Advances in Applied Mathematics 应用数学进展, 2021, 10(11), 3783-3795
Published Online November 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1011401

Hans汉斯

随机约束Liu回归的局部影响分析

田维琦，王孟孟，郑智泉

贵州民族大学数据科学与信息工程学院，贵州 贵阳

收稿日期：2021年10月11日；录用日期：2021年11月1日；发布日期：2021年11月15日

摘　要

在进行回归诊断时，影响点的检测一直是统计学者们研究的一个热点问题，而大多数情况下变量之间会存在自相关性即复共线性，再利用普通最小二乘估计进行影响点的检测会掩盖或掩没一些影响点，得到某些误导性结论。因此，本文考虑利用随机约束Liu估计克服数据间存在复共线时对检测带来的影响，在随机约束Liu回归模型下通过Cook似然距离和Tsai、Billor和Loynes (TBL)的另一种似然距离两种局部影响分析方法来检测影响点，分别在三种扰动模型下得到了影响矩阵、影响曲率和梯度所需的计算公式。最后，通过Longley数据集说明了两种方法都能检测影响点。

关键词

局部影响分析，随机约束Liu，影响曲率，影响矩阵，梯度

Local Influence Analysis for the Liu Regression under Stochastic Linear Restrictions

Weiqi Tian, Mengmeng Wang, Zhiquan Zheng

School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou

Received: Oct. 11th, 2021; accepted: Nov. 1st, 2021; published: Nov. 15th, 2021

Abstract

In regression diagnosis, the detection of influence points has always been a hot issue studied by statisticians. In most cases, there will be autocorrelation between variables, namely complex collinearity, and the detection of influence points by using ordinary least square estimation will cover up or conceal some influence points and get some misleading conclusions. Therefore, in this

文章引用: 田维琦, 王孟孟, 郑智泉. 随机约束 Liu 回归的局部影响分析[J]. 应用数学进展, 2021, 10(11): 3783-3795.
DOI: 10.12677/aam.2021.1011401

paper, we consider using random constrained Liu estimation to overcome the influence of complex collinear data on detection. Under the random constrained Liu regression model, two local impact analysis methods, Cook likelihood distance and another likelihood distance of Tsai, Billor and Loynes (TBL), are used to detect the influence points. The formulas of influence matrix, influence curvature and gradient are obtained under three disturbance models. Finally, Longley data set shows that both methods can detect influence points.

Keywords

Local Influence Analysis, Stochastic Restricted Liu, Influence Matrix, Influence Matrix, Gradient

1. 引言

回归诊断是回归分析中的重要内容，其主要的任务就是检测、识别样本观测数据与统计模型存在偏离的数据点，一般常见的诊断方法有全局影响分析和局部影响分析。全局影响分析是对比完全删除某个样本观察值前后估计量的变化程度来度量相应观察值的影响程度[1]，以此识别出影响点。局部影响分析是 1986 年 Cook [2]首次提出用于识别异常点和强影响点的研究方法，从似然函数出发，利用微分几何观点提出的曲率的概念，再引入微小扰动的思想，通过似然距离来度量数据点的影响。近年来，局部影响分析受到统计学者们的广泛关注和发展，许多诊断方法也相继提出，如 Tsai [3]、Billor 和 Loynes [4]提出了基于新的似然距离用斜率替代曲率的局部影响分析方法。Shi [5]提出了通过定义广义 Cook 统计量的局部影响分析方法。在实际生活中，Belsley [6]等学者研究发现普通最小二乘估计在进行回归诊断时会掩没一些影响点，为了克服复共线性带来的某些不好影响，许多学者考虑用有偏估计替代最小二乘估计来减轻在影响点检测时带来的负面影响。如 Jahufer [7]、Jahufer [8]、Shi [9]、Billor [10]等所做的研究。另一方面，引入约束限制也是克服复共线性的一种方法，如 Paula [11]研究了不等式约束下线性模型的局部影响分析，Liu [12]研究了随机约束线性模型的局部影响分析，Yang [13]等研究了等式约束下椭球线性模型的局部影响分析。

基于前人的研究，使得我们考虑线性模型受随机约束且变量之间存在复共线性时的局部影响分析，因此本文考虑随机约束 Liu 估计应用于局部影响分析中，并分别利用 Cook 似然距离方法和 TBL 方法，给出了在方差扰动、因变量扰动、自变量扰动这三种扰动下的诊断统计量。

2. 模型与估计

考虑线性模型

$$y = X\beta + e \tag{2.1}$$

其中 y 是 $n\times 1$ 的因变量，X 是秩为 p 的 $n\times p$ 阶设计矩阵，e 是 $k\times 1$ 随机误差向量，期望为 $E(e)=0$，协方差矩阵为 $Var(e)=\sigma^2 I_n$，其中 σ^2 是已知常数，I_n 为 n 阶单位矩阵。假定参数估计 β 受如下随机约束

$$r = R\beta + u \tag{2.2}$$

其中 r 是 $q\times 1$ 的已知随机向量，R 是秩为 q 的 $q\times p$ 阶已知矩阵且 $q \le p$，u 是 $q\times 1$ 随机误差向量，期望为

$E(u)=0$，方差矩阵为$Var(u)=\sigma^2V$，V是$q\times q$阶已知正定矩阵。

基于模型(2.1)和(2.2)，便于应用于局部影响分析中利用 Marquardt [14]提出的方法，得到随机约束 Liu 回归的增广模型：

$$z=Z\beta+\varepsilon \tag{2.3}$$

这里

$$z=\begin{bmatrix} y \\ r \\ d\beta_{ols} \end{bmatrix}, Z=\begin{bmatrix} X \\ R \\ I_p \end{bmatrix}, \varepsilon=\begin{bmatrix} e \\ u \\ \xi \end{bmatrix} \tag{2.4}$$

z是的阶数为$m\times 1$，Z的阶数为$m\times p$，ε的阶数为$m\times 1$的随机误差向量，期望为$E(\varepsilon)=0$，协方差矩阵为$Var(\varepsilon)=\sigma^2Q$，$Q=diag(I_n,V,I_p)$，且该模型的各随机误差分量间相互独立。

为了将局部影响方法应用到随机约束 Liu 估计中，我们给出随机约束 Liu 回归模型(2.3)参数估计β、σ^2的极大似然估计。因此，我们假设模型(2.3)的随机误差服从$\varepsilon\sim N(0,\sigma^2Q)$的正态分布。

则对应的似然函数为

$$L(\theta)=-\frac{m}{2}\ln 2\pi-\frac{m}{2}\ln\sigma^2-\frac{1}{2}\ln|Q|-\frac{1}{2\sigma^2}\varepsilon^{\mathrm{T}}Q^{-1}\varepsilon \tag{2.5}$$

其中$\theta=(\beta^{\mathrm{T}},\sigma^2)^{\mathrm{T}}$。

下面我们给出参数β，σ^2的极大似然估计。

定理 1 对于模型(2.3)我们有

$$\hat{\beta}=\left(X^{\mathrm{T}}X+R^{\mathrm{T}}V^{-1}R+I_p\right)^{-1}\left(X^{\mathrm{T}}y+R^{\mathrm{T}}V^{-1}r+d\beta_{ols}\right).$$

$$\hat{\sigma}^2=\frac{1}{m}\left[\left(y-X\hat{\beta}\right)^{\mathrm{T}}\left(y-X\hat{\beta}\right)+\left(r-R\hat{\beta}\right)^{\mathrm{T}}V^{-1}\left(r-R\hat{\beta}\right)+\left(d\beta_{ols}-\hat{\beta}\right)^{\mathrm{T}}\left(d\beta_{ols}-\hat{\beta}\right)\right]$$

证明：对$L(\theta)$分别关于β_i和σ^2求导，得

$$\frac{\partial L(\theta)}{\partial\beta_i}=\frac{1}{\sigma^2}\varepsilon^{\mathrm{T}}Q^{-1}Z_{,i}, \tag{2.6}$$

$$\frac{\partial L(\theta)}{\partial\sigma^2}=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\varepsilon^{\mathrm{T}}Q^{-1}\varepsilon, \tag{2.7}$$

其中$Z_{,i}$表示矩阵Z的第i列。令(2.6)式为零得

$$0=\varepsilon^{\mathrm{T}}Q^{-1}Z=\begin{pmatrix} y-X\beta \\ r-R\beta \\ d\hat{\beta}_{ols}-\beta \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} I_n & & \\ & V^{-1} & \\ & & I_p \end{pmatrix}\begin{pmatrix} X \\ R \\ I \end{pmatrix}$$

$$\hat{\beta}=\left(X^{\mathrm{T}}X+R^{\mathrm{T}}V^{-1}R+I_p\right)^{-1}\left(X^{\mathrm{T}}y+R^{\mathrm{T}}V^{-1}r+d\hat{\beta}_{ols}\right)$$

同理，令(2.7)式为零，我们计算

$$\hat{\sigma}^2=\frac{1}{m}\hat{\varepsilon}^{\mathrm{T}}Q^{-1}\hat{\varepsilon} \tag{2.8}$$

易得

$$\hat{\sigma}^2=\frac{1}{m}\left[\left(y-X\hat{\beta}\right)^{\mathrm{T}}\left(y-X\hat{\beta}\right)+\left(r-R\hat{\beta}\right)^{\mathrm{T}}V^{-1}\left(r-R\hat{\beta}\right)+\left(d\beta_{ols}-\hat{\beta}\right)^{\mathrm{T}}\left(d\beta_{ols}-\hat{\beta}\right)\right]$$

其中 $\hat{\varepsilon}=z-Z\hat{\beta}$，证毕。

3. 局部影响分析

为了评估各扰动模型的局部影响，我们现在推导 Cook 方法下的观测信息矩阵 $-H$ 和矩阵 Δ，以及 TBL 方法的梯度 G，以便我们得到最大影响曲率和最大斜率这两个重要的统计量。

3.1. 观测信息矩阵

为了进一步探讨随机约束 Liu 回归模型下的影响点检测，我们推导其模型的观测影响矩阵，将参数估计 θ 划分为 $\theta=\left(\beta^{\mathrm{T}},\sigma^2\right)^{\mathrm{T}}$，对观测信息矩阵 $-H$ 也进行相应划分。

定理 2 对随机约束 Liu 估计，$(p+1)\times(p+1)$阶矩阵

$$-H=\begin{bmatrix}\dfrac{1}{\hat{\sigma}^2}Z^{\mathrm{T}}Q^{-1}Z & \dfrac{1}{\hat{\sigma}^4}Z^{\mathrm{T}}Q^{-1}\left(z-Z\hat{\beta}\right)\\ \dfrac{1}{\hat{\sigma}^4}\left(z-Z\hat{\beta}\right)^{\mathrm{T}}Q^{-1}Z & \dfrac{m}{2\hat{\sigma}^4}\end{bmatrix} \tag{3.1}$$

证明：对(2.6)式的$\dfrac{\partial L(\theta)}{\partial\beta_i}$分别关于 β_j、$\hat{\sigma}^2$ 求导得

$$\frac{\partial^2 L(\theta)}{\partial\beta_j\partial\beta_i}=-\frac{1}{\sigma^2}Z_{,j}^{\mathrm{T}}Q^{-1}Z_{,i},\quad i,j=1,2,\cdots,p$$

$$\frac{\partial^2 L(\theta)}{\partial\sigma^2\partial\beta_i}=-\frac{1}{\sigma^4}\left(z-Z\beta\right)^{\mathrm{T}}Q^{-1}Z,\quad i=1,2,\cdots,p$$

将其中的未知数取为 $\beta=\hat{\beta}$ 和 $\sigma^2=\hat{\sigma}^2$ 时，即得到矩阵 $-H$ 的前行。再对(2.7)式$\dfrac{\partial L(\theta)}{\partial\sigma^2}$关于 β_j 和 σ^2 分别求导，得

$$\frac{\partial^2 L(\theta)}{\partial\beta_j\partial\sigma^2}=-\frac{1}{\sigma^4}\left(z-Z\beta\right)^{\mathrm{T}}Q^{-1}Z_{,j},\quad j=1,2,\cdots,p$$

$$\frac{\partial^2 L(\theta)}{\partial\sigma^2\partial\sigma^2}=\frac{m}{2\sigma^4}-\frac{1}{\sigma^6}\left(z-Z\beta\right)^{\mathrm{T}}Q^{-1}\left(z-Z\beta\right).$$

在上两式中令 $\beta=\hat{\beta}$ 和 $\sigma^2=\hat{\sigma}^2$，则得到 $-H$ 的最后一行。

特别地，利用(2.8)式可得

$$\left.\frac{\partial^2(\theta)}{\partial\sigma^2\partial\sigma^2}\right|_{\left(\beta=\hat{\beta},\sigma^2=\hat{\sigma}^2\right)}=-\frac{m}{2\hat{\sigma}^4}$$

因此可以写成矩阵形式

$$-\frac{\partial L(\theta\mid\omega)}{\partial\theta\partial\theta^{\mathrm{T}}}=\begin{bmatrix}\dfrac{1}{\hat{\sigma}^2}Z^{\mathrm{T}}Q^{-1}Z & \dfrac{1}{\hat{\sigma}^4}Z^{\mathrm{T}}Q^{-1}Z\left(z-Z\hat{\beta}\right)\\ \dfrac{1}{\hat{\sigma}^4}\left(z-Z\hat{\beta}\right)^{\mathrm{T}}Q^{-1}Z & \dfrac{m}{2\hat{\sigma}^4}\end{bmatrix}$$

证毕。

接下来，我们分别讨论关于方差的扰动、因变量的扰动和自变量的扰动下相应的诊断统计量。

3.2. 方差扰动模型

此处我们考虑方差的扰动。令 $W_\omega^{-1}=diag\left\{\omega_{11}^{-1},\cdots,\omega_{nn}^{-1}\right\}$ 是 $n\times n$ 阶正定扰动矩阵且 $\omega_{ii}=1+\omega_i$，$i=1,2,\cdots,n$。显然 $\omega=\left(\omega_1,\cdots,\omega_n\right)^{\mathrm{T}}=0$，即 $W_0^{-1}=I$，对应于无扰动情形，此时 $L\left(\theta\middle|W_0\right)=L\left(\theta\right)$。这种情形下，扰动的协方差矩阵是 $Q_\omega=diag\left\{W_\omega^{-1},V,I\right\}$。

扰动的对数似然函数为

$$L\left(\theta\right)=-\frac{m}{2}\ln 2\pi-\frac{m}{2}\ln\sigma^2-\frac{1}{2}\ln\left|Q_\omega\right|-\frac{1}{2\sigma^2}\varepsilon^{\mathrm{T}}Q_\omega^{-1}\varepsilon \tag{3.2}$$

定理 3 对随机约束 Liu 估计得 $\left(p+1\right)\times n$ 阶矩阵

$$\Delta=\begin{bmatrix}\dfrac{1}{\hat{\sigma}^2}X^{\mathrm{T}}D\left(\hat{e}\right)\\ \dfrac{1}{2\sigma^4}\hat{e}^{\mathrm{T}}D\left(\hat{e}\right)\end{bmatrix} \tag{3.3}$$

其中 $i=1,\cdots,p$，$j=1,2,\cdots,n$，x_{ji} 表示矩阵 X 的第 i 行第 j 列的元素，$\hat{e}_j$ 表示 $\hat{e}$ 的第 j 个元素。

证明：利用(3.2)式的对数似然函数可得

$$\frac{\partial L\left(\theta\middle|\omega\right)}{\partial\beta_i}=-\frac{1}{\sigma^2}\varepsilon^{\mathrm{T}}Q_\omega^{-1}Z_{,i},\ \ i=1,2,\cdots,p.$$

进一步

$$\begin{aligned}\frac{\partial L\left(\theta\middle|\omega\right)}{\partial\beta_i}&=\frac{1}{\sigma^2}\varepsilon^{\mathrm{T}}\begin{pmatrix}W_\omega^{-1}&&\\&V^{-1}&\\&&I\end{pmatrix}Z_{,i}\\&=\frac{1}{\sigma^2}\left(e_1\left(1+\omega_1\right),\cdots,e_n\left(1+\omega_n\right),u^{\mathrm{T}}V^{-1},\xi^{\mathrm{T}}\right)Z_{,i}\\&=\frac{1}{\sigma^2}\left[\sum_{j=1}^{n}e_j\left(1+\omega_j\right)x_{ji}+u^{\mathrm{T}}V^{-1}R_{,i}+I\xi^{\mathrm{T}}\right]\end{aligned}$$

其中，$R_{,i}$ 表示矩阵 R 的第 i 列，ξ_i 表示 ξ 的第 i 个元素。上式关于 ω_j 求导并在 $\theta=\hat{\theta}$、$\omega=0$ 取值可得

$$\left.\frac{\partial^2 L\left(\theta\middle|\omega\right)}{\partial\omega_j\partial\beta_i}\right|_{\theta=\hat{\theta},\omega=0}=\frac{1}{\hat{\sigma}^2}\hat{e}_j x_{ji}$$

其中 $\hat{e}_j=y_j-X_{j,}\hat{\beta}$，$X_{j,}$ 表示矩阵 X 的第 j 行。

又利用(3.2)式的对数似然函数可得

$$\frac{\partial\left(\theta\middle|\omega\right)}{\partial\sigma^2}=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\varepsilon^{\mathrm{T}}Q_\omega^{-1}\varepsilon=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\left[\sum_{j=1}^{n}e_j\left(1+\omega_j\right)e_j+u^{\mathrm{T}}V^{-1}u+\xi^{\mathrm{T}}\xi\right]$$

上式关于 ω_j 求导并在 $\theta=\hat{\theta}$、$\omega=0$ 取值可得

$$\left.\frac{\partial^2 L\left(\theta\middle|\omega\right)}{\partial\omega_j\partial\sigma^2}\right|_{\left(\theta=\hat{\theta},\omega=0\right)}=\frac{1}{2\sigma^4}\left(e_j\right)^2,$$

因此可以写成矩阵表示

$$\left.\frac{\partial L(\theta\,|\,\omega)}{\partial\theta\partial\omega^{\mathrm{T}}}\right|_{(\theta=\hat{\theta},\omega=0)}=\begin{bmatrix}\dfrac{1}{\hat{\sigma}^2}X^{\mathrm{T}}D(\hat{e})\\ \dfrac{1}{2\sigma^4}\hat{e}^{\mathrm{T}}D(\hat{e})\end{bmatrix}$$

证毕。

在 TBL 方法下，我们得

定理 4 对随机约束 Liu 回归模型得

$$G=\left(\frac{1}{2}-\frac{1}{2\hat{\sigma}^2}(\hat{e}_1)^2,\cdots,\frac{1}{2}-\frac{1}{2\hat{\sigma}^2}(\hat{e}_n)^2\right)^{\mathrm{T}} \tag{3.4}$$

证明： 利用(4.2)式的对数似然函数可得

$$\frac{\partial L(\theta|\omega)}{\partial\omega_j}=\frac{1}{2(1+\omega_i)}-\frac{1}{2\sigma^2}e_i^2.$$

代入 $\theta=\hat{\theta}$、$\omega=0$ 即得

$$\left.\frac{\partial L(\theta|\omega)}{\partial\omega_i}\right|_{(\theta=\hat{\theta},\omega=0)}=\frac{1}{2}-\frac{1}{2\hat{\sigma}^2}(\hat{e}_j)^2,\quad i=1,2,\cdots,n$$

证毕。

3.3. 响应变量的扰动模型

此处我们考虑让因变量 y 的扰动形式记为 $y+\omega$，其中扰动 $\omega=(\omega_1,\cdots,\omega_n)^{\mathrm{T}}$。则无扰动的情形对应于 $\omega=(0,\cdots,0)^{\mathrm{T}}$，而扰动后的对数似然为

$$L(\theta\,|\,\omega)=-\frac{m}{2}\ln 2\pi-\frac{m}{2}\ln\sigma^2-\frac{1}{2}\ln|Q|-\frac{1}{2\sigma^2}\varepsilon_\omega^{\mathrm{T}}Q^{-1}\varepsilon_\omega \tag{3.5}$$

其中 $\varepsilon_\omega=\left(y^{\mathrm{T}}+\omega^{\mathrm{T}},r^{\mathrm{T}},d\hat{\beta}_{ols}\right)^{\mathrm{T}}-Z\beta$。

定理 5 对随机约束 Liu 回归模型得 $(p+1)\times n$ 阶矩阵

$$\Delta=\begin{bmatrix}\dfrac{1}{\hat{\sigma}^2}X^{\mathrm{T}}\\ \dfrac{1}{\sigma^4}\hat{e}^{\mathrm{T}}\end{bmatrix} \tag{3.6}$$

证明： 利用(3.5)式中的对数似然函数可得

$$\frac{\partial L(\theta|\omega)}{\partial\beta_i}=\frac{1}{\sigma^2}\left(y+\omega-X\beta,r-R\beta,d\hat{\beta}_{ols}-\beta\right)Q^{-1}Z,\quad i=1,2,\cdots,p$$

$$\frac{\partial L(\theta|\omega)}{\partial\beta_i}=\frac{1}{\sigma^2}\left(y_1+\omega_1-X_{1,}\beta,\cdots,y_n+\omega_n-X_{n,}\beta,u^{\mathrm{T}}V^{-1},\xi^{\mathrm{T}}\right)Z_{,i}$$

$$=\frac{1}{\sigma^2}\left[\sum_{j=1}^{n}\left(y_j+\omega_j-X_{j,}\beta\right)x_{ji}+u^{\mathrm{T}}V^{-1}R_{,i}+\xi_i\right]$$

上式关于 ω_j 求导并在 $\theta=\hat{\theta}$， $\omega=0$ 取值可得

$$\left.\frac{\partial^2 L(\theta|\omega)}{\partial\omega_j\partial\beta_i}\right|_{(\theta=\hat{\theta},\omega=0)}=\frac{1}{\hat{\sigma}^2}x_{ji}.$$

又利用(3.5)式的对数似然函数可得

$$\frac{\partial L(\theta|\omega)}{\partial\sigma^2}=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\varepsilon_\omega^{\mathrm{T}}Q^{-1}\varepsilon_\omega=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\left[\sum_{j=1}^{n}e_{\omega j}e_{\omega j}+u^{\mathrm{T}}V^{-1}u+\xi^{\mathrm{T}}\xi\right]$$

其中 $e_{\omega j}$ 是 e_ω 的第 j 个元素。上式关于 ω_j 求导并在 $\theta=\hat{\theta}$、 $\omega=0$ 取值可得

$$\left.\frac{\partial^2 L(\theta|\omega)}{\partial\omega_j\partial\sigma^2}\right|_{(\theta=\hat{\theta},\omega=0)}=\frac{1}{\hat{\sigma}^4}\hat{e}_j,$$

因此可写成矩阵形式

$$\left.\frac{\partial L(\theta|\omega)}{\partial\theta\partial\omega^{\mathrm{T}}}\right|_{(\theta=\hat{\theta},\omega=0)}=\begin{bmatrix}\frac{1}{\hat{\sigma}^2}X^{\mathrm{T}}\\ \frac{1}{\sigma^4}\hat{e}^{\mathrm{T}}\end{bmatrix}$$

证毕。

在 TBL 方法下，我们得

定理 6 对随机约束 Liu 回归模型得

$$G=-\frac{1}{\hat{\sigma}^2}\hat{e}\tag{3.7}$$

证明：利用(3.5)式的对数似然函数可得

$$\frac{\partial L(\theta|\omega)}{\partial\omega_i}=-\frac{1}{\sigma^2}\left(y_i+\omega_i-X_{i,}\beta\right)$$

代入 $\theta=\hat{\theta}$、 $\omega=0$ 即得

$$\left.\frac{\partial L(\theta|\omega)}{\partial\omega_j}\right|_{(\theta=\hat{\theta},\omega=0)}=-\frac{1}{\hat{\sigma}^2}\hat{e}_i,\ \ i=1,2,\cdots,n,$$

证毕。

3.4. 解释变量的扰动模型

Cook (1986)指出当自变量之间存在复共线性时，自变量的微小扰动会影响最小二乘回归的结果。我们讨论自变量扰动令 $X_\omega=X+s\omega d$，则第 t 个自变量的扰动为 $X_\omega=X+s_t\omega d_t^{\mathrm{T}}$，这里 $\omega=(\omega_1,\cdots,\omega_n)^{\mathrm{T}}$ 表示 n 维的扰动向量，d 为 $p\times1$的向量，则 d_t 为第 t 个分量为 1、其余为 0 的向量，s 为尺度因子，则 s_t 用于解释 X 各列的不同的测量单位。显然无扰动的情形对于 $\omega=(0,\cdots,0)^{\mathrm{T}}$，而扰动后的对数似然函数 $L(\theta|\omega)$ 为

$$L(\theta|\omega)=-\frac{m}{2}\ln2\pi-\frac{m}{2}\ln\sigma^2-\frac{1}{2}\ln|Q|-\frac{1}{2\sigma^2}\varepsilon_\omega^{\mathrm{T}}Q^{-1}\varepsilon\tag{3.8}$$

而这里的ε_ω为

$$\varepsilon_\omega=\left(y^{\mathrm{T}},r^{\mathrm{T}},I_p\right)^{\mathrm{T}}-Z_\omega\beta, \tag{3.10}$$

其中Z_ω表示将矩阵Z中X代替为X_ω得到的矩阵。

定理 7 对随机约束 Liu 估计模型，扰动设计矩阵的第t列得$(p+1)\times n$阶矩阵

$$\Delta=\begin{bmatrix}-\dfrac{s}{\hat{\sigma}^2}\left(\hat{\beta}^{\mathrm{T}}X^{\mathrm{T}}-\delta_{it}\hat{e}^{\mathrm{T}}\right)\\ -\dfrac{s}{\hat{\sigma}^4}\hat{\beta}\hat{e}^{\mathrm{T}}\end{bmatrix},\ i=1,\cdots,p,\ t=1,\cdots,p,\ \delta_{it}=\begin{cases}1 & t=i\\ 0 & t\neq i\end{cases},$$

证明：利用(4.8)和(4.10)式可得

$$\frac{\partial L\left(\theta|\omega\right)}{\partial\beta_i}=\frac{1}{\sigma^2}\varepsilon_\omega^{\mathrm{T}}Q^{-1}Z_{\omega,i},\ i=1,2,\cdots,p,$$

其中$Z_{\omega,i}$表示Z_ω的第i列。进一步

$$\begin{aligned}\frac{\partial L\left(\theta|\omega\right)}{\partial\beta_i}&=\frac{1}{\sigma^2}\left(y_1-X_{\omega1,}\beta,\cdots,y_n-X_{\omega n,}\beta,u^{\mathrm{T}}V^{-1},\xi^{\mathrm{T}}\right)Z_{\omega,i}\\&=\frac{1}{\sigma^2}\left[\sum_{s=1}^{n}\left(y_s-\sum_{r=1,r\neq t}^{p}x_{sr}\beta_r-\left(x_{st}+s_t\omega_s\right)\beta_t\right)x_{\omega si}+u^{\mathrm{T}}V^{-1}R_{,i}+\xi^{\mathrm{T}}I_p\right]\end{aligned}$$

$x_{\omega si}$是X_ω的第s行第i列的元素。

将上式关于ω_j求导并在$\theta=\hat{\theta}$、$\omega=0$计算得

$$\left.\frac{\partial^2L\left(\theta|\omega\right)}{\partial\omega_j\partial\beta_i}\right|_{\left(\theta=\hat{\theta},\omega=0\right)}=\begin{cases}-\dfrac{s_t}{\hat{\sigma}^2}\hat{\beta}_tx_{ji}, & t\neq i\\ -\dfrac{s_t}{\hat{\sigma}^2}\hat{\beta}_tx_{ji}+\dfrac{1}{\hat{\sigma}^2}s_t\left(y_j-X_{j,}\hat{\beta}\right), & t=i\end{cases},$$

此式可合写为

$$\left.\frac{\partial^2L\left(\theta|\omega\right)}{\partial\omega_j\partial\beta_i}\right|_{\left(\theta=\hat{\theta},\omega=0\right)}=-\frac{s_t}{\hat{\sigma}^2}\hat{\beta}_tx_{ji}+\frac{1}{\hat{\sigma}^2}\delta_{it}s_t\left(y_j-X_{j,}\hat{\beta}\right)=-\frac{s_t}{\hat{\sigma}^2}\hat{\beta}_tx_{ji}+\frac{1}{\hat{\sigma}^2}\delta_{it}s_t\hat{e}_j$$

又由(3.8)和(3.10)式算得

$$\begin{aligned}\frac{\partial L\left(\theta\omega\right)}{\partial\sigma^2}&=-\frac{m}{2\sigma^2}+\frac{1}{2\hat{\sigma}^4}\varepsilon_\omega^{\mathrm{T}}Q^{-1}\varepsilon_\omega\\&=-\frac{m}{2\sigma^2}+\frac{1}{2\sigma^4}\left[\sum_{s=1}^{n}\left(y_s-\sum_{r=1,r\neq t}^{p}x_{sr}\beta_r-\left(x_{st}+s_t\omega_s\right)\beta_t\right)^2+u^{\mathrm{T}}V^{-1}u+\xi^{\mathrm{T}}\xi\right]\end{aligned}$$

上式关于ω_j求导并在$\theta=\hat{\theta}$、$\omega=0$取值可得

$$\left.\frac{\partial^2L\left(\theta|\omega\right)}{\partial\omega_j\partial\sigma^2}\right|_{\left(\theta=\hat{\theta},\omega=0\right)}=-\frac{s_t}{\hat{\sigma}^4}\left(y_j-X_{j,}\hat{\beta}\right)\hat{\beta}_t=-\frac{s_t}{\hat{\sigma}^4}\hat{e}_j,$$

因此可以写成矩阵形式

$$\left.\frac{\partial L(\theta\mid\omega)}{\partial\theta\partial\omega^{\mathrm{T}}}\right|_{(\theta=\hat{\theta},\omega=0)}=\begin{bmatrix}-\dfrac{s}{\hat{\sigma}^2}\left(\hat{\beta}X^{\mathrm{T}}-\delta_{it}\hat{e}^{\mathrm{T}}\right)\\ -\dfrac{s}{\hat{\sigma}^4}\hat{\beta}\hat{e}\end{bmatrix}$$

证毕。

在 TBL 方法下，我们得到

定理 8 对随机约束 Liu 估计模型，扰动设计矩阵的第 t 列得

$$G=\frac{s_t\hat{\beta}_t}{\hat{\sigma}^2}\hat{e}$$

证明：利用(3.8)和(3.10)式可得

$$\frac{\partial L(\theta|\omega)}{\partial\omega_j}=\frac{s_t}{\sigma^2}\left(y_i-\left(X_{i,}+s_t\omega_t d_t^{\mathrm{T}}\right)\beta\right)\beta_t$$

代入 $\theta=\hat{\theta}$、$\omega=0$ 即得

$$\left.\frac{\partial L(\theta|\omega)}{\partial\omega_i}\right|_{(\theta=\hat{\theta},\omega=0)}=\frac{s_t}{\hat{\sigma}^2}\hat{e}_i\hat{\beta}_t,\quad i=1,2,\cdots,n$$

证毕。

4. 实证分析

为了验证在 Cook 和 TBL 的基础上提出的新方法的合理性，考虑引入 Longley [15]宏观经济数据集来检验该新方法。这组数据由就业、国民生产总值内含平减物价、国民生产总值、失业数、军事武装部队规模、14 岁及以上的非机构人口、年份 7 个指标组成。其中 y 是总派生就业率，x_1 是 GNP 隐含价格平减指数，x_2 是国民生产总值，x_3 是失业率，x_4 是武装力量的规模，x_5 是 14 岁及以上的非机构人口，x_6 是年份。Belsley 等人利用最大特征值与最小特征值的比值计算得到条件数值，该数据集下的条件数为 43275，说明 Longley 宏观经济数据集变量之间存在很强的共线性。

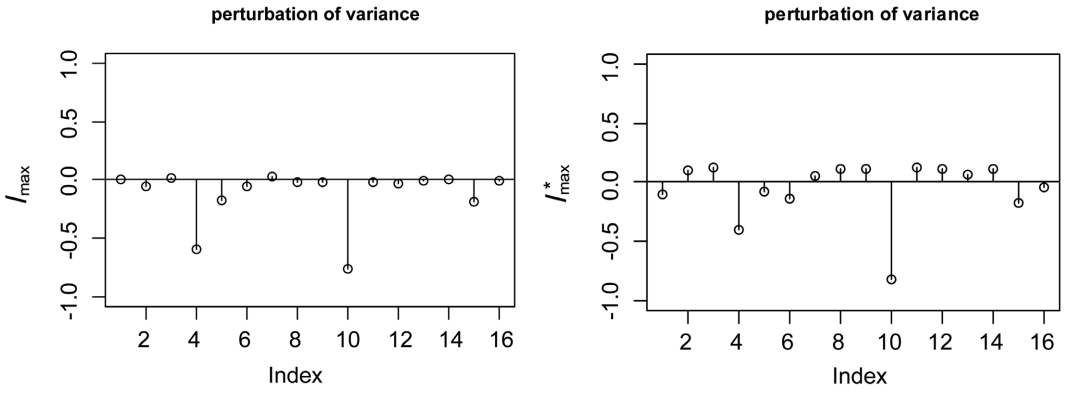
Cook (1977) [16]利用 Longley 基于最小二乘估计的全局影响分析，从大到小依次检测出 5、16、4、10、15 为影响点(本文所有排序都是从大到小)，Walker 和 Birch [17]基于岭估计的全局影响分析，检测出 16、10、4、15、1 为影响点，容易见两种方法检测出的影响点有较大差异，这是因为最小二乘估计在回归诊断时的不稳定性导致的。此外，取岭参数 $k=0.0002$，Shi 和 Wang 考虑岭估计局部影响分析，检测出 10、4、15、16、1 为影响点，当数据集复共线性很强时，Liu 提出了 Liu 估计，该估计在克服复共线性和均方误差准则皆优于岭估计。考虑其优良性，取 Liu 参数 $d=0.9724$，Jahufer 和 Chen (2010)利用 Liu 估计局部影响分析，检测出 4、10、1、5、6 为影响点，

局部影响分析方法探测到了基于 Liu 估计下的影响点，Zhang (2010)基于 Liu 估计局部影响分析，在方差扰动、响应变量扰动、解释变量扰动三种模型下，检测出 4、5、6、10、11、13 为影响点。同样我们选择该数据集，以此来和前人的检测结果进行比较。因此，选择 Ozkale (2009) [18]选择的数据点 2、3 来构成随机约束的(3.5)式。首先对数据点 2、3 进行标准化处理，所以有

$$R=\begin{bmatrix}-0.3154 & -0.3332 & -0.2399 & -0.4269 & -0.3263 & -0.3525\\ -0.3226 & -0.3368 & 0.3150 & -0.3676 & -0.2840 & -0.2983\end{bmatrix}$$

$$r=\begin{bmatrix}-0.3084\\ -0.3783\end{bmatrix}$$

在方差扰动下，取 Liu 参数 $d=0.9$，使用 Cook 的方法所得到的 $l_{\max}$ 向量的检测结果见图 1(a)，影响最大的分量为 10，4。使用 TBL 所得到的 $l_{\max}^{*}$ 向量的检测结果见图 1(b)，同样影响最大的分量为 10，4。这说明在方差扰动下随机约束 Liu 回归在 Cook 方法和 TBL 方法探测到的影响点没有明显差异。

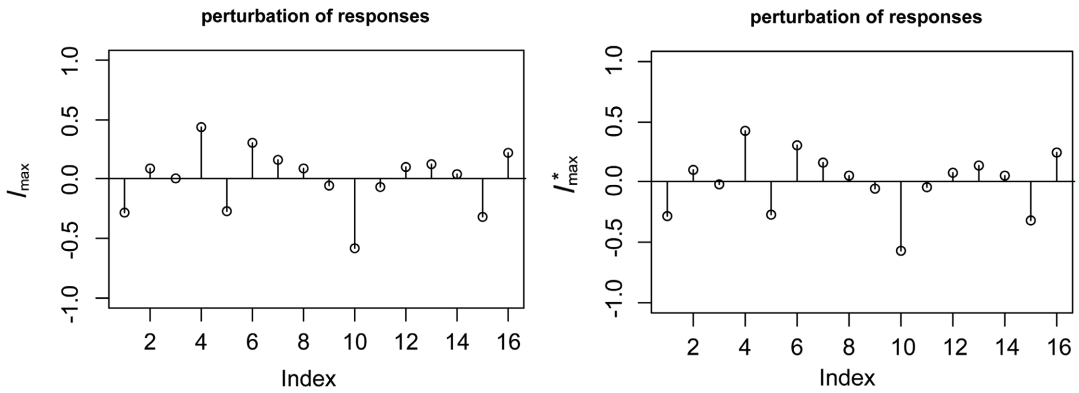


(a) (b)

Figure 1. (a) is the index graph of $l_{\max}$ under variance disturbance; (b) is the index graph of $l_{\max}^{*}$ under variance disturbance

图 1. (a)为方差扰动下 $l_{\max}$ 的指标图；(b)为方差扰动下 $l_{\max}^{*}$ 的指标图

在响应变量扰动下，取 Liu 参数 $d=0.9$，在图 2(a)中，基于 Cook 方法的最大分量 $l_{\max}$ 依次对应点 10、4、15、6、1。在图 2(b)中，基于 TBL 方法的最大分量 $l_{\max}^{*}$ 依次对应点 10、4、15、6、1。根据 Schwarzmann [19]无偏估计下 $l_{\max}$ 正比于残差向量，通过本文定理 6 知 $l_{\max}^{*}$ 是正比于残差向量的，易见两图 $l_{\max}^{*}$、$l_{\max}$ 几乎完全一样，即他们是成比例的。说明随机约束 Liu 回归下 $l_{\max}$ 和 $l_{\max}^{*}$ 是正比于残差向量的



(a) (b)

Figure 2. (a) is the index graph of $l_{\max}$ under the disturbance of response variable; (b) is the index graph of $l_{\max}^{*}$ under the disturbance of response variable

图 2. (a)为响应变量扰动下 $l_{\max}$ 的指标图；(b)为响应变量扰动下 $l_{\max}^{*}$ 的指标图

在单个解释变量扰动下。同样取 Liu 参数 $d=0.9$，基于 Cook 方法的最大分量 $l_{\max}$，见图 3(b)、图 3(c)、图 3(e)的检测结果，发现扰动解释变量 x_2、x_3、x_6 的影响点都为 10、4、15、6、1，见图 3(a)的检测结果，发现扰动解释变量 x_1 的影响点为 10、4、1、15、6，见图 3(d)的检测结果，发现扰动解释变量的 x_4 的影响点为 10、4、15、5、6，见图 3(f)的检测结果，发现扰动解释变量的 x_5 的影响点为 10、4、6、

5、15，易见自变量的改变会导致检测结果的改变。基于 TBL 的最大分量 $l^*_{\max}$，见图 4(a)~(e)的检测结果，发现扰动解释变量扰动 x_1、x_2、x_3、x_4、x_5、x_6 都的影响点都为 10、4、15、6、1。与其他学者的方法相比，检测结果有三个以上是相同的，说明本文基于随机约束下 Liu 估计提出的局部影响分析方法的合理性。

两种方法也反映了 Cook 方法更能为我们提供更多信息，同时，通过检测结果发现我们选取的数据点 2、3 不是影响点，这也说明了我们选取的约束条件不同也会对我们的结果产生负面影响。在这种情形下我们检测到的结论和前人探测到的结论是可比较的。我们也考虑不同的 Liu 参数 d 值检测出的影响点也不完全相同，说明 Liu 参数 d 的选取会影响我们的检测结果。

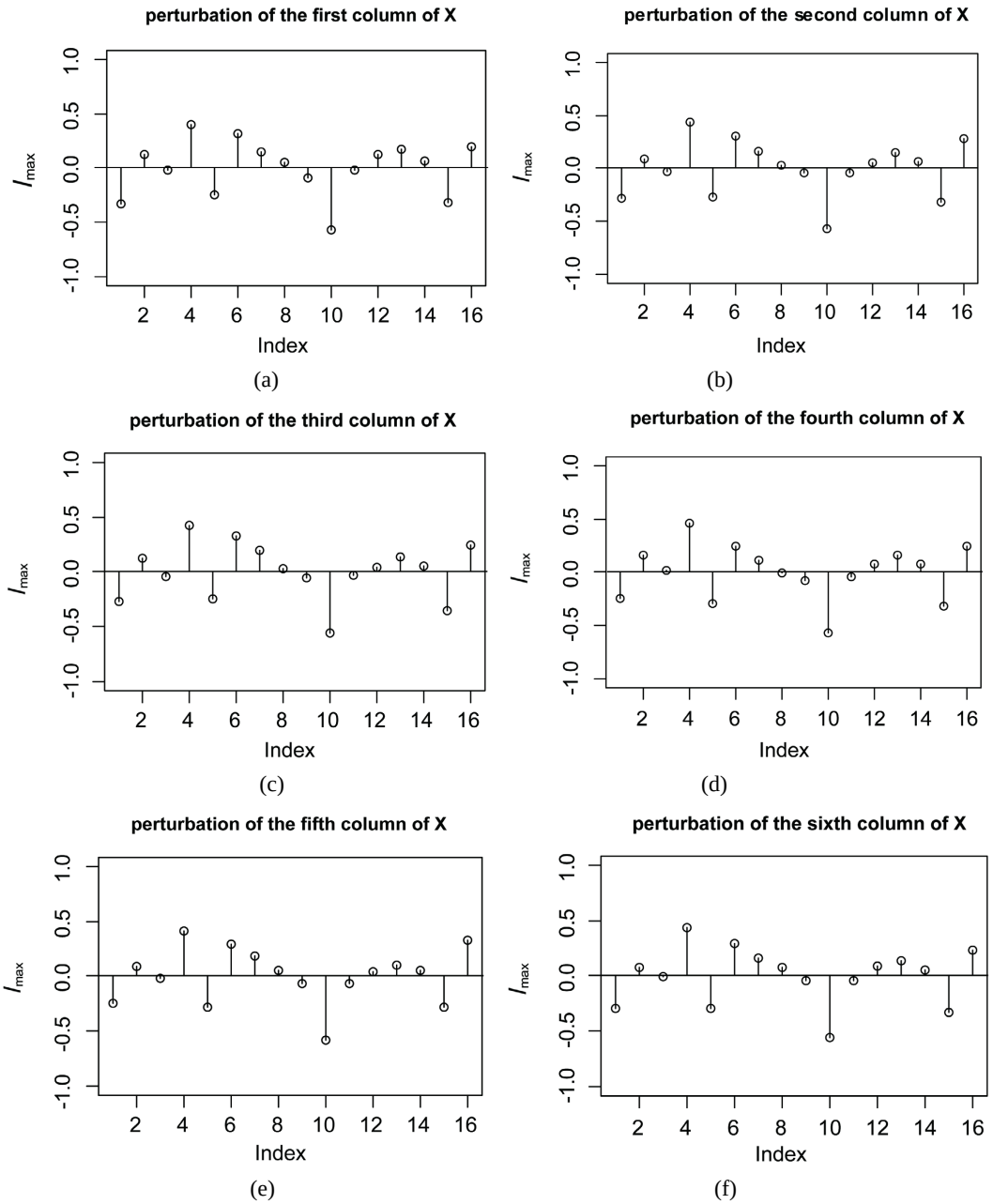


Figure 3. (a)~(f) is the index graph of $l_{\max}$ disturbed by a single explanatory variable

图 3. (a)~(f)为单个解释变量扰动下 $l_{\max}$ 的指标图

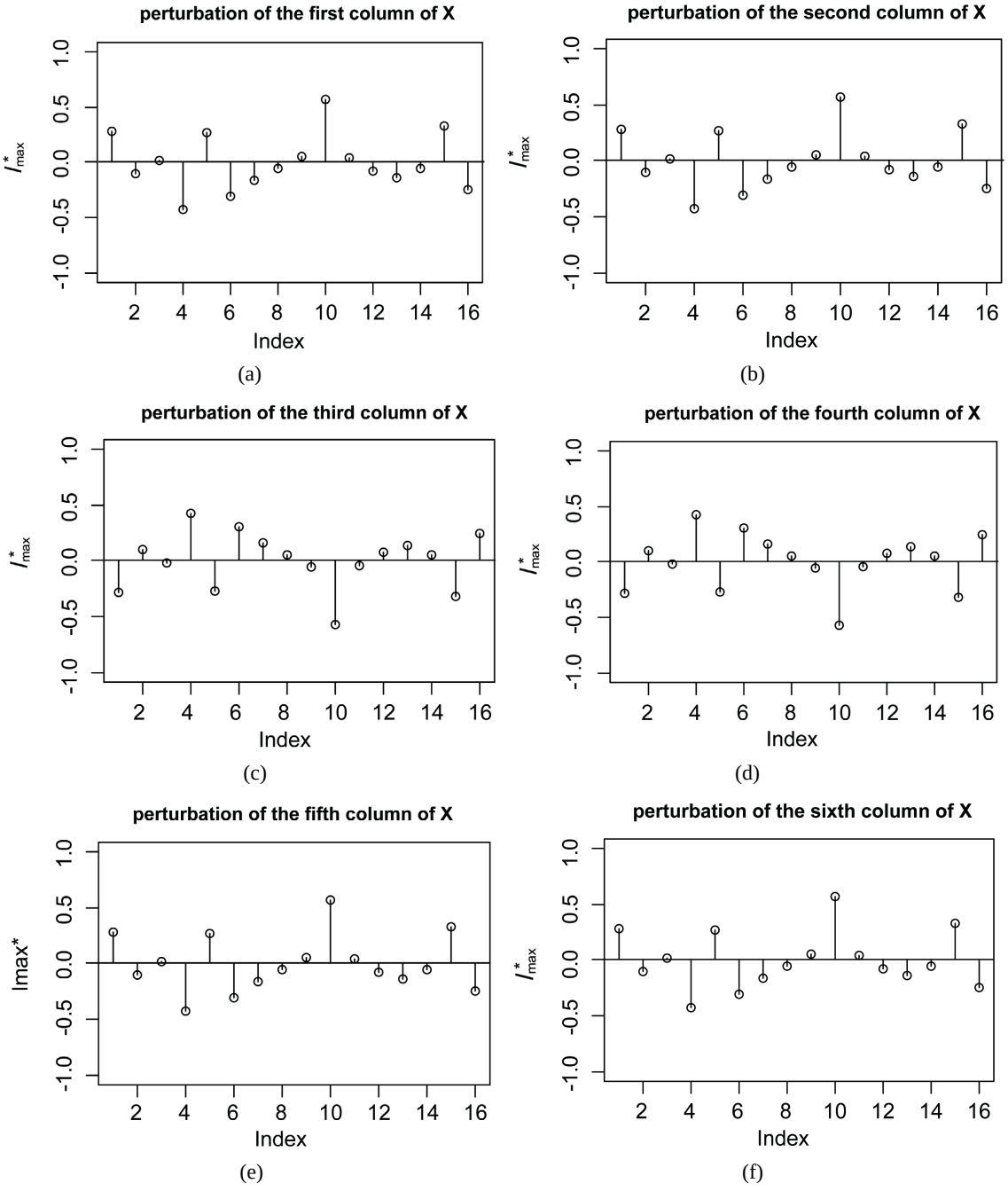


Figure 4. (a)~(f) is the index graph of $l_{\max}^{*}$ disturbed by a single explanatory variable

图 4. (a)~(f)为单个解释变量扰动下 $l_{\max}^{*}$ 的指标图

5. 结论

本文研究了具有随机线性约束 Liu 回归中的影响诊断方法。利用 Marquardt (1970)的方法得到了极大似然随机约束 Liu 估计，我们在三种扰动模型下导出了 Cook 方法的观测信息矩阵 $-H$ 和 Δ，以及 TBL 方法的梯度 G。从而得到 Cook 方法的 $C_{\max}$ 和 $l_{\max}$，以及扰动模型下 TBL 方法的 $S_{\max}^{*}$ 和 $l_{\max}^{*}$。最后，通过数据集进行了验证，说明本文所提出的理论与方法的合理性。

参考文献

[1] 杨莲. 几类统计模型的局部影响分析研究[D]: [博士学位论文]. 重庆: 重庆大学, 2015.

[2] Cook, R.D. (1986) Assessment of Local Influence. *Journal of the Royal Statistical Society*: *Series B*, **48**, 133-155. https://doi.org/10.1111/j.2517-6161.1986.tb01398.x

[3] Tsai, C.L. (1986) Discussion of Assessment of Local Influence by R. D. Cook. *Journal of the Royal Statistical Society*, *Series B*, **48**, 165.

[4] Billor, N. and Loynes, R.M. (1993) Local Influence: A New Approach. *Communications in Statistics—Theory and Methods*, **22**, 1595-1611. https://doi.org/10.1080/03610929308831105

[5] Shi, L. (1997) Local Influence in Principal Components Analysis. *Biometrika*, **84**, 175-186. https://doi.org/10.1093/biomet/84.1.175

[6] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley, New York. https://doi.org/10.1002/0471725153

[7] Jahufer, A. and Chen, J.B. (2010) Identifying Local Influential Observations in Liu Estimator. *Metrika*, **75**, 425-438. https://doi.org/10.1007/s00184-010-0334-4

[8] Jahufer, A. (2013) Detecting Global Influential Observations in Liu Regression Model. *Open Journal of Statistics*, **3**, 5-11. https://doi.org/10.4236/ojs.2013.31002

[9] Shi, L. and Wang, X. (1999) Local Influence in Ridge Regression. *Computational Statistics and Data Analysis*, **31**, 341-353. https://doi.org/10.1016/S0167-9473(99)00019-5

[10] Billor, N. (1999) An Application of the Local Influence Approach to Ridge Regression. *Journal of Applied Statistics*, **26**, 177-183. https://doi.org/10.1080/02664769922511

[11] Paula, G.A. (1993) Assessing Local Influence in Restricted Regression Models. *Computational Statistics and Data Analysis*, **16**, 63-79. https://doi.org/10.1016/0167-9473(93)90245-O

[12] Liu, S., Ahmed, S.E. and Ma, L.Y. (2009) Influence Diagnostics in the Linear Regression Model with Stochastic Linear Restrictions. *Pakistan Journal of Statistics*, **25**, 647-662.

[13] Yang, H. and Yang, L. (2016) Assessing Local Influence for Elliptical Linear Models under Equality Constraints. *Communications in Statistics*: *Theory and Methods*, **45**, 4517-4527. https://doi.org/10.1080/03610926.2013.773351

[14] Marquardt, D.W. (1970) Generalised Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Regression. *Technometrics*, **12**, 591-613. https://doi.org/10.2307/1267205

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

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

[17] Walker, E. and Birch, J.B. (1988) Influence Measures in Ridge Regression. *Technometrics*, **30**, 221-227. https://doi.org/10.1080/00401706.1988.10488370

[18] Ozkale, M.R. (2009) A Stochastic Restricted Ridge Regression Estimator. *Journal of Multivariate Analysis*, **100**, 1706-1716. https://doi.org/10.1016/j.jmva.2009.02.005

[19] Schwarzmann, B. (1991) A Connection between Local-Influence Analysis and Residual Diagnostics. *Technometrics*, **33**, 103-104. https://doi.org/10.1080/00401706.1991.10484773