应用概率统计　第二十五卷
第六期　2009年12月

Chinese Journal of Applied Probability
and Statistics Vol.25 No.6 Dec. 2009

综合报告

ZI数据的统计分析综述 *

解锋昌[1,2]　　韦博成[1]　　林金官[1]

([1]东南大学数学系, 南京, 210096;　[2]南京农业大学数学系, 南京, 210095)

摘　要

ZI (zero-inflated)数据就是含零过多的数据. 从上世纪90年代以来, ZI数据在各个研究领域受到越来越广泛的重视, 现在仍然是数据分析的热点问题之一. 本文首先通过2个实例说明ZI数据的实际意义, 然后介绍ZI数据分析的研究概况和最新进展. 另外文章还系统介绍了各种ZI数据模型、ZI纵向数据模型及其参数估计方法, 同时也介绍了ZI数据的统计诊断等问题, 其中包括作者近年来的一些工作. 最后, 本文列出了若干有待进一步研究的问题.

关键词:　ZI数据, EM算法, 随机效应模型, 统计诊断, score检验, 离差参数, 纵向数据.

学科分类号:　O212.1.

§1.　问题的提出

近年来, 含零过多的数据越来越受到人们的关注, 它们常常出现于公共卫生、生物医学、经济和农业等众多的领域中. 我们不妨称这些数据为ZI (zero-inflated)数据. 例如, 在调查人们一天中吸烟数量时, 可能会出现很多吸烟数量为0的人. 这里有两种情况: 一种是本来吸烟的人碰巧调查时未吸烟; 另一种是本来就不吸烟, 那么他们的吸烟数量当然为0. Gupta et al. (1996)曾经指出, 当观测到额外的取值为0的计数数据时, 如果我们仍用普通的Poisson模型进行拟合, 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. 下面来看两个经典的ZI数据实例.

例 1　HIV数据(Broek, 1995). 该数据涉及98位HIV疾病感染者, 关于每个病人尿道感染次数的情况, 见图1. 从图中可以看出感染0次的人特别多, 约占82.6%. 由于是离散型数据, 通常可用Poisson分布进行拟合, 其拟合结果见图1左边. 但是由图可知, 其拟合时效果很不理想. 该图显示, 拟合预测感染0次的期望频数与实际观测频数有较大差距, 而且对于感染1次和2次的期望频数与观测频数也有很大差距. 因此说明, 应用普通Poisson分布拟合HIV数据效果不好. 所以Broek (1995)建议用ZIP模型(见下一节)拟合HIV数据, 其结果列于图1右边. 由该图可以看出, 经ZIP模型拟合, 由此获得的期望频数与实际观测频数

*国家自然科学基金项目(10671032)资助.
本文2007年5月21日收到, 2008年5月7日收到修改稿.

都相当接近, 特别是对于感染0次的情形. 这表明, 用ZIP模型拟合HIV数据效果得到显著改进. 另外, 两个模型的拟合效果也可以通过Pearson拟合优度统计量χ^2进一步得到说明. 当HIV数据用普通Poisson分布拟合时, 其$\chi^2 = 16.135$, 相应的p值为0.0003, 表明拟合不好; 而用ZIP分布拟合时, $\chi^2 = 1.3723$, 相应的p值为0.2414, 表明拟合优度得到显著改进.

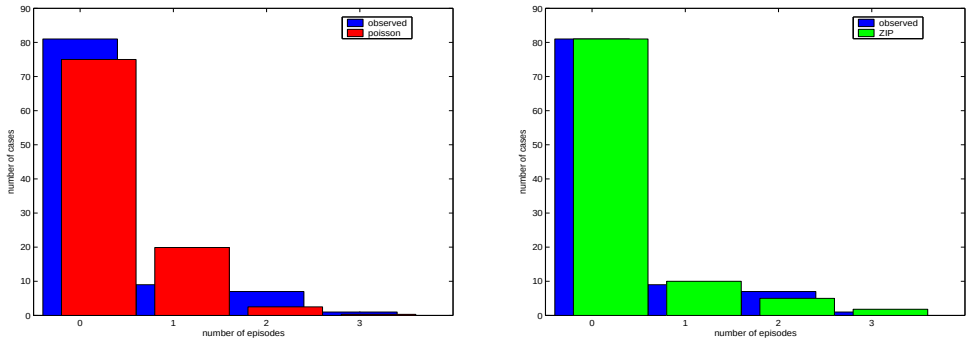

图1 尿道感染的观测频数以及poisson和ZIP模型预测的期望频数

例 2 Accident数据(Greenwood and Yule, 1920; Bohning, 1998). 该数据是关于军工厂中647位女性工人发生事故的次数, 见图2. 其中发生0次事故的人约占70%. 与例1类似, 该数据若用Poisson分布拟合(见图2左边), 其发生0次和1次事故的期望频数与实际观测频数差距较大; 而用ZIP分布拟合时(见图2右边), 其差距明显变小. 另外, 用Poisson分布拟合时, 其Pearson拟合优度统计量$\chi^2 = 115.35$, 相应的p值为< 0.00001, 表明拟合不好; 若用ZIP分布拟合, 则$\chi^2 = 7.838$, 相应的p值为0.0495, 表明拟合优度到显著改进.

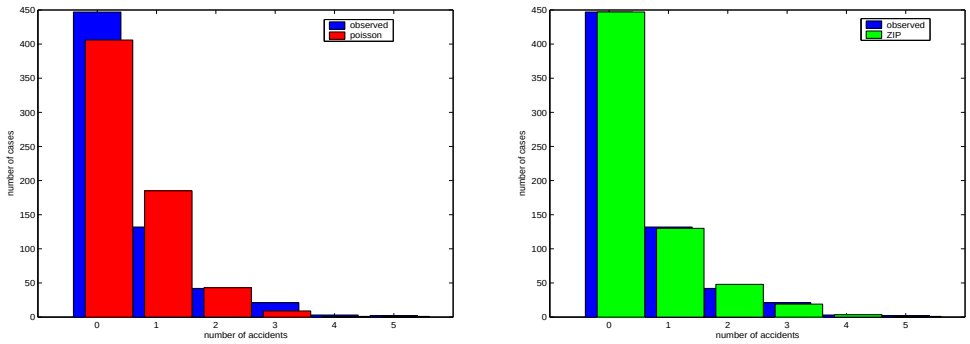

图2 事故的观测频数以及poisson和ZIP模型预测的期望频数

§2. ZI模型及其参数估计和统计诊断

对于含0过多的数据, 很多作者一直在研究和发展相关的统计模型和参数估计方法. Cohen (1954)首次认真考虑了处理含0过多的计数数据, 并提出了调整的Poisson模型; Sigh (1963)和Johnson and Kotz (1969)描述了ZIP分布, 但未考虑协变量; King (1989)和Mullahy (1986)提出了Hurdle Poisson回归模型; Heilbron (1989, 1994)提出了Zero-altered Poisson回归模型, 他们都考虑了协变量. 基于这些工作, Lambert (1992)考虑了Zero-Inflated Poisson

(ZIP)回归模型, 该模型假设取值为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, 组成混合分布. 并且在取值为0的部分和取值为Poisson分布的部分都可以引入协变量, 从而构成ZIP回归模型. 由于这一模型结构上比较合理, 处理上比较方便, 得到理论和应用工作者的广泛认可, 从而成为当前最常用的处理含0过多的计数数据模型. 本文下面也着重介绍近年来发展起来的各种ZI模型.

2.1 ZI模型

Lambert (1992)考虑了Zero-Inflated Poisson (ZIP)混合分布

$$p(y;\phi,\lambda)=\begin{cases}\phi+(1-\phi)\exp(-\lambda), & y=0;\\ (1-\phi)\dfrac{\lambda^y}{y!}\exp(-\lambda), & y>0,\end{cases} \tag{2.1}$$

其中参数ϕ (称为ZI参数)表示取值为0 (也称结构上为0)的非Poisson数据占有的比例. 该模型可以看作为取值为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的混合分布. 同时, 为了考虑ZI数据中因变量和自变量之间的关系, Lambert进一步在ZI参数部分和Poisson部分分别引入协变量, 从而得到ZIP回归模型:

$$\begin{cases}\text{logit}(\phi)=Z^T\gamma;\\ \log(\lambda)=X^T\beta.\end{cases} \tag{2.2}$$

易见, 上述模型的两部分都有明确的解析表达式, 处理上比较方便, 在概率统计上的意义也很清楚, 从而成为当前最常用的处理含0过多数据的统计模型 (Bohning, 1998; Dalrymple et al., 2003; Lee et al., 2001; Bohning et al., 1999; Dietz and Bohning, 2000; Xie et al., 2001; Welsh et al., 1996; Shankar et al., 1997; Street et al., 1999; Cheung, 2002; Ridout et al., 1998).

Lambert的ZIP模型很自然地可以推广到其它离散分布, 近年来, 许多作者考虑了基于其它各种离散分布的ZI模型. 例如, 若在ZIP模型中, (2.1)式的Poisson分布改为二项分布, 即可得ZIB模型(ZI-binomial; Hall, 2000; Vieira et al., 2000).

若离散分布部分为以下负二项分布

$$f_1(y)=\frac{\Gamma(y+\lambda^{1-c}/\tau)}{y!\Gamma(\lambda^{1-c}/\tau)}(1+\tau\lambda^c)^{-\lambda^{1-c}/\tau}(1+\lambda^{-c}/\tau)^{-y},$$

其中负二项分布的期望为λ, 方差为$\lambda(1+\tau\lambda^c)$, 则可得ZINB (ZI-negative binomial)模型 (Ridout et al., 2001; Fahrmeir and Echavarria, 2006).

若离散分布部分为以下广义Poisson分布

$$f_2(y)=\frac{1}{y!}(\frac{\mu}{1+\alpha\mu})^y(1+\alpha y)^{y-1}\exp\Big\{-\frac{\mu(1+\alpha y)}{1+\alpha\mu}\Big\},$$

其中广义Poisson分布的期望为μ, 方差为$\mu(1+\alpha\mu)^2$, 则可得ZIGP (ZI-generalized Poisson) (Gupta et al., 2004; Famoye and Singh, 2006)模型.

若离散分布部分为以下幂级数分布

$$f_3(y)=\frac{b(y)\theta^y}{\sum\limits_{y=0}^{\infty}b(y)\theta^y},$$

则可得ZIPS (ZI-power series)模型(Ghosh et al., 2006).

注意, 对于不同的离散分布, 对应于(2.2)式的回归部分可能会有相应的变化. 由于参数ϕ对应于0-1分布, 因而(2.2)式第一项通常不变, 即为Logistic回归; 而(2.2)式第二项的回归形式要取决于相应离散分布期望参数的形式以及实际问题的需要.

以上我们介绍了含0过多的离散数据的ZI模型, 基于(2.1)、(2.2)两式的ZI模型亦可推广到含0过多的连续型数据. 事实上, (2.1)式表示混合分布, (2.2)式表示回归, 这对连续型数据也可以有类似的结构. 比较常见的, Deng and Paul (2000)考虑了ZI广义线性模型(ZIGLM):

$$p(y;\phi,\lambda)=\begin{cases}\phi+(1-\phi)f_4(0,\theta), & y=0;\\ (1-\phi)f_4(y,\theta), & y>0,\end{cases}\tag{2.3}$$

$$\begin{cases}\text{logit}(\phi)=Z^T\gamma;\\ g(\mu)=X^T\beta,\end{cases}\tag{2.4}$$

其中$f_4(y,\theta)$通常取为指数族分布

$$f_4(y,\theta)=\exp\{a(\theta)y-g(\theta)+c(y)\},$$

$g(\mu)$为联系函数, μ为期望参数.

Deng and Paul (2005)还进一步把以上ZI广义线性模型推广到ZI偏大离差广义线性模型, 这时(2.3)式$f_4(y,\theta)$改为偏大离差指数族分布(Cox, 1983; Chesher, 1984; Dean, 1992)

$$f_5(y,\theta,\alpha)=f_4(y,\theta^*)\Big\{1+\sum_{r=2}^{\infty}\frac{b_r}{r!}D_r(y)\Big\},$$

其中假定对于给定的θ^*, 有$f_4(y,\theta^*)=\exp\{a(\theta^*)y-g(\theta^*)+c(y)\}$, 而且$\theta^*$是连续独立的随机变量, 并有$\mathsf{E}(\theta^*)=\theta(x,\beta)$, $\mathsf{Var}\,(\theta^*)=\alpha b(\theta)>0$, $b_r=\mathsf{E}(\theta^*-\theta)^r$, $D_r(y)=\{(\partial^{(r)}/\partial\theta^{*(r)})\cdot f_4(y;\theta^*)|_{\theta^*=\theta}\}/f_4(y;\theta)$, 其中$\beta$为回归参数, α为偏大离差参数.

可能由于实际问题中经常出现含0过多的离散数据, 而含0过多的连续型数据相对较少, 因而后者的研究还很不充分, 这也给有兴趣的读者留下更多的发挥空间.

2.2 参数估计

对于ZI模型, Newton-Raphson方法或Fisher scoring方法是常用的参数估计方法, 但是对于参数维数较高的情形, 为确保收敛, 也经常可以应用EM算法进行估计(Lambert, 1992;

Hall, 2000; Lee et al., 2001), 以下以ZIP回归模型为例予以说明. 假设$Y_1,\cdots,Y_n$为一组观测数据, 根据(2.1)、(2.2)两式, Y_i满足以下ZIP回归模型

$$\begin{aligned}&\mathsf{P}(Y_i=0)=\phi_i+(1-\phi_i)\exp(-\lambda_i),\\&\mathsf{P}(Y_i=y_i)=(1-\phi_i)\frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i),\qquad y_i=1,2,3,\cdots,\end{aligned}$$

其中参数回归为$\log(\lambda_i)=X_i^T\beta$, $\mathrm{logit}(\phi_i)=Z_i^T\gamma$, X_i, Z_i为协变量, $i=1,\cdots,n$. 为了应用EM算法, 今设$u_i=1$, 如果Y_i来自于退化分布; $u_i=0$, 如果Y_i来自于非退化分布, $i=1,\cdots,n$. 则可将$u=(u_1,\cdots,u_n)^T$看作缺失数据(missing data), 记为Y_m; 而Y_i, X_i, Z_i, $i=1,\cdots,n$可视为观测数据, 并记为Y_o; 所以完全数据为$Y_c=(Y_o,Y_m)$. 因此, 基于完全数据的对数似然函数为

$$\begin{aligned}L_c(\theta|Y_c)\;=\;&\sum_{i=1}^{n}[u_i\log\phi_i(\gamma)+(1-u_i)\log(1-\phi_i(\gamma))]\\&+\sum_{i=1}^{n}(1-u_i)[-\lambda_i+y_i\log\lambda_i-\log y_i!],\end{aligned}$$

其中参数为$\theta=(\beta^T,\gamma^T)^T$, $\phi_i(\gamma)=\exp(Z_i^T\gamma)/[1+\exp(Z_i^T\gamma)]$. 由$L_c(\theta|Y_c)$即可根据EM算法(Dempster et al., 1977)得到参数θ的极大似然估计$\widehat{\theta}$, 细节可见文献(Wu, 1983).

另外, Ghosh et al. (2006)利用Bayes方法给出了ZIPS模型(ZIP是其特例)的点估计与区间估计. 并通过模拟研究发现, 在样本量为50时, 若$\mathsf{P}(Y=0)$不接近于1, 则经典估计方法较好; 反之, 则Bayes估计较好. 另外, Angers and Biswas (2003)也应用Bayes方法分析了ZIGP模型.

2.3　统计诊断

统计诊断的任务就是要检测已知数据是否符合既定模型的假设条件, 其典型问题就是检测数据中的强影响点或异常点(韦博成等, 1991). 关于这方面的问题, ZI模型与非ZI模型没有太多差别, 我们可以基于数据删除模型(Cook, 1977)和局部影响分析方法(Cook, 1986)对ZI回归模型进行影响分析. 李爱萍等(2007)在EM框架下研究了ZIP模型的基于上述两种方法的影响分析问题. 近来, Lee et al. (2004)还基于局部影响分析方法研究了ZIP和ZINB模型中score检验统计量(Broek, 1995; Ridout, 2001)在加权扰动、自变量的加性扰动和乘性扰动下的影响点识别问题.

对于ZI模型, 其更重要的统计诊断是ZI模型拟合中的假设检验问题. 首先, 一个很自然, 也是很基本的问题是: 给定的数据是否需要用ZI模型来进行拟合? 这个问题等价于在(2.1)式或(2.3)式中检验

$$H_0:\phi=0;\qquad H_1:\phi\neq 0.\tag{2.5}$$

这可称为ZI模型的存在性检验. Broek (1995)首先研究了ZIP模型, 得到了检验(2.5)的score统计量; Deng and Paul (2000)对于一般的ZI广义线性模型得到了(2.5)的score检验统计量, 并作为特例, 把他们的结果应用于ZIP和ZIB模型, 得到了(2.5)的score统计量, 并与Broek

(1995)的结果一致. 同时, 他们进行了随机模拟, 并且发现: ZIP(μ)中的μ以及ZIB(10, p)中的p较小(如$p = 0.01$)时, 功效增加很慢; 而相应的值较大时(如$\mu = 2$, $p = 0.2$), 则功效增加很快. 另外Lee et al. (2001)和Jansakul and Hinde (2002)也更深入地考虑了ZIP模型中(2.5)的检验问题. Gupta et al. (2004)考虑了ZIGP模型中(2.5)的检验问题, 得到了相应的score检验统计量.

在实际问题中, 计数数据常常出现偏大离差或偏小离差(over/under dispersion)的情形, ZI模型亦有类似的问题. Ridout et al. (2001)曾经指出, 在ZIP模型中, 若非退化部分有比较严重的偏大离差, 则其参数估计不相合, 从而ZIGP、ZINB, 或其他ZI偏大或偏小离差模型可能比普通的ZIP模型更适合. 因而我们常常需要检验: H_0: 数据用ZIP模型拟合; H_1: 数据用某一个偏大离差或偏小离差ZI模型拟合. 这类检验通常可化为参数检验问题, 例如, 若上述对立假设为ZIGP模型, 则等价于检验广义Poisson分布中的离差参数α是否为0. Ridout et al. (2001)研究了ZINB模型中离差参数的显著性检验(相当于偏大离差检验). 但他们得到的score统计量的抽样分布收敛太慢, 小样本情况下检验的结果可能误导. Jung et al. (2005)进一步给出了Bootstrap score检验统计量. Gupta et al. (2004)研究了ZIGP的离差参数的存在性检验, 得到了score统计量. 对于一般的ZI广义线性模型, Deng and Paul (2005)研究了偏大离差检验、ZI存在性检验以及ZI与偏大离差的同时检验; 得到了相应的score检验统计量.

值得一提的是, ZINB和ZIGP模型中离差参数的存在性检验实际是一种模型选择问题, 当零假设ZIP模型被拒绝时, 对立假设ZINB或ZIGP是否是最佳选择呢? 为此, Thas and Rayner (2005)给出了更一般的可供选择的模型

$$g_k(y;\phi,\lambda,\theta) = C(\theta,\phi,\lambda)\exp\Big\{\sum_{j=1}^{k}\theta_j h_j(y;\phi,\lambda)\Big\}p(y;\phi,\lambda),$$

其中$p(y;\phi,\lambda)$如(2.1)式所示, 代表ZIP模型; $\theta^T = (\theta_1,\cdots,\theta_k) \in \mathbf{R}^k$, $C(\cdot)$是正则化常数, $\{h_j\}$是关于函数$p(y;\phi,\lambda)$正交的函数集, 且满足

$$\sum_{y=0}^{\infty} h_r(y;\phi,\lambda)h_s(y;\phi,\lambda)p(y;\phi,\lambda) = \begin{cases} 1, & r = s; \\ 0, & r \neq s, \end{cases}$$

其中$r, s = 0, 1, \cdots$. Thas and Rayner给出了ZIP模型关于函数$g_k(y;\phi,\lambda,\theta)$的光滑检验(smooth test), 即检验$H_0 : \theta_1 = \cdots = \theta_k = 0$.

关于ZI模型, 变离差检验也是很重要的问题, 在广义回归模型中, 变离差检验主要有两种方法, 即离差参数的参数化方法和随机效应法(Lee and Nelder, 2000). Xie et al. (2009a)应用离差参数的参数化方法研究了ZIGP模型中离差参数α的变离差检验, 同时还考虑了ZI参数ϕ的齐性检验问题. 这时我们假定: $\phi_i = \phi m_1(z_{1i},\gamma_1)$, $\alpha_i = \alpha m_2(z_{2i},\gamma_2)$, 且存在唯一的$\gamma_1^0, \gamma_2^0$使$m_1(z_{1i},\gamma_1^0) = 1$, $m_2(z_{2i},\gamma_2^0) = 1$. 该文研究了以下三种检验并且得到了相应的score检验统计量. $H_{10} : \gamma_1 = \gamma_1^0$; $H_{20} : \gamma_2 = \gamma_2^0$; $H_{30} : \gamma_1 = \gamma_1^0$, $\gamma_2 = \gamma_2^0$.

§3. ZI纵向数据模型的参数估计和统计诊断

近年来, 纵向数据受到各方面的广泛关注, 相应的数据常常是经过重复测量得到的. 这时, 组内与组间相比, 组内常是相关的, 为了正确评价响应变量与协变量之间的关系, 必须考虑组内的相依性, 否则就可能导致错误的结论(Breslow, 1984). 为此, 人们常常选择随机效应模型. 而在这类数据中也常会出现含0过多的情形, 下面的例子就是如此.

例 3 粉虱数据(Hall and Zhang, 2004). 在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(Hall and Zhang, 2004). 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的区组进行6种不同的试验. 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, 其中存活0只昆虫的情形大约占53%, 见下面图3.

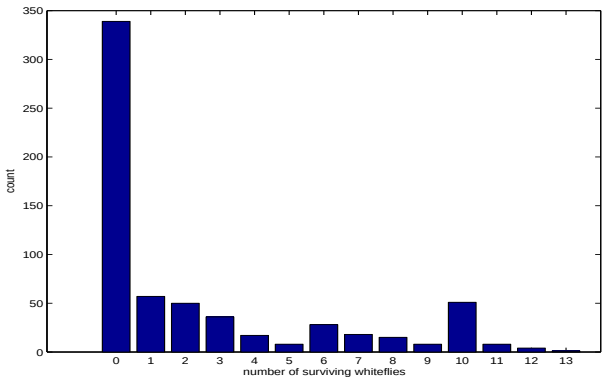

图3 存活昆虫数的观测频数

3.1 ZI混合模型

为了能够拟合这种含0过多的纵向数据, 很多作者提出各种随机效应模型. Hall (2000); Yau and Lee (2001); Wang et al. (2002); Hur et al. (2002)和Xiang et al. (2006, 2007)等都曾研究了下面的ZIP混合效应模型(ZIPM).

今假定第i类中第j个响应Y_{ij}具有下面ZIP分布

$$\begin{aligned}&\mathsf{P}(Y_{ij}=0)=\phi_{ij}+(1-\phi_{ij})\exp(-\lambda_{ij}),\\&\mathsf{P}(Y_{ij}=y_{ij})=(1-\phi_{ij})\frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!}\exp(-\lambda_{ij}),\qquad y_{ij}=1,2,3,\cdots.\end{aligned}$$

且假定

$$\begin{aligned}\mathrm{logit}(\phi_{ij})&=Z_{ij}^T\gamma+u_i,\\\log(\lambda_{ij})&=X_{ij}^T\beta+v_i,\end{aligned}$$

其中u_i和v_i分别表示以上回归中的第i个随机效应.

以上ZIPM模型中的Poisson分布也可以是其他离散型或连续型分布, 则可得到各种ZI混合模型; 诸如: ZIBM (ZIB混合模型) (Hall, 2000); ZINBM (ZINB混合模型) (Xiang et al., 2007; Yau et al., 2003); ZIGPM (ZIGP混合模型) (Xie et al., 2008, 2009b); 基于指数族的ZI混合模型(Wang, 2004; 韦博成, 解锋昌, 2006); ZI-lognormal混合模型(Berk and Lachenbruch, 2002); ZI-正态变换(ZI transformed normal)回归模型(Olsen and Schafer, 2001)等等. 另外, Lee et al. (2006)还给出了多层ZIP混合模型(Multi-level ZIPM), 他们假定Y_{ijk}表示第i类中第j个个体在第k次观测时的响应, Y_{ijk}具有ZIP分布, 且有

$$\begin{aligned}\text{logit}(\phi_{ijk}) &= Z_{ijk}^T\gamma + u_i + s_{ij},\\ \log(\lambda_{ijk}) &= X_{ijk}^T\beta + v_i + w_{ij},\end{aligned}$$

其中u_i和v_i表示第i类效应, s_{ij}和w_{ij}表示第i类中第j个个体的效应.

当观测数据与时间有关时, 考虑数据的相关性是必要的, 这时可将上面模型中s_{ij}和w_{ij}改为s_{ijk}和w_{ijk}, 以便说明个体之间在不同时间测量时的相关性(Lee et al., 2006). 另外, Hall and Zhang (2004)和Dobbie and Welsh (2001)还利用边缘模型(marginal model)研究了ZI重复测量数据.

3.2 参数估计

在应用ZI混合模型拟合含0过多的纵向数据时, 由于随机效应的存在, 使得模型变得十分复杂. 特别, 由于观测数据的似然函数常常涉及高维积分, 由此很难得到精确的参数估计, 为了克服这些困难, 文献中出现了多种多样的方法. Hall and Berenhaut (2002)应用似然函数的Laplace展开进行参数估计, 但是, 若展开的阶数较低, 则精度往往不高, 若展开的阶数较高, 虽然精度提高了, 但涉及的计算工作相当复杂. EM算法是处随理机效应模型的有力工具, 对含0过多数据亦是如此. Hall (2000)采用了EM高斯求积法(EM-Gaussian quadrature)研究了ZIPM和ZIBM模型的参数估计; Wang (2004)采用MCEM方法研究了基于指数族的ZI混合模型的参数估计. 韦博成, 解锋昌(2006)利用EM算法研究了基于指数族的ZI混合模型的参数估计, 并在E步应用Laplace逼近, 以避免高维积分的困难.

另外, Xiang et al. (2007); Lee et al. (2006)和Xie et al. (2008, 2009b)分别采用了EM算法结合约束极大似然估计(EM-REML)方法研究了ZIPM, ZINBM, 多层ZIPM以及ZIGPM等模型的参数估计, 这一方法比较有效, 其具体过程如下:

(1) 给定随机成分的方差的初值;

(2) 利用EM算法估计模型参数, 并对随机效应进行预测, 直至收敛;

(3) 利用REML方法给出随机效应中方差的估计, 直至收敛;

(4) 重复(2)和(3)两步, 直到满足停止条件.

除此之外, Hall and Zhang (2004)还采用EM算法结合广义估计方程(GEE)的方法研究边缘ZI模型中的参数估计问题. 他们首先忽略数据间相关性, 然后在EM算法的M步中, 引入工作相关阵(working correlation matrix)来解释相关性. 他们通过随机模拟说明, EM-GEE方法要优于直接使用GEE方法. 但是, 他们没有说明如何选择恰当的工作相关阵, 以及工作相关阵的选择对参数估计影响等重要问题.

3.3 统计诊断

由于随机效应的存在, ZI混合模型的统计诊断也比较复杂. 关于强影响点或异常点的识别问题, 由于观测数据的似然函数常常涉及高维积分, 没有显式解, 因此无法应用通常的, 基于数据删除模型的方法(Cook, 1977)以及局部影响分析方法(Cook, 1986); 这时还是要借助于EM算法. Zhu and Lee (2001); Zhu et al. (2001)以及解锋昌, 韦博成(2006)提出了基于EM算法和Q函数(即完全数据对数似然的数学期望)的影响分析方法, 这种方法正好可以应用于ZI混合模型. 韦博成, 解锋昌(2006)基于EM-Laplace方法研究了ZI指数族随机效应模型的影响分析问题, 得到了基于数据删除模型和局部影响分析的诊断统计量; 同时他们还基于EM-REML方法研究了ZIGPM模型的影响分析问题(Xie et al., 2008).

对于含0过多的纵向数据, 也有很多模型拟合中的假设检验问题. 例如: ZI参数的存在性检验; 离差参数以及随机效应的显著性检验; 变离差检验; 随机效应的方差齐性等等. Lee et al. (2006)研究了多层ZIP混合模型中ZI参数的存在性检验, 得到了score统计量; Xiang et al. (2007)研究了ZINBM模型离差参数的显著性检验, 得到了score统计量; Xie et al. (2009b)研究了ZIGPM模型离差参数的显著性检验, 同时还研究了ZI部分与GP部分回归系数的显著性检验. Hall and Berenhaut (2002)利用Laplace展开研究了ZIPM和ZIBM模型中随机效应的显著性检验, 得到了score统计量.

§4. 进一步研究的问题

关于ZI模型, 还有很多问题有待进一步研究, 今列举一些如下:

(1) ZI模型中检验统计量的极限分布

在现有的文献中, 通常都假定似然函数满足一定的正则条件(Cox and Hinkley, 1974). 因而在样本容量充分大时, 检验的似然比统计量和score统计量都渐近地服从χ^2分布. 但是Hall and Preatgard (2001)认为, 在某些情形, 统计量的极限分布应该是混合χ^2分布. 关于ZI模型, 特别是ZI混合模型, 检验统计量的渐近分布是一个很值得进一步研究的问题.

(2) 非参数和半参数ZI模型的统计分析

Lam et al. (2006a, b)研究了半参数ZIP模型, 其混合分布仍然如(1)式所示; 回归部分为

$$\begin{cases} \text{logit}(\phi) = Z^T\gamma; \\ \log(\lambda) = X^T\beta + g(t), \end{cases}$$

其中t是连续的可观测的解释变量, $g(\cdot)$是未知的光滑函数.

他们在一定条件下研究了该模型的参数估计方法以及估计量的渐进性质. 还可以进一步研究非参数ZI模型、非参数和半参数ZI混合模型, 及其估计方法、渐进性质、统计诊断等一系列的问题.

(3) 其它问题

时间序列中也会出现ZI数据, Yau et al. (2004)利用ZIP混合自回归模型研究了职业卫生中一个时间序列数据; 这方面的研究工作还很少. 另外, 带有测量误差的ZI模型的研究工作也未见报道. 同时, ZI混合模型中与随机效应有关的估计、检验等问题也有待更深入的研究.

参 考 文 献

[1] Angers, J.F. and Biswas, A., A bayesian analysis of zero-inflated generalized Poisson model, *Computational Statistics and Data Analysis*, **42**(2003), 37–46.

[2] Berk, K.N. and Lachenbruch, P.A., Repeated measures with zeros, *Statistical Methods in Medical Research*, **11**(2002), 303–316.

[3] Bohning, D., Zero-inflated Poisson models and C.A.MAN: a tutorial collection of evidence, *Biometrical Journal*, **40**(1998), 833–843.

[4] Bohning, D., Dietz, E. and Schlattmann, P., The zero-inflation Poisson and the decayed, missing and filled teeth index in dental epidemiology, *J. Roy. Statist. Soc. Ser. A*, **162**(1999), 195–209.

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

[6] Broek, V.J., A score test for zero inflation in a Poisson distribution, *Biometrics*, **51**(1995), 738–743.

[7] Chesher, A., Testing of neglected heterogeneity, *Econometrika*, **52**(1984), 865–872.

[8] Cheung, Y.B., Zero-inflated models for regression analysis of count data: a study of growth and development, *Statistics in Medicine*, **21**(2002), 1461–1469.

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

[10] Cook, R.D., Detection of influential observations in linear regression, *Technometrics*, **19**(1977), 15–8.

[11] Cook, R.D., Assessment of local influence, *J. R. Statist. Soc. B*, **48**(1986), 133–169.

[12] Cox, D.R., Some remarks on overdispersion, *Biometrika*, **70**(1983), 269–274.

[13] Cox, D.R. and Hinkley, D.V., *Theoretical Statistics*, London, Chapman and Hall, 1974.

[14] Dalrymple, M.L., Hudson, I.L. and Ford, R.P.K., Finite mixture, zero-inflated Poisson and hudle models with application to SIDS, *Computational Statistics and Data Analysis*, **41**(2003), 491–504.

[15] Dean, C.B., Testing for overdispersion in Poisson and binomial regression models, *J. Amer. Statist. Assoc.*, **87**(1992), 451–457.

[16] 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.

[17] Deng, D. and Paul, S.R., Score tests for zero-inflation in generalized linear models, *Canad. J. Statist.*, **27**(2000), 563–570.

[18] Deng, D. and Paul, S.R., Score tests for zero-inflation and over-dispersion in generalized linear models, *Statistica Sinica*, **15**(2005), 257–276.

[19] Dietz, E. and Bohning, D., On estimation of the Poisson parameter in zero-modified Poisson models, *Computational Statistics and Data Analysis*, **34**(2000), 441–460.

[20] Dobbie, M.J. and Welsh, A.H., Modelling correlated zero-inflated count data, *Australian and New Zealand Journal of Statistics*, **43**(2001), 431–44.

[21] Fahrmeir, L. and Echavarria, L.O., Structured additive regression for overdispersed and zero-inflated count data, *Applied Stochastic Models in Business and Indutry*, **22**(2006), 351–369.

[22] Famoye, F. and Singh, K.P., Zero-inflated generalized Poisson model with an application to domestic violence data, *Journal of Data Science*, **4 (1)**(2006), 117–130.

[23] Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C., Bayesian analysis of zero-inflated regression models, *Journal Statistical Planning and Inference*, **136**(2006), 1360–1375.

[24] Greenwood, M. and Yule, G.U., An inquiry into the nature of frequency distributions of multiple happenings, etc., *Journal of the Royal Statistical Society*, **83**(1920), 255.

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

[26] Gupta, P.L., Gupta, R.C. and Tripathi, R.C., Score test for zero inflated generalized Poisson regression model, *Comm. Statist. Theory Methods*, **33 (1)**(2004), 47–64.

[27] Hall, D.B., Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics*, **56**(2000), 1030–1039.

[28] 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.

[29] Hall, D.B. and Preatgard, J.T., Order restricted score tests for homogeneity in generalized linear and nonlinear mixed models, *Biometrika*, **88**(2001), 739–751.

[30] Hall, D.B. and Zhang, Z.G., Marginal models for zero inflated clustered data, *Statistical Modelling*, **4**(2004), 161–180.

[31] Heilbron, D., Generalized linear models for altered zero probabilities and over dispersion in count data, Technical Report, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

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

[33] Hur, K., Hedeker, D., Henderson, W., Khuri, S. and Daley, J., Modeling clustered count data with excess zeros in health care outcomes research, *Health Services and Outcomes Research Methodology*, **3**(2002), 5–20.

[34] Jansakul, N. and Hinde, J.P., Score tests for zero-inflated Poisson models, *Coputational Statistics and Data Analysis*, **40**(2002), 75–96.

[35] Johnson, N. and Kotz, S., *Distributions in Statistics: Discrete Distributions*, Houghton Mifflin, Boston, 1969.

[36] Jung, B.C., Jhun, M. and Lee, J.W., Bootstrap tests for overdispersion in a zeros-inflated Poisson regression model, *Biometrics*, **61**(2005), 626–629.

[37] King, G., Event count models for international relations: Generalizations and applications, *International Studies Quarterly*, **33**(1989), 123–147.

[38] Lam, K.F., Xue, H. and Cheung, Y.B., Semiparametric analysis of zero-inflated count data, Research Report, Volume 424, Department of Statistics and Actuarial Science, The University of Hong Kong, 2006a.

[39] Lam, K.F., Xue, H.Q. and Cheung, Y.B., Semiparametric analysis of zero-inflated count data, *Biometrics*, **62**(2006b), 996–1003.

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

[41] Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W. and McLachlan, G.J., Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros, *Statistical Methods in Medical Research*, **15(1)**(2006), 47–61.

[42] Lee, A.H., Wang, K. and Yau, K.K.W., Analysis of zero-inflated Poisson data incorporating extent of exposure, *Biometrical Journal*, **43**(2001), 963–975.

[43] Lee, A.H., Xiang, L. and Fung, W.K., Sensitivity of score tests for zero-inflation in count data, *Statistics in Medicine*, **23**(2004), 2757–2769.

[44] Lee, Y. and Nelder, J.A., Two ways of modeling overdisperssion in non-normal data, *Applied Statistics*, **49**(2000), 591–598.

[45] 李爱萍, 谷政, 解锋昌, ZIP回归模型的数据删除度量和广义杠杆, 南京林业大学学报, **31(6)**(2007), 109–112.

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

[47] Olsen, M.K. and Shafer, J.L., A two-part random-effects model for semicontinuous longitudinal data, *Journal of the American Statistical Association*, **96**(2001), 730–745.

[48] Ridout, M, Demetrio, C.G.B. and Hinde, J., Models for count data with many zeros, Invited paper, The XIXth International Biometric Conference, Cape Town, South Africa, (1998), 179–192.

[49] 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), 219–223.

[50] Shankar, V., Milton, J. and Mannereing, F., Modeling accident frequencies as zero-altered probability processes: An empirical inquiry, *Accident Analysis and Prevention*, **29**(1997), 829–837.

[51] Singh, S., A note on inflated Poisson distribution, *Journal of the Indian Statistical Association*, **1**(1963), 140–144.

[52] Street, A., Jones, A. and Furuta, A., Cost-sharing and pharmaceutical utilisation and expenditure in Russia, *Journal of Health Economics*, **18**(1999), 459–472.

[53] Thas, O. and Rayner, J.C.W., Smooth tests for the zero-inflated Poisson distribution, *Biometrics*, **61**(2005), 808–815.

[54] 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.

[55] Wang, K., Yau, K.K.W. and Lee, A.H., A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays, Computer Methods and Programs in Biomedicine, **68**(2002), 195–203.

[56] Wang, L., Parameter Estimation for Mixtures of Generalized Linear Mixed-effects Models, A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Ful-llment of the Requirements for the Degree Doctor of Philosophy, Athens, Georgia, 2004.

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

[58] 韦博成, 解锋昌, ZI纵向计数数据模型的影响分析, 应用概率统计, **22(3)**(2006), 252–262.

[59] Welsh, A.H., Cunningham, R.B., Donnelly, C.F. and Lindenmayer, D.B., Modelling the abundance of rare species: Statistical models for counts with extra zeros, *Ecological Modelling*, **88**(1996), 297–308.

[60] Wu, C.F.J., On the convergence properties of the EM algorithm, *Ann. Statist.*, **11**(1983), 95–103.

[61] Xiang, L., Lee, A.H., Yau, K.K.W. and Mclachlan, G.L., A score test for zero-inflation in correlated count data, *Statistics in Medicine*, **25**(2006), 1660–1671.

[62] Xiang, L., Lee, A.H., Yau, K.K.W. and Mclachlan, G.L., A score test for overdispersion in zero-inflated Poisson mixed regression model, *Statistics in Medicine*, **26**(2007), 1608–1622.

[63] Xie, M., He, B. and Goh, H., Zero-inflated Poisson model in statistical process control, *Computational Statistics and Data Analysis*, **38**(2001), 191–201.

[64] 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.

[65] Xie, F.C. Lin, J.G. and Wei, B.C., Testing for varying zero-inflation and dispersion in generalized Poisson regression models, *Journal of Applied Statistics*, (2009a), (in press).

[66] Xie, F.C., Wei, B.C. and Lin, J.G., Score tests for zero-inflated generalized Poisson mixed regression models, *Computational Statistics and Data Analysis*, **53**(2009b), 3478–3489.

[67] Xie, F.C., Wei, B.C. and Lin, J.G., Assessing influence for pharmaceutical data in zero-inflated generalized Poisson mixed models, *Statistics in Medicine*, **27**(2008), 3656–3673.

[68] 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), 2907-2920.

[69] Yau, K.K.W., Lee, A.H. and Carrivick, P.J.W., Modeling zero-inflated count series with application to occupational health, *Computer Methods and Programs in Biomedicine*, **74**(2004), 47–52.

[70] Yau, K.K.W., Wang, K. and Lee, A.H., Zero-inflated negative binomial mixed regression modelling of over-dispersed count data with extra zeros, *Biometrical Journal*, **45**(2003), 437–452.

[71] Zhu, H.T. and Lee, S.Y., Local influence for incomplete-data models, *J. R. Statist. Soc. B*, **63**(2001), part 1, 111–126.

[72] Zhu, H.T., Lee, S.Y., Wei, B.C. and Zhu, J., Case-deletion measures for models with incomplete data, *Biometrika*, **88**(2001), 727–737.

Summary of Statistical Analysis for Zero-Inflated Data

Xie Fengchang[1,2] Wei Bocheng[1] Lin Jinguan[1]

([1] *Department of Mathematics, Southest University, Nanjing, 210096*)

([2] *Department of Mathematics, Nanjing Agricultural University, Nanjing, 210095*)

ZI (zero-inflated) data are data with overmuch zeroes. The ZI data have been commonly encountered in a wide variation of disciplines, and have been a hot topic since last decade. In this paper, we first present the actual significance of ZI data via two examples. Then we demonstrate the general situation and latest improvement of statistical analysis for zero-inflated data. Additionally, zero-inflated models, zero-inflated mixed models, and the estimation methods and some diagnostic problems are surveyed systematically. The relevant work done by authors in recent years are also introduced. Finally, several potential topics to be studied are listed.

Keywords: Zero-inflated data, EM algorithm, random effect models, statistical diagnostics, score test, dispersion parameter, longitudinal data.

AMS Subject Classification: 62J99.