Estratégias para análise de contagens sub e superdispersas

Eduardo Elias Ribeiro Junior [†, 1, 2]
Clarice Garcia Borges Demétrio [1]

1 Introdução

Importantes avanços na área de análise de dados na forma de contagens têm sido relatados na literatura. Dentre eles, citam-se, principalmente, métodos para modelar diferentes níveis de dispersão, uma vez que a abordagem via modelo linear generalizado de Poisson supõe equidispersão, o que raramente ocorre em dados reais.

O caso mais comum de falha da suposição de equidispersão e, consequentemente, com mais abordagens possíveis para modelagem, é a superdispersão (média $<$ variância). A subdispersão (média $>$ variância) é menos comum na prática, no entanto têm crescido o número de publicações relatando contagens subdispersas.

As alternativas de análise de contagens não equidispersas estão, geralmente, relacionadas com as causas da não equidispersão. Neste artigo são revisados os modelos COM-Poisson, *Gamma-Count*, Poisson generalizada e Poisson-Tweedie, alternativas bastante flexíveis nesse contexto. A aplicação dos modelos é realizada para análise de um estudo em entomologia.

Estudo de caso: danos em milho

Para ilustrar a análise de dados considerando os modelos citados, considera-se um conjunto de dados provenientes de um estudo em entomologia sobre a espécie *Sitophilus zeamais*, principal praga de milho no Brasil. Nesse estudo, registrou-se o número de progênies (insetos emergentes) da praga em placas de petri, após 60 dias de acompanhamento. As placas foram tratadas com diferentes substratos da planta (Folha, Ramo, Semente) e apenas com água (Controle). Esse estudo foi conduzido no delineamento interamente casualizado com 10 repetições de cada tratamento.

[†]Contato: jreduardo@usp.br
[1]Departamento de Ciências Exatas (LCE) - ESALQ-USP
[2]Laboratório de Estatística e Geoinformação (LEG) - UFPR

2 Modelos probabilísticos

Distribuição COM-Poisson

A distribuição COM-Poisson é a principal representante da família de distribuições Poisson ponderadas (WPD) (Del Castillo & Pérez-Casany 1998). Uma variável aleatória Y pertence à família WPD se sua função massa de probabilidade puder ser escrita como

$$\Pr(Y = y) = \frac{w(y)\exp(-\lambda)\lambda^y}{y!\mathrm{E}_\lambda[w(Y)]}, \quad y \in \mathbb{N}, \tag{1}$$

em que $\mathrm{E}_\lambda(\cdot)$ é o valor médio calculado a partir de uma variável aleatória Poisson de parâmetro λ, chamada de constante de normalização; e $w(y)$ é uma função peso, não negativa e tal que $\mathrm{E}_\lambda[w(Y)]$ seja finita. A função peso $w(y) \equiv w(y, \nu)$, pode depender de um parâmetro adicional de tal forma que sub e superdispersão sejam abrangidas. Obtém-se a distribuição COM-Poisson para $w(y, \nu) = (y!)^{1-\nu}$, para $\nu \geq 0$.

Um inconveniente do modelo COM-Poisson é que os momentos média e variância, em geral, não são obtidos em forma fechada. A partir de uma aproximação para a média, Ribeiro Jr et al. (2018) propõem uma reparametrização, em que $\mu = \lambda^{1/\nu} - (\nu - 1)/2\nu$. Neste artigo considera-se a COM-Poisson reparametrizada para média.

Distribuição *Gamma-Count*

A distribuição *Gamma-Count* é uma generalização da distribuição Poisson que resulta da relação da Poisson com a distribuição do tempo entre eventos. Para a distribuição *Gamma-Count*, assume-se que os tempos entre eventos segue o modelo gama, ao passo que para a Poisson os tempos são exponencialmente distribuídos. Seguindo Winkelmann (1995), a função massa de probabilidade de uma variável aleatória Y que segue o modelo *Gamma-Count* pode ser escrita como

$$\Pr(Y = y) = \int_0^T \frac{\kappa^{y\alpha} t^{y\alpha - 1}}{\Gamma(y\alpha)\exp(\kappa t)} dt - \int_0^T \frac{\kappa^{(y+1)\alpha} t^{(y+1)\alpha - 1}}{\Gamma[(y+1)\alpha]\exp(\kappa t)} dt, \tag{2}$$

que não tem forma fechada, a menos do caso particular $\alpha = 1$, quando a distribuição reduz-se à Poisson. Os momentos da distribuição também não são obtidos em forma fechada. Winkelmann (1995) mostra que para intervalos de observação suficientemente grandes, $T \to \infty$, Y é assintoticamente normal com média $\kappa T/\alpha$ e variância $\kappa T/\alpha^2$. Consequentemente, a distribuição *Gamma-Count* é capaz de modelar superdispersão $(0 < \alpha < 1)$ e subdispersão $(\alpha > 1)$.

Distribuição Poisson generalizada

A distribuição Poisson generalizada é resultante de uma forma limite da distribuição binomial negativa generalizada e pode modelar sub e superdispersão (Zamani & Ismail 2012). Existem duas parametrizações bem conhecidas para a distribuição Poisson generalizada. Sob a parametrização de média, a função massa de probabilidade de uma variável aleatória Y que segue a distribuição Poisson generalizada é dada por

$$\Pr(Y = y) = \left(\frac{\mu}{1 + \sigma\mu}\right)^y \frac{(1 + \sigma y)^{y-1}}{y!} \exp\left[-\mu\frac{(1 + \sigma y)}{(1 + \sigma\mu)}\right], \tag{3}$$

para $\mu > 0$ e $\alpha > \min[-\max(y_i^{-1}), -\max(\mu_i^{-1})]$.

Os momentos da distribuição nessa parametrização são $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \mu(1 + \mu\sigma)^2$, que garantem bastante flexibilidade à distribuição, uma vez que a variância é determinada como uma função cúbica de μ.

Aplicações do modelo de regressão Poisson generalizada são pouco reportadas na literatura. Embora bastante flexível, a grande dificuldade desse modelo reside na complicada restrição do espaço paramétrico, que é difícil de se incorporar, de forma eficiente, no processo de estimação.

Distribuição Poisson-Tweedie

A distribuição Poisson-Tweedie é um caso geral dos modelos Poisson hierárquicos, especificados em dois estágios (Jørgensen 1997, Seção 4.6). Sendo $\mathrm{Tw}_p(\mu, \phi)$ a notação para a ditribuição Tweedie, a distribuição Poisson-Tweedie resulta da especificação

$$Y \mid Z \sim \mathrm{Po}(Z) \quad \text{em que} \quad Z \sim \mathrm{Tw}_p(\mu, \phi), \tag{4}$$

que não tem forma fechada para função de probabilidade, exceto para casos especiais. A esperança e a variância de uma variável aleatória Poisson-Tweedie são $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \mu + \phi\mu^p$. Devido à flexibilidade da distribuição Tweedie, a família Poisson-Tweedie também tem importantes casos particulares que incluem as distribuições Hermite ($p = 0$), Neymann tipo-A ($p = 1$), Pólya-Aeppli ($p = 1,5$), binomial negativa ($p = 2$) e Poisson–inversa gaussiana ($p = 3$) (Bonat et al. 2018).

Pela definição (4), modelos Poisson-Tweedie só modelam superdispersão. Bonat et al. (2018) estendem essa distribuição para contemplar subdispersão, adotando apenas a especificação de momentos e permitindo a estimação de $\phi < 0$ (sujeito a $\mathrm{Var}(Y) > 0$). Essa abordagem é análoga aos modelos de quase-verossimilhança e, a menos dos casos particulares, não se conhece a distribuição completa de Y a partir apenas de sua média e variância, o que impossibilita o cálculo de probabilidades, por exemplo.

Estudo das distribuições

Para explorar a flexibilidade dos modelos apresentados, considera-se o índice de dispersão DI= Var(Y)/E(Y). Na Tabela 1, define-se a notação dos modelos bem seus respectivos parâmetros de dispersão, momentos e a formulações para modelos de regressão.

Tabela 1: Modelos probabilísticos para análise de dados de contagem.

	COM-Poisson	*Gamma-Count*	Poisson generalizada	Poisson-Tweedie
Notação	$\text{CMP}(\mu_i, \nu)$	$\text{GCT}(\kappa_i, \gamma)$	$\text{GPo}(\mu_i, \sigma)$	$\text{PTw}_p(\mu_i, \omega)$
Parâmetro de dispersão	$\phi = \log(\nu)$ $\nu > 0$	$\gamma = \log(\alpha)$ $\alpha > 0$	σ $\sigma > c^*$	ω $\omega > 0$
Esperança	$\approx \mu_i$	$\overset{a}{\approx} \kappa_i/\alpha$	μ_i	μ_i
Variância	$\approx \mu_i/\nu$	$\overset{a}{\approx} \kappa_i/\alpha^2$	$\mu_i(1+\sigma\mu_i)^2$	$\mu_i(1+\omega\mu_i^{p-1})$
Índice de dispersão (DI)	$\approx 1/\nu$	$\overset{a}{\approx} 1/\alpha$	$(1+\sigma\mu_i)^2$	$1+\omega\mu_i^{p-1}$
Regressão	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\kappa_i = \alpha g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$	$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$

$c^* = \min[-\max(y_i^{-1}), -\max(\mu_i^{-1})]$; $\overset{a}{\approx}$ comportamento assintótico quanto $T \to \infty$.

Na Figura 1, são apresentados os índices de dispersão para os modelos COM-Poisson, *Gamma-Count*, Poisson generalizado e Poisson-Tweedie. Os intervalos dos parâmetros de dispersão foram considerados de tal forma que se tenha DI= 4 e $DI = 0,25$ quando E(Y) $= 50$ (para a distribuição Poisson-Tweedie o intervalo começa em 0, pois $\nexists\ \omega \mid$ DI < 1). Para a distribuição Poisson-Tweedie consideram-se 3 diferentes valores para o parâmetro p, $p = 1,1$; $p = 2$ (binomial negativa) e $p = 3$ (Poisson-inversa gaussiana). Diferentes comportamentos do índice de dispersão são observados nos modelos. Destaca-se a similaridade entre a COM-Poisson e *Gamma-Count* e a flexibilidade da Poisson-Tweedie para modelar superdispersão.

3 Análise dos dados

Para análise do número de insetos emergentes, consideram-se preditores com efeitos de tratamento para os quatro modelos apresentados. Os resultados mostram ajustes similares em termos do máximo da função de verossimilhança. Na Figura 2(a), são apresentados o gráfico de dispersão das contagens (em cinza) e os valores ajustados para a média das contagens com intervalos de confiança de 95%. Para o modelo COM-Poisson, o mal condicionamento da função de verossimilhança impossibilitou a aproximação da matriz Hessiana no ponto de máximo e, consequentemente, o cálculo de intervalos de confiança. Na Figura 2(b), são apresentadas as relações média–variância. As variâncias calculadas sob o modelo COM-Poisson foram maiores que as demais em quase todo o intervalo. Para

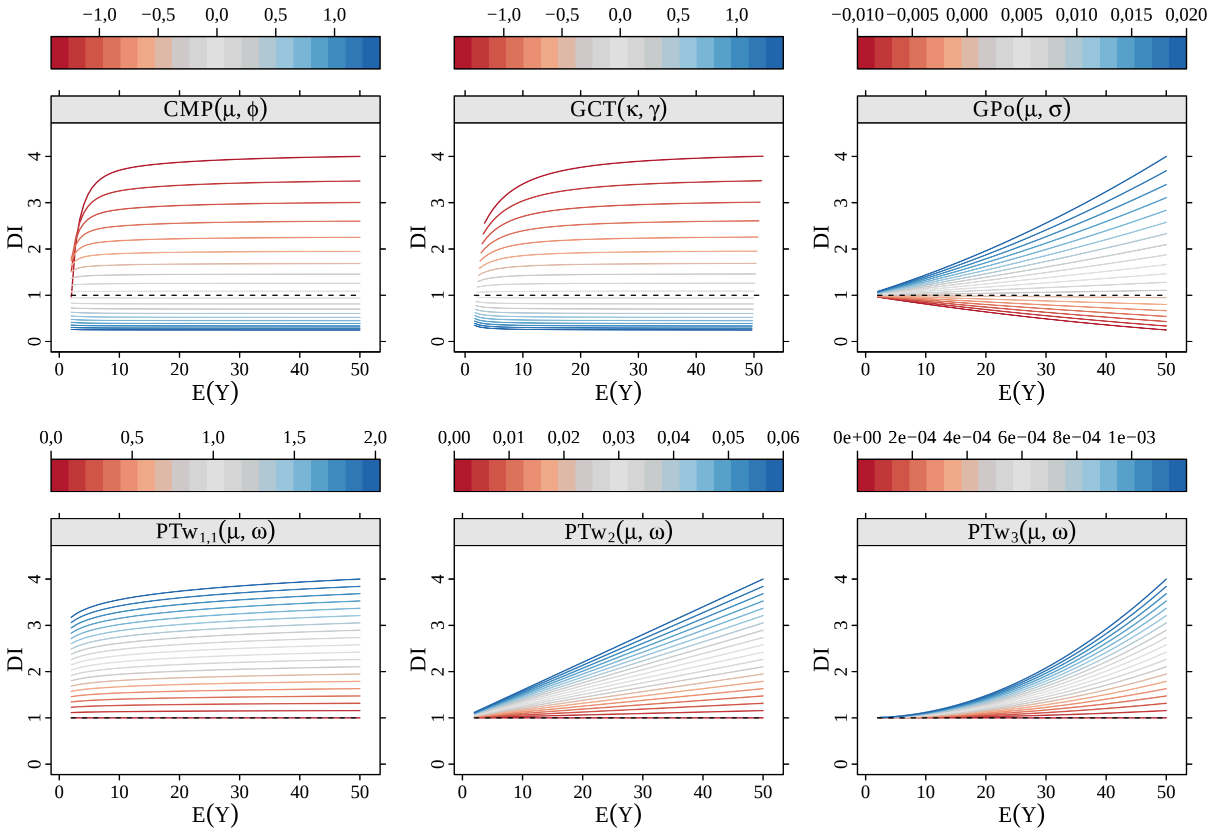

Figura 1: Índices de dispersão para diferentes combinações de parâmetros das distribuições COM-Poisson, *Gamma-Count*, Poisson generalizada e Poisson-Tweedie. As linhas tracejadas representam a equidispersão, DI= 1.

o modelo Poisson-Tweedie, cuja variância é uma função polinomial de ordem p, a relação média–variância é praticamente linear. A estimativa do parâmetro p foi de . No modelo Poisson generalizado, a relação parece ser aproximadamente quadrática ($\hat{\sigma} = 0{,}019$), com menores variâncias para $\mathrm{E}(Y)$ entre 10 e 20 e maiores para $\mathrm{E}(Y)$ próximos a 30, quando comparado aos outros modelos.

4 Considerações finais

Nesse artigo, foram revisadas algumas recentes abordagens para análise de dados na forma de contagens. A gênesis de cada modelo foi apresentada, assim como um resumo comparativo com base no índice de dispersão e no comportamente da relação média–variância. As abordagens apresentadas foram aplicadas para análise do número de insetos emergentes sob diferentes substratos do milho. Os resultados da análise subsidiaram a discussão sobre a comparação dos modelos. Ambos os modelos apresentaram resultados bastante similares, porém, diferenças foram destacadas sobre as esperanças e variâncias obtidas para cada modelo ajustado.

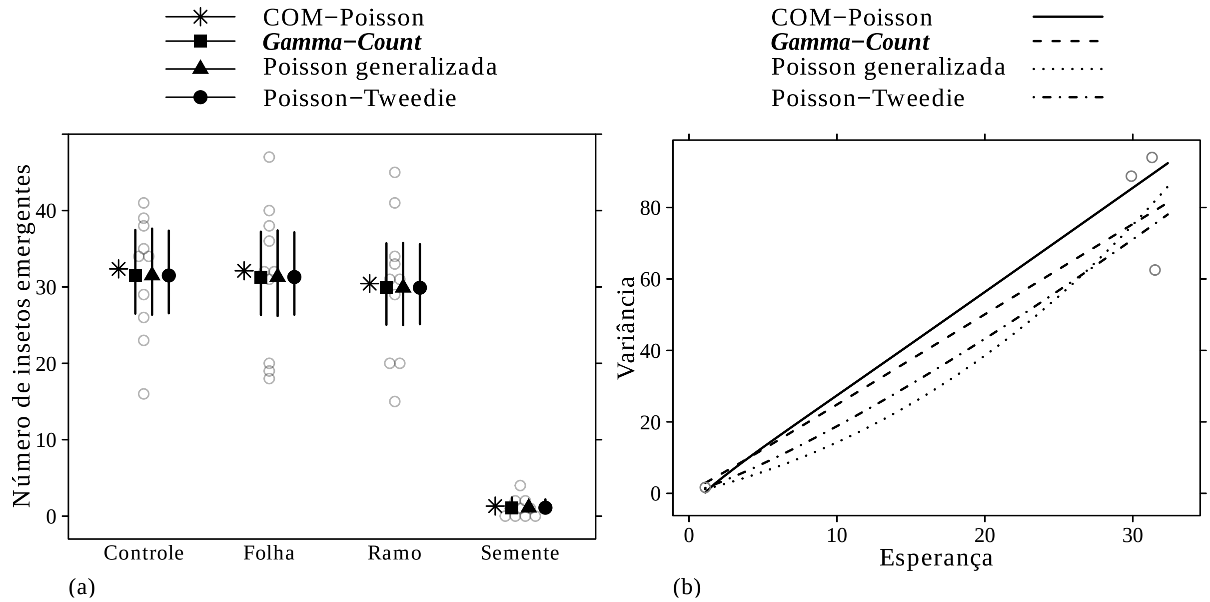

Figura 2: (a) Valores ajustados para a média do número de insetos emergentes com intervalos de confiança de 95% e (b) Relação média–variância para os modelos ajustados. Os pontos em cinza representam os valores observados.

De forma geral, tem-se no artigo uma abrangente revisão de estratégias para análise de dados na forma de contagens, destacando as características dos modelos bem como suas respectivas formulações de forma que o leitor possa se familiarizar com os modelos e identificar possíveis extensões.

Referências

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J. & Démetrio, C. G. B. (2018), 'Extended Poisson-Tweedie: properties and regression model for count data', *Statistical Modelling* **18**(1), 24–49.

Del Castillo, J. & Pérez-Casany, M. (1998), 'Weighted Poisson distributions for overdispersion and underdispersion situations', *Annals of the Institute of Statistical Mathematics* **50**(3), 567–585.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Chapman & Hall, London.

Ribeiro Jr, E. E., Zeviani, W. M., Bonat, W. H., Demétrio, C. G. B. & Hinde, J. (2018), 'Reparametrization of COM-Poisson regression models with applications in the analysis of experimental data', *arXiv (Statistics Applications and Statistics Methodology)* .

Winkelmann, R. (1995), 'Duration Dependence and Dispersion in Count-Data Models', *Journal of Business & Economic Statistics* **13**(4), 467–474.

Zamani, H. & Ismail, N. (2012), 'Functional form for the generalized Poisson regression model', *Communication in Statistics – Theory and Methods* **41**, 3666–3675.