LUCAS SOUZA DA SILVEIRA

MODELOS LINEARES GENERALIZADOS HIERÁRQUICOS MISTOS (HGLMM) AJUSTADOS VIA MÁXIMA VEROSSIMILHANCA HIERÁRQUICA (HIML) E HG-BLUP: OTIMIZAÇÃO DA ANÁLISE ESTATÍSTICA DE VARIÁVEIS CONTÍNUAS E CATEGÓRICAS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Marcos Deon Vilela de Resende

Coorientadores: Camila Ferreira Azevedo
Moysés Nascimento
Rodrigo Silva Alves

VIÇOSA – MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S587m
2022

Silveira, Lucas Souza da, 1991-
Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM) ajustados via Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP: otimização da análise estatística de variáveis contínuas e categóricas / Lucas Souza da Silveira. – Viçosa, MG, 2022.
1 tese eletrônica (88 f.): il. (algumas color.).

Orientador: Marcos Deon Vilela de Resende.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.657
Modo de acesso: World Wide Web.

1. Modelos lineares (Estatística). 2. Modelos multiníveis (Estatística). 3. Teoria bayesiana de decisão estatística.
I. Resende, Marcos Deon Vilela de, 1966-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.542

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

LUCAS SOUZA DA SILVEIRA

MODELOS LINEARES GENERALIZADOS HIERÁRQUICOS MISTOS (HGLMM) AJUSTADOS VIA MÁXIMA VEROSSIMILHANCA HIERÁRQUICA (HIML) E HG-BLUP: OTIMIZAÇÃO DA ANÁLISE ESTATÍSTICA DE VARIÁVEIS CONTÍNUAS E CATEGÓRICAS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 30 de junho de 2022.

Assentimento:

Lucas Souza da Silveira
Autor

Marcos Deon Vilela de Resende
Orientador

Dedico este trabalho a todos aqueles que estiveram comigo nos momentos em que precisei, me dando força e ânimo para continuar.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, que me manteve de pé e me deu a luz para seguir firme em meu propósito enquanto acadêmico.

Aos meus pais, por serem meus pilares durante toda a vida e por me ensinarem, por meio do exemplo, a buscar meus objetivos com honestidade e hombridade.

À minha esposa Paulinha, pelo companheirismo, amor, união e por todo o incentivo que deposita em mim. Obrigado por ser minha companheira!

Aos meus amigos por tornarem essa caminhada mais leve.

Aos meus sogros Cidinha e Edigar, por serem também minha família e, em especial à minha sogra Cidinha, pelas palavras de incentivo e por acreditar em mim como pessoa e profissional.

Aos meus parceiros e amigos de tantos anos de UFV, obrigado pelo convívio enriquecedor.

Aos meus professores do Departamento de Estatística da UFV por todo ensinamento transmitido com dedicação e amor. Obrigado por serem meu exemplo.

Agradeço em especial ao meu orientador Doutor Marcos Deon Vilela de Resende, pela paciência, pelo acolhimento e por todo o ensinamento proporcionado a mim. Obrigado pela oportunidade e confiança ao longo desses anos.

Aos meus coorientadores Doutor Rodrigo Silva Alves, Doutor Moysés Nascimento e Doutora Camila Ferreira Azevedo agradeço igualmente. A ajuda e contribuição de vocês foram muito importantes para que eu conseguisse alcançar meus objetivos.

Aos membros da banca examinadora Doutores (as) Marcos Deon Vilela de Resende, Moysés Nascimento, Leísa Pires Lima, Andrei Caíque Pires Nunes e Guilherme Ferreira Simiqueli, obrigado pela disponibilidade e pelas recomendações engrandecedoras.

Agradeço ainda à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro para realização desse trabalho.

A todos aqueles que de algum modo me auxiliaram e contribuíram para que este trabalho e essa conquista fosse possível, o meu muito obrigado!

“Não fui eu que ordenei a você? Seja forte e corajoso! Não se apavore ou desanime, pois, o Senhor, o seu Deus, estará com você por onde andar. ”
Josué,
1:9.

BIOGRAFIA

LUCAS SOUZA DA SILVEIRA, é filho de Aparecida Marcelino Souza da Silveira e Antônio Lúcio da Silveira e nasceu em 15 de maio de 1991, em Visconde do Rio Branco-Minas Gerais.

No ano de 2010 ingressou no curso de Licenciatura em Matemática, na Universidade Federal de Viçosa, tendo colado grau em janeiro no ano de 2015.

Deu continuidade em seus estudos na mesma instituição, iniciando no mestrado em Estatística Aplicada e Biometria, submetendo-se à defesa em fevereiro de 2017.

Logo após, ingressou no curso de Doutorado no Programa de Pós-Graduação também em Estatística Aplicada e Biometria na UFV, submetendo-se à defesa de tese em junho de 2022.

RESUMO

SILVEIRA, Lucas Souza da, D.Sc., Universidade Federal de Viçosa, junho de 2022. **Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM) ajustados via Maxima Verossimilhanca Hierarquica (HIML) e HG-BLUP: otimização da análise estatística de variáveis contínuas e categóricas.** Orientador: Marcos Deon Vilela de Resende. Coorientadores: Camila Ferreira Azevedo, Moysés Nascimento e Rodrigo Silva Alves.

Este trabalho teve como objetivo principal estudar e comparar os diferentes ajustes de Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM), em função das diferentes classes de variáveis aleatórias (contínuas e categóricas) e de estrutura de modelos. Estudar a flexibilização proporcionada pelos HGLMM frente aos comumente utilizados Modelos Lineares Mistos (LMM) e Modelos Lineares Generalizados Mistos (GLMM). Por vez, o HGLMM é uma classe de modelos mais ampla (a qual abarca todas as demais, via convergência, e, portanto, é a melhor modelagem que pode ser feita para qualquer classe de variáveis aleatórias) que engloba a inferência *bayesiana* e a inferência *fisheriana* (verossimilhancista) por meio da utilização da Verossimilhança Hierárquica. Nessa classe é possível ter como casos particulares de ajustes tanto o Modelo Linear Misto como o Modelo Linear Misto Generalizado. A maior vantagem do HGLMM é a possibilidade de atribuir outras distribuições (além daquela atribuída aos efeitos dos erros) da família exponencial aos fatores de componentes aleatórios do modelo e, assim, flexibilizar o ajuste em diferentes variáveis, modelos e bancos de dados. Para esse estudo foram utilizados três bancos de dados reais que possuíam características de diferentes naturezas como a contínua e categórica. Os ajustes foram comparados pelos valores do *Conditional Akaike Information Criterion* – cAIC (seleção de modelos) e da herdabilidade (via a maximização da acurácia). A herdabilidade foi utilizada para comparar as capacidades das estimativas dos componentes de variância capturarem adequadamente a variação genética e, portanto, propiciarem elevadas acurácias seletivas. Pode-se observar que o HGLMM alternativo teve o melhor ajuste em vários cenários. Para características *contínuas* foi competitivo com o Modelo Linear Misto na estimativa de componentes de variância e em ajuste de acordo com os valores cAIC. Para características categóricas se sobressaiu ao GLMM em diversos cenários. Conclui-se que os HGLMM podem ser usados corriqueiramente na análise estatística, em geral,

propiciando uma triagem eficiente de quais modelos usar na modelagem de cada uma das várias classes de variáveis aleatórias, quais sejam contínuas ou categóricas.

Palavras-chave: Modelos Lineares Generalizados. Modelos Hierárquicos. Seleção de Modelos. Inferência Estatística. Verossimilhança Hierárquica. Quadrados mínimos iterativos ponderados.

ABSTRACT

SILVEIRA, Lucas Souza da, D.Sc., Universidade Federal de Viçosa, June, 2022. **Generalized Hierarchical Mixed Linear Models (HGLMM) fitted via Hierarchical Maximum Likelihood (HlML) and HG-BLUP: optimization of the statistical analysis of continuous and categorical variables**. Advisor: Marcos Deon Vilela de Resende. Co-advisors: Camila Ferreira Azevedo, Moysés Nascimento and Rodrigo Silva Alves.

This work aimed study and compare the different fits of Hierarchical Generalized Linear Mixed Models (HGLMM), as a function of different classes of random variables (continuous and categorical) and model structure. At the same time, aimed the study of the flexibility provided by the HGLMM against the commonly used Linear Mixed Models (LMM) and Generalized Linear Mixed Models (GLMM). The HGLMM is a broader class of models (to which all others converge, and therefore the best modeling that can be done, for any class of random variables) that encompasses Bayesian inference and Fisherian inference through the use of Hierarchical Likelihood. In this class it is possible to have both the Mixed Linear Model and the Generalized Mixed Linear Model as particular cases of adjustments. The biggest advantage of HGLMM is the possibility of assigning others distributions (in addition to the one attributed to error effects) of the exponential family to the others factors of random components of the model and, thus, making the adjustment more flexible in different variables, models and databases. For this study, three real databases were used that had characteristics of different natures, such as continuous and categorical. The adjustments were compared by the values of the conditional Akaike information criterion– cAIC and the heritability (via maximizing accuracy). Heritability was used to compare the ability of the variance component estimates to adequately capture genetic variation. It can be seen that the alternative HGLMM had the best fit in various scenarios. For continuous characteristics it was competitive with the Mixed Linear Model in the estimation of variance components and in adjustment according to cAIC values. For categorical characteristics, GLMM excelled in several scenarios. It is concluded that the HGLMM can be used routinely in statistical analysis, in general, providing an efficient screening of which models to use in the modeling of each of the several classes of random variables, whether continuous and categorical.

Keywords: Generalized Linear Models. Hierarchical Likelihood. Flexible Inference. Model Selection. Selective Accuracy. Heritability.

SUMÁRIO

INTRODUÇÃO GERAL

Os Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM - *Hierarchical Generalized Linear Mixed Model*) foram propostos por Lee e Nelder (1996) para flexibilizar os pressupostos dos modelos tradicionalmente utilizados, como os Modelos Lineares Mistos (LMM - *Linear Mixed Model*) e os Modelos Lineares Generalizados Mistos (GLMM - *Generalized Linear Mixed Model*).

Os LMM são adequados para analisar dados com variáveis explicativas aleatórias cuja distribuição é normal (ou gaussiana). O análogo acontece nos GLMM, em que a distribuição dos resíduos é assumida como normal, mas, nesses modelos é possível assumir qualquer distribuição da família exponencial para distribuição dos resíduos.

Nos GLMM, é comum que as Distribuições Gama, Binomial e Poisson sejam utilizadas para modelar os resíduos. Já nos HGLMM, além dessas possibilidades, os fatores aleatórios do modelo também podem ser modelados com distribuições não gaussianas como as distribuições Gama, Gama-Inversa e Beta. Também é possível combinar diferentes distribuições para os diferentes fatores de efeitos aleatórios do modelo, ou seja, modelar dois ou mais fatores aleatórios com duas ou mais distribuições de probabilidade diferentes.

Todas as combinações possíveis permitem uma extensa lista de ajustes hierárquicos que possuem vantagens computacionais, pois, segundo Lee e Nelder (1996), esses modelos não utilizam a verossimilhança marginal para estimação e, portanto, não exigem integração.

Na prática, com o advento dos HGLMM, o uso da distribuição de probabilidade Gama para a modelagem e análise genética está ganhando espaço e se tornando importante em disputa com a distribuição Normal. A análise genética via distribuição Gama não é corriqueira e universal como a Normal. Assim, temas como a estimação da variância residual e da herdabilidade, o uso das equações de Modelo Misto (MME) com resíduos Gama, a predição BLUP em HGLMM e GLMM em uma distribuição de probabilidade Gama são relevantes.

No presente caso, no modelo para Y (variável fenotípica com distribuição Gama), os erros seguem a distribuição Gama. No caso de testes de progênies com várias plantas por parcela, esses erros estão associados a variação residual dentro de parcela. Essa variação residual (S^2) equivale a variação fenotípica dentro de

parcela (combinação progênie - bloco), a qual tem distribuição qui-quadrado (X^2_{n-1}) com n-1 graus de liberdade, em que n é o número de plantas por parcela, corrigido pela sobrevivência do caráter no experimento. Assim, $S^2 \sim X^2_{n-1}$. Tem-se também que Ln (S^2) ~ X^2_{n-1} (Resende, 2007). Assim, o tratamento das distribuições Gama e Qui-quadrado torna-se cada vez mais importante na genética e melhoramento.

Ademais, a metodologia para estimação de parâmetros presentes nos HGLMM foi apontada por Ronnegard e Lee (2010) como promissora na estimação dos componentes de variância em características binárias, de contagem e de sobrevivência. Os autores ainda citam que as estimativas dos componentes de variância para essas características, quando utilizado o GLMM, produzem estimativas de componentes de variância severamente viesadas, o que pode ser contornado pelos HGLMM.

Devido à escassez de informações sobre a análise genética com o uso da distribuição Gama, no capítulo 2 e capítulo 3 desse trabalho, foi desenvolvido e apresentado alguns tópicos relevantes na *avaliação genética Gama*, a qual propicia a *modelagem HGLMM* em variáveis contínuas e categóricas.

CAPÍTULO 1

REFERENCIAL TEÓRICO

1. Distribuições de probabilidade

Neste tópico apresentam-se todas as possíveis distribuições de probabilidade que foram avaliadas e utilizadas neste trabalho, sendo elas, a Distribuição Gaussiana (ou normal), Beta, Logística, Gama e Gama Inversa para variáveis aleatórias contínuas e distribuição de Poisson para dados categóricos.

Começando pelas distribuições que aportam variáveis contínuas ou variáveis cujos valores pertencem aos conjuntos de números reais, a Distribuição Gaussiana é facilmente amostrada em qualquer conjunto de dados, como por exemplo a altura e o peso de pessoas de uma população, comprimento e a largura das folhas de uma árvore, consumo de combustível de automóveis de uma mesma marca, entre outras infinidades de exemplos.

A Distribuição Gaussiana, segundo Bishop et al, (2006) é amplamente utilizada para distribuição de variáveis contínuas e no caso de uma única variável aleatória X, sua função densidade de probabilidade é descrita como

$$N(x|\mu,\sigma^2) = 1/(2\pi\sigma^2)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},$$

em que μ é a média (ou parâmetro de locação) e σ^2 é a variância (ou parâmetro de dispersão). Se $\mu = 0$ e $\sigma^2 = 1$, essa distribuição torna-se a distribuição normal padrão. Na Figura 1 é apresentada a curva da distribuição normal com diferentes parâmetros de média e dispersão.

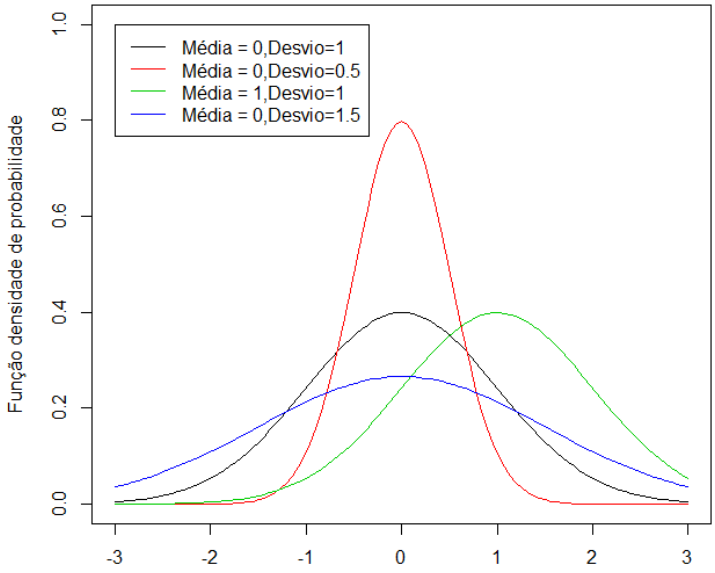

Figura 1. Funções densidade de probabilidades da distribuição normal com diferentes parâmetros de média e dispersão.

A distribuição descrita a seguir refere-se à Distribuição Gama. Essa distribuição é também amplamente utilizada para variáveis aleatórias contínuas e possui algumas propriedades particulares (como a flexibilidade no ajuste de diferentes tipos de dados) que a torna, em muitos casos, preferível em relação à Distribuição Gaussiana.

Segundo Ross (2014), uma variável aleatória X com Distribuição Gama possui função densidade de probabilidade

$$f(x|\alpha,\lambda) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

em que $\lambda > 0$, $\alpha > 0$ e $\Gamma(\alpha)$ é uma função gama definida como $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$. Caso α seja um número inteiro, a integração resulta em $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ ou $\Gamma(\alpha) = (\alpha-1)!$. Ainda, Ross (2014) descreve a relação da Distribuição Gama com outras distribuições de acordo com os valores dos parâmetros assumidos. Como exemplo, se o parâmetro $\lambda = 1/2$ e $\alpha = n/2$, com n inteiro positivo, então essa distribuição se torna uma distribuição χ_n^2 (qui quadrado com *n* graus de liberdade).

A Figura 2 apresenta as curvas das funções densidade de probabilidade de distribuições Gama de acordo com seus respectivos parâmetros. Essa imagem mostra o quanto a função gama é flexível.

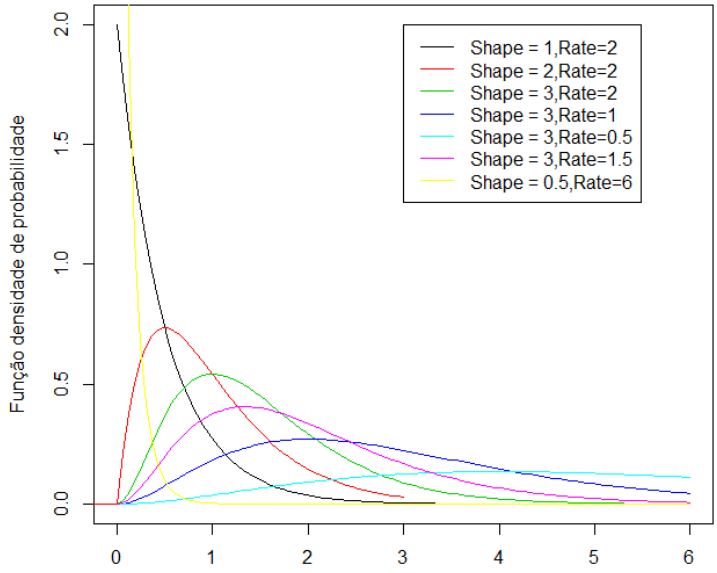

Figura 2. Função densidade de probabilidade de uma variável Gama com diferentes parâmetros de forma e parâmetros de taxa.

Por outro lado, se X é uma variável aleatória com Distribuição Gama(α, λ), então $1/X$ segue Distribuição Gama Inversa com parâmetros α e λ. Assim, seguindo a mesma padronização da Distribuição Gama, a função densidade de probabilidade da função gama inversa é

$$f(x|\alpha,\lambda) = \begin{cases} \frac{\left(\lambda e^{\frac{\lambda}{x}}\right)^{-1}}{\Gamma(\alpha)} \left(\frac{\lambda}{x}\right)^{\alpha+1}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

com $\alpha, \beta > 0$. As curvas da distribuição Gama Inversa com diferentes parâmetros de taxa e escala são desenhadas na Figura 3.

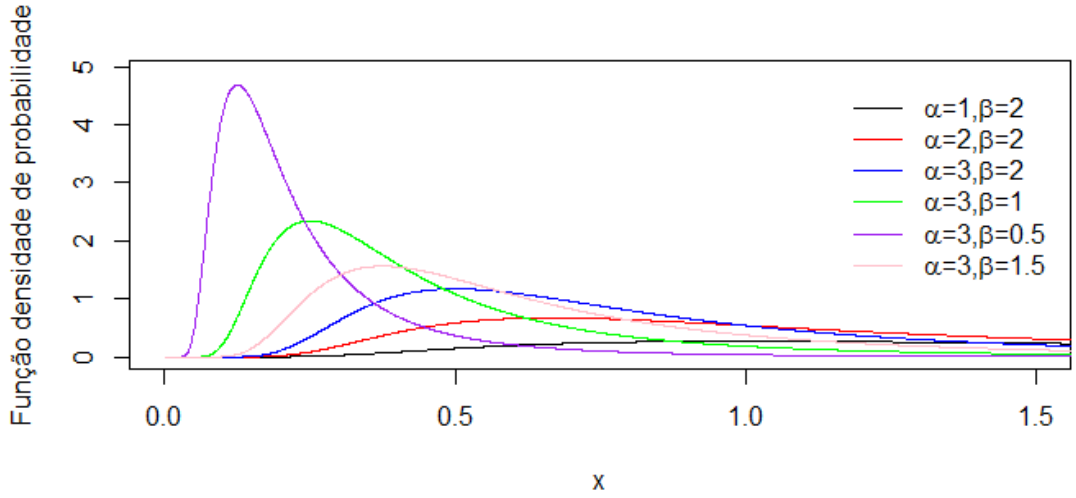

Figura 3. Função densidade de probabilidade de uma variável Gama Inversa

Ainda para dados contínuos, uma variável aleatória com distribuição Beta geralmente é indicada para casos onde existem intervalos limitantes. Segundo Ross (2014), uma variável com distribuição Beta possui a seguinte função densidade de probabilidade,

$$f(x|a,b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1 \\ 0, & x < 0 \end{cases}.$$

Nessa equação, $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. Conforme Ross, a densidade da distribuição é simétrica em ½ e a medida que b se torna maior que a, valores menores se tornam mais prováveis de acontecer. Uma ilustração dessa distribuição com diferentes parâmetros está na Figura 4.

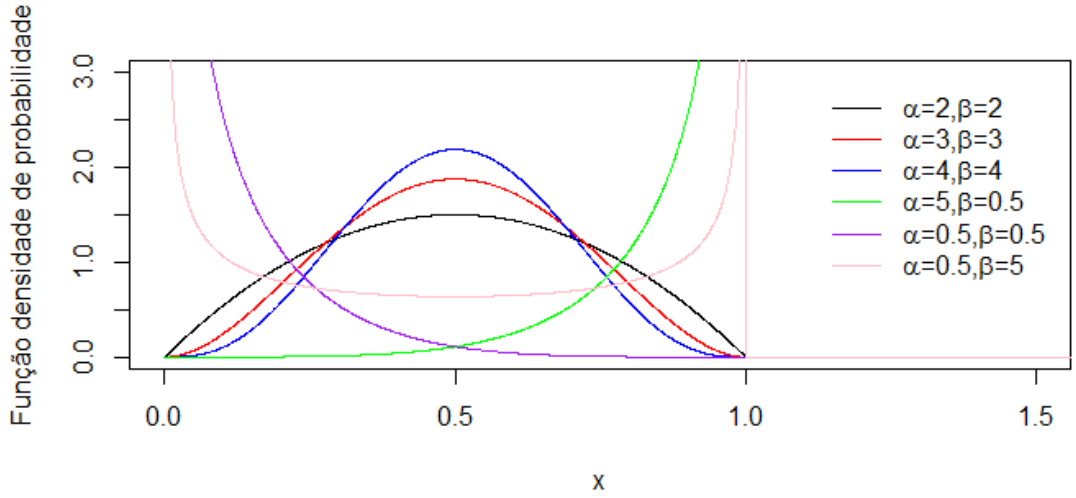

Figura 4. Função densidade de probabilidade de uma variável Beta

Outras distribuições que serão utilizadas neste trabalho dizem respeito à distribuição de variáveis discretas. Mais especificamente à Distribuição Binomial e a Poisson. A Binomial, na escala latente, tem Distribuição Logística. Considerando o parâmetro de locação μ na escala s, a função de densidade de probabilidade da distribuição logística é definida pela função

$$f(x;\mu,s) = \frac{e^{-\frac{(x-\mu)}{s}}}{s\left(1+e^{-\frac{(x-\mu)}{s}}\right)^2}.$$

Uma parametrização alternativa da distribuição Logística pode obtida expressando o parâmetro de escala, s, em termos de desvio padrão σ. Basta substituir $s = q\sigma$, em que $q = \sqrt{3}/\pi$.

A Figura 5 abaixo mostra as diferentes curvas da distribuição logística conforme a mudança de seus parâmetros de média e escala.

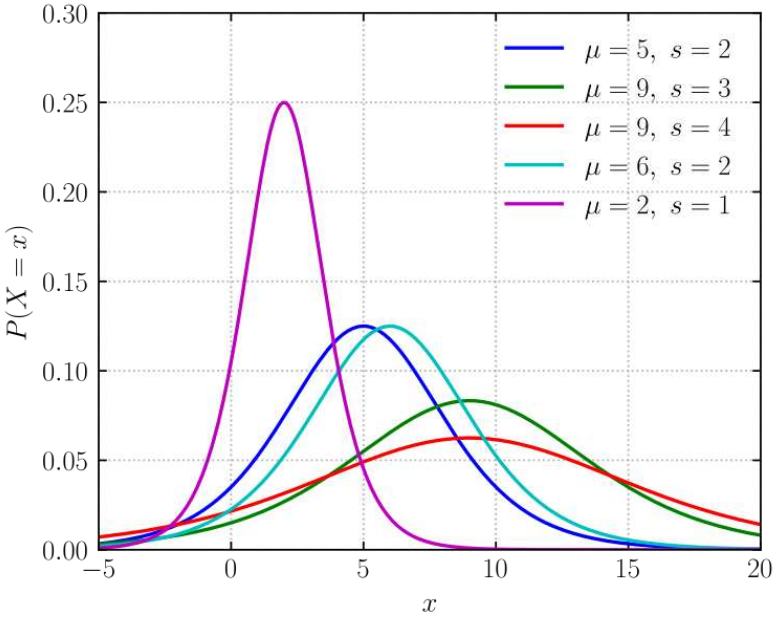

Figura 5. Função densidade de probabilidade de uma variável com distribuição Logística com diferentes parâmetros de média e escala.

Seguindo com distribuições contínuas, a distribuição exponencial dupla é uma variação da distribuição exponencial e, também, é conhecida como distribuição de Laplace. Sua densidade é dada por

$$f(x) = \frac{1}{2\lambda} \exp\left(\frac{-|x-\mu|}{\lambda}\right).$$

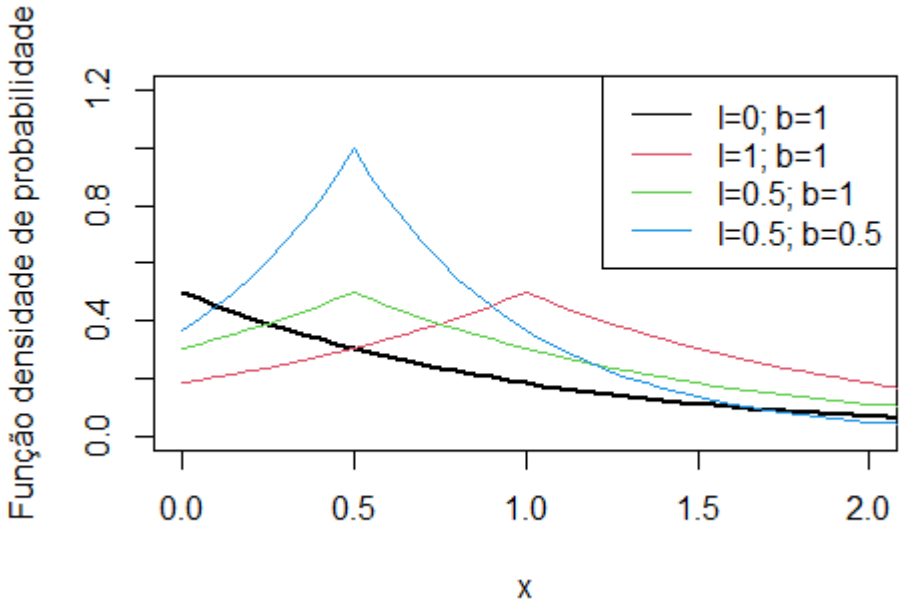

Figura 6 Função densidade de probabilidade de uma variável com distribuição exponencial dupla com diferentes parâmetros de média e escala.

Ao analisar as curvas dessa função densidade de probabilidade na Figura 6, é possível observar que essa distribuição possui simetria (assimetria = 0) no parâmetro de locação (μ) e a forma da curva é conduzida pelo parâmetro λ, também conhecido como parâmetro de escala. A média dessa distribuição é dada por μ e sua variância obtida como $2\lambda^2$.

Já a distribuição Weibull, Segundo Ross (2014), foi originalmente proposta para a interpretação de dados de fadiga e seu uso foi estendido a muitos outros problemas de engenharia.

A função densidade de probabilidade de uma variável com essa distribuição é dada por:

$$f(x) = \frac{\alpha}{\sigma^\alpha} \cdot x^{\alpha-1} \cdot exp\left(-\left(\frac{x}{\sigma}\right)^\alpha\right),$$

com α sendo o parâmetro de forma (shape) e σ o parâmetro de escala. A média dessa distribuição é obtida por $\sigma\Gamma\left(1+\frac{1}{\alpha}\right)$ e a variância por $\sigma^2\left[\Gamma\left(1+\frac{2}{\alpha}\right) - \left(\Gamma\left(1+\frac{1}{\alpha}\right)\right)^2\right]$. Na Figura 7 abaixo pode ser observados várias curvas da função densidade de probabilidade da distribuição Weibull com diferentes parâmetros de shape (a) e escala (b).

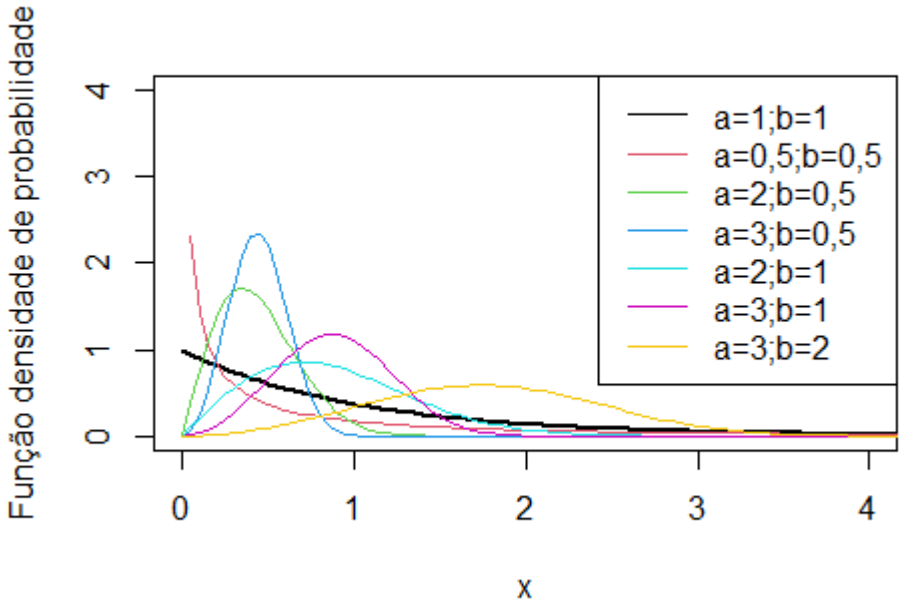

Figura 7. Função densidade de probabilidade de uma variável com distribuição Weibull com diferentes parâmetros de forma e escala.

Outra distribuição interessante é a Gumbel. De acordo com Aydin e Şenoğlu (2015), a distribuição Gumbel é um caso especial da distribuição de valor extremo generalizada, que por vezes, pode ser chamada de distribuição log-Weibull. A função densidade de probabilidade da distribuição Gumbel é descrita por:

$$f(x) = \frac{1}{\sigma} \cdot e^{-(z+e^{-z})},$$

com $z = \frac{x-\mu}{\sigma}$, μ o parâmetro de locação e σ o parâmetro de escala.

A média pode ser obtida a partir de $\mu + \sigma\gamma$, em que γé a constante de Euler (aproximadamente 0,5772) e a variância pode ser obtida por $\frac{\pi^2}{6} \cdot \sigma^2$. Essa distribuição possui assimetria caracterizada em 1,14 e curtose igual a 5,4 (Aydin e Şenoğlu, 2015). Tais características de assimetria e curtose podem ser notadas na ilustração das funções densidade de probabilidade da distribuição Gumbel com diferentes valores para locação e escala na Figura 8 abaixo.

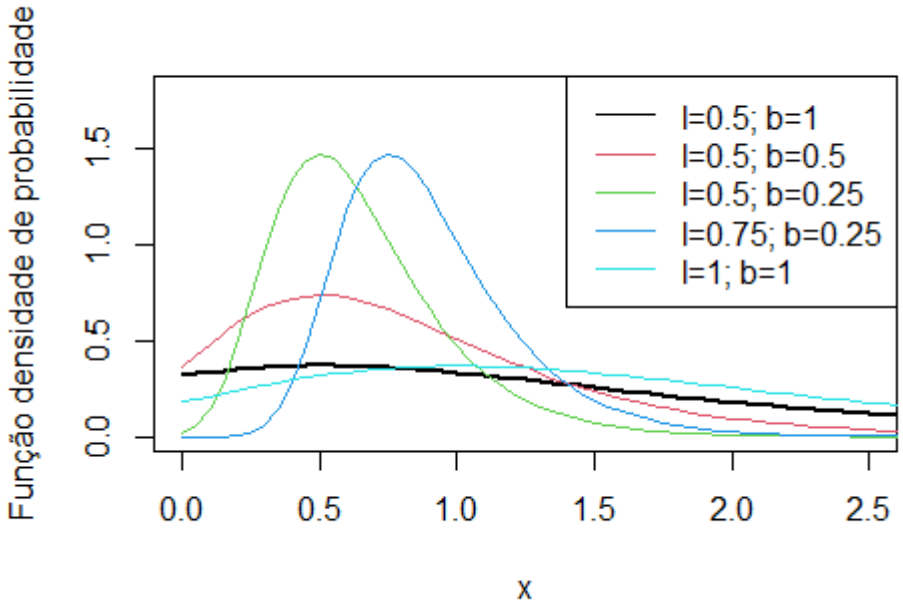

Figura 8. Função densidade de probabilidade de uma variável com distribuição Gumbel com diferentes parâmetros de forma e escala.

Para finalizar as distribuições de variáveis contínuas que serão abordadas nesse trabalho, a distribuição qui-quadrado (χ^2) é um caso particular da distribuição gama com λ = 1/2 e α = n/2, n um número inteiro positivo. Essa distribuição também pode ser obtida como o resultado de n variáveis independentes com distribuição normal padrão elevadas ao quadrado e somadas. A função densidade de probabilidade dessa distribuição é dada por:

$$f(x) = \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)};$$

com x > 0 e n graus de liberdade.

A média de uma variável que segue distribuição χ_n^2 é n e a variância é dada por $2n$. As curvas das funções densidade de probabilidade estão ilustradas na figura 9 abaixo com k diferentes parâmetros de graus de liberdade.

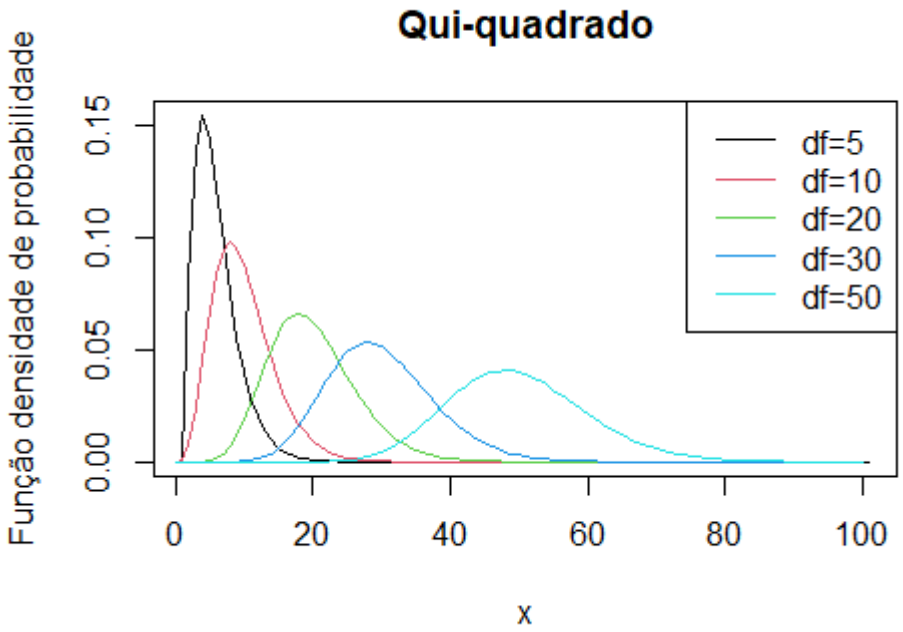

Figura 9 Função densidade de probabilidade de uma variável com distribuição Qui-Quadrado com diferentes parâmetros de graus de liberdade.

Assim como as variáveis contínuas, variáveis discretas também estão amplamente presentes no cotidiano. Como exemplo, é possível mencionar o número de espigas presentes em uma plantação de milho, número de bifurcações em uma árvore de eucalipto, número de tetos em porcas e diversas outras variáveis que podem ser mensuradas com números inteiros ou que permitem que sejam feitas categorizações.

Imagine-se que em um experimento aleatório, uma variável aleatória $X = 1$ represente o "sucesso" do experimento e $X = 0$ represente o "fracasso". Experimentos deste tipo descrevem o comportamento de uma variável aleatória com distribuição de Bernoulli cuja função de probabilidade é dada por $P(X = 1) = p$ e $P(X = 0) = 1 - p$, com $0 \leq p \leq 1$ a probabilidade associada ao "sucesso". Suponha-se que *n* tentativas independentes deste experimento sejam realizadas com probabilidade *p* de "sucesso". Sendo *X* o número de tentativas bem sucedidas deste experimento (sucesso), então a variável aleatória *X* segue uma Distribuição Binomial com parâmetros *n* tentativas com probabilidade *p*, isto é, $X \sim Bin(n, p)$. A função de probabilidade associada a essa variável, segundo Ross (2014) é definida por

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, i = 0, \dots, n,$$

em que i representa o número de sucessos. Se *n* é um valor supostamente grande e *p* um valor pequeno de forma que *np* tenha um valor moderado, então, de acordo com Ross, uma aproximação para a Distribuição Binomial é a distribuição de Poisson com parâmetro $\lambda = np$, cuja função densidade é descrita como

$$P(X = i) = \frac{e^{-\lambda}\lambda^{i}}{i!}.$$

Ademais, Ross apresenta uma série de aplicações onde a distribuição de Poisson é utilizada. Como exemplo, considere-se o número de erros de impressão em uma página ou o número de telefonemas discados incorretamente em um dia.

2. Função de verossimilhança

Antes de introduzir o tópico de verossimilhança hierárquica, ou *h-likelihood* (HL) como definido por Nelder e Lee (1996), é interessante abordar o que é uma função de verossimilhança.

Segundo Resende et al. (2014, p.398) a função de verossimilhança é definida considerando-se uma amostra aleatória $y_1, \dots, y_n$ independente e identicamente distribuída com função densidade de probabilidade ou função de probabilidade dada por $f(y|\theta)$, em que θ é o parâmetro da função. Assim, a função de verossimilhança de $\boldsymbol{y} = (y_1, \dots, y_n)$ é dada por $\ell(\theta; \boldsymbol{y}) = f(\boldsymbol{y}|\theta) = f(y_1, \dots, y_n|\theta)$ que pode ser fatorada nas marginais de forma que $f(y_1, \dots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$. Essa função formaliza a contribuição dos dados amostrais para o conhecimento do parâmetro que são estimados pelo método da máxima verossimilhança.

Tal método baseia-se na obtenção do ponto de máximo de uma função de verossimilhança (Resende et al. 2014). Por vez, o ponto de máximo é constituído pela diferenciação parcial de primeira ordem da função de verossimilhança (em alguns casos é utilizado o *log* da função de verossimilhança no intuito de facilitar os cálculos) em relação aos parâmetros e igualando essas diferenciações parciais a zero, onde a diferenciação parcial de segunda ordem deve ter valor menor que zero. Segundo Resende et al. (2014) esse estimador apresenta propriedades desejáveis como suficiência, eficiência, consistência e invariância a translação. No entanto apresentam estimativas com viés em decorrência da perda de grau de liberdade na estimação de efeitos fixos. Uma correção desse vício é o método de máxima verossimilhança residual (REML). Neste método, somente a proporção da verossimilhança que é

invariante aos efeitos fixos é maximizada, permitindo também a propriedade desejável de estimador sem viés (Resende et al., 2014, p.401).

Embora o método da máxima verossimilhança e máxima verossimilhança restrita tenham boas propriedades estatísticas, na presença de amostras não normais essas propriedades podem não ser mantidas. Contudo, Lee e Nelder (1996) propuseram a verossimilhança hierárquica (HL), definida como o logaritmo da função densidade conjunta de uma componente aleatória e hierárquica v e y, dado por

$$h = l(\theta', \phi; Y|v) + l(\alpha; v),$$

em que $l(\alpha; v)$ é o log da função densidade para v com parâmetro α e $l(\theta', \phi; Y|v)$ é o *log* da função densidade de $Y|u$. A componente aleatória v ocorre na escala na qual o efeito aleatório u ocorre linearmente no preditor linear, ou seja, v possui uma associação linear com a esperança condicional de Y dado u.

Segundo Lee e Nelder (1996), as estimativas derivadas da maximização dessa função, obtidas via $\frac{\partial h}{\partial \beta} = 0$ e $\frac{\partial h}{\partial v} = 0$, são denominados de Estimativas de Máxima Verossimilhança Hierárquica (MHLE).

Conforme Lee e Nelder (2006), as *log*-verossimilhanças marginais podem ser obtidas via integração e quando essas marginais são difíceis de serem obtidas numericamente surge a necessidade de uma nova forma de obtenção das estimativas. Por isso, Lee e Nelder propuseram as verossimilhanças restritas $p_v(h)$ e $p_{\beta,v}(h)$ como aproximações das *log*-verossimilhanças marginais.

De acordo com Resende et al. (2018), os estimadores de efeitos fixos e aleatórios do Modelo Linear Generalizado Hierárquico Misto (HGLMM - *Hierarchical Generalized Linear Mixed Model*) são derivados da maximização da Verossimilhança Hierárquica (HL) e produzem extensões diretas das equações de modelos mistos de Henderson. Os componentes de variância são estimados pela maximização do Perfil de Verossimilhança Hierárquica Ajustada (APHL - *Adjusted Profile Hierarchical Likelihood*), produzindo o método da Máxima Verossimilhança Hierárquica Iterativa (HIML), que é uma extensão direta do método Máxima Verossimilhança Restrita (REML) e equivale, também, a máxima verossimilhança penalizada. Por vez, o APHL é obtido a partir da substituição dos parâmetros de perturbação na função de verossimilhança por suas estimativas de máxima verossimilhança obtidas sob valores fixados dos parâmetros de interesse (Resende et al., 2018, p.65).

3. Método Delta

O Método Delta é uma técnica utilizada para aproximar um vetor aleatório através de uma expansão de Taylor sendo muito útil para deduzir a distribuição assintótica de variáveis (VAN DER VAART, 1998).

Segundo Cox (2005), o método delta está intimamente relacionado ao método de máxima verossimilhança. Isto é, se $\hat{\theta}$ é o estimador de máxima verossimilhança (MLE) de θ e g é uma transformação diferenciável, então $\hat{\phi}$ = g($\hat{\theta}$) é o MLE de ϕ = g(θ), ou seja, a distribuição assintótica de $\hat{\phi}$ obtida pelo método delta é a mesma do MLE.

Segundo Casella e Berger (2006), esse método também pode ser usado para obter uma aproximação para o valor médio e para a variância de uma função de variáveis aleatórias. O método é caracterizado por uma aproximação de Taylor de primeira ordem e pode ser definido como segue: sendo $\boldsymbol{\theta}$ um vetor de parâmetros e g uma função diferenciável e injetora tal que $\hat{g} = g(\widehat{\boldsymbol{\theta}})$, uma aproximação para a variância de $\hat{g}$ pode ser obtida por uma aproximação de Taylor de primeira ordem como

$$g(\widehat{\boldsymbol{\theta}}) \approx g(\boldsymbol{\theta}) + g'(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

em que $g'(\boldsymbol{\theta}) = \left[\frac{\partial \mathrm{g}}{\partial \theta_1}, \ldots, \frac{\partial \mathrm{g}}{\partial \theta_\mathrm{n}}\right]$ é a primera derivada parcial avaliada no vetor $\boldsymbol{\theta}$.

Sob pressuposição de $E[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$, tem-se $E[g(\widehat{\boldsymbol{\theta}})] = g(\boldsymbol{\theta})$ e

$$\begin{aligned} \boldsymbol{V}[\boldsymbol{g}(\hat{\theta})] &= \boldsymbol{E}\left[\boldsymbol{g}(\hat{\theta}) - \boldsymbol{E}[\boldsymbol{g}(\hat{\theta})]\right]^{\boldsymbol{2}} \\ &\approx \boldsymbol{E}[\boldsymbol{g}(\theta) + \boldsymbol{g}'(\theta)(\hat{\theta} - \theta) - \boldsymbol{g}(\theta)]^{\boldsymbol{2}} \\ &= \boldsymbol{E}[\boldsymbol{g}'(\theta)(\hat{\theta} - \theta)]^{\boldsymbol{2}} \\ &= \boldsymbol{V}[\boldsymbol{g}'(\theta)(\hat{\theta} - \theta)] \\ &= [\boldsymbol{g}'(\theta)]\boldsymbol{V}(\hat{\theta})[\boldsymbol{g}'(\theta)]^{\boldsymbol{t}} \end{aligned}$$

Portanto, a variância pelo método delta é dada por $\hat{V}[g(\widehat{\boldsymbol{\theta}})] = [\hat{g}'(\boldsymbol{\theta})]\hat{V}(\widehat{\boldsymbol{\theta}})[\hat{g}'(\boldsymbol{\theta})]^t$, em que $\hat{V}(\widehat{\boldsymbol{\theta}})$ é uma matriz com estimativas de variância e covariância dos parâmetros e $\hat{g}'(\boldsymbol{\theta})$ é o vetor com as derivadas parciais avaliadas na estimativa de $\boldsymbol{\theta}$.

4. Modelos Hierárquicos Generalizados Lineares Mistos

Os Modelos Lineares Mistos (LMM) para variáveis com distribuição Normal foram desenvolvidos por Henderson et al. (1959) e Henderson (1975) e utilizam predição BLUP para os efeitos aleatórios e REML (Patterson e Thompson, 1971; Thompson, 1973) para estimação de componentes de variância. Segundo Resende et al. (2018), quando os componentes de variância são desconhecidos, os estimadores dos componentes de variância conectam-se nas equações de modelos mistos de Henderson, resultando em um estimador não linear para o parâmetro de média.

Seguindo na linha histórica de construção dos modelos, os Modelos Lineares Generalizados (GLM) foram desenvolvidos por Nelder e Wedderburn (1972) para lidar com variáveis descontínuas ou variáveis em que as distribuições eram diferentes da distribuição normal. Anos depois, Thompson e Baker (1981) e Gilmour et al. (1985) desenvolveram os Modelos Lineares Generalizados Mistos (GLMM), acrescentando fatores aleatórios nos GLM. Adiante, Lee e Nelder (1996) estenderam a abordagem BLUP para uma classe ampla de modelos estatísticos com efeitos aleatórios, denominada *Modelos Hierárquicos Generalizados Lineares Mistos* (HGLMM). Segundo Resende et al. (2018), este método baseia-se na *extensão dos métodos de Quase-verossimilhança* de Nelder e Pregibon (1987).

Nos GLMM assume-se que os resíduos podem não apresentar distribuição Normal, mas, os demais efeitos aleatórios do modelo devem seguir distribuição Normal. Entretanto, essa suposição nem sempre é adequada. Resende et al. (2018) cita a situação em que os dados seguem distribuição Poisson e a função de ligação especificada para os resíduos é a logarítmica. Nesse caso, o mais adequado para os fatores aleatórios do modelo seria uma distribuição Gama com função de ligação logarítmica, o que torna o modelo pertencente a classe dos HGLMM por especificar uma distribuição de probabilidade e uma função de ligação para cada fator aleatório. Segundo Resende et al. (2018), para os modelos cujos os fatores aleatórios não são de classificação hierárquica, uma denominação alternativa para os HGLMM são Modelos Lineares Mistos Generalizados Estratificados.

Um preditor BLUP e uma combinação do BLUP com modelos Tweedie de dispersão baseados em distribuição Exponencial para HGLMM foi apresentado por Lee e Ha (2010). Entretanto, segundo Resende et al. (2018), para dados e HGLMM

não Normais o BLUP linear pode não ser tão eficiente.

Conforme Resende et al. (2018) os HGLM são modelos condicionais e incluem modelos não lineares. Além disso, discorrem sobre a vantagem do estimador de moda sobre o estimador de média amostral, que permite a seleção do melhor modelo ao invés do modelo médio. Entretanto, Ma e Jorgensen (2007) não recomendam o uso de estimativas modais para efeitos aleatórios e sim, o uso do método BLUP ortodoxo sob médias proposto por eles. Já, Lee e Ha (2010) mostraram que a estimação pela moda da função de verossimilhança hierárquica via HG-BLUP proporcionaram melhor precisão estatística e manutenção do nível declarado de probabilidade de cobertura em relação BLUP ortodoxo.

A opção do ajuste dos vários fatores de efeitos aleatórios sob diferentes suposições de distribuição pode ser feito via HGLMM, ou seja, a definição dessas distribuições não precisa ficar confinada apenas na distribuição Normal. Essa flexibilidade era disponível apenas para os GLMM, ou seja, apenas para o fator aleatório de erros. Como consequência dessa flexibilização, a maior eficiência preditiva e de seleção pode ser alcançada, principalmente, no melhoramento de plantas que possui muitos fatores de efeitos aleatórios.

Sob a suposição de normalidade, nos LMM, o estimador de Quadrados Mínimos Ponderados (WLS) e o estimador REML, que funcionam sob os modelos condicional e marginal, são estimadores de máxima verossimilhança e, segundo Resende et al. (2018), mesmo sob falha da normalidade e incorreta especificação da matriz de covariância, esses estimadores são consistentes. Nos GLMM essas propriedades não são asseguradas, pois ambos os estimadores serão viesados se a pressuposição de normalidade falhar.

Segundo Resende e Alves (2021), o procedimento HIML supera todos os outros métodos baseados na aproximação de Laplace para a estimação de GLMM para dados binários. Resende et al (2018) dizem que os métodos baseados na aproximação de Laplace como os de Schall (1991) e Breslow e Clayton (1993) não tem boa performance. A maximização da verossimilhança com inobserváveis, embora possa ocorrer com várias técnicas computacionais, são substancialmente viesadas, como os métodos de Séries de Taylor (como PQL - *Penalized Quasi-likelihood*) e aproximações de Laplace. Entretanto, para Mollenberghs et al. (2009), o método de Schall ainda é um eficiente algoritmo de estimação e a Verossimilhança Hierárquica tem um papel importante na literatura.

Segundo Resende e Alves (2021), a HL pode ser usada também na derivação de ferramentas para seleção de modelos como o Critério de informação de Akaike (AIC) condicional da HL, pois esse equivalente ao Critério de Informação da *Deviance* (DIC) aplicada na Estatística Bayesiana (Lee and Noh, 2012).

5. Máxima Verossimilhança Hierárquica (HIML) ou Estendida

Essa abordagem para estimação dos parâmetros foi proposta por Lee e Nelder em 1996. A chamada Verossimilhança Estendida ou Hierárquica é baseada na maximização da verossimilhança conjunta, ou seja, a função é maximizada conjuntamente com respeito aos parâmetros v e $\theta = \beta$, dados os parâmetros de dispersão λ. A função de Verossimilhança Hierárquica ou Estendida é dada por: $L_H(\beta, \lambda, v|y, v) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{\beta,\lambda}(y_{ij}|v_i) f_\lambda(v_i)$, em que: $f_{\beta,\lambda}(y_{ij}|v_i)$ é a verossimilhança condicional da j-ésima (*j=1, ..., n*) observação repetida no i-ésimo (*i=1, ..., N*) indivíduo, isto é, y_{ij}; $f_\lambda(v_i)$ é a verossimilhança do i-ésimo efeito aleatório; e λ contém parâmetros de dispersão dos componentes aleatórios v_i, assim como os parâmetros que descrevem a dispersão residual (φ) da resposta y_{ij}.

Segundo Resende et al. (2018), na HL as estimativas dos parâmetros de dispersão são determinadas pela maximização da Verossimilhança Hierárquica Perfilada Ajustada e no caso de modelos lineares mistos a Verossimilhança Perfilada Ajustada é exatamente igual ao REML. Outra vantagem da HL é a possibilidade de inferência para parâmetros fixos, aleatórios e variáveis não observadas, por possuir o conceito de probabilidade preditiva. Esse conceito é facilmente aceito por bayesianos e frequentistas (Lee e Kim, 2016), por permitir interpretações para os intervalos de credibilidade e de confiança.

Estruturalmente, os efeitos fixos da estrutura de médias podem ser estimados via HL ou via aproximação (da verossimilhança marginal) de Laplace de primeira ordem. Também, conforme Lee e Nelder (2001) é possível fazer uma modelagem conjunta da estrutura de média e dispersão dos modelos em que os parâmetros de superdispersão podem ser estimados ou serem fixados em valores teóricos. No geral, para a estimação dos parâmetros do modelo utiliza-se como algoritmo o método dos Quadrados Mínimos Ponderados Iterativos (IWLS), que, segundo Resende et al. (2018), os efeitos fixos e aleatórios são estimados usando verossimilhança estendida,

e os parâmetros de dispersão são obtidos via maximização da verossimilhança perfilada ajustada.

Adicionalmente, os HGLMM usam distribuições da variável resposta advindas da família exponencial e distribuições para os efeitos aleatórios advindos de distribuições bayesianas conjugadas. As principais distribuições da variável resposta são Gaussiana, Binomial, Poisson e Gama. Já para os demais fatores aleatórios do modelo é possível o ajuste das distribuições Gaussiana, Beta, Gama e Gama Inversa.

Segundo Resende et al. (2018) os métodos de Quase-verossimilhança penalizada (PQL) para GLMM de Schall (1991) e Breslow e Clayton (1993) são o mesmo do método de HL, mas ignorando $\delta v/\delta \varphi$ e $\delta v/\delta \lambda$ na estimação da dispersão. Esse fato acarreta em severo viés, especialmente para dados binários. Já o Viés do estimador de HL para β, conforme Resende et al. (2018), podem ser evitados pela introdução do APHL. Para Resende et al. (2018), robustez é uma questão de parametrização e não de modelo.

Ademais, a abordagem de HL permite que seja trabalhado com quantidades não observáveis sem depender de uma abordagem empírica. Essa fato, permite que seja abordado, por meio da HL, os modelos hierárquicos bayesianos defendidos por Lindley e Smith (1972). Além disso, nesses modelos a otimização conjunta da HL fornece um algoritmo estatisticamente e numericamente eficiente que é expresso como o ajuste de um grupo de GLMM interligados e não requer distribuições *a priori* dos parâmetros e nem quadraturas multidimensionais.

Referências

ALKIMIM, Emilly Ru as et al. Designing the best breeding strategy for Coffea canephora: Genetic evaluation of pure and hybrid individuals aiming to select for productivity and disease resistance traits. **PloS one**, v. 16, n. 12, p. e0260997, 2021.

AYDIN, Demet; ŞENOĞLU, Birdal. Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. **Journal of Modern Applied Statistical Methods**, v. 14, n. 2, p. 12, 2015.

ALVES, Rodrigo Silva et al. Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. **Tree Genetics & Genomes,** v. 16, n. 2, p. 1-8, 2020.

ALVES, Rodrigo Silva et al. Multiple-trait BLUP: a suitable strategy for genetic selection of Eucalyptus. **Tree Genetics & Genomes**, v. 14, n. 5, p. 1-8, 2018.

BRESLOW, N.E.; CLAYTON, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. **Journal of the American Statistical Association**, 88: 9-25.

BISHOP, Christopher M. **Pattern recognition and machine learning**. Springer, 2006.

CHISSOM, Brad S. Interpretation of the kurtosis statistic. **The American Statistician**, v. 24, n. 4, p. 19-22, 1970.

D'AGOSTINO, Ralph B. Transformation to normality of the null distribution of g1. **Biometrika**, p. 679-681, 1970.

HENDERSON, C.R. (1975). Best linear estimation and prediction under a selection model. **Biometrics**, 31: 423-447.

HENDERSON, C.R.; KEMPTHORNE, O.; Searle, S.R.; Von Krosigk, C.M. (1959). The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, 15: 192-218.

LEE, Youngjo; NELDER, John A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society:** Series B (Methodological), v. 58, n. 4, p. 619-656, 1996.

LEE, Y.; NELDER, J.A. (2001). Hierarchical generalized linear models: A synthesis of generalised linear models, random effect models and structured dispersions. **Biometrika**, 88: 987-1006.

LEE, Youngjo; NELDER, John A. Double hierarchical generalized linear models (with discussion). **Journal of the Royal Statistical Society:** Series C (Applied Statistics), v. 55, n. 2, p. 139-185, 2006.

LEE, Y.; HA, I.D. (2010). Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. **Statistics and Computing**, 20: 295-303.

LEE, Y.; KIM, G. (2016). H-likelihood predictive intervals for unobservables. **International Statistical Review**, 84: 487-505.

LEE, Youngjo; RÖNNEGÅRD, Lars; NOH, Maengseok. **Data analysis using hierarchical generalized linear models with R**. Chapman and Hall/CRC, 2017.

LINDLEY, D.V.; SMITH, A.F. (1972). Bayes estimates for the linear model**. Journal of the Royal Statistical Society**: Series B (Methodological), 34: 1-41.

MA, R.; JORGENSEN, B. (2007). Nested generalized linear mixed models: Orthodox best linear unbiased predictor approach. **Journal of the Royal Statistical Society**: Series B, 69: 625–641.

NELDER, J.A.; PREGIBON, D. (1987). An extended quasi-likelihood function. Biometrika, 74: 221-232.

NELDER, J.A.; WEDDERBURN, R.W.M. (1972). Generalized linear models. **Journal of the Royal Statistical Society**, 135: 370-384.

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

RESENDE, M.D.V. (ORG.); SILVA, F.F (ORG.); AZEVEDO, C.F. (ORG.). **Estatística Matemática, Biométrica e Computacional**: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. 1. ed. Visconde do Rio Branco: Suprema, v.1, p.881, 2014.

RESENDE, M.D.V.; AZEVEDO, C.F.; SILVA, F.F.; NASCIMENTO, M.; GOIS, I.B.; ALVES, R.S. **Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP**.1. ed. Visconde do Rio Branco: Suprema, v.1, p.151, 2018.

RESENDE, M.D.V. and ALVES, R.S. (2020) Linear, generalized, hierarchical, bayesian and random regression mixed models in genetics/genomics in plant breeding. **Functional Plant Breeding Journal**, v.2, 1-31.

RONNEGARD, L., SHEN, X. and ALAM, M. (2010) hglm: A Package for Fitting Hierarchical Generalized Linear Models. **The R Journal**, 2(2): 20-28.

ROSS, Sheldon. **A first course in probability**. Pearson, 2014.

SCHALL, H. (1991). Estimation in generalized linear models with random effects. **Biometrika**, 78: 719–727.

THOMPSON, R. (1973). The estimation of variance and covariance components when records are subject to culling. **Biometrics**, 29: 527-550.

THOMPSON, R.; BAKER, R.J. (1981). Composite link functions in generalized linear models. **Applied Statistics**, 30: 125-131.

VRIEZE, S. I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). **Psychological methods**, v. 17, n. 2, p. 228, 2012.

CAPÍTULO 2

FITTING HGLMM MODELS BY HIML/HG-BLUP FOR CONTINUOUS AND DISCRETE DATA IN GENETICS

Resumo

Este trabalho teve como objetivos comparar alguns Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM) que diferem entre si em função da utilização de diferentes suposições de distribuição dos fatores de efeitos aleatórios, em diferentes classes de variáveis aleatórias continuas e discretas. Visa também os comparar aos comumente utilizados Modelos Lineares Mistos (LMM, que se baseiam na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo), e aos Modelos Lineares Generalizados Mistos (GLMM, que se baseiam na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo, exceto os erros, que podem assumir outras distribuições). Uma das maiores vantagens dos HGLMM é a possibilidade de atribuir várias outras distribuições (além da Normal) de probabilidade pertencentes a família exponencial aos demais fatores de efeitos aleatórios do modelo, além dos erros, que já são flexibilizados via GLMM. Os ajustes foram comparados pelos valores do *conditional Akaike Information Criterion* – cAIC (para a seleção de modelos, visando a inferência; sendo condicional devida a associação com clusters ou hierarquias) e da maguinetude da herdabilidade (h^2) (via a maximização da acurácia da predição). A h^2 foi utilizada para comparar as capacidades de as estimativas dos componentes de variância capturarem adequadamente a variação genética e, portanto, propiciarem elevadas acurácias seletivas. Um HGLMM alternativo (com distribuição Gamma) teve o melhor ajuste em vários cenários. Para características *contínuas* foi competitivo (igual ou superior) ao Modelo Linear Misto na estimativa de componentes de variância e em ajuste de acordo com os valores cAIC. Para características *categóricas* e advindas de *contagem* se sobressaiu o GLMM em diversos cenários. Conclui-se que os HGLMM podem ser usados vantajosamente na análise estatística, em geral, propiciando uma triagem eficiente de quais modelos e distribuições usar nos ajustes de cada uma das várias classes de variáveis aleatórias, (continuas, binárias, categóricas ou de contagem).

Palavras-chave: Modelos Lineares generalizados, Modelos hierárquicos, Seleção de modelos, Inferência estatística, Verossimilhança Hierárquica. Quadrados mínimos iterativos ponderados.

1. Introdução

Os Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM - *Hierarchical Generalized Linear Mixed Models*) foram propostos por Lee e Nelder (1996) para flexibilizar os pressupostos dos modelos tradicionalmente utilizados, como os Modelos Lineares Mistos (LMM - *Linear Mixed Models*) e os Modelos Lineares Generalizados Mistos (GLMM - *Generalized Linear Mixed Models*). Os LMM são adequados para analisar dados com variáveis classificatórias e explicativas aleatórias cuja distribuição é normal (ou Gaussiana). Nos GLMM, a distribuição dos resíduos pode ser assumida como normal, mas, é possível assumir qualquer distribuição da família exponencial para distribuição dos resíduos (Resende e Alves 2020).

As famílias exponenciais incluem muitas das distribuições mais comuns tais quais: distribuição normal, exponencial, gama, qui-quadrado, beta, Dirichlet, Bernoulli, distribuição categórica (também chamada distribuição de Bernoulli generalizada, distribuição multinoulli), Poisson, Wishart, Wishart inversa, distribuição geométrica . Várias distribuições comuns são famílias exponenciais, mas somente quando certos parâmetros são fixados e conhecidos. Por exemplo: binomial (com número fixo de tentativas), multinomial (com número fixo de tentativas), binomial negativa (com número fixo de falhas). Exemplos de distribuições comuns que não são da família exponencial são a t de Student, a maioria das distribuições de mistura e até mesmo a família de distribuições uniformes quando os limites não são fixos (Stroup, 2010?).

Nos GLMM é comum o uso das distribuições Gama, Binomial e Poisson para modelar os resíduos. Já nos HGLMM, além dessas possibilidades, os outros fatores aleatórios do modelo também podem ser ajustados com distribuições não gaussianas como as distribuições Gama, Gama-Inversa e Beta. Também é possível combinar diferentes distribuições para os diferentes fatores de efeitos aleatórios do modelo, ou seja, modelar dois ou mais fatores aleatórios com duas ou mais distribuições de probabilidade diferentes. Todas as combinações possíveis permitem uma extensa lista de ajustes hierárquicos que possuem vantagens teóricas e computacionais, pois, segundo Lee e Nelder (1996), esses modelos não utilizam a verossimilhança marginal

ou restrita para estimação e, portanto, não exigem integração. A estimação dos parâmetros de dispersão é feita por meio do perfil ajustado de verossimilhança hierárquica que funciona como uma generalização da máxima verossimilhança restrita (REML). A metodologia para estimação de parâmetros dos HGLMM foi apontada por Ronnegard e Lee (2010) como promissora na estimação dos componentes de variância em características binárias, de contagem e de sobrevivência. Os autores ainda citam que as estimativas dos componentes de variância para essas características discretas, quando utilizado o GLMM, produzem estimativas de componentes de variância severamente viesadas, o que pode ser contornado pelos HGLMM.

De acordo com Resende et al. (2014), Resende et al. (2018) e Resende e Alves (2020), os estimadores de efeitos fixos e aleatórios dos modelos (HGLMM) são derivados da maximização da Verossimilhança Hierárquica (HL) e produzem extensões diretas das equações de modelos mistos de Henderson. Os componentes de variância são estimados pela maximização do Perfil de Verossimilhança Hierárquica Ajustada (APHL - *Adjusted Profile Hierarchical Likelihood*), produzindo o método da Máxima Verossimilhança Hierárquica Iterativa (HIML), que é uma extensão direta do método Máxima Verossimilhança Restrita (REML) e equivale, também, a máxima verossimilhança penalizada. Por vez, o APHL é obtido a partir da substituição dos parâmetros de perturbação na função de verossimilhança por suas estimativas de máxima verossimilhança obtidas sob valores fixados dos parâmetros de interesse (Resende et al., 2018, p.65).

Os Modelos Lineares Mistos (LMM) para variáveis com distribuição Normal foram desenvolvidos por Henderson et al. (1959) e Henderson (1975) e utilizam predição BLUP para os efeitos aleatórios e REML (Patterson e Thompson, 1971; Thompson, 1973) para estimação de componentes de variância. Segundo Resende et al. (2018), quando os componentes de variância são desconhecidos, os estimadores dos componentes de variância conectam-se nas equações de modelos mistos de Henderson, resultando em um estimador não linear para o parâmetro de média.

Seguindo na linha histórica de construção dos modelos, os Modelos Lineares Generalizados (GLM) foram desenvolvidos por Nelder e Wedderburn (1972) para lidar com variáveis descontínuas ou variáveis em que as distribuições eram diferentes da distribuição normal. Anos depois, Thompson e Baker (1981) e Gilmour et al. (1985)

desenvolveram os Modelos Lineares Generalizados Mistos (GLMM), acrescentando fatores aleatórios nos GLM. Adiante, Lee e Nelder (1996) estenderam as abordagens LMM e GLMM para uma classe ampla de modelos estatísticos com efeitos aleatórios, denominada *Modelos Hierárquicos Generalizados Lineares Mistos* (HGLMM), que engloba também a estimação Bayesiana.

Nos GLMM assume-se que os resíduos podem não apresentar distribuição Normal, mas, os demais efeitos aleatórios do modelo devem seguir distribuição Normal. Entretanto, essa suposição nem sempre é adequada. Resende et al. (2018) cita a situação em que os dados seguem distribuição Poisson e a função de ligação especificada para os resíduos é a logarítmica. Nesse caso, o mais adequado para os fatores aleatórios do modelo seria uma distribuição Gama com função de ligação logarítmica, o que torna o modelo pertencente a classe dos HGLMM por especificar uma distribuição de probabilidade e uma função de ligação para cada fator aleatório.

A opção do ajuste dos vários fatores de efeitos aleatórios sob diferentes suposições de distribuição pode ser feita via HGLMM, ou seja, a definição dessas distribuições não precisa ficar confinada apenas na distribuição Normal. Essa flexibilidade era disponível apenas para os GLMM, ou seja, apenas para o fator aleatório de erros. Como consequência dessa flexibilização, a maior eficiência preditiva e de seleção pode ser alcançada, principalmente, no melhoramento de plantas que possui muitos fatores de efeitos aleatórios (Barbosa et al., 2005; Pedroso et al., 2009; Resende, 1999; Resende e Barbosa, 2005).

Os objetivos deste trabalho foram: flexibilizar a definição das distribuições dos fatores aleatórios de vários LMM e GLMM; verificar as possibilidades de um melhor ajuste para variáveis de crescimento de plantas ao considerar a variável resposta (*y*) como tendo Distribuição Gama (aproximação à distribuição Weibull); ajustar e comparar HGLMMs para variáveis contínuas e discretas.

2. Materiais e Métodos

2.1. Conjunto de dados de variáveis contínuas

Foram avaliadas 286 famílias de irmãos completos de eucalipto. O experimento foi conduzido em delineamento de blocos casualizados, com seis plantas por parcela e oito repetições. O plantio foi realizado em novembro de 2003 e o espaçamento foi 3 m entre linhas e 2 m entre árvores. As condições experimentais foram descritas por Alves et al. (2018). Aos dois, cinco e sete anos de idade, foram mensurados a altura (m), diâmetro à altura do peito (DAP) (cm) e volume (m^3). No total foram avaliadas 9 variáveis (3 traits x 3 ages).

2.2. Conjunto de dados de variáveis discretas categóricas associadas a escores

Foram avaliadas 20 famílias híbridas de café canephora. O experimento foi conduzido em delineamento de blocos casualizados, com uma planta por parcela e até 35 repetições. O plantio foi realizado em março de 2011 e o espaçamento foi 3 m entre linhas e 1.5 m entre plantas. As condições experimentais foram descritas por Alkimim et al. (2021).

Por três anos consecutivos (2013 a 2015), foram mensuradas as incidências de ferrugem (causada pelo fungo *Hemileia vastatrix Berk. & Br.*) avaliadas em uma escala de 1 a 5. Para as análises, essa escala foi modificada para variar de 0 a 4, onde 0 foi atribuído a plantas assintomáticas e 4 foi atribuído a plantas altamente suscetíveis aos patógenos.

2.3. Análise exploratória de dados

Visando conhecer os dados em termos de distribuição e estrutura, foram realizadas análises exploratórias usando histogramas e suas curvas de densidade de probabilidade. Para isso empregou-se as funções densidades de probabilidades para as variáveis continuas e função de probabilidade para variáveis discretas.

Adicionalmente, avaliou-se a aderência dos dados às respectivas distribuições. Para verificação de aderência a curva normal foram empregados os testes D' Agostino

(1970) e Anscombe-Glynn (Chissom, 1970) para assimetria e curtose. Os resultados da analise exploratória revelaram quais variáveis apresentaram ou não distribuição normal. Isto remeteu as análises aos diferentes HGLM.

2.4. Ajuste dos modelos HGLMM

2.4.1. Máxima Verossimilhança Hierárquica (HIML)

Segundo Resende et al. (2018), na verossimilhança hierárquica (HL) as estimativas dos parâmetros de dispersão são determinadas pela maximização da Verossimilhança Hierárquica Perfilada Ajustada e no caso de modelos lineares mistos a Verossimilhança Perfilada Ajustada Maximizada é exatamente igual ao REML, podendo ser denominada HIML. Uma vantagem do HIML é a possibilidade de inferência para parâmetros fixos, aleatórios e variáveis não observadas, por possuir o conceito de probabilidade preditiva. Esse conceito é facilmente aceito por bayesianos e frequentistas (Lee e Kim, 2016), por permitir interpretações para os intervalos de credibilidade e de confiança, respectivamente.

Os efeitos fixos da estrutura de médias podem ser estimados via HL ou via aproximação (da verossimilhança marginal) de Laplace de primeira ordem. Também, conforme Lee e Nelder (2001) é possível fazer uma modelagem conjunta da estrutura de média e dispersão dos modelos em que os parâmetros de superdispersão podem ser estimados ou serem fixados em valores teóricos. No geral, para a estimação dos parâmetros do modelo utiliza-se como algoritmo o método dos Quadrados Mínimos Ponderados Iterativos (IWLS), em que, segundo Resende et al. (2018), os efeitos fixos e aleatórios são estimados usando verossimilhança estendida (hierárquica), e os parâmetros de dispersão são obtidos via maximização da verossimilhança perfilada ajustada, produzindo o HIML.

Os modelos lineares generalizados hierárquicos mistos (HGLMM) foram ajustados via máxima verossimilhança hierárquica (HIML) e HG-BLUP pelo algoritmo dos quadrados mínimos ponderados iterativos (IWLS), conforme adaptado de Resende et al. (2018). Para a predição e a estimação via HIML/HG-BLUP os seguintes passos são necessários: definição do modelo estatístico, construção da função de verossimilhança hierárquica, derivação das equações de Modelos Mistos Hierárquicos, derivação dos estimadores de componente de variância via

verossimilhança perfilada ajustada, desenvolvimento de um algoritmo (IWLS) iterativo eficiente e implementação computacional eficaz.

Para a estimação nessa nova classe de modelos, a ideia da Verossimilhança Estendida ou Hierárquica foi introduzida como critério a ser maximizado. Esse método pode então ser denominado **HIML**, embora não tenha sido assim denominada por Lee e Nelder. O método permite um algoritmo (do tipo IWLS) simples expresso na forma de GLMM interconectados. Esse algoritmo não demanda o uso de quadratura no ajuste e nem probabilidades *a priori*; é também, mais rápido do que as demais alternativas.

2.4.2. Equações de modelo hierárquico generalizado misto e algoritmo para HG-BLUP e HIML

Para predição e estimação via HIML/HG-BLUP, são necessários os passos que se seguem.

a. Definição do modelo estatístico: GLMM: Normal-Normal
$y = X\beta + Zv + e$, em que: $Var(v) = D\sigma_v^2$ e $Var(e) = \Sigma\sigma_e^2$.

b. Construção da função de verossimilhança hierárquica: Função de Verossimilhança Estendida (Hierárquica)

$$\begin{aligned}\log L(\beta, v; y, v) &= \log f(y, v)\\ &= \log f(y|v) + \log f(v)\\ &= \tfrac{1}{2}\log|2n\Sigma| - \tfrac{1}{2}(y - X\beta - Zv)^t\Sigma^{-1}(y - X\beta - Zv) - \tfrac{1}{2}\log|2nD| - \tfrac{1}{2}v^tD^{-1}v.\end{aligned}$$

c. Derivação das Equações de Modelos Mistos Hierárquicas: Maximização da função de Verossimilhança Estendida

Passo 1: Derivação da função para β e v.

$dL(\beta,t,v; y,t,v)/d\beta = X^t\Sigma^{-1}(y - X^t\beta - Z^tv)$; e

$dL(\beta,t,v; y,t,v)/dv = Z^t\Sigma^{-1}(y - X^t\beta - Z^tv)$

Passo 2: Equações de Modelo Misto.

$$\begin{bmatrix} X^t\Sigma^{-1}X & X^t\Sigma^{-1}Z \\ Z^t\Sigma^{-1}X & Z^t\Sigma^{-1}Z + D^{-1} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} X^t\Sigma^{-1}y \\ Z^t\Sigma^{-1}y \end{bmatrix}.$$

Passo 3: Informação de Fisher.

A Informação de Fisher para *β* é: $I(\hat{\beta}) = X^t\Sigma^{-1}X$.

A Informação de Fisher para *v* é: $I(\widehat{v}) = (Z^t\Sigma^{-1}Z + D^{-1})$.

Passo 4: Modelo Misto - componentes de média aumentado.

O Modelo Misto aumentado é dado por: $y_a = T\delta + e_a$, em que:

$y_a = \begin{pmatrix} y \\ \psi_M \end{pmatrix}$; $T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}$; $\delta = \begin{pmatrix} \beta \\ v \end{pmatrix}$; e $e_a = \begin{pmatrix} e \\ e_M \end{pmatrix}$.

Assim: $\begin{pmatrix} y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}\begin{pmatrix} \beta \\ v \end{pmatrix} + \begin{pmatrix} e \\ e_M \end{pmatrix}$, em que: $\psi_M = 0$ refere-se ao dado quase-aumentado, com distribuição Normal com média $E(\psi_M) = v$, variância *D* e é independente de *y*. O subscrito *M* denota o modelo para a média e não para a dispersão. A ideia desse modelo linear aumentado não acrescenta algo novo na análise de Modelos Mistos Normais, mas torna-se muito útil na extensão para Modelos Mistos Não Normais.

2.4.3. Equações de Modelo Misto Aumentado

$$\left[\begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}'\begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1}\begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}\right]\begin{pmatrix} \beta \\ v \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}'\begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1}\begin{pmatrix} y \\ \psi M \end{pmatrix}, \text{ e}$$

$T^t\Sigma_a^{-1}T\hat{\delta} = T^t\Sigma_a^{-1}y_a$, em que: $\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}$.

Este sistema pode ser resolvido como uma regressão ponderada ordinária do tipo:

$$X'V^{-1}X\,\hat{\beta} = X'V^{-1}\,y, \text{ em que } V \text{ é substituída por } \Sigma_a.$$

Este sistema de equações é idêntico às equações de Modelo Misto de Henderson. Nas aplicações em melhoramento animal e vegetal, os efeitos aleatórios *v* são correlacionados, de acordo com uma dada matriz de correlação *A*. Esse modelo pode ser transformado pela recomputação de *Z* como $ZA^{1/2}$, em que $A^{1/2}$ é alguma transformação raiz quadrada de *A* (transformação Cholesky). Então, o modelo aumentado descrito acima, contendo efeitos aleatórios independentes e identicamente distribuídos, pode ser aplicado, sem a necessidade de inversão da matriz *A*. Os modelos de análises de fatores contêm parâmetros em *A*, a serem

estimados e $ZA^{1/2}$ contém os carregamentos dos fatores.

d. Derivação dos estimadores de componente de variância via verossimilhança perfilada

Sob essa nova *Z*, tem-se que $Var(y) = (\lambda ZZ^t + \Phi I_n)$, em que *n* é o número de observações. Sob o modelo aumentado tem-se $Var(e_a) = \Sigma_a^{-1} = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} \Phi I_n & 0 \\ 0 & \lambda I_m \end{pmatrix}$, em que *m* é o tamanho do vetor *v*. Neste caso, a distribuição $v \sim N(0, \lambda A)$ torna-se $u \sim N(0, \lambda I)$.

Os componentes de *deviance* correspondentes a *e* são os resíduos quadráticos dados por: $d_i = (y_i - X_i\hat{\beta} - Z_i\hat{v})^2$. Os componentes de *deviance* correspondentes a e_M são os resíduos quadráticos dados por: $d_{Mi} = (\psi M - \hat{v}_i) = \hat{v}_i^2$. E os *leverages* (pontos de influência ou alavancagem) correspondentes são dados pelos elementos diagonais de: $T(T^t\Sigma_a^{-1}T)^{-1}T^t\Sigma_a^{-1}$.

e. Obtenção de um algoritmo iterativo eficiente (IWLS)

A estimação de parâmetros (β, τ, e *v*, em que: τ são parâmetros de variância associados a *v*) no modelo linear misto: $y = X\beta + Zv + e$ (sendo *e* com distribuição normal e variância σ^2), pode ser feito pelo algoritmo IWLS para o modelo linear aumentado como mostrado a seguir:

Passo 1. Comece com uma estimativa do parâmetro de variância τ;

Passo 2. Dado a estimativa corrente de τ, atualize $\hat{\delta}$ resolvendo as equações de quadrados mínimos generalizados (ou de modelos mistos) aumentados: $T^t\Sigma_a^{-1}T\hat{\delta} = T^t\Sigma_a^{-1}y_a$;

Passo 3. Dado o corrente valor de $\hat{\delta}$ obtenha o valor atualizado de τ;

Passo 4. Faça iterações entre os passos 2 e 3 até a convergência. Na convergência os erros padrões de $\hat{\beta}$ e $(\hat{v} - v)$ podem ser computados da inversa da matriz de informação H^{-1} da verossimilhança H e os erros padrões de $\hat{\tau}$ são computados da matriz Hessiana, a qual contem a Informação de Fisher.

4.2.4. Otimização da análise estatística de variáveis continuas e discretas

Para as análises dos dois conjuntos de dados (eucaliptos e cafeeiros), o HGLMM utilizado, em sua forma geral foi dado por:

$y = X\beta + Zg + Wc + e$, (1) em que y é o vetor de dados fenotípicos, β vetor de efeitos fixos (média geral e repetições), g é o vetor dos efeitos aleatórios de progênies (conforme o conjunto de dados), c é o vetor dos efeitos aleatórios de ambiente comum da parcela (eucaliptos) ou blocos (cafeeiros) e e é o vetor de resíduos. X, Z e W são as matrizes de incidência para os efeitos β, g e c, respectivamente.

Partindo desse modelo, vários ajustes foram executados. A diferença entre os ajustes foram as distribuições assumidas para e, no caso, a Gama ou Normal para variáveis contínuas, e Poisson para variáveis categóricas associadas a contagem. Também, as distribuições de g e c variaram entre Gaussiana, Gama, Gama-Inversa e Beta.

Ademais, as funções de ligação foram atribuídas conforme Resende et al. (2018), que sugerem função identidade para distribuição Normal, função logarítmica para as distribuições Gama, Gama Inversa e Poisson.

Para a análise do conjunto de dados de café canéfora (análise conjunta de safras), o HGLMM utilizado foi dado por:

$y = X\beta + Zg + Wc + Tp + e$, (2) em que y é o vetor de dados fenotípicos, β vetor de efeitos fixos (média geral e repetições), g é o vetor dos efeitos aleatórios de progênies (conforme o conjunto de dados), c é o vetor dos efeitos de ambiente comum do bloco (aleatórios), p é o vetor dos efeitos aleatórios de ambiente permanente de indivíduo e e é o vetor de resíduos. X, Z, W e T são as matrizes de incidência para os efeitos β, g, c e p, respectivamente. e e é o vetor dos resíduos (aleatórios). X, Z, W e T são as matrizes de incidência para β, g, c e p, respectivamente.

Análises preliminares de LMM foram realizadas no *software* Selegen (Resende, 2007; 2016).

4.2.5. Seleção de modelos

Os HGLM são modelos condicionais e incluem modelos não lineares. Sob a suposição de normalidade, nos LMM, o estimador de Quadrados Mínimos Ponderados

(WLS) e o estimador REML, que funcionam sob os modelos condicional e marginal, são estimadores de máxima verossimilhança (Resende et al., 2018).

Para a comparação dos diferentes HGLMM, utilizou-se o cAIC – conditional *Akaike information criterion*. O cAIC mede a qualidade dos ajustes e permite a seleção de modelos; a herdabilidade e a acurácia medem a confiabilidade da seleção genética ou de genótipos (Resende e Duarte, 2007). Para uma comparação relativa, foi tomado o valor de cAIC do ajuste *Normal-Normal-Normal (N-N-N,* para genótipos, ambiente comum e resíduo, respectivamente*)* como referência, e desse valor foram subtraídos os valores de cAIC dos demais ajustes. Logicamente, a primeira diferença tem valor 0, pois, neste caso, é o valor de cAIC do ajuste *N-N-N* menos ele mesmo. No âmbito do melhoramento genético, as aplicações podem ser realizadas tanto no contexto fenotípico quanto gnômico (Azevedo et al., 2015; Resende Jr. et al., 2012).

Segundo Resende e Alves (2020), a verossimilhança hierárquica pode ser usada também na derivação de ferramentas para seleção de modelos como o Critério de informação de Akaike (AIC) condicional da verossimilhança hierárquica, pois esse é equivalente ao Critério de Informação da *Deviance* (DIC) aplicada na Estatística Bayesiana (Lee and Noh, 2012).

O critério de informação de Akaike (AIC) é vantajoso para a seleção de modelos por não se limitar à estrutura hierárquica (Resende e Alves, 2020). O AIC é dado por: $AIC = -2LogL + 2p$, em que $LogL$ é o logaritmo do máximo da função de verossimilhança e *p* é o número de parâmetros estimados. Conforme Resende e Alves (2020), o termo $-2LogL$ refere-se à qualidade do ajuste e $2p$ à penalização do modelo a medida em que p aumenta e o modelo fica mais parametrizado. Quanto menor o valor do AIC, melhor é o ajuste. Diferenças significativas entre AICs de dois modelos, devem ser pelo menos 2 unidades (Cavanaugh and Neath, 2019; Resende e Alves, 2020; 2022) ou 1 unidade (Sakamoto et al., 1986). Diferenças significativas entre BICs devem ser pelo menos 2 unidades (Neath and Cavanaugh, 2012; Resende e Alves, 2020; 2022) entre modelos. Ainda, sobre esse critério, Vrieze (2012) menciona que o AIC seleciona assintoticamente o modelo que minimiza o erro quadrático médio de predição e a função de risco ou perda quadrática.

4.2.6. Scripts do *Software* HGLM

O *software* HGLM (Ronnegard et al., 2010; Alam et al., 2014) em R foi usado. As rotinas computacionais desse *software* são apresentadas a seguir. No pacote HGLM, utilizando o default, a convergência é alcançada quando a mudança no *Log-likelihood* é menor que 10^{-6} de uma iteração a outra e a mudança na estimativa do parâmetro de variância individual é menor que 1%.

```
## Roteiro HGLMM
## Leitura da biblioteca
library(hglm)
## Leitura e preparação dos dados
dados <- read.table("")

## Histograma e Curva da f.d.p.
hist(DBH.Amb1,main = "DBH - Ambiente 1", xlab = "DBH",
ylab="Frequencia", prob=TRUE)
curve(dnorm(x, mean=mean(DBH.Amb1), sd=sd(DBH.Amb1)), add=TRUE)

## Avaliação de assimetria e curtose
library(moments)
agostino.test(dados$y)
anscombe.test(dados$y)

## ajuste do modelo
modelo = hglm2(y ~ 1 + fixo + (1|fator1)+ (1|fator2),
               family = gaussian(link = identity),
               rand.family = list(gaussian(link = identity), gaussian(link =
identity)),
               method = "EQL",
               calc.like = TRUE,conv = 1e-6, maxit = 1500,
               data = dados)
## Distribuições e suas respectivas funções de ligação
```

```
# gaussian(link = identity)
# Gamma(link = log)
# inverse.gamma(link=log)
# Beta(link = logit)
# binomial(link = logit)
# binomial(link = probit)
# poisson(link = log)
```

No pacote HGLM os HGLMM usam distribuições da variável resposta advindas da família exponencial (Gaussiana, Binomial, Poisson e Gama) e distribuições para os efeitos aleatórios advindos de distribuições bayesianas conjugadas. Para os demais fatores aleatórios do modelo é possível o ajuste das distribuições Gaussiana, Beta, Gama e Gama Inversa.

4. Resultados e Discussão

4.1. Testes de normalidade das variáveis continuas

Os resultados dos testes de normalidade são apresentados na Tabela 1.

Tabela 1. Resultados dos testes de normalidade para 9 variáveis de crescimento em Eucalipto.

Variáveis	Assimetria (A)	Classificação A	Curtose (C)	Classificação C
ALT2	-1.26	Negativa	5.73	Leptocúrtica
DAP2	-0.53	Negativa	3.48	Leptocúrtica
VOL2	0.00^{ns}	Normal	3.00^{ns}	Normal (Mesocurtica)
ALT5	-0.72	Negativa	5.52	Leptocúrtica
DAP5	0.26	Positiva	3.56	Leptocúrtica
VOL5	0.93	Positiva	5.02	Leptocúrtica
ALT7	-1.23	Negativa	4.45	Leptocúrtica
DAP7	-0.17	Negativa	2.73	Platicúrtica
VOL7	0.64	Positiva	3.46	Leptocúrtica

Assimetria Positiva = distribuição concentrada à esquerda; Assimetria Negativa = distribuição concentrada à direita; Leptocúrtica = distribuição pontiaguda; Platicúrtica = distribuição achatada; ^{ns}: não significativo à 5% de significância.

Dentre as 9 variáveis da Tabela 1, apenas uma (volume aos 2 anos) apresentou normalidade. Isto revela a importância dos HGLMM para flexibilizar a modelagem e a análise estatística e assim, possibilitar a consideração de distribuições de probabilidades mais realísticas.

4.2. Histogramas e funções densidade de probabilidade

As variáveis volume aos 7 anos (Figura 1) e volume aos 5 anos (Figura 2) apresentam distribuição proporcional a uma qui-quadrado. As Figuras 1, 2 e 3 apresenta os histogramas das distribuições aos 7, 5 e 2 anos de idade.

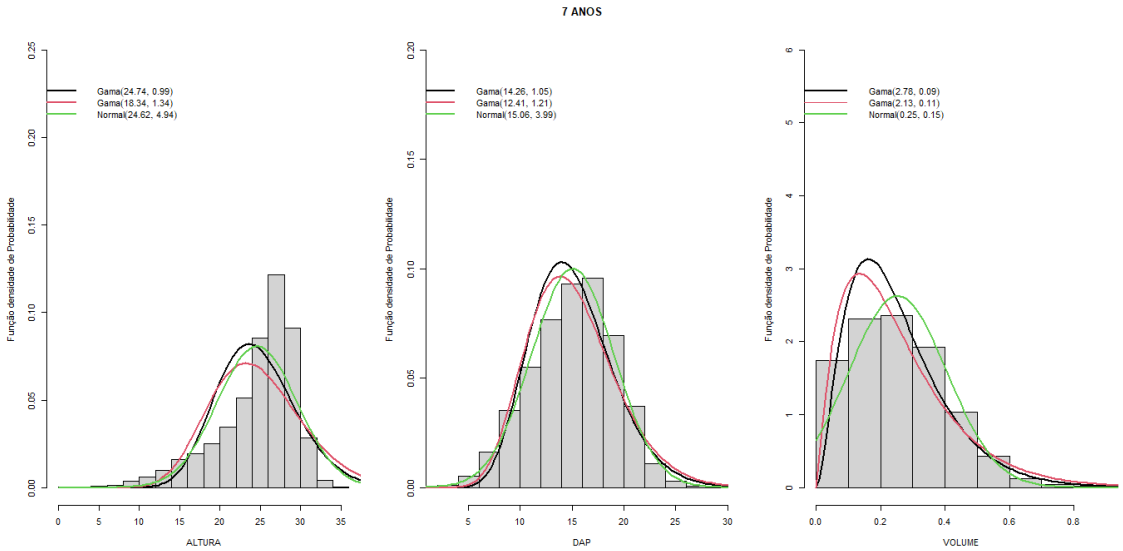

Figura 1. Histograma das características altura, DAP e volume em progênies de Eucalipto na idade de 7 anos e as curvas das funções densidades de probabilidade da distribuição normal com média (24,62) e desvio-padrão (4,95); normal com média (15,07) e desvio-padrão (3,99); normal com média (0,25) e desvio-padrão (0,15), respectivamente.

Verificam-se que as 3 distribuições se assemelham a: Logística para altura, Normal para diâmetro e Qui-quadrado para o volume. Estas 3 distribuições são casos particulares da Gama, que é uma distribuição ampla e generalizada.

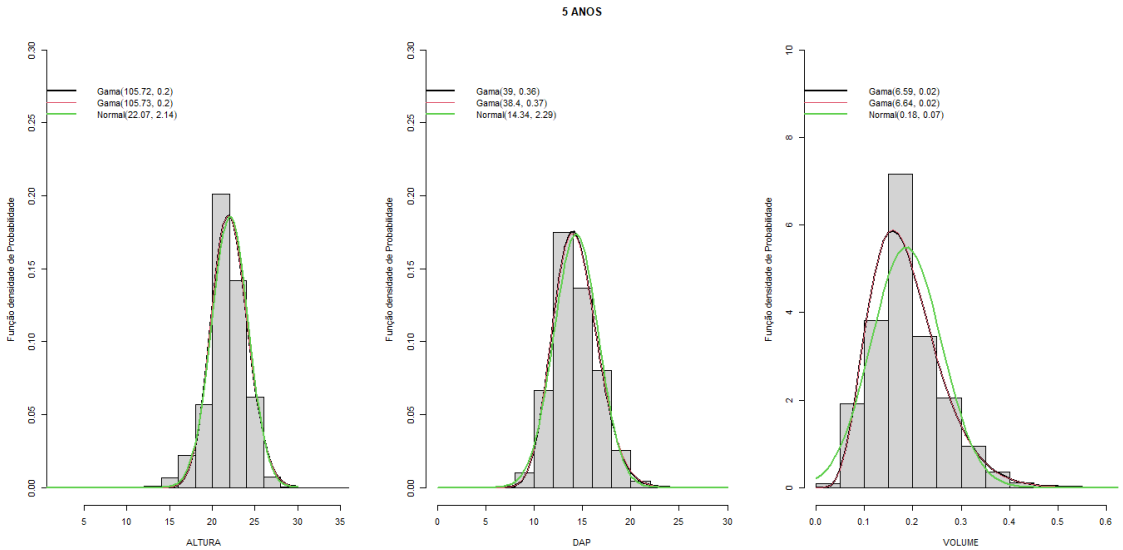

Figura 2. Histogramas das características altura, diâmetro (DAP) e volume em progênies de Eucalipto na idade de 5 anos e as curvas das funções densidades de probabilidade das distribuições Gama(105,72;0,2), Gama(105,73;0,2) e Normal(22,07; 2,14) para a variável altura; Gama(39;0,36), Gama(38,4;0,37) e

Normal(14,34;2,29) para DAP; Gama(6,59;0,02), Gama(6,64;0,02) e Normal(0,18; 0,07) para volume.

Pelos histogramas da Figura 2 verifica-se que as 3 distribuições se assemelham a: Logística para altura, Normal para diâmetro e Qui-quadrado para o volume, aos 5 anos.

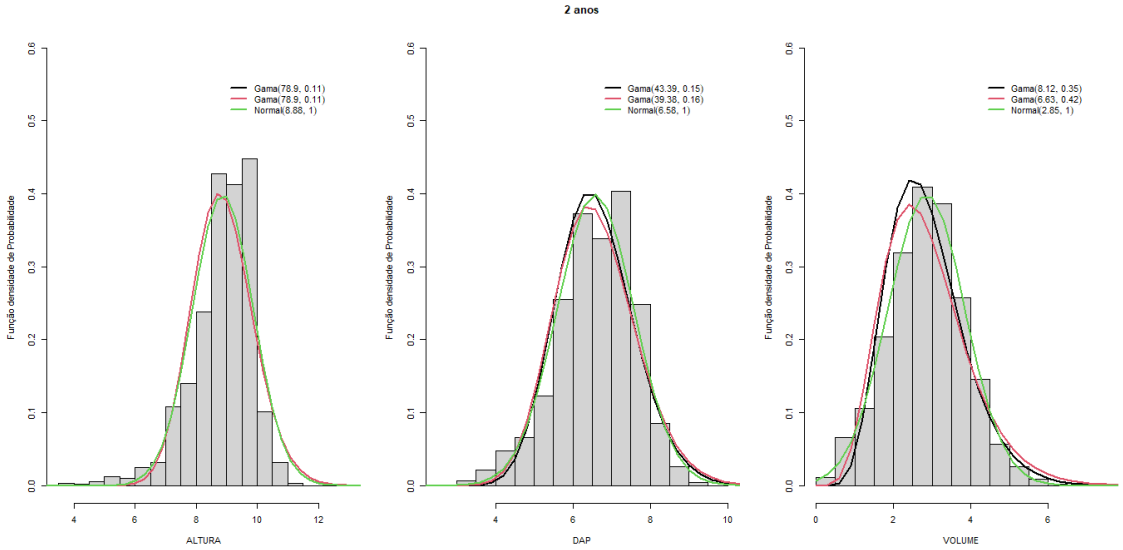

Figura 3. Histogramas das características altura, diâmetro (DAP) e volume em progênies de Eucalipto na idade de 5 anos e as curvas das funções densidades de probabilidade das distribuições Gama(78,9;0,11), Gama(78,9;0,11) e Normal(8,88; 1) para a variável altura; Gama(43,39;0,15), Gama(39,38;0,16) e Normal(6,58;1) para DAP; Gama(8,12;0,35), Gama(6,63;0,42) e Normal(2,85; 1) para volume.

Verificam-se que as 3 distribuições se assemelham a: Logística para altura, Logística para diâmetro e Normal para o volume, aos 2 anos.

4.3. Ajustes das distribuições de probabilidade

A Tabela 2 mostra os ajustes de modelos para as variáveis altura, diâmetro e volume medidos aos 7 anos de idade. Nela encontram-se os valores de cAIC e herdabilidade obtidos para cada variável. Neste cenário, o melhor ajuste foi o modelo Gama (*G-G-G*) em termos de valores menores de cAIC e maiores valores de herdabilidade, para os três (família, parcela e erro) fatores de efeitos aleatórios.

Tabela 2. Informações sobre a distribuição e a função de ligação para cada fator aleatório do modelo para as características volume (V7), diâmetro (D7) e altura (A7)

em eucalipto na idade de 7 anos. Resumo das distribuições adotadas nos fatores aleatórios do modelo. cAIC - *conditional Akaike criterium*, diferença entre os AIC de cada ajuste e AIC do ajuste com distribuição normal para todos os efeitos aleatórios (*N-N-N*). Valores de herdabilidade conforme os componentes estimados em cada uma das combinações de distribuição.

Variá-vel	Fam distr	Fam link	Parc distr	P. link	Erro Distr.	E. Lnk	Resu mo Distr.	AIC	Dif	h^2
V7 Normal	normal	id	Normal	Id	normal	Id	N-N-N	-11505	0	0.09
V7 Normal	gama	log	Gama	Log	normal	Id	G-G-N	-11503	1	0.09
V7 Normal	G-Inv	log	G-Inv	Log	normal	Id	GI-GI-N	-11506	-1	0.09
V7 Normal	gama	log	normal	Id	normal	Id	G-N-N	-11504	1	0.09
V7 Gama	gama	log	gama	Log	gama	Log	G-G-G	-12071	-566	0.10
V7 Gama	G-Inv	log	G-Inv	Log	gama	Log	GI-GI-G	-12062	-557	0.10
V7 Gama	normal	id	normal	Id	gama	Log	N-N-G	-12065	-560	0.10
V7 Gama	G-Inv	log	normal	Id	gama	Log	GI-N-G	-12062	-557	0.10
V7 Gama	G-Inv	log	gama	Log	gama	Log	GI-G-G	-12062	-557	0.10
D7 Normal	normal	id	Normal	Id	normal	Id	N-N-N	63522	0	0.08
D7 Normal	gama	log	Gama	Log	normal	Id	G-G-N	63547	26	0.07
D7 Normal	G-Inv	log	G-Inv	Log	normal	Id	GI-GI-N	63538	16	0.00
D7 Normal	gama	log	Normal	Id	normal	Id	G-N-N	63548	26	0.07
D7 Gama	gama	log	Gama	Log	gama	Log	G-G-G	64696	1174	0.11
D7 Gama	G-Inv	log	G-Inv	Log	gama	Log	GI-GI-G	64697	1175	0.11
D7 Gama	normal	id	Normal	Id	gama	Log	N-N-G	64696	1174	0.12
D7 Gama	G-Inv	log	Normal	Id	gama	Log	GI-N-G	64697	1176	0.11
D7 Gama	G-Inv	log	Gama	Log	gama	Log	GI-G-G	64696	1174	0.11
A7 Normal	Norml	Id	normal	Id	normal	Id	N-N-N	68436	0	0.07
A7 Normal	Gama	log	gama	Log	normal	Id	G-G-N	68351	-85	0.05

A7 Normal	G-Inv	log	G-Inv	Log	normal	Id	GI-GI-N	68488	53	0.00
A7 Normal	Gama	log	normal	Id	normal	Id	G-N-N	68425	-11	0.05
A7 Gama	Gama	log	gama	Log	gama	Log	G-G-G	71875	3439	0.09
A7 Gama	G-Inv	log	G-Inv	Log	gama	Log	GI-GI-G	71879	3443	0.09
A7 Gama	Norml	Id	normal	Id	gama	Log	N-N-G	71877	3441	0.09
A7 Gama	G-Inv	log	normal	Id	gama	Log	GI-N-G	71880	3444	0.09
A7 Gama	G-Inv	log	gama	Log	gama	Log	GI-G-G	71878	3442	0.09

Resumo Dist. - Resumo das distribuições atribuídas a cada fator aleatório do modelo utilizando letras para cada distribuição (*N* - normal, *G* - gama e *GI* - gama inversa) na ordem dos fatores progênie, parcela e erro, respectivamente. Dif - é a diferença entre os AIC do ajuste *N-N-N* (conforme o resumo mencionado acima e o AIC dos demais ajustes.

A Tabela 3 mostra os ajustes de modelos para as variáveis altura, diâmetro e volume medidos aos 5 anos de idade.

Tabela 3. Informações sobre a distribuição e a função de ligação para cada fator aleatório do modelo para as características volume (V7), diâmetro (D7) e altura (A7) em eucalipto na idade de 5 anos. Resumo das distribuições adotadas nos fatores aleatórios do modelo. cAIC – *conditional Akaike criterium*, Diferença entre os AIC de cada ajuste e AIC do ajuste *N-N-N*. Valor de herdabilidade conforme os componentes estimados em cada uma das combinações de distribuição.

Variá-vel	Fam. Distr.	F. link	Parc. Distr.	P. link	Erro distr.	E. link	Resu-mo distr.	AIC	Dif.	h^2
A5 Normal	Normal	Id	Normal	id	normal	id	N-N-N	49577	0	0.09
A5 Normal	Gama	log	Gama	log	normal	id	G-G-N	49539	-38	0.08
A5 Normal	G-Inv	log	G-Inv	log	normal	id	GI-GI-N	49599	22	0.00
A5 Normal	Gama	log	Normal	id	normal	id	G-N-N	49580	3	0.08
A5 Gama	Gama	log	Gama	log	gama	log	G-G-G	50852	1275	0.13
A5 Gama	G-Inv	log	G-Inv	log	gama	log	GI-GI-G	50862	1285	0.13

A5 Gama	Normal	id	Normal	id	gama	log	N-N-G	50857	1280	0.13
A5 Gama	G-Inv	log	Normal	id	gama	log	GI-N-G	50857	1280	0.13
A5 Gama	G-Inv	log	Gama	log	gama	log	GI-G-G	50867	1290	0,08
D5 Normal	Norm.	id	Norm.	id	Normal	id	N-N-N	51783	0	0,10
D5 Normal	Gama	log	gama	log	Normal	id	G-G-N	51821	38	0,09
D5 Normal	G-Inv	log	G-Inv	log	Normal	id	GI-GI-N	51751	-32	0,08
D5 Normal	Gama	log	Norm.	id	Normal	id	G-N-N	51816	33	0,09
D5 Gama	Gama	log	gama	log	gama	log	G-G-G	51864	81	0.15
D5 Gama	G-Inv	log	G-Inv	log	gama	log	GI-GI-G	51862	79	0.15
D5 Gama	Norm.	id	Norm.	id	gama	log	N-N-G	51863	80	0.15
D5 Gama	G-Inv	log	Norm.	id	gama	log	GI-N-G	51862	79	0.15
D5 Gama	G-Inv	log	Gama	log	gama	log	GI-G-G	51863	80	0.15
V5 Normal	Normal	id	normal	id	normal	Id	N-N-N	-29576	0	0.11
V5 Normal	Gama	log	gama	log	normal	Id	G-G-N	-29574	1	0.11
V5 Normal	G-Inv	log	G-Inv	log	normal	Id	GI-GI-N	-29577	-1	0.11
V5 Normal	Gama	log	normal	id	normal	Id	G-N-N	-29575	1	0.11
V5 Gama	Gama	log	gama	log	gama	Log	G-G-G	-30732	-1156	0.17
V5 Gama	G-Inv	log	G-Inv	log	gama	Log	GI-GI-G	-30737	-1161	0.17
V5 Gama	Normal	Id	normal	id	gama	Log	N-N-G	-30735	-1159	0.17
V5 Gama	G-Inv	log	normal	id	gama	Log	GI-N-G	-30737	-1161	0.17
V5 Gama	G-Inv	log	gama	log	gama	Log	GI-G-G	-30738	-1162	0.17

Resumo Dist. - Resumo das distribuições atribuídas a cada fator aleatório do modelo utilizando letras para cada distribuição (*N* - normal, *G* - gama e *GI* - gama inversa) na ordem dos fatores progênie, parcela e erro, respectivamente.

Dif - é a diferença entre os AIC do ajuste *N-N-N* (conforme o resumo mencionado acima e o AIC dos demais ajustes.

Nesse cenário envolvendo o volume aos 5 anos de idade (Tabela 3), os modelos com distribuição do erro gama tiveram vantagem em termos do AIC, o que era esperado, pois a distribuição qui-quadrado (mais plausível para volume aos 5 anos) é um caso particular da distribuição gama. Concomitantemente, os valores de herdabilidade para esses modelos com distribuição do erro gama foram maiores que os modelos cujo o erro tinha distribuição normal.

A Tabela 4 mostra os ajustes de modelos para as variáveis altura, diâmetro e volume medidos aos 2 anos de idade.

Tabela 4. Informações sobre a distribuição e a função de ligação para cada fator aleatório do modelo para as características volume (V7), diâmetro (D7) e altura (A7) em eucalipto na idade de 2 anos. Resumo das distribuições adotadas nos fatores aleatórios do modelo. cAIC- *condicional Akaike criterium*, diferença entre os AIC de cada ajuste e AIC do ajuste *N-N-N*. Valor de herdabilidade conforme os componentes estimados em cada uma das combinações de distribuição.

Alt-2 anos	Fam. Distr.	F. link	Parc. Distr.	Parc. Link	E. Distr..	E. Link	Resu -mo Dist.	AIC	Dif.	h^2
A2 Normal	normal	Id	Normal	id	normal	id	N-N-N	16144	0	0.14
A2 Normal	gama	Log	Gama	log	normal	id	G-G-N	16121	-23	0.13
A2 Normal	G-Inv	Log	G-Inv	log	normal	id	GI-GI-N	16160	16	0.14
A2 Normal	gama	Log	Normal	id	normal	id	G-N-N	16139	-5	0.14
A2 Gama	gama	Log	Gama	log	gama	log	G-G-G	17187	1042	0.15
A2 Gama	G-Inv	Log	G-Inv	log	gama	log	GI-GI-G	17193	1049	0.14
A2 Gama	normal	Id	Normal	id	gama	log	N-N-G	17190	1046	0.14
A2 Gama	G-Inv	Log	Normal	id	gama	log	GI-N-G	17191	1047	0.14
A2 Gama	G-Inv	Log	Gama	log	gama	log	GI-G-G	17196	1051	0.14
D2 Normal	normal	id	Normal	id	normal	id	N-N-N	16762	0	0.09
D2 Normal	gama	log	Gama	log	normal	id	G-G-N	16759	-4	0.09

D2 Normal	G-Inv	log	G-Inv	log	normal	id	GI-GI-N	16764	2	0.09
D2 Normal	gama	log	Normal	id	normal	id	G-N-N	16761	-1	0.09
D2 Gama	gama	log	Gama	log	gama	log	G-G-G	17366	603	0.14
D2 Gama	G-Inv	log	G-Inv	log	gama	log	GI-GI-G	17368	605	0.14
D2 Gama	normal	id	Normal	id	gama	log	N-N-G	17367	604	0.14
D2 Gama	G-Inv	log	Normal	id	gama	log	GI-N-G	17367	605	0.14
D2 Gama	G-Inv	log	Gama	log	gama	log	GI-G-G	17368	606	0.14
V2 Normal	normal	Id	Normal	id	normal	id	N-N-N	16587	0	0.11
V2 Normal	gama	log	Gama	log	normal	id	G-G-N	16591	4	0.11
V2 Normal	G-Inv	log	G-Inv	log	normal	id	GI-GI-N	16579	-8	0.11
V2 Normal	gama	log	Normal	id	normal	id	G-N-N	16588	2	0.11
V2 Gama	gama	log	Gama	log	gama	log	G-G-G	17449	863	0.16
V2 Gama	G-Inv	log	G-Inv	log	gama	log	GI-GI-G	17454	867	0.16
V2 Gama	normal	Id	Normal	id	gama	log	N-N-G	17452	865	0.16
V2 Gama	G-Inv	log	Normal	id	gama	log	GI-N-G	17454	867	0.16
V2 Gama	G-Inv	log	Gama	log	gama	log	GI-G-G	17454	868	0.16

Resumo Dist. - Resumo das distribuições atribuídas a cada fator aleatório do modelo utilizando letras para cada distribuição (*N* - normal, *G*- gama e *GI* - gama inversa) na ordem dos fatores progênie, parcela e erro, respectivamente. “Dif” - é a diferença entre os AIC do ajuste *N-N-N* (conforme o resumo mencionado acima e o AIC dos demais ajustes).

4.4. Resultados para variáveis categóricas

A seguir são apresentados os resultados das análises do conjunto de dados referente à Incidência de Ferrugem (IR) em café canéfora nos anos de 2013 a 2015. A Figura 4 contém os histogramas dos registros de incidência a ferrugem.

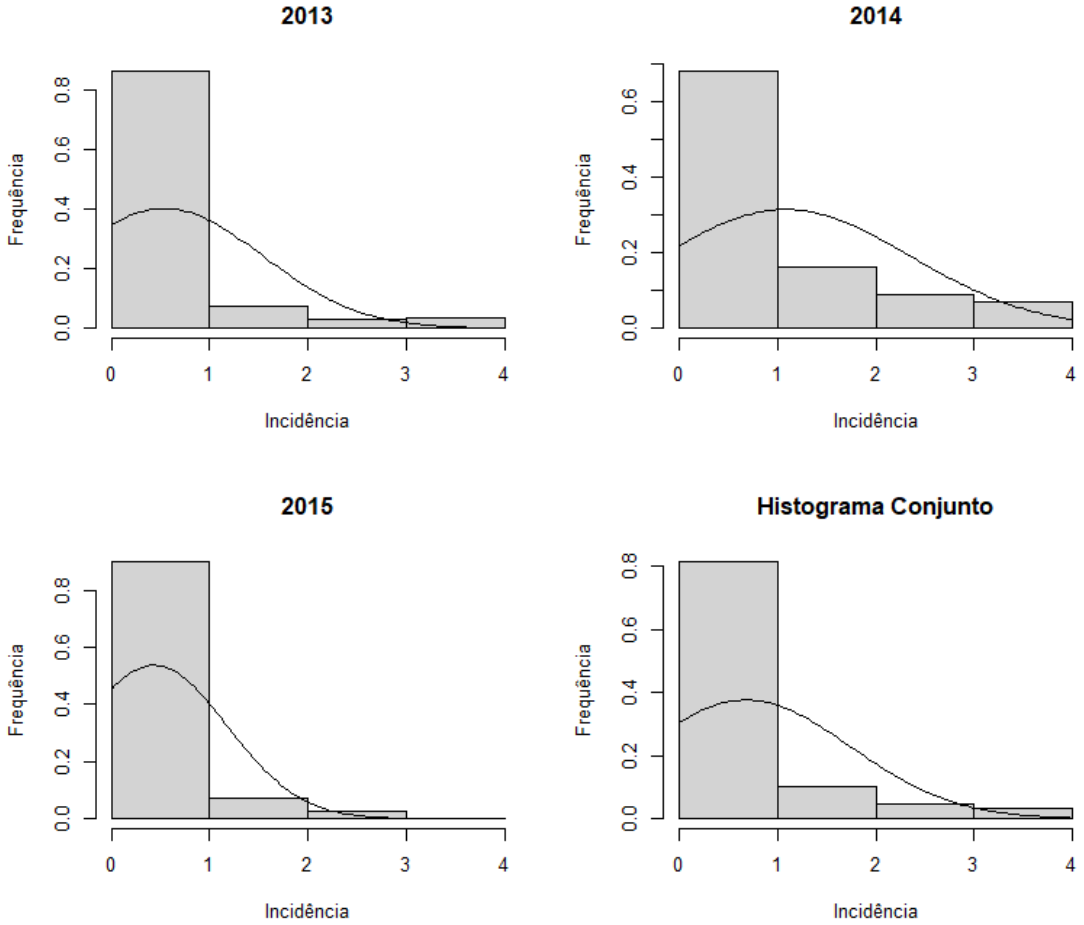

Figura 4. Histogramas da variável incidência de ferrugem medidas em escalas de 0 a 4 e a curva da distribuição normal com média e variância de cada fenótipo em cada safra (anos de 2013, 2014 e 2015) e histograma conjunto do fenótipo nas safras dos três anos conjuntamente.

Na Tabela 5 são apresentados os resultados para essa característica avaliada em todos os anos.

Tabela 5. Informações sobre a distribuição e a função de ligação para cada fator aleatório do modelo para a característica de incidência de ferrugem (*incidence of rust* - R) com 5 níveis (0 a 4) de infecção em café avaliados nos anos de 2013 (R1), 2014 (R2), 2015 (R3) e todos (R4). Resumo das distribuições adotadas nos fatores aleatórios (N = distribuição normal, P= distribuição *poisson*, B = distribuição beta). AIC – conditional *Akaike criterium* e herdabilidade (h^2).

R	Fam. Dist.	Fam. Link	Bloco Dist.	Blo-co link	E. Dist.	E. Link	Resu-mo Dist.	AIC	h^2

R1 Pois.	Normal	Id	Normal	id	Normal	id	N-N-N	629	0.16
R1 Pois.	Normal	Id	Normal	id	Poisson	log	*N-N-P*	*422*	*0.40*
R1 Pois.	Gama	Log	Gama	log	Poisson	log	*G-G-P*	*423*	*0.37*
R1 Pois.	G-Inv	Log	G-Inv	log	Poisson	log	*GI-GI-P*	*422*	*0.40*
R1 Pois.	Beta	Logito	Beta	Logito	Poisson	log	B-B-P	*422*	0.14
R2 Pois.	Normal	id	Normal	Id	Normal	id	N-N-N	679	0.28
R2 Pois.	Normal	id	Normal	Id	Poisson	log	*N-N-P*	*549*	*0.53*
R2 Pois.	Gama	log	Gama	Log	Poisson	log	*G-G-P*	*551*	*0.50*
R2 Pois.	G-Inv	log	G-Inv	Log	Poisson	log	*GI-GI-P*	*548*	*0.49*
R2 Pois.	Beta	Logito	Beta	Logito	Poisson	log	B-B-P	*549*	0.21
R3 Pois.	Normal	Id	Normal	Id	Normal	id	N-N-N	499	0.11
R3 Pois.	Normal	Id	Normal	Id	Poisson	log	*N-N-P*	*402*	*0.23*
R3 Pois.	Gama	Log	Gama	Log	Poisson	log	*G-G-P*	*403*	*0.22*
R3 Pois.	G-Inv	Log	G-Inv	Log	Poisson	log	*GI-GI-P*	*402*	*0.21*
R3 Pois.	Beta	Logito	Beta	Logito	Poisson	log	B-B-P	*403*	0.07

IR Mult.	Fam. Dist.	Fam. link	Bloco Dist.	Bloco link	Perm. Dist	Perm. Link	E. Dist.	E. link	Resumo Dist	AIC	h^2
Pois.	Norm.	id	Norm.	id	Norm.	id	Norm.	id	N-N-N	2807	0.16
Pois.	Norm.	id	Norm.	id	Norm.	id	Pois.	log	*N-N-P*	*2263*	0.45
Pois.	Gama	log	Gama	log	Gama	log	Pois.	log	*G-G-P*	*2269*	0.44
Pois.	G-inv	log	G-inv	log	G-inv	log	Pois.	log	*GI-GI-P*	*2261*	0.41
Pois.	Beta	logit	Beta	logito	Beta	logito	Pois.	log	B-B-P	*2265*	0.22

Nesse cenário, os modelos com erros Poisson foram melhores como esperado, pois os dados são de contagem. Dentre os modelos com erro Poisson, não houveram grandes diferenças em termos de cAIC. Já em relação aos valores de herdabilidades ajustadas conforme as distribuições, o modelo B-B-P (conhecido como Poisson-Beta) teve o menor valor. O N-N-P (GLMM Poisson), o G-G-P (HGLMM Poisson-Gama) e o GI-GI-P (HGLMM) obtiveram os maiores valores de herdabilidade, corroborando com

a justificativa da utilização do HGLMMM como maneira alternativa aos modelos tradicionalmente conhecidos, porém, menos flexíveis que o HGLMM. Para os demais fatores de efeitos aleatórios (genótipos e blocos), as distribuições normal, gama e gama inversa foram similares.

5. Conclusões

A abordagem HGLMM mostrou-se efetiva em modelar os diferentes fatores de efeitos aleatórios, com variadas distribuições;

A modelagem Gama mostrou-se superior a Normal em algumas situações.

O volume de madeira é melhor modelado por uma distribuição Qui-quadrado, em algumas situações.

Para as variáveis categóricas, os dados do cafeeiro arábica revelaram superioridade da distribuição Poisson para os erros, mostrando eficiência do sistema Poisson-Gama e Poisson-Normal, sendo a Gama e a Normal as distribuições mais adequadas aos demais fatores de efeitos aleatórios.

Referências

ALKIMIM, E. R. et al. Designing the best breeding strategy for Coffea canephora: genetic evaluation of pure and hybrid individuals aiming to select for productivity and disease resistance traits. **PloS one**, v. 16, n. 12, p. e0260997, 2021.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, 2015

AYDIN, D; ŞENOĞLU, B. Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. **Journal of Modern Applied Statistical Methods**, v. 14, n. 2, p. 12, 2015.

BARBOSA, M. H. P.; RESENDE, M. D. V.; BRESSIANI, J. A.; SILVEIRA, L. C. L. Selection of sugarcane families and parents by Reml/Blup. **Crop Breeding and Applied Biotechnology**, v. 5, p. 443-450, 2005.

BRESLOW, N.E.; CLAYTON, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. **Journal of the American Statistical Association**, 88: 9-25.

BISHOP, C. M. **Pattern recognition and machine learning**. Springer, 2006.

BULMER, M. G. **The mathematical theory of quantitative genetics**. Oxford: Charedon Press, 1980. 254 p.

CASELLA, G; BERGER, R.L. **Statistical Inference**. Second Edition. 2006. 668 p.
CAVANAUGH, J. E.; NEATH, A. A. 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **Computational Statistics**, 11:2-11.
CHISSOM, B.S. Interpretation of the kurtosis statistic. **The American Statistician**, v. 24, n. 4, p. 19-22, 1970.

D'AGOSTINO, R B. Transformation to normality of the null distribution of g1. **Biometrika**, p. 679-681, 1970.

FISHER, R. A.; MACKENZIE, K. Studies in crop variation II: The manorial response of different potato varieties. **Journal of Agricultural Science**, 1923.
HENDERSON, C.R. (1975). Best linear estimation and prediction under a selection model. **Biometrics**, 31: 423-447.

HENDERSON, C.R.; KEMPTHORNE, O.; Searle, S.R.; Von Krosigk, C.M. (1959). The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, 15: 192-218.

LEE, Y.; NELDER, J.A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society:** Series B (Methodological), v. 58, n. 4, p. 619-656, 1996.

LEE, Y.; NELDER, J.A. (2001). Hierarchical generalized linear models: A synthesis of generalised linear models, random effect models and structured dispersions. **Biometrika**, 88: 987-1006.

LEE, Y.; NELDER, J. A. Double hierarchical generalized linear models (with discussion). **Journal of the Royal Statistical Society:** Series C (Applied Statistics), v. 55, n. 2, p. 139-185, 2006.

LEE, Y.; HA, I.D. (2010). Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. **Statistics and Computing**, 20: 295-303.

LEE, Y.; KIM, G. (2016). H-likelihood predictive intervals for unobservables. **International Statistical Review**, 84: 487-505.

LEE, Y.; RÖNNEGÅRD, L.; NOH, M. **Data analysis using hierarchical generalized linear models with R**. Chapman and Hall/CRC, 2017.

LINDLEY, D.V.; SMITH, A.F. (1972). Bayes estimates for the linear model**. Journal of the Royal Statistical Society**: Series B (Methodological), 34: 1-41.

MA, R.; JORGENSEN, B. (2007). Nested generalized linear mixed models: Orthodox best linear unbiased predictor approach. **Journal of the Royal Statistical Society**: Series B, 69: 625–641.

MCCULLAGH, P. AND NELDER, J. A. (1989). **Generalized Linear Models**. Chapman & Hall, London.

MIRZAEI et al., 2016. Modeling frequency distributions of tree height, diameter and crown area by six probability functions for open forests of Quercus persica in Iran. **Journal of Forestry Research.**

NEATH, A. A.; CAVANAUGH, J. E. 2012. The Bayesian information criterion: background, derivation, and applications. **Computational Statistics**, 4:199-203.

NELDER, J.A.; PREGIBON, D. (1987). An extended quasi-likelihood function. Biometrika, 74: 221-232.

NELDER, J.A.; WEDDERBURN, R.W.M. (1972). Generalized linear models. **Journal of the Royal Statistical Society**, 135: 370-384.

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

PEDROZO, C. A.; BENITES, F. R. G.; BARBOSA, M. H. P.; RESENDE, M. D. V.; SILVA, F. L. Eficiência de índices de seleção utilizando a metodologia REML/BLUP no melhoramento da cana-de-açúcar. **Scientia Agraria**, v.10, n.1, 31-36, 2009.

RESENDE MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informações Tecnológicas, Brasília, 975p.

RESENDE MDV (2007) **Matemática e estatística na análise de experimentos e no melhoramento genético**, Embrapa Florestas, Colombo, 490p.

RESENDE MDV; SILVA, FF; AZEVEDO, CF (2014) **Estatística matemática, biométrica e computacional**. Suprema, Visconde do Rio Branco, 882p.

RESENDE MDV (2015) **Genética quantitativa e de populações**. Visconde do Rio Branco, 463p.

RESENDE, M. D. V. **Melhoramento de essências florestais**. In: Borem, A. **Melhoramento de espécies cultivadas**. Viçosa, MG: UFV, 589-647, 1999.

RESENDE, M. D. V.; BARBOSA, M. H. P. **Melhoramento genético de plantas de propagação assexuada**. Embrapa Florestas.

RESENDE, M. D. V., BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística**, v.20, p.30-65, 2002.

RESENDE MDV; DUARTE JB (2007). Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, 37: 182-194.

RESENDE, M.D.V. (2016). *Software* Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology** 16: 330-339.

RESENDE, M.D.V.; AZEVEDO, C.F.; SILVA, F.F.; NASCIMENTO, M.; GOIS, I.B.; ALVES, R.S. **Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP**.1. ed. Visconde do Rio Branco: Suprema, v.1, p.151, 2018.

RESENDE MDV; ALVES RS (2020). Linear, generalized, hierarchical, Bayesian and random regression mixed models in genetics/genomics in plant breeding. **Functional Plant Breeding Journal**.

RESENDE MDV; ALVES RS (2022). Statistical significance, selection accuracy and experimental precision in plant breeding. **Crop Breeding and Applied Biotechnology**.

RESENDE JR., M.F.R. ; VALLE, P.R.M. ; RESENDE, M. D. V. ; GARRICK, D. J. ; FERNANDO, R. L. ; DAVIS, J.M. ; JOKELA, E. J. ; MARTIN, T. A. ; PETER, G. F. ; KIRST, M. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, v.190, p.1503 - 1510, 2012a.

RONNEGARD, L., SHEN, X. and ALAM, M. (2010) hglm: A Package for Fitting Hierarchical Generalized Linear Models. **The R Journal**, 2(2): 20-28.

ROSS, Sheldon. **A first course in probability**. Pearson, 2014.

SAKAMOTO, Y.; ISHIGURO, M.; KITAGAWA, G. **Akaike information criterion statistics.** KTK, Tokyo, 1986.

SCHALL, H. (1991). Estimation in generalized linear models with random effects. **Biometrika**, 78: 719–727.

STROUP, W. W. (2013). **Generalized linear mixed models**. Boca Raton: CRC Press.
THOM, Herbert CS. A note on the gamma distribution. **Monthly weather review**, v. 86, n. 4, p. 117-122, 1958.

THOMPSON, R. (1973). The estimation of variance and covariance components when records are subject to culling. **Biometrics**, 29: 527-550.

THOMPSON, R.; BAKER, R.J. (1981). Composite link functions in generalized linear models. **Applied Statistics**, 30: 125-131.

VRIEZE, S. I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). **Psychological methods**, v. 17, n. 2, p. 228, 2012.

WOLFINGER, R; & O'CONNELL, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. **Journal of Statistical Computation and Simulation**, 48, 233-243.

CAPÍTULO 3

STATISTICAL GENETICS ANALYSES OF GAMMA DISTRIBUTED DATA VIA HGLMM FITTED BY HIML/HG-BLUP

Resumo

Os Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM) abrangem os Modelos Lineares Mistos (LMM, que se baseiam na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo) e os Modelos Lineares Generalizados Mistos (GLMM, que se baseiam na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo, exceto os erros, que podem assumir outras distribuições) e também a estimação Bayesiana. Uma das maiores vantagens dos HGLMM é a possibilidade de atribuir várias outras distribuições (além da Normal) de probabilidade pertencentes a família exponencial aos demais fatores de efeitos aleatórios do modelo, além dos erros, que já são flexibilizados via GLMM. HGLMMs com distribuição Gamma tem apresentado melhores ajustes em vários cenários. Para características *contínuas* foi competitivo com o Modelo Linear Misto na estimativa de componentes de variância e em ajuste de acordo com os valores cAIC. A abordagem HGLMM mostrou-se efetiva em modelar efeitos aleatórios com variadas distribuições; A modelagem Gama mostrou-se superior a Normal em várias algumas situações. O volume de madeira é melhor modelado por uma distribuição Qui-quadrado, em algumas situações. Pesos para escalar as distribuições foram derivados e avaliados, sendo uteis e necessários para a abordagem HGLMM. Para as variáveis categóricas, os dados do cafeeiro arábica revelaram superioridade da distribuição Poisson para os erros, mostrando eficiência do sistema Poisson-Gama e Poisson-Normal, sendo a Gama e a Normal as distribuições mais adequadas aos demais fatores de efeitos aleatórios.

Palavras-chave: Modelos Lineares generalizados, Modelos hierárquicos, Seleção de modelos, Inferência estatística, Verossimilhança Hierárquica. Quadrados mínimos iterativos ponderados.

1. Introdução

Os Modelos Lineares Generalizados Hierárquicos Mistos (HGLMM) abrangem os Modelos Lineares Mistos (LMM, que baseiam-se na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo) e os Modelos Lineares Generalizados Mistos (GLMM, que baseiam-se na normalidade das distribuições de todos os fatores de efeitos aleatórios do modelo, exceto os erros, que podem assumir outras distribuições) e também a estimação Bayesiana (Lee e Nelder, 1996; Resende et al., 2018; Resende e Alves, 2020). Uma das maiores vantagens dos HGLMM é a possibilidade de atribuir várias outras distribuições (além da Normal) de probabilidade pertencentes a família exponencial aos demais fatores de efeitos aleatórios do modelo, além dos erros, que já são flexibilizados via GLMM. HGLMMs com distribuição Gamma tem apresentado melhores ajustes em vários cenários. Para características *contínuas* foi competitivo com o Modelo Linear Misto na estimativa de componentes de variância e em ajuste de acordo com os valores cAIC (Resende et al., 2023).

Com o advento dos HGLMM, o uso da distribuição de probabilidade Gama para a modelagem e análise genética está ganhando voga e se tornando importante em disputa com a distribuição Normal. A análise genética via distribuição Gama não é corriqueira e universal como a Normal. Assim, temas como a estimação da variância residual e da herdabilidade, o uso das equações de Modelo Misto (MME) com resíduos Gama, a predição BLUP em HGLMM e GLMM sob uma distribuição de probabilidade Gama, são relevantes.

No presente caso, no modelo para Y (variável fenotípica com distribuição Gama), os erros seguem a distribuição Gama. No caso de testes de progênies com várias plantas por parcela, esses erros estão associados a variação residual dentro de parcela. Essa variação residual (S^2) equivale a variação fenotípica dentro de parcela (combinação progênie - bloco), a qual tem distribuição qui-quadrado (X^2_{n-1}) com n-1 graus de liberdade, em que n é o número de plantas por parcela, corrigido pela sobrevivência do caráter no experimento. Assim, $S^2 \sim X^2_{n-1}$. Tem-se também que Ln (S^2) ~ X^2_{n-1} (Resende, 2007). Assim, o tratamento das distribuições Gama e Qui-quadrado torna-se cada vez mais importante na análise estatística em genética e melhoramento.

Devido à escassez de informações sobre a análise genética com o uso da distribuição Gama, o presente artigo visa desenvolver, apresentar e difundir alguns tópicos relevantes na *modelagem HGLMM* de variáveis continuas, a qual propicia a *avaliação genética Gama*. Assim, os objetivos deste trabalho foram: flexibilizar a definição das distribuições dos fatores aleatórios de vários LMM e GLMM; verificar as possibilidades de um melhor ajuste para variáveis de crescimento de plantas ao considerar a variável resposta (*y*) como tendo Distribuição Gama (aproximação às distribuições Weibull e Qui-quadrado e outras); ajustar e comparar HGLMMs para variáveis contínuas e discretas; derivar estimadores para a h^2 sob distribuição Gama.

2. Metodologia

2.1 HGLMM, HIML, HG-BLUP e IWLS

Os modelos lineares generalizados hierárquicos mistos (HGLMM) foram ajustados via máxima verossimilhança hierárquica (HIML) e HG-BLUP pelo algoritmo dos quadrados mínimos ponderados iterativos (IWLS). Para a predição e a estimação via HIML/HG-BLUP os seguintes passos são necessários: definição do modelo estatístico, construção da função de verossimilhança hierárquica, derivação das equações de Modelos Mistos Hierárquicos, derivação dos estimadores de componente de variância via verossimilhança perfilada ajustada, desenvolvimento de um algoritmo (IWLS) iterativo eficiente e implementação computacional eficaz.

Para a estimação nessa nova classe de modelos, a ideia da Verossimilhança Estendida ou Hierárquica foi introduzida como critério a ser maximizado. Esse método pode então ser denominado **HIML**, embora não tenha sido assim denominada por Lee e Nelder. O método permite um algoritmo (do tipo IWLS) simples expresso na forma de GLMM interconectados.

Segundo Resende et al. (2018), na verossimilhança hierárquica (HL) as estimativas dos parâmetros de dispersão são determinadas pela maximização da Verossimilhança Hierárquica Perfilada Ajustada e no caso de modelos lineares mistos a Verossimilhança Perfilada Ajustada Maximizada é exatamente igual ao REML, podendo ser denominada HIML. Uma vantagem do HIML é a possibilidade de inferência para parâmetros fixos, aleatórios e variáveis não observadas, por possuir o conceito de probabilidade preditiva. Esse conceito é facilmente aceito por bayesianos e frequentistas (Lee e Kim, 2016), por permitir interpretações para os intervalos de credibilidade e de confiança, respectivamente.

2.2 Equações de HGLMM e algoritmo para HG-BLUP e HIML

Para predição e estimação via HIML/HG-BLUP, são necessários os passos a seguir.

a. Definição do modelo estatístico: GLMM: Normal-Normal.

$y = X\beta + Zv + e$, em que: $Var(v) = D\sigma_v^2$ e $Var(e) = \Sigma\sigma_e^2$.

b. Construção da função de verossimilhança hierárquica: Função de Verossimilhança Estendida (Hierárquica).

$$\begin{aligned}\log L(\beta, v; y, v) &= \log f(y, v) \\ &= \log f(y|v) + \log f(v) \\ &= \tfrac{1}{2}\log|2\pi\Sigma| - \tfrac{1}{2}(y - X\beta - Zv)^t\Sigma^{-1}(y - X\beta - Zv) - \tfrac{1}{2}\log|2\pi D| - \tfrac{1}{2}v^tD^{-1}v.\end{aligned}$$

c. Derivação das Equações de Modelos Mistos Hierárquicas: Maximização da função de Verossimilhança Estendida.

Passo 1: Derivação da função para β e v:

$$d L(\beta, t, v; y, t, v)/d\beta = X^t\Sigma^{-1}(y - X^t\beta - Z^tv); \text{ e}$$

$$d L(\beta, t, v; y, t, v)/d v = Z^t\Sigma^{-1}(y - X^t\beta - Z^tv)$$

Passo 2: Equações de Modelo Misto.

$$\begin{bmatrix} X^t\Sigma^{-1}X & X^t\Sigma^{-1}Z \\ Z^t\Sigma^{-1}X & Z^t\Sigma^{-1}Z + D^{-1} \end{bmatrix}\begin{bmatrix}\hat{\beta} \\ \hat{v}\end{bmatrix} = \begin{bmatrix} X^t\Sigma^{-1}y \\ Z^t\Sigma^{-1}y \end{bmatrix}.$$

Passo 3: Informação de Fisher.

A Informação de Fisher para β é: $I(\hat{\beta}) = X^t\Sigma^{-1}X$.

A Informação de Fisher para v é: $I(\widehat{v}) = (Z^t\Sigma^{-1}Z + D^{-1})$.

Passo 4: Modelo Misto - componentes de média aumentado.

O Modelo Misto aumentado é dado por:

$y_a = T\delta + e_a$, em que: $y_a = \begin{pmatrix} y \\ \psi_M \end{pmatrix}$; $T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}$; $\delta = \begin{pmatrix}\beta \\ v\end{pmatrix}$; e $e_a = \begin{pmatrix} e \\ e_M \end{pmatrix}$.

Assim: $\begin{pmatrix} y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}\begin{pmatrix}\beta \\ v\end{pmatrix} + \begin{pmatrix} e \\ e_M \end{pmatrix}$, em que: $\psi_M = 0$ refere-se ao dado quase-aumentado, com distribuição Normal com média $E(\psi_M) = v$, variância D e é independente de y. O subscrito M denota o modelo para a média e não para a dispersão. A ideia desse modelo linear aumentado não acrescenta algo novo na análise de Modelos Mistos Normais, mas torna-se muito útil na extensão para Modelos

Mistos Não Normais.

Equações de Modelo Misto Aumentado

$\left[\begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}' \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1} \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}\right] \begin{pmatrix} \beta \\ v \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}' \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1} \begin{pmatrix} y \\ \psi M \end{pmatrix}$, e

$T^t \Sigma_a^{-1} T \hat{\delta} = T^t \Sigma_a^{-1} y_a$, em que: $\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}$.

Esse sistema pode ser resolvido como uma regressão ponderada ordinária do tipo: X'V^{-1}X $\hat{\beta}$ = X'V^{-1} *y*, em que *V* é substituída por Σ_a.

Esse sistema de equações é idêntico às equações de Modelo Misto de Henderson. Nas aplicações em melhoramento animal e vegetal, os efeitos aleatórios *v* são correlacionados, de acordo com uma dada matriz de correlação *A*. Esse modelo pode ser transformado pela recomputação de *Z* como $ZA^{1/2}$, em que $A^{1/2}$ é alguma transformação raiz quadrada de *A* (transformação Cholesky). Então, o modelo aumentado descrito acima, contendo efeitos aleatórios independentes e identicamente distribuídos, pode ser aplicado, sem a necessidade de inversão da matriz *A*. Os modelos de análises de fatores contêm parâmetros em *A*, a serem estimados e $ZA^{1/2}$ contém os carregamentos dos fatores.

d. Derivação dos estimadores de componente de variância via verossimilhança perfilada: sob essa nova *Z*, tem-se que $Var(y) = (\lambda Z Z^t + \Phi I_n)$, em que *n* é o número de observações. Sob o modelo aumentado tem-se $Var(e_a) = \Sigma_a^{-1} = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} \Phi I_n & 0 \\ 0 & \lambda I_m \end{pmatrix}$, em que *m* é o tamanho do vetor *v*. Neste caso, a distribuição $v \sim N(0, \lambda A)$ torna-se $u \sim N(0, \lambda I)$.

Os componentes de *deviance* correspondentes a *e* são os resíduos quadráticos dados por: $d_i = (y_i - X_i\hat{\beta} - Z_i\hat{v})^2$. Os componentes de *deviance* correspondentes a e_M são os resíduos quadráticos dados por: $d_{Mi} = (\psi M - \hat{v}_i) = \hat{v}_i^2$. E os *leverages* (pontos de influência ou alavancagem) correspondentes são dados pelos elementos diagonais de: $T(T^t \Sigma_a^{-1} T)^{-1} T^t \Sigma_a^{-1}$.

e. Obtenção de um algoritmo iterativo eficiente (IWLS): a estimação de parâmetros (β, τ, e *v*, em que: τ são parâmetros de variância associados a *v*) no modelo linear misto: $y = X\beta + Zv + e$ (sendo *e* com distribuição normal e variância σ^2),

pode ser feito pelo algoritmo IWLS para o modelo linear aumentado como mostrado a seguir:

Passo 1. Comece com uma estimativa do parâmetro de variância τ;

Passo 2. Dado a estimativa corrente de τ, atualize $\hat{\delta}$ resolvendo as equações de quadrados mínimos generalizados (ou de Modelos Mistos) aumentados: $T^t\Sigma_a^{-1}T\hat{\delta} = T^t\Sigma_a^{-1}y_a$;

Passo 3. Dado o corrente valor de $\hat{\delta}$ obtenha o valor atualizado de τ;

Passo 4. Faça iterações entre os passos 2 e 3 até a convergência. Na convergência os erros padrões de $\hat{\beta}$ e $(\hat{v} - v)$ podem ser computados da inversa da matriz de informação H^{-1} da verossimilhança H e os erros padrões de $\hat{\tau}$ são computados da matriz Hessiana, a qual contem a Informação de Fisher.

2.3 Ponderação dos erros ou resíduos nas MME em GLMM e HGLMM

Em GLMM e HGLMM, as soluções das equações de Modelo Misto envolvem a ponderação dos erros calculados mediante o ajuste do modelo e dos erros teóricos esperados, associados a distribuição da variável latente e da função de ligação.

Tomando-se *v = a*, o modelo GLMM $y^* = X\beta + Za + (y-\mu)g'(\mu)$ contempla a função de ligação g'(u) e tem a mesma estrutura da primeira e segunda ordem que o modelo LMM $y = X\beta + Za + e$, de forma que os algoritmos de estimação e predição para o caso normal podem ser adaptados, apenas substituindo y por y* e $Cov(e) = R = I\sigma_e^2$ por $Cov[(y-\mu)g'(\mu)] = W^{-1}\sigma_e^2$. Assim, troca-se R por S e define-se S = $W^{-1}\sigma_e^2$, em que W é uma matriz diagonal, que contempla os coeficientes de ponderação ou pesos w_i (Wolfinger e O'Connell, 1993; Resende e Biele, 2002).

Assim, têm-se as seguintes equações de Modelo Misto (Resende e Biele, 2002):

$$\begin{bmatrix} X'S^{-1}X & X'S^{-1}Z \\ Z'S^{-1}X & Z'S^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta}_L \\ \hat{a}_L \end{bmatrix} = \begin{bmatrix} X'S^{-1}y^* \\ Z'S^{-1}y^* \end{bmatrix}, \text{ em que:}$$

S^{-1}: matriz com termos diagonais dados por $p_i\ (1-p_i)\frac{1}{\sigma^2_{e_L}}$, em que $w_i = [p_i\ (1- p_i)]$ são os pesos, no caso da Binomial, com função de ligação identidade.

$\sigma^2_{e_L}$: variância residual na escala contínua latente (*liability*);

$\beta_L\ e\ a_L$ = efeitos fixos e aleatórios na escala latente.

De posse da variável observacional ou dependente ajustada y*, tem-se que o modelo linear misto equivale a $y^* = X\beta + Za + (y-\mu)g'(\mu)$, em que:

$E(y^*) = X\beta,\ Cov(a) = G,\ Cov[(y-\mu)g'(\mu)] = W^{-1}\sigma_e^2$ *e* $Cov(y^*) = ZGZ' + W^{-1}\sigma_e^2$.

Assim, S = $W^{-1}\sigma_e^2$ ou seja, $S = \frac{\hat{\sigma}^2_{eL_1}}{p_1(1-p_1)}$, no caso da Binomial, com função de ligação identidade.

A matriz W representa uma matriz diagonal de pesos conhecidos (w), dados pela variância condicional Var (Y|u) (Wolfinger e O'Connell, 1993; Resende e Biele, 2002; Resende, 2002; 2015). Alguns pesos w são apresentados na Tabela 1, obtidos de Resende (2007, pagina 70) e Resende et al. (2014, pagina 67). A última coluna apresenta os valores numéricos de w usados nas aplicações práticas do presente trabalho. Para a Gama, w foi derivado no presente trabalho, via expansão de primeira ordem (método Delta) em Serie de Taylor, conforme mostrado em tópico seguinte.

Tabela 1. Pesos (w) das distribuições nas MME.

Distribuição do erro	**Ligação canônica g(u)**	**Variância residual dentro progênies (pesos w)**	$\sigma^2 = \phi / w$ = $(\hat{\sigma}_e^2) / w$	w
Normal	Identidade	1	$(\hat{\sigma}_e^2)$ /1	1
Binomial	Logito	$\pi^2/3$	$(\hat{\sigma}_e^2)$ /($\pi^2/3$)	3.29
Poisson	Log	u^{-1}	$(\hat{\sigma}_e^2)$ /u^{-1}	u^{-1}
Gama*	Reciproca ou inversa	(n-1)/2	$(\hat{\sigma}_e^2)$ /[(n-1)/2]	2.00
Gumbel ou Valor Extremo, Weibull	Comp log-log	$\pi^2/6$	$(\hat{\sigma}_e^2)$ / ($\pi^2/6$)	1.645
		w = fator de escala. Sob GLMM, $(\hat{\sigma}_e^2)$ tem que ser substituído por ϕ.	$\phi = (\hat{\sigma}_e^2)$: fator de dispersão estimado via REML	

*: para a Gama, w foi derivado no presente trabalho, via expansão de primeira ordem (método Delta) em Serie de Taylor, conforme mostrado em outro tópico.

Para a distribuição Gama, os pesos são dados pela inversa (reciproca) da variância condicional Var (Y|u) para o caráter Y=ln (S^2) os pesos constantes são dados por w_i= (n_i-1)/2. Possuem w = $\pi^2/6$: *Weibull*, *Gumbel*, Valor extremo, Exponencial,

Log-Gama. A distribuição *Gumbel*, é um caso especial da distribuição de *valor extremo* generalizada, sendo também chamada de distribuição *log-Weibull*. A distribuição exponencial é um caso especial da *Weibull* e da Gama. A Log-Gama é uma generalização da função de valores extremos. Informações complementares as da Tabela 1 são apresentadas na Tabela 2, para um modelo genérico com um só fator aleatório (genótipos), além do erro.

Tabela 2. Distribuições para o componente aleatório (erros, variável observada) e variável latente, funções de ligação, bem como variâncias residuais e herdabilidades da variável latente, para variáveis binárias (Binomiais ou Poisson) e continuas (Normal e Gama).

Componente aleatório	**Variável latente**	**Funções de ligação**	**Variância residual ou peso w**	**Herdabilidade**
Binomial	Logística	Logito: $\eta=log\,[\mu/(1-\mu)]$	$\pi^2/3$	$h^2=\sigma_g^2/(\sigma_g^2+\pi^2/3)$
	Normal Padrão	Probito: $\eta=\Phi^{-1}(\mu)$	1	$h^2=\sigma_g^2/(\sigma_g^2+1)$
	Gumbel	Comp log-log: $\eta=log[-log(1-\mu)]$	$\pi^2/6$	$h^2=\sigma_g^2/(\sigma_g^2+\pi^2/6)$
	Binomial	Identidade: $\eta=\mu$	$\mu(1-\mu)$	$h^2=\sigma_g^2/[(\sigma_g^2+[\mu(1-\mu)]]$
Poisson	Poisson	*Log* (μ)	μ^{-1}**	$h^2=\sigma_g^2/(\sigma_g^2+\mu^{-1})$
Poisson	Poisson	Identidade: $\eta=\mu$	μ	
Normal	Normal	Identidade: $\eta=\mu$	1	$h^2=\sigma_g^2/(\sigma_g^2+\sigma_e^2)$
Gama	Gama	Identidade: $\eta=\mu$	μ^2	
Gama	Gama	Reciproca ou Inversa: μ^{-1}	(n – 1) / 2*	

*: deduzida no presente trabalho pelo método Delta; **: deduzida pelo método Delta, por Foulley et al. (1987).

2.5. Derivação da variância (de amostragem) da variância residual na Distribuição Gama

Média e variância sob Distribuição Gama

O inventor da distribuição gama foi Leonhard Euler em 1729. Mas foi Daniel Bernoulli quem deu em 1729 a primeira representação de uma função de interpolação dos fatoriais na forma de um produto infinito, mais tarde conhecido como função gama. A variável aleatória contínua Y tem distribuição gama quando sua função densidade de probabilidade é dada por:

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot y^{\alpha-1} \cdot e^{-\beta y}, y > 0$$

$$f(y) = 0, y \leq 0,$$

em que α é o parâmetro de forma (α > 0) e β é o parâmetro de escala (β > 0). Assim, Y ~ G(α, β), e se α = 1, a distribuição é Exponencial. Quando α= n/2, n inteiro e β=1/2, tem-se distribuição Qui quadrado.

Uma variável aleatória distribuída de acordo com uma distribuição gama é contínua e não negativa. A distribuição gama é flexível e pode acomodar muitas formas de distribuição dependendo dos valores de α e β. É comumente usado para variáveis de resposta não-negativas e assimétricas com coeficiente de variação constante e quando a alternativa usual, uma distribuição log-normal, é inadequada. Os erros gama são úteis para dados que apresentam um coeficiente de variação constante. A variância de amostragem e a distribuição condicional do caráter não distribuído normalmente podem ser aproximadas por uma distribuição gama.

As definições paramétricas das medias e variâncias da Gama são: media: E(Y) = α / β; variância: Var (Y) = α / β^2, em que α: parâmetro de forma; β: parâmetro de escala; θ = 1/β: parâmetro de taxa: inverso do parâmetro de escala. Como exemplo de estimação, tem-se os cálculos a seguir, associados a função densidade de probabilidade da variável volume aos 7 anos, no teste de progênie de eucalipto. Foram estimadas a media (0.25) e variância (0.15^2) e obtidas as estimativas dos parâmetros da Gama:

β = E(Y) / Var (Y) = 0.25 / 0.15^2 = 11.11

α = E(Y) β = 0.25 x 11.11 = 2.78

θ = 1 / β = 1 / 11.11 = 0.09

De posse dos valores 2.78 para α e 0.09 para θ, plotou-se a curva em preto no histograma abaixo. Os gráficos das funções densidades de probabilidade das três variáveis (altura, DAP e volume) avaliadas aos sete anos de idade no experimento com progênies de eucalipto encontram-se na Figura abaixo.

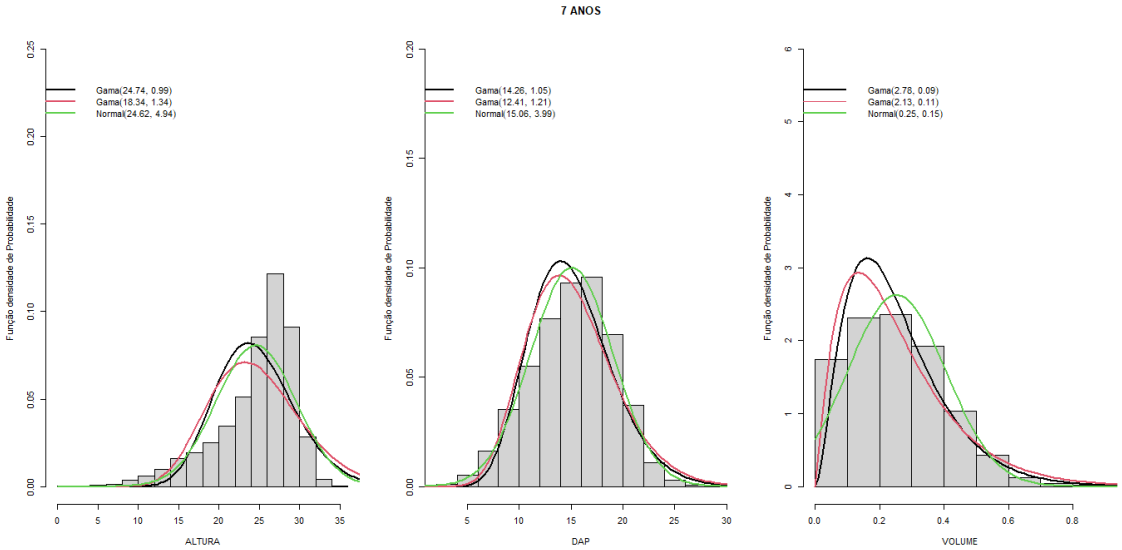

Figura 1. Figura com os histogramas das distribuições de altura, DAP (diâmetro na altura do peito) e volume.

Pelo histograma, verificam-se que as 3 distribuições se assemelham, visualmente, a: Logística para altura, Normal para diâmetro e Qui-quadrado para o volume. Estas 3 distribuições são casos particulares da Gama, que é uma distribuição ampla e generalizada.

Expansão em Série de Taylor para derivação da variância de amostragem da variância residual na distribuição Gama

A variância de amostragem da variância residual, obtida pela Expansão em Serie de Taylor via método Delta foi utilizada e descrita a seguir. O método delta consiste em retirar uma parte dos termos da expansão em série ou polinômio de Taylor de uma função, com vista à obtenção de aproximações, para os momentos de uma estatística de interesse. A prática habitual consiste em truncar a expansão após a primeira derivada. No caso de esta se anular no ponto em que é efetuada a expansão recorre-se ao termo correspondente à segunda derivada. Uma aproximação para o valor médio e para a variância de uma função de variáveis aleatórias pode ser obtida utilizando-se uma aproximação pela série de Taylor. Em geral, utiliza-se uma

aproximação de primeira ordem em torno dos verdadeiros valores paramétricos. A aproximação de Taylor de primeira ordem, quando utilizada para calcular expressões aproximadas para variâncias é denominado método Delta (Casella e Berger, 2006). O nome Serie de Taylor foi introduzido por Brook Taylor, em 1715.

Com base em Bulmer (1980), suponha que $X_1, X_2, \dots, X_n$ seja um conjunto de variáveis aleatórias com média conhecida, variância e covariância dadas por:

$$E[X_i] = \zeta_i$$
$$V[X_i] = V_{ii}$$
$$COV[X_i, X_j] = V_{ij}.$$

Procura-se por uma expressão aproximada para a variância de uma função das X_is, ou seja, $Y = f(\boldsymbol{X})$. Para isso, faz-se a expansão de $f(\boldsymbol{X})$ na série de Taylor sobre $\boldsymbol{X} = \boldsymbol{\zeta}$:

$$Y = f(\zeta) + \sum_{i=1}^{k} (X_i - \zeta_i) \frac{\partial f}{\partial X_i}\Big|_{X=\zeta}.$$

Então,

$$E[Y] \cong f(\zeta)$$
$$V[Y] = E\big[Y - E[Y]\big]^2 \cong \textstyle\sum_{i,j} V_{ij} \frac{\partial f}{\partial X_i} \frac{\partial f}{\partial X_j}\Big|_{\boldsymbol{X}=\boldsymbol{\zeta}}.$$

Essa expressão aproximada de $V[Y]$ é útil em muitos contextos e pode ser generalizada para obtenção de covariância de duas funções de X_is, $Y_1 = f_1(\boldsymbol{X})$ e $Y_2 = f_2(\boldsymbol{X})$:

$$COV[Y_1, Y_2] \cong \sum_{i,j} V_{ij} \frac{\partial f_1}{\partial X_i} \frac{\partial f_2}{\partial X_j}\Big|_{\boldsymbol{X}=\boldsymbol{\zeta}}$$

Para uma variável, usando Expansão da Série de Taylor e pelo teorema de Taylor tem-se:

$f(x) \approx f(a) + f'(a)(x - a)$.

Fazendo $a = \mu_x$ a média de X, uma expansão da série de Taylor de $y = f(x)$ sobre μ_x propicia a aproximação:

$y = f(x) \approx f(\mu_x) + f'(\mu_x)(x - \mu_x)$.

Aplicando a variância em ambos os lados da equação, tem-se:

$Var(y) = Var(f(x)) \approx [f'(\mu_x)]^2 Var(X)$.

Assim, se Y é qualquer função de uma variável aleatória X, torna-se necessário apenas calcular a variância de X e a primeira derivada da função para a aproximação da variância de Y .

Variâncias de funções de componentes de variância via Expansão de Taylor

As funções mais importantes são a raiz quadrada (para computo de desvios padrões) e proporção entre componentes de variância (herdabilidades). Para a raiz quadrada, tem-se a aproximação $Var(y^{1/2}) \approx Var(y) / [4E(y)]$.

A variância do desvio padrão ($y^{1/2} = \sigma$) para os efeitos residuais de um modelo ajustado equivale a (aplicando a variância em ambos os lados da equação):

$$Var(\sigma) \approx Var(\sigma^2) / [4E(\sigma^2)]$$
$$\approx Var(\sigma^2) / (4\sigma^2)$$
$$\approx (2\sigma^4) / (4v\sigma^2)$$
$$\approx \sigma^2 / (2v)$$
$$\approx \sigma^2 / (2(n-1)),$$

em que $v = (n - 1)$ são os graus de liberdade e n é o número de plantas por parcela.

No contexto da variância de amostragem da variância residual dentro de parcelas na distribuição Gama tem-se:

$$(2\sigma^4) / (4v\sigma^2) \approx [1 / (4\sigma^2)] [(2\sigma^4) / v]$$
$$\approx \sigma^2 / (2(n-1)).$$

Obteve-se assim $Var(\sigma)$, ou seja, trabalhando-se com desvios (σ). Trabalhando-se com $Var(\sigma^2)$ obtém-se $Var(\sigma^2) \approx (2/(n-1))$. Usando esse resultado e sendo os pesos w dados pelo inverso (função de ligação reciproca, conforme Wolfinger e O'Connell, 1993) da variância condicional Var (Y|u), tem-se que *w = (n–1)/2.*

Assim, a esperança (E(erro)) e a variância (Var(erro)) foram obtidas pela Expansão em Serie de Taylor, de primeira ordem (método Delta) e as variâncias equivalem a:

$$Var(S^2) = 2 / (n-1);$$
$$Var(Ln\ (S^2)) = 2 / (n-1).$$

A expressão $Var(Ln\ (S^2)) = 2 / (n-1)$ concorda com Resende (2007, pagina 167).

A expressão 2/(n-1) é também a variância de ln (S^2). Tomando-se S como a raiz quadrada de S^2, tem-se que a distribuição de S tem media e variância S = Var de S = 2 (n-1) = 2v, em que v = (n -1) são os graus de liberdade e também é a média. Essa abordagem é apresentada nos livros, os quais atribuem S~ X^2_{n-1} .

A função de ligação Gama com a escala latente é a *link function* canônica reciproca ou inversa de Gama (Wolfinger e O'Connell, 1993; Crawley, 2015). Isto é importante para determinar os pesos ou ponderadores (w) da analise nas MME. Assim, tomando-se os recíprocos das quantidades acima tem-se os pesos:

$$w = 1 / Var(S^2) = (n-1) / 2;$$

$$w = 1 / (Ln\ (S^2)) = (n-1) / 2;$$

Dessa forma, o peso w equivale ao número de graus de liberdade da distribuição X^2_{n-1} dividido por 2.

2.6. Fator de escala, variância residual, herdabilidade e distribuições escaladas sob Gama em HGLM

Quando os dados associados a um caráter em questão têm distribuição normal, as variâncias residuais dentro de progênies têm distribuição qui-quadrado, com número de graus de liberdade v = (n-1) associados a estimação de cada variância residual dentro progênie, em que n é o número de plantas por parcelas.

A variância de amostragem de uma variável com distribuição qui-quadrado equivale a $2u^2/v = 2v = 2\ (n-1)$, em que u = v é a média geral e, v = (n – 1) é o número de graus de liberdade dentro da parcela. A variância do logaritmo da variância dentro de progênies, Var(ln(S^2)), tem distribuição assintótica normal com variância 2/v, ou seja, 2/(n – 1) (Resende, 2007, pagina 167).

Na deviance escalada $D = \frac{(y-u)^2}{s^2} = \frac{\hat{s}^2}{s^2}$, o parâmetro s^2 é a variância esperada (paramétrica) e $\hat{s}^2$ é a variância estimada (Mc Callagh e Nelder, 1989; pag 24-225 e 127-128). O uso da variância residual escalada é fundamental na comparação das variâncias residuais e herdabilidades das análises das diferentes distribuições assumidas para o caráter. As variâncias residuais escaladas são dadas pela razão entre a variância residual estimada e variação residual teórica, esperada segundo cada distribuição de probabilidade paramétrica assumida e sua função de ligação.

Para se obter a mesma escala nas comparações dos resultados entre diferentes distribuições devem ser adotados ponderadores que projetam as estimativas para uma mesma base de comparação (projeção para uma escala comum) (Resende, 2002).

Tabela 3. Estimadores da variância e h^2 sob o modelo do teste de progênie de eucalipto.

Distribuição	**Variância Fenotípica** $\hat{\sigma}_y^2$ $\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / w$	**Herdabilidade Escalada** $\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / w)$
Normal	$\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / 1$	$\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / 1)$
Logística	$\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / (\pi^2/3)$	$\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / (\pi^2/3)$
Weibull	$\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / (\pi^2/6)$	$\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 / (\pi^2/6)$
Gama	$\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 /_{[(n-1)/2]}$	$\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 /_{[(n-1)/2]}$
Qui-quadrado	$\hat{\sigma}_y^2 = \hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2$ /[2(n-1)]	$\hat{h}_a^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_c^2 + \hat{\sigma}_e^2 /$ [2(n-1)]

Na Tabela 4 a seguir são apresentados os estimadores para as variâncias residuais e herdabilidades escaladas utilizados no presente trabalho.

Tabela 4. Fatores de projeção das variâncias residuais e herdabilidades escaladas utilizados no presente trabalho.

Distribuição	**Variância Teórica**	**Variância Residual Escalada**	**Fator de projeção da herdabilidade escalada**
Normal Padrão (N)	Var(Res) ~ N = 1	$\hat{s}_{Res}^2 = \hat{s}_N^2 / S_N^2$	$\hat{\sigma}^2 = \hat{\sigma}_e^2 / 1$
Weibull (W)	Var(Res)~W = $\pi^2/6$=1.645	$\hat{s}_{Res}^2 = \hat{s}_G^2 / S_W^2$	$\hat{\sigma}^2 = \hat{\sigma}_e^2 / 1.645$
Exponencial Dupla (ED)	Var(Res) ~ ED=2	$\hat{s}_{Res}^2 = \hat{s}_G^2 / S_{ED}^2$	$\hat{\sigma}^2 = \hat{\sigma}_e^2 / 2$
Logística (L)	Var(Res)~ L = $\pi^2/3$=3.29	$\hat{s}_{Res}^2 = \hat{s}_G^2 / S_L^2$	$\hat{\sigma}^2 = \hat{\sigma}_e^2 / 3.29$
Qui-quadrado (Q)	Var(Res) ~ Q=2(n-1)=8	$\hat{s}_{Res}^2 = \hat{s}_G^2 / S_Q^2$	$\hat{\sigma}^2 = \hat{\sigma}_e^2 / 8$

Gama (G)*	Var(Res) ~ G=(n-1)/2=2	$\hat{s}^2_{Res} = \hat{s}^2_G / S^2_G$	$\hat{\sigma}^2 = \hat{\sigma}^2_e / 2$

*: deduzida no presente trabalho; n = 5.

2.7. Distribuição Qui-Quadrado (χ^2) e variável volume

A variável $\chi^2 = \frac{(n-1)\hat{s}^2}{s^2}$ tem distribuição qui-quadrado com v=n-1 graus de liberdade. A estimativa $\hat{s}^2$ advém de uma amostra aleatória da distribuição normal y~N(u, s^2). Possui média v, e variância 2v, em que v = (n-1) refere-se aos graus de liberdade, e n é o número de indivíduos na parcela.

A distribuição qui-quadrado pode ser derivada da distribuição normal. Se a variável aleatória y_i segue a distribuição normal padrão, a variável χ^2=soma (y_i^2) segue uma distribuição χ^2 com v graus de liberdade. A média de χ^2 é E(χ^2) = v e a variância $Var(\chi^2) = s^2 = 2v = 2(n-1)$.

Em geral, uma função quadrática de dados de uma distribuição normal, tem distribuição qui-quadrado. Assim, a variável volume de madeira, a qual é função quadrática da variável diâmetro do caule, segue a distribuição qui-quadrado. Devido a isto, o histograma e a curva da função densidade da variável volume exibem um padrão típico da distribuição qui-quadrado, conforme as Figuras mostrada anteriormente. É importante notar que $\hat{s}^2$ é a variância amostral e s^2 a variância verdadeira (paramétrica) da distribuição. A qui-quadrado y não apresenta qualquer eixo de simetria e é definida sobre o domínio real positivo. A variável Y depende de um único parâmetro v e quanto maior o valor de v, mais larga é a distribuição em torno do valor médio y=v.

2.8. Dados experimentais

Conjunto de dados para variáveis contínuas

Foram avaliadas 286 famílias de irmãos completos de eucalipto. O experimento foi conduzido em delineamento de blocos casualizados, com seis plantas por parcela e oito repetições. O plantio foi realizado em novembro de 2003 e o espaçamento foi 3 m entre linhas e 2 m entre árvores. Aos dois, cinco e sete anos de idade, foram

mensurados a altura (m), diâmetro à altura do peito (DAP) (cm) e volume (m^3). No total foram avaliadas 9 variáveis.

Conjunto de dados para variáveis discretas categóricas associadas a escores

Foram avaliadas 20 famílias híbridas de café canephora. O experimento foi conduzido em delineamento de blocos casualizados, com uma planta por parcela e até 35 repetições. O plantio foi realizado em março de 2011 e o espaçamento foi 3 m entre linhas e 1.5 m entre árvores. As condições experimentais foram descritas por Alkimim et al. (2021).

Por três anos consecutivos (2013 a 2015), foram mensuradas as incidências de ferrugem (causada pelo fungo *Hemileia vastatrix Berk. & Br.*) avaliadas em uma escala de 1 a 5. Para as análises, essa escala foi modificada para variar de 0 a 4, onde 0 foi atribuído a plantas assintomáticas e 4 foi atribuído a plantas altamente suscetíveis aos patógenos.

2.9. Analises dos dados experimentais

Para as análises dos dois conjuntos de dados (eucaliptos e cafeeiros), o HGLMM utilizado, em sua forma geral foi dado por:

$y = X\beta + Zg + Wc + e$, (1) em que y é o vetor de dados fenotípicos, β vetor de efeitos fixos (média geral e repetições), g é o vetor dos efeitos aleatórios de progênies (conforme o conjunto de dados), c é o vetor dos efeitos aleatórios de ambiente comum da parcela (eucaliptos) ou blocos (cafeeiros) e e é o vetor de resíduos. X, Z e W são as matrizes de incidência para os efeitos β, g e c, respectivamente.

Partindo desse modelo, vários ajustes foram executados. A diferença entre os ajustes foram as distribuições assumidas para e, no caso, a Gama ou Normal para variáveis contínuas, e Poisson para variáveis categóricas associadas a contagem. Também, as distribuições de g e c variaram entre Gaussiana, Gama, Gama-Inversa e Beta.

Ademais, as funções de ligação foram atribuídas conforme Resende et al. (2018), que sugerem função identidade para distribuição Normal, função logarítmica para as distribuições Gama, Gama Inversa e Poisson.

Para a análise do conjunto de dados de café canéfora (análise conjunta de safras), o HGLMM utilizado foi dado por:

$y = X\beta + Zg + Wc + Tp + e$, (2) em que y é o vetor de dados fenotípicos, β vetor de efeitos fixos (média geral e repetições), g é o vetor dos efeitos aleatórios de progênies (conforme o conjunto de dados), c é o vetor dos efeitos de ambiente comum do bloco (aleatórios), p é o vetor dos efeitos aleatórios de ambiente permanente de indivíduo e e é o vetor de resíduos. X, Z, W e T são as matrizes de incidência para os efeitos β, g, c e p, respectivamente. e e é o vetor dos resíduos (aleatórios). X, Z, W e T são as matrizes de incidência para β, g, c e p, respectivamente.

Análises preliminares de LMM foram realizadas no *software* Selegen (Resende, 2015; 2016). O *software* HGLM (Ronnegard et al., 2010; Alam et al., 2014) em R foi usado. As rotinas computacionais desse *software* são apresentadas por Resende et al. (2023).

Para a comparação dos diferentes HGLMM, utilizou-se o cAIC – conditional *Akaike information criterion*. O cAIC mede a qualidade dos ajustes e permite a seleção de modelos. Quanto menor o valor do AIC, melhor é o ajuste. Diferenças significativas entre AICs de dois modelos, devem ser pelo menos 2 unidades (Cavanaugh and Neath, 2019; Resende e Alves, 2020; 2022) ou 1 unidade (Sakamoto et al., 1986). Diferenças significativas entre BICs devem ser pelo menos 2 unidades (Neath and Cavanaugh, 2012; Resende e Alves, 2020; 2022) entre modelos.

3. Resultados e Discussão

3.1. Herdabilidades Gama e projeção para outras distribuições em variáveis no melhoramento florestal e do cafeeiro.

Variáveis continuas

As análises que serão apresentadas a seguir, referem-se a variáveis contínuas dadas pela medição de altura, diâmetro e volume em 286 famílias de irmãos completos de eucalipto, avaliadas em 8 blocos, com 6 plantas por parcela.

Tabela 5. Avaliação genética de progênies de eucalipto, características altura, DAP e volume na idade de 7 anos em termos da herdabilidade (h^2) no sentido amplo entre progênies, considerando a distribuição Gama para os resíduos e as distribuições projetadas (Weibull, exponencial-dupla, logística e qui-quadrado). Erros Normais foram também apresentados como padrões para comparações.

Erros	**V.**	**Distrib.**	**h^2**	**V.**	**Distrib.**	**h^2**	**V.**	**Distrib.**	**h^2**
Norm.	Vol-7	**Normal**	**0.09**	Vol-5	**Normal**	**0.11**	Vol-2	**Normal**	**0.11**
Gama	Vol-7	Weibull	0.09	Vol-5	Weibull	0.15	Vol-2	Weibull	0.14
Gama	Vol-7	Gama	0.10	Vol-5	Gama	0.17	Vol-2	**Gama**	**0.16**
Gama	Vol-7	Logit	0.16	Vol-5	Logit	0.25	Vol-2	Logit	0.24
Gama	Vol-7	Q-Q	**0.32**	Vol-5	Q-Q	**0.41**	Vol-2	Q-Q	0.41
Norm.	Alt-7	**Normal**	**0.07**	Alt-5	**Normal**	**0.09**	Alt-2	**Normal**	**0.14**
Gama	Alt-7	Weibull	0.05	Alt-5	Weibull	0.12	Alt-2	Weibull	0.13
Gama	Alt-7	**Gama***	**0.09**	Alt-5	**Gama**	**0.13**	Alt-2	**Gama**	**0.15**
Gama	Alt-7	Logit	0.13	Alt-5	Logit	0.18	Alt-2	Logit	0.2
Gama	Alt-7	Q-Q	0.25	Alt-5	Q-Q	0.26	Alt-2	Q-Q	0.29
Norm.	Dap-7	**Normal**	**0.08**	Dap-5	**Normal**	**0.10**	Dap-2	**Normal**	**0.09**
Gama	Dap-7	Weibull	0.10	Dap-5	Weibull	0.13	Dap-2	Weibull	0.12
Gama	Dap-7	**Gama**	**0.11**	Dap-5	**Gama**	**0.15**	Dap-2	**Gama**	**0.14**

Gama	Dap-7	Logit	0.17	Dap-5	Logit	0.22	Dap-2	Logit	0.21
Gama	Dap-7	Q-Q	0.32	Dap-5	Q-Q	0.36	Dap-2	Q-Q	0.35

* os resultados da Gama são validos também para a Exp. Dupla, pois possui o mesmo peso w = 2. Norm. = Distribuição normal; Q-Q = Distribuição Qui-Quadrado; V.= variável.

Com base na Tabela 5, as eficiências das novas análises calcadas na Gama sobre a tradicional abordagem Normal foram: 0.32/0.09; 0.41/0.11; 0.16/0.11 para volume nas três idades; 0.09/0.07; 0.13/0.09; 0.15/0.14 para altura nas três idades; 0.11/0.08; 0.15/0.10; 0.14/0.09 para Dap nas três idades. Verificam-se maiores vantagens da abordagem Gama na variável volume (distribuição aproximada Qui-quadrado), seguida por DAP (distribuição aproximada Normal) e, altura (distribuição aproximada Logística).

A variável volume segue uma distribuição qui-quadrado. No entanto, a totalidade dos trabalhos avaliam o volume segundo uma distribuição normal. A modelagem dessa variável sob distribuição Gama mostrou-se vantajosa, conduzindo a ganhos altos sobre a modelagem tradicional. Admitindo erros Gama, a analise permitiu capturar elementos de distribuições derivadas ou relacionadas a Gama. Assim, modelando-se o erro como Gama e fator de correção ou projeção qui-quadrado, maximiza a acurácia da análise. Outras variáveis como altura e diâmetro apresentam distribuição Weibull e poderiam também ser melhoradas adotando-se função densidade de probabilidade própria para a Weibull a qual está relacionada a Exponencial, Qui-quadrado e Gumbel, todas essas, relacionadas a Gama. As principais variáveis de crescimento em espécies florestais seguem as distribuições Weibull, Normal e Gama (Mirzaei et al., 2016). De todas essas distribuições, a Gama é a mais flexível e abrangente.

Na Tabela 6, referentes aos dados de Eucalipto, são apresentados os componentes de variância genotípicas (Vg), entre parcelas (Vc) e dentro de parcelas (Ve). E também as herdabilidades Normal (h^2_N), Gama (h^2_G) e Qui-quadrado (h^2_Q) para a variável volume 7. Verifica-se que foi muito vantajosa, em termos de herdabilidades, a modelagem Qui-quadrado (h^2_Q) para esse caráter.

Tabela 6. Componentes (Vg), Vc e Ve) de variância e herdabilidades (h^2) para o caráter volume em Eucalipto aos 7 anos.

Volume 7	**h^2**	Parâmetros Normal

Distr.	h^2_N	Vg	Vc	Ve
N-N-N	0.09	0.001975	0.000476	0.020644
G-G-N	0.09	0.001984	0.000473	0.020648
GI-GI-N	0.09	0.001964	0.000479	0.02064
G-N-N	0.09	0.001982	0.000478	0.020644
	h^2_G		Parâmetros Gama	
G-G-G	0.10	0.027877	0.000794	0.477406
GI-GI-G	0.10	0.027311	0.001014	0.477564
N-N-G	0.10	0.027733	0.001043	0.477386
GI-N-G	0.10	0.027291	0.001035	0.477559
GI-G-G	0.10	0.027297	0.000988	0.477584
	h^2_Q		Parâmetros Gama	
G-G-G	0.32	0.027877	0.000794	0.477406
GI-GI-G	0.31	0.027311	0.001014	0.477564
N-N-G	0.31	0.027733	0.001043	0.477386
GI-N-G	0.31	0.027291	0.001035	0.477559
GI-G-G	0.31	0.027297	0.000988	0.477584

Variáveis discretas

Para as variáveis categóricas discretas, os dados do cafeeiro arábica revelaram superioridade da distribuição de Poisson para os erros, mostrando eficiência do sistema Poisson-Gama e Poisson-Normal (Tabela 7).

Tabela 7. Avaliação genética de progênies de cafeeiros para a resistência a ferrugem avaliadas em 2 anos.

Ferrugem 2013				**Ferrugem 2014**			
Distr.	AIC	Dist Erro	h2	Distr.	AIC	Dist Erro	h2
N-N-N	629	Normal	0.16	N-N-N	679	Normal	0.28
N-N-P	**422**	**Poisson**	**0.40**	**N-N-P**	**549**	**Poisson**	**0.53**
G-G-P	423	Poisson	0.37	**G-G-P**	**551**	**Poisson**	**0.50**
GI-GI-P	**422**	**Poisson**	**0.40**	GI-GI-P	548	Poisson	0.49
B-B-P	422	Poisson	0.14	B-B-P	549	Poisson	0.21
Ferrugem2015				**Ferrugem Conjunta**			
Distr.	AIC	Dist Erro	h2	Distr.	AIC	Dist Erro	h2
N-N-N	499	Normal	0.11	N-N-N	1702	Normal	0.16
N-N-P	**402**	**Poisson**	**0.23**	**N-N-P**	**1459**	**Poisson**	**0.45**
G-G-P	**403**	**Poisson**	**0.22**	**G-G-P**	**1467**	**Poisson**	**0.44**
GI-GI-P	402	Poisson	0.21	GI-GI-P	1459	Poisson	0.41

B-B-P	403	Poisson	0.07	B-B-P	1462	Poisson	0.22

Verifica-se que os melhores resultados, em termos de herdabilidades foram proporcionados pelo uso de diferentes distribuições para o erro, e para os outros dois fatores aleatórios (genótipos e parcelas). Sobressaíram-se as combinações erros Poisson (P), e parcelas e genótipos com distribuições Normal (N), Gama (G) e Gama Inversa (GI) (Tabela 7).

4. Conclusões

A abordagem HGLMM mostrou-se efetiva em modelar efeitos aleatórios com variadas distribuições;

A modelagem Gama mostrou-se superior a Normal em várias algumas situações.

O volume de madeira é melhor modelado por uma distribuição Qui-quadrado, em algumas situações.

Pesos para escalar as distribuições foram derivados e avaliados, sendo uteis e necessários para a abordagem HGLMM.

Para as variáveis categóricas, os dados do cafeeiro arábica revelaram superioridade da distribuição Poisson para os erros, mostrando eficiência do sistema Poisson-Gama e Poisson-Normal, sendo a Gama e a Normal as distribuições mais adequadas aos demais fatores de efeitos aleatórios.

Referências

ALKIMIM, E. R. et al. Designing the best breeding strategy for Coffea canephora: genetic evaluation of pure and hybrid individuals aiming to select for productivity and disease resistance traits. **PloS one**, v. 16, n. 12, p. e0260997, 2021.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics**, 2015

AYDIN, D; ŞENOĞLU, B. Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. **Journal of Modern Applied Statistical Methods**, v. 14, n. 2, p. 12, 2015.

BARBOSA, M. H. P.; RESENDE, M. D. V.; BRESSIANI, J. A.; SILVEIRA, L. C. L. Selection of sugarcane families and parents by Reml/Blup. **Crop Breeding and Applied Biotechnology**, v. 5, p. 443-450, 2005.

BRESLOW, N.E.; CLAYTON, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. **Journal of the American Statistical Association**, 88: 9-25.

BISHOP, C. M. **Pattern recognition and machine learning**. Springer, 2006.

BULMER, M. G. **The mathematical theory of quantitative genetics**. Oxford: Charedon Press, 1980. 254 p.

CASELLA, G; BERGER, R.L. **Statistical Inference**. Second Edition. 2006. 668 p.

CAVANAUGH, J. E.; NEATH, A. A. 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. **Computational Statistics**, 11:2-11.

CHISSOM, B.S. Interpretation of the kurtosis statistic. **The American Statistician**, v. 24, n. 4, p. 19-22, 1970.

CRAWLEY, M.J. **Statistics: An Introduction Using R (2nd Edition)**. John Wiley & Sons, Chichester, *2015.*

D'AGOSTINO, R B. Transformation to normality of the null distribution of g1. **Biometrika**, p. 679-681, 1970.

FISHER, R. A.; MACKENZIE, K. Studies in crop variation II: The manorial response of different potato varieties. **Journal of Agricultural Science**, 1923.

FOULLEY, J. L.; GIANOLA, D.; IM, S. Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. **Theoretical and Applied Genetics**, v. 73, p. 870, 1987.

HENDERSON, C.R. (1975). Best linear estimation and prediction under a selection model. **Biometrics**, 31: 423-447.

HENDERSON, C.R.; KEMPTHORNE, O.; Searle, S.R.; Von Krosigk, C.M. (1959). The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, 15: 192-218.

LEE, Y.; NELDER, J.A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society:** Series B (Methodological), v. 58, n. 4, p. 619-656, 1996.

LEE, Y.; NELDER, J.A. (2001). Hierarchical generalized linear models: A synthesis of generalised linear models, random effect models and structured dispersions. **Biometrika**, 88: 987-1006.

LEE, Y.; NELDER, J. A. Double hierarchical generalized linear models (with discussion). **Journal of the Royal Statistical Society:** Series C (Applied Statistics), v. 55, n. 2, p. 139-185, 2006.

LEE, Y.; HA, I.D. (2010). Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. **Statistics and Computing**, 20: 295-303.

LEE, Y.; KIM, G. (2016). H-likelihood predictive intervals for unobservables. **International Statistical Review**, 84: 487-505.

LEE, Y.; RÖNNEGÅRD, L.; NOH, M. **Data analysis using hierarchical generalized linear models with R**. Chapman and Hall/CRC, 2017.

LINDLEY, D.V.; SMITH, A.F. (1972). Bayes estimates for the linear model**. Journal of the Royal Statistical Society**: Series B (Methodological), 34: 1-41.

MA, R.; JORGENSEN, B. (2007). Nested generalized linear mixed models: Orthodox best linear unbiased predictor approach. **Journal of the Royal Statistical Society**: Series B, 69: 625–641.

MCCULLAGH, P. AND NELDER, J. A. (1989). **Generalized Linear Models**. Chapman & Hall, London.

MIRZAEI et al., 2016. Modeling frequency distributions of tree height, diameter and crown area by six probability functions for open forests of Quercus persica in Iran. **Journal of Forestry Research.**

NEATH, A. A.; CAVANAUGH, J. E. 2012. The Bayesian information criterion: background, derivation, and applications. **Computational Statistics**, 4:199-203.

NELDER, J.A.; PREGIBON, D. (1987). An extended quasi-likelihood function. Biometrika, 74: 221-232.

NELDER, J.A.; WEDDERBURN, R.W.M. (1972). Generalized linear models. **Journal of the Royal Statistical Society**, 135: 370-384.

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

PEDROZO, C. A.; BENITES, F. R. G.; BARBOSA, M. H. P.; RESENDE, M. D. V.; SILVA, F. L. Eficiência de índices de seleção utilizando a metodologia REML/BLUP no melhoramento da cana-de-açúcar. **Scientia Agraria**, v.10, n.1, 31-36, 2009.

RESENDE MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informações Tecnológicas, Brasília, 975p.

RESENDE MDV (2007) **Matemática e estatística na análise de experimentos e no melhoramento genético**, Embrapa Florestas, Colombo, 490p.

RESENDE MDV; SILVA, FF; AZEVEDO, CF (2014) **Estatística matemática, biométrica e computacional**. Suprema, Visconde do Rio Branco, 882p.

RESENDE MDV (2015) **Genética quantitativa e de populações**. Visconde do Rio Branco, 463p.

RESENDE, M. D. V. **Melhoramento de essências florestais**. In: Borem, A. **Melhoramento de espécies cultivadas**. Viçosa, MG: UFV, 589-647, 1999.

RESENDE, M. D. V.; BARBOSA, M. H. P. **Melhoramento genético de plantas de propagação assexuada**. Embrapa Florestas.

RESENDE, M. D. V., BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística**, v.20, p.30-65, 2002.

RESENDE MDV; DUARTE JB (2007). Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, 37: 182-194.

RESENDE, M.D.V. (2016). *Software* Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breeding and Applied Biotechnology** 16: 330-339.

RESENDE, M.D.V.; AZEVEDO, C.F.; SILVA, F.F.; NASCIMENTO, M.; GOIS, I.B.; ALVES, R.S. **Modelos Hierárquicos Generalizados Lineares Mistos (HGLMM), Máxima Verossimilhança Hierárquica (HIML) e HG-BLUP**.1. ed. Visconde do Rio Branco: Suprema, v.1, p.151, 2018.

RESENDE MDV; ALVES RS (2020). Linear, generalized, hierarchical, Bayesian and random regression mixed models in genetics/genomics in plant breeding. **Functional Plant Breeding Journal**.

RESENDE MDV; ALVES RS (2022). Statistical significance, selection accuracy and experimental precision in plant breeding. **Crop Breeding and Applied Biotechnology**.

RESENDE JR., M.F.R. ; VALLE, P.R.M. ; RESENDE, M. D. V. ; GARRICK, D. J. ; FERNANDO, R. L. ; DAVIS, J.M. ; JOKELA, E. J. ; MARTIN, T. A. ; PETER, G. F. ; KIRST, M. Accuracy of genomic selection methods in a standard dataset of loblolly pine. **Genetics**, v.190, p.1503 - 1510, 2012a.

RONNEGARD, L., SHEN, X. and ALAM, M. (2010) hglm: A Package for Fitting Hierarchical Generalized Linear Models. **The R Journal**, 2(2): 20-28.

ROSS, Sheldon. **A first course in probability**. Pearson, 2014.

SAKAMOTO, Y.; ISHIGURO, M.; KITAGAWA, G. **Akaike information criterion statistics.** KTK, Tokyo, 1986.

SCHALL, H. (1991). Estimation in generalized linear models with random effects. **Biometrika**, 78: 719–727.

STROUP, W. W. (2013). **Generalized linear mixed models**. Boca Raton: CRC Press.

THOM, Herbert CS. A note on the gamma distribution. **Monthly weather review**, v. 86, n. 4, p. 117-122, 1958.

THOMPSON, R. (1973). The estimation of variance and covariance components when records are subject to culling. **Biometrics**, 29: 527-550.

THOMPSON, R.; BAKER, R.J. (1981). Composite link functions in generalized linear models. **Applied Statistics**, 30: 125-131.

VRIEZE, S. I. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). **Psychological methods**, v. 17, n. 2, p. 228, 2012.

WOLFINGER, R; & O'CONNELL, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. **Journal of Statistical Computation and Simulation**, 48, 233-243.