# 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 $\lambda$, que representa a média e também a variância, o que implica em uma relação identidade ($\lambda = \mathrm{E}(Y) = \mathrm{V}(Y)$). Essa propriedade, chamada de equidispersão, é uma particularidade do modelo Poisson que inadequada a 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* ($\mathrm{E}(Y) < \mathrm{V}(Y)$) que podem ocorrer por diversão razões como heterogeneidade das unidades experimentais, ausência de covariáveis experimentais, diferentes amplitudes de domínio (*offset*) não considerados, correlação entre as observações, excesso de zero ente outras. O caso menos comum, mas que tem ganhado a atenção da comunidade estatística é a *subdispersão* ($\mathrm{E}(Y) > \mathrm{V}(Y)$). Os processos que reduzem a variabilidade das contagens, abaixo do estabelecido pela Poisson, não são tão conhecidos quanto os que produzem variabilidade

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

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 a 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) e Poisson Inversa-Gaussiana (PIG). Para o caso de subdispersão há os modelos da classe *Duration Dependence*, que flexibilizam a distribuição do tempo entre evento, assumida Exponencial no caso Poisson (WINKELMANN, 2008). Um representante desses modelos é o *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 a 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 a 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 do modelo COM-Poisson para análise de quatro experimentos planejados. 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 nós propomos uma reparametrização do modelo COM-Poisson, incorporando um parâmetro de média ao modelo. Essa reparametrização apresenta as vantagens para interpretação e para ajuste dos modelos, uma vez que encontramos relação aproximadamente ortogonal dos parâmetros de média em relação ao parâmetro de precisão. Com aplicações para análise de dados experimentais, contemplando as situação de sup, super e equidispersão nós discutimos 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 em 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*.

# 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

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

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 $\nu$ combinados com grandes valores de $\lambda$, 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 $\lambda$ e $\nu$ [1].

Um inconveniente desse modelo é que os momentos média e variância não tem forma fechada. Shmueli et al. (2005), a partir de uma aproximação para $Z(\lambda,\nu)$ apresentam uma forma aproximada para os dois primeiros momentos da distribuição

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

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 $\frac{1}{\nu}E(Y)$. Na Figura 2 (a) apresentamos a relação média-variância, com médias e variâncias calculadas numericamente para diferentes níveis de dispersão. É interessante notar o aspecto linear dessa relação.

Nesse artigo nos propomos uma reparametrização do modelo COM-Poisson baseada na expressão aproximada para média (2). A reparametrização consiste na introdução do

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

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

| | | | $\boldsymbol{\lambda}$ | | | |
|---|---|---|---|---|---|---|
| $\boldsymbol{\nu}$ | 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 |

parâmetro $\mu$ definido como

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

O parâmetro de precisão da distribuição é tomado na escala do logaritmo neperiano para evitar a restrição no espaço paramétrico, denotamos por $\phi$ esse novo parâmetro, ou seja, $\phi = \log(\nu) \Rightarrow \phi \in \mathbb{R}$. Para esse parâmetro as interpretações são como se segue

$$\phi < 0 \Rightarrow \text{Superdispersão}; \quad \phi = 0 \Rightarrow \text{Equidispersão; e} \quad \phi > 0 \Rightarrow \text{Subdispersão}$$

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

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

Denotamos a distribuição COM-Poisson reparametrizada por COM-Poisson$_{\mu}$.Na Figura 1 apresentamos distribuições de probabilidade para diferentes parâmetros ilustrando a flexibilidade da distribuição. As distribuições em vermelho representam o caso particular quando a COM-Poisson se resume a distribuição Poisson.

Uma avaliação da acurácia das aproximações em (2) e consequentemente da reparametrização, é apresentada na Figura 2 (b) e (c) para valores de $\mu$ variando de 0 a 30 e diferentes níveis de dispersão ($-1.2 < \log(\nu) < 1$). Em (b) temos os erros quadráticos para a média, $(\mu - E[X])^2$ e em (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 bem

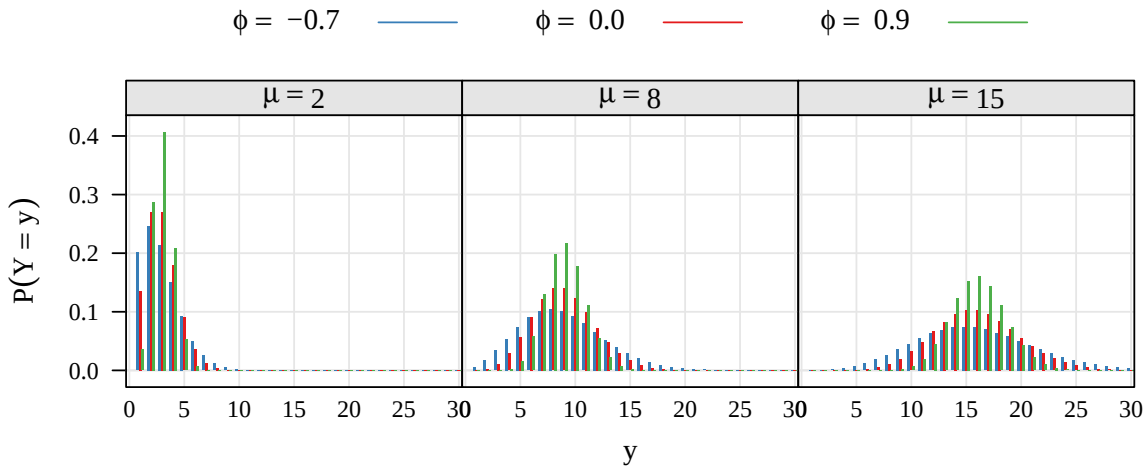


Figura 1: Probabilidades pela distribuição COM-Poisson$_\mu$ para diferentes parâmetros.

acurada (valores menores que 0.03), porém tem perda de acurácia para combinações de $\nu$ pequenos ($< 0.35$) com médias também baixas ($\mu < 10$). Para a variância, gráfico (c), temos que a aproximação $V[X] = \frac{\mu}{\nu}$, proposta por Sellers e Shmueli (2010), não tem bom desempenho para valores pequenos de $\nu$. A restrição que envolve e média da distribuição parece não ter relevância. De forma geral, para média e variância, os valores de $\nu$ parecem ser mais influentes na acurácia do que os valores de $\mu$.

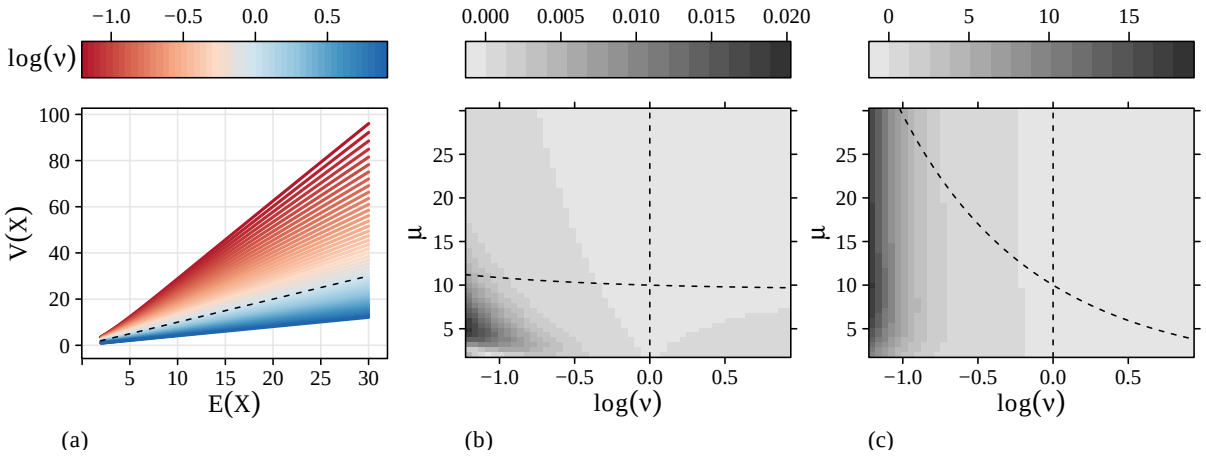


Figura 2: Representação dos momentos da distribuição COM-Poisson oaara diferentes médias e níveis de dispersão. (a) relação média-variância. (b) e (c) erros quadráticos da aproximações para média e variância respectivamente.

Os resultados apresentados na Figura 2 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 independentes de contagens $y_i$, $i = 1, 2 \ldots, n$, as estimativas para $\boldsymbol{\theta} = (\beta\,, \phi)$ são obtidas pelos argumentos que maximizam a função de log-verossimilhança

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

em que $\mu_i = e^{\underline{x}_i^t\beta}$, com $\underline{x}_i$ o vetor $(x_{i1}, x_{i2}, \ldots x_{ip})$ de covariáveis da i-ésima observação, e $(\beta, \phi) \in \mathbb{R}^{p+1}$. A constante $Z(\mu_i, \phi)$ é calculada como

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

Note que a avaliação da log-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 (**??**). Como $\ell(\mu, \phi \mid \underline{y})$ não possui derivada analítica, a maximação é realizada pelo algoritmo BFGS (NOCEDAL; WRIGHT, 1995) que fornece 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 da log-verossimilhança (método de Wald), fazendo $\sqrt{\boldsymbol{v}}$, em que $\boldsymbol{v}$ são os elementos da diagonal da matriz $-\mathcal{H}^{-1}(\boldsymbol{\theta})$. Intervalos de confiança para $\hat{\mu}_i$ são obtidos pelo método delta.

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

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

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

Nas aplicações também utilizamos a abordagem por 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 $\sigma$, $V(y_i) = \sigma V(\mu_i)$. Esses modelos são ajustados no *software* R (R CORE TEAM, 2016), com a função `glm(..., family = quasipoisson)`.

# 3 Análise de Dados Experimentais

## 3.1 Caso subdisperso: Experimento de desfolha artificial em capulhos de algodão

Experimento com plantas de algodão *Gossypium hirsutum* submetidas à 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, maça e capulho) (`est`) com o objetivo de 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) consideramos 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: maça, 5: capulho). As estimativas dos parâmetros sob os modelos Poisson, COM-Poisson, COM-Poisson$_\mu$ e Quasi-Poisson e medidas de qualidade de ajuste são exibidos na Tabela 2.

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

| | Poisson | | COM-Poisson | | COM-Poisson$_\mu$ | | Quasi-Poisson | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP |
| $\phi\,\sigma^2$ | | | 1.585 | 12.417 | 1.582 | 12.392 | 0.241 | |
| $\beta_0$ | 2.190 | 34.572 | 10.897 | 7.759 | 2.190 | 74.640 | 2.190 | 70.420 |
| $\beta_{11}$ | 0.437 | 0.847 | 2.019 | 1.770 | 0.435 | 1.819 | 0.437 | 1.726 |
| $\beta_{12}$ | 0.290 | 0.571 | 1.343 | 1.211 | 0.288 | 1.223 | 0.290 | 1.162 |
| $\beta_{13}$ | -1.242 | -2.058 | -5.750 | -3.886 | -1.247 | -4.420 | -1.242 | -4.192 |
| $\beta_{14}$ | 0.365 | 0.645 | 1.595 | 1.298 | 0.350 | 1.328 | 0.365 | 1.314 |
| $\beta_{15}$ | 0.009 | 0.018 | 0.038 | 0.035 | 0.008 | 0.032 | 0.009 | 0.036 |
| $\beta_{21}$ | -0.805 | -1.379 | -3.725 | -2.775 | -0.803 | -2.961 | -0.805 | -2.809 |
| $\beta_{22}$ | -0.488 | -0.861 | -2.265 | -1.805 | -0.486 | -1.850 | -0.488 | -1.754 |
| $\beta_{23}$ | 0.673 | 0.989 | 3.135 | 2.084 | 0.679 | 2.135 | 0.673 | 2.015 |
| $\beta_{24}$ | -1.310 | -1.948 | -5.894 | -3.657 | -1.288 | -4.095 | -1.310 | -3.967 |
| $\beta_{25}$ | -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 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 modelo Poisson é claramente inferior aos demais. A diferença entre as log-verossimilhanças dos modelos Poisson e COM-Poisson$_\mu$ 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 $\phi$ foi de 1.582, evidenciando a subdispersão das contagens.

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

Na Figura 3 nós apresentamos as curvas de predição com bandas de confiança para todos os modelos. As predições pontuais foram as mesmas para ambos os modelos, porém os intervalos de predição são maiores no modelo Poisson devido a suposição de equidispersão.

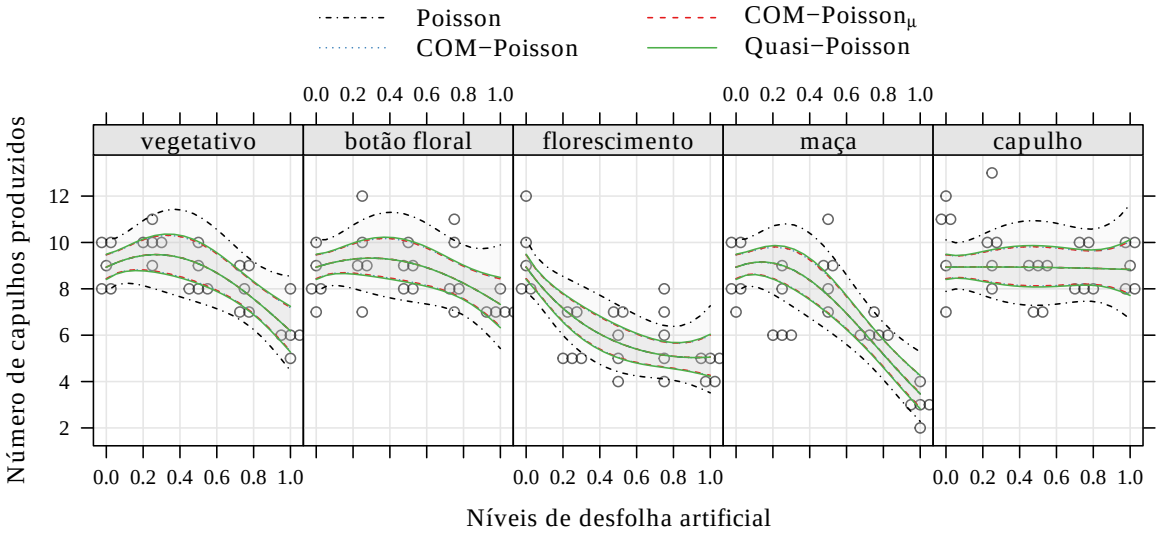


Figura 3: Curva dos valores preditos com intervalos de confiança de (95%) como função do nível de desfolha e estágio fenológico da planta.

Na Tabela 3 apresentamos as correlações do parâmetro $\phi$ com os demais parâmetros de regressão $\boldsymbol{\beta}$, calculadas a partir da matriz de covariâncias, para os modelos COM-Poisson e COM-Poisson$_\mu$. As correlações são praticamente nulas quando considerado o modelo reparametrizado, indicando que esta reparametrização ortogonaliza o parâmetro extra da COM-Poisson com os parâmetros de regressão.

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

| | $\beta_0$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | $\beta_{14}$ | $\beta_{15}$ | $\beta_{21}$ | $\beta_{22}$ | $\beta_{23}$ | $\beta_{24}$ | $\beta_{25}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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$_\mu$ | 0.00 | -0.00 | -0.00 | -0.00 | -0.00 | -0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

## 3.2 Caso superdisperso: Experimento de dose potássica e umidade do solo em produtividade de soja

Nesse experimento estudou-se os componentes de produção da soja com relação à 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 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 é avaliar a produção, mensurada pelo número de grão de soja produzidos (`ngra`) com relação aos diferentes níveis de adubação potássica e umidade do solo.

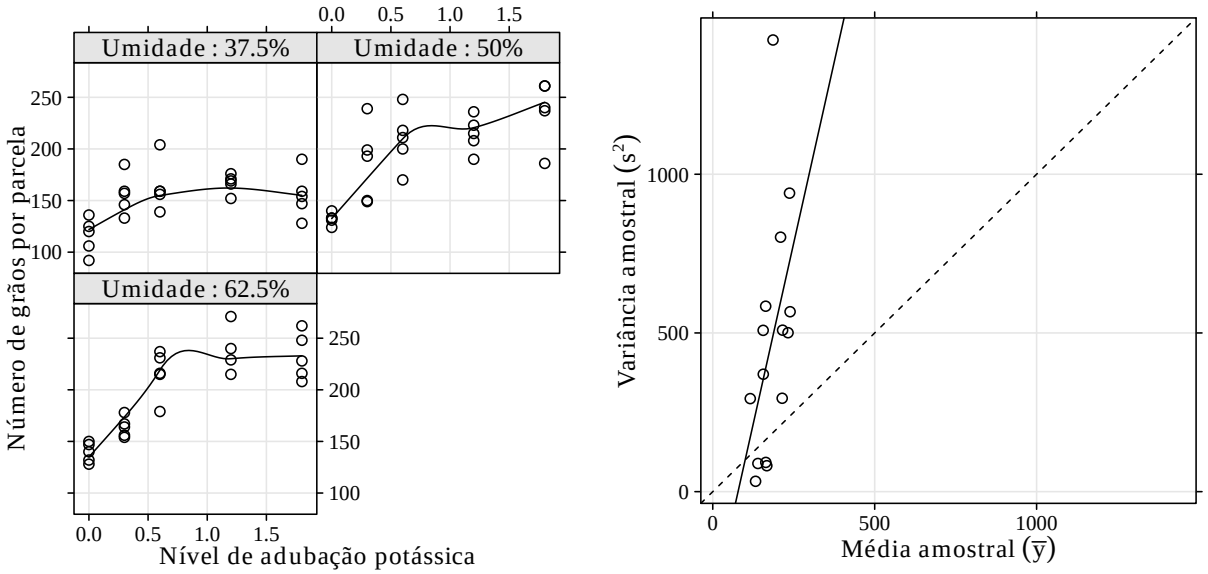


Figura 4: Diagrama de dispersão do número de grãos nos diferentes níveis de adubação potássica e umidade do solo (esquerda) e dispersão das médias e variâncias amostrais calculadas em cada tratamento experimental.

Na Figura 4 à esquerda é apresentado o diagrama de dispersão dos dados onde podemos notar que há um efeito quadrático do nível de adubação potássica. À direita são apresentadas as médias e variâncias amostrais, 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 a superdispersão das contagens.

Na modelagem desse conjunto de dados nós propomos, com base na análise descritiva (Figura 4), o preditor linear

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

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

Na Tabela 4 são apresentas 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 $\phi$ estimado no modelo COM-Poisson$_\mu$ foi de 315.42, e a diferença das log-verossimilhanças dos modelos Poisson e COM-Poisson$_\mu$ foi 29.697 indicando que $\phi$ é significativamente diferente de 0. Quanto as estimativas e razões entre estimativa e erros padrão as interpretações são análogas da seção anterior. Ambos os modelos são concordantes quanto a indicação de significância dos efeitos, porém o modelo Poisson indica efeitos com maior significância por não se adequar a variabilidade extra.

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

| | Poisson | | COM-Poisson | | COM-Poisson$_\mu$ | | Quasi-Poisson | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP |
| $\phi\,\sigma^2$ | | | -0.779 | -4.721 | -0.782 | -4.737 | 2.615 | |
| $\beta_0$ | 4.867 | 144.289 | 2.232 | 6.042 | 4.867 | 97.781 | 4.867 | 89.225 |
| $\gamma_1$ | -0.019 | -0.730 | -0.009 | -0.494 | -0.019 | -0.495 | -0.019 | -0.452 |
| $\gamma_2$ | -0.037 | -1.373 | -0.017 | -0.921 | -0.037 | -0.931 | -0.037 | -0.849 |
| $\gamma_3$ | -0.106 | -3.889 | -0.049 | -2.422 | -0.106 | -2.634 | -0.106 | -2.405 |
| $\gamma_4$ | -0.092 | -3.300 | -0.042 | -2.102 | -0.092 | -2.237 | -0.092 | -2.040 |
| $\tau_1$ | 0.132 | 3.647 | 0.061 | 2.295 | 0.132 | 2.472 | 0.132 | 2.255 |
| $\tau_2$ | 0.124 | 3.432 | 0.057 | 2.177 | 0.124 | 2.326 | 0.124 | 2.122 |
| $\beta_1$ | 0.616 | 11.014 | 0.284 | 4.729 | 0.616 | 7.464 | 0.616 | 6.811 |
| $\beta_2$ | -0.276 | -10.250 | -0.127 | -4.589 | -0.276 | -6.946 | -0.276 | -6.338 |
| $\beta_{31}$ | 0.146 | 4.268 | 0.067 | 2.614 | 0.146 | 2.892 | 0.146 | 2.639 |
| $\beta_{32}$ | 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 | | — | |

Nos casos de superdispersão a avaliação da verossimilhança é mais cara, devido a constante $Z(\mu, \phi)$ necessitar de um número elevado de incrementos para convergência. Nessa aplicação utilizamos 700 incrementos para cálculo das constantes. Para ajuste do modelo COM-Poisson padrão foram necessárias 264 avaliações da verossimilhança, enquanto que para o modelo reparametrizado COM-Poisson$_\mu$ apenas 78. Isso pode ser justificado pelo bom comportamento da log-verossimilhança quando temos parâmetros ortogonais, o que facilita sua maximização. A Tabela 5 apresenta as correlações do parâmetro de dispersão com os parâmetros de regressão, as correlações para o modelo COM-Poisson$_\mu$ são todas iguais a zero, considerando 3 casas decimais, o que indica a ortogonalidade empírica dos parâmetros.

Tabela 5: Correlações empírica entre $\hat{\phi}$ e os parâmetros de locação $\boldsymbol{\beta}$ para os modelos ajustados aos dados superdispersos

| | $\beta_0$ | $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_4$ | $\tau_1$ | $\tau_2$ | $\beta_1$ | $\beta_2$ | $\beta_{31}$ | $\beta_{32}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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$_\mu$ | 0.00 | 0.00 | 0.00 | -0.00 | 0.00 | 0.00 | -0.00 | -0.00 | 0.00 | -0.00 | 0.00 |

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

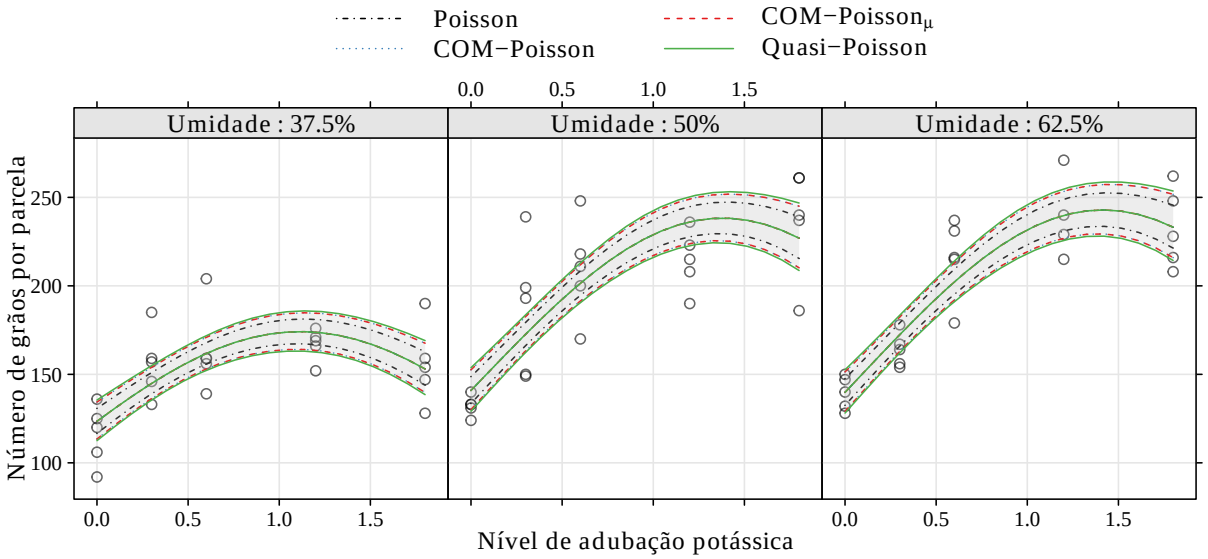


Figura 5: Curva dos valores preditos com intervalos de confiança de (95%) como função do nível de adubação potássica para cada nível de umidade.

## 3.3 Caso equidisperso: Experimento de nitrofeno em ambientes aquáticos

Experimento inteiramente casualizado do tipo dose-resposta com o objetivo de avaliar a toxidade do Nitrofeno, herbicida que foi utilizado extensivamente para o controle de ervas daninhas em cereais e arroz. No experimento avaliou-se 50 animais de uma espécie de zooplâncton (*Ceriodaphnia dubia*) submetidos 5 diferentes dosagens do herbicida Nitrofeno (0, 0.8, 1.6, 2.35 e 3.10 caneca/$10^2$litros) registrando o número total de ovos eclodidos após três ninhadas (BAILER; ORIS, 1994).

Para esse conjunto de dados testamos três preditores lineares a fim de avaliar o desempenho dos modelos em testes de razão de verossimilhanças para seleção de variáveis regressoras. Os preditores são

$$\begin{aligned}
&\text{Preditor 1: } \log(\mu_i) = \beta_0 + \beta_1 \texttt{dose}_i \\
&\text{Preditor 2: } \log(\mu_i) = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 \\
&\text{Preditor 3: } \log(\mu_i) = \beta_0 + \beta_1 \texttt{dose}_i + \beta_2 \texttt{dose}_i^2 + \beta_3 \texttt{dose}_i^3
\end{aligned}$$

Na Tabela 6 apresentamos as medidas de qualidade de ajuste e os testes de razão de verossimilhanças dos modelos constituídos dos três preditores lineares definidos acima. 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 quando omitimos os efeitos de ordem maior no preditor, os modelos passam a acusar superdispersão, isso exemplifica o que foi discutido na seção **??** sobre as causas da superdispersão. Outro aspecto que destacamos é que embora o modelo Quasi-Poisson seja robusto à não equidispersão, quando aplicado sob equidispersão apresentou nível descritivo maior 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 temos o modelo Poisson também adequado para indicação da significância dos efeitos. Note também que os parâmetros do modelo COM-Poisson padrão são comparáveis com os demais, isso ocorre, pois estamos 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 | $\ell$ | AIC | 2(diff $\ell$) | diff np | P($>\chi^2$) | |
|---|---|---|---|---|---|---|---|
| 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 | $\ell$ | AIC | 2(diff $\ell$) | diff np | P($>\chi^2$) | $\hat{\phi}$ |
| 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$_\mu$ | np | $\ell$ | AIC | 2(diff $\ell$) | diff np | P($>\chi^2$) | $\hat{\phi}$ |
| 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 $\ell$, diferença entre log-verossimilhanças; QDev, valor da quasi-deviance, F, estatística F baseada nas quasi-deviances; diff np, diferença entre o np.

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

| | Poisson | | COM-Poisson | | COM-Poisson$_\mu$ | | Quasi-Poisson | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP | Estimate | Est/EP |
| $\phi\,\sigma^2$ | | | 0.048 | 0.236 | 0.047 | 0.232 | 1.031 | |
| $\beta_0$ | 3.477 | 62.817 | 3.649 | 4.850 | 3.477 | 64.308 | 3.477 | 61.860 |
| $\beta_1$ | -0.086 | -0.433 | -0.091 | -0.448 | -0.088 | -0.452 | -0.086 | -0.426 |
| $\beta_2$ | 0.153 | 0.863 | 0.161 | 0.878 | 0.155 | 0.894 | 0.153 | 0.850 |
| $\beta_3$ | -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 $\phi$, estimá-lo não leva à prejuízos para análise.

Finalmente na Tabela 8 apresentamos as correlações empíricas do parâmetro de dispersão $\phi$ com os de locação $\beta$. É interessante notar que mesmo no caso particular em que a COM-Poisson se resumo 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.

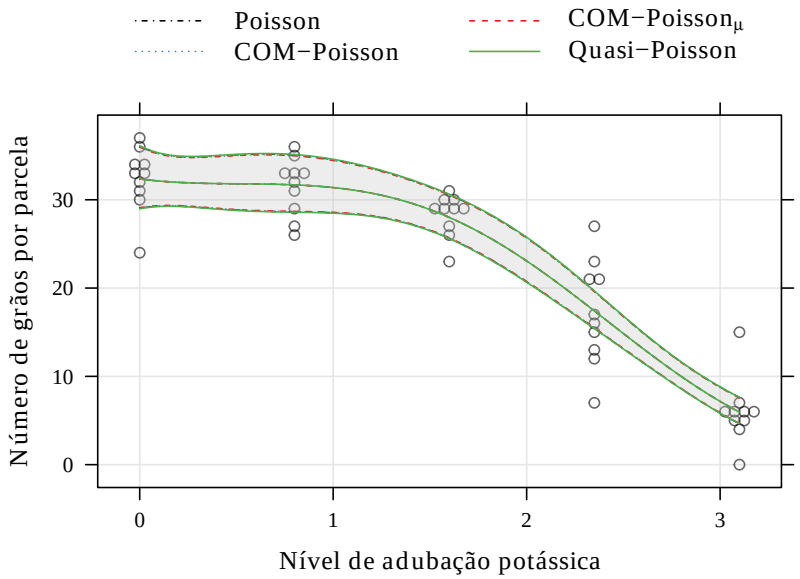


Figura 6: Diagrama de dispersão do número de ovos eclodidos para cada nível de dosagem esperimentadocom curvas de predição e bandas de confiança de 95%.

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

| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|---|---|---|---|---|
| COM-Poisson | 0.997 | -0.077 | 0.156 | -0.422 |
| COM-Poisson$_\mu$ | -0.000 | 0.002 | -0.003 | 0.003 |

# 4 Conclusões

Neste trabalho propomos o modelo COM-Poisson reparametrizado (COM-Poisson$_\mu$) 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 característica de ortogonalidade entre os parâmetros de regressão e de precisão no modelo reparametrizado, 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 equidispersão) apresentaram resultados similares à abordagem semi-paramétrica 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 as abordagens convencionais. Sendo assim nós incentivamos 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 artigo.

Como tópicos para pesquisas futuras sugerimos o estudo de aproximações para a constante de normalização do modelo COM-Poisson, uma vez que para conjunto 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 $\phi$ pode ser útil em casos 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. ANDHINDE, J.; DÉMETRIO, C. G. B. Extended poisson-tweedie: properties and regression model for count data. *Arxiv*, 2016.

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*. Dissertação (Mestrado) — Universidade Federal do Paraná, 2016.

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. ISSN 00359254.

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. ISBN 978-3-540-77648-2.

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. ISSN 0266-4763.