Modelos para analisar dados de contagens com superdispersão em uma plantação de morango

Sidcleide Barbosa de Sousa [1], **Clarice Garcia Borges Demétrio** [2], **Fernanda Canassa** [3], **Italo Delalibera Junior** [3]

Introdução

Resultados experimentais de dados na forma de contagens são bastante recorrentes em diversas áreas da Ciência e os modelos lineares generalizados, propostos por Nelder e Wedderburn (1972), viabilizam um leque de opções para a variável resposta. A generalidade dessa classe de modelos e eficiência computacional do algoritmo de estimação tornaram-se referência metodológica para análise desse tipo de dados.

Para dados de contagens, é natural assumir a distribuição de Poisson como referência probabilística na construção de modelos. No entanto, a relação de igualdade entre média e variância dessa distribuição e a ausência de um parâmetro para quantificar a variabilidade extra-Poisson, torna-se necessária a utilização de modelos específicos para acomodar o efeito de superdispersão. Isso pode ocorrer por diversas razões como heterogeneidade das unidades experimentais, ausência de covariáveis experimentais, correlação entre as observações, excesso de zeros entre outras (HINDE; DEMÉTRIO, 1998).

Para análise de dados de contagens superdispersos existem várias alternativas. Um dos modelos a ser considerado é o quase-Poisson, o qual é baseado em suposições de primeiro e segundo momentos para acomodar a variabilidade extra. Entretanto, em muitos casos, esse modelo também pode não ser adequado. Uma abordagem alternativa são os modelos de dois estágios que é o modelo Poisson com efeito aleatório gama, conhecido como modelo binomial negativo.

Esse trabalho teve como objetivo utilizar modelos para dados de contagens, que possam acomodar a superdispersão.

Material e Métodos

Os dados a serem utilizados são provenientes de um experimento realizado no Laboratório de Patologia e Controle Microbiano de Insetos do Departamento de Entomologia da ESALQ/USP, cujo objetivo foi estudar os diferentes tipos de isolados *Metarhizium spp.*, *B. bassiana*, *I. fumosorosea* em uma plantação de morango.

Os ensaios foram conduzidos em casa-de-vegetação, os fungos entomopatogênicos foram inoculadas nas raízes das plantas de morango e após a imersão das raízes, as mudas foram transplantadas e cultivadas em vasos. Nas plantas, foi feita a adubação de acordo com as exigências nutricionais de morango para cultivo em vasos e a irrigação foi feita por gotejamento, realizada diariamente e avaliado, o ciclo do morangueiro é de seis meses.

O experimento analisado foi com 25 isolados, sendo 15 de *Metarhizium spp.*, 5 de *B. bassiana* e 5 de *I. fumosorosea*, usando o delineamento casualizado em blocos, com cinco repetições.

[1]Sidcleide Barbosa de Sousa. e-mail: sbarbosas1987@usp.br *email*
[2]Clarice Garcia Borges Demétrio. e-mail: clarice.demetrio@usp.br *email*
[3]Fernanda Canassa. e-mail: Fernanda.canassa@usp.br *email*
[3]Italo Delalibera Junior. e-mail: delalibera@usp.br *email*

Sessenta dias após a inoculação dos isolados nas raízes das plantas de morango, um inseto fêmea criada em laboratório foi colocada no folheto de cinco plantas de forma aleatória. Sete dias após a infestação das fêmeas cada folheto infestado foi descolado e o número de ovos foi contado com intuito de avaliar os efeitos benéficos às plantas de morango inoculadas com diferentes isolados e verificar se o mesmo inibia a fêmea de multiplicar seus ovos.

Modelos Lineares Generalizados MLGs

A classe de modelos lineares generalizados (MLG) é uma importante extensão de modelos e métodos conhecidos como os modelos lineares, modelos log-lineares, entre outros. Os MLGs envolvem três componentes: o primeiro é o componente aleatório representado pelo conjunto de variáveis aleatórias $Y_i, \ldots, Y_n$, proveniente da mesma distribuição, pertencente à família exponencial na forma canônica com função de probabilidade ou função densidade de probabilidade expressa por

$$f(\mathrm{y}_i, \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \tag{1}$$

em que, $\phi > 0$ é um parâmetro de dispersão do modelo e a sua inversa ϕ^{-1} é um parâmetro de precisão, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas e θ_i é chamado de parâmetro canônico.

Demonstra-se que a média e a variância são, respectivamente, $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ e $\mathrm{Var}(Y_i) = \phi b"(\theta_i) = \phi \mathrm{V}_i$ sendo, $\mathrm{V}_i = \mathrm{V}(\mu_i) = d\mu_i/d\theta_i$ chamada de função de variância, que depende apenas da média μ_i. O segundo componente é o preditor linear, o qual relaciona as variáveis explanatórias no modelo, isto é,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

em que, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$, sendo $\boldsymbol{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)'$ a matriz do modelo e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ o vetor de parâmetros desconhecidos. O terceiro componente é a função de ligação, que relaciona a média μ_i da variável aleatória ao preditor linear η_i (DEMÉTRIO *et al.*, 2014), ou seja,

$$\eta_i = g(\mu_i).$$

Com base nessa metologia, serão discutidos os modelo para dados de contagens.

Modelo Poisson

Um caso particular do MLGs é quando se assume que $Y \sim \mathrm{Poisson}(\lambda)$ com função de probabilidade (fp)

$$\mathrm{P}(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, y = 0, 1, 2, \ldots$$

em que, λ é a taxa de eventos e tem-se que

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu = \lambda. \tag{2}$$

e a função de ligação logarítmica $g(\mu) = \log(\lambda) = \boldsymbol{x'\beta} = \boldsymbol{\eta}$. Na forma da família exponencial canônica tem-se que $\theta = \log(\lambda)$, $\phi = 1$ e $\mathrm{V}(\mu) = \lambda = \mu$.

A função "deviance" do modelo Poisson é dada por

$$D_p = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right].$$

em que, $\hat{\mu}_i$, $i = 1, 2, \ldots, n$, são valores ajustados para o modelo de interesse. Uma alternativa de medida de qualidade de ajuste é a estatística de Pearson X^2, que assume a seguinte forma

$$X_p^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\text{Var}}(Y_i)}.$$

Tem-se que D_p e X^2 são equivalentes e, assintoticamente, ambas tem distribuição χ^2_{n-p} com $(n-p)$ graus de liberdade, sendo n o número de observações e p o número de parâmetros estimados.

Gráficos de diagnósticos são usados para detectar possíveis falhas no modelo, normalmente comparam-se valores observados e ajustados usando-se os componentes da "deviance" residual. O gráfico meio normal mostra como os resíduos comportam-se caso o modelo esteja bem ajustado aos dados. Se os pontos estiverem fora dos envelopes de simulação, o modelo não é adequado. Para que o modelo tenha um bom ajuste o número de pontos fora do envelope de simulação não ultrapasse 5%, esse gráfico está implementado no pacote `hnp` no R (MORAL *et al.*, 2017).

Na prática, o modelo de Poisson pode não fornecer um bom ajuste para dados de contagens reais, ou seja, quando a variabilidade dos dados é maior do que a média, esse modelo não é capaz de acomodar a superdispersão. Então, extensões do modelo Poisson podem ser usadas para analisar dados com superdispersão, em que o $\phi > 1$. Uma abordagem inicial para acomodar essa variabilidade extra é considerar o modelo quase Poisson, outra é usar a distribuição binomial negativa.

Modelo Quase Poisson

Uma forma de incorporar a superdispersão é por meio da quase verossimilhança, que requer apenas a especificação do primeiro e segundo momentos da distribuição, ou seja, a média e a variância. Para o modelo linear generalizado Poisson, a equação (2) é substituída por

$$\text{Var}(Y_i) = \phi\mu_i.$$

em que, ϕ é chamado parâmetro de dispersão, o mesmo é desconhecido e sua estimação é dada por

$$\hat{\phi} = \frac{X_p^2}{n-p}$$

sendo $X_p^2 = \sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}$ é a estatística de Pearson do modelo Poisson.

Modelo binomial negativo

Uma alternativa para explicar a superdispersão é utilizar o modelo com dois estágios, o qual permite que a média da Poisson varie. Com isso, $Y_i/\lambda_i \sim \text{Poisson}(\lambda_i)$ e assumindo que λ_i é uma variável aleatória, sendo $\lambda_i \sim \text{gama}(\alpha, \beta)$ tem-se a distribuição binomial negativa com média e variância dadas por

$$\mathrm{E}(Y_i) = \mu_i$$

e

$$\mathrm{Var}(Y_i) = \mu_i(1 + \frac{1}{\alpha}\mu_i).$$

Resultados e Discussão

Na Figura 1, são apresentadas as médias e as variâncias amostrais para o número de ovos, sendo que a linha tracejada representa a suposição de equidispersão, ou seja, média e variância iguais. A maioria dos pontos estão acima da linha tracejada, indicando que, possivelmente, esses dados apresentam superdispersão.

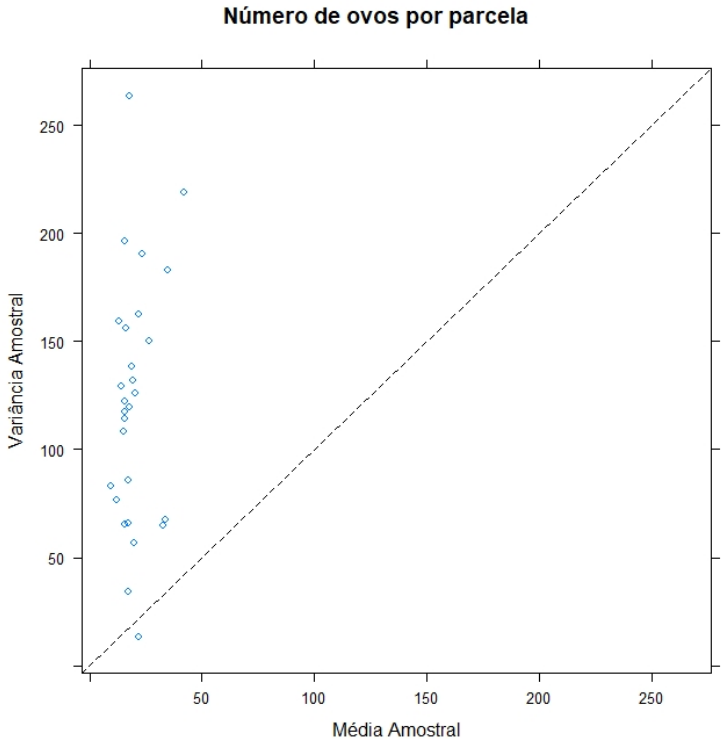

Figura 1: Gráfico de dispersão das médias e variância amostrais calculadas em cada tratamento experimental.

A análise desse conjunto de dados, foi feita utilizando os modelos com função de ligação logarítmica. Inicialmente foi ajustado o modelo Poisson com o preditor linear, dado por

$$\eta_{ij} = \alpha + \beta_{1i} + \beta_{2j}, \quad i = 1, \ldots, 5 \quad \text{e} \quad j = 1, \ldots, 28$$

sendo, α uma constante β_{1i} é o efeito do j-ésimo bloco e β_{2j} é o efeito do i-ésimo isolado.

Por meio dos valores da "deviance" residual e X^2 da tabela 1, há evidência que o modelo Poisson não está bem ajustado. Isso, também, pode ser visto no gráfico meio normal com envelope simulado na Figura 2(a). No modelo Poisson, os resíduos componentes da "deviance" estão completamente fora do envelope de simulação, mostrando evidência que o modelo não foi bem ajustado aos dados. Seguindo a analise com o modelo quase Poisson a estimativa do parâmetro de superdispersão é $\hat{\phi} = 1644,939/248 = 6,6328$. Por meio do

gráfico meio normal na Figura 2(b) observa-se que o modelo não foi capaz de acomodar a variabilidade extra existente.

Tabela 1: Análise da deviance para os dados com plantação de morangueiros , usando o modelo Poisson logarítmica.

Fator de variação	df	Deviance	valor-p	X^2	valor-p
bloco	4	14.94			
Isolado	27	692.23			
Resíduo	248	1704.1	<0.0001	1644.939	<0.0001

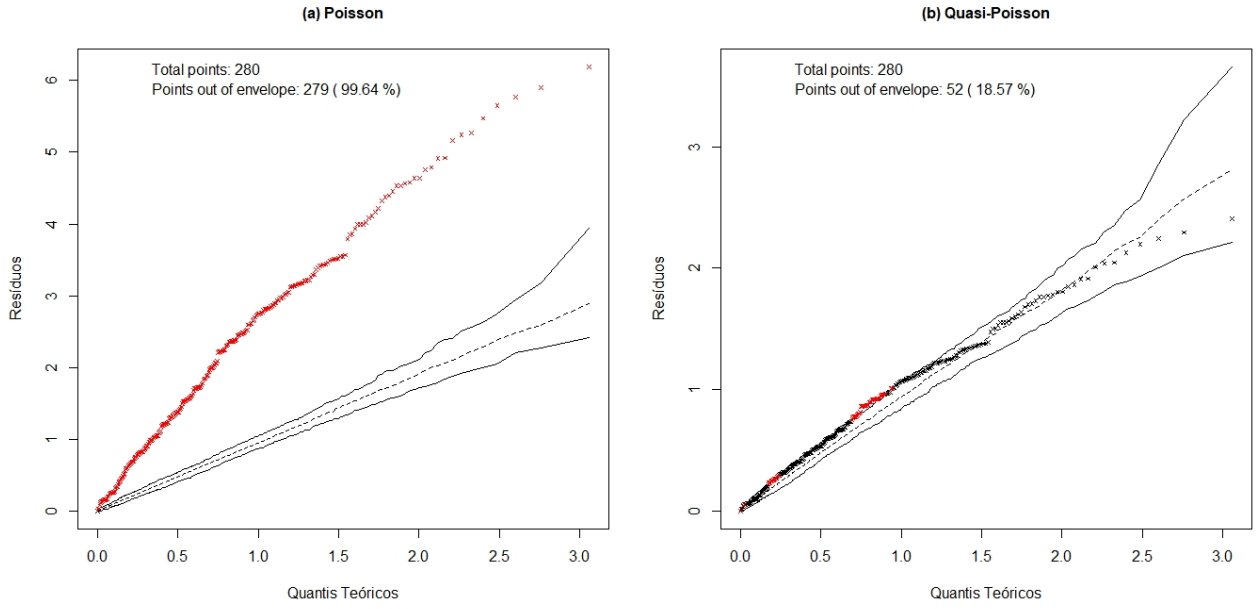

Figura 2: Gráfico meio normal com envelopes de simulação e os resíduos por meio dos modelos (a) Poisson e (b) quase-Poisson.

Ajustando-se o modelo binomial negativo, o valor estimado é $\hat{\theta} = 3.12$, o qual indica uma superdispersão. O gráfico meio normal como mostrado na Figura 3, indica que há evidência que o modelo binomial negativo ajusta-se satisfatoriamente aos dados.

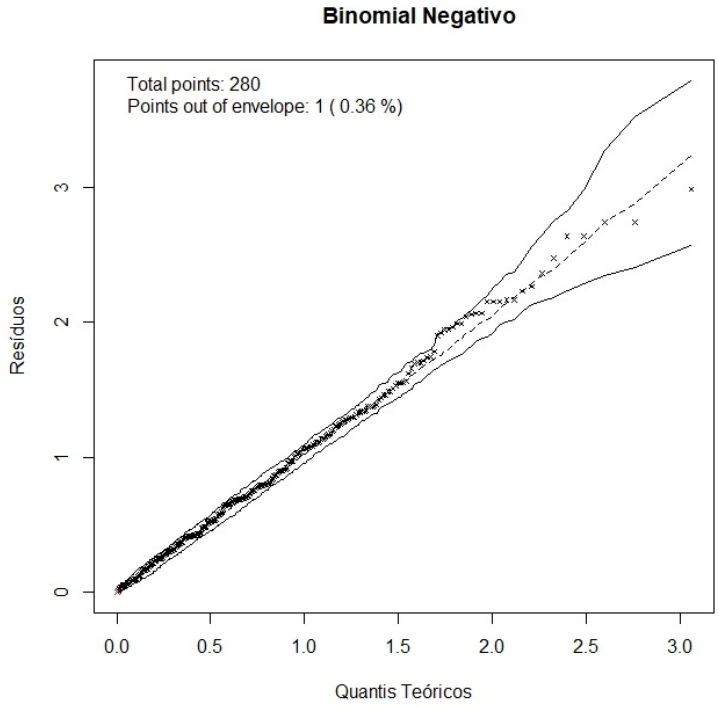

Figura 3: Gráfico meio normal com envelopes de simulação e os resíduos por meio do modelo binomial negativo.

Agradecimentos

Agradecimentos à CAPES e à Esalq pelo apoio financeiro.

Referências

DEMÉTRIO, C. G.; HINDE, J.; MORAL, R. A. Models for overdispersed data in entomology. In: **Ecological modelling applied to entomology**. [S.l.]: Springer, 2014. p. 219–259.

HINDE, J.; DEMÉTRIO, C. G. Overdispersion: models and estimation. **Computational Statistics & Data Analysis**, Elsevier, v. 27, n. 2, p. 151–170, 1998.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. Half-normal plots and overdispersed models in r: The hnp package. **J. Stat. Softw**, v. 81, p. 23, 2017.

NELDER, J.; WEDDERBURN, R. Generalized linear models. **Royal Statistical Society A**, v. 135, n. 3, p. 370–384, 1972.