Modelos duplos COM-Poisson: modelando média e dispersão na análise de contagens

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

1 Introdução

A análise de contagens é comumente realizada considerando a modelagem da média da variável resposta em termos de covariáveis, ou seja, $g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, em que $g(\cdot)$ é uma função monótona e diferenciável, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots x_{ip})^\top$ é o vetor de de covariáveis da i-ésima observação e $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados (Pregibon 1984).

Sob a abordagem paramétrica, obtém-se a relação entre média e variância a partir da especificação da distribuição da variável reposta. Por exemplo, para as distribuições Poisson, Poisson generalizada e Poisson-Tweedie as variâncias são μ_i, $\mu_i(1+\phi\mu_i)^2$ e $\mu_i(1+\phi\mu_i^p)$, respectivamente. Outras distribuições como COM-Poisson e *Gamma-Count*, também são flexíveis para modelar sub e superdispersão, porém, não têm formas fechadas para a média ou para a variância.

Apesar da flexilidade induzida pela especificação da distribuição, a variância da variável resposta depende das covariáveis apenas por meio de μ_i. Nesse artigo, propõem-se os modelos duplos COM-Poisson, em que se modela a média e a dispersão em termos de covariáveis. Dessa forma, garante-se flexibilidade para modelar, por exemplo, casos em que covariáveis levam a um acréscimo da média e decréscimo da dispersão. Esse abordagem é similar à proposta de Smyth (1988), no entanto, para a distribuição COM-Poisson as expressões dos modelos de média e de dispersão não podem ser obtidos em forma fechada.

2 Estudo de caso

Para preservar os recursos de ecossistemas de água sobre os efluentes industriais, estudos em biometria avaliam os efeitos de poluentes de diversas fontes, tais como fertilizantes e pesticidas, sobre o crescimento e a reprodução de determinada espécie. Bailer & Oris (1994) avaliaram o impacto de diferentes doses de nitrofeno (herbicida usado para controle de ervas daninhas) na reprodução de uma espécie de zooplâncton de água doce. Na Figura 1(a), são apresentados os dados observados, notando-se o decréscimo na reprodução do zooplâncton para concentrações maiores de nitrofeno. Na Figura 1(b), tem-se o gráfico

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

de dispersão das médias e variâncias, além da reta identidade, representando a equidispersão. Para as três primeiras concentrações experimentadas, há a indicação de subdispersão, no entanto, para as duas maiores concentrações, a variância amostral é maior que a respectiva média, sugerindo uma variabilidade extra-Poisson. Esse é um exemplo em que a modelagem da variabilidade considerando apenas a relação média–variância da distribuição adotada, pode não ser adequada.

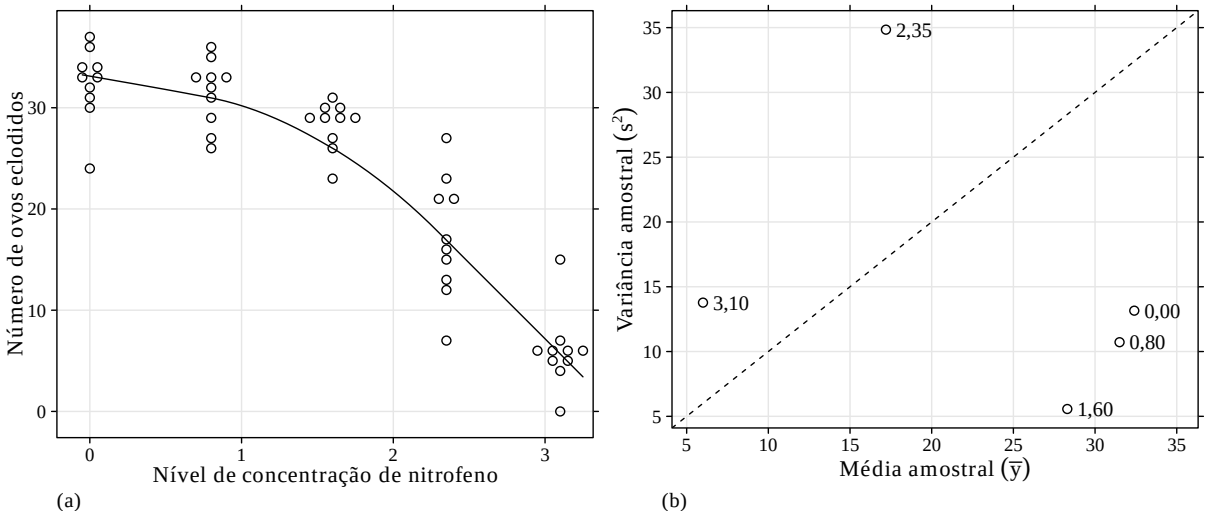

Figura 1: (a) Número de ovos eclodidos para cada dosagen de nitrofen e (b) dispersão das médias e variâncias amostrais para cada dose. As linhas representam (a) a curva de suavização estimada pelo algoritmo *lowess* e (b) a reta identidade, representando a equidispersão.

3 Modelos duplos COM-Poisson

A distribuição COM-Poisson é uma generalização biparamétrica da Poisson que contempla sub e superdispersão (Shmueli et al. 2005). Sua função massa de probabilidade é dada por

$$\Pr(Y = y \mid \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \qquad y = 0, 1, 2, \ldots \tag{1}$$

em que $\lambda > 0$, $\nu \geq 0$ e $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j/(j!)^\nu$ é uma constante de normalização. Esse modelo possui as distribuições Poisson ($\nu = 1$) e geométrica ($\nu = 0$ e $\lambda < 1$) como casos particulares e permite super e subdispersão quando $0 < \nu < 1$ e $\nu > 1$, respectivamente.

A média e a variância para a distribuição COM-Poisson não são obtidos em formas fechadas. Shmueli et al. (2005) apresentam aproximações para a média e para a variância da forma

$$\mathrm{E}(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \qquad \text{e} \qquad \mathrm{Var}(Y) \approx \frac{\lambda^{1/\nu}}{\nu}, \tag{2}$$

que são particularmente acuradas para $\nu \leq 1$ ou $\lambda > 10^\nu$ (Shmueli et al. 2005). Ribeiro

Jr et al. (2018) mostram que a aproximação para a média é satisfatória, justificando sua proposta de reparametrização para média do modelo COM-Poisson, $\text{CMP}_{\mu}(\mu_i, \nu)$, em que $\mu = \lambda^{1/\nu} - (\nu - 1)/2\nu$.

Para definição dos modelos duplos COM-Poisson, considere um conjunto de dados $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$, $i = 1, 2, \ldots, n$, em que y_i são realizações independentes do modelo COM-Poisson e $\boldsymbol{x}_i$ e $\boldsymbol{z}_i$ são subvetores do vetor de covariáveis. Propõe-se a modelagem conjunta da média e dispersão pela especificação

$$Y_i \sim \text{CMP}_{\mu}(\mu_i, \nu_i), \quad \text{em que} \quad g(\mu_i) = \boldsymbol{x}_i^{\top}\boldsymbol{\beta} \quad \text{e} \quad g(\nu_i) = \boldsymbol{z}_i^{\top}\boldsymbol{\gamma},$$

ou seja, ambos os parâmetros de média e dispersão são modelados com covariáveis.

A estimação do vetor de parâmetros $(\boldsymbol{\beta}^{\top}, \boldsymbol{\gamma}^{\top})^{\top}$ é realizada pelo método de máxima verossimilhança, utilizando o algoritmo `BFGS` implementado pelo função `optim` do software `R`. Os códigos para ajuste e análise dos dados nesse artigo são disponibilizados em material suplementar online[1].

4 Resultados e discussões

Para análise da reprodução do zooplâncton sob diferentes concentrações de nitrofeno, consideram-se os preditores

$$\begin{aligned}
&\text{Escalar:} && \log(\nu_i) = \gamma_0, \\
&\text{Linear:} && \log(\nu_i) = \gamma_0 + \gamma_1 \texttt{dose}_i, \\
&\text{Quadrático:} && \log(\nu_i) = \gamma_0 + \gamma_1 \texttt{dose}_i + \gamma_2 \texttt{dose}_i^2 \text{ e} \\
&\text{Cúbico:} && \log(\nu_i) = \gamma_0 + \gamma_1 \texttt{dose}_i + \gamma_2 \texttt{dose}_i^2 + \gamma_3 \texttt{dose}_i^3,
\end{aligned}$$

para dispersão e cúbico para a média.

Na Tabela 1, são apresentadas estimativas e erros padrões para os parâmetros de cada modelo considerado. Note que as estimativas para a estrutura da média são bastante similares para as diferentes estruturas de dispersão. Esse resultado é esperado, uma vez que os parâmetros μ e ν se mostram ortogonais no modelo COM-Poisson Ribeiro Jr et al. (2018).

Na Tabela 2, são apresentadas algumas algumas medidas de qualidade de ajuste seguidas de testes de razão de verossimilhanças entre os modelos considerados. Os resultados indicam que há uma importante melhoria no ajuste ao considerar as estruturas linear e quadrática para modelagem da dispersão. Para a hipótese $\beta_2 = 0$, os testes assintóticos de Wald e de razão de verossimilhanças apresentaram valores de p iguais a 0,103 e 0,057, respectivamente.

[1]Disponível em `http://www.leg.ufpr.br/~eduardojr/papercompanions/rbras2018`

Tabela 1: Estimativas e erros padrões dos parâmetros do modelo duplo COM-Poisson ajustados aos dados do estudo sobre o nitrofeno.

	Estimativa (Erro Padrão)			
Parâmetro	Escalar	Linear	Quadrático	Cúbico
β_0	2,981 (0,035)[a]	2,978 (0,042)[a]	2,972 (0,049)[a]	2,975 (0,047)[a]
β_1	−3,952 (0,287)[a]	−3,980 (0,365)[a]	−4,041 (0,447)[a]	−4,013 (0,418)[a]
β_2	−2,131 (0,260)[a]	−2,161 (0,311)[a]	−2,218 (0,351)[a]	−2,197 (0,330)[a]
β_3	−0,543 (0,221)[a]	−0,573 (0,212)[a]	−0,604 (0,206)[a]	−0,597 (0,195)[a]
γ_0	0,048 (0,205)	0,295 (0,211)	0,243 (0,259)	0,353 (0,227)
γ_1	–	−5,244 (1,363)[a]	−7,013 (2,307)[a]	−5,729 (1,844)[a]
γ_2	–	–	−3,984 (2,444)	−2,918 (1,904)
γ_3	–	–	–	1,522 (1,412)

Est (EP)[a] indica que $|\text{Est/EP}| > 1,96$.

Tabela 2: Medidas de qualidade de ajuste e testes de razão de verossimilhança para os modelos ajustados.

	G.l	Deviance	AIC	χ^2	$\Pr(>\chi^2)$
Escalar	45	288,127	298,127	–	–
Linear	44	274,111	286,111	14,0163	0,0002
Quadrático	43	270,493	284,493	3,6179	0,0572
Cúbico	42	269,503	285,503	0,9898	0,3198

Na Figura 2(a), são apresentados os valores ajustados $\hat{\nu}_i$ com bandas de confiança de 95% para os quatro preditores. Com exceção do preditor escalar, todos os modelos apresentam uma variação da dispersão extra relação média–variância do modelo. Uma constatação desse fato se dá, aproximadamente para concentração de nitrofeno de $2\mu\text{g}/10^2$litros, em que o o número de ovos eclodidos passam de superdispersos para equidispersos.

As correspondentes médias e variâncias obtidas dos modelos ajustados são apresentadas na Figura 2(b). Note que a ordem do polinômio adotado para $\log(\nu_i)$ não representa a ordem da curva para a variância da variável resposta. Isso se dá tanto devido à função de ligação, quanto à relação média–variância da própria distribuição.

5 Considerações finais

Nesse artigo, foram propostos os modelos duplos COM-Poisson para modelagem conjunta da média e dispersão na análise de contagens com diferentes níveis de dispersão. Essa classe de modelos permite modelar a dispersão por covariáveis. Os parâmetros são estimados pelo método da máxima verossimilhança e inferências são realizadas com base nas distribuições assintóticas dos estimadores. A metodologia é aplicada para análise do número de ovos eclodidos sob dosagens de herbicida em um estudo de biometria. A

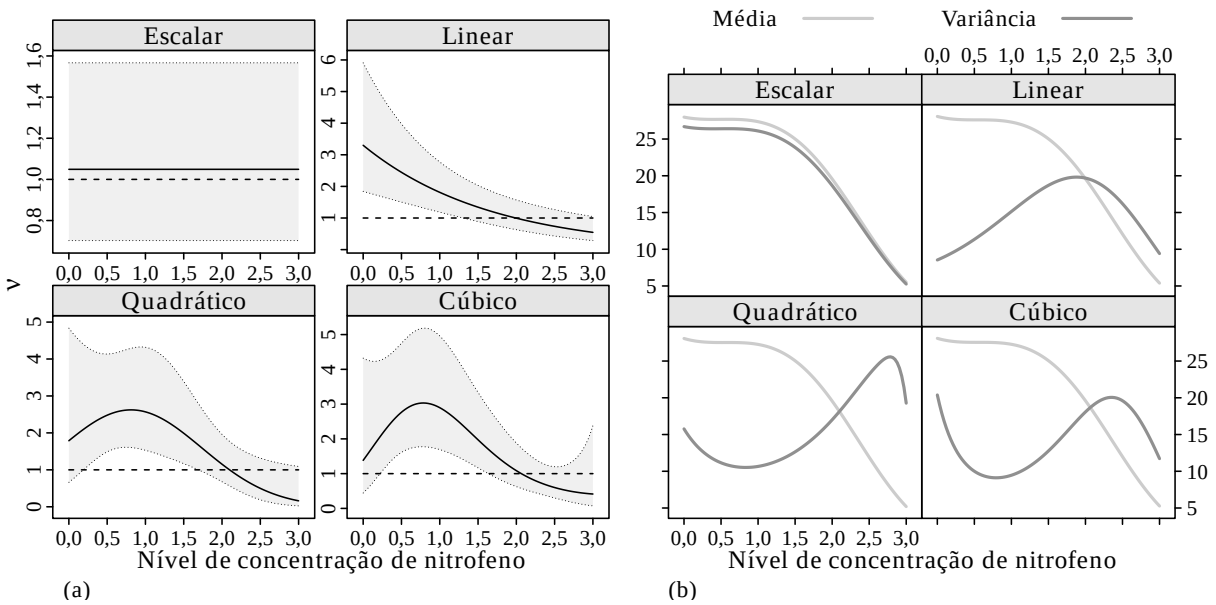

Figura 2: (a) Valores ajustados para o parâmetro de dispersão ν com bandas de confiança de 95% e (b) médias e variâncias obtidas a partir do modelo ajustado. A linha horizontal pontilhada em (a) representa a modelo de poisson ($\nu = 1$).

proposta apresentou melhorias em termos do ajuste quando comparada ao modelo de regressão COM-Poisson convencional. As análises sugerem que melhorias no processo de estimação podem ser feitas, aproveitando a propriedade de ortogonalidade da parametrização utilizada. Para trabalhos futuros, estudos de simulação se fazem necessários a fim de avaliar as propriedades dos estimadores e a robustez do modelo. Além disso, como melhoria na análise do estudo de caso, considerar preditores não lineares para a média pode ser útil a fim de evitar o uso de polinômios cúbicos.

Referências

Bailer, A. & Oris, J. (1994), 'Assessing toxicity of pollutants in aquatic systems', *In Case Studies in Biometry* pp. 25–40.

Pregibon, D. (1984), 'Review: P. McCullagh, J. A. Nelder, generalized linear models"', *The Annals of Statistics* **12**(4), 1589–1596.

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)* .

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. (2005), 'A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **54**(1), 127–142.

Smyth, G. K. (1988), 'Generalized linear models with varying dispersion', *Journal of the Royal Statistical Society. Series B (Methodology)* **51**(1), 47–60.