

# THE GENSYS METHOD FOR GENETIC EVALUATION OF BEEF CATTLE

F. S. Schenkel and F. V. Brito  
 Federal University of Rio Grande do Sul, Porto Alegre, RS.  
 GenSys Associate Consultants S/C Ltda, Porto Alegre, RS.  
 Brasil

## ABSTRACT

A beef cattle genetic evaluation method designed for large data sets and to be executed on a microcomputer environment is described. The GenSys Method, thus named, rose from the necessity of estimating genetic values with desirable properties from real records of herds under selection.

## INTRODUCTION

To establish genetic differences it is important that: 1) there are sufficient records in quantity and quality; 2) these records have a good structure and 3) an adequate methodology of analysis is used.

The objective of this paper is to describe basic and operational concepts of a beef cattle genetic evaluation methodology that is been used by GenSys on some 100 thousand weaning and 50 thousand posweaning records (yearly accumulated) and it is executed on a microcomputer environment.

## THEORETICAL BACKGROUND

To be effective and useful to breeders, a beef cattle genetic evaluation method must have three fundamental characteristics: 1) consider that selection is happening within and across herds; 2) be robust to extreme records and outliers and 3) check and consider data connectedness.

The Mixed Model Methodology (MMM) (Henderson, 1963), proven to provide estimators and predictors with desirable properties given that the assumption of no selection (needed to prove unbiasedness and minimum variance on such setting) is true, is worldwide used in beef cattle evaluations. Henderson (1975 and 1984) proposed different matrix amendments to allow for selection and keep the methodology's properties. Another approach with the same aim, which has shown to be workable under all situations so far encountered, is used as one of the components of the GenSys Method, as proposed by Fries (1990) and, with corrections, by Fries and Schenkel (1993).

Given the model  $y = X\beta + Z\mu + e$

where:

$y$  = vector of observations;

$\beta$  = vector of fixed unknown effects (Contemporary Groups (CG));

$X$  e  $Z$  = incidence matrices;

$\mu$  = vector of random unobserved effects (sire and dams);

$e$  = vector of random unobserved residuals;

and:

$$E \begin{bmatrix} e \\ \mu \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ X\beta \end{bmatrix}, \quad \text{Var} \begin{bmatrix} e \\ \mu \\ y \end{bmatrix} = \begin{bmatrix} R & 0 & R \\ 0 & G & GZ \\ R & ZG & V \end{bmatrix},$$

$$V = \text{Var}(y) = \text{Var}(Z\mu + e) = ZGZ' + R, \quad G = A\sigma_{\mu}^2 \quad \text{and} \quad R = E\sigma_e^2,$$

Fries and Schenkel (1993) showed that the Mixed Model Equations (MME):

$$\begin{aligned} X'R^{-1}X\beta^B + X'R^{-1}Z\mu^B &= X'R^{-1}y \\ Z'R^{-1}X\beta^B + (Z'R^{-1}Z+G^{-1})\mu^B &= Z'R^{-1}y, \end{aligned}$$

provide a solution for the genetics effects  $\mu$ , like a two Steps Least Squares (2SLS) procedure:

$$\mu^B = (Z'MZ + G^{-1})^{-1} Z'MZ \mu^G, \text{ after removing}$$

$$\beta^B = \beta + (X'R^{-1}X)^{-1} X'R^{-1}Z(Z'MZ + G^{-1})^{-1} \mu^B \text{ and that the MME can or should be rewritten as:}$$

$$X'V^{-1}X \beta^B = X'V^{-1}y \text{ and } (Z'MZ + G^{-1})^{-1} \mu^B = Z'My, \text{ where:}$$

$$V^{-1} = R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1} \text{ and } M = R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}.$$

The parametric function  $\beta^B = \beta + (X'R^{-1}X)^{-1} X'R^{-1}Z(\mu - \mu^B)$ , obtained from the previous equations, permits the conclusion that: 1)  $\beta^B$  do not contain a function of  $\mu$  only when  $X'R^{-1}Z\mu = 0$  by experimental design; 2) if  $\mu$  is assumed to be a fixed effect or  $G=0$ , then  $\mu^B$  would not be regressed and  $\mu^B = \mu$  and  $\beta^B = \beta$ ; and 3) if none of the above two situations are satisfied, then it is obligatory to assume that  $E(\mu) = 0$  so that it is possible to demonstrate, algebraically, that  $\beta^B$  is an unbiased estimator of  $\beta$ . But, only if no selective process is or has happened such assumption is reasonable and can be made.

Can be noticed that since  $\beta^B$  is as a function of  $\mu$ , the solution  $\mu^B$  will by its turn be affected (at minimum there is an information loss) by  $\beta^B$  because  $\mu^B = (Z'R^{-1}Z + G^{-1})^{-1} Z'R^{-1}(y - X\beta^B)$ .

Henderson (1975) proved that  $\mu^B$  is unbiased if the selection in  $y$  can be represented by the matrix  $L$  such that  $L'X=0$ . Then, the selection should be realized within the levels of  $X$  for to keep  $\mu^B$  reliable. Thus,  $\mu^B$  would not have its main proclaimed advantage over CG ratios and deviations, which is of allowing for comparisons of estimated genetic values across fixed effects levels (CGs).

The proposed modification means, in terms of equations to solve, applying the same postmultiplication to the block  $X'R^{-1}Z$  by  $(Z'R^{-1}Z)^{-1}(Z'R^{-1}Z + G^{-1})$  as applied to  $Z'R^{-1}Z$ , as it is done to transform the normal equations (NE) into the MME. This operational symmetry was used by Fries and Schenkel (1993) to reparametrize  $\mu$  into Lush's selection criteria ( $\mu^*$ ). The modification allows, even under selection, to obtain unbiased estimators of the fixed effects and nearly unbiased (Schaeffer and Mao, 1987) predictors of the random effects. The new set of equations would then be:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z(Z'R^{-1}Z)^{-1}(Z'R^{-1}Z + G^{-1}) \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta^G \\ \mu^* \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

The multiplication of the original NE by the full rank  $(Z'R^{-1}Z)^{-1}(Z'R^{-1}Z + G^{-1})$  causes the new set of equations to retain the original dependencies. Thus, it necessary to impose suitable restrictions in order to obtain meaningful solutions  $\beta^G$  and  $\mu^*$ . This modification is equivalent to solving the NE considering all effects fixed and then regress  $\mu$  by a function of progeny size and variances ratios.

Using the 13<sup>th</sup> theorem of Zyskind et al. (1964), Fries and Schenkel (1993) showed, in appendix A, a proof that if the var-covariance matrix of the observations is given by  $V = ZGZ' + R$ , for some  $G$  and  $R$  non-negative definite, then the Generalized Least Squares and the Weighted Least Squares Methods provide identical solutions for any  $G$ , known or estimated from the records. This result also showed the extreme importance of a correct definition of the matrix  $R$  in the estimation process.

## DESCRIPTION OF THE GENSYS METHOD

The expected progeny differences of sires are obtained using the following additive genetic model:

$$y_{ijkl} = \mu + G_i + D_j + S_k + E_{ijkl} \quad (1)$$

where, besides a simbol for the overall mean:

$y_{ijkl}$  = Observation on animal  $l$ , born from the dam  $j$  and sire  $k$ , belonging to contemporary group  $i$ , that received the  $l^{\text{th}}$  mendelian sampling;

$G_i$  = Environmental effect of the Contemporary Group (CG)  $i$ ;

$D_j$  = Real Production Ability (RPA) of the dam  $j$ ;

$S_k$  = Expected Progeny Diference (EPD) of the sire  $k$ ;

$E_{ijkl}$  = Residual effect associated with the observation on animal  $l$ , belonging to CG  $i$ , born from the dam  $j$  and sire  $k$ , that received and is confounded with the  $l^{\text{th}}$  mendelian sampling;

and with the assumptions:

$$\begin{aligned}
E(y_{ijk}) &= u + G_i + D_j + S_k \quad v \quad i,j,k,l \\
V(D_j) &= \sigma_d^2 \quad v \quad j \quad \text{and} \quad V(S_k) = \sigma_s^2 \quad v \quad k \\
V(E_{ijk}) &= R, \text{ a diagonal matrix} \\
COV(D_j, S_k) &= 0 \quad v \quad j,k \\
COV(D_j, E_{ijk}) &= 0 \quad \text{and} \quad COV(S_k, E_{ijk}) = 0 \quad v \quad i,j,k,l
\end{aligned}$$

The numerator relationship matrix (A) is not used since the var-covariance matrix of genetic effects is different from  $A^{-1}\sigma_a^2$  under selection (Henderson, 1984).

There are strong indications showing that a correct definition of CG can be the most important single factor in determining genetic results. Brito (1992) found this to have a larger effect than the methodology used: CG deviations vs MMM. In the GenSys method, some other (besides the ones recommended by BIF) possible CG's components are tested and may be used, e.g.: weighing date, individual pasture number, type of reproduction used to obtain next calf and code for heifers managed to calf at 2 years vs later. A compromise has to be made between average size of CGs and practical significance of the effects, even in the case of MM methodologies.

At weaning, the characteristics generally evaluated are weight gain from birth to weaning (WG) and visual scores for conformation, early maturity, muscling and frame. The WG records were used by some environmental modelling studies on different breeds and regions using segmented polynomials. Multiplicative correction factors which use the average WG from each CG are used to adjust for: weaning age (lin+quad effects), cow age (quad-quad within sex of calf) and birth date within season (cub+quad+cub effects). For crossbred animals, the WG are pre-adjusted for individual and maternal heterosis, after partitioning the heterozygosity into ZE (Zebu-European), EE and ZZ components.

Solutions are obtained by an iterative process over the data set, using a Gauss-Seidel-like algorithm. All effects, initially, are estimated as if they are fixed. Using the modification, problems are exhibited that, with the standard MME, remained hidden. When a real data set is not fully connected a meaningful LS solution is not possible. Within MMM framework, each disconnected subgroup of effects has the property of adding to zero and, therefore, the assumption that there are no genetic differences amongst these subsets (of environmental or genetic effects) is made. Small disconnected groups of sires, that in reality may be very good or very bad, will sum to zero, have their group solution added, and will be reported unnoticed.

On the Gensys Method, besides a check on connectedness (an all-or-none condition for each record), an heuristic measure of connectedness for each CG and an information level concept are used. First, a central mass (well connected and with large progenies) group of sires and dams is chosen. The procedure also has to take into account if the sire is unknown or is a Multiple Sire. Then, the number of genetic links are counted for each CG. One genetic link is furnished by each animal in the CG that is the product of a sire and/or dam from the "central mass" group and for each half-sib this animal posses on other connected CGs. A CG is considered to be well enough connected if: 1) there are at least three sires and/or dams connected to the main group and 2) the number of genetic links is at least 10. If a large proportion of records are found to be disconnected, the procedure is restarted within this set. All remaining records are run under selection index procedures and separately reported. Records which are from a connected CG and have, at least, one other of the effects connected are used for the estimation process labeled as level one of information. Indirect ties give rise to increasing information levels (il).

The diagonal matrix R is used to hold the amount of confidence behind each record, using as scalars a robust ponderation factor (ROB) defined by:

$$ROB = \sigma_m / (IAD * \sigma_{gc})$$

where:  $\sigma_m$  = the average standard deviation from all the CGs;  
 $\sigma_{gc}$  = standard deviation from the animal's CG (but, if  $\sigma_{gc} < \sigma_m$ , then  $\sigma_{gc} = \sigma_m$ );  
IAD =  $il + \text{abs}(\text{int}((\text{dev}/\sigma_{gc}))$  ,  $il$  + absolute integer value of ( ); and  
dev = the observation's deviation from its expected value.

The use of ROB as the elements from the diagonal of R is an approach to jointly take care of: (1) heterogeneity of CG's variances, (2) extreme observations and outliers and (3) level of connectedness through the number of direct or indirect ties. The smaller value that IAD can attain is 1, what happens when the information level is 1 and the deviation is smaller than the CG's standard deviation. If the CG's standard deviation is larger than the average standard deviation, the ROB will be smaller than 1. ROB can assume, at maximum, the value of 1 and it is not used to estimate the mendelian segregation effect, hence does not affect directly the calculations of a young animal's EPD.

Begging with the Fixed Solutions (FS) for dam and sire effects, the predictors of these effects or the desired selection criteria (SC) are obtained by:

$$SC_j = FS_j * [ n_j / (n_j + k) ], \text{ being:}$$

$SC_j$  = Selection criterion, EPD of sire j or RPA of the dam j;

$FS_j$  = Fixed solution for the genetic effect of sire/dam j;

$n_j$  = Total robust progeny size from sire/dam j;

$k$  = Ratio of residual variance to sire or dam variance in model (1).

The dams' EPDs are obtained by a re-scaling of the dams' RPA.

A young animal's EPD is obtained by a function of sire and dam EPD's and an estimated mendelian segregation effect. After obtaining the animal's EPDs for each trait they are combined into a (Weaning or Final) Index in the following way: (1) standardized EPDs are obtained by the division of each EPD by its standard deviation (calculated for each herd or set of herds) and then (2) these are multiplied by weighing factors which represent the amount of selection pressure desired to exert on each trait, in herd or client. Every year, genetic and environmental results are evaluated by all personnel involved and the index may be fine tuned.

#### ACKNOWLEDGEMENT

The authors wish to express their gratitude to all GenSys' associates. All discussions, concepts, procedures' development, field testings and continuing reshaping of the Method are done by all members.

#### REFERENCES

- BRITO, F. V. (1992) *Comparações entre Metodologias para Avaliação Genética de Bovinos de Corte*. Porto Alegre, Universidade Federal do RGS, Facul. de Agronomia. 177 p. Dissertação de Mestrado em Zootecnia.
- FRIES, L. A. (1990) *Mixed Model Methodology Revisited*. (Submitted to Revista Brasileira de Genética on Feb. 05, 1990 and rejected)
- FRIES L.A. & SCHENKEL, F.S. (1993) Estimation and Prediction Under a Selection Model. *Anais da 30ª Reunião da Sociedade Brasileira de Zootecnia*. Rio de Janeiro, RJ. (1-22).
- HENDERSON, C.R. (1963) Selection Index and Expected Genetic Advance. In: HANSON, W.D. & ROBINSON, H.F. *Statistical Genetics and Plant Breeding*. Washington, National Academy of Science. National Research Council. p. 141-163 (Publication 982).
- HENDERSON, C.R. (1975) Best Linear Unbiased Estimation and Prediction Under a Selection Model. *Biometrics*, Washington, 31: 423-447.
- HENDERSON, C.R. (1984) *Applications of Linear Models in Animal Breeding*. Guelph, University of Guelph. 462 p.
- SCHAEFFER, L.R. & MAO, I.L. (1987) *Notes on Linear Model Theory and Variance Component Estimation*. In: CURSO DE GENÉTICA ANIMAL, 1988. Zaragoza, Centro Internacional de Altos Estudios Agronomicos Mediterráneos - Instituto Agron Mediterraneo de Zaragoza. 237 p.
- ZYSKIND, G.; KEMPTHORNE, O.; WHITE, R.F.; DAYHOFF, E.E.; DOERFLER, T.E. (1964) *Research on Analysis of Variance and Related Topics*. Ohio, Office of Aerospace Research, USAF. p. 134-138.