INFERÊNCIA BAYESIANA NA ESTIMAÇÃO DE COMPONENTES DE VARIÂNCIA DE BOVINOS SIMENTAL[1]

GILMARA BRUSCHI SANTOS[2], HENRIQUE NUNES DE OLIVEIRA[3], GUILHERME JORDÃO DE MAGAHÃES ROSA[4], LUIS FERNANDO AARÃO MARQUES[5]

[1] Parte da dissertação de mestrado do primeiro autor, bolsista da FAPESP
[2] Doutoranda em produção animal, Faculdade de Medicina Veterinária e Zootecnia, UNESP, Botucatu
[3] Professor da Faculdade de Medicina Veterinária e Zootecnia, UNESP, Botucatu
[4] Professor da Universidade de Michigan
[5] Professor da Universidade Estadual do Espírito Santo

RESUMO - Os pesos à idade de 730 dias foram analisados para verificação da presença de heterogeneidade de variância e posterior estimação dos componentes de variância. O objetivo do trabalho foi comparar modelos gaussianos e um modelo com distribuição normal contaminada para estimação dos componentes de variância. Esta última distribuição é menos sensível a observações com valores discrepantes e pode gerar predições mais acuradas dos valores genéticos. Foi utilizada uma abordagem Bayesiana para implementação das análises. Os resultados sugerem que, embora os cálculos sejam um pouco mais trabalhosos para o modelo misto, ele pode apresentar inferências mais robustas em situações tanto com presença de heterogeneidade de variâncias quanto com presença de observações discrepantes.

PALAVRAS-CHAVE: distribuição normal contaminada, Heterogeneidade de variâncias, modelo robusto

BAYESIAN INFERENCE ON VARIANCE COMPONENTS ESTIMATION OF SIMENTAL CATTLE

ABSTRACT - Weight at day 730 were analysed to account for variance heterogeneity and variance components estimation. The aim of this study was to compare gaussian and a robust model for estimation of variance components. This late distribution is less sensible with outliers and have more accurately predictions of breeding values. It was used a Bayesian approach for analysis. Results sugest that, athouth calcules are a little more dificult for the mixed model, it may have more robust inferences either on situations with variance heterogeneity or outliers.

KEYWORDS: contaminated normal distribution, robust model, Variance heterogeneity

INTRODUÇÃO

A heterogeneidade de variâncias do resíduo ou presença de valores discrepantes (outliers) pode influenciar na distribuição dos dados, o que pode levar a estimativas equivocadas dos componentes de variância do modelo adotado. Algumas alternativas podem ser usadas para se corrigir possíveis erros nas estimativas. Entre elas destaca-se a utilização de modelos robustos que têm sido aplicados através de métodos Bayesianos.

Entre os modelos para estimação robusta deve-se considerar aqueles que utilizam as distribuições normal-independentes, que representam um interessante grupo de distribuições de caudas longas (leptocúrticas) (Rogers e Tukey, 1972). Algumas destas distribuições, tais como a normal contaminada e a t de Student têm sido testadas como alternativas à distribuição normal em modelos mistos (Strandén and Gianola, 1999; Rosa, 1999; Pereira 2001). Será discutida neste trabalho a distribuição normal contaminada.

O objetivo deste trabalho foi verificar a aplicabilidade do modelo robusto, utilizando a distribuição normal contaminada em abordagem Bayesiana, na estimação de componentes de variância, na característica peso aos 730 dias de animais da raça Simental.

MATERIAL E MÉTODOS

Os dados utilizados nas análises são provenientes dos serviços de Genealogia e de Controle de Desenvolvimento Ponderal, dos arquivos da Associação Brasileira de Criadores da Raça Simental (ABCRS). O arquivo de pedigree continha 29.872 animais. O arquivo referente ao peso aos 730 dias continha dados de 3.559 animais, filhos de 526 touros e 1885 vacas, distribuídos em 574 grupos de contemporâneos. Estes dados foram submetidos a uma análise crítica através do programa computacional (*Statistical Analysis System*), versão 6.12 (SAS, 1996). Procedeu-se assim à

eliminação de registros inconsistentes e formação dos grupos de contemporâneos a serem considerados como efeitos fixos (ambiente) nos modelos estatísticos. Estes grupos foram definidos como animais de mesmo sexo, nascidos no mesmo ano-estação, criados sob igual regime alimentar, na mesma fazenda. Em seguida realizou-se a prova de Kolmogorov-Smirnov para verificar a normalidade das curvas de distribuição dos dados. Para verificação da presença de heterogeneidade de variâncias foi utilizado o teste de Levene, numa modificação originalmente proposta por Brown & Forsythe (1974). A qual consiste em usar a mediana no lugar da média para calcular os desvios, o que torna o teste bem mais robusto. A estimação dos componentes de variância foi feita, num primeiro momento pelo método frequentista REML (restricted maximum likelihood) sendo as análises implementadas por meio do *software* MTDFREML (*Multiple Trait Derivative-Free Restricted Maximum Likelihood*) desenvolvido por Boldman et al.(1993). Foi utilizado um modelo animal, supondo-se distribuição gaussiana dos resíduos. Não mais do que três reinícios, utilizando-se os resultados da rodada anterior como valor inicial na rodada subseqüente, foram necessários para garantir a convergência a um máximo global. Para representação das análises por este modelo, a sigla GML foi utilizada. Para o uso da abordagem bayesiana na estimação dos componentes de (co) variância sob modelo animal também com distribuição gaussiana dos resíduos foi usado o *software* MTGSAM (*Multiple Trait Gibbs Sampling in Animal Models*), desenvolvido por Van Tassel e Van Vleck (1995). Assumiu-se que as distribuições a priori para os componentes de variância e efeitos fixos eram desconhecidas (priores flat ou não informativas). Os valores para iteração na rodada inicial foram obtidos na literatura. As densidades marginais dos componentes de variância foram estimadas a partir das amostras geradas pelo amostrador de Gibbs. A inspeção gráfica e o programa Gibanal (VanKaam, 1998) foram usados para determinar a convergência. Foram realizadas no total 750.000 iterações do amostrador de Gibbs. As 1.000 primeiras iterações foram descartadas para permitir que a distribuição inicial, fornecida como priori, não interferisse nos resultados; e para evitar a redundância das informações, causada pela correlação serial entre amostras geradas subseqüentemente, foi tomada apenas uma amostra a cada 350 geradas. Para este modelo utilizou-se a sigla BG para sua representação. Como alternativa aos modelos acima descritos, foi utilizado um programa de computador específico para estas análises com o modelo robusto e distribuição normal contaminada. Este programa é uma modificação efetuada por Pereira (2001) em Fortran 77, a partir de um programa desenvolvido pelo pesquisador Daniel Sorensen no Instituto Dinamarquês de Ciência Animal, para análise Bayesiana com modelos Gaussianos. As mesmas condições usadas para o modelo gaussiano foram também adotadas neste modelo. Para o modelo robusto utilizou-se a sigla BM para sua representação.

RESULTADOS E DISCUSSÃO

A representação gráfica das distribuições da característica peso aos 730 dias de idade e de seus desvios em relação à média dos contemporâneos está na Figura 1. Estão apresentadas as distribuições de freqüência das características na forma de histogramas e as distribuições normais esperadas com média e variância iguais às estimadas para as características na forma de linhas contínuas. Embora não seja aparente na distribuição da característica observada, fica claro, no caso em que são apresentados os desvios em relação à média dos contemporâneos a forma leptocúrtica (caudas longas) da distribuição. Em relação à distribuição normal nota-se que há um excesso na região da moda, o que é típico desta forma de distribuição. Possivelmente existe uma heterogeneidade de variâncias e/ou valores discrepantes na característica estudada. Pode-se inferir, de acordo com estes resultados que a pressuposição de normalidade pode não ser a mais adequada para a estimação dos componentes de variância para a característica. Os valores discrepantes podem influenciar de maneira muito significativa estes resultados. Encontram-se, na Tabela 1, os componentes de variância e herdabilidade estimados com o modelo GML e as médias posteriores pelos modelos BG e BM. As estimativas de ϕ e τ para o modelo BM foram 0,2747 e 0,1083. Esta característica apresentou pequena proporção de indivíduos da população com maior variância, o que pode ser devido ao descarte seletivo que tende a homogeneizar os rebanhos, ou ainda à exclusão de certos rebanhos onde o cuidado com a coleta de dados é menor e onde, possivelmente, as pesagens são interrompidas mais cedo. A variância da população contaminante é 9,23 vezes o valor da população base. Aproximadamente 28% dos animais apresentaram variância residual maior. Pereira (2001), trabalhando com peso ao nascimento de bovinos Simental encontrou diferenças muito maiores de variância entre as subpopulações. Entretanto, o autor atribuiu tal discrepância à baixa qualidade do conjunto de dados. Mais uma vez, as médias das distribuições posteriores dos componentes de variância do modelo BG diferiram das dos outros dois modelos. O componente de

variância genética foi mais alto e o de variância residual mais baixo, resultando em herdabilidade bem mais alta por este modelo. Em cada iteração do amostrador de Gibbs, os animais com registro de produção são classificados como sendo de uma das subpopulações. A observação do número proporcional de vezes que o animal foi classificado na população de variância mais alta permitiu identificar erros na formação de grupos de contemporâneos para o peso aos 730 dias. Os animais classificados erroneamente apareciam como observações discrepantes dentro dos grupos, e nos modelos gaussianos, parte do desvio em relação aos contemporâneos era atribuída ao valor genético destes animais. Este dado empírico indica que o modelo BM pode acomodar melhor as observações discrepantes do que os modelos gaussianos. Pereira (2001) em estudo de simulação já havia observado que este modelo é bem superior ao modelo gaussiano em situações em que há heterogeneidade de variância não sistemática.

CONCLUSÕES

As estimativas de variâncias genéticas produzidas pelo modelo robusto utilizado no presente trabalho foram semelhantes às estimativas de máxima verossimilhança restrita em modelo gaussiano, enquanto que para o resíduo foi identificada uma mistura de distribuições normais com diferentes variâncias.

REFERÊNCIAS BIBLIOGRÁFICAS

PEREIRA, I. G. **Estudo se simulação e aplicação de modelos lineares mistos com distribuição normal contaminada no melhoramento genético animal**. Botucatu, FMVZ/UNESP, 2001. 91p. (Tese – Doutorado em Zootecnia).

ROGERS, W. H.; TUKEY, J. W. Understanding some long-tailed distributions. **Statistica Neerlandia**, v.26, p.211-226, 1972.

ROSA, G. J. M. **Análise bayesiana de modelos lineares mistos robustos via amostrador de Gibbs.** Piracicaba, ESALQ, 1998. 57p. (Tese – Doutorado em Estatística).

SAS . *User's Guide: Statistics,* Cary: SAS INSTITUTE. 956p., 1996

STRANDÉN, I. J. **Robust mixed effects linear models with t distributions and application to dairy cattle breeding**. Madison, 1996. 176p. Thesis (PhD) – University of Wisconsin.

VanKAAM, J. B. C. H. M. (1998). Disponível em:

<http://www.student.wau.nl/~janthijs/breedingsite/eadgibanal.html>

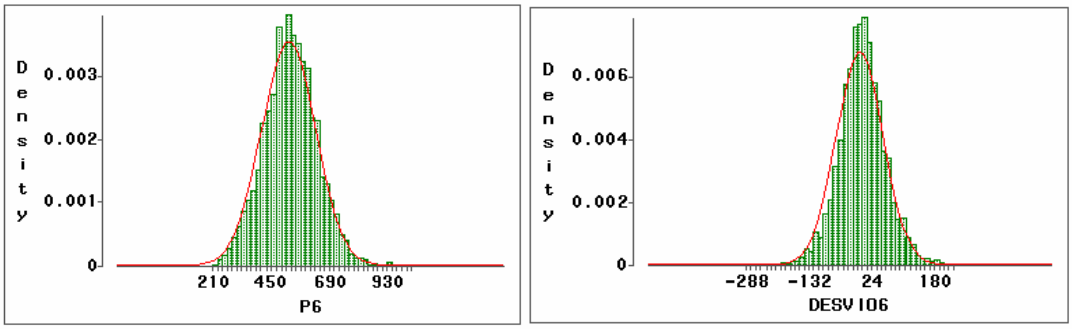

FIGURA 1. Distribuição de freqüência do peso aos 730 dias de bovinos Simental (à esquerda) e das freqüências de seus desvios em relação a media dos contemporâneos

TABELA 1. Estimativas dos componentes de variância e herdabilidade pelo modelo GML e médias a posteriori, pelos modelos BG e BM para a característica peso aos 730 dias. [A] Primeira população; [B] Segunda população; [C] Média ponderada das duas populações

MODELO	σ_g^2	σ_e^2	h^2
GML	1586,29	2747,38	0,37
BG	2330,65	2290,33	0,50
BM[A]	1428,29	919,42	0,60
BM[B]	1428,29	8489,56	0,14
BM[C]	1428,29	2998,93	0,32