FRANCISCO REGILSON SOUZA

MODELAGEM DE EXPERIMENTOS PLANEJADOS COM RESPOSTAS DISCRETAS

LAVRAS - MG

2013

FRANCISCO REGILSON SOUZA

MODELAGEM DE EXPERIMENTOS PLANEJADOS COM RESPOSTAS DISCRETAS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Orientador

Dr. Júlio Sílvio de Sousa Bueno Filho

LAVRAS – MG

2013

Ficha Catalográfica Elaborada pela Coordenadoria de Produtos e Serviços da Biblioteca Universitária da UFLA

Souza, Francisco Regilson.
Modelagem de experimentos planejados com respostas discretas / Francisco Regilson Souza. – Lavras : UFLA, 2013.
92 p. : il.

Tese (doutorado) – Universidade Federal de Lavras, 2013.
Orientador: Júlio Sílvio de Souza Bueno Filho.
Bibliografia.

1. Modelos generalizados mistos. 2. Modelo Poisson. 3. Modelos generalizados. 4. Modelo binomial. I. Universidade Federal de Lavras. II. Título.

CDD – 519.72

FRANCISCO REGILSON SOUZA

MODELAGEM DE EXPERIMENTOS PLANEJADOS COM RESPOSTAS DISCRETAS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 18 de julho de 2013.

Dr. Joel Augusto Muniz	UFLA
Dr. Marcio Balestre	UFLA
Dr. Carlos Alberto da Silva Ledo	EMBRAPA
Dr. Marcelo Tavares	UFU - Universidade Federal de Uberlândia

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

LAVRAS – MG

2013

Ao meu pai Gilvan, *in memoriam*

A minha mãe Regina

e aos meus irmãos

OFEREÇO

À minha esposa Cristina,

aos meus filhos Lucas e Lara

DEDICO

AGRADECIMENTOS

A **DEUS**, por tudo...

Ao IF Baiano – Campus Santa Inês, pela oportunidade, em especial ao Diretor Geral Natanaíldo Barbosa Fernandes.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX) pela oportunidade e acolhimento.

À Coordenadoria de aperfeiçoamento de pessoal do Ensino Superior (CAPES), pela concessão da bolsa de estudos.

À FAPEMIG pelo apoio na participação em Congressos.

Aos professores e funcionários do DEX pelo convívio e transmissão de ensinamentos.

Ao professor Dr. Júlio Sílvio Sousa Bueno Filho pelas orientações, sugestões e paciência na elaboração deste trabalho.

A meus pais, por tudo que fizeram pelo meu crescimento moral e intelectual.

À minha família, em especial meu tio e professor Cordeirinho, pelo incentivo.

Ao amigo José Otaviano, In Memoriam, pela amizade, pelo incentivo e colaboração.

A toda turma do DINTER: Norma, Azly, Nelson, Tânia, Ângela, Valter, Otaviano, Marcio, Edmary, Jaime, Jailson, Vasquez, Cleide e Isabel, pela amizade e companheirismo.

Aos professores membros da banca examinadora: Júlio Sílvio Sousa Bueno Filho, Joel Augusto Muniz, Marcio Balestre, Carlos Alberto da Silva Ledo, Marcelo Tavares, pelas sugestões e críticas apresentadas.

RESUMO

Muitos experimentos planejados apresentam valores discretos para a variável resposta. Nestes casos é comum utilizar os modelos lineares generalizados (MLG) para analisar os dados. Quando se tomam m observações $(m > 1)$ em cada unidade experimental (UE), o modelo deveria considerar a variação entre UE. No entanto, sabe-se que o modelo usual (MLG) não o faz, ocasionando diagnósticos como superdispersão e falta de ajuste do modelo. Neste trabalho se apresenta uma justificativa para utilização dos modelos lineares generalizados mistos (MLGM) como opção para esses casos, adicionando um componente aleatório no preditor linear, para capturar as variações entre UE existentes. Isto é feito comparando as duas análises em experimentos simulados com respostas discretas (oriundas ou da distribuição binomial ou da Poisson). Foi considerado um arranjo experimental de UE em um delineamento inteiramente casualizado e simulados experimentos supondo que os efeitos destas UE eram conhecidos. As respostas dos tratamentos foram combinadas às das UE em um modelo linear $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$. A partir daí, simulou-se respostas discretas usando os inversos das ligações canônicas dos modelos binomial e Poisson. Os experimentos resultantes foram analisados das duas formas (MLG e MLGM). As análises foram feitas usando *Software* R 2.14 com 4.000 simulações para cada configuração, com diferentes valores de m. Foram utilizados como parâmetros de comparação: Taxas de erro tipo I, *Deviance* residual, critério de informação de Akaike – *AIC*, critério bayesiano de Schwarz – *BIC* e o acurácia das estimativas de efeitos (em relação aos valores paramétricos conhecidos). Para o caso binomial, a análise MLGM preservou as taxas de erro tipo I o que não ocorreu com o MLG cujas taxas excederam sistematicamente os valores nominais. Para resposta Poisson, a análise MLGM também foi consideravelmente mais rigorosa que o MLG. Os critérios de informação do ajuste foram sempre melhores no MLGM em comparação com o MLG, o mesmo ocorrendo com a acurácia. Em todas as situações analisadas os MLGM mostraram-se mais bem ajustados aos dados dos experimentos do que os MLG e devem ser utilizados em sua substituição.

Palavras-chave: Modelo binomial. Modelo Poisson. Modelos lineares generalizados. Modelos lineares generalizados mistos.

ABSTRACT

Many planned experiments present discrete values for the response variable. In these cases it is common practice to use Generalized Linear Models (GLM) to analyze the data. When *m* observations are taken (*m>1*) from each experimental unit (EU) the model should consider the variations between EU. However, it is known that that is not the case with GLM, casing diagnostics with overdispersion and lack of model adjustment. In this dissertation we present a justification for the use of Generalized Linear Mixed Model (GLMM) as an option for these cases, adding a random effect component to the linear predictor in order to capture the variation between EU. This is done comparing both analyses in simulated experiments with discrete responses (derived from either Binomial or Poisson distributions). We considered an EU experimental arrangement in a completely randomized design and simulated experiments in which we supposed that the effects of the EU were known. The responses from the treatments were to those of the EU in a $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ linear model. From this, discrete responses were simulated using the inverse of the canonical link functions for Binomial and Poisson distributions. The resulting experiments were analyzed using GLMM and GLM. Analyses were conducted using R 2.14 software with 4000 simulations per configuration, with different *m* values. Comparisons were based on: Type I error rates, Residual Deviance, Akaike´s Information Criterion, Bayesian (Schwarz) Information Criterion and accuracy of estimated effects (related to known parametric values). For the Binomial case, the GLMM analysis preserved the type I error rates which did not occur with the GLM, in which the rates systemically exceeded the nominal values. For the Poisson case, GLMM analysis was also considerably more rigorous than GLM analysis. The information criteria of the adjustment were always better in GLMM than in GLM, the same occurring with the accuracy. In all analyzed situations the GLMM was more well-adjusted to the experimental data and should be used to replace GLM.

Keywords: Binomial model. Poisson model. Generalized linear models. Generalized linear mixed models.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Os Modelos Lineares Generalizados (NELDER; WEDDERBURN, 1972) têm sido uma ferramenta muito utilizada na análise de dados em diferentes áreas. No entanto, para dados na forma de proporções e de contagens, frequentemente, as observações obtidas apresentam maior variabilidade do que é possível explicar pelo modelo, cuja forma padrão de análise envolve o uso dos modelos Binomial e Poisson, respectivamente, necessitando às vezes, modelos mais amplos que incorporem essa variabilidade extra, com vistas na adequação e ajuste do modelo.

Dentre as possíveis razões para esse problema está a falta de ajuste do modelo, como consequência da possível falta de termos no preditor linear e o componente aleatório do MLG que apresenta a variabilidade da variável aleatória maior do que a predita pelos modelos Binomial ou Poisson, fenômeno este denominado de *superdispersão* (MCCULLOCH; SEARLE, 2001).

Nos experimentos aleatorizados em que se medem variáveis aleatórias contínuas, é apenas necessário assumir que os efeitos de tratamento e de Unidades Experimentais são aditivos. Como resultado, segue-se que o modelo Gauss - Markov Normal (GMN) pode ser utilizado como boa aproximação da distribuição nula, para inferir sobre diferenças entre médias de tratamentos. Se em um experimento aleatorizado forem tomadas em cada unidade experimental, várias observações da variável de interesse, certamente a distribuição da variável resposta seguirá o modelo GMN, porém os testes F para tratamentos serão realizados com a estimativa de variância entre unidades experimentais, ou seja, o denominador do teste F deve ser a fonte de variação da unidade experimental.

Consideremos uma situação análoga em que se deseja analisar um experimento em um Delineamento Inteiramente Casualizado - DIC com r repetições por tratamento para uma variável com uma só resposta Bernoulli em

cada unidade experimental. Neste caso, o modelo correto é o modelo binomial e o teste de razão de verossimilhanças (ou diferença na deviance) adequado comparará o modelo com uma só média ao modelo com uma média para cada tratamento. O procedimento é análogo a um modelo generalizado de efeitos fixos para tratamento – MLG (NELDER; WEDDERBURN, 1972). Quando se tomam várias (digamos, m) observações Bernoulli na unidade experimental, o teste adequado deveria levar em conta a variação das unidades experimentais. O modelo usual (MLG) não o faz. Em lugar disso, a análise MLG equivale a descrever a variação em um DIC com $r.m$ repetições para cada tratamento. Tal análise pode levar a maiores taxas de erros do tipo I e às excessivas declarações de superdispersão (estimativas de deviance residual com valor muito superior ao esperado pelas distribuições em apreço com as médias declaradas). Nestas situações, uma alternativa que vem ganhando destaque é a utilização dos modelos lineares generalizados mistos - MLGM (MCCULLOCH; SEARLE, 2001).

Para justificar a adoção desse método (MLGM) recorre-se à derivação de um modelo linear, seguindo procedimento apresentado por Hinkelmann e Kempthorne (2008) para variáveis contínuas, com a partição do erro experimental em componentes capazes de quantificar, de forma isolada, os efeitos entre e dentro de unidades experimentais. Nessa abordagem, não se cria um efeito aleatório para o preditor linear, mas sim um componente capaz de captar as variações aleatórias dos próprios dados.

Para Wilk e Kempthorne (1955), o erro experimental é composto por duas componentes: o erro devido às diferenças existentes entre as unidades experimentais e o erro técnico: erros de tratamento devido à inabilidade de repetir um tratamento, às condições da sua aplicação e erros de medida devido a falhas em medições repetidas na mesma situação física não corresponderem exatamente.

É importante ressaltar que a análise mais comum na prática agrícola continua sendo o uso da transformação estabilizadora (arco-seno da raiz quadrada da proporção observada) na busca de uma aproximação de variâncias estáveis ao modelo Gauss - Markov Normal. O uso dos modelos generalizados fixos (MLG) é, no entanto, a segunda opção mais comum de análise dispondo de rotinas prontas e bem mais estabelecidas em pacotes como o R e o SAS (PINHEIRO; BATES, 2000; WOLFINGER; TOBIAS; SALL, 1994).

1.1 Objetivos

O objetivo deste trabalho é comparar o ajuste de modelos generalizados (MLG) ao de modelos generalizados mistos (MLGM) na análise de experimentos planejados com respostas discretas. As comparações se darão em um estudo de simulação de experimentos com respostas que seguem a distribuição binomial e de Poisson.

2 REFERENCIAL TEÓRICO

Esta seção apresenta uma breve revisão da literatura com o objetivo de contextualizar o tema em estudo e trazer resultados e argumentos relacionados ao desenvolvimento deste trabalho.

2.1 Modelo Linear Clássico

O modelo clássico de regressão teve origem nos trabalhos astronômicos de Gauss entre 1809 e 1821. O método de mínimos quadrados surgiu pela primeira vez em trabalhos de Legendre em 1805 e mais tarde por Gauss em 1809 para predizer a trajetória do asteroide Ceres. A associação do modelo a experimentos planejados foi apresentada por Fisher no período de 1920 a 1935. Atualmente, a análise de modelos lineares é uma das técnicas mais importantes da estatística, sendo utilizada nas mais diversas áreas, por profissionais, pesquisadores, etc.

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \qquad (1)$$

Em que $\mathbf{y}$ representa o vetor de dimensões $nx1$, de dados observados; $\mathbf{X}$ de dimensão $n\,x\,p$, é a matriz de delineamento; $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^t$ de dimensão $px1$, é o vetor de parâmetros desconhecidos de efeitos fixos e $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)^t$ é o vetor de dimensão $nx1$de erros aleatórios, que em geral assume-se $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

O modelo linear clássico modela a média de $\mathbf{y}$, usando o vetor de parâmetros de efeitos fixos. Os componentes do vetor $\boldsymbol{\varepsilon}$ são variáveis aleatórias independentes e identicamente distribuídas com média $\mathbf{0}$ e variância σ^2.

Assumindo-se que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, tem-se que $\mathbf{y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$. No modelo clássico de regressão, o estimador de máxima verossimilhança do vetor de parâmetros $\boldsymbol{\beta}$, coincide com o de mínimos quadrados e representam o melhor estimador linear não viesado. Para a obtenção dos estimadores de $\boldsymbol{\beta}$ pelo método da máxima verossimilhança, escreve-se a função de verossimilhança:

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

As estimativas de máxima verossimilhança dos parâmetros são obtidas maximizando a função, tomando-se a sua derivada e igualando a zero. A solução obtida desse sistema fornece estimativas que maximizam $\boldsymbol{\beta}$. As estimativas de $\boldsymbol{\beta}$ pelo método dos mínimos quadrados, consiste na solução do sistema de equações normais (2.3), obtido pela minimização do erro,

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'y}, \qquad (3)$$

e dada por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \text{ e } \text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X'X})^{-1}\sigma^2. \qquad (4)$$

Desde que $\mathbf{X'X}$ seja não singular, caso contrário utiliza-se uma inversa generalizada e as estimativas dos parâmetros são dadas por:

$$\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \quad \text{e} \quad Var(\boldsymbol{\beta}^0) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-'}\sigma^2 . \tag{5}$$

O uso de modelos lineares clássicos é justificado quando os dados são contínuos e o experimento é aleatorizado, pois desta forma há distribuição aproximadamente normal sob a suposição de aditividade (KEMPTHORNE, 1955). Experimentos aleatorizados com dados de contagens ou proporções são comuns em diversas áreas de aplicação (como experimentos agrícolas, por exemplo), mas em geral não satisfazem as pressuposições do modelo (aditividade, normalidade, variância constante, independência). Uma abordagem comum é a transformação dos dados, com o propósito de que esse problema seja contornado, o que nem sempre se consegue. Uma recomendação mais geral é adotar os MLG's.

2.2 Modelos Lineares Generalizados

Mesmo dispondo de uma vasta literatura sobre os MLG's tais como: Dobson (2001), Lee, Nelder e Pawitan (2006), McCullagh e Nelder (1989), McCulloch e Searle (2001), Molenberghs e Verbeke (2005), Nelder e Wedderburn (1972) e Paula (2013), faz-se necessária uma breve apresentação dessa metodologia, haja vista que, essa referência é utilizada como suporte teórico em todo desenvolvimento deste trabalho.

A técnica dos modelos lineares generalizados aplica-se aos casos em que a variável resposta possui distribuição normal, e ainda, estende-se a qualquer distribuição pertencente à família exponencial (NELDER; WEDDERBURN, 1972). A ideia básica é estimar os parâmetros de um modelo linear usando-se o método da máxima verossimilhança baseado na distribuição dos dados.

2.2.1 Família Exponencial

Muitas das distribuições conhecidas podem ser reunidas em uma família particular denominada família exponencial de distribuições. Assim, por exemplo, pertencem a essa família as distribuições normal, binomial, binomial negativa, gama, Poisson, normal inversa, multinomial, beta, logarítmica, dentre outras. A importância da família exponencial de distribuições teve maior destaque, na área dos modelos de regressão, a partir do trabalho pioneiro de Nelder e Wedderburn (1972) que definiram os modelos lineares generalizados.

Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial uniparamétrica se, para uma amostra de n observações, as variáveis respostas $Y_1, Y_2, ..., Y_n$, independentes e provenientes da mesma distribuição de probabilidade, têm sua função densidade de probabilidade (f.d.p.) na forma:

$$f(y_i;\theta_i,\phi,\omega_i) = \exp\left\{\frac{1}{a_i(\phi)}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i;\phi,\omega_i)\right\}. \qquad (6)$$

Em que $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções reais conhecidas, θ_i é a forma canônica do parâmetro de localização, $a_i(\phi) = \frac{\phi}{\omega_i}$, com ω_i, o peso e $\phi > 0$, o parâmetro de dispersão conhecido.

Quando o parâmetro de dispersão ϕ for desconhecido a distribuição considerada pode ou não fazer parte da família exponencial biparamétrica.

A função de verossimilhança para a amostra de n observações é dada por:

$$L = L(\boldsymbol{\theta}, \phi; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi),$$

ou substituindo (6) obtém-se:

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

Tomando-se o logaritmo da função de verossimilhança, denominada log verossimilhança:

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

m que a média e a variância de Y_i podem ser encontradas pelas expressões:

$$E(Y_i) = b'(\theta_i) = \mu_i \quad \text{e} \quad \mathrm{V}(\mathrm{Y}_i) = a_i(\phi) b''(\theta_i) = \frac{\phi}{\omega_i} b''(\theta_i) = \frac{\phi}{\omega_i} V(\mu_i),$$

Em que $b'(\theta_i)$ e $b''(\theta_i)$ são obtidos pelas derivadas parciais da função log verossimilhança:

$$b'(\theta_i) = \frac{\partial l(\boldsymbol{\theta}, \phi, \mathbf{y})}{\partial \theta_i} \quad \text{e} \quad b''(\theta_i) = \frac{\partial^2 l(\boldsymbol{\theta}, \phi, \mathbf{y})}{\partial \theta_i^2}.$$

Sendo encontrado que $b''(\theta_i) = V(\mu_i)$, uma vez que, $b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i}$, a qual é denominada função de variância, por Nelder e Wedderburn (1972).

2.2.2 Estrutura dos Modelos Lineares Generalizados (MLG)

A teoria dos MLG's comporta uma série de métodos estatísticos de análise de dados univariados, que são tratados como casos particulares: modelos de regressão (linear simples, múltipla, não linear); modelos de análise de variância; modelos de análise de covariância; modelo logístico para o estudo de proporções; modelos log – lineares para análise de dados na forma de contagens, etc.

Em McCullagh e Nelder (1989), os modelos lineares generalizados (MLG) estão estruturados em três componentes, para uma amostra de n observações da mesma variável resposta Y, como segue:

a) Componente aleatório – representado pelas variáveis respostas $Y_1, Y_2, ..., Y_n$, independentes e provenientes da mesma distribuição de probabilidade, pertencente à família exponencial na forma (6);
b) Componente sistemático - as variáveis explicativas $\mathbf{x}'_i = \left[x_{i1}, x_{i2}, ..., x_{ip} \right]$, $i = 1, 2, ..., n$ que dão origem a um vetor de preditores lineares:

$$\eta_i = \sum_{j=1}^{p} x_{ij} \beta_j = \mathbf{x}_i^t \boldsymbol{\beta} \text{ ou } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} . \quad (7)$$

Em que $\boldsymbol{\eta}$, chamado de preditor linear, é um vetor de dimensão $nx1$; $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^t$, com $p < n$, é um vetor de p parâmetros desconhecidos, a serem estimados e $\mathbf{X}$, a matriz de delineamento ou covariáveis de dimensão $n \times p$:

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

c) 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, $g(\mu_i) = \eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$.

As funções de ligação podem ser obtidas diretamente da distribuição de probabilidade da família exponencial na forma (6).

2.2.3 Estimação dos parâmetros por máxima verossimilhança

Para a estimação dos parâmetros do modelo pela técnica dos MLG's utiliza-se o método da máxima verossimilhança com base na amostra. Sendo uma amostra aleatória $\mathbf{y} = (y_1, y_2, ..., y_n)^t$ de n observações de uma distribuição pertencente à família exponencial, para que se obtenham as estimativas $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ pelo método da máxima verossimilhança é necessário

determinar os valores de $\boldsymbol{\beta}$ que maximizam a função $l(\boldsymbol{\beta},\phi;\mathbf{y})$. De acordo com McCullagh e Nelder (1989), a função de verossimilhança de $\boldsymbol{\beta}$ é expressa por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i;\theta_i,\phi,\omega_i) = \prod_{i=1}^{n} \exp\left\{\frac{\omega_i}{\phi}\left[y_i\theta_i - b(\theta_i)\right] + c(y_i,\phi,\omega_i)\right\}$$

ou ainda,

$$L(\boldsymbol{\beta}) = \exp\left\{\frac{1}{\phi}\sum_{i=1}^{n}\omega_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i,\phi,\omega_i)\right\} \tag{8}$$

e portanto o logaritmo da função de verossimilhança que, designada *log verossimilhança* é dada por:

$$\ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{\omega_i(y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi,\omega_i) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}). \tag{9}$$

Em que, a contribuição de cada observação y_i para a verossimilhança é expressa por:

$$l_i(\boldsymbol{\beta}) = \frac{\omega_i(y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi,\omega_i).$$

Os estimadores de máxima verossimilhança para $\boldsymbol{\beta}$ são obtidos pela solução do sistema de equações de verossimilhança:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \text{ com } j = 1,...,p. \tag{10}$$

Em geral o sistema acima não apresenta solução analítica, portanto, recorre-se a um processo iterativo para resolução numérica. Nelder e Wedderburn (1972) sugeriram o uso do método dos mínimos quadrados ponderados, baseado no método dos *scores* de Fisher. Como alternativa, por exemplo, seria o método do algoritmo Newton - Raphson que utiliza a matriz Hessiana, no lugar dos *scores* de Fisher.

2.2.4 Seleção de modelos

Por seleção de modelos, esta etapa constitui-se numa das mais importantes da análise de dados (BOZDANGAN, 1987). No processo de seleção de modelos é importante ressaltar que não existem modelos "verdadeiros" ou que corresponda à situação real. Um modelo bem ajustado deve estar próximo da situação real, que causa perdas de informações dentro de limites aceitáveis e que melhor explique o fenômeno em estudo.

McCullagh e Nelder (1989) consideram que um modelo está bem ajustado a um conjunto de observações $\mathbf{y}$, quando puder substituir $\mathbf{y}$ por um conjunto de valores estimados $\hat{\boldsymbol{\mu}}$, para um modelo com um número relativamente pequeno de parâmetros. Certamente os valores estimados não serão iguais aos observados, no entanto, espera-se de um modelo ajustado, que estas discrepâncias sejam pequenas o suficiente, para serem aceitáveis. Considerando a escolha da distribuição da variável resposta e da função de ligação adequados, o objetivo passa a ser a determinação do menor número de termos necessários, para a componente sistemática, de modo que descreva os dados de forma satisfatória. Um modelo com um grande número de variáveis explanatórias pode explicar melhor os dados, porém aumenta significativamente a complexidade na sua interpretação. Ao contrário, um modelo com poucas variáveis explanatórias pode ser de fácil interpretação, mas não se ajusta bem

aos dados. Portanto a seleção de um modelo consiste em buscar um equilíbrio, entre um modelo que explique bem os dados e que não seja de difícil interpretação.

Dentre os critérios constantes na literatura para seleção de modelos, os mais utilizados se baseiam no máximo da função de verossimilhança (LITTEL et al., 2002; WOLFINGER, 1993), com mais destaque para o Teste da Razão de Verossimilhança, o Critério de Informação de Akaike - AIC (AKAIKE, 1974) e o Critério Bayesiano de Schwarz – BIC (SCHWARZ, 1978). O teste da razão de verossimilhança é indicado para testar dois modelos, sendo que os mesmos tenham uma estrutura hierárquica ou aninhada (encaixados) ou ainda um dos modelos seja um caso especial do outro. Sendo l_S o máximo do logaritmo natural da função de verossimilhança para o modelo mais parametrizado (saturado) e l_M para modelo em investigação, à estatística para o teste da razão de verossimilhança, denominada *deviance,* é dada por $D = -2.(l_M - l_S)$.

O Critério de Informação de Akaike (AIC) considera a existência de um modelo "verdadeiro", desconhecido, que descreve a relação entre a variável dependente e as variáveis explicativas, e tenta escolher dentre os modelos em investigação o que apresenta menor divergência com o modelo "verdadeiro". O critério de informação de Akaike (*AIC)* pode ser calculado por $AIC = deviance + 2p$, sendo p o número de parâmetros livres. Desse modo, o modelo em investigação com o menor AIC será considerado mais ajustado.

O Critério Bayesiano de Schwarz (BIC) tem como princípio a existência de um modelo "verdadeiro" que descreve a relação entre a variável dependente e as variáveis explanatórias, dentre os modelos em investigação. Desse modo o critério é determinado pela estatística que maximiza a probabilidade de se identificar o "verdadeiro" modelo dentre os investigados. A estatística para o critério BIC para um determinado modelo é dada por

$BIC = deviance + p.\log n$, sendo n o número de observações e p o número de parâmetros livres. O modelo com menor BIC é considerado o de melhor ajuste.

Os três critérios, embora apresentem princípios e justificativas diferentes, têm em comum a utilização de funções do máximo da função de verossimilhança como medida de ajuste.

2.2.5 Medidas de qualidade do ajuste

O modelo saturado, representado por S, é útil para julgar a qualidade do ajuste de um modelo em investigação, representado por M, através da introdução de uma medida da distância entre os valores ajustados $\hat{\mu}$ com esse modelo e dos correspondentes valores observados $\mathbf{y}$. Essa medida de discrepância entre o modelo saturado e o modelo corrente é baseada na estatística de razão de verossimilhança de Wilks e denominada *deviance* (desvio). O logaritmo da função de verossimilhança (função *log-verossimilhança*) de um modelo linear generalizado é dado por:

$$\ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i \left[y_i q(\mu_i) - b(q(\mu_i)) \right]}{\phi} + c(y_i, \phi, \omega_i) .$$

Em que se substituiu θ_i por $q(\mu_i)$, para fazer salientar, na função *log-verossimilhança*, a relação funcional existente entre θ_i e μ_i.

Para o modelo saturado S - se tem $\hat{\mu}_i = y_i$, ou seja, $\hat{\mu}_i$ é a estimativa de máxima verossimilhança de y_i e o máximo da função *log-verossimilhança* para esse modelo é dado por:

$$l_S(\hat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \frac{\omega_i \left[y_i q(y_i) - b(q(y_i)) \right]}{\phi} + c(y_i, \phi, \omega_i) .$$

Para o modelo em investigação M - se tem $\hat{\mu}_i = \mu_i$, ou seja, $\hat{\mu}_i$ é a estimativa de máxima verossimilhança de μ_i, para $i = 1, 2, ..., n$, e o máximo da função *log-verossimilhança* para esse modelo com m parâmetros é dado por:

$$l_M(\hat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} \frac{\omega_i \left[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i)) \right]}{\phi} + c(y_i, \phi, \omega_i) .$$

Os índices em $\hat{\boldsymbol{\beta}}$ e l indicam o modelo para o qual estão sendo calculados. Se comparar o modelo em investigação M ao modelo saturado S através da estatística de razão de verossimilhanças, obtém-se:

$$D^*(y; \hat{\mu}) = -2(l_M(\hat{\boldsymbol{\beta}}_M) - l_S(\hat{\boldsymbol{\beta}}_S))$$

$$D^*(y; \hat{\mu}) = -2\sum_{i=1}^{n} \frac{\omega_i}{\phi} \left\{ \left[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i)) \right] - \left[y_i q(y_i) - b(q(y_i)) \right] \right\}$$

$$D^*(y; \hat{\mu}) = \frac{D(y; \hat{\mu})}{\phi} . \qquad (11)$$

A expressão $D^*(y; \hat{\mu})$ obtida em (11) denomina-se desvio reduzido e a estatística $D(y; \hat{\mu})$ é denominada desvio para o modelo corrente. Note que o desvio $D(y; \hat{\mu})$ é função apenas dos dados. Decorre da definição que o desvio é sempre maior ou igual a zero, e decresce medida que covariáveis vão sendo adicionadas ao modelo minimal, tornando-se igual a zero para o modelo saturado.

Tanto melhor será o ajuste do MLG aos dados quanto menor for o valor do desvio $D^*(y_i;\hat{\mu}_i)$ (CORDEIRO, 1986; DEMÉTRIO, 1993). Em geral, testa-se o ajuste de um MLG comparando-se o valor de $D^*(y_i;\hat{\mu}_i)$ com os percentis da distribuição χ^2_{n-p}, com $n-p$ graus de liberdade, sendo n o número de observações e p o número de parâmetros livres (posto da matriz do modelo). Assim, quando se tiver $D^*(y_i;\hat{\mu}_i) \leq \chi^2_{n-p;\alpha}$, ou seja, $D^*(y_i;\hat{\mu}_i)$ inferior ao valor crítico da distribuição $\chi^2_{n-p;\alpha}$, pode-se considerar que existem evidências, a um nível aproximado de $100(1-\alpha)$ de confiança, que o modelo proposto está bem ajustado aos dados. Ou ainda, se o valor de $D^*(y_i;\hat{\mu}_i)$ for próximo do valor esperado $(n-p)$ de uma distribuição χ^2_{n-p}, pode ser um indicativo de que o modelo ajustado aos dados é adequado.

Outras estatísticas de testes de ajuste de modelos são: o Critério de Informação de Akaike (*AIC*), que pode ser obtida pela relação $AIC = deviance + 2p$, sendo p, o número de parâmetros livres e o Critério Bayesiano de Schwarz (*BIC*), $BIC = deviance + p.\log n$, com p, o número de parâmetros livres e n, o número de observações. Para as estatísticas AIC e BIC, na comparação de modelos, tanto melhor será quanto menor for o seu valor. Também se pode utilizar como parâmetro a distribuição Qui-quadrado $\chi^2_{n-p;\alpha}$.

2.3 Distribuições Bernoulli

Na realização de um experimento (ensaio) *E*, associado a uma variável aleatória discreta X, cujos resultados possíveis podem ser sucesso (se ocorrer o evento de interesse) ou fracasso (se ocorrer o evento que não interessa), sendo π

a probabilidade de *sucesso* e $1-\pi$ a probabilidade de *fracasso*, em uma única tentativa, diz-se que esta variável aleatória tem distribuição Bernoulli.

Nessas condições, se a variável aleatória discreta X tem distribuição Bernoulli, e a designar como:

X : nº de sucessos em uma única tentativa no experimento. Deve-se ter:

$$X = \begin{cases} 1, \text{ sucesso} \\ 0, \text{ fracasso} \end{cases} \quad \text{sendo} \quad P(X=1)=\pi \;\; e \;\; P(X=0)=1-\pi .$$

Com sua função de probabilidade dada por:

$$P(X=x)=\pi^{x}(1-\pi)^{1-x}. \tag{12}$$

A esperança (média) e a variância da variável X com distribuição de Bernoulli são dadas por:

$$E(X)=\pi \text{ e } Var(X)=\pi(1-\pi)$$

2.4 Distribuição Binomial

A esperança de uma variável aleatória discreta com função de distribuição Bernoulli, é igual à probabilidade π de ocorrer sucesso. Se for executado um experimento tipo Bernoulli, independentemente, m vezes, o número de sucessos pode variar entre 0 e m:

$$\sum_{i=1}^{m} x_i = y .$$

O número total de possíveis sucessos em m repetições do experimento é dado pela combinação de:

$$\binom{m}{y} = \frac{m!}{y!(m-y)!}.$$

Em que:

m = número de observações por UE

y = vetor com os números de sucessos ocorridos em m observações do experimento em cada UE, $y = 0,1,2,...,m$.

Logo se definir a variável aleatória Y tal que:

Y = número de sucessos ocorridos em m repetições independentes do experimento do tipo Bernoulli, têm-se então que Y tem distribuição Binomial com parâmetros m e π, ou seja:

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

Em que π representa a probabilidade de sucesso do experimento.

Pode-se mostrar que a distribuição Binomial pertence à família exponencial na forma canônica.

Seja Y uma variável aleatória tal que mY tem distribuição binomial com parâmetros m e π, $Y \sim B(m,\pi)/m$, com $0 < \pi < 1$, e m conhecido, a função de probabilidade é dada por:

$$f(y;\pi) = \binom{m}{ym} \pi^{ym}(1-\pi)^{m-ym}.$$

Pertence à família exponencial,

$$f(y;\pi) = \exp\left\{ m\left[y\theta - \ln(1+e^{\theta}) \right] + \ln\binom{m}{ym} \right\}. \quad (13)$$

Sendo que $y \in \left\{0, \frac{1}{m}, \frac{2}{m}, ..., 1\right\}$, com $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$ e esperança e variância de Y dados por:

$$E(Y) = b'(\theta) = \pi \text{ e } \text{var}(Y) = b''(\theta)a(\phi) = \frac{\pi(1-\pi)}{m}.$$

Em que o parâmetro canônico é a função logística,

$$\theta = \eta = \ln\left(\frac{\pi}{1-\pi}\right), \text{ com inversa } \pi = \frac{e^{\eta}}{1+e^{\eta}}. \quad (14)$$

Para o modelo binomial com função de ligação canônica e lembrando que $\phi = 1$, tem-se que o desvio (*deviance*) é dado por:

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

Ou ainda,

$$D^*(y_i;\hat{\mu}_i) = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right]. \quad (15)$$

2.5 Distribuição Poisson

Dados na forma de contagens aparecem com muita frequência nas aplicações. São exemplos disso o número de frutos por planta, o número de insetos por planta, o número de acidentes, o número de chamadas telefônicas, o número de elementos numa fila de espera, etc. O modelo de Poisson desempenha um papel fundamental na análise desses tipos de dados. Esse modelo é um membro da família exponencial que tem a particularidade de o valor médio ser igual à variância. Se considerar que as respostas Y_i são independentes e modeladas por uma distribuição de Poisson com distribuição de probabilidade $P(\mu)$ de parâmetro $\mu > 0$, sua função de probabilidade é dada por:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} \quad \text{com} \quad y = 0,1,2,\ldots, \quad (16)$$

pertence à família exponencial da forma (6)

$$f(y;\mu) = \exp\{y\ln(\mu) - \mu - \ln(\mu!)\}, \quad (17)$$

Em que $\theta = \ln(\mu)$ e a média e função de variância dada por: $\mu(\theta) = e^{\theta}$ *e* $V(\mu) = \mu$.

Como $\theta = \ln(\mu)$, a função de ligação canônica é a função logarítmica $\eta = \ln(\mu)$ e um modelo linear generalizado com função de ligação canônica é conhecido por modelo de regressão de Poisson, ou modelo log-linear.

Para o modelo Poisson com função de ligação canônica, tem-se que a função log-verossimilhança é dada por:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln \mu_i - \mu_i - \ln(y_i!),$$

sendo:

$$l_M(\hat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} y_i \ln \hat{\mu}_i - \hat{\mu}_i - \ln(y_i!) \quad \text{e} \quad l_S(\hat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} y_i \ln y_i - y_i - \ln(y_i!).$$

Lembrando que $\phi = 1$ o *deviance* para o modelo de Poisson é:

$$D^*(y;\hat{\mu}) = 2\left[\sum_{i=1}^{n} y_i \ln \frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)\right]. \tag{18}$$

2.6 Superdispersão

Os MLG's têm sido uma ferramenta muito utilizada na análise de dados em diferentes áreas. No entanto, para dados na forma de proporções e de contagens, frequentemente, as observações obtidas apresentam maior variabilidade do que é possível explicar pelo modelo, cuja forma padrão de análise envolve o uso dos modelos Binomial e Poisson, respectivamente,

necessitando às vezes, modelos mais amplos que incorporem essa variabilidade extra, com o objetivo de inferência.

Dentre as possíveis razões para esse problema está a falta de ajuste do modelo, como consequência da possível falta de termos no preditor linear e o componente aleatório do MLG que apresenta a variância maior do que a predita pelos modelos Binomial ou Poisson, denominada de *superdispersão*.

McCulloch e Searle (2001) mostram que, se y_{ij} são independentes e possuem distribuição Bernoulli (p_i), em que p_i assume valores entre (0, 1), então é razoável que p_i possua distribuição Beta (α, β). Sendo assim, tem-se que:

$$E(p_i) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad V(p_i) = \frac{\alpha.\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}.$$

E decorre que as esperanças e variâncias para y_{ij} é dada por:

$$E(y_{ij}) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad V(y_{ij}) = \frac{\alpha.\beta}{(\alpha + \beta)^2}.$$

Com covariância entre duas observações y_{ij} e $y_{ij'}$, para $i \neq j'$ (dois valores da mesma classe ou UE) calculada por:

$$Cov(y_{ij}, y_{ij'}) = V(p_i) = \frac{\alpha.\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}.$$

E a covariância entre duas observações y_{ij} e $y_{i'j'}$, para $i \neq i'$, é igual a zero (entre classes ou UE), mas a correlação intraclasse é diferente de zero e é calculada por:

$$\rho = Cov(y_{ij}, y_{ij'}) = \frac{1}{(\alpha + \beta + 1)}.$$

Se y_{ij} (com $j = 1, ..., m_i$) é a j-ésima observação tomada na i-ésima UE, ela é uma variável aleatória independente com distribuição Bernoulli com média μ, então o total da UE representado por $y_{i.}$ possui distribuição Binomial (m_i, μ), com variância $m_i \mu(1-\mu)$. Então a variância de $y_{i.}$ é calculada por:

$$Var(y_{i.}) = m_i \mu(1-\mu).\left[1 + \frac{m_i - 1}{2}.\rho\right].$$

Como $\alpha > 0$ e $\beta > 0$, condição da distribuição Beta, tem-se que $\rho > 0$, e conclui-se que o desvio é maior que a variância binomial, que representa a *superdispersão*. Em síntese, tem-se:

a) Para dados na forma de proporção $Var(Y_{i.}) > m_i \mu_i (1-\mu_i)$
b) Para dados na forma de contagens $Var(Y_{i.}) > \mu_i$.

Hinde e Demétrio (1998a, 1998b) apresentam modelos que incorporam a superdispersão, através da composição de distribuição, denominados Modelo Beta-Binomial para dados na forma de proporção e Binomial negativo para dados de contagens.

2.7 Modelos Lineares Generalizados Mistos (MLGM)

Para Searle (1987), os modelos lineares nos parâmetros possuem ao menos um efeito aleatório (erro experimental). E são classificados quanto à natureza dos seus efeitos em fixo – quando todos os seus efeitos do modelo são fixos, exceto o erro (resíduo); aleatório – quando todos os seus efeitos do modelo são aleatórios, exceto a constante (média geral); misto – quando coexistem outros efeitos fixo ou aleatório no modelo, além da constante e do erro.

Aspectos sobre a classificação da natureza dos efeitos:

a) Efeitos de um fator são considerados fixos, quando:
   - Os níveis em estudo forem escolhidos pelo pesquisador, de modo que o interesse está centrado nesses níveis;
   - Inferência restrita aos níveis em estudo.

b) Efeitos de um fator são considerados aleatórios, quando:
   - Os níveis em estudo correspondem a uma amostra aleatória de uma população de referência;
   - Os níveis provêm de uma distribuição de probabilidade;
   - A inferência é extrapolada para a população de referência.

A classificação dos efeitos implica na definição do modelo, objetivo e procedimento de análise.

Assim para um modelo fixo, o interesse está na estimação dos próprios efeitos fixos. O modelo linear fixo conforme mostrado em (1) é dado por:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

com $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, logo $\mathbf{y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$, sendo σ^2 a variância do erro.

Em que:

$\mathbf{y}$: vetor de observações, de dimensão $n\,x\,1$;

$\mathbf{X}$: matriz de incidência, de dimensão $n\,x\,p$, que relaciona as observações aos efeitos fixos do modelo;

$\boldsymbol{\beta}$: vetor de efeitos fixos, de dimensão $p\,x\,1$;

$\mathbf{e}$: vetor de erros aleatórios, de dimensão $n\,x\,1$.

O modelo linear misto é uma extensão do modelo linear fixo. O nome modelo misto deriva do fato de que o modelo contém além dos parâmetros de efeitos fixos $\boldsymbol{\beta}$, apresenta também parâmetros de efeitos aleatórios $\mathbf{u}$. Esses modelos surgem com a necessidade de se incluir parâmetros de efeitos aleatórios por diversas razões: dentre outras quando são tomadas medidas repetidas na mesma unidade experimental (LITTELL et al., 2002). Para o modelo linear misto, o interesse está centrado na estimação dos efeitos fixos, estimação dos componentes de variância e covariância dos efeitos aleatórios e predição dos efeitos aleatórios. Para a obtenção do preditor linear dos efeitos aleatórios necessita-se do conhecimento prévio ou estimação dos componentes de variância. Henderson (1953) apresenta os métodos de estimação dos componentes de variância, através de solução explícita, denominado de mínimos quadrados que, para dados balanceados, corresponde ao procedimento usual da ANAVA – Análise de Variância. Esse método tem como vantagem, estimadores não viciados, porém possibilita a obtenção de estimativas negativas para os componentes de variância e ainda existem situações de desbalanceamentos que sua eficiência não é comprovada. Hartley e Rao (1967) apresentam um algoritmo para obter estimadores de Máxima Verossimilhança – ML, para estimação de componentes de variâncias nos modelos mistos. Patterson e

Thompson (1971) introduziram um novo algoritmo que permite operar com a Verossimilhança residual ou restrita (REML). Melhorias computacionais e generalizações desse método foram apresentadas por Harville (1977). A representação atual dos modelos mistos foi apresentada por Henderson (1984):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}, \tag{19}$$

com $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, sendo $\mathbf{R} = \mathbf{I}\sigma_e^2$ e $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, sendo $\mathbf{G} = \mathbf{A}\sigma_a^2$ a matriz de covariâncias dos efeitos aleatórios, com σ_e^2 a variância do erro e σ_a^2 a variância dos efeitos aleatórios dos tratamentos. E ainda $\mathbf{y}$ tem distribuição normal multivariada com,

$$E(\mathbf{y}) = \mathbf{X\beta} \quad \text{e} \quad Var(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R} = \mathbf{V}. \tag{20}$$

Em que:

$\mathbf{y}$: vetor de observações, de dimensão $n\,x\,1$;

$\mathbf{X}$: matriz de incidência, de dimensão $n\,x\,p$, que relaciona as observações aos efeitos fixos do modelo;

$\boldsymbol{\beta}$: vetor de parâmetros de efeitos fixos, de dimensão $p\,x\,1$;

$\mathbf{Z}$: matriz de incidência, de dimensão $n\,x\,q$, que relaciona as observações aos efeitos aleatórios do modelo;

$\mathbf{u}$: vetor de efeitos aleatórios, de dimensão $q\,x\,1$;

$\mathbf{e}$: vetor de erros aleatórios, de dimensão $n\,x\,1$;

Atualmente o procedimento padrão de estimação de componentes de variância para modelos Gaussianos ou aproximadamente Gaussianos é o REML.

Com exceção dos delineamentos balanceados, o método REML requer processos iterativos nas equações de modelo misto. O estimador $\hat{\beta}$ para os efeitos fixos, obtido por este método é o melhor estimador linear não tendencioso – BLUE do modelo, e $\hat{u}$ o melhor preditor linear não tendencioso - BLUP.

Quando $\mathbf{G}$ e $\mathbf{R}$ são conhecidas, esses estimadores podem ser obtidos pelo método dos mínimos quadrados generalizados:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad \text{e variância } V(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{21}$$

que representa o melhor estimador linear não viesado de $\boldsymbol{\beta}$. De maneira análoga temos o melhor preditor linear não viesado(?) de $\mathbf{u}$:

$$\hat{\mathbf{u}} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \text{ e variância} \tag{22}$$

$$V(\hat{\mathbf{u}}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\,. \tag{23}$$

Outra forma de se obter as estimativas de $\boldsymbol{\beta}$ e $\mathbf{u}$ pode ser através da função de verossimilhança dos dados. Se tiver razões para assumir que $\mathbf{u}$ e $\mathbf{e}$ possuem distribuição normal, as estimativas podem ser obtidas (MAR TINS et al., 1993) pela solução do sistema:

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

que apresenta solução análoga às anteriores.

A principal justificativa de se adotar um modelo linear misto é a possibilidade de se fazer a predição de efeitos aleatórios, na presença de efeitos

fixos. Não é o objetivo deste trabalho a predição dos efeitos aleatórios (das UE), mas considerar seus efeitos no modelo, com vistas no melhor ajuste e melhores estimativas dos efeitos fixo. Para a adoção do MLGM, pode-se assumir que os efeitos aleatórios (de UE) e os erros (resíduos) são independentes, identicamente distribuídos com média zero e são não correlacionados, pois nesse caso a distribuição considerada não tem que ser a normal, podendo ser qualquer distribuição pertencente à família exponencial, neste trabalho, as distribuições em estudo são Binomial e Poisson.

Os MLG têm somente um componente aleatório (o erro), mas podem ser estendidos para ter efeitos aleatórios no preditor linear (BRESLOW; CLAYTON, 1993; MCCULLOCH; SEARLE, 2001). A extensão é conhecida como MLGM. Lee e Nelder (1996, 2001) estenderam o trabalho de Breslow e Clayton (1993) apresentando modelos lineares generalizados hierárquicos (ou simplesmente modelos hierárquicos). Esses modelos caracterizam-se por apresentar uma definição hierárquica da função de verossimilhança (ou densidade conjunta), ao contrário de modelos hierárquicos bayesianos, em que a hierarquia está nas prioris. Esses resultados podem ser encontrados em Lee, Nelder e Pawitan (2006).

Logo, uma forma de se modelar a variabilidade entre observações, da mesma unidade experimental, é com a inclusão de variáveis latentes no preditor linear para captar a variação existente nas mesmas unidades experimentais, assumindo-se uma distribuição de probabilidade, em geral, a distribuição normal, para a variável latente. Para essa abordagem, será feita a decomposição do erro experimental (HINKELMANN; KEMPTHORNE, 2008) em componentes capazes de capturar os efeitos aleatórios, de forma isolada, dentro e entre unidades experimentais. O componente aleatório responsável pelo efeito das unidades experimentais será adicionado ao preditor linear, obtendo-se, por

conseguinte um MLGM. A forma geral do MLGM pode ser representada conforme (19):

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}.$$

Em que $\mathbf{y}$, $\mathbf{X}$ e $\boldsymbol{\beta}$, como definido em (19), $\mathbf{Z}$ é uma matriz diagonal de dimensão n, e $\mathbf{u}$ e $\boldsymbol{\varepsilon}$, são vetores de efeitos aleatórios de dimensão n, sendo: $\mathbf{u}$, o vetor de erros (resíduos) dentro das unidades experimentais e $\boldsymbol{\varepsilon}$, o vetor de erros (resíduos) entre unidades experimentais, com $\mathbf{e} = \mathbf{u} + \boldsymbol{\varepsilon}$.

Para basear o presente trabalho será derivado, inicialmente, a partir da aleatorização, o modelo linear de referência, que contenha um componente que represente os efeitos aleatórios entre e dentro das UE, conforme apresentado em Hinkelmann e Kempthorne (2008) para variáveis contínuas. Mais tarde será observado que para as distribuições binomial e Poisson tal modelo corresponde ao MLG e não ao MLGM.

2.8 Modelo derivado linear para variáveis contínuas

Na experimentação agrícola, em geral, os ensaios contêm em suas parcelas mais de um indivíduo (ou mais de uma amostra), que denominamos de tamanho amostral, e o que se faz normalmente é analisar o total da parcela ou a média por parcela. Com esse procedimento a variação (ou variância) entre indivíduos dentro de uma mesma parcela (UE), é negligenciada. Esta variação dá origem à chamada variância dentro das UE's e a variação residual é a variância entre as UE.

Considere que se dispõe de $N = mrt$ Unidades Observacionais – UO, sendo m destas UO pertencentes a cada uma das $n = rt$ UE. Seja k o contador

das UE, $k = 1,...,n = rt$, em que r são as repetições do tratamento i, indicadas pelo contador j, $j = 1,...,r$. Aleatoriza-se os rótulos dos tratamentos às k UE. Esse procedimento estabelece uma distribuição de referência para o erro entre UE caso não haja efeito de tratamento. O processo de aleatorização pode ser expresso matematicamente com a introdução de variáveis aleatórias de delineamento do tipo Bernoulli (0,1), como segue:

$$\delta_{ij}^{k} = \begin{cases} 1, \text{ se a UE k recebeu a repetição } j \text{ do tratamento } i \\ 0, \text{ caso contrário.} \end{cases}$$

Decorrem da definição as seguintes propriedades estatísticas:

$$P(\delta_{ij}^{k} = 1) = \frac{1}{n}$$
$$P(\delta_{ij}^{k} = 1, \delta_{i'j'}^{k'} = 1) = \frac{1}{n}\frac{1}{n-1}(k \neq k')(ij \neq i'j').$$

Por outro lado,

$$P(\delta_{ij}^{k} = 1, \delta_{i'j'}^{k} = 1) = 0 \quad (ij \neq i'j')$$
$$P(\delta_{ij}^{k} = 1, \delta_{ij}^{k'} = 1) = 0 \quad (k \neq k').$$

Isto porque se a UE k recebeu a repetição j do tratamento i, nenhuma outra UE poderá recebê-la e assim como nenhum outro tratamento poderá ser designado àquela UE k.

Como consequência disto tem-se que:

$$\sum_k \delta_{ij}^k = 1, \quad \sum_{ij} \delta_{ij}^k = 1.$$

Pode-se supor que se o tratamento i é aplicado a UE k, a resposta conceitual em alguma escala contínua pode ser representada por T_{ik} e pode ser expressa como função de componentes, a qual tem identidade representada por:

$$T_{ik} = \bar{T}_{..} + (\bar{T}_{i.} - \bar{T}_{..}) + (\bar{T}_{.k} - \bar{T}_{..}) + (T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}). \quad (24)$$

Nota-se que a soma de todos os componentes é igual à resposta T_{ik}. Vale salientar que, para variáveis contínuas, esta decomposição proporciona argumentos geométricos para usar a análise de variância como uma estatística resume o experimento.

2.8.1 Erro Experimental

Na experimentação agrícola, por mais homogêneas que sejam as condições experimentais, quando as UE se tratam de áreas, plantas, aves, animais, etc. não podem ser tratadas, por exemplo, como se fosse uma placa de petri. Sempre existem variações unitárias, próprias de cada UE. Essas variações ficam mais acentuadas, quando se tem repetidas observações nas UE, com uma possível correlação confundida com o efeito da UE.

Wilk e Kempthorne (1955) definem como erro experimental os aspectos de observações repetidas feitas em condições similares para serem idênticas; as causas são classificadas como:

(1) Erros unitários ou erro de unidade devido a falhas nas unidades experimentais diferentes para produzir identicamente sob as mesmas condições e tratamento;
(2) Erros de tratamento devido à inabilidade de repetir um tratamento e às condições da sua aplicação exatamente;
(3) Erros de medidas devido a falhas em medidas repetidas na mesma situação física não corresponderem exatamente.

As causas (2) e (3) são classificadas aqui como erro técnico. Nesta abordagem, os componentes $(\bar{T}_{.k} - \bar{T}_{..})$ e $(T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..})$ da expressão (24) representam uma partição da variação responsável pelo erro experimental:

$$(T_{ik} - \bar{T}_{i.}) = (\bar{T}_{.k} - \bar{T}_{..}) + (T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}), \tag{25}$$

e ainda, por força da ortogonalidade, a soma das variações captadas individualmente por cada um dos dois componentes é igual à variação observada em $(T_{ik} - \bar{T}_{i.})$, sendo que $(\bar{T}_{.k} - \bar{T}_{..})$ quantifica o efeito do erro de unidade ou unitário, que representa o quanto uma UE difere da sua média e

$$(T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}) = (T_{ik} - \bar{T}_{i.}) - (\bar{T}_{.k} - \bar{T}_{..}). \tag{26}$$

O componente em (26) quantifica o efeito do erro técnico e representa o efeito da interação do *i*-ésimo tratamento com a *k*-ésima UE. No modelo linear clássico, por exemplo, admite-se que o efeito de tratamento e a UE são aditivos, assumindo-se que a variação captada por este componente é aditiva ao longo de todas as UE's, nesse caso, o efeito da interação tratamento vezes UE será negligenciada.

2.8.2 Modelo derivado linear para um DIC

Ao final do período experimental uma observação é feita, a qual será denotada por y_{ij} que representa a repetição j do *i*-ésimo tratamento. Usando as variáveis aleatórias de delineamento δ_{ij}^{k}, podemos estabelecer a relação entre o valor observado y_{ij} e a resposta conceitual T_{ik}, como segue:

$$y_{ij} = \sum_{k} \delta_{ij}^{k} T_{ik} \ . \tag{27}$$

Isto significa que se a UE k recebeu a repetição j do tratamento i, então será observado T_{ik}. Se fizer na resposta conceitual T_{ik} expressa em (24):

$$\mu = \bar{T}_{..} \ , \ t_i = (\bar{T}_{i.} - \bar{T}_{..}) \ , \ p_k = (\bar{T}_{.k} - \bar{T}_{..}) \ e \ \upsilon_{ik} = (T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}),$$

obtém-se o modelo $T_{ik} = \mu + t_i + p_k + \upsilon_{ik}$, que substituindo em (27)

$$y_{ij} = \sum_{k} \delta_{ij}^{k} (\mu + t_i + p_k + \upsilon_{ik})$$

e, finalmente,

$$y_{ij} = \mu + t_i + \sum_{k} \delta_{ij}^{k} p_k + \sum_{k} \delta_{ij}^{k} \upsilon_{ik}, \tag{28}$$

constitui o modelo derivado linear.

Se preferir pode-se fazer $u_j = \sum_k \delta_{ij}^k p_k$; se a UE k recebeu a *j*-ésima repetição do *i*-ésimo tratamento, observa-se o desvio p_k e passa a ser a *j*-ésima contribuição relativa ao tratamento i para o componente u_j e de maneira análoga, $\varepsilon_{ij} = \sum_k \delta_{ij}^k v_{ik}$, o modelo deduzido será:

$$y_{ij} = \mu + t_i + u_j + \varepsilon_{ij}. \tag{29}$$

Com o rigoroso processo de aleatorização imposto e aditividade desses componentes com o auxílio das variáveis aleatórias de delineamento δ_{ij}^k, assumimos que $u_j \square IID(0, \sigma_u^2)$ e $\varepsilon_{ij} \square IID(0, \sigma_\varepsilon^2)$ com $e_{ij} = u_j + \varepsilon_{ij}$.

Hinkelmann e Kempthorne (2008) apresentam o procedimento completo e as propriedades decorrentes desse processo para variável contínua.

No processo exposto acima, cada repetição j do tratamento i recebe o resultado da soma das m mensurações independentes de sucessos ou fracassos, que podem ser codificadas como 1 ou 0 respectivamente, que nesse caso possui distribuição Bernoulli. Dentro da UE, a soma das m Bernoulli's independentes representa uma distribuição Binomial de tamanho amostral m. No caso das m observações serem provindas de contagens, as j repetições do tratamento i representam contagens que são números discretos, cada uma delas seguindo a distribuição Poisson, com médias associadas à UE (e não apenas ao tratamento).

3 MATERIAL E MÉTODOS

Nesta seção será apresentada a metodologia utilizada e rotinas para análise MLGM, para um experimento em um Delineamento Inteiramente Casualizado - DIC, com a utilização de um modelo derivado linear com efeito aleatório de UE, para dados com respostas discretas (binomiais e Poisson). Será apresentada uma análise simulada para dados com resposta binomial e Poisson, com efeitos de tratamentos e UE's conhecidos, e comparados à análise usual do MLG. Foram usadas como medidas de ajustes dos dois modelos a *deviance*, o critério de informação de Akaike (AKAIKE, 1974) - *AIC*, o critério bayesiano de Schwarz (SCHWARZ, 1978) – *BIC*, as taxas de erros tipo I, as taxas de declaração de superdispersão, as distribuições das estimativas de efeitos de tratamentos nos dois modelos e o erro quadrático médio das estimativas de efeitos de tratamentos.

3.1 Modelo Linear Generalizado e Generalizado Misto

No MLG o único componente aleatório são as variáveis respostas $Y_1, Y_2, ..., Y_n$, independentes e provenientes da mesma distribuição de probabilidade. As variáveis explicativas, responsáveis pelas informações a serem analisadas e estimação dos parâmetros, dão origem a um vetor de preditores lineares:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = x_i^t\boldsymbol{\beta} \text{ ou } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} . \quad (30)$$

Em que $\mathbf{X}$ representa a matriz do Delineamento Inteiramente Casualizado (DIC) com 5 tratamentos e 5 repetições por tratamento, $\boldsymbol{\beta}$ é o vetor de parâmetros $p\,x1$ e $\boldsymbol{\eta}$ é o vetor denominado de preditor linear, de dimensão n. Para essa abordagem, a análise é feita seguindo o modelo:

$$Y_{ij} \sim \mathrm{B}(\pi_i, m) \text{ para a binomial e } Y_{ij} \sim \mathrm{P}(\mu_i, m) \text{ para a Poisson.}$$

Ou seja, as observações seguem a distribuição binomial com média associada ao tratamento. No modelo generalizado misto acrescenta-se uma componente aleatória u_j que contabiliza as variações existentes entre as UE. Dessa forma o preditor linear passa a ser descrito por:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \text{ com } \mathbf{u} = [u_1, ..., u_j]^t, \text{ e } u_j = \sum_k \delta_{ij}^k p_k, \text{ para } \quad j,k = 1, ..., n = rt \tag{31}$$

Sendo $\mathbf{Z}$ uma matriz identidade de ordem n e $\mathbf{u}$ um vetor de efeitos aleatórios de dimensão n, mesma de $\mathbf{X}\boldsymbol{\beta}$ definida em (30). O modelo obtido por esse processo será:

$$Y_{ij} \sim \mathrm{B}(\pi_{ij}, m) \text{ para a binomial e } Y_{ij} \sim \mathrm{P}(\mu_{ij}, m) \text{ para a Poisson.}$$

Ou seja, as observações seguem a distribuição binomial com média associada tanto ao tratamento quanto à UE.

Por analogia ao que foi apresentado para variáveis contínuas quando se teria a aproximação normal para a análise paramétrica, no caso das distribuições, binomial e Poisson, pode-se pensar que a variável contínua está ligada à resposta

pela função de ligação. As respostas discretas podem ser desta forma associadas à variável conceitual contínua, produzindo um modelo derivado que não corresponde ao modelo generalizado usual (em que as proporções nas UE dependeriam apenas do efeito de tratamento). A aproximação contínua que parece adequada é o modelo generalizado misto, considerando efeitos aleatórios de UE.

3.2 Medidas da Qualidade do Ajuste

A *Deviance* é baseada na estatística de razão de verossimilhanças de Wilks e calculada como a diferença entre a log-verossimilhança do modelo completo e do modelo sob investigação $D(\boldsymbol{\theta};\mathbf{y}) = -2\left[l(\boldsymbol{\theta};\mathbf{y}) - l(\hat{\boldsymbol{\theta}};\mathbf{y}) \right]$. Como a distribuição das *deviances* é aproximadamente qui-quadrado para o MLG, deve-se compara-la com o número de graus de liberdade do resíduo. Nesta notação, um modelo é considerado melhor (mais bem ajustado aos dados) que outro se tiver menor *Deviance*. O critério de informação de Akaike (AIC) é dado por $AIC = deviance + 2p$ e o critério bayesiano de Schwarz (BIC), $BIC = deviance + p.\log n$, em que p é o número de parâmetros e n, o número de observações.

A taxa de erro tipo I representa a probabilidade de um modelo acusar diferenças entre tratamentos quando de fato não existem e será estimada simulando configurações com efeitos nulos de tratamento e calculando a proporção de análises que rejeitam a hipótese nula.

A taxa de declaração de superdispersão será calculada como a porcentagem das vezes em que a variância do componente aleatório do MLG é maior que a prevista pela distribuição do modelo em análise (binomial ou Poisson).

Outro aspecto registrado é o acerto nas estimativas de efeitos de tratamentos, apresentados através das distribuições dos efeitos de tratamentos e do Erro Quadrático Médio – EQM, dado por: $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$, sendo $\hat{\theta}$ o estimador de θ, que pode ser reescrita (MOOD et al., 1974) por $EQM(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$, em que $[E(\hat{\theta}) - \theta]$ representa o viés das estimativas dos efeitos de tratamentos e mede o quanto a estimativa se aproxima do parâmetro.

Serão apresentadas as distribuições das estimativas resultantes dos 4.000 experimentos simulados, apenas para fins de ilustração. Adicionalmente foram calculados os erros quadráticos médios das estimativas pelos dois métodos.

3.3 Material Simulado nas análises MLG e MLGM

Foi simulada uma área experimental composta de 25 UE com efeitos conhecidos, seguindo um gradiente bem definido. O croqui de campo com os valores das médias das UE na escala do preditor é dado pela seguinte Figura 1 a seguir:

-4	-3	-2	-1	0
-3	-2	-1	0	1
-2	-1	0	1	2
-1	0	1	2	3
0	1	2	3	4

Figura 1 Croqui com efeitos de UE

Foram simulados ainda cinco tratamentos com efeitos conhecidos T ={-2, -1, 0, 1, 2} a serem designados às UE's, com 5 repetições por tratamento, de forma aleatória e representa a parte fixa $\mathbf{X\beta = T}$ do preditor linear. O vetor $\mathbf{u}$ contém os efeitos das UE: $\mathbf{u} = (-4,-3,-2,-1,0,-3,-2,-1,0,1,-2,-1,0,1,2,-1,0,1,2,3,0,1,2,3,4)^t$, e representa a parte aleatória do modelo $\mathbf{Zu = u}$. Estes dois grupos de efeitos foram combinados em um modelo linear, ou seja: $\boldsymbol{\eta} = \mathbf{X\beta + Zu}$. A escolha desses valores (tratamentos e UE) para essa combinação de efeitos, se justifica pela necessidade em tornar evidente os efeitos entre e dentro das UE, para a comparação das análises MLG e MLGM.

3.4 Software

As análises estatísticas referentes ao ajuste dos modelos bem como os gráficos ilustrativos, serão realizados com o auxilio do software livre R versão 2.14.2 (R DEVELOPMENT CORE TEAM, 2009) utilizando a função glmer() do pacote lme4.

3.5 Rotinas de Análise

Para a análise dos resultados de cada experimento foram empregados o modelo generalizado fixo com efeitos de tratamentos e o modelo generalizado misto, com efeito de tratamentos e de UE. O MLG foi ajustado com a função básica do R $glm(resp \sim trat,\ family = Binomial)$, para a análise binomial e $glm(y \sim trat,\ family = Poisson)$ para a análise Poisson. Sendo $resp$ uma matriz

em que na primeira coluna vão os sucessos e na segunda os fracassos para a variável observada, e y é a resposta para a variável observada Poisson.

Para a análise MLGM foi construída uma fonte de variação "parcela" correspondente às UE. A análise foi feita usando o pacote *lme4* do R e sua função $glmer(resp \sim trat + (1|parcela),\ family = Binomial)$, para o modelo binomial e $glmer(y \sim trat + (1|parcela),\ family = Poisson)$, para a Poisson. Nesta especificação, será ajustada uma componente da variância para a distribuição do preditor linear entre parcelas.

3.6 Análises MLG e MLGM para resposta Binomial

Para a realização deste trabalho, foi ajustada uma função do R para analisar um experimento planejado em um DIC com resposta binomial para cada UE, considerando que, a soma de m Unidades Observacionais - UO independentes do tipo Bernoulli (0,1) na UE, equivale a uma resposta binomial com tamanho amostral m.

Num experimento com resposta binomial em que cada tratamento possui r repetições e cada UE recebe m respostas Bernoulli independentes, a análise usual do MLG trata como $r.m$ respostas Bernoulli independentes para cada tratamento, que corresponde a uma análise de dados com uma distribuição binomial com tamanho amostral $r.m$, mas, o que ocorre é r repetições de uma resposta binomial para cada tratamento, ou seja, cada UE recebe uma distribuição binomial com proporção distinta (para as suas m respostas Bernoulli independentes). Nesse caso existem diferentes distribuições entre e dentro de UE.

Quanto às variações existentes entre UO da mesma UE, foi derivado um modelo linear associado a um experimento planejado em um DIC, que apresente

um componente que seja capaz de quantificar de forma isolada as variações das UE's.

Este método segue todo processo de derivação do modelo linear proposto por Hinkelmann e Kempthorne (2008) para variáveis contínuas. O componente com as variações aleatórias existentes e contabilizadas de cada UE foram adicionadas ao preditor linear, produzindo um novo modelo, que doravante será designado de MLGM.

A função de ligação entre o preditor linear e a probabilidade de sucesso da distribuição da variável resposta a ser simulada na parcela é a logística, para a qual, a proporção para cada combinação de tratamento e UE pode ser encontrada pela função inversa:

$$\boldsymbol{\pi} = \frac{\exp(\mathbf{X\beta} + \mathbf{Zu})}{1 + \exp(\mathbf{X\beta} + \mathbf{Zu})}. \tag{32}$$

Tais valores de $\boldsymbol{\pi}$ representam probabilidades de sucesso que foram utilizadas para a simulação de respostas binomiais para cada UE, com diferentes tamanhos amostrais (m).

Desta forma, para cada UE uma proporção diferente surge associada às m respostas Bernoulli's independentes.

Foram simuladas configurações com m variando entre os seguintes valores 1, 5, 10,15 e 20. Foram efetuadas 4.000 simulações para cada configuração.

3.7 Análises MLG e MLGM para resposta Poisson

Para a realização desse método e para analisar um experimento planejado em um DIC para dados com resposta Poisson, em que cada tratamento

i possui r repetições e cada UE possui m UO, foi adotado procedimento análogo ao visto no caso da distribuição binomial.

Se considerar que as respostas Y_i são independentes e modeladas por uma distribuição de Poisson com distribuição de probabilidade $P(\mu)$ de parâmetro $\mu > 0$, sua função de ligação canônica é a logarítmica dada por:

$$\boldsymbol{\eta} = \boldsymbol{\theta} = \ln(\boldsymbol{\mu}), \text{ com inversa dada por: } \boldsymbol{\mu} = \exp(\boldsymbol{\eta}). \qquad (33)$$

A função de ligação entre o preditor linear e a média da distribuição da variável resposta a ser simulada na parcela, para cada combinação de tratamento e UE, é a logarítmica, para a qual, a média pode ser encontrada pela função inversa:

$$\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}). \qquad (34)$$

Desta forma, para cada UE uma média diferente surge associada às m respostas Poisson independentes. Tais valores de $\boldsymbol{\mu}$ representam as médias que foram utilizadas para a simulação das respostas Poisson para cada UE, com diferentes tamanhos amostrais (m).

Foram simuladas configurações com m variando entre os seguintes valores: $m = \{1, 2, 3\}$.

4 RESULTADOS

Nessa seção serão apresentados os resultados para a distribuição binomial e Poisson em duas subseções separadas. Em cada uma das subseções serão apresentados a comparação dos resultados das análises MLG e MLGM, através das distribuições das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike - *AIC*, o critério bayesiano de Schwarz – *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos. Esses são os parâmetros utilizados para o julgamento do ajuste e adequabilidade dos modelos.

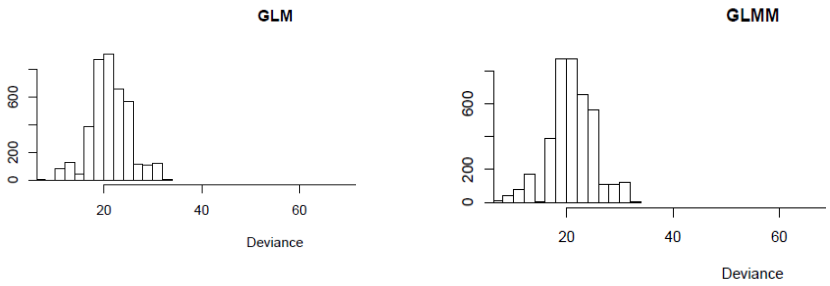
4.1 Resultados para a distribuição Binomial

Nessa subseção os resultados das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike - *AIC*, o critério bayesiano de Schwarz – *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos nas duas análises MLG e MLGM referem-se a dados com respostas binomial.

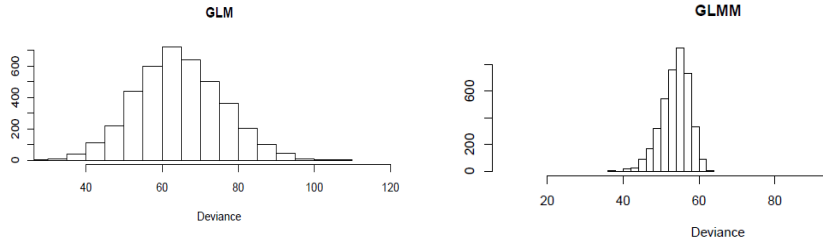
4.1.1 Distribuições das estimativas das *deviances* residuais

Na Figura 2, apresenta-se a comparação das distribuições das estimativas das *deviances* residuais dos dois modelos em análise MLG e MLGM para os tamanhos amostrais $m = 1$, $m = 5$, $m = 10$, $m = 15$ e $m = 20$.

Nota-se que não há diferença nas distribuições das estimativas das *deviances* residuais entre o MLG e o MLGM quando m=1, o que confirma a expressão da variância do componente aleatório do MLG $Var(y_{i.}) = m_i \mu(1-\mu).\left[1+\frac{m_i-1}{2}.\rho\right]$, que nesse caso reduz-se a variância do modelo $Var(y_{i.}) = \mu(1-\mu)$, uma vez, que as repetidas observações dentro das UE's justificam parte da variação das UE's e na ausência de repetições, a análise pelos MLG e MLGM são equivalentes. Obviamente, neste caso, a pressuposição de distribuições Bernoullis independentes entre as UE's é válida (ou seja, há uma só distribuição binomial por tratamento). Pode-se notar também (pela expressão acima) que na medida em que os valores de m(UO) crescem ($m=5$, $m=10$, $m=15$ e $m=20$), a variância do componente aleatório cresce proporcionalmente, tornando-se maior que a variância do modelo, ocasionando aumento na *deviance*. Para os valores de $m=5$, $m=10$, $m=15$ e $m=20$, nota-se que os valores das estimativas das *deviances* residuais para o MLGM são sempre menores que o do MLG. Além de menor média, as estimativas são mais concentradas, o que indica a adequação do modelo.



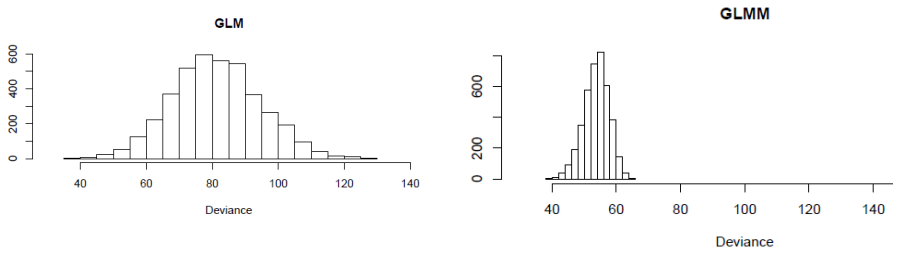
Deviance para $m=1$ para o MLG (à esquerda) e MLGM (à direita).



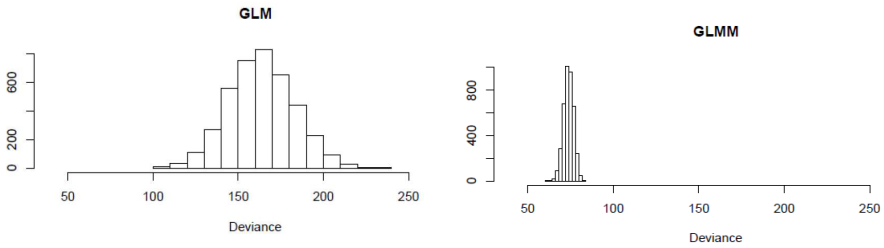
Deviance para $m = 5$ para o MLG (à esquerda) e MLGM (à direita).

Figura 2 Distribuições das estimativas das *deviances* residuais dos modelos MLG e MLGM com os tamanhos amostrais $m = 1$, $m = 5$, $m = 10$, $m = 15$ e $m = 20$

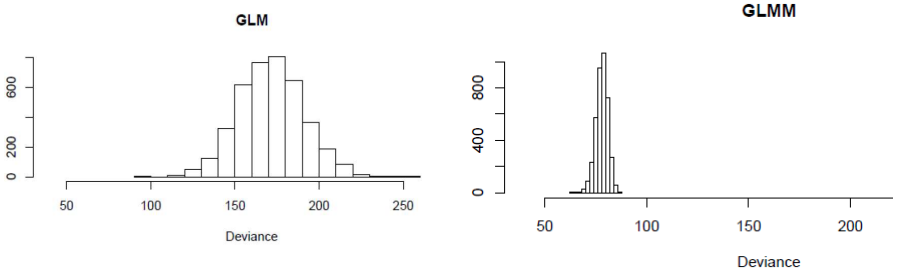
(...continua...)



Deviance para $m = 10$ para o MLG (à esquerda) e MLGM (à direita).



Deviance para $m = 15$ para o MLG (à esquerda) e MLGM (à direita).



Deviance para $m = 20$ para o MLG (à esquerda) e MLGM (à direita).

Figura 2 Distribuições das estimativas das *deviances* residuais dos modelos MLG e MLGM com os tamanhos amostrais $m = 1$, $m = 5$, $m = 10$, $m = 15$ e $m = 20$

(conclusão)

Foi encontrado resultado semelhante para $m = 25$ e apresentado no apêndice A. Efetivamente o aumento do tamanho amostral (valores de m) implica também em aumento na diferença entre as estimativas das *deviances* residuais nos modelos de análise MLG e MLGM. Isto porque, na análise MLG, a *deviance* está relacionada proporcionalmente a m para a distribuição binomial.

4.1.2 Taxas de erro tipo I

A Tabela 1 mostra as taxas de erro tipo I para os níveis de significância $\alpha = 1\%$ e $\alpha = 5\%$ para os dois modelos em análise MLG e MLGM com diferentes tamanhos amostrais $m = 1$, $m = 5$, $m = 10$ *e* $m = 15$. A taxa de erro tipo I é um importante instrumento que auxilia na verificação da adequação ou não de um modelo, uma vez que, as taxas de erro tipo I representam em termos percentuais, a probabilidade da análise acusar diferenças entre tratamentos que de fato não existem. É fácil perceber que o aumento no tamanho amostral m implica no crescimento das taxas de erro tipo I para o MLG, ao contrário acontece no MLGM.

Tabela 1 Taxas de erro tipo I nas análises MLG e MLGM, para o modelo binomial com diferentes tamanhos amostrais $m=1,\ m=5,\ m=10\ e\ \ m=15$

	m=1		m=5		m=10		m=15	
α	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**
MLG	4,975%	0,375%	59,750%	29,475%	95,275%	82,675%	98,725%	91,825%
MLGM	4,375%	0,975%	4,425%	0,200%	0,775%	0,000%	0,025%	0,000%

4.1.3 Taxas de declaração de Superdispersão do MLG

Na Tabela 2 são apresentadas as taxas de declaração de superdispersão para o modelo MLG com diferentes tamanhos amostrais como $m=1,\ m=5,\ m=10\ e\ \ m=15$. A taxa de declaração de superdispersão, mostra em termos percentuais, o quanto a variância do componente aleatório do MLG foi declarada maior que a do modelo em análise (binomial). No caso, trata-se de um estudo de simulação e podemos atribuir às altas taxas de erro tipo I observadas para o MLG à falta de um componente no preditor linear que incorpore ao modelo as variações entre UE.

Tabela 2 Taxas de declaração de superdispersão, na análise MLG, para o modelo binomial com diferentes tamanhos amostrais $m=1,\ m=5,\ m=10\ e\ \ m=15$

	m=1		m=5		m=10		m=15	
α	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**
MLG	39,3%	29,0%	100%	99,98%	100%	100%	100%	100%

4.1.4 Medidas de ajuste AIC e BIC

Na Tabela 3, são mostradas as medidas de ajustes AIC e BIC para os dois modelos em análise MLG e MLGM com diferentes tamanhos amostrais como $m=1,\ m=5,\ m=10\ e\ m=15$.

Exceto para $m=1$, que nesse caso as análises são equivalentes, o MLGM apresentou menores valores para o AIC e BIC que o MLG.

Tabela 3 Medidas de ajustes *AIC* e *BIC*, nas análises MLG e MLGM, para o modelo binomial com diferentes tamanhos amostrais $m=1,\ m=5,\ m=10\ e\ m=15$

	m=1		**m=5**		**m=10**		**m=15**	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
MLG	38,264	43,140	82,284	93,597	134,061	148,147	193,589	209,297
MLGM	38,236	43,112	63,393	74,706	77,272	91,357	86,823	102,530

Na comparação entre dois modelos de análise, considera-se mais ajustado o que apresentar menor AIC e BIC, segundo estes critérios, o MLGM demonstra melhor ajuste que o MLG, com menor AIC e BIC, confirmando resultados anteriores.

4.1.5 Distribuição das estimativas dos efeitos de tratamentos

Na Figura 3 são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação {-2, -1, 0, 1, 2}, nessa ordem, nos experimentos com $m=5$. O viés das estimativas dos efeitos de tratamentos, dado por $[E(\hat{\theta})-\theta]$, sendo o valor de $\hat{\theta}$ a estimativa do parâmetro θ, é uma medida da aproximação da estimativa do parâmetro e quanto menor é o viés,

mais próxima é a estimativa do parâmetro, quando se tem $E(\hat{\theta}) = \theta$, o viés é nulo.

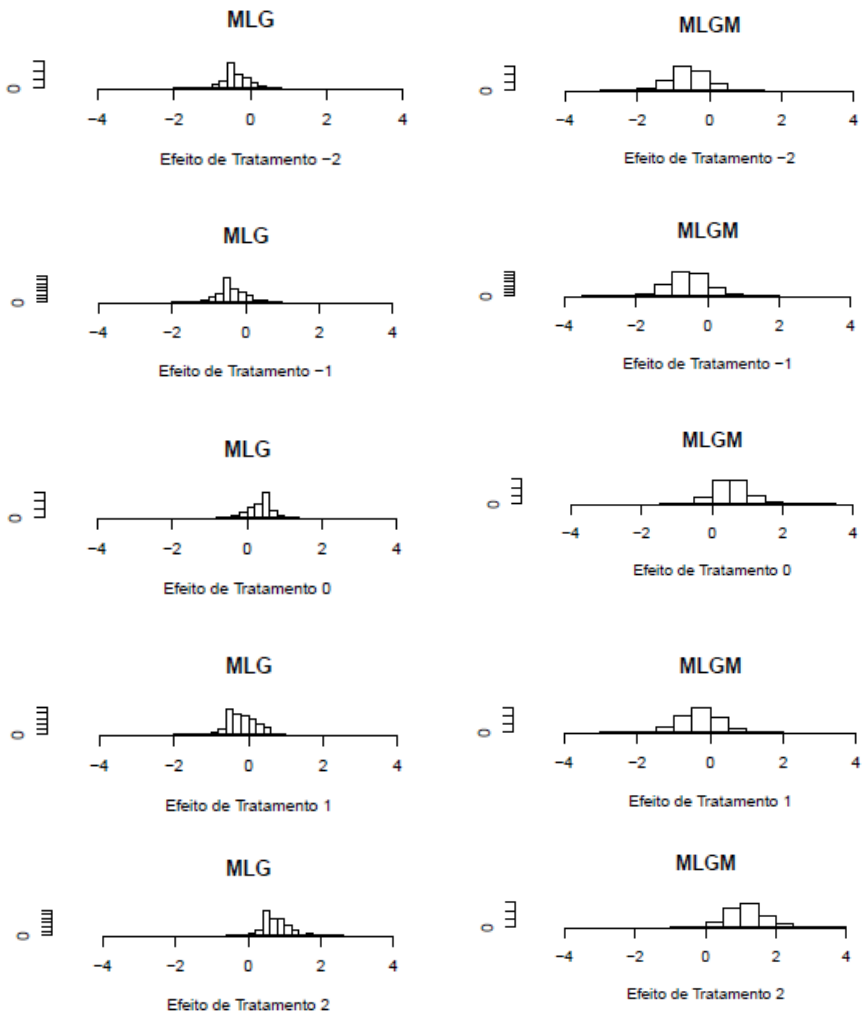


Figura 3 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m = 5$

Pode-se notar que o viés das estimativas do efeito de tratamento, para $m = 5$, é menor no MLGM que no MLG, indicando que o MLGM se aproxima mais do verdadeiro efeito de tratamento e consequentemente apresenta melhor

ajuste. Esta observação pode ser comprovada analiticamente e graficamente pelo Erro Quadrático Médio – EQM, nos resultados que seguem.

Na Figura 4, são apresentadas as distribuições das estimativas de efeitos dos tratamentos que foram utilizados na simulação {-2, -1, 0, 1, 2}, nos experimentos com tamanho amostral $m = 10$. Nota-se que o viés $[E(\hat{\theta}) - \theta]$ das estimativas do efeito de tratamento é menor no MLGM do que no MLG, denotando que o MLGM é mais ajustado que o MLG, confirmando o resultado obtido para $m = 5$.

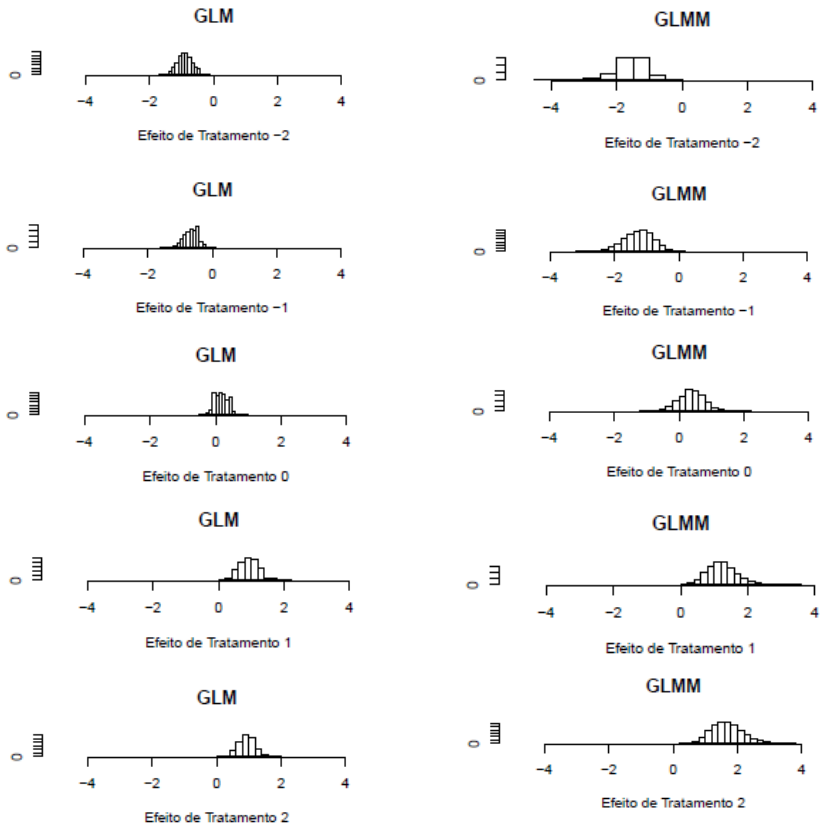


Figura 4 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m = 10$

Resultado semelhante ao anterior foi encontrado na Figura 5, em que são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação {-2, -1, 0, 1, 2}, nos experimentos com tamanho amostral $m=15$. Nota-se que o viés $[E(\hat{\theta})-\theta]$ das estimativas dos efeitos de tratamentos é menor no MLGM do que no MLG, indicando que o MLGM é mais bem ajustado que o MLG.

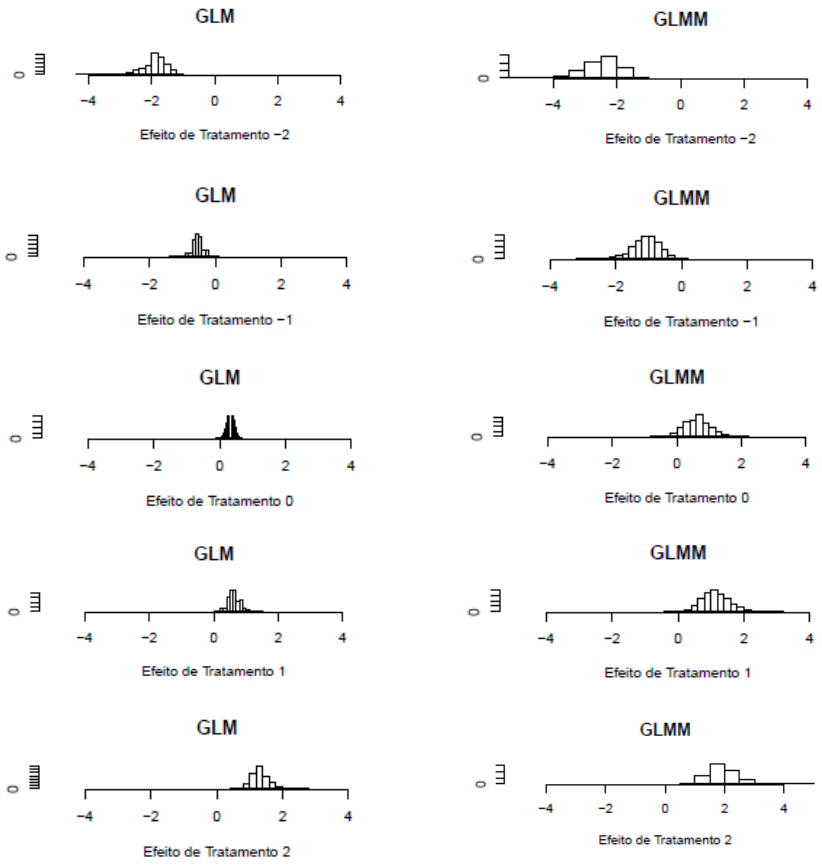


Figura 5 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m=15$

Resultado semelhante foi encontrado para o tamanho amostral $m = 20$ apresentado no apêndice A.

4.1.6 Erro Quadrático Médio das estimativas de efeitos de tratamentos

Na Tabela 4 são apresentadas as médias e desvios dos Erros Quadráticos Médios – EQM das estimativas do efeito de tratamento para o MLG e MLGM com tamanho amostral $m = 1$, $m = 5$, $m = 10$ e $m = 15$, expresso por $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$, sendo o valor de $\hat{\theta}$ a estimativa do parâmetro θ, que representa o desvio quadrático da estimativa em relação ao parâmetro. Desse modo, quanto mais próxima a estimativa for do parâmetro, menor é o EQM e consequentemente, o modelo que apresenta menor EQM é mais bem ajustado. A expressão $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ pode ser reescrita como $EQM(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$, e quando se tem $E(\hat{\theta}) = \theta$ a expressão resume-se a $EQM(\hat{\theta}) = Var(\hat{\theta})$, ou seja, quando o viés do estimador é nulo o erro quadrático é igual à variância do estimador.

Tabela 4 Média e Desvio padrão dos EQM de tratamentos no MLG e MLGM e eficiência relativa das estimativas (EQM-MLG/EQM-MLGM), para $m = 1, m = 5$, $m = 10$ e $m = 15$

	m=1		**m=5**		**m=10**		**m=15**	
	Média	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**
MLG	20,14	32,23	5,51	1,35	5,26	0,95	5,33	0,83
MLGM	20,64	33,49	4,35	1,29	4,21	0,86	4,42	0,68
MLG/MLGM	0,98	0,96	0,27	0,04	0,25	0,10	0,21	0,22

Exceto para $m = 1$, que os modelos MLG e MLGM são equivalentes, o MLGM apresenta menor EQM que o MLG, para os valores de m ($m = 5$, $m = 10$ e $m = 15$), indicando, portanto, que é mais bem ajustado.

4.1.7 Distribuição do Erro Quadrático Médio das estimativas de efeitos de tratamentos

Na Figura 6, são apresentadas as distribuições do Erro Quadrático Médio – EQM das estimativas dos efeitos de tratamentos para os dois modelos de análise MLG e MLGM com os diferentes tamanhos amostrais $m = 5$, $m = 10$, $m = 15$ e $m = 20$. O MLGM está apresentado no gráfico na cor vermelha e o MLG na cor preta. Pode-se notar que, em média, o EQM das estimativas de efeitos de tratamentos é menor para o MLGM que no MLG, para todos os tamanhos amostrais considerados. Conforme esse critério, o MLGM aproxima mais do valor paramétrico que o MLG.

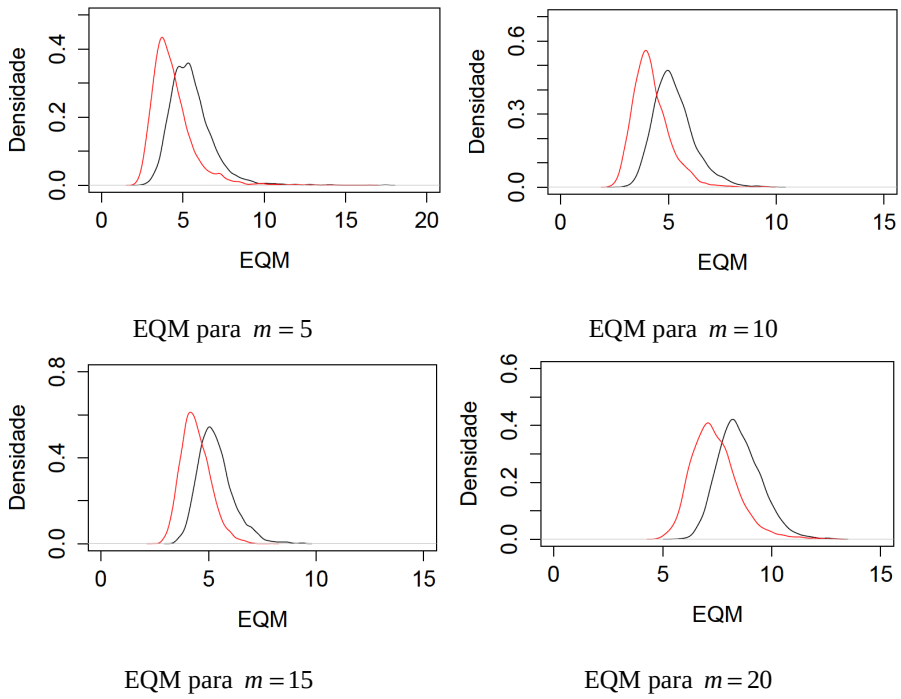


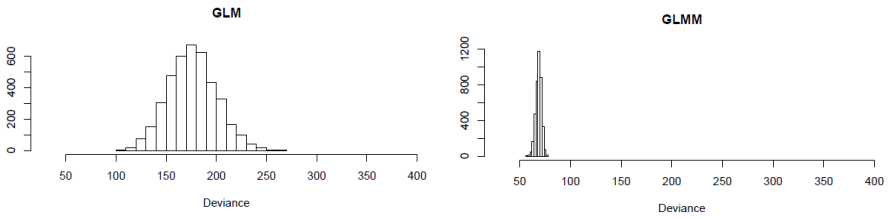
Figura 6 Distribuições dos erros quadráticos médios - EQM no MLG (preto) e MLGM (vermelho) para $m = 5$, $m = 10$, $m = 15$, $m = 20$ *e* $m = 25$

4.2 Resultados para a distribuição Poisson

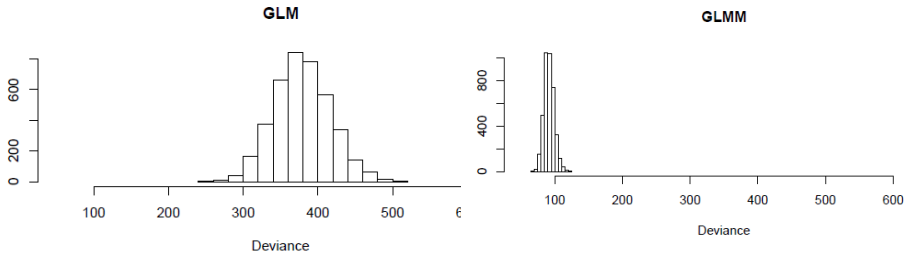
Nessa subseção os resultados das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike - *AIC*, o critério bayesiano de Schwarz – *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos nas duas análises MLG e MLGM referem-se a dados com respostas Poisson.

4.2.1 Distribuições das estimativas das deviances residuais

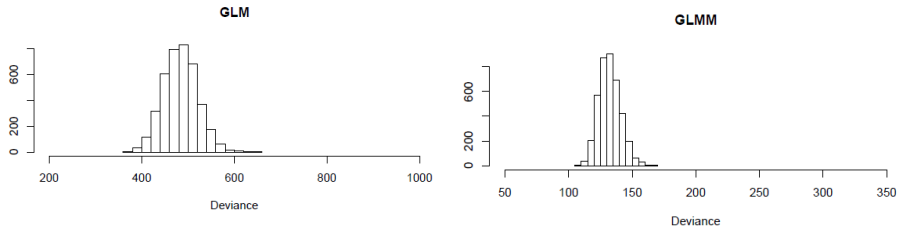
Na Figura 7 são apresentadas as distribuições das estimativas das *deviances* residuais dos dois modelos de análise, para os tamanhos amostrais $m=1$, $m=2$ e $m=3$. Nota-se que há diferença significativa entre as médias das estimativas das *deviances* residuais para dos modelos MLG e o MLGM, sendo que o MLGM apresenta menor *deviance* residual que o MLG, para todos os tamanhos amostrais analisados ($m=1$, $m=2$ e $m=3$).



Deviance residual para $m=1$ para o MLG (à esquerda) e MLGM (à direita).



Deviance para $m=2$ para o MLG (à esquerda) e MLGM (à direita).



Deviance para $m=3$ para o MLG (à esquerda) e MLGM (à direita).

Figura 7 Distribuições das estimativas das *deviances* residuais dos modelos MLG e MLGM com os tamanhos amostrais $m=1$, $m=2$ e $m=3$

É preciso ressaltar que a escala teve que ser reduzida para apresentar os histogramas do MLGM. Com base nisto, conclui-se que para todos os valores de m, inclusive $m=1$, as *deviances* do MLGM são menores e mais concentradas, indicando sua melhor adequação. O aumento de m implica também em aumento na diferença entre as médias das estimativas das *deviances* residuais dos dois modelos. Mesmo para $m=1$, o MLG é inadequado. Isto porque é mais correto modelar uma distribuição de Poisson com média diferente para cada UE. Deve-se notar, no entanto, que para valores de m muito maiores (talvez $m>15$, por exemplo) há um incremento rápido na aproximação pela análise normal (usando, por exemplo, uma transformação estabilizadora da variância). Isto deve ocorrer por um efeito de limite central ao se tomar médias de contagens (o mesmo efeito, embora presente na Binomial, é mais lento por serem médias de variáveis com mesma distribuição Bernoulli).

4.2.2 Taxas de erro tipo I

Na Tabela 5, são apresentadas as taxas de erro tipo I para os dois modelos em análise MLG e MLGM para diferentes tamanhos amostrais como $m=1$, $m=2$ e $m=3$ com níveis de significância $\alpha=1\%$ e $\alpha=5\%$. Nota-se que o MLG apresenta taxa de erro tipo I de 100% para todos os tamanhos amostrais ($m=1$, $m=2$ e $m=3$) que indica, com probabilidade de 100%, de acusar diferenças entre tratamentos que de fato não existem. O contrário acontece com o MLGM, com taxas bem menores, que reflete um modelo melhor ajustado, com resultados mais confiáveis para essa situação.

Tabela 5 Taxas de erro tipo I para as análises MLG e MLGM, para o modelo Poisson com diferentes tamanhos amostrais $m=1,\ m=2\ e\ \ m=3$

	m=1		m=2		m=3	
α	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**
MLG	100%	100%	100%	100%	100%	100%
MLGM	1,725%	0,150%	0,150%	0,025%	0,00%	0,00%

4.2.3 Taxas de declaração de Superdispersão do MLG

Na Tabela 6, são apresentadas as taxas de declaração de superdispersão para o modelo MLG com diferentes tamanhos amostrais $m=1,\ m=2$ e $m=3$, com níveis de significância $\alpha=1\%$ e $\alpha=5\%$. Analogamente ao que acontece na análise binomial, o MLGM é modelado para não apresentar superdispersão, a dispersão declarada para a U.E na análise do MLGM é o parâmetro de canônico do modelo Poisson $\sigma_u^2=1$. As taxas de declaração de superdispersão representa, em termos percentuais, o quanto a variância do componente aleatório do modelo MLG é maior que a do modelo Poisson.

Tabela 6 Taxas de declaração de superdispersão, nas análises MLG, para o modelo Poisson com diferentes tamanhos amostrais $m=1,\ m=2\ e\ \ m=3$

	m=1		m=2		m=3	
α	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**
MLG	100%	100%	100%	100%	100%	100%

O MLG apresenta taxas de declaração de superdispersão de 100% para todos os tamanhos amostrais ($m=1,\ m=2$ e $m=3$) e todos os níveis de significância $\alpha=1\%$ e $\alpha=5\%$, que indica que a variância do componente aleatório do MLG é maior que a do modelo Poisson, necessitando de um

componente no preditor linear para acomodar essa variabilidade extra e reflete a falta de ajuste do MLG para esse tipo de análise.

4.2.4 Medidas de ajuste AIC e BIC

Na Tabela 7, são apresentadas as medidas de ajustes AIC e BIC para os dois modelos em análise MLG e MLGM para diferentes tamanhos amostrais como $m=1, m=2$ e $m=3$. Na comparação de dois modelos de análise, será considerado mais ajustado o que apresentar menor valor para o AIC e BIC. Para todos os tamanhos amostrais ($m=1, m=2$ e $m=3$) o MLGM apresenta menor AIC e BIC que o MLG, portanto se ajusta melhor aos dados para essa situação.

Tabela 7 Medidas de ajustes *AIC* e *BIC*, nas análises MLG e MLGM, para o modelo Poisson com diferentes tamanhos amostrais $m=1, m=2$ e $m=3$

	m=1		**m=2**		**m=3**	
	AIC	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
MLG	273,862	278,738	529,557	537,205	849,458	858,728
MLGM	75,214	80,090	111,930	119,577	142,522	151,792

Na Tabela 7, pode-se notar ainda que na medida em que aumenta o tamanho amostral ($m=1, m=2$ e $m=3$), piora o ajuste do MLG aumentando os valores do AIC e BIC, ampliando ainda mais a diferença para o MLGM.

4.2.5 Distribuição das estimativas dos efeitos de tratamentos

Na Figura 8, são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação {-2, -1, 0, 1, 2}, nessa ordem, nos

experimentos, com tamanho amostral $m = 1$. Nota-se que o viés $[E(\hat{\theta}) - \theta]$ das estimativas de efeito de tratamento é menor no MLGM do que no MLG, que sugere melhor ajuste do MLGM que o MLG, nesse tipo de análise.

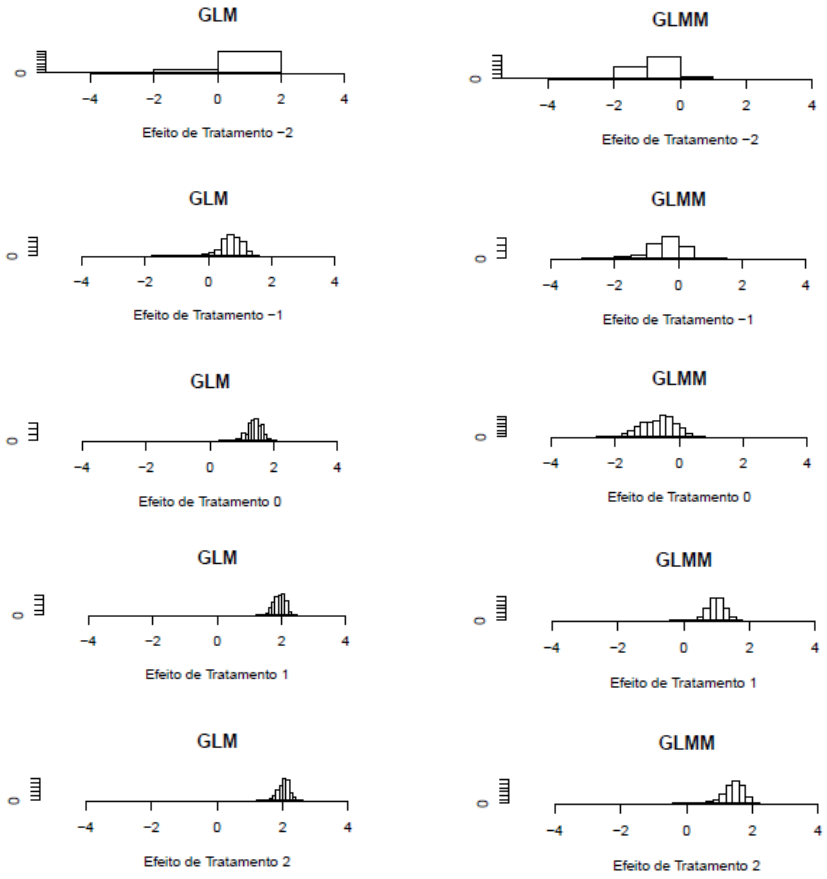


Figura 8 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m = 1$

Na Figura 9, são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação {-2, -1, 0, 1, 2}, nos experimentos com tamanho amostral $m = 2$. Confirmam-se os resultados anteriores para $m = 1$, em

que o viés das estimativas de efeitos de tratamentos $[E(\hat{\theta})-\theta]$ é menor no MLGM do que no MLG, configurando melhor ajuste do MLGM que no MLG.

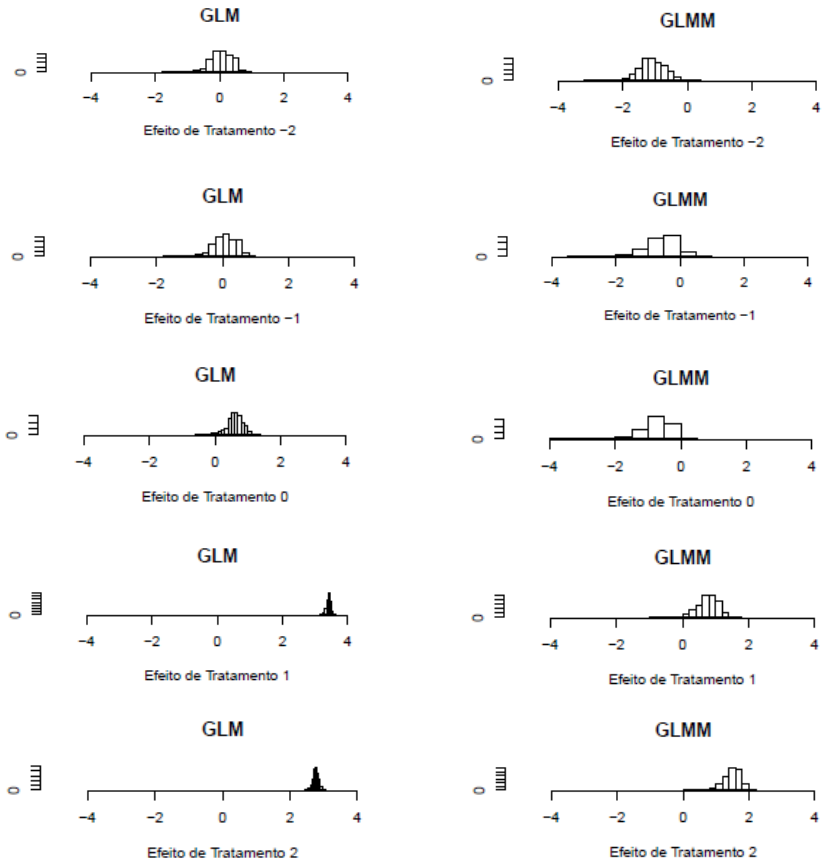


Figura 9 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m=2$

Resultados também semelhantes aos anteriores ($m=1$ e $m=2$) são apresentados na Figura 10, para as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação {-2, -1, 0, 1, 2}, nos experimentos com tamanho amostral com $m=3$. Nota-se que o viés das estimativas dos efeitos de

tratamentos $[E(\hat{\theta})-\theta]$ é menor no MLGM do que no MLG, indicando melhor ajuste do MLGM que no MLG, para esse tipo de análise.

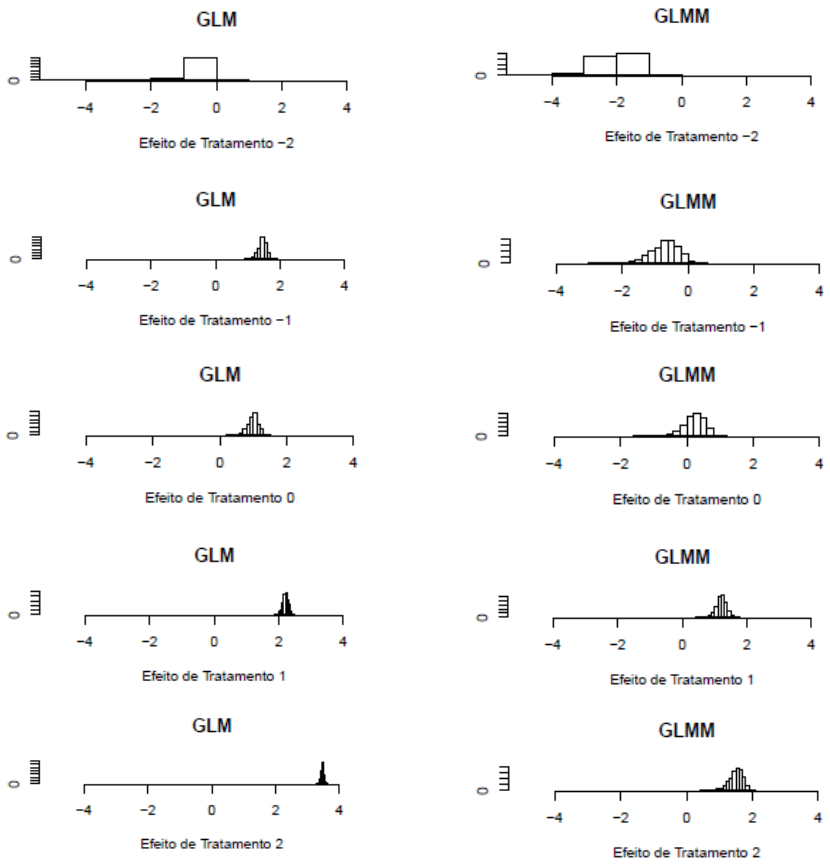


Figura 10 Distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2} nas análises MLGM e MLG, com $m=3$

4.2.6 Erros Quadráticos Médios das estimativas de efeitos de tratamentos

Na Tabela 8, são apresentadas as médias e desvios dos Erros Quadráticos Médios – EQM das estimativas de efeitos de tratamentos

$EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ para o MLG e MLGM com tamanho amostral $m = 1$, $m = 2$ e $m = 3$ e a eficiência relativa das estimativas de efeito de tratamento do MLG em relação ao MLGM. Pode-se notar que para todos os tamanhos amostrais ($m = 1$, $m = 2$ e $m = 3$) o MLGM apresenta menor EQM que o MLG, indicando mais proximidade do valor paramétrico de tratamento e sugerindo melhor ajuste.

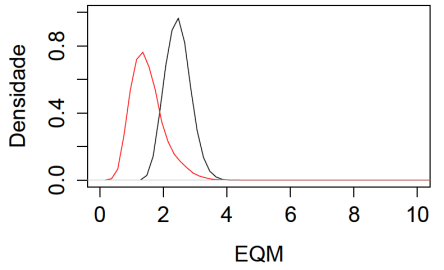
Tabela 8 Média, Desvio padrão dos EQM de tratamentos no MLG e MLGM e eficiência relativa das estimativas (EQM-MLG/EQM-MLGM) para ($m = 1$, $m = 2$ e $m = 3$)

	m=1		**m=2**		**m=3**	
	Média	**Desvio**	**Média**	**Desvio**	**Média**	**Desvio**
MLG	6,86	13,16	11,08	0,76	2,76	0,24
MLGM	5,91	13,14	3,72	0,45	0,88	0,26
MLG/MLGM	0,160	0,001	1,981	0,681	2,140	0,915

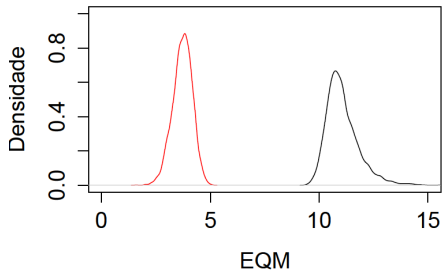
Pode-se observar ainda que o aumento do tamanho amostral (valores de m) implica também em aumento na diferença entre as médias dos EQM das estimativas de efeito de tratamento dos modelos MLG e MLGM. Esse resultado sugere mais eficiência do MLGM em relação ao MLG para valores maiores de m.

4.2.7 Distribuição do Erro Quadrático Médio das estimativas de efeitos de tratamentos

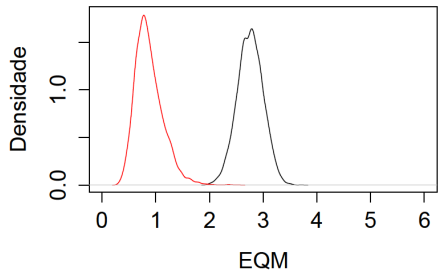
Na Figura 11, são apresentadas as distribuições do Erro Quadrático Médio – EQM das estimativas dos efeitos de tratamentos para os dois modelos de análise MLG e MLGM com os diferentes tamanhos amostrais $m=1$, $m=2$ e $m=3$. O MLGM está apresentado no gráfico na cor vermelha e o MLG na cor preta. Pode-se notar que, em média, o EQM das estimativas de efeitos de tratamentos é menor para o MLGM que no MLG, para todos os tamanhos amostrais considerados. Conforme esse critério, o MLGM aproxima mais do valor paramétrico que o MLG. Note que ao contrário do que se observou no modelo binomial, a diferença entre os EQM não aumenta "sempre" com o aumento do tamanho amostral m.



EQM para $m = 1$



EQM para $m = 2$



EQM para $m = 3$

Figura 11 Distribuições do EQM no MLG (preto) e MLGM (vermelho) para $m = 1$, $m = 2$ e $m = 3$

5 DISCUSSÃO

A *deviance* é uma importante medida de ajuste de modelo e auxilia de forma decisiva na sua escolha, assim como o critério de informação de Akaike (AKAIKE, 1974) - *AIC*, e o critério bayesiano de Schwarz (SCHWARZ, 1978) – *BIC*. *Deviances*, *AIC* e *BIC* baixos são desejáveis e podem sugerir um bom ajuste do modelo. Para analisar dois modelos nas mesmas condições experimentais, submetidos às mesmas observações, certamente, o melhor modelo (mais ajustado) será aquele que apresenta menores *deviance*, *AIC* e *BIC*.

A análise MLG estima o erro experimental $\hat{\sigma}_e^2$ com $(n.m - p)$ graus de liberdade; sendo n= número de UE's, m= número de UO nas UE's e p = número de parâmetros e compara o valor $(n.m-p)\hat{\sigma}_e^2 / \sigma_e^2$ com o quantil da distribuição qui- quadrado com $(n.m - p)$ graus de liberdade, para verificação do ajuste. Para a análise MLGM, o erro experimental foi particionado em componentes $\sigma_e^2 = \sigma_u^2 + \sigma_\varepsilon^2$ para capturar as variações entre e dentro da UE. No entanto a soma de duas variáveis independentes com distribuição qui-quadradas também segue uma distribuição qui-quadrado, com graus de liberdade que corresponde à soma dos graus de liberdade das duas (MOOD et al., 1974), ou seja: $\chi_a^2 + \chi_b^2 = \chi_{a+b}^2$. Para a análise MLGM com as qui- quadradas independentes $(n-p)\hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2$ e $n.(m-1)\hat{\sigma}_u^2 / \sigma_u^2$, em que sua soma $(n-p)\hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2 + n.(m-1)\hat{\sigma}_u^2 / \sigma_u^2$, representa uma qui- quadrado com $(n.m - p)$, pois, $(n-p) + n.(m-1) = n.m - p$, é razoável que se compare o valor da soma, da análise MLGM, com o quantil da distribuição qui- quadrado com $(n.m - p)$ graus de liberdade. Desse modo se as duas análises MLG e MLGM são comparadas ao quantil da distribuição qui-quadrado com os mesmos graus de

liberdade, julga-se mais ajustado o modelo que apresentar menor *deviance* residual. Este modelo foi sempre o MLGM, com exceção da situação $m = 1$ para o modelo binomial, em que eles não se distinguem.

O erro tipo I é outro importante parâmetro no julgamento da adequabilidade de um modelo, pois, quando se faz uma análise, espera-se que os resultados sejam confiáveis e/ou próximos do que seria uma situação real. Uma análise que apresenta altas taxas de erro tipo I indica que o modelo pode sugerir, com alta probabilidade, que existem diferenças entre tratamentos en quanto na verdade estas não existem. Ao contrário, modelos com baixas taxas de erro tipo I são desejáveis, pois, podem apresentar menores chances de errar.

A taxa de declaração de superdispersão é outra importante ferramenta na escolha de um modelo. Ela indica quantas vezes, em termos percentuais, foi estimada *deviance* maior que a esperada para o modelo Binomial ou Poisson. Desta forma, o MLG é considerado não ajustado por ser superdisperso. Nessa simulação, um modelo MLG com altas taxas de declaração de superdispersão indica muitos “erros” ao ignorar os efeitos aleatórios de parcelas no seu ajuste. Note que para os MLGM a *deviance* residual sempre foi a esperada para os respectivos modelos Binomial e Poisson.

Estas importantes ferramentas serão utilizadas no julgamento dos modelos MLG e MLGM para dados com respostas discretas, como segue.

As simulações foram capazes de ilustrar que quando as UE's recebem apenas uma resposta Bernoulli, as duas análises são equivalentes (Figura 2), como já era esperado, pois, nesse caso, não há variações da UE, apenas entre elas(?), que é a forma usual de análise MLG.

O modelo generalizado misto (MLGM) apresenta menor média das estimativas das *deviances* residuais, para a análise de ensaios binomiais do que o modelo generalizado fixo (MLG) para tamanhos amostrais das UE's maiores que 1 ($m > 1$), ou seja, quando cada UE possui mais de uma resposta Bernoulli

(Figura 2), demonstrando, portanto melhor ajuste. O mesmo ocorre com as demais medidas de ajustes *AIC* e *BIC* (Tabela 3), apresentando valores bem menores para o MLGM que o MLG, para todos os tamanhos amostrais maiores que 1 ($m=5$, $m=10$ e $m=15$). No caso do *AIC*, as diferenças do MLG para o MLGM, variam em termos percentuais de 29,8% (menor diferença) a 122,9% (maior diferença) a mais para o MLG. E no *BIC* essas diferenças do MLG para o MLGM, variam em termos percentuais de 25,3% a 104,1% a mais para o MLG.

Na Tabela 1, mostra-se que para todos os tamanhos amostrais (inclusive $m=1$) as taxas de erro tipo I são muito maiores no MLG que no MLGM. Sendo a maior taxa do MLGM 4,4%, en quanto o MLG chega a uma taxa de 98,7%. Que sugere, nesse caso, altíssima probabilidade 98,7% de acusar diferenças não existentes entre tratamentos, associada ao diagnóstico de superdispersão. Mais uma vez constata-se a maior adequabilidade do MLGM para dados com resposta binomial.

Na Tabela 2, confirmam-se os resultados anteriores quando apresenta a menor taxa de superdispersão em 39,3% para $m=1$ e 100% para $m=5$ no MLG. Isto Indica um modelo superdisperso, mal ajustado e pode levar a conclusões equivocadas quanto às diferenças entre médias de tratamentos.

Para as estimativas dos efeitos de tratamentos (Figuras 3,4 e 5) o modelo MLG tende a apresentar estimativas mais agregadas de efeitos de tratamentos, no entanto erra com maior frequência o valor paramétrico dos efeitos, comprovado também através dos EQM's. Isto confirma as maiores taxas de erro tipo I do MLG.

Como se pode notar na Figura 3 ($m=5$), o MLG erra nas estimativas dos efeitos paramétricos $\{-2,-1,0,1,2\}$, em -2 e 2, em relação ao MLGM. O mesmo se repete na Figura 4 ($m=10$), quando MLG erra nos mesmos valores, e na Figura 5 ($m=15$), erra na estimativa 2, enquanto o MLGM quando não acerta, se aproxima bastante do valor paramétrico. Em termos percentuais, o

MLG erra em pelo menos 20% (em relação ao número de estimativas) a mais que o MLGM nas estimativas dos efeitos de tratamentos para dados com resposta binomial.

Como visto o MLGM aproximou-se bastante dos valores paramétricos dos efeitos simulados de tratamentos, gerando, no entanto, estimativas mais dispersas. Isto se justifica por esse modelo levar em conta variação entre UE. Ao contrário do MLG em que as UO's são todas consideradas independentes, o MLGM computa uma componente da variância que reflete a agregação nas UE's. Isto torna potenciais testes de média mais confiáveis e menos sujeitos a erros do tipo I.

O erro quadrático médio é uma medida resumo da precisão (acurácia) das estimativas de efeito de tratamento. Na Tabela 4, observa-se que o MLGM se demonstrou mais preciso que o MLG, para todos os tamanhos amostrais $m > 1$, com precisão superior a 20% em relação ao MLG.

A Figura 6 apresenta as distribuições das estimativas das *deviances* residuais para os tamanhos amostrais $m = 1$, $m = 2$ e $m = 3$. Nota-se que a *deviance* média é muito menor no MLGM que no MLG. O mesmo ocorre com as demais medidas de ajustes *AIC* e *BIC* apresentando valores bem menores para o MLGM que o MLG, para todos os tamanhos amostrais. No caso do *AIC* as diferenças do MLG para o MLGM (Tabela 7), variam em termos percentuais de 264,12% até 496,02% a mais para o MLG. E no *BIC* essas diferenças do MLG para o MLGM, variam em termos percentuais de 248,06% até 465,72% a mais para o MLG. Denota-se, portanto, que o MLGM é mais ajustado para a análise de dados de contagens que a análise usual do MLG. Isto se justifica pela existência de variações entre e dentro das UE's.

Na Tabela 5, apresentam-se as taxas de erro tipo I. Nota-se que para todos os tamanhos amostrais $m = 1$, $m = 2$ e $m = 3$, o MLG apresenta taxas de 100%, em quanto a maior taxa apresentada pelo MLGM é de 1,7%. Mostra-se,

portanto, que o MLG pode acusar, com alta probabilidade, diferenças entre médias de tratamentos, que na realidade não existe. Enquanto a análise MLGM confirma os resultados anteriores, demonstrando-se mais ajustado para a análise de dados com resposta Poisson.

Na Tabela 6, apresentam-se as taxas de superdispersão para o MLG de 100% para todos os tamanhos amostrais $m = 1$, $m = 2$ e $m = 3$. Analogamente ao caso da binomial isto Indica um modelo superdisperso, mal ajustado e pode levar a conclusões equivocadas quanto às diferenças entre médias de tratamentos.

Análogo ao resultado do modelo binomial, no modelo Poisson (Tabela 8), o MLGM também se apresentou mais preciso que o MLG, sendo que o EQM das estimativas do GLMM foi pelo menos 16% menor que o do GLM.

O modelo MLG tende a apresentar estimativas mais agregadas de efeitos de tratamentos (Figuras 8, 9 e 10), no entanto erra com maior frequência o valor paramétrico dos efeitos. Isto confirma as elevadas taxas de erro tipo I ao se realizarem testes de comparações de médias.

Em análise à Figura 8 ($m = 1$), pode-se notar que o MLG erra em todas as estimativas dos efeitos paramétricos $\{-2,-1,0,1,2\}$. O MLGM, mesmo quando não acerta, se aproxima bastante do valor paramétrico. Na Figura 9 ($m = 2$), o MLG erra em 4 das 5 estimativas, enquanto o MLGM acerta em quase todas e finalmente na Figura 10 ($m = 3$), o MLG erra em todas as estimativas dos valores paramétricos ao contrário o MLGM acerta quase todas. Em termos percentuais, o MLG erra em pelo menos 80% (em relação ao número de estimativas) a mais que o MLGM nas estimativas dos efeitos de tratamentos para dados com resposta Poisson.

Conforme visto o MLGM aproximou-se bastante dos valores paramétricos dos efeitos simulados de tratamentos, gerando, no entanto, estimativas mais dispersas. Isto se justifica por esse modelo levar em conta

variação entre UE. Ao contrário do MLG em que as UO’s são todas consideradas independentes, o MLGM computa uma componente da variância que reflete a agregação nas UE's.

Considerando os resultados das duas análises MLG e MLGM e os parâmetros utilizados na comparação: *Deviance*, *AIC* e *BIC*, taxas de erro tipo I e de superdispersão e estimativas dos valores paramétricos e EQM, pode-se concluir que o MLGM, é mais ajustado e apresenta resultados mais confiáveis, para a análise de dados discretos que o MLG, para analisar efeitos fixos de tratamentos.

6 CONCLUSÕES

Em experimentos planejados, a análise de modelos Bernoulli ou Binomiais utilizando a metodologia de modelos lineares generalizados mistos (MLGM) deve ser recomendada em substituição à análise de modelos generalizados (fixos, MLG) sempre que se tiver mais que uma unidade observacional por unidade experimental.

Para os experimentos com resposta Poisson, a conclusão é ainda mais forte: modelos generalizados mistos (MLGM) são mais adequados para a análise do que o modelo generalizado fixo (MLG) ainda que haja apenas uma unidade observacional por unidade experimental.

Em ambos os casos, ao contrário dos GLM, os GLMM permitem controlar a taxa de erro tipo I e não sofrem com diagnósticos de superdispersão.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, Boston, v. 19, n. 6, p. 716-723, 1974.

BOZDONGAN, H. Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. **Psychometrika**, Williamsburg, v. 52, n. 3, p. 345-370, 1987.

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

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

DEMÉTRIO, C. G. B. **Modelos lineares generalizados na experimentação agronômica**. Porto Alegre: UFRGS, 1993. 125 p.

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

HARTLEY, H. O.; RAO, J. N. K. Maximum likelihood estimation for the mixed analysis of variance model. **Biometrika**, London, v. 54, n. 1/2, p. 93-108, 1967.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. **Journal of the American Statistics Association**, Washington, v. 72, p. 320-328, 1977.

HENDERSON, C. R. **Aplications of linear models in animal breeding**. Guelph: University of Guelph, 1984. 462 p.

HENDERSON, C. R. Estimation of variance and covariance components. **Biometrics**, Raleigh, v. 9, p. 226-252, 1953.

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

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 8., 1998, São Paulo. **Anais…** São Paulo: SINAPE, 1998b. 1 CD-ROM.

HINKELMANN, K.; KEMPTHORNE, O. **Design and analysis of experiments**. 2nd ed. New Jersey: J. Wiley, 2008. 631 p.

KEMPTHORNE, O. The randomization theory of experimental inference. **Journal of the American Statistical Association**, New York, v. 50, p. 946-967, 1955.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society. Series B**, London, v. 58, n. 4, p. 619-678, 1996.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models: a synthesis of generalized linear models, random effect models and structured dispersions. **Biometrika**, London, v. 88, n. 4, p. 987-1006, Jan. 2001.

LEE, Y.; NELDER, J. A.; PAWITAN, Y. **Generalized linear models with random effects**. London: Chapman & Hall, 2006. 396 p.

LITTELL, R. C. et al. **SAS system for mixed models**. Cary: Statistical Analysis System Institute, 2002. 633 p.

MARTINS, E. N. et al. **Modelo linear misto**. Viçosa, MG: UFV, 1993. 46 p.

MCCULLOCH, C. E.; SEARLE, S. R. **Generalized, linear, and mixed models**. New York: J. Wiley, 2001. 325 p.

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

MOLENBERGHS, G.; VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer Science, 2005. 683 p.

MOOD, A. M. et al. **Introduction to the theory of statistics**. 3rd ed. Tokyo: McGraw-Hill, 1974. 564 p.

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

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, London, v. 58, p. 545-554, 1971.

PAULA, G. **Modelos de regressão com apoio computacional**. São Paulo: EDUSP, 2013. 428 p.

PINHEIRO, J. C.; BATES, D. M. **Mixed-effects models in S and S-PLUS**. 2nd ed. New York: Springer, 2009. 530 p.

R DEVELOPMENT CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.R-project.org>. Acesso em: 10 mar. 2013.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics**, Hayward, v. 6, n. 2, p. 461-464, 1978.

SEARLE, S. R. **Linear models for unbalanced data**. New York: J. Wiley, 1987. 536 p.

WILK, M. B.; KEMPTHORNE, O. Fixed, mixed and random models. **Journal of the American Statistical Association**, New York, v. 50, p. 1144-1169, 1955.

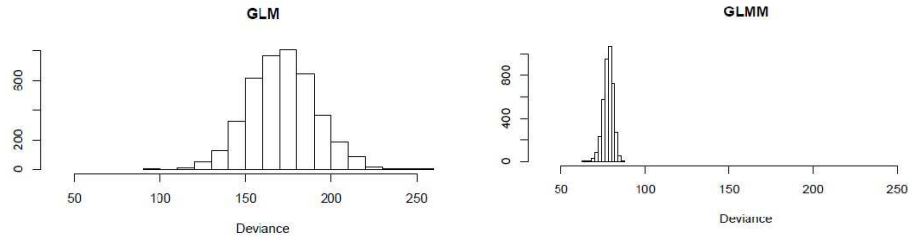
WOLFINGER, R. D. Covariance estruture selection in general mixed models. **Comunications in Statistics**, New York, v. 22, n. 4, p. 1079-1106, Apr. 1993.

WOLFINGER, R. D.; TOBIAS, R. D.; SALL, J. Computing gaussian likelihoods and their derivatives for general linear mixed models. **SIAM Journal on Scientific Computing**, Philadelphia, v. 15, n. 6, p. 1294-1310, 1994.

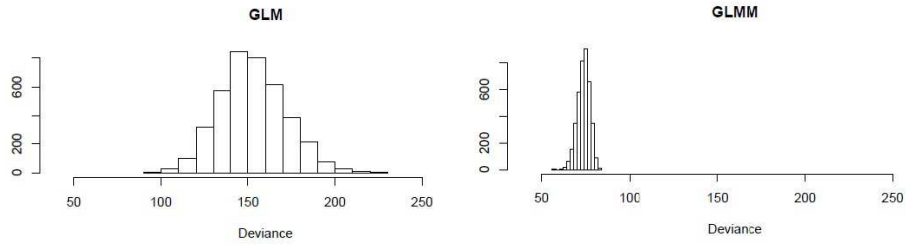
ANEXOS

ANEXO A - Distribuições das deviances residuais para $m = 20$ e $m = 25$ do modelo binomial

Na figura 12 apresentam-se as distribuições das estimativas das deviances residuais dos dois modelos em análise para os tamanhos amostrais $m = 20$ e $m = 25$. Nota-se que há diferença significativa entre o MLG e o MLGM sendo que no MLGM, a deviance média é muito menor que no MLG.



Deviance $m = 20$



Deviance $m = 25$

Figura 12: Deviances do MLG (à esquerda) e MLGM (à direita) para $m = 20$ e $m = 25$.

ANEXO B - Distribuições das estimativas de efeitos de tratamentos para $m = 20$ e $m = 25$ do modelo binomial

Na figura 13 apresentam-se as distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2}, nessa ordem, nos experimentos com $m = 20$. Nota-se que o viés das estimativas é menor no MLGM do que no MLG.

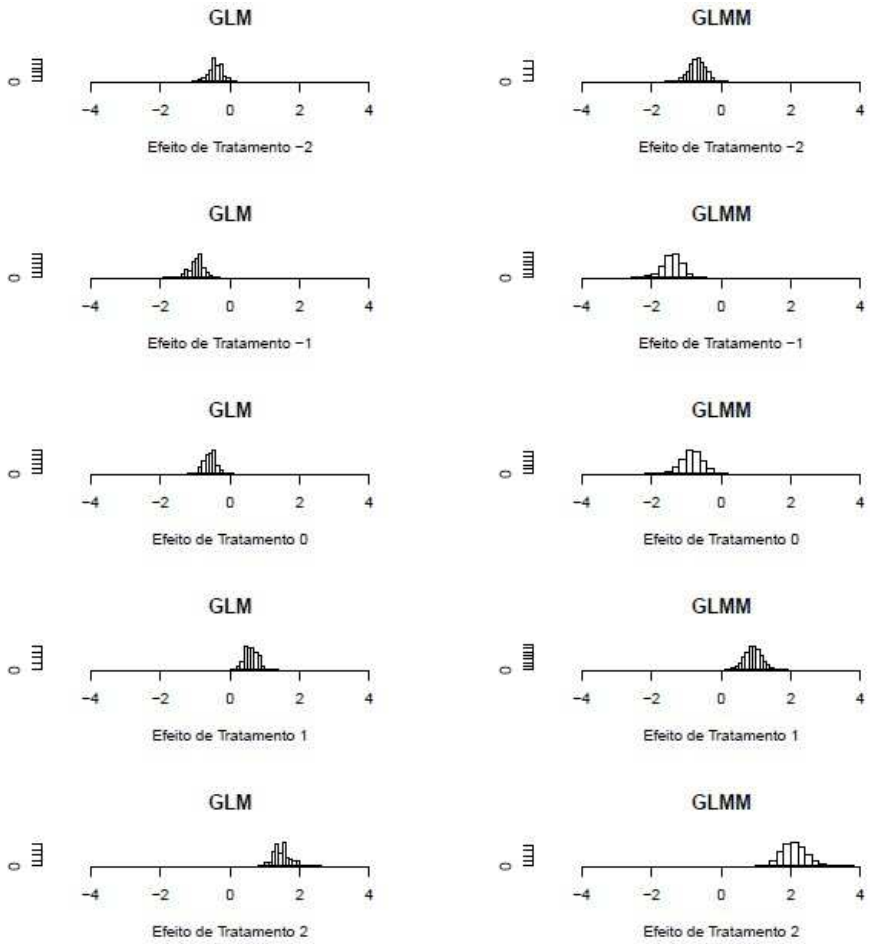


Figura 13: Distribuição de efeitos de tratamentos MLG (à esquerda) e MLGM (à direita) para $m = 20$.

Na figura 14 apresentam-se as distribuições das estimativas de efeitos dos tratamentos {-2, -1, 0, 1, 2}, nessa ordem, nos experimentos com $m = 25$. Nota-se que o viés das estimativas é menor no MLGM do que no MLG.

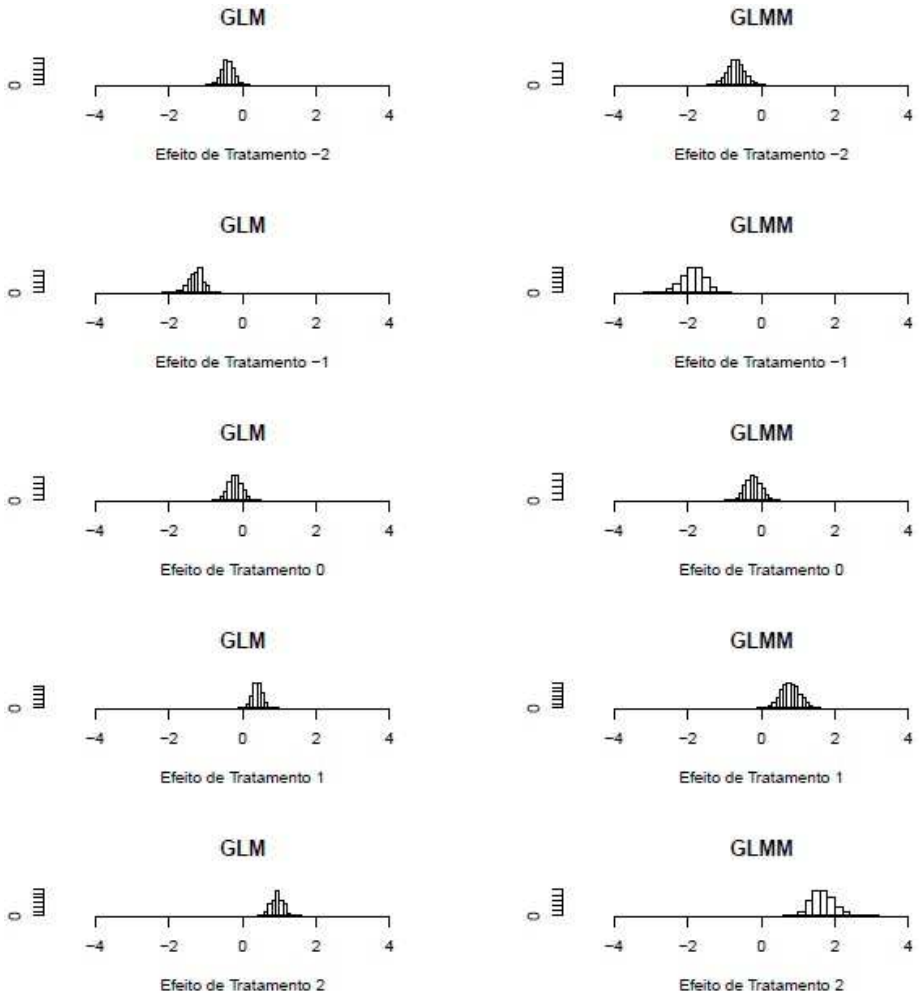


Figura 14: Distribuição de efeitos de tratamentos MLG (à esquerda) e MLGM (à direita) para $m = 25$.

ANEXO C - ROTINAS DE ANÁLISES

BINOMIAL

A função “analisa.Binomial(n,m,p,t)” abaixo ao mesmo tempo gera e analisa o experimento. Os parâmetros desta função são: n (número de observações do experimento), m (número de observações da UE), p (média paramétrica, soma de média da unidade experimental e efeito de tratamento, na escala da proporção) e t (vetor com os rótulos de tratamentos sorteados). Utilizamos em “analisa...” a função glmer() do pacote lme4 do R. A função apresentada neste exemplo devolve os seguintes resumos (respectivamente e para os modelos MLG e MLGM) deviance residual do modelo (dmlg, dmlgm) e coeficientes do preditor linear (fmlg, fmlgm). Outros resumos podem ser solicitados de forma análoga.

```
library(lme4)

analisa.Binomial <- function(n,m,p,t){
 y    <- rbinom(n,m,p)    # gera sucessos
 resp  <- cbind(y, m-y)    # concatena sucessos e fracassos
 Trat  <- factor(t)        # transforma t em fator
 mlg   <- glm(resp ~ -1+Trat, family=binomial) # MLG
 dmlg  <- deviance(mlg)    # Deviance
 fmlg  <- coef(mlg)        # Coeficientes
 parc  <- factor(1:n)      # cria fator para as UE
 mlgm  <- glmer(resp ~ -1+Trat
             +(1|parc), family=binomial) # MLGM
 dmlgm <- deviance(mlgm)   # Deviance
 fmlgm <- fixef(mlgm)      # Coeficientes
 return(list(dmlg, dmlgm, fmlg, fmlgm)) # saida
}
```

A forma de criar e passar as condições experimentais de um experimento para a função é:

```
baseTrat <- c(-2:2)
r   <- 5                    # número de repetições do DIC
t   <- sample(rep(baseTrat,r)) # aleatoriza os tratamentos
n   <- length(t)            # tamanho do experimento
m   <- 5                    # ensaios Bernoulli por parcela
U   <- c(c(-4:0),c(-3:1),c(-2:2),c(-1:3),c(0:4)) # efeito de UE
eta <- t+U                  # preditor linear
p   <- exp(eta)/(1+exp(eta))   # probabilidade de sucesso na UE
```

POISSON

A função “analisa.Poisson(n,m,p,t)” é análoga à da distribuição binomial, mas a variável resposta não precisa estar em duas colunas... O que muda nos parâmetros é que chamamos lambda o preditor. Utilizamos em “analisa” a função glmer() do pacote lme4 do R.

```
library(lme4)

analisa <- function(n,m,lambda,t){
  y     <- rpois(n,lambda)
  Trat  <- factor(t)
  mlg   <- glm(y ~ -1+Trat, family=poisson)
  dmlg  <- deviance(mlg)
  fmlg  <- coef(mlg)
  parc  <- factor(kronecker(1:length(t),rep(1,m)))
  mlgm  <- glmer(y ~ -1+Trat + (1|parc), family=poisson)
  dmlgm <- deviance(mlgm)
  fmlgm <- fixef(mlgm)
  return(list(dmlg, dmlgm, fmlg, fmlgm))
}
```

A forma de criar e passar as condições experimentais é também análoga:

```
baseTrat <- c(-2:2)
```

```
r   <- 5                    # número de repetições do DIC
t   <- sample(rep(baseTrat,r)) # aleatoriza os tratamentos
n   <- length(t)             # tamanho do experimento
m   <- 3                     # observações Poisson
U   <- c(c(-4:0),c(-3:1),c(-2:2),c(-1:3),c(0:4)) # efeito de UE
eta <- t+U                   # preditor linear
lambda <- exp(eta)           # media da Poisson na UE
```