Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas

Keliny Martins de Melo Sousa

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2010

Keliny Martins de Melo Sousa
Licenciada em Matemática

Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas

Orientador:
Prof . Dr. **VITOR AUGUSTO OZAKI**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2010

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Sousa, Keliny Martins de Melo
Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas / Keliny Martins de Melo Sousa. - - Piracicaba, 2010.
66 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2010.
Bibliografia.

1. Modelos lineares generalizados 2. Perdas agrícolas - Modelagem 3. Seguro Agrícola Verossimilhança I. Título

CDD 519.5
S725m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Dedicatória

A **Deus**,

por me dar força interior e coragem para concluir este trabalho, e pela perseverança de não desistir nunca.

Aos meus pais,

Waldinar de Sousa do Nascimento e Edenir Martins de Melo Sousa, *que sempre me apoiaram, estiveram presentes e acreditaram em meu potencial, me incentivando na busca de novas realizações.*

“ Dificuldades e obstáculos são fontes valiosas de saúde e força para qualquer sociedade.”

Albert Einstein

AGRADECIMENTOS

Tentarei expressar nestes poucos parágrafos apenas uma pequena parte da minha gratidão em relação a algumas pessoas que participaram direta ou inderatamente para a conclusão deste trabalho. Foram tantos sorrisos, lágrimas, decepções mas também muitas conquistas, que tornam a transcrição dessas poucas páginas um trabalho árduo. Antecipadamente, expresso minhas desculpas àqueles não mencionados, mas que de maneira alguma, são menos importantes para mim.

Agradeço, dedico, ofereço, em primeiro lugar a Deus, por ter me dado força, saúde, perserverança, humildade e sabedoria.

Àos meus pais, Waldinar de Sousa e Edenir Martins, por serem o meu porto seguro em muitos momentos da minha caminhada. Ao meu pai, uma pessoa simples e perseverante, que sempre incentivou e colocou a nossa formação profissional (minha e das minhas irmãs Keylla e Kennya e irmão Joabe) em primeiro lugar. À minha mãe, meu espelho de vida, de mãe, mulher, amiga, filha, por sua incansável generosidade me possibilitando aprender a cada dia a concretude do amor. Palavras me faltam para expressar toda a minha gratidão.

Aos meus avós, Nilo Martins, Joelina Melo, Nascimento(*in memoriam*) e Rita Sousa. Em especial, ao vozinho (Nilo), por todo incentivo e seu exemplo de caráter, e a vozinha (Joelina), por suas palavras de sabedoria e por ser exemplo de fidelidade a Deus. O amor incondicional de vocês é o que me faz caminhar.

Às minhas irmãs, Keylla Martins, Kennya Martins (minha psicóloga), e irmão Joabe Martins por tantas histórias e pelas diversas demonstrações de amor em cada palavra, gesto, silêncio, carinho, confiança, preocupação e dedicação.

Aos meus cunhados, Rui Evaldo e Daniel Melo e sobrinhas Raquel Yasmim (amor do tamanho do universo), Bianca Vida e Sofia Melo, por me proporcionar tantos momentos ímpares de alegria.

À minha "família" de Piracicaba, Renato Nunes, Ana Paula, Rafael Gregolin, Rose Muniz, Tiago Egydio, pela alegria, por momentos tão únicos e inesquecíveis, e por estarem sempre ao meu lado me mostrando a importância de ser persistente e de sempre lutar. Em especial a Ludmila (flor), por ser a minha "irmãzinha", sempre muito companheira e amiga.

Ao Prof. Dr. Vitor Augusto Ozaki, a compreensão, a paciência, a exigência, e a orientação na elaboração deste trabalho.

À Prof^a. Dr^a. Clarice Garcia Borges Demétrio, por toda a experiência, serenidade, sabedoria transmitidas durante essa etapa final do trabalho e por toda a amizade e paciência. Sua participação foi imprescendível para a conclusão deste trabalho.

Ao conselho do programa de Pós-Graduação em Estatística e Experimentação Agronômica.

Aos professores do Departamento de Ciências Exatas da ESALQ/USP, Prof. Dr. Carlos Tadeu dos Santos Dias, Prof. Dr. Décio Barbin, Prof. Dr. Cesár Gonçalves, Prof^a Dr^a Roseli Aparecida Leandro, Prof. Dr. Silvio Sandoval Zocchi e Prof^a. Dr^a. Sônia Maria Stefano Piedade.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, as secretárias Solange de Assis Paes Sabadin (Sô), e Luciane Brajão e aos técnicos em informática Jorge Alexandre Wiendl e Eduardo Bonilha, os auxílios permanentes.

Ao Antonio Carlos Ricardo Braga Junior, pela amizade, carinho, diversão e longas conversas.

À Mariana Ragassi Urbano, pela amizade e por sempre estar disposta a me ajudar.

Aos colegas de estudo do mestrado e do doutorado, a amizade, e a paciência, principalmente, dos amigos Epaminondas, Rodrigo, Patricia, Marcelino, Tiago de Oliveira, Raphael, Pedro, Adriana, Ana Patricia, Ricardo, Tiago Flor, Diógenes, Marcelo, Elizabeth, Fernanda, Elton, Paula Fontes, Guilherme, Cássio, Paula Klefens, Simone, Glaúcia, Vanderly e Alexandre (*in memoriam*).

Aos amigos de Teresina-PI, Américo Júnior, Áurea Beatriz e Danielle, pela amizade, carinho, compreensão, paciência e companheirismo, apesar da distância.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), o auxílio financeiro prestado.

A todas as pessoas que contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

RESUMO

Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas

O presente trabalho tem por objetivo utilizar a abordagem dos modelos lineares generalizados e os modelos de dispersão no contexto do seguro agrícola. Os modelos lineares generalizados (MLG's) constituem uma extensão dos modelos lineares de regressão múltipla introduzida por Nelder e Wedderburn (1972), que inclui modelos cuja variável resposta pertence à família exponencial de distribuições. O MLG é formado por um componente aleatório, que possui distribuição pertencente à famíia exponencial, um componente sistemático, conectados por uma função de ligação. Jorgensen (1997) estende a utilização dos MLG para uma classe mais ampla de modelos probabilísticos, denominados modelos de dispersão. A estimação dos parâmetros foi baseada no método da máxima verossimilhança, e também, em função da amostra ser relativamente pequena, optou-se pelo método de *bootstrap* não-paramétrico. As duas abordagens foram aplicadas a dois conjuntos de dados de sinistros de 15 municípios do estado do Rio Grande do Sul. Os resultados mostraram que a precipitação acumulada tem influência na ocorrência de sinistros. Entretanto, na modelagem do montante do sinistro não foi encontrada nenhuma variável significativa. Usando o método de *bootstrap*, foi encontrada influência das variáveis precipitação acumulada e a temperatura média no número de sinistros.

Palavras-chave: Seguro agrícola; Modelo linear generalizado; Modelo de dispersão; Superdispersão; Família Tweedie; *Bootstrap*

ABSTRACT

Generalized linear models and model dispersion applied to modelling agricultural claims

The main objective of this work is to use the generalized linear models and dispersion models in the agricultural insurance context. The Generalized Linear Model (GLM) are an extension of the multiple regression linear models presented by Nelder e Wedderburn (1972). This approach include situations in which the response variable can be included in exponencial the family. The GLM is composed of a randomized component, a sistematic component and the link functions. Jϕrgensen (1997) extend the application of the GLM for a more general class of probability models, called dispersion models. Both approaches were applied in two insurance datasets for 15 citys in Rio Grande do Sul. The parameters estimation was based in the maximum likelihood method, in addition, because of the relatively small sample, the non-parametric Bootstrap method was used. This study show, using GLM, that only the accumulated rainfall was statistically significant. However, any of the covariates was significant when modelling the amount of claims. In the analysis using Bootstrap method the accumulated rainfall and average temperature were significant when modelling the number of insurance clains.

keywords: Agricultural insurance; Generalized linear models; Dispersion models; Overdispersion; Tweedie family; Bootstrap

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A agricultura é uma das atividades mais antigas exercidas pelo homem. Por ser uma atividade econômica tipicamente caracterizada pela sua vulnerabilidade a eventos que estão fora do controle do produtor, mecanismos voltados a sua proteção tem sido desenvolvido. Podemos destacar o seguro agrícola, que é a forma mais eficaz de transferir o risco dos produtores para outros agentes econômicos (OZAKI, 2005).

Esse mecanismo de seguro agrícola, permite ao produtor igualar sua renda quando ocorre um evento que cause danos econômicos à situação em que tal evento não ocorra mediante o pagamento de um prêmio (importância paga pelo segurado à seguradora em troca da transferência do risco a que ele está exposto) e o recebimento de uma indenização, caso ocorra o sinistro.

Nesse sentido, é fundamental detectar quais variáveis influenciam a quantidade de sinistros durante a vigência de um contrato. Nos ramos não-agrícolas várias metodologias têm sido utilizadas para a modelagem de sinistros. Dentre elas, destacam-se os modelos lineares generalizados, introduzido por Nelder e Wedderburn (1972) e os modelos de dispersão desenvolvido por Jϕrgensen (1997). No entanto, não há referência de nenhum trabalho que utilize a metodologia supramencionada no contexto agrícola.

1.1 Objetivos

O objetivo geral da pesquisa é utilizar a abordagem dos modelos lineares generalizados e os modelos de dispersão no seguro agrícola testando diversas distribuições.

Neste cenário, dois objetivos específicos são propostos:

- Modelar os dados de sinistros agrícolas;
- Analisar uma série de modelos estatísticos a fim de tirar conclusões a respeito dos fatores geradores de sinistros.

1.2 Seguro agrícola

Em qualquer setor de atividade econômica existem riscos que variam em menor ou maior grau. No setor agrícola, além do risco de mercado, existem diversas outras fontes

que a tornam uma atividade eminentemente arriscada. A principal delas refere-se ao fato de que, a atividade agrícola é altamente dependente de condições ambientais de difícil controle pelo homem, de modo que as variáveis climáticas e sua interação com fatores bióticos podem influenciar sobremaneira o resultado final da safra. Nesse sentido, em diversos países, o seguro agrícola tem se destacado como um dos principais mecanismos de gestão de risco (OZAKI; SHIROTA, 2005).

No entanto, no Brasil, o seguro agrícola tem encontrado sérias dificuldades em se popularizar. Vários são os fatores (OZAKI, 2007):

- alto custo do prêmio: em virtude do elevado risco, as taxas de prêmio são bem mais elevadas quando comparada a outros ramos de seguro;

- risco moral: ocorre quando a seguradora é incapaz de observar se o produtor utilizou ou não o fator de produção;

- seleção adversa: situação em que pessoas mais prováveis em receber um dano coberto pelo seguro são mais propensas a demandá-lo;

- risco sistêmico: um evento causador do sinistro ocorre afetando não apenas um produtor, mas muitos produtores em uma vasta extensão territorial;

- inexistência de séries suficientemente longas de dados de produtividade agrícola individual.

Nesse contexto, o governo federal, por meio do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) sancionou, em meados de 2003, a lei 10823, que subvenciona o prêmio do seguro rural[1]. A lei tem por objetivo principal reduzir o valor do prêmio pago pelo produtor rural. Desde então, o seguro agrícola tem se expandido safra após safra em diversas regiões do país, principalmente na cultura de soja, milho e maçã. Além do subsídio federal, o setor ainda conta com os subsídios estaduais nos estados de Minas Gerais e São Paulo.

[1]O seguro rural é formado pelo seguro agrícola, pecuário, florestas, aqüícola, penhor rural, benfeitorias, vida e cédula do produto rural.

Se pelo lado dos demandantes, o incentivo criado reduziu o prêmio, pelo lado das seguradoras (ofertantes), existe o problema do risco sistêmico, podendo levar a carteira agrícola de uma seguradora à ruína. Para evitar acumular grandes prejuízos as seguradoras utilizam o resseguro, que é o "seguro das seguradoras". Ou seja, as seguradoras dividem entre uma ou mais resseguradoras uma parcela de suas apólices, reduzindo assim a exposição ao risco. Nesse sentido, a seguradora paga um prêmio para as resseguradoras garantindo em troca o ressarcimento das indenizações que forem devidas na proporção das apólices transferidas.

Para alavancar de vez esse mercado, o governo federal prevê, ainda no primeiro semestre de 2010, a sanção do Projeto de Lei Complementar 374, que cria o fundo de catástrofe. Esse fundo protegerá as seguradoras em caso de eventos catastróficos.

Se no campo político e coorporativo existe uma grande movimentação por parte do governo e das empresas, no universo acadêmico existem poucos estudos que se aplicam ao seguro agrícola.

Nesse sentido, esse trabalho pretende estudar o impacto das variáveis climáticas nos sinistros agrícolas por meio dos MLG's e modelo de dispersão.

A próxima seção revisa os modelos lineares generalizados e os modelos de dispersão detalhando a forma de estimação pelo método de máxima verossimilhança e por *bootstrap*.

2 DESENVOLVIMENTO

2.1 Revisão de Literatura

2.1.1 Modelos Lineares Generalizados

2.1.1.1 Definição

Os Modelos Lineares Generalizados (MLG's) constituem uma extensão dos modelos lineares de regressão múltipla introduzida por Nelder e Wedderburn (1972), que inclui modelos cuja variável resposta pertence à família exponencial de distribuições, além de dar maior flexibilidade para a relação funcional entre a média da variável resposta μ e o preditor linear η. A função de ligação entre a média e o preditor linear não é necessariamente a identidade, podendo assumir qualquer forma monótona não-linear.

Os MLG's podem ser usados quando se tem uma única variável aleatória Y associada a um conjunto de variáveis explanatórias $x_1, x_2, ..., x_p$. Para uma amostra de n observações $(y_i, \mathbf{x}_i)$ em que $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$, $i = \{1, 2, \ldots, n\}$ é o vetor coluna de variáveis explicativas, o MLG envolve três componentes (CORDEIRO; DEMÉTRIO, 2007):

i) Componente aleatório: a variável resposta é representada por um conjunto de variáveis aleatórias independentes $Y_1, ..., Y_n$ com distribuição pertencente a família exponencial de distribuições com médias $\mu_1, ..., \mu_n$. A função densidade de probabilidade de Y_i é dada por

$$f(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \tag{1}$$

em que $\phi > 0$, conhecido, é o parâmetro de dispersão e θ_i o parâmetro canônico.

ii) Componente sistemático: as variáveis explicativas entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{r=1}^{p} x_{ir}\beta_j = \mathbf{x}_i^T\beta \quad \text{ou} \quad \boldsymbol{\eta} = X\boldsymbol{\beta}, \tag{2}$$

em que $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})^T$ é a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ o vetor dos parâmetros, $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^T$ o preditor linear e $c(y_i, \phi)$ conhecido.

iii) Função de ligação: relaciona o componente aleatório ao componente sistemático, isto é,

$$\eta_i = g(\mu_i), \tag{3}$$

sendo $g(\cdot)$ uma função monótona e diferenciável.

Nesses termos, um MLG é definido por uma distribuição da família (1), uma estrutura linear (2) e uma função de ligação (3).

A generalização em relação ao modelo linear clássico é que:

- a variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica (distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; poisson e binomial negativa para contagens);

- as variáveis explicativas ou covariáveis, entram na forma de um modelo linear (componente sistemático);

- a ligação entre os componentes aleatório e sistemático é feita através de uma função de ligação.

Existem diversas possibilidades de escolha da função de ligação. Entretanto, essa escolha depende do problema de modelagem em particular. Se a função de ligação é escolhida de tal forma que $g(\mu_i) = \theta_i = \eta_i$, o preditor linear modela diretamente o parâmetro canônico θ_i e a função de ligação η_i é denominada de ligação canônica (MCcCULAGH; NELDER, 1989).

Para o modelo clássico de regressão, a função de ligação canônica é a identidade com preditor linear igual a média. Essa função de ligação é adequada no sentido em que η e μ, podem assumir valores na reta real. Mas quando se tratar, por exemplo, da distribuição de Poisson em que $\mu > 0$ possui restrições, a função de ligação identidade não deve ser usada, pois $\hat{\mu}$ poderá assumir valores negativos, dependendo dos valores obtidos para $\hat{\boldsymbol{\beta}}$.

Ainda, dados de contagem dispostos em tabela de contigência, sob a suposição de independência, conduzem, naturalmente, a efeitos multiplicativos cuja linearização pode ser obtida por meio da função de ligação logarítmica, isto é, $\eta = \log(\mu)$ e, portanto, $\mu = e^{\eta}$.

Seja a variável Y_i definida em (1). O valor esperado e a variância são dadas por:

$$\mathrm{E}(Y_i) = \mu_i = b^{'}(\theta_i) \quad \text{e} \quad \mathrm{Var}(Y_i) = \phi b^{''}(\theta_i).$$

A função que relaciona o parâmetro canônico θ_i com a média μ_i é denotada por $\theta_i = q(\mu_i)$. A função da média μ na variância é representada por $b''(\mu_i) = V(\mu_i)$, e denomina-se $V(\mu_i)$ de função de variância. Sendo assim, as distribuições da família exponencial têm relação conhecida entre a média e a variância, e essa relação é denotada por:

$$\mathrm{Var}(Y_i) = \phi V(\mu_i).$$

2.1.1.2 Exemplos

Como exemplos de distribuições pertencentes à família exponencial, têm-se as distribuições normal, Poisson, binomial, gama, normal inversa e binomial negativa.

- Distribuição Normal

A variável aleatória Y tem distribuição normal com função densidade de probabilidade dada por

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right],$$

com $\mu \varepsilon \mathbb{R}$ e $\sigma > 0$, conhecido. Desenvolvendo a função, tem-se

$$\begin{aligned} f(y;\mu,\sigma^2) &= \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \exp\left[\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right], \end{aligned}$$

e utilizando a notação da família exponencial tem-se: $\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \dfrac{\mu^2}{2} = \dfrac{\theta^2}{2}$ e $c(y,\phi) = -\frac{1}{2}\left[\dfrac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]$.

Portanto, a distribuição normal pertence à família exponencial de distribuições com média $\mathrm{E}(Y) = b'(\theta) = \dfrac{2\theta}{2} = \theta$ e variância $\mathrm{Var}(Y) = \phi b''(\theta) = \sigma^2$.

- Distribuição de Poisson

A variável aleatória Y tem distribuição de Poisson, com parâmetro $\mu > 0$ e função de probabilidade dada por

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad \text{para} \quad y = 0, 1, 2, 3, \ldots$$

Desenvolvendo a função, tem-se

$$f(y;\mu) = \exp[y\log\mu - \mu - \log(y!)],$$

e utilizando a notação da família exponencial tem-se: $\theta = \log(\mu) \Rightarrow \mu = e^{\theta}$, $\phi = 1$, $b(\theta) = \mu = e^{\theta}$ e $c(y,\phi) = -\log(y!)$. Portanto, a distribuição de Poisson pertence à família exponencial de distribuições com média $\mathrm{E}(Y) = b'(\theta) = \mu = e^{\theta}$ e variância $\mathrm{Var}(Y) = \phi b''(\theta) = e^{\theta}$.

O modelo de Poisson é de grande importância para a análise de dados em forma de contagem. Suas características principais são:

- proporciona, em geral, uma descrição satisfatória de dados experimentais cuja variância é proporcional à média;

- pode ser deduzido teoricamente de princípios elementares com um número mínimo de restrições;

- se eventos ocorrem independente e aleatoriamente no tempo, com taxa média de ocorrência constante, o modelo determina o número de eventos, em um intervalo de tempo especificado.

- Distribuição Binomial

A variável aleatória Y tem distribuição binomial com função de probabilidade dada por

$$f(y;\pi) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y}, \quad \pi \in [0,1], \quad y = 0,1,2,\ldots,n.$$

Desenvolvendo a função, tem-se

$$\begin{aligned} f(y;\pi) &= \exp\left[\log\binom{n}{y} + y\log\pi + (n-y)\log(1-\pi)\right] \\ &= \exp\left[y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}\right], \end{aligned}$$

e utilizando a notação da família exponencial tem-se:

$$\theta = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu}{n-\mu}\right) \Rightarrow \mu = \frac{ne^{\theta}}{(1+e^{\theta})},$$

$$\phi = 1, \quad b(\theta) = -n\log(1-\pi) = n\log(1+e^{\theta}) \quad \text{e} \quad c(y,\theta) = \log\binom{n}{y}.$$

Portanto, a distribuição binomial pertence à família exponencial de distribuições, com média e variância dadas, respectivamente, por

$$\mathrm{E}(Y) = b'(\theta) = \frac{ne^{\theta}}{1+e^{\theta}} = \mu \quad \text{e} \quad \mathrm{Var}(Y) = \phi b''(\theta) = \frac{ne^{\theta}}{(1+e^{\theta})^2} = \frac{n\mu - \mu^2}{n}.$$

O modelo binomial é usado, principalmente, no estudo de dados na forma de proporções e na análise de dados binários.

- Distribuição Gama

A variável aleatória Y tem distribuição gama com função densidade de probabilidade dada por:

$$f(y;\mu,v) = \frac{\left(\frac{v}{\mu}\right)^{v}}{\Gamma(v)} y^{v-1} \exp\left(\frac{-yv}{\mu}\right), \quad y > 0.$$

Desenvolvendo a função, tem-se

$$\begin{aligned} f(y;\mu,v) &= \exp\left\{\log\left(\frac{v}{\mu}\right)^{v} - \log\Gamma(v) + (v-1)\log(y) - \frac{yv}{\mu}\right\} \\ &= \exp\left\{v\left[y\left(-\frac{1}{\mu}\right) - \log(\mu)\right] + v\log(vy) - \log\Gamma(v) - \log(y)\right\}, \end{aligned}$$

e utilizando a notação da família exponencial tem-se:

$$\theta = -\frac{1}{\mu} \Rightarrow \mu = -\frac{1}{\theta}, \quad \phi = \frac{1}{v}, \quad b(\theta) = \log(\mu) = -\log(-\theta) \quad \text{e}$$

$$c(y,\phi) = v\log(vy) - \log\Gamma(v) - \log(y).$$

Portanto, a distribuição gama pertence à família exponencial de distribuições, com média e variância dadas, respectivamente, por

$$\mathrm{E}(Y) = b'(\theta) = -\frac{1}{\theta} = \mu \quad \text{e} \quad \mathrm{Var}(Y) = \phi b''(\theta) = \frac{1}{v}\cdot\mu^2 = \frac{\mu^2}{v}.$$

- Distribuição Normal Inversa

A variável aleatória Y tem distribuição normal inversa com função densidade de probabilidade dada por:

$$f(y;\mu,\sigma^2) = (2\pi\sigma^2 y^3)^{-1/2} \exp\left[-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}\right], \quad y > 0$$

Desenvolvendo a função, tem-se

$$\begin{aligned} f(y;\mu,\sigma^2) &= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2 y^3) - \frac{(y-\mu)^2}{2\mu^2\sigma^2 y}\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\left[y\left(-\frac{1}{2\mu^2}\right)+\frac{1}{\mu}\right] - \frac{1}{2}\log(2\pi\sigma^2 y^3) - \frac{1}{2\sigma^2 y}\right\}. \end{aligned}$$

e utilizando a notação da família exponencial, tem-se:

$$\theta = -\frac{1}{2\mu^2}, \quad \phi = \sigma^2, \quad b(\theta) = -\frac{1}{\mu} = -(-2\theta)^{1/2}$$

e

$$c(y,\phi) = -\frac{1}{2}\left[\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y}\right].$$

Portanto, a distribuição normal inversa pertence à família exponencial de distribuições com média e variância, respectivamente, $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2\mu^3$.

- Distribuição Binomial Negativa

A variável aleatória Y tem distribuição binomial negativa com função densidade de probabilidade dada por

$$f(y;\mu,k) = \frac{\Gamma(k+y)}{\Gamma(k)y!}\frac{\mu^y k^k}{(\mu+k)^{k+y}},$$

em que $k > 0$, conhecido, $\mu > 0$ e $y = 0, 1, \ldots$. Desenvolvendo a função, tem-se

$$\begin{aligned} f(y;\mu,k) &= \exp\left\{\log\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right) + y\log(\mu) + k\log(k) - (k+y)\log(\mu+k)\right\} \\ &= \exp\left\{y(\log(\mu) - \log(\mu+k)) + k(\log(k) - \log(\mu+k)) + \log\frac{\Gamma(k+y)}{\Gamma(k)y!}\right\}, \end{aligned}$$

e utilizando a notação da família exponencial tem-se:

$$\theta = \log\left(\frac{\mu}{\mu+k}\right), \quad \phi = 1, \quad b(\theta) = -k\log(1-e^\theta)$$

e

$$c(y,\phi) = \log\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right).$$

Portanto, a distribuição binomial negativa pertence à família exponencial de distribuições com média e variância, respectivamente, $\mathrm{E}(Y) = \frac{ke^\theta}{1-e^\theta}$ e $\mathrm{Var}(Y) = \frac{ke^\theta}{(1-e^\theta)^2}$.

2.1.1.3 Estimação dos parâmetros e verificação do ajuste do modelo

Vários são os métodos de estimação de parâmetros. Iremos basear o estudo no método da máxima verossimilhança e no método de *bootstrap* não-paramétrico em virtude da amostra ser relativamente pequena.

2.1.1.3.1 Estimação via máxima verossimilhança e verificação do ajuste do modelo

Seja um modelo linear generalizado e suponha que os dados a serem analisados sejam representados pelo vetor $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. O logarítmo da função de verossimilhança como função apenas de $\boldsymbol{\beta}$, considerando-se o parâmetro de dispersão ϕ conhecido, dado o vetor $\mathbf{y}$ é definido por (CORDEIRO; DEMÉTRIO, 2007):

$$l(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi),$$

em que $\theta_i = q(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \sum_{r=1}^{p} x_{ir}\beta_j$.

O vetor escore é formado pelas derivadas parciais de primeira ordem do logarítmo da função de verossimilhança, ou seja,

$$U_j = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{dl_i}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j}.$$

A estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ do vetor de parâmetros $\boldsymbol{\beta}$ é obtida igualando-se $U_j = 0$ para $j = 1, ..., p$. Em geral, as equações $U_j = 0$, $j = 1, ..., p$, não são lineares e devem ser resolvidas numericamente pelo método de Newton-Raphson ou o escore de Fisher.

O método de Newton-Raphson para a solução de uma equação $f(x) = 0$ é baseado na aproximação de Taylor para a função de $f(x)$ na vizinhança do ponto x_m, ou seja,

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})},$$

sendo $x^{(m+1)}$ e $x^{(m)}$, o valor de x nos passos $m+1$ e m, respectivamente; $f(x^{(m)})$ a função $f(x)$ avaliada em $x^{(m)}$ e $f'(x^{(m)})$ a derivada da função $f(x)$ avaliada em $x^{(m)}$.

O objetivo é obter a solução do sistema de equações $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$ e, usando-se a versão multivariada do método de Newton-Raphson, tem-se que:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{J}^{(m)})^{-1}\mathbf{U}^{(m)},$$

sendo $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ os vetores de parâmetros estimados nos passos m e $(m+1)$, respectivamente, $\mathbf{U}^{(m)}$ o vetor escore avaliado no passo m, e $(\mathbf{J}^{(m)})^{-1}$ a inversa da negativa da matriz de derivadas parciais de segunda ordem de $l(\boldsymbol{\beta})$, com elementos $-\partial^2 l(\boldsymbol{\beta})/\partial\beta_r\partial\beta_s$ avaliada no passo m.

O método escore de Fisher, envolve a substituição da matriz de informação observada, $\mathbf{J}$, pela matriz esperada de informação esperada de Fisher, $\mathbf{K}$. Logo,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (I^{(m)})^{-1}U^{(m)}.$$

A matriz de informação de Fisher para $\boldsymbol{\beta}$ é dada por

$$\mathbf{K} = \phi^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X},$$

em que $\mathbf{W} = \text{diag}\{w_1, \ldots, w_n\}$ uma matriz diagonal de pesos.

Quando um modelo é ajustado, é necessário verificar se, realmente, ele se ajusta às observações. Nelder e Wedderburn (1972) desenvolveram a estatística desvio, definida por

$$S_p = 2(\hat{l_n} - \hat{l_p}) = 2\phi^{-1}\sum_{i=1}^{n}\{y_i[\tilde{\theta_i} - \hat{\theta_i}] - b(\tilde{\theta_i}) + b(\hat{\theta_i})\} = \phi^{-1}D_p,$$

sendo S_p o desvio escalonado, $\hat{l}_n$ e $\hat{l}_p$ os máximos do logarítmo da função de verossimilhança para os modelos saturados e correntes, respectivamente, $\tilde{\theta_i} = \tilde{\theta_i}(y_i)$ e $\hat{\theta_i} = \hat{\theta_i}(\mu_i)$ as estimativas do parâmetro canônico sob os modelos saturado e corrente, e D_p o desvio que pode ser calculado a partir das observações e das estimativas de suas respectivas médias.

Pode-se, ainda, escrever

$$S_p = \frac{1}{\phi}\sum_{i=1}^{n} d_i^2,$$

sendo que d_i^2 mede a diferença dos logarítmos das funções de verossimilhança observada e ajustada, para a observação i correspondente, e é chamado de componente do desvio. A soma deles mede a discrepância total entre os dois modelos na escala logarítmica da verossimilhança.

Sejam dois modelos encaixados M_q e M_p, com q e p parâmetros, respectivamente, $q < p$, então a diferença

$$S_p - S_q = \phi^{-1}(D_p - D_q) \sim \mathcal{X}^2_{p-q},$$

tem distribuição assintótica qui-quadrado com $(p - q)$ graus de liberdade, dado ϕ conhecido.

Se ϕ for desconhecido, deve-se obter uma estimativa $\hat{\phi}$ consistente, e a inferência é baseada na estatística F, dada por

$$F = \frac{(D_p - D_q)/(p - q)}{\hat{\phi}} \sim \mathrm{F}_{[(p-q),(n-p)]}. \tag{4}$$

para testar o efeito dos termos que estão incluídos no modelo M_p e não estão incluídos no modelo M_q.

Assim, um modelo bem (mal) ajustado aos dados, com uma verossimilhança máxima grande (pequena), tem um pequeno (grande) desvio. Entretanto, um grande número de variáveis explanatórias, visando reduzir o desvio, significa um grau de complexidade na interpretação do modelo. Na prática, procuram-se modelos simples com desvios moderados, situados entre os modelos mais complicados e os que não se ajustam bem aos dados. Portanto, quanto melhor for o ajuste do MLG aos dados tanto menor será o valor do desvio D_p.

Ainda, para um modelo bem ajustado, espera-se que a desvio residual seja aproximadamente igual ao número de graus de liberdade residual, devido à suposição de aproximação da distribuição do desvio pela distribuição $\mathcal{X}^2$. Quando isso não acontece, uma explicação é que a variância pode ser maior que o previsto para o modelo e esse fenômeno é descrito como superdispersão (HINDE; DEMÉTRIO, 1998a, b).

- Superdispersão

A utilização dos modelos lineares generalizados na análise de dados tem se tornado de uso frequente, principalmente com o avanço dos recursos computacionais disponíveis. Entretanto, deve-se tomar cuidado, principalmente no caso de variáveis discretas, com a superdispersão. Os casos mais comuns são dados na forma de proporção e de contagem, cuja forma padrão de análise envolve o uso dos modelos binomial e Poisson.

Para um bom ajuste espera-se que o desvio residual esteja próximo ao número de graus de liberdade do desvio residual. Quando isso não ocorre, uma explicação alternativa

para a falta do ajuste está relacionada ao componente aleatório do MLG, isto é, a variabilidade da variável aleatória que é maior do que a predita pelos modelos binomial ou de Poisson, ou seja,

i) dados na forma de proporção com $\text{Var}(Y_i) > m_i\pi_i(1-\pi_i)$; e,

ii) dados na forma de contagem com $\text{Var}(Y_i) > \mu_i$,

em que Y_i é a variável resposta. Nestes casos, tem-se que $\phi > 1$, fato conhecido por superdispersão. Pode ocorrer, também, a subdispersão, situação em que $\phi < 1$.

Diferentes modelos e métodos de estimação têm sido propostos na literatura para resolver os problemas de superdispersão. Hinde e Demétrio (1998b) dividiram as diferentes formas de abordagens da superdispersão em dois grupos:

i) assumir alguma forma mais geral para a função de variância, possivelmente incluindo parâmetros adicionais que podem ser estimados por métodos como quase-verossimilhança, pseudo-verossimilhança e momentos;

ii) assumir um modelo de dois estágios para a variável resposta, isto é, supondo que o parâmetro da distribuição da variável resposta é uma variável aleatória com alguma distribuição.

No caso em que a variável resposta é uma proporção, e supondo que Y_i tem distribuição binomial com

$$\text{E}(Y_i) = \mu_i = m_i\pi_i \quad \text{e} \quad \text{Var}(Y_i) = m_i\pi_i(1-\pi_i), \tag{5}$$

em que π_i é a probabilidade do sucesso. Se a variância observada for maior do que a prevista por (5), existirá superdispersão e a distribuição binomial não é mais adequada. A forma mais simples de se modelar a superdispersão é especificar apenas os dois primeiros momentos, mantendo $\text{E}(Y_i) = \mu_i = m_i\pi_i$ e substituindo (5) por

$$\text{Var}(Y_i) = \phi m_i\pi_i(1-\pi_i), \quad \phi > 0 \quad \text{constante.}$$

A forma geral é dada por

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)\left[1 + \phi(m_i-1)^{\delta_1}\{\pi_i(1-\pi_i)\}^{\delta_2}\right]. \tag{6}$$

Várias funções de variância são definidas para diferentes valores de δ_1, δ_2 e ϕ em (6), ou seja,

i) se $\phi = 0$, tem-se o modelo binomial padrão, em que

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i);$$

ii) para $\delta_1 = 0$ e $\delta_2 = 0$, tem-se

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi],$$

que é o modelo de superdispersão constante, reparametrizado de forma diferente;

iii) para $\delta_1 = 1$ e $\delta_2 = 0$, tem-se

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+(m_i-1)\phi],$$

que é o modelo II de Williams (WILLIAMS, 1982), sendo que a função de variância da distribuição beta-binomial é um caso especial desse modelo;

iv) para $\delta_1 = 1$ e $\delta_2 = 1$, tem-se

$$\text{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi(m_i-1)\pi_i(1-\pi_i)],$$

que é o modelo III de Williams ou logístico normal.

Para dados de contagem, a forma geral é dada por:

$$\text{Var}(Y_i) = m_i\{1+\phi\mu_i^{\delta}\}. \tag{7}$$

Várias funções de variância são definidas para diferentes valores de δ e ϕ em (7), ou seja,

i) se $\phi = 0$, tem-se o modelo de Poisson padrão, em que

$$\text{Var}(Y_i) = \mu_i;$$

ii) para $\delta = 0$, tem-se

$$\text{Var}(Y_i) = \mu_i[1+\phi],$$

que é o modelo de superdispersão constante, reparametrizado de forma diferente;

iii) para $\delta = 1$, tem-se

$$\text{Var}(Y_i) = \mu_i[1 + \phi\mu_i^{\delta}],$$

que é a função de variância da distribuição binomial negativa.

Já os modelos do tipo (ii), isto é, os modelos em dois estágios, levam a modelos de probabilidade composta para a resposta e, em princípio, todos os parâmetros podem ser estimados usando-se o método da máxima verossimilhança. Em geral, a distribuição composta resultante não tem forma simples e métodos de estimação aproximados podem ser utilizados.

2.1.1.3.2 Tipos de resíduos em modelos lineares generalizados

Os resíduos representam uma ferramenta importante na verificação da adequação do modelo, pois por meio deles é possível saber se o modelo está ou não ajustado às observações.

As técnicas para análise de resíduos e diagnósticos para modelos lineares generalizados são semelhantes às técnicas utilizadas para modelos lineares clássicos, mas necessitam de algumas adaptações.

Assim, ajustado um modelo linear generalizado a um conjunto de observações y_i, $i = 1, \ldots, n$, vários tipos de resíduos podem ser definidos:

i) Resíduos ordinários

$$r_i = y_i - \hat{\mu}_i,$$

em que $\hat{\mu}_i$ é a média ajustada.

ii) Resíduos de Pearson

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\hat{V}^{1/2}}.$$

O resíduo r_i^p, assim definido, corresponde à contribuição de cada observação para o cálculo da estatística de Pearson generalizada. A desvantagem do resíduo de Pearson é que a sua distribuição é, geralmente, bastante assimétrica para modelos não normais.

iii) Resíduos de Anscombe

$$A_i = \frac{N(y_i) - N(\hat{\mu}_i)}{N'(\hat{\mu}_i)\hat{V}_i^{1/2}},$$

em que $N(y_i)$ é a transformação da observação y_i, escolhida para tornar a sua distribuição o mais próximo possível da normal. Barndorff-Nielsen (1978) mostra que a transformação

a considerar nos modelos lineares generalizados é da forma $N(\mu) = \int V^{-1/3} d\mu$ em que V é a função de variância.

iv) Resíduos de Pearson estudentizados

$$r_i^{p'} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - \hat{h}_{ii})}},$$

sendo $\hat{h}_{ii}$ o i-ésimo elemento da diagonal da matriz denotada por $H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$ com matriz de projeção W.

v) Componentes do desvio

$$r_i^D = \text{sinal}(y_i - \hat{\mu}_i) \sqrt{\frac{2\omega_i}{\phi}[y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\tilde{\theta}_i) - b(\hat{\theta}_i)]},$$

sendo que ω_i é um peso a priori.

vi) Componentes do desvio estudentizados

$$r_i^{D'} = \frac{r_i^D}{\sqrt{\hat{\theta}(1 - h_i)}},$$

sendo h_i o elemento da diagonal da matriz H.

2.1.1.3.3 O método de bootstrap e verificação do ajuste do modelo

O método de reamostragem chamado de *bootstrap*, introduzido por Efron (1979), é usado para estimar a distribuição das estatísticas de interesse, que muitas vezes são extremamente difíceis de serem obtidas pelos métodos tradicionais (exatos e assintóticos).

O *bootstrap* pode ser paramétrico ou não-paramétrico. O *bootstrap* não-paramétrico considera que a função de distribuição F, dos dados, é desconhecida e estimada através da distribuição empírica $\hat{F}$. Já o *bootstrap* paramétrico considera que a função de distribuição F pode ser estimada por $\hat{F}_{par}$ a partir de um modelo paramétrico conhecido para os dados.

Suponha que seja observada uma amostra aleatória $r_1, r_2, ..., r_n$ de uma distribuição F estimada pela distribuição $\hat{F}$, que pode ser paramétrica ou não.

Assim, $r = (r_1, r_2, ..., r_n)$ representa o vetor dos dados, para os quais se calcula o estimador $\hat{\beta} = s(\hat{F})$ de um parâmetro de interesse $\beta = s(F)$.

Considera-se que $\hat{F}$ é a distribuição empírica de r. Então, uma amostra *bootstrap* $R^* = (r_1^*, r_2^*, \ldots, r_n^*)$ é construída escolhendo-se aleatoriamente, com reposição, n elementos da amostra $r = (r_1, r_2, \ldots, r_n)$. Por exemplo, com $n = 6$, poder-se-ia pensar em $r^* = (r_5, r_3, r_6, r_1, r_4, r_1)$. A replicação *bootstrap* do parâmetro de interesse para essa amostra *bootstrap* é denotada por $\hat{\beta}^*$. Se forem geradas B amostras *bootstrap* $r^{*1}, r^{*2}, \ldots, r^{*B}$, a replicação *bootstrap* do parâmetro de interesse para a b-ésima amostra é dada por

$$\hat{\beta}^*(b) = s(r^{*b}), \tag{8}$$

ou seja, é o valor de $\hat{\beta}$ para a amostra *bootstrap* r^{*b}.

- Estimativa do erro padrão

A expressão para o estimador *bootstrap* do erro-padrão (EFRON; TIBSHIRANI, 1993) é dada por

$$\hat{\sigma}_{boot} = \sqrt{\sum_{b=1}^{B} \frac{[\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)]^2}{B-1}},$$

em que $\hat{\beta}^*(\cdot) = \sum_{b=1}^{B} \frac{\hat{\beta}^*(b)}{B}$, $\hat{\beta}^*(b)$ é descrita em (8) e B é o número de replicações *bootstrap*.

- Intervalo de confiança Bootstrap

Usando-se *bootstrap* podem-se obter intervalos aproximados de $(1-\alpha)100\%$ de confiança para o parâmetro de interesse β. Existem diferentes métodos para a construção de intervalos de confiança *bootstrap*. Como exemplo, tem-se o de percentis *bootstrap*.

i) Método percentis bootstrap

Um conjunto de dados *bootstrap* r^* é gerado de acordo com $\hat{F} \rightarrow R^*$. Com base nesse conjunto de dados são calculadas replicações *bootstrap* $\hat{\beta}^* = s(r^*)$. Considerando que $\hat{G}$ é a estimativa da função desconhecida da distribuição acumulada de $\hat{\beta}^*$. O intervalo percentil

de $(1-\alpha)100\%$ de confiança é definido por

$$[\hat{\beta}_{\%,inf}, \hat{\beta}_{\%,sup}] = [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]. \tag{9}$$

Por definição, tem-se que $\hat{G}^{-1}(\alpha) = \hat{\beta}^*(\alpha)$ é o $(100-\alpha)$-ésimo percentil da distribuição *bootstrap* de $\hat{\beta}^*$. Então os intervalos percentis podem ser escritos como

$$[\hat{\beta}_{\%,inf}, \hat{\beta}_{\%,sup}] = [\hat{\beta}^{*(\alpha)}, \hat{\beta}^{*(1-\alpha)}]. \tag{10}$$

As expressões (9) e (10) referem-se à situação ideal do *bootstrap* para a qual o número de replicações é infinito.

Na prática, usa-se um número finito B de replicações. Para o processo, geram-se B conjuntos de dados bootstrap $r_{*1}, r_{*2}, \ldots, r_{*B}$ e calculam-se as replicações *bootstrap* $\hat{\beta}(b) = s(r^{*(b)})$, $b = 1, 2, \ldots, B$.

Seja $\hat{\beta}^*_B(\alpha)$ o 100α-ésimo percentil empírico dos valores $\hat{\beta}^*(b)$, ou seja, o valor $B\alpha$-ésimo na lista ordenada das B replicações de $\hat{\beta}^*$. Assim, se $B = 1000$ e $\alpha = 0,05$, $\hat{\beta}^*_B(\alpha)$ é o 100α-ésimo valor ordenado das 1000 replicações. Se $(B \cdot \alpha)$ não é um inteiro, utiliza-se o maior inteiro menor do que ou igual a $(B+1)\alpha$. Assim, quanto maior for B, melhor serão os intervalos estimados. Logo, o intervalo percentil aproximado de $(1-\alpha)100\%$ de confiança é

$$[\hat{\beta}_{\%,inf}, \hat{\beta}_{\%,sup}] = [\hat{\beta}^{*(\alpha)}_B, \hat{\beta}^{*(1-\alpha)}_B].$$

2.1.2 Modelos de Dispersão

Em modelos lineares generalizados, a análise do desvio generaliza o método tradicional da análise de variância para dados normais. Assim, a idéia principal de Jϕrgensen (1997) é estender essa metodologia para uma classe mais ampla de modelos, ou seja, ter a soma dos quadrados dos resíduos da análise de variância generalizada para uma noção de desvio aplicável na análise do desvio.

Jϕrgensen (1997) apresenta a classe de dispersão (MD) que cobre um grande número de distribuições não-normais, incluindo distribuições dos seguintes sete tipos de dados (sendo D o suporte da distribuição):

1. Dados da reta real, $D = \mathbb{R}$.

2. Dados positivos, $D = \mathbb{R}_+$.

3. Dados positivos com zeros, $D = \mathbb{R}_0 = [0, \infty)$.

4. Proporções, $D = (0, 1)$.

5. Direções, $D = [0, 2\pi)$.

6. Dados na forma de contagens, $D = N_0 = 0, 1, 2, ...$

7. Dados Binomiais, $D = 0, ..., m$.

O conceito que fundamenta um modelo de dispersão é que a noção de locação e escala pode ser generalizada para posição e dispersão, respectivamente, para os sete tipos de dados (JϕRGENSEN, 1997).

2.1.2.1 Definição

Um modelo de dispersão MD(μ, σ^2) é uma família de distribuições cuja função densidade de probabilidade pode ser escrita na forma

$$f(y; \mu, \sigma^2) = a(y; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\} y \in C, \tag{11}$$

em que $\Omega \subseteq C \subseteq \mathbb{R}$, $a \geq 0$ é uma função apropriada, μ e σ^2 são, respectivamente, o parâmetro de posição e dispersão com $\mu \in \Omega$ e $\sigma^2 > 0$ e d é o desvio, satisfazendo $d(\mu; \mu) = 0$ para $\mu \in \Omega$ e $d(y; \mu) > 0$ para $y \neq \mu$.

A unidade do desvio é dita regular se $d(y; \mu)$ é diferenciável contínuo duas vezes com respectivo y, μ em $\Omega \times \Omega$ e satisfaz

$$\frac{\partial^2 d}{\partial \mu^2}(\mu, \mu) > 0 \quad \forall \mu \in \Omega.$$

A função de variância $V : \Omega \to \mathbb{R}_+$ do desvio regular é definida por

$$V(\mu) = \frac{2}{\dfrac{\partial^2 d}{\partial \mu^2}(\mu; \mu)}. \tag{12}$$

Os modelos de dispersão são classificados em duas classes: os modelos próprios de dispersão e os modelos exponenciais de dispersão, definidos a seguir.

2.1.2.2 Modelos Próprios de Dispersão

Uma distribuição de probabilidade denotada por PD(μ, σ^2) é chamada de modelo próprio de dispersão, se (11) é da forma:

$$f(y;\mu,\sigma^2) = \frac{a(\sigma^2)}{\sqrt{V(y)}} \exp\left\{-\frac{1}{2\sigma^2} d(y;\mu)\right\}, \quad y \in C \tag{13}$$

onde d é o desvio regular com função de variância $V(\mu)$. Os modelos próprios de dispersão são caracterizados pelo desvio d, pois a função de variância em (12) está em função de d, com $a(\sigma^2)$ em torno da constante normalizada.

As distribuições pertencentes à classe PD são distribuições usadas para análise de dados limitados em um intervalo, como os dados direcionais (ou cíclicos) e as proporções. Exemplos dessas distribuições são a distribuição de Von Mises para dados direcionais e a distribuição Simplex para proporções (JϕRGENSEN, 1997).

2.1.2.3 Modelos Exponenciais de Dispersão

A variável aleatória Y pertence à família de distribuições exponenciais de dispersão (ED), se sua função de densidade é da forma:

$$f(y;\theta) = c(y)\exp\{\theta y - \kappa(\theta)\} \quad y \in \mathbb{R}, \tag{14}$$

em que $c(y)$ é uma função e $\kappa(\theta)$ é a correspondente função de cumulante com o parâmetro canônico θ pertencente ao conjunto $\Theta = \{\theta \in \mathbb{R} | k(\theta) < \infty\}$. A média da variável aleatória Y com distribuição em (14) é $\mu = \tau(\theta)$, com $\tau(\theta) = \kappa'(\theta)$ e $\mu \in \Omega = \tau(\text{int}\Theta)$. A função de variância é $V(\mu) = \tau'\{\tau^{-1}(\mu)\}$. O modelo exponencial de dispersão pode ser aditivo e reprodutivo.

O modelo exponencial de dispersão aditivo, denotado por $Y \sim \text{ED}^*(\mu, \lambda)$, tem função de densidade dada por

$$f^*(y;\theta,\lambda) = c^*(y;\lambda)\exp(\theta y - \lambda\kappa(\theta)), \tag{15}$$

em que μ é a média, $\sigma^2 = 1/\lambda$ o parâmetro de dispersão e λ o parâmetro índice.

A função geradora de cumulantes é dada por

$$K^*(s;\theta,\lambda) = \lambda\{\kappa(\theta + s) - \kappa(\theta)\}.$$

Segue que a função geradora de momentos pode ser escrita como

$$K^*(s;\theta,\lambda) = \exp\{\lambda[\kappa(\theta+s) - \kappa(\theta)]\},$$

podendo-se obter a média

$$\mu = \lambda\kappa^{'}(\theta) = \tau(\theta),$$

e a variância

$$\mathrm{Var}(Y) = \lambda\kappa^{''}(\theta) = \frac{\mathrm{V}(\mu)}{\sigma^2} = \lambda\mathrm{V}(\mu),$$

sendo que a função de variância da média é dada por

$$\mathrm{V}(\mu) = \kappa^{''}(\tau^{-1}(\mu)) \quad \text{com} \quad \tau(\theta) = \kappa^{'}(\theta).$$

O modelo exponencial de dispersão reprodutivo é definido pela aplicação da transformação $X = Y/\lambda$ em (14), conhecida como transformação dual. A forma reprodutiva tem função densidade

$$f(x;\theta,\lambda) = c(x;\lambda)\exp[\lambda\{\theta x - \kappa(\theta)\}], \tag{16}$$

e função geradora de cumulantes dada por

$$K(s;\theta,\lambda) = \lambda\{\kappa(\theta + s/\lambda) - \kappa(\theta)\}.$$

Segue que a função geradora de momentos pode ser escrita como

$$K(s;\theta,\lambda) = \exp\lambda\{\kappa(\theta + s/\lambda) - \kappa(\theta)\}.$$

podendo-se obter a média $\mu = \kappa^{'}(\theta)$ e variância $\mathrm{Var}(X) = \lambda^{-1}\kappa^{''}(\theta)$.

Escrevendo a média μ em função de θ, tem-se

$$\tau(\theta) = \mu = k^{'}(\theta) \quad \text{de forma que} \quad \theta = \tau^{-1}(\mu).$$

Definindo a função de variância $\mathrm{V}(\mu) = \kappa^{''}(\tau^{-1}(\mu))$ e o parâmetro de dispersão $\sigma^2 = 1/\lambda$, a variância em termos da média e do parâmetro de dispersão é dada por

$$\mathrm{Var}(X) = \sigma^2\mathrm{V}(\mu).$$

Dessa forma, escreve-se $X \sim \text{ED}(\mu, \sigma^2)$. A unidade do desvio é dada por

$$d(x;\mu) = 2\left[\sup\{\theta x - \kappa(\theta)\} - x\tau^{-1}(\mu) + \kappa\{\tau^{-1}(\mu)\}\right], \quad \theta \in \Theta.$$

A propriedade de convolução para o modelo de dispersão exponencial aditivo é definida como se segue.

Seja $Y_1, \ldots, Y_n$ variáveis aleatórias independentes e $Y_i \sim \text{ED}^*(\theta, \lambda_i)$ com λ_i pertencente ao conjunto dos números reais positivos, ou seja, $\Lambda = \{\lambda_i | \lambda_i > 0\} = \mathbb{R}_+$ com $i = 1, \ldots, n$. Então, a distribuição de $Y_+ = Y_1 + \ldots + Y_n$ é dada por

$$Y_+ \sim \text{ED}^*(\theta; \lambda_1 + \ldots + \lambda_n).$$

Seja o modelo reprodutivo com transformação dual. Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes e

$$X_i \sim \text{ED}\left(\mu, \frac{\sigma^2}{\omega_i}\right) \quad i = 1, \ldots, n,$$

em que $\omega_1, \ldots, \omega_n$ são pesos positivos tal que $\omega_i/\sigma^2 \in \Lambda$ para todo i e $\mu \in \Omega$.

Escrevendo $\omega_+ = \omega_1 + \ldots + \omega_n$, a fórmula de convolução da forma reprodutiva é dada por

$$\frac{1}{\omega_+}\sum_{i=1}^{n} \omega_i X_i \sim \text{ED}\left(\mu, \frac{\sigma^2}{\omega_+}\right).$$

2.1.2.3.1 Exemplos de Modelos Exponenciais de Dispersão

- Distribuição Normal

Seja $N(\mu, \sigma^2)$ denotando a distribuição normal com média μ e variância σ^2 cuja função de densidade é dada por

$$\begin{aligned} p(y;\mu;\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{1}{2}\mu^2\right)\right\}. \end{aligned}$$

Fazendo $\theta = \mu$ e $\lambda = \dfrac{1}{\sigma^2}$, a distribuição normal é um modelo de dispersão exponencial reprodutivo com função de cumulante $k(\theta) = \dfrac{1}{2}\theta^2$. O valor médio $\tau(\theta) = \theta$ e a função de variância unitária é $\text{V}(\mu) = 1$, definida em $\Omega = \mathbb{R}$.

O desvio unitário da distribuição normal é $d(y;\mu) = (y-\mu)$.

- Distribuição Poisson

Considere a variável Y com distribuição de Poisson denotada por $Y \sim \text{Po}(\mu)$, com média $\mu = \lambda e^{\theta}$, sendo $\lambda > 0$ e $\theta \in \mathbb{R}$. Nessa parametrização, a função densidade de probabilidade é dada por:

$$p^*(y;\theta,\lambda) = \frac{\lambda^y}{y!}\exp(\theta y - \lambda e^{\theta}), \quad y = 0,1,2,\ldots$$

que é um modelo aditivo $\text{ED}^*(\theta,\lambda)$ com função de cumulante $\kappa(\theta) = e^{\theta}$. Os parâmetros λ e θ não são identificáveis nesse caso.

O desvio unitário de Poisson é

$$d(y;\mu) = 2\left\{y\log\frac{y}{\mu} - (y-\mu)\right\},$$

e a função de variância unitária é $V(\mu) = \mu$.

- Distribuição Gama

A distribuição Gama, denotada por $\text{Ga}(\mu,\sigma^2)$, é um modelo de dispersão exponencial reprodutivo, com função densidade dada por

$$p(y;\mu;\sigma^2) = \frac{\lambda^{\lambda}}{\Gamma(\lambda)}y^{\lambda-1}\exp\left\{-\lambda\left(\frac{y}{\mu} + \log\mu\right)\right\},$$

sendo μ a média, $\sigma^2 = \dfrac{1}{\lambda}$ o quadrado do coeficiente de variação, e $\theta = -\dfrac{1}{\mu}$ o parâmetro canônico.

A unidade do desvio é

$$d(y;\mu) = 2\left(\frac{y}{\mu} + \log\frac{\mu}{y} - 1\right),$$

e a função de variância unitária é $V(\mu) = \mu^2$, $\mu > 0$. Portanto, a variância de Y é dada por

$$\text{Var}(Y) = \sigma^2\mu^2.$$

- Distribuição Binomial

Seja $Y \sim \text{Bi}(m, \mu)$ denotando uma variável aleatória binomial, representando o número de sucessos em m tentativas independentes de Bernoulli com parâmetro de probabilidade μ. Escrevendo a função de probabilidade de Y em função do parâmetro logístico de $\theta = \log \mu/(1-\mu)$ obtém-se:

$$\begin{aligned} p^*(y; m, \mu) &= \binom{m}{y} \mu^y (1-\mu)^{m-y} \\ &= \binom{m}{y} \exp\{\theta y - m \log(1 + e^\theta)\}, \end{aligned}$$

para $y = 0, \ldots, m$.

O desvio unitário é

$$d(y; \mu) = 2 \left\{ y \log \frac{y}{\mu} + (1-y) \log 1 - y1 - \mu \right\},$$

e a função de variância unitária é dada por

$$V(\mu) = \mu(1-\mu), \quad \mu \in (0, 1).$$

Expressando-se a variância em termos da média $\mu = m\mu_0$, obtém-se a expressão

$$\text{Var}(Y) = mV(\mu/m) = \mu(1 - \mu/m).$$

- Distribuição Binomial Negativa

Seja Y uma variável aleatória binomial negativa, com parâmetro de probabilidade p e parâmetro de forma λ. A função de probabilidade de Y, para $y = 0, 1, \ldots,$ é dada por

$$\begin{aligned} p^*(y; p, \lambda) &= \binom{\lambda + y - 1}{y} p^y (1-p)^\lambda \\ &= \binom{\lambda + y - 1}{y} \exp\{\theta y + \lambda \log(1 - e^\theta)\}, \end{aligned}$$

com parâmetro canônico $\theta = \log p < 0$ e parâmetro índice $\lambda > 0$.

O desvio unitário é dado por

$$d(y; \mu) = 2 \left\{ y \log \frac{y(1+\mu)}{\mu(1+y)} + \log \frac{1+\mu}{1+y} \right\},$$

e a função de variância unitária é dada por

$$V(\mu) = \mu(1+\mu), \quad \mu > 0.$$

Expressando-se a variância em termos da média $\mu = \dfrac{\lambda p}{1-p}$ obtém-se a expressão

$$\text{Var}(Y) = \lambda V\left(\frac{\mu}{\lambda}\right) = \mu\left(1+\frac{\mu}{\lambda}\right) = \frac{\lambda p}{(1-p)^2}.$$

2.1.2.4 Modelo Exponencial de Dispersão Tweedie

Um caso especial dos modelos exponenciais de dispersão é o modelo Tweedie, denotado $\text{Tw}_p(\mu, \sigma^2)$, com função de variância definida por

$$V(\mu) = \mu^p, \quad \mu \in \Omega_p \tag{17}$$

em que p é o parâmetro com domínio $(-\infty, 0] \cup [1, \infty]$ e Ω definido na tabela 1.

Tabela 1 - Modelos de dispersão exponencial da família Tweedie

Distribuições	p	S	Ω	Θ
Extrema Estável	$p < 0$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{R}_0$
Normal	$p = 0$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
Não existe	$0 < p < 1$	-	$\mathbb{R}_+$	$\mathbb{R}_0$
Poisson	$p = 1$	$\mathbb{N}_0$	$\mathbb{R}_+$	$\mathbb{R}$
Poisson Composta	$1 < p < 2$	$\mathbb{R}_0$	$\mathbb{R}_+$	$\mathbb{R}_-$
Gama	$p = 2$	$\mathbb{R}_+$	$\mathbb{R}_+$	$\mathbb{R}_-$
Positiva Estável	$2 < p < 3$	$\mathbb{R}_+$	$\mathbb{R}_+$	-$\mathbb{R}_0$
Normal Inversa	$p = 3$	$\mathbb{R}_+$	$\mathbb{R}_+$	-$\mathbb{R}_0$
Positiva Estável	$p > 3$	$\mathbb{R}_+$	$\mathbb{R}_+$	-$\mathbb{R}_0$
Extrema Estável	$p = \infty$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}_-$

Fonte: Jϕrgensen (1997).

Por definição, o modelo tem média μ e variância $\text{Var}(Y) = \sigma^2\mu^p, \quad \text{com} \quad \mu \in \Omega_p$. A Figura 1 mostra a forma da distribuição Tweedie para alguns valores de p.

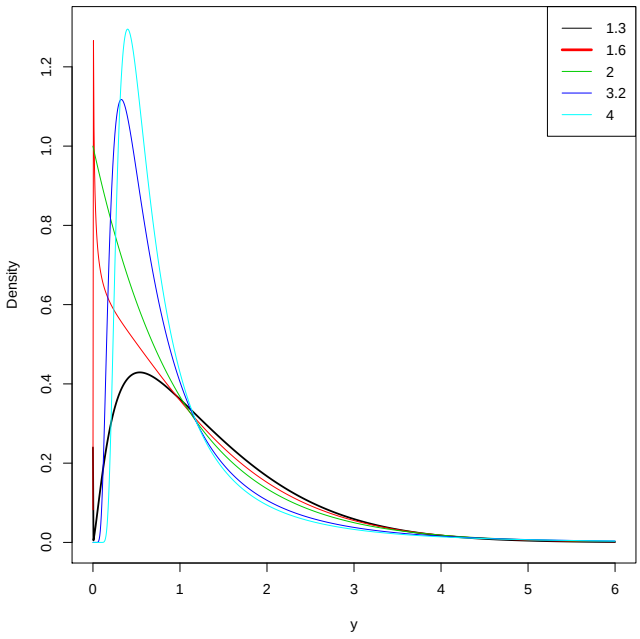

Figura 1 - Distribuição Tweedie para variações do valor de p

Jϕrgensen (1997) mostra, pela tabela 1, que as funções de variância com potência $0 < p < 1$ não correspondem a modelos exponenciais de dispersão. Para $p < 0$ ou $p > 2$, o modelo Tweedie é gerado pelas distribuições extrema estável ou positiva estável com índice $\alpha = \dfrac{p-2}{p-1}$. Para $p \neq 0, 1, 2$, com desvio unitário para a família $\mathrm{Tw}_p(\mu, \sigma^2)$ dada por

$$d_p(y;\mu) = 2\left[\frac{\{\max(y,0)\}^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right].$$

Para os casos em que $p < 0$ e $1 < p < 2$, o desvio unitário é finito para $y = 0$, com valor

$$d_p(0;\mu) = \frac{\mu^{2-p}}{2-p}.$$

O modelo Tweedie é invariante para as transformações de escala (Jϕrgensen, 1997), ou seja,

$$\text{Se } Y \sim \mathrm{Tw}_p(\mu, \sigma^2), \text{ então } cY \sim \mathrm{Tw}_p(c\mu, c^{2-p}\sigma^2),$$

para todo $c > 0$, já que E(cY)= $c\mu$ e Var(cY)= c^2 Var(Y)= $c^2\sigma^2\mu^p = \sigma^2 c^{2-p}(c\mu)^p$.

2.1.2.5 A Distribuição de Poisson Composta

Sejam $N, X_1, ..., X_N$ variáveis aleatórias independentes, tais que $N \sim \text{Po}(m)$ e $X_i's$ são identicamente distribuídas. A distribuição Z dada por

$$Z = \sum_{i=1}^{N} X_i \tag{18}$$

com Z definida em 0 para $N = 0$, é chamada de distribuição de Poisson composta. Seja $m = \lambda\kappa_p(\theta)$ e

$$X_i^* \sim \text{Ga}(\theta, -\alpha), \quad i = 1, 2, \ldots,$$

sendo que $\alpha = \dfrac{p-2}{p-1}$, com $\alpha, \theta < 0$. A fórmula de convolução é dada por

$$Z|N = n \sim \text{Ga}^*(\theta, -n\alpha) \quad \text{com} \quad n \geq 1.$$

Seja a função densidade de probabilidade da distribuição gama $\text{Ga}^*(\theta, \lambda)$ denotada por $p^*(x; \theta, \lambda)$. A função de densidade conjunta de Z e N, com $n \geq 1$ e $z > 0$, é dada por

$$\begin{aligned} p_{Z,N}(z, n; \theta, \lambda, \alpha) &= p^*(z; \theta, -n\alpha)\frac{m^n}{n!}e^{-m} \\ &= \frac{(-\theta)^{-n\alpha}m^n z^{-n\alpha-1}}{\Gamma(-n\alpha)n!}\exp\{\theta z - m\} \\ &= \frac{\lambda^n \kappa_p^n\left(-\frac{1}{z}\right)}{\Gamma(-n\alpha)n!z}\exp\{\theta z - \lambda\kappa_p(\theta)\}. \end{aligned} \tag{19}$$

Portanto, a variável aleatória Z em (18), tem função densidade de probabilidade dada por

$$p_Z^*(z; \theta, \lambda, \alpha) = \frac{1}{z}\sum_{n=1}^{\infty}\frac{\lambda^n \kappa_p^n\left(-\frac{1}{z}\right)}{\Gamma(-n\alpha)n!}\exp\{\theta z - \lambda\kappa_p(\theta)\},$$

em que

$$\kappa_p(\theta) = \frac{\alpha - 1}{\alpha}\left(\frac{\theta}{\alpha - 1}\right)^{\alpha}.$$

A distribuição conjunta de N, Z é definida por

$$p(z, n; \lambda, \alpha) = \frac{\lambda^n \kappa_p^n\left(-\frac{1}{z}\right)}{\Gamma(-n\alpha)n!z}.$$

Para $z > 0$ obtem-se a função de probabilidade condicional, para $n = 1, \ldots,$

$$p_{N|Z}(n|z; \lambda, \alpha) = \frac{p(z, n; \lambda, \alpha)}{\Sigma_{i=1}^{\infty} p(z, i; \lambda, \alpha)}.$$

Em particular, a média condicional de N dado Z é

$$E(N|Z) = \frac{\Sigma_{n=1}^{\infty} np(z, n; \lambda, \alpha)}{\Sigma_{n=1}^{\infty} p(z, n; \lambda, \alpha)}.$$

Generalizando (19), temos para o modelo $\mathrm{Tw}_p(\mu_i, \sigma^2/\omega_i)$, sendo ω_i o peso, a função de densidade conjunta dada por

$$f_{Y,N}(y, n; \theta, \sigma^2, \alpha) = \frac{\{(\lambda w)^{1-\alpha}\kappa_\alpha(-\frac{1}{y})\}^n}{\Gamma(-n\alpha)n!y} \exp[\lambda\omega\{\theta y - \kappa_\alpha(\theta)\}]. \tag{20}$$

Portanto, a função densidade de probabilidade é definida por

$$p(y; \theta, \sigma^2, \alpha) = c_p(y; \lambda) \exp\{\lambda\omega[\theta y - \kappa_p(\theta)]\}, \quad y \geq 0,$$

em que

$$c_p(y; \lambda) = \begin{cases} \dfrac{1}{y}\displaystyle\sum_{n=1}^{\infty} \dfrac{\{\lambda^n \kappa_p^n \left(-\frac{1}{\lambda y}\right)}{\Gamma(-n\alpha)n!} & \text{se } y > 0; \\ 1 & \text{se } y = 0. \end{cases}$$

A figura 2 representa algumas funções de densidade da Poisson composta para $p = 1, 5$.

2.1.2.6 Resíduos

Os resíduos têm uma importante aplicação na verificação e validação dos modelos. São eles:

- Resíduos de Pearson e Wald

Seja um modelo de dispersão $\mathrm{DM}(\mu, \sigma^2)$ com desvio unitário regular d e função de variância V. Seja Ω o domínio para μ e C o suporte. O resíduo de Pearson é definido por

$$r_P = \frac{y - \mu}{V^{1/2}(\mu)}.$$

Observa-se que r_P/σ é a padronização de $y - \mu$, considerando o desvio padrão assintótico $\sigma V^{1/2}(\mu)$ (para σ^2 pequeno). Para modelos exponenciais de dispersão, r_P tem média zero e variância σ^2.

Definindo a inversa do valor médio τ^{-1} para $\mu \in \Omega$ como

$$\tau^{-1}(\mu) = \int_{\mu_0}^{\mu} V^{-1}(t)dt,$$

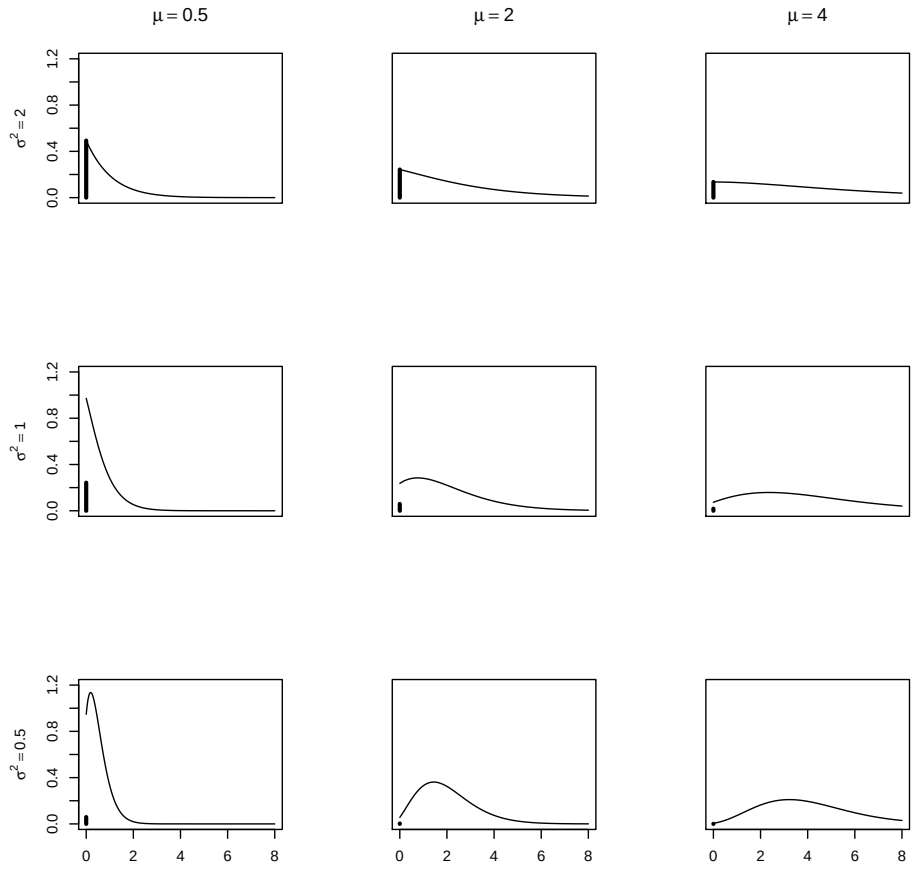

Figura 2 - Algumas funções densidade de probabilidade da Poisson composta. A probabilidade em zero é indicado pela barra

com $\mu_0 \in \Omega$ fixo, porém arbitrário. Considerando $\theta = \tau^{-1}(\mu)$ a generalização do parâmetro canônico define-se o resíduo de Wald como segue:

$$r_W = \{\tau^{-1}(y) - \tau^{-1}(\mu)\}V^{1/2}(y), \quad y \in \Omega.$$

Observa-se que essa definição não depende da escolha de μ_0.

- Resíduos Escore e Escore Dual

O resíduo escore é definido por

$$s = -\frac{\partial d}{2\partial \mu}V^{1/2}(\mu),$$

e s pode ser considerado como uma versão padronizada da estatística escore. O resíduo escore dual é dado por

$$u = \frac{\partial d}{2\partial y} V^{1/2}(y).$$

No caso de modelos exponenciais de dispersão, s e u reduzem-se a r_P e r_W, respectivamente (JϕRGENSEN, 1997).

2.1.2.7 Estimação dos Parâmetros

Sejam $Y_1, \ldots, Y_m$ variáveis aleatórias independentes com distribuição $\text{Tw}(\mu_i, \sigma^2/\omega_i)$, $i = 1, 2, \ldots, m$ e função densidade dada em (19). Considera-se a estimação do vetor de parâmetros $(\beta, \sigma^2, \alpha)$ sob este modelo, baseada na máxima verossimilhança, assumindo a estrutura de regressão

$$g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j, \tag{21}$$

sendo $\beta = (\beta_1, \ldots, \beta_k)^T$ um vetor de parâmetros desconhecidos. Assim, estimar θ significa estimar μ ou que por (21) significa estimar β.

Sejam $y = (y_1, \ldots, y_m)^T$, $\mu = (\mu_1, \ldots, \mu_m)^T$ e $\omega = (\omega_1, \ldots, \omega_m)^T$. A função de verossimilhança tem a forma

$$l(y; \mu, \sigma^2/\omega) = \prod_{i=1}^{m} f(y_i; \mu_i, \sigma^2/\omega_i).$$

e, portanto,

$$L(y; \mu, \sigma^2/\omega) = \log l(y; \mu, \sigma^2/\omega) = \sum_{i=1}^{m} L_i(y_i; \mu_i, \sigma^2/\omega_i),$$

com

$$L_i(y_i; \mu_i, \sigma^2/\omega) = \log f(y_i; \mu_i, \sigma^2/\omega_i).$$

Os estimadores $\hat{\beta}$ de máxima verossimilhança são soluções das equações

$$\frac{\partial L}{\partial \beta_i} = 0 \quad \text{ou seja} \quad \sum_{i=1}^{m} \frac{\partial L_i}{\partial \beta} = 0 \quad j = 1, \ldots, k.$$

Como $\mu = \kappa'(\theta) = \left(\frac{\theta}{\alpha - 1}\right)^{\alpha - 1}$ na família Tweedie, temos que $\theta = (\alpha - 1)\mu^{1/(\alpha-1)}$, e portanto, o logarítmo para a função de verossimilhança para $(\beta, \sigma^2, \alpha)$

é dado por

$$L(\beta,\sigma^2,\alpha)=\sum_{i=1}^{m}[n_i\{(1-\alpha)\log(\omega_i/\sigma^2)+\log\kappa_p(-1/y_i)\}-\log\Gamma(-n_i\alpha)$$
$$-\log(n_i!y_i)+\frac{\omega_i}{\sigma^2}\{y_i(\alpha-1)\mu_i^{1/(\alpha-2)}-\kappa_p((\alpha-1)\mu_i^{1/(\alpha-1)})\}].$$

Supondo β conhecido, os estimadores de máxima verossimilhança $\hat{\sigma}^2,\hat{\alpha}$, são soluções das equações de verossimilhança

$$\frac{\partial L}{\partial\sigma^2}=0. \tag{22}$$

$$\frac{\partial L}{\partial\alpha}=0. \tag{23}$$

Usando-se (21) pode-se isolar σ^2 como uma função de α, ou seja

$$\hat{\sigma}^2_\alpha=\frac{-\sum_{i=1}^{m}w_i\{y_i(\alpha-1)\mu_i^{1/(\alpha-1)}-\kappa_p((\alpha-1)\mu_i^{1/(\alpha-1)})\}}{n_+(1-\alpha)}, \tag{24}$$

com $n_+=n_1+...+n_m$.

Define-se a verossimilhança perfilada para α, supondo β conhecido, por $\tilde{L}_\beta(\alpha)=L(\beta,\alpha,\hat{\sigma}^2_\alpha)$ e usando (24) encontra-se

$$\tilde{L}_\beta(\alpha)=(1-\alpha)\sum_{i=1}^{m}n_i\log(\omega_i/\hat{\sigma}^2_\alpha)+\sum_{i=1}^{m}n_i\log\kappa_p(-1/y_i)$$
$$-\sum_{i=1}^{m}\log\Gamma(-n_i\alpha)-\sum_{i=1}^{m}\log(n_i!y_i)+n_+(\alpha-1).$$

Pela regra da cadeia

$$\frac{\partial\tilde{L}_\beta}{\partial\alpha}(\hat{\alpha})=\frac{\partial L}{\partial\sigma^2}(\hat{\sigma},\hat{\sigma}^2_\alpha)\frac{\partial\sigma^2}{\partial\alpha}+\frac{\partial L}{\partial\alpha}(\hat{\alpha},\hat{\sigma}^2_\alpha)=0.$$

e, portanto, o estimador de máxima verossimilhança $\hat{\alpha}$, é solução da equação

$$\frac{\partial\tilde{L}_\beta}{\partial\alpha}=0.$$

2.2 Material e Métodos

2.2.1 Material

O seguro agrícola é um dispositivo que pode ser usado pelo agricultor para estabilizar seu rendimento contra perdas parciais ou totais, na medida que essa perda é causada por condições climáticas adversas, que estão fora do controle do produtor. Com o objetivo de verificar quais são as variáveis que influenciam a incidência de sinistros, bem como o valor pago, foram analisados dois conjuntos de dados formado por 15 observações, referentes a 15 municípios do estado do Rio Grande do Sul, no ano de 2004.

Os dados foram cedidos pela Secretaria de Agricultura e Abastecimento do estado Rio Grande do Sul, e pelo Banco de Dados Climáticos da Embrapa.

Para o primeiro conjunto de dados, são consideradas as variáveis número de sinistros, precipitação acumulada e temperatura média e para o segundo conjunto de dados são consideradas as variáveis montante (valor) dos sinistros, precipitação acumulada e temperatura média. Para os municípios que não possuíam estações metereológicas em 2004, as observações para a precipitação acumulada e temperatura média foram estimadas pela estação mais próxima a esses municípios. Os dados da tabela 2 e 3 referem-se ao número de sinistros e montante do sinistro, respectivamente, registrados por município no período de 2004.

2.2.2 Métodos

Para a análise estatística dos conjuntos de dados, algumas distribuições sob a abordagem dos modelos lineares generalizados (MCcCULLAGH; NELDER, 1989) e modelos de dispersão (JϕRGENSEN, 1997), são ajustadas e os resultados discutidos.

Para verificar se as estimativas dos coeficientes dos modelos são estatisticamente diferentes de zero, foi assumida a condição assintótica. Além disso, por se tratar de uma amostra pequena (15 municípios), foi utilizado o método de bootstrap não-paramétrico para o cálculo das estimativas.

As análises foram feitas usando-se o *software* R, versão 2.10.1, sendo que os programas são apresentados no Anexo A.

Tabela 2 - Número de sinistros seguro agrícola solidário uva

Municípios	Número de Sinistros	Precipitação(mm)	Temperatura(C)
Flores da Cunha	27	1915	16,3
Caxias do Sul	16	1736	17,2
Monte Belo do Sul	13	1736	17,2
Bento Gonçalves	10	1736	17,2
Santa Tereza	8	1736	17,2
Garibaldi	3	1736	17,2
São Valentim do Sul	2	1692	16,9
Coronel Pilar	2	1736	17,2
Antônio Prado	2	1692	16,9
Dois Lajeados	1	1692	16,9
Liberato Salzano	1	1785	17,5
São Marcos	1	1915	16,3
Planalto	1	1811	19,4
Serafina Corrêa	1	1785	17,5
Monte Alegre dos Campos	1	1416	15,2

Fonte: Secretaria de Agricultura e Abastecimento do Rio Grande do Sul.

1. Modelos para análise do número de sinistros

Seja Y a variável aleatória números de sinistros em 15 minicípios do estado do Rio Grande do Sul. A distribuição padrão a ser assumida é a Poisson, por se tratar de dados de contagem de parâmetro λ, ou seja, $Y \sim \text{Poisson}(\lambda)$.

Assume-se a função de ligação logaritmica (canônica) e preditor linear

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{25}$$

em que, η_i é o preditor linear correspondente à observação y_i com $i = 1, 2, \ldots, 15$; β_0, β_1 e β_2 representam os parâmetros, e x_{1i}, x_{2i} representam, respectivamente, precipitação acumulada e temperatura média correspondente ao i-ésimo município.

Tabela 3 - Montante seguro agrícola solidário uva

Municípios	Montante (R$)	Número de Sinistros	Precipitação(mm)	Temperatura(C)
Flores da Cunha	3.798,6	27	1915	16,3
Caxias do Sul	2.572,8	16	1736	17,2
Monte Belo do Sul	2.263	13	1736	17,2
Bento Gonçalves	1.963,9	10	1736	17,2
Santa Tereza	1.632,6	8	1736	17,2
Garibaldi	3.045,3	3	1736	17,2
São Valentim do Sul	3.927	2	1692	16,9
Coronel Pilar	1.247,6	2	1736	17,2
Antônio Prado	86,3	2	1692	16,9
Dois Lajeados	5056,3	1	1692	16,9
Liberato Salzano	4.981,2	1	1785	17,5
São Marcos	1.423,7	1	1915	16,3
Planalto	1.399,3	1	1811	19,4
Serafina Corrêa	1.351,5	1	1785	17,5
Monte Alegre dos Campos	902.5	1	1416	15,2

Fonte: Secretaria de Agricultura e Abastecimento do Rio Grande do Sul.

Para um modelo bem ajustado, espera-se que a deviance residual seja aproximadamente igual ao número de graus de liberdade residual, devido à suposição de aproximação da distribuição do desvio pela distribuicao qui-quadrado $\mathcal{X}^2$. Quando isso não acontece, uma explicação é que a variância pode ser maior do que a prevista para o modelo e esse fenômeno é descrito como superdispersão (HINDE; DEMÉTRIO, 1998a, b).

Uma forma de modelar a superdispersão é substituir a variância do modelo original por uma forma mais geral, ou seja, $\text{Var}(Y_i) = \phi\mu$. Um modelo alternativo pode ser o binomial negativo para o número de sinistros.

Assim, assumindo uma nova estrutura para modelar Y, foi ajustado o modelo binomial negativo com variável resposta descrito pelo número de sinistros, função de ligação logarítmica, com o mesmo preditor linear dado em (25).

Para efetuar o ajuste do modelo generalizado proposto para o conjunto de dados

descrito, foi utilizado o método da máxima quase-verossimilhança para estimar os parâmetros de regressão e o parâmetro de dispersão ϕ.

A verificação do ajuste do modelo, foi realizado com base no gráfico de probabilidade normal com envelope simulado, sendo que a verificação da significância das variáveis explanatórias precipitação acumulada e temperatura média foi feita utilizando-se a estatística

$$F = \frac{(D_2 - D_1)/(f_2 - f_1)}{\hat{\phi}} \sim F_{f_2 - f_1, f_3},$$

em que f_1, f_2 denotam os números de graus de liberdade dos desvios D_1 e D_2, e $\hat{\phi}$ é estimado do modelo maximal com f_3 graus de liberdade.

Em virtude da amostra ser relativamente pequena utilizou-se como forma alternativa de estimação, o método de *bootstrap* não-paramétrico com 1000 replicações utilizando o modelo binomial negativo. A verificação da significância dos parâmetros é feita através do intervalo com 95% de confiança.

2. Modelos para análise do montante de sinistros

Seja $Z(\omega)$ descrito em (18) um modelo para o montante de sinistros correspondendo a uma exposição ao risco ω, com $N(\omega)$ representando o número de sinistros e Z_i o valor do i-ésimo sinistro. Essa exposição é definida como sendo $\omega = vt$, em que v é o valor segurado, e t o intervalo de tempo em que ω permanece exposto ao risco.

Seja $Y(\omega) = Z(\omega)/\omega$ a taxa de sinistros observada por unidade de exposição, e $\mu = \mathrm{E}[Y(\omega)]$ a taxa média de sinistros, assumida como sendo igual para todo ω.

A exposição dentro da análise, de forma apropriada, leva em consideração que a precisão de $Y(\omega)$, como estimador de μ aumenta com ω. Assim, assumindo a condição assintótica, temos que

$$Y_i \sim \mathrm{Tw}\left(\mu_i, \frac{\sigma^2}{\omega_i}\right).$$

A função de ligação utilizada será a logarítmica com preditor linear

$$\eta_i \;\; = \;\; \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{26}$$

em que, η_i é o preditor linear correspondente à observação y_i com $i = 1, 2, \ldots, 15$; β_0, β_1 e β_2 representam os parâmetros, e x_{1i}, x_{2i} representam precipitação acumulada e temperatura média correspondente ao i-ésimo município.

Para a estimação dos parâmetros $(\beta, \sigma^2, \alpha)$ utilizou-se o método da máxima verossimilhança assumindo a estrutura de regressão descrita em (21) e para a verificação do ajuste do modelo uma análise gráfica dos resíduos.

Em particular, a estimação para o parâmetro α é feita usando-se função de máxima verossimilhança perfilada, implementado no *software* R. Para verificação da significância das variáveis explanatórias precipitação acumulada e temperatura média, é utilizada a estatística F.

Da mesma forma que o caso anterior, optou-se pelo método de *bootstrap* não-paramétrico com 1000 replicações utilizando o modelo Tweedie. A verificação da significância dos parâmetros é feita através do intervalo com 95% de confiança.

2.3 Resultados e Discussão

Os resultados obtidos para a análise dos dois conjuntos de dados são apresentados e discutidos a seguir.

1. Modelos para análise do número de sinistros

O histograma da variável resposta (número de sinistros) do conjunto de dados da tabela 2 é representado pela figura 3.

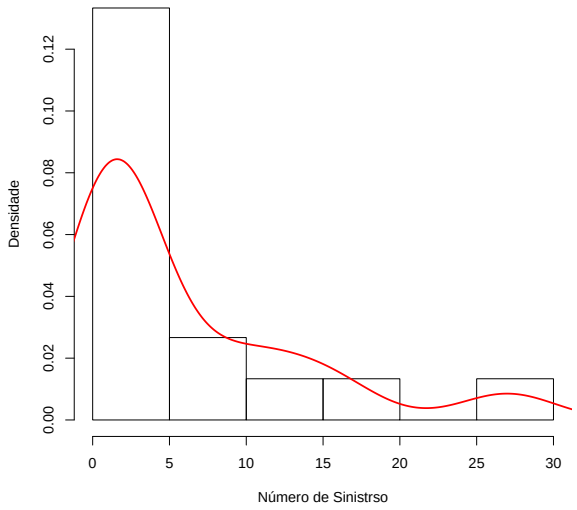

Figura 3 - Histograma da variável número de sinistros

Ajustando-se o modelo Poisson log-linear ao número de sinistros com preditor linear dado em (25), verificou-se que o desvio residual foi de 75,95 com 13 graus de liberdade, indicando a falta de ajuste desse modelo. Os gráficos dos resíduos versus valores ajustados e o gráfico normal de probabilidade com envelope simulado, confirmam esse resultado (Figura 4a, 4b, respectivamente).

Por outro lado, usando-se o modelo binomial negativo com preditor linear dado em (25), o gráfico normal de probabilidade com envelope simulado evidencia o ajuste do modelo, conforme pode-se verificar na figura 5a.

(a) (b)

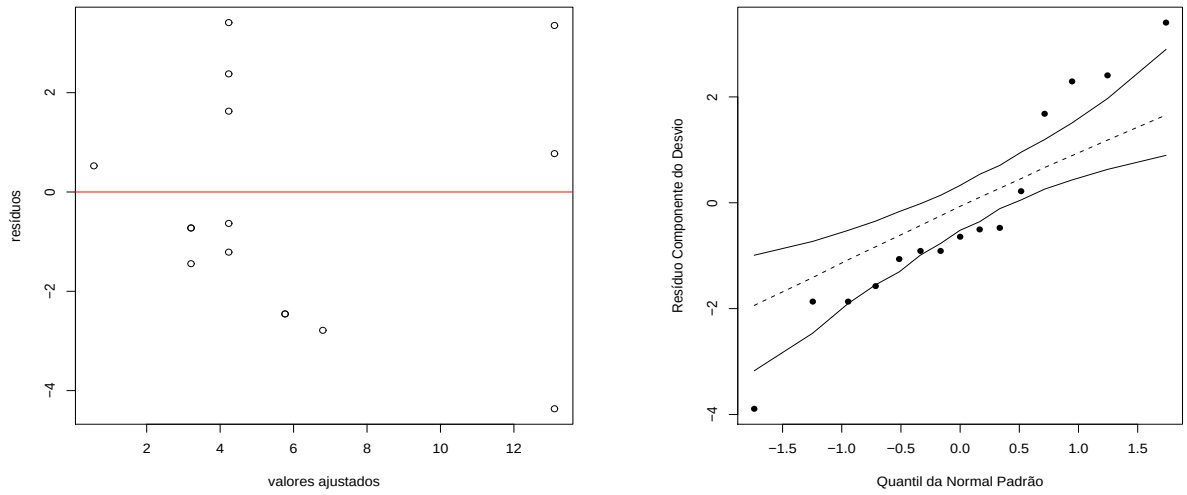

Figura 4 - (a) Gráfico dos resíduos versus valores ajustados o modelo de Poisson (b) Gráfico normal de probabilidade com envelope simulado para o modelo de Poisson

O teste F indica que há evidência significativa apenas para a precipitação acumulada. A equação do modelo ajustado é dada por

$$y_i = -0.004409 + 0,00557(\text{Precipitação})$$

Portanto, observa-se uma tendência crescente no número de sinistros como função da precipitação, ou seja, quanto maior a precipitação acumulada, maior será o número de sinistros. A figura 5b confirma essa tendência.

Para a análise através do método de *bootstrap*, as respectivas estimativas e o erro padrão dos parâmetros, são descritos na tabela 4.

De acordo com os resultados de *bootstrap*, as estimativas dos coeficientes de regressão para a precipitação acumulada e temperatura média, são estatisticamente diferente de zero com 95% de confiança, de acordo com o intervalo do percentil, pois o zero não está incluído nos intervalos de confiança construídos através desse método. A tabela 5 confirma esse resultado.

(a) (b)

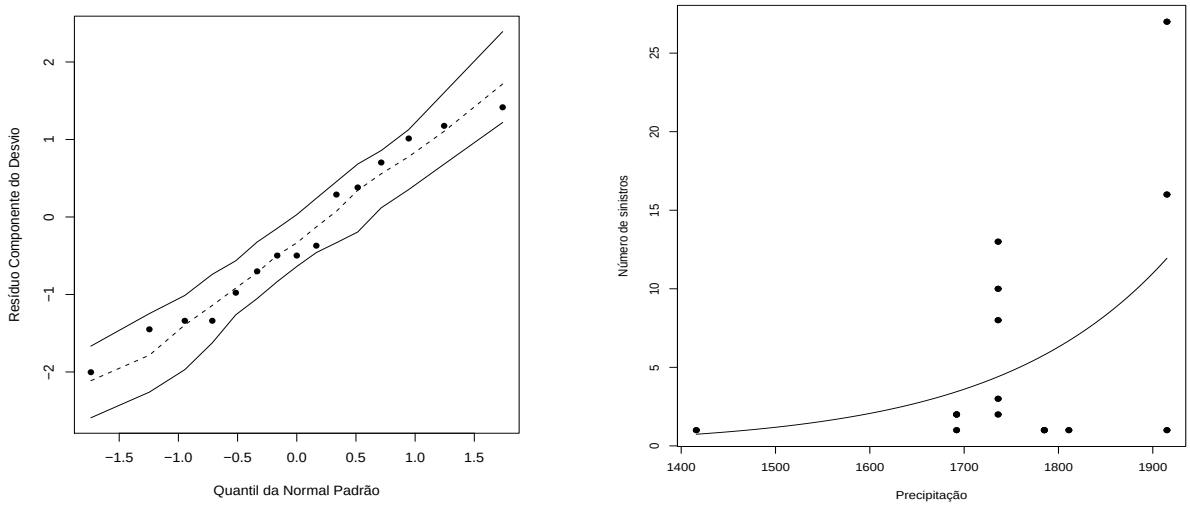

Figura 5 - (a) Envelopes de confianca para $\alpha = 5\%$ (b) Número de sinistros segundo a precipitação: valores observados e curva ajustada

Tabela 4 - Estimativas do *bootstrap* para o modelo binomial negativo

Parâmetros	Estimativa	Erro padrão
dispersão	2,634	3,893
intercepto	-0,265	0,439
precipitação	0,006	0,002
temperatura	-0,525	0,219

Fonte: Recurso do estudo.

Tabela 5 - Intervalo de confiança do *bootstrap* para o modelo binomial negativo

Parâmetro	Limite inferior	Limite superior
intercepto	-0,888	0,662
precipitação	0,001	0,011
temperatura	-1,042	-0,030

Fonte: Recurso do estudo.

Assim, para o método de *bootstrap* tem-se a precipitação acumulada e temperatura média explicando o número de sinistros.

2. Modelos para análise do montante de sinistros

O histograma da variável resposta (montante de sinistros) do conjunto de dados da tabela 3 é apresentado na figura 6.

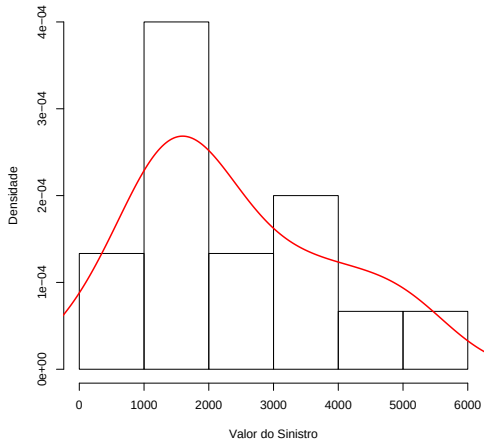

Figura 6 - Histograma do valor do sinistro

A estimativa de p, para o modelo proposto, obtida pelo método da verossimilhança perfilada foi 1,57, conforme indica a figura 7.

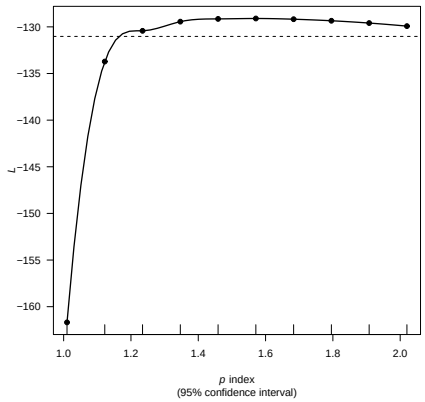

Figura 7 - Gráfico da verossimilhança perfilada para o modelo Tweedie

Para verificar se existe alguma relação entre número de sinistros e precipitação acumulada, foi ajustado a distribuição de Poisson composta com preditor linear descrito em

(26). Os gráficos dos resíduos representados na figura 8 confirmam esse resultado. Usando-se o teste F, verifica-se a não-significância para a precipitação acumulada e temperatura média.

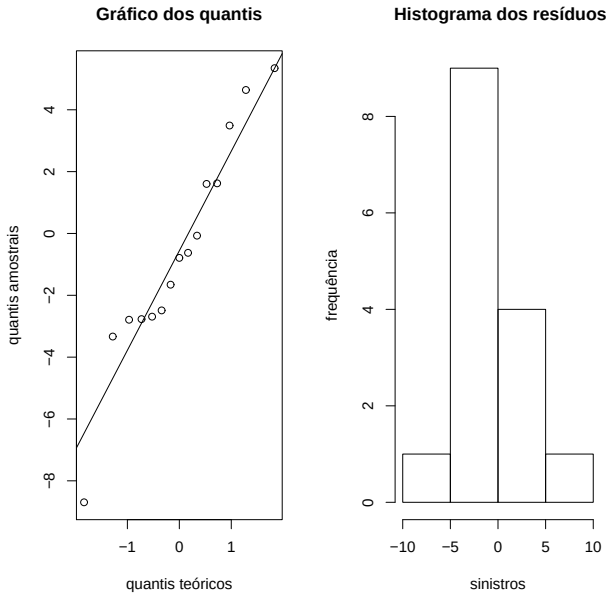

Figura 8 - Gráfico normal de probabilidades e histograma dos resíduos, ajustando-se o modelo Tweedie aos dados de seguro agrícola

Usando-se o método de *bootstrap*, o viés e o erro padrão dos parâmetros com base nos parâmetros originais, são apresentados na tabela 6.

Tabela 6 - Estimativas de *bootstrap* para o modelo Tweedie

Parâmetro	Original	Viés	Erro padrão
intercepto	2,278803e-02	-8,957139e-04	1,288518e-01
precipitação	-7,614930e-06	-1,879672e-06	5,742233e-05
temperatura	1,499804e-04	2,461980e-04	9,064080e-03

Fonte: Recurso do estudo.

De acordo com os resultados de *bootstrap*, as estimativas dos coeficientes de regressão para a precipitação acumulada e temperatura média, são estatisticamente iguais a zero com 95% de confiança, de acordo com o intervalo normal, pois o zero está incluído nos intervalos de confiança. A tabela 7 confirma esse resultado.

Tabela 7 - Intervalo de confiança de *bootstrap* para o modelo Tweedie

Parâmetro	Limite inferior	Limite superior
intercepto	-0.,289	0,2762
precipitação	-0,0001	0,0001
temperatura	-0,0179	0,0177

Fonte: Recurso do estudo.

Assim, para o método de *bootstrap* tem-se que a precipitação acumulada e temperatura média não explicam o montante de sinistros, confirmando assim, os resultados obtidos com o modelo Tweedie e usando-se a teoria assintótica.

3 CONSIDERAÇÕES FINAIS

O objetivo principal desse trabalho foi utilizar a abordagem dos modelos lineares generalizados e os modelos de dispersão para a análise de dados de seguro agrícola. Foram feitas duas aplicações em conjunto de dados reais, sendo que para o primeiro conjunto de dados devido à superdispersão observada, utilizou-se a distribuição binomial negativa. No segundo conjunto de dados, fez-se o uso da família de Tweedie.

Além disso, por se tratar de uma amostra pequena (15 municípios) foi usada a abordagem do *bootstrap* não-paramétrico para verificar a consistência das estimativas. A análise dos modelos foi desenvolvida utilizando o *software* R, através dos pacotes MASS, TWEEDIE e BOOT, e o *software* Ox.

Levando em consideração o método da máxima verossimilhança, observou-se o ajuste do conjunto de dados número de sinistros, por meio do envelope simulado (Fig.5a) com a precipitação acumulada sendo significativa, usando-se o teste F.

Para o conjunto de dados montante de sinistro, os gráficos de resíduos confirmam o ajuste (Fig.8), porém a precipitação acumulada e a temperatura média não se mostraram significativas, ou seja, as variáveis precipitação acumulada e temperatura média não explicam o montante do sinistro.

Usando-se *bootstrap* não-paramétrico, foi encontrada significância das variáveis precipitação acumulada e temperatura média, sobre o modelo binomial negativo, ou seja, essas variáveis explicam o número de sinistros. Entretanto, para o modelo Tweedie não foi encontrada significância dessas variáveis. Nesse sentido, a precipitação acumulada e a temperatura média não explicam o montante de sinistros.

Para uma análise acurada da sinistralidade, outras variáveis de grande relevância teriam que ser analisadas, por exemplo, a taxa média de subvenção do governo, o uso de maquinárias, mão-de-obra qualificada, uso de fertilizantes e inseticidas. Para tanto, seria necessária uma quantidade razoável de dados para captar precisamente o risco de cada produtor.

Portanto, se adotar um modelo para análise de uma situação qualquer deve-se estar ciente de suas limitações. Infelizmente, a falta de um conjunto de dados com maior número de observações e uma maior quantidade de variáveis, impossibilitou uma mode-

lagem mais precisa. No entanto, de maneira alguma procurou-se encerrar com este trabalho a discussão sobre o assunto, mas incentivar uma pesquisa mais aprofundada.

REFERÊNCIAS

BAXTER, L.; COUTTS, S.; ROSS, G. Applications of linear models in motor insurance. In: INTERNACIONAL CONGRESS OF ACTUARIES, 21., 1980, Zurich. **Proceedings...**Zurich: s.ed., 1980 p.11–29.

BARNDORFF-NIELSEN, O. E. **Information and exponencial families in statistical theory**. New York: John Wiley and Sons, 1978. 248 p.

CORDEIRO, G. M. Modelos lineares generalizados. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 7., Campinas, 1986. **Minicurso...** Campinas: UNICAMP, 1986, 286p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. Modelos Lineares Generalizados. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 12., Santa Maria, 2007. **Minicurso**. Santa Maria: UFSM, 2007. 161p.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. New York: Cambridge University Press, 1997. 592 p.

EFRON, B. Bootstrap methods: another look at the jackknife. **Annals of Statistics**, Hayward, v.7, p.1-26, 1979.

EFRON, B.; TIBSHIRANI, R. **An Introduction to the bootstrap**. London; New York: Chapman and Hall, 1993. 436p.

PENG, F.; DEY, D. K.; GELFAND, A.E.Overdispersed Generalized Linear Models. **Statistical Planning and Inference**, Holanda, v.64, n.64, p.93-108, 1997.

FERREIRA, P. P. **Modelos de precificação e ruína para seguros de curto prazo**. Rio de Janeiro: FUNENSEG, 2005. 210p.

GANIO, L.M.; SCHAFER, D.W. Diagnostics for Overdispersion.**Journal of the American Statistical Association**, London, v.87, p.795-804, 1992.

HABERMAN† S.; RENSHAW A. E. Generalized linear model and actuarial science. **The Statistician**, London, v.45, n.4, p.407-436, Jul./Aug. 1996

HALCROW H. G.; Actuarial structures for crop insurance In: **Journal of Farm Economics**, Lancaster, v.31, n.3, p.418-443, Aug. 1949

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models an estimation. In: **SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA**, 13., 1998a, São Paulo, **Anais...**, São Paulo, 1998a. 73 p.

_____ Overdispersion: models an estimation. **Computational Statistics and Data Analysis**, New York, v. 27, p.151-170, 1998b.

JϕRGENSEN, B. **The theory of dispersion models**, London; New York: Chapman and Hall, 1997. 337p.

SMYTH, G.K.;JϕRGENSEN, B. Fitting Tweedie's compound model to insurance claims data In: **Journal ASTIN Bulletin**, Denmark, v.32, p.143-157

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**, London: Chapman and Hall, 1989. 511p.

MIRANDA, M. J.; GLAUBER J. W. Systemic risk, reinsurance, and the failure of crop insurance markets. **American Journal of Agricultural Economics**, Ames, v.79, n.1, p.206-215, Feb., 1997.

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

OZAKI, V. A. **Métodos atuariais aplicados à determinação da taxa de prêmios de contratos de seguro agrícola:** um estudo de caso. 2005. 324 p. (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2005.

_____ O papel do seguro na gestão do risco agrícola e os empecilhos para o seu desenvolvimento. **Revista Brasileira de Risco e Seguro**, Rio de Janeiro, v.2, n.4, p.75-92, out./mar. 2007.

_____ Em busca de um novo paradigma para o seguro rural no Brasil. **Revista Brasileira de Risco e Seguro**, Rio de Janeiro, v.46, n.1, p. 97-119, jan./mar., 2008.

OZAKI, V.A. ; SHIROTA, R. Um estudo da viabilidade de um programa de seguro agrícola baseado em um índice de produtividade regional em Castro (PR). **Revista de Economia e Sociologia Rural**, Rio de Janeiro, v.43, n.3, p.485-503, 2005.

PETERS, G. W.; SHEVCHENKO, P. V.; MARIO V. W. Model uncertainty in claims reserving within Tweedie's compound Poisson models In: **Journal ASTIN Bulletin**, Denmark, v.39, p.1-33

RENSHAW A. E.; HABERMAN† S. **Statistical analysis of life assurance lapses**. London: J.I.A., 1986., p.459-497.

VERBEEK, H. G. An Approach to the Analysis of claims experience in motor liability excess of loss reinsurance In: **Journal ASTIN Bulletin**, Denmark, v.6, p.195—202

WEDDERBURN, R. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. **Biometrika**, London, v.61, n.3, p.439–447, 1974.

WILLIANS, D. A. Extra-binomial variation in logistic linear models. **Applied Statistics**, London, v.31, p.144-148, 1982.

ANEXOS

ROTINA DO SOFTWARE R

```
################################BANCO DE DADOS UVA#################################

dados<-read.table('uva.txt',header=T)
attach(dados)

hist(sin)
plot(sin)

mod1<-glm(sin~prec+temp,family=poisson)

anova(mod1)
summary(mod1)


#Gráfico de resíduos
res<-residuals(mod1,type="deviance")
qqnorm(res,xlab="quantis amostrais",ylab="quantis teóricos")
qqline(res)
plot(fitted(mod1),res,,xlab="valores ajustados",ylab="resíduos")
abline(h=0,col="red")

#Modelo binomial negativo

library(MASS)

mod2<-glm.nb(sin~prec+temp)
a<-density(sin)
hist(sin,prob=T,xlab="Número de Sinistro",ylab="Densidade",main="")
lines(a,col='red',lwd=2)


#Gráfico de resíduos
res<-residuals(mod2,type="deviance")
a<-density(res)
par(mfrow=c(1,2))
hist(res)
lines(a,col='red',lwd=2)
qqnorm(res,xlab="quantis amostrais",ylab="quantis teóricos")
qqline(res)
plot(fitted(mod2),res)
abline(h=0)


=======================================FAMILIA TWEEDIE==============================

##################BANCO DE DADOS UVA##############################
```

```
require(DAAG)
require(tweedie)
dados1<-read.table('uva1.txt',header=T)
attach(dados1)

dados.Kel<-dados3[,1]

#Perfilada

Kel4<-tweedie.profile(dados.Kel~prec+temp, p.vec=seq(1.01,1.99,length=10),fit.glm=TRUE,
   do.plot=TRUE,method="interpolation", do.smooth=TRUE,do.ci=TRUE,phi.method= "mle")
Kel4$p.max


#Modelo ajustado

mod.ajust<-glm(mont~prec+temp,family=tweedie(var.power=1.57))
summary(mod.ajust)
anova(mod.ajust)

#Resíduos

res<-residuals(mod.ajust,type="deviance")
a<-density(res)
par(mfrow=c(1,2))
lines(a,col='red',lwd=2)
qqnorm(res,xlab="quantis teóricos",ylab="quantis amostrais")
qqline(res)
hist(res,xlab="sinistros",ylab="frequência",main=" Histograma dos resíduos")
```