MODELOS PARA PROPORÇÕES COM SUPERDISPERSÃO E EXCESSO DE ZEROS - UM PROCEDIMENTO BAYESIANO

Adriano Ferreti BORGATTO[1]
Clarice Garcia Borges DEMÉTRIO[2]
Roseli Aparecida LEANDRO[2]

▪ RESUMO: Neste trabalho, três modelos foram apresentados para analisar dados obtidos a partir de um ensaio de controle biológico para *Diatraea saccharalis*, uma praga comum em plantações de cana-de-açúcar. Usando-se a distribuição binomial como modelo de probabilidade, um ajuste adequado não pôde ser obtido, devido à variabilidade extra-Binomial e pelo excesso de zeros. O uso do modelo beta-binomial permitiu incorporar parte dessa variabilidade adicional enquanto que o modelo beta-binomial inflacionado de zeros (ZIBB) permitiu modelar também o excesso de zeros. Para a estimação dos parâmetros foi utilizada a abordagem Bayesiana aliada às técnicas de simulação Monte Carlo com Cadeias de Markov (MCMC), através do algoritmo Metropolis-Hastings e o DIC (*Deviance Information Criterion*) para seleção de modelos. A convergência da cadeia foi monitorada através da análise gráfica dos valores gerados e dos diagnósticos apropriados implementados no módulo CODA, disponíveis no *software* R (Cowles and Carlin, 1995).

▪ PALAVRAS-CHAVE: Análise Bayesiana; modelo binomial; superdispersão; excesso de zeros; métodos MCMC.

1 Introdução

O controle biológico aplicado, um importante ramo da entomologia, busca métodos para controle de pragas sem prejudicar o meio ambiente e o ecossistema, utilizando inimigos naturais da praga. A broca-da-cana (*Diatraea saccharalis*), principal praga da cana-de-açúcar no Brasil, vem sendo combatida por parasitóide

[1]Departamento de Informática e Estatística, Universidade Federal de Santa Catarina - UFSC, CEP 88040-900, Florianópolis, SC, Brasil. E-mail: *borgatto@inf.ufsc.br*
[2]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo - ESALQ/USP, CEP 13418-900, Piracicaba, SP, Brasil. E-mail: *clarice@carpa.ciagri.usp.br* / *raleandr@carpa.ciagri.usp.br*

de ovos, evitando assim os danos causados pela lagarta, além de reduzir a infestação da praga na própria safra em que é liberado. A motivação deste trabalho surgiu a partir de dados de um ensaio de controle biológico cujo objetivo era relacionar a proporção de ovos parasitados com números diferentes de fêmeas do parasitóide para otimização da produção, em laboratório, de ovos parasitados com *Trichogramma galloi* para liberações inundativas no campo.

Para a análise desses dados foram considerados três modelos, sendo seus parâmetros estimados sob o enfoque Bayesiano que permite incorporar informações prévias a respeito dos parâmetros. Inicialmente, considerou-se o modelo de regressão binomial logística. Entretanto, dados desse tipo, em geral, apresentam uma variabilidade maior do que a esperada pelo modelo padrão (superdispersão) em função da variabilidade da probabilidade de parasitismo (por algum tipo de correlação ou diferença entre indivíduos) e da presença de um excesso de zeros (Hinde e Demétrio, 1998, Ridout et al., 1998, Vieira et al., 2000, Hall, 2000 e Brooks et al., 1997). Modelos alternativos foram o modelo beta-binomial, que pode incorporar parte dessa variabilidade, e o beta-binomial inflacionado de zeros, pois pode-se supor que há uma mistura de populações (fêmeas com e sem capacidade de parasitar), mas que uma só produz zeros e a outra, contagens não-negativas.

2 Ensaio de controle biológico

Os dados deste trabalho são provenientes de um ensaio completamente casualizado com 10 repetições, realizado no Departamento de Entomologia da ESALQ/USP. Fêmeas (2, 4, 8, 16, 32, 64 e 128) do parasita *Trichogramma galloi*, foram colocadas para parasitar 128 ovos de *Anagasta kuehniella*, um hospedeiro alternativo (mais econômico) a *D. saccharalis*. A variável resposta observada foi o número de ovos parasitados dentre os 128 ovos, sendo que o interesse estava na determinação do número necessário de fêmeas para produzir o número máximo de ovos parasitados.

A distribuição de frequências do número de ovos parasitados é apresentada na Tabela 1.

Tabela 1 - Frequências observadas do número de ovos parasitados em relação ao número de fêmeas

Número de ovos	Número de fêmeas do *T. galloi*						
parasitados	2	4	8	16	32	64	128
0	5	7	5	5	5	6	4
1 a 25	0	2	2	0	0	0	1
26 a 50	3	0	1	2	0	0	2
50 a 75	2	1	2	3	0	1	0
75 a 100	0	0	0	0	3	2	3
100 a 125	0	0	0	0	2	1	0

Essa distribuição e um diagrama de dispersão da proporção de ovos parasitados de *A. kuehniella* em relação ao número de fêmeas (Figura 1) de *T. galloi* mostraram que existe uma grande variabilidade dos dados e um excesso de zeros, provocados, provavelmente, pela variabilidade natural dos ovos dentro de cada grupo e pela variabilidade entre fêmeas.

3 Modelos

Seja Y_i a variável aleatória "número de ovos parasitados"em m_i ovos, com observações $(y_i,\ m_i)$ e $p_i = \dfrac{y_i}{m_i}$ representando a proporção de ovos parasitados, $i = 1, 2, \cdots, n$, sendo n o número de unidades experimentais.

Modelo binomial

Assume-se, inicialmente, que a variável aleatória Y_i tem distribuição binomial, com

$$P(Y_i = y_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i - y_i}, \qquad 0 \leq p_i \leq 1$$

e que p_i está relacionada às variáveis explanatórias através de $g(p_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, sendo $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)$ um vetor de parâmetros a ser estimado, $\boldsymbol{x}_i^T = (1, x_i, x_i^2, \ldots, x_i^q)$ o vetor de variáveis explanatórias correspondente à unidade experimental i, com polinômio de grau q e $g(.)$ uma função de ligação adequada. A esperança e a variância de Y_i são, respectivamente, $E(Y_i) = m_i p_i$ e $Var(Y_i) = m_i p_i(1-p_i)$.

Para o modelo de regressão binomial logística admitiu-se que

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \ldots + \beta_q x_i^q, \tag{1}$$

sendo β_j, $j = 0, 1, \ldots, q$, parâmetros a serem estimados, $x_i = z_i - \bar{z}$, $i = 1, 2, \ldots, n$, $z_i = \log_2$ (número de fêmeas), $\bar{z} = \sum_{i=1}^{n} \dfrac{\log_2 \text{ (número de fêmeas)}}{\text{número de fêmeas}}$.

Sob o enfoque Bayesiano, assumiu-se que os parâmetros β_j têm distribuições de probabilidade a priori $N(\mu_j, \sigma_j^2)$, isto é,

$$\beta_j \sim N(\mu_j, \sigma_j^2);\ j = 0, 1, \ldots, q \tag{2}$$

com as constantes μ_j e σ_j^2 conhecidas. Em particular, como não existem informações do pesquisador a respeito desses parâmetros, admitiram-se prioris vagas, com os valores dos hiperparâmetros dados por $\mu_j = 0$ e $\sigma_j^2 = 10^2$.

Considerando-se a função de verossimilhança para o modelo com covariáveis e independência entre os parâmetros, obteve-se a distribuição conjunta a posteriori $\pi(\beta_0, \ldots, \beta_q \mid \boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{m})$ e a partir dela as distribuições condicionais completas a posteriori para os parâmetros β_j dadas por:

$$\pi(\beta_j \mid \boldsymbol{\beta}_{-j}, \boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \exp\left\{\frac{-1}{2\sigma_j^2}(\beta_j - \mu_j)^2\right\} \prod_{i=1}^{n}\left[\frac{e^{y_i x_i \beta}}{(1+e^{x_i\beta})^{m_i}}\right], \tag{3}$$

sendo $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)$ e $\boldsymbol{\beta}_{-j} = (\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_q)$.

Modelo beta-binomial

Uma possibilidade para incorporar a dispersão extra-Binomial apresentada pelos dados é assumir uma distribuição de probabilidade para o parâmetro p_i da distribuição binomial (Hinde e Demétrio, 1998, Collett, 2002). Supondo que $Y_i \mid p_i \sim Bin(m_i, p_i)$ e $p_i \sim Beta(a_i, b_i)$, $a_i > 0$ e $b_i > 0$, a função de probabilidade para a variável Y_i é dada por

$$P(Y_i = y_i) = \binom{m_i}{y_i} \frac{\Gamma(a_i + y_i)\Gamma(m_i + b_i - y_i)\Gamma(a_i + b_i)}{\Gamma(m_i + a_i + b_i)\Gamma(a_i)\Gamma(b_i)}. \tag{4}$$

Fazendo-se $p_i = \dfrac{a_i}{a_i + b_i}$ e $\delta_i = \delta = \dfrac{1}{a_i + b_i}$ (impondo-se que $a_i + b_i$ é constante; Williams, 1982), o modelo beta-binomial dado em (4) fica

$$P(Y_i = y_i) = \binom{m_i}{y_i} \frac{\Gamma\left(\frac{p_i}{\delta} + y_i\right)\Gamma\left(\frac{1-p_i}{\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{p_i}{\delta}\right)\Gamma\left(\frac{1-p_i}{\delta}\right)}$$

com esperança $E(Y_i) = m_i p_i$ e variância dada por

$$Var(Y_i) = m_i p_i (1 - p_i) \left(\frac{m_i \delta + 1}{\delta + 1}\right). \tag{5}$$

O termo $\left(\dfrac{m_i \delta + 1}{\delta + 1}\right)$ em (5) é sempre maior do que 1, pois $\delta > 0$ e quando $\delta \to 0$ a variância do modelo beta-binomial tende à variância do modelo binomial.

Admitiu-se, ainda, função de ligação logística e preditor linear dado por (1) e para o enfoque Bayesiano, as distribuições a priori utilizadas para os parâmetros β_j são as mesmas descritas em (2) e com os mesmos hiperparâmetros e considerou-se também

$$\delta \sim Gama(c, d), \tag{6}$$

com as constantes c e d conhecidas. De forma semelhante ao caso anterior, consideram-se prioris vagas com $c = 1$ e $d = 0,01$.

Considerando-se a função de verossimilhança para o modelo com covariáveis e independência entre os parâmetros, obteve-se a distribuição conjunta a posteriori $\pi(\beta_0, \ldots, \beta_q, \delta \mid \boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{m})$ e a partir dela obtiveram-se as distribuições condicionais completas a posteriori para β_j e δ dadas por:

$$\pi(\beta_j \mid \delta, \boldsymbol{\beta}_{-j}, \boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \exp\left\{\frac{-1}{2\sigma_j^2}(\beta_j - \mu_j)^2\right\} \prod_{i=1}^{n} \left[\frac{\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta} + y_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta}\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right] \tag{7}$$

para $j = 0, 1, \ldots, q$ e

$$\pi(\delta \mid \boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \delta^{c-1} \exp\{-d\delta\} \prod_{i=1}^{n} \left[\frac{\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta} + y_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta}\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)} \right]. \quad (8)$$

Modelo ZIBB

A explicação para o excesso de zeros pode estar na incapacidade natural de algumas fêmeas em parasitarem ovos do hospedeiro alternativo, caracterizando uma mistura de fêmeas de duas populações diferentes. Seguindo Ghosh et al. (2006), a variável aleatória Y no modelo beta-binomial inflacionado de zeros (ZIBB) pode ser representada por $Y = V(1-B)$ em que B é uma variável aleatória com distribuição Ber(w) e V é uma variável aleatória com distribuição beta-binomial, sendo que w representa a probabilidade de uma fêmea não ser capaz de parasitar o hospedeiro alternativo. Mostrou-se que a função de probabilidade para Y_i fica dada por:

$$P(Y_i = y_i) = \begin{cases} w_i + (1-w_i)\frac{\Gamma\left(\frac{1}{\delta}\right)\Gamma\left(\frac{1-p_i}{\delta}+m_i\right)}{\Gamma\left(\frac{1}{\delta}+m_i\right)\Gamma\left(\frac{1-p_i}{\delta}\right)} & y_i = 0 \\ (1-w_i)\binom{m_i}{y_i}\frac{\Gamma\left(\frac{p_i}{\delta}+y_i\right)\Gamma\left(\frac{1-p_i}{\delta}+m_i-y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta}+m_i\right)\Gamma\left(\frac{p_i}{\delta}\right)\Gamma\left(\frac{1-p_i}{\delta}\right)} & y_i > 0 \end{cases}$$

em que $0 \leq w_i < 1$, $0 \leq p_i < 1$ e $\delta > 0$.

A esperança e variância desse modelo são dadas, respectivamente, por $E(Y_i) = (1-w_i)m_ip_i = \mu_i$ e $Var(Y_i) = \mu_i\left[(1-p_i)\left(\frac{m_i\delta+1}{\delta+1}\right) + \left(\frac{w_i}{1-w_i}\right)\mu_i\right]$.

Para o caso particular em que $w_i = 0$, esse modelo reduz-se ao beta-binomial. Admitiu-se, ainda, função de ligação logística para ambos os modelos e preditor linear dado por (1) para o modelo beta-binomial, enquanto que para o modelo Bernoulli

$$g(w_i) = \log\left(\frac{w_i}{1-w_i}\right) = \gamma_0 + \gamma_1 g_i + \ldots + \gamma_{q_2} g_i^{q_2},$$

sendo γ_k o coeficiente do modelo de regressão para $k = 0, 1, \ldots, q_2$, $g_i = (z_i - \bar{z})$, $i = 1, 2, \ldots, 70$.

Para o enfoque Bayesiano, as distribuições a priori utilizadas para os parâmetros β_j e δ são as mesmas descritas em (2) e (6), respectivamente. Para os parâmetros γ_k usou-se

$$\gamma_k \sim N(\eta_k, \nu_k^2);\ k = 1, \ldots, q_2. \quad (9)$$

com as constantes η_k e ν_k^2 conhecidas. Como não existem informações do pesquisador a respeito desses parâmetros, admitiram-se prioris vagas, com os valores dos hiperparâmetros dados por $\mu_j = \eta_k = 0$, $\sigma_j^2 = \nu_k^2 = 10^2$, $c = 1$ e $d = 0,01$.

Considerando-se a função de verossimilhança para o modelo com covariáveis, independência dos parâmetros e H_i uma variável latente, com $h_i = 1$ se $y_i = 0$ e $h_i = 0$ se $y_i > 0$, obteve-se a distribuição conjunta a posteriori $\pi(\beta_0, \ldots, \beta_q, \gamma_k, \delta \mid \boldsymbol{x}_i, \boldsymbol{g}_i, \boldsymbol{y}, \boldsymbol{m})$ e a partir dela obtiveram-se as distribuições condicionais completas a posteriori para β_j, γ_k e δ dadas por:

$$\pi(\beta_j \mid \boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \delta, \boldsymbol{x}_i, \boldsymbol{g}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \exp\left\{\frac{-1}{2\sigma_j^2}(\beta_j - \mu_j)^2\right\}$$
$$\left\{\prod_{i=1}^{n} h_i \left[\frac{1}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right] + \right.$$
$$\left. + \prod_{i=1}^{n}(1-h_i)\left[\frac{e^{g_i\gamma}}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta} + y_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta}\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right]\right\}$$

sendo $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{q_2})$,

$$\pi(\gamma_k \mid \boldsymbol{\gamma}_{-k}, \boldsymbol{\beta}, \delta, \boldsymbol{x}_i, \boldsymbol{g}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \exp\left\{\frac{-1}{2\nu_k^2}(\gamma_k - \eta_k)^2\right\}$$
$$\left\{\prod_{i=1}^{n} h_i \left[\frac{1}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right] + \right.$$
$$\left. + \prod_{i=1}^{n}(1-h_i)\left[\frac{e^{g_i\gamma}}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta} + y_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta}\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right]\right\}$$

sendo $\boldsymbol{\gamma}_{-k} = (\gamma_0, \ldots, \gamma_{k-1}, \gamma_{k+1}, \ldots, \gamma_{q_2})$ e

$$\pi(\delta \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{x}_i, \boldsymbol{g}_i, \boldsymbol{y}, \boldsymbol{m}) \propto \delta^{c-1} \exp\{-d\delta\}$$
$$\left\{\prod_{i=1}^{n} h_i \left[\frac{1}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right] + \right.$$
$$\left. + \prod_{i=1}^{n}(1-h_i)\left[\frac{e^{g_i\gamma}}{(1+e^{g_i\gamma})} \frac{\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta} + y_i\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta} + m_i - y_i\right)\Gamma\left(\frac{1}{\delta}\right)}{\Gamma\left(\frac{1}{\delta} + m_i\right)\Gamma\left(\frac{e^{x_i\beta}}{(1+e^{x_i\beta})\delta}\right)\Gamma\left(\frac{1}{(1+e^{x_i\beta})\delta}\right)}\right]\right\}.$$

Obtenção de uma amostra da distribuição a posteriori conjunta

As distribuições a posteriori condicionais completas dadas em (3), (7), (8), (**??**), (10) e (10) não possuem forma fechada e por esse motivo foi usado o algoritmo Metropolis-Hastings para obtenção de uma amostra da respectiva distribuição conjunta e, através dela, da distribuição marginal a posteriori de cada parâmetro (Casella e George, 1992) com estimativas dadas pela média de cada amostra.

A implementação computacional foi feita usando-se o procedimento IML (*Iteractive Matrix Linear*) do programa SAS (*Statistical Analysis System*, Sas Institute, 2000) gerando uma cadeia com 100.000 valores para cada parâmetro, desprezando os primeiros 2.000 valores e selecionando um a cada 40 dos 98.000 valores restantes, formando uma amostra de 2.400 valores.

A convergência das estimativas dos parâmetros foi monitorada através da análise gráfica dos valores gerados além de terem sido feitos os testes de diagnósticos de Raftery & Lewis (1992) e de Heidelberger & Welch (1983), implementados no módulo CODA do programa R.

Seleção de modelos

Para a seleção de modelos foi usado o critério proposto por Spiegelhalter et al. (2002), que é uma generalização do critério de informação de Akaike, com base na distribuição a posteriori da estatística *deviance* e é dado por

$$DIC = \bar{D}(\mu, y) + p_D,$$

sendo que $p_D = \bar{D}(\mu, y) - D(\bar{\mu}, y)$ representa o número efetivo de parâmetros, $\bar{D}(\mu, y) = E_{\mu|y}[D(\mu, y)]$ e $D(\bar{\mu}, y)$ é a *deviance* obtida fixando-se as estimativas dos parâmetros da esperança a posteriori. Um menor valor do DIC indica um modelo melhor ajustado.

4 Resultados e discussões

Os três modelos propostos foram ajustados aos dados, variando-se os graus dos polinômios, tendo-se verificado graficamente que as condições para convergência foram satisfeitas. Considerou-se, no entanto, uma constante como preditor linear para o modelo para a proporção de zeros (w), devido ao fato que a quantidade de zeros é constante para todos os tratamentos (Figura 1).

Na Tabela 2 estão apresentados os valores do DIC para os diversos modelos ajustados. Observa-se que o modelo ZIBB, tendo os polinômios de terceiro e quarto graus como preditores lineares para as proporções de ovos parasitados e uma constante para os zeros, apresentaram valores de DIC inferiores aos demais modelos, indicando melhor ajuste. Entretanto, os valores do DIC desses dois modelos estão muito próximos o que permite optar-se pelo modelo com menor número de parâmetros.

Nota-se, através do gráfico dos valores preditos versus o logaritmo do número de fêmeas (Figura 2), usando o modelo ZIBB com polinômio de terceiro grau como preditor linear para as proporções de ovos parasitados e uma constante para as proporções de zeros, que os valores médios observados estão dentro do intervalo de credibilidade a 95%. Portanto, existem evidências de que esse modelo ajusta-se adequadamente ao conjunto de dados. As estimativas de seus parâmetros são mostradas na Tabela 3.

Tabela 2 - Valores do DIC para os modelos binomial, beta-binomial e ZIBB

Modelos	Preditores lineares	DIC
	$logit(p) = \beta_0$	$4.747,50$
	$logit(p) = \beta_0 + \beta_1 x_i$	$4.538,60$
binomial	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$	$4.538,60$
	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$	$4.375,82$
	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$	$4.295,41$
	$logit(p) = \beta_0$	$446,42$
	$logit(p) = \beta_0 + \beta_1 x_i$	$447,24$
beta	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$	$449,15$
binomial	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$	$440,55$
	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$	$442,53$
	Fixando-se: $logit(w) = \gamma_0$	
	$logit(p) = \beta_0$	$417,95$
	$logit(p) = \beta_0 + \beta_1 x_i$	$413,47$
ZIBB	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$	$414,54$
	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$	$398,00$
	$logit(p) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4$	$398,03$

Tem-se, então, que

$$g(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -0,1844 + 2,2370x_i - 0,0885x_i^2 - 0,9714x_i^3 \qquad (10)$$

e

$$g(\hat{w}_i) = \log\left(\frac{\hat{w}_i}{1-\hat{w}_i}\right) = 0,0972.$$

De (10) tem-se que

$$\hat{p}_i = \frac{exp\{-0,1844 + 2,2370x_i - 0,0885x_i^2 - 0,9714x_i^3\}}{1 + exp\{-0,1844 + 2,2370x_i - 0,0885x_i^2 - 0,9714x_i^3\}}$$

que maximizada em relação a x_i resulta que o número de fêmeas do parasita *T. galloi* que dá o número máximo de ovos parasitados é $\zeta_{max} = 2^{x_{max}+\bar{z}} = 52$. A probabilidade de uma fêmea não parasitar é dada por

$$\hat{w} = \frac{e^{0,0972}}{1 + e^{0,0972}} = 0,5243.$$

Análise de sensibilidade

Foram adotadas diferentes distribuições a priori para os parâmetros de interesse para o modelo ZIBB: N(0,50), N(0,25), N(0,10), N(0,5) e N(0,1) para os parâmetros β_0, β_1, β_2, β_3 e γ_0 enquanto que para o parâmetro δ foram utilizadas as distribuições Gama(1,0.1) e Gama(1,1). Foi gerada uma amostra com 2.400 valores

Tabela 3 - Estimativas dos parâmetros para o modelo ZIBB, tendo um polinômio de terceiro grau como preditor linear para as proporções de ovos parasitados e uma constante para as proporções de zeros

			Intervalos de credibilidade	
Parâmetros	Médias	Desvios Padrões	$2,5\%$	$97,5\%$
β_0	-0,1844	0,2149	-0,6185	0,2290
β_1	2,2370	0,4236	1,4264	3,1029
β_2	-0,0885	0,1545	-0,4034	0,2215
β_3	-0,9714	0,2154	-1,4091	-0,5436
γ_0	0,0972	0,2376	-0,3699	0,5651
δ	0,1698	0,0491	0,0967	0,2870

para cada combinação entre elas. Observou-se pelo comportamento das estimativas a posteriori dos parâmetros, que não houve sensibilidade às prioris (informativas e vagas) adotadas.

Conclusões

Neste artigo, foram usados modelos para dados na forma de proporção com superdispersão e excesso de zeros. Usando-se como critério a estatística DIC, verificou-se que o modelo que melhor se ajustou aos dados foi o ZIBB, tendo um polinômio de terceiro grau como preditor linear para as proporções e uma constante para as proporções de zeros. A partir desse modelo, obteve-se que 52 fêmeas do parasita *T. galloi* é o número ideal para maximizar o número de ovos parasitados de *A. kuehniella*. Estudos adicionais são necessários considerando formas mais complexas para a superdispersão.

Agradecimentos

Os autores agradecem ao CNPQ e à FAPESP pelo suporte financeiro.

BORGATTO, A. F.; DEMÉTRIO, C. G. B.; LEANDRO, R. A. Models for zero-inflated and overdispersed proportion data - A Bayesian approach. *Rev. Mat. Estat.,* São Paulo, v.24, n.2, p.121-131, 2006.

▪ *ABSTRACT: In general, standard binomial regression models do not fit well to proportion data from biological control assays, mainly when there is excess of zeros and overdispersion. In this study, beta-binomial and zero-inflated beta-binomial models are applied to a data set obtained from a biological control assay to produce parasitized eggs to control Diatraea saccharalis, a common pest in sugar cane. A parasite (Trichogramma galloi) was put to parasitize 128 eggs of Anagasta kuehniella, an economically suitable alternative host, with a variable number of female parasites (2, 4, 8,..., 128), each with 10 replicates in a completely randomized experiment. A Bayesian procedure was formulated using a simulation technique (Metropolis Hastings) for estimation of the parameters of interest. The convergence of the Markov Chain generated was monitored by visualization of the trace plot and using Raftery & Lewis and Heidelberg & Welch diagnoses presented in module CODA of software R.*

▪ *KEYWORDS: Generalized linear models; Bayesian analysis; binomial model; overdispersion; excess of zeros; MCMC method.*

Referências

BROOKS, S. P.; MORGAN, B.J.T.; RIDOUT, M. S.; PACK, S. E. Finite mixture models for proportions. *Biometrics*, Washington, v.53, p.1097 – 1115, 1997.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs sampler. *Am. Stat.*, Washington, v.46, p.167 – 170, 1992.

COLLETT, D. *Modelling binary data*. 2nd ed. London: Chapman-Hall, 2002. 387p.

COWLES, M. K.; CARLIN, B. P. Markov chain Monte Carlo diagnostic: A comparative review. *J. Am. Stat. Assoc.*, New York, v.91, p.883 – 904, 1995.

GHOSH, S. K.; MUKHOPADHYAY, P.; LU, J. C. Bayesian analysis of zero-inflated regression models. *J. Stat. Plann. Infer.* Amsterdam, v.136, p.1360 – 1375, 2006.

HALL, D. B. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, Washington, v.56, p.1030 – 1039, 2000.

HEIDELBERGER, P.; WELCH P. D. Simulation run length control in the presence of an initial transient. *Oper. Res.*, Baltimore, v.31, p.1109 – 1144, 1983.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. *Comp. Stat. Data Anal.*, Amsterdam, v.27, p.151 – 170, 1998.

RAFTERY, A. E.; LEWIS, S. M. One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Stat. Sci.*, Hayward, v.7, p.493 – 497, 1992.

RIDOUT, M. S.; DEMÉTRIO, C. G. B.; HINDE, J. P. Models for count data with many zeros. in: INTERNATIONAL BIOMETRIC CONFERENCE, 1998, Cape Town. *Proceedings...*, Cape Town: IBC, 1998, p.179 – 192

SAS INSTITUTE. *SAS/IML:user´s guide, version 8*. Cary, 2000. Disponível em: <http://www.sfu.ca/sasdoc/sashtml/iml/index.htm >.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P., LINDE, A. V. D. Bayesian measures of model complexity and fit. *J. R. Stat. Soc., Ser. B*, London, v.64, p. 583–639, 2002.

VIEIRA, A. M. C., HINDE, J. ; DEMÉTRIO, C. G. B. Zero-inflated proportion data models applied to a biological control assay. *J. Appl. Stat.*, Abingdon, v.27 n.3, p.373 – 389, 2000.

WILLIAMS, D. A. Extra-binomial variation in logistic linear models. *J. Appl. Stat.*, Abingdon, v.31, p.144 – 148, 1982.

Recebido em 09.05.2004.

Aprovado após revisão em 24.06.2006.

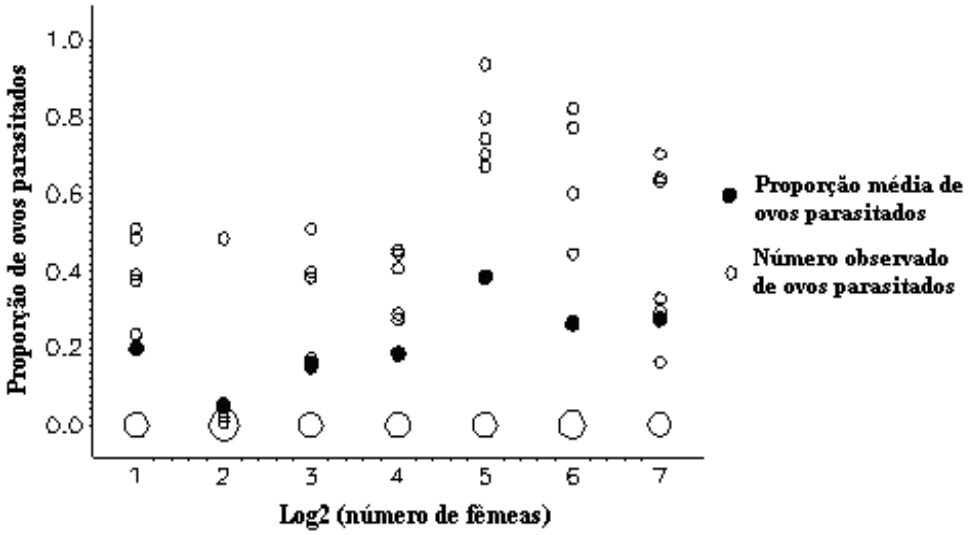

Figura 1 - Proporção de ovos de *A. kuehniella* parasitados, em função do logaritmo do número de fêmeas do *T. galloi* (o diâmetro do círculo é proporcional à grandeza da frequência).

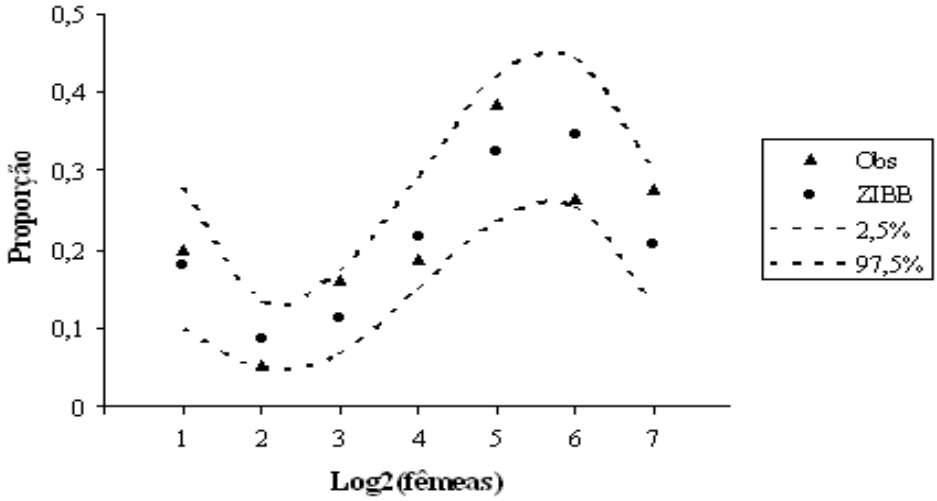

Figura 2 - Proporções observadas, preditas e intervalos de credibilidade a 95% de probabilidade para o modelo ZIBB com preditor linear para as proporções o polinômio de terceiro grau e para os zeros uma constante.