ISSN: 2317-0840

Uso das distribuições Poisson, Poisson-Gama, Poisson-Inversa Gaussiana e Poisson-Lindley Generalizada para dados de contagem

Silvia M. Freitas, Caroline G. Duarte[†]

Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará (UFC).

Resumo: *A análise usual para dados discretos é atráves de uma distribuição de Poisson, Binomial ou Binomial Negativa, via Modelos Lineares Generalizados (MLG). Entretanto, um dos cuidados que se deve ter ao fazer a análise de dados discretos, é com a superdispersão. Esse termo é utilizado quando a presença de variação nos dados excede à variância nominal estipulada pelo modelo proposto. Dessa forma, a utilização de um modelo com base apenas na distribuição Poisson, que tem como suposição a equidispersão, não se apresenta como uma opção adequada. Uma alternativa para dados com essa característica é o uso das distribuições compostas, através dos modelos em dois estágios, ou hierárquicos, como uma forma para modelar essa superdispersão. A metodologia dos modelos em dois estágios associa, uma distribuição à resposta condicionada a sua média e, posteriormente, uma distribuição ao parâmetro de média, de forma que, incondicionalmente, se tem uma distribuição composta para a variável resposta. Neste trabalho é utilizada a distribuição clássica de Poisson, para dados de contagem, e as distribuições Gama, Inversa Gaussiana e Lindley Generalizada para o parâmetro de média da Poisson, que implicam nas distribuições compostas Poisson-Gama, Poisson-Inversa Gaussiana e a Poisson-Lindley Generalizada. Assim, o objetivo principal deste trabalho é apresentar esses modelos hierárquicos, que permitem a modelagem de dados de contagem com superdispersão. Também foram abordados alguns tipos de resíduos da estrutura dos MLGs, adaptados para as distribuições compostas.*

Palavras-chave: *Poisson, Superdispersão, Distribuições Compostas, Modelos Lineares Generalizados, Modelos Hierárquicos.*

Abstract: *The usual analysis for discrete data is through a Poisson, Binomial or Negative Binomial distribution, via Generalized Linear Models (GLM). However, one of the precautions to be taken when analyzing discrete data is overdispersion. This term is used when the presence of variation in the data exceeds the nominal variance stipulated by the proposed model. Thus, the use of a model based on the Poisson distribution, which assumes equidispersion, would be unfounded in the presence of overdispersion. An alternative to this problem is to use mixed distributions, through models in two stages, or hierarchical, as a way to accommodate this overdispersion. The methodology of the two-stage models associates a distribution to the response conditioned to its average and, later, a distribution to the average parameter, so that, unconditionally, there is a compound distribution for the response variable. In this work, the classical Poisson distribution is used for counting data, and the Gama, Inverse Gaussian and Generalized Lindley distributions for the Poisson mean parameter, thus generating the Poisson-Gama, Poisson-Inverse Gaussian and Generalized Poisson-Lindley. So, the main objective of this work is to present these hierarchical models, or models in two stages, that allow the modeling of count data with overdispersion. Futhermore, some types of residues of the structure of the MLGs were also approached, adapted for the composite distributions.*

Keywords: *Poisson Distribution, Overdispersion, Compound Distribution, Generalized Linear Models, Hierarchical Models.*

[†]Autora correspondente: carolgduarte@hotmail.com.

Introdução

Em muitas áreas é frequente deparar-se com a investigação de características, feitas em unidades experimentais, que se apresentem na forma de contagem, por exemplo, o número de defeitos que podem aparecer em um processo de produção (Engenharia) ou o número de sinistros associados a uma carteira de seguros (Atuária). Dados deste tipo são denominados como dados discretos, pois são expressos em termos de contagens associados a uma característica de interesse (BICKEL; DOKSUM, 1977). De forma geral, dados de contagem podem ser modelados através da distribuição Poisson. Entretanto, ao se utilizar o modelo probabilístico Poisson, acabamos por assumir que a média e variância são iguais. Sendo que a restrição de modelar média igual à variância faz com que tal modelo tenha uma forte limitação de aplicação em diversas áreas. Visto que, na prática, é bastante comum que dados de contagem apresentem uma variância superior à variância nominal da distribuição Poisson assumida pelo modelo. Tal comportamento é denominado de superdispersão.

Uma forma de contemplar essa superdispersão é utilizar os modelos para dados inflacionados de zeros, conhecidos como ZIP - Poisson Inflacionada de Zeros e ZINB - Binomial-Negativa Inflacionada de Zeros (LAMBERT, 1992; HINDE e DEMÉTRIO, 1998; HILBE, 2014). Os modelos em dois estágios também podem ser utilizados como outra alternativa para modelar a superdispersão (HINDE; DEMETRIO, 1998). De forma que, segundo Hinde e Demétrio (1998) e, utilizado em Mendes (2017), se associa uma distribuição à resposta condicionada à sua média, e uma distribuição ao parâmetro de média, obtendo-se então uma distribuição composta para a variável resposta.

A proposta do trabalho é fazer um estudo dos modelos através da modelagem com o uso de distribuições compostas, Poisson-Gama (GREENWOOD; YULE, 1920), Poisson-Inversa Gaussiana (HOLLA, 1967) e a Poisson-Lindley Generalizada (WONGRIN; BODHISUWAN, 2016), para dados de contagem inflacionados de zeros. Será estudada a performance de cada abordagem e o impacto dessas modelagens na qualidade dos ajustes, usando-se conjuntos de dados reais. Além de abordar alguns tipos de resíduos da estrutura dos MLGs. Para o desenvolvimento dos algoritmos necessários para o uso dos métodos acima descritos, foi-se utilizado o *software* estatístico `R` (Rstudio Team, 2020).

Modelos Lineares Generalizados

No modelo clássico de regressão linear, a distribuição Normal tem um papel fundamental, visto que se assume que a fonte de variação e a variável resposta seguem uma distribuição Normal, permitindo que o procedimento de inferência exatas destes modelos sejam realizados (MYERS et al., 2010). Entretanto, muitas vezes na prática, algumas pressuposições como aditividade do componente sistemático do modelo, normalidade ou variância constante, não são atendidas, fazendo com que um modelo linear clássico não seja apropriado (COSTA, 2003). Dessa forma, o MLG, introduzido por Nelder e Wedderburn (1972), nos permite ajustar um modelo de regressão onde a distribuição considerada não precisa ser necessariamente Normal, podendo seguir qualquer distribuição da classe de distribuições chamada família exponencial.

Superdispersão para dados de contagem com excesso de zeros

Um dos cuidados que se deve ter ao fazer a análise de dados de contagem é a possibilidade da existência de superdispersão. Esse termo é usado quando a razão entre variância e a média é maior do que um. Isso pode ocorrer em diversas áreas, tanto que McCullagh e Nelder (1989) citam que, na prática, não é incomum obter superdispersão, a dispersão nominal sendo a exceção.

Uma atenção maior deve ser utilizada em situações com a superdispersão visto que a não consideração desta, na análise dos dados, pode levar a sub ou superestimação dos erros-padrão das estimativas dos parâmetros do modelo em questão, causando assim uma estimação incorreta destes e, como consequência, podendo levar a uma tomada de decisão errônea (HINDE; DEMETRIO, 1998). E ela pode ocorrer por várias razões, uma delas é quando existe excesso de zeros no conjunto de dados implicando em uma

variabilidade maior do que a esperada. Dessa forma, utilizar um modelo com base em uma distribuição de Poisson não seria adequado, visto que esta possui a propriedade de equidispersão.

Algumas alternativas são propostas na literatura para contornar o problema da inflação de zeros no conjunto. Entre eles, podemos citar a utilização de distribuições alternativas, como mostrado por Breslow (1984); abordar uma forma mais geral para a função de variância, conforme McCullagh e Nelder (1989); ou, assumir distribuições compostas para a variável resposta, conforme descrito por Hinde e Demétrio (1998).

Neste trabalho são utilizados, como alternativa, os modelos hierárquicos. Sendo abordado da seguinte forma: em primeiro estágio, associa-se uma distribuição à variável resposta condicionada à sua média e, em segundo estágio, associa-se uma distribuição à média, de modo que, incondicionalmente, tem-se uma distribuição composta para a variável resposta.

De forma específica, são abordados os modelos Poisson, Poisson-Gama, Poisson-Inversa Gaussiana e Poisson-Lindley Generalizada, todos derivados de misturas da distribuição de Poisson com as distribuições Gama, Inversa Gaussiana (IG) e a Lindley Generalizada (LG), para modelar o parâmetro de média (λ) da Poisson. Dessa forma se tem:

$$\begin{cases} Y_i|\theta_i \sim \text{Pois}(\lambda_i) \\ \lambda_i \sim \text{Gama}(\mu, \phi) \text{ ou } \text{IG}(\mu, \sigma) \text{ ou } \text{LG}(\alpha, \theta, \beta = 1). \end{cases}$$

Modelo Poisson

Proposta por Siméon-Denis Poisson (1937), o modelo Poisson possui um papel importante na análise de dados de contagem. Ele possui algumas características como: proporcionar uma descrição dos dados experimentais cuja variância é proporcional à média, ser deduzido teoricamente de princípios elementares com um número mínimo de restrições e, determinar o número de eventos em um intervalo, dado que ocorreram independentemente e aleatoriamente (CORDEIRO; DEMÉTRIO, 2008).

Considere $Y_1, \ldots, Y_n$ variáveis respostas condicionalmente independentes e um vetor de covariáveis $\boldsymbol{x}_i^\top$. Assumindo que a distribuição condicional $Y_i|\boldsymbol{x}_i^\top$ segue uma $\text{Pois}(\mu_i)$, a função de probabilidade de $Y_i|\boldsymbol{x}_i^\top$, é dada na forma

$$f(y_i|\boldsymbol{x}_i^\top) = \frac{e^{-\mu_i}(\mu_i)^{y_i}}{y_i!} \quad \text{com} \;\; \mu_i > 0, y_i = 0, 1, \ldots \tag{1}$$

Como propriedade, temos que (1) pertence à chamada Família Exponencial, então, utilizando uma função de ligação canônica sendo a logarítmica para a média, a relação do preditor linear com μ_i é definido por

$$\eta_i = \log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \,,$$

sendo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ o vetor de parâmetros, de dimensão $p \times 1$. Dessa forma, temos como média e variância do modelo, respectivamente

$$\text{E}(Y_i|\boldsymbol{x}_i^\top) = \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \qquad e \qquad \text{Var}(Y_i|\boldsymbol{x}_i^\top) = \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \,. \tag{2}$$

Distribuição Poisson-Gama

Na literatura também é conhecida como distribuição de Pascal, quando r é um natural positivo, ou Binomial Negativa, quando r é um real positivo, esta distribuição apresenta formas especiais que foram surgidas por Pascal e Fermat (DEMÉTRIO; CORDEIRO, 2008). Em 1907, Gosset usou a distribuição Binomial Negativa para modelar dados de contagem no lugar da distribuição de Poisson, visto que ela apresenta uma maior flexibilidade por não possuir a propriedade de equidispersão.

Esta distribuição também foi obtida por Greenwood e Yule (1920) como consequência da suposição de modelos a propensão de acidentes. Os autores consideraram que o número de acidentes seguia uma

distribuição de Poisson, com parâmetro λ, em que λ variava de acordo com uma distribuição Gama, de parâmetros α e β. Assim, considerando Y uma variável aleatória com distribuição Gama, com parâmetros α e β, é possível expressar a função de probabilidade de $Y \sim \text{Gama}(\alpha, \beta)$ como

$$g(y;\alpha,\beta) = \frac{\left(\frac{\beta}{\alpha}\right)^{\beta}}{\Gamma(\beta)} y^{\beta-1} \exp\left\{-\frac{y\beta}{\alpha}\right\}, \;\; y > 0 \tag{3}$$

em que $\alpha, \beta > 0$. A esperança e variância são dadas, respectivamente, por

$$\text{E}(Y) = \alpha \qquad e \qquad \text{Var}(Y) = \frac{\alpha^2}{\beta}. \tag{4}$$

Assumindo que a variável aleatória condicional de Y dado o parâmetro λ segue uma distribuição Poisson, $Y|\lambda \sim \text{Pois}(\lambda)$, e que o parâmetro em si, λ, segue uma distribuição gama, $\lambda \sim \text{Gama}(\alpha, \beta)$, a função de probabilidade marginal de Y pode ser obtida utilizando a condicional e a distribuição do parâmetro, ou seja

$$f(y) = \int_{\lambda} f(Y|\lambda) g(\lambda) d\lambda.$$

Substituindo-se pelas respectivas funções de probabilidades, obtém-se a marginal

$$\begin{aligned} f(y) &= \int_0^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} \frac{\left(\frac{\beta}{\alpha}\right)^{\beta}}{\Gamma(\beta)} \lambda^{\beta-1} \exp\left\{-\frac{\beta\lambda}{\alpha}\right\} d\lambda = \frac{1}{y!} \frac{\left(\frac{\beta}{\alpha}\right)^{\beta}}{\Gamma(\beta)} \int_0^{\infty} e^{-\lambda}\lambda^y \lambda^{\beta-1} \exp\left\{-\frac{\beta\lambda}{\alpha}\right\} d\lambda \\ &= \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^{\beta} \int_0^{\infty} e^{-\lambda} \lambda^{y+\beta-1} e^{-\left(\frac{\beta\lambda}{\alpha}\right)} d\lambda = \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^{\beta} \int_0^{\infty} \lambda^{y+\beta-1} e^{-\lambda\left(\frac{\beta}{\alpha}+1\right)} d\lambda, \end{aligned}$$

e dado que $\Gamma(\gamma) = \int_0^{\infty} t^{\gamma-1} e^{-t} \text{dt}$, então tem-se que

$$f(y) = \frac{1}{y!\Gamma(\beta)} \left(\frac{\beta}{\alpha}\right)^{\beta} \frac{\Gamma(y+\beta)}{\left(\frac{\beta}{\alpha}+1\right)^{y+\beta}}.$$

Portanto

$$f(y;\alpha,\beta) = \frac{\Gamma(y+\beta)}{\Gamma(y+1)\Gamma(\beta)} \left(\frac{\beta}{\alpha+\beta}\right)^{\beta} \left(\frac{\alpha}{\alpha+\beta}\right)^{y}, \; y = 0, 1, \ldots \text{ e } \alpha, \beta > 0. \tag{5}$$

Assim, Y segue uma distribuição Poisson-Gama, $Y \sim \text{PG}(\alpha, \beta)$, ou binomial negativa. Dado que a distribuição PG é uma distribuição composta, é possível encontrar o valor esperado de Y a partir da propriedade da esperança condicional, levando em consideração que $\lambda \sim \text{Gama}(\alpha, \beta)$. Assim, respectivamente, temos média e variância dadas por

$$\begin{aligned} \text{E}(Y) &= \text{E}[\text{E}(Y|\lambda)] = \text{E}(\lambda) = \alpha \\ \text{Var}(Y) &= \text{E}[\text{Var}(Y|\lambda)] + \text{Var}[\text{E}(Y|\lambda)] = \text{E}(\lambda) + \text{Var}(\lambda) = \alpha + \frac{\alpha^2}{\beta}. \end{aligned}$$

Observa-se que a variância de Y, escrita em função do valor esperado, cresce mais rapidamente com relação à média do que para a distribuição Poisson. Isso permite uma maior flexibilidade entre média e variância.

Modelo Poisson-Gama

Sejam $Y_1, \ldots, Y_n$ variáveis condicionalmente independentes, dado um vetor de variáveis explicativas $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Considerando que a distribuição condicional $Y_i|\boldsymbol{x}_i^\top$ segue uma $\mathrm{PG}(\mu_i, \phi)$, a função de probabilidade de $Y_i|\boldsymbol{x}_i^\top$, conforme (5), é definida por

$$f(y_i|\boldsymbol{x}_i^\top) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}, \tag{6}$$

com $y_i = 0, 1, \ldots$ e $\mu_i, \phi > 0$. A função de probabilidade definida (6) pertence a família exponencial uniparamétrica, sendo um caso particular dos modelos lineares generalizados.

A esperança e variância do modelo podem ser expressos, utilizando-se uma função de ligação logarítmica, para uma comparação direta com o modelo Poisson, respectivamente por:

$$\mathrm{E}(Y_i|x_i^T) = \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \quad \text{e} \quad \mathrm{Var}(Y|x_i^T) = \frac{\mu_i^2}{\phi} + \mu_i = \mu_i\left(\frac{\mu_i}{\phi}+1\right). \tag{7}$$

Distribuição Poisson-Inversa Gaussiana

A distribuição Poisson-Inversa Gaussiana (PIG) foi proposta (HOLLA, 1967) como uma alternativa para a distribuição de Poisson para casos com superdispersão. Em alguns estudos de seguro e medicamento, a distribuição PIG foi proposta como uma boa alternativa pala modelar dados com superdispersão ou de cauda longa do que a Poisson-Gama (PUTRI et al., 2020). Holla (1965) realizou um estudo acerca da distribuição PIG tanto para o caso univariado como para o caso multivariado.

Seja X uma variável aleatória com distribuição Inversa Gaussiana, ou Normal Inversa, de parâmetros μ e σ. Podemos expressar a função de probabilidade de $X \sim \mathrm{IG}(\mu, \sigma)$ como

$$g(x; \mu, \sigma) = \sqrt{\frac{\sigma}{2\pi x^3}}\exp\left\{-\frac{\sigma(x-\mu)^2}{2\mu^2 x}\right\}, \tag{8}$$

em que $x > 0$ e $\mu > 0$, sendo a média. A esperança e variância da PIG são dadas, respectivamente

$$\mathrm{E}(X) = \mu \qquad e \qquad \mathrm{Var}(X) = \frac{\mu^3}{\sigma}. \tag{9}$$

Assumindo que uma variável aleatória $Y|\lambda \sim \mathrm{Pois}(\lambda)$ e $\lambda \sim \mathrm{IG}(\mu, \sigma)$, expressa em (8), então, marginalmente, a função de probabilidade marginal de Y, é dada por

$$\begin{aligned} f(y) &= \int_0^\infty p(y;\lambda)g(\lambda;\mu,\sigma)\mathrm{d}\lambda = \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\left(\frac{\sigma}{2\pi\lambda^3}\right)^{1/2}\exp\left\{\frac{\sigma(\lambda-\mu)^2}{2\mu^2\lambda}\right\}\mathrm{d}\lambda \\ &= \left(\frac{2\sigma}{\pi}\right)^{\frac{1}{2}}\frac{1}{y!}\,e^{\sigma/\mu}\left[\frac{\sigma}{2\left(1+\frac{\sigma}{2\mu^2}\right)}\right]^{\frac{1}{2}\left(y-\frac{1}{2}\right)} K_{y-1/2}\left[\sqrt{2\sigma\left(1+\frac{\sigma}{2\mu^2}\right)}\right], \end{aligned} \tag{10}$$

com $y = 0, 1, \ldots$ e $\mu, \sigma > 0$. Além disso, K_α representa a função de Bessel, obtida segundo a solução de uma equação diferencial de segunda ordem (HOLLA, 1967). Portanto, temos que (10) representa a função de probabilidade de de $Y \sim PIG(\mu, \sigma)$.

Dado que a distribuição PIG é uma distribuição composta, o valor esperado de Y também pode ser ser expresso partindo da propriedade da esperança condicional. Da mesma forma que temos, respectivamente:

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|\lambda)] = \mathrm{E}(\lambda) = \mu$$

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|\lambda)] + \mathrm{Var}[\mathrm{E}(Y|\lambda)] = \mathrm{E}(\lambda) + \mathrm{Var}(\lambda) = \mu\left(\frac{\mu^2}{\sigma}+1\right).$$

Modelo Poisson-Inversa Gaussiana

Considere $Y_1, \ldots, Y_n$ variáveis condicionalmente independentes, dado um vetor de variáveis explicativas $\boldsymbol{x}_i^\top$. Assumindo que a distribuição condicional $Y_i|\boldsymbol{x}_i^\top$ segue uma PIG(μ_i, σ), a função de probabilidade, conforme (10), é definida por

$$f(y_i|\boldsymbol{x}_i^\top) = \left(\frac{2\sigma}{\pi}\right)^{\frac{1}{2}} \frac{1}{y_i!}\, e^{\sigma/\mu_i} \left[\frac{\sigma}{2\left(1+\frac{\sigma}{2\mu_i^2}\right)}\right]^{\frac{1}{2}\left(y_i-\frac{1}{2}\right)} K_{y_i-1/2}\left[\sqrt{2\sigma\left(1+\frac{\sigma}{2\mu_i^2}\right)}\right] \tag{11}$$

em que $y_i = 0, 1, \ldots$ e $\mu_i, \sigma > 0$.

A esperança e variância da distribuição incondicional da resposta, associada ao modelo, pode ser expressa, utilizando uma função de ligação logarítmica, respectivamente

$$\mathrm{E}(Y_i|\boldsymbol{x}_i^\top) = \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \quad \text{e} \quad \mathrm{Var}(Y_i|\boldsymbol{x}_i^\top) = \frac{\mu_i^3}{\sigma} + \mu_i = \mu_i\left(\frac{\mu_i^2}{\sigma} + 1\right).$$

Distribuição Poisson-Lindley Generalizada

A distribuição Lindley foi introduzida por Lindley (1958), para estudos realizados sobre vida útil. Segundo Ghitany *et al.* (2011), esta distribuição pode ser especialmente útil para modelagens em estudos de mortalidade. Assim como o caso da distribuição PG, que pode ser expressa por meio de uma mistura entre as distribuições Poisson e Gama, alguns autores sugeriram ramificações dessa natureza em relação a distribuição Lindley. Podemos citar duas, a Poisson-Lindley introduzida por Sankaran (1970) e a Poisson-Lindley Generalizada apresentada por Wongrin e Bodhisuwan (2016). Esta última sendo uma mistura entre a distribuição Poisson e a distribuição Lindley Generalizada apresentada por Elbatal *et al.* (2013).

A distribuição Lindley Generalizada é uma distribuição para dados de natureza contínua, especificada por três parâmetros (α, β e θ) e, dado sua maior flexibilidade, comparada à Lindley, ela é bastante usada como alternativa para dados de tempo de vida.

Distribuição Poisson-Lindley Generalizada com dois parâmetros

A distribuição Poisson-Lindley Generalizada como citado, possui três parâmetros. Entretanto, a Poisson-Lindley Generalizada pode ser reduzida a dois parâmetros, o que facilita a parte computacional e prática para a aplicação desta distribuição.

Primeiramente, consideramos a distribuição Lindley Generalizada (LG) apresentada por Mahmoudi e Zekerzadeh (2010), onde eles focaram na redução de parâmetros da distribuição Lindley generalizada expressa por Zakerzadeh e Dolati (2009). Considerando os parâmetros α e θ, positivos, e $\beta = 1$, obtemos a expressão da função de probabilidade da distribuição Lindley com dois parâmetros como sendo

$$g(y; \alpha, \beta = 1, \theta) = \frac{1}{\theta+1}\frac{\theta^{\alpha+1} y^{\alpha-1}(\alpha+y)}{\Gamma(\alpha+1)} e^{-\theta y}, \tag{12}$$

em que $y > 0$ e $\alpha, \theta > 0$, denotada como $Y \sim \mathrm{LG}(Y; \alpha, \beta = 1, \theta)$. O valor esperado e a variância são denotados, respectivamente, por:

$$\mathrm{E}(Y) = \frac{\alpha(\theta+1)+1}{\theta(\theta+1)} \quad \text{e} \quad \mathrm{Var}(Y) = \frac{(\alpha+1)[\alpha(\theta+1)+2]}{\theta^2(\theta+1)} - \left[\frac{\alpha(\theta+1)+1}{\theta(\theta+1)}\right]^2. \tag{13}$$

A função de probabilidade marginal de Y, pode ser expressa por meio da integral da função de probabilidade da Poisson juntamente com a LG (12), onde $Y|\lambda \sim Pois(\lambda)$ e $\lambda \sim LG(\alpha, \beta = 1, \theta)$, sendo especificada por:

$$
\begin{aligned}
f(y;\alpha,\beta=1,\theta) &= \int_0^\infty p(y;\lambda)g(y;\alpha,\beta=1,\theta)\mathrm{d}\lambda = \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\frac{1}{\theta+1}\frac{\theta^{\alpha+1}\lambda^{\alpha-1}(\alpha+\lambda)}{\Gamma(\alpha+1)}e^{-\theta\lambda}\mathrm{d}\lambda \\
&= \frac{\theta^{\alpha+1}}{y!(\theta+1)\Gamma(\alpha+1)}\int_0^\infty(\alpha+\lambda)e^{-\lambda(\theta+1)}\lambda^{y+\alpha-1}\mathrm{d}\lambda \\
&= \frac{\theta^{\alpha+1}}{y!(\theta+1)\Gamma(\alpha+1)}\left[\frac{\alpha\Gamma(y+\alpha)}{(\theta+1)^{y+\alpha}}+\frac{\Gamma(y+\alpha+1)}{(\theta+1)^{y+\alpha+1}}\right].
\end{aligned}
$$

Portanto, temos que a marginal de Y segue uma distribuição Poisson-Lindley Generalizada com parâmetros α e θ, ou seja, $Y \sim PLG(\alpha,\beta=1,\theta)$, com função de probabilidade

$$
f(y;\alpha,\theta) = \frac{\theta^{\alpha+1}\Gamma(y+\alpha)[\alpha(\theta+1)+(y+\alpha)]}{y!\Gamma(\alpha+1)(\theta+1)^{y+\alpha+2}}\,, com\ y=0,1,\ldots \text{ e } \alpha,\theta>0. \tag{14}
$$

Propriedades

Propriedade 0.1 *Suponha que $Y \sim PLG(\alpha,\beta=1,\theta)$, o valor esperado e a variância podem ser expressos como, utilizando as propriedades de esperança e variância condicionais*

$$
\begin{aligned}
\mathrm{E}(Y) &= \mathrm{E}[\mathrm{E}(Y|\lambda)] = \mathrm{E}(\lambda) = \frac{\alpha(\theta+1)+1}{\theta(\theta+1)}\,. \\
\mathrm{Var}(Y) &= \mathrm{Var}[\mathrm{E}(Y|\lambda)] + \mathrm{E}[\mathrm{Var}(Y|\lambda)] = \mathrm{Var}(\lambda)+\mathrm{E}(\lambda) \\
&= \frac{(\alpha+1)[\alpha(\theta+1)+2]}{\theta^2(\theta+1)} - \left[\frac{\alpha(\theta+1)+1}{\theta(\theta+1)}\right]^2 + \frac{\alpha(\theta+1)+1}{\theta(\theta+1)}\,.
\end{aligned}
$$

Propriedade 0.2 *Considere que $Y \sim PLG(\alpha,\beta=1,\theta)$, a função log-verossimilhança a ser maximizada, é dada por*

$$
\begin{aligned}
l(y;\alpha,\theta) &= \sum_{i=1}^n\left\{\log\left[\frac{\theta^{\alpha+1}\Gamma(y_i+\alpha)[\alpha(\theta+1)+(y_i+\alpha)]}{y_i!\Gamma(\alpha+1)(\theta+1)^{y_i+\alpha+2}}\right]\right\} \\
&= \sum_{i=1}^n\left\{\log\left[\theta^{\alpha+1}\Gamma(y_i+\alpha)\big[\alpha(\theta+1)+(y_i+\alpha)\big]\right] - \log\left[y_i!\Gamma(\alpha+1)(\theta+1)^{y_i+\alpha+2}\right]\right\} \\
&\quad + \log\big[\alpha(\theta+1)+(y_i+\alpha)\big]\Big\}.
\end{aligned} \tag{15}
$$

Propriedade 0.3 *A função densidade acumulada de $Y \sim PLG(\alpha,\beta=1,\theta)$ pode ser escrita em função de uma mistura de duas binomiais negativas, denotado como:*

$$
F(y;\alpha,\beta=1,\theta) = \frac{\theta}{1+\theta}\,G_1(y) + \frac{1}{1+\theta}\,G_2(y)\,, \tag{16}
$$

em que $G_1(y)$ e $G_2(y)$ são função acumuladas de uma binomial negativa de parâmetros $r=\alpha$ e $r=\alpha+1$, respectivamente, e $p=\frac{\theta}{1+\theta}$.

Seja a função de probabilidade dada conforme (14), podemos expressá-la como:

$$
\frac{\theta^{\alpha+1}\alpha\Gamma(y+\alpha)}{y!(\theta+1)^{y+\alpha+1}\Gamma(\alpha+1)} + \frac{\theta^{\alpha+1}(\alpha+y)\Gamma(y+\alpha)}{y!(\theta+1)^{y+\alpha+2}\Gamma(\alpha+1)}\,.
$$

Considere uma binomial negativa com $r>0$ e $0<p<1$, ou seja, $BN(r,p)$. A sua função de probabilidade por ser expressa como:

$$P(y) = \frac{\Gamma(y+r)}{y!\Gamma(r)} p^r (1-p)^y \, .$$

Se admitirmos que $r = \alpha$ e $p = \frac{\theta}{1+\theta}$, obtém-se:

$$g_1(y) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{\theta}{1+\theta}\right)^{\alpha} \left(\frac{1}{1+\theta}\right)^{y} = \frac{\alpha\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)} \frac{\theta^{\alpha}}{(1+\theta)^{\alpha+y}} \, . \tag{17}$$

Admitindo que $r = \alpha + 1$ e $p = \frac{\theta}{1+\theta}$, obtém-se:

$$g_2(y) = \frac{\Gamma(y+\alpha+1)}{y!\Gamma(\alpha+1)} \left(\frac{\theta}{1+\theta}\right)^{\alpha+1} \left(\frac{1}{1+\theta}\right)^{y} = \frac{(\alpha+y)\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)} \frac{\theta^{\alpha+1}}{(1+\theta)^{\alpha+1+y}} \, . \tag{18}$$

Assim, utilizando as parametrizações (17) e (18) e, multiplicando cada uma por $\frac{\theta}{1+\theta}$ e $\frac{1}{1+\theta}$, respectivamente, encontramos a função de probabilidade da PLG:

$$\begin{aligned} p(y;\alpha,\beta=1,\theta) &= \frac{\theta}{1+\theta}\, g_1(y) + \frac{1}{1+\theta}\, g_2(y) \\ &= \frac{\theta}{(1+\theta)} \frac{\alpha\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)} \frac{\theta^{\alpha}}{(1+\theta)^{\alpha+y}} + \frac{1}{(1+\theta)} \frac{(\alpha+y)\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)} \frac{\theta^{\alpha+1}}{(1+\theta)^{\alpha+1+y}} \end{aligned}$$

$$p(y;\alpha,\beta=1,\theta) = \frac{\theta^{\alpha+1}\alpha\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)(1+\theta)^{\alpha+y+1}} + \frac{\theta^{\alpha+1}(\alpha+y)\Gamma(y+\alpha)}{y!\Gamma(\alpha+1)(1+\theta)^{\alpha+y+2}} \, , \; y = 0, 1, \ldots$$

Essa é uma propriedade importante, visto que na utilização de resíduos quantílicos tem-se a necessidade da geração de números aleatórios. E pode-se reescrever a PLG com base em uma distribuição conhecida, simplificando o trabalho computacionalmente.

Modelo Poisson-Lindley Generalizada com dois parâmetros

Seja $Y_1, \ldots, Y_n$ as variáveis respostas dado um conjunto de covariáveis $\boldsymbol{x}_i^{\top}$. Assumindo que a distribuição condicional de Y_i dado $\boldsymbol{x}_i^{\top}$ segue uma distribuição $\mathrm{PLG}(\alpha, \beta = 1, \theta)$ com parâmetros α e $\theta > 0$, temos como valor esperado do modelo e os componentes da variância, dados por, respectivamente

$$\mathrm{E}(Y_i|\boldsymbol{x}_i^{\top}) = \mathrm{E}\left\{\mathrm{E}\left[(Y_i|\boldsymbol{x}_i^{\top})|\lambda_i\right]\right\} = \frac{\alpha_i(\theta+1)+1}{\theta(\theta+1)} = \mu_i \, . \tag{19}$$

$$\begin{aligned} \mathrm{Var}(Y_i|\boldsymbol{x}_i^{\top}) &= \mathrm{E}\left\{\mathrm{Var}\left[(Y_i|\boldsymbol{x}_i^{\top})|\lambda_i\right]\right\} + \mathrm{Var}\left\{\mathrm{E}\left[(Y_i|\boldsymbol{x}_i^{\top})|\lambda_i\right]\right\} \\ &= \mu_i + \frac{(\alpha_i+1)[\alpha_i(\theta+1)+2]}{\theta^2(\theta+1)} - \mu_i^2 \, . \end{aligned} \tag{20}$$

Assim, como os modelos anteriormente citados, no modelo PLG a função de ligação logarítmica é adotada para fazer a ligação entre a média da variável resposta e as covariáveis. Reparametrizando a média da distribuição PLG com $\mu_i = e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}$ temos:

$$\frac{\alpha_i(\theta+1)+1}{\theta(\theta+1)} = \mu_i \Rightarrow \; \alpha_i(\theta+1)+1 = \mu_i\theta(\theta+1) \Rightarrow \alpha_i = \frac{\mu_i\theta(\theta+1)-1}{\theta+1} \, ,$$

portanto,

$$\alpha_i = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}\,. \tag{21}$$

Portanto, a função de probabilidade de $Y_i|\boldsymbol{x}_i^\top \sim PLG(\alpha,\theta)$ pode ser expressa na forma de um modelo linear por substituir α_i na função de probabilidade

$$f(Y_i|\boldsymbol{x}_i^\top) = \frac{\Gamma\left(y_i + \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}\right)}{y_i!\Gamma\left(\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}+1\right)} \times \frac{\theta^{\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}+1}}{(\theta+1)^{y_i+\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}+2}} \times \left[e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1 + y_i + \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\theta(\theta+1)-1}{\theta+1}\right]. \tag{22}$$

Resíduos

Análise de resíduos

A análise de diagnóstico é uma etapa fundamental no ajuste de modelos de regressão, pois ela verifica possíveis afastamentos das suposições feitas para o modelo, bem como a existência de observações influentes que causam alguma interferência nos resultados do ajuste. Boa parte dos métodos de análise de resíduos são semelhantes aos procedimentos usados para o modelo clássico de regressão, com algumas adaptações. Entretanto, deve-se usar com cautela, visto que alguns resultados dependem fortemente das propriedades do modelo proposto.

Alguns trabalhos citados na literatura apresentam extensões da análise de resíduos para os modelos lineares generalizados. Podemos citar, por exemplo, Cox e Snell (1968) e Fahrmeir e Tutz (1994). Para mais detalhes, veja (PAULA, 2013).

Tipos de resíduos

Um resíduo, geralmente denotador por r_i, é uma medida que expressa a distância entre uma observação (y_i) e o seu valor ajustado ($\hat{\mu}_i$), isto é, $r_i = d_i(y_i, \hat{\mu}_i)$, sendo d_i uma função adequada de fácil interpretação usualmente escolhida para estabilizar a variância ou induzir simetria na distribuição amostral de r_i. Garantindo comparabilidade dos resíduos e detecção de resíduos discrepantes (COX; SNELL, 1968).

A seguir, apresentam-se alguns tipos de resíduos mais comumente utilizados ao se tratar dos MLGs, descritos em Cordeiro e Demétrio (2008), McCullagh e Nelder (1989) e Paula (2013).

a) **Resíduo de Pearson padronizado**: O resíduo de Pearson, componentes da estatística χ^2 de Pearson generalizada, têm uma versão padronizada, sendo definido como

$$r_i^{P^*} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1-\hat{h}_{ii})}}, \tag{23}$$

em que h_{ii} é o i-ésimo elemento da diagonal da matriz de projeção $\mathbf{H}$

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}\mathbf{W}^{\frac{1}{2}}, \tag{24}$$

sendo $\mathbf{W}$ a matriz de pesos. A matriz $\mathbf{H}$ mede a influência em unidades estudentizadas de $\mathbf{y}$ sobre $\hat{\boldsymbol{\mu}}$, tendo como propriedades que $\mathrm{tr}(\mathbf{H}) = p$ e $0 \leq h_{ii} \leq 1$. Assim, além de consideramos a função de variância do modelo, como no caso dos resíduos de Pearson, agora levamos em conta a medida de *leverage* estimada, $\hat{h}_{ii}$.

b) **Resíduo componente da *deviance* padronizado**: O resíduo componente da *deviance* é definido pela contribuição de cada observação para a *deviance*, ou desvio, do modelo, sendo uma medida de distância de y_i em relação a $\hat{\mu}_i$ na escala do logaritmo da verossimilhança (log-verossimilhança). Considerando um modelo saturado, ou seja, um modelo onde o número de parâmetros de regressão é igual ao número de observações, $p = n$, tem-se que este modelo atribui toda a variação dos dados ao componente sistemático, ajustando-se perfeitamente e reproduzindo os próprios dados, entretanto, tal modelo é de difícil interpretação. A medida desvio é expressa como sendo a diferença entre os máximos da função de log-verossimilhança do modelo saturado e do modelo sob pesquisa, isto é,

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\left[l(\mathbf{y};\mathbf{y}) - l(\hat{\boldsymbol{\mu}};\mathbf{y})\right] = \sum_{i=1}^{n} d_i. \tag{25}$$

Portanto, o resíduo componente da *deviance* fica definido por: $r_i^D = sinal(y_i - \hat{\mu}_i)\sqrt{d_i}$, em que $sinal(\alpha) = -1$, se $\alpha < 0$ e $sinal(\alpha) = 1$, se $\alpha > 0$. Um valor grande para r_i^D indica que a i-ésima observação é mal ajustada pelo modelo. Sua forma padronizada, levando em conta a medida de *leverage*, é definida como

$$r_i^{D*} = \frac{r_i^D}{\sqrt{(1-\hat{h}_{ii})}}. \tag{26}$$

c) **Resíduo quantílico aleatorizado**: Sugerido por Dunn e Smyth (1996), os resíduos quantí-licos aleatorizados se baseiam no teorema da inversa da função de distribuição acumulada e tem como proposta apresentar uma distribuição Normal, independentemente da distribuição da variável resposta. O resíduo quantílico é definido por

$$r_i^q = \begin{cases} \Phi^{-1}[\mathrm{F}(y_i;\hat{\mu}_i,\phi)], \text{ se F é contínua} \\ \Phi^{-1}[\mathrm{F}(y_i^-;\hat{\mu}_i,\phi) - u_i \mathrm{p}(y_i;\hat{\mu}_i,\phi)], \text{ se F é discreta} \end{cases} \tag{27}$$

sendo $\mathrm{F}(y_i;\hat{\mu}_i,\phi)$ a função de distribuição acumulada de Y_i, $\mathrm{p}(y_i;\hat{\mu}_i,\phi)$ a função de probabilidade, $\mathrm{F}(y_i^-;\hat{\mu}_i,\phi) = \lim_{y\to y_i^-} \mathrm{F}(y;\hat{\mu}_i,\phi)$, u_i uma variável aleatória uniforme no intervalo $(0,1]$ e $\Phi(\cdot)$ a função de distribuição acumulada de uma Normal padrão.

A seguir são apresentados de forma algébrica, dois destes adaptados para a forma dos modelos apresentados. Levando em consideração que a distribuição Poisson e Poisson-Gama pertencem à família exponencial de distribuição, com parâmetro de dispersão $a(\phi) = 1$, então vamos utilizar esta mesma propriedade para os demais dois modelos. Além disso, vamos utilizar a função de variância, $V(\hat{\mu}_i)$ escrita apenas na forma da própria variância do modelo, $\mathrm{Var}(\hat{y}_i)$.

I. **Resíduos de Pearson padronizado**: Para a composição dos resíduos de Pearson padronizado, é necessário utilizarmos a matriz de pesos $\mathbf{W}$. Considerando, por exemplo, uma função de ligação logarítmica, $\eta_i = \log(\mu_i)$, o i-ésimo elemento da diagonal de $\mathbf{W}$ (matriz de pesos), é dado por

– Poisson:

$$w_i = \left[\left(\frac{\partial \log(\mu_i)}{\partial \mu_i}\right)^2 V(y_i)\right]^{-1} = \left[\left(\frac{1}{\mu_i}\right)^2 \mu_i\right]^{-1} = \left[\frac{1}{\mu_i^2}\,\mu_i\right]^{-1} = \mu_i\,.$$

– Poisson-Gama:

$$w_i = \left[\left(\frac{\partial \log(\mu_i)}{\partial \mu_i}\right)^2 V(y_i)\right]^{-1} = \left[\left(\frac{1}{\mu_i}\right)^2 \mu_i \left(\frac{\mu_i}{\phi}+1\right)\right]^{-1} = \frac{\phi\mu_i}{\mu_i+\phi}.$$

– Poisson-Inversa Gaussiana:

$$w_i = \left[\left(\frac{\partial \log(\mu_i)}{\partial \mu_i}\right)^2 V(y_i)\right]^{-1} = \left[\left(\frac{1}{\mu_i}\right)^2 \mu_i \left(\frac{\mu_i^2}{\sigma}+1\right)\right]^{-1} = \frac{\mu_i \sigma}{\mu_i^2+\sigma}.$$

– Poisson-Lindley Generalizada:

$$\begin{aligned} w_i &= \left[\left(\frac{\partial \log(\mu_i)}{\partial \mu_i}\right)^2 V(y_i)\right]^{-1} = \left[\left(\frac{1}{\mu_i}\right)^2 \left(\mu_i + \frac{(\alpha_i+1)[\alpha_i(\theta+1)+2]}{\theta^2(\theta+1)} - \mu_i^2\right)\right]^{-1} \\ &= \left[\frac{1}{\mu_i} + \frac{(\alpha_i+1)[\alpha_i(\theta+1)+2]}{\mu_i^2\theta^2(\theta+1)} - 1\right]^{-1} \\ &= \frac{\mu_i^2\theta^2(\theta+1)}{\mu_i\theta^2(\theta+1)(1-\mu_i)+(\alpha+1)[\alpha_i(\theta+1)+2]}\ . \end{aligned}$$

Assim, utilizando, o resíduo padronizado é descrito para cada modelo, é dado como

– Poisson: $r_i^P = \frac{y_i-\hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{h}_{ii})}}$;

– Poisson-Gama: $r_i^P = \frac{y_i-\hat{\mu}_i}{\sqrt{\left(\frac{\hat{\mu}_i^2}{\phi}+\hat{\mu}_i\right)(1-\hat{h}_{ii})}}$;

– Poisson-Inversa Gaussiana: $r_i^P = \frac{y_i-\hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-\hat{h}_{ii})}}$;

– Poisson-Lindley Generalizada: $r_i^P = \frac{y_i-\hat{\mu}_i}{\sqrt{\left(\hat{\mu}_i+\frac{(\hat{\alpha}_i+1)[\hat{\alpha}_i(\theta+1)+2]}{\theta^2(\theta+1)}-\hat{\mu}_i^2\right)(1-\hat{h}_{ii})}}$.

em que $\hat{h}_{ii}$ é o i-ésimo elemento da diagonal da matriz de projeção $\widehat{\mathbf{H}}$, de cada distribuição, com a matriz de pesos sendo $\widehat{\mathbf{W}} = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_n)$.

II. **Resíduo componente da *deviance* padronizado**: Utilizando a função de probabilidade definida de cada modelo e considerando a função logarítmica da verossimilhança temos

– Poisson:

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{n} \log(f(y_i|\boldsymbol{x}_i^\top)) = -\sum_{i=1}^{n}\mu_i + y_i\sum_{i=1}^{n}\log(\mu_i) - \sum_{i=1}^{n}\log y_i!,$$

e, pela definição em (25) para os desvios, obtemos

$$\begin{aligned} d_i &= 2\big[-y_i + y_i\log(y_i) - \log y_i! - \big(-\hat{\mu}_i + y_i\log(\hat{\mu}_i) - \log y_i!\big)\big] \\ &= 2\big[-y_i + y_i\log(y_i) + \hat{\mu}_i - y_i\log(\hat{\mu}_i)\big] = 2\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i-\hat{\mu}_i)\right],\ \text{para } y_i > 0. \end{aligned}$$

Caso $y_i = 0$, a substituição no logaritmo seria indefinido,assim, obtemos o desvio como

$$\begin{aligned} f(y_i = 0|\boldsymbol{x}_i^\top) &= \frac{e^{-\mu_i}(\mu_i)^0}{0!} = \frac{e^{-\mu_i}\times 1}{1} = \exp(-\mu_i)\ , \\ \Rightarrow d_i &= 2\big[l(y_i;y_i) - l(\hat{\mu}_i;y_i)\big] = 2\big[\log\big(\exp\{-y_i\}\big) - \log\big(\exp\{-\hat{\mu}_i\}\big)\big] = 2\hat{\mu}_i \end{aligned}$$

Portanto, temos como resíduo componente da *deviance* para a Poisson

$$r_i^D = \begin{cases} sinal(y_i - \hat{\mu}_i)\sqrt{2}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]^{\frac{1}{2}} & \text{se } y_i > 0 \\ sinal(y_i - \hat{\mu}_i)\sqrt{2}\left[\hat{\mu}_i\right]^{\frac{1}{2}} & \text{se } y_i = 0 \end{cases}$$

– Poisson-Gama:

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{n}\left[\log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right) + \phi\log\phi + y_i\log\mu_i - (\phi+y_i)\log(\mu_i+\phi)\right]$$

Utilizando a definição para os desvios, obtemos

$$\begin{aligned} d_i &= 2\left[\log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right) + \phi\log(\phi) + y_i\log y_i - (\phi+y_i)\log(y_i+\phi)\right] \\ &\quad - \left[\log\left(\frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\right) + \phi\log(\phi) + y_i\log\hat{\mu}_i - (\theta+y_i)\log(\hat{\mu}_i+\phi)\right] \\ &= 2\left[\phi\log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right) + y_i\log\left(\frac{y_i}{\hat{\mu}_i}\frac{(\hat{\mu}_i+\phi)}{(y_i+\phi)}\right)\right], \text{ para } y_i > 0. \end{aligned}$$

Caso $y_i = 0$ se obtém o desvio como sendo

$$\begin{aligned} f(y_i = 0|\boldsymbol{x}_i^\top) &= \sum_{i=1}^{n}\left[\phi\log\phi - \phi\log(\mu_i+\phi)\right] \\ \Rightarrow d_i &= 2\big[l(y_i;y_i) - l(\hat{\mu}_i;y_i)\big] = 2\big[-\phi\log(\phi) + \phi\log(\hat{\mu}_i+\phi)\big] = 2\phi\log\left(\frac{\hat{\mu}_i+\phi}{\phi}\right). \end{aligned}$$

Portanto, o resíduo componente da *deviance* para a Poisson-Gama é dado por:

$$r_i^D = \begin{cases} sinal(y_i - \hat{\mu}_i)\sqrt{2}\left[\phi\log\left(\frac{\hat{\mu}_i+\phi}{y_i+\phi}\right) + y_i\log\left(\frac{y_i}{\hat{\mu}_i}\frac{(\hat{\mu}_i+\phi)}{(y_i+\phi)}\right)\right]^{\frac{1}{2}} & \text{, se } y_i > 0 \\ sinal(y_i - \hat{\mu}_i)\sqrt{2}\left[\phi\log\left(\frac{\hat{\mu}_i+\phi}{\phi}\right)\right]^{\frac{1}{2}} & \text{, se } y_i = 0 \end{cases}$$

– Poisson-Inversa Gaussiana:

$$\begin{aligned} l(\boldsymbol{\mu},\sigma) = \sum_{i=1}^{n}\Bigg\{ &\frac{1}{2}\log\left(\frac{2\sigma}{\pi}\right) - \log y_i! + \frac{\sigma}{\mu_i} + \frac{1}{2}\left(y_i - \frac{1}{2}\right)\log\sigma \\ &-\frac{1}{2}\left(y_i - \frac{1}{2}\right)\log\left[2\left(1+\frac{\sigma}{2\mu_i^2}\right)\right] + \log \mathrm{K}_{y_i-1/2}\left[\sqrt{2\sigma\left(1+\frac{\sigma}{2\mu_i^2}\right)}\right]\Bigg\}. \end{aligned}$$

O desvio será dado de forma a considerar a log-verossimilhança do modelo estimado ($\hat{\mu}_i = \hat{\mu}_i$) e do modelo saturado ($\tilde{\mu}_i = y_i$), definido como:

$$\begin{aligned} d_i &= 2[l(y_i,y_i) - l(\hat{\mu}_i,y_i)] \\ &= 2\Bigg\{\sigma\left(\frac{1}{y_i} + \frac{1}{\hat{\mu}_i}\right) + \left(\frac{y_i}{2} - \frac{1}{4}\right)\left(\log\left[2\left(1+\frac{\sigma}{2\hat{\mu}_i^2}\right)\right] - \log\left[2\left(1+\frac{\sigma}{2y_i^2}\right)\right]\right) \\ &\quad + \mathrm{K}_{y_i-1/2}(y_i) - \mathrm{K}_{y_i-1/2}(\hat{\mu}_i)\Big\}, \end{aligned} \tag{28}$$

em que $\mathrm{K}_{y_i-1/2}(\alpha)$ é dado por: $\mathrm{K}_{y_i-1/2}(\alpha) = \mathrm{K}_{y_i-1/2}\left[\sqrt{2\sigma\left(1+\frac{\sigma}{2\alpha^2}\right)}\right]$.

Portanto, temos como resíduo *deviance* para a Poisson-Inversa Gaussiana: $r_i^D = sinal(y_i - \hat{\mu}_i)\sqrt{d_i}$ em que d_i é expresso por (28).

– Poisson-Lindley Generalizada:

Seja $\Theta = (\boldsymbol{\beta}^\top, \theta)^\top$ um vetor de parâmetros. A função de log-verossimilhança para o modelo pode ser escrita da seguinte forma, considerando a substituição de α_i (21)

$$
\begin{aligned}
l(\Theta) = \sum_{i=0}^{n} \Bigg\{ & \log\left[\Gamma\left(y_i + \frac{\mu_i\theta(\theta+1)-1}{\theta+1}\right)\right] - \log y_i! - \log\left[\Gamma\left(1 + \frac{\mu_i\theta(\theta+1)-1}{\theta+1}\right)\right] \\
& + \left(\frac{\mu_i\theta(\theta+1)-1}{\theta+1} + 1\right)\log\theta - \left(y_i + \frac{\mu_i\theta(\theta+1)-1}{\theta+1} + 2\right)\log(\theta+1) \\
& + \log\left[\mu_i\theta(\theta+1) - 1 + y_i + \frac{\mu_i\theta(\theta+1)-1}{\theta+1}\right]\Bigg\}, \quad \text{com} \quad \mu_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}).
\end{aligned}
$$

O desvio será dado de forma a considerar a log-verossimilhança do modelo estimado e do modelo saturado ($\tilde{\mu}_i = y_i$), de forma que: $d_i = 2[l(y_i, y_i) - l(\hat{\mu}_i, y_i)]$.

Assim, para utilizar os resíduos componente da *deviance* de forma padronizada, basta dividirmos os resíduos por $(1-\hat{h}_{ii})^{1/2}$, onde $w_i = \hat{\mu}_i$, citado anteriormente.

Aplicação - Abelhas

Os dados considerados foram retirados de Marciano (2009) e Mendes (2017). Um estudo foi realizado na área de apicultura, com o intuito de verificar o número de abelhas que polinizam determinada espécie de planta no decorrer do tempo. Para isto, realizou-se quatro coletas em um intervalo de tempo variável segundo a hora do dia. Os horários das coletas forma: 4,5,6,8,10,12,14,16 e 18 horas, tendo no total 36 observações. Os dados foram ajustado considerando a variável resposta o número de abelhas coletando pólen.

Ao realizar uma análise inicial, é possível verificar que o número mínimo de abelhas observadas corresponde a zero, algo que ocorre até o primeiro quartil, ou seja, pelo menos 25% dos dados. Ao comparar a variância com a média apresentada dos dados, vemos que a variância é bem superior à média amostral ($s^2 = 184,8 > \bar{x} = 11,1$), caracterizando um conjunto de dados com superdispersão, influenciado pelo excesso de zeros.

Na Tabela 1 são apresentados os valores do valor-p em relação ao teste Anderson-Darling (AD) para cada distribuição em relação a estimação da frequência dos vaores observados.

Tabela 1: Teste AD em relação aos valores estimados para a frequência de abelhas.

	P	PG	PIG	PLG
Valor-p	<0,001	0,6095	0,0895	0,6063

Fonte: Autores.

Modelagem dos dados

Observando o comportamento da variável resposta, número de abelhas coletando pólen, dado a covariável tempo. temos que ela não segue uma tendência linear nem quadrática. Assim, segundo o modelo

definido por Marciano (2009), ajustou-se um modelo cúbico, usando os t tempos, para o conjunto de dados. Como função de ligação, será utilizada a logarítmica. Além disso, afim de obter o melhor ajuste, usou-se os modelos Poisson, Poisson-Gama, Poisson-Inversa Gaussiana e Poisson-Lindley Generalizada. Utilizou-se o *software* R com as seguintes funções *glm*(), *glm.nb*(), *gamlss*() e *optim*() para a modelagem dos dados, respectivamente.

Como o modelo Poisson-Lindley Generalizada não possui função pronta no *software* R para estimação, foi utilizada a função *optim* atráves do método *SANN*, que consiste em utilizarmos a verossimilhança para a estimação dos parâmetros do modelo (ver detalhes em Béslisle, 1992). Como as estimativas variam de acordo com os valores iniciais propostos, escolheu-se recorrer como valores iniciais para o vetor de β's as estimativas dadas pelos modelos Poisson, Poisson-Gama e Poisson-Inversa Gaussiana e, para o parâmetro θ, avaliou-se uma grade de parâmetros possíveis conforme feito por Mendes (2017). A Tabela 2, estão apresentados os critérios Akaike (AIC) e Bayesiano (BIC), respectivamente, para cada modelo.

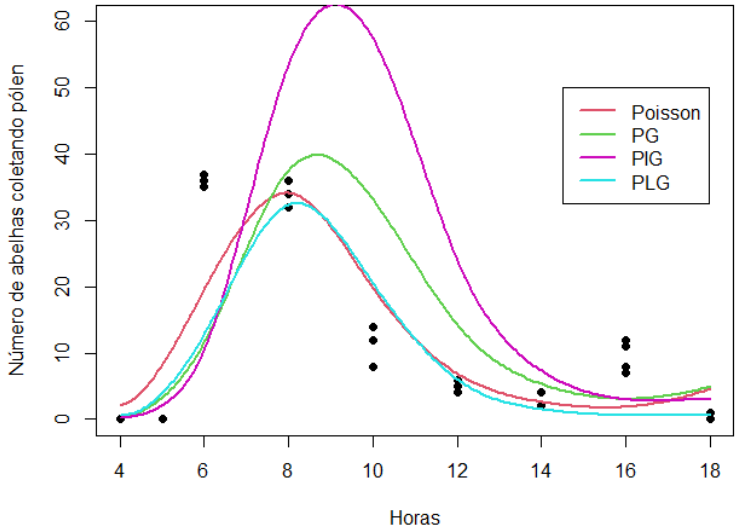
Tabela 2: Critérios AIC e BIC para cada modelo ajustado ao conjunto das abelhas.

Critérios	Poisson	PG	PIG	PLG
AIC	356,80	227,78	222,87	260,58
BIC	363,14	235,70	230,79	266,92

Fonte: Autores.

A Figura 1 retorna uma visualização gráfica referente ao comportamento do ajuste para cada um dos modelos.

Figura 1: Modelos do 3º grau ajustados e valores observados



Fonte: Autores.

Na Tabela 3 tem-se as estimativas e os erros padrão dos parâmetros de cada modelo ajustado, em que cada β_i está relacionado as horas do modelo cúbico. Nota-se que as estimativas de β_1, β_2 e β_3 não diferenciam muito em relação ao modelo utilizado, observando-se mais diferença ao se tratar do intercepto. Além disso, não há mudança de sinais nas estimativas.

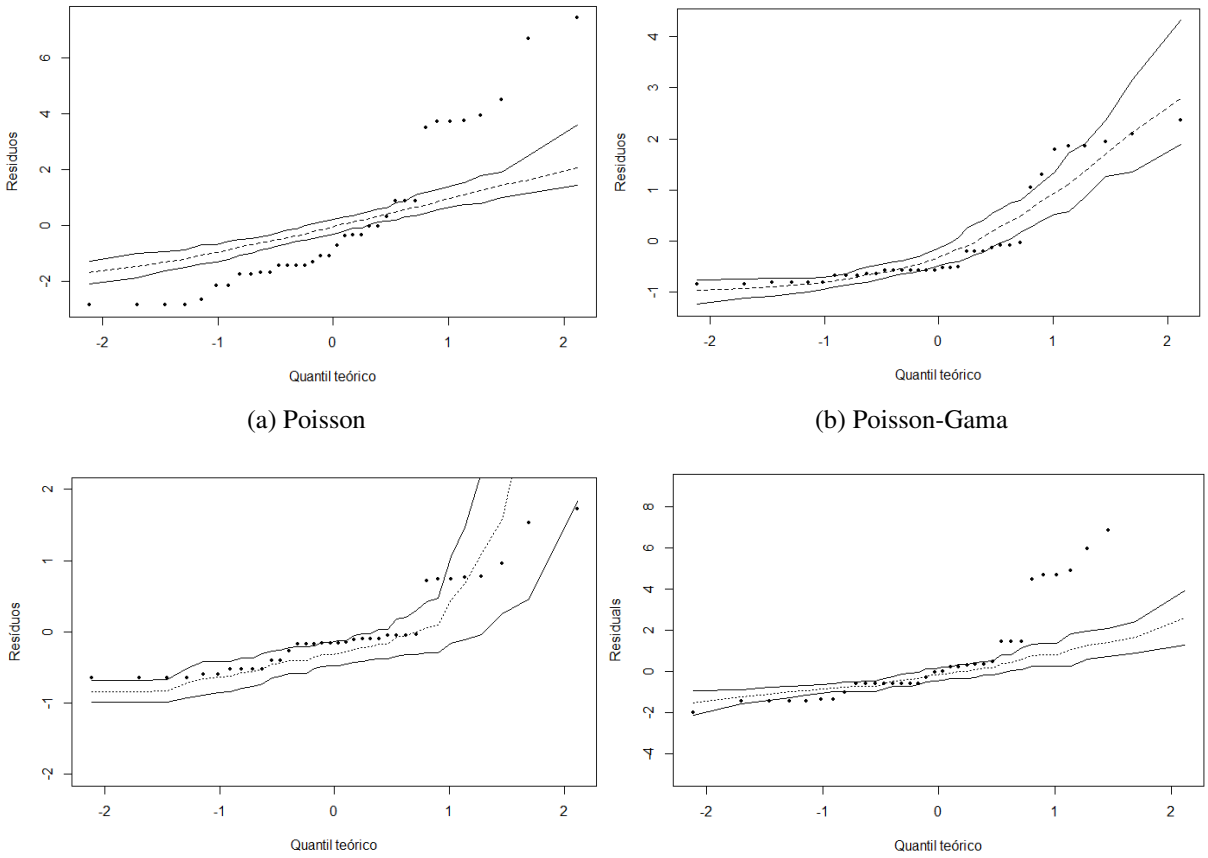
Os gráficos apresentados na Figura 2 são referentes aos resíduos de Pearson padronizados. Utilizando como base Atkinson (1981), que sugeriu o uso de uma espécie de banda para a flutuação dos pontos, foram construídos os resíduos simulados para o modelo PIG e PLG. Para o modelo Poisson e PG, utilizou-se a função *hnp*() do R.

Tabela 3: Estimativas e erros padrão dos modelos ajustados.

Parâmetros	Poisson	PG	PIG	PLG
Intercepto	$-11,120$ ***	$-14,658$ ***	$-18,037$ **	$-14,896$ ***
	(1,119)	(3,501)	(6,558)	(0,005)
β_1	$4,486$ ***	$5,127$ ***	$5,950$ **	$5,506$ ***
	(0,360)	(1,127)	(2,050)	(0,015)
β_2	$-0,429$ ***	$-0,451$ ***	$-0,502$ *	$-0,513$ ***
	(0,036)	(0,109)	(0,193)	(0,003)
β_3	$0,012$ ***	$0,012$ ***	$0,012$ *	$0,014$ ***
	(0,001)	(0,003)	(0,006)	(<0,001)
θ				0,996
σ/ϕ		0,835	0,898	

Fonte: Autores.

Figura 2: Envelope simulado para os resíduos de Pearson padronizado dos modelos ajustados



(a) Poisson

(b) Poisson-Gama

(c) Poisson-Inversa Gaussiana

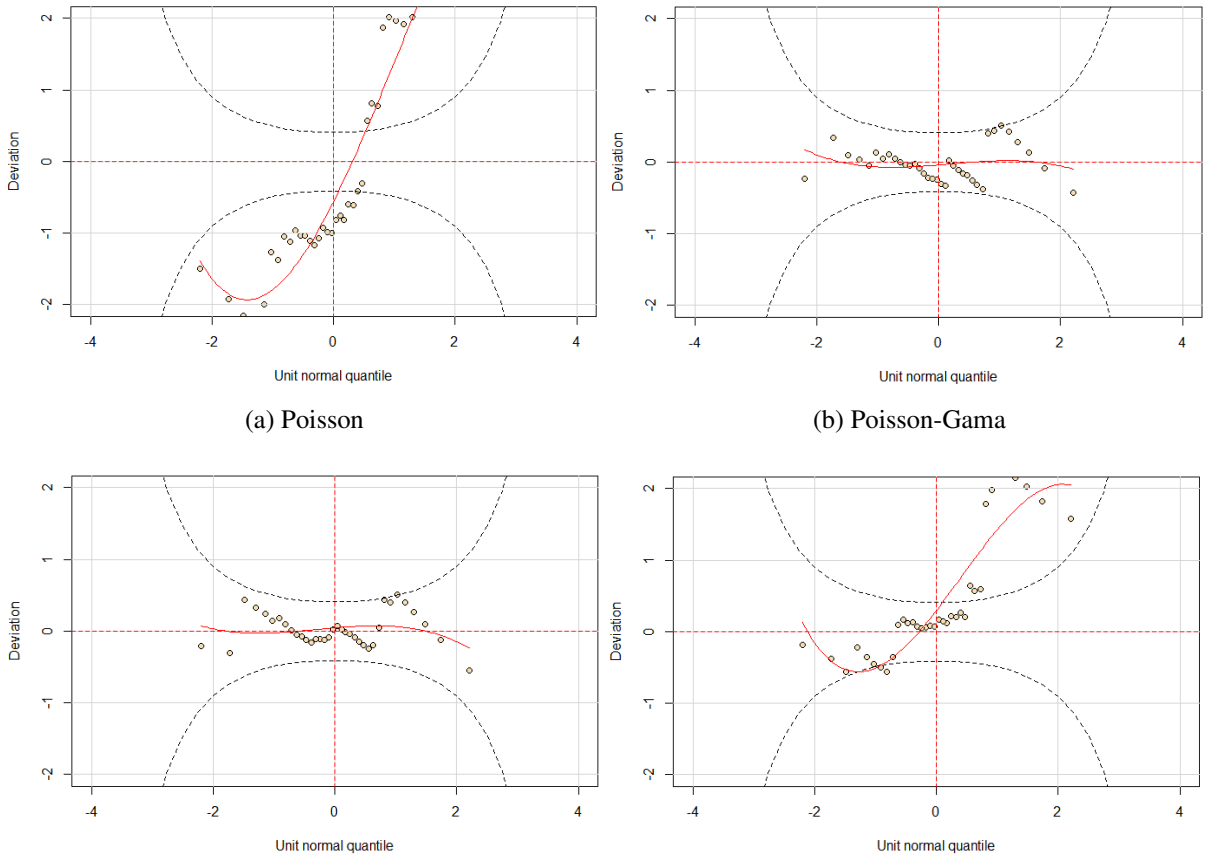
(d) Poisson-Lindley Generalizada

Fonte: Autores.

Apesar de ser esperado que haja alguns pontos fora da banda de confiança, ao se utilizar o ajuste do Modelo Poisson, mais de 95% das observações estão fora região de confiança. Em contraste, temos que para os demais modelos, aproximadamente 30% das observações estão fora das bandas de confiança.

Na Figura 3 é apresentado o *worm plot* para os resíduos. Este gráfico consiste em utilizar os resíduos quantílicos. Caso o modelo esteja bem ajustado, se espera que os resíduos estejam distribuídos sobre a reta que passa em zero. Assim, anulamos completamente a possibilidade de bom ajuste da Poisson pelos resíduos quantílicos. Apesar de não estarem tão distribuídos sobre a linha horizontal, o gráfico mostra que a PG e PIG aparentam estar bem ajustadas.

Figura 3: Envelope simulado para os resíduos quantílicos dos modelos ajustados



(a) Poisson

(b) Poisson-Gama

(c) Poisson-Inversa Gaussiana

(d) Poisson-Lindley Generalizada

Fonte: Autores.

Considerações Finais

O propósito deste trabalho foi apresentar alternativas para a análise de dados de contagem com excesso de zero, por meio de distribuições compostas para os modelos em dois estágios, que tem como base a distribuição de Poisson. Comparou-se os resultados obtidos das distribuições PG, PIG e PLG, de forma a verificar se realmente há uma melhora nos ajustes quando comparada com a distribuição de Poisson. Além disso, realizou-se uma análise dos resíduos para cada distribuição, tendo como foco implementar um código no *software* R, de forma que retornasse os resíduos e os gráficos, respectivamente, para a distribuição PLG, visto que esta não possui implementação pronta. Nas aplicações verificamos que o modelo Poisson-Gama e Poisson-Inversa Gaussiana apresentaram melhores resultados. Entretanto, a PLG também apresentou um ótimo valor nos critérios de informação e comportamente quando comparada com o ajuste de Poisson. Sendo um modelo simples quando comparado com algumas outras alternativas de misturas para dados com excesso de zeros, podendo servir como alternativa para a distribuição de Poisson quando trabalhado com dados com superdispersão.

Agradecimentos

Agradecemos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC - UFC) pela concessão da Bolsa de Iniciação Científica e auxílio financeiro ao qual possibilitou a realização do começo deste estudo. Agradeço à Dra. Sílvia Maria, minha orientadora, que com sua sabedoria e experiência me ensinou e auxiliou durante a pesquisa. Também agradeço ao Prof. Dr. Maurício Mota, pela colaboração na minha composição deste trabalho.

Referências

ATKINSON, A. C. *Plots, transformations, and regression : an introduction to graphical methods of diagnostic regression analysis*. Oxford New York: Clarendon Press Oxford University Press, 1985.

BÉLISLE, C. Convergence theorems for a class of simulated annealing algorithms on d. *Journal of Applied Probability*, v. 29, p. 885–895, 1992.

BICKEL, P.; DOKSUM, K. *Mathematical Statistics*. Oakland: Holden-day, Inc., 1977.

BRESLOW, N. Extra-poisson variation in log-linear models. *Applied Statistics*, v.33, 1984.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. *Modelos lineares generalizados e extensões*. Departamento de Estística e Informática, Recife, PE. Departamento de Ciências Exatas, USP, 2008.

COSTA, S. C. *Modelos Lineares Generalizados Mistos para Dados Longitudinais*. Tese (Doutorado em Agronomia) Escola Superior de Agricultura, USP, 2003.

COX, D.; SNELL, E. A general definition of residuals. *Journal of the Royal Statistical Society. Series B*, v.30, 07 1968.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, [American Statistical Association, Taylor Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America], v.5, n.3, p.236–244, 1996.

ELBATAL, I.; MEROVCI, F.; ELGARHY, M. A new generalized lindley distribution. *Mathematical Theory and Modeling*, v.3, p.30–47, 2013.

FAHRMEIR, L.; TUTZ, G. Multivariate statistical modelling based on generalized linear models. 2nd ed. *Journal of the American Statistical Association*, v.91, 1996.

GREENWOOD, M.; YULE, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, [Wiley, Royal Statistical Society], v. 83, n. 2, p. 255–279, 1920.

HILBE, J. M. *Modeling Count Data*. Cambridge University Press, 2014.

HINDE, J.; DEMETRIO, C. Overdispersion: Models and estimation. *Computational Statistics Data Analysis*, v. 27, p. 151–170, 1998.

HOLLA, M. On a poisson-inverse gaussian distribution. *Metrika*, v. 11, p. 115–121, 1967.

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, v. 34, 1992.

LINDLEY, D. Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society*. Series B, v. 20, 1958.

MAHMOUDI, E.; ZAKERZADEH, H. Generalized poisson–lindley distribution. *Communications in Statistics—Theory and Methods*, v. 39, p. 1785–1798, 2010.

MARCIANO, F. W. P. *Principais tipos de resíduos utilizados na análise de diagnóstico em MLG com aplicações para os modelos: Poisson, ZIP e ZINB*, 2009.

MCCULLAGH, J. A.; NELDER, P. *Generalized linear models*. Boca Raton London New York: Chapman and Hall, 1989.

MENDES, A. M. F. *Modelo Poisson-Lindley Generalizada para dados de contagem com superdispersão*. Monografia (Graduação em Estatística) – Departamento de Estatística e Matemática Aplicada, UFC, Fortaleza, 2017.

NELDER, J.A; WEDDERBURN, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, v.135, n.3, p.370-384, 1972.

MYERS, R.; MONTGOMERY, D.; VINING, G.; ROBINSON, T. *Generalized linear models: With applications in engineering and the sciences: Second edition*. 2010.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. 2013.

POISSON, S. Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés. *Paris, France: Bachelier*, 1837.

PUTRI, G.; NURROHMAH, s.; FITHRIANI, I. Comparing poisson-inverse gaussian model and negative binomial model on case study: horseshoe crabs data. *Journal of Physics: Conference Series*, 2020.

RSTUDIO TEAM. *RStudio: Integrated Development Environment for R*. Boston, MA, 2020. Disponível em: http://www.rstudio.com/.

SANKARAN, M. The discrete poisson-lindley distribution. *Biometrics*, v.26, p.145, 1970.

WONGRIN, W.; BODHISUWAN, W. *The poisson-generalized lindley distribution and its applications*. v. 38, p. 645–656, 2016.

ZAKERZADEH, H.; DOLATI, A. Generalized lindley distribution. *Journal of Mathematical Extension*, v.3, p.13–25, 2009.