Can a model with Genetic Groups for Mendelian Sampling Deviations Correct for pre-selection bias?

W.F. Fikse[1]

[1] Department of Animal Breeding and Genetics, Swedish university of Agricultural Sciences, Uppsala, Sweden

ABSTRACT: Pre-selection of animals in breeding schemes invalidates some of the assumptions necessary to maintain the unbiasedness property in genetic evaluations. In particular, the expected value of Mendelian sampling deviations is no longer zero, and the Mendelian sampling variance is reduced. This study investigated, using simulation, the possibility to attenuate bias due to pre-selection using a genetic evaluation model with genetic groups for Mendelian sampling deviations proposed some years ago. Two generations of data were simulated. Daughter yield deviations and yield deviations were simulated for males and females, resp., in generations 1 and 2. Males in generation 1 were pre-selected on estimated Mendelian sampling deviations. Accounting for pre-selection by including genetic groups for Mendelian sampling deviations did not remove bias in estimated breeding values. The reason was that pre-selection not only led to selection differential for Mendelian sampling deviations but also for parent averages, which the model failed to account for.

Keywords: Pre-selection; Genetic groups; Mendelian sampling deviations.

INTRODUCTION

Pre-selection is a topic that has engaged animal breeders for a long time (e.g. Robertson (1966)). Most recently, the issue has received renewed attention as a result of the introduction of genomic selection. Pre-selection typically leads to violation of the important condition that data on which selection has been based must be included in the genetic evaluation to obtain unbiased breeding values (Sorensen and Kennedy (1984)). Failing to include those data leads to that assumptions for breeding values and Mendelian sampling deviations no longer are met. Patry and Ducrocq (2009) proposed a genetic evaluation model including genetic groups for Mendelian sampling deviations as a means to attenuate bias due to genomic pre-selection. However, their model has thus far not been applied to simulated or field data, and the performance of the proposed model is largely unknown. The aim of this study was to illustrate, with simulation, some properties of that approach.

Materials and Methods

Model. In a standard linear animal model the mean and variance of breeding values are assumed to be 0 and $\mathbf{A}\sigma_a^2$, resp. This model can be reformulated in terms of Mendelian sampling deviations (e.g., Mrode (2005)):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad (eq\ 1)$$
$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZTm} + \mathbf{e} \qquad (eq\ 2)$$

where $\mathbf{y}$ is the vector observations, $\mathbf{b}$ the vector of fixed effects, $\mathbf{a}$ the vector of random genetic effects

($\sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$), $\mathbf{m}$ the vector of random Mendelian sampling deviations and $\mathbf{e}$ the vector of random residuals ($\sim N(\mathbf{0}, \mathbf{R})$), $\mathbf{X}$ and $\mathbf{Z}$ design matrices relating observations to fixed and random genetic effects, respectively, $\mathbf{T}$ a lower triangular matrix satisfying $\mathbf{TDT'} = \mathbf{A}$ (Kennedy et al. (1988)) and $\mathbf{A}$ the numerator relationship matrix. If the mean and variance of Mendelian sampling deviations ($\mathbf{m}$) are specified to be $\mathbf{0}$ and $\mathbf{D}\sigma_a^2$ then both models (eqs 1 and 2) are statistically equivalent.

A consequence of pre-selection is that the mean and variance of Mendelian sampling deviations no longer comply with the assumptions of the standard linear animal model. Thus, Patry and Ducrocq (2009) proposed a model with the following assumptions: $E(\mathbf{m}) = \mathbf{Q\Delta}$ and $Var(\mathbf{m}) = \mathbf{\Gamma}$, where $\mathbf{Q}$ is a matrix connecting pre-selected animals to a group of animals going through the same pre-selection step, $\mathbf{\Delta}$ a vector of nonzero expectations of Mendelian sampling deviations and $\mathbf{\Gamma}$ a diagonal matrix with a priori variances of Mendelian sampling deviations. Note that $\mathbf{D} \neq \mathbf{\Gamma}$, since the variance of Mendelian sampling deviations of selected animals is expected to be reduced; a modified relationship matrix $\widetilde{\mathbf{A}}$ is then defined as: $\widetilde{\mathbf{A}} = \mathbf{T\Gamma T'}$.

Patry and Ducrocq (2009) presented two alternative forms for the mixed model equations for this model: one where the solutions for genetic groups for Mendelian sampling deviations ($\mathbf{\Delta}$) are estimated from the data:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{0} & \mathbf{X'R^{-1}Z} \\ \mathbf{0} & \sigma_a^{-2}\mathbf{Q'T'\widetilde{A}^{-1}TQ} & \sigma_a^{-2}\mathbf{Q'T'\widetilde{A}^{-1}} \\ \mathbf{Z'R^{-1}X} & \sigma_a^{-2}\mathbf{\widetilde{A}^{-1}TQ} & \mathbf{Z'R^{-1}Z} + \sigma_a^{-2}\mathbf{\widetilde{A}^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{\Delta} \\ \mathbf{a} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{0} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

$$(eq\ 3)$$

and one where those are assumed to be known, i.e. derived from realized selection differentials due the pre-selection $\widehat{\mathbf{\Delta}}$:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \sigma_a^{-2}\mathbf{\widetilde{A}^{-1}} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} + \sigma_a^{-2}\mathbf{\widetilde{A}^{-1}TQ\widehat{\Delta}} \end{bmatrix}$$

$$(eq\ 4)$$

Simulation. Three generations of animals were simulated. Generation 0 consisted of 500 males and 2500 females, and all animals were unrelated. Males and females in generation were mated at random, and each mating

resulted in two offspring. Twenty percent of the males in generation 1 were selected, and mated at random to the females to produce generation 2. Again, each mating resulted in two offspring, one of each gender.

Daughter yield deviations (DYD) were simulated for males in generations 1 and 2, assuming a progeny group size of 50. All females in generations 1 and 2 were assumed to have one record (yield deviation, YD). The simulated heritability and genetic variance were 0.5 and 1.0, respectively.

Pre-selection of the males in generation 1 was done on the basis of estimated Mendelian sampling deviations. The estimated Mendelian sampling deviations were calculated as the difference between the EBV and parent average, where EBVs were obtained from a BLUP genetic evaluation using phenotypes (DYD and YD) for generation 1 animals and simulated variance components.

Genetic evaluation. Four alternative genetic evaluations were carried out:

M1) A genetic evaluation, applying the model specified in equation 1, using phenotypes (DYD and YD) for all animals in generation 1 and 2. That is, data for unselected generation 1 males were included. As the assumptions for the standard linear model were met, this alternative was expected to yield unbiased breeding values.

M2) As alternative M1), but excluding the data for unselected generation 1 males. In this scenario, the condition that all data on which selection had been based was not met, and it was expected that breeding values would be biased.

M3) A genetic evaluation with genetic groups for Mendelian sampling deviations and excluding the data for unselected generation 1 males. In this scenario, genetic group effects were estimated from the data (equation 3).

M4) As alternative M3), but supplying known selection differential ($\hat{\mathbf{\Delta}}$; equation 4). The selection differential was computed using the mean estimated Mendelian sampling deviations for selected and unselected generation 1 males.

Simulated variance components were used in all genetic evaluations. The simulation was repeated 10 times.

RESULTS

As expected, using all data in the genetic evaluation resulted in unbiased breeding values for all categories of animals (Table 1). Discarding data for unselected generation 1 males and ignoring that in the genetic evaluation (M2) led to bias in breeding values. Including genetic groups for Mendelian sampling deviations, and estimating those genetic group effects from the data (M3) did not work satisfactory and resulted in overestimated breeding values in generation 2. When known selection differentials were supplied into the mixed model equations (M4) then breeding values were underestimated.

Table 1. Mean prediction error[1] for estimated breeding values by generation and gender for four alternative genetic evaluation models[2].

Category of animal	M1	M2	M3	M4
Gen 0, ♀	0.00	0.00	0.00	0.00
Gen 0, ♂	-0.01	-0.01	-0.01	-0.01
Gen 1, ♀	-0.01	-0.09	-0.01	-0.06
Gen 1, ♂	-0.01	0.07	0.29	0.01
Gen 2, ♀	-0.01	-0.32	0.19	-0.13
Gen 2, ♂	-0.01	-0.39	0.19	-0.17

[1] Mean of 10 replicates; SE of means are all < 0.02
[2] M1: standard animal model, all data
M2: standard animal model, no data for unselected generation 1 males
M3: model with genetic groups for Mendelian sampling deviations; genetic group effects estimated
M4: model with genetic groups for Mendelian sampling deviations; genetic group effects supplied

Pre-selection of males in generation 1 led to genetic superiority of slightly more than half a unit genetic standard deviation in generation 2 (Table 2). Even though pre-selection was on estimated Mendelian sampling deviations, there was also a selection differential in parent averages: the true parent average of selected males was approximately half a unit genetic standard deviation higher compared to unselected males.

Table 2. Mean of true breeding value (BV), parent average (PA) and Mendelian sampling deviation (MS) by generation and gender.

Category of animal	True BV	True PA	True MS
Gen 0, ♀	0.00	0.00	0.00
Gen 0, ♂	0.01	0.00	0.01
Gen 1, ♀	0.01	0.00	0.00
Gen 1, ♂, unselected	-0.28	-0.10	-0.17
Gen 1, ♂, selected	1.16	0.43	0.74
Gen 2, ♀	0.58	0.58	0.00
Gen 2, ♂	0.59	0.58	0.00

DISCUSSION

Contrary to expectation, including genetic groups for Mendelian sampling deviations did not attenuate bias due to pre-selection. The reason becomes apparent from Table 2, which shows that pre-selection on the basis of estimated Mendelian sampling deviations also led to a selection differential in parent averages. In other words, the parents of selected animals were better than average. The genetic group model only included genetic groups for Mendelian sampling deviations of selected animals, but erroneously assumed no difference in parent average of selected versus unselected animals.

Estimating or supplying selection differentials (M3 versus M4) affected bias in breeding values differently. In M3, Mendelian sampling deviations of the pre-selected animals were overestimated and their estimated parent averages were close to zero (not shown). Thus, the superiority of selected generation 1 males was attributed to the wrong source (Mendelian sampling deviations), and

propagation of the too optimistic estimated Mendelian sampling deviations to the next generation led to overestimated breeding values of generation 2 animals. On the other hand, the estimated Mendelian sampling deviations were close to their true values for M4 but the selection differential in parent averages was severely underestimated (not shown). As a consequence, the genetic level of both generation 1 and 2 animals was underestimated.

Increasing the number of progeny per sire (from 50 to 100) increased the selection differential for Mendelian sampling deviation and reduced it for parent averages (not shown). Thus, data were in better agreement with the assumptions in the genetic group model for this scenario, and the bias in estimated breeding values for M3 and M4 indeed reduced. The underlying cause for the change in selection differentials is that the prediction error covariance between Mendelian sampling deviation and parent averages decreases with increasing daughter group sizes.

Results presented in Table 1 were for the situation with $\mathbf{D} = \mathbf{\Gamma}$, i.e. assuming that selection did not alter the variance of Mendelian sampling deviations of selected and unselected animals. Using smaller values in $\mathbf{\Gamma}$ halved the bias in breeding values estimated with M4, but had no effect on the bias of breeding values estimated with M3.

Theory predicted that genomic prediction shifts the emphasis of estimated breeding values towards Mendelian sampling deviations (e.g. Daetwyler et al. 2007). However, experience from practical applications point toward genomic breeding values being considerably influenced by markers capturing genetic relationships (cf. Habier et al., 2007). Consequently, it is expected that genomic pre-selection leads to selection differential not only for Mendelian sampling deviations but for parent averages as well, similar to the pre-selection in the present simulation study. Therefore, the model with genetic groups for Mendelian sampling deviations proposed by Patry and Ducrocq (2009) cannot be expected to fully correct for the effects of genomic pre-selection. Instead, it remains important to include data for all genomically tested animals, both selected and discarded, to achieve unbiased genetic evaluations.

CONCLUSION

Accounting for pre-selection by inclusion of genetic groups for Mendelian sampling deviations in the genetic evaluation model reduced but did not remove bias in breeding values. Pre-selection typically leads to superiority of Mendelian sampling deviation and parent average of selected animals. The later is not accounted for in a model with genetic groups for Mendelian sampling deviations, explaining why bias was not completed removed.

LITERATURE CITED

Daetwyler, H.D., Villanueva, B., Bijma, P., et al. (2007). J. Anim. Breed. Genet. 124: 369-376.

Habier, D., Fernando, R.L., Dekkers, J.C.M. (2007). Genetics 177: 2389-2397.

Kennedy, B.W., Schaeffer, L.R., Sorensen, D.A. (1988). J. Dairy Sci. 72(suppl. 2), 17-26.

Mrode, R.A. (2005). Linear models for the prediction of animal breeding values. Wallingford, Oxon, UK; CAB International.

Patry, C., and Ducrocq, V. (2009). Interbull Bull. 39: 77-82.

Robertson, A. (1966). Anim. Prod. 8: 95-108.

Sorensen, D.A. and B.W. Kennedy. (1984). J. Anim. Sci. 58: 1213-1223.