应用概率统计　第二十二卷
第三期　2006 年 8 月

Chinese Journal of Applied Probability
and Statistics Vol. 22 No. 3 Aug. 2006

ZI 纵向计数数据模型的影响分析 *

韦博成　　解锋昌

(东南大学数学系，南京，210096)　(南京农业大学数学系，南京，210095)

摘　要

基于 EM 算法和 Laplace 逼近，本文给出了研究 ZI (即含 0 较多的) 纵向计数数据模型的影响分析方法．为了识别含 0 较多的分组计数数据中的强影响点，本文将 ZI 纵向数据模型中取值为 0 的数据赋予一定的权重；而把随机效应看作缺失数据；在此基础上引入 EM 算法，从而应用完全数据对数似然函数的条件期望以及相应的 Q 距离函数进行影响分析；并进一步应用 Laplace 逼近方法简化 EM 算法中的积分计算．在此基础上，基于数据删除模型和局部影响分析方法导出了适用于 ZI 纵向计数数据模型的诊断统计量．本文也通过实际计数数据的例子验证了诊断统计量的有效性．

关 键 词：ZI 随机效应模型，Laplace 逼近，EM 算法，数据删除模型，局部影响分析．

学 科 分 类 号：O212.2.

§1. 引　　言

在生物医学，公共卫生，经济，农业，道路安全等众多领域的数据分析中，计数数据是十分常见的情形．对此类数据，常利用 Poisson 回归模型来揭示作为响应变量的计数数据和一组协变量之间的内在联系．然而，在某些数据中，往往含有比普通 Poisson 回归模型更多的取值为 0 的数据．例如，考虑病人手术后的抱怨次数，一般来说，若病人处于低风险状态，则抱怨的次数往往为 0, 若病人处在高风险状态，则会有比较多抱怨次数．Gupta et al[1] 指出，当观测到额外的取值为 0 的计数数据时，如果仍用普通 Poisson 回归进行拟合，则对于计数数据中取值较小的数据的预测将会产生较大误差．针对含 0 较多的计数数据，很多作者一直在研究和发展相关的统计模型．Cohen[2] 首次考虑了处理含 0 较多的计数数据的调整的 Poisson 模型 (但未考虑协变量); 文献 [3] 提出了 Hurdle Poisson 回归模型；文献 [4] 提出了 Zero-altered Poisson 回归模型；这些都考虑了协变量的情形．基于这些工作，Lambert[5] 首次提出了 Zero-Inflated Poisson (ZIP) 模型，该模型假定取值为 0 的计数数据和取值服从 Poisson 分布的计数数据各占一定比例而组成混合分布．并且在取值为 0 和取值服从 Poisson 的部分均可以引入协变量；由于这一模型比较合理，处理上也较方便，从而成为最常用的处理含 0 较多的计数数据模型．另外，Hall[6] 和 Vieira et al[7] 讨论了 ZIB (ZI binomial) 模型；Ridout et al[8] 研究了 ZINB (ZI negative binomial) 模型．他们分别考虑了含 0 较多的二项分布和负二项分布的计数数据模型．

另外，纵向数据近年来受到各方面的广泛关注，计数数据也常常是经过分组重复测量得到的．这时，组内与组间相比，组内常是相关的，为正确评价响应变量与协变量之间关系，必须

* 国家社会科学基金资助项目 (基金号：04BTJ002).
本文 2006 年 3 月 1 日收到.

考虑组内的相依性，否则就可能导致错误结论 [9]. 为此，人们常常选择随机效应模型．对计数数据， Breslow[9] 建议利用随机效应 Poisson 模型； Thall[10] 利用 Poisson-Gamma 回归模型分析纵向区间计数数据．另外， Siddiqui[11] 利用 Poisson 随机效应回归模型处理分组计数数据．

最近，为了能够同时拟合含 0 较多和重复测量的纵向计数数据， Hall[6] 和 Yau and Lee[12] 将随机效应引入 ZIP 和 ZIB 模型中．由于随机效应的引入，使得 ZI 模型变得复杂，若直接由观测数据的对数似然进行统计分析将很困难，于是出现了多种处理方法，如文献 [6] 采用 EM 算法结合高斯求积 (Gaussian quadrature) 方法进行参数估计；文献 [13] 利用二阶 Taylor 展开进行异方差和偏大离差的 Score 检验；文献 [14] 给出了约束极大似然估计；等等．

本文研究 ZI 纵向计数数据模型的影响分析问题，这也是数据分析的一个重要环节．由于同时存在 ZI 和随机效应，很难应用 Cook 等人的传统方法 [15,16] 进行影响分析．本文应用 Zhu et al[17,18] 提出的 Q 函数方法，研究了模型的影响分析问题；这方面的研究在文献中还未见报道．在这里，我们将随机效应看作缺失数据，引进 EM 算法 [19] 和 Laplace 逼近 [20,21], 从而基于完全数据对数似然函数的条件期望进行影响分析．本文分别得到了基于数据删除模型的诊断统计量以及在各种扰动方案下基于正则曲率的诊断统计量．

§2. ZI 随机效应模型及其参数估计

为了研究纵向计数数据, 我们假定数据集包含响应变量 y_{ij} 和协变量 $X_{ij}(p_1\times 1)$, $Z_{ij}(p_2\times 1)$, $W_{ij}(p_3\times 1)$, 他们分别是第 i $(i=1,\cdots,n)$ 组中第 j $(j=1,\cdots,t_i)$ 个数据，同时假定

$$Y_{ij}|b_i\sim\begin{cases}0, & \text{若具有概率 } p_{ij};\\ F_1(y_{ij}|b_i), & \text{若具有概率 } 1-p_{ij},\end{cases}\tag{1}$$

其中 F_1 是指数族分布, 具有密度 $f_1(y_{ij}|b_i)=\exp\{\phi\rho_{ij}[y_{ij}\theta_{ij}-a(\theta_{ij})]+c(y_{ij},\phi)\}$. 对于纵向计数数据， F_1 通常取为 Poisson 分布或二项分布，但是在下面的推导中本文并未对 F_1 加以太多限制，而是结合一般的指数族分布进行讨论．另外, 式中 ρ_{ij} 是已知常数， ϕ 是尺度参数, 且 $Y_{ij}|b_i$ 具有条件均值 $\mu_{ij}=\dot a(\theta_{ij})$ 和条件方差 $\ddot a(\theta_{ij})/(\phi\rho_{ij})$, 其中 $\dot a(\cdot)$, $\ddot a(\cdot)$ 是 $a(\cdot)$ 的一二阶导数．按照广义线性模型的方法, 假定 $g(\mu_{ij})=\eta_{ij}=X_{ij}^T\beta+Z_{ij}^Tb_i$, 或 $\theta_{ij}=k(X_{ij}^T\beta+Z_{ij}^Tb_i)=\dot a^{-1}\{g^{-1}(\eta_{ij})\}$, 其中 $\beta=(\beta_1,\cdots,\beta_{p_1})^T$ 是回归系数， k 和 g 连续可微， $\dot a^{-1}$, g^{-1} 分别是函数 $\dot a$, g 的反函数. 同时, 假定混合概率 p_{ij} 满足 $g_p(p_{ij})=W_{ij}^T\gamma$, 其中 g_p 是联系函数, 如 g_p 可取为 logit 等形式， γ 是 $p_3\times 1$ 未知参数．另外, 假定随机效应 $b_1,\cdots,b_n$ 相互独立, 且有 $b_i\sim N(0,\Sigma)$, 其中 $\Sigma=\Sigma(\delta)$, δ 是 $p_4\times 1$ 未知方差成分参数．

记 $\psi=(\beta^T,\phi,\gamma^T,\delta^T)^T$ 表示 ZI 随机效应模型 (1) 的参数， Y_o 表示可观测数据集．则基于 Y_o 的对数似然 $l_o(\psi|Y_o)$ 可表示为

$$l_o(\psi|Y_o)=C_0+\sum_{i=1}^n\log\Big\{\int\prod_{j=1}^{t_i}f(y_{ij}|b_i)\phi_{p_2}(b_i)\mathrm{d}b_i\Big\},\tag{2}$$

式中 C_0 是常数， $f(y_{ij}|b_i)=\{p_{ij}+(1-p_{ij})f_1(y_{ij}|b_i)\}^{I_{\{y_{ij}=0\}}}\{(1-p_{ij})f_1(y_{ij}|b_i)\}^{1-I_{\{y_{ij}=0\}}}$, 其中 $I_A=1$, 若条件 A 成立； $I_A=0$, 若条件 A 不成立．另外 $\phi_{p_2}(\cdot)$ 表示 p_2 维正态密度函数， (2)

式是在 $(-\infty,\infty)\times\cdots\times(-\infty,\infty)$ 上的积分.

对于 ZI 随机效应模型 (1), 若利用 (2) 式进行参数估计, 则相当困难, 因为 (2) 式涉及到高维积分, 且难以求出解析解. 同样, 从 (2) 式出发进行影响分析也很困难. 为此, 可应用 EM 算法, 将模型中随机效应 $\{b_1,\cdots,b_n\}$ 视为缺失数据, 记为 Y_m, 并用 $Y_c=\{Y_o,Y_m\}$ 表示完全数据. 则基于完全数据的对数似然 $l_c(\psi|Y_c)$ 可表示为

$$\begin{aligned} l_c(\psi|Y_c) &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\{I_{\{y_{ij}=0\}}\log(p_{ij}+(1-p_{ij})\exp[-\phi\rho_{ij}a(\theta_{ij})+c(0,\phi)]) \\ &\quad +[1-I_{\{y_{ij}=0\}}][\log(1-p_{ij})+\phi\rho_{ij}y_{ij}\theta_{ij}-\phi\rho_{ij}a(\theta_{ij})+c(y_{ij},\phi)]\} \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}b_i^T\Sigma^{-1}b_i-\frac{n}{2}\log|\Sigma|. \end{aligned} \tag{3}$$

(3) 式中常数被略去. 下面基于 $l_c(\psi|Y_c)$ 利用 EM 算法 [19] 来估计参数 ψ. 标准的 EM 算法包含两步, 即 E-step: $Q(\psi|\psi^{(h)})=\mathsf{E}\{l_c(\psi|Y_c)|Y_o,\psi^{(h)}\}$, M-step: $\psi^{(h+1)}=\arg\max\limits_{\psi}Q(\psi|\psi^{(h)})$, 式中 $\psi^{(h)}$ 是 EM 算法第 h 步参数估计. 可以证明 EM 算法中获得的序列 $\{\psi^{(h)}\}$ 收敛到参数 ψ 的极大似然估计 $\widehat{\psi}$, 以上的细节可参见文献 [22].

值得注意的是, 在 E-step 中, 需计算条件期望, 但难以得到解析式. 对此类积分的求解, 多数是采用数值逼近或 MCMC 方法, 文献 [21] 建议利用 Laplace 一阶或二阶逼近来求解条件期望, 其优点在于计算量相对于其他数值计算方法要少得多. 本文将首次利用 Laplace 逼近结合 EM 算法来探讨 ZI 随机效应模型的计算方法并应用于影响分析问题.

Laplace 逼近方法 [20] 是多维积分近似的一般方法, 在相对较弱条件下, 有

$$\int_A\zeta(t)\exp\{-\xi m(t)\}\mathrm{d}t\approx\zeta(\widehat{t})\exp\{-\xi m(\widehat{t})\}\Big\{\frac{(2\pi)^d}{|\xi M(\widehat{t})|}\Big\}^{1/2}\qquad(\xi\to+\infty),$$

其中 A 是 $\mathbf{R}^d$ 上开集, $\xi>0$, $M(t)$ 是 $m(t)$ 的二阶导数矩阵, $\widehat{t}$ 是 $m(t)$ 在 A 上的唯一最小点.

而在上面 E-step 的条件期望中, 涉及到下面的积分形式

$$\mathsf{E}\{\zeta(b_i)|Y_o,\psi^{(h)}\}=\frac{\displaystyle\int\zeta(b_i)\exp\Big\{\sum_{j=1}^{t_i}\log f(y_{ij}|b_i)+\log\phi_{p_2}(b_i)\Big\}\mathrm{d}b_i}{\displaystyle\int\exp\Big\{\sum_{j=1}^{t_i}\log f(y_{ij}|b_i)+\log\phi_{p_2}(b_i)\Big\}\mathrm{d}b_i}.$$

利用上面的 Laplace 公式, 有 $\mathsf{E}\{\zeta(b_i)\}\approx\zeta(\widehat{b}_i)$, 其中 $\widehat{b}_i=\arg\min\limits_{b_i}\Big\{-\sum\limits_{j=1}^{t_i}\log f(y_{ij}|b_i)-\log\phi_{p_2}(b_i)\Big\}$.

§3. 基于数据删除的影响分析

数据删除模型是统计诊断最基本的模型之一. 对于分组数据, 一般有两种删除方案: 即删除一组中的某一个数据, 以及删除一个组的数据. 在下文中, 凡是带有下标 "$[i]$" 或 "$[ij]$" 的量表示将原来数据中第 i 组数据或第 i 组中第 j 个数据予以删除.

3.1 删除一个数据时的诊断统计量

设完全数据中删除第 i 组中第 j 个数据，相应的对数似然为 $l_c(\psi|Y_{c[ij]})$. 又设删除前后的 Q 函数分别为 $Q(\psi|\widehat{\psi}) = \mathsf{E}\{l_c(\psi|Y_c)|Y_o, \widehat{\psi}\}$ 和 $Q_{[ij]}(\psi|\widehat{\psi}) = \mathsf{E}\{l_c(\psi|Y_{c[ij]})|Y_{o[ij]}, \widehat{\psi}\}$, 其中 $\widehat{\psi}$ 是 ψ 的极大似然估计. 令 $\widehat{\psi}_{[ij]} = \arg\max_{\psi} Q_{[ij]}(\psi|\widehat{\psi})$. 为了诊断第 i 组中第 j 个数据对极大似然估计 $\widehat{\psi}$ 的影响，基本方法就是比较 $\widehat{\psi}$ 与 $\widehat{\psi}_{[ij]}$ 的差异. 若删除第 i 组中第 j 个数据时对参数估计有严重影响，即 $\widehat{\psi}$ 与 $\widehat{\psi}_{[ij]}$ 相差较大，则认为第 i 组中第 j 个数据是强影响点. 但是，对于每个数据点都要计算 $\widehat{\psi}_{[ij]}$, 所涉及到的计算量非常大，特别是当 $\sum_{i=1}^{n} t_i$ 很大时. 因此，通常采用下面的一步近似 [17,23] 减轻负担.

$$\widehat{\psi}^{1}_{[ij]} = \widehat{\psi} + \{-\ddot{Q}(\widehat{\psi}|\widehat{\psi})\}^{-1}\dot{Q}_{[ij]}(\widehat{\psi}|\widehat{\psi}), \tag{4}$$

其中，$\dot{Q}_{[ij]}(\widehat{\psi}|\widehat{\psi}) = \partial Q_{[ij]}(\psi|\widehat{\psi})/\partial\psi|_{\psi=\widehat{\psi}}$, $\ddot{Q}(\widehat{\psi}|\widehat{\psi}) = \partial^2 Q(\psi|\widehat{\psi})/\partial\psi\partial\psi^T|_{\psi=\widehat{\psi}}$. 为了计算 $\dot{Q}_{[ij]}(\widehat{\psi}|\widehat{\psi})$, $\ddot{Q}(\widehat{\psi}|\widehat{\psi})$, 需要求 $l_c(\psi|Y_c)$ 在删除前后关于参数的导数.

根据 $l_c(\psi|Y_c)$ 的表达式，通过计算，可得下面二阶导数 (未列出的表达式为 0)

$$\begin{aligned}
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\beta\partial\beta^T} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{p_{ij}(1-p_{ij})f_1(0|b_i)(\phi\rho_{ij}\dot{a}(\theta_{ij})\dot{k})^2}{v_{ij}^2}X_{ij}X_{ij}^T \\
&\quad + I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)(-\phi\rho_{ij})}{v_{ij}}[\ddot{a}(\theta_{ij})\dot{k}^2 + \dot{a}(\theta_{ij})\ddot{k}]X_{ij}X_{ij}^T \\
&\quad + [1-I_{\{y_{ij}=0\}}]\phi\rho_{ij}[y_{ij}\ddot{k} - \ddot{a}(\theta_{ij})\dot{k}^2 - \dot{a}(\theta_{ij})\ddot{k}]X_{ij}X_{ij}^T\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\beta\partial\phi} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{p_{ij}f_1(0|b_i)(-\rho_{ij}a(\theta_{ij}) + \dot{c}_\phi(0,\phi))}{v_{ij}^2}(-\phi\rho_{ij})\dot{a}(\theta_{ij})\dot{k}X_{ij} \\
&\quad + I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\rho_{ij}\dot{a}(\theta_{ij})\dot{k}X_{ij} + [1-I_{\{y_{ij}=0\}}]\rho_{ij}[y_{ij} - \dot{a}(\theta_{ij})]\dot{k}X_{ij}\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\beta\partial\gamma^T} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{f_1(0|b_i)\phi\rho_{ij}\dot{a}(\theta_{ij})\dot{k}}{v_{ij}^2}X_{ij}\dot{p}_{ij}^T\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\phi^2} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{p_{ij}(1-p_{ij})f_1(0|b_i)}{v_{ij}^2}[-\rho_{ij}a(\theta_{ij}) + \dot{c}_\phi(0,\phi)]^2 \\
&\quad + I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\ddot{c}_\phi(0,\phi) + [1-I_{\{y_{ij}=0\}}]\ddot{c}_\phi(y_{ij},\phi)\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\phi\partial\gamma^T} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{-f_1(0|b_i)}{v_{ij}^2}[-\rho_{ij}a(\theta_{ij}) + \dot{c}_\phi(0,\phi)]\dot{p}_{ij}^T\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\gamma\partial\gamma^T} &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{v_{ij}[1-f_1(0|b_i)]\ddot{p}_{ij} - [1-f_1(0|b_i)]^2\dot{p}_{ij}\dot{p}_{ij}^T}{v_{ij}^2} \\
&\quad - [1-I_{\{y_{ij}=0\}}]\frac{(1-p_{ij})\ddot{p}_{ij} + \dot{p}_{ij}\dot{p}_{ij}^T}{(1-p_{ij})^2}\Big\}, \\
\frac{\partial^2 l_c(\psi|Y_c)}{\partial\delta_{t_1}\partial\delta_{t_2}} &= \frac{n}{2}\mathrm{tr}\{\Sigma^{-1}\dot{\Sigma}(t_1)\Sigma^{-1}\dot{\Sigma}(t_2)\Sigma^{-1}(\Sigma - 2S_b) + \Sigma^{-1}(S_b - \Sigma)\Sigma^{-1}\ddot{\Sigma}(t_1,t_2)\}.
\end{aligned}$$

其中，$v_{ij} = p_{ij} + (1-p_{ij})\exp((-\phi\rho_{ij}a(\theta)ij) + c(0,\phi))$, $\dot{k}$, $\ddot{k}$ 是函数 k 的一二阶导数，如 $\dot{k}(u) = \mathrm{d}k(u)/\mathrm{d}u$, $\ddot{k}(u) = \mathrm{d}^2k(u)/\mathrm{d}u^2$, $\dot{c}_\phi(0,\phi) = \mathrm{d}c(0,\phi)/\mathrm{d}\phi$, $\ddot{c}_\phi(0,\phi) = \mathrm{d}^2c(0,\phi)/\mathrm{d}\phi^2$, $\dot{p}_{ij} = \partial p_{ij}/\partial\gamma$, $\ddot{p}_{ij} = \partial^2 p_{ij}/\partial\gamma\partial\gamma^T$, $\dot{\Sigma}(t_1) = \partial\Sigma/\partial\delta_{t_1}$, $\ddot{\Sigma}(t_1,t_2) = \partial^2\Sigma/\partial\delta_{t_1}\partial\delta_{t_2}$, $S_b = \sum_{i=1}^{n} b_i b_i^T/n$.

根据 $l_c(\psi|Y_{c[ij]})$ 的表达式可得下面的导数

$$\begin{aligned}
\frac{\partial l_c(\widehat{\psi}|Y_{c[ij]})}{\partial \beta} &= -\Big\{ -I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\phi\rho_{ij}\dot{a}(\theta_{ij})\dot{k}X_{ij} \\
&\quad + [1-I_{\{y_{ij}=0\}}]\phi\rho_{ij}[y_{ij}-\dot{a}(\theta_{ij})]\dot{k}X_{ij}\Big\}_{\widehat{\psi}}, \\
\frac{\partial l_c(\widehat{\psi}|Y_{c[ij]})}{\partial \phi} &= -\Big\{ I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}[-\rho_{ij}a(\theta_{ij})+\dot{c}_\phi(0,\phi)] \\
&\quad + [1-I_{\{y_{ij}=0\}}][\rho_{ij}y_{ij}\theta_{ij}-\rho_{ij}a(\theta_{ij})+\dot{c}_\phi(y_{ij},\phi)]\Big\}_{\widehat{\psi}}, \\
\frac{\partial l_c(\widehat{\psi}|Y_{c[ij]})}{\partial \gamma} &= -\Big\{ I_{\{y_{ij}=0\}}\frac{\dot{p}_{ij}(1-f_1(0|b_i))}{v_{ij}} - [1-I_{\{y_{ij}=0\}}]\frac{\dot{p}_{ij}}{1-p_{ij}}\Big\}_{\widehat{\psi}}, \\
\frac{\partial l_c(\widehat{\psi}|Y_{c[ij]})}{\partial \delta_{t_1}} &= -\Big\{ \frac{1}{2}(b_i^T\Sigma^{-1}\dot{\Sigma}(t_1)\Sigma^{-1}b_i) - \frac{1}{2}\mathrm{tr}(\dot{\Sigma}(t_1)\Sigma^{-1})\Big\}_{\widehat{\psi}}.
\end{aligned}$$

但由于 $\widehat{\psi}_{[ij]}-\widehat{\psi}$ 是一个向量, 不便比较, 必须选择一个合适的距离, 以便定量地比较影响的大小. 这可考虑在影响分析中经常使用的广义 Cook 距离 $\mathrm{GD}_{ij}=(\widehat{\psi}_{[ij]}-\widehat{\psi})^T\{-\ddot{Q}(\widehat{\psi}|\widehat{\psi})\}(\widehat{\psi}_{[ij]}-\widehat{\psi})$ 以及类似于似然距离的 Q 距离 $\mathrm{QD}_{ij}=2\{Q(\widehat{\psi}|\widehat{\psi})-Q(\widehat{\psi}_{[ij]}|\widehat{\psi})\}$. 结合 (4) 式, 亦可得到一步近似公式 $\mathrm{GD}_{ij}^1=\dot{Q}_{[ij]}(\widehat{\psi}|\widehat{\psi})^T\{-\ddot{Q}(\widehat{\psi}|\widehat{\psi})\}^{-1}\dot{Q}_{[ij]}(\widehat{\psi}|\widehat{\psi})$ 和 $\mathrm{QD}_{ij}^1=2\{Q(\widehat{\psi}|\widehat{\psi})-Q(\widehat{\psi}_{[ij]}^1|\widehat{\psi})\}$.

3.2 删除一组数据时的诊断统计量

对于删除一组数据的情形, 同样有一步近似公式

$$\widehat{\psi}_{[i]}^1=\widehat{\psi}+\{-\ddot{Q}(\widehat{\psi}|\widehat{\psi})\}^{-1}\dot{Q}_{[i]}(\widehat{\psi}|\widehat{\psi}), \tag{5}$$

其中, $\dot{Q}_{[i]}(\widehat{\psi}|\widehat{\psi})=\partial Q_{[i]}(\psi|\widehat{\psi})/\partial\psi|_{\psi=\widehat{\psi}}$. 另外, 为了得到 $\dot{Q}_{[i]}(\widehat{\psi}|\widehat{\psi})$, 需要计算 $l_c(\psi|Y_{c[i]})$ 关于参数的导数. 通过简单推导可得 $\partial l_c(\psi|Y_{c[i]})/\partial\psi=\sum\limits_{j=1}^{t_i}[\partial l_c(\psi|Y_{c[ij]})/\partial\psi]$, 从而结合 (5) 式有类似于 3.1 节中的一步近似 GD_i^1 和 QD_i^1.

§4. 局部影响分析

令 ω 是一个定义在 $\Omega\subset\mathbf{R}^N$ 上的 N 维向量, 表示对模型的扰动因素; $l_o(\psi,\omega|Y_o)$ 和 $l_c(\psi,\omega|Y_c)$ 分别是受扰动模型的基于观测数据的对数似然和基于完全数据的对数似然. 我们假定存在 ω^0 使得 $l_o(\psi,\omega^0|Y_o)=l_o(\psi|Y_o)$ 和 $l_c(\psi,\omega^0|Y_c)=l_c(\psi|Y_c)$ 对于所有 ψ 成立. 设 $\widehat{\psi}$, $\widehat{\psi}(\omega)$ 分别是无扰动模型和扰动模型中参数 ψ 的极大似然估计. 则由 EM 算法知, $\widehat{\psi}$ 和 $\widehat{\psi}(\omega)$ 分别使得 $Q(\psi|\widehat{\psi})=\mathsf{E}\{l_c(\psi|Y_c)|Y_o,\widehat{\psi}\}$ 和 $Q(\psi,\omega|\widehat{\psi}(\omega))=\mathsf{E}\{l_c(\psi,\omega|Y_c)|Y_o,\widehat{\psi}(\omega)\}$ 达到最大值.

对于不完全数据问题, 考虑 Q 距离函数 [18,24,25] $f_Q(\omega)=2[Q(\widehat{\psi}|\widehat{\psi})-Q(\widehat{\psi}(\omega)|\widehat{\psi})]$, 其中 $Q(\widehat{\psi}|\widehat{\psi})=Q(\widehat{\psi},\omega^0|\widehat{\psi})$, $Q(\widehat{\psi}(\omega)|\widehat{\psi})=Q(\widehat{\psi}(\omega),\omega^0|\widehat{\psi})$. 由 [18] 可知 $f_Q(\omega)$ 与 Cook[16] 提出的似然距离函数有着密切联系和相似的统计性质. 根据 [16] 中的思想, 我们考虑影响图 $\eta(\omega)=(\omega^T,f_Q(\omega))^T$, 其影响曲率可表示为

$$c_{f_Q,d}=-2d^T\ddot{Q}_{\omega^0}d=-2d^T\triangle_{\omega^0}^T\ddot{Q}_\psi^{-1}(\widehat{\psi})\triangle_{\omega^0}d, \tag{6}$$

其中

$$\ddot{Q}_{\omega^0}=\frac{\partial^2 Q\{\widehat{\psi}(\omega)|\widehat{\psi}\}}{\partial\omega\partial\omega^T}\Big|_{\omega=\omega^0},\qquad \ddot{Q}_{\psi}(\widehat{\psi})=\frac{\partial^2 Q(\psi|\widehat{\psi})}{\partial\psi\partial\psi^T}\Big|_{\psi=\widehat{\psi}},\qquad \triangle_{\omega}=\frac{\partial^2 Q(\psi,\omega|\widehat{\psi})}{\partial\psi\partial\omega^T}\Big|_{\psi=\widehat{\psi}(\omega)}.$$

以上结果表明，最大影响曲率可表示为 $c_{\max}=\lambda_1$，λ_1 为影响矩阵 $\ddot{Q}_{\omega^0}$ 的绝对值最大的特征值，对应于 λ_1 的特征向量 $d_{\max}$ 为最大影响曲率方向．

下面分别考虑组内加权扰动，组间加权扰动，随机效应方差发生扰动和解释变量发生扰动四种情形．为了根据公式 (6) 计算四种情形下的影响曲率，主要是求出 $\partial^2 l_c(\psi,\omega|Y_c)/\partial\psi\partial\omega^T$, 从而可得 $\triangle_{\omega}$, 而公式 (6) 中的 $\ddot{Q}_{\psi}(\widehat{\psi})$ 可参见前一节．

情形 1　组内加权扰动

现考虑加权扰动模型，设 $\omega=(\omega_{11},\cdots,\omega_{1t_1},\omega_{21},\cdots,\omega_{nt_n})^T$ 为加权扰动向量，$\omega^0=(1,1,\cdots,1)^T$ 对应于无扰动情形；则基于完全数据的组内加权扰动模型的对数似然函数为

$$\begin{aligned}l_c(\psi,\omega|Y_c) &= \sum_{i=1}^{n}\sum_{j=1}^{t_i}\omega_{ij}\{I_{\{y_{ij}=0\}}\log(p_{ij}+(1-p_{ij})\exp[-\phi\rho_{ij}a(\theta_{ij})+c(y_{ij},\phi)])\\ &\quad+[1-I_{\{y_{ij}=0\}}][\log(1-p_{ij})+\phi\rho_{ij}y_{ij}\theta_{ij}-\phi\rho_{ij}a(\theta_{ij})+c(y_{ij},\phi)]\}\\ &\quad-\frac{1}{2}\sum_{i=1}^{n}b_i^T\Sigma^{-1}b_i-\frac{n}{2}\log|\Sigma|.\end{aligned}\tag{7}$$

利用式 (7) 可得

$$\begin{aligned}\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\beta\partial\omega_{ij}} &= -I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\phi\rho_{ij}\dot{a}(\theta_{ij})\dot{k}X_{ij}\\ &\quad+[1-I_{\{y_{ij}=0\}}]\phi\rho_{ij}[y_{ij}-\dot{a}(\theta_{ij})]\dot{k}X_{ij},\\ \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\phi\partial\omega_{ij}} &= I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}[-\rho_{ij}a(\theta_{ij})+\dot{c}_{\phi}(0,\phi)]\\ &\quad+[1-I_{\{y_{ij}=0\}}][\rho_{ij}y_{ij}\theta_{ij}-\rho_{ij}a(\theta_{ij})+\dot{c}_{\phi}(y_{ij},\phi)],\\ \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\gamma\partial\omega_{ij}} &= I_{\{y_{ij}=0\}}\frac{\dot{p}_{ij}[1-f_1(0|b_i)]}{v_{ij}}-[1-I_{\{y_{ij}=0\}}]\frac{\dot{p}_{ij}}{1-p_{ij}},\\ \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\delta\partial\omega_{ij}} &= 0.\end{aligned}$$

情形 2　组间加权扰动

设 $\omega=(\omega_1,\cdots,\omega_n)^T$ 为加权扰动向量，$\omega^0=(1,1,\cdots,1)^T$ 对应于无扰动情形；则基于完全数据的组间加权扰动模型的对数似然函数为

$$\begin{aligned}l_c(\psi,\omega|Y_c) &= \sum_{i=1}^{n}\omega_i\sum_{j=1}^{t_i}\{I_{\{y_{ij}=0\}}\log(p_{ij}+(1-p_{ij})\exp[-\phi\rho_{ij}a(\theta_{ij})+c(y_{ij},\phi)])\\ &\quad+[1-I_{\{y_{ij}=0\}}][\log(1-p_{ij})+\phi\rho_{ij}y_{ij}\theta_{ij}-\phi\rho_{ij}a(\theta_{ij})+c(y_{ij},\phi)]\}\\ &\quad-\frac{1}{2}\sum_{i=1}^{n}\omega_i(b_i^T\Sigma^{-1}b_i+\log|\Sigma|).\end{aligned}\tag{8}$$

利用式 (8) 可得

$$\begin{aligned}
\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\beta\partial\omega_i} &= \sum_{j=1}^{t_i}\Big\{-I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\phi\rho_{ij}\dot a(\theta_{ij})\dot k X_{ij}\\
&\quad +[1-I_{\{y_{ij}=0\}}]\phi\rho_{ij}[y_{ij}-\dot a(\theta_{ij})]\dot k X_{ij}\Big\},\\
\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\phi\partial\omega_i} &= \sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}[-\rho_{ij}a(\theta_{ij})+\dot c_\phi(0,\phi)]\\
&\quad +[1-I_{\{y_{ij}=0\}}][\rho_{ij}y_{ij}\theta_{ij}-\rho_{ij}a(\theta_{ij})+\dot c_\phi(y_{ij},\phi)]\Big\},\\
\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\gamma\partial\omega_i} &= \sum_{j=1}^{t_i}\Big\{I_{\{y_{ij}=0\}}\frac{\dot p_{ij}[1-f_1(0|b_i)]}{v_{ij}}-[1-I_{\{y_{ij}=0\}}]\frac{\dot p_{ij}}{1-p_{ij}}\Big\},\\
\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\delta\partial\omega_i} &= \frac12 b_i^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\delta}\Sigma^{-1}b_i-\frac12\mathrm{tr}\Big(\frac{\partial\Sigma}{\partial\delta}\Sigma^{-1}\Big).
\end{aligned}$$

情形 3 随机效应方差的扰动

文中假定随机效应 $b_i\sim N(0,\Sigma)$, 为了研究此假定中方差矩阵发生扰动时的效应, 我们设 $\mathsf{Var}\,(b_i)=\omega_i^{-1}\Sigma(\delta)$, $i=1,\cdots,n$. 则可得 $l_c(\psi,\omega|Y_c)$ 关于参数的导数为

$$\frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\beta\partial\omega_i}=0,\quad \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\phi\partial\omega_i}=0,\quad \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\gamma\partial\omega_i}=0,\quad \frac{\partial^2 l_c(\psi,\omega|Y_c)}{\partial\delta\partial\omega_i}=\frac12 b_i^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\delta}\Sigma^{-1}b_i.$$

情形 4 解释变量的扰动

为了简单, 仅考虑解释变量 X_{ij} 的扰动. 设 $s_1,\cdots,s_p$ 是解释变量中不同度量单位的尺度因子, 则 $X_{ij}(\omega)=X_{ij}+S\omega_{ij}$, ω_{ij} 是 $p_1\times1$ 扰动向量, $S=\mathrm{diag}(s_1,\cdots,s_p)$. $\omega^0=0$ 表示模型没有扰动. 注意此扰动对示性变量是没意义的. 于是可得 $l_c(\psi,\omega|Y_c)$ 关于参数 ψ 和 ω 的导数

$$\begin{aligned}
\frac{\partial^2 l_c(\psi,\omega^0|Y_c)}{\partial\beta\partial\omega_{ij}} &= I_{\{y_{ij}=0\}}\frac{p_{ij}(1-p_{ij})f_1(0|b_i)}{v_{ij}^2}[\phi\rho_{ij}\dot a(\theta_{ij})\dot k]^2X_{ij}(S\beta)^T\\
&\quad -I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)\phi\rho_{ij}}{v_{ij}}[(\ddot a(\theta_{ij})\dot k^2+\dot a(\theta_{ij})\ddot k)X_{ij}(S\beta)^T+\dot a(\theta_{ij})\dot kS]\\
&\quad +[1-I_{\{y_{ij}=0\}}][\phi\rho_{ij}y_{ij}(\ddot kX_{ij}(S\beta)^T+\dot kS)\\
&\quad -\phi\rho_{ij}((\ddot a(\theta_{ij})\dot k^2+\dot a(\theta_{ij})\ddot k)X_{ij}(S\beta)^T+\dot a(\theta_{ij})\dot kS)],\\
\frac{\partial^2 l_c(\psi,\omega^0|Y_c)}{\partial\phi\partial\omega_{ij}} &= -I_{\{y_{ij}=0\}}\frac{p_{ij}(1-p_{ij})f_1(0|b_i)\phi\rho_{ij}\dot a(\theta_{ij})\dot k}{v_{ij}^2}[-\rho_{ij}a(\theta_{ij})+\dot c_\phi(0,\phi)](S\beta)^T\\
&\quad -I_{\{y_{ij}=0\}}\frac{(1-p_{ij})f_1(0|b_i)}{v_{ij}}\rho_{ij}\dot a(\theta_{ij})\dot k(S\beta)^T\\
&\quad +[1-I_{\{y_{ij}=0\}}]\rho_{ij}[y_{ij}-\dot a(\theta_{ij})]\dot k(S\beta)^T,\\
\frac{\partial^2 l_c(\psi,\omega^0|Y_c)}{\partial\gamma\partial\omega_{ij}} &= I_{\{y_{ij}=0\}}\frac{\phi\rho_{ij}\dot a(\theta_{ij})\dot kf_1(0|b_i)}{v_{ij}^2}\dot p_{ij}(S\beta)^T,\\
\frac{\partial^2 l_c(\psi,\omega^0|Y_c)}{\partial\delta\partial\omega_{ij}} &= 0.
\end{aligned}$$

§5. 实例：粉虱数据

以下应用粉虱数据 [26] 来说明本文导出的影响分析统计量的有效性．这批数据来源于园艺试验中利用杀虫剂控制温室栽培的一品红上的银叶粉虱．试验设计是完全随机分组的，每周重复测量，共计 12 周．试验中每三株一品红作为一个试验单位，共有 18 个试验单位，它们被随机分成三个不同区组进行 6 种不同试验．当粉虱出现于固定在叶子上的笼子里两天后，开始计量其中存活的昆虫数，然后重复下去．文献 [14, 26] 详细研究了这组数据，其中文献 [14] 假定在不同株一品红上具有随机效应，利用 ZIB 随机效应模型拟合数据，但是该文并未对数据进行影响分析．以下就着重讨论这方面的问题．

设 y_{ijkl} 为第 k $(k=1,2,\cdots,54)$ 株一品红上昆虫在第 i $(i=1,\cdots,6)$ 个试验条件 (treatment) 下，第 j $(j=1,2,3)$ 个区组 (block)，第 l $(l=1,\cdots,12)$ 周 (week) 观测中存活数．设 n_{ijkl} 是第 k 株一品红上昆虫在第 i 个试验条件下，第 j 个区组，第 l 周观测中总的昆虫数．进一步，设 β_{2i} 是第 i 个试验效应，β_{3j} 是第 j 个区组效应，β_{4l} 是第 l 周效应，b_k 是第 k 株植物的一维随机效应，服从标准正态分布．为了简单，我们只考虑模型包含主效应 (试验，区组和周)．具有主效应的 ZIB 随机效应模型可表示为

$$Y_{ijkl}|b_k \sim \{p_{ijkl}+(1-p_{ijkl})(1-\pi_{ijkl})^{n_{ijkl}}\}^{I_{\{y_{ijkl}=0\}}}\{(1-p_{ijkl})B(n_{ijkl},\pi_{ijkl}|b_k)\}^{1-I_{\{y_{ijkl}=0\}}},$$

其中，$B(n_{ijkl},\pi_{ijkl}|b_k)$ 表示二项概率，另外 $\mathrm{logit}(\pi_{ijkl})=\beta_1+\beta_{2i}\mathrm{treatment}_i+\beta_{3j}\mathrm{block}_j+\beta_{4l}\mathrm{week}_l+\sigma b_k$, $\mathrm{logit}(p_{ijkl})=\gamma_1+\gamma_{2i}\mathrm{treatment}_i+\gamma_{3j}\mathrm{block}_j+\gamma_{4l}\mathrm{week}_l$, 利用 EM 算法和 Laplace 一阶逼近可以得到参数的极大似然估计 (记为 ELML)，表 1 分别列出了此估计与文献 [14] 中所给的参数的极大似然估计 (ML)．从表 1 可以看出，本文提出的 ELML 估计和文献 [14] 中所给的 ML 估计基本一致，相差较小．另外，本文的 ELML 估计的标准差最大为 0.0214，最小为 0.0030，平均标准差为 0.0121．这些都表明本文方法在估计参数时是有效的．

表 1　ZIB 随机效应模型拟合粉虱数据时的参数估计

参数	ML	ELML	参数	ML	ELML	参数	ML	ELML
β_1	-0.5733	-0.5245	β_{46}	-0.4278	-0.4702	γ_{31}	0.0504	0.0432
β_{21}	-1.0577	-1.0930	β_{47}	-0.0111	-0.0404	γ_{32}	0.1584	0.1481
β_{22}	-0.6270	-0.6535	β_{48}	-0.4898	-0.5051	γ_{41}	-0.5168	-0.5240
β_{23}	-1.0894	-1.1582	β_{49}	-0.1609	-0.1918	γ_{42}	-0.3754	-0.3787
β_{24}	-0.5479	-0.5680	$\beta_{4,10}$	-0.0322	-0.0367	γ_{43}	0.5711	0.5668
β_{25}	2.6677	2.6509	$\beta_{4,11}$	0.6855	0.6940	γ_{44}	1.2579	1.2497
β_{31}	0.3960	0.3893	σ	0.5417	0.6804	γ_{45}	1.1664	1.1550
β_{32}	0.2544	0.2213	γ_1	-0.4261	-0.4078	γ_{46}	0.3691	0.3408
β_{41}	0.1636	0.1584	γ_{21}	0.0796	0.0626	γ_{47}	0.4247	0.4132
β_{42}	-0.2342	-0.2333	γ_{22}	0.4118	0.4172	γ_{48}	1.0606	1.0500
β_{43}	0.5705	0.5477	γ_{23}	0.4552	0.4091	γ_{49}	0.4266	0.4061
β_{44}	-0.2271	-0.2445	γ_{24}	0.6156	0.6188	$\gamma_{4,10}$	0.8028	0.8074
β_{45}	-0.1457	-0.1866	γ_{25}	-3.3269	-3.3181	$\gamma_{4,11}$	-0.3339	-0.3260

下面考虑广义 Cook 距离，Q 距离分别在删除一个数据和一组数据情形下的影响诊断，以及局部影响中组内加权扰动和组间加权扰动下的影响诊断．通过对第 4 、第 5 节相关统计量的计算，得到相应的诊断结果，分别列于散点图 1-3. 由图 1-3(左) 可知，第 43 株植物影响最大，第 15 株植物也有较大影响．而由图 1-3(右) 可知，第 497 号数据点是强影响点，第 503 号点影响也较大．注意，第 497 号数据点和第 503 号数据点都在上面提到的影响最大的第 43 组数据中，这显然是合理的．另外，通过计算发现试验 5 中的植物 (编号为 $i = 1, 2, 3$ 和 $43, 44, 45$) 上昆虫死亡比例都普遍较小，而唯有第 43 株植物上昆虫死亡比例较大，为 0.5315, 并且其它试验中植物上的死亡比例都较大．同时，在第 43 株植物这一组数据里，第 497 号数据点中， 12 只昆虫死亡了 11 只，是最多的．这些都与上面所得到的强影响点以及影响大的植物是一致的．

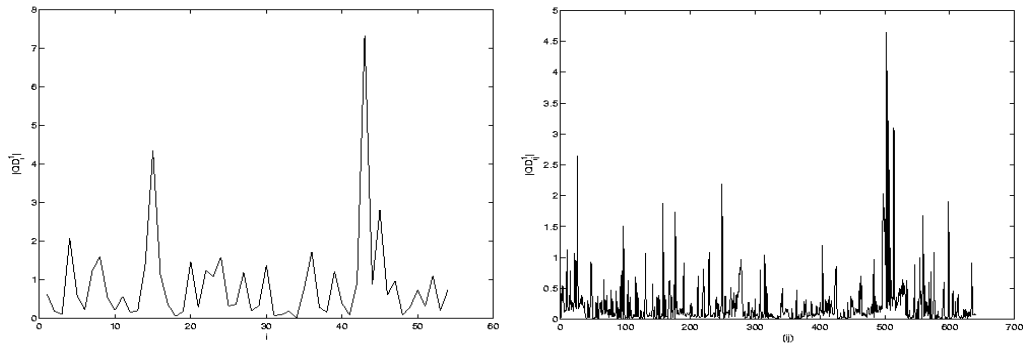

图 1　删除一组 (左)/ 一个 (右) 数据时 $|\mathrm{QD}^1|$ 的散点图

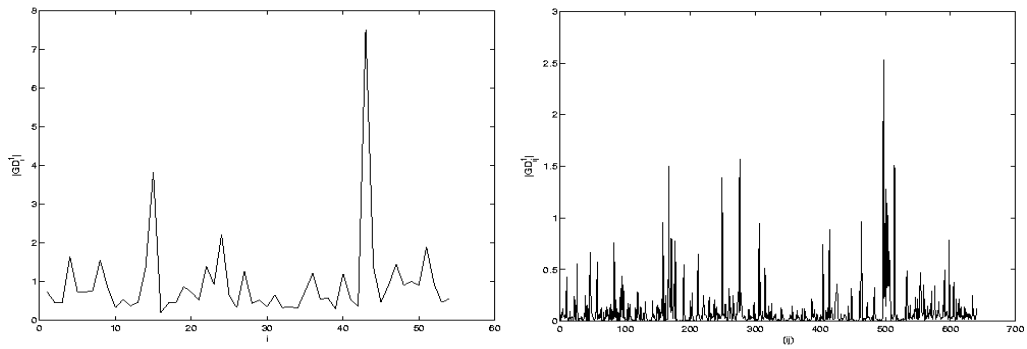

图 2　删除一组 (左)/ 一个 (右) 数据时 $|\mathrm{GD}^1|$ 的散点图

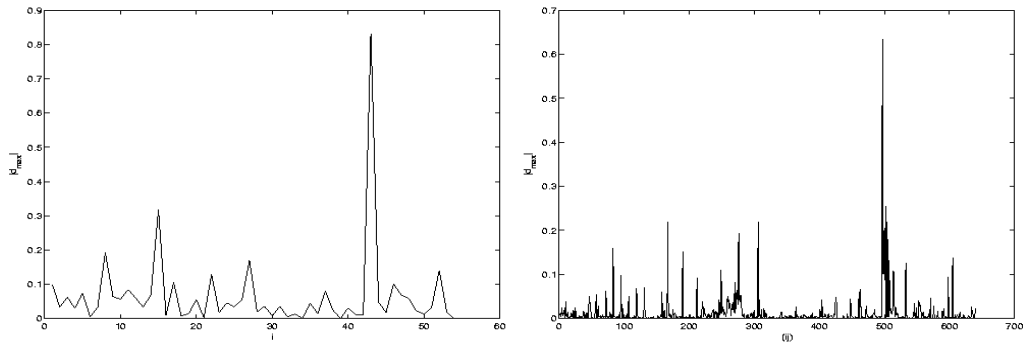

图 3　组间 (左)/ 组内 (右) 加权扰动模型中 $|d_{\max}|$ 的散点图

§6. 讨　　论

本文基于 ZI 随机效应模型研究了含 0 较多的纵向计数数据中强影响点的识别问题，本文的推导并未对 (1) 式的基本分布 F_1 加以太多限制，而是结合一般的指数族分布进行讨论．因此，本文的主要结果不仅适合于离散情形，也适合于连续情形．对于离散情形，不仅包括典型的 ZIP, ZIB 随机效应模型，也包括适用于偏大离差情形的 ZINB (ZI-negative binomial) 随机效应模型．在这些模型中，随机效应可以同时出现在非退化部分以及混合概率部分，而本文仅考虑了非退化部分出现随机效应的情形，对于两部分都有随机效应的情形，也可应用本文的方法进行研究．另外，由于随机效应的引入，使得模型复杂化，若直接应用 Cook 等作者传统的影响分析方法则有很大困难；为此，本文将模型中的随机效应看作缺失数据，利用 EM 算法及相应的 Q 函数进行影响分析，克服了积分上的困难．在算法的 E-step 过程中，涉及到了关于随机效应的积分，而这一般都得不到解析解，本文借助于 Laplace 一阶逼近得到此类积分的一个形式简单的表示，且计算量相对于其它的方法要少得多．一阶逼近的误差阶数为 $O(t_i^{-1})$, 为了获得更好的逼近效果，还可以利用 Taylor 二阶展开获得 Laplace 二阶逼近；同时，也可以应用 MCEM 算法或 MCMC 算法，以及高斯求积 (Gaussian quadrature) 方法来逼近此类积分，但是，这些方法的计算量都较大，特别是 MCMC 算法，其计算量更大，并且还涉及到收敛的判断问题．但是对影响分析来说，一阶近似通常已经能够满足识别强影响点的需要．

后记　陈希孺院士生前曾多次访问南京和东南大学，对发展南京地区的数理统计事业发挥了重要的促进作用．值此陈老师逝世一周年之际，仅以此文遥寄我们对他的怀念与哀思．

参 考 文 献

[1] Gupta, P., Gupta, R. and Tripathi, R., Analysis of zero-adjusted count data, *Computational Statistics and Data Analysis*, **23**(1996), 207–218.

[2] Cohen, A., Estimation of the Poisson parameter from truncated samples and from censored samples, *Journal of the American Statistical Association*, **49**(1954), 158–168.

[3] Mullahy, J., Specification and testing of some modified count data models, *Journal of Econometrics*, **33**(1986), 341–365.

[4] Heilbron, D., Zero-altered and other regression models for count data with added zeros, *Biometrics Journal*, **36**(1994), 531–547.

[5] Lambert, D., Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**(1992), 1–14.

[6] Hall, D.B., Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics*, **56**(2000), 1030–39.

[7] Vieira, A.M.C., Hinde, J.P. and Demetrio, C.G.B., Zero-inflated proportion data models applied to a biological control assay, *Journal of Applied Statistics*, **27**(2000), 373–389.

[8] Ridout, M., Hinde, J. and Demetrio, C.G.B., A score test for testing a zero-inflated Poisson regression model against zero-inflated negative alternatives, *Biometrics*, **57**(2001), 219–23.

[9] Breslow, N.E., Extra Poisson variation in log-linear models, *Applied Statistics*, **33**(1984), 38–44.

[10] Thall, P.F., Mixed Poisson likelihood regression models for longitudinal interval count data, *Biometrics*, **44**(1992), 197–209.

[11] Siddiqui, O., Modeling clustered count and survival data with an application to a school-based smoking prevention study, PhD Dissertation, University of Illinois at Chicago, 1996.

[12] Yau, K.K.W. and Lee, A.H., Zero-inflated Poisson regression with random effects to evaluate an occupations injury prevention programme, *Statistics in Medicine*, **20**(2001), 2907–20.
[13] Hall, D.B. and Berenhaut, K.S., Score tests for heterogeneity and overdispersion in zero-inflated poisson and binomial regression models, *The Canadian Journal of Statistics*, **30(3)**(2002), 1–16.
[14] Wang, L.H., Parameter estimation for mixtures of generalized linear mixed-effects models, A dissertation submitted to the Graduate Faculty of the University of Georgia in partial fulfillment of the requirements for the Degree Doctor of Philosophy, Athens, Georgia, 2004.
[15] Cook, R.D., Detection of influential observations in linear regression, *Technometrics*, **19**(1977), 15–8.
[16] Cook, R.D., Assessment of local influence, *J. R. Statist. Soc. B*, **48**(1986), 133–169.
[17] Zhu, H.T., Lee, S.Y., Wei, B.C., Zhu, J., Case-deletion measures for models with incomplete data, *Biometrika*, **88**(2001), 727–737.
[18] Zhu, H.T. and Lee, S.Y., Local influence for incomplete-data models, *J. R. Statist. Soc. B*, **63**(2001), Part 1, 111–126.
[19] Dempster, A.P., Laird, N.M. and Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Statist. Soc. B*, **39**(1977), 1–38.
[20] Evans, M. and Swartz, T., *Approximating Integrals Via Monte Carlo and Deterministic Methods*, Oxford University Press, Oxford, 2000.
[21] Cortinas Abrahantes, J. and Burzykowski, T., A version of the EM algorithm for proportional hazard model with random effects, Technical Report 0455, Interuniversity Attraction Pole, 2004.
[22] Wu, C.F.J., On the convergence properties of the EM algorithm, *Ann. Statist.*, **11**(1983), 95–103.
[23] Cook, R.D., Weisberg, S., *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
[24] Zhu, H.T. and Lee, S.Y., Local influence for generalized linear mixed models, *The Canadian Journal of Statistics*, **31(3)**(2003), 293–309.
[25] 解锋昌，韦博成，多元 t 分布数据的局部影响分析， 应用概率统计, **22(2)**(2006), 173–183.
[26] Hall, D.B., and Zhang, Z.G., Marginal models for zero inflated clustered data, *Statistical Modelling*, **4**(2004), 161–180.

Influence Analysis in ZI Longitudinal Count Data Models

Wei Bocheng
(*Department of Mathematics, Southest University, Nanjing, 210096*)
Xie Fengchang
(*Department of Mathematics, Nanjing Agricultural University, Nanjing, 210095*)

Based on the EM algorithm and Laplace approximation, this paper presents a method of influence analysis for zero inflated longitudinal count data models. To detect the influential observations in clustered count data with excess zeros, we regard the random effects as the missing data and put certain weight to the data with zero values in ZI longitudinal data models. According to this fact, we develop the influence method for the model based on the conditional expectation of the complete-data log-likelihood function and the associated Q-distance function under the EM algorithm. The Laplace approximation is also employed for integral computing in E-step. Then the case-deletion model and the local influence analysis are investigated for the model and several diagnostic measures are obtained. Finally, a numerical example of the real count data is given to illustrate the results in this paper.