Statistical and Application 统计学与应用, 2012, 1, 37-43
http://dx.doi.org/10.12677/sa.2012.12008 Published Online December 2012 (http://www.hanspub.org/journal/sa.html)

Influence Analysis of Semivarying Coefficient Reproductive Dispersion Mixed Models

Rong Jiang, Xiaohan Yang, Weimin Qian

Department of Mathematics, Tongji University, Shanghai
Email: jrtrying@126.com, xiaohyang@tongji.edu.cn, wmqian2003@yahoo.com.cn

Received: Nov. 19th, 2012; revised: Nov. 26th, 2012; accepted: Dec. 13th, 2012

Abstract: This paper proposes several case-deletion as well as local influence measures for assessing the influence of an observation for semivarying coefficient reproductive dispersion mixed models. The essential idea is to treat the latent random effects in the model as missing data and estimate unknown parameters by acceleration of Monte Carlo EM algorithm. On the basis of the Q-function which is associated with the conditional expectation of the complete-data log-likelihood, we generate generalized Cook Distance. Moreover, three different perturbation schemes are discussed. Finally，one real illustrative example is presented to prove the methodology.

Keywords: Semivarying Coefficient Reproductive Dispersion Mixed Models; Local Influences; Penalized Spline; Cook Distance; Acceleration of Monte Carlo EM Algorithm

半变系数再生散度混合效应模型的影响分析

姜　荣，杨筱菡，钱伟民

同济大学应用数学系，上海
Email: jrtrying@126.com, xiaohyang@tongji.edu.cn, wmqian2003@yahoo.com.cn

收稿日期：2012 年 11 月 19 日；修回日期：2012 年 11 月 26 日；录用日期：2012 年 12 月 13 日

摘　要：本文把随机效应看作缺失数据并利用 P-样条拟合非参数部分，应用 Monte Carlo EM 加速算法得到半变系数再生散度混合效应模型的未知参数的估计，同时利用 Q 函数，得到了模型的广义 Cook 距离。此外，本文还研究了三种不同扰动情形的局部影响分析，得到了相应的影响矩阵。最后，通过一个实际例子验证了所提出的诊断统计量的有效性。

关键词：半变系数再生散度混合效应模型；局部影响；P-样条；广义 Cook 距离；Monte Carlo EM 加速算法

1. 引言

统计诊断从 20 世纪 70 年代中期受到统计学家的广泛关注，经过近 40 年的发展，异常点识别、残差分析、影响分析和数据变换等内容现已成为统计诊断的主要课题。特别地，基于数据删除模型和局部影响的诊断分析方法现已成为统计诊断的通用方法，它们可广泛地应用于各种统计模型的影响分析。例如，线性模型(Cook and Weisberg[1])，非线性回归模型(Seber and Wild[2])，半参数非线性模型(姜荣，邵明江，钱伟民[3])，线性混合效应模型(Beckman et al. [4])，半参数广义线性混合效应模型(张浩，朱仲义[5])。Jorgensen[6]首次提出了再生散度模型(RDM)，并指出广义线性模型的理论可以推广到以 RDM 为随机误差的模型。唐年胜和韦博成[7]研究了非线性再生散度随机效应

模型，讨论了该模型的几何结构、渐近性质和统计诊断等问题。

变系数模型在生物医学、公共卫生、经济、农业、制造业、道路安全等众多领域的数据分析中有广泛的应用。本文研究的半变系数再生散度混合效应模型是变系数模型的推广，对于此类模型的参数和非参数的估计关键在于条件期望的计算，Lin and Zhang[8]提出将随机效应当成参数从而用条件众数代替条件期望，但是这种方法对于非正态的模型估计效果很差。本文根据 McCulloch[9]将随机效应看作缺失数据，进而引入 EM 算法，并在 E 步中使用 MCMH 方法来计算条件期望，再利用 P-样条对非参数部分进行逼近。EM 算法和 Monte Carlo EM 算法，其收敛速度都是线性的，被缺损信息的倒数所控制，当缺损数据的比例很高时，收敛速度就非常缓慢。Monte Carlo EM 加速算法(罗季[10])在后验众数附近具有二次收敛速度。本文应用 Monte Carlo EM 加速算法估计全部未知参数。并通过一个实际例子验证了所提出的诊断统计量的有效性。

2. 主要结果

2.1. 模型介绍

假设第 i 个接受试验单元第 j 次的观察值 y_{ij} 关于随机效应 b_i 的条件密度为：

$$p_{y_{ij}|b_i}\left(y_{ij}\left|b_i,\beta,\alpha,\sigma^2\right.\right)=a\left(y_{ij};\sigma^2\right)\exp\left\{-\frac{1}{2\sigma^2}d\left(y_{ij};u_{ij}\right)\right\}$$

$$b_i\sim N\left(0,\sum(\gamma)\right)$$

$$\eta\left(u_{ij}\right)=x_{ij}^T\alpha\left(w_{ij}\right)+v_{ij}^T\beta+z_{ij}^Tb_i$$

其中 $\left(x_{ij},w_{ij},v_{ij},y_{ij},z_{ij}\right),i=1,\cdots,m;\ j=1,\cdots,n_i$ 表示 n 个独立的观察数据点，$\alpha(\cdot)$ 为未知函数，u_{ij} 是位置参数，σ^2 是散度参数，$a(\cdot;\cdot)$ 为已知函数，$d(\cdot;\cdot)$ 为已知单位偏差度函数，$\eta(\cdot)$ 为联系函数。根据唐年胜和韦博成[7]不妨设假设 $\eta(\cdot)$ 为典则联系，即 $\eta\left(u_{ij}\right)=u_{ij}$。

2.2. 非参数函数的 P-样条估计

对于未知单变量函数 $\alpha(\cdot)$，本文采用 P-样条估计。根据 Yu, et al.[11]，假设：

$$\alpha\left(w_i\right)=\delta_0+\delta_1w_i+\cdots+\delta_lw_i^l+\sum_{r=1}^{K}\delta_{l+r}\left(w_i-\kappa_r\right)_+^l$$

其中 $\{\kappa_r\}_{r=1}^K$ 为 K 个样条节点，$l\geq1$ 且为整数。Yu, et al.[11]详细研究了节点的选择方法，对于光滑函数(取 $l=2$ 并固定选取 5~10 个节点)。通常情况下，取预测变量的等分位点为节点。如果函数有不连续点，则在其附近要有一个节点；如果函数有限多极大值和极小值，则需要取 10 个以上的节点。设样条系数为 $\delta=\left(\delta_0,\delta_1,\cdots,\delta_{l+K}\right)^T$，样条基为：

$$B\left(w_i\right)=\left(1,w_i,\cdots,w_i^l,\left(w_i-\kappa_1\right)_+^l,\cdots,\left(w_i-\kappa_K\right)_+^l\right)^T$$

则函数 $\alpha(\cdot)$ 的样条函数为 $\alpha\left(w_i\right)=B^T\left(w_i\right)\delta$。

将上述向量结合写成矩阵的形式，

$X=\left(x_1,x_2,\cdots,x_m\right)^T$，$w=\left(w_1,w_2,\cdots,w_m\right)^T$，

$V=\left(v_1,v_2,\cdots,v_m\right)^T$，$Y=\left(y_1^T,y_2^T,\cdots,y_m^T\right)^T$，

$Z=\mathrm{diag}\left(z_1,z_2,\cdots,z_m\right)$，其中

$y_i=\left(y_{i1},y_{i2},\cdots,y_{in_i}\right)^T$，$x_i,w_i,v_i,z_i$ 类似的定义。

$B(w)=\left(B(w_1),B(w_2),\cdots,B(w_m)\right)^T$ 和

$\alpha(w)=B^T(w)\delta$，则模型可写成如下矩阵形式：

$$p_{y|b}\left(y\left|b,\beta,\alpha,\sigma^2\right.\right)=a\left(y;\sigma^2\right)\exp\left\{-\frac{1}{2\sigma^2}d\left(y;u\right)\right\}$$

$$b\sim N\left(0,\sum(\gamma)\right)$$

$$u=\left(BX\right)^T\delta+V^T\beta+Zb$$

由 Yu, et al.[11]，上述模型的惩罚对数似然函数为：

$$PL_o\left(\beta,\delta,\sigma^2,\gamma\left|Y_o\right.\right)=L_o\left(\beta,\delta,\sigma^2,\gamma\left|Y_o\right.\right)-\frac{1}{2}n\lambda\delta^TK\delta,$$

$\lambda>0$

$$L_o\left(\beta,\delta,\sigma^2,\gamma\left|Y_o\right.\right)=$$

$$\sum_{i=1}^{m}\log\int\prod_{j=1}^{n_i}p_{y_{ij}|b_i}\left(y_{ij}\left|b_i,\beta,\sigma^2,\gamma\right.\right)\left|\sum(\gamma)\right|^{-1/2}$$

$$\cdot\exp\left(-\frac{1}{2}b_i^T\sum{}^{-1}(\gamma)b_i\right)\mathrm{d}b_i$$

其中：Y_o 表示观测到的数据集，$\lambda>0$ 是光滑参数，K 是与节点 t 有关的矩阵，这里取 K 为对角矩阵，且只取最后 K 个对角线元素的值为 1，其它为 0。

[注]：参照 Yu, et al.[11]，我们可通过 GCV 方法选取光滑参数 λ，GCV 的具体算法可以参见 Yu, et al.[11]的(21)式。具体计算时，可用格子点方法获得最优的 λ。

2.3. 模型的估计

可根据文献 McCulloch[9]和 Zhu, et al.[12]，用 EM 算法对模型的参数和非参数进行估计。具体的做法是

将随机效应 b_i 看作缺失数据 Y_m，并用 $Y_c=\{Y_o,Y_m\}$ 表示完全数据，则完全数据的惩罚对数似然函数为：

$$PL_c\left(\beta,\delta,\sigma^2,\gamma\middle|Y_c\right)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[\log a\left(y_{ij};\sigma^2\right)-\frac{1}{2\sigma^2}d\left(y_{ij};u_{ij}\right)\right]-\sum_{i=1}^{m}\left[\frac{1}{2}\ln\left|\sum(\gamma)\right|+\frac{1}{2}b_i^T\sum{}^{-1}(\gamma)b_i\right]-\frac{1}{2}n\lambda\delta^TK\delta$$

利用 EM 算法求解。标准的 EM 算法包含 E 步和 M 步，给定初值 $\psi^{(0)}$；

$$E-\text{step}:Q_\psi\left(\psi\middle|\psi^{(m)}\right)=E\left\{PL_c\left(\psi\middle|Y_c\right)\middle|Y_o,\psi^{(m)}\right\}$$

$$M-\text{step}:\psi^{(m+1)}=\text{Max}Q_\psi\left(\psi\middle|\psi^{(m)}\right)$$

其中 $\psi=\left(\beta^T,\delta^T,\sigma^2,\gamma^T\right)^T$，$m$ 表示 EM 算法中迭代的次数。E 步是对分布 $p\left(b_i\middle|y_i,\psi\right)$ 求条件期望，而 M 步是求解 $\psi^{(m)}$ 使得 $Q_\psi\left(\psi\middle|\psi^{(m)}\right)$ 达到最大。在相对较弱的条件下，通过反复的迭代，$\psi^{(m)}$ 能收敛到 ψ 的极大似然估计 ψ。根据以上的 EM 算法，可以得到每次迭代计算 ψ 的极大似然估计方程为：

$$\dot{Q}\left(\psi\middle|\psi^{(m)}\right)=E\left[\frac{\partial PL_c\left(\psi\middle|Y_c\right)}{\partial\psi}\middle|Y_o,\psi^{(m)}\right]=0$$

条件分布 $p\left(b_i\middle|y_i,\psi\right)$ 无法直接获得积分的解析表达式，因此用 MH 方法来抽取 b_i 的样本 $\left\{b_i^{(n)}:n=1,\cdots,N\right\}(i=1,\cdots,m)$。基本思想是，选取一个适当的建议分布，通常选正态分布为建议分布。假设抽样链处于第 $t-1$ 时刻的状态为 $b_i^{(t-1)}$，从建议分布 $N\left(b_i^{(t-1)},\sigma_b^2\Omega_b\right)$ 中随机抽取一个潜在的转移值 b_i^*，从均匀分布 $U(0,1)$ 中随机抽取一个数 u，如果 $u\le\alpha\left(b_i^{(t-1)},b_i^*\right)$，则接受 b_i^* 作为链在下一时刻的状态值，即 $b_i^{(t)}=b_i^*$；否则，设 $b_i^{(t)}=b_i^{(t-1)}$，其中

$$\alpha\left(b_i^{(t-1)},b_i^*\right)=\min\left\{1,\frac{p\left(b_i^*\middle|y_i,\psi\right)}{p\left(b_i^{(t-1)}\middle|y_i,\psi\right)}\right\},$$

$$\Omega_b=E\left\{-\frac{\partial^2}{\partial b_i\partial b_i^T}\left[\log p\left(b_i\middle|y_i,\psi\right)\right]\right\}$$

并选取 σ_b^2 的值，以使得整个迭代过程从建议分布中产生的潜在转移值的接受率位于区间[0.25,0.34](Gelman, et al.[13])。

Monte Carlo EM 加速算法

1) 选取初值 $\psi^{(0)}$，令 $m=0$。

2) 用 MH 算法生成 N 个随机抽样 $b^{(1)},\cdots,b^{(N)}$，然后用这 N 个样本近似计算条件期望(Monte Carlo)，记为 $\hat{Q}\left(\psi\middle|\psi^{(m)},Y\right)$。

3) 将 $\hat{Q}\left(\psi\middle|\psi^{(m)},Y\right)$ 极大化，解出 $\psi_{EM}^{(m)}$，使得 $\hat{Q}\left(\psi_{EM}^{(m)}\middle|\psi^{(m)},Y\right)=\max_{\psi\in\Theta}\hat{Q}\left(\psi\middle|\psi^{(m)},Y\right)$。

4) 求解 γ，使得下式达到最小

$$\frac{1}{2N}\sum_{k=1}^{N}\left(\ln\left|\sum(\gamma)\right|+b_i^{(k)}\sum{}^{-1}(\gamma)b_i\right)。$$

5) N-R 步：令

$$\psi^{(m+1)}=\psi^{(m)}+\ddot{Q}^{-1}\left(\psi\middle|\psi^{(m)}\right)\dot{Q}\left(\psi\middle|\psi^{(m)}\right).$$

6) 重复步骤(2)~(5)，直到存在某个 M 使得 $\Pi\psi^{(M)}-\psi^{(M-1)}\Pi<\delta$ 时停止迭代，并取 $\hat{\psi}=\psi^{(M)}$，其中 δ 为指定的充分小的正数。

[**注**]：如果迭代收敛，则 $\hat{\psi}$ 就是 ψ 的极大似然估计。

定理 1：设 $Q\left(\psi\middle|\psi^{(m)}\right)\in C^{(2)},\psi^{(m)}$ 充分靠近 $\hat{\psi}$，$\dot{Q}\left(\hat{\psi}\middle|\psi^{(m)}\right)=0$。如果 $\ddot{Q}\left(\hat{\psi}\middle|\psi^{(m)}\right)$ 正定，且其 Hesse 矩阵满足 Lipschitz 条件。则对一切 m，当 n 充分大时，所得序列 $\left\{\psi^{(m)}\right\}$ 收敛到最优解 $\hat{\psi}$，并且序列具有二阶收敛速度。

证明：见文献[11]的定理 2.1。

3. 影响分析

由于本文是基于纵向数据，因此对模型讨论个体删除和组删除。在局部影响分析中，讨论了组内加权扰动、组间加权扰动和随机效应方差的扰动。

3.1. 个体删除模型的影响分析

设 $PL_c\left(\psi\middle|y_{c[ij]}\right)$ 是删除模型中第 i 组第 j 个观测值后所得到的完全数据的惩罚对数似然函数，相应的 Q 函数为 $Q_{[ij]}\left(\psi\middle|\hat{\psi}\right)=E\left\{PL_c\left(\psi\middle|Y_{c[ij]}\middle|Y_o,\hat{\psi}\right)\right\}$，且假定 $\hat{\psi}$ 和 $\hat{\psi}_{[ij]}$ 分别是 $Q\left(\psi\middle|\hat{\psi}\right)$ 和 $Q_{[ij]}\left(\psi\middle|\hat{\psi}\right)$ 达到最大值时候的取值。如果 $\hat{\psi}$ 和 $\hat{\psi}_{[ij]}$ 相差很大，则认为第 i 组第 j 个观测值为强影响点。实际计算中，如果对每一个 $\psi_{[ij]}$

都要进行迭代计算，则计算量非常大，因此根据 Zhu, et al.[12]，用 $\hat{\psi}_{[ij]}$ 的一步近似 $\hat{\psi}_{[ij]}^{1}$ 来减少计算量：

$$\hat{\psi}_{[ij]}^{1}=\hat{\psi}+\left\{-\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)\right\}^{-1}\dot{Q}_{[ij]}\left(\hat{\psi}\middle|\hat{\psi}\right) \tag{1}$$

其中 $\dot{Q}_{[ij]}\left(\hat{\psi}\middle|\hat{\psi}\right)=\partial Q_{[ij]}\left(\psi\middle|\hat{\psi}\right)/\partial\psi\Big|_{\psi=\hat{\psi}}$，

$\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)=\partial^{2}Q\left(\psi\middle|\hat{\psi}\right)/\partial\psi\partial\psi^{T}\Big|_{\psi=\hat{\psi}}$。

又由于 $\hat{\psi}_{[ij]}-\hat{\psi}$ 不能定量地表达影响的小，仿照 Cook 距离，用 Q 函数构造广义 Cook 距离；

$$CD_{ij}\left(\psi\right)=\left(\hat{\psi}_{[ij]}-\hat{\psi}\right)^{T}\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)\left(\hat{\psi}_{[ij]}-\hat{\psi}\right)$$

根据(1)式，可得到一步近似公式：

$$CD_{ij}\left(\psi\right)^{1}=\dot{Q}_{[ij]}\left(\hat{\psi}\middle|\hat{\psi}\right)^{T}\left\{-\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)\right\}^{-1}\dot{Q}_{[ij]}\left(\hat{\psi}\middle|\hat{\psi}\right)。$$

3.2. 组删除模型的影响分析

在纵向数据模型中，同一组中的观测值通常有相同的协变量，因此有必要研究一组数据对于模型的影响。对于组删除模型，同样可以推导出一步近似公式：

$$\hat{\psi}_{[i]}^{1}=\hat{\psi}+\left\{-\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)\right\}^{-1}\dot{Q}_{[i]}\left(\hat{\psi}\middle|\hat{\psi}\right) \tag{2}$$

其中

$$\dot{Q}_{[i]}\left(\hat{\psi}\middle|\hat{\psi}\right)=\partial Q_{[i]}\left(\psi\middle|\hat{\psi}\right)/\partial\psi\Big|_{\psi=\hat{\psi}}=\sum_{j=1}^{n_i}\partial E\left\{PL_{c}\left(\psi\middle|Y_{c[ij]}\middle|Y_{o},\hat{\psi}\right)\right\}\Big/\partial\psi\Big|_{\psi=\hat{\psi}}$$

根据(2)，可得到广义 Cook 距离的一步近似公式：

$$CD_{i}\left(\psi\right)^{1}=(\dot{Q}_{[i]}\left(\psi\middle|\hat{\psi}\right)^{T}\left\{-\ddot{Q}\left(\hat{\psi}\middle|\hat{\psi}\right)\right\}^{-1}\dot{Q}_{[i]}\left(\psi\middle|\hat{\psi}\right)。$$

3.3. 局部影响分析

设 $\omega=\left(\omega_{1},\cdots,\omega_{q}\right)^{T}$ 是定义在开区间域 $\Omega\subset R^{q}$ 上的向量。令 $PL_{c}\left(\psi,\omega\middle|Y_{c}\right)$ 为扰动模型的完全惩罚对数似然函数。假定存在 ω^{0} 使得

$PL_{o}\left(\psi,\omega^{0}\middle|Y_{o}\right)=PL_{o}\left(\psi\middle|Y_{o}\right)$ 和

$PL_{c}\left(\psi,\omega^{0}\middle|Y_{c}\right)=PL_{c}\left(\psi\middle|Y_{c}\right)$ 对所有的 ψ 都成立。设 $\hat{\psi}$ 和 $\hat{\psi}\left(\omega\right)$ 分别使得 Q 函数

$Q\left(\psi\middle|\hat{\psi}\right)=E\left\{PL_{c}\left(\psi\middle|Y_{c}\right)\middle|Y_{o},\hat{\psi}\right\}$ 和

$Q\left(\psi,\omega\middle|\hat{\psi}\left(\omega\right)\right)=E\left\{PL_{c}\left(\psi,\omega\middle|Y_{c}\right)\middle|Y_{o},\hat{\psi}\left(\omega\right)\right\}$ 达到最大值。以上的条件期望均是对条件分布 $p\left(Y_{m}\middle|Y_{o},\hat{\psi}\right)$ 求积分。

根据 Zhu, et al.[12]，构造模型的 Q 函数距离：

$$L_{Q}\left(\omega\right)=2\left\{Q\left(\psi\middle|\hat{\psi}\right)-Q\left(\psi\left(\omega\right)\middle|\hat{\psi}\right)\right\},$$

其二阶近似为：

$$L_{Q}^{II}\left(\omega\right)=-h^{T}\Delta^{T}\ddot{Q}^{-1}\left(\psi\middle|\hat{\psi}\right)\Delta h。$$

下文，将分别研究三种加权扰动情况：组内加权扰动，组间加权扰动，随机效应方差的扰动。

3.3.1. 组内加权扰动

在不考虑数据结构的情况下，在所有观测数据中找强影响点或异常点，比较常用的方法是给每个数据加权。令 $\omega=\left(\omega_{11},\cdots,\omega_{1n_{1}},\omega_{21},\cdots,\omega_{mn_{m}}\right)^{T}$ 为扰动向量，当 $\omega^{0}=1_{n}$ 时，模型为无扰动模型，其中 1_{n} 为所有元素为 1 的 $n\times1$ 维向量，则组内加权扰动的似然函数可表示为：

$$PL_{c}\left(\psi,\omega\middle|Y_{c}\right)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\omega_{ij}\left[\log a\left(y_{ij};\sigma^{2}\right)-\frac{1}{2\sigma^{2}}d\left(y_{ij};u_{ij}\right)\right]-\sum_{i=1}^{m}\left[\frac{1}{2}\ln\left|\sum\left(\gamma\right)\right|+\frac{1}{2}b_{i}^{T}\sum{}^{-1}\left(\gamma\right)b_{i}\right]-\frac{1}{2}n\lambda\delta^{T}K\delta$$

通过对上式求导我们有：

$$\Delta_{\omega_{ij}}=\left(\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\beta^{T}},\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\delta^{T}},\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\sigma^{2}},\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\gamma^{T}}\right)^{T}$$

其中：

$$\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\beta^{T}}=-\frac{1}{2\sigma^{2}}v_{ij}^{T}\dot{d}_{ij}$$

$$\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\delta^{T}}=-\frac{1}{2\sigma^{2}}\left(B\left(w_{ij}\right)x_{ij}\right)^{T}\dot{d}_{ij}$$

$$\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\sigma^{2}}=\frac{1}{2\sigma^{4}}d_{ij}+\frac{1}{a\left(y_{ij};\sigma^{2}\right)}\frac{\partial a\left(y_{ij};\sigma^{2}\right)}{\partial\sigma^{2}}$$

$$\frac{\partial^{2}PL_{c}\left(\psi,\omega\middle|Y_{c}\right)}{\partial\omega_{ij}\partial\gamma^{T}}=0$$

由此可得到 $\Delta_{\omega}=\left(\Delta_{\omega_{11}},\cdots,\Delta_{\omega_{mn_{m}}}\right)$。

3.3.2. 组间加权扰动

在组间数据中找强影响数据组或异常数据组，比较常用的方法是给每个数据组加权。令

$\omega=(\omega_1,\cdots,\omega_m)^T$ 扰动向量，当 $\omega^0=(1,1,\cdots,1)^T$ 为无扰动模型，则组间加权扰动的似然函数可表示为：

$$PL_c(\psi,\omega|Y_c)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\omega_i\left[\log a(y_{ij};\sigma^2)-\frac{1}{2\sigma^2}d(y_{ij};u_{ij})\right]$$
$$-\sum_{i=1}^{m}\left[\frac{1}{2}\ln|\sum(\gamma)|+\frac{1}{2}b_i^T\sum{}^{-1}(\gamma)b_i\right]-\frac{1}{2}n\lambda\delta^TK\delta$$

通过对上式求导我们有：

$$\Delta_{\omega_i}=\left(\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\beta^T},\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\delta^T},\right.$$
$$\left.\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\sigma^2},\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\gamma^T}\right)^T$$

其中：

$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\beta^T}=-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}v_{ij}^T\dot{d}_{ij}$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\delta^T}=-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\left(B(w_{ij})x_{ij}\right)^T\dot{d}_{ij}$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\sigma^2}=\sum_{j=1}^{n_i}\left(\frac{1}{2\sigma^4}d_{ij}+\frac{1}{a(y_{ij};\sigma^2)}\frac{\partial a(y_{ij};\sigma^2)}{\partial\sigma^2}\right)$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\gamma^T}=0$$

由此可得到 $\Delta_\omega=\left(\Delta_{\omega_1},\cdots,\Delta_{\omega_m}\right)$。

3.3.3. 随机效应方差的扰动

在模型中，随机效应 b_i 是从 $N(0,\sum(\gamma))$ 中随机抽取的，为了进一步研究协方差阵中随机效应的扰动影响，假设 $Var(b_i)=\omega_i\sum(\gamma),i=1,\cdots,m$ 则组间加权扰动的似然函数可表示为：

$$PL_c(\psi,\omega|Y_c)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left[\log a(y_{ij};\sigma^2)-\frac{1}{2\sigma^2}d(y_{ij};u_{ij})\right]$$
$$-\sum_{i=1}^{m}\left[\frac{1}{2}\ln|\Sigma(\gamma)|+\omega_i\frac{1}{2}b_i^T\sum{}^{-1}(\gamma)b_i\right]-\frac{1}{2}n\lambda\delta^TK\delta$$

通过对上式求导我们有：

$$\Delta_{\omega_i}=\left(\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\beta^T},\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\delta^T},\right.$$
$$\left.\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\sigma^2},\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\gamma^T}\right)^T$$

其中：

$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\beta^T}=0$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\delta^T}=0$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\sigma^2}=0$$
$$\frac{\partial^2PL_c(\psi,\omega|Y_c)}{\partial\omega_i\partial\gamma^T}=\frac{1}{2}b_i^T\sum{}^{-1}(\gamma)\frac{\partial\sum(\gamma)}{\partial\gamma^T}\sum{}^{-1}(\gamma)b_i$$

由此可得到 $\Delta_\omega=\left(\Delta_{\omega_1},\cdots,\Delta_{\omega_m}\right)$。

4. 实例分析

结合药物血浆渗透数据(Davidian and Giltinan[14])。来说明本文给出的统计诊断方法的可操作性和有效性。

对 6 个病人自愿者，在 8 小时内通过 11 次静脉注射相同剂量的药物，测得每位病人血液中药物浓度数据。本节用半变系数再生散度混合效应模型拟合该数据。假设 $y_{ij}|b_i$ 服从单参数 I 型极值分布，则 $y_{ij}|b_i$ 的概率密度函数为

$$p(y_{ij}|b_i)=\exp\left\{y_{ij}-\mu_{ij}-\exp(y_{ij}-\mu_{ij})\right\}$$

其中：$\mu_{ij}=x_{ij}\alpha(x_{ij})+x_{ij}\beta+b_i$，则 $d(y_{ij};\mu_{ij})$ 满足单位偏差度函数及再生散度定义的条件。

实际计算中，对于变系数部分，选取固定的 5 个节点且阶数为 3 阶，并通过 GCV 的方法得到光滑参数的估计 $n\lambda=1.342\times10^{-4}$，然后用 Monte Carlo EM 加速算法得到诊断统计量的值。下面列出广义 Cook 距离和局部影响的图形：

1) 个体删除模型

从图 1 中可以发现第 1 号和第 23 号点的影响比较大，其中第 23 号点是数据中数值最大的点。且以上结果与韦博成，林金官，解锋昌[15]的结果一致。

2) 组删除模型

从图 2 中可以发现第 3 组数据的影响最大，因为它包含了强影响点第 23 号点，同时第 1 组数据的影响也较大，因为包含了强影响点第 1 号点。

3) 局部影响分析

从图 3、图 4 和图 5 可以发现，局部影响分析和广义 Cook 距离的结果基本一致，从另一个角度证明了广义 Cook 距离的有效性。

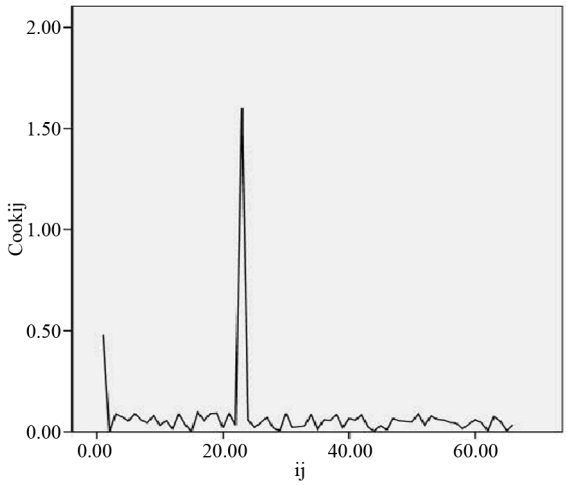

Figure 1. Generalized Cook distance of individual delete model
图 1. 个体删除模型的广义 Cook 距离

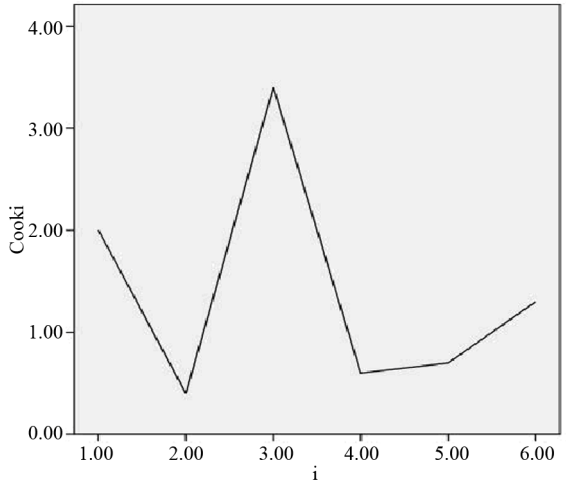

Figure 2. Generalized Cook distance of group delete model
图 2. 组删除模型的广义 Cook 距离

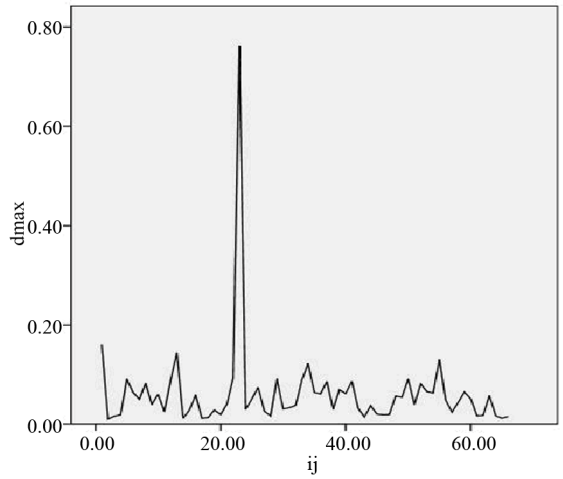

Figure 3. Within group disturbance
图 3. 组内扰动

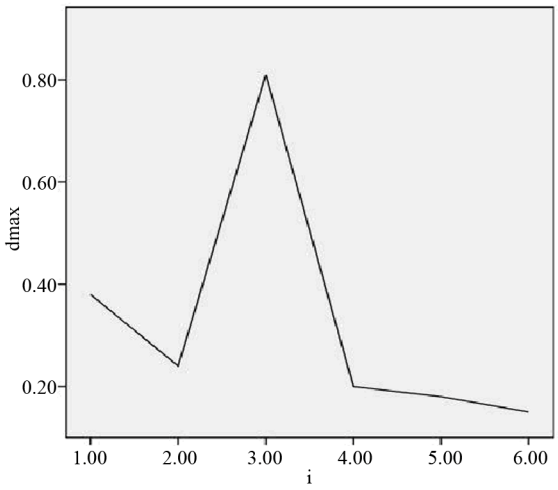

Figure 4. Between group disturbance
图 4. 组间扰动

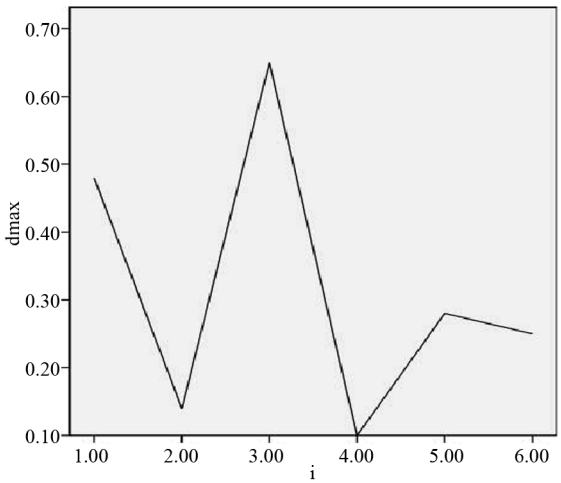

Figure 5. Random effect variance disturbance
图 5. 随机效应方差的扰动

5. 致谢

感谢审稿的宝贵意见以及编辑的帮助。

参考文献 (References)

[1] R. D. Cook, S. Weisberg. Residual and influence in regression. New York: Chapman and Hall, 1982.
[2] G. Seber, C. J. Wild. Nonlinear Regression. New York: Wiley, 1989.
[3] 姜荣, 邵明江, 钱伟民. 半参数非线性模型中的 t-型估计和影响分析[J]. 华东师范大学学报(自然科学版), 2011, 3: 1-11.
[4] R. J. Beckman, C. J. Nachtsheim and R. D. Cook. Diagnostics for mixed-model analysis of variance. Technometrics, 1987, 29(4): 413-426.
[5] 张浩, 朱仲义. 半参数广义线性混合效应模型的影响分析[J]. 应用数学学报, 2007, 30(4): 773-756.
[6] B. Jorgensen. The theory of dispersion models. London: Chap-

man and Hall, 1997
[7] 唐年胜,韦博成. 非线性再生散度模型[M]. 北京: 科学出版社, 2007.
[8] X. H. Lin, D. W. Zhang. Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical Society, Series B, 1999, 61(2): 381-400.
[9] C. E. McCulloch. Maximum likelihood algorithm for generalized linear mixed models. Journal of the American Statistical Association, 1997, 92(437): 162-170.
[10] 罗季. Monte Carlo EM 加速算法[J]. 应用概率统计, 2008, 24(3): 311-318.
[11] Y. Yu, D. Ruppert. Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, 2002, 97(460): 1042-1054.
[12] H. T. Zhu, S. Y. Lee, B. C. Wei and J. Zhu. Case-deletion measures for models with incomplete data. Biometrika, 2001, 88(3): 727-737.
[13] A. Gelman, G. O. Roberts and W. R. Gilks. Efficient metropolis jumping rules. Bayesian statistics 5, Oxford: Oxford University Press, 1995.
[14] M. Davison, D. M. Giltinan. Nonlinear models for repeated measurement data. London: Chapman and Hall, 1995.
[15] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京: 高等教育出版社, 2009.

