

Aplicações dos modelos lineares mistos na pesquisa agropecuária

Mixed linear models in the research agriculture and livestock

Érica M Pereira^{1*}; Geraldo A Gravina¹ e José Tarcísio L Thiébaud³

1. Estudante de Pós-Graduação - LEAG/CCTA/UENF, Campos, RJ, ericamirre@hotmail.com; 2. Professor Associado - LEAG/CCTA/UENF, Campos, RJ; 3; Professor Associado - LEAG/CCTA/UENF, Campos, RJ.

*Autor para correspondência: ericamirre@hotmail.com

Resumo Os modelos mistos são tradicionalmente estudados com o uso do procedimento da análise de variância, quando os componentes de variância são estimados pela solução de um sistema linear de equações. Para a análise dos modelos lineares mistos, alguns pontos são relevantes: a predição dos efeitos aleatórios, testes e estimação dos componentes de variância, estimação e testes de hipóteses para os efeitos definidos como fixos. Na análise dos modelos lineares mistos desbalanceados, o problema não é tão simples e o uso do pacote estatístico SAS (Statistical Analysis System) pode ser utilizado, demandando um volume maior de conhecimento teórico, em razão da opção pela estrutura da matriz de variância e co-variância e da escolha do método de estimação dos componentes. No procedimento MIXED estão disponíveis os métodos ML Maximum Likelihood, REML (Restricted Maximum Likelihood) e MIVQUE (Minimum Variance Quadratic Unbiased), descritos respectivamente por Hartley e Rao (1967), Rao (1971a) e Patterson e Thompson (1971). Para uso otimizado do SAS é necessário também que se defina o tipo de soma de quadrados mais adequado: ss1, ss2, ss3 ou ss4, segundo o SAS. Por suas propriedades, o método REML é o mais indicado e, para dados com desbalanceamento, a soma de quadrados do tipo ss3 deve ser escolhida. Quando os dados estão balanceados, os quatro tipos de soma de quadrados apresentam os mesmos resultados.

Palavras-chaves: Modelos mistos, delineamento inteiramente casualizado, Bbalanceados, SAS e genes.

Abstract The mixed models have been traditionally studied by means of variance analysis in which variance components are estimated by a linear model solution. For mixed model analysis some points are relevant: the prediction of random effects, tests upon and estimation of variance components, and estimation and tests of hypothesis upon fixed effects. In the analysis of unbalanced linear mixed models the problem is not that simple and the use of the SAS program (SAS System Inc., Cary, NC, USA) is preferable, which demand a proper

understanding of the theoretical background because of a proper modeling of the structure of the variance-covariance matrix and the choice of the estimation method is required. In the MIXED procedure are available the ML (Maximum Likelihood), REML (Restricted Maximum Likelihood), and MIVQUE (Minimum Variance Quadratic Unbiased) as described respectively by Hartley e Rao (1967), Rao (1971a), and Patterson e Thompson (1971). For an optimized use of SAS it is rather necessary to define the appropriate type of sum of squares: SS1, SS2, SS3, or SS4. Because of its properties the use of REML is recommended and for unbalanced data the SS3 type is the appropriate choice. With balanced data the four types of sum of squares yield the same results.

Keywords: mixed models, randomized design, balanced, and SAS genes

Introdução

Ao se estabelecer um modelo estatístico objetiva-se buscar a explicação de uma ou mais variáveis dependentes por meio de efeitos fixos e ou aleatórios. Desta forma, o êxito da modelagem está diretamente ligado às estruturas de variâncias e covariâncias das variáveis aleatórias que, nos modelos lineares mistos, vão ser explicadas por fatores aleatórios, após a eliminação dos efeitos fixos e da caracterização dos efeitos residuais.

Os modelos lineares, nos parâmetros, possuem pelo menos um efeito aleatório designado por erro experimental. Se o modelo apresenta todos os demais efeitos fixos ele é denominado de modelo linear fixo. Quando o modelo linear apresenta, além do erro experimental, outros efeitos aleatórios em comum com outros efeitos fixos, além da média, é denominado de modelo misto.

Nos modelos lineares mistos, a análise de variância apresenta algumas peculiaridades que devem ser mencionadas. A composição

das esperanças matemáticas dos quadrados médios permite a estruturação correta dos testes de hipóteses e, caso o interesse seja a estimação dos componentes de variância, os métodos foram amplamente estabelecidos por Henderson (1953), Rao (1970), Hartley e Rao (1967) e Patterson e Thompson (1971).

Estimadores de Restricted Maximum Likelihood (REML)

Segundo Searle (1982), Searle *et al.* (1992), Schaeffer (1978) e Patterson e Thompson (1971), considerando o modelo linear misto para uma característica:

$$Y = Xb + Zg + e, \text{ em que:}$$

y = vetor de dimensão ($n \times 1$) de observações medidas em n animais,
 X = matriz de dimensão ($n \times p$) ou de incidência de p efeitos fixos,
 b = vetor coluna de dimensão ($p \times 1$) ou de parâmetros desconhecidos de efeitos fixos,
 Z = matriz de dimensão ($n \times n$) ou de incidência de valores genéticos, matriz diagonal com 1 ou 0 na diagonal principal, indicando se a característica y foi medida ou não no animal,
 g = vetor de dimensões ($n \times 1$) de Valores Genéticos
 e = vetor de dimensão ($n \times 1$) de resíduos

As pressuposições da distribuição conjunta dos vetores (y, g, e) são descritas como:

$$\begin{bmatrix} y \\ g \\ e \end{bmatrix} \sim N \left\{ \begin{bmatrix} xb \\ \phi \\ \phi \end{bmatrix}, \begin{bmatrix} ZGZ' + R & ZG & R \\ GZ' & G & \phi \\ R & \phi & R \end{bmatrix} \right\}, \text{ em que:}$$

$$G = A \times G_0$$

$$R = I \times R_0$$

A = matriz do número do coeficiente de parentesco entre os indivíduos e de dimensões ($n \times n$),

I = matriz Identidade de ordem n ,

G_0 = matriz de variância genética aditiva.

R_0 = matriz de variância residual.

n = número de animais da amostra.

Se q características múltiplas são consideradas, as matrizes R_0 e G_0 serão simétricas e de ordem q e denominadas, respectivamente de variâncias e covariâncias genéticas aditivas e variâncias e (co) variâncias residuais das q características múltiplas. O objetivo, em qualquer situação, é estimar R_0 e G_0 .

Quando os vetores y, g têm distribuição conjunta normal multivariada, os vetores b e g são estimados e preditos, respectivamente, por:

$$b^0 = [X'R^{-1}X - X'R^{-1}Z(ZR^{-1}Z + G^{-1})^{-1}ZR^{-1}X]^{-1} * [X'R^{-1}y - XR^{-1}Z(ZR^{-1}Z + G^{-1})^{-1}Z'R^{-1}y]$$

$$\hat{g} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - Xb^0), \text{ em que:}$$

$$G = A \otimes G_0$$

$$R = I \otimes R_0$$

As matrizes já definidas anteriormente e \otimes = operador do produto direto das matrizes.

Para q características múltiplas a função densidade de probabilidade do vetor y é:

$$f(y) = \frac{1}{(2\pi)^{nq/2} |ZGZ' + R|^{0,5}} e^{-0,5[(y - Xb)'(ZGZ' + R)^{-1}(y - Xb)]}$$

A transformação logaritmo Neperiano de $f(y)$ é dada por:

$$LN[f(y)] = (-0,5 n q) LN(2\pi) - 0,5 LN|ZGZ' + R| - 0,5 [y'(ZGZ' + R)^{-1}y - 2y'(ZGZ' + R)^{-1}Xb + b'X'(ZGZ' + R)^{-1}Xb]$$

A transformação acima pode ser desdobrada em duas parcelas: $LN[f(y)]'$ e $LN[f(y)]''$ denominadas, respectivamente, por logaritmo neperiano da função densidade probabilidade dos contrastes ortogonais entre os efeitos fixos e logaritmo neperiano da função densidade probabilidade dos contrastes ortogonais entre os efeitos aleatórios.

$$LN[f(y)]' = -0,5 p[x] LN(2\pi) - 0,5 LN[X'(ZGZ' + R)^{-1}X] - 0,5 \{y'(ZGZ' + R)^{-1}X[X'(ZGZ' + R)^{-1}X]^{1/2} X'(ZGZ' + R)^{-1}y - 2y'(ZGZ' + R)^{-1}X[X'(ZGZ' + R)^{-1}X]^{1/2} X'(ZGZ' + R)^{-1}Xb + b'X'(ZGZ' + R)^{-1}X[X'(ZGZ' + R)^{-1}X]^{1/2} X'(ZGZ' + R)^{-1}Xb\}$$

$$LN[f(y)]'' = -0,5 p \{k'k(ZGZ' + R)^{-1}k\} LN(2\pi) - 0,5 LN|K(ZGZ' + R)^{-1}K| - 0,5 \{y'K'[K(ZGZ' + R)^{-1}K]^{-1}Ky\}, \text{ em que:}$$

$LN[f(y)]'$ = logaritmo neperiano da função densidade probabilidade dos contrastes ortogonais entre os efeitos fixos e, portanto, constante e livre de derivação por não influir nos componentes;

$LN[f(y)]''$ = logaritmo neperiano da função densidade probabilidade dos contrastes ortogonais entre os efeitos aleatórios;

p = indicação de posto da matriz;

K = é a matriz quadrada dos contrastes ortogonais entre os efeitos aleatórios;

Para a estimação dos componentes de (Co) variância, apenas a parcela $LN[p(y)]''$ fornecerá as derivadas em relação aos elementos das matrizes R_0 e G_0 e no ponto extremo da função de verossimilhança, todas as derivadas serão identicamente nulas.

$$\frac{\partial LN[f(y)]''}{\partial \sigma_{ei}^2} = -0,5tr\{K'[K(ZGZ' + R)KR]^{-1}\} + 0,5y'K'[K(ZGZ' + R)K]^{-1}$$

$$KR_iK'[K(ZGZ' + R)K]^{-1}Ky \equiv \phi$$

$$\text{para } i = [1, q]$$

$$\frac{\partial LN[f(y)]''}{\partial \sigma_{ei} \partial \sigma_{e'i'}} = -0,5tr\{K'[K(ZGZ' + R)K]^{-1}KR_{ii'}\} + 0,5y'K'[K(ZGZ' + R)K]^{-1}$$

$$KR_{ii'}K'[K(ZGZ' + R)K]^{-1}Ky \equiv \phi$$

$$\text{para } i \neq i' \text{ e todo } i = [1, q]$$

$$\frac{\partial LN[f(y)]'}{\partial \sigma_{gi}^2} = -0,5tr\{K'[K(ZGZ' + R)KG_i]^{-1}\} + 0,5y'K'[K(ZGZ' + R)K]^{-1}KG_iK'[K(ZGZ' + R)K]^{-1}Ky = \phi$$

para $i = [1, q]$

$$\frac{\partial LN[f(y)]'}{\partial \sigma_{gi} \partial \sigma_{g'i'}} = -0,5tr\{K'[K(ZGZ' + R)K]^{-1}KG_{ii'}\} + 0,5y'K'[K(ZGZ' + R)K]^{-1}KG_{ii'}K'[K(ZGZ' + R)K]^{-1}Ky = \phi$$

para $i \neq i'$ e todo $i = [1, q]$, em que:

$$R_i = Z_i \otimes Q_i$$

$$R_{i'} = Z_{i'} \otimes Q_{i'}$$

$$G_i = (Z_i A Z_i') \otimes Q_i$$

$$G_{i'} = (Z_{i'} A Z_{i'}) \otimes Q_{i'}$$

Fazendo as simplificações nas $2(C_p^2 + q)$ derivadas parciais e sabendo que $K[K(ZGZ' + R)K]^{-1}K = P$ é o projetor ortogonal da parte aleatória das observações no espaço coluna da matriz X :

$$Py = y - Xb^0$$

$$tr(PR_i) = y'PR_iPy$$

$$tr(PR_{ii'}) = y'PR_{ii'}Py$$

$$tr(PG_i) = y'PG_iPy$$

$$tr(PG_{ii'}) = y'PG_{ii'}Py \quad \text{em que:}$$

Z_i = matriz de incidência ($n \times n$) da característica $i = [1, q]$;
 Q_i = matriz ($q \times q$) que contém 1 na posição (i, i) e 0 nas demais posições;
 $Z_{ii'}$ = matriz de incidência simultânea ($n \times n$) das características i e i' ;
 $Q_{ii'}$ = matriz ($q \times q$) que contém 1 nas posições simétricas de i e i' ;
 0 nas demais posições;

Outra forma da matriz P é:

$$P = (ZGZ' + R)^{-1} - (ZGZ' + R)^{-1}X[X'(ZGZ' + R)^{-1}X]^{-1}X'(ZGZ' + R)^{-1}$$

$$P(ZGZ' + R)P = P$$

Fazendo $ZGZ' + R = V$, ter-se-á:

$$PVP = [V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}]V[V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}]V$$

$$PVP = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} + V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} + V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

$$* (X'V^{-1}X)^{-1}X'V^{-1}$$

Sendo $(X'V^{-1}X)$ matriz simétrica, a escolha apropriada da inversa generalizada levará à igualdade:

$$(X'V^{-1}X)^{IC} = (X'V^{-1}X)^{IC}(X'V^{-1}X)(X'V^{-1}X)^{IC}$$

$$PVP = V^{-1} - V^{-1}X(X'V^{-1}X)^{IC}X'V^{-1}$$

$$PVP = P, \text{ ter-se-á:}$$

$$tr(PR_i) = tr[P(ZGZ' + R)PR_i]$$

$$tr(PR_{ii'}) = tr[P(ZGZ' + R)PR_{ii'}]$$

$$tr(PG_i) = tr[P(ZGZ' + R)PG_i]$$

$$tr(PG_{ii'}) = tr[P(ZGZ' + R)PG_{ii'}]$$

para todo $i \neq i'$ e $i = [1, q]$.

Fatorando os seguintes membros das $2(C_p^2 + q)$ equações acima ter-se-á:

$$tr(PR_i) = t_i' e s$$

$$tr(PR_{ii'}) = t_{ii'}' e s$$

$$tr(PG_i) = t_i' g s$$

$$tr(PG_{ii'}) = t_{ii'}' g s = \begin{bmatrix} \sigma_{e1}^2 \\ \sigma_{e1e2}^2 \\ \vdots \\ \sigma_{e1eq}^2 \\ \vdots \\ \sigma_{eq}^2 \\ \sigma_{g1}^2 \\ \sigma_{g1g2}^2 \\ \vdots \\ \sigma_{g1gq}^2 \\ \vdots \\ \sigma_{gq}^2 \end{bmatrix} t_{ie} = \begin{bmatrix} tr(PR_i PR_{i1}) \\ tr(PR_i PR_{i2}) \\ \vdots \\ tr(PR_i PR_{iq}) \\ tr(PR_i PG_{i1}) \\ tr(PR_i PG_{i2}) \\ \vdots \\ tr(PR_i PG_{iq}) \end{bmatrix}$$

todo $i = [1, q]$, $R_{ii} = Ri$ e $G_{ii} = Gi$

$$t_{ij} = \begin{bmatrix} tr(PR_{ij} PR_i) \\ tr(PR_{ij} PR_{i2}) \\ \vdots \\ tr(PR_{ij} PR_{(q-1)q}) \\ tr(PR_{ij} PR_q) \\ tr(PR_{ij} PR_{G1}) \\ tr(PR_{ij} PR_{G12}) \\ \vdots \\ tr(PR_{ij} PR_{(q-1)q}) \\ tr(PR_{ij} PR_q) \end{bmatrix}$$

De forma análoga para t_{je} e t_{jg} .

Na forma matricial o sistema é:

$Ts = v$, em que:

$$T = \begin{bmatrix} t'_1 e \\ t'_{12} e \\ \vdots \\ t'_{(q-1)q} e \\ t'_q e \\ \vdots \\ t'_1 g \\ t'_{12} g \\ \vdots \\ t'_{(q-1)q} g \\ t'_q g \end{bmatrix} \quad v = \begin{bmatrix} y'PR_1Py \\ y'PR_{12}Py \\ \vdots \\ y'PR_{1q}Py \\ \vdots \\ y'PR_{(q-1)q}Py \\ y'PR_qPy \\ y'PG_1Py \\ y'PG_{12}Py \\ \vdots \\ y'PG_{1q}Py \\ \vdots \\ y'PG_{(q-1)q}Py \\ y'PG_qPy \end{bmatrix}$$

As soluções para o vetor s são obtidas iterativamente:

- 1-obter os elementos T e v , com estimativas iniciais de s ;
- 2-Resolver o sistema e obter novas aproximadamente para s ;
- 3-Repetir o procedimento até a obtenção da convergência.

Matricialmente a função de verossimilhança quando o vetor y tem uma distribuição normal multivariada é dada por:

$$f(y) = (2\pi)^{-0,5N} |V|^{-0,5} \exp(-0,5)[(y - Xb)' v^{-1}(y - Xb)]$$

Onde,

$$N = n_1 + n_2 + \dots + n_q$$

$N = nq$ se todos os grupos de heterogeneidade de variância têm o mesmo tamanho.

No REML, quando um grupo de contrastes $K'y$, e quando $K'X = f$ e K' tem posto $[N - p(X)]$.

$$f(k'y) = (2\sigma)^{-0,5[N-p(x)]} |K'VK|^{-0,5} \exp(-0,5)[(K'y)' (K'VK)^{-1}(K'y)]$$

A transformação logaritmo neperiano de $f(k'y)$ é dada por:

$$LN[f(y)] = (-0,5)[N - p(X)]LN(2\pi) - (0,5)LN|K'VK| - (0,5)[y'K(K'VK)^{-1}K'y]$$

A constante $(-0,5)[N-p(X)]LN(2\pi)$ não interfere nos componentes de (Co)variância e pode ser ignorada.

Searle (1971) demonstrou que:

$$LN |K'VK| = LN |V| + LN |X'V^{-1}X|$$

$$y'K(K'VK)^{-1}K'y = y'Py = (y - Xb^0)' V^{-1}(y - Xb^0)$$

$$LN[f(K'y)] = (-0,5)LN |V| - (0,5)LN |X'V^{-1}X| - (0,5)[(y - Xb^0)' V^{-1}(y - Xb^0)]$$

Quando A e D são matrizes de mesma ordem e, portanto, o produto é possível, então: $|AD| = |A| |D|$

$$|V| = |ZGZ' + R|$$

$$|V| = |R(I + R^{-1}ZGZ')|$$

Quando $R^{-1}ZGZ'$ é idempotente:

$$|V| = |R| |I + ZR^{-1}Z'G|$$

$$|V| = |R(G^{-1} + Z'R^{-1}Z)G|$$

$$|V| = |R| |G^{-1} + Z'R^{-1}Z| |G|$$

e, conseqüentemente, tem-se:

$$LN |V| = LN |R| + LN |G^{-1} + Z'R^{-1}Z| + LN |G|$$

Para matrizes quadradas e não singulares:

$$\begin{bmatrix} A & -B \\ Q & D \end{bmatrix} = |A| |D + QA^{-1}B| = |A| |D + A^{-1}BQ| = |AD| + |BQ| = |D| |A + BD^{-1}Q|$$

Se A e D são matrizes identidade:

$$|I + QB| = |I + BQ| = |I + B'Q| = |I + Q'B|$$

se θ é uma constante ou escalar:

$$|\theta A| = \theta^m |A|, \text{ quando } m \text{ é a ordem da matriz } A.$$

Da mesma forma, para:

$$C = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}$$

$$|C| = |Z'R^{-1}Z + G^{-1}| |X'(R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1})X|$$

fazendo, $S = (R^{-1} - R^{-1}X(X'R^{-1}X)^C X'R^{-1})$

$$|C| = |X'R^{-1}X| |G^{-1} + Z'SZ|$$

$$|C| = |Z'R^{-1}Z + G^{-1}| |X'V^{-1}X|$$

Como $LN |X'V^{-1}X| = LN |C| - LN |Z'R^{-1}Z + G^{-1}|$ e com as simplificações:

$$LN[f(K'y)] = -0,5 LN |R| - (0,5) LN |G| - (0,5) LN |C| - (0,5) y'Py$$

$$LN |R| = LN |I \sigma_0^2| = N LN(\sigma_0^2)$$

$$LN |G| = \sum_{i=1}^q g_i LN(\sigma_i^2)$$

$$LN |C| = LN |X'R^{-1}X| + LN |Z'SZ + G^{-1}|$$

que é o LN da parcela da função de verossimilhança, sem a constante, e que afeta os componentes de variância.

$$C^* = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}S_0^2 \end{bmatrix}$$

A matriz de (Co) variância de y é

$$V = \sum_{i=1}^q Z_i Z_i' S_{gi}^2 + I S_e^2$$

E o projetor da forma quadrática $y'Py$ é:

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$$

Derivando parcialmente $\ln [f(k'y)]$ em relação a cada σ_i^2 ($i = 1, 2, \dots, q$) obtém-se, segundo Searle (1971):

$$\frac{\partial \ln [f(k'y)]}{\partial \sigma_{gi}^2} = -(0,5) \text{tr}(P) + 0,5 y' P P y \text{ para todo } i = [1, q]$$

$$\text{tr}(P Z_i Z_i') = \frac{q_i}{\sigma_i^2} - \text{tr} \left(\frac{C_{ii} \sigma_o^2}{\sigma_{gi}^4} \right)$$

$$\text{tr}(P) = [N - p(X)] - \sum_{i=1}^q \frac{(\hat{g}_i' \hat{g}_i)}{\sigma_{gi}^2}$$

Do sistema de equações normais, obtém a solução abaixo.

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} b^0 \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

$$b^0 = \{X[R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}]\} X^{lc} * \\ * \{X[R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}]\} y$$

$$b^0 = (XV^{-1}X)'(XV^{-1}y)$$

Solução múltipla obtida diretamente do sistema de equações normais, e que coincide com a apresentada por (Schaffer 1993).

Sendo,

$$Py = (V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})y$$

$$b^0 = (XV^{-1}X)'(XV^{-1}y)$$

Em continuidade à solução do sistema de equações normais, obtém o resultado para os valores genéticos preditos, em que:

$$\hat{g} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - Xb^0)$$

Henderson, citado por Schaeffer (1993), apresenta a solução.

$$\hat{g}_i = G_i Z_i' P y, \text{ em que:}$$

$$P y = V^{-1}(y - Xb^0)$$

Henderson demonstrou também, que a forma quadrática.

$$y' P P y = y'y - b^0 X'y - \sum_{i=1}^q \hat{g}_i' Z_i' y - \sum_{i=1}^q \hat{g}_i' \hat{g}_i \left(\frac{\alpha_i}{\sigma_e^2} \right)$$

E, no ponto de máximo da função de máxi-verossimilhança, ter-se-á:

$$\hat{\sigma}_{gi}^2 = [\hat{g}_i' \hat{g}_i + \text{tr}(C_{ii} \sigma_e^2)] / q_i$$

$$\hat{\sigma}_e^2 = y' P y / (N - p(X))$$

Heterogeneidade da variância

A preocupação com os efeitos da heterogeneidade de variância, em classes de efeitos fixos, fez com que Lush (1945) recomendasse a produção de animais nos locais de uso desses animais, quando os genes desejáveis têm maior probabilidade de se expressar. A maior expressão dos genes de interesse, favorece a obtenção de estimadores mais acurados e precisos. Assim, variações genéticas e ou ambientais estão relacionadas com o desempenho dos animais (Dickerson 1962).

Falconer (1952) deu início ao estudo das correlações genéticas entre as características fenotípicas em ambientes distintos, e concluiu pela significância da interação genótipo ambiente: o melhor genótipo em um ambiente pode não ser tão bom em outro ambiente. Robertson *et al.* (1960) recomendam a comparação das herdabilidades, em classes de efeitos fixos, como forma de evitar as alterações na classificação de animais submetidos à seleção.

No melhoramento animal, a seleção de animais tem sido praticada em determinadas condições ambientais. Em características limitadas pelo sexo, a estimativa dos valores dos pais é obtida considerando o desempenho dos filhos e, neste caso, é importante caracterizar o ambiente onde as progênies estão atuando, bem como a distribuição delas nos vários ambientes. Por exemplo, a semelhança entre filhos de um reprodutor, em um determinado rebanho, é o reflexo tanto da covariância ambiental entre meio-irmãs, como da interação genótipo ambiente (Meyer 1989, Short *et al.* 1990). A covariância ambiental entre meio-irmãs paternas é função de fatores biológicos, de manejo e, até mesmo estatísticos, como a não remoção do efeito de rebanho, de outro fator fixo ou de um modelo inadequado para a estimação dos valores genéticos (Meyer 1987).

Se o efeito da interação reprodutor-rebanho for considerado no modelo linear de avaliação genética, a avaliação de progenitores é menos acurada e precisa em poucos rebanhos. A interação pode não afetar significativamente a avaliação de reprodutores, com progênies em vários rebanhos, como é o caso de reprodutores utilizados na inseminação artificial. Ignorar a interação pode significar a superestimação da acurácia nas avaliações, principalmente, quando poucos rebanhos são considerados. Quando as diferenças entre progênies de um mesmo reprodutor não são as mesmas em diferentes rebanhos, a interação reprodutor-rebanho, pode não ser significativa se

a heterogeneidade de variâncias explicarem parte dessas diferenças.

Variâncias heterogêneas entre rebanhos e aumento da variância associados ao aumento de médias de características produtivas importantes, têm sido verificados em várias oportunidades e por vários autores como Meyer (1987), Boldman e Freeman (1990), Boldman e van Vleck (1991), Dong e Mad (1990), Stanton *et al.* (1991), Dodenhoff e Swalve (1998) e Torres (1998).

Segundo Torres (1998), quando a heterogeneidade de variância não é levada em conta, embora presente, a produção das filhas de reprodutores deverá ser ponderada pelos desvios padrão dos rebanhos que as contêm. As filhas oriundas de rebanhos com variâncias maiores influenciarão mais a avaliação de reprodutores. Os vieses serão maiores, na seleção das mães, quando elas tendem a expressar suas características em um único rebanho.

Bereskin e Lush (1965) concluíram que os resultados da avaliação de reprodutores em rebanhos de alto nível, podem não estar de acordo com avaliação de reprodutores, no teste de progênes, em rebanhos de baixo nível de produção.

Mesmo que o número de dados seja insuficiente para a estimação dos componentes de variâncias, rebanhos similares têm sido agrupados e os componentes têm sido obtidos dentro de cada grupo (Mirande e Van Vleck 1985). Os estudos evidenciaram a existência de variâncias heterogêneas entre os diferentes ambientes.

Mohammad *et al.* (1982) estimaram os valores genéticos de reprodutores para a produção de leite com modelos reprodutores com ou sem a interação reprodutor-rebanho e interação reprodutor- peso ideal de novilhas. A correlação de SPEARMAN entre os modelos com e sem interação foram significativos, indicando que não houve alteração, também, significativa na ordenação dos valores genéticos preditos nos vários modelos utilizados.

Van Tassel e Berger (1994), com dados simulados, estudaram as consequências da eliminação da interação reprodutor-rebanho e da matriz de parentesco na estimação dos componentes de variância em três casos distintos de distribuição de filhas nos rebanhos. Os componentes de variância foram estimados pelo método das estimativas quadráticas não-viesadas de variância mínima (MIVQUE). Quando o parentesco entre os reprodutores foi eliminado, as variâncias do efeito de reprodutor e da interação reprodutor-rebanho tenderam à subestimação, quando os reprodutores eram mais próximos. A variância atribuída aos reprodutores foi superestimada quando, no modelo linear, a interação reprodutor-rebanho não foi considerada. Para os dados resultantes de inseminação artificial, quando não foi considerada a matriz de parentesco, a variância da interação reprodutor-rebanho foi subestimada.

O efeito da interação reprodutor-rebanho quando as variâncias entre rebanhos eram heterogêneas, e seis diferentes combinações de heterogeneidades de variâncias entre rebanho – ano - estação foram simulados. Quando a característica não é influenciada pela interação reprodutor-rebanho a acurácia das estimativas de valores genéticos pode ser subestimada.

Spike e Freeman (1976) verificaram a importância da acurácia

na avaliação genética animal. Concluíram que com o aumento da acurácia definida como a eficiência do processo de avaliação na identificação das diferenças observadas nos fenótipos, a mudança genética esperada foi ampliada significativamente.

Ojala *et al.* (1984), trabalhando com dados simulados, avaliaram o efeito do número de amostras nas classes rebanho-reprodutor, sobre a acurácia da avaliação genética. Os resultados demonstraram que a acurácia das estimativas melhora significativamente quando o número de amostras aumenta de 2 para 3 observações por classe. O aumento de número de progênes por reprodutor teve, no entanto, maior efeito sobre acurácia.

Conexidade de dados

Segundo Searle (1971) para modelos fixos de classificação cruzada, a conexidade entre os dados existe quando todas as funções lineares entre dois níveis dos fatores considerados são estimáveis, ou seja, quando existem observações em todas as combinações de níveis ou classes de fatores. Nos modelos mistos, se a pressuposição de esperança dos efeitos aleatórios, dentro de níveis dos efeitos fixos não se verifica, as comparações entre os efeitos aleatórios em diferentes níveis dos efeitos fixos serão viesadas.

A simulação de dados tem sido usada com frequência em estudo de problemas de conexidade e o efeito na avaliação genética animal, nos diversos tipos de modelos lineares. Tong *et al.* (1980), Kennedy (1975, 1988) e Analla *et al.* (1995) foram alguns dos grupos que trabalharam com simulação nos problemas de conexidade dos dados experimentais.

Foulley *et al.* (1987) discutiram o impacto da conexidade em relação ao tipo de modelo, fixo ou misto, utilizado e o efeito nas propriedades de um estimador de mérito genético, não-viesado e com quadrado médio do erro mínimo. Segundo os autores, o problema de conexidade restringe o uso do método dos quadrados mínimos regressivos, quando há falta de conexidade o processo iterativo é interrompido. Quando a metodologia dos modelos mistos é convenientemente aplicada, o sistema de equações normais sempre apresentará solução, sem levar em conta o grau de conexidade.

Schenkel (1991) avaliou as vantagens do uso dos modelos mistos quando o problema de conexidade existe. Nos modelos mistos o processo faz com que as soluções dentro de grupos desconectados somem zero e não provoquem a quebra do processo de convergência.

A teoria do índice de seleção só admite comparações entre animais de um mesmo grupo contemporâneo ou que compartilham o mesmo momento reprodutivo. O método dos quadrados mínimos permite a comparação entre grupos contemporâneos, quando existe a conexidade dos dados.

O efeito de três tipos de falta de conexidade em um modelo linear com 2 fatores classificatórios sobre a estimação dos componentes de variância, usando o método da estimação quadrática não-viesada de norma mínima (MINQUE). O autor concluiu que a falta de conexidade

acarreta a falta de associação dos graus de liberdade com as várias formas quadráticas. Ainda segundo Schenkel (1991), parece impróprio estimar os componentes de variância sem antes eliminar as subclasses desconectadas. Em pequenos grupos de dados o problema é detectado de imediato, mas, para um grande volume de dados, o problema não é evidenciado e pode ser a causa de ocorrência de estimativas negativas dos componentes de variância.

Referências

- Analla M, Sanchez-Palma A, Munoz-Serrano A, Serradilla JM (1995) Simulation analysis with BLUP methodology of different data structures in goat selection schemes in Spain. **Small Ruminant Research** 17: 51-55.
- Bereskin B, Lush JL (1965) Genetic and environmental factors in dairy sire evaluation. III. Influence of environmental and others extraneous correlations among the daughters. **Journal of Dairy Science** 48: 356-360.
- Boldman KG, Freeman AE (1990) Adjustment for heterogeneity of variances by herd production level in dairy cow and sire evaluation. **Journal of Dairy Science** 73: 503-512.
- Boldman KG, van Vleck LD (1991) Derivative-free restricted maximum likelihood estimation in animal models with a sparse matrix solver. **Journal of Dairy Science** 74: 4337-4343.
- Dickerson GE (1962) Implications of genetic-environmental interactions in animal breeding. **Animal Production** 4: 47-57.
- Dodenhoff J, Swalve HH (1998) Heterogeneity of variances across regions of northern Germany and adjustment in genetic evaluation. **Livestock Production Science** 53: 225-236.
- Dong MC, Mad IL (1990) Heterogeneity of (co)variance and heritability in different levels of intra herd milk production and of herd average. **Journal of Dairy Science** 73: 843-851.
- Falconer DS (1952) The problem of environment and selection. **American Naturalist** 86: 293-298.
- Foulley JL, Im S, Gianola D, Hoeschele I (1987) Empirical Bayes estimation of parameters for n polygenic binary traits. **Genetics, Selection, Evolution** 19: 197-224.
- Hartley HO, Rao JNK (1967) Maximum-likelihood estimation for the mixed analysis of variance model. **Biometrika** 54: 93-108.
- Henderson CR (1953) Estimation of variance and covariance components. **Biometrika** 9: 226-52.
- Kennedy BW (1975) Comparison of genetic group and relationship methods for mixed model sire evaluation. **Journal of Dairy Science** 58: 1507-1514.
- Kennedy BW (1988) **Workshop on estimation of breeding values with the animal model**. Guelph, Ontario: University of Guelph.
- Lush J. (1945) The Built Index problem in the light of modern genetics. **Journal of Dairy Science** 16: 501-522.
- Meyer DJ (1987) The management of cytology specimens. **Compendium on Continuing Education - Small Animal Practice** 9: 10-17.
- Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. **Genetics, Selection Evolution** 21: 317-340.
- Mirande SL, Van Vleck LD (1985) Trends in genetic and phenotypic variances for milk production. **Journal of Dairy Science** 68: 2278-2286.
- Mohammad WA, Lee AJ, Grossman M (1982) Genotype-environment interaction in sire evaluation. **Journal of Dairy Science** 65: 857-860.
- Ojala M, Syvajarvi J, Hellman T (1984) Effect of the size of the comparison on the accuracy of a sire's evaluation. **Journal of Animal Breeding** 102: 91-94.
- Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. **Biometrika** 58: 545-554.
- Rao CR (1970) Estimation of heteroscedastic variance in linear models. **Journal of the American Statistical Association** 65: 161-172.
- Rao CR (1971a) Minimum variance quadratic unbiased estimation of variance components. **Journal of Multivariate Analysis** 1: 445-456.
- Rao CR (1971b) Estimation of variance and covariance components minque theory. **Journal of Multivariate Analysis** 1: 257-275.
- Rao R (1972) Estimation of variance and covariance components in linear models. **Journal of the American Statistical Association** 67: 112-115.
- Robertson A, O'Connor LK, Edwards J (1960) Progeny testing dairy bulls at different management levels. **Animal Production** 2: 141-152.
- Schaeffer LR (1978) Simultaneous estimation of variance and covariances components from multitrait mixed model equations. **Biometrics** 34: 199-208.
- Schenkel FS (1991) **Aplicação da metodologia de modelos mistos na avaliação genética de suínos**. Dissertação Mestrado. Porto Alegre, Universidade Federal do Rio Grande do Sul.
- Searle SR (1971) Topics in variance component estimation. **Biometrics** 27: 1-76.
- Searle SR (1982) **Matrix Algebra Useful to Statistics**. New York, John Wiley and Sons.
- Searle S, Casella G, McCulloch CE (1992) **Variance Components**. New York, John Wiley and Sons.
- Short RE, Bellows RA, Staigmiller RB, Berardinelli JG, Custer EE (1990) Physiological mechanisms controlling anestrus and infertility in postpartum beef cattle. **Journal of Animal Science** 68: 799-816.
- Spike PL, Freeman AE (1978) Prediction of genetic differences among herds with estimates of breeding values. **Journal of Dairy Science** 61: 1476-1482.
- Stanton TL, Blake RW, Van Vleck LD, Quaas RL (1991) Genotype by environment interaction for Holstein milk yield in Colombia, Mexico, and Porto Rico. **Journal of Dairy Science** 74: 1700-1714.
- Tong AKW, Kennedy BW, Moxley JE (1980) Potential errors in sire evaluation from regional genetic differences. **Journal of Dairy Science** 63: 627-633.
- Torres RA (1998) **Efeito da heterogeneidade de variância na avaliação genética de bovinos da raça holandesa no Brasil**. Belo Horizonte, UFMG.