文章编号：1009 − 444X(2022)02 − 0212 − 06

0−1 膨胀负二项回归模型在 COVID−19 疫情分析中的应用

马巧玲，肖　翔

（上海工程技术大学 数理与统计学院，上海 201620）

摘要：在公共卫生等应用领域，经常会同时出现零观测值、一观测值较多的情况. 为更好地拟合这类数据，采用 0−1 膨胀负二项分布及其回归模型进行分析. 在数据扩充基础上，结合 Pólya−Gamma 潜变量对模型参数进行贝叶斯推断. 最后，对我国湖北省 2019 冠状病毒病 (COVID−19) 死亡数据集进行分析. 研究表明，0−1 膨胀负二项回归模型能够达到更好的拟合效果.

关键词：0−1 膨胀负二项回归模型；2019 冠状病毒病；Pólya−Gamma 潜变量；贝叶斯推断

中图分类号：O212　　　　**文献标志码：**A

Application of zero-and-one-inflated negative binomial regression model in COVID−19 epidemic analysis

MA Qiaoling，*XIAO Xiang*

（School of Mathematics，Physics and Statistics，Shanghai University of Engineering Science，Shanghai 201620，China）

Abstract：Count datas with excess zeros and ones arise frequently in the field of public health. In order to fit the kind of data, a zero-and-one-inflated negative binomial (ZOINB) distribution and its regression model were adopted for analysis. Based on data augmentation strategy and Pólya−Gamma latent variables Bayesian inference was used to estimate the parameters of ZOINB regression model. Finally, one corona virus disease 2019 (COVID−19) death data-set from Hubei Province in China was analyzed. The result illustrates that ZOINB regression model can achieve better fitting effect.

Key words：zero-and-one-inflated negative binomial (ZOINB) regression model; corona virus disease 2019 (COVID−19); Pólya−Gamma latent variables; Bayesian inference

计数数据一直是统计学研究的热点，在医疗卫生、金融证券、保险精算、工业生产等众多领域中存在着大量的计数数据. 在实际应用中，时常会遇到 0 和 1 过多 (称之为“0−1 膨胀”) 的数据样本[1]. 例如，在 2019 冠状病毒病 (COVID−19) 大流行中，个体在感染过一次 COVID−19 后，自身就会产生抗体，使得其感染的次数可能最多为一次. 又如，当前网络购物非常普遍，人们很少在实体服装店

收稿日期：2021 − 10 − 29

基金项目：全国统计科学研究项目资助(2020LY080)

作者简介：马巧玲(1998 −)，女，在读硕士，研究方向为统计学. E-mail: qiaolingma@126.com

通信作者：肖　翔(1980 −)，男，讲师，硕士，研究方向为统计学. E-mail: xiaoxiang@sues.edu.cn

买衣服, 即使在大型商场购物时, 大家的心态只是看看款式, 货比三家, 因而, 很多顾客选择不购买衣服或者只购买了一件衣服.

近年来, 国内学者对 0−1 膨胀泊松 (Zero-and-One-Inflated Poisson, ZOIP) 分布进行了深入研究, 取得丰富的研究成果. 例如, 田震[2] 研究了 0-1 膨胀回归模型及其参数估计, 并基于数据删失模型对 ZOIP 模型进行统计诊断. Tang 等[3] 引入隐变量构造 ZOIP 模型新的结构形式, 采用极大似然估计与贝叶斯方法对模型进行参数估计, 并对新加坡军团菌感染数据进行研究, 取得了较好的拟合效果. Liu 等[4] 通过重参数化的方法, 计算 ZOIP 模型中参数的 Jeffreys 先验和 reference 先验, 并进行客观贝叶斯分析, 拟合效果比使用 naive flat 先验更好. 夏丽丽等[5] 采用局部多项式核回归法对 ZOIP 模型进行参数估计, 结合 EM 算法和 Newton-Raphson 迭代法对参数进行近似求解, 最后对糖尿病患者数据的实例分析, 验证该方法的有效性. 刘娱等[6] 基于 ZOIP 模型分别构建了 Wald 检验、LR 检验以及 Score 检验的检验统计量, 并对暴风雪发生次数进行实证研究.

但在实际应用中, 有些 0−1 膨胀数据集存在很大的变异性, 这时 ZOIP 模型的拟合效果并不是很理想. 相比泊松分布及其回归模型, 负二项分布及其回归模型在模拟方差与均值之间关系时具有更大的灵活度, 可以看作是泊松分布及其回归模型的一种拓展. 因此, 有学者在负二项分布及其回归模型的基础上, 提出零膨胀负二项分布及其回归模型. Faroughi 等[7] 建立嵌套二元零膨胀负二项回归模型, 它的优势在于可以进行似然比检验来选择最佳模型, 且具有灵活的边际均值−方差形式关系, 该模型可以拟合具有正相关或负相关的二元零膨胀计数数据, 允许额外两个因变量的过度分散. Saffari 等[8] 建立右删失二元负二项模型来模拟零过多和具有极端值的计数数据, 采用共轭梯度最优的极大似然估计法对模型的参数进行估计, 该模型在估计频率的拟合优度方面表现出优越的性能. Kang 等[9] 利用零膨胀负二项模型研究韩国中学生网络欺凌行为的风险因素与预测因子, 提出预防青少年网络犯罪的防范措施.

零膨胀负二项分布及其回归模型尽管可以较好地解释零膨胀现象, 但它们还是存在一定的局限性, 不能解释"一膨胀"现象产生的内在机理. 因此, 李蒙[10] 最早对零膨胀负二项分布模型进行推广, 提出 0−1 膨胀负二项分布 (Zero-and-One-Inflated Negative Binomial, ZOINB) 及其回归模型, 用来拟合变异过大的 0-1 膨胀数据集. 然而, 近年来国内外学者关于 ZOINB 及其回归模型的研究非常少, 主要原因在于 ZOINB 回归模型的结构比较复杂, 不能设计有效的抽样算法, 导致抽样效率偏低, 实际数据的拟合效果不好. 本研究主要贡献在于: 第一, 修正了文献 [10] 中基于隐变量所构造完全似然函数的表达式. 第二, 详细阐述了基于 Pólya-Gamma 潜变量 ZOINB 回归模型中后验样本的抽样机制. 第三, 采用贝叶斯方法对模型参数进行估计, 并对 COVID−19 爆发初期湖北省死亡数据集进行统计推断, 寻找响应变量与协变量之间的关系, 为政府部门进行疫情的分析与预测提供有价值的参考依据.

1　0−1 膨胀负二项分布及其回归模型

设一个非负随机变量 Y 表示为 $Y=V(1-B_1)+B_1(1-B_2)$, 其中, B_1、B_2、V 相互独立, B_1 服从成功概率为 p_1 的伯努利分布, B_2 服从成功概率为 p_2 的伯努利分布, V 服从参数为 θ 的负二项分布, 即

$$P(V=k)=\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}\theta^k(1-\theta)^r,\ \ k=0,1,\cdots$$

则 Y 与 (B_1,B_2,V) 之间的关系具体表现为

$$\begin{cases}(Y=0)\Leftrightarrow(V=0,B_1=0)\cup(B_1=1,B_2=1)\\(Y=1)\Leftrightarrow(V=1,B_1=0)\cup(B_1=1,B_2=0)\\(Y=k)\Leftrightarrow(V=k,B=0),\ \ k=2,3,\cdots\end{cases}\quad(1)$$

相应的分布律为

$$P(Y=k)=\begin{cases}p_1p_2+(1-p_1)(1-\theta)^r, & k=0\\p_1(1-p_2)+(1-p_1)r\theta(1-\theta)^r, & k=1\\(1-p_1)\dfrac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}\theta^k(1-\theta)^r, & k=2,3,\cdots\end{cases}\quad(2)$$

式 (2) 为 0−1 膨胀负二项分布模型, 记为 $Y\sim\text{ZOINB}(p_1,p_2,\theta)$. 其中, $0\leqslant p_1\leqslant 1$, $0\leqslant p_2\leqslant 1$, $0\leqslant\theta\leqslant 1$[10]. 可知, p_1 和 $1-p_1$ 分别为一个伯努利分布与一个负二项分布的混合比例. 特别地, 当 $p_2=1$ 时, ZOINB 分布变成零膨胀负二项分布; 当 $p_1=0$, ZOINB 分布退化成传统意义下的负二项分布. 当 $r=1$

时，0−1 膨胀负二项分布模型变成了 0−1 膨胀几何分布模型[11 − 12].

设$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)$是来自 0−1 膨胀负二项分布容量为$n$的观测值，则$(p_1,p_2,\theta)$的似然函数为

$$
\begin{aligned}
L(p_1,p_2,\theta|\boldsymbol{Y}) &= [p_1p_2+(1-p_1)(1-\theta)^r]^{S_0}\times \\
&[p_1(1-p_2)+(1-p_1)r\theta(1-\theta)^r]^{S_1}\times \\
&\left[(1-p_1)\frac{\Gamma(k+r)}{\Gamma(k+1)\Gamma(r)}\right]^{n-S_0-S_1}\times \\
&\theta^S(1-\theta)^{r(n-S_0-S_1)}
\end{aligned}
\tag{3}
$$

式中：S_0为$\{i:Y_i=0\}$中元素的个数；S_1为$\{i:Y_i=1\}$中元素的个数；$S=\sum_{Y_i\geqslant 2}Y_i$.

式 (3) 结构复杂，不利于后续研究，因此，结合式 (1) 中的隐变量$\boldsymbol{B}_1=(B_{11},B_{12},\cdots,B_{1n})$、$\boldsymbol{B}_2=(B_{21},B_{22},\cdots,B_{2n})$、$\boldsymbol{V}=(V_1,V_2,\cdots,V_n)$，构建数据扩充下$(p_1,p_2,\theta)$的完全似然函数，并对文献 [10] 中相应完全似然函数的表达式进行修正，公式为

$$
\begin{aligned}
L(p_1,p_2,\theta|\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V}) &= \\
&\prod_{i=1}^{n}[p_1p_2^{B_{2i}}(1-p_2)^{1-B_{2i}}]^{B_{1i}} \\
&\left[(1-p_1)\frac{\Gamma(V_i+r)}{\Gamma(V_i+1)\Gamma(r)}\theta^{V_i}(1-\theta)^r\right]^{1-B_{1i}} = \\
&\prod_{i=1}^{n}p_1^{B_{1i}}(1-p_1)^{1-B_{1i}}p_2^{B_{1i}B_{2i}}(1-p_2)^{B_{1i}(1-B_{2i})}\times \\
&\left[\frac{\Gamma(V_i+r)}{\Gamma(V_i+1)\Gamma(r)}\theta^{V_i}(1-\theta)^r\right]^{1-B_i}
\end{aligned}
\tag{4}
$$

设相互独立的响应变量$Y_i\sim \mathrm{ZOINB}(p_1,p_{2i},\theta_i)$，$i=1,2,\cdots,n$，将模型 (2) 中的参数向量$\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_n)$和$\boldsymbol{p}_2=(p_{21},p_{22},\cdots p_{2n})$与协变量矩阵$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$、$\boldsymbol{Z}=(z_1,z_2,\cdots,z_n)$建立连接关系为

$$
\begin{cases}
\ln\dfrac{r\theta_i}{1-\theta_i}=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\
\mathrm{logit}(p_{2i})=\ln\dfrac{p_{2i}}{1-p_{2i}}=z_i^{\mathrm{T}}\boldsymbol{\gamma}
\end{cases}
\tag{5}
$$

式中：$\boldsymbol{x}_i=(x_{i0},x_{i1},\cdots,x_{i(s-1)})$为一个长度为$s$的向量，且$x_{i0}=1$；$z_i=(z_{i0},z_{i1},\cdots,z_{i(t-1)})$为一个长度为$t$的向量，且$z_{i0}=1$；$\boldsymbol{\beta}$与$\boldsymbol{\gamma}$分别为$\boldsymbol{x}_i$与$z_i$−1 膨胀负二项回归模型.

2 贝叶斯估计

对式 (5) 进行变形得到

$$
\begin{cases}
\theta_i=\dfrac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})} \\
p_{2i}=\dfrac{\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})}{1+\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})}
\end{cases}
\tag{6}
$$

其中，$\tilde{\beta}_0=\beta_0-\ln r$，$\tilde{\beta}_j=\beta_j$，$j=1,2,\cdots,s-1$. 将式 (6) 代入式 (5)，得到数据扩充下$(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1)$的完全似然函数

$$
\begin{aligned}
&L(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1|\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\propto \\
&\prod_{i=1}^{n}\frac{\left\{\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{V_i(1-B_{1i})}}{\left\{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{(V_i+r)(1-B_{1i})}}\times \\
&\prod_{i=1}^{n}\frac{\left\{\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}B_{2i}}}{\left\{1+\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}}}\times \\
&p_1^{\sum_{i=1}^{n}B_{1i}}(1-p_1)^{n-\sum_{i=1}^{n}B_{1i}}
\end{aligned}
\tag{7}
$$

假设参数向量$\tilde{\boldsymbol{\beta}}$和$\boldsymbol{\gamma}$的先验分布为多维正态分布，即$\tilde{\boldsymbol{\beta}}\sim N_s(\mu_{\tilde{\beta}},\sigma_{\tilde{\beta}}^2I_s)$，$\boldsymbol{\gamma}\sim N_t(\mu_\gamma,\sigma_\gamma^2I_t)$，其中，$\boldsymbol{\mu}_{\tilde{\beta}}$和$\boldsymbol{\mu}_{\gamma}$是已知向量，$\sigma_{\tilde{\beta}}^2$和$\sigma_\gamma^2$是已知常数. 另外，参数$p_1$服从区间$[0,1]$上的均匀分布，即$p_1\sim U(0,1)$. 进一步假设$\tilde{\boldsymbol{\beta}}$、$\boldsymbol{\gamma}$和$p_1$相互独立，则$(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1)$的联合先验分布$\pi(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1)=\pi(\tilde{\boldsymbol{\beta}})\pi(\boldsymbol{\gamma})\pi(p_1)$. 结合式 (7)，得到数据扩充下$(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1)$的后验分布为

$$
\begin{aligned}
&\pi(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1|\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\propto \\
&\quad L(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1|\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\pi(\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},p_1)\propto \\
&\prod_{i=1}^{n}\frac{\left\{\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{V_i(1-B_{1i})}}{\left\{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{(V_i+r)(1-B_{1i})}}\times \\
&\prod_{i=1}^{n}\frac{\left\{\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}B_{2i}}}{\left\{1+\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}}}\times \\
&p_1^{\sum_{i=1}^{n}B_{1i}}(1-p_1)^{n-\sum_{i=1}^{n}B_{1i}}\times\pi(\tilde{\boldsymbol{\beta}})\pi(\boldsymbol{\gamma})\pi(p_1)
\end{aligned}
\tag{8}
$$

由式 (8) 可得各参数的满条件分布式为

$$
\begin{aligned}
&\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\propto \\
&\pi(\tilde{\boldsymbol{\beta}})\prod_{i=1}^{n}\frac{\left\{\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{V_i(1-B_{1i})}}{\left\{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{(V_i+r)(1-B_{1i})}}
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma}|\tilde{\boldsymbol{\beta}},p_1,Y,B_1,B_2,V)\propto \\
&\pi(\boldsymbol{\gamma})\prod_{i=1}^{n}\frac{\left\{\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}B_{2i}}}{\left\{1+\exp(z_i^{\mathrm{T}}\boldsymbol{\gamma})\right\}^{B_{1i}}}
\end{aligned}
\tag{10}
$$

$$\pi(p_1|\tilde{\boldsymbol{\beta}},\boldsymbol{\gamma},\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\propto \pi(p_1)p_1^{\sum_{i=1}^{n}B_{1i}}(1-p_1)^{n-\sum_{i=1}^{n}B_{1i}} \tag{11}$$

满条件分布式 (9) 和式 (10) 并不是常见的分布, 李蒙[10] 直接利用 Metropolis−Hastings 方法进行抽样, 但抽样效率低, 效果不尽如人意. 本研究引入 Pólya−Gamma 潜变量, 结合 Pólya−Gamma 分布和条件高斯分布进行抽样, 得到高效率的后验样本, 具体方法如下.

引 理[13]　设 $p(\omega)$为 Pólya−Gamma 分布 $\mathrm{PG}(b,0)$, $(b>0)$ 的概率密度函数, 对于任意的实数 $a\in R$, 有

$$\frac{\{\exp(\psi)\}^a}{\{1+\exp(\psi)\}^b}=2^{-b}\exp(\kappa\psi)\int_0^{\infty}\exp\left(-\frac{\omega\psi^2}{2}\right)p(\omega)\mathrm{d}\omega \tag{12}$$

其中, $\kappa=a-\dfrac{b}{2}$.

将上述引理运用于式 (9) 得到

$$\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V})\propto \pi(\tilde{\boldsymbol{\beta}})\prod_{i=1}^{n}\frac{\left\{\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{V_i(1-B_{1i})}}{\left\{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\right\}^{(V_i+r)(1-B_{1i})}}\propto \pi(\tilde{\boldsymbol{\beta}})\prod_{i=1}^{n}\exp(\tilde{\kappa}_i\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})\int_0^{\infty}\exp\left(-\frac{\omega_i(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})^2}{2}\right)p(\omega_i)\mathrm{d}\omega_i \tag{13}$$

式中: $\tilde{\kappa}_i=\dfrac{(V_i-r)(1-B_{1i})}{2}$, $\omega_i\sim\mathrm{PG}((V_i+r)(1-B_{1i}),0)$.

记 $\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_n)$为 Pólya−Gamma 变量, 若$\omega_i$已经从Pólya−Gamma 分布$\mathrm{PG}((V_i+r)(1-B_{1i}),0)$抽样得到, 对于给定的$\omega$, 有

$$\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V},\omega)\propto \pi(\tilde{\boldsymbol{\beta}})\prod_{i=1}^{n}\exp\left\{\tilde{\kappa}_i\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}-\frac{\omega_i(\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}})^2}{2}\right\}\propto \pi(\tilde{\boldsymbol{\beta}})\prod_{i=1}^{n}\exp\left\{-\frac{\omega_i}{2}\left(\frac{\tilde{\kappa}_i}{\omega_i}-\boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}\right)^2\right\}\propto \pi(\tilde{\boldsymbol{\beta}})\exp\left\{-\frac{1}{2}(\lambda-X\tilde{\boldsymbol{\beta}})^{\mathrm{T}}\Omega(\lambda-X\tilde{\boldsymbol{\beta}})\right\} \tag{14}$$

式中: $\boldsymbol{\lambda}=\left(\dfrac{\tilde{\kappa}_1}{\omega_1},\dfrac{\tilde{\kappa}_2}{\omega_2},\cdots,\dfrac{\tilde{\kappa}_n}{\omega_n}\right)$, $\boldsymbol{\Omega}=\mathrm{diag}(\omega_1,\omega_2,\cdots,\omega_n)$, $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$.

由式 (14) 可以看出, $\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V},\omega)$为条件高斯分布. 因此, 由式 (9) 抽样得到$\tilde{\boldsymbol{\beta}}$的后验样本, 公式为

$$\omega_i\sim\mathrm{PG}((V_i+r)(1-B_{1i}),0),\quad i=1,2,\cdots,n \tag{15}$$

$$\tilde{\boldsymbol{\beta}}|\boldsymbol{\gamma},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V},\omega\sim N(\tilde{\boldsymbol{M}}_\omega,\tilde{\boldsymbol{H}}_\omega) \tag{16}$$

其中

$$\tilde{\boldsymbol{H}}_\omega=(\boldsymbol{X}^{\mathrm{T}}\Omega\boldsymbol{X}+\sigma_{\tilde{\beta}}^{-2}\boldsymbol{I}_s^{-1})^{-1}$$

$$\tilde{\boldsymbol{M}}_\omega=\tilde{\boldsymbol{H}}_\omega(\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{\kappa}}+\sigma_{\tilde{\beta}}^{-2}\boldsymbol{I}_s^{-1}\mu_{\tilde{\beta}})$$

$$\tilde{\boldsymbol{\kappa}}=\left(\frac{(V_1-r)(1-B_{11})}{2},\frac{(V_2-r)(1-B_{12})}{2},\cdots,\frac{(V_n-r)(1-B_{1n})}{2}\right)$$

同理, 由式 (10) 中抽样得到$\boldsymbol{\gamma}$的后验样本, 公式为

$$\omega_i\sim PG(B_{1i},0),\quad i=1,2,\cdots,n \tag{17}$$

$$\boldsymbol{\gamma}|\tilde{\boldsymbol{\beta}},p_1,\boldsymbol{Y},\boldsymbol{B}_1,\boldsymbol{B}_2,\boldsymbol{V},\omega\sim N(\boldsymbol{M}_\omega,\boldsymbol{H}_\omega) \tag{18}$$

其中

$$\boldsymbol{H}_\omega=(\boldsymbol{Z}^{\mathrm{T}}\Omega\boldsymbol{Z}+\sigma_{\gamma}^{-2}\boldsymbol{I}_t^{-1})^{-1}$$

$$\boldsymbol{M}_\omega=\boldsymbol{H}_\omega(\boldsymbol{Z}^{\mathrm{T}}\kappa+\sigma_{\gamma}^{-2}\boldsymbol{I}_t^{-1}\mu_{\gamma})$$

$$\kappa=\left(B_{11}B_{21}-\frac{B_{11}}{2},B_{12}B_{22}-\frac{B_{12}}{2},\cdots,B_{1n}B_{2n}-\frac{B_{1n}}{2}\right)$$

综上, 设计 Gibbs 抽样机制, 对后验分布式 (8) 进行抽样, 具体步骤如下.

1) 设定参数初始值$\tilde{\beta}^{(0)},\gamma^{(0)},p_1{}^{(0)}$.

2) 对$t=1,2,\cdots$进行迭代更新:

(a) 给定 $\tilde{\boldsymbol{\beta}}^{(t-1)},\boldsymbol{\gamma}^{(t-1)}$, 由式 (6) 得到$\theta_i^{(t-1)},p_{2i}^{(t-1)}$;

(b) 利用$(p_{1i}^{(t-1)},p_{2i}^{(t-1)},\theta_i^{(t-1)})$的条件分布[3], 得到样本$(B_{1i}^{(t)},B_{2i}^{(t)},V_i^{(t)})$, $i=1,2,\cdots,n$;

(c) 通过式 (15) 和式 (16), 借助 R 软件中 BayesLogit 程序包, 抽样得到 $\tilde{\boldsymbol{\beta}}^{(t)}$;

(d) 通过式 (17) 和式 (18), 借助 R 软件中 BayesLogit 程序包, 抽样得到$\boldsymbol{\gamma}^{(t)}$;

(e) 借助 R 软件, 从贝塔分布 $\mathrm{Beta}\left(1+\sum_{i=1}^{n}B_{1i}^{(t)},n+1-\sum_{i=1}^{n}B_{1i}^{(t)}\right)$ 抽样得到 $p_{1i}^{(t)}$.

3　数值模拟

本节通过数值模拟, 对 0−1 膨胀负二项回归模

型进行参数估计. 假设回归模型是一元线性的，公式为

$$\begin{cases}\ln\dfrac{\theta_i}{1-\theta_i}=\tilde{\beta}_0+\beta_1x_{i1}\\ \mathrm{logit}(p_{2i})=\gamma_0+\gamma_1z_{i1},\ i=1,2,\cdots,n\end{cases}$$

式中：$\tilde{\beta}_0=\beta_0-\ln r$，$x_{i1}$由成功概率为 0.5 的贝努利分布产生，$z_{i1}$由标准正态分布产生. 设置$r=3$，$\tilde{\boldsymbol{\beta}}=(2-\ln3,\ 1.5)$，$\boldsymbol{\gamma}=(1,\ 2)$，样本容量分别为 50 和 100. 对于先验分布的超参数，假设$\mu_{\tilde{\beta}}=\mu_{\gamma}=(0,0)$，$\sigma^2_{\tilde{\beta}}=\sigma^2_{\gamma}=100$，每次模拟重复 1 000 次. 参数估计量的均值、中位数、均方误差和置信区间覆盖率，见表 1. 可以看出，随着样本容量的增加，参数的估计值越来越接近真值，误差也在不断地减少.

表 1　ZOINB 回归模型的参数估计

Table 1　Parameter estimation of ZOINB regression model

样本容量	统计量	p_1	$\tilde{\beta}_0$	β_1	γ_0	γ_1
50	均值	0.288 7	0.888 6	1.432 5	0.978 9	1.940 4
	中位数	0.290 1	0.888 1	1.461 3	0.961 3	1.956 7
	均方误差	0.003 5	0.003 1	0.041 9	0.042 4	0.050 4
	覆盖率	0.960 2	0.953 2	0.942 1	0.933 2	0.953 1
100	均值	0.292 7	0.891 3	1.491 5	0.991 5	1.973 1
	中位数	0.296 5	0.894 8	1.491 1	0.981 1	1.927 3
	均方误差	0.002 3	0.001 3	0.023 4	0.021 3	0.019 3
	覆盖率	0.954 1	0.948 2	0.949 2	0.950 3	0.950 4

4　实例分析

从湖北省卫生健康委员会官方网站上获得 2020 年 1 月 23 日至 2 月 28 日，湖北省除武汉市之外其他 30 个城市 COVID−19 死亡病例数，运用 ZOINB 回归模型对 COVID−19 死亡数据集进行分析，如图 1 所示. 图中横坐标为每个城市死亡病例数，纵坐标为城市数. 由图可见，死亡人数为 0 或 1 的城市很多，数据集出现 0−1 膨胀现象. 考虑以下 4 个协变量：x_1为该城市距离武汉市最短的空间距离；x_2为该城市铁路网密度，即该城市铁路总路线除以城市的总面积；x_3为乘客密度，即该城市乘客总数除以城市的总人口数；x_4为人均病床数，即该城市医院总床位数除以城市的总人口数.

令协变量矩阵$\boldsymbol{X}=(1,x_1,x_2,x_3,x_4)$，$\boldsymbol{Z}=(1,z_1,z_2)=(1,x_2,x_3)$，$\boldsymbol{\mu}_{\tilde{\beta}}=(0,0,0,0,0)$，$\boldsymbol{\mu}_{\gamma}=(0,0,0)$，$\sigma^2_{\tilde{\beta}}=\sigma^2_{\gamma}=100$，$r$分别为2,3,4,5. ZOINB 回归模型中的预测项采用线性形式，即

$$\begin{cases}\ln\dfrac{\theta_i}{1-\theta_i}=\tilde{\beta}_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}\\ \mathrm{logit}(p_{2i})=\gamma_0+\gamma_1z_{i1}+\gamma_2z_{i2},\ \ i=1,2,\cdots,n\end{cases}$$

参数向量的估计结果见表 2. 观测频数与拟合频数的比较见表 3 和如图 1 所示.

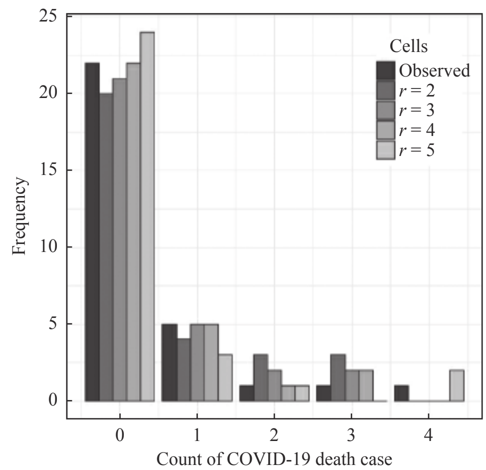

图 1　ZOINB 回归模型中 COVID−19 死亡数据的观测频数与拟合频数

Fig. 1　Observation frequency and fitted frequency for COVID−19 death data in ZOINB regression model

从表 2 可以看出，β_2、β_3为正数，说明x_2、x_3分别与 COVID−19 死亡人数呈正相关；β_1、β_4为负数，说明x_1、x_4分别与 COVID−19 死亡人数呈负相关，这与现实情况是吻合的. 另根据表 2 中赤池信息量准则（Akaike Information Criterion, AIC）值和表 3 中观测频数与拟合频数的接近程度，当

表 2　ZOINB 回归模型中参数估计均值的比较

Table 2　Comparison of parameter estimation mean in ZOINB regression model

参数	$r=2$	$r=3$	$r=4$	$r=5$
$\tilde{\beta}_0$	0.616 6	0.615 3	0.614 4	0.613 8
β_1	−0.472 2	−0.453 6	−0.443 4	−0.476 6
β_2	0.346 5	0.345 8	0.350 1	0.351 1
β_3	0.230 1	0.231 5	0.233 4	0.233 7
β_4	−1.260 5	−1.282 5	−1.333 2	−1.353 6
γ_0	0.065 2	0.074 4	0.075 8	0.069 7
γ_1	0.145 7	0.418 7	0.502 8	0.558 4
γ_2	0.348 7	0.366 3	0.382 8	0.408 7
AIC	1 536.314	1 537.843	1 526.173	1 548.801

表 3　ZOINB 回归模型中的观测频数与拟合频数

Table 3　Comparison of observation frequency and fitted frequency in ZOINB regression model

观测值	观测频数	拟合频数			
		$r=2$	$r=3$	$r=4$	$r=5$
0	22	20	21	22	24
1	5	4	5	5	3
2	1	3	2	1	1
3	1	3	2	2	0
4	1	0	0	0	2

$r=4$时, ZOINB 回归模型拟合效果最好.

5　结　语

本研究针对 0−1 膨胀变异过大的数据集, 提出 0−1 膨胀负二项回归模型, 在数据扩充的基础上, 通过隐变量的条件分布, 对样本数据的膨胀部分进行解释, 并将复杂的似然函数形式转化为简单的表达形式, 巧妙地引入 Pólya−Gamma 潜变量, 在贝叶斯推断中获得效率更高的后验样本, 实现更好的拟合结果.

参考文献:

[1]　张良超, 周金亮, 温利民. 零膨胀泊松模型中风险参数的贝叶斯估计 [J]. **江西师范大学学报 (自然科学版)**, 2020, 44(3): 269 − 274.

[2]　田震. 零—膨胀回归模型及其统计诊断 [D]. 昆明: 云南大学, 2016.

[3]　TANG Y C, LIU W C, XU A C. Statistical inference for zero-and-one-inflated Poisson models [J]. **Statistics Theory and Related Fields**, 2017, 1(2): 216 − 226.

[4]　LIU W C, TANG Y C, XU A C. A zero-and-one inflated Poisson model and its application [J]. **Statistics and Its Interface**, 2018, 11(2): 339 − 351.

[5]　夏丽丽, 田茂再. 零—膨胀泊松回归模型的非参数统计分析及其应用 [J]. **数理统计与管理**, 2019, 38(2): 235 − 246.

[6]　刘娱, 安博文, 田茂再. 零—膨胀泊松模型的似然检验及模型比较 [J]. **统计与决策**, 2021, 37(577): 20 − 24.

[7]　FAROUGHI P, ISMAIL N. Bivariate zero-inflated negative binomial regression model with applications [J]. **Journal of Statistical Computation and Simulation**, 2017, 87(3): 457 − 477.

[8]　SAFFARI S E, ALLEN J C. Bivariate negative binomial regression model with excess zeros and right censoring: an application to Indonesian data [J]. **Journal of Applied Statistics**, 2020, 47(10): 1901 − 1914.

[9]　KANG K I, KANG K, KIM C. Risk factors influencing cyberbullying perpetration among middle school students in Korea: Analysis using the zero-inflated negative binomial regression model [J]. **International Journal of Environmental Research and Public Health**, 2021, 18(5): 2224 − 2224.

[10]　李蒙. 0−1膨胀负二项模型及其统计分析 [D]. 上海: 华东师范大学, 2018.

[11]　肖翔. 0−1膨胀几何分布回归模型及其应用 [J]. **系统科学与数学**, 2019, 39(9): 1486 − 1499.

[12]　XIAO X, TANG Y C, XU A C, et al. Bayesian inference for zero-and-one-inflated geometric distribution regression model using Pólya-Gamma latent variables [J]. **Communication in Statistics-Theory and Method**, 2020, 49(15): 3730 − 3743.

[13]　NICHOLAS G P, JAMES G S, JESSE W. Bayesian inference for logistic models using Pólya−Gamma latent variables [J]. **Journal of the American Statistical Association**, 2013, 108(504): 1339 − 1349.

（编辑: 韩琳）