Advances in Applied Mathematics 应用数学进展, 2021, 10(4), 923-930
Published Online April 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.104100

Hans汉斯

基于Pena距离的KL估计的影响分析

王孟孟，田维琦

贵州民族大学数据科学与信息工程学院，贵州 贵阳
Email: 1418277781@qq.com

收稿日期：2021年3月14日；录用日期：2021年4月3日；发布日期：2021年4月16日

摘　要

利用Pena距离对KL估计的影响分析进行讨论，得到了KL估计的Pena统计量的表达式，并对其性质进行讨论分析，从而得到高杠异常点的判别方法。本文对Pena统计量与Cook统计量的性质进行了比较，得出在一定条件下Pena统计量是优于Cook统计量的结论。通过实例对比分析，得到研究结果表明本文提出的理论和方法是科学合理的。

关键词

KL估计，Pena距离，影响分析

Influence Analysis of KL Estimation Based on Pena Distance

Mengmeng Wang, Weiqi Tian

School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou
Email: 1418277781@qq.com

Received: Mar. 14th, 2021; accepted: Apr. 3rd, 2021; published: Apr. 16th, 2021

Abstract

In this paper, the influence analysis of KL estimation is discussed on the Pean distance. The expression of Pena statistics of KL estimation is obtained. The properties of Pena statistics are discussed and analyzed; meanwhile the discrimination of high-leverage outlier is obtained. In this paper, the properties of Pena statistic and Cook statistic are compared, and it is concluded that Pena statistic is better than Cook statistic under certain conditions. Through the example analysis, the research results show that the theory and method proposed in this paper are scientific and reasonable.

文章引用: 王孟孟, 田维琦. 基于 Pena 距离的 KL 估计的影响分析[J]. 应用数学进展, 2021, 10(4): 923-930.
DOI: 10.12677/aam.2021.104100

Keywords

KL Estimates, Pena Distance, Influence Analysis

1. 引言

在统计学中，统计诊断是数据分析的第一步，主要目的就是对样本数据中异常点或强影响点的识别和诊断，传统的判断异常点或强影响点的常用统计量有 Cook 距离、似然距离、W-K 统计量和 AP 统计量等。美国统计学家 Daniel Pena [1]于 2005 年提出的一种新的诊断统计量 Pena 距离，该统计量是对诊断统计量的重要补充，Pena 距离是一种度量线性回归模型影响的新方法，这种方法与传统的诊断方法有较大的区别。之前的方法是研究删除一个点(组)对回归分析的影响及对模型预测值的影响，或是某个样本点(组)的微小扰动对参数估计的影响及对模型预测的影响；而 Pena 距离这一统计量是研究样本中的某一点受其余各点的影响，也即度量样本中各点删除对某一特定样本点回归值及预测值的影响。孟丽丽等[2]基于 Pena 距离研究了加权最小二乘估计的影响分析，胡江等[3] [4] [5] [6]基于 Pena 距离研究了非线性回归模型、广义线性回归模型和 t-回归模型的影响分析；Semra Türkan 等[7]研究了基于 Pena 距离的岭估计和改进岭估计的影响分析；Hadi Emami 等[8]研究了基于 Pena 距离的岭估计的影响分析；Muhammad Kashif 等[9] [10]研究了基于 Pena 距离的 Liu 估计和改进岭估计的影响分析。

本文将 Pena 统计量推广到 Kibria-Lukman 估计的影响分析问题，给出 Kibria-Lukman 估计影响分析的 Pena 统计量的表达式，并对其性质进行了讨论，从而得到高杠异常点的判别方法。在一定条件下对 Pena 统计量与 Cook 统计量的性能进行了比较分析，并通过实例分析对该方法的有效性进行验证。

2. 理论基础

考虑一般线性回归模型：

$$y = X\beta + \varepsilon \tag{1}$$

其中 $y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，$\beta=(\beta_0,\beta_1,\cdots,\beta_{p-1})^{\mathrm{T}}$，$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$，$\varepsilon\sim N(0,\sigma^2 I)$，$\sigma^2>0$，$I$ 是 n 阶单位矩阵，X 为 $n\times p$ 阶的已知设计矩阵，其中第 i 行为 $(1,x_{i1},x_{i2},\cdots,x_{ip-1})$。当模型(1)满足高斯－马尔可夫条件时，此时最小二乘估计(OLS) $\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ 为 β 的最佳线性无偏估计。$\hat{y}=X\hat{\beta}=Hy$，其中 $H=X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ 为对角元素 $h_{ii}=x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i$ 的帽子矩阵，$s^2=\dfrac{e^{\mathrm{T}}e}{n-p}$ 是 σ^2 的无偏估计，其中 $e=y-\hat{y}=y-X\hat{\beta}=(I-H)y$。

当解释变量存在复共线性时，最小二乘估计往往表现出不稳定性，其优良性就会被破坏，若再基于最小二乘估计方法做影响分析很明显是不合适。为了解决这个问题，B. M. Golam Kibria 和 Adewale F. Lukman [11]提出了一种新的 Ridge-Type 估计，它称为 Kibria-Lukman (KL)估计，该估计是在岭估计和刘估计类中提出了一种新的单参数估计，因此它具有岭估计和刘估计的很多特征。其 KL 估计的表示如下：

$$\hat{\beta}_{KL}=(X^{\mathrm{T}}X+\lambda I_p)^{-1}(X^{\mathrm{T}}X-\lambda I_p)\hat{\beta}=W(k)M(k)\hat{\beta}$$

其中 λ 是非负常数，$W(k)=\left[I_p+\lambda\left(X^{\mathrm{T}}X\right)^{-1}\right]^{-1}$，$M(k)=\left[I_p-\lambda\left(X^{\mathrm{T}}X\right)^{-1}\right]$。

根据 Daniel Pena [1]提出的 Pena 距离，得出基于 KL 估计的 Pena 距离可表示为：

$$S_{KL,i}=\frac{s_{KL,i}^{\mathrm{T}}s_{KL,i}}{p\sigma^2_{(\hat{y}_{KL,i})}},\quad i=1,2,\cdots,n \tag{2}$$

其中 $s_{KL,i}=\left(\hat{y}_{KL,i}-\hat{y}_{KL,i(-1)},\hat{y}_{KL,i}-\hat{y}_{KL,i(-2)},\cdots,\hat{y}_{KL,i}-\hat{y}_{KL,i(-n)}\right)^{\mathrm{T}}$，$\hat{y}_{KL,i}-\hat{y}_{KL,i(-j)}=\dfrac{h_{KL,ji}e_{KL,i}}{1-h_{KL,jj}}$，$\sigma^2_{\hat{y}_{KL,i}}=\hat{\sigma}^2h_{KL,ii}$，其中 $e_{KL,i}=y_{KL,i}-\hat{y}_{KL,i}$，$\hat{y}_{KL,i}$ 为第 i 个点 $y_{KL,i}$ 的拟合值，$\hat{y}_{KL,i(-j)}$ 是删除第 j 个点后第 i 个点的拟合值，$h_{KL,ii}$ 是 $H_{KL}=X\left(X^{\mathrm{T}}X+\lambda I_p\right)^{-1}\left(X^{\mathrm{T}}X-\lambda I_p\right)\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$ 的对角元素。因此，对于公式(2)又可以重新表示为：

$$S_{KL,i}=\frac{1}{p\hat{\sigma}^2h_{KL,ii}}\sum_{j=1}^{n}\frac{h^2_{KL,ji}e^2_{KL,j}}{\left(1-h_{KL,jj}\right)^2} \tag{3}$$

其中，$\hat{\sigma}^2$ 的估计为 $s^2=\dfrac{e_{KL}^{\mathrm{T}}e_{KL}}{n-p}$。

定理 2.1 当样本中不含异常点时，有 $E\left(S_{KL,i}\right)\to\dfrac{1}{p}\ (n\to\infty)$

证明 因为 $e_{KL,j}=\left(1-h_{KL,jj}\right)y_{KL,j}$，所以 $\mathrm{var}\left(e_{KL,j}\right)=E\left(e^2_{KL,j}\right)=\left(1-h_{KL,jj}\right)\hat{\sigma}^2$，故

$$E\left(S_{KL,i}\right)=\frac{1}{p\hat{\sigma}^2h_{KL,ii}}\sum_{j=1}^{n}\frac{h^2_{KL,ji}E\left(e^2_{KL,j}\right)}{\left(1-h_{KL,jj}\right)^2}=\frac{1}{ph_{KL,ii}}\sum_{j=1}^{n}\frac{h^2_{KL,ji}}{1-h_{KL,jj}}$$

记 $h^*=\max\limits_{1\le i\le n}h_{KL,ii}$，对于上式有

$$E\left(S_{KL,i}\right)\le\frac{1}{p\left(1-h^*\right)}=\frac{1}{p}+\frac{h^*}{p\left(1-h^*\right)}\to\frac{1}{p}\left(h^*\to0\right)$$

而当 $h_{KL,jj}\ge\dfrac{1}{n}$，有

$$E\left(S_{KL,i}\right)=\frac{1}{ph_{KL,ii}}\sum_{j=1}^{n}\frac{h^2_{KL,ji}}{1-h_{KL,jj}}\ge\frac{1}{p\left(1-\dfrac{1}{n}\right)}\to\frac{1}{p}\left(n\to\infty\right)$$

定理 2.1 表明：当 $h^*\to0$，$n\to\infty$ 时，所有样本点的数学期望影响趋于 $\dfrac{1}{p}$，即 $E\left(S_{KL,i}\right)\to\dfrac{1}{p}$，所以当样本点的期望和 $\dfrac{1}{p}$ 相差很大时，可以判断出样本点中的异常点。与 Cook 距离相比较，Pena 距离是优于 Cook 距离的，因为 Cook 距离的数学期望为 $\dfrac{h_{KL,ii}}{p\left(1-h_{KL,ii}\right)}$，其数学期望是依赖于 $h_{KL,ii}$，随 $h_{KL,ii}$ 的变化而变化。

定理 2.2 在 $h_{KL,ii}$ 很小的样本中，当 $n\to\infty$，$p\to\infty$，但 $\dfrac{p}{n}\to0$，则 $S_{KL,i}$ 的分布近似于正态分布。

证明 利用中心极限定理，假设没有异常值，$h^*=\max\limits_{1\le i\le n}h_{KL,ii}\le c\bar{h}$，$c>0$，其中 $\bar{h}=\sum\limits_{i=1}^{n}\dfrac{h_{KL,ii}}{n}$，当 $n\to\infty$，

$p\to\infty$，但 $\frac{p}{n}\to 0$，对于公式(3)可以写成 $S_{KL,i}=\sum_{j=1}^{n}m_{ij}\left(\frac{e_{KL,j}}{\hat{\sigma}}\right)^2$，其中 $m_{ij}=\frac{h_{KL,ji}^2}{ph_{KL,ii}\left(1-h_{KL,jj}\right)^2}$，$e_{KL,j}$ 是 $\mathrm{cov}\left(e_{KL,j}\right)=\sigma^2\left(I-2H_{KL}+H_{KL}H_{KL}^{\mathrm{T}}\right)$ 的正态随机变量。因此，当 $n\to\infty$，$h_{KL,ij}\to 0$，$S_{KL,i}$ 是自由度为 1 的卡方独立变量的加权组合。$m_{ij}>0$，下证 $\frac{m_{ij}}{\sum_{j=1}^{n}m_{ij}}\to 0$。

因为

$$m_{ij}\le\frac{h_{KL,jj}}{p\left(1-h_{KL,jj}\right)^2}\approx\frac{h_{KL,jj}\left(1+2h_{KL,jj}\right)}{p},$$

则有

$$\frac{m_{ij}}{\sum_{j=1}^{n}m_{ij}}\le\frac{h_{KL,jj}\left(1+2h_{KL,jj}\right)}{p+2\sum_{j=1}^{n}h_{KL,jj}^2}\le\frac{h_{KL,jj}\left(1+2h_{KL,jj}\right)}{p}$$

所以，当 $p\to\infty$，$\frac{m_{ij}}{\sum_{j=1}^{n}m_{ij}}\to 0$，故在这些假设下，$S_{KL,i}$ 的分布是近似于正态分布。

定理 2.2 表明：对于大样本和解释变量比较多时，Pena 距离 $S_{KL,i}$ 的分布近似于正态分布，而 Cook 距离的分布[12]是偏态分布，由此可知，Pena 距离的这一性质是优于 Cook 距离。

$S_{KL,i}$ 的这一性质，对于大样本和解释变量比较多时，$S_{KL,i}$ 的分布将近似于正态分布，其截断点可以通过这一性质来寻找。因此，当样本观测点的值远远大于 $\frac{S_{KL,i}-E\left(S_{KL,i}\right)}{SD\left(S_{KL,i}\right)}$，则该样本观测点就可以视为异常点或影响点，但是当样本中存在异常点或影响点时，$S_{KL,i}$ 的均值和标准差很容易受影响。于是 Daniel Pena [1]提出使用 Pena 距离的中位数和中位数绝对偏差来代替均值和标准差。所以，如果 $S_{KL,i}$ 满足

$$\left|S_{KL,i}\right|\ge Median\left(S_{KL,i}\right)+4.5MAD\left(S_{KL,i}\right)\tag{4}$$

则称该样本点是异常点或影响点。

其中 $MAD\left(S_{KL,i}\right)=\frac{Median\left\{\left|S_{KL,i}-Median\left(S_{KL,i}\right)\right|\right\}}{0.645}$ 是正态数据标准差的稳健估计量，$Median\left(S_{KL,i}\right)$ 是 $S_{KL,i}$ 的中位数。

下面考虑 Pena 统计量 $S_{KL,i}$ 在含有一组相同的高杠异常点的样本点中具有的性质。

设有 n 个样本点 $\left(y_1,x_1^{\mathrm{T}}\right),\left(y_1,x_2^{\mathrm{T}}\right),\cdots,\left(y_1,x_n^{\mathrm{T}}\right)$，记 $X_0^{\mathrm{T}}=\left(x_1,x_2,\cdots,x_n\right)$，$y_0^{\mathrm{T}}=\left(y_1,y_2,\cdots,y_n\right)$，$\hat{\beta}_{KL(0)}=\left(X_0^{\mathrm{T}}X_0+kI\right)^{-1}\left(X_0^{\mathrm{T}}y_0-k\left(X_0^{\mathrm{T}}X_0\right)^{-1}X_0^{\mathrm{T}}y_0\right)$，$u_i=y_i-x_i^{\mathrm{T}}\hat{\beta}_{KL0}$。假设在样本点中含有 k 个相同的高杠异常点 $\left(y_a,x_a^{\mathrm{T}}\right)$，令 $u_a=y_a-x_a^{\mathrm{T}}\hat{\beta}_{KL(0)}$，$X_T^{\mathrm{T}}=\left(X_0^{\mathrm{T}},x_a1_k^{\mathrm{T}}\right)$，$y_T^{\mathrm{T}}=\left(y_0,y_a1_k^{\mathrm{T}}\right)$，其中 $1_k^{\mathrm{T}}=\overbrace{(1,1,\cdots,1)}^{k个}$，$e_{KL,i}=y_i-x_i^{\mathrm{T}}\hat{\beta}_{KL(T)}$，$\hat{\beta}_{KL(T)}=\left(X_T^{\mathrm{T}}X_T+\lambda I\right)^{-1}\left(X_T^{\mathrm{T}}y_T-\lambda\left(X_T^{\mathrm{T}}X_T\right)^{-1}X_T^{\mathrm{T}}y_T\right)$，

$H_{KL(T)}=X_T\left(X_T^{\mathrm{T}}X_T+\lambda I\right)^{-1}\left(X_T^{\mathrm{T}}X_T-\lambda I\right)\left(X_T^{\mathrm{T}}X_T\right)^{-1}X_T^{\mathrm{T}}$ 是对 $n+k$ 个样本点的投影矩阵，其中 $H_{KL(T)}$ 的元素

记为 $h_{KL(ij)}$，记 $H_{KL(0)}=X_0\left(X_0^{\mathrm{T}}X_0+\lambda I\right)^{-1}\left(X_0^{\mathrm{T}}X_0-\lambda I\right)\left(X_0^{\mathrm{T}}X_0\right)^{-1}X_0^{\mathrm{T}}$ 是对 n 个正常点的投影矩阵，其中的元素记为 $h_{KL(ij)}^0$。

设投影矩阵 $H_{KL(T)}$ 的分块形式为：$H_{KL(T)}=\begin{bmatrix}H_{KL(T)11} & H_{KL(T)12}\\ H_{KL(T)21} & H_{KL(T)22}\end{bmatrix}$，则有 $H_{KL(T)11}$，$H_{KL(T)12}$，$H_{KL(T)22}$ 分别是 $n\times n$，$n\times k$，$k\times k$ 矩阵，$H_{KL(T)21}$ 是 $H_{KL(T)12}$ 的转置矩阵，并且有

$$H_{KL(T)11}=H_{KL(0)}-\frac{k}{kh_{KL(a)}^0+1}h_{KL(1a)}^0\left(h_{KL(1a)}^0\right)^{\mathrm{T}},\tag{5}$$

其中 $h_{KL(a)}^0=x_a^{\mathrm{T}}\left(X_0^{\mathrm{T}}X_0+\lambda I\right)^{-1}\left(X_0^{\mathrm{T}}X_0-\lambda I\right)\left(X_0^{\mathrm{T}}X_0\right)^{-1}x_a$，

$$h_{KL(1a)}^0=X_0\left(X_0^{\mathrm{T}}X_0+\lambda I\right)^{-1}\left(X_0^{\mathrm{T}}X_0-\lambda I\right)\left(X_0^{\mathrm{T}}X_0\right)^{-1}x_a.$$

同理可求得

$$H_{KL12}=H_{KL21}=\frac{1}{kh_{KL(a)}^0+1}h_{KL(1a)}^0 1_k^{\mathrm{T}},\tag{6}$$

$$H_{KL22}=\frac{1}{kh_{KL(a)}^0+1}h_{KL(a)}^0 1_k 1_k^{\mathrm{T}}\tag{7}$$

又因为 $e_{KL,i}=y_i-x_i^{\mathrm{T}}\hat{\beta}_{KL(T)}$，$u_{KL,i}=y_i-x_i^{\mathrm{T}}\hat{\beta}_{KL(0)}$

所以

$$e_{KL,i}=u_{KL,i}-kh_{KL(ia)}u_{ia},\ i=1,2,\cdots,n,\tag{8}$$

对于异常点的 $e_{KL,a}$ 有

$$e_{KL,a}=\frac{1}{kh_{KL(a)}^0+1}u_a,\ i=1,2,\cdots,n\tag{9}$$

对于正常点，利用公式(8)，有 Cook 距离为：

$$D_{KL,i}=\frac{\left(u_i-kh_{KL(ia)}u_a\right)^2h_{KL,ii}}{p\hat{\sigma}^2\left(1-h_{KL,ii}\right)^2}\tag{10}$$

对于异常点，利用公式(8)，有 Cook 距离为：

$$D_{KL(ia)}=\frac{u_a^2h_{KL(a)}}{p\hat{\sigma}^2\left(1+(k-1)h_{KL(a)}\right)^2\left(1+kh_{KL(a)}\right)}\tag{11}$$

假设样本中有高杠异常点，即由 $h_{KL(a)}^0\to\infty$，则有 $H_{KL12}=H_{KL21}\to 0$，由此可得 $h_{KL(ja)}\to 0,(j=1,2,n)$，$H_{KL22}\to\frac{1_k1_k^{\mathrm{T}}}{k}$，即 H_{KL22} 中的元素 $h_{KL(a)}\to\frac{1}{k}$。

又因为

$$\alpha_{ja}^2=\frac{h_{KL,(ja)}^2}{h_{KL,jj}h_{KL(a)}}\to\begin{cases}0, & j=1,2,\cdots,n\\ \frac{k}{n}, & j=n+1,n+2,\cdots,n+k\end{cases}$$

所以对于正常点，有

$$S_{KL,i}=\sum_{j=1}^{n}\alpha_{ji}^{2}D_{KL,j}\ ,\ \ i=1,2,\cdots,n \tag{12}$$

而对于异常点，有

$$S_{KL,i}=\frac{k^{2}}{n}D_{KL(a)},\ \ i=n+1,n+2,\cdots,n+k \tag{13}$$

综上所述可以得到：对于正常点的样本点，当 $h_{KL(a)}^{0}\to\infty$，$h_{KL(ja)}\to 0$ 时，利用公式(9)，有

$$e_{KL,i}\to u_{i},\ \ E\left(S_{KL,i}\right)\to\frac{1}{p}$$

对于异常点，当 $h_{KL(a)}^{0}\to\infty$，$e_{KL,a}\to 0$，$D_{KL(ia)}\to 0$，$S_{KL,i}\to 0$。即有

定理 2.3 当样本中含有高杠杆异常点时，Pena 统计量 $S_{KL,i}$ 的数学期望，有

$$E\left(S_{KL,i}\right)\to\begin{cases}0, & \text{高杠杆异常点}\\ \dfrac{1}{p}, & \text{正常点}\end{cases}$$

定理 2.3 表明：当数据中包含有一群相同的高杠异常点时，可以根据 $S_{KL,i}$ 的值很容易把它们识别出来，而这一点 Cook 距离是不能做到的。

3. 实证分析

案例数据来自文献 Longley 数据集[13]，是强共线性的宏观经济数据，其中包含 GNP deflator (GNP 平减指数)、GNP (国民生产总值)、Unemployed (失业率)、Armed Forces (武装力量)、Population (14 岁以上的非机构人口)、year (年份)，Employed (就业率)。回归模型(1)给出如下：

$$y=X\beta+\varepsilon$$

其中 $X=\left(x_{1},x_{2},x_{3},x_{4},x_{5},x_{6}\right)$，$y$ 是就业率，x_{1} 是 GNP 平减指数，x_{2} 是国民生产总值，x_{3} 是失业率，x_{4} 是武装力量，x_{5} 是 14 岁以上的非机构人口，x_{6} 是年份。该数据集的条件数为 43,275 [14]，则说明该数据集回归变量之间存在严重的多重共线性。

Cook [15]使用该数据集基于数据删除法得到最小二乘估计的 Cook 距离，识别出样本点 5，16，4，10 和 15 为影响点，Walker 和 Birch [16]基于岭估计的数据删除法使用 Cook 距离、W-K 统计量、杠杆值和残差，识别出最有影响的五个点，即点 16、10、4、15 和 5，Jahufer 和 Jianbao [17]基于改进岭估计得到 Cook 距离和 W-K 统计量，确定点 10、4、15、16 和 1 为影响点，Semra Türkan 等[7]基于岭估计和改进岭估计得到 Pena 统计量，当 $k=0$ 时，识别出影响点为 5、16、6、15 和 10，当 $k=0.0002$ 时，岭估计识别出的影响点为 16、15、10、4 和 1，改进岭估计识别出的影响点为 16、15、5、4 和 10；Kashif 等[9]基于 Liu 估计得到 Pena 统计量，当 $d=0.1$ 时，识别出的影响点为 3、10、11、4 和 5，$d=0.5$ 时，识别出的影响点为 10、3、4、15 和 16，$d=0.9$ 时，识别出的影响点为 10、4、15、5 和 16，$d=1$ 时，识别出的影响点为 15、10、5、4 和 16。在本文中，我们使用相同的数据集基于 KL 估计得到的 Pena 统计量来识别影响点，当 $\lambda=0$ (OLS)和 $\lambda=0.0002$，通过公式(3)计算得到 $S_{LK,i}$，结果见表 1。

由表 1 显示结果可以看出，基于 KL 估计所提出的 Pena 统计量，当 $\lambda=0$ 时，KL 估计退化为最小二乘估计，其识别出最有影响的五个样本点分别为：5、16、6、15 和 10，与 Semra Türkan 等[7]人识别出

的影响点是一样的；当 $\lambda = 0.0002$ 时，其 Pena 统计量识别出的最有影响的五个样本点分别为 16、5、15、6 和 4，与其它作者相比，至少有三个影响点是一样的，验证了本文基于 KL 估计所提出来的 Pena 统计量是合理可行的。

Table 1. The five most influential observations: Longley data
表 1. Longley 数据集中：最可能的 5 个影响点

$\lambda = 0$		$\lambda = 0.0002$	
case	$S_{LK,i}$	case	$S_{LK,i}$
5	0.6976	16	0.6087
16	0.5701	5	0.5017
6	0.5270	15	0.4451
15	0.4308	6	0.4004
10	0.3364	4	0.3322

4. 结论

在本文中，综合考虑了复共线性和影响诊断问题，Belsley 等[14]建议在检测异常点或影响点时，应处理复共线性问题。因此本文在基于 KL 估计下一般线性回归模型中，使用 Pena 距离来讨论 KL 估计的影响诊断，得到了基于 KL 估计下的 Pena 距离的表达式，并对其性质进行证明，得到 Pena 距离的分布在一定条件下近似于正态分布，并通过该统计量能识别出数据中高杠异常点，从而得到高杠异常点的判别方法。在文中将 Pena 距离与 Cook 距离的性质进行了比较，得出在一定条件下 Pena 统计量是优于 Cook 统计量的。最后，通过实例研究的结果验证，说明本文所提出的理论与方法是合理可行的。

参考文献

[1] Pena, D. (2005) A New Statistic for Influence in Linear Regression. *Technometrics*, **47**, 1-12. https://doi.org/10.1198/004017004000000662

[2] 孟丽丽. 基于 Pena 距离的加权最小二乘估计的影响分析[J]. 数理统计与管理, 2009, 28(2): 252-257.

[3] 胡江. 基于 Pena 距离的非线性回归模型的影响分析[J]. 大学数学. 2012, 28(5): 80-85.

[4] 胡江, 林金官, 赵彦勇. 基于 Pena 距离的广义线性回归模型的影响分析[J]. 应用数学, 2017, 30(3): 539-546.

[5] 胡江. 基于 Pena 距离的几种回归模型的影响分析[D]: [硕士学位论文]. 南京: 东南大学, 2012.

[6] Hu, J., Lin, J.G. and Zhao, Y.Y. (2017) Influence Analysis of Generalized Linear Regression Model Based on Pena Distance. *Applied Mathematics*, **30**, 539-546.

[7] Türkan, S. and Toktamıs, Ö. (2012) Detection of Influential Observations in Ridge Regression and Modified Ridge Regression. *Model Assisted Stats & Applications*, **7**, 91-97. https://doi.org/10.3233/MAS-2011-0215

[8] Emami, H. and Emami, M. (2016) New Influence Diagnostics in Ridge Regression. *Journal of Applied Statistics*, **43**, 476-489. https://doi.org/10.1080/02664763.2015.1070804

[9] Kashif, M., Amanullah, M. and Aslam, M. (2018) Pena's Statistic for the Liu Regression. *Journal of Statistical Computation and Simulation*, **88**, 2473-2488. https://doi.org/10.1080/00949655.2018.1468444

[10] Kashif, M., Ullah, M.A. and Aslam, M. (2019) Influential Diagnostics with Pena's Statistic for the Modified Ridge Regression. *Communications in Statistics*: *Simulation and Computation*. https://doi.org/10.1080/03610918.2019.1634204

[11] Kibria, B.M.G. and Lukman, A.F. (2020) A New Ridge-Type Estimator for the Linear Regression Model: Simulations and Applications. *Scientifica*, **2020**, Article ID: 9758378. https://doi.org/10.1155/2020/9758378

[12] Muller, K.E. and Mok, M.C. (1997) The Distribution of Cook's D Statistic. *Communications in Statistics—Theory and*

Methods, **26**, 525-546. https://doi.org/10.1080/03610927708831932

[13] Longley, J.W. (1967) An Appraisal of Least Squares Programs for Electronic Computer from the Point of View of the User. *Journal of American Statistical Association*, **62**, 819-841. https://doi.org/10.1080/01621459.1967.10500896

[14] Belsley, D.A., Kuh, E. and Welsch, R.E. (1989) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

[15] Cook, R.D. (1977) Detection of Influential Observation in Linear Regression. *Technometrics*, **19**, 15-18. https://doi.org/10.1080/00401706.1977.10489493

[16] Walker, E. and Birch, J. (1989) Influence Measures in Ridge Regression. *Technometrics*, **30**, 221-227. https://doi.org/10.1080/00401706.1988.10488370

[17] Jahufer, A. and Chen, J.B. (2009) Assessing Global Influential Observations in Modified Ridge Regression. *Statistics & Probability Letters*, **79**, 513-518. https://doi.org/10.1016/j.spl.2008.09.019