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

Modelos para dados de contagem com superdispersão: uma aplicação em um experimento agronômico

Douglas Toledo Batista

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

Piracicaba
2015

Douglas Toledo Batista
Estatístico

Modelos para dados de contagem com superdispersão: uma aplicação em um experimento agronômico

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **IDEMAURO ANTONIO RODRIGUES DE LARA**

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

Piracicaba
2015

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

Batista, Douglas Toledo
Modelos para dados de contagem com superdispersão: uma aplicação em um experimento agronômico / Douglas Toledo Batista. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2015.
69 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura “Luiz de Queiroz”.

1. Dados discretos 2. MLG 3. Quase-verossimilhança 4. Medidas de ajuste 5. HNP I. Título

CDD 634.31
B288m

AGRADECIMENTOS

Meus agradecimentos vão a todos os colegas, amigos e familiares que acompanharam essa longa trajetória, contribuindo para concretização deste trabalho. Em especial:

Ao meu orientador Prof. Dr. Idemauro Antonio Rodrigues de Lara, pela disposição e competência na condução deste trabalho;

Aos professores do departamento de Matemática e Estatística da ESALQ-USP, pelos ensinamentos;

À minha mãe, pelo amor, dedicação, incentivo e presença constante em todas as etapas da minha vida;

À minha namorada Cristiane, pelo apoio e incentivo para a conclusão deste trabalho;

À minha turma de mestrado, em especial Daniel, Erasnilson, Fernando, Otávio, Rick, Simone, Valiana e Patrícia;

À Vanessa Vogit, pelo fornecimento dos dados;

À Comissão de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos;

Aos funcionários do departamento de Matemática e Estatística da ESALQ-USP, pela prontidão e atenção dispensada;

Finalmente, a todos que de alguma forma contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

RESUMO

Modelos para dados de contagem com superdispersão: uma aplicação em um experimento agronômico

O modelo de referência para dados de contagem é o modelo de Poisson. A principal característica do modelo de Poisson é a pressuposição de que a média e a variância são iguais. No entanto, essa relação de média-variância nem sempre ocorre em dados observacionais. Muitas vezes, a variância observada nos dados é maior do que a variância esperada, fenômeno este conhecido como superdispersão. O objetivo deste trabalho constitui-se na aplicação de modelos lineares generalizados, a fim de selecionar um modelo adequado para acomodar de forma satisfatória a superdispersão presente em dados de contagem. Os dados provêm de um experimento que objetivava avaliar e caracterizar os parâmetros envolvidos no florescimento de plantas adultas da laranjeira variedade "x11", enxertadas nos limoeiros das variedades "Cravo" e "Swingle". Primeiramente ajustou-se o modelo de Poisson com função de ligação canônica. Por meio da *deviance*, estatística X^2 de Pearson e do gráfico *half-normal plot* observou-se forte evidência de superdispersão. Utilizou-se, então, como modelos alternativos ao Poisson, os modelos Binomial Negativo e Quase-Poisson. Verificou-se que o modelo Quase-Poisson foi o que melhor se ajustou aos dados, permitindo fazer inferências mais precisas e interpretações práticas para os parâmetros do modelo.

Palavras-chave: Dados discretos; MLG; Quase-verossimilhança; Medidas de ajuste; HNP

ABSTRACT

Models for count data with overdispersion: application in an agronomic experiment

The reference model for count data is the Poisson model. The main feature of Poisson model is the assumption that mean and variance are equal. However, this mean-variance relationship rarely occurs in observational data. Often, the observed variance is greater than the expected variance, a phenomenon known as overdispersion. The aim of this work is the application of generalized linear models, in order to select an appropriated model to satisfactorily accommodate the overdispersion present in the data. The data come from an experiment that aimed to evaluate and characterize the parameters involved in the flowering of orange adult plants of the variety "x11" grafted on "Cravo" and "Swingle". First, the data were submitted to adjust by Poisson model with canonical link function. Using deviance, generalized Pearson chi-squared statistic and half-normal plots, it was possible to notice strong evidence of overdispersion. Thus, alternative models to Poisson were used such as the negative binomial and Quasi-Poisson models. The Quasi-Poisson model presented the best fit to the data, allowing more accurate inferences and practices interpretations for the parameters.

Keywords: Discrete data; GLM; Quasi-likelihood; Adjustment measures; HNP

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O Brasil é considerado o maior produtor mundial de laranja e no ano de 2013 sua produção foi de aproximadamente 18 milhões de toneladas. Em segundo e terceiro lugar encontram-se os Estados Unidos e a China, respectivamente, com uma produção de cerca de 8 milhões de toneladas cada (FAO, 2013). A citricultura brasileira está altamente concentrada na produção de laranja doce, correspondendo a aproximadamente 90% do total da produção de citrus do país, sendo que destes 90%, cerca de 80% são produzidos na região sudeste (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE, 2013).

As frutas cítricas tiveram origem no sudeste da Ásia e se espalharam durante a idade média para o sul da Europa e norte da África e, mais tarde, para os demais continentes. A produção dos citrinos (Laranja, Clementinas, Tangerinas, Limões, Limas, Toranja) tem despertado interesse mundial, pois os frutos apresentam características singulares como vida longa, aparência atraente e ótimas propriedades medicinais, o que facilita a exportação (SPIEGEL-ROY; GOLDSCHMIDT, 1996). Uma vez que a citricultura representa um importante componente da produção agrícola brasileira, cada vez mais estudos recebem incentivos no que se refere ao aperfeiçoamento e ampliação dos conhecimentos técnicos sobre a produção citrícola.

No presente trabalho apresenta-se uma aplicação relatada por Voigt (2013), que trata de um estudo sobre a laranjeira variedade “x11”, um mutante espontâneo da laranja doce que possui como principal característica o período juvenil curto, apresentando florescimento precoce a partir de um ou dois anos de cultivo. O mutante “x11” foi enxertado em duas variedades de citros, o limão “Cravo” e o citrumelo “Swingle”, de modo a verificar o comportamento da planta em relação à produção dos ramos de acordo com as classificações predefinidas (ramos uniflorais, multiflorais, sem flores e com flores abortadas), após as podas eventuais. A floração é uma etapa crucial para a produção de frutos, e a indução floral é provocada pela temperatura e pelas variações ambientais (SPIEGEL-ROY; GOLDSCHMIDT, 1996). Uma característica deste experimento é a natureza da variável mensurada: uma contagem.

Em experimentos na área agronômica, biológica e demais áreas afins, geralmente utilizam-se dados de contagem, especialmente em subáreas como entomologia, microbiologia, fitopatologia, dentre outras. Quando as observações são independentes, pode-se

utilizar inicialmente, o modelo de regressão de Poisson, que é um modelo linear generalizado (CAMERON; TRIVEDI, 2013). No modelo de Poisson a principal característica é que a média e a variância são iguais. No entanto, essa relação nem sempre ocorre, podendo a variância observada ser maior que a variância esperada e esse fenômeno é chamado de superdispersão (MCCULLAGH; NELDER, 1989).

O avanço em análises de regressão fornecidos pelos modelos lineares generalizados (MLGs) pode ser considerado um importante desenvolvimento estatístico dos últimos 40 anos. Essa classe de modelos incorpora a modelagem de dados normais e não normais, incluindo regressão linear múltipla, ANOVA (análise de variância), regressão logística, regressão de Poisson e modelos log-lineares para tabelas de contingência. Por esse motivo, os MLGs têm sido uma ferramenta essencial na análise de dados, em diferentes áreas do conhecimento.

Dessa forma, o presente trabalho visa descrever e aplicar algumas técnicas dos MLGs para dados de contagem do experimento relatado por Voigt (2013), com a finalidade de ajustar um modelo da produção do número de ramos de acordo com as classificações supracitadas para os porta-enxertos limão “Cravo” e citrumelo “Swingle”. Os objetivos específicos são:

i) Comparar alguns modelos aplicados aos problemas da superdispersão;

ii) Utilizar métodos gráficos de diagnóstico para avaliar a qualidade do ajuste do modelo, com ênfase do gráfico *half-normal plot*;

iii) Identificar qual porta-enxerto é mais eficiente no que diz respeito à produtividade, em termos de maior produção de ramos com flores.

2 REVISÃO BIBLIOGRÁFICA

O modelo de regressão linear clássico teve origem na astronomia, desenvolvido por Gauss no período de 1809 a 1821 (CORDEIRO; LIMA, 2006). Esse modelo é de grande importância para descrever uma série de fenômenos aleatórios, entretanto, se limitam à suposição de normalidade para a variável resposta, bem como homogeneidade de variância e independência dos erros. Quando essa suposição não é alcançada, algum tipo de transformação pode ser sugerida a fim de obter a normalidade, sendo a transformação de Box e Cox (1964) a mais conhecida. Com objetivo de expandir a distribuição da variável resposta, no início dos anos 70, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (MLGs), para os quais não há necessidade de aplicar qualquer tipo de transformação na variável resposta em busca de atingir as pressuposições de normalidade e variância constante (PAULA, 2013).

Os MLGs podem ser vistos como uma extensão dos modelos lineares clássicos, ou ainda, uma classe mais ampla que permite uma generalização dos modelos clássicos e uma ampliação da distribuição da variável resposta, desde que a mesma pertença à família exponencial de distribuições. Além disso, também permite uma padronização no algoritmo de estimação dos parâmetros do modelo. Assim, os modelos log-lineares para tabelas de contingência, regressão logística, modelo probito, regressão Poisson, modelos de delineamento com análise de variância (ANOVA), modelos de análise de sobrevivência, dentre outros, podem ser vistos como casos particulares de um MLG. Estes modelos têm determinadas propriedades similares (e.g. linearidade e os princípios de diagnóstico do modelo por meio da análise de resíduo).

Atualmente, dispõe-se de uma ampla literatura sobre os MLGs. Com o objetivo de apresentar a temática e explorar os meios pelos quais é possível obter um maior conhecimento a respeito dos MLGs, a seguir faz-se uma breve explanação sobre os trabalhos mais representativos disponíveis. O livro de McCullagh e Nelder (1989) é considerado a principal referência sobre o assunto, apresentando uma discussão vasta sobre o tema, expondo aplicações práticas e uma série de demonstrações analíticas.

Olsson (2002) e Dobson (2010) apresentam uma introdução aos MLGs, por meio de um conteúdo unificado teórico e prático. Temas como aplicações e o uso de modelos de regressão múltipla, ANOVA para dados contínuos, modelos log-lineares para dados de

contagem na forma de tabelas de contingência, regressão logística para dados binários e modelos para análise de sobrevivência fazem parte destas obras. Cordeiro (1986), Demétrio (2002), Cordeiro e Demétrio (2010) e Paula (2013) são referências na língua portuguesa, proporcionando uma leitura essencial para a introdução ao tema, apresentando aplicações em conjunto de dados de diversas áreas do conhecimento, como Agronomia, Ciências Atuarias, Biologia, Medicina, Pesca, Odontologia e Economia.

Para análise de dados categorizados pode-se citar Agresti (2002) como referência. Hosmer e Lemeshow (2004), por sua vez, apresentam um conteúdo amplo para análise de dados cuja variável resposta assume apenas dois valores possíveis (dicotômica), destacando o modelo de regressão logística, que permite incluir uma ou mais variáveis explicativas (ou covariável) na relação funcional, com interpretações relativamente simples. Vieira (1998) exibe um estudo para dados de proporção aplicado ao controle biológico, no qual seis modelos são utilizados, sendo eles Binomial, Binomial superdisperso, Binomial truncado, Binomial truncado superdisperso, mistura de Bernoulli-Binomial e mistura de Bernoulli-Binomial superdisperso. Questões como comparação de modelos e superdispersão fazem parte dessa obra.

No que diz respeito à análise dos dados, por meio de um MLG, frequentemente o pesquisador recorre ao uso de *software* estatístico. Diversos *softwares* estão disponíveis como SAS (`http://www.sas.com`), S-Plus (`http://www.insightful.com`), STATISTIC (`http://www.statsoft.com`), *R* (`http://www.r-project.org`), dentre outros. O *software* *R* se destaca por ser livre e permitir ao usuário o controle sobre a análise estatística. Dobson (2010) e Paula (2013) apresentam exemplos e rotinas do uso dessa ferramenta. Maiores detalhes podem ser encontradas em Faraway (2005).

Em experimentação agronômica é comum o pesquisador obter observações em escala discreta, na forma de contagem, sejam eles em testes de germinação, análise microbiológica de solos, avaliação entomológica, teste de toxicidade, dentre outros. Nestas situações, as contagens se referem ao número de eventos específicos que ocorrem em um intervalo e, por definição, consistem apenas em números inteiros positivos (KARAZSIA; DULMEN, 2008). O desenvolvimento de modelos para dados de contagem foi impulsionado pelos MLGs. O modelo de regressão Poisson é um caso especial, descrito pela primeira vez por Nelder e Wedderburn (1972) e detalhado em McCullagh e Nelder (1989). Winkelmann (2008) e Cameron e Trivedi (2013) são referências adicionais para modelos de contagem.

Tópicos como modelos de Poisson, violação da suposição de Poisson, estimação dos parâmetros, modelos de misturas, estimação por simulação, modelos inflacionários de zero, dados longitudinais, técnicas de diagnósticos, qualidade do ajuste, entre outros, são descritos por estes autores.

A principal característica do modelo de regressão Poisson é a de que a média e a variância são iguais. No entanto, essa relação nem sempre ocorre, podendo a variância observada ser maior do que a variância esperada e esse fenômeno é chamado de superdispersão, também conhecido como sobredispersão ou variação extra-Poisson. Neste caso, o uso da regressão Poisson não é adequado, pois o erro padrão obtido a partir do modelo de regressão Poisson será impreciso e pode ser gravemente subestimado, sendo, portanto, a interpretação e previsão do modelo fadada ao erro (HINDE; DEMÉTRIO, 1998a).

2.1 Modelos Lineares Generalizados

Os MLGs desenvolvidos por Nelder e Wedderburn (1972), vem ganhando cada vez mais espaço entre os pesquisadores, pois permitem interpretações práticas dos parâmetros, além da existência de todo um aparato computacional que viabiliza o uso dessa teoria. Em diversos estudos, sejam eles de natureza experimental ou observacional, o principal objetivo do pesquisador é analisar as influências que uma ou mais variáveis explicativas $(x_1, x_2, ..., x_p)$, exercem sobre a variável resposta de interesse (Y). As variáveis explicativas compõem a estrutura linear do modelo e estas podem ser qualitativas, quantitativas ou ambas. Já a variável resposta pode ser discreta, contínua e categórica, e estas assumem diferentes distribuições de probabilidade, tais como Normal, Normal Inversa, Gama, Binomial, Binomial Negativa e Poisson. Assim, para uma variável resposta contínua simétrica, uma escolha plausível é a distribuição Normal, enquanto que para dados contínuos assimétricos pode-se optar pelas distribuições Gama ou Normal Inversa. Para dados de proporção, a distribuição Binomial mostra-se aconselhável, enquanto que para dados de contagem pode-se utilizar a distribuição Poisson ou Binomial Negativa (MCCULLAGH; NELDER, 1989).

De acordo com Nelder e Wedderburn (1972), McCullagh e Nelder (1989) e Agresti (2002) os MLGs apresentam três componentes:

i) Componente aleatório: responsável por identificar a variável resposta e sua distribuição de probabilidade. Sejam Y_i, $i = 1, 2, ...n$, sendo n observações independentes da variável resposta com $\mathrm{E}(Y_i) = \mu_i$ e sua distribuição de probabilidade pertencente à

família exponencial, em que Y_i pode ser discreta ou contínua, com função de probabilidade ou função de densidade de probabilidade, respectivamente, descrita sob a forma:

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{1}$$

em que $a(.)$, $b(.)$ e $c(.)$ são funções específicas e $a(\phi) > 0$. Quando o parâmetro $a(\phi)$ for conhecido, a distribuição da variável aleatória Y_i (1) pertence à família exponencial na forma canônica, e θ_i é um parâmetro natural ou canônico. Usualmente $a(\phi) = \phi/w_i$, no qual w_i são pesos a *priori* (em geral $w_i = 1$) e ϕ representa um parâmetro de dispersão constante. A média e a variância de Y_i são dadas por:

$$\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i$$

$$\mathrm{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu_i).$$

Dessa forma $b'(\theta_i)$ e $b''(\theta_i)$ são as derivadas de 1^a e 2^a ordens em relação a θ_i, respectivamente. A função $V(\mu_i)$ depende apenas de μ_i e é conhecida como função de variância;

ii) Componente sistemático: responsável por especificar as variáveis explicativas $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i)^T$ com $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{ip})^T$ ao vetor $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)^T$, o preditor linear, por meio de um modelo linear dado por:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \quad j = 1, 2, ..., p \qquad \text{ou} \qquad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

sendo $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ um vetor de p parâmetros desconhecidos, e $\mathbf{X}$ é uma matriz de delineamento ou uma matriz de covariáveis de dimensões $n \times p$, no qual $\mathbf{x}_{i1} = 1$ para todo i;

iii) Função de ligação: responsável por associar o componente sistemático ao componente aleatório, sendo uma função monótona e diferenciável, dada por:

$$\eta_i = g(\mu_i) \quad \Rightarrow \quad \mu_i = g^{-1}(\eta_i).$$

Registra-se que a função de ligação que transforma a média no parâmetro natural, ou

seja, $g(\mu_i) = \theta_i$ é denominada como função de ligação canônica, e apresenta algumas vantagem teóricas e práticas ao modelo, como por exemplo, garante a existência de um conjunto de estatísticas suficientes [$\mathbf{X}^T\mathbf{Y}$ em notação matricial] para o vetor $\boldsymbol{\beta}$, simplificando o algorítimo de estimação e interpretação dos parâmetros. Entretanto, a função de ligação canônica não garante qualidade no ajuste do modelo (MYERS et al., 2010, cap.5).

Dessa forma, pode-se sumarizar o processo de escolha dos MLGs em três etapas:

i) Qual é a distribuição de probabilidade da variável resposta?

ii) Qual é a função de ligação mais adequada?

iii) Quais são as covariáveis que devem ser selecionadas para descrever o comportamento real dos dados?

2.1.1 Estimação dos parâmetros $\beta's$

Como já mencionado, em um MLG uma escolha fundamental é a definição da distribuição de probabilidade da variável resposta (componente aleatório), a matriz do modelo que contém as variáveis explicativas (componente sistemático) e, por último, a função de ligação. O método principal para estimar os $\beta's$ no MLG é o da máxima verossimilhança, pois possui propriedades excelentes, como consistência e eficiência assintótica (CORDEIRO; DEMÉTRIO, 2010). O logaritmo da função de verossimilhança de uma amostra aleatória da família (1) é:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} Li = \sum_{i=1}^{n} \ln f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i,\theta_i).$$

Uma propriedade importante da forma exponencial de distribuição é que seus elementos satisfazem às condições suficientes para assegurar o máximo do logaritmo da função de verossimilhança $L(\beta)$, seja dado unicamente pelo sistema de equações da função *score*:

$$U_j = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial Li}{\partial \beta_j} = 0$$

para todo j. E aplicando a regra da cadeia tem-se:

$$\frac{\partial Li}{\partial \beta_j} = \frac{\partial Li}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

e assim a função *score* será:

$$U_j = \sum_{i=1}^{n}\left[\frac{y_i - \mu_i}{a(\phi)}\frac{1}{V(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij}\right] \quad j = 1, ..., p.$$

As estimativas de máxima verosimilhança dos $\beta's$ são obtidas por meio do algoritmo numérico de *Newton-Raphson* (mínimos quadrados iterativos reponderados) e deve ser executado considerando a m-ésima iteração, até que um critério de convergência seja atingido. Um critério plausível de convergência pode ser:

$$\sum\nolimits_{j=1}^{p}\left(\frac{\beta_j^m - \beta_j^{(m+1)}}{\beta_j^m}\right) < \xi$$

assumindo para ξ um valor satisfatoriamente pequeno. Dessa forma o algoritmo é elaborado da seguinte forma (DEMÉTRIO, 2002):

1° Passo:

$$\eta_i{}^{(m)} = \sum_{j=1}^{p} x_{ij}\beta_j^{(m)}, \quad \mu_i = g^{-1}(\eta_i{}^{(m)});$$

2° Passo:

$$z_i^{(m)} = \eta_i + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)}); \quad \boldsymbol{Z}^{(m)} = \left[z_1^{(m)}, z_2^{(m)}, ..., z_n^{(m)}\right]^T$$

$$W_i^{(m)} = \frac{w_i}{V(\mu_i^{(m)})\left[g'(\mu_i^{(m)})\right]^2}; \quad \boldsymbol{W}^{(m)} = \text{diag}(W_i^{(m)});$$

3° Passo:

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{Z}^{(m)} \tag{2}$$

sendo $\boldsymbol{X}$ a matriz de planejamento, $\boldsymbol{W}$ a matriz de pesos e $\boldsymbol{Z}$ um vetor para a variável dependente ajustada;

4° Passo: Retornar ao 1° Passo com $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m+1)}$ e repetir o processo até convergência,

obtendo-se $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$.

O estimador de $\widehat{\boldsymbol{\beta}}$ tem distribuição assintótica dada por:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N_p(\mathbf{0}, (\boldsymbol{X^T W X})^{-1}),$$

portanto, assintoticamente, $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X^T W X})^{-1}$ (PAULA, 2013). Com base na matriz $\text{Var}(\widehat{\boldsymbol{\beta}})$, pode-se construir o intervalo de confiança (IC) para os parâmetros do modelo:

$$\text{IC}(\beta_j, \alpha\%) = \left(\widehat{\beta}_j \pm z_{\alpha/2}\sqrt{\text{Var}(\widehat{\beta}_{jj})}\right)$$

sendo, z o quantil da distribuição normal e $\text{Var}(\widehat{\beta}_{jj})$ são os elementos da diagonal da matriz $\text{Var}(\widehat{\boldsymbol{\beta}})$.

2.1.2 Função desvio e estatística X^2 de Pearson

Após a seleção das covariáveis do modelo faz-se necessário verificar a qualidade do ajuste. Segundo McCullagh e Nelder (1989), o ajuste de um modelo a um conjunto de dados pode ser considerado como uma forma de substituir um conjunto Y_i por um conjunto de μ_i, no qual o ideal é um modelo com um número relativamente pequeno de parâmetros. As medidas de discrepância para os MLGs que generalizam a soma de quadrados dos resíduos do modelo linear clássico são: estatística generalizada X^2 de Pearson e a função desvio (*deviance*). A estatística X^2 é dada por:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

sendo $V(\widehat{\mu}_i)$ a função de variância estimada do modelo sob pesquisa.

Nelder e Wedderburn (1972) desenvolveram a *deviance*, que foi definida como duas vezes a diferença entre o logaritmo da função de verossimilhança do modelo completo (ou saturado) $L(y, \phi)$, com n parâmetros, e o máximo do logaritmo da função de verossimilhança do modelo sob pesquisa, com p parâmetros $L(\widehat{\mu}, \phi)$. Assim, a expressão é dada por:

$$\frac{D\left(y;\widehat{\mu}\right)}{\phi}=\frac{1}{\phi}\sum_{i=1}^{n}2w_i\left\{y_i\left[\tilde{\theta}_i-\widehat{\theta}_i\right]-b(\tilde{\theta}_i)+b(\widehat{\theta}_i)\right\}=2\left(L(y,\phi)-L(\widehat{\mu},\phi)\right),$$

no qual $\tilde{\theta}_i=\tilde{\theta}_i(y_i)$ e $\widehat{\theta}_i=\widehat{\theta}_i(\mu_i)$ são as estimativas dos parâmetros canônicos do modelo completo e sob pesquisa, respectivamente, $D(y;\widehat{\mu})$ é a *deviance* calculada a partir das observações e das estimativas de suas respectivas médias e, $D^*(y;\widehat{\mu})=D(y;\widehat{\mu})/\phi$ é a *deviance* padronizada. Esta medida é utilizada para fazer inferência do modelo sob pesquisa. Para distribuições como Poisson e Binomial, a *deviance* e a *deviance* padronizada são idênticas, pois $\phi=1$. Assim, se a diferença entre as *deviances* do modelo completo e do modelo sob pesquisa for pequena, pode-se concluir que o modelo com o menor número de parâmetros é tão informativo quanto o modelo completo (MCCULLAGH; NELDER, 1989).

Quando Y_i tem distribuição normal, então $D^*(y;\widehat{\mu})$ e X^2/ϕ são semelhantes a soma de quadrados dos resíduos e seguem uma distribuição qui-quadrado exata ${\chi^2}_{n-p}$. Para outras distribuições tem-se resultados assintóticos, em alguns casos específicos a *deviance* não poderá ser utilizada como uma estatística de falta de ajuste, pois não se sabe a sua convergência (LEE; NELDER; PAWITAN, 2006).

Especificamente quando se tem dados de contagem, isto é, se $Y_i\sim\text{Poisson}(\mu_i)$ e $\mu_i\to\infty$ então $D(y;\widehat{\mu})\sim{\chi^2}_{n-p}$, dessa forma, o valor esperado desta variável é igual ao número de graus de liberdade. Assim, para um modelo bem ajustado espera-se que $D(y;\widehat{\mu})\simeq n-p$ (PAULA, 2013). Caso contrário, existem evidências de que o modelo é inadequado, e quando $D(y;\widehat{\mu})\gg n-p$ o pesquisador pode considerar a existência da superdispersão (HINDE; DEMÉTRIO, 1998b).

A *deviance* tem ainda a vantagem geral como uma medida de discrepância pois ela é aditiva para um conjunto de modelos encaixados (MCCULLAGH; NELDER, 1989).

2.2 Superdispersão para Dados de Contagem

Em MLGs quando a variância da variável resposta Y, (variância empírica ou amostrada) excede a variação nominal (variância esperada conforme o modelo probabilístico estabelecido), diz-se que pode existir o problema da superdispersão (MCCULLAGH; NELDER, 1989). Dobson (2010), por sua vez, considera superdispersão quando $\text{Var}(Y)>$

E(Y), e esta pode ocorrer devido à falta de independência entre as observações ou a ausência de covariáveis capazes de explicar a heterogeneidade entre as observações.

Ao contrário do que se postula, em dados de contagem, o fenômeno da superdispersão é muito comum em diversas áreas do conhecimento como Epidemiologia (LEE et al., 2012), Psicologia (GARDNER; MULVEY; SHAW, 1995), Ecologia (HOEF; BOVENG, 2007; RICHARDS, 2008), Entomologia (MORAL, 2013), dentre outras. Na área agronômica, por exemplo, diversidades ambientais como fertilidade do solo e densidade de pragas, podem afetar significativamente o experimento. Em tais circunstâncias há possibilidade de se obterem dados superdispersos (SURIYAGODA et al., 2012). McCullagh e Nelder (1989) salientam: "A superdispersão não é incomum na prática. Na verdade alguns sustentam que a superdispersão é normal na prática e a dispersão nominal é a exceção."

Cameron e Trivedi (1986), Hinde e Demétrio (1998a) exibem uma revisão de modelos superdispersos para dados de contagem e proporção, no qual tópicos como métodos de estimação, qualidade de ajuste e testes de hipóteses são explorados. Ademais, Hinde e Demétrio (1998a) descrevem os modelos para superdispersão em duas categorias:

i) Modelos que assumem uma forma mais geral para a função de variância, admitindo um parâmetro adicional. Os parâmetros podem ser estimados por diferentes métodos de estimação como Quase-verossimilhança, Pseudo-verossimilhança e Momentos;

ii) Modelos de dois estágios para a variável resposta. Assim, assume-se uma distribuição de probabilidade para o parâmetro do modelo a ser estimado. Modelos de probabilidade compostos, como Binomial Negativo e Poisson-normal, são utilizados com frequência para dados de contagem. Os parâmetros são estimados pelo método da máxima verossimilhança.

Uma classe de modelos para a superdispersão são os chamados de Quase-verossimilhança. Estes modelos foram propostos por Wedderburn (1974) e consistem em uma técnica de ajuste na qual não se precisa especificar a distribuição da variável resposta. Para tal propósito, uma relação entre a média e variância é utilizada, ou seja, a função de variância é uma função da média. Esta função inclui um fator multiplicativo conhecido como parâmetro de superdispersão, que é estimado a partir dos dados.

A estimação dos modelos de Quase-verossimilhança surge como uma alternativa ao modelo de Poisson, na qual a relação média-variância é dada pela forma $V(\mu_i) = \phi\mu_i$,

para alguma constante ϕ. Quando $\phi = 1$ representa o modelo de Poisson, e quando $\phi < 1$ demonstra a presença de subdispersão, ou seja, uma "pequena" variação e esse fenômeno teoricamente é possível, porém incomum (OLSSON, 2002). Por outro lado, quando $\phi > 1$ representa a superdispersão. O parâmetro ϕ pode ser estimado por:

$$\widehat{\phi} = \frac{X^2}{n-p}, \tag{3}$$

em que X^2 é a estimativa da estatística de Pearson ajustada para o modelo, com n observações e p parâmetros. McCullagh e Nelder (1989) apresentam uma ampla discussão sobre a vantagem em utilizar a estatística X^2 de Pearson, ao invés de utilizar a *deviance* para estimar o parâmetro de dispersão. O modelo de Quase-verossimilhança leva as mesmas estimativas do modelo de Poisson, no entanto, os erros padrões obtidos serão multiplicados por $\sqrt{\widehat{\phi}}$, corrigindo a variabilidade das estimativas. Na equação (3), é imediato observar que quando X^2 é maior do que os graus de liberdade do modelo $(n-p)$, há indícios de superdispersão.

Williams (1982), propõe uma modificação no algoritmo de mínimos quadrados iterativos reponderados de um modelo de regressão logística, para acomodar a superdispersão de forma satisfatória para dados de proporção. Posteriormente Breslow (1984) propôs uma modificação neste algoritmo para incorporar o efeito da superdispersão de forma satisfatória para dados de contagem. Este método consiste em fazer uma série de ajustes para o modelo de Poisson com pesos $w_i = (1 + \widehat{\phi}\widehat{\mu}_i)^{-1}$, em que $\widehat{\phi}$ é a estimativa do parâmetro de superdispersão e $\widehat{\mu}_i$ são os valores ajustados da i-ésima observação. Assim o algoritmo é dado por:

i) Ajustar o modelo de Poisson com pesos $w_i = 1$. Se a estimativa da *deviance* for próxima aos graus de liberdade do modelo, o processo é finalizado, sendo a variação residual explicada adequadamente. Caso contrário deve-se prosseguir ao passo ii;

ii) Calcular a estimativa de $\widehat{\phi}$;

iii) Definir os novos pesos wi e retornar ao passo i.

Para maiores detalhes ver (BRESLOW, 1984).

Ridout, Demétrio e Hinde (1998) apresentam o índice de superdispersão

$$\text{IS} = \frac{\text{variância} - \text{média}}{\text{média}},$$

e essa, medida auxilia a identificar a superdispersão presente nos dados, sendo que quando IS > 0 indica a presença de superdispersão, IS < 0 subdispersão e IS $= 0$ indica que a média e a variância são iguais.

2.2.1 Causas e implicações da superdispersão

Olsson (2002) sugere que antes de qualquer tentativa de modelar a superdispersão, deve-se examinar todas as possíveis razões para o ajuste insatisfatório do modelo. Algumas das possíveis causas podem ser:

i) Escolha errada do preditor linear, como termos de interação ou termos não lineares foram omitidas do preditor linear ($\boldsymbol{\eta}$);

ii) Presença de *outliers* nos dados;

iii) Escolha inadequada da função de ligação;

iv) Insuficiência de dados, podendo os pressupostos da teoria assintótica não serem satisfeitos, motivando um modelo insatisfatório.

Com o auxílio da análise de resíduos descrita em McCullagh e Nelder (1989, Cap.12), é possível diagnosticar se a superdispersão é provocada pelas situações i, ii e iii descritas. A existência de alguma anormalidade nos dados, como um erro de digitação ou de medida, contagem incorreta da variável resposta, dentre outros, também podem ser responsáveis pela superdispersão, de modo que, com a simples exclusão desse(s) *outlier*(es), pode-se contornar o problema da superdispersão. No entanto, existem outras fontes de variação, como as descritas em Hinde e Demétrio (1998b) que podem acarretar na superdispersão. Algumas possibilidades são:

i) Variabilidade do material experimental, que produzem um componente adicional na variância e que não são quantificadas no modelo pressuposto;

ii) Omissão de covariáveis capazes de explicar a heterogeneidade;

iii) Correlação entre as respostas individuais;

iv) Amostragem por conglomerados.

Também é possível obter dados superdispersos devido ao excesso de zeros. Ridout, Demétrio e Hinde (1998) apresentam uma revisão de literatura e discutem uma metodologia comum para dados de contagem inflacionados de zeros, com uma aplicação em horticultura. Em Lambert (1992) é proposto um modelo de mistura de distribuições Bernoulli-Poisson com intuito de modelar de forma satisfatória o excesso de zeros na contagem de defeitos de fabricação de placas de circuito impresso.

A surperdispersão acarreta em desvios padrões imprecisos. Dessa forma, como consequência, as estimativas dos erros padrões estarão subestimadas, causando assim um comprometimento das inferências do modelo, aumentando a probabilidade de cometer o erro tipo I em teste de hipóteses: tornando-se possível obter interpretações incorretas e algumas predições inadequadas (HINDE; DEMÉTRIO, 1998a, 1998b).

2.3 Modelos para dados de contagem

2.3.1 Modelo Poisson

O modelo de referência para dados de contagem é o modelo de regressão Poisson. A distribuição de Poisson foi desenvolvida por Siméon-Denis Poisson, em 1837 e publicada em seu trabalho *Recheches sur la probabilité des jugements em matières criminelles et matière civile* (POISSON, 1842). A distribuição de Poisson fornece um modelo de referência para diversos fenômenos aleatórios associados a processo de contagem. Os valores assumidos pela variável aleatória Y são números inteiros positivos e, assim, qualquer fenômeno aleatório proveniente de contagem é um candidato para a modelagem, assumindo distribuição de Poisson (*e.g.* números de sementes germinadas, números de insetos mortos devido a aplicação de um inseticida, número de telefonemas em um central de atendimento ao consumidor, número de defeitos por unidade de algum material etc.).

A distribuição de Poisson é uma distribuição de probabilidade discreta, com apenas um parâmetro μ, $\mu > 0$, que corresponde a uma taxa média de ocorrência em um determinado intervalo. Se Y é uma variável aleatória com distribuição de Poisson, $Y \sim \text{Poisson}(\mu)$, sua função de probabilidade é dada por:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, ...$$

Sabe-se que a distribuição de Poisson pertence à família exponencial canônica, ou seja:

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\right\}.$$

Assim, $\theta_i = \ln(\mu_i) \Rightarrow \mu_i = \exp(\theta_i)$ é o parâmetro canônico e $b(\theta_i) = \exp(\theta_i) = \mu_i$ é uma função monótona e derivável. Então verifica-se que $\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i$ e $\mathrm{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu_i) = \mu_i$, sendo $V(\mu_i)$ a função de variância e $a(\phi) = 1$. Para a função de ligação canônica $\ln(\mu_i)$ tem-se que: $\eta_i = \theta_i \Leftrightarrow g(\mu_i) = \ln(\mu_i)$.

Para o modelo de Poisson a *deviance* e a *deviance* padronizada são dadas por:

$$D(y; \widehat{\mu}) = D^{*}(y; \widehat{\mu}) = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right].$$

Pode-se observar que $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i)$, porém, conforme já salientado, nem sempre isso ocorre. Quando $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i)$ pode-se caracterizar como um problema de superdispersão. Nesse contexto, a distribuição Binomial Negativa surge como uma alternativa à distribuição de Poisson (DOBSON, 2010).

2.3.2 Modelo Binomial Negativo

Se há indícios de superdispersão nos dados implicando na inadequação do modelo de Poisson, uma alternativa é a utilização do modelo Binomial Negativo (BN), pois permite uma maior flexibilidade na modelagem da variância (CAMERON; TRIVEDI, 2013).

A distribuição BN ou de Pascal é uma combinação entre a distribuição Gamma e a Poisson. Segundo Winkelmann (2008), uma forma de obter a distribuição BN é considerar Y uma variável aleatória cuja distribuição condicionada a uma variável aleatória Z, tenha um distribuição de Poisson com média z, ou melhor,

$$Y|Z \sim \text{Poisson}(z), \quad i.e, \quad f(y/z) = \frac{e^{-z}z^{y}}{y!} \quad y = 0, 1, 2, ...$$

Supondo que Z seja uma variável aleatória não observável (ou latente) com distribuição

Gamma e a $\mathrm{E}(Z) = \mu$ e $\mathrm{Var}(Z) = \dfrac{\mu^2}{k}$, ou seja,

$$Z \sim \text{Gamma}(\mu, k), \quad i.e, \quad f(z; k, \mu) = \frac{\left(\frac{zk}{\mu}\right)^k \exp\left(\frac{-zk}{\mu}\right)}{\Gamma(k)} \frac{1}{z}, \quad z > 0$$

em que $\Gamma(.)$ é a função Gamma definida por: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. Desta forma, a distribuição conjunta de Y e Z é dada por:

$$f(y, z) = f(z; k, \mu) f(z) = \frac{\left(\frac{zk}{\mu}\right)^k \exp\left(\frac{-zk}{\mu}\right)}{\Gamma(k)} \frac{1}{z} \left(\frac{\exp(-z) z^y}{y!}\right)$$

Assim, pode-se obter a distribuição marginal de Y:

$$f(y; k, \mu) = \int_0^\infty \frac{\left(\frac{zk}{\mu}\right)^k \exp\left(\frac{-zk}{\mu}\right)}{\Gamma(k)} \frac{1}{z} \left(\frac{\exp(-z) z^y}{y!}\right) dz$$

$$= \frac{\left(\frac{k}{\mu}\right)^k}{y! \Gamma(k)} \int_o^\infty \exp\left\{-z\left(1 + \frac{k}{\mu}\right)\right\} z^{y+k-1} dz$$

levando a:

$$f(y; k, \mu) = \frac{\Gamma(y + k)}{y! \Gamma(k)} \left(\frac{k}{\mu + k}\right)^k \left(\frac{\mu}{\mu + k}\right)^y, \qquad y = 0, 1, 2, ... \tag{4}$$

Maiores detalhes são dados em Winkelmann (2008). Se uma variável aleatória discreta Y_i, $i = 1, 2, ..., n$ tem distribuição BN denotada por $Y_i \sim \text{BN}(\mu_i, k)$, sua função de probabilidade é dada pela Equação (4). A distribuição BN tem dois parâmetros $k > 0$ e μ, no qual a $\mathrm{E}(Y_i) = \mu_i$ e $\mathrm{Var}(Y_i) = \mu_i + \mu_i^2/k$, sendo k^{-1} o parâmetro de dispersão (AGRESTI, 2002). Nota-se que a função de variância é quadrática em vez de linear como no modelo de Poisson.

Quando $k^{-1} \to 0$, $Var \to \mu$, a distribuição Binomial Negativa converge para distribuição de Poisson (WINKELMANN, 2008; CAMERON; TRIVEDI, 2013). Normalmente k^{-1} é desconhecido e quando estimado ajuda a resumir a influência da superdispersão (AGRESTI, 2002).

A distribuição Binomial Negativa, quando k é fixo, pertence à família expo-

nencial canônica, isto é:

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left\{y_i \ln\left(\frac{\mu_i}{\mu_i+k}\right) - k\ln\left(\frac{k}{\mu_i+k}\right) + \ln\left(\frac{\Gamma(k+y_i)}{\Gamma(k)y_i!}\right)\right\}$$

logo, assumindo que $\theta_i = \ln({}^{\mu_i}/_{\mu_i+k})$ é o parâmetro canônico e $b(\theta_i) = -k\ln({}^{k}/_{\mu_i+k}) = -k\ln(1-e^{\theta_i})$ é uma função monótona e derivável, portanto, verifica-se que $\mathrm{E}(Y_i) = b'(\theta_i) = ({}^{e^{\theta_i}k}/_{1-e^{\theta_i}})$, e $\mathrm{Var}(Y_i) = a(\phi)b''(\theta_i) = a(\phi)V(\mu_i) = \left(\frac{{\mu_i}^2}{k} + \mu_i\right)$ e $V(\mu_i)$ é a função de variância e $a(\phi) = 1$.

A principal vantagem da distribuição BN em relação à Poisson, é o parâmetro adicional (k) que introduz consideravelmente uma maior flexibilidade para modelar a superdispersão por meio da função de variância. Assim, a distribuição BN não apresenta a mesma restrição da distribuição de Poisson, $\mathrm{Var}(Y) = \mathrm{E}(Y)$ (WINKELMANN, 2008). No ajuste do modelo BN o processo iterativo descrito na seção 2.1.1 alterna a estimação de $\boldsymbol{\beta}$ e k a cada passo. Assim, os valores de $\boldsymbol{\beta}$ e k são atualizados até a convergência do algoritmo. Uma vez estimado o parâmetro k, este é considerado uma constante para todas as observações.

Para o modelo BN a *deviance* padronizada é dada por:

$$D^*(y;\widehat{\mu}) = 2\sum_{i=1}^{n} w_i\left[y_i\ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i+k)\ln\left(\frac{y_i+k}{\widehat{\mu}_i+k}\right)\right],$$

e

$$w_i = \frac{\mu_i^2}{\mu_i^2 k^{-1} + \mu_i}.$$

2.3.3 Modelo de Quase-verossimilhança

Para os modelos descritos anteriormente, assume-se uma distribuição de probabilidade pertencente à família exponencial. No entanto, em alguns casos não é adequado adotar uma distribuição de probabilidade a priori para a variável resposta, pois os dados podem não seguir determinadas distribuições de probabilidade. Nesses casos, Wedderburn (1974) propôs o método de estimação de Quase-verossimilhança (QV), no qual não há necessidade de especificar uma distribuição para a variável resposta, mas sim uma relação entre a média e a variância. Assim, foi proposta uma função de estimação que, sob hipó-

teses gerais, leva a estimadores consistentes e assintoticamente normais dos parâmetros do modelo. Esses modelos representam uma extensão mais flexível dos MLGs, pois apresentam apenas um componente sistemático e uma função de ligação que relaciona a média ao preditor linear.

Desse modo, Y_i é a variável de interesse, em que se assume $\mathrm{E}(Y_i) = \mu_i$ e uma variância definida por $\mathrm{Var}(Y_i) = \phi V(\mu_i)$, no qual a função de variância $V(\mu_i) = \mu_i$ é uma função conhecida da média e $\phi > 0$ é o parâmetro de dispersão constante. O logaritmo da função de Quase-verossimilhança Q é definido como:

$$Q(y;\mu) = \frac{1}{\phi}\int_{y}^{\mu} \frac{y-t}{V(t)}dt$$

em que $V(t)$ é uma função positiva conhecida. Logo, esta variável aleatória Q apresenta a forma:

$$Q(y_i;\mu_i) = \frac{1}{\phi}\int_{y_i}^{\mu_i} \frac{y_i-t}{t}dt = \frac{1}{\phi}\left[y_i \ln\left(\frac{\mu_i}{y_i}\right) + y_i - \mu_i\right].$$

Deste modo, segundo Paula (2013), tem-se que $Q(y_i;\mu_i)$ é proporcional ao logaritmo da função de verossimilhança do modelo de Poisson quando $\phi = 1$ e $y > 0$. Segundo McCullagh e Nelder (1989) um estimador consistente para ϕ é $\widehat{\phi} = X^2/(n-p)$, no qual X^2 é a estatística de Pearson e $(n-p)$ é o grau de liberdade do modelo maximal. Este modelo é computacionalmente mais simples, já que não é necessário estimar o parâmetro de dispersão por meio do algoritmo iterativo.

O método de estimação dos parâmetros por Quase-verossimilhança com suposição Poisson (aqui chamado de modelo Quase-Poisson) é análogo ao descrito para o modelo de Poisson, porém a $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ é corrigida, pois $w_i = \widehat{\mu}_i \big/ \widehat{\phi}$. Segundo Hoef e Boveng (2007) ao comparar a matriz $\boldsymbol{W}$ do modelo Quase-Poisson percebe-se que os pesos são diretamente proporcionais com a média, já o modelo BN a média tem uma relação quadrática, desse modo, valores médios baixos indicam pouco peso na matriz $\boldsymbol{W}$, e altos valores médios apresentam um peso maior, sendo essa a principal diferença entre esses dois métodos.

Para o modelo Quase-Poisson a função Quase-desvio não padronizada é dada por:

$$D(y;\widehat{\mu}_i) = -2\phi^2 Q(\widehat{\mu};y) = 2\sum_{i=1}^{n}\int_{\widehat{\mu}_i}^{y_i} \frac{y_i-t}{V(t)}dt = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right\}.$$

2.4 Técnicas de diagnóstico

As técnicas clássicas de diagnóstico baseiam-se na análise de resíduos e podem ser formais (como testes) ou informais (como gráficos). As técnicas formais para modelos clássicos consistem em testes de hipóteses para verificar a normalidade dos resíduos (SHAPIRO; WILK; CHEN, 1968), a homocedasticidade (homogeneidade de variância dos resíduos) (BARTLETT, 1937) e independência dos resíduos (DURBIN; WATSON, 1950). Essas técnicas são de grande importância para a validação de um modelo, pois se as pressuposições do modelo forem violadas, os resultados não serão confiáveis, assim a previsão do modelo estará comprometida (MYERS et al., 2010).

A análise de resíduos para os MLGs é semelhante às usadas em modelos clássicos, porém com algumas adaptações. Para verificar a pressuposição de linearidade para o modelo clássico utiliza-se os vetores $\mathbf{Y}$ e $\widehat{\boldsymbol{\mu}}$, enquanto no MLGs utiliza-se a variável dependente ajustada $\mathbf{z}$ e o preditor linear $\widehat{\boldsymbol{\eta}}$. A variância residual é substituída por uma estimativa de ϕ e a matriz de projeção $\mathbf{H}$ ou matriz "hat", é definida por: $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{\frac{1}{2}}$ (MCCULLAGH; NELDER, 1989).

Na classe dos MLGs pode-se citar basicamente 5 tipos de resíduos:

i) Resíduos ordinários $r_i = (y_i - \widehat{\mu}_i)$, em que y_i representa a variável reposta e $\widehat{\mu}_i$ é sua estimativa correspondente;

ii) Resíduo de Pearson generalizado

$$r_{Pi} = (y_i - \widehat{\mu}_i)\sqrt{\frac{w_i}{V(\widehat{\mu}_i)}}$$

no qual $V(\widehat{\mu})$ é a função de variância e w_i é um peso a priori;

iii) Resíduo de Pearson generalizado estudentizados

$$rpe_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\phi}V(\widehat{\mu})(1-h_i)}},$$

sendo $\widehat{\phi}$ uma estimativa consistente do parâmetro ϕ, h_i são os elementos da diagonal da matriz de projeção, $\mathbf{H}$. Esta matriz também mede a influência, em unidades estudentizadas, do vetor $\boldsymbol{Y}$ sobre $\widehat{\boldsymbol{\mu}}$ (MCCULLAGH; NELDER, 1989);

iv. Componentes de *Deviance* residual, definida como a raiz quadrada de cada elemento da *deviance* com o sinal do resíduo:

$$\mathrm{dr}_i = \sqrt{\mathrm{di}} \times \mathrm{sinal}(y_i - \widehat{\mu}_i)$$

em que, $\mathrm{di} = \left\{ y_i \left[\tilde{\theta}_i - \widehat{\theta}_i \right] - b(\tilde{\theta}_i) + b(\widehat{\theta}_i) \right\}$;

v. *Deviance* estudentizada

$$de_i = \frac{\sqrt{di} \times \mathrm{sinal}(y_i - \mu_i)}{\sqrt{\widehat{\phi}(1 - h_i)}}.$$

Todas essas medidas de discrepância são de grande importância para mensurar a diferença entre o valor observado e o valor estimado pelo modelo. McCullagh e Nelder (1989, Cap.12) apresentam uma arguição referente ao(s) ponto(s) discrepante(s), no qual a simples exclusão desse(s) ponto(s) pode representar uma melhoria para a qualidade do ajuste. Todas essas medidas auxiliam o pesquisador na escolha do "melhor" modelo.

2.4.1 Métodos gráficos

As técnicas gráficas mais utilizadas para os MLGs são:

i) Gráfico dos resíduos contra valores ajustados: pode mostrar a existência de valores discrepantes como também heterogeneidade de variância. O padrão do gráfico deve ser uma distribuição dos resíduos em torno de zero com amplitude constante;

ii) Gráfico dos resíduos contra variáveis explicativas: pode mostrar se existe uma relação sistemática (relação não linear, não normalidade, heterocedasticidade etc) entre os resíduos e uma variável explicativa. O padrão do gráfico deve ser uma distribuição dos resíduos em torno de zero com amplitude constante;

iii) Gráfico dos resíduos contra a ordem das observações: pode auxiliar na detecção de alguma variável altamente correlacionada com o sequência de tempo (ou espaço) que as observações foram coletadas;

iv) Gráfico da variável dependente ajustada contra preditor linear: ajuda a verificar a adequação da função de ligação e o padrão esperado é uma tendência linear, indicando

a adequação da função de ligação;

v) Gráfico dos resíduos absolutos contra valores ajustados: serve para identificar se a função de variância adotada é adequada, o comportamento esperado dos pontos é uma distribuição aleatória em torno de zero e amplitude constante;

vi) Gráfico normal e meio normal de probabilidades (*normal plot e half-normal plot*): são úteis para verificar se a função de variância foi corretamente especificada, e detectar a presença de *outlier*. O comportamento esperado dos resíduos para um modelo adequado é aproximadamente uma reta.

Para maior detalhamento quanto à análise gráfica pode-se ver (MCCULLAGH; NELDER, 1989; DEMÉTRIO, 2002; PAULA, 2013). Paula (2013) apresenta rotinas computacionais com o auxilio do *software* R, para a construção dos gráficos descritos, pelos itens i a vi (exceto o gráfico *half-normal plot*).

Hinde e Demétrio (1998a) destacam o uso do gráfico *half-normal plot* (hnp) com envelope simulado para verificar a qualidade do ajuste do modelo, especialmente quando os dados apresentam superdispersão. Vieira (1998) utiliza o hnp para comparar diferentes modelos para dados de proporção com superdispersão, utilizando o hnp como uma ferramenta de diagnóstico e decisão na escolha do "melhor" modelo com o auxílio do *software* GLIM.

O hnp é obtido plotando-se os valores absolutos ordenados de uma medida de diagnóstico adequada (Resíduo de Pearson, *deviance* padronizada, entre outras) contra os correspondentes valores esperados das estatísticas de ordem, em valor absoluto, da distribuição meio-normal dada por:

$$\Phi^{-1}\left[\frac{(i+n-\frac{1}{8})}{2n+\frac{1}{2}}\right]$$

em que Φ^{-1} é a função acumulada da distribuição normal padrão, sendo $i = 1, .., n$, sendo n a dimensão da amostra.

Para facilitar a análise gráfica, Atkinson (1985) propôs a construção de um "envelope" simulado, com objetivo de diminuir a subjetividade desta análise. Dessa forma, para um modelo adequado espera-se que os valores observados se distribuam satisfatoriamente dentro do "envelope". Assim é possível identificar se a distribuição do erro foi corretamente especificada e detectar a presença de *outlier* (HINDE; DEMÉTRIO, 1998b).

Passos para a construção de um *half-normal plot*:

i. Ajustar o modelo e calcular o resíduo pertinente representado por r_i, em valor absoluto e colocá-los em ordem crescente;

ii. Simular 99 amostras para a variável resposta, utilizando a mesma matriz das variáveis regressoras;

iii. Retornar aos modelos ajustados e para cada amostra simulada calcular os novos $r_{j(i)}$, em valores absolutos, $j = 1, ..., 99$, $i = 1, ..., n$ e dispor esses valores em ordem crescente;

iv. Para cada modelo ajustado calcular os percentis 5%, 50% e 95%;

v. Plotar os valores desses percentis r_i observados contra as estatísticas esperadas da distribuição meio-normal.

Moral (2013) implementou a função **hnp()** no *software* R, permitindo o uso do hnp para diversas funções de probabilidade, diferentes funções de ligação e para diferentes tipos de resíduos.

2.4.2 Técnica formal para adequacidade da função de ligação

Segundo McCullagh e Nelder (1989) uma técnica formal para examinar adequabilidade da função de ligação é adicionar ao modelo ajustado o preditor linear ao quadrado ($\boldsymbol{\eta}^2$), como uma covariável extra e analisar a redução na *deviance*, que é equivalente a realizar o teste da razão de verossimilhança (diferença entre as *deviances*). Assim tem-se:

$$\xi_{rv} = \left(D(y; \widehat{\mu})_{f2} - D(y; \widehat{\mu})_{f1} \right)$$

em que a $D(y; \widehat{\mu})_{f1}$ é a *deviance* do modelo (sem o preditor linear adicionado no modelo) encaixado em um modelo maior com a *deviance* $D(y; \widehat{\mu})_{f2}$ (com $\boldsymbol{\eta}^2$ adicionado no modelo). Sob hipótese nula assintoticamente temos que $\xi_{rv} \sim {\chi^2}_{f2-f1}$, no qual $f1$ e $f2$ são os graus de liberdade associados aos modelos. Dessa forma, se a redução for expressiva há indícios que a função de ligação é inadequada.

2.4.3 Técnica formal para a função de variância

Segundo McCullagh e Nelder (1989) uma técnica formal para verificar se a função de variância é adequada, é assumir que a $V(\zeta) = \mu^{\zeta}$ em que ζ é uma constante qualquer. O próximo passo é ajustar os modelos para diferentes valores de ζ e observar o comportamento da *deviance*. Para essa comparação é necessário aplicar os conceitos de Quase-verrossimilhança estendida, discutida em McCullagh e Nelder (1989, Cap.9), pois permite a comparação de diferentes funções de variância.

3 MATERIAL

Os dados cedidos para a realização deste trabalho são provenientes de um estudo experimental desenvolvido na área de concentração "Biologia na Agricultura e no Ambiente". Esse experimento foi realizado no ano de 2011 e teve como produto final a dissertação intitulada "Caracterização fenotípica e avaliação da expressão de genes envolvidos na indução e no florescimento da laranjeira "x11" ", defendida por Vanessa Voigt, pelo Programa de Pós-graduação em Ciências do Centro de Energia Nuclear na Agricultura-CENA\USP. O objeto de estudo do experimento é a laranjeira variedade "x11", que é um mutante espontâneo da laranja doce, que possui como principal característica o período juvenil curto, apresentando florescimento precoce a partir de um ou dois anos de cultivo.

Ainda, segundo Voigt (2013), a variedade "x11" dispõe de outras características de interesse, tais como: (i) capacidade de florescer várias vezes em um mesmo ano, após o procedimento de podas, (ii) desenvolvimento de uma flor na porção terminal de ramos em crescimento, (iii) plantas de baixo porte e ramos curtos e (iv) frutos com seis sementes em média. Outras espécies do gênero *Citrus* são plantas perenes com ciclo vegetativo longo e levam muito tempo para a análise do fenótipo resultante, o que representa problemas no que diz respeito à execução de programas de melhoramento convencional. Dessa forma, a variedade "x11" representa uma alternativa para estudos de investigação genética, pois apresenta ciclo juvenil curto.

Um dos objetivos do trabalho original visava avaliar e caracterizar os parâmetros envolvidos no florescimento de plantas adultas da laranjeira variedade "x11", enxertadas nos limoeiros porta-enxertos das variedades "Cravo" (*Citrus limonia* Osbeck) e "Swingle" (*Citrus paradisi* Macf. x *Poncirus trifoliata* (L.) Raf.), quando submetidas a podas realizadas nas quatro estações do ano. Os porta-enxertos são muito utilizados para o cultivo de citros, pois influenciam na produtividade do enxertado, podendo induzir variações no crescimento, na qualidade e produção dos frutos, na absorção de nutrientes, e conferir tolerância a determinados intempéries. O porta-enxerto limoeiro "Cravo" é o mais utilizado no Brasil devido às suas qualidades, dentre elas, rápido crescimento, precocidade, alta produção e maior tolerância à seca. O porta-enxerto "Swingle", por sua vez, é o segundo mais utilizado, apresentando produção de frutos de maior qualidade, tolerância ao frio e maior resistência a alguns fitopatógenos (VOIGT, 2013).

O estudo realizado por Voigt (2013) utilizou dezesseis plantas adultas de

laranjeira “x11”, sendo nove delas enxertadas em limoeiro “Cravo” e sete enxertadas em citrumeleiro “Swingle”. Todas as plantas apresentavam aproximadamente três anos de idade e foram conduzidas em vasos de 20 L, contendo uma mistura de substrato comercial e solo, com fertirrigação em sistema automático. As plantas foram incubadas em estufa sob condições naturais no Centro APTA Citros Sylvio Moreira/IAC. As podas não severas (aproximadamente 1 cm de comprimento) foram realizadas nos ramos resultantes do último surto de crescimento, ou seja, aproximadamente com um ano de idade ou menos, em todos os ramos da planta, de forma a induzir novas brotações e florescimentos na laranjeira “x11”. As podas se iniciaram no outono e primavera, e verão e inverno no ano de 2012. Semanalmente após cada poda, foram avaliadas as variáveis relacionadas ao crescimento e desenvolvimento das brotações das gemas axilares, desde o início das brotações até o florescimento completo.

Para efeito de ilustração, consideram-se neste trabalho, os dados referentes ao período da primavera, no qual a natureza da variável resposta mensurada foi obtida na forma de contagem, segundo a classificação de ramos de cada uma das plantas, sendo esta dividida em quatro categorias mutuamente exclusivas, conforme abaixo:

i) Número de ramos multiflorais (com flores terminais e laterais);

ii) Número de ramos sem flor (vegetativos);

iii) Número de ramos com flores abortadas (com flores terminais não desenvolvidas).

iv) Número de ramos uniflorais (com flores terminais);

As classificações i, ii, iii e iv serão denominadas a partir de agora como multifloral, sem flor, abortada e unifloral respectivamente. Na Figura 1 encontram-se as imagens ilustrativas para todas as classificações de ramos.

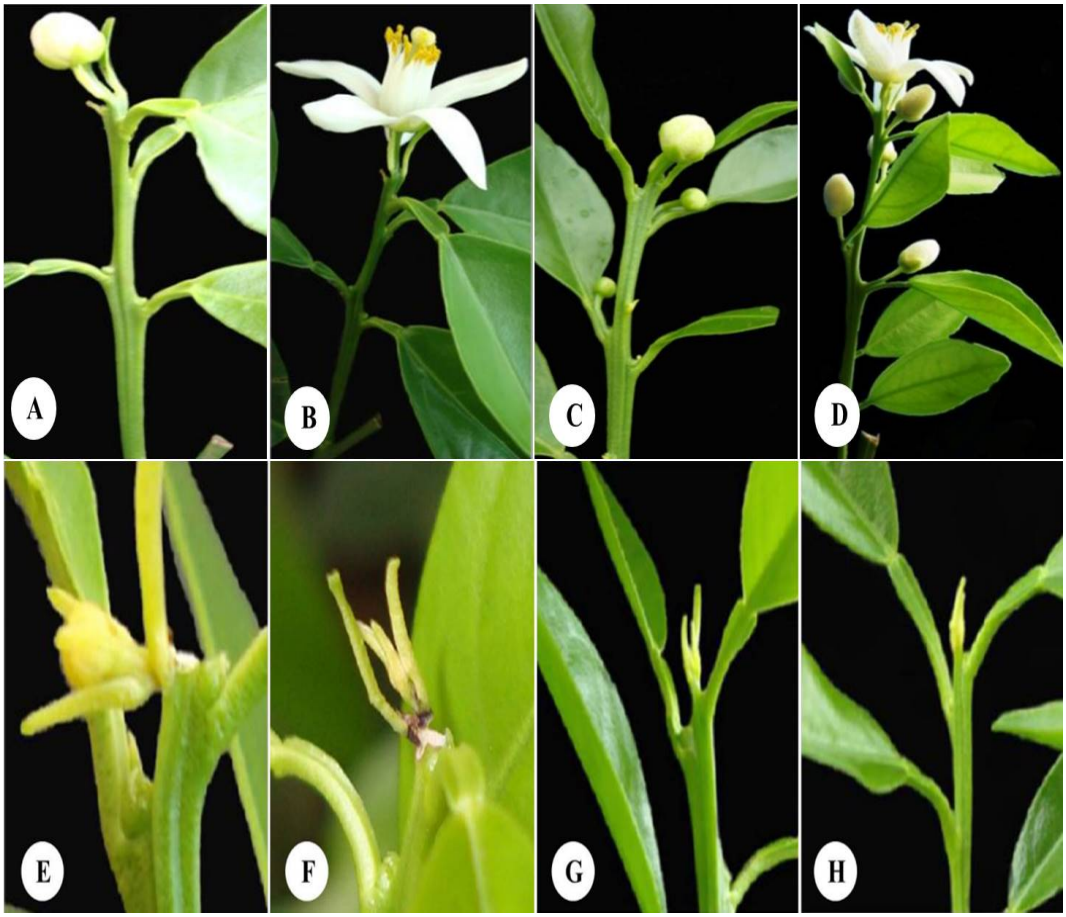


Figura 1 - A e B ramos uniflorais com flores fechadas e abertas, respectivamente; C e D ramos multiflorais com flores fechadas e abertas, respectivamente; E e F ramos com flores abortadas; G e H ramos sem flores

Fonte: Adaptado de Voigt (2013)

4 MÉTODOS

Para estabelecer a notação considere a variável resposta, Y_{ijk}, como o número de ramos da i-ésima planta, na j-ésima classificação de ramos e no k-ésimo porta enxerto, em que $i = 1, 2, ..., 16$, $j = 1, 2, 3, 4$ (multifloral, sem flor, abortada e unifloral) e $k = 1, 2$ (porta-enxerto "Cravo" e "Swingle"). A estrutura do componente sistemático refere-se a um delineamento inteiramente casualizado. Inicialmente considerou-se o modelo em que a variável resposta Y_{ijk} tem distribuição de Poisson com função de ligação logarítmica. Para a situação experimental descrita e, considerando os objetivos do estudo, tem-se como estrutura para o preditor linear os modelos:

$$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j + \beta_{2k}\text{porta-enxerto}_k. \tag{5}$$

Como as covariáveis são fatores, fixou-se como categoria de referência no processo de estimação o porta enxerto limão "Cravo", com a classificação de ramos multifloral. Dessa forma, o vetor de parâmetros estimado neste modelo é:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_{12} \\ \widehat{\beta}_{13} \\ \widehat{\beta}_{14} \\ \widehat{\beta}_{22} \end{pmatrix} \begin{cases} \text{casela de referência} \\ \text{efeito da classificação sem flor no porta enxerto limão "Cravo"} \\ \text{efeito da classificação flor abortada no porta enxerto limão "Cravo"} \\ \text{efeito da classificação unifloral no porta enxerto limão "Cravo"} \\ \text{efeito do porta enxerto citrumelo "Swingle"} \end{cases}$$

Por meio do modelo 5 pode-se estimar o número médio de ramos de acordo com cada classificação e efeito de porta-enxerto, tem-se então que:

$$\widehat{\mu}_{jk} = \exp(\widehat{\eta}_{jk}) = \exp(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j + \widehat{\beta}_{2k}\text{porta-enxerto}_k)$$

no qual, tem-se as restrições $\widehat{\beta}_{11} = \widehat{\beta}_{21} = 0$.

Outra opção de modelo a ser testado não considerará o efeito de porta-enxerto:

$$\eta_j = \ln(\mu_j) = \beta_o + \beta_{1j}\text{classificação}_j \tag{6}$$

no qual fixou-se como categoria de referência no processo de estimação a classificação de

ramos multifloral. Dessa forma, o vetor de parâmetros estimado neste modelo é:

$$\widehat{\boldsymbol{\beta}} = \left(\begin{array}{c} \widehat{\beta}_0 \\ \widehat{\beta}_{12} \\ \widehat{\beta}_{13} \\ \widehat{\beta}_{14} \end{array} \right) \left\{ \begin{array}{l} \text{casela de referência} \\ \text{efeito da classificação sem flor} \\ \text{efeito da classificação flor abortada} \\ \text{efeito da classificação unifloral} \end{array} \right.$$

Por meio do modelo 6 pode-se estimar o número médio de ramos de acordo com cada classificação, tem-se então que:

$$\widehat{\mu}_j = \exp(\widehat{\eta}_j) = \exp(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j)$$

no qual, tem-se a restrição $\beta_{11} = 0$.

Outra opção de modelo a ser testado é considerando-se apenas o efeito de porta-enxerto:

$$\eta_k = \ln(\mu_k) = \beta_0 + \beta_{2k}\text{porta-enxerto}_k \tag{7}$$

no qual fixou-se como categoria de referência no processo de estimação o porta-enxerto limão “Cravo”. Dessa forma, o vetor de parâmetros estimados neste modelo é:

$$\widehat{\boldsymbol{\beta}} = \left(\begin{array}{c} \widehat{\beta}_0 \\ \widehat{\beta}_{22} \end{array} \right) \left\{ \begin{array}{l} \text{casela de referência} \\ \text{efeito de porta enxerto “Swingle”} \end{array} \right.$$

Por meio do modelo 7 pode-se estimar o número médio de ramos de acordo com cada porta-enxerto, tem-se então que:

$$\widehat{\mu}_k = \exp(\widehat{\eta}_k) = \exp(\widehat{\beta}_0 + \widehat{\beta}_{2k}\text{porta-enxerto}_k)$$

no qual, tem-se a restrição $\beta_{21} = 0$.

O ajuste de um MLG para os dados pode ser avaliado por meio da *deviance*. O modelo descrito em (5), representa o efeito da classificação de ramos e porta-enxerto, no qual tem-se o número médio de ramos de acordo com cada classificação e porta-enxerto. O modelo descrito em (6), por sua vez, representa o efeito das classificações de ramos, assim tem-se o número médio de ramos para cada classificação. Finalmente, no modelo descrito em (7), tem-se o efeito de porta-enxerto, no qual tem-se o número médio de ramos para

cada porta-enxerto.

O modelo de Poisson requer algumas condições, como a $\text{Var}(Y_{jk}) = \text{E}(Y_{jk}) = \mu_{jk}$ além da condição de independência entre as observações. Porém, conforme já citado, quando assumida a distribuição de Poisson para a variável resposta, pode ocorrer uma variância maior do que a média, além daquela esperada pelo modelo. Esse fato pode ser indicativo de superdispersão e os motivos para esse fenômeno foram discutidos na Seção 2.2. Se for constatada a presença de superdispersão, faz-se necessário o uso de modelos que levem em conta esta variação extra-Poisson. Sendo assim, as estruturas dos preditores lineares (5), (6) e (7) serão consideradas, porém por meio dos modelos:

i) Modelo BN para a variável resposta;

ii) Modelo Quase-Poisson,

cujos procedimentos bem como vantagens para os casos de superdispersão foram discutidas nas seções 2.3.2 e 2.3.3.

4.1 Seleção dos modelos

Para comparar os ajustes dos modelos com diferentes preditores lineares e a mesma função de ligação, usa-se o teste da razão de verossimilhança (diferença entre as *deviances*). Para ϕ conhecido, tem-se que a estatística do teste:

$$\xi_{RV} = \frac{1}{\phi}\left((D(y;\widehat{\mu})_{f2} - D(y;\widehat{\mu})_{f1}\right),$$

em que a $D(y;\widehat{\mu})_{f1}$ é a *deviance* de um modelo encaixado em um modelo maior com a *deviance* $D(y;\widehat{\mu})_{f2}$. Sob a hipótese nula, assintoticamente tem-se que $\xi_{RV} \sim {\chi^2}_{f_2-f_1}$, no qual f_1 e f_2 são os graus de liberdade associados aos modelos. Sendo assim, é possível verificar se a inclusão de uma covariável é significativa ou não para o modelo.

Para o modelo Quase-Poisson ϕ é desconhecido, porém estimado a partir dos dados. Assim, para comparar os modelos encaixados utiliza-se a seguinte estatística:

$$F = \frac{D(y;\widehat{\mu})_{f_2} - D(y;\widehat{\mu})_{f_1}}{\widehat{\phi}(f_2 - f_1)}$$

que é distribuída assintoticamente como $F_{(f_2-f_1,n-p)}$ (JØRGENSEN, 2002).

As análises foram desenvolvidas com auxílio do *software* R (R Core Team, 2014). O modelo BN e análise de diagnóstico, foram desenvolvidos por intermédio de funções disponíveis nos pacotes MASS (VENABLES; RIPLEY, 2002) e hnp (MORAL; HINDE; DEMETRIO, 2014), respectivamente. A qualidade do ajuste do modelo foi avaliada mediante a *deviance* e estatística X^2 de Pearson. A comparação da qualidade de ajuste dos modelos com diferentes distribuições de probabilidade foi realizada por meio do *half-normal plot*. O nível de significância adotado foi $\alpha = 0,05$.

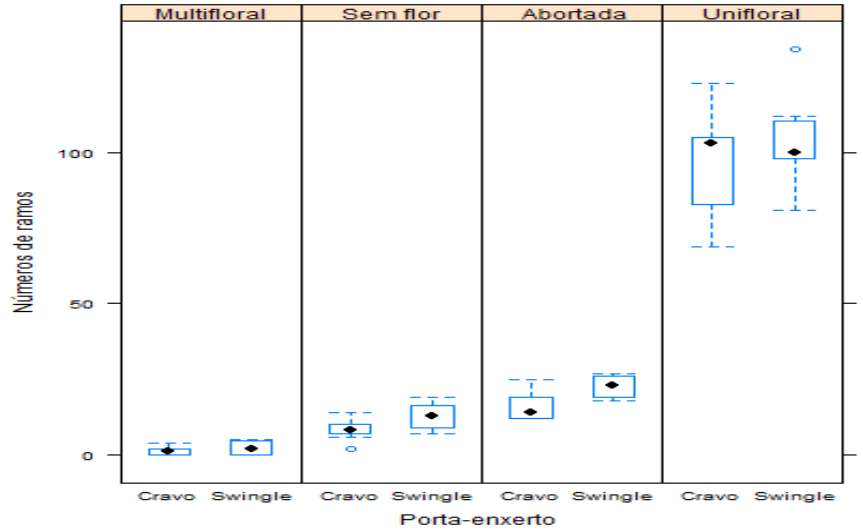
5 RESULTADOS E DISCUSSÃO

Primeiramente, uma análise exploratória foi realizada a fim de conhecer o comportamento dos dados. Algumas medidas descritivas para uma averiguação inicial, estão disponíveis na Tabela 1. Os resultados preliminares mostram que existe uma relação média-variância, ou seja, a medida que a média aumenta a variância também aumenta, e além disso, é possível observar que a variância é maior do que a média.

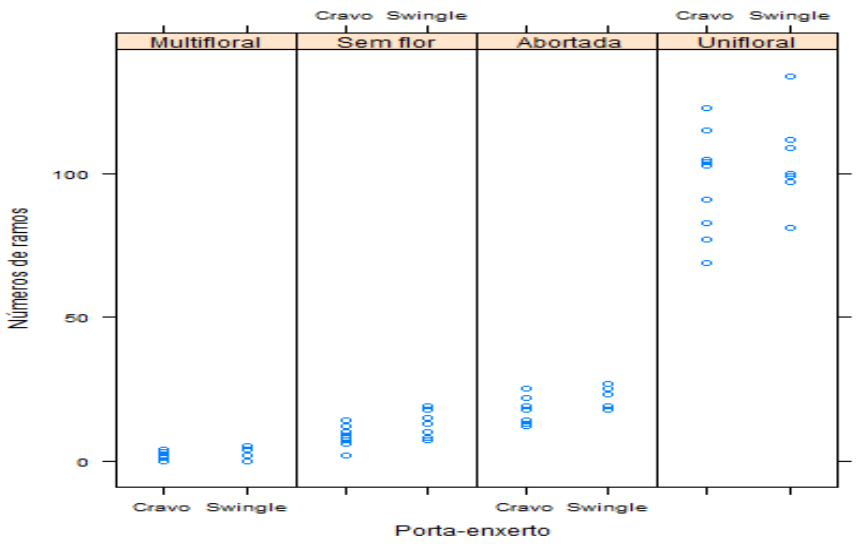
Tabela 1 - Medidas descritivas da contagem do número de ramos em relação às classificações e ao efeito de porta-enxertos para a estação primavera

Classificação dos ramos	Porta-enxerto limão "Cravo"				
	Amplitude	Amplitude interquartil	Média	Variância	Índice de superdispersão
Multifloral	4,00	2,00	1,33	2,25	0,69
Sem flor	12,00	3,00	8,33	12,25	0,47
Flor abortada	13,00	7,00	16,33	23,75	0,45
Unifloral	54,00	22,00	96,67	320,50	2,32
Total	123,00	32,25	30,67	1604,46	51,32
Classificação dos ramos	Porta-enxerto citrumelo "Swingle"				
	Amplitude	Amplitude interquartil	Média	Variância	Índice de superdispersão
Multifloral	5,00	4,50	2,29	5,57	1,44
Sem flor	12,00	7,50	12,86	22,48	0,75
Flor abortada	9,00	7,00	22,57	15,29	-0,32
Unifloral	53,00	12,50	104,57	267,62	1,56
Total	134,00	34,00	35,57	1768,25	48,71

De modo geral, o porta-enxerto citrumelo "Swingle" apresentou maiores médias em relação ao porta-enxerto limão "Cravo". Por meio da Figura 2 nota-se uma predominância de ramos uniflorais, para ambos os porta-enxertos na estação primavera.



(a)



(b)

Figura 2 - *Boxplot* (a) e gráfico de pontos (b) referente ao porta-enxerto e classificação de ramos para a estação primavera

A seguir, foram ajustados os modelos propostos na seção 4, começando pelo modelo de Poisson com função de ligação canônica. A Tabela 2 apresenta as *deviances*, graus de liberdade (g.l) e o nível descritivo da estatística χ^2. Sendo assim, pela análise das *deviances*, que coincide com o teste da razão de verossimilhanças para os modelos encaixados, apresentados na seção 4.1, é possível verificar que houve efeito de porta-enxerto

e classificação de ramos, bem como da interação, embora a mesma não seja de interesse prático. No modelo de Poisson tem-se que $D(y;\widehat{\mu}) = D^*(y;\widehat{\mu}) = 125,86$ e $X^2 = 116,84$ (com 59 graus de liberdade), portanto pode-se testar as hipóteses:

$$\begin{cases} H_0 : \text{O modelo de Poisson apresenta um ajuste satisfátorio;} \\ H_1 : \text{O modelo de Poisson não apresenta um ajuste satisfátorio,} \end{cases}$$

verificando-se, para ambas as medidas de qualidade de ajuste a um nível descritivo menor do que 0,001, indicando um ajuste insatisfatório do modelo de Poisson.

Tabela 2 - Análise da *deviance* para o modelo de Poisson

Modelos Poisson	*deviances*	g.l	p-valor
$\eta_k = \ln(\mu_k) = \beta_0 + \beta_{2k}\text{porta-enxerto}_k$	2.822,89	62	–
$\eta_j = \ln(\mu_j) = \beta_o + \beta_{1j}\text{classificação}_j$	137,35	60	<0,001
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j + \beta_{2k}\text{porta-enxerto}_k$	125,86	59	<0,001
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j * \beta_{2k}\text{porta-enxerto}_k$	117,18	56	0,03

Na Tabela 3 apresentam-se as estimativas dos parâmetros para o modelo de Poisson. Todas as classificações de ramos e os porta-enxertos foram altamente significativas. No entanto, observa-se que a *deviance* é muito maior do que os graus de liberdade e, além disso, $\widehat{\phi} = 1,98$, indicando a existência do fenômeno da superdispersão. Por meio do gráfico *half-normal plot*, de fato, verifica-se que praticamente todos os pontos estão fora do envelope simulado, com nível de confiança igual a 95%, evidenciando que o modelo de Poisson não se ajustou de forma satisfatória Figura 3(a).

Tabela 3 - Estimativas e erros padrões dos parâmetros do modelo de Poisson para a estação primavera

Parâmetro	Estimativa	Erro - Padrão	p-valor
β_o (multifloral"Cravo")	0,49	0,19	<0,001
β_{12} (sem flor "Cravo")	1,77	0,20	<0,001
β_{13} (abortada "Cravo")	2,39	0,20	<0,001
β_{14} (unifloral"Cravo")	4,05	0,19	<0,001
β_{22} ("Swingle")	0,15	0,04	<0,001

Para acomodar a superdispersão de forma satisfatória, outros modelos foram ajustados, tais como o modelo BN e o modelo Quase-Poisson.

Na Tabela 4 encontram-se os valores, 2×(log-verossimilhança maximizada) 2(log L), graus de liberdade (g.l), e o nível descritivo da estatística χ^2, para os modelos BN. É possível verificar que houve efeito de porta-enxerto e classificações de ramos e o mesmo não ocorre com a interação.

Tabela 4 - Análise do logaritmo da função de verossimilhança maximizada para o modelo Binomial Negativo

Modelos BN	2(log L)	g.l	p-valor
$\eta_k = \ln(\mu_k) = \beta_0 + \beta_{2k}\text{porta-enxerto}_k$	-563,24	62	–
$\eta_j = \ln(\mu_j) = \beta_o + \beta_{1j}\text{classificação}_j$	-391,87	60	<0,001
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j + \beta_{2k}\text{porta-enxerto}_k$	-382,51	59	<0,001
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j * \beta_{2k}\text{porta-enxerto}_k$	-376,92	56	0,13

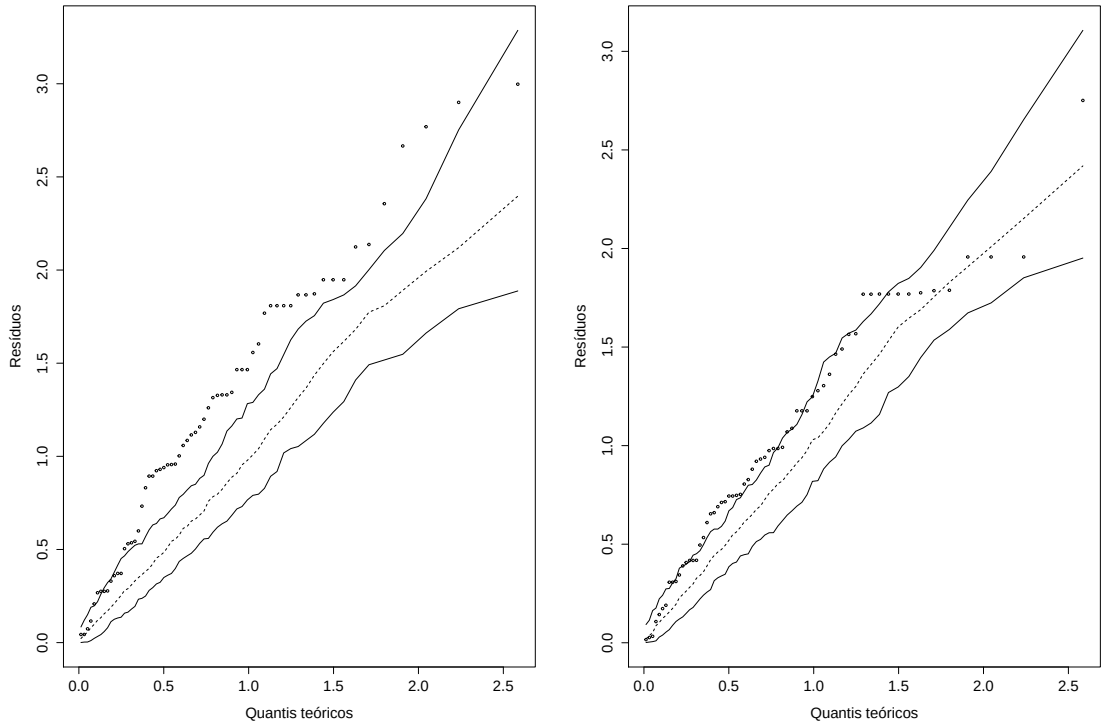
Os coeficientes estimados para o modelo BN, selecionado, encontram-se na Tabela 5. A estimativa para o parâmetro de dispersão foi de $\widehat{k} = 48,8$ com erro padrão de $23,7$. Obteve-se que $D^*(y;\widehat{\mu}) = 85,29$ e $X^2 = 76,22$ (com 59 graus de liberdade), portanto pode-se testar as hipóteses:

$$\begin{cases} H_0 : \text{O modelo BN apresenta um ajuste satisfátorio;} \\ H_1 : \text{O modelo BN não apresenta um ajuste satisfátorio,} \end{cases}$$

obtendo-se para estas medidas de qualidade de ajuste a um nível descritivo de p-valor = $0,01$ e p-valor = $0,06$, respectivamente. Considerando um nível de significância de 5%, e utilizando a *deviance* como medida de falta de ajuste, conclui-se que o modelo BN apresenta um ajuste insatisfatório. No entanto, pela estatística de Pearson, conclui-se que o modelo BN apresenta um ajuste adequado. Contudo, ao observar o gráfico *half-normal plot* Figura 3(b) verifica-se que o modelo BN não apresentou um ajuste adequado, uma vez que a maioria dos pontos estão fora da do envelope simulado, considerando um nível de confiança igual a 95%.

Tabela 5 - Estimativas e erros padrões dos parâmetros do modelo BN para a estação primavera

Parâmetro	Estimativa	Erro - Padrão	p-valor
β_o (multifloral "Cravo")	0,46	0,20	0,017
β_{12} (sem flor "Cravo")	1,77	0,21	<0,001
β_{13} (abortada "Cravo")	2,39	0,20	<0,001
β_{14} (unifloral "Cravo")	4,05	0,20	<0,001
β_{22}("Swingle")	0,21	0,07	0,002



(a) (b)

Figura 3 - *Half-normal plot* para o modelo de Poisson (a) e modelo BN (b)

Por último, foi utilizado o modelo Quase-Poisson com função de variância dada por $V(\mu_{ijk}) = \phi\mu_{ijk}$. A Tabela 6 apresenta as *deviances*, graus de liberdade (g.l) e o nível descritivo da estatística F. Sendo assim, é possível verificar que houve efeito de porta-enxerto e classificação de ramos e o mesmo não ocorre com a interação. Dessa forma a $D^*(y;\widehat{\mu}) = D(y;\widehat{\mu})/\widehat{\phi} = 125,86/1,98 = 63,56$ e $X^2/\widehat{\phi} = 116,84/1,98 = 59,01$ (com 59

graus de liberdade), portanto pode-se testar as hipóteses:

$$\begin{cases} H_0 : \text{O modelo Quase-Poisson apresenta um ajuste satisfátorio;} \\ H_1 : \text{O modelo Quase-Poisson não apresenta um ajuste satisfátorio,} \end{cases}$$

constatando-se, para ambas medidas de qualidade de ajuste, a um nível descritivo de p-valor=0,32 e p-valor=0,47, indicando um ajuste adequado ao nível de 5% de significância. Segundo Paula (2013) essas medidas devem ser olhadas apenas de forma descritiva já que no modelo de Quase-verossimilhança a distribuição da variável resposta é em geral desconhecida.

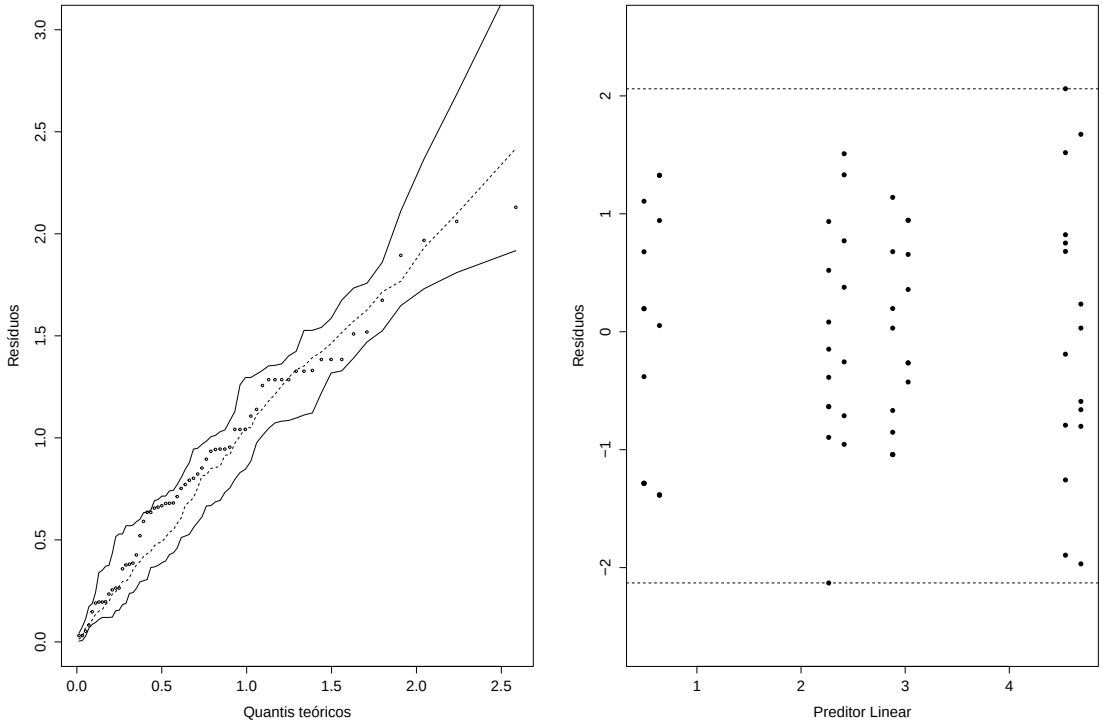
Tabela 6 - Análise da *deviance* para o modelo de Quase-Poisson

Modelos Quase-Poisson	*deviance* padronizada	g.l	p-valor
$\eta_k = \ln(\mu_k) = \beta_0 + \beta_{2k}\text{porta-enxerto}_k$	1.425,70	62	–
$\eta_j = \ln(\mu_j) = \beta_o + \beta_{1j}\text{classificação}_j$	69,36	60	<0,001
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j + \beta_{2k}\text{porta-enxerto}_k$	63,56	59	0,01
$\eta_{jk} = \ln(\mu_{jk}) = \beta_o + \beta_{1j}\text{classificação}_j * \beta_{2k}\text{porta-enxerto}_k$	59,18	56	0,22

Os erros padrões já multiplicados por $\sqrt{\widehat{\phi}}$ e as estimativas dos parâmetros são apresentados na Tabela 7. Para o modelo Quase-Poisson, observa-se por meio do gráfico *half-normal plot* Figura 4(a) que os pontos se acomodam dentro do envelope simulado. Além disso, verifica-se que os resíduos se distribuem satisfatoriamente entre -2,13 e 2,06 Figura 4(b). Estes indícios mostram que este modelo apresenta então, um melhor ajuste do que os demais. O mesmo resultado foi encontrado por Hoef e Boveng (2007), ao comparar a Quase-Poisson *versus* BN para dados de contagem com superdispersão.

Tabela 7 - Estimativas e erros padrões dos parâmetros do modelo Quase-Poisson para a estação primavera

Parâmetro	Estimativa	Erro - Padrão	p-valor
β_o (multifloral “Cravo”)	0,49	0,27	0,071
β_{12} (sem flor “Cravo”)	1,77	0,29	<0,001
β_{13} (abortada “Cravo”)	2,39	0,28	<0,001
β_{14} (unifloral “Cravo”)	4,05	0,27	<0,001
β_{22} (“Swingle”)	0,15	0,06	0,019



(a) (b)

Figura 4 - *Half-normal plot* para o modelo Quase-Poisson (a) e gráfico dos resíduos contra o preditor linear (b)

Comparando os valores estimados pelas Tabela 3, 5 e 7 observa-se que não há grandes alterações em termos das estimativas pontuais dos parâmetros dos modelos, com exceção dos erros padrões que são distintos de acordo com cada modelo. Assim, as estimativas dos erros padrões do modelo de Poisson e BN são subestimadas e portanto suas estimativas podem não ser suficientemente confiáveis, podendo afetar as inferências do modelo, como por exemplo os intervalos de confiança.

Levando em consideração os critérios adotados neste trabalho para verificar a qualidade de ajuste dos modelos, o modelo Quase-Poisson foi o mais adequado para os dados. Dessa forma, pode-se iniciar as interpretações dos parâmetros. A classificação de ramos multiflorais e o porta-enxerto limão "Cravo" foi tomada como categoria de referência no processo de estimação. Verifica-se, que na primavera, a tendência é a produção de ramos unifloral, sendo seguida pela ocorrência de ramos com flores abortadas e sem flores e, por último, a produção de ramos multifloral. Em contrapartida, espera-se um aumento na for-

mação de ramos quando o porta-enxerto é o citrumelo “Swingle”, para todas as classificações de ramos.

Por meio dos resultados da Tabela 7, é possível obter os intervalos de confiança para os parâmetros do modelo Quase-Poisson. Na Tabela 8 encontra-se os intervalos com 95% de confiança.

Tabela 8 - Estimativas dos intervalos de confiança, para os parâmetros do modelo Quase-Poisson para a estação primavera

Parâmetros	Estimativa	Intervalo de confiança de 95%	
		Limite inferior	Limite superior
β_o (multifloral “Cravo”)	0,49	-0,08	0,98
β_{12} (sem flor “Cravo”)	1,77	1,24	2,38
β_{13} (abortada “Cravo”)	2,39	1,88	2,98
β_{14} (unifloral “Cravo”)	4,05	3,56	4,62
β_{22} (“Swingle”)	0,15	0,03	0,27

Para comparar o número médio de flores referente ao tipo de porta-enxerto, realiza-se uma razão de médias, tal que:

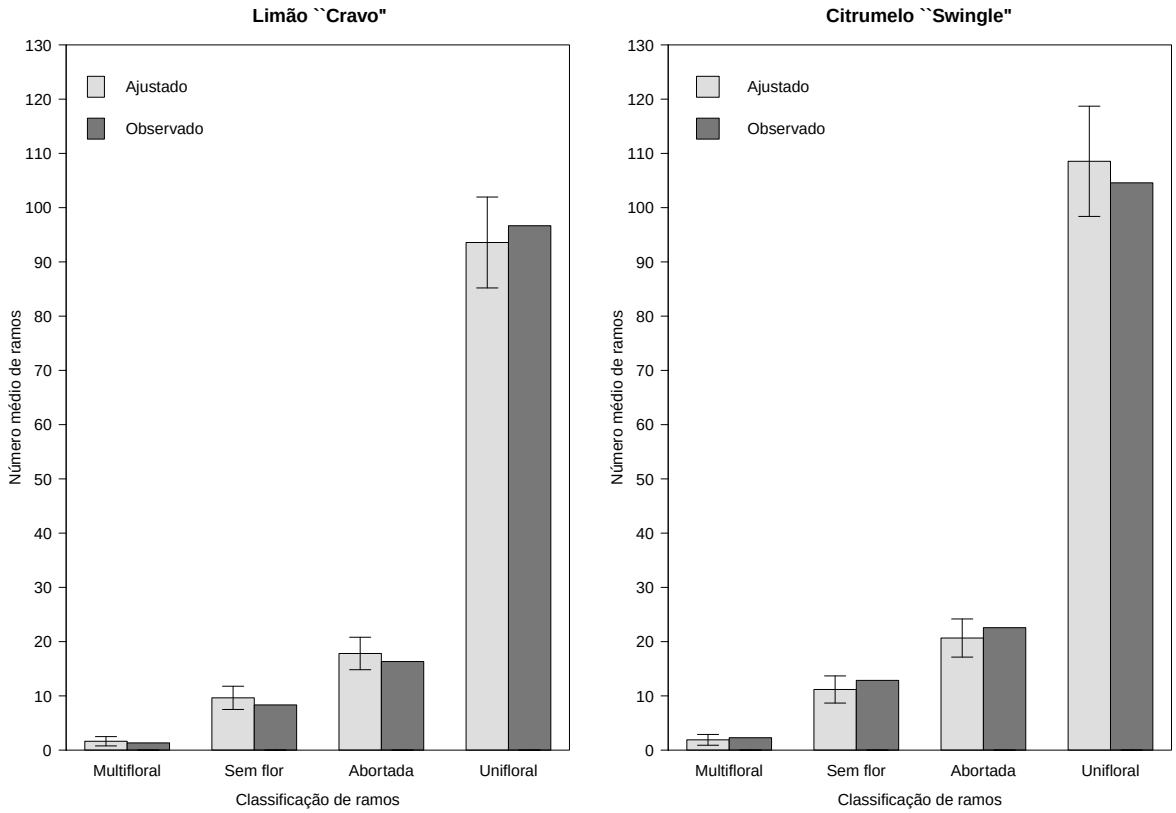
“Swingle”: $\widehat{\mu}_s = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j + \widehat{\beta}_{2k}\text{porta-enxerto}_k\right) \quad j = 2,3,4, \quad k = 2$

“Cravo”: $\widehat{\mu}_c = \exp\left(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j + \widehat{\beta}_{2k}\text{porta-enxerto}_k\right) \quad j = 2,3,4, \quad k = 1$

com isso,

$$\frac{\widehat{\mu}_s}{\widehat{\mu}_c} = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j + \widehat{\beta}_{2k}\text{porta-enxerto}_k)}{\exp(\widehat{\beta}_0 + \widehat{\beta}_{1j}\text{classificação}_j + \widehat{\beta}_{2k}\text{porta-enxerto}_k)} = \exp(0,15) \cong 1,16$$

Portanto, quando o cultivar é o citrumelo “Swingle”, tem-se um aumento médio na produção de ramos de 16% em relação ao limão “Cravo”. De forma análoga é possível concluir com 95% de confiança que o porta-enxerto citrumelo “Swingle” terá um produção de 3% a 30% maior no número de ramos quando comparado com o porta-enxerto limão “Cravo” na estação primavera. Na Figura 5(a) e Figura 5(b) encontram-se o número médio de ramos observados e o número médio ramos estimado pelo modelo e seus respectivos intervalos de confiança. Nota-se a predominância de ramos uniflorais para a estação primavera.



(a) (b)

Figura 5 - Comparação de valores ajustados e observados do modelo Quase-Poisson (a) Limão "Cravo" (b) citrumelo "Swingle", com seus respectivos intervalos com 95% de confiança para a estação primavera

6 CONCLUSÃO

Neste trabalho os dados foram analisados utilizando três modelos, sendo dois deles apropriados para dados de contagem com superdispersão. Segundo os critérios de adequabilidade do ajuste dos modelos adotados, ou sejam o gráfico *half-normal plot* e as medidas de qualidade de ajuste, conclui-se que os modelos Poisson e o Binomial Negativo não apresentaram um ajuste satisfatório aos dados. O modelo Quase-Poisson, por sua vez, foi o que melhor se ajustou aos dados e, por meio deste, foi possível obter estimativas pontuais e intervalares, assim como estimar o número médio de ramos para cada porta-enxerto dentro de sua respectiva classificação, auxiliando na compreensão da indução floral e, consequentemente, na produção de frutos.

Os demais modelos citados não se adequaram satisfatoriamente aos nossos dados. No entanto, esses modelos podem ser úteis para outras aplicações. Esse estudo mostra que é importante a aplicação de diferentes ajustes de modelos aos dados de modo a encontrar aquele que melhor responda aos objetivos propostos. Um tema bastante interessante para pesquisas futuras é considerar um estudo de simulação para dados de contagem com superdispersão (não inflacionado de zeros), envolvendo médias altas e baixas, para comparar o desempenho dos modelos Binomial Negativo e Quase-Poisson. A ideia parte do princípio de que quando tratamentos apresentam médias baixas o melhor ajuste parece ser o Quase-Poisson, ao passo que, quando os tratamentos apresentam médias altas, o Binomial Negativo parece ser uma escolha plausível em função do peso na matriz $\mathbf{W}$.

Os modelos lineares generalizados têm recebido especial atenção nos últimos anos, uma vez que sua aplicação é plausível em diversas áreas do conhecimento. Este trabalho proporcionou uma oportunidade singular de testar os diferentes modelos a fim de verificar qual a melhor forma de tratamento dos dados.

REFERÊNCIAS

AGRESTI, A. **Categorical data analysis**. 2nd ed. New Jersey: John Wiley & Sons, 2002. 709 p.

ATKINSON, A. C. **Plots, transformations, and regression:** an introduction to graphical methods of diagnostic regression analysis. Oxford: Clarendon , 1985. 282 p.

BARTLETT, M. S. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences**, JSTOR, London, v. 160, n. 901, p. 268–282, 1937.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society,** London, v. 26, n. 2, p. 211–252, 1964.

BRESLOW, N. E. Extra-poisson variation in log-linear models. **Applied statistics**, London, v. 33, n. 1, p. 38–44, 1984.

CAMERON, A. C.; TRIVEDI, P. K. **Regression analysis of count data**. 2nd ed. New York: Cambridge University Press, 2013. 411 p.

CAMERON, A. C.; TRIVEDI, P. K. Econometric models based on count data. comparisons and applications of some estimators and tests. **Journal of applied econometrics**; Cambridge, v. 1, n. 1, p. 29–53, 1986.

CORDEIRO, G. M. **Modelos lineares generalizados**. Campinas, VII SINAPE, 1986. 286p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. B. **Modelos lineares generalizados e Extensões**, 2010. Disponível em: <http://www.lce.esalq.usp.br/arquivos/aulas/2010/LCE5868/livro.pdf>. Acesso em: 29 jan. 2014.

CORDEIRO, G. M.; LIMA, E. A. N. **Modelos paramétricos**, 2006. Disponível em:<http://www.ufjf.br/clecio_ferreira/files/_2013/05/Livro-Gauss-e-Eufrasio.pdf>. Acesso em: 29 jan. 2014.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**, 2002. Disponível em:<http://www.lce.esalq.usp.br/clarice/Apostila.pdf>. Acesso em: 29 jan. 2014.

DOBSON, A. J. **An introduction to generalized linear models**. 2nd ed. New York: Chapman & Hall/CRC, 2010. 221 p.

DURBIN, J.; WATSON, G. S. Testing for serial correlation in least squares regression: I. **Biometrika**, London, v. 37, n. 3/4, p. 409–428, 1950.

FAO. **Food and Agricultural commodities production Countries by commodity**. 2013. Disponível em: <http://faostat3.fao.org/browse/rankings>. Acesso: 02 de março de 2015.

FARAWAY, J. J. **Extending the linear model with R:** generalized linear, mixed effects and nonparametric regression models. New York: Chapman & Hall/CRC, 2005. 331 p.

GARDNER, W.; MULVEY, E. P.; SHAW, E. C. Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. **Psychological bulletin**; New York, v. 118, n. 3, p. 392, 1995.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. **Computational Statistics & Data Analysis**, Amsterdam, v. 27, n. 2, p. 151–170, 1998a.

_____. **Overdispersion:** Models and Estimation, 1998b. Disponível em:<http://pointer.esalq.usp.br/departamentos/lce/arquivos/aulas/2011/LCE5868/OverdispersionBook.pdf>. Acesso em: 29 jan. 2014.

HOEF, J. M. V.; BOVENG, P. L. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? **Ecology**, London, v. 88, n. 11, p. 2766–2772, 2007.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. 2nd ed. New York: John Wiley & Sons, 2004. 375 p.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE. **Produção Agrícola Municipal**. 2013. Disponível em: <http://www.sidra.ibge.gov.br>. Acesso: 02 de março de 2015.

JØRGENSEN, B. Generalized linear models. **Encyclopedia of environmetrics**. Chichester: Jonh Wiley, 2002. p. 873-880.

KARAZSIA, B. T.; DULMEN, M. H. Regression models for count data: Illustrations using longitudinal predictors of childhood injury. **Journal of pediatric psychology**, Atlanta, v. 33, n. 10, p. 1076–1084, 2008.

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. **Technometrics**, Abingdon, v. 34, n. 1, p. 1–14, 1992.

LEE, J. H.; HAN G.; FULP, W. J.; GIULIANO A. R. Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. **Epidemiology and infection**, Cambridge Univ Press, v. 140, n. 6, p. 1087–1094, 2012.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized linear models with random effects:** unified analysis via H-likelihood. New York: John Wiley & Sons, 2006. 326 p.

McCULLAGH, P.; NELDER, J. A. **Generalized linear models**, Londres: Chapman and hall, 1989. 511 p.

MORAL, R. A. **Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais**. 2013. 173 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2013.

MORAL, R. A.; HINDE, J.; DEMETRIO, C. G. B. **hnp: Half-Normal Plots with Simulation Envelopes**. 2014. R package version 1.0. Disponível em: <http://CRAN.R-project.org/package=hnp>. Acesso em: 01 abr. 2015

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G.; ROBINSON, T. J. **Generalized linear models: with applications in engineering and the sciences**. 2nd ed. New Jersey: John Wiley & Sons, 2010. 496 p.

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

OLSSON, U. **Generalized linear models:** an applied approach. Lund: Studentlitteratur, 2002. 232 p.

PAULA, G. A. **Modelos de Regressão:** com apoio computacional, 2013. Disponível em: <https://www.ime.usp.br/ giapaula/texto_2013.pdf>. Acesso em: 14 fev. 2014.

POISSON, S. D. Recherches sur la probabilité des jugements en matières criminelles et matière civile (4to, 1837). all published at Paris. **A translation of Poisson's Treatise on Mechanics was published in London in**, 1842.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 20 Jul. 2014.

RICHARDS, S. A. Dealing with overdispersed count data in applied ecology. **Journal of Applied Ecology**, Nova Jersey, v. 45, n. 1, p. 218–227, 2008.

RIDOUT, M.; DEMÉTRIO, C. G. B.; HINDE, J. Models for count data with many zeros. In: **Proceedings of the XIXth International Biometric Conference**, Cape Town, 1998. v. 19, p. 179–192.

SHAPIRO, S. S.; WILK, M. B.; CHEN, H. J. A comparative study of various tests for normality. **Journal of the American Statistical Association**, Taylor & Francis Group, Boston, v. 63, n. 324, p. 1343–1372, 1968.

SPIEGEL-ROY, P.; GOLDSCHMIDT, E. E. **The biology of citrus**. Cambridge: Cambrige University Press, 1996. 230 p.

SURIYAGODA, L. D. B.; RYAN, M. H.; LAMBERS, H.; RENTON, M. Comparison of novel and standard methods for analysing patterns of plant death in designed field experiments. **Journal of Agricultural Science-London**, Cambridge Univ Press, v. 150, n. 3, p. 319, 2012.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4>. Acesso em: 01 abr. 2015

VIEIRA, A. M. C. **Modelos para dados de proporções com superdispersão aplicados ao controle biológico**. 1998. 61 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1998.

VOIGT, V. **Caracterização fenotípica e avaliação da expressão de genes envolvidos na indução e no florescimento da laranjeira'x11'**. 2013. 109 p. Dissertação (Mestrado em Biologia na Agricultura e no Ambiente) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2013.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. **Biometrika**, Biometrika Trust, London v. 61, n. 3, p. 439–447, 1974.

WILLIAMS, D. A. Extra-binomial variation in logistic linear models. **Applied statistics**, Hoboken, v. 31, n. 2, p. 144–148, 1982.

WINKELMANN, R. **Econometric analysis of count data**. Berlin: Springer-Verlag, 2008. 332 p.

APÊNDICE

Linhas de comando R

```
rm(list=ls(all=TRUE))
dados <- read.table("dados.txt", header=TRUE, quote="\"")
str(banco)
ptex1=factor((banco$ptex),labels=c("Cravo","Swingle"))
catg1=factor((banco$catg),labels=c("Unifloral","Multifloral","Sem flor"," Abortada"))
dados=cbind(banco,ptex1,catg1)
attach(dados)
str(dados)
summary(dados)

############################################################
#                    Análise descritiva                    #
############################################################
tapply(resp,list(catg1),sum)
tapply(resp,list(ptex1),sum)
tapply(resp,list(catg1),var)
tapply(resp,list(catg1),length)
tapply(resp,list(ptex1,catg1),median)
tapply(resp,list(ptex1,catg1),var)
tapply(resp,list(ptex1,catg1),IQR)
media=tapply(resp,list(ptex1,catg1),mean)
n=tapply(resp,ptex1, length);n
soma= tapply(resp,ptex1, sum); soma
media= tapply(resp,ptex1, mean);media
variancia= tapply(resp,ptex1, var); variancia
dist.int= tapply(resp,ptex1, IQR); dist.int
f1=function(x) max(x) - min(x)
amplitude= tapply(resp,catg1, f1); amplitude
resumo1=rbind(n,soma, media, variancia, dist.int,amplitude)
round(resumo1,4)
mean(resp)
var(resp)
var(resp)/mean(resp)
n=tapply(resp,list(catg1,ptex1), length);n
soma= tapply(resp,list(catg1,ptex1), sum); soma
media= tapply(resp,list(catg1,ptex1), mean);media
variancia= tapply(resp,list(catg1,ptex1), var); variancia
dist.int= tapply(resp,list(catg1,ptex1), IQR); dist.int
f1=function(x) max(x) - min(x)
amplitude= tapply(resp,list(catg1,ptex1), f1); amplitude
resumo2=cbind(n,soma, media, variancia, dist.int,amplitude)
round(resumo2,2)
library(lattice)
bwplot(resp ~ ptex1|catg1,
horizontal = FALSE ,
,xlab="Ponta enxerto ",
ylab="Números de ramos",layout=c(4,1))
xyplot(resp~ptex1|catg1,type=c("p"),
xlab="Tipo de ponta enxerto ",
ylab="Números de ramos",layout=c(4,1),
points = F)
############################################################
#                           GLM                            #
############################################################
```

```
############################################################
#                    modelo de Poisson                     #
############################################################
mod1_poi=glm(resp~1,family=poisson,data=dados);summary(mod1_poi)
mod2_poi=glm(resp~ptex1,family=poisson,data=dados);summary(mod2_poi)
mod3_poi=glm(resp~catg1,family=poisson(link = "log"),data=dados);summary(mod3_poi)
mod4_poi=glm(resp~catg1+ptex1,family=poisson(link="log"));summary(mod4_poi)
anova(mod1_poi,mod2_poi,mod3_poi,mod4_poi, test="LRT")
deviance(mod4_poi)
(pvalor=1-pchisq(deviance(mod4_poi),df.residual(mod4_poi)))
(phi_hat = (sum(residuals(mod4_poi,type="pearson")^2)/(mod4_poi$df.res)))
(X2<-sum(residuals(mod4_poi, 'pearson')^2))
(pvalor=1-pchisq(X2, df.residual(mod4_poi)))
(est <-(cbind(Estimate = coef(mod4_poi),
confint(mod4_poi,level=.95))))
round(exp(est),2)
require(hnp)
hnp(mod4_poi)
############################################################
#                          quasipoisson                    #
############################################################
mod1_Qpoi=glm(resp~1,family=quasipoisson);summary(mod1_Qpoi)
mod2_Qpoi=glm(resp~ptex1,family=quasipoisson);summary(mod2_Qpoi)
mod3_Qpoi=glm(resp~catg1,family=quasipoisson);summary(mod3_Qpoi)
mod4_Qpoi=glm(resp~ptex1+catg1,family=quasipoisson);summary(mod4_Qpoi)
anova(mod1_Qpoi,mod2_Qpoi,mod3_Qpoi,mod4_Qpoi,test="F")
deviance(mod4_Qpoi)
(pvalor=1-pchisq(deviance(mod4_Qpoi),df.residual(mod4_Qpoi)))
(phi_hat = (sum(residuals(mod4_Qpoi,type="pearson")^2)/(mod4_Qpoi$df.res)))
(X2<-sum(residuals(mod4_Qpoi, 'pearson')^2))
(pvalor=1-pchisq(X2, df.residual(mod4_Qpoi)))
(est <-(cbind(Estimate = coef(mod4_Qpoi),
confint(mod4_Qpoi,level=.95))))
round(exp(est),2)
require(hnp)
hnp(mod4_Qpoi)
############################################################
#                    binomial negativo                     #
############################################################
library(MASS)
mod1_bn=glm.nb(resp~1,data=dados,link="log");summary(mod1_bn)
mod2_bn=glm.nb(resp~ptex1,data=dados,link="log");summary(mod2_bn)
mod3_bn=glm.nb(resp~catg1,data=dados,link="log");summary(mod3_bn)
mod4_bn=glm.nb(resp~ptex1+catg1,link=("log"),data=dados);summary(mod4_bn)
anova(mod1_bn,mod2_bn,mod3_bn,mod4_bn, test="Chisq")
(deviace <- cbind(mod2_bn$deviance,mod3_bn$deviance,mod4_bn$deviance))
deviance(mod4_bn)
(pvalor=1-pchisq(deviance(mod4_bn),df.residual(mod4_bn)))
(X2<-sum(residuals(mod4_bn, 'pearson')^2))
(pvalor=1-pchisq(X2, df.residual(mod4_bn)))# nao rejeitamos o modelo
require(hnp)
hnp(mod4_bn)
####################################################################
# grafico valores observados X valores preditos e IC
####################################################################
media_o=as.matrix(tapply(resp,list(ptex,catg1),mean))
mod5_Qpoi=glm(resp~catg1+ptex1,family=quasipoisson);summary(mod5_Qpoi)
valor_predito_pm=as.matrix((predict(mod5_Qpoi,type="response")))
```

```
# Limite inferiro
inflim_pm <- as.matrix(with(predict(mod5_Qpoi, se=TRUE, type="response"),
fit - qnorm(0.975)*se.fit))
# Limite superior
suplim_pm <- as.matrix(with(predict(mod5_Qpoi, se=TRUE, type="response"),
fit + qnorm(0.975)*se.fit))

m_ajus=as.matrix(mod5_Qpoi$fitted.values)
medias=(tapply(resp,list(ptex1,catg1),mean))
(mediasA=matrix(c(m_ajus[17,],
m_ajus[33,],
m_ajus[49,],
m_ajus[1,],
m_ajus[26,],
m_ajus[42,],
m_ajus[58,],
m_ajus[10,]),
nrow = 2, ncol = 4,byrow = TRUE,
dimnames = list(c("Cravo", "Swingle"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))
tapply(resp,list(ptex1,catg1),mean)
##################################################
#         grafico de barras limao cravo
##################################################
(mediascravo=matrix(c(valor_predito_pm[17,],
valor_predito_pm[33,],
valor_predito_pm[49,],
valor_predito_pm[1,],
media_o[1,1],
media_o[1,2],
media_o[1,3],
media_o[1,4]),
nrow = 2, ncol = 4,byrow = TRUE,
dimnames = list(c("ajustadoCravo", "observadocravo"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))
(LI_cravo=matrix(c(inflim_pm[17,],
inflim_pm[33,],
inflim_pm[49,],
inflim_pm[1,],0,0,0,0),
nrow = 2, ncol = 4,byrow = T,
dimnames = list(c("LI_cravo","observado"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))
(LS_cravo=matrix(c(suplim_pm[17,],
suplim_pm[33,],
suplim_pm[49,],
suplim_pm[1,],0,0,0,0),
nrow = 2, ncol = 4,byrow = T,
dimnames = list(c("LI_cravo","observado"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))

library(gplots)
tapply(resp,list(ptex1,catg1),mean)
colcamb <- c("gray87","gray47")
rb=barplot2(mediascravo,
beside=TRUE,
xlab="Classificação de ramos", ylab="Número médio de ramos",
col=colcamb,main=" Limão ‘‘Cravo’’",
plot.ci=TRUE,
```

```
ci.l=LI_cravo,
ci.u=LS_cravo,
ylim=c(0,130),axes=F)
ylabel=seq(from=0,to=140,by=10)# criando o eixo
axis(2,at=ylabel,las=2)
legend("topleft", legend=c("Ajustado","Observado"),
fill=colcamb, bty="n")
box()
############################################################
#         grafico de barras limao swingle
############################################################
(mediasswigle=matrix(c(valor_predito_pm[26,],
valor_predito_pm[42,],
valor_predito_pm[58,],
valor_predito_pm[10,],
media_o[2,1],
media_o[2,2],
media_o[2,3],
media_o[2,4]),
nrow = 2, ncol = 4,byrow = TRUE,
dimnames = list(c("ajustadoswingle", "observadoswingle"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))
##inflim_pm#################################################
(LI_swigle=matrix(c(inflim_pm[26,],
inflim_pm[42,],
inflim_pm[58,],
inflim_pm[10,],0,0,0,0),
nrow = 2, ncol = 4,byrow = T,
dimnames = list(c("LI_Swingle","observado"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))
(LS_swigle=matrix(c(suplim_pm[26,],
suplim_pm[42,],
suplim_pm[58,],
suplim_pm[10,],0,0,0,0),
nrow = 2, ncol = 4,byrow = T,
dimnames = list(c("LI_Swingle","observado"),
c("Multifloral", "Sem flor", "Abortada","Unifloral"))))

############################################################
library(gplots)
tapply(resp,list(ptex1,catg1),mean)
colcamb <- c("gray87","gray47")
barplot2(mediasswigle,
beside=TRUE,
xlab="Classificação de ramos", ylab="Número médio de ramos",
col=colcamb,main="Citrumelo ‘‘Swingle’’",
plot.ci=TRUE,
ci.l=LI_swigle,
ci.u=LS_swigle,
ylim=c(0,130),axes=F)
ylabel=seq(from=0,to=140,by=10)# criando o eixo
axis(2,at=ylabel,las=2)
box()
legend("topleft", legend=c("Ajustado","Observado"),
fill=colcamb, bty="n")
############################################################
#     BRESLOW   glm.poisson.disp {dispmod}              #
############################################################
```

```
require(dispmod)
require(MASS)
mod.disp <- glm.poisson.disp(mod4_poi,maxit=100)
summary(mod.disp)
mod.disp$dispersion
mod.disp$disp.weights
hnp(mod.disp)
deviance(mod.disp)
(pvalor=1-pchisq(deviance(mod.disp),df.residual(mod.disp)))
(phi_hat = (sum(residuals(mod.disp,type="pearson")^2)/(mod.disp$df.res)))
(X2<-sum(residuals(mod.disp, 'pearson')^2))
(pvalor=1-pchisq(X2, df.residual(mod.disp)))
```