应用概率统计　第二十八卷
第三期　2012年6月

Chinese Journal of Applied Probability
and Statistics Vol.28 No.3 Jun. 2012

一类零膨胀半参数混合效应模型的惩罚似然估计 *

施　红　星

(楚雄师范学院初等教育学院, 楚雄, 675000)

《应用概率统计》版权所用

摘　要

本文将半参数线性混合效应模型推广应用到一类具有零膨胀的纵向数据或集群数据的研究中, 提出了一类新的半参数混合效应模型, 然后利用广义交叉核实法选取光滑参数, 通过最大惩罚似然函数方法与EM算法给出了模型参数部分与非参数部分的估计方法, 最后, 通过模拟和实例说明了本文方法的有效性.

关键词:　零膨胀, 半参数模型, 惩罚似然, P-样条.

学科分类号:　O212.7.

§1.　引　　言

计数数据(count data)广泛存在于生物医学、遗传学、保险精算等多个研究领域中, Poisson分布、负二项分布、广义Poisson分布是常用的拟合分布, 但如果零观测值出现的概率大于其分布所允许的概率, 则认为该数据出现了零膨胀(zero-inflation)现象. 在实际问题中, 具有零膨胀的计数数据普遍存在. 例如在保险研究中, 由于损失过小导致保户放弃索赔的情况显著增加, 零索赔率的增大导致总体索赔次数出现零膨胀; 在流行病学研究中, 由于个体观测自身携带抗体导致其感染某病毒的次数为零, 这也使得总体感染次数的零膨胀. 近年来, 统计学家们对具有零膨胀的计数数据进行了多方面的研究, Fahrmeir和Echavarrri (2006)研究了一类零膨胀的可加性模型; Yip和Yau (2005)考查了各类零膨胀回归模型在保户赔付中的应用; Xie et al. (2009, 2008)研究了带有零膨胀的广义Poisson混合效应回归模型的Score检验问题和统计诊断与局部影响分析; Ghosh et al. (2006)讨论了零膨胀回归模型的Bayesian方法. 从已有研究来看, 目前对零膨胀数据的研究大多是建立参数回归模型或具有混合效应的参数模型, 而对于该类数据的半参数或非参数回归模型的相关研究较少, 有待进一步研究.

众所周知, 半参数线性混合效应模型和非参数广义线性混合效应模型是统计学研究的热门课题之一, 是处理纵向数据或集群数据的重要方法, 在生物医学、心理学、经济学等领域有广泛的应用, 有众多的研究成果. 如Davidian和Giltinan (1995), Diggle et al. (2002), Lin和Carroll (2001), He et al. (2002), 张浩和朱仲义(2006)等等. 本文将借鉴Zhang et al.

*国家社会科学基金(10BTJ001)和国家自然科学基金(11171105)资助.
本文2011年11月25日收到, 2011年12月28日收到修改稿.

(1998), 张浩和朱仲义(2006)和Lin和Carroll (2001)等研究中提出的最大惩罚似然函数方法给出了一类零膨胀半参数混合效应模型的估计方法, 并通过随机模拟和实例分析说明方法的有效性.

本文安排如下: 第二节介绍零膨胀半参数广义Poisson混合效应模型; 第三节给出P-样条及节点选取的方法; 第四节研究模型的估计方法和EM算法; 第五节通过模拟和实例说明本文方法的有效性.

§2. 模型与记号

首先, 我们给出零膨胀广义Poisson分布(Zero-Inflated Generalized Poisson Distribution)的定义, 简记为ZIGP. 假设Y服从ZIGP(ϕ,λ,α), 若

$$\mathsf{P}\{Y=y\}=\begin{cases}\phi+(1-\phi)p(0;\lambda,\alpha), & y=0;\\ (1-\phi)p(y;\lambda,\alpha), & y>0,\end{cases} \tag{2.1}$$

其中$p(y;\lambda,\alpha)$为广义Poisson分布的概率分布函数, 即为

$$p(y;\lambda,\alpha)=\frac{1}{y!}\Big(\frac{\lambda}{1+\alpha\lambda}\Big)^y(1+\alpha y)^{y-1}\exp\Big\{-\frac{\lambda(1+\alpha y)}{1+\alpha\lambda}\Big\},\qquad y=0,1,2,\ldots,$$

其中α为散度参数, λ为均值参数.

若Y服从广义泊松分布, 则Y的均值和方差分别为

$$\mathsf{E}[Y|\lambda,\alpha]=\lambda,\qquad \mathsf{Var}\,[Y|\lambda,\alpha]=\lambda(1+\alpha\lambda)^2,$$

显然当$\alpha\to 0$时, 上述分布即为Poisson分布.

若Y服从ZIGP(ϕ,λ,α), 则Y的均值和方差分别为

$$\mathsf{E}[Y|\phi,\lambda,\alpha]=(1-\phi)\lambda,\qquad \mathsf{Var}\,[Y|\phi,\lambda,\alpha]=(1-\phi)\lambda[(1+\alpha\lambda)^2+\phi\lambda,$$

当$\phi=0$时, ZIGP分布变为GP分布.

下面, 我们给出纵向数据或集群数据下一类零膨胀半参数混合效应模型的定义, 假设Y_{ij}为感兴趣的响应变量, y_{ij}为其观察值, 表示第i个接受试验单元在第j次观察或第i个群第j个样本的观察计数值, 对应的观察时间为t_{ij}, 相应的协变量为x_{ij},z_{ij}, $i=1,2,\ldots,n$, $j=1,2,\ldots,n_i$, 于是$(y_{ij},x_{ij},z_{ij},t_{ij})$为样本观测值.

进一步, 我们假设$Y_{ij}\sim$ ZIGP$(\phi_{ij},\lambda_{ij},\alpha)$, 一般情况下, 散度参数$\alpha$可以单独估计或事先给定, 我们分别针对零膨胀部分$\phi_{ij}$与模型均值部分$\lambda_{ij}$建立如下半参数混合效应模型(参考Wu和Zhang (2006)).

$$\begin{cases}\mathrm{logit}(\phi_{ij})=z_{ij}^T\gamma+g(t_{ij})+u_i,\\ \log(\lambda_{ij})=x_{ij}^T\beta+h(t_{ij})+v_i,\end{cases} \tag{2.2}$$

《应用概率统计》版权所用

其中γ与β分别为对应于协变量z_{ij}与x_{ij}的参数部分的系数, $g(\cdot)$与$h(\cdot)$为两个关于时间的未知光滑函数, 半参数部分为固定效应, u_i与v_i为随机效应, 假设$(u_i^T, v_i^T) \sim N(0, \Sigma)$.

在实际应用中, x_{ij}与z_{ij}往往是相同的协变量, 或者其中一个是另一个的一部分, 模型(2.1)和(2.2)就称为零膨胀半参数广义Poisson混合效应模型.

§3. P-样条及节点的选取

众所周知, 对于未知函数的近似已经有很多非参数回归技术和光滑方法, 其中较常用的有核光滑方法, 局部多项式拟合, 回归样条, 光滑样条和惩罚样条等. 这里, 为了兼顾计算复杂性和近似程度, 我们采用惩罚样条(P-样条)作为非参数光滑工具. 根据Green和Silverman (1994)的讨论, 给出模型(2.1)和(2.2)的惩罚对数似然函数为

$$\mathrm{PL}(\gamma, \beta, g, h) = l(y; \phi, \lambda) - \frac{1}{2}\lambda_1 \int (g''(t))^2 - \frac{1}{2}\lambda_2 \int (h''(t))^2 \mathrm{d}t, \tag{3.1}$$

其中$\lambda_1, \lambda_2 > 0$是光滑参数, 而

$$\begin{aligned} l(y; \phi, \lambda) \quad = \quad & \sum_{i=1}^{n} \sum_{j=1}^{n_i} \Big\{ I(y_{ij} = 0) \log \int [\phi_{ij} + (1 - \phi_{ij}) p(0; \lambda_{ij}, \alpha)] f(\omega_i; \Sigma) \mathrm{d}\omega_i \\ & + I(y_{ij} > 0) \log \int [1 - \phi_{ij} p(y_{ij}; \lambda_{ij}, \alpha)] f(\omega_i; \Sigma) \mathrm{d}\omega_i \Big\}, \end{aligned} \tag{3.2}$$

其中$\omega_i = (u_i^T, v_i^T)$, $f(\omega_i; \Sigma)$为对应的正态分布的密度函数.

进一步, 由Yu et al. (2002), Ruppert et al. (2003), 假设

$$g(t) = \tau_0 + \tau_1 t + \cdots + \tau_q t^q + \sum_{k=1}^{K} \tau_{q+k} (t - h_k)_+^q,$$
$$h(t) = \delta_0 + \delta_1 t + \cdots + \delta_l t^l + \sum_{k=1}^{K} \delta_{l+k} (t - h_k)_+^l,$$

其中$w_+^k = [\max(0, w)]^k$, $\{h_k\}_{k=1}^K$是样条节点, 分别记样条系数向量为

$$\tau = (\tau_0, \tau_1, \ldots, \tau_{q+k})^T, \qquad \delta = (\delta_0, \delta_1, \ldots, \delta_{l+k})^T,$$

基函数为

$$B_1(t) = \{1, t, \ldots, t^q, (t - h_1)_+^q, \ldots, (t - h_k)_+^q\}^T,$$
$$B_2(t) = \{1, t, \ldots, t^l, (t - h_1)_+^l, \ldots, (t - h_k)_+^l\}^T,$$

于是$g(\cdot)$与$h(\cdot)$的近似样条模型可记为$g(t) = \tau^T B_1(t)$, $h(t) = \delta^T B_2(t)$.

关于样条节点的选取采用Yu et al. (2002)的方法, 对于光滑、单调或单峰的函数一般取$5 \sim 10$个节点, 通常情况下, 取预测变量的等分点为节点; 若函数有很多极大值和极小值, 则需取10个以上的节点, 在本文的模拟及实例应用中, 我们均选取$K = 15$为样条节点个数.

《应用概率统计》版权所用

在选取样条节点后, 由Yu et al. (2002)可知$\int(\ddot{g}(t))^2\mathrm{d}t$与$\int(\ddot{h}(t))^2\mathrm{d}t$可以表示为$\tau^TK\tau$与$\delta^TK\delta$, 其中$K$是与节点$t$有关的对角矩阵; 得到惩罚矩阵后, 样条阶的确定采用Wu和Zhang (2006)的方法, 选取$[\mathrm{rank}(K^TK)/3]$为样条的阶, 本文的实际计算中, 我们由此可选$q=l=3$为样条的阶.

§4. 惩罚似然估计与EM算法

对于模型(2.1)和(2.2), 若直接利用(3.1)与(3.2)式对模型的未知成分进行参数估计, 是相当困难的, 因为涉及到高维积分, 难以求出解析解, 为此, 我们应用EM算法, 将模型中的随机效应$\omega=(u,v)$视为缺失数据. 记完全数据$D_c=(D_o,D_m)$, 其中D_m即为随机效应ω, 表示缺失数据, D_o为观察数据. 则基于完全数据的惩罚对数似然函数$\mathrm{PL}(\gamma,\beta,\delta,\tau,\Sigma|D_c)$可以表示为

$$\mathrm{PL}(\gamma,\beta,\delta,\tau,\Sigma|D_c)=l_1(\phi,\lambda)+\sum_{i=1}^n\log f(\omega_i;\Sigma)-\frac{1}{2}N\lambda_1\tau^TK\tau-\frac{1}{2}N\lambda_2\delta^TK\delta, \tag{4.1}$$

其中

$$\begin{aligned} l(\phi,\lambda) \quad = \quad & \sum_{i=1}^n\sum_{j=1}^{n_i}\{I(y_{ij}=0)\log[\phi_{ij}+(1-\phi_{ij})p(0;\lambda_{ij},\alpha)] \\ & +I(y_{ij}>0)\log[1-\phi_{ij}p(y_{ij};\lambda_{ij},\alpha)]\}, \end{aligned}$$

$N=\sum\limits_{i=1}^n n_i$, λ_1,λ_2为光滑参数.

现在基于(4.1)式构造EM算法.

E-step:

$$Q(\psi|\psi^{(m)})=\mathsf{E}[\mathrm{PL}(\gamma,\beta,\delta,\tau,\Sigma|D_c)|D_o,\psi^{(m)}];$$

M-step:

$$\psi^{(m+1)}=\arg\max_{\psi}[Q(\psi|\psi^{(m)})].$$

其中$\psi=(\gamma,\beta,\delta,\tau,\Sigma)$, $\psi^{(m)}$表示第m次的迭代值.

在计算E步时, 由于无法直接得到积分的解析表达式, 故先通过Metropolis算法得到条件分布$f_{\omega|y}(\cdot)$的随机样本, 然后再利用MonterCarlo方法近似计算条件期望. 具体来说, 选取$f(\omega)$为建议分布, 重新生成随机变量ω^*, 然后以下列概率接受ω^*, 否则保留ω,

$$q(\omega,\omega^*)=\min\Big\{1,\frac{f_{\omega|y}(\omega^*|y,\gamma,\beta,\tau,\delta,\Sigma)f(\omega)}{f_{\omega|y}(\omega|y,\gamma,\beta,\tau,\delta,\Sigma)f(\omega^*)}\Big\}.$$

在计算M步时, 根据模型针对随机效应分布的假设, 首先获得参数Σ的估计为

$$\widehat{\Sigma}=\frac{1}{J}\sum_{j=1}^J(\omega^{(j)})^T\cdot\omega^{(j)},$$

《应用概率统计》版权所用

其中$\{\omega^{(j)}\}_{j=1}^{J}$为上述Metropolis算法中获得的随机样本. 而参数$\theta=(\gamma,\beta,\delta,\tau)$, 则通过估计方程

$$\mathsf{E}\Big[\frac{\partial \mathrm{PL}(\theta,\widehat{\Sigma}|D_c)}{\partial\theta}\Big]=0,$$

用Newton-Raphson方法获得参数θ的估计.

光滑参数λ_1,λ_2的最优值由广义交叉核实法得到, 取

$$\mathrm{GCV}(\lambda_1,\lambda_2)=\frac{(2N)^{-1}\|Y_1-HY_1\|^2}{\{1-(2N)^{-1}\mathrm{tr}(H)\}^2},$$

其中$Y_1=(M^T,Y^T)$, Y^T为$\{y_{ij}\}$ $(i=1,2,\ldots,n,\ j=1,2,\ldots,n_i)$拉长的$N$维向量, M^T为对应于Y^T的指示向量, 当$y_{ij}=0$时, $M_{ij}=1$, 否则为0. 帽子矩阵H为

$$H=SC(C^TSC+\Delta)^{-1}C^T,$$

其中

$$C=\begin{pmatrix} B_1 & 0\\ 0 & B_2\end{pmatrix},\qquad S=\begin{pmatrix} S_1 & 0\\ 0 & I\end{pmatrix},\qquad \Delta=\begin{pmatrix} \lambda_1K & 0\\ 0 & \lambda_2K\end{pmatrix},$$

$$S_1=\mathrm{diag}(\exp\{z_{ij}^T\gamma+g(t_{ij})+u_i\}/[1+\exp\{z_{ij}^T\gamma+g(t_{ij})+u_i\}]^2).$$

计算时首先用格子点法固定λ_1后得到最小化GCV的λ_2^*, 然后令$\lambda_2=\lambda_2^*$, 再求使GCV最小化的$\lambda_1=\lambda_1^*$, 于是得到$(\lambda_1^*,\lambda_2^*)$为最佳光滑参数.

综上所述可以得到构造模型惩罚似然估计的算法如下:

(1) 选取参数的初值$\psi^{(0)}$, 令$m=0$;

(2) 用GCV法得到极小化GCV的最佳光滑参数$\lambda_1^{(m)},\lambda_2^{(m)}$;

(3) 由EM算法得到$\psi^{(m+1)}$的估计;

(4) 令$m=m+1$, 返回(2)迭代, 直至迭代收敛.

§5. 随机模拟与实际例子

为了验证本文提出的惩罚似然估计方法的有效性, 我们对以下模拟和实际例子进行分析.

5.1 模拟分析

在该部分中, 我们考虑一个简化的随机模拟模型, 考虑如下的一个零膨胀Poisson模型: $Y_{ij}\sim \mathrm{ZIP}(\phi_{ij},\lambda_{ij})$, 而

$$\begin{cases}\mathrm{logit}(\phi_{ij})=\gamma_0+\gamma_1x_{1ij}+\gamma_2x_{2i}+u_i+g(t_{ij});\\ \log(\lambda_{ij})=\beta_0+\beta_1x_{1ij}+\beta_2x_{2i}+v_i+h(t_{ij}),\end{cases}$$

《应用概率统计》版权所用

其中$i=1,2,\ldots,n$, $j=1,2,\ldots,n_i$, $x_{1ij}\sim U[-1,1]$, x_{2i}服从$\{-1,1\}$上的离散均匀分布, $t_{ij}\sim U[-2,2]$, $(u_i,v_i)^T\sim N(0,\Sigma)$, 取$\Sigma$为

$$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}=\begin{pmatrix}1 & 0.3\\ 0.3 & 1\end{pmatrix}.$$

而

$$h(t)=\frac{\sin(\pi t/2)}{1+2t^2[\mathrm{sign}(t)+1]},\qquad g(t)=\sin(\pi t/2).$$

参数真值取为$\beta_0=1.5$, $\beta_1=-1$, $\beta_2=2$, $\gamma_0=-2$, $\gamma_1=0.5$, $\gamma_2=-1$, 在每个数据集中令$n=50$, $n_i=20$, 重复1000次模拟后, 得到的参数部分的偏差与标准误差结果见表1. 最优的光滑参数λ_1介于3.25 ~ 3.56之间, λ_2介于1.62 ~ 2.32之间, 非参数部分的估计见图1, 图2所示(图中实线为估计值, 虚线为真实值, 点画线为5%与95%分位值).

表1 随机模拟参数估计

参数	β_0	β_1	β_2	γ_0	γ_1	γ_2
偏差	0.036	0.066	0.023	0.012	0.058	0.032
标准误	0.397	0.175	0.367	0.271	0.305	0.290

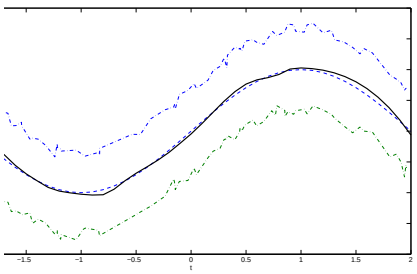

图1 模拟数据$g(t)$估计值

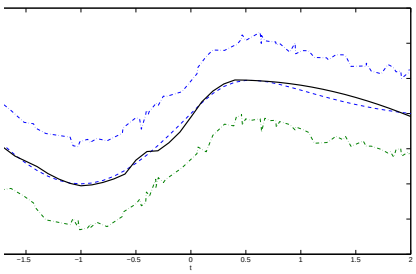

图2 模拟数据$h(t)$估计值

5.2 实际例子

本小节, 我们以上海市青少年体质测试抽查数据(国家社科基金10BTJ001资助项目)为例利用本文方法进行分析. 该项测试于2008–2009年间在上海市抽取了21所中小学7岁至17岁的中小学生4687名, 检测了包括身体形态, 身体功能和身体素质等方面的47项指标, 为验证本文提出的方法, 我们仅针对身体素质中的一项运动能力进行研究. 根据相关的体质测试标准, 选取“引体向上”作为响应变量(y), 学生的性别(x_1), 体重指数BMI(x_2, 体重(kg)/身高的平方(m^2))以及年龄(t)作为协变量, 同时考虑学校的效应, 随机抽取10所学校, 每所200名学生作为集群数据进行分析.

《应用概率统计》版权所用

初步的探索性分析可以发现y具有明显的零膨胀特征, 且针对$y>0$的情况用Poisson分布拟合也存在超散度的情况, 单独估计的散度参数$\alpha=1.311$. 综上我们建立如下的ZIGP半参数混合效应模型, $y_{ij}\sim \mathrm{ZIGP}(\phi_{ij},\lambda_{ij},\alpha)$, $i=1,2,\ldots,10$ (学校), $j=1,2,\ldots,200$ (学生).

$$\begin{cases}\mathrm{logit}(\phi_{ij})=\gamma_0+\gamma_1x_{1ij}+\gamma_2x_{2ij}+u_i+g(t_{ij});\\ \log(\lambda_{ij})=\beta_0+\beta_1x_{1ij}+\beta_2x_{2ij}+v_i+h(t_{ij}).\end{cases}$$

实例中参数部分的估计结果如表2所示, 非参数部分的估计结果如图3, 图4所示.

表2　实际例子参数估计

参数	β_0	β_1	β_2	γ_0	γ_1	γ_2
估计值	1.304	1.108	0.544	-4.065	-0.588	0.082
标准差	0.640	0.170	0.302	1.155	0.055	0.081

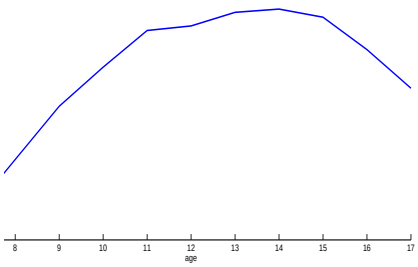

图3　实例数据$g(\cdot)$估计值

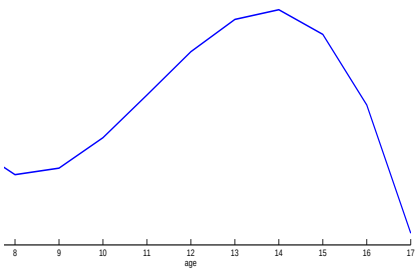

图4　实例数据$h(\cdot)$估计值

参 考 文 献

[1] Fahrmeir, L. and Echavarrri, L.O., Structured additive regression for overdispersed and zero-inflated count data, *Applied Stochastic Models in Business and Industry*, **22(5)**(2006), 351–369.

[2] Yip, K.C.H. and Yau, K.K.W., On modeling claim frequency data in general insurance with extra zeros, *Insurance: Mathematics and Economics*, **36(2)**(2005), 153–163.

[3] Xie, F.C., Wei, B.C. and Lin, J.G., Score test for zero-inflated generalized Poisson mixed regression models, *Computational Statistics & Data Analysis*, **53(9)**(2009), 3478–3489.

[4] 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(8)**(2008), 3656–3673.

[5] Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C., Bayesian analysis of zero-inflated regression models, *Journal of Statistical Planning and Inference*, **136(4)**(2006), 1360–1375.

[7] Davidian, M. and Giltinan, D.M., *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall, 1995.

[7] Diggle, P.J., Liang, K.Y. and Zeger, S.L., *Analysis of Longitudinal Data*, Cambridge University Press, 2002.

《应用概率统计》版权所用

[8] Lin, X.H. and Carroll, R.J., Semiparametric regression for clustered data, *Biometrika*, **88(4)**(2001), 1179–1185.

[9] He, X.M., Zhu, Z.Y. and Fung, W.K., Estimation in semiparametrics model for longitudinal data with unspecificed dependence structure, *Biometrika*, **89(3)**(2002), 579–590.

[10] 张浩, 朱仲义, 半参数广义线性混合效应模型的估计及其渐近性质, 应用概率统计, **25(2)**(2009), 171–184.

[11] Zhang, D., Lin, X. and Sower, M.F., Semiparametric stochastic mixed model for longitudinal data, *Journal of the American Statistical Association*, **93(442)**(1998), 710–719.

[12] Wu, H. and Zhang, J.T., *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley Interscience, 2006.

[13] Green, P.J. and Silverman, B.W., *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, 1994.

[14] Yu, Y. and Ruppert, D., Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association*, **97(460)**(2002), 1042–1054.

[15] Ruppert, D., Wand, M.P. and Carroll, R.J., *Semiparametric Regression*, Cambridge University Press, 2003.

《应用概率统计》版权所用

Penalized Likelihood Estimation of a Class of Zero-Inflated Semiparametric Mixed-Effect Model

Shi Hongxing

(*School of Primary Education, Chuxiong Normal University, Chuxiong, 675000*)

In this paper, we consider an extension of semiparametric linear mixed-effect model to a class of longitudinal data or clustered data with zero-inflation and propose a novel semiparametric mixed model. Based on the maximum penalized likelihood estimation and EM algorithm, we give a method for our proposed model which may estimate both of the parameters and nonparameters simultaneously. In the method, we use GCV to choose the smoothing parameter.Finally, we study a simulation analysis and one real data set to illustrate the proposed method.

Keywords: Zero-inflation, semiparametric model, penalized likelihood, P-spline.

AMS Subject Classification: 62G05.