

# 零膨胀集群数据层次回归模型的贝叶斯推断 \*

施 红 星

(楚雄师范学院初等教育学院, 楚雄, 675000)

## 摘要

零膨胀Poisson回归模型是研究零观测值过多的计数数据的常用工具, 本文提出了一类拟合具有这类特征的集群数据的层次零膨胀泊松回归模型, 并给出了相应的贝叶斯推断方法, 参数估计通过Gibbs抽样获得, 模型比较与选择则通过拟合优度检验与BIC准则实现. 最后, 利用一个船舶受损事故数据来展示本文方法的实现及应用.

关键词: 零膨胀, 层次回归模型, Gibbs抽样, BIC准则.

学科分类号: O212.8.

## §1. 引 言

计数数据(count data)广泛存在于金融、保险、临床医学、遗传学以及抽样调查等多个研究领域中, 常用的拟合分布有负二项分布、Poisson分布、广义Poisson分布等. 如果零观测值出现的概率大于其分布所允许的概率, 我们则认为该数据出现了零膨胀(zero-inflation)现象. 在实际问题中, 普遍存在具有零膨胀特征的计数数据. 例如在流行病学研究中, 由于个体观测者自身携带抗体导致其感染某病毒的次数为零, 使得总体感染次数出现零膨胀现象; 在保险精算研究中, 由于损失过小导致保户放弃索赔的情况频繁发生, 零索赔率的增大导致总体索赔次数出现零膨胀. 近年来, 统计学家对具有零膨胀特征的计数数据进行了多方面研究, 如Fahrmeir和Echavarría (2006)研究了一类零膨胀的可加性模型; Yip和Yau (2005)讨论了各类零膨胀回归模型在保户赔付中的应用; Xie等人(2009, 2008)系统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等人(2006)研究了零膨胀回归模型的贝叶斯方法. 从目前已有研究来看, 对零膨胀数据的研究主要集中在基于各种分布的参数回归模型或具有混合效应的参数模型的建模方面.

近年来, 层次回归模型的研究取得了重大进展, 这类模型是一族灵活、广泛的统计模型, 综合了线性回归模型和随机效应模型的优势, 日益发展成为分析集群数据、重复测量数据或层次数据等复杂数据的标准工具. 相关研究成果参见文献[6]-[10], 回顾已有成果, 关于零膨胀集群数据的层次回归模型的研究较少, 有待于进一步讨论.

\*国家社会科学基金(10BTJ001)和国家自然科学基金(11171105)资助.

本文2012年4月9日收到, 2012年5月2日收到修改稿.

本文借鉴Lee等人(2006)的工作, 针对具有零膨胀特征的集群数据提出两层次的回归模型, 然后利用数据添加的思想给出模型的贝叶斯推断方法, 进而在贝叶斯拟合优度检验与BIC准则之下讨论模型的比较与选择问题, 并通过一个实例分析说明方法的实现及应用.

本文安排如下: 第二节介绍零膨胀集群数据层次回归模型; 第三节研究贝叶斯建模与推断; 第四节讨论模型的拟合检验和模型选择; 第五节通过实例展示本文方法的实现及应用.

## §2. 模型设定

首先, 我们给出零膨胀Poisson分布(Zero-Inflated Poisson Distribution)的定义, 简记为ZIP. 假设 $Y$ 服从ZIP( $\phi, \lambda$ ), 若

$$P\{Y = y\} = \begin{cases} \phi + (1 - \phi) \exp(-\lambda), & y = 0; \\ (1 - \phi) \exp(-\lambda) \lambda^y / y!, & y > 0, \end{cases} \quad (2.1)$$

其中 $0 < \phi < 1$ 为零膨胀参数, 显然, 当 $\phi = 0$ 时, ZIP分布变为Poisson分布.

进一步, 我们给出集群数据下零膨胀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*个群第*j*个样本的观察计数值, 相应的协变量为 $x_{ij}, z_{ij}, i = 1, 2, \dots, m, j = 1, 2, \dots, n_i, N = \sum_{i=1}^m n_i$ , 若 $Y_{ij} \sim \text{ZIP}(\phi_{ij}, \lambda_{ij})$ , 考虑集群间的相关性, 我们分别针对零膨胀参数 $\phi_{ij}$ 与模型均值部分 $\lambda_{ij}$ 建立如下的混合效应模型:

$$\begin{cases} \xi_{ij} = \log(\lambda_{ij}) = x_{ij}^T \beta + u_i; \\ \eta_{ij} = \log\left(\frac{\phi_{ij}}{1 - \phi_{ij}}\right) = z_{ij}^T \gamma + v_i, \end{cases} \quad (2.2)$$

其中 $\beta$ 与 $\gamma$ 分别为对应于协变量 $x_{ij}$ 与 $z_{ij}$ 的回归系数,  $u_i$ 与 $v_i$ 为随机效应, 按照通常的假设, 设 $u_i$ 与 $v_i$ 相互独立, 分别服从均值为0, 方差为 $\sigma_u^2, \sigma_v^2$ 的正态分布.

在实际问题中, 由于数据收集过程或集群间嵌套关系的影响, 上述的混合效应模型还可能呈现内在关联或具有层次结构, 受Lee等人(2006)结果的启发, 为了更加准确且灵活地刻画数据的这类关系, 我们定义基于上述ZIP混合效应模型的两层次回归模型为

$$\xi_{ij} = \log(\lambda_{ij}) = \beta_{0i} + x_{ij}^T \beta_{1i} + \varepsilon_{ij}^{(1)}, \quad (2.3)$$

其中

$$\begin{cases} \beta_{0i} = \alpha_{00} + w_i^T \alpha_{01} + u_{0i}, \\ \beta_{1i} = \alpha_{10} + w_i^T \alpha_{11} + u_{1i}, \end{cases} \quad (2.4)$$

而

$$\eta_{ij} = \text{logit}(\phi_{ij}) = \gamma_{0i} + z_{ij}^T \gamma_{1i} + \varepsilon_{ij}^{(2)}, \quad (2.5)$$

其中

$$\begin{cases} \gamma_{0i} = \delta_{00} + w_i^T \delta_{01} + v_{0i}, \\ \gamma_{1i} = \delta_{10} + w_i^T \delta_{11} + v_{1i}, \end{cases} \quad (2.6)$$

其中(2.3)与(2.5)为第一层回归方程,  $\beta_{0i}, \gamma_{0i}$ 与 $\beta_{1i}, \gamma_{1i}$ 分别是个体层次回归分析的截距项与斜率项,  $\varepsilon_{ij}^{(1)}, \varepsilon_{ij}^{(2)}$ 是个体层次回归分析的误差项, 分别服从 $N(0, \sigma_1^2)$ 与 $N(0, \sigma_2^2)$ , 而(2.4)与(2.6)为第二层回归方程, 分别针对个体层次的截距项和斜率项进行回归,  $w_i^T = (w_{1i}, \dots, w_{pi})$ 是集群层次的协变量, 表示群之间的嵌套或关联关系, 往往以哑变量或二分变量为主, 而 $u_i^T = (u_{0i}, u_{1i})$ ,  $v_i^T = (v_{0i}, v_{1i})$ 分别是集群层次回归模型的随机效应, 且假设  $u_i \sim N(0, \Sigma_u)$ ,  $v_i \sim N(0, \Sigma_v)$ , 其中 $\Sigma_u$ 与 $\Sigma_v$ 为未知的协方差矩阵.

若记 $\alpha = (\alpha_0^T, \alpha_1^T)$ , 而 $\alpha_0^T = (\alpha_{00}, \alpha_{01})$ ,  $\alpha_1^T = (\alpha_{10}, \alpha_{11})$ ;  $\delta = (\delta_0^T, \delta_1^T)$ , 而 $\delta_0^T = (\delta_{00}, \delta_{01})$ ,  $\delta_1^T = (\delta_{10}, \delta_{11})$ 为感兴趣参数, 则上述(2.3)-(2.6)对应的层次回归模型可以表示为如下的矩阵、向量形式:

$$\begin{cases} \xi = X\beta + \varepsilon^{(1)}; \\ \eta = Z\gamma + \varepsilon^{(2)}, \end{cases} \quad \begin{cases} \beta = W\alpha + u; \\ \gamma = W\delta + v, \end{cases} \quad (2.7)$$

其中 $\xi, \eta$ 为(2.3)、(2.5)式中对应的向量, 其余记号有类似含义.

### §3. 贝叶斯建模与推断

相比于似然方法, 贝叶斯方法综合了样本中的先验信息, 对于某些分布与复杂模型的建模具有特别的灵活性. 下面, 我们具体讨论本文模型的贝叶斯建模.

#### 3.1 关于潜变量的数据添加方法

根据Ghosh等人(2006)的工作, 零膨胀回归模型中的响应变量 $Y_{ij}$ 可以表示为 $Y_{ij} = D_{ij}(1 - B_{ij})$ , 其中 $B_{ij}$ 是具有参数 $\phi_{ij}$ 的二点分布随机变量, 而 $D_{ij}$ 服从参数为 $\lambda_{ij}$ 的Poisson分布, 于是 $Y_{ij}$ 作为观测数据可以分解为由潜变量 $(D_{ij}, B_{ij})$ 组成的完全数据, 而基于完全数据 $(D_{ij}, B_{ij})$ 的建模可以通过下列条件分布进行数据添加来实现.

给定 $Y_{ij} = y_{ij}$ ,  $(D_{ij}, B_{ij})$ 的联合条件分布为:

当 $y_{ij} > 0$ 时,

$$P(D_{ij} = d_{ij}, B_{ij} = 0 | Y_{ij} = y_{ij}) = 1,$$

当 $y_{ij} > 0$ 时,

$$P(D_{ij} = d_{ij}, B_{ij} = 1 | Y_{ij} = y_{ij}) = \frac{\phi_{ij} P(D_{ij} = d_{ij})}{\phi_{ij} + (1 - \phi_{ij}) P(D_{ij} = 0)},$$

$$P(D_{ij} = 0, B_{ij} = 0 | Y_{ij} = y_{ij}) = \frac{(1 - \phi_{ij}) P(D_{ij} = 0)}{\phi_{ij} + (1 - \phi_{ij}) P(D_{ij} = 0)},$$

对应于观测数据  $\{Y_{ij} = y_{ij} : i = 1, \dots, m, j = 1, \dots, n_i\}$ , 我们可以得到潜变量  $\{D_{ij}, B_{ij}\}$  的样本, 用  $(D, B)$  表示其向量形式.

### 3.2 先验分布的选择

设  $\theta = (\alpha, \delta, \sigma^2, \Sigma)$  为全体参数组成的向量, 其中  $\alpha, \delta$  为感兴趣参数,  $\sigma^2 = (\sigma_1^2, \sigma_2^2)$ ,  $\Sigma = (\Sigma_u, \Sigma_v)$  为未知的多余参数, 根据参数的意义, 在后面的贝叶斯分析中选择如下的独立先验分布, 即

$$\pi(\theta) = \pi(\alpha)\pi(\delta)\pi(\sigma^2)\pi(\Sigma),$$

其中  $\pi(\alpha)$  选择为正态先验  $N(0, \Omega_\alpha)$ ,  $\pi(\delta)$  为  $N(0, \Omega_\delta)$ , 而  $\Omega_\alpha = \text{diag}(\sigma_\alpha^2)$ ,  $\Omega_\delta = \text{diag}(\sigma_\delta^2)$ ,  $\sigma_\alpha^2$  与  $\sigma_\delta^2$  为给定的超参数.  $\pi(\sigma^2)$  选择为逆Gamma分布, 即

$$\sigma_1^{-2} \sim \text{Gamma}(a_1, b_1), \quad \sigma_2^{-2} \sim \text{Gamma}(a_2, b_2),$$

$\pi(\Sigma)$  选择为逆Wishart分布, 即

$$\Sigma_u^{-1} \sim \text{Wishart}(\rho, (\rho R_u)^{-1}), \quad \Sigma_v^{-1} \sim \text{Wishart}(\rho, (\rho R_v)^{-1}),$$

而其中  $a_1, b_1, a_2, b_2, \rho, R_u, R_v$  均为事先给定的超参数.

超参数的选取可通过给定的先验信息来实现, 若无任何先验信息, 通常的选择为  $\sigma_\alpha^2 = \sigma_\delta^2 = 1000$ ,  $a_i = b_i = 0.001$  ( $i = 1, 2$ ),  $R_u = R_v = I$ ,  $I$  为单位阵,  $\rho = 2$ .

### 3.3 Gibbs抽样与M-H算法

基于上述先验设定与数据添加方法, 我们给出下列通过满条件分布来获得随机样本的 Gibbs 抽样过程:

第一步: 设置初始值  $(\theta^{(0)}, u^{(0)}, v^{(0)}, \varepsilon^{(0)})$ , 然后计算

$$\phi_{ij}^{(0)} = \frac{\exp(\eta_{ij}^{(0)})}{1 + \exp(\eta_{ij}^{(0)})},$$

其中  $\eta_{ij}^{(0)}$  可以通过(2.5)式计算得到.

第二步: 在当前值  $(\theta^{(t)}, u^{(t)}, v^{(t)}, \varepsilon^{(t)})$  之下, 由(2.3)与(2.5)式计算  $\phi_{ij}^{(t)}$  与  $\lambda_{ij}^{(t)}$ , 然后, 对所有  $\{y_{ij} : i = 1, \dots, m, j = 1, \dots, n_i\}$ , 产生潜变量  $(D_{ij}^{(0)}, B_{ij}^{(0)})$  如下:

若  $y_{ij} = 0$ , 抽取  $B_{ij}^{(t)} \sim \text{Bernoulli}(\phi_{ij}^{(t)})$ , 然后若  $B_{ij}^{(t)} = 0$ , 则令  $D_{ij}^{(t)} = 0$ ; 否则, 抽取  $D_{ij}^{(t)} \sim \text{Poisson}(\lambda_{ij}^{(t)})$ ;

若  $y_{ij} > 0$ , 则令  $B_{ij}^{(t)} = 0$  及  $D_{ij}^{(t)} = y_{ij}$ .

第三步: 在当前值 $(\theta^{(t)}, u^{(t)}, v^{(t)}, \varepsilon^{(t)})$ 及 $(D^{(t)}, B^{(t)})$ 之下, 分别从下列满条件后验分布中抽取所需后验样本,

$$P((\varepsilon^{(1)})^{(t+1)}|\text{rest}) \propto \exp \left\{ -\frac{1}{2(\sigma_1^2)^{(t)}} (\xi^{(t)} - X\beta^{(t)})^T (\xi^{(t)} - X\beta^{(t)}) \right\},$$

其中“rest”表示除 $\varepsilon^{(1)}$ 之外的所有 $\theta$ 与 $u, v$ 中的参数及未观测量.

$$\begin{aligned} P(\beta^{(t+1)}|\text{rest}) &\propto \exp \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} [D_{ij}^{(t)} \cdot \xi_{ij}^{(t)} - \exp(\xi_{ij}^{(t)})] \right. \\ &\quad \left. - \frac{1}{2} (\beta^{(t)} - W\alpha^{(t)})^T [\Sigma_u^{(t)}]^{-1} (\beta^{(t)} - W\alpha^{(t)}) \right\}, \\ P(u^{(t+1)}|\text{rest}) &\sim N(\beta^{(t)} - W\alpha^{(t)}, \Sigma_u^{(t)}), \\ P((\sigma_1^{-2})^{(t+1)}|\text{rest}) &\sim \text{Gamma}\left(\frac{N}{2} + a_1, \left(b_1^{-1} + \frac{1}{2} \|(\varepsilon^{(1)})^{(t)}\|\right)^{-1}\right), \\ P(\Sigma_u^{(t+1)}|\text{rest}) &\sim \text{Wishart}\left(m + \rho, \left(\sum_{i=1}^m u_i^{(t)} \cdot u_i^{(t)T} + \rho R_u\right)^{-1}\right). \end{aligned}$$

类似地, 我们有

$$\begin{aligned} P((\varepsilon^{(2)})^{(t+1)}|\text{rest}) &\propto \exp \left\{ -\frac{1}{2(\sigma_2^2)^{(t)}} (\eta^{(t)} - Z\gamma^{(t)})^T (\eta^{(t)} - Z\gamma^{(t)}) \right\}, \\ P(\gamma^{(t+1)}|\text{rest}) &\propto \exp \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} [B_{ij}^{(t)} \cdot \eta_{ij}^{(t)} - \exp(\eta_{ij}^{(t)})] \right. \\ &\quad \left. - \frac{1}{2} (\gamma^{(t)} - W\delta^{(t)})^T [\Sigma_v^{(t)}]^{-1} (\gamma^{(t)} - W\delta^{(t)}) \right\}, \\ P(v^{(t+1)}|\text{rest}) &\sim N(\gamma^{(t)} - W\delta^{(t)}, \Sigma_v^{(t)}), \\ P((\sigma_2^{-2})^{(t+1)}|\text{rest}) &\sim \text{Gamma}\left(\frac{N}{2} + a_2, \left(b_2^{-1} + \frac{1}{2} \|(\varepsilon^{(2)})^{(t)}\|\right)^{-1}\right), \\ P(\Sigma_v^{(t+1)}|\text{rest}) &\sim \text{Wishart}\left(m + \rho, \left(\sum_{i=1}^m v_i^{(t)} \cdot v_i^{(t)T} + \rho R_v\right)^{-1}\right). \end{aligned}$$

在上述抽取过程中,  $P(\varepsilon^{(1)}|\text{rest}), P(\varepsilon^{(2)}|\text{rest}), P(\beta|\text{rest}), P(\gamma|\text{rest})$ 是复杂的非标准分布, 需要利用M-H算法来实现, 具体过程参见Gelman和Rubin (1992), 得到上述抽样样本后, 参数 $\alpha^{(t+1)}$ 与 $\delta^{(t+1)}$ 可以根据模型(2.7)利用最小二乘估计得到, 综上, 样本 $(\theta^{(t+1)}, u^{(t+1)}, v^{(t+1)}, \varepsilon^{(t+1)})$ 得到.

第四步: 重复第二步与第三步, 直到样本收敛.

### 3.4 贝叶斯推断

在后验样本的抽样过程中, 抽样的收敛情况可以通过Gelman & Rubin统计量来监测(参见Gelman和Rubin (1992)), 在实际分析过程中, 也可以通过不同初始值的样本分布图来判断, 且Gibbs抽样的顺序不会影响Bayes估计的结果, 当样本收敛后, 样本观察值的序

《应用概率统计》

列 $\{(\theta^{(j)}, u^{(j)}, v^{(j)}, \varepsilon^{(j)}) : j = 1, \dots, J\}$ 也视为所需的后验样本的观察值(参见Geyer (1992)),于是可以得到相应感兴趣参数的估计值为

$$\hat{\theta} = \frac{1}{T} \sum_{t=1}^T \theta^{(t)}, \quad \hat{\Sigma}_u = \frac{1}{T} \sum_{t=1}^T u^{(t)} \cdot u^{(t)T}, \quad \hat{\Sigma}_v = \frac{1}{T} \sum_{t=1}^T v^{(t)} \cdot v^{(t)T},$$

其中 $t = 1, \dots, T$ 对应于收敛样本中足够大的子集.

类似地, 我们也可以得到相应估计值的标准误的估计值.

## §4. 模型的拟合检验与模型选择

### 4.1 拟合检验统计量

根据讨论数据与设定模型的特点, 我们采用Johnson (2004)给出的贝叶斯拟合检验统计量来评价模型对数据的拟合程度, 定义

$$G(\tilde{\theta}) = \sum_{k=1}^K [m_k(\tilde{\theta}) - Np_k]^2 / Np_k,$$

其中 $K$ 为等概率分组的组数,  $N$ 为所有观察值的总和, 令 $p_k = 1/K$ , 而 $m_k(\tilde{\theta})$ 是给定模型之下的第 $k$ 个分组中的实际观测的样本点数,  $\tilde{\theta}$ 是前述的参数的贝叶斯估计.  $m_k(\tilde{\theta})$ 可以通过如下步骤计算:

第一步: 对所有的 $\{y_{ij} : i = 1, \dots, m, j = 1, \dots, n_i\}$ , 计算参数估计值 $\tilde{\theta}$ 之下的累积概率 $F_{ij}(y_{ij}|\tilde{\theta})$ , 即

$$\sum_{l=0}^{y_{ij}} [\phi_{ij} + (1 - \phi_{ij})e^{-\lambda_{ij}}]^{I(l=0)} \left[ (1 - \phi_{ij})e^{-\lambda_{ij}} \frac{\lambda_{ij}^l}{l!} \right]^{I(l>0)}.$$

第二步: 对所有的 $\{y_{ij} : i = 1, \dots, m, j = 1, \dots, n_i\}$ , 按如下规则抽取相应的 $u_{ij}$ : 若 $y_{ij} = 0$ , 则抽取 $u_{ij} \sim U(0, F_{ij}(0|\tilde{\theta}))$ ; 若 $y_{ij} > 0$ , 则抽取 $u_{ij} \sim U(F_{ij}(y_{ij}-1|\tilde{\theta}), F_{ij}(y_{ij}|\tilde{\theta}))$ .

第三步: 计算

$$m_k(\tilde{\theta}) = \#\left\{ \frac{k-1}{K} < u_{ij} < \frac{k}{K} \right\},$$

对所有 $i = 1, \dots, m, j = 1, \dots, n_i, k = 1, \dots, K$ . 通常情况下, 我们选取 $K \approx N^{0.4}$ .

在得到检验统计量的基础上, 我们可以通过计算如下的 $p$ -值来评估模型的拟合程度,  $P[\chi_{K-1}^2 \geq G(\tilde{\theta})]$ , 即上述 $p$ -值表示自由度为 $(K-1)$ 的 $\chi^2$ 分布对应的上尾概率,  $p$ -值越小表明拟合程度越差.

### 4.2 模型选择准则

我们给出下列BIC作为模型选择的准则:

$$\text{BIC} = -2 \log l(\tilde{\theta}|Y) + \log N \cdot \text{Par},$$

其中  $\log l(\tilde{\theta}|Y)$  是基于贝叶斯估计  $\tilde{\theta}$  的观察数据对应的对数似然值, Par 是给定模型中的所有参数的个数. BIC 值越小, 表明模型对数据的拟合越好, 也就越能够被选择.

在实际应用中, 与本文给出的层次ZIP模型一起, 一般的ZIP模型, ZIP混合效应模型, 以及分栏Poisson模型(Hurdle Poisson Model)也常常作为候选模型对数据进行拟合, 我们可以综合上述的拟合检验  $p$ -值和BIC值来对模型进行比较和选择.

## §5. 实际例子

本节中我们以McCullagh和Nelder (1989)中讨论过的船舶受损数据为实际例子, 说明本文给定模型与方法的应用.

### 5.1 数据及模型

该数据是Lloyd社记录的34条船只在一个5年使用时期中发生事故导致受损的情况, 是一组重复测量数据, 同时, 船只的种类, 建造时间及已服务年限是重要的协变量. 对数据的基本分析可以发现零膨胀的特征. 为了深入分析上述协变量对受损情况及零膨胀情况的影响, 确定潜在的非同质风险因素, 我们建立如下的两层次ZIP模型:

对应个体观测层次, 假设

$$\begin{cases} [y_{ij}|\phi_{ij}, \lambda_{ij}] \sim \text{ZIP}(\phi_{ij}, \lambda_{ij}) \\ \log(\lambda_{ij}) = \beta_{0i} + \beta_{1i}t_{ij} + \varepsilon_{ij}^{(1)} & i = 1, \dots, 34, j = 1, 2, \dots, n_i \\ \text{logit}(\phi_{ij}) = \gamma_{0i} + \gamma_{1i}t_{ij} + \varepsilon_{ij}^{(2)} \end{cases}$$

对于集群层次, 假设

$$\begin{cases} \beta_{mi} = \alpha_{m0} + \sum_{k=1}^5 x_{ki}^{(1)} \alpha_{mk}^{(1)} + \sum_{l=1}^4 x_{li}^{(2)} \alpha_{ml}^{(2)} + \sum_{s=1}^2 x_{si}^{(3)} \alpha_{ms}^{(3)} + u_{mi} \\ \gamma_{mi} = \delta_{m0} + \sum_{k=1}^5 x_{ki}^{(1)} \delta_{mk}^{(1)} + \sum_{l=1}^4 x_{li}^{(2)} \delta_{ml}^{(2)} + \sum_{s=1}^2 x_{si}^{(3)} \delta_{ms}^{(3)} + v_{mi} \end{cases} \quad m = 0, 1.$$

其中  $y_{ij}$  表示第  $i$  条船在 5 年间第  $j$  段观测期间的受损次数,  $t_{ij}$  为对应的累积使用时间取  $\log$  后的值,  $x_{ki}^{(1)}, x_{li}^{(2)}, x_{si}^{(3)}$  ( $k = 1, \dots, 5, l = 1, \dots, 4, s = 1, 2$ ) 是 0-1 变量, 分别表示第  $i$  条船属于的船舶类型(A, B, C, D, E), 建造年代( $Y_1, \dots, Y_4$ ) 和服务年限( $S_1, S_2$ ).

### 5.2 模型比较与估计结果

除了上述的两层次ZIP模型之外, 我们还利用一般的ZIP模型, ZIP混合效应模型, 以及分栏Poisson模型(简记为HPM)来对该数据进行拟合, 通过贝叶斯拟合优度检验与BIC准则来选择模型, 计算结果如表1所示.

表1 候选模型的 $p$ -值与BIC值

模型	层次ZIP	ZIP	ZIP混合效应	HPM
$p$ -值	0.673	0.198	0.519	0.028
BIC值	12443.17	13209.08	13141.26	14532.83

从表1可以看出, 层次ZIP模型的 $p$ -值最大而BIC值最小, 说明在所给候选模型中最为适合, 其次依次为ZIP混合效应模型, ZIP模型与HPM模型.

具体的层次ZIP模型估计结果见表2.

表2 层次ZIP模型的参数估计值

参数	Poisson成分				logistic成分			
	截距		斜率		截距		斜率	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差
常数	2.116	0.180	-3.472	0.163	-1.565	0.214	-10112	0.113
A	0.869	0.144	0.817	0.192	0.382	0.057	1.047	0.236
B	-0.289	0.125	0.370	0.038	-0.218	0.013	0.403	0.019
C	-1.360	0.140	1.313	0.096	0.378	0.049	0.951	0.045
D	-1.195	0.141	-0.525	0.040	0.384	0.092	1.462	0.131
E	0.367	0.121	0.457	0.095	-0.457	0.074	1.471	0.097
$Y_1$	1.504	0.129	1.462	0.269	-0.767	0.165	0.645	0.042
$Y_2$	0.749	0.182	1.741	0.214	1.535	0.103	0.417	0.051
$Y_3$	1.623	0.142	0.658	0.254	0.292	0.083	0.499	0.156
$Y_4$	-0.292	0.082	1.316	0.169	-0.901	0.125	1.242	0.069
$S_1$	2.793	0.181	0.535	0.013	-0.862	0.043	0.939	0.171
$S_2$	-1.676	0.201	0.471	0.025	0.832	0.033	1.432	0.149

### 5.3 结果分析及应用

由表2中的估计结果计算后可知, 最大与最小的零膨胀参数分别为0.2624( $\phi_{10,3}$ )与0.1035( $\phi_{24,2}$ ), 均离0较远, 说明数据集的零膨胀现象很明显, 且变化较大, 模型的设定是合理的.

根据表2的结果可以对三个协变量定义的不同属性进行风险评价和估计, 例如第2条记录是A类船舶,  $Y_2$ 年建造,  $S_1$ 服务年限, 对这类船舶, 其个体层次的Poisson成分的截距与斜

率的估计值为

$$\begin{aligned}\beta_{02} &= 2.116 + 0.869x^{(1)} - 0.749x^{(2)} + 2.793x^{(3)} = 5.029, \\ \beta_{12} &= -3.472 + 0.817x^{(1)} + 1.741x^{(2)} + 0.471x^{(3)} = -0.443.\end{aligned}$$

对应的logist回归成分的 $\gamma_{02} = -1.721$ ,  $\gamma_{12} = 0.341$ , 进而可以估计这类风险分类的船舶的平均受损次数为

$$\left[1 - \frac{1}{1 + \exp(-1.721 + 0.341 \times 1.2)}\right] \exp(5.029 - 0.443 \times 1.2) = 19.051.$$

其余风险分类可类似讨论.

## 参 考 文 献

- [1] Fahrmeir, L. and Echavarría, L.O., Structured additive regression for overdispersed and zero-inflated count data, *Applied Stochastic Models in Business and Industry*, **22**(4)(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.
- [6] Afshartous, D. and Michailidis, G., Distributed multilevel modeling, *Journal of Computational and Graphical Statistics*, **16**(4)(2007), 901–924.
- [7] Gelman, A., Multilevel (hierarchical) modeling, *Technometrics*, **48**(3)(2006), 432–435.
- [8] Crainiceanu, C.M., Staicu, A.M. and Di, C.Z., Generalized multilevel functional regression, *Journal of the American Statistics Association*, **104**(488)(2009), 1550–1561.
- [9] Dunson, D.B., Bayesian nonparametric hierarchical modeling, *Biometrical Journal*, **51**(2)(2009), 273–284.
- [10] Di, C.Z. and Roche, K.B., Multilevel latent class models with dirichlet mixing distribution, *Biometrics*, **67**(1)(2011), 86–96.
- [11] Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W. and McLachlan, G.J., Multi-level zero-inflated Poisson regression modeling of correlated count data with excess zeros, *Statistical Methods in Medical research*, **15**(1)(2006), 47–61.
- [12] Gelman, A. and Rubin, G.O., Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**(4)(1992), 457–472.
- [13] Geyer, C.J., Practical Markov chain Monte Carlo, *Statistical Science*, **7**(4)(1992), 473–483.
- [14] Johnson, V.E., A Bayesian  $\chi^2$  test for goodness-of-fit, *The Annals of Statistics*, **32**(6)(2004), 2361–2384.

[15] McCullagh, P. and Nelder, J.A., *Generalized Linear Models* (2nd Edition), Chapman and Hall, 1989.

## Bayesian Inference of Hierarchical Regression Model for Zero-Inflated Clustered Count Data

SHI HONGXING

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

Zero-inflated Poisson (ZIP) regression model is a popular tool for analyzing count data with excess zeros. In this paper, a flexible hierarchical ZIP regression model is proposed to handle with such data with cluster and Bayesian approach is develop. A Gibbs sampler is employed to produce the Bayesian estimate, a goodness-of-fit and a Bayesian information criterion (BIC) are used for model comparison and selection. Finally, an application of data from a ship damage incident study illustrates the proposed method.

**Keywords:** Zero-inflation, hierarchical regression model, Gibbs sampler, BIC.

**AMS Subject Classification:** 62F15.