

Reparametrização dos Modelos de Regressão COM-Poisson com Aplicações na Análise de Dados de Contagem Experimentais

Eduardo Elias Ribeiro Junior ^{† 1 3}

Walmes Marques Zeviani ^{2 3}

Wagner Hugo Bonat ^{2 3}

Clarice Garcia Borges Demétrio ¹

1 Introdução

Variáveis aleatórias de contagem são de natureza discreta e representam o número de ocorrências de um evento em um domínio discreto ou contínuo. Dados dessa natureza são frequentes em estudos experimentais, por exemplo número de grãos por planta, número de frutos por árvore, número de insetos na parcela experimental entre outros. Com a introdução dos modelos lineares generalizados por Nelder e Wedderburn (1972), dados de contagem passaram a ser comumente analisados considerando a distribuição Poisson, uma vez que das distribuições mais conhecidas da família exponencial é a única para tratar dados dessa natureza.

O modelo Poisson possui apenas um parâmetro, denotado por λ , que representa a média, e também, a variância, o que implica em uma relação identidade ($\lambda = E(Y) = V(Y)$). Essa propriedade, chamada de equidispersão, é uma particularidade do modelo Poisson que é inadequada para diversas situações. Quando aplicado sob negligência dessa suposição, o modelo Poisson apresenta erros padrões inconsistentes para as estimativas dos parâmetros e, por consequência, para toda função desses parâmetros (WINKELMANN, 1995).

Na prática, o caso mais comum de fuga da suposição de equidispersão e, consequentemente com um maior número de abordagens possíveis, é a *superdispersão* ($E(Y) < V(Y)$) que podem ocorrer por diversas razões como heterogeneidade das unidades experimentais, ausência de covariáveis experimentais, diferentes amplitudes de domínio (*offset*) não considerados, correlação entre as observações, excesso de zero entre outras (HINDE; DEMÉTRIO, 1998). O caso menos comum, mas que tem ganhado a atenção da comunidade estatística é a *subdispersão* ($E(Y) > V(Y)$). Os processos que reduzem a variabilidade das contagens, abaixo do estabelecido pela Poisson, não são tão conhecidos

[†]Contato jreduardo@usp.br

¹Departamento de Ciências Exatas (LCE) - ESALQ-USP

³Laboratório de Estatística e Geoinformação (LEG) - UFPR

²Departamento de Estatística (DEST) - UFPR

quanto os que produzem variabilidade extra. Pela mesma razão, são poucas as abordagens descritas na literatura capazes de tratar subdispersão. Podemos justificar o caso de subdispersão quando o processo de Poisson é violado, ou seja, quando os tempos entre eventos não são exponencialmente distribuídos (ZEVIANI et al., 2014). Nesses casos, as contagens resultantes podem ser sub ou superdispersas.

Para análise de dados de contagens não equidispersos existem diversas alternativas. No caso de superdispersão destacam-se os modelos que incluem efeitos aleatórios em nível de observação, considerando assim a heterogeneidade não observada. Um exemplo bem conhecido dessa prática é o modelo Poisson com efeitos aleatórios gama, que resulta no modelo binomial negativo. Porém, outras escolhas para a distribuição dos efeitos aleatórios podem ser tomadas como, por exemplo, o modelo Poisson-Tweedie (BONAT et al., 2016). Para o caso de subdispersão, modelos da classe *Duration Dependence* flexibilizam a distribuição do tempo entre eventos, assumida exponencial no caso Poisson (WINKELMANN, 2008). Um representante dessa classe é o modelo *gamma-count*, cujo a distribuição assumida é gama (ZEVIANI et al., 2014). Uma outra abordagem para dados sub e superdispersos, que têm sido de interesse em pesquisas da comunidade estatística, é o modelo COM-Poisson (SHMUELI et al., 2005). Esse modelo pertence à classe de modelos de distribuição ponderada de Poisson (SELLERS; BORLE; SHMUELI, 2012) que flexibilizam a suposição de linearidade da razão de probabilidades consecutivas, permitindo caudas mais pesadas ou mais leves à distribuição (RIBEIRO JR, 2016).

O modelo COM-Poisson pertence à família exponencial e tem como casos particulares os modelos Poisson e geométrico e como caso limite o modelo binomial. Algumas aplicações recentes do modelo COM-Poisson são apresentadas por Lord, Geedipally e Guikema (2010), para análise do número de acidentes de trânsito; por Sellers e Shmueli (2010), na modelagem do número de ampolas quebradas em fretes aéreos; e por Ribeiro Jr (2016) que apresenta aplicações para dados dados experimentais como o número de capulhos de algodão sob efeito de desfolha, número de capulhos de algodão sob efeito da infestação induzida de mosca branca, ocorrência de ninfas de mosca branca em lavoura de soja e número de grãos e vagens para difentes níveis de adubação potássica e umidade do solo. A principal desvantagem desse modelo é que não há um parâmetro que represente a média da distribuição o que dificulta a interpretação dos coeficientes em um estrutura de modelo de regressão.

Nesse artigo, propõe-se uma reparametrização do modelo COM-Poisson, reescrevendo a média da distribuição como função dos parâmetros originais, a partir de uma aproximação dada por Shmueli et al. (2005). Essa reparametrização apresenta vantagens para interpretação e para o ajuste do modelo, uma vez que encontrou-se uma relação, aproximadamente, ortogonal entre parâmetros de média e precisão. Com aplicações para análise de dados experimentais, contemplando as situação de sub, super e equidispersão discutem-se os aspectos do modelo COM-Poisson reparametrizado comparando-o com os modelos

Poisson, COM-Poisson na parametrização original e quasi-Poisson. Toda a análise é realizada no *software* R (R CORE TEAM, 2016) e os códigos para ajuste e inferência dos modelos, bem como os conjuntos de dados, são disponibilizados em material suplementar *online*¹.

2 Metodologia

2.1 Distribuição COM-Poisson

A distribuição de probabilidades COM-Poisson Sellers e Shmueli (2010), generaliza a distribuição Poisson em termos da razão de probabilidades sucessivas, ao custo da adição de um parâmetro. Sendo Y uma variável aleatória que segue o modelo COM-Poisson, então

$$\frac{\Pr(Y = y - 1)}{\Pr(Y = y)} = \frac{y^\nu}{\lambda}$$

ao passo que sob a distribuição Poisson essa razão resulta em $\frac{y}{\lambda}$, uma função linear em y . Essa característica permite caudas mais pesadas ou mais leves à distribuição acomodando os casos de sub, super e equidispersão. A função massa de probabilidade do modelo COM-Poisson é dada por

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

em que $\lambda > 0$, $\nu \geq 0$ e $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ é uma constante de normalização.

Observe que a série $Z(\lambda, \nu)$ é teoricamente divergente somente quando $\nu = 0$ e $\lambda \geq 1$, mas numericamente para valores pequenos de ν combinados com grandes valores de λ , a soma assume valores tão elevados que excedem a capacidade de representação dos computadores usuais (*overflow*). Na Tabela 1, são apresentadas as somas calculadas com incrementos de 1000, ou seja, $\sum_{j=0}^{1000} \lambda^j / (j!)^\nu$ para diferentes valores de λ e ν ².

Um inconveniente desse modelo é que os momentos média e variância não têm formas fechadas. Shmueli et al. (2005), a partir de uma aproximação para $Z(\lambda, \nu)$ apresentam uma forma aproximada para a média e variância de uma variável aleatória que segue o modelo COM-Poisson

$$E(Y) \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \quad \text{e} \quad V(Y) \approx \frac{\lambda^{1/\nu}}{\nu} \quad (2)$$

os autores ressaltam que essa aproximação é satisfatória para $\nu \leq 1$ ou $\lambda > 10^\nu$. Sellers e Shmueli (2010) descrevem ainda que a relação média-variância pode ser aproximada por

¹Disponível em <<http://www.leg.ufpr.br/~eduardojr/papercompanions/rbras2017>>

²Para $\lambda = 1$ e $\nu = 0$ exibimos Inf, embora o valor numericamente calculado seja 1000, pois a série é claramente divergente nessa situação. Inf representa ∞ .

Tabela 1: Valores calculados numericamente de $Z(\lambda, \nu)$ para diferentes valores de λ (0,5 a 50) e ϕ (0 a 1).

ν	λ					
	0,5	1	5	10	30	50
0	2,00E+00	Inf	Inf	Inf	Inf	Inf
0,1	1,92E+00	7,64E+00	Inf	Inf	Inf	Inf
0,2	1,86E+00	5,25E+00	3,17E+273	Inf	Inf	Inf
0,3	1,81E+00	4,32E+00	1,60E+29	2,54E+282	Inf	Inf
0,4	1,77E+00	3,80E+00	4,71E+10	1,33E+56	Inf	Inf
0,5	1,74E+00	3,47E+00	1,34E+06	3,67E+22	3,32E+196	Inf
0,6	1,72E+00	3,23E+00	2,05E+04	4,99E+12	1,73E+76	4,63E+177
0,7	1,70E+00	3,06E+00	2,37E+03	3,69E+08	4,93E+39	6,93E+81
0,8	1,68E+00	2,92E+00	6,49E+02	2,70E+06	5,09E+24	3,43E+46
0,9	1,66E+00	2,81E+00	2,74E+02	1,47E+05	1,80E+17	2,19E+30
1	1,65E+00	2,72E+00	1,48E+02	2,20E+04	1,07E+13	5,18E+21

$\frac{1}{\nu}E(Y)$. Na Figura 1(a), apresenta-se um gráfico para estudo da relação média-variância, com médias e variâncias calculadas numericamente para diferentes níveis de dispersão. É interessante notar o aspecto linear dessa relação.

Nesse artigo, propõe-se uma reparametrização do modelo COM-Poisson baseada na expressão aproximada para média (2). A reparametrização consiste na introdução do parâmetro μ definido como

$$\mu = h(\lambda, \nu) = \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \Rightarrow \lambda = h^{-1}(\mu, \nu) = \left(\mu + \frac{(\nu + 1)}{2\nu} \right)^\nu. \quad (3)$$

O parâmetro de precisão da distribuição, denotado por ν , é tomado na escala do logaritmo³ para evitar a restrição no espaço paramétrico, denota-se por ϕ esse novo parâmetro, ou seja, $\phi = \log(\nu)$, $\phi \in \mathbb{R}$. Esse parâmetro possui a mesma interpretação de ν , porém em uma escala diferente. Para $\phi < 0$ temos o caso de superdispersão, para $\phi > 0$ subdispersão e para $\phi = 0$ equidispersão, caso particular em que a COM-Poisson se resume à Poisson.

Substituindo os novos parâmetros, definidos em (3), na função massa de probabilidade (1) tem-se

$$\Pr(Y = y \mid \mu, \phi) = \left(\mu + \frac{e^\phi - 1}{2e^\phi} \right)^{ye^\phi} \frac{(y!)^{-e^\phi}}{Z(\mu, \phi)}, \quad y = 0, 1, 2, \dots \quad (4)$$

Denota-se a distribuição COM-Poisson reparametrizada por COM-Poisson $_\mu$. Na Figura 2 são apresentadas distribuições de probabilidade para diferentes parâmetros ilustrando a flexibilidade da distribuição. As distribuições para $\phi = 0$ representam o caso particular Poisson.

³Nesse artigo, por simplicidade, denota-se logaritmo (log) como o logaritmo neperiano (\log_e).

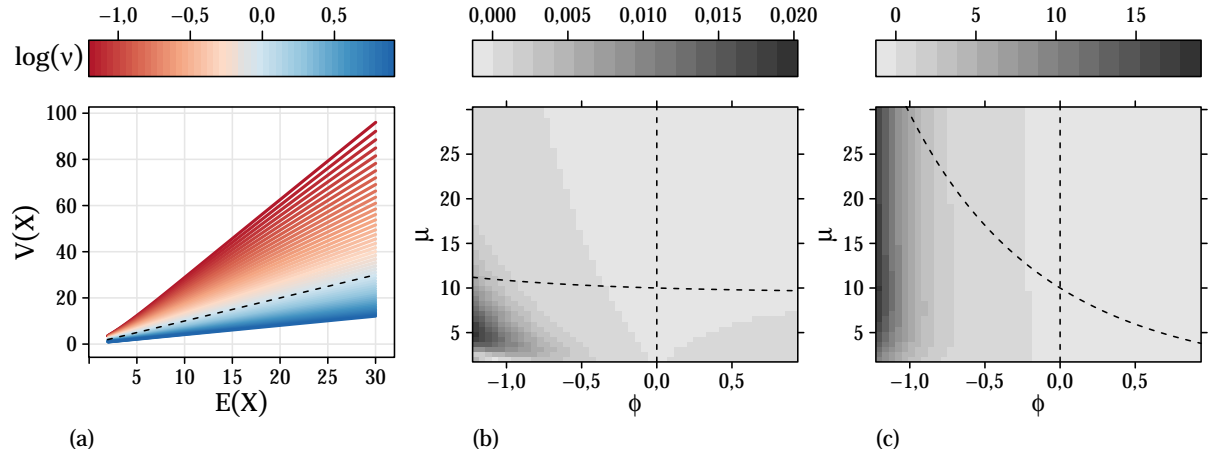


Figura 1: Representação dos momentos da distribuição COM-Poisson para diferentes médias e níveis de dispersão: (a) relação média-variância, (b) e (c) erros quadráticos da aproximações para média e variância, respectivamente.

Uma avaliação da acurácia das aproximações em (2) e, conseqüentemente, da reparametrização é apresentada na Figura 1(b) e 1(b)(c) para valores de μ variando de 0 a 30 e diferentes níveis de dispersão ($-1,2 < \phi < 1$). Em 1(b) têm-se um gráfico dos erros quadráticos para a média, $(\mu - E[X])^2$ e em 1(c) para a variância $(\frac{\mu}{\nu} - V[X])^2$, em que $E[X]$ e $V[X]$ são calculados numericamente usando a definição de momentos. As linhas tracejadas representam as restrições $\nu \leq 1$ ou $\lambda > 10^\nu$. Observa-se que a aproximação para a média é acurada (valores menores que 0,03), porém tem perda de acurácia para combinações de ν pequenos ($< 0,35$) com médias também baixas ($\mu < 10$). Para a variância, gráfico 1(c), têm-se que a aproximação $V[X] = \frac{\mu}{\nu}$, proposta por Sellers e Shmuli (2010), não tem bom desempenho para valores pequenos de ν . A restrição que envolve a média da distribuição parece não ter relevância. De forma geral, para média e variância, os valores de ν parecem ser mais influentes na acurácia do que os valores de μ .

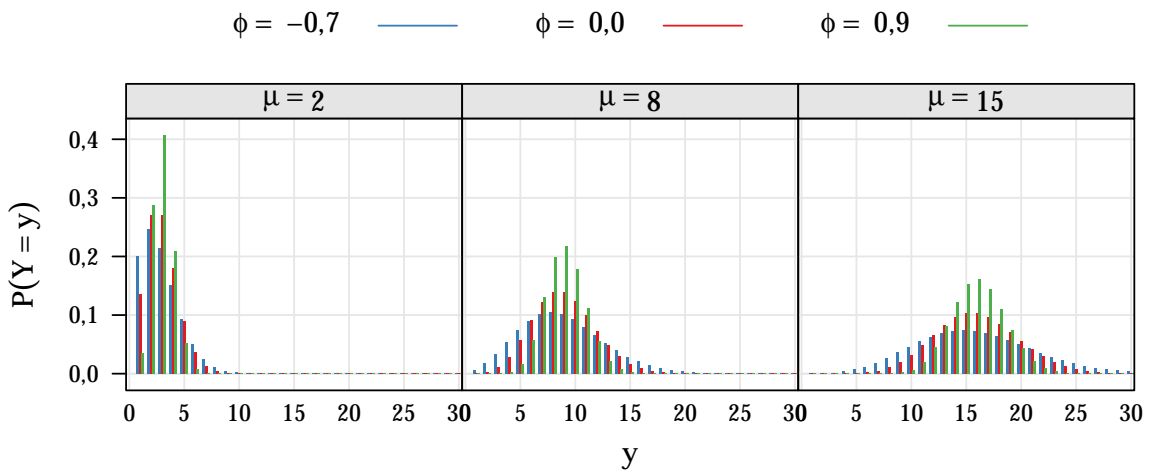


Figura 2: Probabilidades pela distribuição COM-Poisson_μ para diferentes parâmetros.

Os resultados apresentados na Figura 1 mostram que a aproximação para o primeiro momento central da distribuição é acurada, e sendo assim a reparametrização proposta adequada.

2.2 Estimação e Inferência

Os modelos COM-Poisson padrão e reparametrizado são ajustados via maximização da verossimilhança. Para uma amostra independente de contagens y_i , $i = 1, 2, \dots, n$, as estimativas para $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ são obtidas pelos argumentos que maximizam o logaritmo da função de verossimilhança

$$\ell(\boldsymbol{\beta}, \phi \mid \mathbf{y}) = e^\phi \left[\sum_{i=1}^n y_i \log \left(\mu_i + \frac{e^\phi - 1}{2e^\phi} \right) - \sum_{i=1}^n \log(y_i!) \right] - \sum_{i=1}^n \log(Z(\mu_i, \phi)), \quad (5)$$

onde $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$, sendo \mathbf{x}_i o vetor $(x_{i1}, x_{i2}, \dots, x_{ip})$ de covariáveis da i -ésima observação, e $(\boldsymbol{\beta}, \phi) \in \mathbb{R}^{p+1}$. A constante $Z(\mu_i, \phi)$ é calculada como

$$Z(\mu_i, \phi) = \sum_{j=0}^{\infty} \left[\left(\mu_i + \frac{e^\phi - 1}{2e^\phi} \right)^{je^\phi} \frac{1}{(j!)^{e^\phi}} \right]$$

A avaliação do logaritmo da função de verossimilhança requer o cálculo de uma série infinita para cada observação, o que torna sua computação cara para regiões do espaço paramétrico cuja soma demora a convergir.

A estimação dos parâmetros requer a maximização numérica de (5). Como $\ell(\mu_i, \phi \mid \mathbf{y})$ não possui derivada analítica, a maximização é realizada pelo algoritmo BFGS (NOCEDAL; WRIGHT, 1995) que utiliza estimativas numéricas para o vetor gradiente $\mathcal{U}(\boldsymbol{\theta})$ e matriz hessiana $\mathcal{H}(\boldsymbol{\theta})$. Os erros padrão das estimativas são obtidos pela aproximação normal do logaritmo da função de verossimilhança (método de Wald), fazendo $\sqrt{\mathbf{v}}$, em que \mathbf{v} são os elementos da diagonal da matriz $-\hat{\mathcal{H}}^{-1}(\boldsymbol{\theta})$, estimada pelo algoritmo BFGS. Intervalos de confiança para $\hat{\mu}_i$ são obtidos pelo método delta (RIBEIRO JR et al., 2012).

Para o modelo COM-Poisson padrão, o procedimento de ajuste é análogo ao apresentado, porém, considerando a Equação (5) em termos de λ . Mesmo para a distribuição COM-Poisson padrão, o parâmetro de dispersão é mantido na escala do logaritmo, para evitar a restrição do espaço paramétrico.

A comparação dos modelos é realizada pelo valor maximizado do logaritmo da função de verossimilhança e pelos critérios de Akaike (AIC) e Bayesiano (BIC) que penalizam a log-verossimilhança pelo número de parâmetros, ou seja,

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}, \mathbf{y}) + 2p \quad \text{e} \quad \text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}, \mathbf{y}) + \log(n)p.$$

Nas aplicações, também, utiliza-se a abordagem quasi-Poisson (WEDDERBURN,

1974) como modelo de referência. Essa abordagem é uma especificação de momentos e acomoda sub e superdispersão corrigindo a variância de y_i pelo parâmetro adicional σ , $V(Y_i) = \sigma\mu_i$. Esses modelos são ajustados no *software* R (R CORE TEAM, 2016), com a função `glm(..., family = quasipoisson)`.

3 Estudos de caso

3.1 Desfolha artificial em capulhos de algodão

Um experimento com plantas de algodão *Gossypium hirsutum* submetidas a diferentes níveis de desfolha artificial de remoção foliar, (0, 25, 50, 75 e 100%) (**des**), em combinação com o estágio fenológico no qual a desfolha foi aplicada, (vegetativo, botão floral, florescimento, maçã e capulho) (**est**) teve como objetivo avaliar o número de capulhos de algodão produzidos (**ncap**). Esse experimento foi conduzido sob delineamento inteiramente casualizado com cinco repetições, em casa de vegetação. Esse conjunto de dados foi analisado em Zeviani et al. (2014) usando a distribuição *gamma-count*.

Seguindo os resultados de Zeviani et al. (2014) considera-se o preditor linear

$$\log(\mu_{ij}) = \beta_0 + \beta_{1j}\text{def}_i + \beta_{2j}\text{def}_i^2$$

em que i varia nos níveis de desfolha (1: 0%, 2: 25%, 3: 50%, 4: 75% e 5: 100%) e j nos estágios fenológicos da planta (1: vegetativo, 2: botão floral, 3: florescimento, 4: maçã, 5: capulho). As estimativas dos parâmetros para os modelos Poisson, COM-Poisson, COM-Poisson _{μ} e quasi-Poisson e medidas de qualidade de ajuste são exibidas na Tabela 2.

Os resultados apresentados na Tabela 2 mostram as medidas de qualidade de ajuste (verossimilhança maximizada, AIC e BIC) dos modelos COM-Poisson praticamente idênticas. O ajuste do modelo Poisson é claramente inferior aos demais. A diferença entre o logaritmo da verossimilhanças dos modelos Poisson e COM-Poisson _{μ} foi de 94,811 que, quando comparada com a distribuição qui-quadrado com 1 grau de liberdade, mostram que os modelos são significativamente diferentes ($\phi \neq 0$). O valor estimado para ϕ foi de 1,582, evidenciando subdispersão.

Ainda na Tabela 2, pode-se ver a vantagem do modelo COM-Poisson _{μ} (reparametrizado), pois as estimativas pontuais são muito similares às obtidas no modelo Poisson enquanto que, para o COM-Poisson essas estimativas não estão em uma escala interpretável. As razões entre estimativa e erro padrão sob os modelos COM-Poisson são muito similares à abordagem quasi-Poisson, porém com a vantagem de se ter uma distribuição de probabilidades associada às contagens.

Na Figura 3, apresentam-se as curvas de predição com bandas de confiança para todos

Tabela 2: Estimativas dos parâmetros (Est) e razões entre as estimativa e erro padrão (EP) para os quatro modelos ajustados aos dados subdispersos.

	Poisson		COM-Poisson		COM-Poisson _{μ}		Quasi-Poisson	
	Est	Est/EP	Est	Est/EP	Est	Est/EP	Est	Est/EP
ϕ, σ			1,585	12,417	1,582	12,392	0,241	
β_0	2,190	34,572	10,897	7,759	2,190	74,640	2,190	70,420
β_{11}	0,437	0,847	2,019	1,770	0,435	1,819	0,437	1,726
β_{12}	0,290	0,571	1,343	1,211	0,288	1,223	0,290	1,162
β_{13}	-1,242	-2,058	-5,750	-3,886	-1,247	-4,420	-1,242	-4,192
β_{14}	0,365	0,645	1,595	1,298	0,350	1,328	0,365	1,314
β_{15}	0,009	0,018	0,038	0,035	0,008	0,032	0,009	0,036
β_{21}	-0,805	-1,379	-3,725	-2,775	-0,803	-2,961	-0,805	-2,809
β_{22}	-0,488	-0,861	-2,265	-1,805	-0,486	-1,850	-0,488	-1,754
β_{23}	0,673	0,989	3,135	2,084	0,679	2,135	0,673	2,015
β_{24}	-1,310	-1,948	-5,894	-3,657	-1,288	-4,095	-1,310	-3,967
β_{25}	-0,020	-0,036	-0,090	-0,076	-0,019	-0,074	-0,020	-0,074
LogLik	-255,803		-208,250		-208,398		—	
AIC	533,606		440,500		440,795		—	
BIC	564,718		474,440		474,735		—	

os modelos. As predições pontuais foram as mesmas para ambos os modelos porém, os intervalos de predição são maiores no modelo Poisson devido à suposição de equidispersão.

Na Tabela 3, apresentam-se as correlações entre $\hat{\phi}$ e $\hat{\beta}$, calculadas a partir da matriz de covariâncias, para os modelos COM-Poisson e COM-Poisson _{μ} . As correlações são praticamente nulas quando considerado o modelo reparametrizado, indicando que esta reparametrização praticamente ortogonaliza o parâmetro extra da COM-Poisson com os parâmetros de regressão, nesse particular conjunto de dados.

Tabela 3: Correlações empírica entre $\hat{\phi}$ e $\hat{\beta}$ para os modelos ajustados aos dados subdispersos.

	$\hat{\beta}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$
COM-Poisson	1,00	0,22	0,15	-0,49	0,16	0,00	-0,35	-0,23	0,26	-0,46	-0,01
COM-Poisson _{μ}	0,00	-0,00	-0,00	-0,00	-0,00	-0,00	0,00	0,00	0,00	0,00	0,00

3.2 Dose potássica e umidade do solo em produtividade de soja

Nesse experimento, estudaram-se os componentes de produção da soja com relação a diferentes níveis de adubação potássica (K) aplicada ao solo (0, 0,3, 0,6, 1,2 e 1,8 100mg dm⁻³) e diferentes níveis de umidade do solo (umid) (37,5, 50, 62,5%, que representam pouca água, água em quantidade ideal e água em abundância respectivamente), caracterizando um experimento fatorial 5 × 3 (SERAFIM et al., 2012). O experimento foi

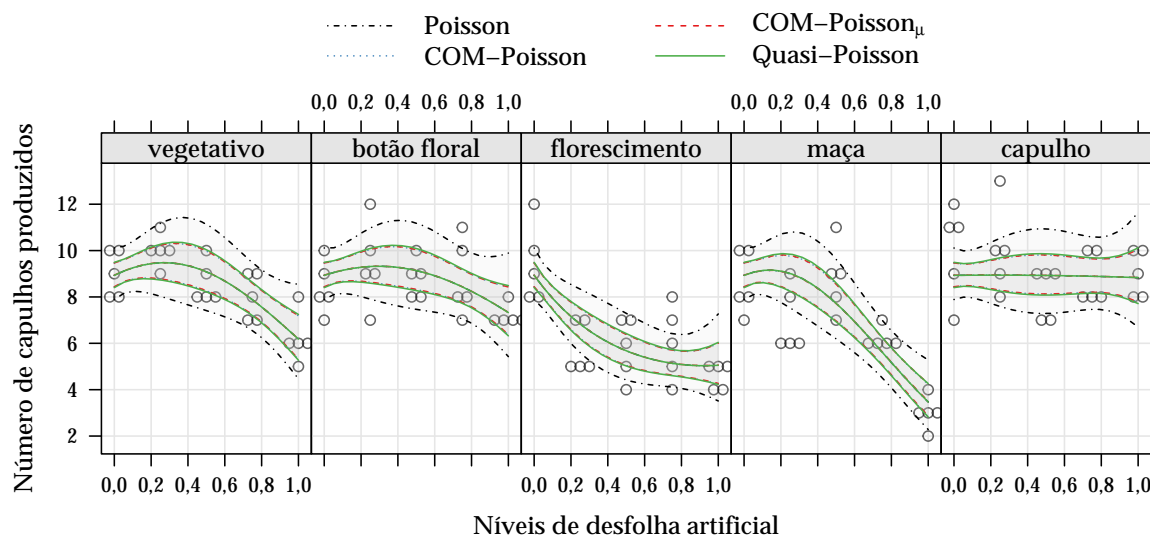


Figura 3: Curva dos valores preditos com intervalos de confiança de (95%) como função do nível de desfolha e estágio fenológico da planta.

instalado em casa de vegetação, no delineamento de blocos casualizados completos e a unidade experimental foi um vaso com duas plantas de soja. O objetivo do experimento foi avaliar a produção, mensurada pelo número de grãos de soja produzidos (**ngra**) com relação aos diferentes níveis de adubação potássica e umidade do solo.

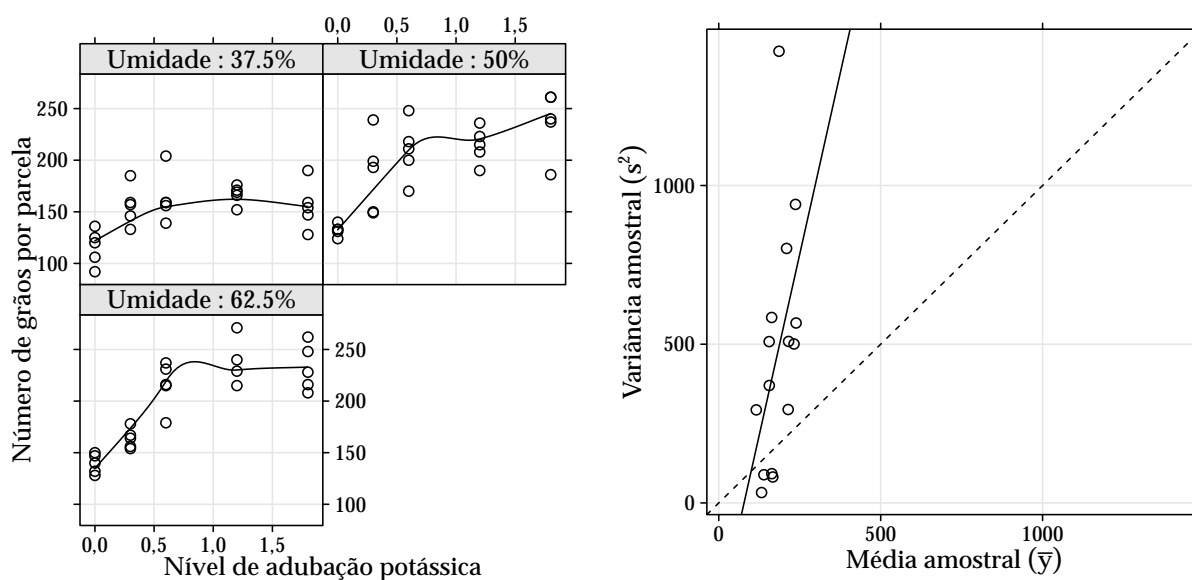


Figura 4: Diagrama de dispersão do número de grãos nos diferentes níveis de adubação potássica e umidade do solo (esquerda) e dispersão das médias e variâncias amostrais calculadas em cada tratamento experimental.

Na Figura 4 à esquerda, é apresentado o diagrama de dispersão dos dados na qual pode-se notar que há um efeito quadrático do nível de adubação potássica. No gráfico

à direita, são plotadas as médias e variâncias amostrais, sendo que a linha tracejada representa a reta identidade (suposição de equidispersão) e a contínua a reta de mínimos quadrados. A maioria dos pontos estão dispostos acima da linha tracejada, indicando superdispersão.

Para análise desse conjunto de dados propõe-se, com base na análise descritiva (gráfico à esquerda na Figura 4), o seguinte preditor linear

$$\log(\mu_{ijk}) = \beta_0 + \gamma_i + \tau_j + \beta_1 K_k + \beta_2 K_k^2 + \beta_{3j} K_k$$

com $i=1$: bloco II, 2: bloco III, 3: bloco IV e 4: V e $j=1$: 50% e 2: 62,5%, em que γ_i é o efeito do i -ésimo bloco, τ_j o efeito do j -ésimo nível de umidade aplicado e β_{3j} o efeito de primeira ordem de adubação potássica (K) para o j -ésimo nível de umidade do solo (umid).

Na Tabela 4, são apresentadas as estimativas dos parâmetros bem como os valores padronizados pelos seus respectivos erros padrão e medidas de qualidade de ajuste para todos os modelos em estudo. Assim como na Seção 3.2, os modelos COM-Poisson apresentaram medidas de qualidade de ajuste muito próximas e melhores do que as obtidas pelo modelo Poisson. O parâmetro de precisão ϕ estimado no modelo COM-Poisson $_{\mu}$ foi de -0,782, e a diferença do logaritmo das verossimilhanças dos modelos Poisson e COM-Poisson $_{\mu}$ foi 29,697 indicando que ϕ é significativamente diferente de 0. Quanto às estimativas e razões entre estimativas e erros padrão, as interpretações são análogas da seção anterior. Ambos os modelos são concordantes quanto à indicação de significância dos efeitos porém, o modelo Poisson indica efeitos com maior significância por não se adequar à variabilidade extra.

Nos casos de superdispersão a avaliação da função de verossimilhança é mais cara, devido à constante $Z(\mu, \phi)$ necessitar de um número elevado de incrementos para convergência. Nessa aplicação utilizaram-se 700 incrementos para cálculo das constantes. Para ajuste do modelo COM-Poisson padrão foram necessárias 264 avaliações da função de verossimilhança, enquanto que para o modelo reparametrizado COM-Poisson $_{\mu}$ apenas 78. Isso pode ser justificado pelo bom comportamento do logaritmo função de verossimilhança quando temos parâmetros ortogonais, o que facilita sua maximização. Os tempos para ajuste dos modelos COM-Poisson e COM-Poisson $_{\mu}$ foram de 15,831 e 7,267 segundos, respectivamente. A Tabela 5 apresenta as correlações entre o estimador para o parâmetro de dispersão os estimadores para os parâmetros de regressão, sendo as correlações para o modelo COM-Poisson $_{\mu}$ todas iguais a zero, considerando 3 casas decimais, o que indica a ortogonalidade empírica dos parâmetros.

Tabela 4: Estimativas dos parâmetros (Est) e razões entre as estimativa e erro padrão (EP) dos quatro modelos ajustados aos dados de soja.

	Poisson		COM-Poisson		COM-Poisson _μ		Quasi-Poisson	
	Est	Est/EP	Est	Est/EP	Est	Est/EP	Est	Est/EP
ϕ, σ			-0,779	-4,721	-0,782	-4,737	2,615	
β_0	4,867	144,289	2,232	6,042	4,867	97,781	4,867	89,225
γ_1	-0,019	-0,730	-0,009	-0,494	-0,019	-0,495	-0,019	-0,452
γ_2	-0,037	-1,373	-0,017	-0,921	-0,037	-0,931	-0,037	-0,849
γ_3	-0,106	-3,889	-0,049	-2,422	-0,106	-2,634	-0,106	-2,405
γ_4	-0,092	-3,300	-0,042	-2,102	-0,092	-2,237	-0,092	-2,040
τ_1	0,132	3,647	0,061	2,295	0,132	2,472	0,132	2,255
τ_2	0,124	3,432	0,057	2,177	0,124	2,326	0,124	2,122
β_1	0,616	11,014	0,284	4,729	0,616	7,464	0,616	6,811
β_2	-0,276	-10,250	-0,127	-4,589	-0,276	-6,946	-0,276	-6,338
β_{31}	0,146	4,268	0,067	2,614	0,146	2,892	0,146	2,639
β_{32}	0,165	4,829	0,076	2,884	0,165	3,272	0,165	2,986
LogLik	-340,082		-325,241		-325,233		—	
AIC	702,164		674,482		674,467		—	
BIC	727,508		702,130		702,116		—	

Tabela 5: Correlações empírica entre $\hat{\phi}$ e $\hat{\beta}$ para os modelos ajustados aos dados super-dispersos.

	$\hat{\beta}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{31}$	$\hat{\beta}_{32}$
COM-Poisson	1,00	-0,08	-0,15	-0,40	-0,34	0,38	0,36	0,77	-0,75	0,43	0,47
COM-Poisson _μ	0,00	0,00	0,00	-0,00	0,00	0,00	-0,00	-0,00	0,00	-0,00	0,00

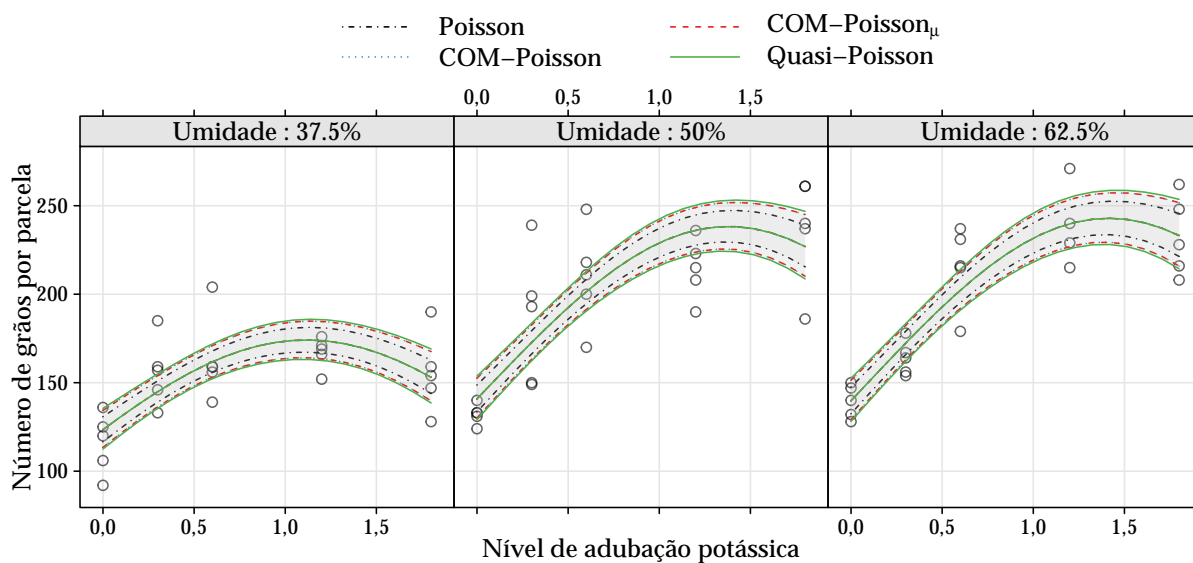


Figura 5: Curva dos valores preditos com intervalos de confiança de (95%) como função do nível de adubação potássica para cada nível de umidade.

Finalizando a análise desse conjunto de dados apresenta-se na Figura 5, as curvas de predição com bandas de confiança de 95% para todos os modelos. As médias pontuais são praticamente idênticas levando às mesmas interpretações para todos os modelos. Entretanto, os intervalos de predição são mais estreitos no caso Poisson, não por se acomodar melhor aos dados, mas sim pela restrição de equidispersão.

3.3 Nitrofenol em ambientes aquáticos

Um experimento inteiramente casualizado do tipo dose-resposta foi realizado com o objetivo de avaliar a toxicidade do nitrofenol, herbicida que foi utilizado extensivamente para o controle de ervas daninhas em cereais e arroz. No experimento, avaliaram-se 50 animais de uma espécie de zooplâncton (*Ceriodaphnia dubia*) submetidos 5 diferentes dosagens do herbicida nitrofenol (0, 0,8, 1,6, 2,35 e 3,10 caneca/10²litros) registrando-se o número total de ovos eclodidos após três ninhadas (BAILER; ORIS, 1994).

Para esse conjunto de dados foram propostos três preditores lineares a fim de avaliar o desempenho dos modelos usando o teste de razão de verossimilhança para seleção de covariáveis,

$$\text{Preditor 1: } \log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i$$

$$\text{Preditor 2: } \log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{dose}_i^2$$

$$\text{Preditor 3: } \log(\mu_i) = \beta_0 + \beta_1 \text{dose}_i + \beta_2 \text{dose}_i^2 + \beta_3 \text{dose}_i^3.$$

Na Tabela 6, apresentam-se as medidas de qualidade de ajuste e os testes de razão de verossimilhanças dos modelos constituídos dos três preditores lineares definidos. Todos os modelos indicam o efeito cúbico da dosagem do herbicida nas contagens de ovos eclodidos. Sob este preditor há evidências de equidispersão, estimativas $\hat{\phi}$ próximas a zero e $\hat{\sigma}$, da quasi-Poisson, próxima a 1. É interessante notar que se omitem os efeitos de ordem maior no preditor, os modelos passam a acusar superdispersão, isso exemplifica o que foi discutido na Seção 1 sobre as causas de superdispersão. Outro aspecto que se destaca é que embora o modelo quasi-Poisson seja robusto à não equidispersão, quando aplicado sob equidispersão apresentou nível descritivo maior do que os modelos paramétricos, ou seja, os testes realizados sob o modelo quasi-Poisson são menos poderosos.

Na Tabela 7, são apresentadas as estimativas dos parâmetros de regressão para os modelos considerando o efeito cúbico de **dose**. As interpretações são similares às seções anteriores porém, nesse caso têm-se o modelo Poisson, também, adequado para indicação da significância dos efeitos. Note, também, que os parâmetros do modelo COM-Poisson padrão são comparáveis aos demais, isso ocorre, pois se está no caso particular $\phi = 0$ o que implica que $\lambda = \mu$.

Tabela 6: Medidas de ajuste para avaliação e comparação entre preditores e modelos ajustados.

Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	P(> χ^2)	
Preditor 1	2	-180,667	365,335				
Preditor 2	3	-147,008	300,016	67,319	1	2,31E-16	
Preditor 3	4	-144,090	296,180	5,835	1	1,57E-02	
COM-Poisson	np	ℓ	AIC	2(diff ℓ)	diff np	P(> χ^2)	$\hat{\phi}$
Preditor 1	3	-167,954	341,908				-0,893
Preditor 2	4	-146,964	301,929	41,980	1	9,22E-11	-0,059
Preditor 3	5	-144,064	298,129	5,800	1	1,60E-02	0,048
COM-Poisson _{μ}	np	ℓ	AIC	2(diff ℓ)	diff np	P(> χ^2)	$\hat{\phi}$
Preditor 1	3	-167,652	341,305				-0,905
Preditor 2	4	-146,950	301,900	41,405	1	1,24E-10	-0,069
Preditor 3	5	-144,064	298,127	5,773	1	1,63E-02	0,047
Quasi-Poisson	np	QDev	AIC	F	diff np	P(> F)	$\hat{\sigma}$
Preditor 1	2	123,929					2,262
Preditor 2	3	56,610		60,840	1	5,07E-10	1,106
Preditor 3	4	50,774		5,659	1	2,16E-02	1,031

np, número de parâmetros; diff ℓ , diferença entre o logaritmo das verossimilhanças; QDev, valor da quasi-deviance, F, estatística F baseada nas quasi-deviances; diff np, diferença entre o np.

Tabela 7: Estimativas dos parâmetros (Est) e razões entre as estimativa e erro padrão (EP) dos quatro modelos ajustados aos dados de ovos eclodidos.

	Poisson		COM-Poisson		COM-Poisson _{μ}		Quasi-Poisson	
	Est	Est/EP	Est	Est/EP	Est	Est/EP	Est	Est/EP
ϕ, σ			0,048	0,236	0,047	0,232	1,031	
β_0	3,477	62,817	3,649	4,850	3,477	64,308	3,477	61,860
β_1	-0,086	-0,433	-0,091	-0,448	-0,088	-0,452	-0,086	-0,426
β_2	0,153	0,863	0,161	0,878	0,155	0,894	0,153	0,850
β_3	-0,097	-2,398	-0,102	-2,229	-0,098	-2,464	-0,097	-2,361

Na Figura 6, são exibidas as curvas de predição com bandas de confiança de cobertura de 95%. As curvas são totalmente sobrepostas, tanto para os valores preditos como para os intervalos de confiança. Isso indica que mesmo no caso, em que não é necessária a estimação do parâmetro ϕ , estimá-lo não leva a prejuízos para análise.

Finalmente, na Tabela 8 apresentam-se as correlações empíricas de $\hat{\phi}$ com $\hat{\beta}$. É interessante notar que mesmo no caso particular em que a COM-Poisson se resume à Poisson, as correlações empíricas para modelo padrão não são nulas. Para o modelo reparametrizado, assim como nos outras análises apresentadas, as correlações são praticamente nulas.

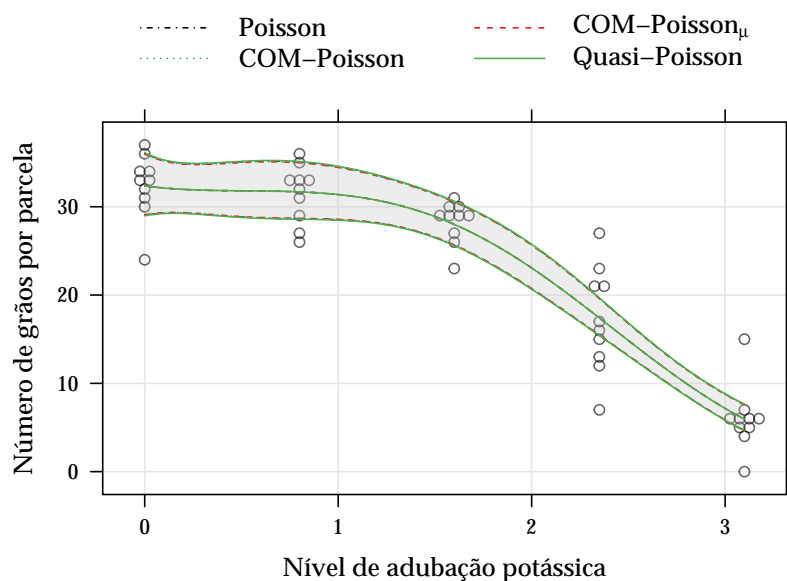


Figura 6: Diagrama de dispersão do número de ovos eclodidos para cada nível de dosagem experimentado, com curvas de predição e bandas de confiança de 95%.

Tabela 8: Correlações empírica entre $\hat{\phi}$ e $\hat{\beta}$ para os modelos ajustados aos dados equidispersos.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
COM-Poisson	0,997	-0,077	0,156	-0,422
COM-Poisson _{μ}	-0,000	0,002	-0,003	0,003

4 Conclusões

Neste artigo propõe-se o modelo COM-Poisson reparametrizado (COM-Poisson _{μ}) com base na aproximação da esperança da distribuição, para análise de dados de contagem sub, super e equidispersos provenientes de três experimentos planejados. Nas três aplicações o modelo reparametrizado é comparado com sua parametrização original e com os modelos Poisson e quasi-Poisson.

Aspectos da aproximação para média e consequentemente da reparametrização proposta foram apresentados, mostrando que a reparametrização é adequada. Nas análises de dados observou-se a característica de ortogonalidade empírica entre os estimadores para os parâmetros de regressão e de precisão no modelo reparametrizado, observado pelas correlações praticamente nulas entre os estimadores, algo que não ocorre para o modelo em sua parametrização original. Devido a essa característica o procedimento computacional é mais rápido sob o modelo reparametrizado. Outra vantagem, que favorece a reparametrização é que os parâmetros de regressão estimados têm interpretação de razão de taxas, assim como ocorre no modelo Poisson.

Os resultados sob os modelos COM-Poisson para as três situações (sub, super e equi-

dispersão) apresentaram resultados similares à abordagem baseado em suposições de momentos quasi-Poisson porém, com a vantagem de se adotar uma distribuição para as contagens possibilitando o cálculo de probabilidades, por exemplo.

De forma geral, os resultados apresentados pelos modelos COM-Poisson reparametrizados foram satisfatórios e superiores às abordagens convencionais. Sendo assim incentiva-se seu uso na análise de dados de contagem. As rotinas computacionais para ajuste dos modelos COM-Poisson e COM-Poisson reparametrizado são disponibilizadas no complemento online do artigo¹.

Como tópicos para pesquisas futuras sugere-se o estudo de aproximações para a constante de normalização do modelo COM-Poisson, uma vez que para conjuntos de dados com muitas observações o tempo computacional para avaliação da verossimilhança é elevado. Como forma de flexibilizar o modelo, a adoção de uma estrutura de regressão para o parâmetro ϕ pode ser útil nos casos em que a dispersão esteja relacionada às covariáveis. Como última sugestão, estudos de simulação para verificar a robustez do modelo à má especificação da distribuição da variável resposta, serão de grande valia.

Referências

- BAILER, A.; ORIS, J. Assessing toxicity of pollutants in aquatic systems. *In Case Studies in Biometry*, 1994.
- BONAT, W. H.; JØRGENSEN, B.; KOKONENDJI, C. C.; HINDE, J.; DÉMETRIO, C. G. B. Extended poisson-tweedie: properties and regression model for count data. *Statistical Modelling*, to appear, 2016.
- HINDE, J.; DÉMETRIO, C. G. B. Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, v. 27, n. 2, p. 151–170, 1998.
- LORD, D.; GEEDIPALLY, S. R.; GUIKEMA, S. D. Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis*, v. 30, n. 8, p. 1268–1276, 2010. ISSN 02724332.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, v. 135, p. 370–384, 1972.
- NOCEDAL, J.; WRIGHT, S. J. *Numerical optimization*. [S.l.]: Springer, 1995. 636 p. ISSN 0011-4235. ISBN 0387987932.
- R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016.
- RIBEIRO JR, E. E. *Extensões e Aplicações dos Modelos de Regressão COM-Poisson*. Monografia (Bacharelado) — Departamento de Estatística, Universidade Federal do Paraná, 2016.
- RIBEIRO JR, P. J.; BONAT, W. H.; KRAINSKI, E. T.; ZEVIANI, W. M. Métodos computacionais para inferência com aplicações em R. In: *20º Simpósio Nacional de Probabilidade e Estatística*. [S.l.: s.n.], 2012. p. 282.

- SELLERS, K. F.; BORLE, S.; SHMUELI, G. The com-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, v. 28, n. 2, p. 104–116, 2012.
- SELLERS, K. F.; SHMUELI, G. A flexible regression model for count data. *Annals of Applied Statistics*, v. 4, n. 2, p. 943–961, 2010. ISSN 19326157.
- SERAFIM, M. E.; ONO, F. B.; ZEVIANI, W. M.; NOVELINO, J. O.; SILVA, J. V. Umidade do solo e doses de potássio na cultura da soja. *Revista Ciência Agronômica*, v. 43, n. 2, p. 222–227, 2012. ISSN 1806-6690.
- SHMUELI, G.; MINKA, T. P.; KADANE, J. B.; BORLE, S.; BOATWRIGHT, P. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, v. 54, n. 1, p. 127–142, 2005.
- WEDDERBURN, R. W. M. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, v. 61, n. 3, p. 439, 1974. ISSN 00063444.
- WINKELMANN, R. Duration Dependence and Dispersion in Count-Data Models. *Journal of Business & Economic Statistics*, v. 13, n. 4, p. 467–474, 1995. ISSN 0735-0015.
- WINKELMANN, R. *Econometric Analysis of Count Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 342 p.
- ZEVIANI, W. M.; Ribeiro Jr, P. J.; BONAT, W. H.; SHIMAKURA, S. E.; MUNIZ, J. A. The Gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics*, p. 1–11, 2014.