail:hrz_2991@163.com.

$$\boldsymbol{Y}(i)=\boldsymbol{X}(i)\boldsymbol{\beta}+\boldsymbol{\varepsilon}(i),\boldsymbol{E}(\boldsymbol{\varepsilon}(i))=0,\mathrm{Cov}(\boldsymbol{\varepsilon}(i))=\sigma^2\boldsymbol{I}_{n-1}, \tag{4}$$

其中,$\boldsymbol{Y}(i),\boldsymbol{X}(i),\boldsymbol{\varepsilon}(i)$ 为模型(1)的 $\boldsymbol{Y},\boldsymbol{X},\boldsymbol{\varepsilon}$ 中去掉第 i 行后得到的向量或矩阵[11].

引理 1[9] 在模型(1)下提出了未知参数 $\boldsymbol{\beta}$ 的 Stein 岭型主成分估计,记为:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

其中,$\boldsymbol{L}=\boldsymbol{V}(\boldsymbol{A}^{-1}-1)^{1/2}\wedge\boldsymbol{V}',\boldsymbol{V}=\mathrm{diag}\left(\left(\frac{\lambda_1}{p}\right)^{\frac{1}{2}},\left(\frac{\lambda_2}{p}\right)^{\frac{1}{2}},\cdots,\left(\frac{\lambda_n}{p}\right)^{\frac{1}{2}}\right),\boldsymbol{A}=\mathrm{diag}(\frac{\lambda_1}{\lambda_1+k},\frac{\lambda_2}{\lambda_2+k},\cdots,\frac{\lambda_r}{\lambda_r+k},\frac{\lambda_{r+1}}{1+k},\cdots,\frac{\lambda_n}{1+k}),\boldsymbol{X}=\boldsymbol{U}\wedge\boldsymbol{V}'(\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I},\boldsymbol{V}'\boldsymbol{V}=\boldsymbol{I})$.

1 模型(2)与模型(3)Stein 岭型主成分估计的关系

用 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_G$ 表示模型(2)中 $\boldsymbol{\beta}$ 的 Stein 岭型主成分估计,$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(J)}$ 表示模型(3)中 $\boldsymbol{\beta}$ 的 Stein 岭型主成分估计,因此有

定理 1 对于 m 组数据方差扰动,令 $\boldsymbol{G}=\boldsymbol{I}-\sum_{j\in\boldsymbol{J}}(1-g_j)d_jd'_j,0<g_j\leq1,j\in\boldsymbol{J},1<m\leq n$,则有 $\lim\limits_{\substack{g_j\to0^+\\ j\in\boldsymbol{J}}}\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(\boldsymbol{J})}$,其中 d_j 表示第 j 个元素为 1,其余元素都是 0 的列向量.

证明 由引理 1 给出的定义可知,Stein 岭型主成分估计是最小二乘估计的一种压缩,即存在 $c(0<c<1)$,有 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})=c\hat{\boldsymbol{\beta}}$.

模型(2)可变形为

$$\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Y}=\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{\varepsilon},\boldsymbol{E}(\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{\varepsilon})=0,\mathrm{Cov}(\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{\varepsilon})=\sigma^2\boldsymbol{I}. \tag{5}$$

(5)式满足 G-M 条件.

此时 Stein 岭型主成分估计的目标函数为

$$\begin{cases}(\hat{\boldsymbol{\beta}}_G-b)'\boldsymbol{X}'\boldsymbol{G}\boldsymbol{X}(\hat{\boldsymbol{\beta}}_G-b),\\ \|b\|^2=c^2\|\hat{\boldsymbol{\beta}}_G\|^2.\end{cases} \tag{6}$$

使方程组(6)达到最小值的解就是模型(2)的 Stein 岭型主成分估计,且 $\hat{\boldsymbol{\beta}}_G=(\boldsymbol{X}'\boldsymbol{G}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}\boldsymbol{Y}$ 表示模型(5)下的最小二乘估计[13].

同理,模型(3)的 Stein 岭型主成分估计的目标函数为

$$\begin{cases}(\hat{\boldsymbol{\beta}}_G-b)'\boldsymbol{X}'_{(\boldsymbol{J})}\boldsymbol{X}_{(\boldsymbol{J})}(\hat{\boldsymbol{\beta}}_{(\boldsymbol{J})}-b),\\ \|b\|^2=c^2\|\hat{\boldsymbol{\beta}}_{(\boldsymbol{J})}\|^2.\end{cases} \tag{7}$$

使方程组(7)达到最小值的解就是模型(3)的 Stein 岭型主成分估计,且 $\hat{\boldsymbol{\beta}}_{(\boldsymbol{J})}=(\boldsymbol{X}'_{(\boldsymbol{J})}\boldsymbol{X}_{(\boldsymbol{J})})^{-1}\boldsymbol{X}'_{(\boldsymbol{J})}\boldsymbol{Y}_{(\boldsymbol{J})}$ 表示模型(3)下的最小二乘估计.

当 $\boldsymbol{G}=\boldsymbol{I}-\sum_{j\in\boldsymbol{J}}(1-g_j)d_jd'_j$ 时,方程组(6)中可变形为

$$\begin{cases}(\hat{\boldsymbol{\beta}}_G-b)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_G-b)-\sum_{j\in\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_G-b)'(1-g_j)x_jx'_j(\hat{\boldsymbol{\beta}}_G-b),\\ \|b\|^2=c^2\|\hat{\boldsymbol{\beta}}_G\|^2,\end{cases} \tag{8}$$

其中 $\hat{\boldsymbol{\beta}}_G=(\boldsymbol{X}'\boldsymbol{X}-\sum_{j\in\boldsymbol{J}}(1-g_j)x_jx'_j)^{-1}(\boldsymbol{X}'\boldsymbol{Y}-\sum_{j\in\boldsymbol{J}}(1-g_j)x_jy'_j)$,若令 $g_j\to0^+,j\in\boldsymbol{J}$,则方程组(8)与方程组(7)等价.即 $g_j\to0^+,j\in\boldsymbol{J}$ 时,方程组(6)与方程组(7)同解.

证毕.

定理 1 说明了,当模型(2)中的 $\boldsymbol{G}=\boldsymbol{I}-\sum_{j\in\boldsymbol{J}}(1-g_j)d_jd'_j$,且 $g_j\to0^+$ 时,方差阵扰动等于删除了 $\boldsymbol{J}=\{j_1,j_2,\cdots,j_m\}$ 中的各行数据.这是因为当 $g_j\to0^+$ 时,$\mathrm{Var}(y_j)=\mathrm{Cov}(e_j)=\frac{\sigma^2}{g_j}\to\infty$,即模型(2)中自动删

掉方差很大的点，从而可得到最优估计.

推论 1　对于一组数据方差扰动，令 $\boldsymbol{G}=\boldsymbol{I}-(1-g_i)d_i d'_i$，则有 $\lim\limits_{g_i\to 0^+}\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(J)}$.

证明　考察方程组(6)，把 $\boldsymbol{G}=\boldsymbol{I}-(1-g_i)d_i d'_i$ 带入方程组中得

$$\begin{cases}(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)-(1-g_i)(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)'x'_i x_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b),\\ \|b\|^2=c^2\|\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}\|^2.\end{cases}\tag{9}$$

此时，$\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}=(\boldsymbol{X}'\boldsymbol{X}-(1-g_i)x_i x'_i)^{-1}(\boldsymbol{X}'\boldsymbol{Y}-(1-g_i)x_i y_i)$.

当 $g_j\to 0^+$ 时，方程组(9)可写成

$$\begin{cases}(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)-(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b)'x'_i x_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{G}}-b),\\ \|b\|^3=c^2\|\hat{\boldsymbol{\beta}}\|^2.\end{cases}\tag{10}$$

方程组(10)就是模型(5)删除第 i 行后所得的数据删除模型的目标函数.

证毕.

根据定理 1 和推论 1，对于删除一组或多组数据对 $\boldsymbol{\beta}$ 的 Stein 岭型主成分估计所引起的影响分析时，可以通过协方差阵扰动模型的影响分析中令 $\boldsymbol{G}=\boldsymbol{I}-\sum\limits_{j\in\boldsymbol{J}}(1-g_j)d_j d'_j$，且 $g_j\to 0^+$ 得到.

2　模型(2)与模型(1)Stein 岭型主成分估计的关系

模型(1)的 Stein 岭型主成分估计如引理 1 所示，模型(2)的 Stein 岭型主成分估计 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{G}\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{G}\boldsymbol{Y}$.

记 $\bar{\boldsymbol{G}}=\boldsymbol{I}-\boldsymbol{G}$，$\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'$ 因此有

定理 2　若 $\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}}$ 可逆，则

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\delta,\tag{11}$$

其中 $\delta=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$.

证明　注意矩阵和式求逆公式与公式 $\boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\boldsymbol{H}\bar{\boldsymbol{G}}=(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}$，有

$$\begin{aligned}&\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{G}\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{G}\boldsymbol{Y}=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L}-\boldsymbol{X}'\bar{\boldsymbol{G}}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}-\boldsymbol{X}'\bar{\boldsymbol{G}}\boldsymbol{Y})=\\&\boldsymbol{P}[(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}+(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}]\boldsymbol{X}'(\boldsymbol{I}-\bar{\boldsymbol{G}})\boldsymbol{Y}=\\&\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{Y}-\\&\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}\boldsymbol{Y}-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}\boldsymbol{Y}=\\&\hat{\boldsymbol{\beta}}(\boldsymbol{P})+\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\boldsymbol{H}\boldsymbol{Y}-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}\boldsymbol{Y}-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\\&\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}[(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}-\boldsymbol{I}]\boldsymbol{Y}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\delta.\end{aligned}$$

证毕.

下面分别讨论 m 组或一组数据方差扰动对 Stein 岭型主成分估计的变化情况.

推论 2　对于 m 组数据方差扰动，令 $\boldsymbol{G}=\boldsymbol{I}-\sum\limits_{j\in\boldsymbol{J}}(1-g_j)d_j d'_j$，$\boldsymbol{J}=\{j_1,j_2,\cdots,j_m\}$，则

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})[\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})^{-1}]\delta_{\boldsymbol{J}},$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(\boldsymbol{J})}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}})^{-1}\delta_{\boldsymbol{J}},$$

其中 $\delta_{\boldsymbol{J}}=(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}})\boldsymbol{Y}_{\boldsymbol{J}}$，$\boldsymbol{H}_{\boldsymbol{J}}=\boldsymbol{X}_{\boldsymbol{J}}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{\boldsymbol{J}}$.

证明　不失一般性，设 $\boldsymbol{J}=\{1,2,\cdots,m\}$，则

$$\boldsymbol{G}=\begin{pmatrix}\boldsymbol{G}_{\boldsymbol{J}}&0\\0&\boldsymbol{I}_{n-m}\end{pmatrix},\bar{\boldsymbol{G}}=\begin{pmatrix}\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}}&0\\0&0\end{pmatrix},\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_{\boldsymbol{J}}\\\boldsymbol{X}_{(\boldsymbol{J})}\end{pmatrix},\boldsymbol{H}=\begin{pmatrix}\boldsymbol{H}_{\boldsymbol{J}}&*\\\boldsymbol{X}_{(\boldsymbol{J})}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{(\boldsymbol{J})}&*\end{pmatrix}.$$

于是有 $\boldsymbol{X}'\bar{\boldsymbol{G}}=\boldsymbol{X}_{\boldsymbol{J}}((\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})\quad 0)$，$\boldsymbol{H}\bar{\boldsymbol{G}}=\begin{pmatrix}\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})&0\\\boldsymbol{X}_{(\boldsymbol{J})}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{(\boldsymbol{J})}(\boldsymbol{I}-\boldsymbol{G}_{\boldsymbol{J}})&0\end{pmatrix}$.因此有$(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}=$

$$\begin{pmatrix}\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})&0\\-\boldsymbol{X}_{(\boldsymbol{J})}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{(\boldsymbol{J})}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})&\boldsymbol{I}_{(\boldsymbol{J})}\end{pmatrix}^{-1}=\begin{pmatrix}[\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})]^{-1}&0\\*&\boldsymbol{I}_{n-m}\end{pmatrix}.$$

带入(11)式可得:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\delta=$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}_{\boldsymbol{J}}\left((\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})\quad -0\right)\begin{pmatrix}[\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})]^{-1} & 0\\ * & \boldsymbol{I}_{n-m}\end{pmatrix}\delta_{\boldsymbol{J}}=$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})[\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{G}_{\boldsymbol{J}})]^{-1}\delta_{\boldsymbol{J}}.$$

另一方面,根据定理 1 可知 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(\boldsymbol{J})}=\lim\limits_{\substack{g_j\to 0^+\\ j\in\boldsymbol{J}}}\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_{\boldsymbol{J}}(\boldsymbol{I}_{\boldsymbol{J}}-\boldsymbol{H}_{\boldsymbol{J}})^{-1}\delta_{\boldsymbol{J}}.$

这与文献[12]中结果相同.

证毕.

推论 3 对于一组数据方差扰动,令 $\boldsymbol{G}=\boldsymbol{I}-(1-g_i)d_id'_i$,则

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\frac{\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i(1-g_i)\delta_i}{1-(1-g_i)h_{ii}},\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(\boldsymbol{J})}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\frac{\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i\delta_i}{1-h_{ii}},$$

其中 $\delta_i=y_i-x'_i(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\boldsymbol{Y}$,$h_{ii}$ 为 $\boldsymbol{H}$ 的第 i 个对角元素,$h_{ii}=x_i(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})x'_i$.

证明 $\bar{\boldsymbol{G}}=\boldsymbol{I}-\boldsymbol{G}=(1-g_i)d_id'_i$,$(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}=[1-(1-g_i)x_i(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i]^{-1}$. 带入(11)式可得:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'\bar{\boldsymbol{G}}(\boldsymbol{I}-\boldsymbol{H}\bar{\boldsymbol{G}})^{-1}\delta=\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\frac{\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i(1-g_i)\delta_i}{1-(1-g_i)x_i(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i}=$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\frac{\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i(1-g_i)\delta_i}{1-(1-g_i)h_{ii}}.$$

再由推论 1 可得 $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{(\boldsymbol{J})}=\lim\limits_{g_i\to 0^+}\hat{\boldsymbol{\beta}}(\boldsymbol{P})_{\boldsymbol{G}}==\hat{\boldsymbol{\beta}}(\boldsymbol{P})-\frac{\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}x'_i\delta_i}{1-h_{ii}}.$

这与文献[11]中结果相同.

证毕.

定理 2 描述了一般的协方差阵扰动前后 Stein 岭型主成分估计之间的关系以及关系成立的条件,定理 2 说明了残差 δ 越大,均方差扰动对回归系数的 SRPCE 的影响越大.推论 2 和推论 3 讨论了多组或一组协方差阵扰动前后 SRPCE 之间的关系,并根据定理 1 和推论 1 的结论得出删除多组或一组数据前后 SRPCE 之间的关系.

由于 $\boldsymbol{X}$ 已经中心化,因此在推论 3 中 $h_{ii}=x_i(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})x'_i$ 表示了 x_i 到中心的某种距离.推论 3 的结论表明,x_i 越远离中心(即 h_{ii} 越接近于 1),数据 (y_i,x'_i) 对 Stein 岭型主成分估计的影响越大.

3 影响的度量

Cook 距离是由 Cook[14] 于 1977 年提出的用于分析和识别线性回归分析中强影响点的一种距离.

引理 2[3] 类似于 Cook 距离,定义协方差阵扰动对有偏估计的影响测度度量为

$$D(\boldsymbol{M},C)=\frac{(\hat{\boldsymbol{\beta}}(\boldsymbol{Z})-\hat{\boldsymbol{\beta}})'\boldsymbol{M}(\hat{\boldsymbol{\beta}}(\boldsymbol{Z})-\hat{\boldsymbol{\beta}})}{C},\tag{12}$$

其中 $\boldsymbol{M}$ 为正定阵,C 为正数,对不同的 $\boldsymbol{M}$,C 和 $\hat{\boldsymbol{\beta}}(\boldsymbol{Z})$,可给出 $D(\boldsymbol{M},C)$ 的各种表示式和统计意义.

为了方便叙述,用 $\lambda_i,\mu_i(i=1,2,\cdots,k)$ 分别记为 $\boldsymbol{H},\bar{\boldsymbol{G}}$ 的特征值,$\tau_i=\lambda_i\mu_i$,$\boldsymbol{\Gamma}'$ 是 $\boldsymbol{H}$ 的特征向量为列所成的正交矩阵,记 $\boldsymbol{A}=\boldsymbol{\Gamma}\delta$,$\hat{V}_{\boldsymbol{G}}=\frac{1}{n-q}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{G}))'\boldsymbol{G}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{G}))$,$\boldsymbol{L}=\boldsymbol{A}\hat{\boldsymbol{V}}_{\boldsymbol{G}}\boldsymbol{A}'=(l_{ij})$,其他记号的意义同前.

定理 3 模型(2)中,令 $\boldsymbol{G}=\boldsymbol{I}-\sum\limits_{j\in\boldsymbol{J}}(1-g_j)d_jd'_j$,$0<g_j\leq 1$,$\boldsymbol{J}=\{j_1,j_2,\cdots,j_m\}$,$1\leq m\leq n$,取 $\boldsymbol{M}=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})$,$C=q\hat{V}_{\boldsymbol{G}}$,若 $\boldsymbol{H}_{\boldsymbol{J}}\bar{\boldsymbol{G}}_{\boldsymbol{J}}=\bar{\boldsymbol{G}}_{\boldsymbol{J}}\boldsymbol{H}_{\boldsymbol{J}}$,则 Stein 岭型主成分估计的 Cook 距离可表示为

$$D(\boldsymbol{P})_{\boldsymbol{G}}=\frac{\boldsymbol{P}^3}{q}\sum_{i=1}^{m}\frac{\mu_i l_{ii}}{1-\tau_i}\cdot\frac{\tau_i}{1-\tau_i},\tau_i<1,i=1,2,\cdots,m,\tag{13}$$

若有某个 $\tau_i=1$,则令 $D(\boldsymbol{P})_G=\infty$.

证明　由定理 2 及 $D(\boldsymbol{P})_G$ 的定义可得

$$D(\boldsymbol{P})_G=\frac{[\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_J\bar{\boldsymbol{G}}_J(\boldsymbol{I}_J-\boldsymbol{H}_J\bar{\boldsymbol{G}}_J)^{-1}\delta_J]'\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})[\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}_J\bar{\boldsymbol{G}}_J(\boldsymbol{I}_J-\boldsymbol{H}_J\bar{\boldsymbol{G}}_J)^{-1}\delta_J]}{q\hat{\boldsymbol{V}}_G}=$$

$$\frac{\delta'_J(\boldsymbol{I}_J-\bar{\boldsymbol{G}}_J\boldsymbol{H}_J)^{-1}\bar{\boldsymbol{G}}_J\boldsymbol{X}_J\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{X}'_J\bar{\boldsymbol{G}}_J(\boldsymbol{I}_J-\boldsymbol{H}_J\bar{\boldsymbol{G}}_J)^{-1}\delta_J}{q\hat{\boldsymbol{V}}_G}=$$

$$\frac{\boldsymbol{P}^3\delta'_J(\boldsymbol{I}_J-\bar{\boldsymbol{G}}_J\boldsymbol{H}_J)^{-1}\bar{\boldsymbol{G}}_J\boldsymbol{H}_J\bar{\boldsymbol{G}}_J(\boldsymbol{I}_J-\boldsymbol{H}_J\bar{\boldsymbol{G}}_J)^{-1}\delta_J}{q\hat{\boldsymbol{V}}_G}. \tag{14}$$

注意到 $\boldsymbol{H}_J\bar{\boldsymbol{G}}_J=\bar{\boldsymbol{G}}_J\boldsymbol{H}_J$,故存在正交阵 $\boldsymbol{\Gamma}$,使得 $\boldsymbol{H}_J=\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$,$\bar{\boldsymbol{G}}_J=\boldsymbol{\Gamma}'\boldsymbol{U}\boldsymbol{\Gamma}$.

其中 $\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_m\}$,$0<\lambda_i\leq 1$,$\boldsymbol{U}=\mathrm{diag}\{\mu_1,\mu_2,\cdots,\mu_m\}$,$0<\mu_i\leq 1$.

于是 $D(\boldsymbol{P})_G$ 可简化为

$$D(\boldsymbol{P})_G=\frac{\boldsymbol{P}^3}{q}\boldsymbol{A}'\mathrm{diag}\left(\frac{\lambda_1\mu_1^2}{(1-\lambda_1\mu_1)^2},\frac{\lambda_2\mu_2^2}{(1-\lambda_2\mu_2)^2},\cdots,\frac{\lambda_m\mu_m^2}{(1-\lambda_m\mu_m)^2}\right)\boldsymbol{A}\hat{\boldsymbol{V}}_G^{-1}=$$

$$\frac{\boldsymbol{P}^3}{q}\mathrm{diag}\left(\frac{\lambda_1\mu_1^2}{(1-\lambda_1\mu_1)^2},\frac{\lambda_2\mu_2^2}{(1-\lambda_2\mu_2)^2},\cdots,\frac{\lambda_m\mu_m^2}{(1-\lambda_m\mu_m)^2}\right)\boldsymbol{A}\hat{\boldsymbol{V}}_G^{-1}\boldsymbol{A}'=$$

$$\frac{\boldsymbol{P}^3}{q}\sum_{i=1}^{m}\frac{\lambda_i\mu_i^2l_{ii}}{(1-\lambda_i\mu_i)^2}=\frac{\boldsymbol{P}^3}{q}\sum_{i=1}^{m}\frac{\mu_il_{ii}}{1-\tau_i}\cdot\frac{\tau_i}{1-\tau_i}.$$

若有某个 $\tau_i=1$,则令 $D(\boldsymbol{P})_G=\infty$.

证毕.

定理 4　在定理 3 中,若取 $\boldsymbol{M}=\boldsymbol{P}(\boldsymbol{X}'_J\boldsymbol{X}_J)$,其他条件不变,则有

$$D(\boldsymbol{P})_G=\frac{\boldsymbol{P}^3}{q}\sum_{i=1}^{m}\frac{\tau_il_{ii}}{1-\tau_i}\cdot\frac{\tau_i}{1-\tau_i},\tau_i<1,i=1,2,\cdots,m, \tag{15}$$

若有某个 $\tau_i=1$,则令 $D(\boldsymbol{P})_G=\infty$.

证明方法类似于定理 3 的证明.

由(12)式关于 $D(\boldsymbol{M},C)$ 的定义式,有 $D(\boldsymbol{P}(\boldsymbol{X}'_J\boldsymbol{X}_J),q\hat{V})=\frac{\boldsymbol{P}}{q}\frac{(\boldsymbol{X}_J\hat{\boldsymbol{\beta}}(\boldsymbol{Z})-\boldsymbol{X}_J\hat{\boldsymbol{\beta}})'(\boldsymbol{X}_J\hat{\boldsymbol{\beta}}(\boldsymbol{Z})-\boldsymbol{X}_J\hat{\boldsymbol{\beta}})}{\hat{\boldsymbol{V}}}$,可知,当 $\boldsymbol{M}=\boldsymbol{P}(\boldsymbol{X}'_J\boldsymbol{X}_J)$ 时,$D(\boldsymbol{M},C)$ 度量了协方差阵扰动时 $\boldsymbol{X}_J$ 处对 $\boldsymbol{Y}$ 的预测影响大小.

由定理 3 和定理 4 可知,对不同的 m,$\boldsymbol{M}$,C 可以得到不同的 Cook 距离,根据文献[3,15]的方法得出协方差阵扰动下 Stein 岭型主成分估计的一些常见的 Cook 距离的具体形式,这里不一一列举.

4　实例分析

为了讨论本文定义的 Cook 距离在检测异常点的效果,选择文献[1]中的外贸数据来做分析,该组数据存在着共线性.把样本数据 3 组合为 1 组,并求得各组的协方差阵分别为:

$$\mathrm{Cov}_1=\begin{pmatrix}123.4&-5.98&83.58\\-5.98&0.37&-4.28\\83.58&-4.28&57.24\end{pmatrix},\mathrm{Cov}_2=\begin{pmatrix}59.52&-2.23&41.19\\-2.23&1.00&-2.33\\41.19&-2.33&29.17\end{pmatrix},\mathrm{Cov}_3=\begin{pmatrix}145.0&16.24&97.83\\16.24&3.50&11.78\\97.83&11.78&66.42\end{pmatrix}$$

比较可得 3 组协方差阵均不相等[16],说明该组数据的协方差阵出现扰动.

为此,本文主要在 Stein 岭型主成分估计下研究协方差阵扰动模型的拟合程度,以及实验数据中某组数据的影响程度,并根据文章中所推断出来的统计量检测异常点的效果,结果如下:

首先根据岭迹法(见图 1)以及文献[9]给出的偏参数 $\boldsymbol{P}$ 的最优值计算公式,求得 $k=0.04$,$\boldsymbol{P}=0.634$(其中 $\hat{\alpha}_i^2$ 和 σ^2 用最小二乘估计的参数来代替).其次计算出 δ',$\hat{\boldsymbol{V}}$ 分别为

$$\delta'=\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{H})(-1.64,-2.61,-0.51,-0.72,-0.47,-0.63,0.53,1.75,1.95,0.49,0.20),$$

$$\hat{\boldsymbol{V}}=\frac{1}{n-q}\boldsymbol{Y}'(\boldsymbol{I}_n-\boldsymbol{H})\boldsymbol{Y}=2.30.$$

于是求得表 1 中的 Cook 距离 D_1,D_2 和 D_3(见表 2).D_i 越大,说明(y_i,x'_i)对$(\hat{\boldsymbol{\beta}}(\boldsymbol{P}),\hat{\sigma}^2)$的影响越大.从表1可以看出,三种不同的影响度量均指示第11号观测值的影响最大.事实上,第11号观测值的杠杆值h_{ii}也是最大的,说明该点距离数据集中心最远,这是造成第 11 号观测点影响最大的原因.由 D_i 的计算式可知,D_i 值综合了杠杆值h_{ii}与学生化残差(r_i 或 r_i^*)的模的大小,因此用 D_i 值来度量影响程度是合理的.

在表 1 中 $D_1=D(\boldsymbol{M}_1,q\hat{V})$,$D_2=D(\boldsymbol{M}_1,q\hat{V}(i))$,$D_3=D(\boldsymbol{M}_2,q\hat{V}(i))$,其中 $\boldsymbol{M}_1=\boldsymbol{P}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{L}'\boldsymbol{L})$,$\boldsymbol{M}_2=\boldsymbol{P}(\boldsymbol{X}'_J\boldsymbol{X}_J)$.

表 1 外贸数据的影响度量

Tab.1 The impact measurements of foreign trade data

No.	h_{ii}	D_1	D_2	D_3
1	0.162	0.020	0.183	0.030
2	0.179	0.022	0.203	0.036
3	0.072	0.008	0.081	0.006
4	0.103	0.012	0.117	0.012
5	0.477	0.093	0.540	0.258
6	0.161	0.020	0.182	0.029
7	0.210	0.027	0.237	0.049
8	0.297	0.043	0.336	0.100
9	0.174	0.021	0.198	0.034
10	0.369	0.060	0.418	0.154
11	0.795	0.396	0.901	0.716

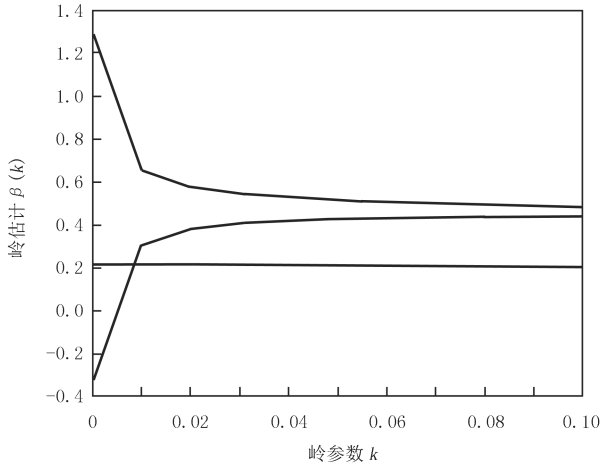
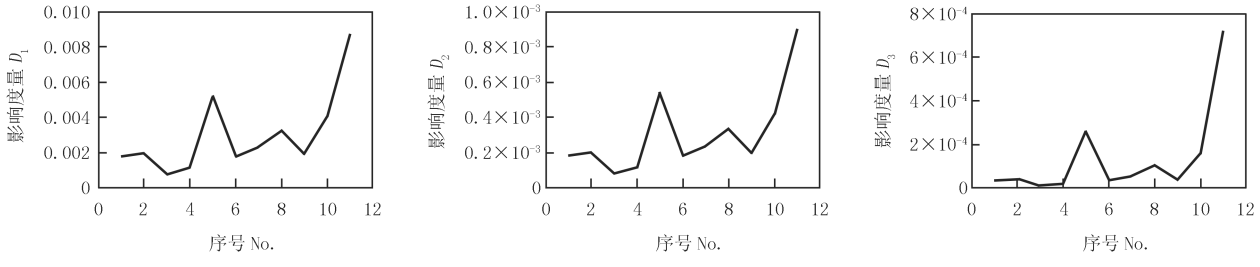


图 1 外贸数据回归岭迹图

Fig.1 Regression ridge trace of foreign trade data

以 No.为横坐标,分别以 D_1,D_2 和 D_3 为纵坐标,用表 1 中的数值做影响图(见图 2),从图 2 中可以直观看出第 11 号观测点是影响最大的点.



(a) 以 D_1 为影响度量的影响图 (b) 以 D_2 为影响度量的影响图 (c) 以 D_3 为影响度量的影响图

图 2 不同影响度量的影响图

Fig.2 Influence diagrams of different influence measures

5 结 论

本文研究了线性统计模型的影响分析问题.证明了协方差阵扰动下的 Stein 岭型主成分估计与 G-M 模型下的 Stein 岭型主成分估计和数据删除模型下的 Stein 岭型主成分估计存在关系,三者可以通过表达式互相表示;并给出了度量扰动影响的 Cook 距离.

通过上面的实例,可以得出结论:第 11 号观测点是强影响点,表 1 中 3 种度量值对数据的影响方面总体效果具有一致性,都可以作为判断强影响点的度量,对诊断数据是否为强影响点具有统计意义.

参 考 文 献

[1] 王松桂.线性统计模型[M].北京：高等教育出版社，1999：34-40.

[2] 韦博成，鲁国斌，史建清.统计诊断引论[M].南京：东南大学出版社，1991：38-78.

[3] 朱秀娟，李宁.协方差阵扰动模型的影响分析[J].应用概率统计，1993(4)：356-370.

[4] 林路.协方差阵扰动模型岭估计的影响分析[J].工程数学学报，1995，12(3)：83-88.

[5] Zhang S L，Qin H.Influence analysis of covariance matrix disturbance in a linear model with respect to a restricting condition[J].Acta Mathematica Scientia，2006，26(26)：621-628.

[6] Zhang S L，Qin H.Influence analysis on a ridge estimator in a multivariate regression model with covariance matrix disturbance[J].J Math，2010(1)：157-162.

[7] 杨莲.几类统计模型的局部影响分析研究[D].重庆：重庆大学，2014.

[8] 赵帅.带约束线性模型 LIU 估计的影响分析[D].北京：北京交通大学，2015.

[9] 朱宁，李建军，李兵.一种有偏岭-压缩组合估计的新形式[C]//第八届中国青年运筹信息管理学者大会论文集.桂林：[出版者不详]，2006：287-290.

[10] 朱宁，刘庆华，周桂兰，等.均方误差准则下的几乎无偏 Stein 岭型主成分估计的优良性[J].河南师范大学学报(自然科学版)，2017，45(5)：1-6.

[11] 朱宁，严冠东.Stein 岭型主成分估计下的单个数据删除模型的研究[J].统计与决策，2015(14)：16-18.

[12] 朱宁，严冠东，刘庆华.Stein 岭型主成分估计下多个数据删除模型的强影响分析[J].汕头大学学报(自然科学版)，2015，30(2)：20-27.

[13] 王松桂.线性模型的理论及其应用[M].合肥：安徽教育出版社，1987：165-172.

[14] Cook R D.Detection of Influential Observation in Linear Regression[J].Technometrics，1977，42(1)：65-68.

[15] 王松桂.线性回归诊断(I)[J].数理统计与管理，1985(6)：38-49.

[16] 同济大学数学系.工程数学线性代数[M].北京：高等教育出版社，2014：29-33.

Influence analysis of covariance matrix disturbance on stein ridge type principal component estimator

Zhu Ning[a,b]，Huang Rongzhen[a]，Zhang Maojun[a]，Deng Chaohai[a]

(a.School of Mathematics and Computing Science；b.Institute of Information Technology of Guet，Guilin University of Electronic Science and Technology，Guilin 541004，China)

Abstract：In this paper，the issue of influence analysis of covariance matrix disturbance on stein ridge type principal components estimator (SRPCE) in linear regression model is studied. We prove that in the data deletion model，some limit of SRPCE which in the regression model with covariance matrix disturbance is SRPCE. Then，we set up the relationships among $\hat{\boldsymbol{\beta}}(\boldsymbol{P})_G$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{P})$. Next，we define the distance measure D_G，which can be assessed the disturbing influence. Afterwards，we give several calculation formulas of D_G. Finally，a practical example is presented to illustrate the effectiveness of this method.

Keywords：Stein ridge type principal components estimator；covariance matrix disturbance；data deletion model；influence analysis；Cook distance

[责任编校　陈留院]