Efeito do descarte de valores discrepantes na avaliação genética animal

José Marques Carneiro Júnior[1], Priscila Ferreira Wolter[3], Giselle Mariano Lessa de Assis[2], Williane Maria de Oliveira Martins[4], Francisco Aloísio Cavalcante[5]

[1]Pesquisador Embrapa Acre. E-mail: marques@cpafac.embrapa.br
[2]Pesquisador Embrapa Acre. E-mail: giselle@cpafac.embrapa.br
[3]Graduanda de Ciências Biológicas da União Educacional do Norte. E-mail: priscilawolter18@yahoo.com.br
[4]Graduanda de Engenharia Agronômica da Universidade Federal do Acre. E-mail: williane_martins@yahoo.com.br
[5]Pesquisador da Embrapa Acre. E-mail: aloisio@cpafac.embrapa.br

Resumo: O objetivo deste trabalho foi avaliar o efeito da presença e da exclusão de *outliers*, na avaliação genética de touros. Foram simuladas cinco estruturas de população com 0%, 4%, 7%, 10% e 15% de *outliers*, herdabilidade média de 0,40 e três efeitos fixos para formação dos grupos de contemporâneos. Os *outliers* foram simulados inserindo-se valores discrepantes nos efeitos genéticos das vacas, de acordo com a porcentagem desejada de observações discrepantes. Foi utilizado o programa MTDFREML para predição dos valores genéticos e estimação dos efeitos de grupos de contemporâneos. As variáveis utilizadas para verificação da acurácia na avaliação genética foi o Quadrado Médio do Erro na predição dos valores genéticos (QME_VG), o Quadrado Médio do Erro na estimação dos Grupos de Contemporâneos (QME_GC) e a Correlação de Spearman. Verificou-se que a Correlação de Spearman foi de 0,78 para a estrutura de população com 0% de *outliers* e 0,69 para a estrutura com 15% de *outliers*. Os QME_VG aumentaram conforme o nível de valores discrepantes, sendo de 344,07 para 0% de *outliers* e 437,46 para a estrutura com 15% de *outliers*. Os QME_GC não apresentaram tendência definida de comportamento. Concluiu-se que a presença de *outliers* prejudica a classificação de touros e reduz a acurácia da predição dos valores genéticos; o descarte de observações discrepantes (três desvio padrão) não melhora a hierarquização dos touros nem a estimação dos efeitos fixos.

Palavras–chave: simulação de dados, avaliação genética, valores discrepantes, pressuposição do modelo

Discard Effects of discrepant values on Animal Genetic Evaluation

Abstract: The objective was to evaluate the presence and exclusion of outliers in the genetic evaluation of bulls. Five population structures were simulated with 0%, 4%, 7%, 10% and 15% outliers, average heritability of 0.40 and three fixed effects for formation of contemporary groups. The outliers were simulated inserting discrepant values in the cows genetic effects, according to the desired percentage of discrepant observations. The MTDFREML program was used to prediction genetic values and estimation of contemporary groups effects. The variables used for verification of genetic evaluation accuracy was the mean square error in breeding values prediction (QME_VG), the mean square error in contemporary groups estimation (QME_GC) and Spearman correlation. It was found that the Spearman correlation was 0.78 for population structure with 0% of outliers and 0.69 for structure with 15% of outliers. The QME_VG increased as the discrepant values level, being 0% to 344.07 and 437.46 of outliers to structure with 15% of outliers. The QME_GC showed no defined behavior trend. It was concluded that the outliers presence affect the bulls ranking and reduces the accuracy of breeding values prediction, the discard of discrepant observations (three standard deviation) does not improve the bulls ranking or the of effects fixed estimation.

Keywords: data simulated, genetic evaluation, outliers, model presupposition

Introdução

A avaliação genética animal utilizando modelos lineares mistos tem como pressuposição básica a distribuição normal tanto para os resíduos quanto para os demais efeitos aleatórios o que torna essas análises muito sensíveis à presença de valores discrepantes (*outliers*) nas observações (Rogers & Tukey, 1972). Barnett e Lewis (1995) definem *outliers* em um conjunto de dados como sendo as observações que parecem ser inconsistentes com o conjunto de dados remanescentes. Assim, considera-se como valor discrepante a observação cuja magnitude apresenta baixa probabilidade de pertencer à distribuição normal do conjunto de dados que se está trabalhando.

As possíveis correções para os *outliers* incluem a checagem e retirada de valores extremos; a redefinição da população de interesse ou a reespecificação do modelo. O critério mais utilizado tem sido o descarte de observações fora do intervalo de três desvios padrão. Entretanto, para Draper e Smith (1981), a simples exclusão de *outliers* pode não ser o procedimento correto e as regras propostas para rejeição de *outliers* deveriam incluir a reanálise sem essas observações as quais, podem ser portadoras de informações dos indivíduos de uma população.

Dentro deste contexto, o objetivo do presente trabalho é avaliar o efeito do descarte de observações discrepantes, na predição dos valores genéticos de touros e na estimação dos efeitos de grupos de contemporâneos.

Material e Métodos

Os dados utilizados neste trabalho foram simulados utilizando o software SAS (2000). Foram simuladas cinco estruturas de população com cinco repetições cada e diferentes porcentagens de valores discrepantes (0%, 4%, 7%, 10% e 15%). A herdabilidade média utilizada foi de 0,40 e três efeitos fixos foram utilizados para formar os grupos de contemporâneos. A presença de *outliers* foi simulada inserindo-se valores discrepantes nos efeitos genéticos das vacas no processo de simulação, de acordo com a porcentagem desejada.

A estimação dos efeitos fixos, dos componentes de variâncias e a predição dos valores genéticos foram realizadas por meio do programa MTDFREML (*Multiple Trait Derivative-Free Restricted Maximum Likelihood*), descrito por Boldman et al. (1995). Foram realizadas duas estratégias de análise: análise com e sem descarte de observações discrepantes. Como critério de descarte foi adotado a exclusão de observações fora do limite de três desvios padrão para mais e para menos.

Para verificar o efeito dos diferentes níveis de observações discrepantes e do descarte destas observações, foi calculado o Quadrado Médio do Erro na predição dos valores genéticos (QME_VG), como a média dos quadrados das diferenças entre os valores genéticos verdadeiros e dos valores genéticos preditos e o Quadrado Médio do Erro na estimação dos GC (QME_GC), como a média dos quadrados das diferenças entre os valores dos grupos de contemporâneos verdadeiros e dos estimados, conforme Schenkel (1998):

$$QME = \sum_{i=1}^{n} \frac{1}{n} (\hat{a}_i - a_i)^2$$

em que QME é o quadrado médio do erro, n é o número de indivíduos, âi refere-se ao valor genético predito e ai ao valor genético verdadeiro. Quanto mais próximo de zero maior é a acurácia da predição dos valores genéticos. Foi calculada também a Correlação de Spearman (CS) entre os valores genéticos verdadeiros e os preditos.

Resultados e Discussão

Na Tabela 1 encontram-se a média dos Quadrados Médios dos Erros para os valores genéticos preditos (QME_VG), a média dos Quadrados Médios dos Erros para os efeitos de grupos de contemporâneos (QME_GC) e a Correlação de Spearman (CS) entre os valores genéticos verdadeiros e preditos para as diferentes porcentagens de valores discrepantes. De forma geral, verificou-se o QME_VG aumentou conforme o nível de valores discrepantes, sendo de 344,07 para nível zero de *outliers* e 437,46 para a estrutura com 15% de *outliers*, indicando perda de eficiência da metodologia na predição dos valores genéticos. Verificou-se que a Correlação de Spearman diminuiu com o aumento da porcentagem de *outliers*, sendo de 0,78 para a estrutura de população sem *outliers* e 0,69 para a estrutura com 15% de *outliers*. Este resultado era esperado uma vez que a menor eficiência na predição dos valores genéticos pode resultar em pior classificação dos indivíduos. O QME_GC não apresentou tendência definida de comportamento.

Na Tabela 1 também está presente os resultados para QME_VG, QME_GC e CS para as análises com descarte de observações discrepantes. De modo geral é possível observar que a estratégia de descarte de observações discrepantes não resultou em melhoras na obtenção dos valores genéticos e na classificação dos touros. Observou-se também que a estratégia de descarte de valores fora de três desvios padrões eliminou indivíduos da população que não apresentava originalmente valores discrepantes. Este fato não é desejável uma vez que pode ocorrer a eliminação de indivíduos de genética superior da população.

Tabela 1 Médias dos Quadrados Médios dos Erros para os valores genéticos preditos (QME_VG), Médias dos Quadrados Médios dos Erros para efeitos dos grupos de contemporâneos (QME_GC) e Correlação de Spearman (CS) e Desvios Padrão (DP) entre valores genéticos verdadeiros e preditos, para os diferentes níveis de *outliers*.

Análises sem descarte de *outliers*			
Porcentagens de *outliers (%)*	QME_VG ± DP	QME_GC ± DP	CS ± DP
0	344,07 ± 28,04	1.438,57 ± 514,01	0,78 ± 0,02
4	403,79 ± 19,23	1.623,40 ± 358,84	0,74 ± 0,03
7	403,25 ± 45,83	1.560,20 ± 571,17	0,73 ± 0,03
10	431,46 ± 37,92	1.417,40 ± 178,35	0,72 ± 0,02
15	437,46 ± 36,34	1.436,10 ± 199,96	0,69 ± 0,02
Análises com descarte de *outliers*			
0	359,67 ± 25,86	1.456,55 ± 557,22	0,77 ± 0,01
4	413,95 ± 25,71	1.548,23 ± 365,61	0,73 ± 0,03
7	392,42 ± 35,03	1.512,64 ± 274,06	0, 73 ± 0,01
10	454,32 ± 27,92	1.310,66 ± 302,87	0, 70 ± 0,01
15	440,97 ± 28,24	1.390,78 ± 233,20	0, 69 ± 0,03

Conclusões

A presença de valores discrepantes prejudica a avaliação genética de touros, reduzindo a acurácia da predição dos valores genéticos.

A estratégia de descarte de valores discrepantes fora do intervalo de três desvios padrão não resulta em melhoras na hierarquização dos touros e nem na estimação dos efeitos de grupos de contemporâneos.

Literatura citada

BARNETT, V.; LEWIS, T. **Outliers in statistical data**. Chichester, John Wiley, 1995. 584p.

BOLDMAN, K.G., KRIESE, L.A., VAN VLECK, L.D. et al. **A manual for use of MTDFREML**. **A set of programs to obtain estimates of variances and covariances** [DRAFT].Lincoln: USDA/ARS, 1995.

DRAPER, N. R., SMITH, H. **Applied regression analysis.** New York, John Wiley, 2. ed., 1981. 709p.

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

SAS Institute Inc. SAS/STAT: User guide. Cary: SAS Institute Inc., 2000. USA.

SCHENKEL, F.S. **Studies on effects of parental selection of estimation of genetic parameters an breeding values of metric traits**. Ghelph: University of Ghelph, 1998. 191p. (Ph.d. Thesis) - University of Ghelph, 1998.