应用概率统计　第二十八卷
第二期　2012年4月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.2 Apr. 2012

基于M估计的线性混合模型的局部影响分析 *

孙慧慧　　林金官

(盐城师范学院数学科学学院, 盐城, 224002)　　(东南大学数学系, 南京, 210096)

《应用概率统计》版权所用

摘　要

对纵向数据的线性混合模型, 用Fisher得分法得到了参数的M估计(稳健估计), 给出了其渐近性质, 利用影响曲率研究了M估计下的随机误差方差扰动的局部影响分析问题, 并通过葡萄糖数据的实例进行了分析论证.

关键词: M估计, 线性混合模型, 局部影响分析, 影响曲率.

学科分类号: O212.2.

§1. 引　　言

局部影响分析是Cook (1986)首先提出来的一种统计诊断方法, 他从似然函数出发, 借助于统计曲率概念提出了评价模型微小扰动的局部影响分析方法, 也是识别数据集中异常点和强影响点的重要方法. 关于纵向数据模型局部影响分析的文献已有不少, 如Beckman和Nachtsheim (1987)研究了混合效应模型中对误差方差和随机效应方差进行扰动的局部影响分析; Wei和Zhong (2001)研究了非线性随机效应模型的影响分析问题; 杨爱军等(2006)研究了利用影响曲率对非线性随机效应模型自相关系数扰动的统计诊断.

传统的极大似然估计是不稳健的, 当观察值中含有异常点时, 会对参数的估计产生影响, 进而对统计分析产生影响. 为了消除异常值对估计的影响, 可采用M估计的方法. M估计最早是由Huber引入回归问题, 是目前应用最广泛的稳健估计方法. Paramjit (2000)用得分方法对纵向数据线性混合模型的参数进行了稳健估计; Huggins (1993)介绍了重复测量的稳健分析方法. 前人大多研究的是稳健估计的算法及其稳健性, 本文将M估计引入统计诊断, 研究M估计下的局部影响分析问题.

§2. 模型介绍和参数估计

本文研究的模型如下

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta} + \boldsymbol{C}_k\boldsymbol{\tau}_k + \boldsymbol{e}_k, \qquad k = 1, 2, \ldots, N, \tag{2.1}$$

*国家自然科学基金(11171065)、江苏省自然科学基金(BK2011058)和盐城师范学院科研项目基金(11YCKL019)资助.
本文2011年8月5日收到, 2011年12月29日收到修改稿.

《应用概率统计》版权所用

其中$\boldsymbol{y}_k=(y_{k1},\ldots,y_{kn_k})^T$是第$k$个个体的$n_k$次观测结果组成的向量, $\boldsymbol{X}_k$是q维未知固定效应向量$\boldsymbol{\beta}$的$n_k\times q$阶设计矩阵, 且$\boldsymbol{X}_k=(x_{k1},\ldots,x_{kn_k})^T$. $\boldsymbol{C}_k$是随机效应$\boldsymbol{\tau}_k$的$n_k\times r$阶设计矩阵, 且假设$\boldsymbol{\tau}_k\sim N(0,\sigma^2\Gamma)$. $\boldsymbol{e}_k=(e_{k1},\ldots,e_{kn_k})^T$是$n_k\times 1$维不可观测的随机误差向量, 假设$\boldsymbol{e}_k\sim N(0,\sigma^2\Omega_k)$, 同时假设$\boldsymbol{\tau}_k$与$\boldsymbol{e}_k$相互独立. 从而$\mathsf{Cov}\,(\boldsymbol{y}_k)=\sigma^2\Sigma_k=\sigma^2\boldsymbol{C}_k\Gamma\boldsymbol{C}_k^T+\sigma^2\Omega_k$, 此处及以后, T表示矩阵或向量的转置. 我们用$\boldsymbol{\alpha}$表示Σ_k中的未知参数向量, 则模型的对数似然函数为

$$l(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2|\boldsymbol{y})=\text{constant}-\frac{1}{2}M\log\sigma^2-\frac{1}{2}\sum_{k=1}^{N}\log|\Sigma_k|-\sum_{k=1}^{N}\frac{1}{2}\varepsilon_k^T\varepsilon_k,\tag{2.2}$$

其中$\varepsilon_k=\sigma^{-1}\Sigma_k^{-1/2}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})$. 我们看到(2.2)式的最后一项是一个平方和的形式, 随偏差增长很快. 如果采用增长较慢的函数代替这个二次函数, 就可以提高其稳健性, 从而限制异常观察值的影响. 基于这个观点的稳健估计算法主要是M估计, 是目前应用最广泛的稳健估计方法. 本文我们选用Huber函数来代替二次函数.

$$\rho(\varepsilon)=\begin{cases}\dfrac{1}{2}\varepsilon^2, & \text{当 } |\varepsilon|\le c \text{ 时;}\\[2mm] c|\varepsilon|-\dfrac{1}{2}c^2, & \text{当 } |\varepsilon|>c \text{ 时,}\end{cases}$$

其中c是固定常数, 通常$c\in[0.7,2]$, 本文我们取$c=1.4$. 对$\rho(\varepsilon)$求导得

$$\psi(\varepsilon)=\frac{\partial\rho(\varepsilon)}{\partial\varepsilon}=\begin{cases}\varepsilon, & \text{当 } |\varepsilon|\le c \text{ 时;}\\ c\,\mathrm{sign}(\varepsilon), & \text{当 } |\varepsilon|>c \text{ 时.}\end{cases}$$

因此(2.2)的稳健形式为

$$\eta(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2|\boldsymbol{y})=\text{constant}-\frac{1}{2}\kappa_1 M\log\sigma^2-\frac{1}{2}\kappa_1\sum_{k=1}^{N}\log|\Sigma_k|-\sum_{k=1}^{N}\sum_{j=1}^{n_k}\rho(\varepsilon_{jk}),\tag{2.3}$$

其中$\kappa_1=\mathsf{E}(\varepsilon\psi(\varepsilon))=\mathsf{P}(|\varepsilon|\le c)$为相合修正因子.

现在我们基于(2.3)式, 用Fisher得分迭代法对参数进行稳健估计. 关于$\boldsymbol{\beta}$有

$$\frac{\partial\eta}{\partial\boldsymbol{\beta}}=\sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}\psi[\sigma^{-1}\Sigma_k^{-1/2}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})],$$

$$H_{\boldsymbol{\beta}\boldsymbol{\beta}^T}=\mathsf{E}\Big(-\frac{\partial^2\eta}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\Big)=\nu\sigma^{-2}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1}\boldsymbol{X}_k,$$

其中Λ为对角阵, 对角元素$\Lambda_{ii}=\partial\psi(\varepsilon)/\partial\varepsilon$, 当$|\varepsilon|\le c$时, 其值为1, 当$|\varepsilon|<c$时, 其值为0. $\nu=\mathsf{E}(\Lambda_{ii})=\mathsf{P}(|\varepsilon|\le c)=\displaystyle\int_{-c}^{c}(2\pi)^{-1/2}e^{-1/2\varepsilon^2}\mathrm{d}\varepsilon=\kappa_1$. 由此得到$\boldsymbol{\beta}$的稳健估计的迭代公式为

$$\widehat{\boldsymbol{\beta}}^{(h+1)}=\widehat{\boldsymbol{\beta}}^{(h)}+(\widehat{H}^{(h)}_{\boldsymbol{\beta}\boldsymbol{\beta}^T})^{-1}\frac{\partial\eta^{(h)}}{\partial\boldsymbol{\beta}}.\tag{2.4}$$

当迭代序列收敛时, 得到$\widehat{\boldsymbol{\beta}}$即为$\boldsymbol{\beta}$的稳健极大似然估计(RMLE). 注意到, 当$c=\infty$时, $\boldsymbol{\beta}$的稳健估计$\widehat{\boldsymbol{\beta}}$就是传统的极大似然估计(MLE). 用类似的方法, 可以得到方差分量的稳健极大似然估计迭代公式.

众所周知, 参数的极大似然估计(MLE)具有渐近正态性, 根据Domowitz和White(1982)以及Sanjoy (2004), 稳健极大似然估计(RMLE)也具有相合性和渐近正态性.

§3. 随机误差的方差加权扰动模型局部影响分析

本节我们研究稳健情形下对随机误差方差进行扰动的局部影响分析问题.

令$\eta(\boldsymbol{\theta})$表示原模型$M$的稳健对数似然函数, 参数$\boldsymbol{\theta}$的RMLE记为$\widehat{\boldsymbol{\theta}}$. 假设原模型$M$受到扰动, 成为$M(\boldsymbol{\omega})$, 其中$\boldsymbol{\omega}=(\omega_1,\omega_2,\ldots,\omega_N)^T$为描述扰动因素的$N$维向量, 其定义域$\Omega$为$\mathbb{R}^N$上的某一开集. 受到扰动后的模型用$M(\boldsymbol{\omega})$表示, 相应的稳健对数似然函数为$\eta(\boldsymbol{\theta}|\boldsymbol{\omega})$, 相应的RMLE为$\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$. 特别, 假定存在$\boldsymbol{\omega}=\boldsymbol{\omega}_0\in\Omega$, 使$M(\boldsymbol{\omega}_0)=M$表示未受扰动的模型, 因此有$\eta(\boldsymbol{\theta}|\boldsymbol{\omega}_0)=\eta(\boldsymbol{\theta})$, $\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0)=\widehat{\boldsymbol{\theta}}$, 并且有$\eta(\widehat{\boldsymbol{\theta}})=\eta(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}_0)|\boldsymbol{\omega}_0)$. 我们考虑模型(2.1)的随机误差方差加权模型, 令$\mathsf{Var}\,(\boldsymbol{e}_k)=\sigma^2\omega_k I_{n_k}$, 即通过扰动标量$\omega_k$对第$k$个个体随机误差的方差进行扰动. 这时$\boldsymbol{\omega}=(\omega_1,\omega_2,\ldots,\omega_N)^T$, $\boldsymbol{\omega}_0=(1,\ldots,1)^T$, 而相应的扰动模型的对数似然函数为

$$l(\boldsymbol{\theta}|\boldsymbol{\omega})=\text{constant}-\frac{1}{2}M\log\sigma^2-\frac{1}{2}\sum_{k=1}^{N}\log|\Sigma_k(\omega_k)|-\sum_{k=1}^{N}\frac{1}{2}\varepsilon_k(\omega_k)^T\varepsilon_k(\omega_k),\tag{3.1}$$

其中$\Sigma_k(\omega_k)=\boldsymbol{C}_k\Gamma\boldsymbol{C}_k^T+\omega_k I_{n_k}$, $\varepsilon_k(\omega_k)=\sigma^{-1}\Sigma_k^{-1/2}(\omega_k)(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})$. 则扰动模型的稳健对数似然函数为

$$\eta(\boldsymbol{\theta}|\boldsymbol{\omega})=\text{constant}-\frac{1}{2}\kappa_1 M\log\sigma^2-\frac{1}{2}\kappa_1\sum_{k=1}^{N}\log|\Sigma_k(\omega_k)|-\sum_{k=1}^{N}\sum_{j=1}^{m_k}\rho(\varepsilon_{jk}(\omega_k)).\tag{3.2}$$

为评价扰动对于统计分析的影响, 通常需要比较扰动前后估计量之间的差异. 为度量这种差异, 似然距离仍然是行之有效的. 对于上述模型M及其扰动模型$M(\boldsymbol{\omega})$, M估计$\widehat{\boldsymbol{\theta}}$与$\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega})$之间的似然距离定义为$\mathrm{LD}(\boldsymbol{\omega})=2\{\eta(\widehat{\boldsymbol{\theta}})-\eta(\widehat{\boldsymbol{\theta}}(\boldsymbol{\omega}))\}$. 由于$\eta(\widehat{\boldsymbol{\theta}})$为全局最大值, 因此恒有$\mathrm{LD}(\boldsymbol{\omega})\geq 0$, $\mathrm{LD}(\boldsymbol{\omega})$越大, 表明扰动对于估计量$\widehat{\boldsymbol{\theta}}$的影响越大. 我们采用Cook定义的影响曲率方法来研究扰动的影响, 首先定义似然距离$z=\mathrm{LD}(\boldsymbol{\omega})$形成的曲面在$\boldsymbol{\omega}_0$处沿方向$\boldsymbol{d}=(d_1,\ldots,d_N)$的曲率, $\boldsymbol{d}$可取为单位向量, 由韦博成等(1991)得

定理 3.1　对于上述扰动模型, 曲面$z=\mathrm{LD}(\boldsymbol{\omega})$在$\boldsymbol{\omega}_0$处沿$\boldsymbol{d}$方向的影响曲率可表示为

$$C_d=2\boldsymbol{d}^T(-\ddot{F})\boldsymbol{d}=2\boldsymbol{d}^T\Delta^T[-\ddot{\eta}(\widehat{\boldsymbol{\theta}})]^{-1}\Delta\boldsymbol{d}=2\mathrm{LD}^{\mathrm{II}}(\boldsymbol{\omega}),$$

其中$-\ddot{F}$称为影响矩阵, $\ddot{\eta}$为$\eta(\boldsymbol{\theta})$关于$\boldsymbol{\theta}$的二阶导数, $\mathrm{LD}^{\mathrm{II}}(\boldsymbol{\omega})$为似然距离$\mathrm{LD}(\boldsymbol{\omega})$的二阶近似,

《应用概率统计》版权所用

《应用概率统计》版权所用

$\ddot{F}$的表达式如下

$$\ddot{F}=\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\Big|_{(\boldsymbol{\omega}_0,\widehat{\boldsymbol{\theta}})}=\Delta^T\ddot{\eta}(\widehat{\boldsymbol{\theta}})^{-1}\Delta,\qquad \Delta=\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T}\Big|_{(\boldsymbol{\omega}_0,\widehat{\boldsymbol{\theta}})}.$$

我们用$C_{\max}$表示C_d的最大值, 即似然距离沿$\boldsymbol{d}$方向的变化率的最大值, $\boldsymbol{d}_{\max}$表示似然距离对于扰动最敏感的方向. 由韦博成等(1991)知, $C_{\max}=2\lambda_1$, 其中λ_1为矩阵$-\ddot{F}$的最大特征值, 而$\boldsymbol{d}_{\max}$为λ_1对应的特征向量. 我们可以对$(i,(\boldsymbol{d}_{\max})_i)$, $i=1,\ldots,N$进行列表或作散点图, $(\boldsymbol{d}_{\max})_i$为$\boldsymbol{d}_{max}$的第$i$个分量, 其中数值明显比其他点大的, 就是该扰动影响特别大的点, 因而就对应于强影响个体或异常个体.

下面我们根据(3.2)推导影响曲率的具体表达式

$$\begin{aligned}
\frac{\partial\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}} &= \sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\psi(\varepsilon_k(\omega_k)),\\
\frac{\partial\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\sigma^2} &= \frac{1}{2}\sigma^{-2}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\psi(\varepsilon_k(\omega_k))-\frac{1}{2}M\kappa_1\sigma^{-2},\\
\frac{\partial\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\alpha_i} &= \frac{1}{2}\sum_{k=1}^{N}\psi[\sigma^{-1}\Sigma_k^{-1/2}(\omega_k)(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})]^T\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\alpha_i}\\
&\quad\cdot[\sigma^{-1}\Sigma_k^{-1/2}(\omega_k)(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{\beta})]-\frac{1}{2}\kappa_1\sum_{k=1}^{N}\mathrm{tr}\Big(\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Big),\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= -\sigma^{-2}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\Lambda\Sigma_k^{-1/2}(\omega_k)\boldsymbol{X}_k,\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\sigma^4} &= \frac{1}{2}M\kappa_1\sigma^{-4}-\frac{3}{4}\sigma^{-4}\sum_{k=1}^{N}\varepsilon_k^T(\omega_k)\psi(\varepsilon_k(\omega_k))-\frac{1}{4}\sigma^{-4}\sum_{k=1}^{N}\varepsilon_k^T(\omega_k)\Lambda\varepsilon_k(\omega_k),\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\delta_i\partial\delta_j} &= \frac{1}{2}\kappa_1\sum_{k=1}^{N}\mathrm{tr}\Big(\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_j}\Big)\\
&\quad-\frac{3}{4}\sum_{k=1}^{N}\psi(\varepsilon_k(\omega_k))^T\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_j}\varepsilon_k(\omega_k)\\
&\quad-\frac{1}{4}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_j}\Sigma_k^{-1}(\omega_k)\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\varepsilon_k(\omega_k),\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}\partial\delta_i} &= -\frac{1}{2}\sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Sigma_k^{-1/2}(\omega_k)\psi(\varepsilon_k(\omega_k))\\
&\quad-\frac{1}{2}\sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\varepsilon_k(\omega_k),\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\beta}\partial\sigma^2} &= -\frac{1}{2}\sigma^{-3}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\psi(\varepsilon_k(\omega_k))-\frac{1}{2}\sigma^{-3}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\Lambda\varepsilon_k(\omega_k),\\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\sigma^2\partial\delta_i} &= -\frac{1}{4}\sigma^{-2}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\frac{\partial\Sigma_k(\omega_k)}{\partial\alpha_i}\Sigma_k^{-1}(\omega_k)\psi(\varepsilon_k(\omega_k))\\
&\quad-\frac{1}{4}\sigma^{-2}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\alpha_i}\varepsilon_k(\omega_k).
\end{aligned}$$

《应用概率统计》版权所用

以上各二阶导数均在$(\boldsymbol{\omega}_0, \widehat{\boldsymbol{\theta}})$处计值, 经计算得到$\ddot{\eta}(\widehat{\boldsymbol{\theta}})$. Δ的第k行为

$$\Delta_i^T = \Big\{\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\beta_1}, \ldots, \frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\beta_q}, \frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\sigma^2}, \frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\delta_1}, \ldots, \frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\delta_r'}\Big\},$$

其中

$$\begin{aligned}
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\boldsymbol{\beta}} &= -\frac{1}{2}\sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\Sigma_k^{-1/2}(\omega_k)\psi(\varepsilon_k(\omega_k)) \\
&\quad -\frac{1}{2}\sigma^{-1}\sum_{k=1}^{N}\boldsymbol{X}_k^T\Sigma_k^{-1/2}(\omega_k)\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\varepsilon_k(\omega_k), \\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\sigma^2} &= -\frac{1}{4}\sigma^{-2}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\Sigma_k^{-1}(\omega_k)\psi(\varepsilon_k(\omega_k)) \\
&\quad -\frac{1}{4}\sigma^{-2}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\varepsilon_k(\omega_k), \\
\frac{\partial^2\eta(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_k\partial\delta_i} &= \frac{1}{2}\kappa_1\sum_{k=1}^{N}\mathrm{tr}\Big(\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\Big) \\
&\quad -\frac{3}{4}\sum_{k=1}^{N}\psi(\varepsilon_k(\omega_k))^T\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\varepsilon_k(\omega_k) \\
&\quad -\frac{1}{4}\sum_{k=1}^{N}\varepsilon_k(\omega_k)^T\frac{\partial\Sigma_k(\omega_k)}{\partial\omega_k}\Sigma_k^{-1}(\omega_k)\Lambda\Sigma_k^{-1}(\omega_k)\frac{\partial\Sigma_k(\omega_k)}{\partial\delta_i}\varepsilon_k(\omega_k).
\end{aligned}$$

以上各式均在$(\boldsymbol{\omega}_0, \widehat{\boldsymbol{\theta}})$处计值.

§4. 实例分析

葡萄糖数据由美国科罗拉多州医疗中心大学小儿科临床研究病房提供, 该数据通过对13个控制病人和20个肥胖病人测试其标准葡萄糖忍耐力. 实验过程为: 让33个病人服用葡萄糖, 分别在0, 0.5, 1, 1.5, 2, 3, 4, 5小时候后测其血样. 实验目的是为了研究比较控制组病人和肥胖组病人是否有显著区别. 本文对该数据采用以下的线性纵向数据模型来进行分析

$$\boldsymbol{y}_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \mathbf{1}\boldsymbol{\tau}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, 33,$$

其中$\boldsymbol{y}_i = (y_{ij})$, $j = 1, 2, \ldots, 8$; y_{ij}为第i个病人第j次的测量结果, 根据Pan和Fang (2002), 采用$\boldsymbol{X}_i$的设计阵如下

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0.5 & 1 & 1.5 & 2 & 3 & 4 & 5 \\ 0 & 0.25 & 1 & 2.25 & 4 & 9 & 16 & 25 \\ 0 & 0.125 & 1 & 3.375 & 8 & 27 & 64 & 125 \end{bmatrix}.$$

《应用概率统计》版权所用

下面对葡萄糖数据进行随机误差方差扰动局部影响的诊断, 并与基于MLE的对应诊断统计量比较. 经过计算, 我们考虑最大影响曲率方向关于各分量的散点图.

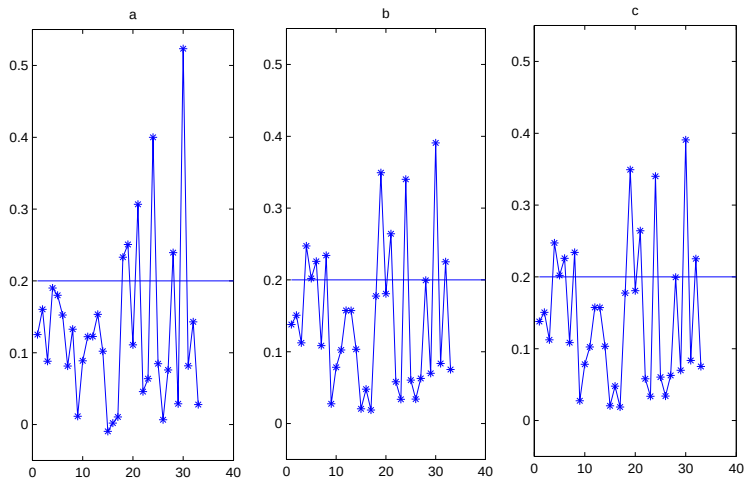


图1 随机误差方差扰动局部影响分析图

其中图1(a)表示Huber函数中常数$c = 1.4$时经过计算所得的局部影响分析图, 图1(b)表示$c = 4$情况下的散点图, 图1(c)为$c = \infty$, 即一般极大似然估计下计算所得的散点图. 可以看出, 图1(b)与图1(c)已相当接近. 从散点图(a)可以看出基于M估计时第30个受试个体对随机误差的方差扰动最为敏感, (b), (c)也反映出同样的情况. 另外, 在(b)、(c)中第19和第24个个体也对扰动较为敏感, 而(a)只反映第24个个体较为敏感. 而且(b), (c)散点图中0.2以上的数值明显比(a)多. 由此可见, Huber函数中常数c愈大, 限制作用愈小, 稳健估计情形下的影响分析愈接近极大似然估计的情形.

参 考 文 献

[1] Cook, R.D., Assessment of local influence, *Journal of the Royal Statistical Society, Ser. B*, **48**(1986), 133–169.

[2] Beckman, R.J., Nachtsheim, C.J. and Cook, R.D., Diagnostics for mixed-model analysis of variance, *Technometrics*, **29(4)**(1987), 413–426.

[3] Wei, B.C. and Zhong, X.P., Influence analysis in nonlinear models with rondom effects, *Applied Mathematics – A Journal of Chinese Universities*, **16B(1)**(2001), 35–44.

[4] 杨爱军, 林金官, 韦博成, 具有AR(1)误差的非线性随机效应模型中自相关系数的扰动诊断, 应用数学, **19(4)**(2006), 818–822.

[5] Paramjit, S.G., A robust mixed linear model analysis for longitudinal data, *Statistics in Medicine*, **19**(2000), 975–987.

[6] Huggins, R.M., A robust approach to the analysis of repeated measures, *Biometrics*, **49**(1993), 715–720.

[7] Domowitz, I. and White, H., Misspecified models with dependent observations, *Journal of Econometrics*, **20**(1982), 35–58.

[8] Sanjoy, K.S., Robust analysis of generalized linear mixed models, *Journal of the American Statistical Association*, **466**(2004), 451–460.

[9] 韦博成, 鲁国斌, 史建清, 统计诊断引论, 东南大学出版社, 1991.

[10] Pan, J.X. and Fang, K.T., *Growth Curve Models and Statistical Diagnostics*, New York: Springer-Verlag, 2002.

Local Influence Analysis of Mixed Effect Linear Models Based on M-Estimation

Sun Huihui

(*School of Mathematics and Science, Yancheng Teachers University, Yancheng, 224002*)

Lin Jinguan

(*Department of Mathematics, Southeast University, Nanjing, 210096*)

In this paper, the Fisher scoring method is applied to get M-estimator (robust estimator) in the mixed effects linear model for longitudinal data. And its asymptotic property is given later. Then the local influence analysis of variance of the error is studied based on M-estimation. At last the method is illustrated by the grape sugar data example.

Keywords: M-estimation, mixed effect linear model, local influence analysis, influence curvature.

AMS Subject Classification: 62J20.

《应用概率统计》版权所用