MODELOS LINEARES GENERALIZADOS MISTOS PARA DADOS LONGITUDINAIS

SILVANO CESAR DA COSTA

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Doutor em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

P I R A C I C A B A
Estado de São Paulo - Brasil
Janeiro - 2003

MODELOS LINEARES GENERALIZADOS MISTOS PARA DADOS LONGITUDINAIS

SILVANO CESAR DA COSTA

Licenciado em Matemática

Orientadora: Prof^{a} Dr^{a} **CLARICE GARCIA BORGES DEMÉTRIO**

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Doutor em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

P I R A C I C A B A

Estado de São Paulo - Brasil

Janeiro - 2003

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

Costa, Silvano Cesar da
Modelos lineares generalizados mistos para dados longitudinais / Silvano Cesar da Costa. - - Piracicaba, 2003.
110 p.

Tese (doutorado) - Escola Superior de Agricultura Luiz de Queiroz, 2003.
Bibliografia.

1. Análise de dados longitudinais 2. Distribuição binomial 3. Distribuição de Poisson 4. Modelos lineares generalizados 5. SAS (programa de computador) I. Título

CDD 511.8

DEDICATÓRIA

A

DEUS

força maior de todo ser humano.

Aos meus pais,
João Candido da Costa e
Maria Martins da Silva Costa,
(*in memoriam*), os inúmeros bons exemplos que propiciaram, os quais me impulsionam na batalha do dia-a-dia.

À minha esposa **Simoni** o amor, a compreensão, o apoio e, principalmente, por se manter sempre forte e zelar pela saúde e bem-estar de nossos filhos, nos momentos em que estive ausente.

Aos meus filhos **Bruno (o "amigão")** e **Amanda (minha "lindezura")**, que alegram minha vida dando-me motivos e estímulos para progredir.

AGRADECIMENTOS

À Prof^{a} Dr^{a} Clarice Garcia Borges Demétrio a orientação, a amizade e a confiança depositada.

À CAPES o fundamental suporte financeiro concedido.

À Universidade Estadual de Londrina e aos professores do Departamento de Matemática Aplicada o afastamento concedido para a realização deste trabalho.

Aos meus sogros Afrânio Gomes Patriota e Mirian de Macedo Patriota o apoio, o incentivo, a confiança e por se fazerem presentes nos momentos difíceis.

Aos professores e funcionários do Departamento de Ciências Exatas da ESALQ/USP que me propiciaram condições para a realização deste trabalho.

Aos colegas e amigos de doutorado, em especial à Maria Cristina Neves de Oliveira (Embrapa Soja - Londrina-PR), Suely Ruiz Giolo (UFPR - Curitiba-PR), Claudia Cristina Paro de Paz (IZ - Colina - SP), João Mauricio de Araújo Mota (UFC - Fortaleza-CE) e Adriano Ferreti Borgatto a força, a amizade, a troca de conhecimentos e atenção recebida em todos os momentos.

Ao pesquisador Eduardo Suguino e à FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo que cederam os dados sobre Camu-Camu utilizados neste trabalho.

Ao pesquisador Odnei Fernandes e a empresa Monsanto do Brasil S/A que cederam os dados sobre milho geneticamente modificado utilizados neste trabalho.

A todos que, de forma direta ou indireta, contribuiram para a realização deste trabalho.

SUMÁRIO

Página

LISTA DE FIGURAS

Página

LISTA DE TABELAS

Página

MODELOS LINEARES GENERALIZADOS MISTOS PARA DADOS LONGITUDINAIS

Autor: SILVANO CESAR DA COSTA

Orientadora: Profª Drª CLARICE GARCIA BORGES DEMÉTRIO

RESUMO

Experimentos cujas variáveis respostas são proporções ou contagens, são muito comuns nas diversas áreas do conhecimento, principalmente na área agrícola. Na análise desses experimentos, utiliza-se a teoria de modelos lineares generalizados, bastante difundida (McCullagh & Nelder, 1989; Demétrio, 2001), em que as respostas são independentes. Caso a variância estimada seja maior do que a esperada, estima-se o parâmetro de dispersão, incluindo-o no processo de estimação dos parâmetros. Quando a variável resposta é observada ao longo do tempo, pode haver uma correlação entre as observações e isso tem que ser levado em consideração na estimação dos parâmetros. Uma forma de se trabalhar essa correlação é aplicando a metodologia de equações de estimação generalizada (EEG), discutida por Liang & Zeger (1986), embora, neste caso, o interesse esteja nas estimativas dos efeitos fixos e a inclusão da matriz de correlação de "trabalho"sirva para se obter um melhor

ajuste. Uma outra alternativa é a inclusão, no preditor linear, de um efeito latente para captar variabilidades não consideradas no modelo e que podem influenciar nos resultados. No presente trabalho, usa-se uma forma combinada de efeito aleatório e parâmetro de dispersão, incluídos conjuntamente na estimação dos parâmetros. Essa metodologia é aplicada a um conjunto de dados obtidos de um experimento com camu-camu, com objetivo de se avaliarem quais os melhores métodos de enxertia e tipos de porta-enxertos que podem ser utilizados, através da proporção de pegamentos da muda. Vários modelos são ajustados, desde o modelo em parcelas subdivididas (supondo independência), até o modelo em que se considera o parâmetro de dispersão e efeito aleatório conjuntamente. Há evidências de que o modelo em que se inclui o efeito aleatório e o parâmetro de dispersão, conjuntamente, resultam em melhores estimativas dos parâmetros. Outro conjunto de dados longitudinais, com milho transgênico MON810, em que a variável resposta é o número de lagartas (*Spodoptera frugiperda*), é utilizado. Neste caso, devido ao excesso de respostas zero, emprega-se o modelo de regressão Poisson inflacionado de zeros (ZIP), além do modelo Poisson padrão, em que as observações são consideradas independentes, e do modelo Poisson inflacionado de zeros com efeito aleatório. Os resultados mostram que o efeito aleatório incluído no preditor foi não significativo e, assim, o modelo adotado é o modelo de regressão Poisson inflacionado de zeros. Os resultados foram obtidos usando-se os procedimentos NLMIXED, GENMOD e GPLOT do SAS - *Statistical Analysis System*, versão 8.2.

GENERALIZED LINEAR MIXED MODELS IN LONGITUDINAL DATA

Author: SILVANO CESAR DA COSTA
Adviser: Prof[a] Dr[a] CLARICE GARCIA BORGES DEMÉTRIO

SUMMARY

Experiments which response variables are proportions or counts are very common in several research areas, specially in the area of agriculture. The theory of generalized linear models, well diffused (McCullagh & Nelder, 1989; Demétrio, 2001), is used for analyzing these experiments where the responses are independent. If the estimated variance is greater than the expected variance, the dispersion parameter is estimated including it on the parameter estimation process. When the response variable is observed over time a correlation among observations might occur and it should be taken into account in the parameter estimation. A way of dealing with this correlation is applying the methodology of generalized estimating equations (GEEs) discussed by Liang & Zeger (1986) although, in this case, the interest is on the estimates of the fixed effect being the inclusion of a "working" correlation matrix useful to obtain more accurate estimates. Another alternative is the inclusion

of a latent effect in the linear predictor to explain variabilities not considered in the model that might influence the results. In this work the random effect and the dispersion parameter are combined and included together in the parameter estimation. Such methodology is applied to a data set obtained from an experiment realized with camu-camu to evaluate, through proportion of grafting well successful of seedling, which kind of grafting and understock are suitable to be used. Several models are fitted, since the split plot model (with independence assumption) up to the model where the dispersion parameter and the random effect are considered together. There is evidence that the model including the random effect and the dispersion parameter together, produce better estimates of the parameters. Another longitudinal data set used here comes from an experiment realized with the MON810 transgenic corn where the response variable is the number of caterpillars (*Spodoptera frugiperda*). In this case, due to the excessive number of zeros obtained, the zero inflated Poisson regression model (ZIP) is used in addition to the standard Poisson model, where observations are considered independent, and the zero inflated Poisson regression model with random effect. The results show that the random effect included in the linear predictor was not significant and, therefore, the adopted model is the zero inflated Poisson regression model. The results were obtained using the procedures NLMIXED, GENMOD and GPLOT available on SAS - Statistical Analysis System, version 8.2.

1 INTRODUÇÃO

O uso de modelos lineares clássicos não é, em geral, apropriado para analisar dados de contagens ou proporções, que são muito freqüentes nas mais diversas áreas, pois as pressuposições do modelo (aditividade, normalidade, variância constante, independência) não são atendidas. Uma abordagem comum é a transformação dos dados, com o propósito de que esse problema seja contornado, o que nem sempre ocorre.

Na análise desse tipo de dados, pode-se utilizar a teoria de modelos lineares generalizados (MLG) que também envolve uma transformação (não dos dados, mas da esperança condicional; nesse contexto a transformação é conhecida como função de ligação), mas o objetivo principal é encontrar uma escala sobre a qual um modelo linear aditivo ocorra. As distribuições padrões mais utilizadas para a análise de dados de contagens e proporções são, respectivamente, Poisson e binomial, sendo casos particulares de modelos lineares generalizados (Nelder & Wedderburn, 1972).

Dados de proporções podem ser obtidos, por exemplo, a partir de experimentos realizados para verificar o pegamento de porta-enxerto, em que a planta apresenta uma de duas respostas possíveis: pegou ou não pegou. Assim, a variável resposta, número de pegamentos, pode ser estudada como a soma de eventos Bernoulli independentes. A proporção de pegamentos pode ser usada na comparação de diferentes porta-enxertos.

Já, dados de contagem surgem de várias formas, podendo ser, por exemplo, o número de lagartas observadas na cultura de milho para se verificar a eficácia do milho geneticamente modificado no controle da praga.

Quando a mesma unidade experimental é observada ao longo do tempo

(longitudinal), espera-se que haja uma correlação entre essas unidades, levando à violação da suposição de independência. Uma forma de se levar em consideração os dados correlacionados é modelar explicitamente a estrutura de correlação, utilizando-se a abordagem de equação de estimação generalizada (EEG), dada por Liang & Zeger (1986). Este método permite modelar a variabilidade entre as observações incluindo na análise uma matriz de correlação de "trabalho". Os modelos lineares generalizados têm somente um componente aleatório mas podem ser estendidos para ter efeitos aleatórios no preditor linear. A extensão é conhecida como modelos lineares generalizados mistos (MLGM).

Logo, uma forma alternativa de se modelarem dados longitudinais é com a inclusão de variáveis latentes no preditor linear para se levar em conta a correlação entre as mesmas unidades experimentais, assumindo-se uma distribuição de probabilidade, em geral, a distribuição normal, para a variável latente.

Os objetivos principais deste trabalho são:

i) revisão da literatura sobre a modelagem de proporções medidas longitudinalmente;

ii) a extensão dos modelos binomial e Poisson, utilizando-se de variáveis latentes e parâmetros de superdispersão na análise dos dados;

iii) implementação dos métodos utilizados em programas computacionais usando o SAS - *Statistical Analysis System* e

iv) aplicação dos modelos propostos através da utilização de dois conjuntos de dados reais, um exemplificando a distribuição binomial com efeito aleatório e outro a distribuição Poisson inflacionada de zeros com efeito aleatório.

2 REVISÃO DE LITERATURA

2.1 Modelos lineares

O modelo linear clássico utilizado na análise de dados é definido por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

em que, $\mathbf{y}$ representa o vetor de dimensões $n \times 1$, de dados observados; $\boldsymbol{X}$, de dimensões $n \times p$, é a matriz de delineamento; $\boldsymbol{\beta}$, de dimensões $p \times 1$, é um vetor de parâmetros desconhecidos de efeitos fixos e $\boldsymbol{\epsilon}$ é o vetor de dimensões $n \times 1$, de erros aleatórios.

O objetivo do modelo linear clássico é modelar a média de $\boldsymbol{y}$, usando-se o vetor de parâmetros de efeitos fixos $\boldsymbol{\beta}$. Os componentes do vetor $\boldsymbol{\epsilon}$ são variáveis aleatórias independentes e identicamente distribuídas com média 0 e variância σ^2.

Assumindo-se que $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$, tem-se que $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, cuja função de verossimilhança é dada por:

$$L = L(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)'\left(\frac{\boldsymbol{I}}{\sigma^2}\right)\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right]}{(2\pi\sigma^2)^{n/2}}.$$

As estimativas de máxima verossimilhança dos parâmetros são obtidas resolvendo-se o sistema de equações normais: $\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'y}$, cuja solução é $\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}X'y}$, desde que $\boldsymbol{X'X}$ seja não singular. Tem-se, também, que $V(\hat{\boldsymbol{\beta}}) = \boldsymbol{(X'X)}^{-1}\sigma^2$.

Caso $(\boldsymbol{X'X})^{-1}$ não exista, utiliza-se uma inversa generalizada e as estimativas dos parâmetros são dadas por:

$$\boldsymbol{\beta}^o = (\mathbf{X'X})^{-}\mathbf{X'y} \text{ e } \mathbf{V}(\boldsymbol{\beta}^{\mathbf{o}}) = \boldsymbol{(X'X)^{-}X'X(X'X)^{-'}\sigma^2}.$$

2.2 Modelos lineares mistos

O nome modelo misto vem do fato de que o modelo contém parâmetros de efeitos fixos, $\boldsymbol{\beta}$, e parâmetros de efeitos aleatórios, $\boldsymbol{b}$. Estes modelos são usados para modelar a parte aleatória através da inclusão de uma matriz de variâncias-covariâncias (Littell et al., 1996). A necessidade de se incluírem parâmetros de covariâncias pode surgir por várias razões, dentre elas:

i) as unidades experimentais sobre as quais os dados são medidos, podem ser colocadas em grupos e os dados de um grupo comum são correlacionados. Isso pode ocorrer com dados de famílias, ninhadas, colônias e pessoas que habitam a mesma casa e

ii) medidas repetidas são tomadas sobre a mesma unidade experimental e são correlacionadas. A natureza dessas medidas pode ser multivariada. Exemplos comuns são dados observados ao longo do tempo, chamados dados longitudinais.

Em princípio, todo modelo linear que contenha a média geral ou uma constante μ, tomada como fixa, e um termo referente ao erro, assumido como aleatório, é um modelo linear misto. Contudo, tal denominação é, geralmente, reservada a modelos lineares que contenham efeitos fixos, além de μ, e qualquer outro termo aleatório, além do erro (Martins et al., 1993). Assim, pode-se considerar como misto o seguinte modelo:

$$\boldsymbol{y = X\beta + Zb + \epsilon}$$

em que $\boldsymbol{y}$ é o vetor de observações e assume-se que $\boldsymbol{\beta}$ é um vetor de parâmetros de efeitos fixos desconhecidos, com matriz de delineamento conhecida $\boldsymbol{X}$, $\boldsymbol{b}$ é um vetor de parâmetros de efeitos aleatórios desconhecidos, com matriz de delineamento conhecida $\boldsymbol{Z}$ e $\boldsymbol{\epsilon}$ um vetor de erros aleatórios desconhecidos. Admitindo-se que $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G})$ e $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{R})$ e, além disso, que são variáveis não-correlacionadas, tem-se que,

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} \qquad \text{e} \qquad V = V(\boldsymbol{Y}) = \boldsymbol{ZGZ'} + \boldsymbol{R}, \tag{2}$$

sendo $\boldsymbol{G}$ a matriz de variâncias e covariâncias dos efeitos aleatórios presentes no vetor $\boldsymbol{b}$ e $\boldsymbol{R}$ a matriz de variâncias e covariâncias residual. Assim, pode-se modelar a variância dos dados, em (2), especificando a estrutura (ou forma) de $\boldsymbol{Z}, \boldsymbol{G}$ e $\boldsymbol{R}$.

Note que, quando $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ e $\boldsymbol{Z} = \boldsymbol{0}$, o modelo misto reduz-se ao modelo linear padrão, o qual foi definido na equação (1).

2.2.1 Estimação de β e b quando G e R são conhecidas

Se $\boldsymbol{G}$ e $\boldsymbol{R}$ são conhecidas, então, usando-se o estimador de mínimos quadrados generalizados, o melhor estimador linear não viesado de $\boldsymbol{\beta}$, é dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V^{-1}}\mathbf{X})^{-}\mathbf{X}'\mathbf{V^{-1}}\mathbf{y}, \tag{3}$$

com matriz de variâncias e covariâncias dada por:

$$V(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-1},$$

sendo $\mathbf{V}$ dada pela equação (2). Se $\boldsymbol{V}$ for singular, utiliza-se uma inversa generalizada

$$V(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X'V}^{-1}\boldsymbol{X}\right)^{-}.$$

Como $\boldsymbol{V = ZGZ' + R}$, tem-se que sua inversa é dada por:

$$\boldsymbol{V}^{-1} = \left[\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\left(\boldsymbol{Z'R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z'R}^{-1}\right],$$

assim,

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = \left[\boldsymbol{X'R}^{-1}\boldsymbol{X} - \boldsymbol{X'R}^{-1}\boldsymbol{Z}\left(\boldsymbol{Z'R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z'R}^{-1}\boldsymbol{X}\right]^{-1}.$$

Ao se substituir $\boldsymbol{V}$ por $\boldsymbol{ZGZ'} + \boldsymbol{R}$ e $\boldsymbol{V}^{-1}$ por $\left[\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}\left(\boldsymbol{Z'R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z'R}^{-1}\right]$, reduzem-se as dificuldades computacionais, dado que as dimensões de $\boldsymbol{R}$ e $\boldsymbol{G}$ são menores do que as de $\boldsymbol{V}$.

De forma análoga, o melhor preditor linear não viesado de $\boldsymbol{b}$ é:

$$\hat{\boldsymbol{b}} = (\boldsymbol{Z'R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z'R}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{4}$$

cuja variância é dada por:

$$V\left(\hat{\boldsymbol{b}}\right) = \boldsymbol{GZ'V}^{-1}\boldsymbol{ZG} - \boldsymbol{GZ'V}^{-1}\boldsymbol{X}\left(X'\boldsymbol{V}^{-1}X\right)^{-1}\boldsymbol{X'V}^{-1}\boldsymbol{ZG}.$$

Outra maneira de se obterem as estimativas de $\boldsymbol{\beta}$ e $\boldsymbol{b}$, seria através do uso de funções baseadas na função de verossimilhança dos dados, explorando-se a suposição de que $\boldsymbol{b}$ e $\boldsymbol{\epsilon}$ são normalmente distribuídos e determinando-se as equações de modelos mistos. Assim, de acordo com Martins et al. (1993), se

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{ZGZ'} + \boldsymbol{R}|^{1/2}} \exp\left\{\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta})'\left(\boldsymbol{ZGZ'} + \boldsymbol{R}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right]\right\},$$

então, a função de densidade de probabilidade conjunta, $f(\boldsymbol{y}, \boldsymbol{b}) = f(\boldsymbol{y}|\boldsymbol{b})\, f(\boldsymbol{b})$, é dada por:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{b}) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{R}|^{1/2}} \exp\left\{\frac{1}{2}\left[(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})'\,\boldsymbol{R}^{-1}\,(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})\right]\right\} \\ &\quad \frac{1}{(2\pi)^{n/2}|\boldsymbol{G}|^{1/2}} \exp\left\{\frac{1}{2}\left[(\boldsymbol{b} - \boldsymbol{0})'\,\boldsymbol{G}^{-1}\,(\boldsymbol{b} - \boldsymbol{0})\right]\right\}. \end{aligned}$$

Portanto, tem-se que o logaritmo da função de verossimilhança é dado por

$$\begin{aligned} \ell &= \frac{1}{2}2n\log(2\pi) - \frac{1}{2}\left(\log|\boldsymbol{R}| + \log|\boldsymbol{G}|\right) - \frac{1}{2}\left(\boldsymbol{y'R}^{-1}\boldsymbol{y} - 2\boldsymbol{y'R}^{-1}\boldsymbol{X\beta} - 2\boldsymbol{y'R}^{-1}\boldsymbol{Zb} + \right. \\ &\quad \left. + 2\boldsymbol{\beta'X'R}^{-1}\boldsymbol{Zb} + \boldsymbol{\beta'X'R}^{-1}\boldsymbol{X\beta} + \boldsymbol{b'Z'R}^{-1}\boldsymbol{Zb} + \boldsymbol{b'G}^{-1}\boldsymbol{b}\right). \end{aligned} \tag{5}$$

Derivando-se ℓ em relação a $\boldsymbol{\beta}$ e $\boldsymbol{b}$ e igualando-se as expressões resultantes a $\mathbf{0}$, obtêm-se as Equações de Modelos Mistos (MME)

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

que permitem obter soluções para os efeitos fixos $\boldsymbol{\beta}$ e predições para os efeitos aleatórios $\boldsymbol{b}$, como dados em (3) e (4).

2.2.2 Estimação de β e b quando G e R são desconhecidas

Se as matrizes $\boldsymbol{G}$ e/ou $\boldsymbol{R}$ (ou $\boldsymbol{V}$) são desconhecidas, a estimação dos componentes destas matrizes pode ser feita usando-se métodos baseados na função de verossimilhança padrão sob a suposição de normalidade multivariada conjunta de $\boldsymbol{b}$ e $\boldsymbol{\epsilon}$.

Sendo $\hat{\boldsymbol{G}}$ e $\hat{\boldsymbol{R}}$ os estimadores de $\boldsymbol{G}$ e $\boldsymbol{R}$, respectivamente, as equações de modelos mistos são dadas por:

$$\begin{bmatrix} \boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X} & \boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z}+\hat{\boldsymbol{G}}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{y} \end{bmatrix},$$

cujas estimativas são:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} \text{ e} \\ \hat{\boldsymbol{b}} &= \left(\boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}\boldsymbol{Z}+\hat{\boldsymbol{G}}^{-1}\right)^{-1}\boldsymbol{Z}'\hat{\boldsymbol{R}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}). \end{aligned}$$

Na prática, assume-se uma estrutura de covariâncias para $\boldsymbol{G}$ ou $\boldsymbol{R}$, ou para ambas, assim, $\boldsymbol{V}$ é uma função de poucos parâmetros desconhecidos. Para valores fixos de $\boldsymbol{V}$, obtém-se um estimador de $\boldsymbol{\beta}$, como dado em (3).

Da mesma forma que para modelos de regressão, esses resultados fornecem o fundamento para inferência estatística para combinações lineares do vetor de parâmetros $\boldsymbol{\beta}$, e podem ser usados para testar hipóteses e construir intervalos de confiança para as combinações lineares dos parâmetros.

2.2.3 Dados longitudinais

O termo medidas repetidas refere-se a casos em que conjuntos de dados são obtidos a partir de múltiplas mensurações sobre a mesma unidade experimental ou indivíduo (Littell et al., 1996).

O estudo básico de medidas repetidas consiste de delineamentos experimentais completamente aleatorizados em que os tratamentos são alocados às unidades experimentais e os dados são coletados mais de uma vez para cada unidade experimental. Desta forma, têm-se dois fatores a serem estudados, tratamentos e tempo. Assim, pode-se dizer que todo experimento com medidas repetidas são experimentos com pelo menos dois fatores, em que tratamento é o fator entre-indivíduos e tempo, o fator dentro de indivíduos ou intra-indivíduos.

Estudos longitudinais constituem um caso especial dos estudos de medidas repetidas, que abrange o delineamento conhecido como parcelas sub-divididas e *crossover*, assim como outros tipos de pesquisa em que cada unidade amostral é observada sob diferentes tratamentos e têm como objetivos:

i) o estudo do comportamento da variável resposta ao longo do tempo e

ii) a verificação da existência de dependência da variável resposta em relação às covariáveis.

A característica distinta de estudos longitudinais é a dimensão ordenada com que os dados são coletados e o fato de que as observações repetidas para um indivíduo tendem a ser correlacionadas. Tal correlação pode ser modelada através de uma estrutura de covariâncias dos dados observados sendo que, para outros tipos de dados, é usual assumir que os erros sejam independentes. Modelar uma estrutura de covariâncias apropriada é essencial para que as inferências sobre as médias sejam válidas.

Há semelhanças entre os experimentos com medidas repetidas e os experimentos em parcelas subdivididas, em que o fator tratamento e o fator tempo correspondem, respectivamente, à parcela e subparcela no experimento em parcelas

subdivididas. A diferença entre eles está no fato de que, em experimentos em parcelas subdivididas, os níveis da subparcela são aleatoriamente atribuídos às unidades de subparcela dentro das unidades de parcelas. Sendo assim, as respostas de diferentes subparcelas na mesma parcela são igualmente correlacionadas umas com as outras. Já em experimentos com medidas repetidas as respostas de tempos mais próximos são, em geral, mais fortemente correlacionadas do que as respostas de tempos mais distantes (Xavier, 2000).

Assim, os modelos para análise de dados longitudinais devem levar em conta a relação entre as observações seriais sobre a mesma unidade e os modelos de efeitos aleatórios em dois estágios podem ser usados. Nos modelos de dois estágios, as distribuições de probabilidades para vetores respostas de diferentes indivíduos pertencem a uma única família, mas alguns parâmetros de efeitos aleatórios variam através de indivíduos, com uma distribuição especificada para o segundo estágio. Laird & Ware (1982) propõem o modelo de dois-estágios:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{6}$$

em que, $\boldsymbol{\beta}$ é um vetor de dimensões $p \times 1$ de parâmetros, desconhecido, $\boldsymbol{X}_i$ é uma matriz de delineamento conhecida, específica para o i-ésimo indivíduo de dimensões $n_i \times p$, $\boldsymbol{b}_i$ é um vetor de dimensões $k \times 1$ de efeitos individuais, desconhecido, $\boldsymbol{Z}_i$ é uma matriz de dimensões $n_i \times k$ de delineamento, conhecida e $\boldsymbol{\epsilon}_i$ é distribuído como $N(\mathbf{0}, \boldsymbol{R}_i)$, sendo $\boldsymbol{R}_i$ uma matriz de covariância positiva-definida de dimensões $n_i \times n_i$.

Para o primeiro estágio, $\boldsymbol{\beta}$ e $\boldsymbol{b}_i$ são considerados fixos e os $\boldsymbol{\epsilon}_i$ são assumidos independentes, de forma que, condicional sobre $\boldsymbol{b}_i$, tem-se, de (6), que

$$\begin{aligned} E\big(Y_i \mid \boldsymbol{b}_i\big) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i \\ V\big(Y_i \mid \boldsymbol{b}_i\big) &= \boldsymbol{R}_i. \end{aligned}$$

No segundo estágio, assume-se que os $\boldsymbol{b}_i$ têm distribuição $N(\mathbf{0}, \mathbf{G})$, independentemente um dos outros e dos $\boldsymbol{\epsilon}_i$; $\boldsymbol{G}$ é uma matriz de covariância positiva-definida de dimensões $k \times k$ e os parâmetros populacionais, $\boldsymbol{\beta}$, são tratados como efeitos fixos.

Marginalmente, os $\boldsymbol{Y}_i$ são independentes e têm distribuição com média $\boldsymbol{X}_i\boldsymbol{\beta}$ e matriz de covariância $\boldsymbol{R}_i + \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{'}$. Essa família de modelos de dois-estágios inclui modelos de crescimento e modelos de medidas repetidas como casos especiais.

Note que em (2) a atenção é focada na inferência da distribuição marginal, que Zeger, Liang & Albert (1988) designaram por modelos PA (*Population-Average*) equanto em (7) é focada na distibuição condicional, obtendo-se os parâmetros específicos de indivíduos, chamados modelos SS (*Subject-Specific*)

2.3 Critério de seleção para a estrutura de covariâncias

Muitas são as estruturas de covariâncias que se podem utilizar e Xavier (2000) apresenta alguns exemplos dessas estruturas e, dentre elas, encontram-se:

i) componentes de variância: caracterizada por variâncias iguais e observações independentes,

$$V = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix};$$

ii) simetria composta (variância comum mais diagonal): igualdade de variâncias e covariâncias, ou seja, covariâncias constantes entre quaisquer observações de uma mesma unidade devido a erros independentes,

$$V = \begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix};$$

iii) não-estruturada: todas as variâncias e covariâncias podem ser desiguais. Especifica uma matriz completamente geral, parametrizada em termos de

variâncias e covariâncias. As variâncias são restritas a valores não negativos e as covariâncias não têm restrições,

$$V = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_{22} & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix};$$

iv) auto-regressiva de $1^{\underline{a}}$ ordem - AR(1): dados de séries temporais igualmente espaçados e correlações diminuindo exponencialmente, ou seja, a covariância entre duas observações decresce à medida em que aumenta o intervalo de tempo entre elas e se denota por ρ o parâmetro auto-regressivo, de forma que, para um processo estacionário, assume-se que $\mid \rho \mid < 1$,

$$V = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix};$$

v) espacial: covariâncias são funções da distância Euclidiana entre os vetores especificados pelas coordenadas. Utilizada para dados de séries temporais desigualmente espaçados,

$$\sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{31}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix},$$

entretanto, para uma escolha adequada da melhor estrutura é necessário utilizar algum critério de seleção. Dentre esses critérios, destacam-se o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC) que são, na verdade, valores para os logaritmos das funções de verossimilhanças penalizadas para o número de

parâmetros estimados, isto é,

$$\begin{aligned} AIC &= -2\ell + 2p \\ BIC &= -2\ell + p \log n, \end{aligned}$$

sendo ℓ o máximo do logaritmo da função de verossimilhança, p o número de parâmetros estimados pelo modelo e n o número de observações.

A estrutura de covariâncias com valores do critério mais próximos de zero é considerada mais adequada aos dados.

2.4 Modelos lineares generalizados e extensões

Na teoria de modelos lineares mistos assume-se, em geral, que:

i) os erros têm distribuição normal, e

ii) a variável resposta é uma combinação linear de efeitos fixos e aleatórios.

Muitas são as situações em que a variável resposta de interesse não tem distribuição normal. Em outras situações, pode haver restrições sobre a variação de valores apropriados para as funções que modelam diretamente a variável resposta. Uma alternativa para a análise de dados nesses casos é dada por Nelder & Wedderburn (1972). A idéia básica é estimar os parâmetros de um modelo linear usando-se o método da máxima verossimilhança baseado na distribuição dos dados.

Embora haja uma vasta literatura sobre Modelos Lineares Generalizados, como McCullagh & Nelder (1989), Cordeiro (1986), Collett (1991), Demétrio (2001), Dobson (2001), Paula (2001), faz-se necessária uma introdução ao assunto para desenvolvimentos em Modelos Lineares Generalizados Mistos (MLGM). Como definido por Nelder & Wedderburn (1972), os modelos lineares generalizados (MLG) para uma amostra de n observações de uma única variável resposta Y têm três componentes:

i) componente aleatório - as variáveis respostas $Y_1, Y_2, \ldots, Y_n$, são independentes e seguem uma distribuição que pertence à família exponencial na forma canônica, isto é,

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{1}{a_i(\phi)}\Big[y_i\theta_i - b(\theta_i)\Big] + c(y_i;\phi)\right\} \tag{7}$$

em que $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas, θ_i, o parâmetro natural ou canônico e, em geral, $a_i(\phi) = \dfrac{\phi}{w_i}$, com w_i o peso a priori e $\phi > 0$, conhecido, o parâmetro de dispersão ou escala. De acordo com Azzalini (1996), o parâmetro ϕ somente, não é responsável pela variabilidade de Y_i, mas sim, a razão $\dfrac{\phi}{w_i}$, que varia de observação para observação;

ii) componente sistemático - as variáveis explanatórias ou explicativas $\boldsymbol{x'}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$, $i = 1, 2, \ldots, n$ que dão origem a um vetor de preditores lineares:

$$\boldsymbol{\eta} = \boldsymbol{X\beta}$$

em que $\boldsymbol{\eta}$, chamado preditor linear, é um vetor de dimensões $n \times 1$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)', p < n$, é um vetor de p parâmetros desconhecidos a serem estimados e

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x'_1} \\ \boldsymbol{x'_2} \\ \cdots \\ \boldsymbol{x'_n} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

é uma matriz de delineamento ou covariáveis de dimensões $n \times p$ e

iii) função de ligação - faz a ligação entre o componente aleatório e o componente sistemático por meio de uma função conhecida $g(\cdot)$, monótona e diferenciável, que liga a média μ_i em (i) ao preditor linear em (ii), isto é,

$$g(\mu_i) = \eta_i = \boldsymbol{x}'_i\beta. \tag{8}$$

As funções de ligação podem ser obtidas diretamente da forma da distribuição de probabilidade na forma da família exponencial ou de algum modelo físico ou biológico de $\boldsymbol{\mu}$. Como casos particulares dos modelos lineares generalizados podem ser citados, por exemplo, a regressão logística e o modelo log-linear. Firth (1991) discute sobre a especificação da função de ligação e suas propriedades.

Assim, sendo $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ uma amostra aleatória de n observações de uma distribuição pertencente à família exponencial (7), tem-se a função de verossimilhança dada por:

$$\begin{aligned} L = L(\boldsymbol{\theta}, \phi; \boldsymbol{y}) &= \prod_{i=1}^{n} f(y_i; \theta_i, \phi) \\ &= \exp\left\{ \sum_{i=1}^{n} \left[\frac{1}{a_i(\phi)} \Big[y_i\theta_i - b(\theta_i) \Big] + c(y_i; \phi) \right] \right\} \end{aligned}$$

e o logaritmo da função de verossimilhança correspondente,

$$\ell = \ell(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \log\, L(\boldsymbol{\theta}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{1}{a_i(\phi)} \Big[y_i\theta_i - b(\theta_i) \Big] + c(y_i; \phi) \right\}. \tag{9}$$

Pode-se mostrar que a média e a variância de Y_i são dadas por:

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = \mu_i \quad \text{e} \\ V(Y_i) &= a_i(\phi) b''(\theta_i) = \frac{\phi}{w_i} b''(\theta_i) = \frac{\phi}{w_i} V(\mu_i) \end{aligned} \tag{10}$$

sendo $b'(\theta_i) = \dfrac{\partial \ell(\boldsymbol{\theta}, \phi; \boldsymbol{y})}{\partial \theta_i}$ e $b''(\theta_i) = \dfrac{\partial^2 \ell(\boldsymbol{\theta}, \phi; \boldsymbol{y})}{\partial \theta_i^2}$.

Tem-se, ainda, que $b''(\theta_i) = V(\mu_i)$, pois $b''(\theta_i) = \dfrac{\partial \mu_i}{\partial \theta_i}$, e é chamada função de variância, desempenhando um papel muito importante na família exponencial, uma vez que ela caracteriza a distribuição (Paula, 2001).

Pode-se mostrar que as distribuições normal, Poisson, binomial, binomial negativa, gama e normal inversa pertencem à família exponencial na forma canônica (Demétrio, 2001).

2.4.1 Estimação por máxima verossimilhança

Para se obterem as estimativas $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ pelo método da máxima verossimilhança é necessário determinar os valores de $\boldsymbol{\beta}$ que maximizam $\ell(\boldsymbol{\beta}, \phi, \boldsymbol{y})$.

Derivando-se (9) em relação a β_j, e usando-se a regra da cadeia obtêm-se as equações de estimação para β_j, dadas por

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)\left(\frac{\partial \theta_i}{\partial \eta_i}\right) x_{ij}, \qquad j = 1, 2, \ldots, p$$

e, fazendo-se $\Delta_i = \dfrac{\partial \theta_i}{\partial \eta_i}$, tem-se

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)\Delta_i \, x_{ij}, \qquad j = 1, 2, \ldots, p$$

que na forma matricial, fica

$$\boldsymbol{U} = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X'\Delta(y - \mu)}, \tag{11}$$

conforme notação de Liang & Zeger (1986), sendo $\boldsymbol{U}$ o vetor escore de dimensões $1 \times p$ e elementos $U_j, j = 1, \ldots, p$; $\boldsymbol{\Delta} = diag\{\Delta_i\}$; $\boldsymbol{y_i}$ o vetor das observações y_i, $i = 1, \ldots, n$ e $\boldsymbol{\mu}$ o vetor de médias μ_i.

Note que Δ_i, pode ser escrito como

$$\Delta_i = \frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right).$$

Logo,

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)x_{ij}\frac{1}{V(\mu)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \qquad j = 1, 2, \ldots, p \tag{12}$$

e, fazendo-se $W_i = \dfrac{1}{V(Y_i)}\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2$, tem-se

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n}(y_i - \mu_i)W_i\left(\frac{\partial \eta_i}{\partial \mu_i}\right) x_{ij} \qquad j = 1, 2, \ldots, p$$

que, na forma matricial, fica

$$\boldsymbol{U = X'W\Delta^*(y - \mu)}$$

sendo $\Delta^* = diag\left\{\dfrac{\partial \eta_i}{\partial \mu_i}\right\}$.

Em geral, as equações em (12) são não-lineares em $\boldsymbol{\beta}$ (Dobson, 2001), de forma que para a sua solução são necessários procedimentos iterativos, sendo Newton-Raphson e o escore de Fisher os mais utilizados.

Usando-se Newton-Raphson, tem-se

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\boldsymbol{I}^{(m)}\right]^{-1} \boldsymbol{U}^{(m)},$$

sendo $\boldsymbol{U}^{(m)}$ o vetor escore, com elementos $\left(\dfrac{\partial \ell}{\partial \beta_j}\right)$ e $\boldsymbol{I}^{(m)}$ a matriz de informação observada com elementos $\left(-\dfrac{\partial^2 \ell}{\partial \beta_j \partial \beta_{j'}}\right)$, avaliados em $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$. Trocando-se a matriz de informação observada I pela matriz de informação esperada de Fisher $\Im$, tem-se a solução pelo método escore de Fisher, isto é,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left[\Im^{(m)}\right]^{-1} \boldsymbol{U}^{(m)}.$$

Para modelos lineares generalizados, o método escore de Fisher é particularmente simples devido à relação

$$E\left(-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_{j'}}\right) = E\left[\left(\frac{\partial \ell}{\partial \beta_j}\right)\left(\frac{\partial \ell}{\partial \beta_{j'}}\right)'\right] = E\big[U_j\, U_j'\big] = \Im$$

e, portanto,

$$E\left(-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_{j'}}\right) = \sum_{i=1}^{n} \frac{x_{ij}\; x_{ij'}}{V(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,$$

que na forma matricial, é $\boldsymbol{X'WX}$.

Logo,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \big[\boldsymbol{X'W^{(m)}X}\big]^{-1} \boldsymbol{U}^{(m)}$$

ou, então,

$$\boldsymbol{\beta}^{(m+1)} \;\; = \;\; \big[\boldsymbol{X'W^{(m)}X}\big]^{-1} \boldsymbol{X'W^{(m)}t^{(m)}},$$

sendo, $\boldsymbol{t}^{(m)} = \boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{\Delta}^{*(m)}(\boldsymbol{y} - \boldsymbol{\mu})^{(m)}$ conhecido como vetor de variáveis dependentes ajustadas, e cujos elementos são dados por

$$t_i = \boldsymbol{x}_i'\boldsymbol{\beta}_j + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right).$$

Este procedimento, apresentado em McCullagh & Nelder (1989) e Demétrio (2001), é conhecido como método dos mínimos quadrados ponderados iterativos. Ele exige alguns valores iniciais para $\boldsymbol{\beta}$ que podem ser obtidos a partir das estimativas de μ_i baseadas nos valores observados y_i. Considerando-se a expansão de $g(y_i)$ em série de Taylor em torno de μ_i, isto é

$$\begin{aligned} g(y_i) &\approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) \\ &\approx \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \\ &\approx t_i \end{aligned}$$

vê-se que t_i é uma aproximação local para $g(y_i)$.

Para se obterem as distribuições amostrais assintóticas, usam-se as aproximações em série de Taylor até termos de 1ª ordem, assim:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{U}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\mathfrak{F}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$$

mas, $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = 0$, pois $\widehat{\boldsymbol{\beta}}$ é o estimador que maximiza o logaritmo da função de verossimilhança.

Portanto,

$$\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{\mathfrak{F}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\mathfrak{F}}^{-1}\boldsymbol{U}.$$

Logo,

$$E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = E(\boldsymbol{\mathfrak{F}}^{-1}\boldsymbol{U}) = \boldsymbol{\beta},$$

ou seja, $\widehat{\boldsymbol{\beta}}$ é um estimador assintoticamente não viesado para $\boldsymbol{\beta}$.

A matriz de variâncias e covariâncias de $\boldsymbol{\beta}$ é dada por:

$$\begin{aligned} V = Cov(\hat{\boldsymbol{\beta}}) &= E\Big[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\Big] = E\Big[\boldsymbol{\mathfrak{F}}^{-1}\boldsymbol{U}\boldsymbol{U}'\boldsymbol{\mathfrak{F}}^{-1}\Big] \\ &= \boldsymbol{\mathfrak{F}}^{-1}\boldsymbol{E}\Big[\boldsymbol{U}\boldsymbol{U}'\Big]\boldsymbol{\mathfrak{F}}^{-1} = \boldsymbol{\mathfrak{F}}^{-1}. \end{aligned}$$

Também, pode-se escrever que:

$$\Im = E\Big[UU'\Big] = E\Bigg\{ \left[\sum_{i=1}^{n} \frac{1}{a_i(\phi)} X_i'\Delta_i(Y_i-\mu_i)\right]\left[\sum_{i=1}^{n} \frac{1}{a_i(\phi)}(Y_i-\mu_i)\Delta_i X_i\right]\Bigg\}$$

$$= \sum_{i=1}^{n} \frac{1}{\Big[a_i(\phi)\Big]^2}\, X_i'\Delta_i\, E\Big[Y_i-\mu_i\Big]^2\, \Delta_i X_i.$$

Como, $E\Big[Y_i-\mu_i\Big]^2 = V(Y_i) = Cov\big(Y_i,Y_i\big) = a_i(\phi)b''(\theta_i)$, então

$$\Im = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\, X_i'\Delta_i\; b''(\theta_i)\; \Delta_i X_i$$

e, fazendo-se $A_i = diag\Big\{b''(\theta)\Big\}$, tem-se que

$$\Im = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}\, X_i'\Delta_i\; A_i\; \Delta_i X_i.$$

Portanto, $V(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Im^{-1} E\Big[UU'\Big]\Im^{-1}}$ pode ser escrita como

$$V(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{n} \frac{1}{a_i(\phi)}\, X_i'\Delta_i\; A_i\; \Delta_i X_i\right]^{-1} \left[\sum_{i=1}^{n} \frac{1}{[a_i(\phi)]^2}\, X_i'\Delta_i\; Cov(Y_i)\; \Delta_i X_i\right] \left[\sum_{i=1}^{n} \frac{1}{a_i(\phi)}\, X_i'\Delta_i\; A_i\; \Delta_i X_i\right]^{-1}.$$

A distribuição amostral assintótica para $\widehat{\boldsymbol{\beta}}$ (Demétrio, 2001) é dada por

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\, \Im(\widehat{\boldsymbol{\beta}})\; (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p,$$

ou, de forma equivalente,

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \Im^{-1})$$

que é a base para a construção de testes e intervalos de confiança para os modelos lineares generalizados.

As propriedades assintóticas do estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ e as condições de regularidades, dado em Liang & Zeger (1986), podem ser encontradas em Artes (1997).

2.4.2 *Deviance* e seleção de modelo

Segundo McCullagh & Nelder (1989), o ajuste de um modelo a um conjunto de dados observados $\boldsymbol{y}$ pode ser considerado como uma maneira de se substituir $\boldsymbol{y}$ por um conjunto de valores estimados $\widehat{\boldsymbol{\mu}}$ para um modelo com um número de parâmetros relativamente pequeno.

Dadas n observações podem-se ajustar modelos contendo até n parâmetros. O modelo mais simples, o modelo nulo, tem apenas um parâmetro, representando um μ comum para todos os $y's$; o modelo saturado tem n parâmetros, um por observação. O modelo saturado atribui toda a variabilidade dos $y's$ ao componente sistemático e é, portanto, não-informativo.

O ajuste de um modelo a um conjunto de dados pode ser medido por:

$$D = -2\Big[\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})\Big]$$

sendo $\ell(\boldsymbol{y}, \phi, \boldsymbol{y})$ o valor do logaritmo da função de verossimilhança do modelo saturado e $\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y})$ o valor do logaritmo da função de verossimilhança do modelo corrente (sob pesquisa), maximizados sob $\boldsymbol{\beta}$ para um valor fixo do parâmetro de dispersão ϕ. Se $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{\mu})$ e $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{y})$ são as estimativas do parâmetro canônico sob os dois modelos, então

$$\frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{\phi} = \sum_{i=1}^{n} \frac{2w_i}{\phi} \Big\{ y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i) \Big\}$$

sendo $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ conhecida como *deviance* para o modelo corrente. Note-se que para as distribuições binomial e Poisson tem-se $\phi = 1$.

Quando ϕ não é conhecido, pode-se estimá-lo e usá-lo para calcular a *scaled deviance*, dada por:

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = \frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{\widehat{\phi}}. \tag{13}$$

Outra medida importante de discrepância é a estatística X^2 generalizada de Pearson, que tem a forma

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

sendo $V(\widehat{\mu})$ uma função de variância estimada para a distribuição sob estudo.

Tanto a *deviance* como a X^2 generalizada de Pearson têm distribuição χ^2 exata para modelos lineares clássicos (distribuição normal) e resultados assintóticos são obtidos para outras distribuições.

A *deviance* tem uma vantagem geral como uma medida de discrepância, em que ela é aditiva para conjuntos aninhados de modelos.

2.4.3 Superdispersão

A utilização de MLG na análise de dados tem se tornado de uso freqüente, principalmente com o avanço dos recursos computacionais disponíveis. Um dos cuidados que se deve tomar na análise de dados, principalmente no caso de variáveis discretas, é com a superdispersão que pode ocorrer. Quando se assume que as observações seguem uma distribuição na família exponencial, a função de variância tem uma forma conhecida, por exemplo, $Var(Y_i) = \mu_i$ para distribuição de Poisson e $Var(Y_i) = \pi_i(1 - \pi_i)$ para dados binários, casos em que $\phi = 1$. Entretanto, para dados envolvendo contagens ou proporções, é comum ocorrer uma variabilidade observada maior do que aquela explicada pela distribuição na família exponencial. Hinde & Demétrio (1998b) apresentam algumas das possíveis causas de superdispersão, que são dadas a seguir.

i) variabilidade do material experimental, que pode ser devida à variabilidade individual, gerando um componente aleatório adicional que não é levado em conta na análise do modelo básico;

ii) correlação entre respostas individuais, que pode ocorrer entre indivíduos do mesmo grupo, como por exemplo, no estudo de doenças em plantas, pode haver uma correlação entre plantas da mesma unidade experimental;

iii) agrupamentos amostrais;

iv) dados de nível agregado, sendo que o processo de agregação pode levar a distribuições compostas e

v) variáveis não-observáveis omitidas e de alguma forma todas as outras categorias são casos especiais desta, geralmente, de uma forma mais complexa.

O fato de não se considerar a superdispersão na análise dos dados pode levar à estimação incorreta dos erros padrões, sendo os mesmos super ou sub-estimados e, conseqüentemente, uma avaliação incorreta da significância dos parâmetros de regressão individual.

A distribuição padrão para análise de dados de proporção é a distribuição binomial, enquanto que para contagens é a Poisson. Essas distribuições têm como pressuposições:

i) independência entre as observações e

ii) a mesma probabilidade de sucesso no caso de proporções, ou a mesma média no caso de contagens, para todos os indivíduos.

Se uma destas suposições não é satisfeita, a variação residual pode ser maior do que aquela predita pelo modelo, ou seja,

i) dados de proporção com $V(Y_i) > m_i\pi_i(1-\pi_i)$ e

ii) dados de contagem com $V(Y_i) > \mu_i$,

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

Diferentes modelos e métodos de estimação têm sido propostos na literatura para resolver o problema da superdispersão. Hinde & Demétrio (1998b) dividiram as diferentes formas de abordagens da superdispersão em dois grupos, a saber

i) assumir alguma forma mais geral para a função de variância, possivelmente incluindo parâmetros adicionais e

ii) assumir um modelo em dois-estágios para a resposta, ou seja, assumir que o parâmetro do modelo básico tenha alguma distribuição.

Modelos do tipo (i), em geral, não correspondem a qualquer distribuição de probabilidade, mas são vistos como extensões do modelo básico. A estimação dos parâmetros de regressão é feita usando-se métodos de quase-verossimilhança, que pressupõem apenas uma relação média-variância para a resposta.

No caso em que a variável resposta é uma proporção, se for identificada a superdispersão, pode-se incluir o parâmetro extra-binomial da seguinte forma: considere que a i-ésima resposta ($1 \leq i \leq n$) é uma contagem de Y_i sucessos e $(m_i - Y_i)$ falhas. Associada a esta resposta tem-se uma matriz $\boldsymbol{X}$, de dimensões $n \times p$, de variáveis explanatórias. O modelo logístico-linear assume que os Y_i são independentes e têm distribuição binomial (m_i, π_i), em que

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \quad e \quad \eta_i = \mathbf{X}'\boldsymbol{\beta}.$$

Para se introduzir variação extra-binomial, usam-se variáveis contínuas não observáveis P_i independentemente distribuídas sobre (0, 1) com $E(P_i) = \pi_i$ e $V(P_i) = \phi\pi_i(1 - \pi_i)$ e assume-se que, condicional a $P_i = p_i$, Y_i é distribuído como binomial (m_i, p_i). Incondicionalmente,

$$\begin{aligned} E(Y_i) &= m_i\pi_i \quad e \\ V(Y_i) &= m_i\pi_i(1 - \pi_i)[1 + \phi(m_i - 1)] = m_i\pi_i(1 - \pi_i)\sigma^2, \end{aligned}$$

sendo $\sigma^2 = [1 + \phi(m_i - 1)]$, o fator de inflação da variância binomial. Um caso especial deste modelo é assumir que P_i tem uma distribuição beta e, assim, Y_i terá uma distribuição beta-binomial e o modelo pode ser ajustado pelo método da máxima verossimilhança.

Hinde & Demétrio (1998b) apresentam formas de variância geral para dados de proporção e contagens.

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

$$Var(Y_i) = m_i\pi_i(1-\pi_i)\left[1+\phi(m_i-1)^{\delta_1}\Big\{\pi_i(1-\pi_i)\Big\}^{\delta_2}\right]. \tag{14}$$

Várias funções de variância são definidas para diferentes valores de δ_1, δ_2 e ϕ em (14), a saber:

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

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

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

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

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

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

$$Var(Y_i) = m_i\pi_i(1-\pi_i)\Big[1+(m_i-1)\phi\Big], \tag{15}$$

que é o Modelo II de Williams (Williams, 1982), sendo que a função de variância da distribuição beta-binomial é um caso especial deste modelo e

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

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

que é o Modelo III de Williams ou logístico normal.

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

$$Var(Y_i) = m_i\Big\{1+\phi\mu_i^{\delta}\Big\}. \tag{16}$$

Várias funções de variância são definidas para diferentes valores de δ e ϕ em (16), a saber:

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

$$Var(Y_i) = \mu_i;$$

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

$$Var(Y_i) = \mu_i\Big[1 + \phi\Big],$$

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

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

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

que é a função de variância da distribuição binomial negativa.

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

Esses modelos são discutidos em Hinde & Demétrio (1998a,b) e também em McCulloch & Searle (2000). Alguns exemplos desses modelos são dados a seguir.

2.4.3.1 Modelo Binomial Negativo

Neste caso, assume-se que $Y_i \mid \theta_i \sim Poisson(\theta_i)$ e que os $\theta_i's$ são variáveis aleatórias com $E(\theta_i) = \mu_i$ e $V(\theta_i) = \sigma_i^2$. Um caso particular é considerar que $\theta_i \sim \Gamma(k, \lambda_i)$, levando à uma distribuição binomial negativa para os $Y_i's$, cuja esperança e variância são dadas por

$$E(Y_i) = \frac{k}{\lambda_i} = \mu_i \quad e \quad V(Y_i) = \mu_i + \frac{\mu_i^2}{k}.$$

Para valores fixos de k, esta distribuição pertence à família exponencial na estrutura de modelos lineares generalizados.

2.4.3.2 Modelo Beta-binomial

Assume-se que $Y_i \mid P_i \sim Binomial(m_i, P_i)$, sendo que os $P_i's$ são condiderados variáveis aleatórias de forma que $P_i \sim Beta(\alpha_i, \beta_i)$. Assim, tem-se que, incondicionalmente

$$E(Y_i) = E\Big[E(Y_i \mid P_i)\Big] = m_i \left(\frac{\alpha_i}{\alpha_i + \beta_i}\right) = m_i \pi_i$$

e

$$\begin{aligned} V(Y_i) &= V\Big[E(Y_i \mid P_i)\Big] + E\Big[V(Y_i \mid P_i)\Big] \\ &= m_i \left(\frac{\alpha_i}{\alpha_i + \beta_i}\right)\left(\frac{\beta_i}{\alpha_i + \beta_i}\right)\left[1 - (1 - m_i)\frac{1}{\alpha_i + \beta_i + 1}\right]. \end{aligned}$$

Fazendo-se,

$$\left(\frac{\alpha_i}{\alpha_i + \beta_i}\right) = \pi_i \;\text{ e }\; \left(\frac{\beta_i}{\alpha_i + \beta_i}\right) = 1 - \pi_i,$$

e usando-se o fato de que, em aplicações do modelo beta-binomial é comum fazer $\alpha_i + \beta_i = c$, ou seja, constante sobre i, tem-se que:

$$V(Y_i) = m_i \pi_i (1 - \pi_i)\big[1 + (m_i - 1)\phi\big],$$

que é a mesma dada em (15).

2.4.3.3 Logístico-normal

A razão pela qual a distribuição normal não foi assumida para P_i em 2.4.3.2 é que ele é restrito ao intervalo $(0, 1)$. Uma abordagem alternativa é transformar P_i, usando $logit\ (P_i) \equiv \log\left(\frac{P_i}{1 - P_i}\right)$, sendo que a variação do $logit\ (P_i)$ é de $(-\infty, \infty)$, logo, a distribuição normal pode ser assumida para ele, isto é,

$logit\ (P_i) \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$. Seja $\eta_i = logit(P_i) = \log\left(\frac{P_i}{1-P_i}\right)$. A média e a variância de Y_{ij} são dadas por

$$\begin{aligned} E(Y_{ij}) &= E\Big[E(Y_{ij} \mid \pi)\Big] \\ &= \int_{-\infty}^{\infty} \frac{e^{\eta_i}}{1+e^{\eta_i}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\pi\sigma^2}(\eta_i-\mu)^2} d\eta_i, \end{aligned}$$

sendo η_i o preditor linear, $\eta_i = x_i'\beta + \sigma z_i$. Fazendo-se uma transformação de variável em função de z_i, tem-se

$$E(Y_{ij}) = \int_{-\infty}^{\infty} \frac{e^{\mu+\sigma z_i}}{1+e^{\mu+\sigma z_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i,$$

que não pode ser calculado de forma explícita, mas pode ser aproximado por métodos numéricos como, por exemplo, o método de quadratura Gauss-Hermite.

A covariância por $Cov\Big(Y_{ij}, Y_{il}\Big) = E(Y_{ij}Y_{il}) - E(Y_{ij})E(Y_{il})$, sendo que

$$E(Y_{ij}Y_{il}) = \int_{-\infty}^{\infty} \left[\frac{e^{\mu+\sigma z_i}}{1+e^{\mu+\sigma z_i}}\right] \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i.$$

Hinde & Demétrio (1998) consideram um modelo em dois-estágios para o $logit(P_i)$, escrevendo

$$U_i = \log\left(\frac{P_i}{1-P_i}\right) \quad \Rightarrow \quad P_i = \frac{e^{U_i}}{1+e^{U_i}}.$$

Usando a expansão em série de Taylor para P_i em torno de $U_i = E(U_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, tem-se

$$E(P_i) \approx \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} := \pi_i,$$

e

$$V(P_i) \approx \left[\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right]^2 V(U_i) := \sigma^2\pi_i^2(1-\pi_i)^2.$$

Os autores aproximam a função de variância para o modelo logístico-normal por

$$V(Y_{ij}) \approx m_i\pi_i(1-\pi_i)\Big[1+\phi(m_i-1)\pi_i(1-\pi_i)\Big],$$

que é a função de variância tipo III de Williams.

2.4.4 Quase-verossimilhança

A estimação dos efeitos fixos dos modelos lineares generalizados é baseada na função de verossimilhança. Uma extensão do método de máxima verossimilhança para ajustar efeitos aleatórios é a quase-verossimilhança. Ela é introduzida por Wedderburn (1974) e é definida da forma que se segue. Suponha que y_i (i = 1, 2,..., n) seja um conjunto de observações com $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ e $V(\boldsymbol{Y}_i) \propto V(\mu_i)$, em que V($\boldsymbol{\mu}$) é alguma função conhecida. Também suponha que $\boldsymbol{\mu}_i$ seja uma função de um conjunto de parâmetros $\beta_1, \ldots, \beta_p$. A função de quase-verossimilhança Q(μ_i, y_i) é definida pela relação

$$\frac{\partial Q(\mu_i, y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)}, \tag{17}$$

ou, eqüivalentemente,

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \frac{y_i - \mu_i^{'}}{V(\mu_i^{'})}\ d\mu_i^{'}.$$

O logaritmo da função de verossimilhança é um caso especial da função de quase-verossimilhança. Wedderburn (1974) mostra que se pode usar qualquer função $Q(\mu_i, y_i)$ que satisfaça (17) como uma base para definir um modelo linear generalizado e obter estimativas de β_i pelos procedimentos conhecidos. A teoria de quase-verossimilhança está baseada apenas na suposição de formas para o primeiro e segundo momentos. Assim, não é necessário especificar completamente a distribuição de probabilidade da variável resposta.

A relação entre a média e a variância de Y_i permite a definição de uma quase-verossimilhança que é maximizada em relação aos parâmetros $\boldsymbol{\beta}$ pelo uso iterativo das equações de mínimos quadrados ponderados:

$$\boldsymbol{X'W\Delta X\hat{\beta}} = \boldsymbol{X'W\Delta Y},$$

em que $\boldsymbol{W} = diag(W_i)$.

Logo, a quase-verossimilhança estende os modelos lineares generalizados a dados cuja distribuição não pertença à família exponencial ou mesmo que não tenha uma descrição exata.

Williams (1982) mostra como uma estimativa, $\hat{\phi}$, do parâmetro ϕ pode ser determinada, igualando-se o valor da estatística X^2, para o modelo, ao seu valor esperado aproximado. O valor de X^2, para um dado modelo, depende do valor de $\hat{\phi}$, sendo assim, o processo de estimação iterativo (Collett, 1991).

Hinde & Demétrio (1998a,b) mostram que o parâmetro de dispersão constante estimado para o modelo binomial é dado por:

$$\widehat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - m_i\widehat{\pi}_i)^2}{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)},$$

enquanto que para o modelo de Poisson é dado por:

$$\widehat{\phi} = \frac{1}{(n-p)} \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}.$$

2.4.5 Método de quadratura Gaussiana

Seja y_{ij} a j-ésima observação ($j = 1, \ldots, k$), correspondendo ao i-ésimo nível ($i = 1, \ldots, n$) do efeito aleatório, de forma que

$$\begin{aligned}
Y_{ij} \mid \boldsymbol{u} &\sim indep. f_{Y_{ij}|\mathbf{U}}(y_{ij} \mid \boldsymbol{u}) \\
f_{Y_{ij}|\boldsymbol{u}}(y_{ij} \mid \boldsymbol{u}) &= \exp\left\{ \frac{1}{a_i(\phi)}\Big[y_{ij}\theta_{ij} - b(\theta_{ij})\Big] + c(y_{ij};\phi) \right\} \\
E[y_{ij} \mid \boldsymbol{u}] &= \mu_{ij} \\
\eta_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u} \\
u_i &\sim \text{ i.i.d. } N(0, \sigma_u^2)
\end{aligned}$$

Note que u_i pode assumir qualquer distribuição de probabilidade mas que, em geral, assume-se distribuição normal. A função de verossimilhança para este modelo é dada por:

$$\begin{aligned}
L &= \prod_{i,j} f(y_{ij}, u_i) = \int \prod_{i,j} f_{Y_{ij}|U_i}(y_{ij} \mid u_i) f_{U_i}(u_i) du_i \qquad (18) \\
&= \prod_{i} \int_{-\infty}^{\infty} h_i(u_i) \frac{e^{-u_i^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}} du_i,
\end{aligned}$$

sendo

$$h_i(u_i) = e^{\sum_j \frac{1}{a_i(\phi)}\left[y_i\theta_i - b(\theta_i)\right] + \sum_j c(y_i;\phi)}.$$

Deve-se notar que o logaritmo da função de verossimilhança é o produto de integrais unidimensionais da forma:

$$\int_{-\infty}^{\infty} h(u)\frac{e^{-u^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}}du,$$

que pode ser escrita, após a transformação $u = \sqrt{2}\sigma_u v$, como:

$$\int_{-\infty}^{\infty} h(\sqrt{2}\sigma_u v)\frac{e^{-v^2}}{\sqrt{\pi}}dv \equiv \int_{-\infty}^{\infty} h^*(v)e^{-v^2}dv, \tag{19}$$

sendo $h^*(\cdot) \equiv h(\sqrt{2}\sigma_u v)/\sqrt{\pi}$.

McCulloch e Searle (2001) apontam que a integração em (19) é, geralmente, complicada de se fazer, principalmente se houver muitos efeitos aleatórios a serem estimados. Uma forma de contornar o problema é utilizar o método de quadratura Gauss-Hermite (ou Gaussiana), que aproxima a integral definida em (19) como uma soma ponderada:

$$\int_{-\infty}^{\infty} h^*(v)e^{-v^2}dv \doteq \sum_{k=1}^{d} h^*(x_k)w_k,$$

em que os pesos w_k e os pontos de quadratura x_k são atribuídos para fornecer uma aproximação precisa.

No caso da função de densidade normal, tem-se que a aproximação da integral fica:

$$\int_{-\infty}^{\infty} h(u)\frac{e^{-u^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}}du \doteq \sum_{k=1}^{d} h(\sqrt{2}\sigma x_k)w_k/\sqrt{\pi}.$$

Observe que para a utilização do algoritmo EM, declara-se $\boldsymbol{u}$ como um vetor de dados perdidos, de forma que os dados completos são representados por $\boldsymbol{w}' = (\boldsymbol{y}', \boldsymbol{u}')$. No algoritmo EM utiliza-se o logaritmo da função de verossimilhança dos dados completos, calculando-se sua esperança em relação à distribuição condicional de $\boldsymbol{u}$ dado $\boldsymbol{y}$ e então maximizando-se a função em relação aos parâmetros.

A distribuição dos dados completos $\boldsymbol{w}$, pode ser fatorada de forma que o logaritmo da função de verossimilhança dos dados completos, L_w, seja,

$$\begin{aligned}\log L_w &= \log f_{\mathbf{Y}|\mathbf{U}} + \log f_{\mathbf{U}} \\ &= \sum_{i=1}^{n} \log f_{Y_i|\mathbf{U}} + \log f_{\mathbf{U}} \\ &= \sum_i \frac{1}{a_i(\phi)}\Big[y_i\theta_i - b(\theta_i)\Big] + \sum_i c(y_i;\phi) + \log f_{\mathbf{U}}.\end{aligned}$$

Esta escolha tem duas vantagens

1) condicional a $\boldsymbol{u}$, os y_i são independentes e

2) $\boldsymbol{\beta}$ *e* $a(\phi)$ entram somente na primeira parte do logaritmo da função de verossimilhança (parte do GLM), enquanto $\boldsymbol{G}$, a matriz de variâncias-covariâncias dos efeitos aleatórios, entra somente através de $f_{\mathbf{U}}$, a parte dos efeitos aleatórios.

Algoritmo EM para o modelo logístico-normal

O algoritmo EM pode ser utilizado no caso do modelo logístico-normal para se obter as estimativas dos parâmetros da seguinte forma: seja o modelo dado por

$$\begin{aligned}Y \mid X_1, X_2, \cdots, X_p, Z &\sim Binomial(n, \pi) \\ Z &\sim N(0,1),\end{aligned}$$

com

$$\log\left(\frac{P_i}{1-P_i}\right) = \boldsymbol{X'\beta} + \sigma Z \quad \Rightarrow \quad P_i = \frac{e^{\boldsymbol{X'\beta}+\sigma Z}}{1+e^{\boldsymbol{X'\beta}+\sigma Z}},$$

e seja um conjunto de dados com n observações $y_i, i = 1, 2, \cdots, n$, para a variável resposta Y e com valores $x_{i1}, x_{2i}, \cdots, x_{pi}$ para as variáveis explanatórias $X_1, X_2, \cdots, X_p$, sendo Z uma variável latente.

Seja $\boldsymbol{\psi} = (\boldsymbol{\beta'}, \sigma)'$, o vetor de parâmetros desconhecidos do modelo

composto. Logo, a distribuição condicional conjunta das observações é dada por:

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\psi}) &= \prod_{i=1}^{n} \binom{n}{y} \pi^{y} (1-\pi)^{n-y} \\ &= \prod_{i=1}^{n} \binom{n}{y} \left[\frac{e^{\boldsymbol{X'\beta}+\sigma Z}}{1+e^{\boldsymbol{X'\beta}+\sigma Z}} \right]^{y} \left[1 - \frac{e^{\boldsymbol{X'\beta}+\sigma Z}}{1+e^{\boldsymbol{X'\beta}+\sigma Z}} \right]^{n-y}. \end{aligned}$$

O logaritmo da função de verossimilhança dos dados completos é dada por:

$$\begin{aligned} \log L(y, z \mid x, \psi) &= \log f(\boldsymbol{Y} \mid x, z, \psi) + \log g(z) \\ &= \sum_{i=1}^{n} \left\{ y \log \left(\frac{e^{\eta}}{1+e^{\eta}} \right) + (n-y) \log \left(1 - \frac{e^{\eta}}{1+e^{\eta}} \right) \right\} + \\ &\quad \log \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} z_i^2} \right] \\ &= \sum_{i=1}^{n} \left\{ y\eta - n \log \left(1 + e^{\eta} \right) \right\} - \frac{1}{2} \sum_{i=1}^{n} \left\{ z_i^2 + \log(2\pi) \right\}. \end{aligned}$$

Assim, os passos E e M consistem em:

Passo E: Cálculo da esperança do $\log f(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{X}, \boldsymbol{\psi})$:

$$\begin{aligned} Q(\psi \mid \psi^{s}) &= E\Big[\log f(y, z \mid x, \psi) \mid y, x, \psi^{s} \Big] \\ &= \sum_{i=1}^{n} E\Big[\log f(y, z \mid x, \psi) \mid y, x, \psi^{s} \Big] \\ &= \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \Big[\log f(y, z \mid x, \psi) f(z \mid y, x, \psi^{(s)}) dz \Big] \right\}, \end{aligned}$$

mas,

$$f(z \mid y, x, \psi^{(s)}) = \frac{f(y, z \mid x, \psi^{(s)})}{f(y \mid x, \psi^{(s)})} = \frac{f(y \mid x, z, \psi)\phi(z)}{f(y \mid x, \psi^{(s)})},$$

portanto,

$$Q(\psi \mid \psi^{(s)}) = \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \Big[\log f(y, z \mid x, \psi) \Big] \frac{f(y \mid x, z, \psi)\phi(z)}{f[y \mid x, \psi^{(s)}]} \, dz_i \right\},$$

sendo que ϕ representa a função de densidade de probabilidade da distribuição normal padrão, e,

$$f\Big[y \mid x, \psi^{(s)} \Big] = \int_{-\infty}^{\infty} f\Big[y \mid x, z, \psi^{(s)} \Big] \phi(z) dz_i.$$

Logo,

$$Q(\psi \mid \psi^{(s)}) = \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \left[y\eta - n\log\left(1+e^{\eta}\right) - \frac{1}{2}\left\{z_i^2 + \log(2\pi)\right\} \right] \right\} \times \frac{f(y \mid x, z, \psi)\phi(z)\, dz}{\int_{-\infty}^{\infty} f(y \mid x, z, \psi^{(s)})\phi(z)\, dz}.$$

Usando-se $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} z_i^2}$, tem-se

$$Q(\psi \mid \psi^{(s)}) = \sum_{i=1}^{n} \left\{ \int_{-\infty}^{\infty} \left[y\eta - n\log\left(1+e^{\eta}\right) - \frac{1}{2}\left\{z_i^2 + \log(2\pi)\right\} \right] \right\} \times \frac{f(y \mid x, z, \psi)\, e^{-\frac{1}{2} z_i^2}}{\int_{-\infty}^{\infty} f(y \mid x, z, \psi^{(s)})\, e^{-\frac{1}{2} z_i^2} dz}\, dz. \tag{20}$$

Usando-se quadratura para aproximar a integral, tem-se

$$Q(\psi \mid \psi^{(s)}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k} \left[y\eta - n\log\left(1+e^{\eta}\right) - \frac{1}{2}\left[z_i^2 + \log(2\pi)\right] \frac{f(y \mid x, z, \psi)\, \pi_j}{\sum_{j=1}^{k} f(y \mid x, z, \psi^{(s)})\, \pi_j} \right] \right\}.$$

Passo M: Consiste em se igualarem a zero as derivadas parciais de $Q(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(s)})$ em relação aos parâmetros que constituem o vetor $\boldsymbol{\psi}$, logo:

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q\left(\psi \mid \psi^{(s)}\right) = \sum_i \sum_j \left\{ \left[y - n\left(\frac{e^{\eta}}{1+e^{\eta}}\right) \right] x \frac{f(y \mid x, z, \psi)\, \pi_j}{\sum_{j=1}^{k} f(y \mid x, z, \psi^{(s)})\, \pi_j} \right\}$$

e

$$\frac{\partial}{\partial \boldsymbol{\sigma}} Q\left(\psi \mid \psi^{(s)}\right) = \sum_i \sum_j \left\{ \left[y - n\left(\frac{e^{\eta}}{1+e^{\eta}}\right) \right] z \frac{f(y \mid x, z, \psi)\, \pi_j}{\sum_{j=1}^{k} f(y \mid x, z, \psi^{(s)})\, \pi_j} \right\}.$$

2.4.6 Dados longitudinais

Na seção (2.2.3) discute-se o caso de medidas repetidas, ou de dados longitudinais, na análise de dados com distribuição normal. Tais estruturas também têm que ser levados em consideração quando se utiliza a teoria de modelos lineares generalizados para dados longitudinais em que a distribuição assumida não seja normal.

Os modelos lineares generalizados são propostos para análise de dados com apenas uma observação por indivíduo os quais assumem independência entre as observações. Um dos primeiros estudos considerando dependência entre as observações e incluindo essa correlação no modelo é feito por Zeger, Liang & Self (1985) que consideram extensões da regressão logística para o caso em que a variável resposta binária é observada repetidamente para cada indivíduo. Eles propõem dois modelos que levam a estimativas consistentes dos parâmetros de regressão e de suas variâncias sob suposições moderadas sobre a dependência do tempo dentro de cada indivíduo. O primeiro modelo é deduzido a partir de uma suposição de que observações repetidas para um indivíduo são independentes umas das outras, dando estimativas que são consistentes, dado qualquer conjunto de processos binários estacionários, tal que $logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i'\boldsymbol{\beta}$, em que x_i' é um vetor $1 \times p$ de covariáveis para o i-ésimo indivíduo (i = 1, 2,...,n) e $\boldsymbol{\beta}$ é o vetor de parâmetros a ser estimado. O segundo modelo é obtido supondo-se que cada série é uma Cadeia de Markov estacionária de ordem um, dando estimativas consistentes quando para as séries é feita a suposição de função de ligação *logit* e seus elementos têm uma autocorrelação de primeira ordem comum, $\rho = corr(Y_{it}, Y_{i,t-1})$, sendo t, o tempo. Essa abordagem para análise de dados longitudinais binários exige covariáveis de tempo independentes, ou seja, as covariáveis observadas no início do estudo não se alteram até o final do mesmo.

Dados perdidos, intervalos não eqüidistantes e especificação de condições iniciais podem causar problemas quando se usam os modelos de Markov, dificultando a interpretação dos efeitos das covariáveis. Devido a esse problema,

Liang & Zeger (1986) introduzem as Equações de Estimação Generalizadas (EEG), em que apenas o primeiro e segundo momentos da função densidade marginal precisam ser definidos. A abordagem é feita para a distribuição marginal ao invés da condicional dadas as covariáveis, embora a distribuição condicional possa ser mais apropriada para alguns problemas. As equações de estimação são extensões das equações de quase-verossimilhança (Wedderburn, 1974) para o caso em que o segundo momento não pode ser completamente especificado em termos da média, mas parâmetros adicionais de correlação devem ser incluídos. Um aspecto das EEG é a necessidade de especificação de uma estrutura de variâncias-covariâncias de trabalho (ou manipulada), ou seja, em alguns casos suspeita-se que uma forma específica de covariância pode ocorrer nos dados. Ela é introduzida nos cálculos da forma que se segue.

Para estabelecer notação, seja $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iti})'$ o vetor de dimensões $n_i \times 1$ de respostas e $\boldsymbol{X_i} = (x_{i1}, x_{i2}, \ldots, x_{it_i})'$ a matriz de dimensões $n_i \times p$ de covariáveis para o i-ésimo indivíduo $(i = 1, 2, \ldots, n)$ para a j-ésima observação (ou tempo), $j = 1, \ldots, t_i$. Assume-se que a densidade marginal de y_{ij} é semelhante à equação dada por (7) e que os dois primeiros momentos de Y_{ij} são semelhantes àqueles dados por (10). Deve-se notar a inclusão do índice j, que se refere às observações repetidas.

As equações de estimação podem levar em conta a correlação para aumentar (melhorar) a eficiência dos estimadores. Os estimadores de $\boldsymbol{\beta}$ permanecem consistentes. Além disso, estimativas consistentes de variância estão disponíveis sob a suposição fraca de que uma média ponderada da matriz de correlação estimada converge para uma matriz fixa.

Seja $\boldsymbol{R(\alpha)}$ uma matriz simétrica de dimensões $n_i \times n_i$ que satisfaz as condições de uma matriz de correlação e, seja α um vetor de dimensões $n_i \times 1$ que caracteriza $\boldsymbol{R(\alpha)}$ completamente. $\boldsymbol{R(\alpha)}$ é conhecida como matriz correlação de "trabalho".

Define-se

$$\boldsymbol{V}_i = a_i(\phi)\boldsymbol{A}_i^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2} \tag{21}$$

que será igual a $Cov(\boldsymbol{Y}_i)$, se $\boldsymbol{R}(\boldsymbol{\alpha})$ é, de fato, a matriz de correlação verdadeira para os $\boldsymbol{Y}_{i's}$.

Assim, definem-se as equações de estimação como sendo

$$\sum_{i=1}^{n} \boldsymbol{D'}_i\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_{it} - \boldsymbol{\mu}_{it}) = \boldsymbol{0}, \tag{22}$$

com $\boldsymbol{D}_i = \dfrac{\partial b'(\theta_i)}{\partial \beta_i} = \dfrac{\partial b'(\theta_i)}{\partial \theta_i}\dfrac{\partial \theta_i}{\partial \eta_i}\dfrac{\partial \eta_i}{\partial \beta_i} = \boldsymbol{A}_i\boldsymbol{\Delta}_i\boldsymbol{X}_i$. Deve-se observar que (22) reduz-se às equações de independência em (11) se $\boldsymbol{R}(\alpha)$ for a matriz identidade e que, para cada i, $\boldsymbol{U}_i(\boldsymbol{\beta}, \alpha) = \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{S}_i$ é similar à função de quase-verossimilhança (Wedderburn, 1974), exceto que os $\boldsymbol{V}_{i's}$ já não são somente em função de $\boldsymbol{\beta}$, mas também de $\boldsymbol{\alpha}$.

Substituindo-se $\boldsymbol{\alpha}$ em (21) e (22) por $\hat{\boldsymbol{\alpha}}(\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\phi})$, um estimador $n^{1/2}$ - consistente de $\boldsymbol{\alpha}$, quando $\boldsymbol{\beta}$ e ϕ são conhecidos. Para completar o processo, troca-se ϕ por $\hat{\phi}(\boldsymbol{Y}, \boldsymbol{\beta})$, um estimador $n^{1/2}$- consistente quando $\boldsymbol{\beta}$ é conhecido. Conseqüentemente, (22) tem a forma

$$\sum_{i=1}^{n} \boldsymbol{U}_i\left[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\left\{\hat{\boldsymbol{\beta}}, \hat{\phi}(\boldsymbol{\beta})\right\}\right] = 0, \tag{23}$$

e $\hat{\boldsymbol{\beta}}_G$, as estimativas dos parâmetros para dados correlacionados, é a solução da equação (23). Liang & Zeger (1986) apresentam, de forma suscinta, as condições de regularidade necessárias para a robustez e consistência desse estimador e mostram que $n^{\frac{1}{2}}\left(\hat{\boldsymbol{\beta}}_G - \boldsymbol{\beta}\right)$ é assintoticamente normal multivariado com média zero e matriz de variâncias-covariâncias, $\boldsymbol{V}_G$, dada por:

$$\boldsymbol{V}_G = \lim_{n\to\infty} n\left[\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1}\left\{\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}Cov(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right\}\left[\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1},$$

conhecida na literatura especializada como estimador (robusto) sanduíche empírico e que pode ser consistentemente estimada por

$$\hat{\boldsymbol{V}}_G = \lim_{n\to\infty} n\left[\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1}\left\{\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V_i}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right\}\left[\sum_{i=1}^{n} \boldsymbol{D'}_i\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right]^{-1}.$$

uma prova mais detalhada é encontrada em Artes (1997).

Escolha da matriz de correlações de trabalho

Os parâmetros de correlação $\boldsymbol{\alpha}$ e o parâmetro de escla ϕ podem ser estimados, em cada passo do processo iterativo, do corrente resíduo de Pearson

$$\hat{r}_{ij} = \frac{y_{ij} - b'(\hat{\theta}_{ij})}{\sqrt{b''(\hat{\theta}_{ij})}},$$

com $\hat{\boldsymbol{\theta}}$ dependendo da corrente estimativa de $\boldsymbol{\beta}$.

A estimativa de ϕ é dada por

$$\hat{\phi}^{-1} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2}{\sum_{i=1}^{n} t_i - p}.$$

A estimativa de $\boldsymbol{\alpha}$ depende da escolha de $\boldsymbol{R(\alpha)}$ e, em geral, $\boldsymbol{\alpha}$ é uma função simples de

$$\hat{\boldsymbol{R}}_{jj'} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}\hat{r}_{ij'}}{\sum_{i=1}^{n} t_i - p}.$$

O procedimento das EEGs para estimar $\boldsymbol{\beta}$ permite que a estrutura de correlação entre as observações da mesma unidade experimental seja especificada de diferentes formas. Algumas especificações para a estrutura da matriz de correlação de trabalho têm sido sugeridas por Liang & Zeger (1986). Considere o caso em que $n_i = 4$.

i) Independência: quando a matriz de correlação $\boldsymbol{R}(\alpha)$ é a identidade, isto é,

$$\boldsymbol{R}(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

obtém-se, então, equações de estimação independentes.

ii) Esférica: essa estrutura assume que $Corr(y_{ij}, y_{i,j'}) = \begin{cases} 1, \text{ se } j = j' \\ \alpha, \text{ se } j \neq j' \end{cases}$, assim

$$\boldsymbol{R}(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}.$$

Essa é a estrutura de correlação assumida em um modelo de efeitos aleatórios com um intercepto aleatório. Dado ϕ, α pode ser estimado por:

$$\hat{\alpha} = \phi \sum_{i=1}^{n} \sum_{j>j'} \frac{\hat{r}_{ij}\hat{r}_{ij'}}{\sum_{i=1}^{n} \frac{1}{2} t_i(t_i - 1) - p}.$$

Note que números arbitrários de observações e tempos de observações para cada indivíduo são possíveis com esta suposição.

iii) Não-estruturada: quando a matriz de correlação é totalmente não-estruturada, haverá $\frac{n_i(n_i - 1)}{2}$ parâmetros a serem estimados da forma:
$Corr(y_{ij}, y_{i,j'}) = \begin{cases} 1, \text{ se } j = j' \\ \alpha_{ij}, \text{ se } j \neq j' \end{cases}$, assim

$$\boldsymbol{R}(\alpha) = \begin{bmatrix} 1 & \alpha_{21} & \alpha_{31} & \alpha_{41} \\ \alpha_{21} & 1 & \alpha_{32} & \alpha_{42} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{41} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{bmatrix}.$$

Essa estrutura fornece o estimador mais eficiente para β, mas é útil somente quando há poucos tempos de observação.

iv) m-dependente: essa estrutura especifica que $Corr(y_{ij}, y_{i,j+1}) = \alpha_j$,

$j = 1, \ldots, t-1$ e $\alpha_j = (\alpha_1, \ldots, \alpha_{t-1})'$, assim

$$Corr(y_{ij}, y_{i,j+1}) \begin{cases} 1, \text{ se } j = 0 \\ \alpha_j, \text{ se } j = 1, 2, \ldots, t-1 \\ 0, \text{ se } j > n \end{cases} .$$

Com a estrutura m-dependente, a correlação depende das distâncias entre medidas; eventualmente elas tendem a zero para $t \geq n$. Um estimador natural para α_j, dado β e ϕ é

$$\hat{\alpha}_j = \phi \sum_{i=1}^{n} \frac{\hat{r}_{ij}\hat{r}_{i,j+1}}{n-p}.$$

v) Auto-regressiva (AR-1): seja $Corr(y_{ij}, y_{i,j'}) = \alpha^{|j-j'|}$, logo

$$\boldsymbol{R}(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^3 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}. \tag{24}$$

Com uma estrutura de correlação auto-regressiva, as correlações também dependem das distâncias entre as medidas, elas diminuem com o aumento das distâncias.

Para qualquer $\boldsymbol{R(\alpha)}$, $\hat{\boldsymbol{\beta}}_G$ e $\hat{\boldsymbol{V}}_G$ serão consistentes. Claro que a escolha de $\boldsymbol{R(\alpha)}$ mais próximo da correlação verdadeira aumenta a eficiência dos estimadores.

Considerando-se que a suposição de normalidade dos resíduos de Pearson não necessariamente ocorre para respostas discretas, tais como respostas Poisson e binárias, Park et al. (1998) estudam os resíduos de Ascombe e *deviances* para estimação dos parâmetros de correlação e comparam com os resíduos de Pearson. Os resultados mostram que a escolha do resíduo tinha pouco efeito (ou nenhum) nas propriedades das estimativas resultantes e, portanto, recomendam o uso dos resíduos de Pearson.

Embora as EEG estejam sendo usadas quando o objetivo principal é a análise de dados correlacionados com mais de uma observação por indivíduo, elas também podem ser usadas na análise de variáveis com uma única resposta por indivíduo para modelar a superdispersão. Com a EEG, está se usando uma medida que varia de indivíduo para indivíduo para estimação da variância ao invés da estimação baseada no modelo propriamente. Esta abordagem é usada por Stokes et al. (2000) no estudo da incidência de doença respiratória em crianças.

O uso das equações de estimação generalizadas tornou-se freqüente e os programas computacionais estatísticos trazem sua implementação devido aos estimadores robustos das variâncias dos parâmetros serem consistentes, mesmo se a matriz de correlação utilizada (por isso o nome: matriz de correlação de trabalho) for mal especificada. Testes tipo Wald podem ser usados para testar a homogeneidade entre grupos.

2.5 Modelos lineares generalizados mistos

A teoria de modelos lineares generalizados considera apenas o estudo de variáveis com efeitos fixos. Uma extensão natural são modelos que se ajustem a dados obtidos a partir de experimentos em que os níveis de um fator foram selecionados de uma população de níveis, isto é, são aleatórios.

Piepho (1999) discute a análise de dados de incidência de bolor penugento da uva (*Plasmopara viticola*). O autor acrescenta efeitos aleatórios ao modelo para controlar a superdispersão que existe em seus dados devido à incidência de doenças ocorrer de forma agrupada, levando à violação da suposição de independência. Um modelo logístico-normal para uma pesquisa clínica multicentro é apresentada por Wolfinger (1993) sendo as clínicas consideradas como fator aleatório.

Considere-se a distribuição condicional de $\boldsymbol{Y}$ dado $\boldsymbol{u}$, sendo $\boldsymbol{Y}$ o vetor de respostas assumido consistir de elementos condicionalmente independentes (não

necessariamente) com densidade pertencente à família exponencial.

$$\begin{aligned} Y_i \mid \boldsymbol{u} &\sim \quad indep.\ f_{Y_i|\boldsymbol{u}}(y_i \mid \boldsymbol{u}) \\ f_{Y_i|\boldsymbol{u}}(y_i \mid \boldsymbol{u}) &= \exp\left\{\frac{w_i}{\phi}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i;\phi)\right\}. \end{aligned} \tag{25}$$

De (10) tem-se que

$$\mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

portanto,

$$E(Y_i \mid \boldsymbol{u}) = \mu_i$$

e

$$g(\mu_i) = \boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u} \tag{26}$$

que é a função de ligação em (8) acrescentada a parte aleatória, sendo $\boldsymbol{x}_i'$ é a i-ésima linha da matriz do modelo para os efeitos fixos; $\boldsymbol{\beta}$ é o vetor de parâmetros com efeitos fixos; $\boldsymbol{z}_i'$ é a i-ésima linha da matriz do modelo para os efeitos aleatórios e $\boldsymbol{u}$ é o vetor de efeitos aleatórios.

Neste caso, μ_i é a média condicional de y_i dado $\boldsymbol{u}$. Para completar a especificação do modelo, atribui-se uma distribuição aos efeitos aleatórios

$$\boldsymbol{u} \sim f_U(u). \tag{27}$$

É usual atribuir distribuição normal aos efeitos aleatórios, ou seja, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{G})$. Uma discussão mais aprofundada sobre os efeitos aleatórios não-normais é dada por Lee & Nelder (1996), que usam distribuições conjugadas.

Considerando o modelo condicional (25), tem-se

$$E(Y_i) \quad = \quad E\big[E(Y_i \mid \boldsymbol{u})\big] = E[\mu_i] = E\big[g^{-1}(\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u})\big]$$

que, em geral, não pode ser simplificado devido à presença de funções não-lineares em $g^{-1}(\cdot)$.

A variância marginal de $\boldsymbol{y}$ é dada por

$$\begin{aligned} V(Y_i) &= V\big[E(Y_i \mid \boldsymbol{u})\big] + E\big[V(Y_i \mid \boldsymbol{u})\big] \\ &= V[\mu_i] + E\big[a(\phi)V(\mu_i)\big] \\ &= V\big[g^{-1}(\boldsymbol{x_i'\beta} + \boldsymbol{z_i'u})\big] + E\Big\{a_i(\phi)V\big[g^{-1}(\boldsymbol{x}_i'\beta + \boldsymbol{z}_i'\boldsymbol{u})\big]\Big\}, \end{aligned}$$

sendo $a_i(\phi) = \dfrac{w_i}{\phi}$, não sendo possível simplificações sem fazer suposições específicas sobre a forma de $g(\cdot)$ e/ou a distribuição condicional de $\boldsymbol{Y}$.

O uso de efeitos aleatórios também introduz uma correlação entre observações que tenham algum efeito em comum. Assumindo-se independência condicional dos elementos de $\boldsymbol{y}$, tem-se

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov\big[E(Y_i \mid \boldsymbol{u}), E(Y_j \mid \boldsymbol{u})\big] + E\big[Cov(Y_i, Y_j \mid \boldsymbol{u})\big] \\ &= Cov(\mu_i, \mu_j) + E(0) \\ &= Cov\big[g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}), g^{-1}(\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{z}_j'\boldsymbol{u})\big]. \end{aligned}$$

Os estimadores resultantes dependem da função geradora de momentos da variável aleatória. Aplicações utilizando distribuições específicas são apresentadas em McCulloch e Searle (2000).

2.5.1 Estimação por máxima verossimilhança

Várias abordagens para estimar os parâmetros do modelo (25) são desenvolvidas na literatura. Schall (1991) sugere estimação de máxima verossimilhança similar ao que é usado para modelos mistos geral, Breslow & Clayton (1993) estudam um tipo de estimador de máxima verossimilhança marginal, McGilchrist (1994) recomenda o melhor preditor linear não-viesado enquanto que Lee & Nelder (1996) introduzem um método geral chamado estimação de máxima verossimilhança hierárquica.

De (25), (26) e (27) pode-se escrever a fórmula para a verossimilhança

$$L = \int \prod_i f_{Y_i|\boldsymbol{u}}(y_i \mid \boldsymbol{u}) f_U(u)\, du = \prod_i \int f(\boldsymbol{y}, u)\, du = \prod_{i=1}^{n} f_Y(\boldsymbol{y}) \tag{28}$$

sendo que a integração é sobre a distribuição de $\boldsymbol{u}$, de dimensões $q \times 1$. Nos casos mais simples, a integração numérica para o cálculo da verossimilhança é direta e, conseqüentemente, a maximização numérica da função de verossimilhança não é difícil, já que o logaritmo da função de verossimilhança é a soma das contribuições independentes de cada agrupamento, que envolve apenas uma integral de dimensão única, que pode ser calculada usando-se técnicas de quadratura.

Equações de verossimilhança para parâmetros de efeitos fixos

Embora uma solução para as equações de verossimilhança seja numericamente difícil, pode-se escrevê-las em uma forma mais simples. De (28), tem-se

$$\ell = \log \quad f_Y(\boldsymbol{y}). \tag{29}$$

Assim,

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[\log \int f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_U(u) du \right] \\
&= \frac{1}{\int f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_U(u) du} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \int f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_U(u) du \right] \\
&= \frac{1}{f_Y(\boldsymbol{y})} \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) \right] f_U(u) du
\end{aligned} \tag{30}$$

pois $f_U(\boldsymbol{u})$ não envolve $\boldsymbol{\beta}$. Mas,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) &= \frac{1}{f_{\mathbf{Y}|\boldsymbol{u}}(\boldsymbol{y} \mid \boldsymbol{u})} \left[\frac{\partial}{\partial \boldsymbol{\beta}} f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) \right] f_{\mathbf{Y}|\mathbf{U}}(\boldsymbol{y} \mid \boldsymbol{u}) \\
&= \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log \quad f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) \right] f_{\mathbf{Y}|\mathbf{U}}(\boldsymbol{y} \mid \boldsymbol{u}),
\end{aligned}$$

então,

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \frac{1}{f_{\mathbf{Y}}(\mathbf{y})} \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log \quad f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_{\mathbf{Y}|\mathbf{U}}(\boldsymbol{y} \mid \boldsymbol{u}) \right] f_U(u) du \\
&= \int \left[\frac{\partial \ell}{\partial \boldsymbol{\beta}} \log \quad f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) \right] \frac{f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) f_U(\boldsymbol{u}) du}{f_{\mathbf{Y}}(\boldsymbol{y})} \\
&= \int \left[\frac{\partial \ell}{\partial \boldsymbol{\beta}} \log \quad f_{Y|u}(\boldsymbol{y} \mid \boldsymbol{u}) \right] f_{U|Y}(\boldsymbol{u} \mid \boldsymbol{y}) du.
\end{aligned} \tag{31}$$

Mas

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X'W\Delta(y-\mu)}, \text{ com } \boldsymbol{W} = diag\left\{W_i\right\},$$

sendo

$$\boldsymbol{W} = diag\left\{W_i\right\} = diag\left[a_i(\phi)V(\mu_i)\frac{\partial \eta_i}{\partial \mu_i}\right]^2 \text{ e } \Delta = diag\left\{\frac{\partial \eta_i}{\partial \mu_i}\right\}.$$

Portanto,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \int \boldsymbol{X'W^*(y-\mu)} f_{U|Y}(u \mid y)du$$

em que $W^* = diag\left[a(\phi)V(\mu_i)\frac{\partial \eta}{\partial \mu}\right]^{-1}$. Logo,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X'y}E[\boldsymbol{W^*} \mid \boldsymbol{y}] - \boldsymbol{X'}E[\boldsymbol{W^*\mu} \mid \boldsymbol{y}].$$

Então,

$$\boldsymbol{X'y}E[\boldsymbol{W^*} \mid \boldsymbol{y}] = \boldsymbol{X'}E[\boldsymbol{W^*\mu} \mid \boldsymbol{y}].$$

Equações de verossimilhança para parâmetros de efeitos aleatórios

Um resultado similar à equação (31) pode ser encontrado para equações de máxima verossimilhança para os parâmetros na distribuição de $f_{\mathbf{U}}(\boldsymbol{u})$. Seja φ denotando os parâmetros dos efeitos aleatórios. De (29) tem-se:

$$\begin{aligned}
\frac{\partial \ell}{\partial \varphi} &= \frac{1}{f(y)}\frac{\partial f(y)}{\partial \varphi} \\
&= \frac{1}{f(y)}\frac{\partial}{\partial \varphi}\left[\int f(y \mid u)f(u)du\right] \\
&= \int \left[\frac{\partial}{\partial \varphi}f(y \mid u)\right]\frac{f(u)}{f(y)}du + \int \frac{f(y \mid u)}{f(y)}\left[\frac{\partial}{\partial \varphi}f(u)\right]du \\
&= \int \frac{f(y,u)}{f(y)}\frac{1}{f(u)}\left[\frac{\partial}{\partial \varphi}f(u)\right]du \\
&= \int f_{U|y}(\boldsymbol{u} \mid \boldsymbol{y})\left[\frac{\partial}{\partial \varphi}\ \log\ \ f_U(\boldsymbol{u})\right]du \\
&= E\left[\frac{\partial}{\partial \varphi}\ \log\ \ f_U(\boldsymbol{u}) \mid \boldsymbol{y}\right],
\end{aligned}$$

que não pode ser simplificada sem que se especifique uma forma para a distribuição dos efeitos aleatórios.

Um algoritmo iterativo para calcular as estimativas de máxima verossimilhança é o EM - *Expectation-Maximization* apresentado por Dempster, Laird e Rubin (1977). O nome é esse pois ele alterna entre calcular valores esperados condicionais e maximizar funções de verossimilhança simplificada.

O algoritmo EM toma a seguinte forma geral:

1) no passo inicial, isto é, $m = 0$, escolhem-se os valores iniciais para $\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}$ e $\boldsymbol{D}^{(0)}$;

2) calculam-se:

   a) $\boldsymbol{\beta}^{(m+1)}$ e $\boldsymbol{\theta}^{(m+1)}$ para maximizar $E\left[\ \log\ f_{\mathbf{Y}\ |\ \mathbf{U}}\ (\boldsymbol{y}\ \ |\ \ \boldsymbol{u}, \boldsymbol{\beta},\ \boldsymbol{\theta})\ |\ \boldsymbol{y}\right]$;

   b) $\boldsymbol{D}^{(m+1)}$ para maximizar $E\Big[\log f_{\mathbf{U}}(\mathbf{u}\ |\ \mathbf{D})\ |\ \mathbf{y}\Big]$;

   c) Faz-se $m = m + 1$ e

3) se a convergência ocorre, os valores obtidos são as estimativas de máxima verossimilhança, caso contrário, retorna-se ao passo 2.

No processo de estimação dos parâmetros, em que o algoritmo EM é utilizado, podem ocorrer algumas dificuldades no processo de integração devido, por exemplo, à quantidade de parâmetros envolvidos. Quando a dimensão da integração envolvida no passo E do algoritmo EM é 1 ou 2, técnicas de integração numérica podem ser utilizadas tal como, por exemplo, o método da quadratura gaussiana. Se dimensões de ordem maior estiverem envolvidas, outros métodos, tais como o de Markov chain Monte Carlo (MCMC), podem ser utilizados.

2.6 Modelo Poisson inflacionado de zero com efeito aleatório (ZIP)

O excesso de zeros em dados de contagem é uma ocorrência, de certa forma, comum em experimentos nas mais diversas áreas, ocasionando o fenômeno

conhecido por superdispersão que implica que a variância observada é muito maior do que a esperada pelo modelo. O excesso de zeros pode ser causado por uma combinação dos chamados zeros estruturais e zeros amostrais, sendo que no primeiro caso, os zeros ocorrerão independentemente do tipo de tratamento e o segundo caso está relacionado com a probabilidade de ocorrência de resposta zero no experimento. Ridout et al. (1998) citam como exemplo a resposta zero de uma planta resistente a uma determinada doença como zero estrutural e, como zero amostral, a resposta zero de uma planta suscetível não doente, mas que poderia ter a doença se tivesse contato com o esporo.

O modelo Poisson inflacionado de zeros sem covariáveis é discutido por vários autores, entre eles, Cohen (1960) e Johnson & Kotz (1969), mas é Lambert (1992) que apresenta este modelo associado ao uso de covariáveis em um exemplo sobre o número de defeitos em um processo industrial. Ridout et al. (1998) apresentam uma revisão sobre modelos que se ajustam a dados de contagens inflacionados de zeros e usam o modelo Poisson inflacionado de zeros (ZIP), e também o modelo binomial negativo inflacionado de zeros (ZINB), para a análise de um conjunto de dados do número de raízes produzidas por 270 brotos de maçã da cultivar *Trajan*. Estudos recentes estendem a metodologia apresentada por Lambert (1992) incorporando um efeito aleatório ao modelo, para levar em consideração a correlação entre respostas e outras fontes de variação não observadas. Esta metodologia é apresentada por Yau & Lee (2001) para avaliar um programa de prevenção de lesões em funcionários da área de limpeza de um hospital. A inclusão de efeitos aleatórios, tanto na parte logística quanto na parte Poisson do modelo ZIP, melhora o ajuste do mesmo. Na área de horticultura, Hall (2000) apresenta, além do modelo ZIP com efeito aleatório, o modelo binomial inflacionado de zeros (ZIB) com efeito aleatório. Ambos os modelos são aplicados a um experimento sobre o uso de sistema de subirrigação em casa de vegetação em que o inseticida *imidacloprid* é aplicado para controlar a mosca branca, sendo duas as variáveis respostas observadas com o objetivo de medir a eficácia do pesticida na mortalidade e na inibição de reprodução da mosca branca.

A primeira variável observada é a proporção de moscas brancas sobreviventes e a segunda variável é o número de insetos imaturos, sendo adotado um modelo ZIB com efeito aleatório e um modelo ZIP com efeito aleatório, respectivamente. Em ambos os casos, o ajuste do modelo foi superior aos modelos ZIP e ZIB padrões.

2.6.1 Caso univariado

Seja $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)'$ um conjunto de variáveis aleatórias independentes em que:

$$Y_i \sim \begin{cases} 0, \text{ com probabilidade } \omega_i \\ Poisson(\lambda_i), \text{ com probabilidade } (1-\omega_i), \end{cases}$$

gerando a distribuição do modelo Poisson inflacionado de zeros, que é dada por:

$$P(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\lambda_i}, & y_i = 0 \\ (1-\omega_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, & y_i = 1, 2, \ldots \end{cases} \tag{32}$$

cujas esperança e variância são dadas por:

$$E(Y_i) = (1-\omega_i)\lambda_i = \mu_i \qquad \text{e} \qquad V(Y_i) = \mu_i + \left(\frac{\omega_i}{1-\omega_i}\right)\mu_i^2.$$

A inclusão de covariáveis no modelo ZIP e a aplicação da teoria de modelos lineares generalizados (McCullagh & Nelder, 1989) é feita com a definição das funções de ligação logística e logarítmica como em Lambert (1992), isto é,

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X\beta} \qquad \text{e} \qquad \log\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = \boldsymbol{G\gamma}, \tag{33}$$

em que $\boldsymbol{X}$ e $\boldsymbol{G}$ são as matrizes associadas às covariáveis, que podem ser, ou não, iguais, e $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são os vetores de parâmetros. Se elas forem iguais, modelos mais parcimoniosos podem ser desenvolvidos supondo-se que os dois preditores lineares sejam relacionados de alguma forma. Lambert (1992) refere-se a esse modelo como ZIP(τ) e o define como

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X\beta} \qquad \text{e} \qquad \log\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = \tau\boldsymbol{X\beta},$$

sendo τ um escalar e implicando que $\omega = (1 + \lambda^{-\tau})^{-1}$.

A estimação dos parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ pode ser feita pelo método de máxima verossimilhança via o algoritmo EM (Lambert, 1992). Seja Z_i uma variável indicadora tal que

$$Z_i = \begin{cases} 1, \text{ se} & Y_i = 0 \quad \text{zero estrutural} \\ 0, \text{ se} & Y_i \sim Poisson(\lambda_i). \end{cases} \tag{34}$$

Logo, pode-se admitir que $Z_i \sim Bernoulli(\omega_i)$, isto é,

$$P(Z_i = z_i) = \omega_i^{z_i}(1 - \omega_i)^{1-z_i}, \quad z_i = 0, 1.$$

Como não se têm informações suficientes sobre o processo de geração dos zeros, admite-se que a variável aleatória Z_i faz o papel de dados perdidos no algoritmo EM e o logaritmo da função de verossimilhança baseado nos dados completos $\ell_c = \ell_c(\boldsymbol{\lambda}, \boldsymbol{\omega}; \boldsymbol{y}, \boldsymbol{z})$, é dado por:

$$\begin{aligned} \ell_c &= \sum_i \log(y_i, z_i; \lambda_i, \omega_i) = \sum_i \log\Big[P(z_i \mid \omega_i)\ \ P(y_i \mid z_i, \lambda_i, \omega_i)\Big] \\ &= \sum_i \log\Big[\omega_i^{z_i}(1 - \omega_i)^{1-z_i}\Big] + \sum_i \log\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right)^{1-z_i} \\ &= \sum_i \left\{ z_i \log\left(\frac{\omega_i}{1-\omega_i}\right) + \log(1-\omega_i)\right\} + \sum_i \Big\{(1 - z_i)\Big[y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\Big]\Big\} \\ &= \ell(\boldsymbol{\omega}; \boldsymbol{z}) + \ell(\boldsymbol{\lambda}; \boldsymbol{y}, \boldsymbol{z}). \end{aligned} \tag{35}$$

Escrevendo-se ℓ_c em termos dos parâmetros de regressão $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, tem-se:

$$\begin{aligned} \ell_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y}) &= \sum_{i=1}^{n} \Big\{ z_i \boldsymbol{G_i\gamma} - \log\left(1 + e^{\boldsymbol{G_i\gamma}}\right) + (1 - z_i)\left[y_i \boldsymbol{X_i\beta} - e^{\boldsymbol{X_i\beta}} - \log(y_i!)\right]\Big\} \\ &= \ell(\boldsymbol{\gamma}; \boldsymbol{z}) + \ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{z}), \end{aligned} \tag{36}$$

o que mostra que a estimação de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ pode ser feita maximizando-se os dois termos de $\ell_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y})$, separadamente.

Do logaritmo da função de verossimilhança para os dados completos em (35), nota-se que Z_i é uma estatística suficiente para ω_i e Z_iY_i é suficiente para

λ_i, sendo que o passo E, do algoritmo EM, reduz-se a estimar o valor esperado condicional de Z_i dado Y_i, λ_i e ω_i.

Passo E:

De (34) tem-se que:

$$\begin{aligned}(Y_i \mid Z_i = 1) &\equiv 0\\ (Y_i \mid Z_i = 0) &\sim Poisson(\lambda_i).\end{aligned}$$

Logo,

$$E\Big[Z_i \mid Y_i, \lambda_i, \omega_i\Big] = \sum_{z_i=0}^{1} z_i P\Big[Z_i = z_i \mid Y_i, \lambda_i, \omega_i\Big] = P\Big[Z_i = 1 \mid Y_i, \lambda_i, \omega_i\Big].$$

Usando-se o Teorema de Bayes, tem-se:

$$E\Big[Z_i \mid Y_i, \lambda_i, \omega_i\Big] = \frac{P(Z_i = 1)P\Big[Y_i = y_i \mid Z_i = 1; \lambda_i, \omega_i\Big]}{\sum_{j=0}^{1} P(Z_i = j)P\Big[Y_i = y_i \mid Z_i = j; \lambda_i, \omega_i\Big]}. \tag{37}$$

Considerando-se o denominador apresentado em (37), e fazendo-se $y_i = 0$, tem-se:

$$\begin{aligned}P(Z_i = 0)P\Big[Y_i = 0 \mid Z_i = 0; \lambda_i, \omega_i\Big] + P(Z_i = 1)P\Big[Y_i = 0 \mid Z_i = 1; \lambda_i, \omega_i\Big] &=\\ = (1 - \omega_i)\frac{e^{-\lambda_i}\lambda_i^0}{0!} + \omega_i = (1 - \omega_i)e^{-\lambda_i} + \omega_i,&\end{aligned}$$

portanto,

$$E\Big[Z_i \mid Y_i, \lambda_i, \omega_i\Big] = \begin{cases} \dfrac{\omega_i}{\omega_i + (1 - \omega_i)e^{-\lambda_i}}, & y_i = 0 \\ 0, & y_i = 1, 2, \ldots \end{cases}$$

Para o passo M o procedimento de estimação pode ser dividido em duas partes; o passo para $\boldsymbol{\beta}$ e o passo para $\boldsymbol{\gamma}$:

Passo M para $\boldsymbol{\beta}$: O parâmetro $\boldsymbol{\beta}$ pode ser estimado pela maximização de $\ell(\boldsymbol{\lambda; y, z})$

via modelo log-linear Poisson ponderado com peso $(1 - \hat{\boldsymbol{z}})$, sendo $\hat{\boldsymbol{z}}$ obtido do passo E, ou seja, $\hat{z}_i = \dfrac{\omega_i}{\omega_i + (1-\omega_i)e^{-\lambda_i}}$;

Passo M para $\boldsymbol{\gamma}$: Vê-se de (35), que o logaritmo da função de verossimilhança de Z é uma Bin(1, ω). Portanto, $\boldsymbol{\gamma}$ pode ser estimado pela maximização de $\ell(\boldsymbol{\omega}; \boldsymbol{z})$ usando-se uma regressão logística não ponderada com variável resposta $\hat{\boldsymbol{z}}$ obtida do passo E.

Os passos E e M devem ser alternados até ocorrer a convergência, usando-se algum critério de parada.

2.6.2 Caso multivariado

No modelo ZIP univariado, as observações são supostas independentes e a forma de ajuste é como a descrita. Quando a variável resposta é observada ao longo do tempo, espera-se que haja uma correlação entre as observações e, sendo assim, existe a necessidade de se levar em conta este fato, incluindo-se um efeito aleatório ao modelo ZIP multivariado.

Seja $\boldsymbol{Y}$ o vetor de respostas contendo dados de n grupos independentes, de forma que $\boldsymbol{Y} = (\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \cdots, \boldsymbol{Y}'_k)'$ em que $\boldsymbol{Y_i} = (Y_{i1}, \cdots, Y_{it_i})'$. Assume-se que, condicional sobre um efeito aleatório b_i, tem-se:

$$Y_{ij} \sim \begin{cases} 0, & \text{com probabilidade } \omega_{ij} \\ Poisson(\lambda_{ij}), & \text{com probabilidade } (1-\omega_{ij}), \end{cases}$$

para $i = 1, \ldots, k$ e $j = 1, \ldots, t_i$, sendo $\boldsymbol{\lambda}_i = (\lambda_{i1}, \cdots, \lambda_{it_i})'$ e $\boldsymbol{\omega}_i = (\omega_{i1}, \cdots, \omega_{it_i})'$. A inclusão do efeito aleatório pode ser feita em um dos dois preditores lineares apresentados em (33) ou, em ambos. Assim, admitindo-se efeito aleatório apenas no modelo de média, têm-se os modelos de regressão log-linear e logístico dados,

respectivamente, por:

$$\log(\boldsymbol{\lambda_i}) = \boldsymbol{X_i\beta} + \sigma b_i \qquad \text{e,}$$

$$\log\left(\frac{\boldsymbol{\omega_i}}{1-\boldsymbol{\omega_i}}\right) = \boldsymbol{G_i\gamma}, \quad i = 1, \ldots, k. \tag{38}$$

sendo $\boldsymbol{X}$ e $\boldsymbol{G}$ as matrizes do delineamento e $b_i \sim N(0, \sigma^2)$, $i = 1, \ldots, k$ e independentes.

Seja $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \sigma)'$ o vetor de parâmetros. O logaritmo da função de verossimilhança para o modelo ZIP com efeito aleatório é dado por:

$$\ell(\boldsymbol{\psi}; \boldsymbol{y}) = \sum_{i=1}^{k} \log \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{t_i} P(Y_{ij} = y_{ij} \mid b_i)\right] \phi(b_i) db_i, \tag{39}$$

sendo,

$$\begin{aligned} P(Y_{ij} = y_{ij} \mid b_i) &= \left[\omega_{ij} + (1-\omega_{ij})e^{-\lambda_{ij}}\right]^{z_{ij}} \left[\frac{(1-\omega_{ij})e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{y_{ij}!}\right]^{1-z_{ij}} \\ &= \left(1 + e^{\boldsymbol{G}_{ij}\gamma}\right)^{-1} \left\{ z_{ij}\left[\exp\left(\boldsymbol{G}_{ij}\gamma\right) + \exp\left(-e^{\boldsymbol{X}_{ij}\beta} + \sigma b_i\right)\right] \right. \\ &\quad \left. + \ (1-z_{ij}) \frac{\exp\left[y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i) - e^{\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i}\right]}{y_{ij}!} \right\} \end{aligned}$$

em que ϕ denota a função de densidade de probabilidade da distribuição normal padrão e $z_{ij} = 1$ se $y_{ij} = 0$ e $z_{ij} = 0$, em caso contrário.

A maximização de (39) em relação a $\boldsymbol{\psi}$ é complicada pela integração em relação aos efeitos aleatórios. Uma forma de maximização utilizada por vários autores é o de empregar o algoritmo EM com quadratura gaussiana. Detalhes são apresentados por Hall (2000).

Assim, seja $Z_{ij} = 1$ quando Y_{ij} vem das respostas zero e $Z_{ij} = 0$ quando Y_{ij} vem do estado Poisson(λ_{ij}). O logaritmo da função de verossimilhança

dos dados completos é:

$$\begin{aligned}\ell_c(\boldsymbol{\psi};\boldsymbol{y},\boldsymbol{z},\boldsymbol{b}) &= \log f(\boldsymbol{b};\boldsymbol{\psi}) + \log f(\boldsymbol{y},\boldsymbol{z} \mid \boldsymbol{b};\boldsymbol{\psi}) \\ &= \sum_{i=1}^{k} \log \phi(b_i) + \sum_{i=1}^{k}\sum_{j=1}^{t_i} \Bigg\{ \Big[z_{ij}\boldsymbol{G}_{ij}\gamma - \log\left(1+e^{\boldsymbol{G}_{ij}\gamma}\right)\Big] + \\ &\quad (1-z_{ij})\Big[y_{ij}(\boldsymbol{B}_{ij}\boldsymbol{\beta}+\sigma b_i) - \exp(\boldsymbol{B}_{ij}\boldsymbol{\beta}+\sigma b_i) - \log(y_{ij}!)\Big]\Bigg\}.\end{aligned}$$

A (r+1)-ésima iteração do algoritmo EM consiste dos seguintes passos:

Passo E: O passo E consiste do cálculo de $Q(\psi \mid \psi^{(s)}) = E\Big[\log f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{b};\boldsymbol{\psi}) \mid \boldsymbol{y},\boldsymbol{\psi}^{(r)}\Big]$, sendo o cálculo da esperança realizado em relação a distribuição conjunta de $\boldsymbol{z},\boldsymbol{b}$ dado $\boldsymbol{y}$ e $\boldsymbol{\psi}^{(r)}$. Esta esperança pode ser calculada em dois passos,

$$Q(\psi \mid \psi^{(r)}) = E\Big[E\big(\log f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{b} \mid \boldsymbol{\psi}) \mid \boldsymbol{y},\boldsymbol{b},\boldsymbol{\psi}^{(r)}\big) \mid \boldsymbol{y},\boldsymbol{\psi}^{(r)}\Big], \tag{40}$$

sendo a esperança interna dada em (40) considerada somente em relação a $\boldsymbol{z}$, uma vez que $\log f(\boldsymbol{y},\boldsymbol{z},\boldsymbol{b} \mid \boldsymbol{\psi})$ é linear em relação a $\boldsymbol{z}$. Esta esperança resulta em $\log f\big(\boldsymbol{y},\boldsymbol{Z}^{(r)},\boldsymbol{b} \mid \boldsymbol{\psi}\big)$, em que $\boldsymbol{Z}^{(r)}$ contém elementos

$$Z_{ij}^{(r)} = E\left(Z_{ij} \mid \boldsymbol{y},\boldsymbol{b},\boldsymbol{\psi}^{(r)}\right) = \begin{cases} 0, \ \text{ se } y_{ij} > 0\ ; \\ \Big[1+\exp\Big(-\boldsymbol{G}_{ij}\boldsymbol{\gamma}^{(r)} - e^{\boldsymbol{B}_{ij}\boldsymbol{\beta}^{(r)}+\sigma^{(r)}b_i}\Big)\Big]^{-1}, \ \text{se } y_{ij}=0\ . \end{cases}$$

Note que $Z_{ij}^{(r)}$ depende de b_i, e que esta dependência será denotada por $Z_{ij}^{(r)}(b_i)$. Para completar o passo E, necessita-se tomar a esperança em relação à distribuição de $\boldsymbol{b} \mid \boldsymbol{y},\boldsymbol{\psi}^{(r)}$.

Logo, tem-se que:

$$
\begin{aligned}
Q\Big(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\Big) \;=\; & \sum_{i=1}^{k}\sum_{j=1}^{t_i}\int_{-\infty}^{\infty}\Bigg\{\Big[Z_{ij}^{(r)}(b_i)\boldsymbol{G}_{ij}\boldsymbol{\gamma} - \log\Big(1+e^{\boldsymbol{G}_{ij}\boldsymbol{\gamma}}\Big)\Big] + \\
& \Big[1 - Z_{ij}^{(r)}(b_i)\Big] \times \\
& \Big[y_{ij}\Big(\boldsymbol{B}_i\boldsymbol{\beta} + \sigma b_i\Big) - \exp\Big(\boldsymbol{B}_{ij}\boldsymbol{\beta} + \sigma b_i\Big)\Big]\Bigg\} \times \\
& \frac{f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_i\Big)\phi(b_i)\; db_i}{\int_{-\infty}^{\infty} f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_i\Big)\phi(b_i)\; db_i},
\end{aligned}
\tag{41}
$$

em que se usa o fato de que:

$$
f\Big(b_i; \boldsymbol{\psi}^{(r)} \mid \boldsymbol{y}_i\Big) = \frac{f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_i\Big)\phi(b_i)}{\int_{-\infty}^{\infty} f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_i\Big)\phi(b_i)\; db_i}.
$$

Usando-se m pontos de quadratura Gaussiana para aproximar as integrais em (41), obtém-se:

$$
\begin{aligned}
Q\Big(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\Big) \;\approx\; & \sum_{i=1}^{k}\sum_{j=1}^{t_i}\Bigg\{\frac{\sum_{\ell=1}^{m}\Big[Z_{ij}^{(r)}(b_\ell)\boldsymbol{G}_{ij}\boldsymbol{\gamma} - \log\Big(1+e^{\boldsymbol{G}_{ij}\boldsymbol{\gamma}}\Big)\Big]}{\sum_{\ell=1}^{m} f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_\ell\Big)w_\ell} \times \\
& f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_\ell\Big)w_\ell + \frac{\sum_{\ell=1}^{m}\Big[1 - Z_{ij}^{(r)}(b_\ell)\Big]}{\sum_{\ell=1}^{m} f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_\ell\Big)w_\ell} \times \\
& \Big[y_{ij}\Big(\boldsymbol{B}_i\boldsymbol{\beta} + \sigma b_\ell\Big) - \exp\Big(\boldsymbol{B}_{ij}\boldsymbol{\beta} + \sigma b_\ell\Big)\Big] \times \\
& f\Big(\boldsymbol{y}_i; \boldsymbol{\psi}^{(r)} \mid b_\ell\Big)w_\ell\Bigg\},
\end{aligned}
$$

em que b_ℓ são os pontos de quadratura e w_ℓ os pesos associados.

Passo M para $\boldsymbol{\gamma}$: Note que $Q\Big(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\Big)$ se decompõe na soma de uma termo envolvendo γ e um segundo termo envolvendo somente $\boldsymbol{\beta}$. Logo, maximiza-se

$Q\big(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\big)$ em relação a $\boldsymbol{\gamma}$, maximizando-se o primeiro termo. Tal maximização nada mais é do que uma regressão logística ponderada de $\boldsymbol{Z}_{ij}^{(r)}(b_\ell)$ sobre $\boldsymbol{G}$ com pesos $\dfrac{f\big(\boldsymbol{y}_i;\boldsymbol{\psi}^{(r)} \mid b_\ell\big)w_\ell}{\sum_{\ell=1}^{m} f\big(\boldsymbol{y}_i;\boldsymbol{\psi}^{(r)} \mid b_\ell\big)w_\ell}$.

Passo M para $\boldsymbol{\beta}, \boldsymbol{\sigma}$: A maximização do segundo termo de $Q\Big(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(r)}\Big)$ em relação à $\boldsymbol{\beta}$ e σ pode ser feita simultaneamente. Também neste passo, a maximização nada mais é do que uma regressão log-linear ponderada de $\boldsymbol{y} \otimes \mathbf{1}_{m\times 1}$ sobre $\boldsymbol{B^*} = \Big[(\boldsymbol{B} \otimes \mathbf{1}_m), \Big(\mathbf{1}_N \otimes (b_1, \ldots, b_m)'\Big)\Big]$ com pesos $\Big[1 - Z_{ij}^{(r)}(b_\ell)\Big] \dfrac{f\big(\boldsymbol{y}_i;\boldsymbol{\psi}^{(r)} \mid b_\ell\big)w_\ell}{\sum_{\ell=1}^{m} f\big(\boldsymbol{y}_i,\boldsymbol{\psi}^{(r)} \mid b_\ell\big)w_\ell}$; $i = 1, \ldots, k$, $j = 1, \ldots, t_i$, $\ell = 1, \ldots, m$.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Experimento 1 - Comparação de métodos de enxertia e tipos de porta-enxertos para camu-camu

O camu-camu (*Myrciaria dubia* (H.B.K.) McVaugh), espécie frutífera nativa das várzeas e cursos de rio da região amazônica, tem despertado grande interesse comercial devido ao alto teor de vitamina C de seu fruto, sendo a maior fonte natural de vitamina C conhecida. Devido à carência de informações sobre seu manejo em locais diferentes de seu ambiente natural, Suguino (2002) conduziu um experimento na área experimental do Departamento de Produção Vegetal da Escola Superior de Agricultura "Luiz de Queiroz" - da Universidade de São Paulo (ESALQ/USP), em Piracicaba-SP, com o objetivo de avaliar quais os melhores métodos de enxertia, por garfagem, e tipos de porta-enxertos que podem ser utilizados para propagar essa planta em terrenos não inundáveis.

O experimento foi instalado no delineamento inteiramente casualizado com 5 repetições e 12 tratamentos no esquema fatorial 3×4 $\big($3 tipos de porta-enxertos: camu-camu, goiabeira (*Psidium guajava* L.) e pitangueira (*Eugenia uniflora* L.) e 4 métodos de enxertia: fenda cheia, fenda lateral, inglês simples e fenda de colo $\big)$. As parcelas foram constituídas de 12 plantas. A notação utilizada para representar os diferentes tratamentos está na Tabela 1.

A variável resposta de interesse foi porcentagem de pegamento, medida mensalmente, durante oito meses (dezembro de 2000 a julho de 2001). Na Tabela 2 são mostrados os números de pegamentos observados de 12 garfos vivos de camu-

Tabela 1. Combinação dos níveis de cada fator, tipos de porta-enxertos e métodos de enxertia.

Métodos de Enxertia	Porta-enxertos		
	Camu-camu	Goiabeira	Pitangueira
Fenda cheia	T1	T2	T3
Fenda lateral	T4	T5	T6
Inglês simples	T7	T8	T9
Colo	T10	T11	T12

camu, por tratamento e por repetição, durante os oito meses de avaliação, enquanto que a Figura 1 representa esses dados graficamente.

Para a análise desses dados devem ser levados em consideração:

i) a natureza discreta da variável resposta (proporções de pegamentos);

ii) a variabilidade entre repetições dentro de cada tratamento, conforme pode ser visto pela Tabela 2 e Figura 1 e

iii) a dependência entre observações ao longo do tempo.

Além disso, pode-se verificar pelas Tabela 2 e Figura 1, apesar da variabilidade entre repetições, que, em geral:

i) o camu-camu como porta-enxerto apresenta uma maior estabilidade ao longo dos meses e uma porcentagem maior de pegamento, sendo T1 o tratamento com maior variabilidade;

ii) a goiabeira como porta-enxerto apresenta uma ligeira estabilidade até o mês de março, sendo que, a partir desse mês, o número de pegamentos tende a zero e

iii) a pitangueira como porta-enxerto foi a que menos se adaptou, pois o número de pegamentos tende a zero de forma mais acentuada, desde o início.

Tabela 2. Números observados de pegamentos a partir de 12 garfos vivos de camu-camu enxertados em diferentes porta-enxertos, durante o período de dezembro/2000 a julho/2001.

Tratamentos	Repetições	Dezembro	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Julho
	r1	9	10	10	10	10	10	10	10
	r2	7	10	10	10	10	10	10	8
T1	r3	8	9	9	9	8	8	8	8
	r4	1	1	1	1	1	1	1	1
	r5	0	0	0	0	0	0	0	0
Média		5,0	6,0	6,0	6,0	5,8	5,8	5,8	5,4
Variância		17,5	25,5	25,5	25,5	24,2	24,2	24,2	20,8
	r1	9	8	8	7	3	1	1	0
	r2	4	7	9	9	9	7	6	2
T2	r3	3	4	4	4	3	2	2	1
	r4	4	7	7	5	4	2	1	0
	r5	1	6	7	5	4	4	2	1
Média		4,2	6,4	7,0	6,0	4,6	3,2	2,4	0,8
Variância		8,7	2,3	3,5	4,0	6,3	5,7	4,3	0,7
	r1	5	6	5	1	0	0	0	0
	r2	8	5	4	1	0	0	0	0
T3	r3	1	3	1	0	0	0	0	0
	r4	7	2	0	0	0	0	0	0
	r5	8	2	0	0	0	0	0	0
Média		5,8	3,6	2,0	0,4	0,0	0,0	0,0	0,0
Variância		8,7	3,3	5,5	0,3	0,0	0,0	0,0	0,0
	r1	10	10	10	10	9	9	9	8
	r2	10	11	11	11	11	11	11	11
T4	r3	12	12	12	12	12	12	12	11
	r4	6	10	9	9	9	8	8	8
	r5	6	7	7	7	7	7	7	7
Média		8,8	10,0	9,8	9,8	9,6	9,4	9,4	9,0
Variância		7,2	3,5	3,7	3,7	3,8	4,3	4,3	3,5
	r1	7	6	6	5	5	3	3	2
	r2	11	10	11	11	8	7	4	2
T5	r3	9	6	6	6	5	2	1	1
	r4	9	2	1	1	1	1	1	0
	r5	8	6	5	4	2	1	1	1
Média		8,8	6,0	5,8	5,4	4,2	2,8	2,0	1,2
Variância		2,2	8,0	12,7	13,3	7,7	6,2	2,0	0,7
	r1	7	5	5	1	0	0	0	0
	r2	9	8	5	4	0	0	0	0
T6	r3	9	3	4	2	1	0	0	0
	r4	5	0	0	0	0	0	0	0
	r5	6	2	0	0	0	0	0	0
Média		7,2	3,6	2,8	1,4	0,2	0,0	0,0	0,0
Variância		3,2	9,3	6,7	2,8	0,2	0,0	0,0	0,0

Tabela 2. Números observados de pegamentos a partir de 12 garfos vivos de camu-camu enxertados em diferentes porta-enxertos, durante o período de dezembro/2000 a julho/2001.

Tratamentos	Repetições	Dezembro	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Julho
	r1	0	0	0	0	0	0	0	0
	r2	0	2	4	3	2	2	0	0
T7	r3	1	2	2	2	1	1	1	1
	r4	3	4	6	6	6	6	5	5
	r5	3	9	10	9	7	7	7	7
Média		1,4	3,4	4,4	4,0	3,2	3,2	2,6	2,6
Variância		2,3	11,8	14,8	12,5	9,7	9,7	10,3	10,3
	r1	1	3	3	3	2	2	0	0
	r2	1	1	2	1	1	1	1	0
T8	r3	1	4	2	0	0	0	0	0
	r4	2	2	3	1	0	0	0	0
	r5	0	2	2	2	1	1	1	0
Média		1,0	2,4	2,4	1,4	0,8	0,8	0,4	0,0
Variância		0,5	1,3	0,3	1,3	0,7	0,7	0,3	0,0
	r1	3	1	1	0	0	0	0	0
	r2	3	1	0	0	0	0	0	0
T9	r3	3	3	2	1	0	0	0	0
	r4	7	5	2	1	0	0	0	0
	r5	0	0	0	0	0	0	0	0
Média		3,2	2,0	1,0	0,4	0,0	0,0	0,0	0,0
Variância		6,2	4,0	1,0	0,3	0,0	0,0	0,0	0,0
	r1	3	8	9	9	8	8	8	8
	r2	5	10	10	10	10	10	9	9
T10	r3	5	9	8	8	8	8	8	8
	r4	2	5	5	5	5	5	5	5
	r5	2	6	3	3	3	3	3	3
Média		3,4	7,6	7,0	7,0	6,8	6,8	6,6	6,6
Variância		2,3	4,3	8,5	8,5	7,7	7,7	6,3	6,3
	r1	4	8	6	7	6	3	3	1
	r2	4	5	5	3	2	0	0	0
T11	r3	4	3	2	0	0	0	0	0
	r4	6	7	4	3	3	3	2	0
	r5	4	4	4	4	1	1	1	0
Média		4,4	5,4	4,2	3,4	2,4	1,4	1,2	0,2
Variância		0,8	4,3	2,2	6,3	5,3	2,3	1,7	0,2
	r1	8	8	3	1	0	0	0	0
	r2	7	6	2	1	0	0	0	0
T12	r3	6	3	3	1	0	0	0	0
	r4	8	8	3	2	1	0	0	0
	r5	5	1	0	0	0	0	0	0
Média		6,8	5,2	2,2	1,0	0,2	0,0	0,0	0,0
Variância		1,7	9,7	1,7	0,5	0,2	0,0	0,0	0,0

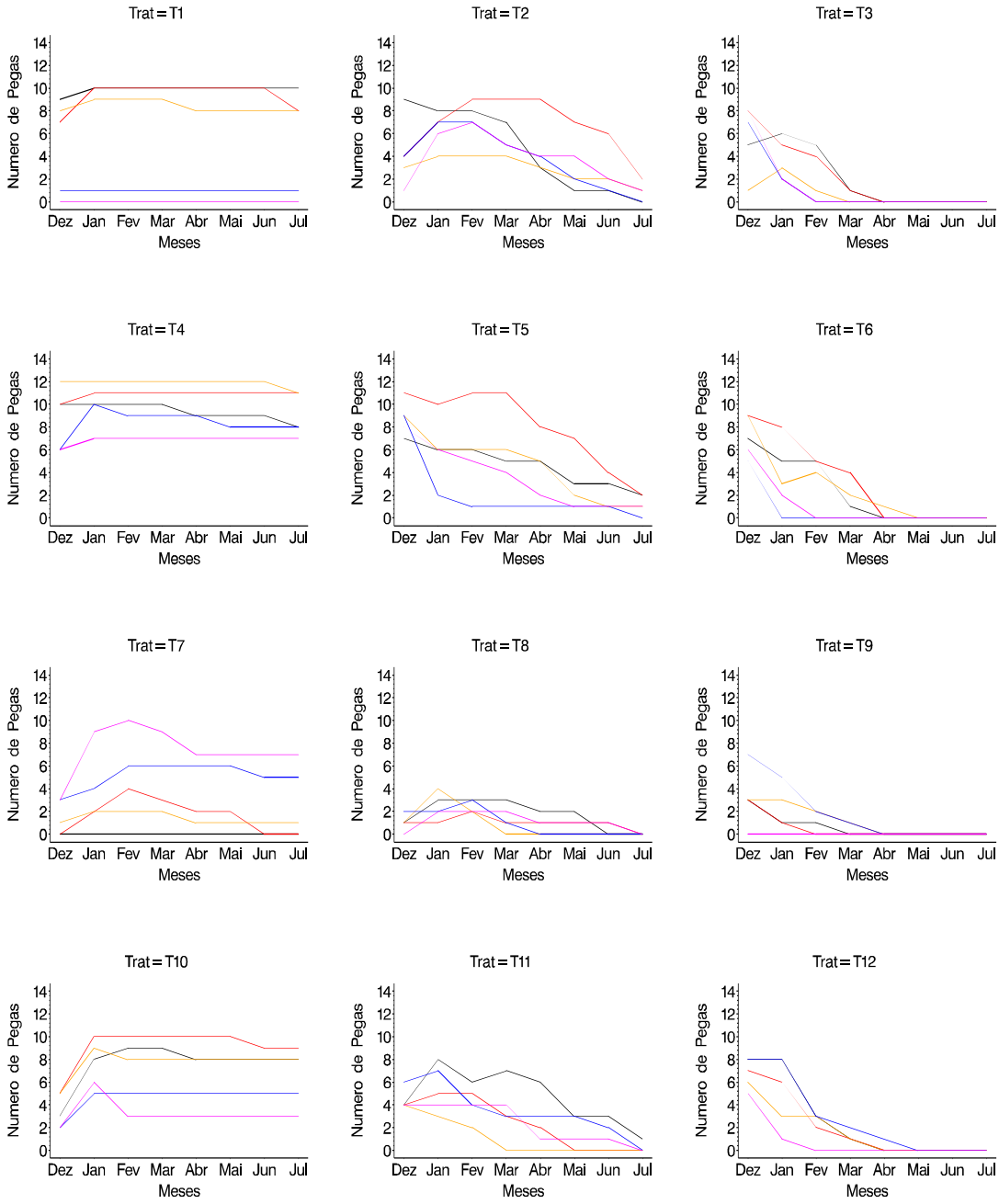

Figura 1 - Gráficos de dispersão dos números observados de pegamentos a partir de 12 garfos vivos de camu-camu enxertados em camu-camu, goiabeira e pitangueira (colunas 1, 2 e 3, respectivamente) por 4 métodos de enxertia, fenda cheia, fenda lateral, inglês simples e colo (linhas 1, 2, 3 e 4, respectivamente).

3.1.2 Experimento 2 - Comparação de milho geneticamente modificado MON810 e milho convencional (híbrido DKB909)

O milho, uma das maiores fontes de alimento, é cultivado em vários países do mundo. Movimenta um mercado de, aproximadamente, U$40 bilhões anuais, distribuídos entre indústrias de produção de alimentos para consumo humano, rações e matéria-prima para centenas de produtos industrializados. O Brasil produz mais de 30 milhões de toneladas de milho anualmente, em 13 milhões de hectares. Seu cultivo é possível desde o Equador até o limite das terras temperadas e desde o nível do mar até altitudes superiores a 3600m, graças aos programas de melhoramento genético.

Uma das pragas mais conhecidas do milho é a *Spodoptera frugiperda*, conhecida como lagarta-do-cartucho do milho, uma espécie polífaga que ataca dezenas de culturas economicamente importantes em vários países. Até hoje, o seu controle tem sido feito, principalmente, com a aplicação de produtos químicos. É sabido, no entanto, que o uso abusivo de qualquer método de controle pode ter efeitos negativos, como, por exemplo, a evolução da resistência.

Uma forma alternativa para o controle da praga é a utilização de milho geneticamente modificado. Assim, um experimento, realizado pela empresa Monsanto do Brasil Ltda., sob a responsabilidade do pesquisador Odnei D. Fernandes, foi conduzido em Rolândia, Estado do Paraná, confome processo deferido pela CNTBio - Comissão Técnica Nacional de Biossegurança. O experimento foi instalado em 11 de março de 2001 e teve como objetivo avaliar a eficiência do milho geneticamente modificado MON810 em relação ao milho convencional (híbrido DKB909) no controle de *Spodoptera frugiperda*.

O ensaio foi conduzido no delineamento inteiramente casualizado, com 3 tratamentos e 8 repetições com parcelas de $1250m^2$, sendo avaliadas durante 9 semanas. Os tratamentos foram:

Trat. 1 : milho geneticamente modificado MON810;

Trat. 2 : milho convencional com aplicação de inseticidas e

Trat. 3 : milho convencional sem a aplicação de inseticidas.

Para a aplicação de inseticidas foi estabelecido que, sempre que 20 a 30% das plantas estivessem com sintomas de ataque de *S. frugiperda*, a mesma seria realizada, seguindo-se as recomendações técnicas para uso de inseticidas. Os inseticidas foram aplicados com o auxílio de equipamento convencional tratorizado.

A variável resposta foi o número de lagartas grandes e a Tabela 3 mostra as freqüências observadas por tratamento e total, enquanto que as Figuras 2 e 3 representam os dados graficamente.

Tabela 3. Freqüências observadas do número de lagartas por tratamento.

Número de	Geral		Trat. 1		Trat. 2		Trat. 3	
lagartas	Freq.	%	Freq.	%	Freq.	%	Freq.	%
0	154	71,30	63	87,50	53	73,61	38	52,78
1	21	9,72	7	9,72	12	16,67	2	2,78
2	11	5,09	1	1,39	5	6,94	5	6,94
3	7	3,24	1	1,39	1	1,39	5	6,94
4	11	5,09			1	1,39	10	13,89
5	5	2,31					5	6,94
6	2	0,93					2	2,78
7	2	0,93					2	2,78
8	2	0,93					2	2,78
9	1	0,46					1	1,39
Médias	0,8518519		0,1666667		0,4027778		1,9861111	
Variâncias			0,2535211		0,6383020		6,2955790	

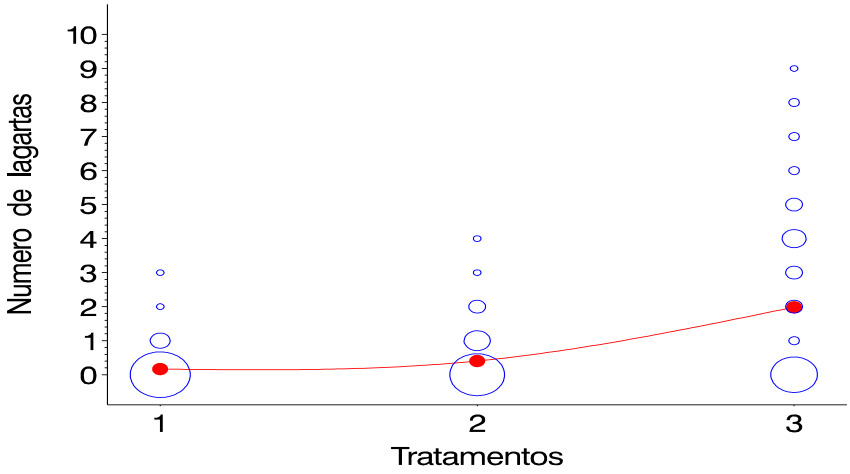

Figura 2 - Freqüências observadas do número de lagartas por tratamento (os círculos são proporcionais às freqüências e • representa a média).

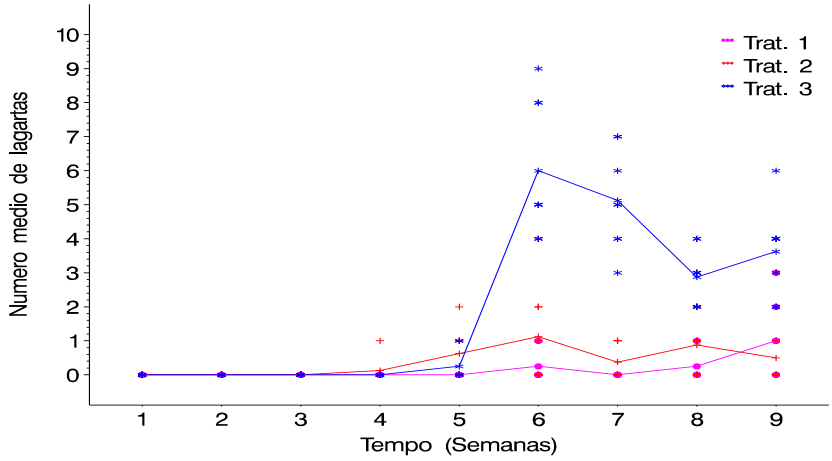

Figura 3 - Números observados de lagartas, ao longo do tempo, por tratamento.

Para a análise desses dados devem ser levados em consideração:

i) a natureza discreta da variável resposta (contagens);

ii) a heterogeneidade na variabilidade entre repetições dentro de cada tratamento conforme mostra a Figura 3;

iii) a dependência entre observações ao longo do tempo e

iv) o excesso de zeros, conforme pode ser visto pelas Tabela 3 e 4 e Figura 2.

Tabela 4. Totais de zeros observados, por tratamento, ao longo do tempo.

Tratamentos	Semanas									Totais
	1	2	3	4	5	6	7	8	9	de zeros
Trat. 1	8	8	8	8	8	6	8	6	3	63
Trat. 2	8	8	8	7	4	4	5	3	6	53
Trat. 3	8	8	8	8	6	0	0	0	0	38

Além disso, tem-se, pelo exame das Tabela 3 e 4 e Figuras 2 e 3 que:

i) o excesso de zeros é maior para os tratamentos 1 (87,50%) e 2 (73,61%) o que é desejável, pois mostra a ação deles na redução populacional do inseto;

ii) o tratamento 3 apresenta menor porcentagem de zeros e, portanto, maior quantidade de lagartas;

iii) até a terceira semana nenhuma lagarta é observada e que estas surgem a partir da sexta, quarta e quinta semanas, respectivamente, para os tratamentos 1, 2 e 3;

iv) os números médios de lagartas para os tratamentos 1 e 2 estão próximos das variâncias, enquanto que para o tratamento 3 o número médio de lagartas é bem menor do que a variância;

v) ao longo do tempo, os tratamentos 2 e 3 são bastante semelhantes, produzindo uma redução do número de lagartas grandes, em relação ao tratamento 1.

3.2 Métodos

3.2.1 Experimento 1

Para a análise estatística do Experimento 1, alguns modelos sob a abordagem de modelos lineares generalizados (McCullagh & Nelder, 1989) são ajustados, discutidos e comparados.

Vários ajustes são feitos, desde o modelo sem considerar a superdispersão, o efeito aleatório e a correlação entre as observações até o modelo em que todos esses efeitos são considerados. As análises foram feitas usando-se o sistema SAS - *Statistical Analysis System*, versão 8.2, através dos procedimentos GENMOD, NLMIXED e GPLOT, sendo que os programas são apresentados no Anexo A.

Para a seleção dos modelos são utilizados o Critério Bayesiano de Schwarz (BIC) e o Critério de Informação de Akaike (AIC), dados por:

$$\begin{aligned} AIC &= -2\ell + 2p \\ BIC &= -2\ell + p \log n, \end{aligned}$$

sendo ℓ o máximo do logaritmo da função de verossimilhança, p o número de parâmetros do modelo e n o número de observações. O modelo com menor valor do critério é o escolhido.

Os modelos considerados para esse experimento são apresentados a seguir.

3.2.1.1 Modelo em parcelas subdivididas

Seja Y a variável aleatória número de pegamentos de enxertos em parcelas de $m = 12$ plantas. A distribuição padrão a ser assumida é a binomial, de

índice m e parâmetro π (probabilidade de sucesso), isto é,

$$Y \sim Bin(m, \pi).$$

Assume-se a função de ligação logística e, inicialmente, independência das observações. Tem-se, então, como preditor linear o esquema em parcelas subdivididas, em que as parcelas estão no delineamento inteiramente casualizado com 5 repetições e os tratamentos no esquema fatorial 3×4 (3 tipos de porta-enxertos e 4 métodos de enxertia) e as subparcelas são as observações no tempo (8 tempos), isto é,

$$\eta_s = logit(\pi_s) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + (\alpha\beta\tau)_{ij\ell} \quad (42)$$

sendo $s = 1, \ldots, n$, μ o efeito associado à média geral, α_i o efeito associado ao i - ésimo método de enxertia, $i = 1, \ldots, 4$, β_j o efeito associado ao j-ésimo porta-enxerto, $j = 1, \ldots, 3$, γ_k o efeito associado às repetições, $k = 1, \ldots, 5$, δ_{ijk} o efeito associado ao erro de parcelas (representa a soma de $\gamma_k, \alpha\gamma_{ik}, \beta\gamma_{jk}$ e $\alpha\beta\gamma_{ijk}$), τ_ℓ o efeito associado ao ℓ-ésimo tempo em meses, $\ell = 1, \ldots, 8$ e $(\alpha\beta)_{ij}$, $(\alpha\tau)_{i\ell}$, $(\beta\tau)_{j\ell}$ e $(\alpha\beta\tau)_{ij\ell}$, os efeitos associados às interações.

Para a descrição do ajuste sequencial do modelo apresentado em (42), será usada a notação de Wilkinson & Rogers (1973):

Modelo	Notação
μ	1
α_i	A
$\alpha_i + \beta_j$	A + B
$\alpha_i + \beta_j + \tau_\ell$	A + B + D
$\alpha_i + \beta_j + (\alpha\beta)_{ij}$	A*B
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk}$	A*B*C
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell$	A*B*C + D
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell}$	A*(B*C + D)
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\beta\tau)_{j\ell}$	B*(A*C + D)
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell}$	A*B*C + D*(A + B)
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + (\alpha\beta\tau)_{ij\ell}$	(A*B)*(C + D)
$\alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + (\alpha\beta\tau)_{ij\ell} + \varepsilon_{ij\ell k}$	A*B*C*D

sendo que $\varepsilon_{ij\ell k}$ representa a soma de $\alpha\tau_{i\ell}, \alpha\tau\gamma_{i\ell k}, \beta\tau\gamma_{j\ell k}$ e $\alpha\beta\tau\gamma_{ij\ell k}$, no modelo saturado.

Para a verificação do ajuste do modelo, é utilizado o gráfico meio-normal (*half-normal plot*) com envelope simulado (Collett, 1991). A comparação entre modelos é feita usando-se o Critério Bayesiano de Schwarz (BIC) e o Critério de Informação de Akaike (AIC) sendo que a verificação da significância dos efeitos é feita através da estatística *deviance*.

3.2.1.2 Modelo de superdispersão com heterogeneidade constante

Segundo o pesquisador, alguns procedimentos para a coleta e manipulação do material experimental foram realizados por diferentes pessoas e, também, a coleta do material não foi feita na mesma árvore-fonte do camu-camu. Esses fatos podem acarretar em uma variância observada maior do que a esperada pelo modelo e, sendo assim, o modelo apresentado em (42) pode não se ajustar satisfatoriamente aos dados.

Uma alternativa é o modelo de superdispersão com heterogeneidade

constante, conforme apresentado na seção 2.4.3, em que:

$$E(Y) = m\pi \text{ e } V(Y) = \phi m\pi(1-\pi),$$

sendo ϕ, o fator de inflação da variância, estimado por:

$$\widehat{\phi} = \frac{1}{(n-p)} \sum_{s=1}^{n} \frac{(y_s - m\widehat{\pi}_s)^2}{m\widehat{\pi}_s(1-\widehat{\pi}_s)},$$

em que $\hat{\pi}$ é estimado a partir do modelo (42), pelo método da quase-verossimilhança.

Para a verificação do ajuste do modelo, é utilizado o gráfico meio-normal (*half-normal plot*) com envelope simulado (Collett, 1991). A comparação entre modelos é feita usando-se o Critério Bayesiano de Schwarz (BIC) e o Critério de Informação de Akaike (AIC) sendo que a verificação da significância dos efeitos dos fatores é feita pela estatística

$$F = \frac{D_1 - D_2}{r\hat{\phi}},$$

em que D_1 e D_2 são valores da estatística *deviance* para modelos encaixados com $p_1 < p_2$ parâmetros e $r = p_2 - p_1$.

Para a obtenção das estatísticas F é usado o proc GENMOD, enquanto que para a obtenção dos valores do Critério Bayesiano de Schwarz e Critério de Informação de Akaike é usado o proc NLMIXED. No Proc NLMIXED, o fator de dispersão é incluído diretamente no modelo e, portanto, basta fazer a diferença entre os valores da função de $-2\times$ {logaritmo da verossimilhança} dividido pela diferença entre os graus de liberdade para a obtenção das estatísticas F.

3.2.1.3 Aplicação das equações de estimação generalizadas

Um modelo alternativo leva em consideração a correlação entre as observações ao longo do tempo, conforme mostrado na seção 2.4.6, através das equações de estimação generalizadas.

Assume-se, então, o modelo logístico com preditor linear:

$$\eta_s = logit(\pi_s) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{43}$$

sendo μ o efeito associado à média geral, α_i o efeito associado ao i-ésimo método de enxertia, $i = 1, \ldots, 4$, β_j o efeito associado ao j-ésimo porta-enxerto, $j = 1, \ldots, 3$, $(\alpha\beta)_{ij}$ o efeito da interação entre α_i e β_j e $s = 1, \ldots, n$.

O processo de estimação dos parâmetros é semelhante ao caso em que se supõem observações independentes, sendo que na matriz de variâncias e covariâncias é introduzida uma matriz de correlação de "trabalho", $\boldsymbol{R(\alpha)}$, para se levar em consideração a correlação entre as observações ao longo dos meses, da seguinte forma:

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2},$$

sendo $\boldsymbol{A}_i$ a matriz diagonal com elementos de $V(Y_{ik})$.

Essa matriz é utilizada na obtenção das equações de estimação generalizadas, dadas por:

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}.$$

O algoritmo para a estimação dos parâmetros segue os passos:

1. calcula-se uma estimativa inicial de $\boldsymbol{\beta}$ considerando-se um modelo linear generalizado padrão, assumindo-se independência;

2. obtém-se a estimativa da matriz de correlação de trabalho $\boldsymbol{R}$ baseada nos resíduos e estrutura assumida;

3. calcula-se uma estimativa da matriz de covariância $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2}$;

4. atualiza-se o valor de $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{r+1} = \boldsymbol{\beta}_r + \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right]_r^{-1} \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right]_r \quad \text{e}$$

5. repete-se o processo, a partir de (2), até a convergência.

Para composição da matriz de correlação AR(1), do tipo dada em (24), usa-se o estimador

$$\hat{\alpha} = \frac{1}{n-p} \sum_{i=1}^{n} r_{ijk} r_{i,j,k+1},$$

em que r_{ij} é o resíduo de Pearson, obedecendo-se a lei de formação da matriz que é:

$$Corr(Y_{ijk}, Y_{i,j,k+1}) = \alpha^{|k-k'|}.$$

Para inferências a respeito dos parâmetros, usa-se o estimador "sanduíche"robusto, como discutido na seção 2.4.4, dado por:

$$\boldsymbol{V}_r(\widehat{\boldsymbol{\beta}}) = \Im^{-1}\boldsymbol{C}\Im^{-1},$$

sendo $\boldsymbol{D}_i = \dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_i}$, $\Im = \sum_{s=1}^{n} \boldsymbol{D}_i' \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{D}_i$, a matriz de informação e $\boldsymbol{C} = \sum_{s=1}^{n} \boldsymbol{D}_i' \widehat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)' \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{D}_i$.

Assintoticamente, $\widehat{\boldsymbol{\beta}}$ tem distribuição $N\big(\boldsymbol{\beta}, \boldsymbol{V}_r(\widehat{\boldsymbol{\beta}})\big)$ e, inferências podem ser feitas usando-se a estatística de Wald.

A seleção da matriz de correlação pode ser feita ajustando-se modelos com estruturas de covariâncias alternativas e comparando-se o Critério de Informação de Akaike (AIC). A checagem do modelo pode ser completada com gráficos de resíduos.

3.2.1.4 Modelo logístico normal

Outra forma de se incorporar no modelo a correlação entre observações ao longo do tempo, é considerar para o modelo logístico o preditor linear com a inclusão de uma variável latente, isto é,

$$\eta_s = logit(\pi_s) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + (\alpha\beta\tau)_{ij\ell} + u_{ij\ell k} \qquad (44)$$

sendo μ o efeito associado à média geral, α_i o efeito associado ao i-ésimo método de enxertia, $i = 1, \ldots, 4$, β_j o efeito associado ao j-ésimo porta-enxerto, $j = 1, \ldots, 3$, τ_ℓ o efeito associado ao ℓ-ésimo tempo em meses, $\ell = 1, \ldots, 8$, $u_{ij\ell k}$ o efeito aleatório associado ao i-ésimo indivíduo, δ_{ijk} o efeito associado ao erro de parcelas, $s = 1, \ldots, n$ e $(\alpha\beta)_{ij}$, $(\alpha\tau)_{i\ell}$, $(\beta\tau)_{j\ell}$ e $(\alpha\beta\tau)_{ij\ell}$ interações entre os efeitos principais. Assume-se que $u_{ij\ell k} = \sigma Z_{ij\ell k}$, sendo que $Z_{ij\ell k} \sim N(0,1)$ e, portanto, $u_{ij\ell k} \sim N(0, \sigma^2)$.

A função de verossimilhança nesse caso

$$L = \prod_{i=1}^{n} \int_{-\infty}^{+\infty} \binom{n}{y} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)^{n_i - y_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\eta-\mu)^2} d\eta,$$

não pode ser resolvida explicitamente. Assim, usa-se o método de quadratura Gaussiana para estimação dos parâmetros, dado que tem-se apenas um efeito aleatório no preditor linear.

Estudos adicionais precisam ser realizados para a verificação do ajuste do modelo, pois ainda não existem técnicas disponíveis.

3.2.1.5 Modelo considerando fator de dispersão e efeito aleatório

A fim de serem incorporados, simultaneamente, a variabilidade entre indivíduos e a correlação entre observações ao longo do tempo, considera-se, inicialmente, o modelo de superdispersão com heterogeneidade constante, como em 3.2.1.2, obtendo-se $\hat{\phi}$ a partir do preditor dado por (42). Em seguida, fixa-se $\phi = \hat{\phi}$ que é incorporado na função de quase-verossimilhança, com preditor linear dado por (44).

Também nesse caso, não se conseguem obter equações de quase-verossimilhança de forma explícita e, assim, usa-se o método de quadratura Gaussiana, considerando-se que há apenas um efeito aleatório no preditor linear a ser estimado.

Estudos adicionais precisam ser realizados para a verificação do ajuste do modelo, pois ainda não existem técnicas disponíveis.

3.2.2 Experimento 2

Os modelos utilizados para a análise desse experimento são descritos a seguir, sendo que os programas, em SAS, são apresentados no Anexo B.

3.2.2.1 Modelo em parcelas subdivididas

Seja Y a variável aleatória número de lagartas grandes. A distribuição padrão a ser assumida, neste caso, é a Poisson, de parâmetro λ, ou seja,

$$Y \sim Poi(\lambda).$$

Assume-se a função de ligação logarítmica e, inicialmente, independência das observações. Também neste caso, tem-se como preditor linear o esquema em parcelas subdivididas, em que as parcelas estão no delineamento inteiramente casualizado com 8 repetições, sendo formadas por 3 tratamentos e as subparcelas são as observações no tempo (9 tempos), isto é,

$$\eta = log\ (\lambda) = \mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell}, \tag{45}$$

sendo μ o efeito associado à média geral, α_i o efeito associado ao i-ésimo tratamento, $i = 1, \ldots, 3$, δ_{ik} o efeito associado ao erro de parcelas, τ_ℓ o efeito associado ao ℓ-ésimo tempo (semanas), $\ell = 1, \ldots, 9$, e $(\alpha\tau)_{i\ell}$ o efeito associado à interação.

Para a verificação do ajuste do modelo, é utilizada a *deviance residual*, que para um modelo bem ajustado deverá ser próxima ao número de graus de liberdade do modelo, e o gráfico meio-normal com envelope simulado. A verificação da significância dos efeitos é feita através da estatística *deviance*.

3.2.2.2 Modelo Poisson inflacionado de zeros

Devido ao excessivo número de zeros do experimento 2, apresentado na Tabela 3, acredita-se que o modelo Poisson padrão não se ajuste aos dados. Sendo assim, utiliza-se o modelo Poisson inflacionado de zeros como dado pela equação 32.

Os preditores lineares para o modelo de médias e modelo de zeros são dados, respectivamente, por:

$$\log(\lambda) = \mu + \alpha_i + (\alpha\gamma)_{ik} + \tau_\ell + (\alpha\tau)_{i\ell}$$

e

$$\log\left(\frac{\omega}{1-\omega}\right) = \mu + \alpha_i + \tau_\ell$$

sendo que os efeitos para esse modelo já foram definidos em (45).

O logaritmo da função de verossimilhança para esse modelo é apresentado pela equação 35. Deve-se observar que Z_i é uma variável dicotômica, podendo assumir:

$$Z_i = \begin{cases} 1, \text{ se} & Y_i = 0 \quad \text{zero estrutural} \\ 0, \text{ se} & Y_i \sim Poisson(\lambda_i) \end{cases},$$

ou seja, $Z_i \sim Bernoulli(\omega_i)$. A explicação para o zero estrutural é que a lagarta nunca será encontrada na planta, assim tem-se uma função de probabilidade degenerada, assumindo sempre o valor 1 e o zero Poisson, quer dizer que há uma probabilidade λ_i de ocorrer resposta zero.

Pelo fato de não se terem informações suficientes sobre o processo de geração dos zeros, a variável Z faz o papel de dados perdidos no algoritmo EM, discutido na seção (2.6.1).

A equação (35) mostra que a estimação dos parâmetros pode ser feita maximizando-se os dois termos de $\ell_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \boldsymbol{y})$, separadamente.

A matriz de variâncias e covariâncias assintótica pode ser estimada usando-se a inversa da matriz de informação, e inferências podem ser feitas usando-se testes da razão de verossimilhança.

3.2.2.3 Modelo Poisson inflacionado de zeros com efeito aleatório

Devido aos dados serem observados ao longo de 9 semanas, acredita-se que haja correlação entre as observações no tempo. Uma forma de se levar em consideração essa correlação é incluir uma variável latente no modelo. Assim, considera-se que os preditores lineares para o modelo de médias e modelo de zeros são dados, res-

pectivamente, por:

$$\begin{aligned} \log(\lambda) &= \mu + \alpha_i + (\alpha\gamma)_{ik} + \tau_\ell + (\alpha\tau)_{i\ell} + (\alpha\tau\gamma)_{i\ell k} + \sigma Z \\ \log\left(\frac{\omega}{1-\omega}\right) &= \mu + \alpha_i \end{aligned} \tag{46}$$

sendo μ o efeito associado à média geral, α_i o efeito associado ao i-ésimo tratamento, $i = 1, \ldots, 3$, τ_ℓ o efeito associado ao ℓ-ésimo tempo em semanas, $\ell = 1, \ldots, 9$, γ_k a repetição, $k = 1, \ldots, 8$, $Z \sim N(0, \sigma^2)$ e $(\alpha\gamma)_{ik}$, $(\alpha\tau)_{i\ell}$ e $(\alpha\tau\gamma)_{i\ell k}$, os efeitos associados às interações.

Na estimação dos parâmetros desse modelo, usa-se o algoritmo EM com quadratura Gaussiana.

Para saber se a inclusão do efeito aleatório é significativo, basta fazer a diferença dos valores máximos das funções de verossimilhanças entre o modelo sem e com o efeito aleatório. Sob $H_0 : \sigma = 0$, $-2\times$ {diferença entre os valores máximos das funções de verossimilhanças} tem distribuição assintótica que é uma mistura $50 : 50$ de χ^2_0 e χ^2_1.

Para seleção de modelos usa-se o Critério Bayesiano de Schwarz (BIC) e o Critério de Informação de Akaike.

4 RESULTADOS E DISCUSSÃO

4.1 Experimento 1 - Comparação de métodos de enxertia e tipos de porta-enxertos para camu-camu

4.1.1 Ajuste do modelo usando parcelas subdivididas

Analisando-se, inicialmente, os dados da Tabela 2, como se as observações fossem independentes, em um esquema em parcelas subdivididas, supondo-se a distribuição binomial para a variável resposta, os resultados obtidos para os diferentes submodelos estão apresentados na Tabela 5.

Embora esses modelos não sejam adequados aos dados, pois não levam em consideração a correlação entre as observações ao longo do tempo, pode-se observar que o modelo com menores AIC e BIC tem preditor linear dado por:

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\tau\beta)_{\ell j}.$$

A partir da Tabela 5 pode-se obter a análise de *deviance* que é apresentada na Tabela 6. Note que o valor de -2 log *ver* do modelo saturado é simplesmente o termo constante da distribuição binomial, ou seja, $\log \binom{n}{y} = 847,72$.

Pode-se ver que existem evidências de efeito significativo da interação Método × Porta-enxerto, de Tempo e da interação de Porta-enxerto × Tempo, não sendo significativo, porém, o efeito da interação tripla Método × Porta-enxerto × Tempo nem da interação Método × Tempo, confirmando a escolha do modelo pelos critérios AIC e BIC.

Deve-se notar que o efeito da interação Método×Tempo apresenta-se não significativo se ajustado para a interação Porta-enxerto × Tempo, enquanto que

Tabela 5. Valores de $-2\log\ ver$, critérios AIC e BIC, com os respectivos graus de liberdade, para diversos modelos ajustados aos dados da Tabela 2.

Modelos (η)	g.l.	-2 log *ver.*	AIC	BIC
Constante	479	3791,6	3793,6	3797,8
A	476	3460,1	3468,1	3484,8
B	477	3074,0	3080,0	3092,5
A + B	474	2696,2	2708,2	2733,2
A + B + D	467	2419,8	2445,8	2500,0
A*B	468	2619,5	2643,5	2693,6
A*B*C	420	1847,0	1967,0	2217,4
A*B*C + D	413	1511,1	1645,5	1925,1
B*(A*C + D)	399	1072,6	**1234,6**	**1572,6**
A*(B*C + D)	392	1473,7	1649,6	2016,9
A*B*C + D*(A + B)	378	1045,3	1249,3	1675,0
(A*B)*(C + D)	336	1021,6	1309,6	1910,7
A*B*C*D	0	847,2	1552,3	2854,5

é significativo, se não ajustado. Isso mostra a não ortogonalidade das colunas da matriz $\boldsymbol{X'WX}$, o que é uma característica quando se usa outra distribuição para os dados que não a distribuição normal.

Essas conclusões, porém, têm que ser olhadas com cautela, pois, como pode ser visto pela Figura 4, existem evidências de que mesmo o modelo com interação tripla não se ajusta bem aos dados. Isso já era esperado, pois a dependência entre observações ao longo do tempo não foi considerada. Nota-se, também, que a *deviance* residual é bastante menor do que o número de graus de liberdade, mostrando evidências de subdispersão.

Tabela 6. Análise de *deviance* para os dados da Tabela 2.

Causa de variação	g.l.	*Deviance*	Valor de p
Método	3	331,5	$< 0,0001$
Porta|Método	2	763,9	$< 0,0001$
Porta	2	717,7	$< 0,0001$
Método|Porta	3	377,8	$< 0,0001$
Método×Porta	6	76,7	$< 0,0001$
Resíduo (A)	48	772,5	$< 0,0001$
Tempo	7	335,9	$< 0,0001$
Tempo×Porta|Tempo×Método	14	428,4	$< 0,0001$
Tempo×Método|Tempo×Porta	21	27,3	$< 0,1616$
Tempo×Método×Porta	42	23,6	$< 0,9901$
Resíduo (B)	336	174,5	

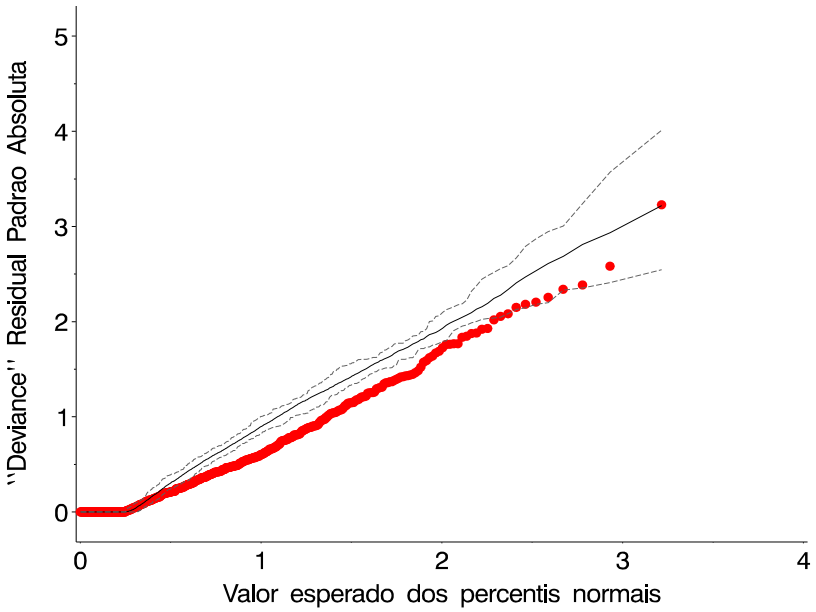

Figura 4 - Gráfico meio normal não considerando a superdispersão.

4.1.2 Ajuste do modelo considerando a subdispersão

Como o modelo binomial padrão não se ajusta bem aos dados, ajusta-se o modelo com fator de heterogeneidade constante, usando-se o método de quase-

verossimilhança para a estimação dos parâmetros, sendo o preditor linear o mesmo. Tem-se que $\hat{\phi} = 0,4640 (< 1)$, confirmando as evidências de subdispersão. Os resultados obtidos pelo ajuste dos diferentes submodelos estão apresentados na Tabela 7, enquanto que o gráfico meio-normal com envelope simulado é mostrado na Figura 5, evidenciando a falta de ajuste desse modelo, também. Pode-se, porém, ver pela Tabela 7 que o modelo com menor AIC é aquele com preditor linear dado por

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell},$$

enquanto que o modelo com menor BIC é aquele com preditor linear

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\beta\tau)_{j\ell}.$$

A partir da Tabela 7 pode-se construir a Análise de *Deviance* que é mostrada na Tabela 8. Pode-se ver que existem evidências de efeito significativo de quase todos os fatores do modelo, só não sendo significativo, porém, o efeito da interação tripla Método × Porta-enxerto × Tempo, o que confirma o resultado obtido pelo critério AIC.

Novamente, deve-se ter cuidado nas conclusões, já que, como pode ser visto pela Figura 5, existem evidências de que mesmo o modelo com interação tripla não se ajusta bem aos dados. Isso já era esperado, pois a dependência entre observações ao longo do tempo não foi considerada.

Tabela 7. Valores de $-2\log$ *ver*, critérios AIC e BIC, com os respectivos graus de liberdade, para diversos modelos ajustados aos dados da Tabela 2, considerando-se fator de dispersão constante.

Modelo	g.l.	-2 log *ver.*	AIC	BIC
Constante	479	8171,0	8173,0	8177,2
A	476	7456,5	7464,5	7481,2
B	477	6624,4	6630,4	6642,9
A + B	474	5810,3	5822,3	5847,3
A*B	468	5645,1	5669,1	5719,2
A*B*C	420	3980,2	4100,2	4350,7
A*B*C + D	413	3257,3	3391,3	3670,9
B*(A*C + D)	399	2311,4	2473,4	**2811,4**
A*(B*C + D)	392	3175,7	3351,7	3719,0
A*B*C + D*(A + B)	378	2252,6	**2456,6**	2882,3
(A*B)*(C + D)	336	2201,6	2489,6	3090,7
A*B*C*D	0	1825,9		

Tabela 8. Análise de *deviance*, considerando fator de dispersão constante.

Causa de variação	g.l.	*Deviance*	*Valor F*	Valor p
Método	3	714,5	238,2	$< 0,0001$
Porta | Método	2	1646,4	823,2	$< 0,0001$
Porta	2	1546,7	773,4	$< 0,0001$
Método | Porta	3	814,2	271,4	$< 0,0001$
Método×Porta	6	165,2	27,5	$< 0,0001$
Resíduo (A)	48	1664,9		
Tempo	7	723,1	103,3	$< 0,0001$
Tempo×Método | Tempo×Porta	21	58,8	2,8	$< 0,0001$
Tempo×Porta | Tempo×Método	14	923,3	66,0	$< 0,0001$
Tempo×Método×Porta	42	50,9	1,21	$< 0,1811$
Resíduo (B)	336	174,5		

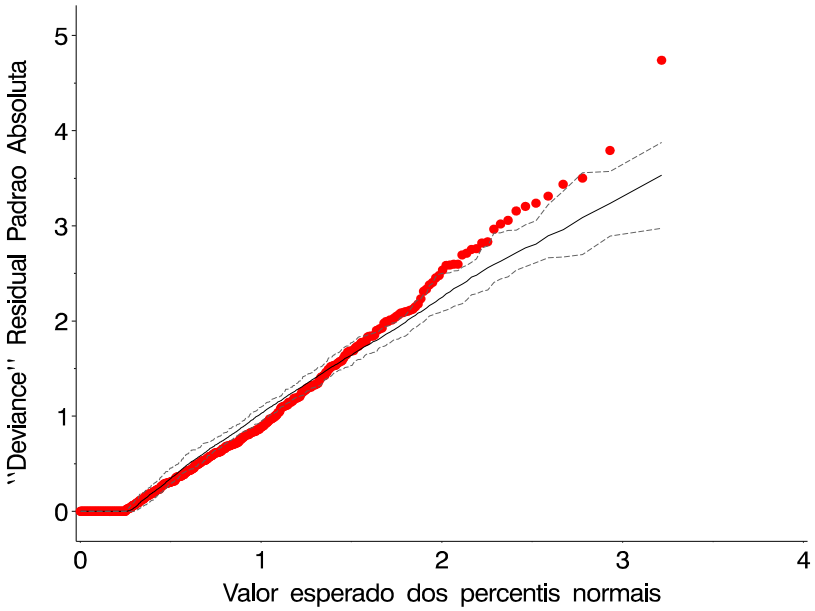

Figura 5 - Gráfico meio normal, levando-se em consideração o parâmetro de dispersão

4.1.3 Modelo incorporando matriz de correlação

Outra forma de se analisarem os dados é não incluir o tempo no preditor e escolher uma matriz de correlação que melhor represente a dependência das observações ao longo do tempo. Assim, o preditor linear é dado por:

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

sendo que os efeitos já foram definidos anteriormente.

Para se ter uma idéia para a matriz de correlação de trabalho, pode-se calcular a correlação observada para a variável resposta ao longo dos meses. Para os dados da Tabela 2, tem-se que a matriz de correlação, eliminados os efeitos de Método, Porta-enxerto e interação Método×Porta-enxerto, está apresentada na Tabela 9.

Tabela 9. Matriz de correlação observada

	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Dez	1							
Jan	0,6835	1						
Fev	0,6256	0,8951	1					
Mar	0,6287	0,8429	0,9318	1				
Abr	0,5369	0,7382	0,8450	0,9307	1			
Mai	0,5414	0,7117	0,8217	0,8960	0,9636	1		
Jun	0,5689	0,7030	0,7934	0,8668	0,9380	0,9635	1	
Jul	0,5809	0,7018	0,7885	0,8293	0,8812	0,9069	0,9551	1

Observa-se, pela Tabela 9, que a correlação decresce ao longo do tempo. Sendo assim, há indícios de que uma estrutura de correlação de trabalho adequada seja a AR(1), embora se tenha ajustado um modelo baseado em três estruturas

de correlação diferentes: AR(1), Simetria Composta e Independência. As matrizes de correlação de trabalho correspondentes, estimadas no SAS para o processo de estimação dos $\boldsymbol{\beta}'s$, são apresentadas nas Tabelas 10, 11 e 12, respectivamente.

Tabela 10. Matriz de correlação de trabalho baseada na estrutura AR(1)

	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Dez	1							
Jan	0,4110	1						
Fev	0,1689	0,4110	1					
Mar	0,0694	0,1689	0,4110	1				
Abr	0,0285	0,0694	0,1689	0,4110	1			
Mai	0,0117	0,0285	0,0694	0,1689	0,4110	1		
Jun	0,0048	0,0117	0,0285	0,0694	0,1689	0,4110	1	
Jul	0,0020	0,0048	0,0117	0,0285	0,0694	0,1689	0,4110	1

Tabela 11. Matriz de correlação de trabalho baseada na estrutura Simetria Composta

	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Dez	1							
Jan	0,3201	1						
Fev	0,3201	0,3201	1					
Mar	0,3201	0,3201	0,3201	1				
Abr	0,3201	0,3201	0,3201	0,3201	1			
Mai	0,3201	0,3201	0,3201	0,3201	0,3201	1		
Jun	0,3201	0,3201	0,3201	0,3201	0,3201	0,3201	1	
Jul	0,3201	0,3201	0,3201	0,3201	0,3201	0,3201	0,3201	1

As estimativas dos efeitos do modelo, bem como os erros padrões empírico e baseado no modelo, para diferentes estruturas de correlação, são apre-

Tabela 12. Matriz de correlação de trabalho baseada na estrutura de Independência

	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Dez	1							
Jan	0	1						
Fev	0	0	1					
Mar	0	0	0	1				
Abr	0	0	0	0	1			
Mai	0	0	0	0	0	1		
Jun	0	0	0	0	0	0	1	
Jul	0	0	0	0	0	0	0	1

sentados na Tabela 13. Observa-se que as estimativas dos efeitos e dos erros padrões empíricos são iguais para Simetria Composta e Independência, diferindo apenas no erro padrão baseado no modelo. Também, o erro estimado baseado no modelo para a estrutura AR(1) é menor do que as outras estruturas, refletindo a diminuição da correlação ao longo do tempo.

Pela análise de *deviance*, baseada nas estatísticas escores para a estrutura AR(1), apresentada na Tabela 14, observa-se que a interação Método de enxertia×tipo de Porta-enxerto não é significativa, havendo significância dos efeitos principais.

Os contrastes para os efeitos principais são apresentados na Tabela 15, e mostram que o método da Fenda Cheia e Fenda Lateral não diferem entre si mas diferem dos métodos de Inglês Simples e Fenda de Colo, ao nível de significância de 5% e, ainda, que Fenda de Colo difere de Inglês Simples. Em relação aos Porta-enxertos, vê-se que o camu-camu difere da Goiabeira e da Pitangueira, sendo que essas também diferem entre si.

Estudos adicionais precisam ser realizados para a verificação do ajuste do modelo, pois ainda não existem técnicas disponíveis.

Tabela 13. Estimativas e erros padrões empírico e baseado no modelo, para diferentes estruturas de correlação

Efeitos	AR(1)	Simetria Composta	Independente
Constante	-1,7823 (0,1999;0,3574)	-1,6551 (0,1952;0,4221)	-1,6551 (0,1952;0,2344)
Método 1	-0,3380 (0,2964;0,5407)	-0,3100 (0,2897;0,6330)	-0,3100 (0,2897;0,3516)
Método 2	0,0057 (0,3440;0,5049)	-0,0156 (0,3319;0,5985)	-0,0156 (0,3319;0,3324)
Método 3	-0,9514 (0,4229;0,6344)	-0,9509 (0,4182;0,7435)	-0,9509 (0,4182;0,4130)
Porta 1	1,9622 (0,3906;0,4373)	1,8138 (0,3810;0,5241)	1,8138 (0,3810;0,2911)
Porta 2	0,6095 (0,3416;0,4636)	0,4772 (0,3194;0,5580)	0,4772 (0,3194;0,3100)
Método 1×Porta 1	0,0751 (0,7861;0,6472)	0,0596 (0,7776;0,7703)	0,0596 (0,7776;0,4279)
Método 1×Porta 2	0,9893 (0,4604;0,6685)	0,9144 (0,4364;0,7988)	0,9144 (0,4364;0,4437)
Método 2×Porta 1	1,1537 (0,6248;0,6435)	1,1793 (0,6126;0,7740)	1,1793 (0,6126;0,4300)
Método 2×Porta 2	0,6797 (0,5525;0,6395)	0,6916 (0,5218;0,7704)	0,6916 (0,5218;0,4280)
Método 3×Porta 1	-0,2511 (0,7553;0,7395)	-0,2624 (0,7503;0,8801)	-0,2624 (0,7503;0,4889)
Método 3×Porta 2	-0,1274 (0,5323;0,8201)	-0,1155 (0,5066;0,9813)	-0,1155 (0,5066;0,5451)

Tabela 14. Análise de *deviance* para os efeitos, considerando-se a estrutura de correlação AR(1)

Causa de variação	g.l.	*deviance*	Valor p
Método	3	14,41	0,0024
Porta	2	18,25	0,0001
Método×Porta	6	7,67	0,2633

4.1.4 Ajuste considerando o efeito aleatório

Considerando-se o modelo logístico normal para os dados da Tabela 2, obtêm-se os resultados apresentados na Tabela 16. É interessante observar que o BIC para todos os efeitos do modelo aleatório, foram menores do que quando considerado apenas o fator de dispersão. Isso indica que esse modelo ajusta-se melhor do que o modelo que considera apenas a inclusão do parâmetro de dispersão constante.

Pode-se observar que o modelo com menores AIC e BIC tem preditor

Tabela 15. Teste escore para os contrastes entre os efeitos principais, considerando-se a estrutura de correlação AR(1)

Contrastes	g.l.	X^2	Valor p
Método 1 e 2 vs Método 3 e 4	1	12,39	0,0004
Método 1 vs Método 2	1	3,24	0,0720
Método 3 vs Método 4	1	9,84	0,0017
Porta 1 vs Porta 2 e 3	1	10,58	0,0011
Porta 2 vs Porta 3	1	13,56	0,0002

Tabela 16. Ajuste dos modelos com a inclusão do efeito aleatório

Modelo	g.l.	-2 log *ver.*	AIC	BIC	$\hat{\sigma}$(erro)
Constante	479	3791,6	3793,6	3797,8	
Constante + σZ	478	2119,6	2123,6	2127,8	1,6236 (0,1682)
A + σZ	475	2105,6	2115,6	2126,1	1,4374 (0,1492)
A + B + σZ	473	2079,1	2093,1	2107,8	1,0735 (0,1195)
B + σZ	476	2100,5	2108,5	2116,8	1,3197 (0,1427)
B + A + σZ	473	2079,1	2093,1	2107,8	1,0735 (0,1195)
A*B + σZ	467	2070,1	2096,7	2123,9	0,9951 (0,1112)
A*B*C + σZ	419	1969,1	2043,1	2120,6	0,3594 (0,0551)
A*B*C + D + σZ	412	1637,2	1725,2	1817,3	0,3933 (0,0580)
A*(B*C + D) + σZ	391	1599,0	1729,0	1865,1	0,3944 (0,0584)
A*B*C + D*(A + B)+ σZ	377	1181,8	**1339,8**	**1505,2**	0,4597 (1,0175)
(A*B)*(C + D) + σZ	335	1157,2	1399,2	1652,6	0,4553 (1,0164)
A*B*C*D + σZ	0				

linear dado por:

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + \sigma Z.$$

A partir da Tabela 16 pode-se obter a análise de *deviance* que é apresentada na Tabela 17.

Tabela 17. *Deviances* para os efeitos do modelo considerando-se a inclusão do efeito aleatório

Causa de variação	g.l.	*Deviance*	Valor p
σ	1	1672,0	$< 0,0001$
Método	3	14,0	$< 0,0001$
Porta|Método	2	26,5	$< 0,0001$
Porta	2	19,1	$< 0,0001$
Método|Porta	3	21,4	$< 0,0001$
Método×Porta	6	9,0	$< 0,0001$
Resíduo (A)	48	101,0	$< 0,0001$
Tempo	7	331,9	$< 0,0001$
Método×Tempo	21	38,2	$< 0,0001$
Porta×Tempo	14	417,2	$< 0,0001$
Método×Porta×Tempo	42	24,6	$< 0,0001$
Resíduo (B)	335		

É interessante observar que, embora os critérios AIC e BIC concordem quanto aos modelos a serem selecionados, indicando não haver interação entre Método×Porta-enxerto, Tempo×Métodos e tambémda interação tripla Método×Porta-enxerto×Tempo, a análise de *deviance* não detectou tais diferenças. Deve-se olhar com cuidado para esses resultados.

4.1.5 Considerando fator de dispersão e efeito aleatório

Devido à forma como foi conduzido o experimento, acredita-se que é possível haver mais de uma fonte de variabilidade, além daquela atribuída à correlação relativa ao tempo. Assim, inclui-se no modelo o parâmetro de dispersão cuja estimativa é dada por $\hat{\phi} = 0,4640$, calculada pelo ajuste do modelo maximal:

$$\eta = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + (\alpha\beta\tau)_{ij\ell}.$$

Os resultados obtidos para os diferentes modelos encontram-se na Tabela 18.

Tabela 18. Valores de $-2\log\ ver$, critérios AIC e BIC, com os respectivos graus de liberdade, para diversos modelos ajustados aos dados da Tabela 2, considerando-se a inclusão do efeito aleatório.

Modelos	g.l.	-2 log *ver.*	AIC	BIC	σ (erro)
σZ	478	1042,2	1046,2	1050,2	1,5084 (0,1601)
A + σZ	475	1027,5	1037,5	1048,3	1,3339 (0,1443)
A + B + σZ	473	994,2	1008,2	1023,4	0,9303 (0,1120)
A*B + σZ	467	985,2	1011,2	1039,2	0,8610 (0,1047)
A*B*C + D + σZ	412	842,4	882,4	925,6	0,9418 (0,1135)
A*(B*C + D) + σZ	391	826,4	908,4	996,9	0,9394 (0,1130)
A*B*C + D*(A + B)+ σZ	377	635,1	**745,1**	**863,9**	0,9653 (0,1162)
(A*B)*(C + D) + σZ	335	624,4	818,4	1027,8	0,9549 (1,0211)

Pode-se ver que o modelo com menores AIC e BIC é aquele com preditor linear dado por:

$$\eta_s = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ijk} + \tau_\ell + (\alpha\tau)_{i\ell} + (\beta\tau)_{j\ell} + \sigma Z. \tag{47}$$

A partir da Tabela 18 pode-se construir a análise de *deviance* que é mostrada na Tabela 19. Pode-se ver que existem evidências de efeito significativo para os efeitos principais e da interação Porta×Tempo, confirmando o resultado obtido pelo AIC e BIC.

A interação Porta-enxerto×Tempo é significativa, o que, de certa forma, confirma os resultados gráficos dos valores observados, como apresentados nas Figuras 6 e 7. Os testes para os contrastes dos métodos de enxertia são apresentados na Tabela 20. Os resultados apenas confirmam as diferenças observadas na Figura 7, ou seja, fenda cheia não difere da fenda lateral e nem de colo, sendo que todas diferem do inglês.

Na Figura 6 observa-se uma estabilidade maior para o porta-enxerto camu-camu, sendo que os pegamentos dos porta-enxertos Goiabeira e Pitangueira

Tabela 19. Análise de *deviances* para o modelo considerando a correlação e o efeito aleatório conjuntamente

Causa de variação	g.l.	*Deviance*	Valor F	Valor de p
σ	1			
Método	3	14,7	4,9	0,0047
Porta	2	33,3	16,65	$< 0,0001$
Método×Porta	6	9,0	1,5	0,1984
Tempo	7	142,8	20,40	$< 0,0001$
Método×Tempo	21	16,0	0,76	$< 0,7683$
Porta×Tempo	14	191,3	13,66	$< 0,0001$
Método×Porta×Tempo	42	10,7	0,25	$0,999$

Tabela 20. Valores das probabilidades para os contrastes dos efeitos dos métodos de enxertia

	Fenda cheia	Fenda lateral	Inglês
Fenda lateral	0,1026		
Inglês	0,0005	<0,0001	
Colo	0,0984	0,6182	<0,0001

tendem a zero, como pode ser observado para o mês de julho. Na Tabela 21 são apresentadas as médias e as proporções (entre parênteses) de pegamentos, considerando-se os tempos e os tipos de porta-enxertos. Nota-se o decréscimo das médias de pegamento, sendo o porta-enxerto camu-camu mais estável, atingindo média final de pegamento de 5,90, o que corresponde a 49,17% enquanto as demais tendem a zero. O comportamento do número de pegamentos em relação aos métodos de enxertia pode ser observado na Figura 7. Deve-se notar uma maior variabilidade inicial para os métodos de enxertia sendo menor no final do experimento. A variabilidade é decrescente ao longo do tempo mas o comportamento não se altera.

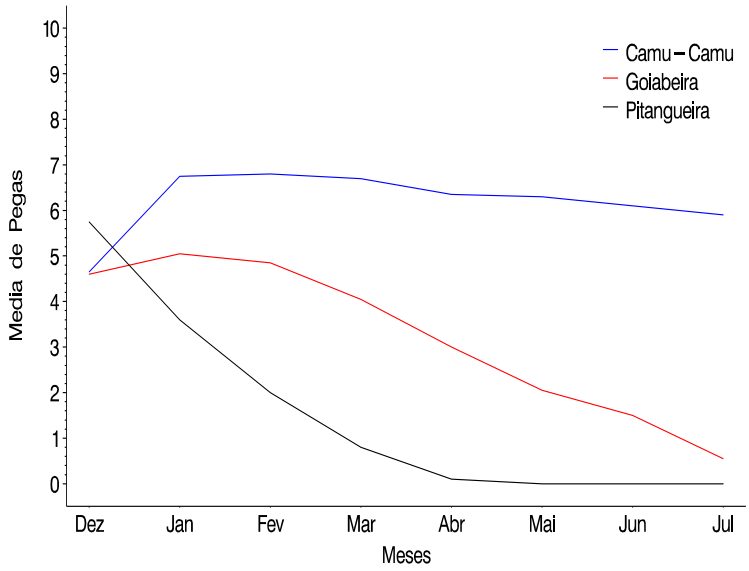

Figura 6 - Totais observados para espécies

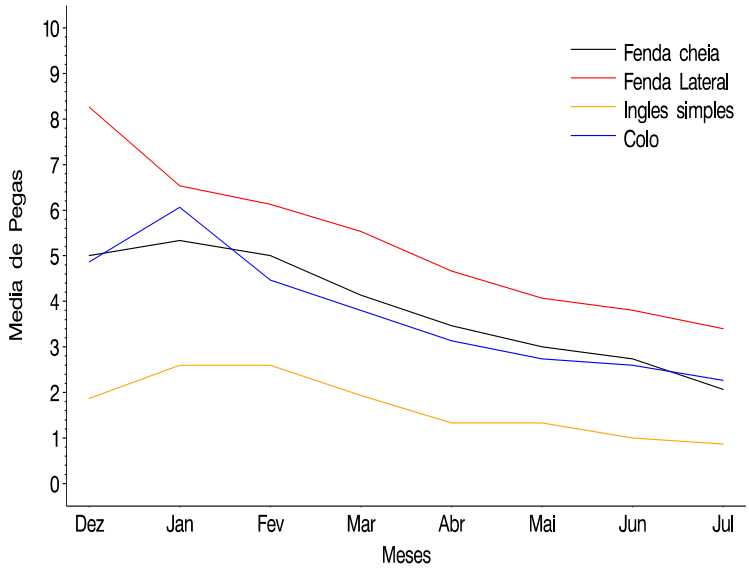

Figura 7 - Totais observados para métodos de enxertias

Tabela 21. Médias (proporções) de pegamentos dos porta-enxertos

Meses	Camu-camu	Goiabeira	Pitangueira
Dez	4,65 (0,3875)	4,60 (0,3822)	5,75 (0,4792)
Jan	6,75 (0,5625)	5,05 (0,4208)	3,60 (0,3000)
Fev	6,80 (0,5667)	4,85 (0,4042)	2,00 (0,1667)
Mar	6,70 (0,5583)	4,05 (0,3375)	0,80 (0,0667)
Abr	6,35 (0,5292)	3,00 (0,2500)	0,10 (0,0083)
Mai	6,30 (0,5250)	2,05 (0,1708)	0,00 (0,0000)
Jun	6,10 (0,5083)	1,50 (0,1250)	0,00 (0,0000)
Jul	5,90 (0,4917)	0,55 (0,0458)	0,00 (0,0000)

Tabela 22. Médias (proporções) de pegamentos dos métodos de enxertia

Meses	Fenda cheia	Fenda lateral	Inglês simples	Colo
Dez	5,00 (0,4167)	8,27 (0,6889)	1,87 (0,1556)	4,87 (0,4056)
Jan	5,33 (0,4444)	6,53 (0,5444)	2,60 (0,2167)	6,07 (0,5056)
Fev	5,00 (0,4167)	6,13 (0,5111)	2,60 (0,2167)	4,47 (0,3722)
Mar	4,13 (0,3444)	5,53 (0,4611)	1,93 (0,1611)	3,80 (0,3167)
Abr	3,47 (0,2889)	4,67 (0,3889)	1,33 (0,1111)	3,13 (0,2611)
Mai	3,00 (0,2500)	4,07 (0,3389)	1,33 (0,1111)	2,73 (0,2278)
Jun	2,73 (0,2278)	3,80 (0,3167)	1,00 (0,0833)	2,60 (0,2167)
Jul	2,07 (0,1722)	3,40 (0,2833)	0,87 (0,0722)	2,27 (0,1889)

4.2 Experimento 2 - Comparação de milho geneticamente modificado MON810 e milho convencional (híbrido DKB909)

4.2.1 Ajuste do modelo usando parcelas subdivididas

Analisando-se, inicialmente, os dados da Tabela 3, como se as observações fossem independentes, em um esquema em parcelas subdivididas, supondo-se a distribuição Poisson para a variável resposta, os resultados obtidos para os diferentes submodelos estão apresentados na Tabela 23.

Tabela 23. Valores de $-2\log$ *ver*, critérios AIC e BIC, com os respectivos graus de liberdade, para diversos modelos ajustados aos dados da Tabela 3.

Modelos	g.l.	-2 log *ver.*	AIC	BIC
μ	215	713,6	715,6	718,9
$\mu + \alpha_i$	213	554,1	560,1	570,2
$\mu + \alpha_i + \delta_{ik}$	192	521,5	569,5	650,5
$\mu + \alpha_i + \delta_{ik} + \tau_\ell$	184	264,3	328,3	**436,3**
$\mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell}$	168	229,4	**325,4**	487,4
$\mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell} + (\tau\gamma)_{\ell k} + (\alpha\tau\gamma)_{i\ell k}$	0	184,9		

Embora esses modelos não levem em consideração a correlação entre as observações ao longo do tempo, pode-se observar que o modelo com menor AIC é aquele com preditor linear dado por

$$\eta = \mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell},$$

enquanto que o modelo com menor BIC é aquele com preditor linear

$$\eta = \mu + \alpha_i + \delta_{ik} + \tau_\ell.$$

A partir da Tabela 23 pode-se construir a análise de *deviance* que é mostrada na Tabela 26. Os resultados da análise mostram que existem evidências

da interação significativa entre Tratamentos×Tempo, concordando com o resultado obtido pelo critério AIC.

Tabela 24. Análise de *deviances* para os dados da Tabela 3

Causa de variação	g.l.	*Deviance*	Valor de p
Tratamentos	2	159,5	$< 0,0001$
Resíduo (A)	21	32,6	
Tempo	8	257,2	$< 0,0001$
Tratamentos×Tempo	16	34,9	0,0041
Resíduo (B)	168	44,5	

O gráfico meio-normal, Figura 8, evidencia problemas na cauda, com grande concentração de zeros, indicando problemas no ajuste, devido ao excesso de respostas iguais a zero, como pode ser notado das Tabelas 3 e 4. Uma forma de se contornar esse problema é usar o modelo Poisson inflacionado de zeros.

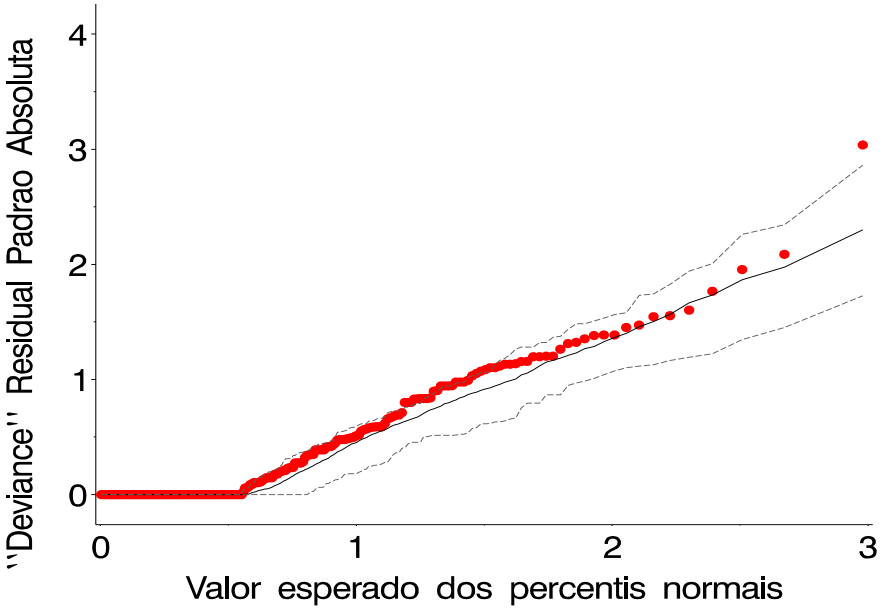

Figura 8 - Ajuste da distribuição Poisson ao número de lagartas

4.3 Ajuste do modelo usando ZIP

Como o modelo Poisson padrão evindencia problemas na cauda, devido à grande quantidade de zeros, opta-se pelo modelo Poisson inflacionado de zeros, como dado pela equação (33). Os preditores lineares para o modelo de médias e de zeros são, respectivamente

$$\log(\lambda) = \mu + \alpha_i + (\alpha\gamma)_{ik} + \tau_\ell + (\alpha\tau)_{i\ell}$$

e

$$\log\left(\frac{\omega}{1-\omega}\right) = \mu + \alpha_i + \tau_\ell$$

Os resultados, considerando fixo o modelo de zeros e variando o modelos de médias, são apresentados na Tabela 25.

Tabela 25. Valores de $-2\log$ *ver*, critérios AIC e BIC, com os respectivos graus de liberdade, para modelos ZIP ajustados aos dados da Tabela 3.

Modelos	g.l.	-2 log *ver.*	AIC	BIC
Fixando-se: $\log\left(\frac{\omega}{1-\omega}\right) = \mu + \alpha_i + \tau_\ell$				
μ	204	377,3	401,3	441,8
$\mu + \alpha_i$	202	294,0	**322,0**	**369,3**
$\mu + \alpha_i + \delta_{ik}$	181	261,2	331,2	449,4
$\mu + \alpha_i + \delta_{ik} + \tau_\ell$	173	245,6	331,6	476,7
$\mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell}$	157	229,4	347,4	546,5
$\mu + \alpha_i + \delta_{ik} + \tau_\ell + (\alpha\tau)_{i\ell} + \varepsilon_{i\ell k}$	0	184,9		

Pelos resultados apresentados na Tabela 25, fixado o modelo de zeros, os valores dos critérios AIC e BIC resultam no modelo:

$$\begin{aligned} \log(\lambda) &= \mu + \alpha_i \\ \log\left(\frac{\omega}{1-\omega}\right) &= \mu + \alpha_i + \tau_\ell. \end{aligned}$$

A partir da Tabela 25 pode-se construir a análise de *deviance* que é mostrada na Tabela 26. Os resultados da análise mostram que existem evidências da interação significativa entre Tratamentos×Tempo, concordando com o resultado obtido pelo critério AIC.

Tabela 26. Análise de *deviance* para o modelo Poisson inflacionado de zeros

Causa de variação	g.l.	*Deviance*	Valor de p
Tratamentos	2	83,3	$< 0,0001$
Resíduo (A)	21	32,8	
Tempo	8	15,6	0,0485
Tratamentos×Tempo	16	16,2	0,4391
Resíduo (B)	168	44,5	

Note que, fixado o modelo de zeros, a interação entre Tratamentos×Tempo é não significativa, havendo significância dos efeitos principais. Além disso, como a *deviance* residual é muito menor do que o número de graus de liberdade, tem-se evidência de uma subdispersão.

4.4 Ajuste do modelo usando ZIP com efeito aleatório

O modelo Poisson inflacionado de zeros discutido na seção 4.3 não leva em consideração a correlação que pode existir entre as observações ao longo do tempo. Uma forma de se incluir essa correlação no modelo é com a adição de um efeito aleatório. Assim, os preditores lineares ficam definidos por:

$$\log(\lambda) = \mu + \alpha_i + (\alpha\gamma)_{ik} + \tau_\ell + (\alpha\tau)_{i\ell} + \phi$$

e

$$\log\left(\frac{\omega}{1-\omega}\right) = \mu + \alpha_i + \tau_\ell.$$

Também nesse caso, fixa-se o modelo de zeros, variando apenas o modelo de médias. Os resultados obtidos para os diferentes submodelos estão apresentados na Tabela 27.

Tabela 27. Valores de $-2\log$ *ver*, critérios AIC e BIC, com os respectivos graus de liberdade, para diversos modelos ajustados com efeito aleatório.

Modelos	g.l.	-2 log *ver.*	AIC	BIC	ϕ (erro)
$\mu+\phi$	214	329,9	355,9	371,2	1,2699 (0,3483)
$\mu+\alpha_i+\phi$	212	291,2	**321,2**	**338,8**	0,0000 (0,2841)
$\mu+\alpha_i+\delta_{ik}+\phi$	191	261,2	333,2	375,6	0,0000 (0,0740)
$\mu+\alpha_i+\delta_{ik}+\tau_\ell+\phi$	183	245,6	333,6	385,5	0,0000 (0,0741)
$\mu+\alpha_i+\delta_{ik}+\tau_\ell+(\alpha\tau)_{i\ell}+\phi$	167	228,4	348,4	419,1	0,0000 (0,0739)
$\mu+\alpha_i+\delta_{ik}+\tau_\ell+(\alpha\tau)_{i\ell}+(\tau\gamma)_{\ell k}+(\alpha\tau\gamma)_{i\ell k}+\phi$	0	184,9			

Pode-se notar, da Tabela 27, que os valores dos critérios AIC e BIC para o modelo Poisson inflacionado de zeros com efeito aleatório, são menores do que os valores obtidos para o modelo inflacionado de zeros, embora a estimativa do efeito aleatório seja aproximadamente, zero.

Tabela 28. Análise de *deviances* para o modelo Poisson inflacionado de zeros

Causa de variação	g.l.	*Deviance*	Valor de p
Tratamentos	2	38,7	$< 0,0001$
Resíduo (A)	21	30,0	
Tempo	8	15,6	0,0485
Tratamentos$\times$Tempo	16	17,2	0,3728
Resíduo (B)	167	43,5	

Da Tabela 28 percebe-se pouca alteração em relação aos resultados da Tabela 26. Sendo assim, dá-se preferência ao modelo Poisson inflacionado de zeros sem o efeito aleatório, uma vez que ele não é significativo.

As médias estimadas pela parte do modelo Poisson, chamado modelo de médias, são:

i) Trat. 1 : milho geneticamente modificado MON810: 0,17;

ii) Trat. 2 : milho convencional com aplicação de inseticidas: 0,40 e

iii) Trat. 3 : milho convencional sem a aplicação de inseticidas: 1,98.

Os contrastes para os parâmetros estimados do modelo de médias são apresentados na Tabela 29. Percebe-se que os tratamentos MON810 e Milho Convencional com Inseticida diferem do tratamento Milho Convencional sem Inseticida, mas não diferem entre si.

Tabela 29. Estimativas dos contrastes das médias dos tratamentos para o modelo de Poisson

Contrastes	Estimativas	Erro padrão	Valor de t	Pr > | t |
Trat. 1 e Trat. 2 vs Trat. 3	-5,9348	0,3835	-15,48	$< 0,0001$
Trat. 1 vs Trat. 3	-0,3226	0,3450	-0,94	0,3507

As estimativas das proporções de zeros, seus erros padrões e intervalos de confiança, estimadas pelo modelo de zeros são apresentados na Tabela 30.

Tabela 30. Estimativas das proporções de zeros, erros padrões e intervalos de confiança.

Tratamentos	Estimativas (erros)	Intervalo de confiança
Trat. 1	0,8572 (0,0375)	[0,7834 - 0,9310]
Trat. 2	0,7291 (0,0401)	[0,6500 - 0,8081]
Trat. 3	0,5443 (0,0530)	[0,4399 - 0,6488]

Percebe-se que as proporções de zeros estimadas pelo modelo de zeros, são aproximadamente iguais às proporções de zeros observadas, apresentadas na Tabela 3. Da Tabela 30, observando-se os intervalos de confiança, pode-se dizer que o Tratamento 3 difere dos Tratamentos 1 e 2 sendo que os Tratamentos 1 e 2 não diferem entre si.

A estimativa do componente aleatório, colocado no modelo para captar a variabilidade existente nas medidas repetidas observadas ao longo do tempo, foi aproximadamente zero. Esse fato quer dizer que o modelo Poisson inflacionado de zeros consegue, neste caso, captar toda informação relevante ao experimento, não necessitando de um parâmetro adicional.

5 CONCLUSÕES

As análises realizadas para os experimentos 1 e 2, mostraram que, quando a variável resposta é observada ao longo do tempo, há necessidade de se levar em consideração a correlação existente entre as observações.

Entre as várias formas de se modelar essa correlação, há indícios de que a inclusão de um fator aleatório, para o qual se pressupõe uma distribuição a priori, no preditor linear, faz com que o modelo ajustado seja melhor, captando a variabilidade não considerada quando da inclusão do fator de dispersão constante.

No caso do experimento 1, a inclusão do efeito aleatório e do parâmetro de dispersão constante, conjuntamente, mostram que o modelo se ajusta melhor aos dados, embora ainda haja necessidade de maiores estudos em relação a selecionar o melhor modelo. As técnicas para seleção de modelos lineares generalizados mistos ainda são campos férteis de pesquisa, embora tenha sido possível mostrar que, não foi detectada interação entre métodos de enxertia e porta-enxertos. Dos métodos de enxertia utilizados para o pegamento das mudas de Camu-camu, Fenda Cheia e Fenda Lateral não diferem entre si mas diferem de Inglês Simples e Fenda de Colo, sendo, portanto os melhores métodos de enxertia. Em relação aos porta-enxertos, Camu-camu mostrou-se o melhor deles e Pitangueira, o pior.

No experimento 2, a inclusão do efeito aleatório não se mostrou significativa, indicando que a correlação entre as observações ao longo do tempo é aproximadamente zero. O modelo Poisson inflacionado de zeros mostrou-se eficiente em estimar o número de lagartas e as proporções de zeros. Os resultados mostram que os tratamentos com milho geneticamente modificado MON810 e milho convencional com aplicação de inseticidas diferem do tratamento convencional sem a aplicação de

inseticida, mas não diferem entre si.

Estudos adicionais são necessários para seleção de modelos.

ANEXOS

ANEXO A - Programa SAS para as análises do Experimento 1.

```
/************************************************
*  E1 - Fenda cheia;       C1 - Camu-camu;       *
*  E2 - Fenda lateral;     C2 - Goiabeira;       *
*  E3 - Inglês simples;    C3 - Pitangueira;     *
*  E4 - Fenda de colo;                           *
*  Os totais das pegas vem da soma de 12 plantas  *
************************************************/

Data Pegas;
input Metod $ Porta $ Trat $ Mes $ Rep $ parc $ pegas
brotos;
n=12; prop=pegas/12; Meses=mes;
datalines;
 E4       C2      T11     Dez    r1      1         4        8
 E4       C2      T11     Dez    r2      2         4        7
 E4       C2      T11     Dez    r3      3         4        7
...       ...     ...     ...    ...     ...      ...      ...
 E2       C3      T6      Jul    r2      57        0        0
 E2       C3      T6      Jul    r3      58        0        0
 E2       C3      T6      Jul    r4      59        0        0
 E2       C3      T6      Jul    r5      60        0        0
;
/************************************************
*  Half-normal plot para a distibuição Binomial    *
*  não considerando a superdispersão e nem o tempo *
************************************************/

goptions reset=global gunit=pct cback=white vpos=250
ftitle=swissb ftext=swiss htitle=12 htext=4 rotate=landscape
target=pslepsfc nodisplay;
filename graphout 'c:\assessor\tese\graficos\Halfsd.eps';

goptions device=pslepsfc target= gsfname=graphout gsfmode=replace
display;
%inc 'c:\temp\halfnorm.sas';
%inc 'c:\temp\label.sas';
halfnorm(data=pegas, resp=pegas, trials=n, class = Metod porta
trat Mes rep, model = Metod porta Metod*porta rep metod*rep porta*rep
Metod*porta*rep mes metod*mes porta*mes metod*porta*mes,
dist=bin, mopt=str(scale=pearson) type1 type3);
run; quit; title2; title1;
```

ANEXO A - Continuação do programa SAS para as análises do Experimento 1.

```
/************************************************************
*  Esta é a análise inicial para se saber se a interação é   *
*  significativa ou não. Incluiu-se o parâmetro de superdis- *
*  persão e a a correlação entre meses.                      *
************************************************************/
proc sort data=pegas; by parc; run;

title1 'Esta é a análise inicial para se saber se a interação é
significativa ou não.';

title2 'Incluiu-se o parâmetro de superdispersão e a a
correlação entre meses.';
options ls=120;

Proc genmod data=pegas; class trat metod porta mes rep;
model pegas/n = metod porta metod*porta metod*porta*rep mes metod*mes
porta*mes metod*porta*mes / dist = bin type1 type3 /*scale=pearson*/;
repeated subject=trat / type=ar(1);
/* Lsmeans Metod porta / diff;
contrast 'Tipo de Metodo E1 vs E2' Metod 1 -1 0  0;
contrast 'Tipo de Metodo E1 vs E3' Metod 1  0 -1 0;
contrast 'Tipo de Metodo E1 vs E4' Metod 1  0  0 -1;
contrast 'Tipo de Metodo E2 vs E3' Metod 0  1 -1  0;
contrast 'Tipo de Metodo E2 vs E4' Metod 0 1  0 -1;
contrast 'Tipo de Metodo E3 vs E4' Metod 0  0 1 -1;
contrast 'Metodto C1 vs C2' porta 1 -1  0;
contrast 'Metodto C1 vs E3' porta 1  0 -1;
contrast 'Metodto C2 vs E3' porta 0  1 -1;
*/ output out=saida p=predito*/; run; title2; title1;

/************************************************************
* Para se usar o NLMIXED, os níveis das variáveis tem que *
* ser numéricos. O programa abaixo converte os níveis da  *
* variáveis para numéricos.                                *
************************************************************/

Data Pegas;
set pegas;
Do;
if mes='Dez' then mes=1;
if mes='Jan' then mes=2;
if mes='Fev' then mes=3;
if mes='Mar' then mes=4;
```

ANEXO A - Continuação do programa SAS para as análises do Experimento 1.

```
if mes='Abr' then mes=5;
if mes='Mai' then mes=6;
if mes='Jun' then mes=7;
if mes='Jul' then mes=8;
if Metod='E1' then Metod=1;
if Metod='E2' then Metod=2;
if Metod='E3' then Metod=3;
if Metod='E4' then Metod=4;
if porta='C1' then porta=1;
if porta='C2' then porta=2;
if porta='C3' then porta=3;
if rep='r1' then repet=1;
if rep='r2' then repet=2;
if rep='r3' then repet=3;
if rep='r4' then repet=4;
if rep='r5' then repet=5;
end; output; run;

/*************************************************
*  Análises utilizando o procedimento NLMIXED que   *
*  são reproduções das análises utilizando o GENMOD *
*  ANÁLISES SEQUENCIAIS                             *
*************************************************/

* Devido ao grande número de análises, optou-se por
* colocar apenas o modelo saturado e os modelos.

Title1 'Modelo 10 - ';
Title2 'Modelo Logit-Normal considerando a CONSTANTE, M, P, M*P,
        M*P*Rep, T, MT, PT, MPT, MPTRepet';
%macro lnorm_model(lev_Metod=, lev_porta=, lev_repet=, lev_mes= );
Proc nlmixed data=Pegas maxiter=2000;
eta = beta0
%do i=1 %to %eval(&lev_Metod-1);  + (Metod=&i)*Metod_&i %end;
%do j=1 %to %eval(&lev_porta-1);  + (porta=&j)*porta_&j  %end;
    %do i=1 %to %eval(&lev_Metod-1);
    %do j=1 %to %eval(&lev_porta-1);
        + (Metod=&i & porta=&j)*metpor_&i._&j %end;  %end;
%do k=1 %to %eval(&lev_repet-1); + (repet=&k)*repet_&k %end;
    %do i=1 %to %eval(&lev_Metod-1);
    %do k=1 %to %eval(&lev_repet-1);
        + (Metod=&i & repet=&k)*metrep_&i._&k %end;  %end;
    %do j=1 %to %eval(&lev_porta-1);
```

ANEXO A - Continuação do programa SAS para as análises do Experimento 1.

```
    %do k=1 %to %eval(&lev_repet-1);
        + (porta=&j & repet=&k)*porrep_&j._&k %end;  %end;
    %do i=1 %to %eval(&lev_Metod-1);
    %do j=1 %to %eval(&lev_porta-1);
    %do k=1 %to %eval(&lev_repet-1);
        + (Metod=&i & porta=&j & repet=&k)*metporrep_&i._&j._&k
        %end; %end; %end;
%do l=1 %to %eval(&lev_mes-1);  + (mes=&l)*Mes_&l %end;
    %do i=1 %to %eval(&lev_Metod-1);
    %do l=1 %to %eval(&lev_mes-1);
        + (Metod=&i & mes=&l)*metmes_&i._&l %end; %end;
    %do j=1 %to %eval(&lev_porta-1);
    %do l=1 %to %eval(&lev_mes-1);
        + (porta=&j & mes=&l)*pormes_&j._&l %end; %end;
  %do i=1 %to %eval(&lev_Metod-1);
  %do j=1 %to %eval(&lev_porta-1);
  %do l=1 %to %eval(&lev_mes-1);
        + (Metod=&i & porta=&j & mes=&l)*metpormes_&i._&j._&l
        %end; %end; %end;
  %do l=1 %to %eval(&lev_mes-1);
  %do l=1 %to %eval(&lev_mes-1);
        + (mes=&l & repet=&k)*mesrep_&l._&k %end; %end;
 %do i=1 %to %eval(&lev_Metod-1);
 %do l=1 %to %eval(&lev_mes-1);
 %do k=1 %to %eval(&lev_repet-1);
        + (Metod=&i & mes=&l & repet=&k)*metmesrep_&i._&l._&k
        %end; %end;  %end;
 %do j=1 %to %eval(&lev_porta-1);
 %do l=1 %to %eval(&lev_mes-1);
 %do k=1 %to %eval(&lev_repet-1);
        + (Porta=&j & mes=&l & repet=&k)*pormesrep_&j._&l._&k
        %end; %end;  %end;
  %do i=1 %to %eval(&lev_Metod-1);
  %do j=1 %to %eval(&lev_porta-1);
  %do l=1 %to %eval(&lev_mes-1);
  %do k=1 %to %eval(&lev_repet-1);
        + (Metod=&i & porta=&j & mes=&l &
        repet=&k)*residuob_&i._&j._&l._&k %end;  %end; %end; %end;
          ;
  expeta = exp(eta); p = expeta/(1+expeta);
  model pegas ~ binomial(n, p); run;
%mend; %lnorm_model(lev_Metod=4, lev_porta=3, lev_repet=5, lev_mes=8);
Title1; Title2;
```

ANEXO A - Continuação do programa SAS para as análises do Experimento 1.

```
/*****************************************************************
*  Para se levar em consideração a superdispersão usando o NLMIXED, *
*  multiplica-se a função de probabilidade pelo mesmo, como dado a  *
*  seguir, pela ANÁLISE SEQUENCIAL. phi = 0.6812                    *
*  Com superdispersão, sem correlação e sem efeito aleatório        *
*****************************************************************/

expeta = exp(eta);
p = expeta/(1+expeta);
  prob = comb(n,pegas)*(p**pegas)*(1-p)**(n-pegas);
  prob = prob**(0.6812**-2);
    logver = log(prob);
  model pegas ~ general(logver);
run;
%mend;
%lnorm_model(lev_Metod=4, lev_porta=3, lev_repet=5, lev_mes=8);
Title1; Title2;

/*****************************************************************
*  Inclusão da correlação entre meses -                           *
*  A inclusão do comando random, abaixo, é semelhante a usar o  *
*  comando Repeated no Genmod com Type=CS.                        *
*  Quando utilizamos o comando Random, a entrada dos dados deve *
*  estar ordenada pelo Subject = .                                *
*****************************************************************/

Proc nlmixed data=Pegas;
      eta = beta0 + aleat

* idem aos outros *

          ;
expeta = exp(eta);
p = expeta/(1+expeta);
    model pegas ~ general(logver);
    random aleat ~ normal(0, sigma**2) subject=parc; run;
%mend;
%lnorm_model(lev_Metod=4, lev_porta=3, lev_repet=5, lev_mes=8 );
Title3; Title2; Title1;
```

ANEXO A - Continuação do programa SAS para as análises do Experimento 1.

```
/*******************************************************
*  Análise considerando a superdispersão, a correlação *
*  entre meses e os efeitos aleatórios                 *
*******************************************************/
Title1 'Modelo 10 - ';
Title2 'Modelo Logit-Normal saturado considerando a
        superdispersão e o efeito aleatório';
      eta = beta0 + aleat

* idem aos outros *

          ;
expeta = exp(eta);
p = expeta/(1+expeta); prob =
    comb(n,pegas)*(p**pegas)*(1-p)**(n-pegas); prob =
    prob**(0.6840**-2); logver = log(prob);
    model pegas ~ general(logver);
    random aleat ~ normal(0, sigma**2) subject=parc;
run;
%mend;
%lnorm_model(lev_Metod=4, lev_porta=3);
Title3; Title2; Title1;
```

ANEXO B - Programa SAS para se ajustar o Modelo Poisson Inflacionado de Zeros.

```
Data odnei; input trat $ repet $ tempo $ parc $ y;
 trat1=trat;
 tempo1=tempo;
datalines;
1 1 1  1 0
1 1 2  1 0
1 1 3  1 0
. . .  . .
. . .  . .
3 8 8 24 2
3 8 9 24 4
;
Title1 'Modelo ZIP com efeito aleatório, ajustado para';
Title2 'tratamento no modelo Poisson com efeito aleatório';
%macro zip_model(lev_trat=, lev_trat1=);
Proc nlmixed data=odnei npoints=20 noad noadscale tech=newrap;
%macro zip_model(lev_trat=, lev_trat1=, lev_tempo=, lev_tempo1= );
 Proc nlmixed data=odnei maxiter=500 ;
 *npoints=20 noad noadscale tech=newrap;
    parms var=1; bounds var>0;
    eta1 = beta0 + u
      %do i=1 %to %eval(&lev_trat-1);
                + (trat=&i)*trat_&i  %end;
      %do j=1 %to %eval(&lev_repet-1);
                + (repet=&j)*repet_&j  %end;
          %do i=1 %to %eval(&lev_trat-1);
          %do j=1 %to %eval(&lev_repet-1);
                + (trat=&i & repet=&j)*tratrep_&i._&j %end; %end;
      %do k=1 %to %eval(&lev_tempo-1);
                + (tempo=&k)*tempo_&k  %end;
          %do i=1 %to %eval(&lev_trat-1);
          %do k=1 %to %eval(&lev_tempo-1);
                + (trat=&i & tempo=&k)*trattem_&i._&k  %end;
          %do k=1 %to %eval(&lev_tempo-1);
          %do j=1 %to %eval(&lev_repet-1);
                + (tempo=&k & repet=&j)*temrep_&k._&j  %end;
          %do i=1 %to %eval(&lev_trat-1);
          %do j=1 %to %eval(&lev_repet-1);
          %do k=1 %to %eval(&lev_tempo-1);
              + (trat=&i & repet=&j & tempo=&k)*residuo_&i._&j._&k %end; %end; %end;
             ;
```

ANEXO B - Continuação do programa SAS para se ajustar o Modelo Poisson Inflacionado de Zeros.

```
lambda=exp(eta1);
    eta2 = beta1
           %do i=1 %to %eval(&lev_trat1-1);
               + (trat1=&i)*trat1_&i  %end;
           %do j=1 %to %eval(&lev_tempo-1);
               + (tempo1=&j)*tempo_&j  %end;
           ;
    w = exp(eta2)/(1+exp(eta2));
    if y=0 then prob=w + (1-w)*exp(-lambda);
    if y=0 then logver=log(prob);
           else logver=log(1-w)+y*log(lambda) - lambda - lgamma(y+1);
    model y ~ general(logver);
  random u ~ normal(0, var**2) subject=parc;
  run;
  %mend;
  %zip_model(lev_trat=3, lev_trat1=3, lev_tempo=9, lev_tempo1=9);
title1; title2; title3;
```

REFERÊNCIAS BIBLIOGRÁFICAS

ARTES, R. Extensões da teoria das equações de estimação generalizadas a dados circulares e modelos de dispersão. São Paulo, 1997. 130p. Dissertação (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo.

AZZALINI, A. **Statistical inference.** London: Chapman & Hall, 1996. 341p.

BRESLOW, N.E.; CLAYTON, D.G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, v.88, n.421, p.9-25, 1993.

CARNEIRO, A.A; CARNEIRO, N.P.; CARVALHO, C.H.S.; VASCONCELOS, M.J.V.; PAIVA, E.; LOPES, M.A. **Milho transgênico**: melhoria da qualidade nutricional do grão. www.biotecnologia.com.br/bio/bio15/15_g.htm. (10 jan. 2003)

COLLETT, D. **Modelling binary data.** London: Chapman & Hall, 1991. 369p.

COHEN, A.C. Estimation in mixtures of discrete distributions. In: INTERNATIONAL SYMPOSIUM ON DISCRETE DISTRIBUTIONS, Montreal, 1963. **Proceedings** Montreal, s.ed., 1963. p.373-378.

CORDEIRO, G.M.**Modelos lineares generalizados**. Campinas: UNICAMP, IMECC, 1986. 286p.

DEMÉTRIO, C.G.B. **Modelos lineares generalizados em experimentação agronômica**. Piracicaba: ESALQ, Departamento de Ciências Exatas, 2001. 113p.

DEMPSTER, A.P.; LAIRD, N.M.; RUBIN, D.B. Maximum likelihood from incomplete observations, **Journal of the Royal Statistical Society**, Ser. B, v.39, p.1-38, 1977.

DOBSON, A.J. **An introduction to generalized linear models.** 2.ed. London: Chapman & Hall, 2001. 225p.

FIRTH, D. Generalized linear models. In: HINKLEY, D.V.; REID, N.; SNELL, E.J. (Ed.) **Statistical theory and modelling**. London: Chapman and Hall, 1991. cap.3, p.55-82.

HALL, D.B. Zero-inflated Poisson and binomial regression with random effects: a case study. **Biometrics**, v.56, p.1030-1039. 2000.

HINDE, J.P.; DEMÉTRIO, C.G.B. Overdispersion: models and estimation, **Computation Statistics and Data Analysis**, v.27, p.151-170. 1998a.

HINDE, J.P.; DEMÉTRIO, C.G.B. **Overdispersion**: models and estimation. Caxambu: ABE, 1998. 73p.

JOHNSON, N.L.; KOTZ, S. **Distributions in statistics**: discrete distributions. Boston: Houghton Mifflin, 1969

LAIRD, N.M.; WARE, J.H. Random-effects models for longitudinal data. **Biometrics**, v.38, p.963-974, 1982.

LAMBERT, D. Zero-inflated Poisson regression, with an application to defects in manufacturing **American Statistical Association**, v.34, n.1, p.1-14, 1992.

LIANG, K.Y.; ZEGER, S.L. Longitudinal data analysis using generalized linear models. **Biometrika**, v.73, n.1, p.13-22, 1986.

LITTEL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER, R.D. **SAS system for mixed models**, Cary, 1996. 633p.

MARTINS, E.N, LOPES, P.S., SILVA, M.A., REGAZZI,A.J. **Modelo linear misto**. Viçosa: Imprensa Universitária, 1993. 46p.

McCULLAGH, P.; NELDER, J.A. **Generalized linear models.** 2.ed. London: Chapman & Hall, 1989. 511p.

McGILCHRIST, C.A. Estimation in generalized mixed models, **Journal Royal Statistical Society B** v.56, n.1, p.61-69, 1994.

NELDER, J.A; WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society**, Series A, v.135, p.370-84, 1972.

PARK, T., DAVIS, C. S., LI, N. Alternative gee estimation procedures for discrete longitudinal data. **Computational Statistics & Data Analysis**, v.28, p.243-256, 1998.

PAULA, G.A. **Modelos de regressão com apoio computacional.** Instituto de matemática e estatística, Universidade de São Paulo, 2001. 252p.

PROGRAMA IRAC-BR **Manejo de resistência de Spodoptera frugiperda a inseticidas na cultura do milho.** www.irac-br.org.br/arquivos/Milho.pdf. (10 jan. 2003)

RIDOUT, M.; DEMÉTRIO, C.G.B.; HINDE, J. Models for count data with many zeros. IN: INTERNATIONAL BIOMETRIC CONFERENCE, Cape Town, 1998. **Proceedings**. Cape Town: IBC, 1998. p.1-13

STOKES, M.E.; DAVIS, C.S.; KOCH, G.G **Categorical data analysis using the SAS system.** 2.ed. Cary, NC, 2000. 626p

XAVIER, L.H. Modelos univariado e multivariado para análise de medidas repetidas e verificação da acurácia do modelo univariado por meio de simulação. Piracicaba, 2000. 91p. Dissertação (Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

YAU, K.K.W.; LEE, A.H. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. **Statistics in Medicine**, v.20, p.2907-2920, 2001.

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

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

WILKINSON, G.N.; ROGERS, C.E. Symbolic description of factorial models for analysis of variance. **Applied Statistics**, v.22, p.392-399, 1973.

WOLFINGER, R.D. **Fitting nonlinear mixed models with the new NLMIXED procedure**. www.sas.com/rnd/app/papers/nlmixedsugi.pdf. (14 nov. 2002)

ZEGER, S.; LIANG, K.Y.; SELF, S.G. The analysis of binary longitudinal data with time-independent covariates, **Biometrika**, v.72, p.31-38, 1985.

ZEGER, S.; LIANG, K.Y.; ALBERT, P. Models for longitudinal data: a generalized estimating equation approach, **Biometrics**, v.44, p.1049-1060, 1988.