ESTIMAÇÃO E PREDIÇÃO EM MODELOS LINEARES GENERALIZADOS MISTOS COM VARIÁVEIS BINOMIAIS

Marcos Deon Vilela de RESENDE[1]
Jonathan BIELE[2]

▪ RESUMO**:** Variáveis binomiais tais como a presença ou ausência de determinados atributos nos indivíduos não são bem descritas por modelos lineares clássicos. Para variáveis deste tipo, a metodologia de modelos lineares generalizados mistos deve ser a metodologia preferida. O presente trabalho teve como objetivos apresentar os fundamentos teóricos e práticos da análise de modelos não lineares fixos e mistos para variáveis binomiais. Foram usados dados referentes à sobrevivência de plantas de duas espécies florestais (*Eucalyptus grandis* e *Ilex paraguariensis*). Concluiu-se que: os estimadores, preditores e algoritmos apresentados permitem uma acurada predição de efeitos aleatórios e uma precisa estimação de componentes de variância associados a variáveis binomiais, através do uso da técnica de modelos lineares generalizados mistos; as funções de ligação logito e probito mostraram-se adequadas na análise dos dados de sobrevivência de plantas de espécies perenes; as funções de ligação complemento log-log e identidade mostraram-se inadequadas aos dados de proporções analisados; a função de ligação logito mostrou-se a mais adequada, pois conduziu a coeficientes de determinação (h^2) dos efeitos aleatórios similares aos obtidos pela função probito e refere-se à base canônica ou natural; as estimativas do parâmetro de dispersão ou fator de heterogeneidade apresentaram magnitudes da ordem de 0,90; quando os experimentos são balanceados e as incidências (sobrevivências) nos diferentes níveis dos efeitos fixos são aproximadamente iguais não existe grande prejuízo em se usar uma análise direta dos dados na escala observada; na análise de dados desbalanceados provenientes de diferentes experimentos com diferentes taxas de sobrevivência recomenda-se fortemente o uso de modelos não lineares e dos algoritmos apresentados.

1 Pesquisador da EMBRAPA, Caixa Postal 319, 83411-000 – Colombo – PR.
2 Professor do Departamento de Estatística – UFPR – Curitiba – PR.

▪ PALAVRAS-CHAVE: dados discretos, transformação de dados, REML, BLUP, distribuição logística, probito, componentes de variância.

Introdução

Variáveis categorizadas ou binomiais tais como a presença ou ausência de determinados atributos nos indivíduos não são bem descritas por modelos estatísticos lineares. Para estas variáveis, os modelos não lineares para dados dicotômicos podem ser mais apropriados.

A técnica de modelos lineares generalizados (MLG), desenvolvida por Nelder & Wedderburn (1972), permite a generalização ou flexibilização dos modelos lineares clássicos de variáveis contínuas, de forma que toda a estrutura para a estimação e predição em modelos lineares normais, pode ser estendida para os modelos não lineares. Os modelos lineares clássicos são, em verdade, casos especiais de modelos lineares generalizados.

Variáveis binomiais relevantes na área biológica são aquelas provenientes de experimentos do tipo dose-resposta, em que os indivíduos sobrevivem, ou não, em função da dosagem do elemento adverso (produto químico, nível de estresse etc). Neste caso, geralmente são atribuídos, por exemplo, os valores 1 para os indivíduos sobreviventes e zero para os mortos. Entretanto, a probabilidade de ocorrer uma resposta 1 (escore 1) é nula para valores altos da dosagem e unitária para valores baixos da dosagem, de forma que tal probabilidade é uma função estritamente decrescente da dosagem. Tal função matemática é a função sigmoidal, em que os modelos são não lineares nos parâmetros, de forma que é recomendada uma transformação da curva sigmóide em uma reta, a fim de que os procedimentos lineares possam ser aplicados na estimação dos parâmetros. Uma das transformações lineares mais recomendadas para os dados binomiais refere-se ao uso do modelo logístico (Demétrio, 1993; Rodrigues-Zas et al., 1997).

Um modelo linear generalizado é definido por: (i) um componente aleatório associado à distribuição da variável resposta; (ii) um componente sistemático linear nos parâmetros, denominado preditor linear ou estrutura linear do modelo; (iii) uma função de ligação, a qual combina o componente aleatório e o componente sistemático. No presente caso, em que o componente observacional tem distribuição binomial, pode-se assumir que a distribuição da tolerância (escala contínua) em função da dosagem tem distribuição logística e a função de ligação é a própria transformação logito a qual lineariza a função

sigmoidal. Pode-se, também, assumir que a tolerância (escala contínua) tem distribuição normal e a função de ligação que lineariza a distribuição é a função probito (Cordeiro, 1986; Demétrio, 1993).

No âmbito dos modelos mistos com efeitos fixos e aleatórios, o componente sistemático dos MLG refere-se aos próprios modelos lineares mistos clássicos exceto o termo de erro, sendo que, para o caso de variáveis binomiais, a função de ligação deve ser incorporada aos preditores (melhores preditores lineares não viciados – BLUP) dos efeitos aleatórios e estimadores (de máxima verossimilhança restrita – REML) dos componentes de variância dos efeitos aleatórios, conforme Gilmour et al. (1985) e Schall (1991). A função de ligação logito é a mais indicada para dados binomiais, pois refere-se a uma ligação canônica ou natural, ou seja, uma função de ligação especial para a qual existe uma estatística suficiente de mesma dimensão que β em um preditor linear $\eta = \theta = X\beta$ (Demétrio, 1993). A incorporação da função de ligação nas equações de modelos lineares mistos para a estimação dos componentes de variância e predição de variáveis aleatórias gera a denominação de modelo não linear (Thompson, 1990), devido à relação não linear que existe entre a escala de dosagem e a probabilidade de um indivíduo pertencer a uma determinada categoria da variável binária.

Na análise de dados binomiais desbalanceados, a metodologia de modelos lineares generalizados mistos deve ser a metodologia preferida (Thompson, 1990; Knuiman & Laird, 1990; Tempelman, 1998; Gianola, 2000). Assim, estudos neste sentido devem ser estimulados.

O presente trabalho tem como objetivos: apresentar os fundamentos teóricos e práticos da análise de modelos não lineares fixos e mistos associada a variáveis binomiais; apresentar estimadores, preditores e algoritmos para a análise dos referidos modelos não lineares; implementar análises completas envolvendo dados da variável sobrevivência de plantas de eucalipto (*Eucalyptus grandis*) e erva-mate (*Ilex paraguariensis*), avaliadas em dois experimentos.

Material e Métodos

Variáveis binomiais e funções de transformação

Em função da correlação (pois a variância $\mu(1-\mu)$ é completamente determinada pela média μ) existente entre médias e variâncias de variáveis binárias (Y), o uso das técnicas padrões de modelos esta-

tísticos lineares não é recomendado. No caso específico da variável sobrevivência, a mesma pode ser modelada em termos de uma escala base contínua, que é a escala da tolerância ou susceptibilidade em resposta à dosagem do ambiente adverso.

Desta forma, pode-se assumir que a tolerância tem uma distribuição logística, a qual possui forma similar a da distribuição normal.

A probabilidade μ, de uma planta sobreviver (receber o escore 1) é função decrescente da dosagem do ambiente adverso, obedecendo à função sigmóide dada por:

$$\mu = \frac{1}{1+e^{-\theta}} = \frac{e^{\theta}}{1+e^{\theta}}$$, em que θ é a variável na escala logística.

A transformação da curva sigmóide em uma reta, visando à aplicação de procedimentos lineares é realizada através da transformação ou função logito dada por:

$$\eta = log \left[\frac{\mu}{1-\mu}\right] = \theta$$

em que log é o logaritmo natural.

Assim, o logito de μ eqüivale a $\eta = \theta$, sendo que um modelo linear pode ser associado a η.

Substituindo-se os valores obtidos de θ na função sigmóide $[\mu = e^{\theta} /(1+e^{\theta})]$, produz-se μ no intervalo entre *0* e *1*.

Modelos lineares generalizados

Nelder & Wedderburn (1972) introduziram a idéia de modelos lineares generalizados visando permitir maior flexibilidade de análise. Tal idéia relaxa a suposição de que *Y* segue distribuição normal e permite *Y* seguir qualquer distribuição que pertença à família exponencial com um parâmetro, na forma canônica. As generalizações ocorrem em duas direções: (i) permitem que a esperança μ, de *Y*, seja uma função monotonicamente diferenciável do preditor linear $\eta = \sum x_i\beta_i$ de forma que $\mu = f(\eta) = f(\sum x_i\beta_i)$; (ii) ou, por inversão, $g(\mu) = \eta$, em que g é a função de ligação, a qual liga a média ao preditor linear.

Para dados binomiais, $0 < \mu < 1$, funções de ligação, tal qual o logito são utilizadas para satisfazer esta restrição natural. As transformações são importantes para: (i) estender a amplitude da variável analisada de $(0,1)$ para a reta real; (ii) fazer a variância constante através da amplitude dos efeitos fixos (na escala de tolerância). A função de ligação descreve, então, a relação existente entre o preditor linear (η) e o valor esperado μ de Y. No modelo linear clássico tem-se $\eta = \mu$ que é chamada de ligação identidade, sendo que esta ligação é adequada no sentido em que ambos η e μ podem assumir valores na reta real (Mc Cullagh & Nelder, 1989).

O modelo linear generalizado (MLG) é, então, definido por (Demétrio, 1993):

(a) um componente aleatório representado pelas variáveis aleatórias independentes Y_1, Y_2, ..., Y_n provenientes de uma mesma distribuição que faz parte da família exponencial na forma canônica;

(b) um componente sistemático ou determinístico linear nos parâmetros, chamado preditor linear,

$\underset{\sim}{\eta} = X \underset{\sim}{\beta}$, por exemplo, em que:

$\underset{\sim}{\eta}' = (\eta_1, \ \eta_2, \ ..., \ \eta_n)$

$\underset{\sim}{\beta}' = (\beta_1, \ \beta_2, \ ..., \ \beta_p)$: vetor de parâmetros,

$X' = (x_1, \ x_2, \ ..., \ x_n)$, sendo x'$_i$ vetor de variáveis explanatórias;

$x_i' = (x_{i1}, \ x_{i2}, \ ..., \ x_{ip}) \forall i$.

uma função de ligação $g(.)$ tal que:

$\eta_i = g \ (\mu_i)$, em que

$\mu_i = E \ (Y_i)$

Em resumo, na definição de um MLG, são fundamentais: (i) definição da distribuição da variável resposta; (ii) definição do preditor linear ou matriz do modelo; (iii) definição da função de ligação.

Segundo Cordeiro (1986), a palavra generalizado significa uma distribuição mais ampla do que a normal para a variável resposta e uma

função não-linear conectando a média desta variável com a parte determinística do modelo.

Estimação e predição em modelos não lineares

As distribuições a serem assumidas para a escala de tolerância e correspondentes funções de ligação devem ser capazes de transformar o intervalo $(0,1)$ em $(-\infty,\infty)$. Neste sentido, as distribuições logística, normal padrão e Gumbel (ou distribuição de valor extremo) para a tolerância e suas correspondentes funções de ligação denominadas logito, probito e complemento log-log são apropriadas para o modelo binomial. A função de ligação complemento log-log ($log[-log\ (1-\mu)]$) pode ser mais indicada para valores de μ próximos de zero.

Existe pouca diferença entre as distribuições normal e logística para a tolerância ou escala contínua (Cordeiro, 1986). Entretanto, a função de ligação logito é a mais usada para dados binomiais pois refere-se a uma ligação canônica ou natural, tendo uma interpretação mais simples.

Considerando a função de ligação logito e a variável sobrevivência, em que os indivíduos vivos recebem o valor *1* e os mortos o valor *0*, tem-se:

$\mu_i = Prob\ (Y_i = 1)$ = probabilidade de o indivíduo sobreviver;

$$g\ (\mu_i) = \eta_i = log\ \left[\frac{\mu_i}{1-\mu_i}\right]$$

$\eta = X\beta$ = modelo linear imposto ao logito, em que:

β = vetor de parâmetros na escala logito (escala de tolerância);

X = matriz de incidência para β.

A variância residual, ou seja, a variância das observações dado η_i *(ou* μ_i*)*, advém da amostragem binomial, ou seja:

$$Var\ (Y_i|\mu_i) = \mu_i\ (1-\mu_i)\ .$$

Modelos não lineares fixos

No caso da variável de trabalho definida pelo modelo $y = X\beta + Za + (y-\mu)\ g'(\mu)$ (que com erro normal e função de ligação identidade equivale ao modelo linear clássico $y = X\beta + Za + e$) de efeitos fixos,

onde os *a's* são tratados como parâmetros fixos, se cada Y_i é uma binomial independente (*1,* μ_i), Nelder & Wedderburn (1972) mostraram que as soluções para as estimativas de máxima verossimilhança de (β, *a*) são obtidas resolvendo-se iterativamente:

$$\begin{bmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'W & y^* \\ Z'W & y^* \end{bmatrix}, \text{ em que:}$$

$y^* = \eta + U\,(y - \mu)$ = variável de trabalho, a qual é uma combinação linear do preditor linear η e da discrepância entre os valores observados e ajustados;

$\mu = E\,(Y)$;

U = uma matriz diagonal com elementos $u_i = (\partial \eta_i / \partial \mu_i)$;

W = uma matriz diagonal com elementos

$w_i = (\partial \mu_i / \partial \eta_i)^2 / [\mu_i\ (1 - \mu_i)]$.

Especificamente para a ligação logito, tem-se:

$$\mu_i = \frac{e^{\eta_i}}{(1 + e^{\eta_i})}; \quad w_i = \mu_i (1 - \mu_i); \quad u_i = 1 / w_i.$$

Assim, $y_i^* = \eta_i + \dfrac{y_i - \mu_i}{\mu_i\ (1 - \mu_i)}$.

O procedimento iterativo é também denominado de mínimos quadrados ponderados iterativos (IWLS). Um algoritmo empregando o método numérico de Newton-Raphson é dado por:

1) Forneça os valores iniciais para $\hat{\mu}_i^{(o)}$ e;

2) Dado $\hat{\mu}_i^{(t)} e\ \hat{\eta}_i^{(t)}$, forme a variável dependente ajustada,

$\hat{y}_i^{*(t)} = \hat{\eta}_i^{(t)} + (y_i - \mu_i^{(t)})\ g'(\mu_i^{(t)})\ e\ w_i^{(t)} = \mu_i^{(t)}(1 - \mu_i^{(t)})$;

3) Reunindo β e *a* em um único vetor ϕ e *X* e *Z* em uma única matriz de incidência *Q*, obtenha $\hat{\phi}^{(t+1)}$ por:

$\hat{\phi}^{(t+1)} = (Q'\hat{W}^{(t)}Q)^{-1}\ Q'\hat{W}^{(t)}\ y^{*(t)}$, em que $\hat{W}^{(t)} = diag\ (w_i^{(t)})$.

Defina $\hat{\eta}^{(t+1)} = Q\hat{\phi}^{(t+1)}\ e\ \hat{\mu}^{(t+1)} = g^{-1}(\hat{\eta}^{(t+1)})$.

4) Repita os passos 2 e 3 até que se atinja a convergência segundo o critério desejado.

Conforme Firth (1991), os valores iniciais óbvios são $\hat{\mu}_i^{(o)} = y_i$, $\hat{\eta}_i^{(o)} = g\ (\mu_i^{(o)})$, em que y_i é a proporção observada.

Modelos não lineares mistos

Em $\eta = X\beta + Za$, define-se:

β e a = vetores de efeitos fixos e aleatórios, respectivamente;

X e Z = matrizes de incidência para β e a, respectivamente.

Todos os efeitos são assumidos como não correlacionados e os efeitos aleatórios são assumidos como tendo esperança zero e matriz de covariância G.

A variável de trabalho pode ser definida pelo modelo $y = \mu + (y - \mu)g'(\mu)$, em que $(y - \mu)g'(\mu)$ refere-se ao vetor de erros aleatórios.

No caso especial em que Y segue uma distribuição normal e $g(\mu)$ é uma ligação identidade obtém-se o modelo linear clássico misto $y = X\beta + Za + e$, em que $E(Y) = X\beta, Cov(a) = G, Cov(e) = R = I\sigma_e^2$ e, consequentemente,

$$Cov(Y) = ZGZ' + R = ZGZ' + I\sigma_e^2.$$

Para este modelo linear misto tradicional têm-se as seguintes equações de modelo misto para a estimação de β e predição de a:

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

Na situação em que os efeitos aleatórios são não correlacionados, os estimadores para os componentes de variância pelo método da máxima verossimilhança (*ML*) são dados por (Fellner, 1986; 1987):

$$\hat{\sigma}_a^2 = \frac{\hat{a}'\hat{a}}{q - tr\ C_{22}^{-1}/\sigma_a^2};$$

$$\hat{\sigma}_e^2 = \frac{(y - X\hat{\beta} - Z\hat{a})'(y - X\hat{\beta} - Z\hat{a})}{N - r(x) - q + tr\ C_{22}^{-1}/\sigma_a^2}\ \sigma_e^2.$$

Pelo método de máxima verossimilhança restrita (REML), os estimadores são dados por:

$$\hat{\sigma}_a^2 = \frac{\hat{a}'\hat{a}}{q - tr\ C^{22}/\sigma_a^2};$$

$$\hat{\sigma}_e^2 = \frac{(y - X\hat{\beta} - Z\hat{a})'(y - X\hat{\beta} - Z\hat{a})}{N - r(x) - q + tr\ C^{22}/\sigma_a^2}\ \sigma_e^2$$ em que:

C_{22} = partição da matriz dos coeficientes das equações de modelo misto, referentes aos efeitos aleatórios;

C^{22} = partição da inversa da matriz dos coeficientes das equações de modelo misto, referentes aos efeitos aleatórios;

q = número de níveis do efeito aleatório a;

$r(x)$ = posto da matriz X;

tr = operador traço matricial;

σ_e^2 e σ_a^2 = valores correntes ou atuais (obtidos na iteração anterior) de $\hat{\sigma}_e^2$ e $\hat{\sigma}_a^2$, respectivamente.

Estes estimadores são idênticos aos apresentados por Dempster et al. (1977) e Harville (1977), referentes ao algoritmo EM (Expectation – Maximization).

Para o caso em que Y segue uma distribuição binomial, tem-se o caso da estimação em modelos lineares generalizados com efeitos fixos e aleatórios, conforme Schall (1991).

A função de ligação logito aplicada aos dados y é linearizada, conforme a expansão em série de Taylor de primeira ordem fornecendo y^*, da seguinte forma:

$$y^* = g(y) = g(\mu) + (y - \mu)\ g'(\mu).$$

Assim tem-se:

$$y_i^* = \eta_i + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)} \qquad = log\left(\frac{\mu_i}{1-\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i(1-\mu_i)}.$$

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

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

O modelo $y^* = X\beta + Za + (y - \mu)g'(\mu)$, que define a variável de trabalho, tem a mesma estrutura de primeira e segunda ordem que o modelo $y = X\beta + Za + e$, de forma que os algoritmos de estimação e predição para o caso normal podem ser adaptados, apenas substituindo y por y^* e $Cov(e) = R$ por $Cov[(y-\mu)g'(\mu)] = W^{-1}\sigma_e^2$.

Assim, têm-se as seguintes equações de modelo misto:

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

S^{-1} = matriz com termos diagonais dado por

$$\mu_i\ (1-\mu_i)\frac{1}{\sigma_{e_L}^2};$$

$\sigma_{e_L}^2$ = variância residual na escala contínua de tolerância ("liability");

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

Os estimadores dos componentes de variância pelo método ML são dados por:

$$\hat{\sigma}_{a_L}^2 = \frac{\hat{a}'_L\,\hat{a}_L}{q - tr\ C_{22}^{-1}/\sigma_{a_L}^2};$$

$$\hat{\sigma}^2_{e_L} = \frac{(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)'\;\; S^{-1}(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)}{N - r(\,x\,) - q + tr\;\; C_{22}^{-1} / \sigma^2_{a_L}}\;\; \sigma^2_{e_L}.$$

Pelo método REML, os estimadores são dados por:

$$\hat{\sigma}^2_{a_L} = \frac{\hat{a}'_L\, \hat{a}_L}{q - tr\;\; C^{22} / \sigma^2_{a_L}};$$

$$\hat{\sigma}^2_{e_L} = \frac{(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)'\;\; S^{-1}(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)}{N - r(\,x\,) - q + tr\;\; C^{22} / \sigma^2_{a_L}}\;\; \sigma^2_{e_L}.$$

No caso em que *a* refere-se a um vetor de valores genéticos aditivos tem-se que $Cov(\,a\,) = G = A\sigma^2_{a_L}$, em que *A* é a matriz de correlação genética aditiva entre os indivíduos e $\sigma^2_{a_L}$ é a variância de a_L. Neste caso, os estimadores REML são dados por:

$$\hat{\sigma}^2_{a_L} = \frac{\hat{a}'_L\, A^{-1}\; \hat{a}_L}{q - tr\;\; (\, A^{-1}\;\; C^{22}\,) / \sigma^2_{a_L}};$$

$$\hat{\sigma}^2_{e_L} = \frac{(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)'\;\; S^{-1}(\, y - X\hat{\beta}_L - Z\hat{a}_L\,)}{N - r(\,x\,) - q + tr\;\; (\, A^{-1}\;\; C^{22}\,) / \sigma^2_{a_L}}\;\; \sigma^2_{e_L}.$$

O processo iterativo é repetido até a convergência, com o valor predito de $\hat{\theta} = X\hat{\beta}_L + Z\hat{a}_L$, transformado, usando a função de ligação para obtenção do novo valor predito de μ através de $\hat{\mu} = \frac{e^{\hat{\theta}}}{1 + e^{\hat{\theta}}}$, o qual é utilizado para atualização de S^{-1} *e y**.

Em resumo, o processo de estimação envolve:

(a) estimação de $\mu = n_1 / N$, em que n_1 é o número de indivíduos que recebem o escore *1*, dentre *N* indivíduos avaliados;
(b) obtenção de *y**, a partir de *y* e μ;

(c) estimação de $\hat{\beta}_L$ *e* $\hat{a}_L$, dados os valores atuais ou correntes de μ, $\sigma^2_{e_L}$ *e* $\sigma^2_{a_L}$;

(d) obtenção de $\hat{\sigma}^2_{e_L}$ *e* $\hat{\sigma}^2_{a_L}$ iterativamente e, após a convergência, proceder à obtenção atualizada de $\hat{\beta}_L$ e $\hat{a}_L$;

(e) obtenção de $\hat{\eta} = \hat{\theta} = X\hat{\beta}_L + Z\hat{a}_L$;

(f) obtenção de novo valor predito de μ, usando a função de ligação, através de $\hat{\mu}_L = \dfrac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}$ (neste passo, a variável volta ao intervalo *(0,1)*);

(g) atualização de S^{-1} via $S^{-1} = \hat{\mu}_1(1-\hat{\mu}_1)\ \dfrac{1}{\hat{\sigma}^2_{eL_1}}$; e de y^* via

$$y^* = log\ \left(\frac{\hat{\mu}_1}{1-\hat{\mu}_1}\right) + \frac{y-\hat{\mu}_1}{\hat{\mu}_1(1-\hat{\mu}_1)} = \hat{\theta} + \frac{y-\hat{\mu}_1}{\hat{\mu}_1(1-\hat{\mu}_1)};$$

(h) voltar ao passo (c), enquanto não se atingir a convergência.

É interessante notar que este algoritmo é essencialmente hierárquico, havendo a necessidade de convergência no passo (d). Algoritmo hierárquico similar (mas não envolvendo variáveis binomiais) para modelos não lineares foi apresentado por Gregoire & Schabemberger (1996).

Os algoritmos apresentados são do tipo EM. Outros algoritmos relatados em literatura para a análise de modelos lineares generalizados mistos (GLMM) são o PQL (*penalized quasi-likelihood*) de Breslow & Clayton (1993) e o IRREML (*iterated re-weighted REML*) de Engel & Keen (1994). Para GLMM's, os procedimentos de estimação e predição de Schall (1991), Breslow & Clayton (1993) e Engel & Keen (1994) são eqüivalentes (Keen & Engel, 1996). Sob suposições de normalidade, para componentes de variância fixos, o procedimento IRREML é equivalente ao procedimento Bayesiano MAP (máximo a posteriori), de forma que o IRREML fornece uma alternativa, não Bayesiana, de derivação do MAP (Engel & Keen, 1996).

O uso de modelos não lineares é especialmente indicado, quando os indivíduos a serem avaliados pertencem a diferentes níveis dos efeitos fixos, com diferentes valores de incidência para a variável em análise. Neste caso, a obtenção do coeficiente de herdabilidade

$[\, h^2 = \sigma_a^2 /(\sigma_a^2 + \sigma_e^2)]$ na escala base contínua, através da transformação de probito é inapropriada pois a transformação é função da incidência, a qual difere para os indivíduos dos diferentes níveis dos efeitos fixos. Uma vez que as diferenças entre os níveis dos efeitos fixos correspondem a mudanças na escala base, uma função para ligação das mudanças nas duas escalas necessita ser incorporada nas equações de modelo misto. Para cômputo de y^* e S^{-1}, diferentes valores de μ_i devem ser computados para os diferentes níveis dos efeitos fixos.

Teste de Ajustamento de um Modelo

Além de sua utilidade na estimação, o princípio da verossimilhança também permite comparar a adequabilidade de vários modelos, desde que os mesmos tenham uma estrutura hierárquica ou aninhada. Dados dois modelos U e V com máximos das funções de verossimilhança (restrita ou não) $L(U)$ e $L(V)$ e correspondentes números de parâmetros n_u e n_v, $D = -2\ log\ (L(U)/L(V))$ possui, aproximadamente e sob certas condições de regularidade, distribuição proporcional a uma χ^2 com $n_v - n_u$ graus de liberdade (assumindo U como hierárquico ou um caso especial de V). Quando o modelo V é saturado e eqüivale a uma exata reprodução dos dados, D recebe o nome especial de *deviance* do modelo U (Crosbie & Hinch, 1985). Sob normalidade de erros, a *deviance* eqüivale a soma de quadrados dos resíduos, a qual segue uma distribuição múltipla de uma χ^2.

O teste de ajustamento de um modelo linear generalizado é feito através da estatística S_p definida por Nelder & Wedderburn (1972) como $S_p = 2(\hat{L}_n - \hat{L}_p)$ em que S_p é a própria *deviance* e $\hat{L}_n$ e $\hat{L}_p$ são os máximos do logaritmo da função de verossimilhança para os modelos saturado e corrente (sob pesquisa). Assim, o modelo saturado é usado como base de medida do ajuste de um modelo sob pesquisa (Demétrio, 1993).

A *deviance* é sempre maior ou igual a zero e à medida que entram variáveis explanatórias (ou covariáveis) no componente sistemático, a *deviance* decresce até se tornar zero para o modelo completo ou saturado. Um modelo bem ajustado aos dados apresenta uma verossimilhança máxima grande e uma pequena *deviance*. Quanto melhor for o ajuste do modelo aos dados tanto menor será o valor de S_p (Cordeiro, 1986; Demétrio, 1993).

Para testar a adequação de um modelo linear generalizado, a *deviance* (com n – p graus de liberdade, em que p é o posto da matriz do modelo) deve ser comparada com alguma distribuição de probabilidade de referência. Não se conhece, em geral, a distribuição de probabilidade da *deviance* e apenas resultados assintóticos são disponíveis. Quando o modelo em investigação é verdadeiro, a *deviance* não é, em geral, distribuída como χ^2_{n-p}. Entretanto, na prática, comtenta-se em testar um modelo linear generalizado, sem muito rigor, comparando-se o valor S_p com o valor crítico de χ^2 com (n - p) graus de liberdade da distribuição χ^2 a um nível de significância α. Assim, se $S_p \geq \chi^2_{n-p;\alpha}$ o modelo é rejeitado e se $S_p \leq \chi^2_{n-p;\alpha}$ o modelo é aceito (Cordeiro, 1986; Demétrio, 1993).

Conforme Mc Cullagh & Nelder (1989), $L(\hat{\mu}, \phi; y)$ equivale ao logaritmo da função verossimilhança maximizada para um valor fixado do parâmetro de dispersão ϕ. Para dados binários, ϕ é assumido conhecido e igual a *1*. Entretanto, na prática, os dados podem exibir algum grau de superdispersão ($\phi > 1$) ou subdispersão ($\phi < 1$). Segundo Lindsey (1997), o fato de ϕ ser diferente de *1* pode ser explicado pelo não atendimento das suposições (eventos independentes, por exemplo) assumidas para que determinado conjunto de dados tenha distribuição binomial. Neste caso, quebra-se a dependência original existente entre a média (μ) e variância [μ (*1*-μ)] de uma distribuição Bernoulli. Em presença de sub ou superdispersão, uma correção visando a realização de inferências aproximadas, refere-se à multiplicação da variância μ (*1*-μ) por uma estimativa $\hat{\phi}$ de ϕ (Firth, 1991; Paul & Islam, 1995).

No caso em que ϕ é desconhecido admite-se que ele é o mesmo para todas as observações. A estimação de ϕ é necessária para obtenção dos erros – padrões das estimativas. Segundo Lindsey (1997), a *deviance* média para o modelo maximal pode propiciar uma estimativa para o parâmetro de dispersão ϕ. Esta *deviance* média é denominada fator de heterogeneidade ou fator de heterogeneidade de variância (Lindsey, 1997; Gilmour et al., 2000). Tal fator de heterogeneidade pode ser calculado por $\hat{\phi} = D / gl$, em que D é a *deviance* como uma medida da falta de ajuste da distribuição binomial e gl refere-se à diferença entre o número de parâmetros dos dois modelos para calcular a *deviance*.

Para dados com distribuição normal, uma análise de variância baseada nas somas dos quadrados para uma seqüência de modelos pode ser utilizada para pesquisar o ajuste de vários modelos e os efeitos das variáveis explicativas, fatores e suas interações. Para os modelos lineares generalizados, um procedimento análogo (Nelder & Wedderburn, 1972), baseado nos valores de *deviance* e não em soma de Quadrados dos resíduos, tem sido utilizado para construção de uma análise de *deviance*.

A análise de *deviance* é uma generalização da análise de variância para modelos lineares generalizados, visando obter, a partir de uma sequência de modelos, cada um incluindo mais termos do que os anteriores, os efeitos de fatores, covariáveis e suas interações. Dada uma sequência de modelos encaixados, utiliza-se a *deviance* como uma medida de discrepância do modelo e forma-se uma tabela de diferenças de *deviances*. Considerando os modelos M_p e M_q (p<q) com p e q parâmetros independentes, pode-se usar $\frac{(n-q)(S_p - S_q)}{(q-p)S_q} \cap F_{q-p,\ n-q}$ como aproximação para teste de ajuste dos modelos (Demétrio, 1993).

Aplicação a dados experimentais

Foram considerados dados de dois experimentos conduzidos pela Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA. Um deles refere-se a um teste de progênies de *Eucalyptus grandis*, instalado no delineamento de blocos ao acaso com 33 tratamentos (progênies), 6 repetições e 6 plantas por parcela. O outro refere-se a um teste de progênies de erva-mate (*Ilex paraguariensis*) instalado no delineamento de blocos ao acaso com 20 tratamentos (progênies), 8 repetições e 6 plantas por parcela. Em ambos os experimentos foi avaliada a variável sobrevivência das plantas, nas idades de dois anos para o experimento com erva mate e cinco anos para o experimento com eucalipto. Na avaliação da sobrevivência foram atribuídos os escores 1 para as plantas vivas e 0 para as mortas.

O modelo misto associado às observações dos dois experimentos, quando o erro é normal e a função de ligação identidade, eqüivale a:

$y = X\beta + Za + e$, em que:
y = vetor de dados;
β = vetor dos efeitos fixos de blocos;
a = vetor paramétrico dos efeitos aleatórios de plantas;

e = vetor de erros aleatórios.
A estrutura de médias e variâncias do modelo eqüivale a:

$$E\begin{bmatrix} y \\ a \\ e \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \\ 0 \end{bmatrix}; \qquad Var\begin{bmatrix} y \\ a \\ e \end{bmatrix} = \begin{bmatrix} V & ZA\sigma_a^2 & I\sigma_e^2 \\ A\sigma_a^2 Z' & A\sigma_a^2 & 0 \\ I\sigma_e^2 & 0 & I\sigma_e^2 \end{bmatrix}, \text{ em que:}$$

A = matriz de correlação genética aditiva entre as plantas.
I = matriz identidade.

Este modelo é parcimonioso pois não contém o efeito de parcela, o qual mostrou-se não significativo pelo teste da razão de verossimilhança, aplicado considerando os máximos do logaritmo das funções de verossimilhança associados à estimação de efeitos fixos e componentes de variância e predição dos efeitos aleatórios, de modelos incluindo e não incluindo o efeito de parcela.

As análises foram realizadas empregando-se os modelos descritos anteriormente e utilizando-se o software ASREML (Gilmour et al., 2000), o qual emprega o algoritmo AI-REML (Average Information REML) desenvolvido por Gilmour et al. (1995) e Johnson & Thompson (1995), e a metodologia de Schall (1991).

Resultados e Discussão

Experimento 1 – Eucalipto

Na Tabela 1 são apresentados os resultados referentes aos efeitos fixos de blocos, para o experimento com eucalipto, considerando a função de ligação logito, com os resultados referentes às médias de blocos apresentados na escala original. Tais resultados permitem inferir sobre a discrepância entre as incidências nos diferentes níveis dos efeitos fixos, a qual relaciona-se com a superioridade da abordagem de modelos não lineares sobre a de modelos lineares (Foulley et al., 1990).

Tabela 1 – Estimativas para as médias de blocos, para a variável sobrevivência no experimento de *Eucalyptus grandis*

Blocos	Média ± desvio padrão
1	0,8990 ± 0,0256
2	0,8687 ± 0,0256
3	0,8535 ± 0,0256
4	0,8939 ± 0,0256
5	0,8737 ± 0,0256
6	0,8889 ± 0,0256

Significância dos efeitos de blocos: F = 0,57, o qual não é significativo.

Verifica-se que os blocos apresentaram médias de magnitudes próximas, as quais são estatisticamente iguais (Tabela 1).

Na Tabela 2 são apresentados os resultados referentes aos componentes de variância associados aos efeitos aleatórios de plantas e residuais, considerando a técnica de modelos lineares generalizados mistos, usando diferentes funções de ligação.

Tabela 2 – Estimativas (pelo método REML) dos componentes de variância referentes aos efeitos aleatórios de plantas (σ_a^2) e do erro (σ_e^2), bem como do coeficiente de determinação (herdabilidade) dos efeitos aleatórios de plantas $[h^2 = \sigma_a^2 /(\sigma_a^2 + \sigma_e^2)]$, associados a diferentes modelos de análise de acordo com a função de ligação empregada. Experimento de *Eucalyptus grandis*

Distribuição da Tolerância	Função de Ligação	$(\hat{\sigma}_a^2)$	$(\hat{\sigma}_e^2)$	$(\hat{h}^2)$
Logística	Logito	1,60922	2,08298	0,43584
Normal Padrão	Probito	0,45854	0,65609	0,41138
Valor Extremo (Gumbel)	Complemento log-log	0,28659	1,42999	0,16696
Distribuição da Escala Original	Função de Ligação	$(\hat{\sigma}_a^2)$	$(\hat{\sigma}_e^2)$	$(\hat{h}^2)$
Binominal	Identidade	0,01881	0,08701	0,1776
Normal	Identidade	0,01834	0,08795	0,1725

Verifica-se que a análise dos dados na própria escala observada ou original, assumindo as distribuições normal ou binomial (ambas com

função de ligação identidade) conduziram a estimativas de h^2 próximas (na faixa de 0,17 a 0,18) (Tabela 2). É importante relatar que a função de ligação identidade não é indicada pois pode levar a proporções fora do intervalo (0, 1). Por outro lado, assumindo as distribuições logística, normal padrão e de Gumbel para a variável na escala transformada (usando as funções de ligação logito, probito e complemento log-log, respectivamente), as estimativas de h^2 obtidas foram de aproximadamente 0,44; 0,41 e 0,17, respectivamente (Tabela 2).

Constata-se que o uso das funções de ligação logito e probito são altamente favoráveis pois conduziram a maior precisão (determinação) dos efeitos aleatórios de plantas, conforme revelado pelas estimativas de h^2. Este fator é muito relevante tendo em vista que o objetivo do experimento é a seleção das plantas com os maiores efeitos aleatórios preditos. Por outro lado, a transformação usando o complemento log-log mostrou-se inadequada pois conduziu a h^2 de mesma magnitude que a análise dos dados na escala observada.

Os resultados obtidos pelo uso das funções de ligação logito e probito foram similares, conforme tem sido verificado em outros experimentos (Villalobos & Garrick, 1999). Entretanto, pode-se optar pelas predições obtidas pelo uso da função logito, tendo em vista a maior estimativa de h^2 obtida (0,44 contra 0,41) e também o fato de ser a ligação canônica e, portanto, a mais recomendada para dados binomiais.

Quando a sobrevivência média é a mesma em todos os níveis dos efeitos fixos (blocos) o coeficiente de determinação h^2 na escala normal padrão (h_N^2) pode ser obtido a partir do h^2 na escala observada (h_B^2) pela expressão (Thompson, 1990): $h_N^2 = h_B^2 \ [\mu \ (1-\mu)] / Z^2$, em que μ é a sobrevivência média e Z é a ordenada da curva normal padrão no ponto igual a μ. Neste experimento, $h_B^2 = 0{,}1725$ *e* μ = *0,8796* e, portanto, *Z = 0,20121 e* $h_N^2 = 0{,}4512$. Este valor difere (superestima) um pouco do valor mais realístico (0,41138), o qual foi obtido considerando que as sobrevivências diferem através dos blocos. Assim, apesar de as médias dos blocos não terem apresentado diferenças estatisticamente significativas, o uso da técnica de modelos lineares generalizados mostrou-se ainda vantajosa.

Os resultados referentes às predições das médias de tratamentos (progênies ou genitores) são apresentados na Tabela 3, considerando três abordagens: como efeitos fixos na escala logística, como efeitos

aleatórios na escala observada e como efeitos aleatórios na escala logística.

Assumindo-se as predições na escala logística como as mais realísticas e adequadas verifica-se que a predição direta com dados na escala observável conduziu a resultados bastante similares aos obtidos pelo procedimento ótimo (Tabela 3). Este fato já era esperado tendo em vista a não significância do efeito de blocos e o balanceamento do experimento. Conforme Foulley et al. (1990), a superioridade dos preditores não lineares aumenta com o aumento do desbalanceamento dos experimentos e com o aumento da discrepância entre as incidências nos diferentes níveis dos efeitos fixos.

Tabela 3 – Médias de tratamentos estimadas (EB) e preditas (PB) na escala observada (binomial) e preditas na escala logística e transformadas para a escala original (PL), no experimento de *Eucalyptus grandis*

Tratamento (Genitor)	EB	PB	PL
1	0,9444	0,9197	0,9202
2	0,8333	0,8510	0,8554
3	0,7778	0,8166	0,8187
4	0,9167	0,9025	0,9054
5	0,9167	0,9025	0,9054
6	0,8889	0,8853	0,8896
7	0,9722	0,9369	0,9339
8	0,9444	0,9197	0,9202
9	0,8889	0,8853	0,8896
10	0,8611	0,8681	0,8729
11	0,9444	0,9197	0,9202
12	0,8611	0,8681	0,8729
13	0,8333	0,8510	0,8554
14	0,9444	0,9197	0,9202
15	0,9444	0,9197	0,9202
16	0,9722	0,9369	0,9339
17	0,8056	0,8338	0,8373
18	1,0000	0,9541	0,9463
19	0,8611	0,8681	0,8729

20	0,6944	0,7650	0,7599
21	0,9444	0,9197	0,9202

Continuação

Tratamento (Genitor)	EB	PB	PL
22	0,9722	0,9369	0,9339
23	0,9167	0,9025	0,9054
24	0,8611	0,8681	0,8729
25	0,7778	0,8166	0,8187
26	0,7500	0,7994	0,7995
27	0,6667	0,7478	0,7395
28	0,8611	0,8681	0,8729
29	0,7778	0,8166	0,8187
30	0,9444	0,9197	0,9202
31	1,0000	0,9541	0,9463
32	0,8056	0,8338	0,8373
33	0,9444	0,9197	0,9202

Experimento 2 – Erva-mate

Na Tabela 4 são apresentados os resultados relativos aos efeitos fixos de blocos, para o experimento com erva-mate, considerando a função de ligação logito, com os resultados das médias de blocos apresentados na escala original.

Tabela 4 – Estimativas para as médias de blocos, para a variável sobrevivência no experimento de erva-mate (*Ilex paraguariensis*)

Blocos	Média $\pm$ desvio padrão
1	0,8250 $\pm$ 0,0386
2	0,8833 $\pm$ 0,0386
3	0,8083 $\pm$ 0,0386
4	0,7333 $\pm$ 0,0386
5	0,8833 $\pm$ 0,0386
6	0,8583 $\pm$ 0,0386
7	0,8833 $\pm$ 0,0386

8	0,6750 ± 0,0386

Significância dos efeitos de blocos: F = 16,22, o qual é altamente significativo.

Constata-se que os blocos apresentaram médias estatisticamente diferentes entre si (Tabela 4).

Na Tabela 5 são apresentados os resultados relativos aos componentes de variância associados aos efeitos aleatórios de plantas e do erro, considerando a técnica de modelos lineares generalizados mistos, usando diferentes funções de ligação.

Conforme verificado também para o experimento com *Eucalyptus grandis*, a análise direta dos dados na própria escala observada (assumindo distribuição normal e função de ligação identidade) e o uso das funções de ligação identidade (assumindo distribuição binomial) e complemento log-log não conduziram a resultados satisfatórios em termos de h^2, produzindo as menores magnitudes das estimativas deste parâmetro. Por outro lado, o uso das funções de ligação logito e probito levou a resultados similares e de maior magnitude para h^2, mostrando-se mais adequadas (Tabela 5).

Tabela 5 – Estimativas (pelo método REML) dos componentes de variância referentes aos efeitos aleatórios de plantas (σ_a^2) e do erro (σ_e^2), bem como do coeficiente de determinação (herdabilidade) dos efeitos aleatórios de plantas $[h^2 = \sigma_a^2 /(\sigma_a^2 + \sigma_e^2)]$, associados a diferentes modelos de análise de acordo com a função de ligação empregada. Experimento de erva-mate (*Ilex paraguariensis*).

Distribuição da Tolerância	Função de Ligação	$(\hat{\sigma}_a^2)$	$(\hat{\sigma}_e^2)$	$(\hat{h}^2)$
Logística	Logito	1,22625	2,37018	0,3410
Normal Padrão	Probito	0,37500	0,71875	0,3429
Valor Extremo (Gumbel)	Complemento log-log	0,25092	1,45674	0,1469
Distribuição da Escala Original	Função de Ligação	$(\hat{\sigma}_a^2)$	$(\hat{\sigma}_e^2)$	$(\hat{h}^2)$
Binominal	Identidade	0,02305	0,12569	0,1550
Normal	Identidade	0,02728	0,11734	0,1886

A obtenção de h^2 na escala normal padrão, considerando a sobrevivência média μ = *0,81873*, pela expressão: $h_N^2 = h_B^2 \ [\mu\ (1-\mu)] / Z^2$, conduziu a $\hat{h}_N^2 = 0{,}4026$. Este valor é

bastante superestimado em relação ao valor mais realístico de *0,3429*, o qual foi obtido considerando que a sobrevivência difere ao longo dos blocos. Estes resultados, em comparação com os resultados obtidos para o experimento com eucalipto, permitem confirmar que o uso da abordagem de modelos lineares generalizados mistos é tanto mais recomendada quanto maiores forem as diferenças entre as incidências da variável através dos níveis dos efeitos fixos.

Os resultados relativos às predições dos efeitos de tratamentos (progênies ou genitores) são apresentados na Tabela 6.

Tabela 6 – Médias de tratamentos estimadas (EB) e preditas (PB) na escala observada (binomial) e preditas na escala logística e transformadas para a escala observável (PL), no experimento de erva-mate *(Ilex paraguariensis)*

Tratamento (Genitor)	EB	PB	PL
1	0,9375	0,9023	0,9044
2	0,8333	0,8290	0,8415
3	0,8542	0,8436	0,8550
4	0,9375	0,9023	0,9044
5	0,8333	0,8290	0,8415
6	0,7917	0,7996	0,8135
7	0,8125	0,8143	0,8277
8	0,6458	0,6970	0,7061
9	0,9375	0,9023	0,9044
10	0,6667	0,7117	0,7222
11	0,8958	0,8730	0,8807
12	0,7083	0,7410	0,7536
13	0,8750	0,8583	0,8681
14	0,8333	0,8290	0,8415
15	0,8958	0,8730	0,8807
16	0,8333	0,8290	0,8415
17	0,8125	0,8143	0,8277
18	0,8125	0,8143	0,8277
19	0,5833	0,6530	0,6571
20	0,8750	0,8583	0,8681

Verificam-se predições similares pelos procedimentos PB (escala observável) e PL (escala logística) também para o experimento com erva-mate (Tabela 6). A consideração dos efeitos de tratamentos como fixos conduziram a superestimativas das medias de tratamentos para ambos os experimentos (Tabelas 3 e 6).

De maneira genérica, pode-se inferir que o uso da metodologia de modelos não lineares mistos associado à técnica de modelos lineares generalizados é um procedimento conceitualmente mais adequado. Na pior das hipóteses, tal procedimento conduz a resultados similares aos obtidos pela abordagem linear normal. Na análise simultânea de vários experimentos desbalanceados, o uso dos modelos não lineares é fortemente recomendada, em relação ao uso dos modelos lineares e posterior obtenção de h_N^2 via transformação de probito usando um valor médio da sobrevivência através dos experimentos.

No presente trabalho, as estimativas do parâmetro de dispersão ϕ eqüivaleram a aproximadamente *0,90* para ambos os experimentos. Tendo em vista que para dados binários tal parâmetro é assumido como conhecido e igual a *1*, pode-se inferir que o grau de subdispersão foi pequeno e passível de negligência.

Conclusões

- os estimadores, preditores e algoritmos apresentados permitem uma acurada predição de efeitos aleatórios e uma precisa estimação de componentes de variância associados a variáveis binomiais, através do uso da técnica de modelos lineares generalizados mistos;
- as funções de ligação logito e probito mostraram-se adequadas na análise dos dados de sobrevivência de plantas de espécies perenes;
- as funções de ligação complemento log-log e identidade mostraram-se inadequadas aos dados binomiais analisados;
- a função de ligação logito mostrou-se a mais adequada, pois conduziu a coeficientes de determinação (h^2) dos efeitos aleatórios similares aos obtidos pela função probito e refere-se à base canônica ou natural;
- as estimativas do parâmetro de dispersão ou fator de heterogeneidade apresentaram magnitudes da ordem de *0,90*, indicando uma pequena subdispersão a qual é passível de negligência;

- quando os experimentos são balanceados e as incidências (sobrevivências) nos diferentes níveis dos efeitos fixos são aproximadamente iguais não existe grande prejuízo em se usar uma análise direta dos dados na escala observada;
- na análise de dados desbalanceados provenientes de diferentes experimentos com diferentes sobrevivências recomenda-se fortemente o uso de modelos lineares generalizados mistos e dos algorítmos apresentados.

RESENDE, M. D. V. de; BIELE, J. Prediction and estimation in mixed non-linear models for binomial traits using generalized linear models. *Rev. Mat. Estat.* (São Paulo), v.20, p.39-65, 2002.

▪ *ABSTRACT: Binomial traits such as the presence or absence of atributes in individuals are not well described by linear statistical models. For this kind of variables the non-linear mixed model methodology associated to the generalized linear models technique is the preferred approach. This paper presents theoretical and practical aspects of the mixed and fixed non-linear models for binomial variables. Data concerning to survival of two perennial plants species (Eucalyptus grandis and Ilex paraguariensis) were used. The following conclusions were obtained: the estimators, predictors and algorithms presented provide acurate and precise prediction of random variables and estimation of variance components, respectively, for binomial variables through the generalized linear mixed model (GLMM) technique; the logit and probit link functions performed very well; the identity and complementary log-log link functions did not provide good results; the logit link functions was preferred as it provide the highest coefficient of determination (h^2) for random effects; the dispersion parameter presented estimates about 0.90, showing the adequacy of the binomial distribution in modelling the data; when estimates for fixed effects don't differ through its levels, direct analysis in observed scale can be used; for unbalanced data and different estimates of fixed effects through its levels the GLMM approach should be used.*

▪ *KEYWORDS: discrete data, data transformation, REML, BLUP, logistic distribution, probit, variance components.*

Referências

BRESLOW, N. E.; CLAYTON, D.G. Aproximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, v.88, p.9-25, 1993.

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

CROSBIE, S. F.; HINCH, G.N. An intuitive explanation of generalized linear models. *N. Z. J. Agric.Res.*, v.28, p.19-29, 1985.

DEMÉTRIO, C. G. B. *Modelos lineares generalizados na experimentação agronômica*. Porto Alegre: Universidade Federal do Rio Grande do Sul., 1993. 125p.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J.R. Stat. Soc.*, v.39, p.1-38, 1977.

ENGEL, B.; KEEN, A. A simple approach for the analysis of generalized linear mixed models. *Stat. Neerlandica*, v.48, n.1, p.1-22, 1994.

ENGEL, B.; KEEN, A. An introduction to generalized linear mixed models. In: INTERNATIONAL BIOMETRIC CONFERENCE, 18., 1996, Boston. *Invited Papers.*. p.125-135.

FELLNER, W. H. Robust estimation of variance components. *Technometrics*, v.28, p.51-60, 1986.

FELLNER, W. H. Sparce matrix and the estimation of variance components by likelihood methods. *Commun. Stat.: Theor. Meth.*, B, v.16, p.439-463, 1987.

FIRTH, D. Generalized linear models. In: HINKLEY, D. V.; REID, N.; SNELL, E. J. *Statistical Theory and Modelling*. London: Chapman & Hall, , 1991. p.55-82.

FOULLEY, J. L.; GIANOLA, D.; IM, S. Genetic evaluation for discrete polygenic traits in animal breeding. In: GIANOLA, D.; HÁMMOND, K. (Ed.). *Advances in statistical methods for genetic improvement of livestock*. Berlin: Springer-Verlag, 1990. p.361-409.

GIANOLA, D. Statistics in animal breeding. *J. Am. Stat. Assoc.*, v.95, n.449, p.296-299, 2000.

GILMOUR, A.R.; ANDERSON, R.D.; RAE, A.L. The analysis of binomial data by a generalized linear mixed model. *Biometrika*, v.72, p.593-599, 1985.

GILMOUR, A. R.; THOMPSON, R., CULLIS, B.R. Average information REML: An efficient algorithm for parameter estimation in linear mixed models. *Biometrics*, v.51, p.1440-1450, 1995.

GILMOUR, A. R.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. ASREML. *Reference Manual*. Orange: NSW Agriculture, 2000. 218p.

GREGOIRE, T. G.; SCHABENBERGER, O. Nonlinear mixed-effects modeling of cumulative bole volume with spatially correlated within-tree data. *J. Agric. Biol. Environm. Stat*., v.1, n.1, p.107-119, 1996.

HARVILLE, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc*., v.72, p.320-328, 1977.

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci*., v.78, p.449-456, 1995.

KEEN, A.; ENGEL, B. Analysis of a mixed model for ordinal data by iterative re-weighted REML. *Stat. Neerlandica*, v.51, n.2, p.129-144, 1997.

KNUIMAN, M. W.; LAIRD, N. M. Parameter estimation in variance component models for binary response data. In: GIANOLA, D.; HAMMOND, K. (Ed.). *Advances in statistical methods for genetic improvement of livestock*. Berlin: Springer-Verlag,. 1990. p.177-206.

LINDSEY, J. K. *Applying generalized linear models*. New York: Springer-Verlag, 1997. 256p.

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

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *J. R. Stat. Soc., Series* A, v.135, p.370-384, 1972.

PAUL, S. R.; ISLAM, A. S. Analysis of proportions in the presence of over-/under dispersion. *Biometrics*, v.51, n.4, p.1400-1410, 1995.

RODRIGUES-ZAS, S.L; GIANOLA, D.; SHOOK, G. E. Factors affecting susceptibility to intramammary infection and mastitis: an approximate Bayesian analysis. *J. Dairy Sci.*, v.80, p.75-85, 1997.

SCHALL, R. Estimation in generalized linear models with random effects. *Biometrika*, v.78, p.719-727, 1991.

TEMPELMAN, R.J. Generalized linear mixed models in dairy cattle breeding. *J. Dairy Sci.*, v.81, p.1428-1444, 1998.

THOMPSON, R. Generalized linear models and applications to animal breeding. In: GIANOLA, D.; HAMMOND, K. (Ed.). *Advances in statistical methods for genetic improvement of livestock*. Berlin: Springer-Verlag, 1990. p.312-328.

VILLALOBOS, N. L.; GARRICK, D. J. Genetic parameters estimates for lamb survival in Romney sheep. *N. Z. Soc. Anim. Prod*., v.29, p.41, 1999.

Recebido em 11.09.2000