Revista Brasileira de Zootecnia
© 2009 Sociedade Brasileira de Zootecnia
ISSN 1516-3598 (impresso)
ISSN 1806-9290 (*on-line*)
www.sbz.org.br

R. Bras. Zootec., v.38, p.72-80, 2009 (supl. especial)

Modelos hierárquicos bayesianos para estimação robusta e análise de dados censurados em melhoramento animal

Fernando Flores Cardoso[1], Guilherme Jordão de Magalhães Rosa[2], Robert John Tempelman[3], Roberto Augusto de Almeida Torres Junior[4]

[1] Embrapa Pecuária Sul/Bagé, RS 96401-970. Bolsista de Produtividade do CNPq.
[2] Department of Dairy Science, University of Wisconsin, Madison, WI 53706, EUA.
[3] Department of Animal Science, Michigan State University, East Lansing, MI 48824, EUA.
[4] Embrapa Gado de Corte/Campo Grande, MS.

RESUMO - Dados extremos influenciados por fatores não considerados no modelo estatístico, podem enviesar as estimativas dos parâmetros e valores genéticos. Além disso, diversas características de importância econômica não seguem uma distribuição normal ou apresentam dados censurados. O objetivo deste trabalho é descrever e ilustrar a aplicação de modelos hierárquicos bayesianos para a detecção e mitigação de dados extremos e para análise de dados censurados. Primeiro, é apresentada a especificação tradicional do modelo animal em estágios hierárquicos sob o enfoque bayesiano, para dados não censurados com distribuição Normal. A seguir, esse modelo é generalizado pela introdução de uma variável de ponderação independente, que permite a especificação de densidades residuais de caudas longas da família de distribuições Normal/ independente. Finalmente, para contemplar a análise de dados censurados, o modelo básico é ampliado pela inclusão de uma variável com distribuição normal truncada no limite inferior do valor observado da característica no momento da avaliação, para aqueles animais que ainda não completaram sua vida reprodutiva no momento da avaliação.

Palavras-chave: análise robusta, dados censurados, dados extremos, inferência bayesiana

Hierarchical Bayesian models for robust estimation and censored data analysis in animal breeding

ABSTRACT- Data strongly influenced by factors not accounted for by the statistical model can bias estimates of genetic parameters and values. Moreover, several traits of economic importance do not follow a normal distribution or have censored data. The objective of this study is to describe and illustrate the application of hierarchical Bayesian models for the detection and muting of outliers and for the analysis of censored data. First, the traditional specification of the animal model in hierarchical stages is presented under the Bayesian approach for normally distributed uncensored data. Then, this model is extended by introducing an independent weighting variable, which allows for the specification of thick tail residual densities from the Normal/independent distribution family. Finally, to cover censored data analysis, the basic model is extended by the inclusion of a variable with truncated normal distribution based on the lower limit in the observed value of the trait at the evaluation time, for those animals that have not yet completed their reproductive life at the evaluation time.

Key Words: Bayesian inference, censored data, outliers, robust analysis

Introdução

A precisão das predições de valores genéticos utilizados para seleção e, conseqüentemente, o progresso genético em programas de melhoramento animal, depende da qualidade da informação fenotípica e de pedigree disponível e da adequação do modelo adotado às características dos dados analisados. O desempenho dos animais de produção é geralmente observado em sistemas e ambientes de produção diversos, com qualidade de dados muitas vezes comprometida pela ocorrência de erros de registro e identificação, de tratamento preferencial a alguns animais, de lesões ou doenças e de formação inadequada de grupos de manejo. Portanto, os conjuntos de dados freqüentemente contém observações extremas, influenciadas por fatores não considerados no modelo estatístico, que podem enviesar as estimativas dos parâmetros e valores genéticos (Stranden & Gianola, 1998).

A edição de dados nos programas de avaliação genética geralmente envolve a eliminação de observações que são consideradas muito afastadas da média fenotípica de sua subclasse (usualmente três ou mais desvios padrão) ou se

Correspondências devem ser enviadas para: fcardoso@cppsul.embrapa.br

a relação observação/média de subclasse se encontra fora do intervalo 60% a 140% (Bertrand & Wiggans, 1998). Apesar de essas estratégias parecerem adequadas para erros óbvios de coleta de informação, ela pode ser ineficiente se a verdadeira distribuição dos resíduos se desvia da normalidade e as variâncias das subclasses são heterogêneas (Cardoso et al., 2005). Modelos pressupondo densidades de caudas longas da família Normal/ independente (Lange & Sinsheimer, 1993; Rosa et al., 2003), como as distribuições t de Student, Slash e Normal Contaminada, são alternativas para mitigar o efeito de dados extremos, sem a necessidade de eliminação de observações que pode ser consideradas extremas com respeito a pressuposição de normalidade (Stranden & Gianola, 1998, 1999)

Além disso, diversas características de importância econômica apresentam dados censurados, isto é, que não são observados integralmente para todos os animais no momento da avaliação genética. Características como longevidade, prolificidade e produtividade total de fêmeas são exemplos de dados censurados, pois muitos animais ainda estão em reprodução no momento da avaliação e, portanto, somente o limite inferior do seu valor fenotípico é conhecido.

Análise de sobrevivência é uma metodologia estatística para ajustar dados de longevidade, combinando registros completos de animais mortos ou que completaram seu ciclo de vida útil e dados censurados de animais ainda em produção, particularmente utilizando-se os modelos de riscos proporcionais (p. ex., Ducrocq & Casella, 1996; Ducrocq & Sölkner, 1998). Nesses modelos, a relação do risco proporcional do animal ser descartado ou morrer em um determinado tempo é não linear com a longevidade e, portanto, a implementação e interpretação é mais difícil que nos modelos lineares convencionais baseados na metodologia dos modelos mistos (Henderson, 1984), amplamente utilizados pelos programas de melhoramento genético animal.

Uma alternativa utilizada para permitir o uso de modelos lineares para características com dados censurados, é substituir esses registros censurados por valores projetados baseados na informação disponível no momento da avaliação (Vanraden & Klaaskate, 1993). Por outro lado, a informação disponível pode ser melhor utilizada, assumindo-se uma distribuição normal truncada para a característica e, sob o enfoque bayesiano, gerar os registros censurados considerando os demais efeitos fixos e aleatórios do modelo (Guo et al., 2001; Donoghue et al., 2004b).

Os modelos hierárquicos bayesianos proporcionam uma metodologia geral e flexível para ajustar a complexidade de fatores genéticos e ambientais que afetam o desempenho de animais de produção em características biológicas complexas, considerando o conhecimento *a priori* e a informação contida nos dados (Sorensen & Gianola, 2002). O objetivo neste trabalho é descrever e ilustrar a aplicação de modelos hierárquicos bayesianos para a detecção e mitigação de dados extremos e para análise de dados censurados.

Metodologia

Construção hierárquica do modelo animal sob enfoque bayesiano

Na formulação em estágios hierárquicos do modelo animal tradicional para dados com distribuição normal e não censurados, o primeiro estágio corresponde à distribuição condicional de amostragem do vetor de dados **y**, de ordem *n* × 1. Em análise univariada, considerando a resposta de um determinado animal *j*, temos o seguinte modelo linear misto:

$$y_j = \mathbf{x}'_j\boldsymbol{ß} + \mathbf{z}'_j\mathbf{u} + e_j, \qquad [1]$$

em que $\boldsymbol{ß}$ é um vetor de ordem p × 1 de efeitos fixos, u é um vetor de ordem q × 1 de efeitos aleatórios, e $\mathbf{x}'_j$ e $\mathbf{z}'_j$ são vetores linha de incidência conhecidos conectando y_j a e u, respectivamente. Finalmente, $e_j \sim N\left(0, s_e^2\right)$ representa o erro aleatório com distribuição normal independente e variância residual s_e^2 idêntica para todo j = 1,2,…, n. Se especificamos que a média da distribuição para y_j é $m_j = \mathbf{x}'\boldsymbol{ß} + \mathbf{z}'\mathbf{u}$, temos que a densidade da distribuição condicional dos registros é:

$$p\left(\mathbf{y}|\boldsymbol{ß}, \mathbf{u}, s_e^2\right) = \left(2ps_e^2\right)^{-n/2} \exp\left[\frac{-1}{2s_e^2}\sum_{j=1}^{n}\left(y_j - m_j\right)^2\right], \qquad [2]$$

No segundo estágio desse modelo, especificam-se as pressuposições *a priori* para as incógnitas usadas para determinar a média e a variância da distribuição normal definida no primeiro estágio. Para os para os parâmetros de locação adotaram-se densidades a priori uniformes para os efeitos fixos em b (com limites inferiores e superiores finitos para assegurar que seja uma priori própria):

$$p(\boldsymbol{ß}) \propto \text{constante}, \qquad [3]$$

e normal multivariada para os efeitos aleatórios, isto é $\mathbf{u} \mid s_u^2 \sim N\left(\mathbf{0}, \mathbf{A}s_u^2\right)$, com densidade,

$$p\left(\mathrm{u} \mid s_u^2\right) = \left(2ps_u^2\right)^{-q/2} \exp\left(-\frac{\mathrm{u}'\mathrm{A}^{-1}\mathrm{u}}{2s_u^2}\right). \qquad [4]$$

No caso do modelo animal, u contém as soluções para os valores genéticos aditivos individuais, A representa a matriz do numerador do parentesco e s_u^2 a variância genética aditiva.

O segundo estágio é concluído pela pressuposição de uma priori conjugada para o componente de variância residual, através de uma distribuição qui-quadrado invertida escalonada, $s_e^2 | n_e, s_e^2 \sim n_e s_e^2 c_{n_e}^{-2}$, com densidade:

$$p\left(s_e^2 | n_e, s_e^2\right) \propto \left(s_e^2\right)^{-\left(\frac{n_e+2}{2}\right)} \exp\left(-\frac{n_e s_e^2}{2 s_e^2}\right), \quad [5]$$

em que n_e são os graus de liberdade ou grau de confiança no valor *a priori* de s_e^2, representado acima por s_e^2.

Finalmente, o terceiro estágio do modelo contempla a pressuposição de uma priori conjugada para s_u^2, também por meio de uma distribuição qui-quadrado invertida escalonada, $s_u^2 | n_u, s_u^2 \sim n_u s_u^2 c_{n_u}^{-2}$, com graus de liberdades, fator escalar $n_u s_u^2$ e densidade:

$$p\left(s_u^2 | n_u, s_u^2\right) \propto \left(s_u^2\right)^{-\left(\frac{n_u+2}{2}\right)} \exp\left(-\frac{n_u s_u^2}{2 s_u^2}\right) \quad [6]$$

Acima o valor *a priori* de s_u^2 é dado por s_u^2.

A distribuição posterior conjunta de todos os parâmetros do modelo é obtida pelo produto de todas as densidades especificadas nos três estágios:

$$p\left(\beta, \mathbf{u}, s_e^2, s_u^2 | \mathbf{y}, n_e, s_e^2, n_u, s_u^2\right) = p\left(\mathbf{y} | \beta, \mathbf{u}, s_e^2\right) p(\beta)\, p\left(\mathbf{u} | s_u^2\right)\, p\left(s_e^2 | n_e, s_e^2\right)\, p\left(s_u^2 | n_u, s_u^2\right) \quad [7]$$

A partir dessa distribuição em [7] são derivadas as distribuições condicionais completas (DCC) de cada parâmetro ou conjunto de parâmetros necessárias para implementar a inferência por meio de métodos Monte Carlo via Cadeias de Markov (MCMC) (Wang et al., 1994). Para os parâmetros de locação $\theta' = [\beta' \quad \mathbf{u}']$, pode ser demonstrado que a DCC é normal multivariada (Gianola & Fernando, 1986):

$$\beta, \mathbf{u} | \mathbf{y}, \sigma_e^2, \sigma_u^2, v_e, s_e^2, v_u, s_u^2 : N\left(\hat{\theta}, \mathbf{C}^{-1}\sigma_e^2\right), \quad [8]$$

em que $\hat{\theta}$ é a solução para o sistema de equações $\mathbf{C}\theta = \mathbf{b}$, em que C é uma matriz de coeficientes,

$$\mathbf{C} = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' & \sum_{j=1}^{n} \mathbf{x}_j \mathbf{z}_j' \\ \sum_{j=1}^{n} \mathbf{z}_j \mathbf{x}_j' & \sum_{j=1}^{n} \mathbf{z}_j \mathbf{z}_j' + \mathbf{A}^{-1} l \end{bmatrix} \quad \text{para} \quad l = s_e^2 / s_u^2, \quad \text{e}$$

$$\mathbf{b} = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{x}_j y_j \\ \sum_{j=1}^{n} \mathbf{z}_j y_j \end{bmatrix}.$$

Já para os parâmetros de dispersão temos as seguintes DCC qui-quadrado invertidas escalonadas:

$$s_e^2 \left| \mathbf{y}, \beta, \mathbf{u}, s_u^2, n_e, s_e^2, n_u, s_u^2 \right. \sim \left(\sum_{j=1}^{n} \left(y_j - m_j\right)^2 + n_e s_e^2 \right) c_{(n_e+n)}^{-2}, \text{ e } [9]$$

$$s_u^2 \left| \mathbf{y}, \beta, \mathbf{u}, s_e^2, n_e, s_e^2, n_u, s_u^2 \right. \sim \left(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + n_u s_u^2\right) c_{(n_u+q)}^{-2} \quad [10]$$

A estratégia para implementar a amostragem de Gibbs é baseada da geração de uma cadeia de amostras sucessivas de [8], [9] e [10], descartando-se um conjunto inicial de amostras para permitir a convergência da cadeia para a sua distribuição estacionária e para eliminar o efeito dos valores iniciais. Essas amostras após o período de descarte são representativas das densidades posteriores marginais dos parâmetros do modelo, das quais é derivada a inferência sobre esses parâmetros ou quaisquer funções de interesse envolvendo tais parâmetros.

Generalização do modelo animal para estimação robusta

Mantendo-se o enfoque univariado e as demais pressuposições acima, o modelo da equação [1] pode ser generalizado para permitir estimação robusta a dados extremos, pela introdução de uma variável de ponderação independente , de modo que:

$$y_j = m_j + e_j / \sqrt{w_j}, \quad [11]$$

para $w_j > 0$ e com densidade $p\left(w_j | n\right)$, em que n é um parâmetro de valor positivo.

A escolha da forma de $p\left(w_j | n\right)$ permite a especificação de diferentes densidades marginais de caudas longas da família de distribuições Normal/independente (Lange & Sinsheimer, 1993; Rosa et al., 2003) para os registros , capazes de melhor ajustar dados extremos não contemplados pela pressuposição de normalidade. Consideramos especificamente duas alternativas, as distribuições *t* de Student e de Slash. A distribuição residual *t* de Student tem sido recomendada para mitigar os efeitos de tratamento preferencial e de observações discrepantes (Stranden & Gianola, 1998; Pinheiro et al., 2001), enquanto que a distribuição de Slash demonstrou ter, em alguns casos, um ajuste melhor que a *t* de Student em conjuntos com dados extremos (Rosa et al., 2003).

Primeiro, deve-se notar que agora $e_j | s_e^2, w_j \sim N\left(0, s_e^2 / w_j\right)$ e que a distribuição Normal ou Gaussiana é um caso especial de [11], fixando $w_j = 1$ para todos os j=1,2,...,n. Segundo que uma especificação distribucional em w_j representa uma forma objetiva de

modelar a falta de ajuste da densidade marginal de y_j em relação distribuição Gaussiana. Dada uma especificação distribucional $p(w_j|n)$, condicional em n, a densidade marginal $p(e_j|s_e^2,n)$ de e_j é determinada por $p(e_j|s_e^2,n)=\int p(e_j|s_e^2,w_j)\,p(w_j|n)dw_j$.

Entretanto, a implementação da inferência via MCMC é facilitada pela especificação da densidade da distribuição de amostragem dos registros também condicional em w, um vetor contendo todos os w_j's, para $j=1,2,\ldots,n$:

$$p(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},s_e^2,\mathbf{w})=(2\pi s_e^2)^{-n/2}\left(\prod_{j=1}^{n}w_j\right)^{-1/2}\exp\left[\frac{-1}{2s_e^2}\sum_{j=1}^{n}w_j(y_j-m_j)^2\right],$$

[12]

Note que a contribuição de cada observação y_j em [12] é ponderada por w_j e, deste modo, dados extremos ao serem associados com valores próximos a zero de w_j, praticamente não afetarão as estimativas dos demais parâmetros do modelo.

Agora se faz necessária a inclusão da especificação de $p(w_j|n)$ no segundo estágio da hierarquia do modelo. Se assumirmos que $w_j|n \sim Gama(n/2,n/2)$, isto é:

$$p(w_j|n)=\frac{(n/2)^{n/2}\,w_j^{n/2-1}\exp(-w_j n/2)}{\Gamma(n/2)},\ j=1,2,\ldots,n,\ n>0,\ w_j>0,$$

[13]

pode-se demonstrar que a densidade marginal $p(e_j|s_e^2,n)$ é uma densidade *t* de Student com parâmetro escalar s_e^2 e graus de liberdade n tal que a variância residual marginal $s_E^2=\mathrm{var}(e_j|s_e^2,n)$ é $\frac{n}{n-2}s_e^2$ (Lange & Sinsheimer, 1993).

Alternativamente w_j pode ser especificado como tendo a seguinte densidade:

$$p(w_j|n)=n\,w_j^{n-1},\ j=1,2,\ldots,n,\ n>0,\ 0<w_j\leq 1,$$

[14]

desse modo conduzindo a especificação da distribuição de Slash para $p(e_j|s_e^2,n)$ com parâmetro escalar s_e^2 e grau de liberdade n, tal que a variância marginal de e_j é $s_E^2=\frac{n}{n-1}s_e^2$ (Lange & Sinsheimer, 1993).

As demais pressuposições do segundo e terceiro estágios do modelo hierárquico para os parâmetros de locação e dispersão em [3], [4], [5] e [6] permanece iguais. Entretanto, o terceiro estágio deve também incluir agora $p(n)$, a distribuição *a priori* dos graus de liberdade ou parâmetro de robustez, onde menores valores de n estão associados com caudas mais longas e maior robustez a dados extremos. Podem ser adotados valores fixos, por exemplo, n = 4 ou 8 para os graus de liberdade, mas esses valores podem também ser estimados no modelo, indicando o grau de afastamento da normalidade no conjunto de dados de interesse (Rosa et al., 2003; Cardoso et al., 2005). A distribuição posterior conjunta de todos os parâmetros do modelo robusto é:

$$p(\boldsymbol{\beta},\mathbf{u},s_e^2,s_u^2,\mathbf{w},n|\mathbf{y},n_e,s_e^2,n_u,s_u^2)$$
$$=p(\mathbf{y}|\mathbf{w},\boldsymbol{\beta},\mathbf{u},s_e^2)\,p(\mathbf{w}|n)\,p(\boldsymbol{\beta})\,p(\mathbf{u}|s_u^2)\,p(s_e^2|n_e,s_e^2)\,p(s_u^2|n_u,s_u^2)\,p(n)$$

[15]

Alguns ajustes e derivações adicionais são necessários nas DCC para implementar a inferência por meio MCMC em relação ao modelo animal. A forma da DCC para os parâmetros de locação permanece normal multivariada, mas temos que introduzir ajustes no sistema de equações que refletem a inclusão das variáveis de ponderação. Agora temos $\tilde{\mathbf{C}}\tilde{\boldsymbol{\theta}}=\tilde{\mathbf{h}}$, onde $\tilde{\mathbf{C}}$ $\begin{bmatrix}\sum_{j=1}^{n}w_j\mathbf{x}_j\mathbf{x}_j' & \sum_{j=1}^{n}w_j\mathbf{x}_j\mathbf{z}_j'\\ \sum_{j=1}^{n}w_j\mathbf{z}_j\mathbf{x}_j' & \sum_{j=1}^{n}w_j\mathbf{z}_j\mathbf{z}_j'+\mathbf{A}^{-}\lambda\end{bmatrix}$ e $\tilde{\mathbf{h}}-\begin{bmatrix}\sum_{j=1}^{n}w_j\mathbf{x}_j y_j\\ \sum_{j=1}^{n}w_j\mathbf{z}_j y_j\end{bmatrix}$, tal que DCC é:

$$\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\sigma_e^2,\sigma_u^2,\mathbf{w},\nu,\nu_e,s_e^2,\nu_u,s_u^2:\ N(\tilde{\boldsymbol{\theta}},\tilde{\mathbf{C}}^{-1}\sigma_e^2),\qquad [16]$$

em que $\tilde{\boldsymbol{\theta}}$ é uma solução robusta a dados extremos do sistema de equações $\tilde{\mathbf{C}}\tilde{\boldsymbol{\theta}}=\tilde{\mathbf{h}}$.

A forma das DCC para os parâmetros de dispersão também permanece a mesma, apenas refletindo a inclusão dos ponderadores:

$$s_e^2|\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{w},n,s_u^2,n_e,s_e^2,n_u,s_u^2 \sim \left(\sum_{j=1}^{n}w_j(y_j-m_j)^2+n_e s_e^2\right)c_{(n_e+n)}^{-2}\ [17]\ \text{e}$$

$$s_u^2|\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{w},n,s_e^2,n_e,s_e^2,n_u,s_u^2 \sim \left(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}+n_u s_u^2\right)c_{(n_u+q)}^{-2}\ [18]$$

As DCCs para os ponderadores w_j's dependem da escolha de $p(w_j|n)$. Adotando a densidade em [13], que conduz à distribuição *t* de Student, essas DDC correspondem a uma série de densidades Gama,

$$w_j|\mathbf{y},\boldsymbol{\beta},\mathbf{u},n,s_u^2,n_e,s_e^2,n_u,s_u^2 \sim Gama\left(\frac{n+1}{2},\frac{1}{2}\left(n+\frac{(y_j-m_j)^2}{s_e^2}\right)\right),$$

$j=1,2,\ldots,n$. [19]

Por outro lado, adotando a densidade em [14], que conduz a uma densidade marginal Slash, as DCC's correspondem a densidades Gama-truncadas,

$$w_j|\mathbf{y},\boldsymbol{\beta},\mathbf{u},n,s_u^2,n_e,s_e^2,n_u,s_u^2 \sim Gama-truncada\left(n+\frac{1}{2},\frac{(y_j-m_j)^2}{2s_e^2}\right)$$

[20]

para j =1,2,...,n e $0 < w_j < 1$.

Finalmente para o parâmetro de robustez n, a DCC depende das escolhas de $p(w_j|n)$ e $p(n)$. Sob a especificação *t* de Student em [13], a DCC de n é dado por:

$$p(n|\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{w},s_u^2,n_e,s_e^2,n_u,s_u^2) \propto \frac{\left(\frac{n}{2}\right)^{n/2}}{\left(\Gamma\left(\frac{n}{2}\right)\right)^n}\left(\prod_{j=1}^{n} w_j^{\frac{n}{2}-1}\right)\exp\left(-\frac{n}{2}\left(\sum_{j=1}^{n} w_j\right)\right)p(n),$$

[21]

a qual não tem a forma de uma distribuição conhecida, independentemente da especificação de $p(n)$. Entretanto, podem ser geradas amostras de [21] usando-se um passo de Metropolis-Hastings (Chib & Greenberg, 1995).

Alternativamente, adotando a especificação Slash em [14] e *uma priori* $n \sim Gama(a_n, b_n)$, convenientemente conjugada para que a DCC de tenha uma forma conhecida, também Gama, dada por:

$$n|\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{w},s_u^2,n_e,s_e^2,n_u,s_u^2 \sim Gama\left(n+a_n, b_n - \sum_{j=1}^{n}\log(w_j)\right)$$

[22]

Similarmente ao modelo animal, a estratégia de amostragem é baseada na geração de uma cadeia de amostras sucessivas de [16], [17], [18], [19] ou [20] e [21] ou [22], derivando a inferência nos parâmetros ou funções desses parâmetros do modelo a partir de amostras salvas após um período inicial de aquecimento da cadeia. Mais detalhes da derivação e implementação desse modelo podem ser encontrados, por exemplo, em Rosa et al. (2003) e Cardoso et al., (2005).

Análise robusta do desempenho pós-desmama de bovinos cruzados Hereford-Nelore

O exemplo a seguir foi originalmente apresentado em Cardoso et al. (2005) e corresponde à análise de n = 22.717 registros de ganho pós-desmame (GPD), coletados entre 1974 e 2000, em uma população de bovinos das raças Hereford, Nelore e suas cruzas, dentro do programa de avaliação genética da Conexão Delta G. Os efeitos fixos e aleatórios do modelo linear misto na equação [11] foram baseados no modelo multirracial descrito por Cardoso & Tempelman (2004), sendo o mesmo para três especificações alternativas da densidade marginal dos resíduos: Normal ou Gaussiana, *t* de Student e Slash.

Os dados foram analisados usando o Programa INTERGEN (Cardoso, 2008), adotando um comprimento da cadeia de MCMC de 200.000 ciclos, após 15.000 ciclos de aquecimento para os três modelos. As amostras foram salvas a cada dez ciclos, tal que 20.000 amostras foram utilizadas para inferência. As médias, modas, e os desvios-padrão dos parâmetros foram obtidos de suas respectivas densidades posteriores marginais.

O gráfico de dispersão dos resíduos padronizados por grupo contemporâneo (GC), derivado da análise do GPD utilizando o modelo Gaussiano (Figura 1), apresenta 0,25% dos dados com resíduos fora do intervalo ± 4,0 erros padrão, sendo esta percentagem 30 vezes maior que a esperada neste modelo. Portanto, a utilização de modelos robustos é indicada para identificar e atenuar o efeito de observações extremas na estimação dos parâmetros a partir desse conjunto de dados.

A densidade posterior da variável de ponderação (w_j) pode ser usada para verificar o grau de afastamento da observação correspondente (y_j) em relação a pressuposição de distribuição normal (Rosa et al., 2003). Valores próximos a zero de w_j sugerem que a observação y_j é um dado extremo, enquanto que w_j próximo a um indica que y_j é praticamente igual ao valor predito pelo modelo. Para ilustrar este ponto, foram selecionados deliberadamente as observações de três machos F_1 pertencentes ao mesmo GC de 160 animais. Baseado no modelo Gaussiano, a primeira observação (y_1) representa a um dado moderadamente extremo, estando 3,08 desvios-padrão (DP) acima de zero, a segunda observação y_2 representa um resíduo próximo a zero ou ajuste perfeito do modelo (0,02 DP), e a terceira observação corresponde a um dado extremo, estando -5,57 DP de zero (Figura 1).

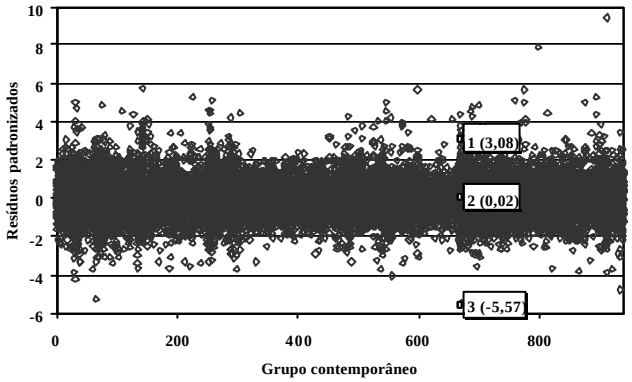

Figura 1 - Gráfico de disperção de resíduos padronizados para ganho pós-desmame dentro de grupo contemporâneo utilizando o modelo Gaussiano. Três resíduos no mesmo grupo contemporâneo estão em destaque: 1. Representa um dado moderadamente extremo +3,08 desvios padrão (DP) acima de zero; 2. Representa um resíduo próximo a zero (ajuste perfeito) e 3. Consiste em um dado extremo, –5,57 DP de zero.

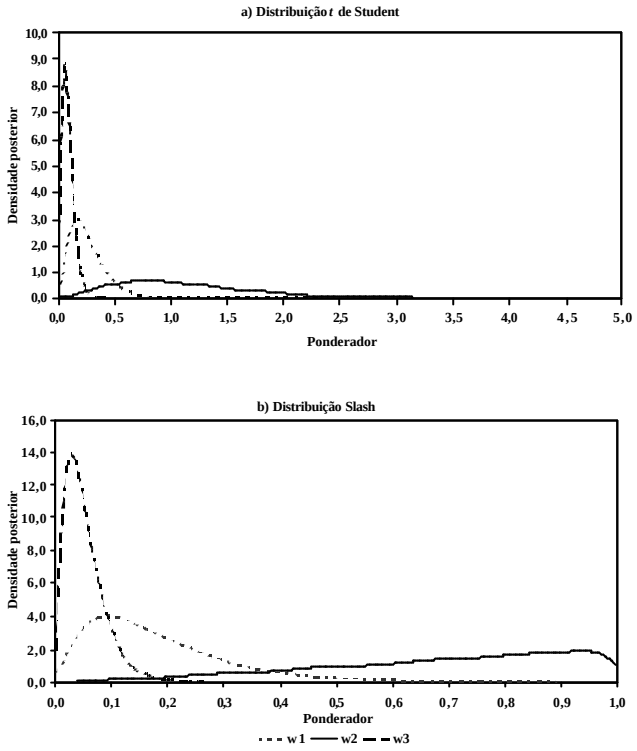

Figura 2 - Distribuição a posteriori de variáveis de ponderação correspondentes à observação 1 (w1 – dado moderadamente extremo); observação 2 (w2 –ajuste quase perfeito do modelo) e observação 3 (w3 – dado extremo) sob os modelos robustos a) modelo *t* de Student e b) Modelo Slash.

Densidades posteriores das variáveis de ponderação w_1, w_2, e w_3, correspondentes às três observações y_1, y_2, e y_3, respectivamente, são apresentadas para os modelos *t* de Student e Slash na Figura 2. Diferenças na forma da distribuição posterior de w_1, w_2, e w_3 entre esses modelos são devidas a que w_j assume qualquer valor real positivo no modelo *t* de Student, enquanto que w_j é definida somente entre 0 e 1 no modelo Slash (Lange & Sinsheimer, 1993). Não obstante, ambos modelos identificaram a observação y_3 como sendo extrema, visto que a moda posterior e o intervalo de probabilidade *a posteriori* (IPP) de 95% para w_3 no modelo *t* de Student foram 0,06 e [0,02; 0,22], respectivamente, e valores correspondentes para o modelo Slash foram 0,03 e [0,01; 0,14]. Ambas as densidades concentraram-se em torno de valores baixos para w_3, detectando y_3 com um dado extremo e automaticamente mitigando seu efeito na inferência sobre os parâmetros do modelo, uma vez que, na prática, sua contribuição foi reduzida a 6 e 3%, respectivamente nos modelos *t* de Student e Slash, daquela observada no modelo Gaussiano, onde todos os ponderadores são iguais a um. Já w_2 teve uma distribuição plana no espaço amostral com moda posterior e o IPP de 95% de 1,17 e [0,23; 2,87], e 0,94 e [0,17; 0,99], respectivamente para os mesmos modelos, indicando que y_2 não é um dado extremo. Finalmente, a moda posterior de 0,17 e o IPP de 95% de [0,03; 0,69} para w_1 no modelo *t* de Student, demonstra que a influencia de y_1 na inferência estatística também é reduzida neste modelo, embora uma menor extensão relativa a y_3. Isto também foi observado no modelo Slash com moda posterior e IPP de 95% para w_1 de 0,15 e [0,02; 0,52], respectivamente.

Note que no caso da distribuição *t* de Student, a expectativa a priori para w_j na equação [13] é o valor neutro de 1,00 e, portanto, a inclusão desse valor no IPP proporciona um critério objetivo para julgar se um dado é extremo em relação à especificação Gaussiana. Por exemplo, acima o IPP de 95% desse modelo contem 1,00 para w_2, mas não para w_1 e w_3, indicando que somente y_2 não poderia ser considerado um dado extremo por este critério.

Esses resultados ilustram que os modelos de caudas longas utilizam uma ponderação específica a cada observação, de tal forma que os registros discrepantes ($w_j << 1$) naturalmente fornecem uma contribuição proporcionalmente menor na estimação de parâmetros quando comparados aos demais registros. Porconseguinte, estes modelos fornecem um tratamento estatístico muito mais adequado aos conjuntos com dados extremos, que a alternativa de detectar e eliminar (ou ajustar) os registros discrepantes por métodos empíricos.

Generalização do modelo animal para análise de dados censurados

Para contemplar a análise de dados censurados, vamos reformular o primeiro estágio do modelo básico. Neste caso, o vetor de dados y pode ser particionado em dois subvetores y_o contendo os m registros observados (isto é, não censurados) e y_c contendo os $n - m$ registros censurados, portanto $\mathbf{y}' = [\mathbf{y}'_o \quad \mathbf{y}'_c]$. A forma da densidade da distribuição condicional de y_o é igual àquela apresentada na equação [2], somente que agora se refere a apenas m registros observados:

$$p\left(\mathbf{y}_o \middle| \boldsymbol{ß}, \mathbf{u}, s_e^2\right) = \left(2ps_e^2\right)^{-m/2} \exp\left[\frac{-1}{2s_e^2}\sum_{j=1}^{m}\left(y_j - m_j\right)^2\right] \quad [23]$$

E para os registros censurados é:

$$p\left(y_{cj} > c_j \middle| \boldsymbol{ß}, \mathbf{u}, s_e^2\right) = 1 - \Phi\left(\frac{c_j - m_j}{s_e}\right), \quad [24]$$

em que y_{cj} é o valor não observado da variável resposta,

c_j é o valor observado dessa variável no momento em que foi censurada e $\Phi(\cdot)$ é a função de densidade cumulativa da distribuição normal padrão. Desde forma, a densidade condicional de todos os registros considerando é dada por:

$$p\left(\mathbf{y}|\boldsymbol{\beta},\mathbf{u},s_e^2\right)=\left(2ps_e^2\right)^{-m/2}\exp\left[\frac{-1}{2s_e^2}\sum_{j=1}^{m}\left(y_j-m_j\right)^2\right]\times\prod_{j=m+1}^{n}\left[1-\Phi\left(\frac{c_j-m_j}{s_e}\right)\right]$$

[23]

Entretanto, a inferência a partir dessa densidade pode ser simplificada pela sua ampliação através da inclusão de uma variável v_j para representar os registros não observados dos animais com dados censurados (Tanner, 1993; Guo et al., 2001). Se $\mathbf{v}=\{v_j\}$, com $v_j > c_j$, para $j = m+1,m+2, ...,n$, tal que o vetor de dados ampliados é dado por $\mathbf{y}_A'=[\mathbf{y}_o' \quad \mathbf{v}']$ com densidade condicional nos parâmetros:

$$p\left(\mathbf{y}_A|\boldsymbol{\beta},\mathbf{u},s_e^2\right)=\left(2ps_e^2\right)^{-m/2}\exp\left\{\frac{-1}{2s_e^2}\left[\sum_{j=1}^{m}\left(y_j-m_j\right)^2+\sum_{j=m+1}^{n}\left(v_j-m_j\right)^2\right]\right\},$$

[24]

em que $v_j > c_j$ para os animais com dados censurados, ou seja, os valores não observados têm distribuição normal truncada no limite inferior do valor observado da característica no momento da avaliação.

As demais pressuposições do segundo e terceiro estágios do modelo hierárquico para os parâmetros de locação e dispersão são iguais as apresentadas para o modelo animal com dados completos em [3], [4], [5] e [6]. A distribuição posterior conjunta de todos os parâmetros do modelo, inclusive v, é obtida pelo produto de todas as densidades especificadas nos três estágios:

$$p\left(\boldsymbol{\beta},\mathbf{u},\mathbf{v},s_e^2,s_u^2|\mathbf{y}_o,n_e,s_e^2,n_u,s_u^2\right)$$
$$=p\left(\mathbf{y}_A|\boldsymbol{\beta},\mathbf{u},s_e^2\right)p(\boldsymbol{\beta})\,p(\mathbf{u}|s_u^2)\,p\left(s_e^2|n_e,s_e^2\right)p\left(s_u^2|n_u,s_u^2\right)\prod_{j=m+1}^{n}I\left(v_j>c_j\right)$$

[25]

Neste caso $I(\cdot)$ é uma variável que indica que os registros $j = m+1, m+2, ..., n$, são censurados. Todas as distribuições condicionais completas (DCC) dos parâmetros derivadas de [25] têm formas conhecidas facilitando a implementação da inferência via amostragem de Gibbs (Guo et al., 2001). Como antes, para os parâmetros de locação $\boldsymbol{\theta}'=[\boldsymbol{\beta} \quad \mathbf{u}']$, a DCC é normal multivariada:

$$\boldsymbol{\beta},\mathbf{u}|\mathbf{y},\sigma_e^2,\sigma_u^2,\mathbf{v},s_e^2,n_e,s_u^2,n_u \sim N\left(\hat{\boldsymbol{\theta}},\mathbf{C}^{-1}\sigma_e^2\right), \qquad [26]$$

em que $\hat{\boldsymbol{\theta}}$ é a solução para o sistema de equações $\mathbf{C}\boldsymbol{\theta}=\hat{\mathbf{h}}$, sendo que C é a mesma matriz de coeficientes do modelo animal e $\hat{\mathbf{b}}=\begin{bmatrix}\sum_{j=1}^{m}\mathbf{x}_j y_j+\sum_{j=m+1}^{n}\mathbf{x}_j v_j\\ \sum_{j=1}^{m}\mathbf{z}_j y_j+\sum_{j=m+1}^{n}\mathbf{z}_j v_j\end{bmatrix}$.

Já para os registros não observados, a DCC tem a seguinte forma:

$$v_j\left|\boldsymbol{\beta},\mathbf{u},\mathbf{v}_{-j},s_e^2,s_u^2,\mathbf{y}_o,n_e,s_e^2,n_u,s_u^2\right. \sim N\left(m_j,s_e^2\right)I\left(v_j>c_j\right)$$

[27]

Neste caso, os v_j's, $j = m+1, m+2, ..., n$, são simulados a partir de uma distribuição normal, com média determinada pelos efeitos fixos e aleatórios no modelo e variância s_e^2. A função $I\left(v_j>c_j\right)$ indica que essa distribuição é truncada, tal que somente são possíveis valores de v_j maiores que o valor observado para a variável censurada no momento da análise c_j. O vetor corresponde a $\mathbf{v}_{-j}$, exceto v_j.

A forma das DCC para os parâmetros de dispersão também permanece a mesma do modelo animal, apenas refletindo a ampliação de registros:

$$s_e^2\left|\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{v},s_u^2,n_e,s_e^2,n_u,s_u^2\right. \sim\left(\sum_{j=1}^{m}\left(y_j-m_j\right)^2+\sum_{j=m+1}^{n}\left(v_j-m_j\right)^2+n_e s_e^2\right)c_{(n_e+n)}^{-2}$$

[28] e $s_u^2\,|\,\mathbf{y},\boldsymbol{\beta},\mathbf{u},\mathbf{v},s_e^2,n_e,s_e^2,n_u,s_u^2 \sim\left(\mathbf{u}'\mathbf{A}^{-1}\mathbf{u}+n_u s_u^2\right)c_{(n_u+q)}^{-2}$

[29].

A estratégia para implementar a amostragem de Gibbs é baseada na geração de uma cadeia de amostras sucessivas de [26], [27], [28] e [29], descartando-se as amostras iniciais. Das amostras após o período de descarte é derivada a inferência nos parâmetros do modelo.

Análise dados censurados de dias para parir em bovinos de corte

Em bovinos de corte, o tempo em dias entre o início da estação de monta e o parto subseqüente da vaca, determina o valor fenotípico da característica dias para parir, que é considerada adequada para avaliação genética do desempenho reprodutivo de vacas em programas de larga escala em nível de fazenda, quando se adota um período fixo de estação de monta (Macgregor & Casey, 1999). Entretanto, como algumas vacas não concebem durante a estação, os dias para parir não são observados para todas as vacas e essa característica é um exemplo de dados censurados.

Alguns autores propuseram a atribuição a todas as vacas que não pariram um valor projetado único com base no seu grupo contemporâneo, por exemplo, penalizando esses registros com 21 dias além do último parto observado (Johnston & Bunter, 1996). Neste caso, todas as fêmeas de

um mesmo grupo de contemporâneas que não pariram recebem o mesmo valor para a característica.

Donoghue et al. (2004a,b), compararam por meio de simulação e analisando registros de dias para o primeiro parto de 33.176 novilhas em rebanhos Angus australianos, a adequação de duas estratégias para lidar com dados censurados de dias para parir, a penalização dentro de grupo contemporâneo com utilização do modelo animal tradicional e o modelo hierárquico com ampliação de dados descrito na seção acima. Os autores não observaram diferenças significativas entre os modelos em termos de correlações de Pearson entre os valores genéticos preditos e verdadeiros. Entretanto, o modelo hierárquico de ampliação de dados proporcionava um melhor ajuste aos dados, estimativas mais precisas da variância residual e da herdabilidade da característica e também uma melhor identificação de quais touros estavam entre os 10% piores quando o percentual de registros censurados era de 20%. Esses resultados permitem concluir que o modelo de ampliação de dados é uma alternativa adequada para a avaliação genética de dias para parir em bovinos de corte.

Conclusões

Neste artigo evidenciamos a flexibilidade dos modelos hierárquicos bayesianos para ajustar características de desempenho animal, que pela sua complexidade não podem ser adequadamente analisadas por meio dos modelos lineares usuais. Neste artigo, consideramos a robustez a dados extremos e a análise de dados censurados, mas muitas outras situações podem ser contempladas, com, por exemplo, a modelos de limiar para características categóricas ordenadas (Sorensen et al., 1995), dados de contagens (Tempelman & Gianola, 1996), acasalamento com reprodutores múltiplos (Cardoso & Tempelman, 2003) e modelos estruturais para heterogeneidade de variância (Foulley et al., 1992; Rekaya et al., 2003). Além disso, a construção hierárquica desses modelos permite combinar as especificações acima de acordo com as características dos dados a serem analisados, por exemplo, formando um modelo limiar robusto (Kizilkaya et al., 2003) ou robusto com heterogeneidade de variância (Cardoso et al., 2005; Kizilkaya & Tempelman, 2005). Note ainda que, no presente caso, pode-se derivar um modelo robusto para análise de dados censurados combinando as especificações em [12] e [24], tal que o vetor de dados ampliados $\mathbf{y}_A' = [\mathbf{y}_o' \quad \mathbf{v}']$ tem a seguinte densidade condicional nos parâmetros:

$$p(\mathbf{y}_A|\beta, \mathbf{u}, s_e^2, \mathbf{w}) = (2\pi s_e^2)^{-m/2} \exp\left\{\frac{-1}{2s_e^2}\left[\sum_{j=1}^{m} w_j (y_j - m_j)^2 + \sum_{j=m+1}^{n} (v_j - m_j)^2\right]\right\}$$

[30]

Desta forma os registros não censurados seriam ponderados por w_j dando robustez a dados extremos e os registros não observados em $\mathbf{v}$ seria simulados a partir de um DCC análoga a em [27].

Finalmente, a intensa demanda computacional das análises baseadas em MCMC tem limitado seu uso em maior escala nos programas de avaliação genética, entretanto, com evolução crescente da capacidade dos computadores e a queda dos custos desses equipamentos, espera-se que nos próximos anos um número crescente de programas adotem métodos MCMC e modelos hierárquicos mais apropriados para características de importância econômica em produção animal.

Literatura Citada

BERTRAND, J.K.; WIGGANS, G. R. Validation of data and review of results from genetic evaluation systems for US beef and dairy cattle. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. **Proceedings...** Armidale, 1998. p.425-432.

CARDOSO, F. F. **Manual de utilização do programa INTERGEN – Versão 1.0 em estudos de genética quantitativa animal**. Bagé: Embrapa Pecuária Sul, 2008. p.48. (Documentos, 74).

CARDOSO, F.F.; ROSA, G.J.M.; TEMPELMAN, R.J. Multiple-breed genetic inference using heavy-tailed structural models for heterogeneous residual variances. **Journal of Animal Science**, v.83, n.8, p.1766-1779, 2005.

CARDOSO, F.F.; TEMPELMAN, R.J. Bayesian inference on genetic merit under uncertain paternity. **Genetics Selection Evolution**, v.35, n.5, p.469-487, 2003.

CARDOSO, F.F.; TEMPELMAN, R.J. Hierarchical Bayes multiple-breed inference with an application to genetic evaluation of a Nelore-Hereford population. **Journal of Animal Science**, v.82, n.6, p.1589-1601, 2004.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hastings algorithm. **American Statistician**, v.49, n.4, p.327-335, 1995.

DONOGHUE, K.A.; REKAYA, R.; BERTRAND, J.K. Comparison of methods for handling censored records in beef fertility data: field data. **Journal of Animal Science**, v.82, n.2, p.357-361, 2004a.

DONOGHUE, K.A.; REKAYA, R.; BERTRAND, J.K. Comparison of methods for handling censored records in beef fertility data: Simulation study. **Journal of Animal Science**, v.82, n.2, p.351-356, 2004b.

DUCROCQ, V.; CASELLA, G. A Bayesian analysis of mixed survival models. **Genetics Selection Evolution**, v.28, n.6, p.505-529, 1996.

DUCROCQ, V.; SÖLKNER, J. "The Survival Kit V3.0' - a package for large analyses of survival data. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. **Proceedings...** Armidale: 1998. p.447-448.

FOULLEY, J.L.; CRISTOBAL, M.S.; GIANOLA, D. et al. Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian

models. **Computational Statistics & Data Analysis**, v.13, n.3, p.291-305, 1992.

GIANOLA, D.; FERNANDO, R.L. Bayesian methods in animal breeding theory. **Journal of Animal Science**, v.63, n.1, p.217-244, 1986.

GUO, S.F.; GIANOLA, D.; REKAYA, R. et al. Bayesian analysis of lifetime performance and prolificacy in Landrace sows using a linear mixed model with censoring. **Livestock Production Science**, v.72, n.3, p.243-252, 2001.

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

JOHNSTON, D.J.; BUNTER, K.L. Days to calving in Angus cattle: Genetic and environmental effects, and covariances with other traits. **Livestock Production Science**, v.45, n.1, p.13-22, 1996.

KIZILKAYA, K.; CARNIER, P.; ALBERA, A. et al. Cumulative t-link threshold models for the genetic analysis of calving ease scores. **Genetics Selection Evolution**, v.35, n.5, p.489-512, 2003.

KIZILKAYA, K.; TEMPELMAN, R.J. A general approach to mixed effects modeling of residual variances in generalized linear mixed models. **Genetics Selection Evolution**, v.37, n.1, p.31-56, 2005.

LANGE, K.; SINSHEIMER, J.S. Normal/independent distributions and their applications in robust regression. **Journal of the American Statistical Association**, v.2, n.2, p.175-198, 1993.

MACGREGOR, R.G.; CASEY, N.H. Evaluation of carving interval and calving date as measures of reproductive performance in a beef herd. **Livestock Production Science**, v.57, n.2, p.181-191, 1999.

PINHEIRO, J.C.; LIU, C.H.; WU, Y.N. Efficient algo rithms for robust estimation in linear mixed- effects models using the multivariate t distribution. **Journal of Computational and Graphical Statistics**, v.10, n.2, p.249-276, 2001.

REKAYA, R.; WEIGEL, K.A.; GIANOLA, D. Bayesian estimation of parameters of a structural model for genetic covariances between milk yield in five regions of the United States. **Journal of Dairy Science**, v.86, n.5, p.1837-1844, 2003.

ROSA, G.J.M.; PADOVANI, C.R.; GIANOLA, D. Robust linear mixed models with Normal/Independent distributions and Bayesian MCMC implementation. **Biometrical Journal**, v.45, n.4, p.573-590, 2003.

SORENSEN, D.A.; ANDERSEN, S.; GIANOLA, D. et al. Bayesian-inference in threshold models using Gibbs sampling. **Genetics Selection Evolution**, v.27, n.3, p.229-249, 1995.

SORENSEN, D.A.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. 1.ed. New York: Springer-Verlag, 2002. 740p.

STRANDEN, I.; GIANOLA, D. Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study. **Genetics Selection Evolution**, v.30, n.6, p.565-583, 1998.

STRANDEN, I.; GIANOLA, D. Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. **Genetics Selection Evolution**, v.31, n.1, p.25-42, 1999.

TANNER, M.A. **Tools for statistical inference**: methods for the exploration of posterior distributions and likelihood functions. 2.ed. New York: Springer-Verlag, 1993. 156p.

TEMPELMAN, R.J.; GIANOLA, D. A mixed effects model for overdispersed count data in animal breeding. **Biometrics**, v.52, n.1, p.265-279, 1996.

VANRADEN, P.M.; KLAASKATE, E.J.H. Genetic evaluation of length of productive life including predicted longevity of live cows. **Journal of Dairy Science**, v.76, n.9, p.2758-2764, 1993.

WANG, C.S.; RUTLEDGE, J.J.; GIANOLA, D. Bayesian-analysis of mixed linear-models via Gibbs sampling with an application to litter size in Iberian pigs. **Genetics Selection Evolution**, v.26, n.2, p.91-115, 1994.