ANÁLISE BAYESIANA DE DADOS DE CONTAGEM COM EXCESSO DE ZEROS E UNS

Milton Yoshio SAITO[1]
Josemar RODRIGUES[1]

▪ RESUMO: Neste texto as distribuições discretas com Zeros e Uns Inflacionadas (ZOID) são estudadas do ponto de vista Bayesiano. São introduzidas duas variáveis latentes de fácil interpretação e que simplificam bastante a expressão da função de verossimilhança. Um exemplo ilustrativo é apresentado utilizando o modelo Poisson e dados associados a um estudo de seguro dentário realizado na Suécia. O algoritmo MCMC é adaptado para solução Bayesiana do problema, obtendo-se facilmente amostras *a posteriori* dos parâmetros do modelo Poisson com Zeros e Uns Inflacionados.

▪ PALAVRAS-CHAVE: Inferência Bayesiana; excesso de zeros e de uns; mistura de modelos.

1 Introdução

Há certas aplicações envolvendo dados discretos onde $Y = 0$ é observado com freqüência significativamente alta em relação ao predito pelo modelo adotado. Alguns exemplos da distribuição Poisson com zeros inflacionados (distribuição ZIP - *Zero Inflated Poisson*) podem ser encontrados em Goraski (1977), Cohen (1960), Martin e Katti (1965) e Kemp (1986).

Recentemente, Melkersson e Olsson (1999) apresentaram uma aplicação em que o modelo ZIP é adaptado também para o excesso de uns nos dados. Nesta aplicação a distribuição ZOIP (*Zeros and Ones Inflated Poisson*) é utilizada num estudo de contagem de visitas ao dentista de cidadãos suecos.

Em dados de contagem do número de visitas ao dentista, muitas vezes, observa-se uma grande quantidade de zero visita na amostra (Rosenqvist et al., 1995; Olsson, 1998). Um indivíduo pode não ir ao dentista por vários motivos: por apresentar

[1]Departamento de Estatística, Universidade Federal de São Carlos - UFSCar, Caixa Postal 676, CEP 13565-905, São Carlos, SP, Brasil. E-mail: *vjosemar.ufscar.br*

dentes saudáveis, por ter medo de dentista, por não poder comparecer (pode estar doente ou não ter tempo, por exemplo), ou simplesmente por livre-arbítrio (ele não é obrigado a ir ao dentista).

Entretanto, na aplicação apresentada, uma visita ao dentista é observada com maior freqüência do que zero visita. Isso pode ser explicado por ter-se costume de ir ao dentista como uma forma de prevenção, por exemplo, em muitos países, crianças em idade escolar são condicionadas a irem ao dentista pelo menos uma vez ao ano. Existem problemas dentários de natureza irreversível e alto custo onde a prevenção geralmente é recomendada já na infância. Além disso, o Sistema Nacional de Seguro Dentário Sueco, encarrega-se de enviar lembretes aos cidadãos para comparecerem ao dentista para fazerem avaliações periódicas. Os sistemas particulares costumam imitar este mesmo tipo de procedimento.

O estudo feito por Melkersson e Olsson (1999) tinha a preocupação de verificar se o aumento de investimentos no tratamento dentário na infância e na adolescência, reduzem os gastos do sistema de seguro dentário na idade adulta. Segundo as normas do Sistema Nacional de Seguro Dentário Sueco, atualmente, o tratamento para cárie dentária é gratuita para crianças e jovens com até 20 anos de idade. A partir de janeiro de 1999 os tratamentos dentários para pessoas com idade entre 20 e 29 anos passaram a ser subsidiados, fato que motivou o estudo realizado.

Neste texto será apresentada uma aproximação bayesiana para o problema de excesso de zeros e uns, similar à realizada por Rodrigues (2003), que descreveu um modelo Bayesiano para distribuições discretas somente com excesso de zeros. Rodrigues (2003) considerou uma mistura de modelos, onde os zeros poderiam ter origem de uma distribuição degenerada no zero ou uma outra distribuição discreta qualquer.

A distribuição com zeros e uns inflacionados (ZOID) para o modelo $p(y|\theta)$ assumido inicialmente pode ser representada pela seguinte expressão:

$$p(y|\theta,\omega) = \omega_0 I_{\{0\}}(y) + \omega_1 I_{\{1\}}(y) + \omega_2 p(y|\theta), \qquad y = 0,1,2,... \tag{1}$$

, onde $\omega = (\omega_0, \omega_1, \omega_2)$, $\quad \omega_0$, ω_1 e $\omega_2 \geq 0 \quad$ e $\quad \omega_0 + \omega_1 + \omega_2 = 1$.

Na verdade ω_0 e ω_1 poderiam assumir valores menores que zero, o que corresponderia ao caso deflacionado. Porém, como o caso de dados inflacionados é mais freqüente, neste texto serão considerados somente valores positivos de ω_0 e ω_1.

A esperança e a variância para ZOID são respectivamente:

$$E(y|\theta,\omega) = \omega_1 + \omega_2 E(y|\theta), \tag{2}$$

$$Var(y|\theta,\omega) = \omega_1 + \omega_2 E\left(y^2|\theta\right) - [\omega_1 + \omega_2 E(y|\theta)]^2. \tag{3}$$

Logo para o modelo Poisson temos:

$$\begin{aligned} E(y|\theta,\omega) &= \omega_1 + \omega_2\theta, \\ Var(y|\theta,\omega) &= \omega_1 + \omega_2\theta(\theta+1) - [\omega_1 + \omega_2\theta]^2. \end{aligned}$$

2 A função de verossimilhança baseada nos dados ampliados

Suponha que $Y = (Y_1, Y_2, ..., Y_n)$ seja um vetor de n variáveis aleatórias independentes obtidas do modelo ZOID. Seja $A = \{y_i : y_i = 0, i = 1, ..., n\}$, $B = \{y_i : y_i = 1, i = 1, ..., n\}$, $n_0 = \#(A)$ e $n_1 = \#(B)$, então a função de verossimilhança pode ser expressa por:

$$L(\theta, \omega) = [\omega_0 + \omega_2 p(0|\theta)]^{n_0} [\omega_1 + \omega_2 p(1|\theta)]^{n_1} \omega_2^{n-n_0-n_1} \prod_{\substack{y_i \notin A \\ y_i \notin B}} p(y_i|\theta). \quad (4)$$

Os elementos do conjunto A podem vir de dois diferentes grupos, a distribuição degenerada em zero ou $p(0|\theta)$. Já os do conjunto B podem vir da distribuição degenerada em um ou $p(1|\theta)$. Nesta situação é natural a introdução de duas variáveis latentes:

$$I_i = \begin{cases} 1 & p(\theta, \omega_0, \omega_2) \\ 0 & [1 - p(\theta, \omega_0, \omega_2)], \end{cases} \quad (5)$$

onde $i = 1, 2, ..., n_0$ e

$$p(\theta, \omega_0, \omega_2) = \frac{\omega_0}{\omega_0 + \omega_2 p(0|\theta)},$$

variável que simula quando o i-ésimo elemento do conjunto A é retirado da distribuição degenerada em zero ou não, e

$$J_j = \begin{cases} 1 & p(\theta, \omega_1, \omega_2) \\ 0 & [1 - p(\theta, \omega_1, \omega_2)], \end{cases} \quad (6)$$

onde $j = 1, 2, ..., n_1$ e

$$p(\theta, \omega_1, \omega_2) = \frac{\omega_1}{\omega_1 + \omega_2 p(1|\theta)},$$

que simula quando o i-ésimo elemento do conjunto B é retirado da distribuição degenerada em um ou não.

Então, a função de verossimilhança baseada nos dados ampliados $D = \{Y, I, J\}$, onde $I = \{I_1, ..., I_{n_0}\}$ e $J = \{J_1, ..., J_{n_1}\}$, introduzida por Tanner e Wong (1987), é dada por:

$$\begin{aligned} L(\theta, \omega|D) = \quad & L(\theta, \omega) \prod_{i=1}^{n_0} \left[p(\theta, \omega_0, \omega_2)^{I_i} (1 - p(\theta, \omega_0, \omega_2))^{1-I_i} \right] \\ & \prod_{j=1}^{n_1} \left[p(\theta, \omega_1, \omega_2)^{J_j} (1 - p(\theta, \omega_1, \omega_2))^{1-J_j} \right] \\ & \underbrace{(\omega_2)^{n-n_0-n_1} \omega_0^{S_0} \omega_1^{S_1}}_{\text{0's e 1's inflacionados}} \underbrace{p(0|\theta)^{n_0-S_0} p(1|\theta)^{n_1-S_1} \prod_{\substack{y_i \notin A \\ y_i \notin B}} p(y_i|\theta)}_{\text{dados do modelo}}, \end{aligned} \quad (7)$$

onde

$$S_0 = \sum_{i=1}^{n_0} I_i \sim Binomial\left[n_0, p\left(\theta, \omega_0, \omega_2\right)\right] \text{ e}$$

$$S_1 = \sum_{j=1}^{n_1} J_j \sim Binomial\left[n_1, p\left(\theta, \omega_1, \omega_2\right)\right].$$

Nota-se que para os dados ampliados, a expressão da verossimilhança pode ser dividida em duas partes, a primeira responsável pelo excesso de zeros e uns e a segunda pelos dados do modelo original (sem considerar inflação). Assumindo uma *priori* conjunta, $\pi\left(\theta, \omega\right)$, *a posteriori* conjunta de (θ, ω), dado D, é

$$\pi\left(\theta, \omega | D\right) \propto L\left[\theta, \omega | D\right] \pi\left(\theta, \omega\right). \tag{8}$$

Na próxima Seção será apresentado um procedimento que pode ser implementado para um modelo específico para obter as *posterioris* marginais de θ e ω, dado os dados D.

3 A distribuição Poisson com zeros e uns inflacionados

A partir de agora a teoria apresentada na Seção anterior será aplicada ao modelo Poisson, cuja função de probabilidades é dada por:

$$p\left(y | \theta\right) = \frac{\theta^y e^{-\theta}}{y!}, \qquad y = 0, 1, 2, ... \tag{9}$$

Contudo, a metodologia é genérica e pode ser usada para outras distribuições discretas.

A função de verossimilhança de (θ, ω), dado D, para o modelo ZOIP (*Zero and One Inflated Poisson*) é

$$L\left(\theta, \omega | D\right) \propto \omega_0^{S_0} \omega_1^{S_1} \omega_2^{n - S_0 - S_1} \theta^{n_1 - S_1 + Z} e^{-\theta(n - S_0 - S_1)}, \tag{10}$$

onde
$Z = \sum_{\substack{y_i \notin A \\ y_i \notin B}} y_i$, $S_0 \sim Binomial\left[n_0, p\left(\theta, \omega_0, \omega_2\right)\right]$ e $S_1 \sim Binomial\left[n_1, p\left(\theta, \omega_1, \omega_2\right)\right]$,
$p\left(\theta, \omega_0, \omega_2\right) = \dfrac{\omega_0}{\omega_0 + \omega_2 e^{-\theta}}$ e $p\left(\theta, \omega_1, \omega_2\right) = \dfrac{\omega_1}{\omega_1 + \omega_2 \theta e^{-\theta}}$.

A função de verossimilhança sugere as seguintes *prioris* independentes, para obter-se um modelo conjugado:

$$\pi\left(\theta\right) \sim Gama\left(a, b\right) \quad \text{e} \quad \pi\left(\omega\right) \sim Dirichlet\left(\alpha_0, \alpha_1, \alpha_2\right).$$

Então, a distribuição *posteriori* conjunta para (θ, ω) dados D, é

$$\pi(\theta,\omega|D) \propto \omega_0^{S_0+\alpha_0-1}\omega_1^{S_1+\alpha_1-1}\omega_2^{n-S_0-S_1+\alpha_2-1} \times\theta^{n_1-S_1+Z+a-1}e^{-\theta(n-S_0-S_1+b)} \quad (11)$$

Nota-se que (11) pode ser escrita como o produto de duas distribuições independentes, uma *Dirichlet* para ω e outra *Gama* para θ. Os parâmetros de ambas *posterioris* são constituídos por mistura de informações provenientes da função de verossimilhança (dados) e das *prioris*. O fato de terem sido implementadas *prioris* conjugadas facilita a escolha dos valores de a, b, α_0, α_1 e α_2, no caso de desejar-se torná-las informativas.

Simulação da *posteriori* usando algoritmo MCMC

O procedimento usado tem apenas dois passos:

Passo 1: Dado $\left(\theta^{(j-1)},\omega_0^{(j-1)},\omega_1^{(j-1)},\omega_2^{(j-1)}\right)$ no estágio $(j-1)$ gera-se uma amostra aleatória $S_0^{(j)}$ e $S_1^{(j)}$ de:

$$Binomial\left[n_0, p\left(\theta^{(j-1)},\omega_0^{(j-1)},\omega_2^{(j-1)}\right)\right]$$

e

$$Binomial\left[n_1, p\left(\theta^{(j-1)},\omega_1^{(j-1)},\omega_2^{(j-1)}\right)\right],$$

respectivamente.

Passo 2: Dado $S_0^{(j)}$ e $S_1^{(j)}$, gera-se uma amostra aleatória das densidades

$$\begin{aligned}\left(\omega_0^{(j)},\omega_1^{(j)},\omega_2^{(j)}\right) &\sim Dirichlet\left(S_0^{(j)}+\alpha_0, S_1^{(j)}+\alpha_1, n-S_0^{(j)}-S_1^{(j)}+\alpha_2\right)\\ \theta^{(j)} &\sim Gama\left(Z+n_1-S_1^{(j)}+a, n-S_0^{(j)}-S_1^{(j)}+b\right).\end{aligned}$$

O procedimento é exatamente o mesmo método MCMC apresentado por Diebolt & Robert (1994), que mostraram que o algoritmo realmente converge para $\pi(\theta,\omega|Y)$.

Como existem programas que não conseguem gerar amostras da distribuição Dirichlet diretamente, pode-se introduzir as seguintes instruções no passo 2:

Gerar amostras de:

$$\begin{aligned}Z_0^{(j)} &\sim Gama\left(S_0^{(j)}+\alpha_0, c\right)\\ Z_1^{(j)} &\sim Gama\left(S_1^{(j)}+\alpha_1, c\right)\\ Z_2^{(j)} &\sim Gama\left(n-S_0^{(j)}-S_1^{(j)}+\alpha_2, c\right),\end{aligned}$$

onde c é uma constante positiva qualquer, e assim

$$\omega_i^{(j)} = \frac{Z_i^{(j)}}{\sum_{k=0}^{2} Z_k^{(j)}} \sim Dirichlet\left(S_0^{(j)} + \alpha_0, S_1^{(j)} + \alpha_1, n - S_0^{(j)} - S_1^{(j)} + \alpha_2\right).$$

Usando o programa WinBugs 1.4 Beta elaborou-se a seguinte representação gráfica do procedimento MCMC para o modelo ZOIP:

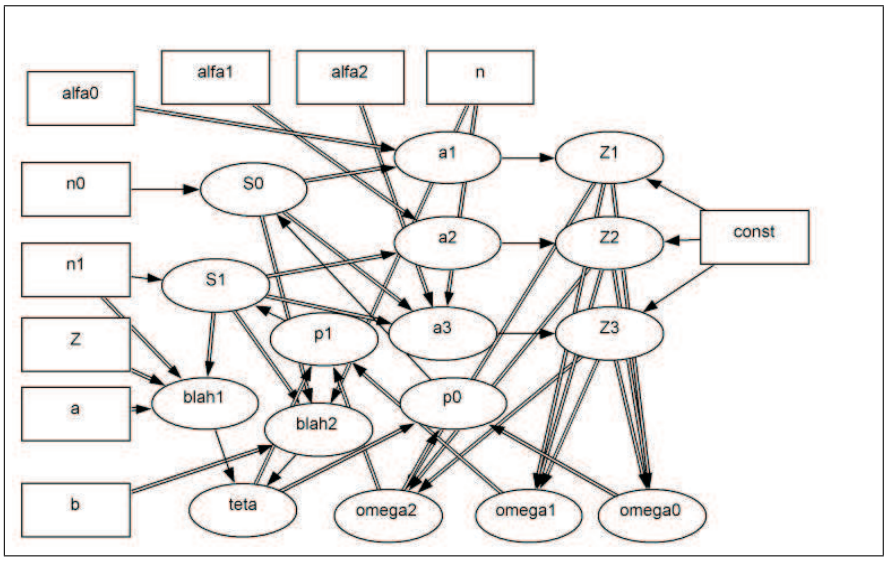

FIGURA 1 - Representação gráfica do procedimento MCMC - Doodle.

4 Exemplo ilustrativo para o modelo Zoip

Como um exemplo ilustrativo do modelo ZOIP considerou-se os dados da Tabela 1 utilizado por Melkersson e Olsson (1999), em que, inicialmente, houve uma tentativa de ajuste do modelo Poisson. Os dados representam a freqüência de visitas ao dentista de cidadãos suecos em 1990.

Os indivíduos visitaram o dentista entre 0 e 20 vezes em um ano, com média de duas visitas, e mediana e moda de uma visita apenas. A amostra é formada por 17,5% de zero visita e 41% de uma visita. Percebe-se que a freqüência de zeros e uns é muito elevada. Nota-se também que a média e variância amostrais diferem significativamente fornecendo um indício de que o ajuste de uma Poisson teórica não é apropriado.

Tabela 1 – Número de visitas ao dentista em 1990

N° de visitas	Freqüência	Poisson teórica
0	134	111
1	314	214
2	149	207
3	69	133
4	32	65
5	26	25
6	14	8
7	6	2
8	1	1
9	0	0
10	11	0
12	3	0
15	3	0
20	4	0
Total	766	766

Utilizou-se, então o algoritmo computacional apresentado na Seção anterior para gerar amostras das densidades marginais a posteriori dos parâmetros da distribuição ZOIP. Utilizando prioris não informativas para ω, isto é, considerando $\alpha_0 = \alpha_1 = \alpha_2 = 1$, e uma priori aproximadamente não informativa para θ com $a = b = 0,0001$. Obteve-se então os gráficos da Figura 2 para as posterioris marginais.

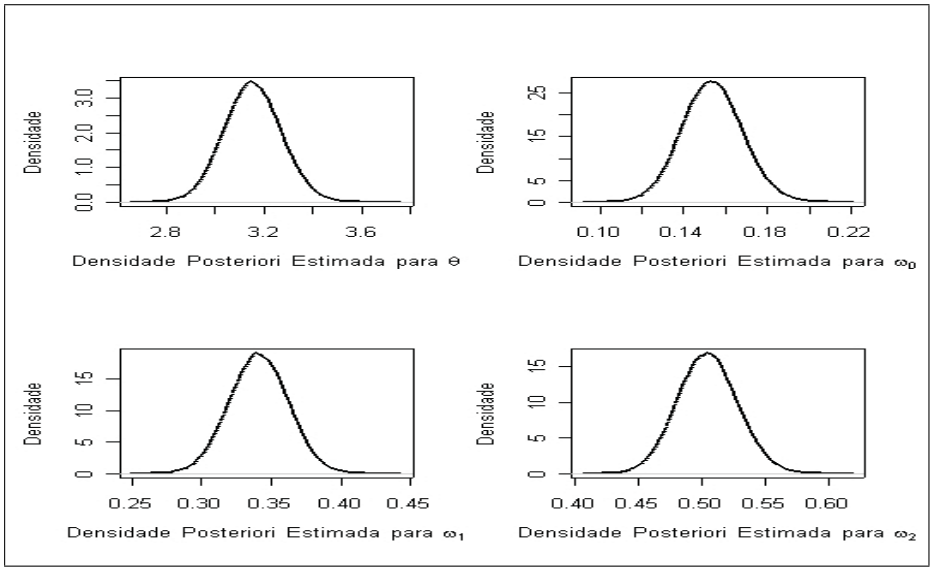

FIGURA 2 – Densidades posterioris marginais.

Na Figura 3 observam-se seqüências de 10.000 amostras das densidades *a posteriori* de cada um dos parâmetros. Tem-se um indício gráfico de convergência do procedimento MCMC implementado, já que os valores aparentam distribuirem-se aleatoriamente em torno de um ponto médio e sem tendência.

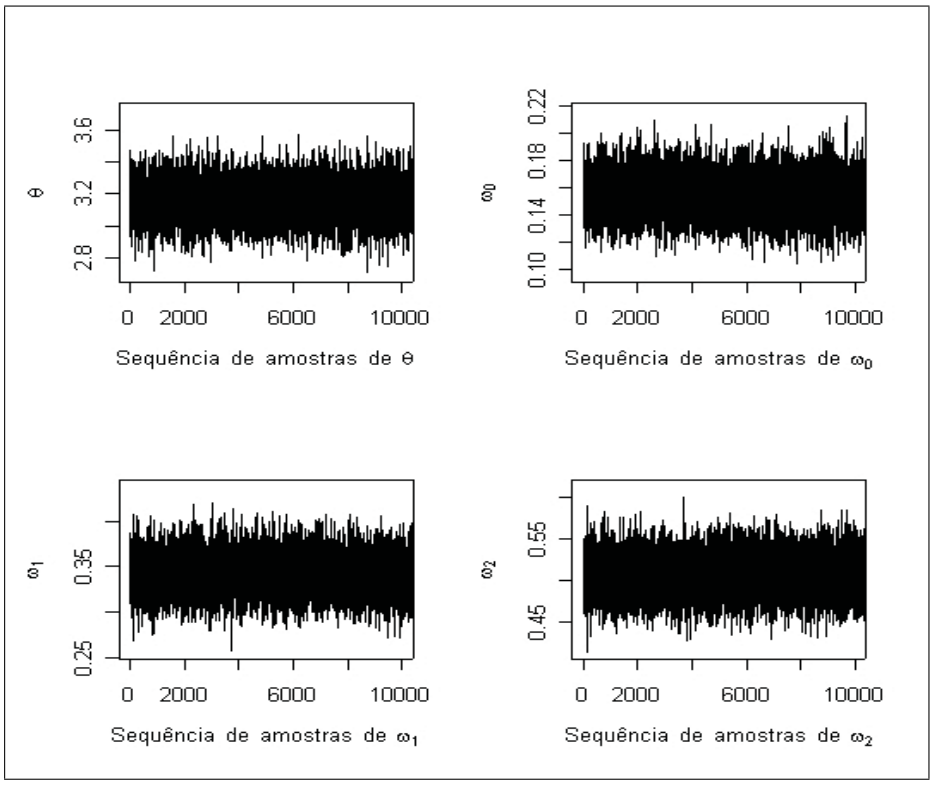

FIGURA 3 – Seqüências de amostras de θ e ω.

A partir das 10.000 amostras, obtidas utilizando-se um período de aquecimento de 5.000 iterações e saltos de 5, obtiveram-se as seguintes estatísticas de resumo para os parâmetros do modelo ZOIP dispostas na Tabela 2:

Tabela 2 – Estatísticas de resumo

Parâmetro	Média	Moda	DP	IC 95%	Z-score	GR
θ	$3,1544$	$3,1478$	$0,1164$	$[2,9270\ ;\ 3,3790]$	$-1,4430$	$1,0$
ω_0	$0,1540$	$0,1530$	$0,0144$	$[0,1272\ ;\ 0,1827]$	$-1,2160$	$1,0$
ω_1	$0,3415$	$0,3402$	$0,0210$	$[0,3000\ ;\ 0,3828]$	$-0,2702$	$1,0$
ω_2	$0,5044$	$0,5053$	$0,0237$	$[0,4584\ ;\ 0,5516]$	$0,8950$	$1,0$

O Z-score é a estatística do teste de Geweke, a um nível de significância de $0,05$, aceita-se a convergência da cadeia para cada um dos parâmetros do modelo. A partir de 3 cadeias diferentes aplicou-se também o teste de Gelman Rubin (GR), e novamente aceitou-se a convergência do procedimento.

O procedimento computacional para realização da simulação foi adaptado para o software WinBugs versão 1.4 Beta e foi colocado em anexo no final deste artigo.

Na Tabela 3 tem-se o ajuste da distribuição ZOIP via metodologia Bayesiana e da distribuição Poisson tradicional. Pela medida de ajuste χ^2_{OBS} observa-se que o modelo Poisson com zeros e uns inflacionados apresenta melhor aderência, adptando-se muito bem as quantidades excessivas de zeros e uns observados no dados. Esta foi a mesma conclusão obtida por Melkersson & Olsson (1999), entretanto, utilizando a abordagem clássica e introduzindo variáveis explanatórias no modelo.

Tabela 3 – Comparação entre o ajuste dos modelos Poisson e ZOIP

N° de visitas	Observado	Modelo Poisson	Modelo ZOIP
0	134	$110,7$	$133,9$
1	314	$214,1$	$312,9$
2	149	$207,1$	$82,4$
3	69	$133,6$	$86,4$
4	32	$64,6$	$68,0$
5	26	$25,0$	$42,8$
6	14	$8,0$	$22,5$
7	6	$2,2$	$10,1$
$8-20$	22	$0,7$	$6,0$
$\chi^2_{obs} = \sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i}$		$774,7$	$120,5$

5 Considerações finais

A abordagem Bayesiana para a distribuição ZOIP pode ser realizada facilmente. A convergência do algoritmo para obter amostras das densidades *posterioris* é bastante rápida, pois os saltos entre as iterações e o período de aquecimento são pequenos como foram utilizadas *prioris* conjugadas, amostras das *posterioris* são obtidas de forma quase imediata. A metodologia apresentada neste texto pode ser utilizada em outras distribuições discretas além desta distribuição. Para trabalhos futuros pode-se introduzir variáveis explanatórias no modelo, o que permitiria a essa abordagem fornecer resultados de maior interesse para o problemas de Melkersson & Olsson (1999). Além disso, existe interesse de aplicar a mesma metodologia utilizando a distribuição generalizada de Poisson.

SAITO, M. Y.; RODRIGUES, J. A Bayesian analysis of zero and one inflated distributions. *Rev. Mat. Estat.*, São Paulo, v.23, n.1, p.47-57, 2005.

▪ *ABSTRACT: In this paper, we study zero and one inflated discrete distributions under a Bayesian point of view. We introduce two simple latent variables which simplify the expression of the likelihood function. An ilustrative example using the Poisson model and a data set of Swedish dental insurance system are presented. The MCMC algorithm is adapted to a Bayesian solution to the problem and later samples are easily obtained.*

▪ *KEYWORDS: Bayesian Inference; zero and one excess; mixture.*

Referências

COHEN, A. C., Estimating the parameteres of modified Poisson distribution. *J. Am. Stat. Assoc.*, Alexandria, v.55, p.139-143, 1960.

DIEBOLT, J.; ROBERT, C. P. Estimation of finite mixture distributions though Bayesian sampling. *J. R. Stat.* B., Cambridge, v.56, p.363-375, 1994.

GORASKI, A. *Distribution Z-Poisson.* Paris: Institut de Statistique de l'Université de Paris.1977. p.43-45. (Publication, 12).

HEILBRON, D. Zero-altered and other regression models for count data with added zeros. *Biom. J.*, Berlin, v.36, p.531-547, 1994.

KEMP, A. W.Weighted discrepancies and maximum likelihood estimation for discrete distribution. *Commun. Stat.*, New York, v.15, p.783-803, 1986.

LAMBERT, D. Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, Washington, v.34, p.1-14, 1992.

MARTIN, D. C.; KATTI, S.K. Fitting of some contagious distributions to some available data by the maximum likelihood method. *Biometrics*, Washington, v.21, p.34-48, 1965.

MELKERSSON, M.; OLSSON, C. Is visiting the dentist a good habit? analysing count data with excess zeros and excess ones. 1999. Disponível em: htttp://www.econ.umu.se¿. Acesso em: 27 dez.2003.

OLSSON, C. *Supplier induced demand: an analysis of the Swedish dental care market.* Umeã University, 1998. (Umed Economic studies, 490).

RODRIGUES, J. Bayesian analysis of zero-inflated distributions. *Commun. Stat.*, New York, v.32, p.281-289, 2003.

ROSENQVIST, G.; ARIEN, S-S.; SINTONEN, H. *Modified cout data models with an application to demand for dental care.* Swedish School of Economics and Business Administration, 1995. (Working Paper, 293).

TANNER, M. A.; WONG, W.W. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, New York, v.82 p.528-540, 1997.

Apêndice

Programa em WinBugs

Programa em WinBugs para o procedimento MCMC de Diebolt & Robert (1994) aplicado ao modelo ZOIP:

```
Modelo
{
Z1 ~dgamma(a1,const)
Z2 ~dgamma(a2,const)
Z3 ~dgamma(a3,const)
S0 ~dbin(p0,n0)
S1 ~dbin(p1,n1)
p0 <- omega0 / (omega0 + omega2 * exp( -teta))
p1 <- omega1 / (omega1 + (omega2 * teta) * exp( -teta))
teta ~dgamma(blah1,blah2)
omega0 <- Z1 / (Z1 + Z2 + Z3)
omega1 <- Z2 / (Z1 + Z2 + Z3)
omega2 <- Z3 / (Z1 + Z2 + Z3)
a1 <- alfa0 + S0
a2 <- alfa1 + S1
a3 <- n - S0 - S1 + alfa2
blah1 <- Z + n1 - S1 + a
blah2 <- n - S0 - S1 + b
}
Dados
list(n=766,n0=134,n1=314,Z=1168,a=0.001,b=0.001,const=1,alfa0=1,alfa1=1,alfa2=1)
Valores Iniciais
list(teta=2,Z1=100,Z2=100,Z3=566,S1=3,S0=5)
list(teta=3,Z1=2,Z2=9,Z3=586,S1=7,S0=50)
list(teta=5,Z1=4,Z2=10,Z3=66,S1=6,S0=5)
```

Recebido em 09.04.2004.

Aprovado após revisão em 17.03.2005.