马巧玲,肖翔. 零一膨胀负二项模型的客观贝叶斯分析[J]. 江西师范大学学报(自然科学版),2023,47(1):8-14.
MA Qiaoling,XIAO Xiang. The objective Bayesian analysis of zero-and-one inflated negative binomial model[J]. Journal of Jiangxi Normal University(Natural Science),2023,47(1):8-14.

文章编号:1000-5862(2023)01-0008-07

零一膨胀负二项模型的客观贝叶斯分析

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: 该文建立了贝叶斯模型,对零一膨胀负二项分布进行了客观贝叶斯估计. 采用数据增广策略,基于完全似然函数,推导出零一膨胀负二项模型不同形式的 reference 先验,进一步证明了相应的后验分布是恰当的. 在不同的样本量和不同的参数真值下,对 3 种 reference 先验的性能进行仿真与评估. 最后,对于生物化学博士生发表论文数量的数据集,零一膨胀负二项模型能够达到较好的拟合效果.

关键词: 零一膨胀负二项模型; 数据增广策略; 客观贝叶斯
中图分类号: O 212　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2023.01.02

0　引言

计数数据广泛应用于保险精算、道路交通安全、传染病防控、气象灾害预测等领域,一直以来都是统计学研究的前沿领域和热点问题. 泊松模型和负二项模型是用来处理计数数据的常用统计模型,泊松分布往往要求总体均值与总体方差相等,而负二项分布对总体均值与总体方差的关系没有限制,在模型选择上更加灵活.

在日常生活中,人们经常会遇到数据零过多的样本数据,称之为“零膨胀”现象,这时若仍然用单个分布模型去分析,则会出现较大的偏差,拟合效果不好. 零膨胀模型是一种特殊的混合模型,研究成果十分丰富. 如 D. Lambert[1] 对传统的泊松回归模型进行了推广,首先提出了零膨胀泊松回归模型,并用于对在工业品焊接中出现的瑕疵数据进行分析. D. Bohning等[2] 通过对龋病数据的分析,证明了零膨胀泊松回归模型比传统的泊松回归模型能更好地拟合数据. He Xuming 等[3] 运用 Seive 极大似然估计法对零膨胀泊松模型的参数进行估计. 张良超等[4] 采用平方损失函数、Linex 损失函数和 Stein 损失函数,对在零膨胀泊松模型中的风险参数进行贝叶斯估计,并对估计的收敛性进行了比较.

在某些特定的场合下也会产生数据 0 和数据 1 同时过多的样本数据,如在新冠疫情常态化防控下,随着新冠疫苗接种的普及和人们防护意识的提高,绝大多数个体最多感染 1 次新冠肺炎病毒. 再如,在健康中国战略下,上海市建立了多层次的养老保障体系,老年人的生活质量与健康水平进一步提高,大部分老年人在 1 年内住院的次数可能最多只有 1 次. 因此,国内外学者将零膨胀泊松模型推广到零一膨胀泊松(ZOIP)模型. 如 Tang Yincai 等[5] 引入隐变量,构建了 ZOIP 模型的新型结构,采用极大似然估计法和贝叶斯估计法对参数进行估计,并对新加坡军团菌的感染数据集进行拟合分析. Liu Wenchen 等[6] 对 ZOIP 模型进行客观贝叶斯估计,推导出 Jeffreys 先验和 reference 先验,并证明它们是 2 阶概率匹配先验. 夏丽丽等[7] 采用局部多项式核回归方法对 ZOIP 模型进行参数估计,通过对糖尿病患者的数据分析,验证了该方法的有效性. 肖翔[8] 针对 ZOIP 模型,基于数据扩充策略和完全似然函数,推导出不同形式的客观先验,证明了后验分布的恰当性.

对于零一膨胀数据集,若非零部分数据过于分散,变异较大,则人们往往倾向于选择更灵活的零一膨胀负二项(ZOINB)模型来进行统计建模. 目前,学者们主要集中在对零膨胀负二项模型的研究[9-11],

收稿日期:2022-10-06
基金项目:国家自然科学基金(62072296)和全国统计科学研究课题(2020LY080)资助项目.
通信作者:肖　翔(1980—),男,江西樟树人,讲师,主要从事贝叶斯统计的研究. E-mail:xiaoxiang@sues.edu.cn

而对零一膨胀负二项模型的研究成果相对较少. 李蒙[12] 建立了零一膨胀负二项分布及其回归模型,采用在 EM 算法下极大似然估计法和朴素贝叶斯估计方法对模型参数进行估计. 本文基于数据增广的完全似然函数,将 Fisher 信息矩阵写成对角矩阵(或分块对角矩阵)的形式,为计算客观先验带了极大的方便,对参数的估计结果比文献[12]中的结果更加稳健和有效.

1　零一膨胀负二项模型

设非负随机变量 Z 表示为 $Z = X(1-B) + B(1-Y)$,且 X、Y、B 相互独立. 随机变量 X 服从成功概率为 θ、失败次数为 r 的负二项分布 $NB(r,\theta)$,即 $P(X=k) = C_{k+r-1}^{k}\theta^{k}(1-\theta)^{r}, k=0,1,\cdots$. 随机变量 B 服从成功概率为 p 的伯努利分布 Bernoulli(p),而随机变量 Y 服从成功概率为 q 的另一个伯努利分布 Bernoulli(q). Z 与 (X,Y,B) 的关系为

$$\begin{cases}(Z=0)\Leftrightarrow(X=0,B=0)\cup(B=1,Y=1),\\(Z=1)\Leftrightarrow(X=1,B=0)\cup(B=1,Y=0),\\(Z=k)\Leftrightarrow(X=k,B=0),k=2,3,\cdots.\end{cases}$$

相应的概率分布为

$$P(Z=k)=\begin{cases}pq+(1-p)(1-\theta)^{r}, k=0,\\p(1-q)+(1-p)r\theta(1-\theta)^{r}, k=1,\\(1-p)C_{k+r-1}^{k}\theta^{k}(1-\theta)^{r}, k=2,3,\cdots.\end{cases}\tag{1}$$

称式(1)为零一膨胀负二项模型,记为 $Z\sim \mathrm{ZOINB}(p,q,\theta)$,其中 $0\leqslant p\leqslant 1, 0\leqslant q\leqslant 1, 0\leqslant\theta\leqslant 1$,且 p 和 $1-p$ 分别是伯努利分布 Bernoulli(q) 与负二项分布 $NB(r,\theta)$ 的混合比例. 当 $q=1$ 时,ZOINB 模型变成零膨胀负二项(ZINB)模型; 当 $p=0$ 时,ZOINB 模型退化的负二项模型. 特别地,当 $r=1$ 时,ZOINB 模型变成了零一膨胀几何分布模型[13-14].

设 $Z=(Z_1,Z_2,\cdots,Z_n)$ 是来自 ZOINB 模型容量为 n 的样本观测值,则 (p,q,θ) 的似然函数为

$$L(p,q,\theta\mid Z) = (pq+(1-p)(1-\theta)^{r})^{N_0}(p(1-q)+(1-p)r\theta(1-\theta)^{r})^{N_1}((1-p)C_{k+r-1}^{k})^{n-N_0-N_1}\cdot\theta^{N}(1-\theta)^{r(n-N_0-N_1)},\tag{2}$$

其中 $N=\sum_{Z_i\geqslant 2}Z_i$,$N_0$ 是在 $\{i: Z_i=0\}$ 中元素的个数,N_1 是在 $\{i: Z_i=1\}$ 中元素的个数.

若采用似然函数式(2)进行客观贝叶斯分析,则推导 reference 先验的过程会十分复杂和冗长. 本节将隐变量 $X=(X_1,X_2,\cdots,X_n)$、$Y=(Y_1,Y_2,\cdots,Y_n)$、$B=(B_1,B_2,\cdots,B_n)$ 进行数据增广,构造出 (p,q,θ) 的完全似然函数:

$$L(p,q,\theta\mid Z,X,Y,B) = \prod_{i=1}^{n}(pq^{Y_i}(1-q)^{1-Y_i})^{B_i}\cdot((1-p)C_{X_i+r-1}^{X_i}\theta^{X_i}(1-\theta)^{r})^{1-B_i} = \prod_{i=1}^{n}p^{B_i}(1-p)^{1-B_i}q^{B_iY_i}(1-q)^{B_i(1-Y_i)}(C_{X_i+r-1}^{X_i}\theta^{X_i}(1-\theta)^{r})^{1-B_i}.\tag{3}$$

相应的对数完全似然函数为

$$l(p,q,\theta\mid Z,X,Y,B) = \sum_{i=1}^{n}(B_i\ln p+(1-B_i)\cdot\ln(1-p)) + \sum_{i=1}^{n}B_i(Y_i\ln q+(1-Y_i)\ln(1-q)) + \sum_{i=1}^{n}(1-B_i)(X_i\ln\theta+r\ln(1-\theta)) + \sum_{i=1}^{n}(1-B_i)\cdot\ln(C_{X_i+r-1}^{X_i}).\tag{4}$$

2　客观贝叶斯估计

2.1　Fisher 信息矩阵

定理 1　对于 ZOINB 模型,基于隐变量 (X,Y,B),$\Psi=(p,q,\theta)$ 的 Fisher 信息矩阵为

$$\boldsymbol{I}=\begin{pmatrix}\frac{n}{p(1-p)} & 0 & 0\\ 0 & \frac{np}{q(1-q)} & 0\\ 0 & 0 & \frac{nr(1-p)}{\theta(1-\theta)^{2}}\end{pmatrix}.$$

证　对数完全似然函数式(4)分别关于 p、q 和 θ 的 1 阶偏导数及 2 阶偏导数为

$$\partial l/\partial p = \sum_{i=1}^{n}(B_i/p-(1-B_i)/(1-p)),$$

$$\partial l/\partial q = \sum_{i=1}^{n}B_i(Y_i/q-(1-Y_i)/(1-q)),$$

$$\partial l/\partial\theta = \sum_{i=1}^{n}(1-B_i)(X_i/\theta-r/(1-\theta)),$$

$$\partial^2 l/\partial p^2 = -\sum_{i=1}^{n}(B_i/p^2+(1-B_i)/(1-p)^2),$$

$$\partial^2 l/\partial q^2 = -\sum_{i=1}^{n}B_i(Y_i/q^2+(1-Y_i)/(1-q)^2),$$

$$\partial^2 l/\partial\theta^2 = -\sum_{i=1}^{n}(1-B_i)(X_i/\theta^2+r/(1-\theta)^2),$$

$$\partial^2 l/(\partial p\partial q) = \partial^2 l/(\partial q\partial p) = 0,$$

$$\partial^2 l/(\partial p\partial\theta) = \partial^2 l/(\partial\theta\partial p) = 0,$$

$$\partial^2 l/(\partial q\partial\theta) = \partial^2 l/(\partial\theta\partial q) = 0.$$

由于隐变量的分布是常见分布,即 $B_i\sim$

Bernoulli(p),$Y_i \sim$ Bernoulli(q),$X_i \sim NB(r,\theta)$,它们的数学期望分别为 $E(B_i)=p$,$E(Y_i)=q$,$E(X_i)=r\theta/(1-\theta)$,$i=1,2,\cdots,n$,所以较为便捷地计算出 Fisher 信息矩阵的各个元素

$$I_{11}=-E(\partial^2 l/\partial p^2)=n/(p(1-p)),$$
$$I_{22}=-E(\partial^2 l/\partial q^2)=np/(q(1-q)),$$
$$I_{33}=-E(\partial^2 l/\partial \theta^2)=nr(1-p)/(\theta(1-\theta)^2),$$
$$I_{12}=I_{21}=I_{13}=I_{31}=I_{23}=I_{32}=0.$$

因此 $\Psi=(p,q,\theta)$ 的 Fisher 信息矩阵为 $\boldsymbol{I}$,这是一个对角矩阵,为后续计算 reference 先验带来了极大的方便.

2.2 reference 先验

J. M. Bernardo[15] 将参数分为感兴趣参数和讨厌参数,当感兴趣参数的先验分布与其后验分布的 Kullback-Liebler 距离达到最大时的先验分布被称为 reference 先验. 统计模型的参数往往有多个,根据参数之间重要性的不同,reference 先验的形式也不同. 本节列举了 7 种 reference 先验,并选择 2 种有代表性参数组合来计算出它们的 reference 先验.

定理 2 对于 ZOINB 模型,$\Psi=(p,q,\theta)$ 的 reference 先验如下:

(a) 当参数组合为$\{(p,q),\theta\}$ 和$\{\theta,(p,q)\}$时,Ψ 的 reference 先验为

$$\pi_{R_1}\propto(1-p)^{-1/2}q^{-1/2}(1-q)^{-1/2}\theta^{-1/2}(1-\theta)^{-1}.$$

(b) 当参数组合为$\{(p,\theta),q\}$ 和$\{q,(p,\theta)\}$时,Ψ 的 reference 先验为

$$\pi_{R_2}\propto p^{-1/2}q^{-1/2}(1-q)^{-1/2}\theta^{-1/2}(1-\theta)^{-1}.$$

(c) 当参数组合为$\{(q,\theta),p\}$,$\{p,(q,\theta)\}$ 和$\{p,q,\theta\}$ 时,Ψ 的 reference 先验为

$$\pi_{R_3}\propto p^{-1/2}(1-p)^{-1/2}q^{-1/2}(1-q)^{-1/2}\theta^{-1/2}(1-\theta)^{-1}.$$

证 以$\{(p,q),\theta\}$ 和$\{p,q,\theta\}$ 为例,采用 2 种不同的方法推导出相应的 reference 先验.

1) 参数组合$\{(p,q),\theta\}$ 表示有 2 组参数,p 和 q 的重要性相同,它们比 θ 更重要. 此时将 Fisher 信息矩阵写成 $\boldsymbol{I}=\begin{pmatrix}\Sigma_1 & \boldsymbol{0}\\ \boldsymbol{0} & I_{33}\end{pmatrix}$,其中$\Sigma_1=\begin{pmatrix}I_{11} & 0\\ 0 & I_{22}\end{pmatrix}$.

根据文献[16],需要找到 2 个函数 $h_1(p,q,\theta)$ 和 $h_2(p,q,\theta)$:

$$h_1=\det(\Sigma_1)=n^2/((1-p)q(1-q)),$$
$$h_2=I_{33}=nr(1-p)/(\theta(1-\theta)^2).$$

Step 1 令$\Omega_{12i}=\{(p,q)\mid \xi_{1i}<p<\eta_{1i},\xi_{2i}<q<\eta_{2i}\}$,$\Omega_{3i}=\{\theta\mid \xi_{3i}<\theta<\eta_{3i}\}$,其中,当 $i\to\infty$ 时,$\xi_{1i},\xi_{2i},\xi_{3i}\to 0$,$\eta_{1i},\eta_{2i},\eta_{3i}\to 1$,则 $\Omega_i=\Omega_{12i}\times\Omega_{3i}$ 是参数空间的一组紧集.

Step 2 当(p,q) 给定时,θ 条件先验为

$$\pi^i_{R_1}(\theta\mid p,q)=\sqrt{h_2}\Omega_{3i}/\int_{\Omega_{3i}}\sqrt{h_2}\,\mathrm{d}\theta\propto\theta^{-1/2}(1-\theta)^{-1}\Omega_{3i}.$$

Step 3 (p,q) 的边缘先验为

$$\pi^i_{R_1}(p,q)=\exp\Big(\int_{\Omega_{3i}}\pi^i_{R_1}(\theta\mid p,q)\log(h_1)\,\mathrm{d}\theta/2\Big)\Omega_{12i}/\iint_{\Omega_{12i}}\exp\Big(\int_{\Omega_{3i}}\pi^i_{R_1}(\theta\mid p,q)\log(h_1)\,\mathrm{d}\theta/2\Big)\mathrm{d}p\mathrm{d}q\propto(1-p)^{-1/2}q^{-1/2}(1-q)^{-1/2}\Omega_{12i}.$$

Step 4 Ψ 的 reference 先验为

$$\pi_{R_1}=\lim_{i\to\infty}\pi^i_{R_1}(p,q)\,\pi^i_{R_1}(\theta\mid p,q)/(\pi^i_{R_1}(p^*,q^*)\cdot\pi^i_{R_1}(\theta^*\mid p^*,q^*))\propto(1-p)^{-1/2}q^{-1/2}(1-q)^{-1/2}\theta^{-1/2}\cdot(1-\theta)^{-1},$$

其中 p^*、q^* 和 θ^* 是在参数空间中给定的值.

2) 参数组合$\{p,q,\theta\}$ 表示有 3 组参数,重要性程度为 p、q、θ 依次递减. 将 Fisher 信息矩阵 $\boldsymbol{I}$ 的逆矩阵记为

$$\boldsymbol{I}^{-1}=\begin{pmatrix}s_{11} & 0 & 0\\ 0 & s_{22} & 0\\ 0 & 0 & s_{33}\end{pmatrix}.$$

令 $\boldsymbol{s}_1=(s_{11})$,$\boldsymbol{s}_2=\begin{pmatrix}s_{11} & 0\\ 0 & s_{22}\end{pmatrix}$,$\boldsymbol{s}_3=\boldsymbol{I}^{-1}$. 根据文献[17],$\boldsymbol{H}_j=\boldsymbol{s}_j^{-1}$,$h_j$ 是 $\boldsymbol{H}_j$ 右下角的元素,$j=1,2,3$,进一步得到

$$h_1=n/(p(1-p)),\ h_2=np/(q(1-q)),\ h_3=nr(1-p)/(\theta(1-\theta)^2).$$

令 $\Omega_i=\Omega_{1i}\times\Omega_{2i}\times\Omega_{3i}=\{p\mid \varepsilon_{1i}<p<\delta_{1i}\}\times\{q\mid \varepsilon_{2i}<q<\delta_{2i}\}\times\{\theta\mid \varepsilon_{3i}<\theta<\delta_{3i}\}$,当 $i\to\infty$ 时,$\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i}\to 0$,$\delta_{1i},\delta_{2i},\delta_{3i}\to 1$. 令 $\varphi_1=p$,$\varphi_2=q$,$\varphi_3=\theta$,对于 $j=1,2,3$,记 $\varphi_{[j]}=(\varphi_1,\varphi_2,\cdots,\varphi_j)$,$\varphi_{[-j]}=(\varphi_{j+1},\varphi_{j+2},\cdots,\varphi_3)$. 特别地,$\varphi_{[-0]}=(\varphi_1,\varphi_2,\varphi_3)$,而 $\varphi_{[0]}$ 是空的. 由于 h_j 只依赖于 $\varphi_{[j]}$,所以,根据文献[17]得

$$\pi_{R_3}(\varphi_1,\varphi_2,\varphi_3)=\Big(\prod_{j=1}^{3}(|h_j|^{1/2}/A_j)\Big)\Omega_i,$$

其中

$$A_1=\int_{\varepsilon_{1i}}^{\delta_{1i}}\sqrt{n/(p(1-p))}\,\mathrm{d}p=c_1,$$
$$A_2=\int_{\varepsilon_{2i}}^{\delta_{2i}}\sqrt{np/(q(1-q))}\,\mathrm{d}q=p^{1/2}c_2,$$
$$A_3=\int_{\varepsilon_{3i}}^{\delta_{3i}}\sqrt{n(1-p)/(\theta(1-\theta)^2)}\,\mathrm{d}\theta=(1-p)^{1/2}c_3,$$

且 c_1、c_2、c_3 都是常数. 因此,Ψ 的 reference 先验为

$\pi_{R_3}=\lim\limits_{i\to\infty}\pi^i_{R_3}(\varphi_1,\varphi_2,\varphi_3)/(\pi^i_{R_3}(\varphi_1^*,\varphi_2^*,\varphi_3^*))\propto p^{-1/2}(1-p)^{-1/2}q^{-1/2}(1-q)^{-1/2}\theta^{-1/2}(1-\theta)^{-1}$,

其中 φ_1^*、φ_2^* 和 φ_3^* 是在参数空间中给定的值.

2.3　后验分析

定理 3　$\Psi=(p,q,\theta)$ 在先验分布 π_{R_1}、π_{R_2} 和 π_{R_3} 下的后验分布都是恰当的.

证　$\Psi=(p,q,\theta)$ 的后验分布具有以下形式:

$$\pi(p,q,\theta\mid Z,X,Y,B)\propto L(p,q,\theta\mid Z,X,Y,B)\cdot\pi(p,q,\theta),\tag{5}$$

其中 $Z=(Z_1,Z_2,\cdots,Z_n)$ 为观测数据,$X=(X_1,X_2,\cdots,X_n)$、$Y=(Y_1,Y_2,\cdots,Y_n)$、$B=(B_1,B_2,\cdots,B_n)$ 为隐变量,$L(p,q,\theta\mid Z,X,Y,B)$ 为在式(3)中的完全似然函数.

以先验分布 π_{R_3} 为例,代入式(5),则(p,q,θ)的后验分布为

$$\pi_{R_3}(p,q,\theta\mid Z,X,Y,B)\propto p^{\sum_{i=1}^n B_i-1/2}(1-p)^{\sum_{i=1}^n(1-B_i)-1/2}\cdot q^{\sum_{i=1}^n B_iY_i-1/2}(1-q)^{\sum_{i=1}^n B_i(1-Y_i)-1/2}\theta^{\sum_{i=1}^n X_i(1-B_i)-1/2}(1-\theta)^{r\sum_{i=1}^n(1-B_i)-1}.\tag{6}$$

当 Z 和(X,Y,B) 给定时,根据式(6),计算以下积分

$$\int_0^1 p^{\sum_{i=1}^n B_i-1/2}(1-p)^{\sum_{i=1}^n(1-B_i)-1/2}\mathrm{d}p=\Gamma\Big(\sum_{i=1}^n B_i+1/2\Big)\Gamma\Big(\sum_{i=1}^n(1-B_i)+1/2\Big)/\Big(\Gamma(n+1)\Big)<\infty,$$

$$\int_0^1 q^{\sum_{i=1}^n B_iY_i-1/2}(1-q)^{\sum_{i=1}^n B_i(1-Y_i)-1/2}\mathrm{d}q=\Gamma\Big(\sum_{i=1}^n B_iY_i+1/2\Big)\Gamma\Big(\sum_{i=1}^n B_i(1-Y_i)+1/2\Big)/\Big(\Gamma\Big(\sum_{i=1}^n B_i+1\Big)\Big)<\infty,$$

$$\int_0^1\theta^{\sum_{i=1}^n X_i(1-B_i)-1/2}(1-\theta)^{r\sum_{i=1}^n(1-B_i)-1}\mathrm{d}\theta=\Gamma\Big(\sum_{i=1}^n X_i\cdot(1-B_i)+1/2\Big)\Gamma\Big(r\sum_{i=1}^n(1-B_i)\Big)/\Big(\Gamma\Big(\sum_{i=1}^n(r+X_i)(1-B_i)+1/2\Big)\Big)<\infty.$$

由于$\int_0^1\int_0^1\int_0^1\pi_{R_3}(p,q,\theta\mid Z,X,Y,B)\,\mathrm{d}p\mathrm{d}q\mathrm{d}\theta<\infty$,所以 $\Psi=(p,q,\theta)$ 在先验分布 π_{R_3} 下的后验分布是恰当的. 即在(Z,X,Y,B) 给定的条件下,$\pi_{R_3}(p,q,\theta\mid Z,X,Y,B)$ 能够成为一个联合概率密度函数,实施抽样,获得后验样本. 在贝叶斯分析中,验证后验分布的恰当性是一个重要环节,若后验分布不是恰当的,则无法进行抽样获得后验样本. 同理可得,$\Psi=(p,q,\theta)$ 在先验分布 π_{R_1} 和 π_{R_2} 下的后验分布也均是恰当的.

2.4　抽样机制

给定观测数据 $Z=(Z_1,Z_2,\cdots,Z_n)$,通过抛硬币试验,设计隐变量(X,Y,B) 产生样本的过程.

1) 当 $Z_i=0$ 时,抛掷一枚正面朝上概率为 $pq/(pq+(1-p)(1-\theta))$ 的硬币. 若正面朝上,则 $B_i=1$,$Y_i=1$,X_i 从负二项分布抽样. 若反面朝上,则 $B_i=0$,$X_i=0$,随后再抛掷另一枚正面朝上概率为 q 的硬币,若正面朝上,则 $Y_i=1$,否则,$Y_i=0$.

2) 当 $Z_i=1$ 时,抛掷一枚正面朝上概率为 $p(1-q)/(p(1-q)+(1-p)r\theta(1-\theta))$ 的硬币. 若正面朝上,则 $B_i=1$,$Y_i=0$,X_i 从负二项分布抽样. 若反面朝上,则 $B_i=0$,$X_i=1$,随后再抛掷另一枚正面朝上概率为 q 的硬币,若正面朝上,则 $Y_i=1$,否则,$Y_i=0$.

3) 当 $Z_i=k(k=2,3,\cdots)$ 时,抛掷一枚正面朝上概率为 q 的硬币. 若正面朝上,则 $B_i=0$,$Y_i=1$,$X_i=k$,否则,$B_i=0$,$Y_i=0$,$X_i=k$.

从式(6)中,得到各个参数的条件分布

$$\pi_{R_3}(p\mid q,\theta,Z,X,Y,B)\propto p^{\sum_{i=1}^n B_i-1/2}(1-p)^{\sum_{i=1}^n(1-B_i)-1/2},\tag{7}$$

$$\pi_{R_3}(q\mid p,\theta,Z,X,Y,B)\propto q^{\sum_{i=1}^n B_iY_i-1/2}(1-q)^{\sum_{i=1}^n B_i(1-Y_i)-1/2},\tag{8}$$

$$\pi_{R_3}(\theta\mid p,q,Z,X,Y,B)\propto\theta^{\sum_{i=1}^n X_i(1-B_i)-1/2}(1-\theta)^{r\sum_{i=1}^n(1-B_i)-1},\tag{9}$$

且可以通过 R 软件获得相应的样本.

后验分布式(6)的 Gibbs 抽样机制设计如下:

1) 设置参数的初始值为 $p^{(0)}$,$q^{(0)}$,$\theta^{(0)}$;

2) 对于 $t=1,2,\cdots$,进行迭代更新.

(a) 给定$(p^{(t-1)},q^{(t-1)},\theta^{(t-1)})$,通过上述抛硬币试验,得到样本$(X_i^{(t)},Y_i^{(t)},B_i^{(t)})$;

(b) 由式(7),从 Beta$\Big(\sum_{i=1}^n B_i+1/2,\sum_{i=1}^n(1-B_i)+1/2\Big)$ 中抽样得到样本 $p_i^{(t)}$;

(c) 由式(8),从 Beta$\Big(\sum_{i=1}^n B_iY_i+1/2,\sum_{i=1}^n B_i(1-Y_i)+1/2\Big)$ 中抽样得到样本 $q_i^{(t)}$;

(d) 由式(9),从 $\mathrm{Beta}\left(\sum_{i=1}^{n} X_i(1-B_i)+1/2,\ r\sum_{i=1}^{n}(1-B_i)\right)$ 中抽样得到样本 $\theta_i^{(t)}$.

3 数值模拟

本节进行数值模拟,基于 3 种不同 reference 先验,对 ZOINB 模型的参数进行估计与评价. 假设在负二项分布中失败次数 $r=3$,样本容量分别设置为 50 和 100,p 的真值分别设置为 0.3 和 0.7,q 的真值分别设置为 0.4 和 0.6,θ 的真值分别设置为 0.3 和 0.8,置信水平设置为 0.95,所有模拟重复 1 000 次. 参数估计量的均值如表 1 所示,参数估计量的均方误差如表 2 所示,参数估计量的置信区间覆盖率如表 3 所示. 从总体上看,随着样本容量的增加,所有估计量的效果不断提高,表 1 中估计值越来越接近真实值,表 2 中参数估计量的均方误差不断变小,表 3 中置信区间覆盖率越来越接近 95%. 对于 p 的估计,基于先验分布 π_{R_3} 的贝叶斯估计效果更优于 π_{R_1} 和 π_{R_2} 的贝叶斯估计效果,这是因为在先验分布 π_{R_3} 中所体现 p 的信息比在 π_{R_1} 和 π_{R_2} 中所体现 p 的信息更加全面. 当混合比例 p 的真值变大时,来自负二项分布的样本数据变少,这或许会影响对 θ 的估计效果.

表 1 参数估计量的均值

θ	p	q	n	θ_{R_1}	p_{R_1}	q_{R_1}	θ_{R_2}	p_{R_2}	q_{R_2}	θ_{R_3}	p_{R_3}	q_{R_3}
0.3	0.3	0.4	50	0.345	0.287	0.296	0.335	0.348	0.403	0.311	0.303	0.412
			100	0.314	0.296	0.312	0.311	0.317	0.386	0.304	0.302	0.402
		0.6	50	0.332	0.283	0.556	0.331	0.278	0.553	0.308	0.298	0.581
			100	0.313	0.295	0.587	0.323	0.284	0.591	0.301	0.297	0.595
	0.7	0.4	50	0.328	0.634	0.325	0.317	0.646	0.345	0.289	0.712	0.405
			100	0.314	0.678	0.379	0.312	0.684	0.376	0.302	0.705	0.397
		0.6	50	0.336	0.628	0.546	0.331	0.644	0.577	0.313	0.646	0.576
			100	0.328	0.675	0.575	0.322	0.665	0.581	0.301	0.624	0.597
0.8	0.3	0.4	50	0.773	0.287	0.386	0.774	0.275	0.383	0.773	0.287	0.383
			100	0.785	0.295	0.393	0.786	0.292	0.386	0.787	0.297	0.398
		0.6	50	0.771	0.282	0.587	0.776	0.273	0.591	0.774	0.288	0.588
			100	0.787	0.292	0.603	0.785	0.295	0.605	0.787	0.295	0.595
	0.7	0.4	50	0.774	0.682	0.385	0.773	0.666	0.384	0.783	0.682	0.376
			100	0.795	0.692	0.393	0.783	0.681	0.392	0.791	0.695	0.394
		0.6	50	0.777	0.691	0.612	0.783	0.683	0.604	0.787	0.697	0.576
			100	0.796	0.695	0.596	0.788	0.696	0.595	0.799	0.697	0.586

表 2 参数估计量的均方误差

θ	p	q	n	θ_{R_1}	p_{R_1}	q_{R_1}	θ_{R_2}	p_{R_2}	q_{R_2}	θ_{R_3}	p_{R_3}	q_{R_3}
0.3	0.3	0.4	50	0.082	0.063	0.089	0.061	0.035	0.066	0.042	0.043	0.066
			100	0.063	0.038	0.076	0.035	0.026	0.059	0.031	0.025	0.051
		0.6	50	0.072	0.073	0.094	0.057	0.038	0.055	0.045	0.044	0.063
			100	0.064	0.056	0.061	0.032	0.025	0.050	0.027	0.038	0.043
	0.7	0.4	50	0.082	0.063	0.094	0.073	0.024	0.072	0.055	0.046	0.062
			100	0.056	0.048	0.063	0.066	0.019	0.063	0.041	0.023	0.054
		0.6	50	0.054	0.056	0.082	0.071	0.036	0.073	0.036	0.056	0.062
			100	0.052	0.045	0.064	0.055	0.018	0.066	0.017	0.029	0.046
0.8	0.3	0.4	50	0.048	0.062	0.082	0.045	0.056	0.056	0.042	0.054	0.044
			100	0.042	0.055	0.063	0.037	0.042	0.048	0.026	0.046	0.031
		0.6	50	0.046	0.073	0.077	0.036	0.062	0.072	0.044	0.047	0.031
			100	0.043	0.062	0.056	0.025	0.043	0.063	0.019	0.027	0.025
	0.7	0.4	50	0.043	0.059	0.047	0.038	0.053	0.058	0.032	0.043	0.037
			100	0.039	0.022	0.037	0.026	0.043	0.048	0.027	0.030	0.023
		0.6	50	0.041	0.062	0.045	0.033	0.055	0.043	0.032	0.037	0.035
			100	0.033	0.042	0.035	0.028	0.032	0.028	0.023	0.022	0.018

表 3　参数估计量的置信区间覆盖率

θ	p	q	n	θ_{R_1}	p_{R_1}	q_{R_1}	θ_{R_2}	p_{R_2}	q_{R_2}	θ_{R_3}	p_{R_3}	q_{R_3}
0.3	0.3	0.4	50	0.876	0.911	0.913	0.945	0.932	0.941	0.942	0.947	0.962
			100	0.916	0.925	0.921	0.953	0.943	0.954	0.945	0.952	0.954
		0.6	50	0.876	0.924	0.923	0.962	0.941	0.962	0.933	0.953	0.945
			100	0.922	0.943	0.935	0.956	0.946	0.961	0.947	0.952	0.953
	0.7	0.4	50	0.894	0.923	0.925	0.906	0.944	0.965	0.903	0.944	0.947
			100	0.913	0.933	0.937	0.939	0.952	0.958	0.948	0.951	0.952
		0.6	50	0.839	0.927	0.934	0.941	0.937	0.936	0.939	0.953	0.945
			100	0.924	0.935	0.946	0.947	0.951	0.943	0.947	0.952	0.949
0.8	0.3	0.4	50	0.932	0.937	0.928	0.943	0.944	0.945	0.942	0.943	0.936
			100	0.943	0.942	0.942	0.952	0.954	0.947	0.948	0.949	0.943
		0.6	50	0.918	0.923	0.937	0.932	0.939	0.941	0.943	0.936	0.939
			100	0.941	0.936	0.942	0.942	0.943	0.948	0.945	0.947	0.945
	0.7	0.4	50	0.927	0.936	0.937	0.912	0.938	0.947	0.946	0.942	0.938
			100	0.936	0.943	0.943	0.954	0.942	0.953	0.949	0.948	0.946
		0.6	50	0.931	0.936	0.937	0.938	0.932	0.942	0.943	0.936	0.943
			100	0.955	0.944	0.943	0.944	0.954	0.948	0.947	0.949	0.949

4　实例分析

考察在 R 软件 pscl 包中的数据集 bioChemists，它记录了 915 名生物化学博士生学习期间发表论文的数量、性别、婚姻状况、孩子的数量、学校的荣誉等信息，其中有 275 人没有发表过论文(记为“0”) ，有 246 人发表了 1 篇论文，178 人发表了 2 篇论文，该数据集零一膨胀现象十分显著. 选择 ZOINB 模型进行拟合分析，基于 3 种先验 π_{R_1}、π_{R_2}、π_{R_3}，计算参数的估计值和 95% 的估计区间，并与文献 [2] 中的结果进行比较(见表 4) .

表 4　当 $r = 2$ 时，ZOINB 模型的参数估计值和 95% 的估计区间

参数	MLE	MLE(EM)	Bayes	Bayes_{R_1}	Bayes_{R_2}	Bayes_{R_3}
p	0.048 1	0.048 1	0.044 0	0.047 8	0.047 8	0.0481
	(0.009,0.163)	(0.024,0.165)	(0.002,0.123)	(0.019,0.121)	(0.023,0.132)	(0.031,0.168)
q	0.651 4	0.651 4	0.776 0	0.651 2	0.651 3	0.651 4
	(0.310,0.890)	(0.396,0.768)	(0.483,0.983)	(0.391,0.715)	(0.393,0.714)	(0.426,0.703)
θ	0.468 2	0.468 2	0.467 9	0.468 1	0.468 2	0.468 2
	(0.436,0.498)	(0.458,0.501)	(0.118,0.491)	(0.418,0.534)	(0.422,0.529)	(0.456,0.501)
AIC	2 608.968	2 608.963	2 653.816	2 620.715	2 618.834	2 606.143

表 4 中第 2 ~ 4 列是文献 [2] 的结果，第 5 ~ 7 列是本文的结果. 由表 4 可以看出: 基于 reference 先验 π_{R_1}、π_{R_2}、π_{R_3} 的估计结果要明显优于第 4 列朴素贝叶斯估计的结果，这说明 reference 先验比 naive flat 先验体现了更多的参数信息. 第 7 列基于 π_{R_3} 的效果与第 3 列在 EM 算法下极大似然估计法的效果相当，但相应的 AIC(Akaike information criterion) 要更小一些，相比之下，基于 π_{R_3} 的估计效果要稍胜一筹.

5　结论

本文对零一膨胀负二项模型进行了客观贝叶斯分析，基于隐变量的完全似然函数，构造了参数的 Fisher 信息矩阵，推导出不同形式的 reference 先验，ZOINB 模型的数值模拟和实例分析效果都较为理想. 采用本文提出的方法，引入恰当的隐变量，使得 Fisher 信息矩阵具有对角矩阵或分块矩阵的形式，这对解决高维或者超高维参数空间的统计推断问题提供了一种新的研究思路.

6　参考文献

[1] LAMBERT D. Zero-inflated Poisson regression with an application to defects in manufacturing [J]. Technometrics, 1992,34(1):1-14.

[2] BOHNING D,DIETZ E,SCHLATTMANN P,et al. Corrigendum: the zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology [J].

Journal of the Royal Statistical Society: Series A, 2000, 160(1): 195-209.

[3] HE Xuming, XUE Hongqi, SHI Ningzhong. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated Poisson models [J]. Journal of Multivariate Analysis, 2010, 101(9): 2026-2038.

[4] 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.

[5] TANG Yincai, LIU Wenchen, XU Ancha. Statistical inference for zero-and-one-inflated poisson models [J]. Statistics Theory and Related Fields, 2017, 1(2): 216-226.

[6] LIU Wenchen, TANG Yincai, XU Ancha. A zero-and-one inflated Poisson model and its application [J]. Statistics and Its Interface, 2018, 11(2): 339-351.

[7] 夏丽丽,田茂再. 零一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.

[8] 肖翔. 0-1 膨胀泊松分布的客观贝叶斯分析及其应用 [J]. 数理统计与管理, 2022, 41(3): 413-426.

[9] GARAY A M, HASHIMOTO E M, ORTEGA E M M, et al. On estimation and influence diagnostics for zero-inflated negative binomial regression models [J]. Computational Statistics and Data Analysis, 2011, 55(3): 1304-1318.

[10] MWALILI S M, LESAFFRE E, DECLERCK D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research [J]. Statistical Methods in Medical Research, 2008, 17(2): 123-139.

[11] YAU K K W, WANG Kui, LEE A H. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros [J]. Biometrical Journal, 2003, 45(4): 437-452.

[12] 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.

[13] 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.

[14] XIAO Xiang, TANG Yincai, XU Ancha, 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.

[15] BERNARDO J M. Reference posterior distributions for Bayesian inference [J]. Journal of the Royal Statistical Society (Series B): Methodological, 1979, 41(2): 113-128.

[16] 茆诗松,汤银才. 贝叶斯统计 [M]. 2 版. 北京: 中国统计出版社, 2012.

[17] BERGER J O, BERNARDO J M. On the development of the reference prior method [J]. Bayesian Statistics, 1992, 4(4): 35-60.

The Objective Bayesian Analysis of Zero – and – One Inflated Negative Binomial Model

MA Qiaoling, XIAO Xiang*

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

Abstract: The Bayesian model is established and the objective Bayesian analysis of zero-and-one inflated negative binomial distribution is discussed in the paper. Using the data augmentation strategy and the full likelihood function, different reference priors of zero-and-one inflated negative binomial model are calculated and it is further shown that the corresponding posterior distributions are proper. For different sample sizes and parameter true values of the parameters, three different reference priors are evaluated through simulations. Finally, a doctoral dissertation data set is analyzed to illustrate the practicability of the proposed model and method.

Key words: zero-and-one inflated negative binomial model; data augmentation strategy; objective Bayesian analysis

(责任编辑: 曾剑锋)