MedCrave
Step into the World of Research
International Journal of Avian & Wildlife Biology

Research Article

Open Access
CrossMark

Theory of estimation of parameters and genetic values under mixed models

Abstract

In animal breeding, it is essential to know genetic parameters such as heritability, with the aim of being able to predict genetic values (GV) and efficiently direct selection programs. A mixed model refers to those cases where the researcher considers fixed and random factors in a statistical model. Models widely used in the area of animal genetic improvement are the reproductive model and the animal model, which consider the reproductive or animal factor as random and a group of non-genetic effects as fixed. These mixed models allow us to obtain both heritability values (h^2) for a trait, as well as genetic predictions such as the expected progeny difference (EPDs) or the predicted transmission ability (PTA) for each animal. An example of birth weight (BW) in cattle was used to calculate the VG, h2 and e2 using a mixed model, with a fixed and a random factor. The ANOVA, ML and REML methods were used to calculate h^2, e^2 and the VG first using all the information and subsequently assuming the last lost data, under a reproductive model and an animal model. The results found using the 3 methods were the same for REML and ANOVA in balanced data and different for the 3 methods in unbalanced data, where in the unbalanced case the ANOVA estimated a negative variance component, therefore, it can be concluded that estimate genetic values and parameters using ANOVA, ML and REML, but with the risk of estimating negative variance components using ANOVA or null (or overestimated) heritabilities with likelihood-based methods when the data structure or model is not the same correct.

Keywords: heritability, ANOVA, REML, ML, mixed models

Volume 8 Issue 1 - 2024

Pérez González José Raúl, Morales Valladares, David Daniel
Maracaibo Territorial Polytechnic University of Maracaibo, Venezuela

Correspondence: Pérez González José Raúl, Maracaibo Territorial Polytechnic University of Maracaibo,Venezuela, Email josejrp1995@gmail.com

Received: February 20, 2024 | **Published:** March 20, 2024

Introduction

In animal breeding, it is essential to know genetic parameters such as heritability in order to be able to predict genetic values (GV) and efficiently conduct selection programs. Genetic parameters are ratios between estimated population variances, known as variance components, which are calculated using linear models containing fixed and random factors, generally known as mixed models.[1] For the correct estimation of parameters and genetic values, it is necessary to have a broad knowledge of estimation using mixed models. Therefore, this article reviews the estimation of variance components and genetic values using ANOVA, ML and REML under a reproductive model and an animal model, explaining the virtues and limitations of each method in balanced and unbalanced data.

Theoretical framework mixed models

A mixed model refers to those cases where the researcher considers both fixed and random factors in a statistical model.[2] A model widely used in the area of animal breeding is the reproductive model or Sire Model, which considers the reproductive factor as random and a group of non-genetic effects as fixed.[2] The reproductive model allows obtaining both heritability values (h^2) for a trait, as well as genetic predictions such as the expected difference of progeny (DEPs) or the predicted transmission ability (PTA) for each breeder.[3] In matrix algebra the reproductive model takes the following form:

$$y = Xb + Zs + e$$

Where y is a vector for the data, X is an incidence matrix relating the data to the fixed effects, b is a vector of unknown parameters for the fixed effects, Z is an incidence matrix relating the data to the random effects, s is a vector of unknown predictions for each player, and e is a vector of residuals.

The covariance structure of the above model is:

$$VAR\begin{bmatrix} s \\ e \end{bmatrix} = \begin{bmatrix} I\sigma^2{}_s & 0 \\ 0 & I\sigma^2{}_s \end{bmatrix}$$

Where I is an identity matrix, σ_s^2 is the variance between breeders and σ_e^2 is the residual variance.

The Henderson normal equations, necessary to find the genetic values of the breeders, for the above model are given by[3]:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\alpha \end{bmatrix}\begin{bmatrix} b_i \\ s_i \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Where α is a ratio of the residual variance to the variance between breeders:

$$\alpha = \frac{\sigma_e^2}{\sigma_s^2}$$

According to Román and Aranguren,[4] it is possible to substitute $I\alpha$ by $A^{-1}\alpha$ in the normal Henderson equations, with the objective of improving predictions using all the parentage information between males, therefore, the new equations are:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix}\begin{bmatrix} b_i \\ s_i \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Where A^{-1} is the inverse of the kinship matrix.

And the covariance structure taking into account the introduction of A is[5]:

$$VAR\begin{bmatrix} s \\ e \end{bmatrix} - \begin{bmatrix} A\sigma_s^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix}$$

Another model widely used in genetic evaluation is the animal model, which uses all the parentage information in the pedigree, and unlike the reproductive model, allows obtaining genetic predictions of all the animals in the herd, whether or not data is present or not:

$$y = Xb + Za + e$$

Where a is a vector of genetic predictions for each animal the covariance structure of the above model is as follows:

$$VAR\begin{bmatrix} a \\ e \end{bmatrix} = \begin{bmatrix} A\sigma_a^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

Where *G* is a variance and covariance matrix for the random effects and *R* is a matrix of residuals.

The Henderson normal equations for this model are given by:[6]

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix}\begin{bmatrix} b_i \\ a_i \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Where α in this model is a ratio of the residual variance to the additive variance:

$$\alpha = \frac{\sigma_e^2}{\sigma_a^2}$$

Where σa^2 is additive genetic variance

Genetic parameters

Using mixed models, it is possible to estimate the variance components, and from them calculate the hereability, which is given by:[5]

$$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

Where h^2 is heritability, and σp^2 is phenotypic variance, therefore, heritability is defined as a quotient between the additive variance and the phenotypic variance. The additive component (additive variance) of the numerator of the formula of h^2 can be estimated using several procedures, a well-known one is to use a reproductive model to estimate the variance between breeders, which is ¼ of the additive variance, therefore, a formula to estimate σa^2 is:

$$\sigma_a^2 = 4\sigma_s^2$$

Where $4\sigma s^2$ is four times the variance among breeders, therefore, heritability can be calculated as:[7]

$$h^2 = \frac{4_s^2}{\sigma_p^2}$$

If the heritability is known, the heritability component can be σa^2 component can be calculated using the following formula:[8]

$$\sigma_a^2 = h^2\sigma_p^2$$

Another parameter of interest is the environmental proportion coefficient, which indicates how much of the differences observed in the phenotype (data) of the animals are due to non-genetic (environmental) factors, this coefficient has the following mathematical formula:[9]

$$e^2 = \frac{\sigma_{en}^2}{\sigma_p^2}$$

Where σen^2 is the environmental variance. The variance component σen^2 is calculated using the difference between σp^2 σa^2 therefore, the formula for σen^2 is:[8]

$$\sigma_{en}^2 = \sigma_p^2 - \sigma_a^2$$

Finally, the variance σp^2 is the sum of the variance components:

$$\sigma_p^2 = \sigma_a^2 + \sigma_{en}^2 = \sigma_s^2 + \sigma_e^2$$

Variance component estimation using a reproductive model analysis of variance

There are several classical methods for estimating the variance components needed to compute h^2 y e^2including analysis of variance (ANOVA), maximum likelihood (ML) and restricted maximum likelihood (REML).

ANOVA is a technique that attempts to separate out different sources of variability. σp^2 into different sources of variability, this involves the separation of sums of squares (SC), degrees of freedom (GL) and mean squares (MS) for each source of variation. Variance components estimated using ANOVA are calculated by equating the expected values of the CM (E (CM)) for each source of variation, with their respective CM and solving the resulting system of equations.[10] CMs are a ratio of SC to GLs for each source of variation:[10]

$$CM = \frac{SC}{GL}$$

In the case of a fixed factor and a random factor, without interaction, the reproductive model, in elementary algebra, is given by:

$$y_{ijk} = \mu + s_i + b_j + e_{ijk}$$

And the ANOVA square for the above model is presented in Table 1.

Table 1 ANOVA for Henderson's method III

FV	SC	GL	CM	E (CM)
Factor Fijo	SCb	n_{fijo} -1	$\frac{SCb}{n_{fijo}-1}$	
Padres	SCs	n_s -1	$\frac{SCs}{n_s-1}$	$E(CM_s) = \sigma_e^2 + n_{fijo}k\sigma_s^2$
residual	$SCtotal - \sum SCresto$	$GLtotal - \sum GLresto$	$\frac{SCe}{GLtotal - GLresto}$	$E(CM_e) = \sigma_e^2$
Total	$y'y - R(\mu)$	n-1		

Citation: Raúl PGJ, Valladares M, Daniel D. Theory of estimation of parameters and genetic values under mixed models. *Int J Avian & Wildlife Biol.* 2024;8(1):27–33. DOI: 10.15406/ijawb.2024.08.00210

Where k is the number of replicates of the design, $nfijo$ is the number of levels of the fixed effect and ns is the number of levels of the random factor. The variance components are calculated by equating the CM to their E (CM):

$$CMs = \sigma_e^2 + n_{fijo} k \sigma_s^2$$

$$CMe = \sigma_e^2$$

And the unique solution of this system of equations is:

$$\sigma_s^2 = \frac{CMs - CMe}{n_{fijo} k}$$

$$\sigma_e^2 = CM_e$$

In balanced data, the CS can be estimated directly without the need for adjustment, for a model with two non-interacting factors:

$$SCs = \frac{\sum y_{s.}^2}{k} - \frac{(\sum y)^2}{n}$$

$$SCb = \frac{\sum y_{b.}^2}{w} - \frac{(\sum y)^2}{n}$$

Where $\sum ys^2$ is the sum of the sum of the sum of the data for each player squared, $\sum yb^2$ is the sum of the sum of the sum of the data for each level of the fixed effect and w is the number of replicates for the fixed effect. For the unbalanced case, the SCs have to be calculated using the type III SCs for the random factor (sire), since type III calculates the SCs of an effect by correcting them with respect to any other effect that does not contain it and orthogonal to any effect (if it exists) that contains it. Type III CS can be expressed as:[11]

$$SCs = SC(\mu, s, b) - SC(\mu, b)$$

The SCs is corrected for the effects of μ y $b$where μ is the intercept or herd mean effect. In order to find the values of SCs it is necessary to fit a complete model and calculate (μ, s, b,) and subtract $SC(\mu, b)$ a reduced model.

Maximum likelihood

The maximum likelihood (ML) method is a classical method of parameter estimation proposed by Fisher,[12] but it was not until Hartley and Rao,[13] that it was used for mixed models in general. Knowing the likelihood function as a function of the parameters of a statistical model given some data, in ML we try to obtain estimators of the variance components that maximize the likelihood function, that is, that have the maximum probability of representing the population parameters.

The likelihood function is defined as the product of the likelihood function of the data, but in practice, the natural logarithm of the likelihood function is used because it is more manageable, if the distribution of the data is normal, in matrix algebra the natural logarithm of the likelihood function is defined as:[11]

$$Ln(L) = -0.5(n).In(2\pi) - 0.5In|V| - 0.5(y - Xb)'V^{-1}(y - Xb)$$

Where (L) is the natural logarithm of the likelihood function and $V = ZGZ' + R$ is the variance and phenotypic covariance matrix of the model. To find the estimators that maximize the likelihood, we need to find the maximum of equation (L)This is achieved with different methodologies, for example, if the data structure is balanced and we have a mixed model, with a random effect and a fixed one with no interaction, the derivative of $Ln(L)$with respect to the parameters to be estimated σ^2_s y σ^2_e will lead us to a system of equations whose solution is:

$$\sigma_s^2 = \frac{\frac{SCs}{n_s} - \sigma_e^2}{n_{fijo} k}$$

$$\sigma_e^2 = \left[1 - \frac{n_{fijo} - 1}{n_s (n_{fijo} k - 1)}\right] CMe$$

An important point of ML estimation, for this model, is that even with balanced data, it is possible to find estimators different from the ones presented above, since these solutions will be valid if the inequality $CMs > CMe$ is met, but on the other hand, if the inequality is $CMs < CMe$ ML estimates for this model and balanced data are given by:[11]

$$\sigma_e^2 = \frac{SCtotal}{(n_s)(n_{fijo})(k)}^{\sigma_s^2=0}$$

That is all phenotypic variability is residual, which may indicate that the model used is incorrect or that the number of data is insufficient, thus increasing the variability of the error. The variance σ^2_p is the sum of the variance components σ^2_e y σ^2_s whose sum gives an estimate of σ^2_p given mathematically by:

$$\sigma_p^2 = \sigma_s^2 + \sigma_e^2 = \frac{(y - Xb)'(y - Xb)}{n}$$

Which is biased, since it is associated with n degrees of freedom. If the structure of the daros is unbalanced, the partial derivatives of (L) lead to nonlinear maximum likelihood equations for the parameters to be estimated, therefore, the system of equations cannot be solved with direct methods. Faced with this problem, iterative number methods are used to try to approximate the maximum of (L) which are applied to the logarithmic likelihood itself and not to the equations resulting from its first derivative, in order to be able to simultaneously calculate the variance components and (L)which we can use to find fit criteria for our model, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).

Restricted maximum likelihood

The restricted maximum likelihood method (REML) is a method proposed by Paterson and Thompson,[13] which takes into account the loss of degrees of freedom by including fixed effects in the statistical model, therefore, the estimation of variability components are unbiased, since they are associated to degrees of freedom, which leads to an estimation of variance of the model. $n - (X)$ degrees of freedom, which leads to an estimate of the variance, which is defined as σ^2_pwhich is defined as:

$$\sigma_p^2 = \frac{(y - Xb)'(y - Xb)}{n - Rango(X)}$$

Where (X) is the rank of the incidence matrix for the fixed effects of the model. For the case where the only fixed effect is $\mu$the variance σ^2_p is associated with $n - 1$ degrees of freedom.

As in ML, in REML, the objective is to maximize the logarithm of a function of the parameters, but in this case restricted, which is known as restricted likelihood function, which in matrix algebra is defined as:[14]

$$Ln(Lr) = -0.5(n - p).In(2\pi) - 0.5In|V| - 0.5|X'V^{-1}X| - 0.5(y - Xb)'V^{-1}(y - Xb)$$

Where (Lr) is the logarithm of the restricted likelihood function. If we have a balanced data structure and by deriving (Lr) as a function of the variability components of the model (model above), we can solve a system of equations that give rise to estimates given by:[11]

$$\sigma_s^2 = \frac{CMs - CMe}{n_{fijo}k}$$

$$\sigma_e^2 = CMe$$

These are identical to estimates using an ANOVA, since a property of REML is that in a balanced data structure, REML estimates = ANOVA as long as the inequality is satisfied. $CMs > CMe$ Otherwise, estimates via ANOVA would be negative and in REML all phenotypic variability is residual. In unbalanced data structure, the derivative of (Lr) with respect to the variance components, gives rise to nonlinear equations, which cannot be solved directly, therefore, in these cases, as in ML, iterative numerical methods are used to approximate the value of the variance components.

REML estimates using kinship information in an animal model

In the case of a simple animal model, where each animal has only one data (and there are animals without data), the ANOVA method cannot be applied, since it is not possible to estimate the variation within groups using this methodology, because the classification variable is each animal that has a unique record, but the ML and REML estimations are applicable since they allow introducing kinship information in the matrix. A. In a mixed model, maximizing (Lr) is equivalent to minimize -2 (Lr) Therefore, the objective function to be minimized, in matrix algebra, can be defined as:[14]

$$-2Ln(Lr) = (n-p).In(2\pi) + Ln|R| + Ln|G| + Ln|C| + y'Py$$

Where $Ln|C|$ is the natural logarithm of the determinant of the coefficient matrix of the normal Henderson equations and $'Py$ is the generalized residual sum of squares. Obviously to minimize $-2Ln(Lr)$ iterative numerical methods are needed, but it has the advantage that it is easier than maximizing $Ln(Lr)$ Therefore, most specialized REML programs use sparse matrix algorithms and numerical methods to try to find estimators resulting from the minimization of $-2Ln(L_r)$.

Materials and methods

An example of birth weight (BW) in cattle was used to calculate the VG, h^2 y e^2 using a mixed model, with a fixed and a random factor. The database is presented in Table 2. In this problem, we want to eliminate the variability that exists between the sexes, therefore, the sex factor is considered as fixed and the father factor as random, which leads us to the statistical model for this problem:

$$PN = media + padre + sexo + error$$

Table 2 Database of animal records, sex and NP

Father	Animal	Sex	y
1	3	Male	36
1	4	Male	35
1	5	Female	33
1	6	Female	28
2	7	Female	31
2	8	Female	29
2	9	Male	28
2	10	Male	36
3	11	Male	38
3	12	Male	37
3	13	Female	29
3	14	female	35

ANOVA, ML, and REML methods were used to calculate h^2 , e^2 and GVs using the data in Table 2, first using all the information and then assuming the last missing data. For the animal model, a similar model was used:

$$PN = media + animal + sexo + error$$

Where all the kinship information and the value of the variance components found in the previous model were used to solve the Henderson normal equations.

Results and discussion

Balanced data in a reproductive model

To calculate the CM, it is necessary to calculate the SC and GL for each source of variation, for this model and our data structure, we can calculate them using the formulas in Table 1:

$$GL_s = 3-1 = 2$$

$$GL_{sexo} = 2-1 = 1$$

$$GL_{total} = 12-1 = 11$$

$$GL_e = 11-1-2 = 11-3 = 8$$

And since the design is balanced, the SCs are:

$$SC_{total} = 36^2 + 35^2 + 33^2 + \ldots + 35^2 - \frac{395^2}{12} = 152.916$$

$$SC_s = \frac{(132)^2 + (124)^2 + (139)^2}{4} - \frac{395^2}{12} = 28.166$$

$$SC_{sexo} = \frac{(185)^2 + (210)^2}{6} - \frac{395^2}{12} = 52.083$$

$$SC_e = 152.916 - (28.166 + 52.083) = 72.667$$

And the CMs come from:

$$CM_S = \frac{28.166}{2} = 14.083$$

$$CM_e = \frac{72.667}{8} = 9.083$$

And from the CM we can calculate the variance components:

$$\sigma_s^2 = \frac{14.083 - 9.083}{2(2)} = 1.25$$

Therefore, h^2 using ANOVA is:

$$h^2 = \frac{4(1.25)}{1.25 + 9.083} = 0.483$$

$$e^2 = \frac{10.33 - 5}{10.33} = 0.515$$

And these ANOVA estimates, too, are REML, since the data structure is balanced and the $CMs > CMe$. Now the calculation of the GVs, using the REML estimates, comes from the solutions of the normal Henderson equations, using the estimated value of the variance components, and calculating the value of α we have that:

$$\alpha = \frac{9.083}{1.25} = 7.2664$$

Therefore, the equations are:

$$\begin{pmatrix} & & & & & \\ 6 & 0 & 2 & 2 & 2 & \\ 0 & 6 & 2 & 2 & 2 & \\ 2 & 2 & 4+7.266 & 0 & 0 & \\ 2 & 2 & 0 & 4+7.266 & 0 & \\ 2 & 2 & 0 & 0 & 4+7.266 & \end{pmatrix}\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix} \rightarrow \begin{bmatrix} & & & & & \\ 6 & 0 & 2 & 2 & 2 & \\ 0 & 6 & 2 & 2 & 2 & \\ 2 & 2 & 11.266 & 0 & 0 & \\ 2 & 2 & 0 & 11.266 & 0 & \\ 2 & 2 & 0 & 0 & 11.266 & \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix}$$

In the previous equations the value of μ was forced to be zero in order to break the linear dependence between the rows and columns of the coefficient matrix. The solution of this system of equations is given by:

$$\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \left[\begin{bmatrix} & & & & & \\ 6 & 0 & 2 & 2 & 2 & \\ 0 & 6 & 2 & 2 & 2 & \\ 2 & 2 & 11.266 & 0 & 0 & \\ 2 & 2 & 0 & 11.266 & 0 & \\ 2 & 2 & 0 & 0 & 11.266 & \end{bmatrix}\right]^{-1}\begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix} = \begin{bmatrix} 30.83 \\ 35 \\ 0.029 \\ -0.680 \\ 0.650 \end{bmatrix}$$

ML estimates are:

$$\sigma_s^2 = \frac{\frac{28.166}{3} - 8.073}{2(2)} = 0.328$$

$$\sigma_e^2 = \left[1 - \frac{2-1}{3(2(2)-1)}\right]9.083 = 8.073$$

Y h^2y e^2using ML is:

$$h^2 = \frac{4(0.328)}{0.328 + 8.073} = 0.156$$

$$e^2 = \frac{8.401 - 1.3212}{8.401} = 0.842$$

And the equations using ML are:

$$\begin{bmatrix} & & & & & \\ 6 & 0 & 2 & 2 & 2 & \\ 0 & 6 & 2 & 2 & 2 & \\ 2 & 2 & 28.612 & 0 & 0 & \\ 2 & 2 & 0 & 28.612 & 0 & \\ 2 & 2 & 0 & 0 & 28.612 & \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix}$$

Therefore the solution is:

$$\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} & & & & & \\ 6 & 0 & 2 & 2 & 2 & \\ 0 & 6 & 2 & 2 & 2 & \\ 2 & 2 & 28.612 & 0 & 0 & \\ 2 & 2 & 0 & 28.612 & 0 & \\ 2 & 2 & 0 & 0 & 28.612 & \end{bmatrix}^{-1}\begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix} = \begin{bmatrix} 30.83 \\ 35 \\ 0.011 \\ -0.267 \\ 0.256 \end{bmatrix}$$

Introduction of the parentage matrix in a reproductive model

Now it is assumed that animal 1 is the father of animal 2, therefore, the equations take into account all the genealogy between males. First we have to calculate A^{-1}. Applying Henderson's rules,[5] we have:

$$A^{-1} = \begin{bmatrix} 1+1/3 & -2/3 & 0 \\ -2/3 & 1/4 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/4 & -2/3 & 0 \\ -2/3 & 1/4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Therefore,

$$A^{-1}\alpha = \begin{bmatrix} \frac{1}{4} & -\frac{2}{3} & 0 \\ -\frac{2}{3} & \frac{1}{4} & 0 \\ 0 & 0 & 1 \end{bmatrix}(7.2664) = \begin{bmatrix} 1.8166 & -4.8442 & 0 \\ -4.8442 & 1.8166 & 0 \\ 0 & 0 & 7.2664 \end{bmatrix}$$

And adding the Z'Z matrix:

$$Z'Z + A^{-1}\alpha = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} + \begin{bmatrix} 1.8166 & -4.8442 & 0 \\ -4.8442 & 1.8166 & 0 \\ 0 & 0 & 7.2664 \end{bmatrix} = \begin{bmatrix} 5.8166 & -0.8442 & 0 \\ -0.8442 & 5.8166 & 0 \\ 0 & 0 & 11.2664 \end{bmatrix}$$

Therefore, the Henderson normal equations are:

$$\begin{bmatrix} & & & & & \\ 6 & 0 & 2.0000 & 2.0000 & 2.0000 & \\ 0 & 6 & 2.0000 & 2.0000 & 2.0000 & \\ 2 & 2 & 5.8166 & -0.8442 & 0.0000 & \\ 2 & 2 & -.08442 & 5.8166 & 0.0000 & \\ 2 & 2 & 0.0000 & 0.0000 & 11.2664 & \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix}$$

And the solution of these equations is:

$$\begin{bmatrix} b_1 \\ b_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} & & & & & \\ 6 & 0 & 2.0000 & 2.0000 & 2.0000 & \\ 0 & 6 & 2.0000 & 2.0000 & 2.0000 & \\ 2 & 2 & 5.8166 & -0.8442 & 0.0000 & \\ 2 & 2 & -0.8442 & 5.8166 & 0.0000 & \\ 2 & 2 & 0.0000 & 0.0000 & 11.2664 & \end{bmatrix}^{-1}\begin{bmatrix} 185 \\ 210 \\ 132 \\ 124 \\ 139 \end{bmatrix} = \begin{bmatrix} 31.6285 \\ 35.7952 \\ -0.7765 \\ -1.9776 \\ 0.3685 \end{bmatrix}$$

In this solution, the fixed effect (sex) is adjusted for the random example, and the random effect is adjusted for the fixed effect.

Unbalanced data in a reproductive model

Assuming the last missing data, the $SCtotal$ is:

$$SCtotal = SC_{total} = 36^2 + 35^2 + 33^2 + \ldots + 29^2 - \frac{360^2}{11} = 148.1818182$$

The GLs are:

$$GL_s = 3 - 1 = 2$$

$$GL_{sexo} = 2 - 1 = 1$$

$$GL_{total} = 11 - 1 = 10$$

$$GL_e = 10 - 1 - 2 = 11 - 3 = 7$$

To calculate the ordinary least squares solutions for the reduced model (without the sire factor) must be estimated:

$$b = \begin{bmatrix} 11 & 5 & 6 \\ 5 & 5 & 0 \\ 6 & 0 & 6 \end{bmatrix}^{-1}\begin{bmatrix} 360 \\ 150 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1666 & -0.1666 & 0 \\ -0.1666 & 0.3666 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 360 \\ 150 \\ 0 \end{bmatrix} = \begin{bmatrix} 35 \\ -5 \\ 0 \end{bmatrix}$$

Citation: Raúl PGJ, Valladares M, Daniel D. Theory of estimation of parameters and genetic values under mixed models. *Int J Avian & Wildlife Biol.* 2024;8(1):27–33. DOI: 10.15406/ijawb.2024.08.00210

Therefore, the $SCsexo$ for the reduced model is:

$$SC_{sexo(reducido)} = \begin{bmatrix}35 & -5 & 0\end{bmatrix}\begin{bmatrix}360\\150\\210\end{bmatrix} - \frac{360^2}{11} = (35)(360)+(-5)(150)-\frac{360^2}{11} = 68.1818$$

To obtain the adjusted SCs, we calculate the SC for the full model and subtract from it, respectively ($reducido$):

$$SCs = SC(\mu, s, sexo,) - SC(\mu, sexo) = 83.6818182 - 68.1818 = 15.5$$

And the SC_e is:

$$SC_e = 148.1818182 - (15.5 + 68.1818) = 64.5$$

Therefore, the CMs are:

$$CM_s = \frac{15.5}{2} = 7.75$$

$$CM_e = \frac{64.5}{7} = 9.214286$$

Finally, the variance components are:

$$\sigma_s^2 = \frac{7.75 - 9.214286}{3.6} = -0.4067$$

$$\sigma_e^2 = 9.214286$$

The component σ^2_s component is a negative estimate of the variance, because the $CMs < CMe$is negative, therefore, the ML and REML estimates for σ^2_s are:

$$\sigma^2_s = 0$$

In other words, all the total variability is residual. Table 3 shows the ML and REML iterations for the calculation of the residual variance with the Newton-Rapson method: As shown in Table 3, when an unbalanced database is used, the estimates for the variance components are different in ANOVA and REML, since with ANOVA we have σ^2_e = 9.2142 and with REML σ^2_e = 8.88. For this particular case, $h^2 = 0$ because the variance of the numerator is zero, obviously to find a more credible estimate, one should increase the number of data used in the genetic evaluation or try another model and compare the AIC.

Table 3 Iteration of variance components for ML and REML

ML				REML			
Iteración	**-2ln(L)**	σ_s^2	σ_e^2	**iteración**	**-2ln(L_r)**	σ_s^2	σ_e^2
1	53.0420	0	7.2727	1	48.6053	0	8.88
2	53.0420	0	7.2727	2	48.6053	0	8.88

Table 4 Genetic values using an animal model

Animal	VG
1	0.422982
2	-0.984574
3	1.10566
4	0.217214
5	0.809321
6	-0.651756
7	-0.273233
8	-0.857664
9	-2.32642
10	0.113062E-01
11	1.33545
12	1.04324
13	-0.117947
14	1.63535

Animal model

The solutions of the Henderson normal equations, using the values of σ^2_e = 9.083 y σ^2_a = 5 are presented in Table 4 Henderson's normal equations are not presented for this case due to its large dimensions.

Conclusion

Genetic values and parameters can be estimated using ANOVA, ML and REML, but with the risk of estimating negative variance components using ANOVA or zero (or overestimated) heritabilities with likelihood-based methods when the data structure or model is not correct. When the data structure is unbalanced, mathematical calculations with ANOVA, ML and REML are more complex and require computational algorithms with higher performance.

Acknowledgments

None.

Conflicts of interest

The authors declared that there are no conflicts of interest.

References

1. Caballero J, Pablo E, Martinez C. Restricted maximum likelihood estimation of variance and covariance components of multiple traits under designs i and ii of North Carolina. *Rev Fitotec Mex.* 2003;26(1):53–66.
2. Castejón, Osiris. *Design and analysis of experiments with statistix*. Venezuela: Rafael Urdaneta University; 2008.
3. Elzo M, Garay O. *Evolution of genetic improvement practices in domestic animal populations*. Gainesville: University of Florida; 2012.
4. Román R, Aranguren A. Genetic evaluation of breeding stock: achievements and challenges From: Achievements and challenges in dual purpose cattle breeding; 2014.
5. Gutiérrez P. Introduction to animal genetic assessment methodology adapted to the EHEA. Madrid: Complutense; 2010.
6. Aranguren A, Román R. The simple animal model: a methodology for geneticists from: achievements and challenges of dual purpose cattle breeding; 2014.
7. Perez J, Montiel N. Heritability of the IBMI index of Italian buffaloes used in artificial insemination. *Rev Cientif FCV-LUZ XXXIII*. 2023;205–207.
8. Román R, Aranguren J, Garcidueñas R, et al. Association between reproductive characteristics and milk production in crossbred heifers. *Revista Espamciencia*. 2023;14(2):63–70.
9. Montgomery D. *Experiments with random factors. From: Design and analysis of experiments*. 2nd edn. Limusa-wiley; 2015.
10. Searle R, Casella G, Mcculloch C. *Variance Components*. Wiley series; 1992.

11. Aldrich J. R. A. Fisher and the making of maximum likelihood 1912 - 1922. *Statist Sci*. 1997;12(3):162–176.

12. Hartley HO, Rao JNK. "Maximum likelihood estimation for the mixed analysis of variance model". *Biometrika*. 1967;54(1):93–108.

13. Patterson HD, Thompson R. "Recovery of inter-block information when block sizes are unequal". *Biometrika*. 1971;58:545–554.

14. Meyer K. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genet Sel Evol.* 1989;21:317–340.

15. Román R, Aranguren J, Villasmil Y, et al. Analysis of fertility at first service in dual-purpose heifers under an animal model. *Revista Científica*. 2010;20(4):383–389.