Properties of mendelian residuals when regressing breeding values using a genomic covariance matrix

Rodolfo J.C. Cantet, Zulma Vitezica

Properties of Mendelian Residuals when regressing Breeding Values using a Genomic Covariance Matrix

R. J.C. Cantet[1] **Z. G. Vitezica**[2]
[1]Universidad de Buenos Aires, Facultad de Agronomía – CONICET, Buenos Aires, Argentina
[2]INRA, UMR1388, Toulouse, France

ABSTRACT: Properties of Mendelian residuals when predicting breeding values (BV) with a positive definite genomic covariance matrix are presented. It is well known that in an infinitesimal model with an additive relationship matrix built from pedigree data, the variance of the Mendelian residuals (MR) from the regression of BV on those from the ancestors, explains half the additive variance without inbreeding (*F*), and a little less than that if *F* of the parents is not zero. We show that: 1) the residual variance of BV regression using a genomic covariance matrix is always less or equal, than Mendelian variance obtained from predictions calculated without using genomic information; 2) MR are independent if BV of ancestors, parents and collateral related non-descendants animals (i.e. full and half-sibs, uncles, cousins) are included in the regression equation.

Key words: genomic covariance matrix; mendelian and regression residuals; residual independence.

INTRODUCTION

While celebrating a symposium in honor of D. Gianola, M. Soller stressed the fact that prediction of breeding values (BV) from an infinitesimal animal model can only explain a little less than 50% of the additive genetic variance if inbreeding is present, so that the accuracy of prediction of BV is limited when using individual phenotype and pedigree. On the other hand, genomic information from a large number of markers is nowadays available, so that reduction of residual variance from predictions using genomic data onto the covariance matrix of BV may be reduced when compared with those from the classical animal model. The fact that the residual additive variance in the regression of BV is reduced by the use of genomic information is recognized by Patry and Ducrocq (2011) when saying that "the usual assumptions on Mendelian sampling expected value and variance are no longer valid". The Mendelian sampling expected value is no longer zero so that the resulting relationship matrix is no longer the correct one". The goal of this note is to present the properties of Mendelian residuals from the regression of BV of an individual into the BV of ancestors, parents and related collateral non-descendants animals (i.e. uncles, full and half-sibs, cousins), when using a genomic covariance matrix.

METHODS

IBD and Additive Relationships when Genomic Markers are Available

The infinitesimal model uses the additive genetic covariance between relatives, which in turn rests on the assumption of independent segregation of loci throughout the genome, so that the additive relationships involved can be calculated as probabilities of genes shared identical by descent (IBD) by relatives. However, genomes are inherited by segments rather than by individual bases, and there is variability among pairs of full and half-sibs in the fractions of genome shared IBD (Guo, 1995; Hill and Weir, 2011). Therefore, the usual calculation of additive relationships using pedigrees is just the expected value of the relationship matrix that results from the realized IBD process. Donnelly (1983) proposed to model the IBD process throughout the genome as a Markov process. Guo (1995) observed that the true probability function is difficult to obtain, but not its mean and variance. After him, we define the pairwise IBD relationship between two animals as the estimated expected value of the shared IBD process. The expectation can be estimated using a large number of markers such as SNPs. Animal breeders use the proportion of SNPs shared by two individuals (Van Raden, 2008), as an estimator of that expectation.

Prediction of Breeding Values and the Additive Relationship Matrix under the Classical Setting

To express the BV of any individual (say *i*) on the BVs of its parents *S* and *D* without genomic information, Foulley and Chevalet (1981) set up the following regression model

$$\boldsymbol{a}_i = b_S\,\boldsymbol{a}_S + b_D\,\boldsymbol{a}_D + \boldsymbol{\phi}_i \qquad [1]$$

where b_S and b_D are the regression coefficients of the BV of the father ($\boldsymbol{a}_S$) and of the mother ($\boldsymbol{a}_D$), on the BV of *i* ($\boldsymbol{a}_i$), respectively. The "error term" ($\boldsymbol{\phi}_i$, i.e. the MR) represents the deviation from the mid parental BVs, and originates in random Mendelian sampling of the grand-parental gametes present in the parents (segregation), possibly coupled to recombination of the grand-paternal gametes in the parents due to crossing over during meiosis. Let $\boldsymbol{b} = [b_S, b_D]'$ and $\boldsymbol{a}_A = [\boldsymbol{a}_S, \boldsymbol{a}_D]'$, Foulley and Chevalet (1981) estimated the regression coefficients as follows

$$\boldsymbol{b} = \left[\mathrm{Var}\begin{bmatrix}\boldsymbol{a}_S\\ \boldsymbol{a}_D\end{bmatrix}\right]^{-1}\mathrm{cov}\left(\begin{bmatrix}\boldsymbol{a}_S\\ \boldsymbol{a}_D\end{bmatrix}, \boldsymbol{a}_i'\right) = \left(\sigma^2_A \boldsymbol{A}_P\right)^{-1}\boldsymbol{g}\;\sigma^2_A = \boldsymbol{A}_P^{-1}\boldsymbol{g} \qquad [2]$$

where

$$\mathrm{Var}\begin{bmatrix}a_S\\ a_D\end{bmatrix} = \begin{bmatrix}1+F_S & 2F_i\\ 2F_i & 1+F_D\end{bmatrix}\sigma^2_A \qquad [3]$$

$$\mathrm{Cov}\begin{bmatrix}a_S\\ a_D\end{bmatrix}, \boldsymbol{a}_i = \begin{bmatrix}0.5\left(1+F_S+2F_i\right)\\ 0.5\left(1+F_D+2F_i\right)\end{bmatrix}\sigma^2_A$$

After a bit of algebra, solutions for [2] are $b_S = b_D = 0.5$, and [1] becomes

$$a_i = \tfrac{1}{2}\,a_S + \tfrac{1}{2}\,a_D + \boldsymbol{\phi}_i \qquad [4]$$

Therefore, in classic theory without genomic information the BV of an individual is regressed to half its parental BV, plus the MR (i.e. Bulmer, 1985, page 125; Quaas, 1988). After [4], we can write $\phi_i = a_i - 0.5\,[a_S + a_D]$, and obtain the variance of ϕ_i as the residual variance of the regression model [1], thus yielding

$$\mathrm{Var}(\phi_i) = \left[(1+F_i) - \boldsymbol{b}'\boldsymbol{A}_P\,\boldsymbol{b}\right]\sigma^2_A = \left[(1+F_i) - \boldsymbol{b}'\boldsymbol{g}\right]\sigma^2_A \quad [5]$$

Bulmer (1985) observed that ϕ_i is independent of any other MR, so that it is also independent of the BV of any other individual. On recalling that $F_i = 0.5\,A_{SD}$, Foulley and Chevalet (1981) observed that:

$$\mathrm{Var}(\phi_i) = \frac{1}{2}\left[1 - \frac{F_S + F_D}{2}\right]\sigma^2_A \quad [6]$$

In absence of inbreeding $\mathrm{Var}(\phi_i) = 0.5\,\sigma^2_A$, i.e. the "within family" additive variance explained by the regression in the infinitesimal model (Walsh and Lynch, 2013; chapter 22), is half the additive variance. However, marker data can help in reducing the value of $\mathrm{Var}(\phi_i)$ as we will shown below.

Reduction of Residual Variance due to using Genomic Information

Consider the regression of the BV of an individual in the BV of other animals in a pedigree. Without genomic information, the MR are uncorrelated. However, the use of genomic information introduces a lack of independence among those fractions of the genome that are shared IBD between individuals over the expected value of the relationship that is calculated only with the pedigree, i.e. the elements of $\boldsymbol{A}$. In a regression setting this can be viewed as comparing two regression equations (Greene, 2012). The "short regression", i.e. model [4], or the regression of the individual BV (a) on the parental BV (from now on the 2 × 1 vector $\boldsymbol{a}_P$ defined as $\boldsymbol{a}_P' = [a_S, a_D]'$. On the other hand, the second model is the "long regression" and corresponds to the regression of BV with a genomic covariance matrix that has as elements relationships from the fraction of genome shared IBD. In this case, a is regressed to both $\boldsymbol{a}_P$ and $\boldsymbol{a}_O$. The latter vector comprises the BV of animals being either ancestors or contemporaries, save its parents and descendants. More formally the joint distribution of BV is taken to be

$$\begin{bmatrix}\boldsymbol{a}_O\\ \boldsymbol{a}_P\\ a\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ 0\end{bmatrix}, \begin{bmatrix}\boldsymbol{G}_{OO} & \boldsymbol{G}_{OP} & \boldsymbol{g}_O\\ \boldsymbol{G}_{PO} & \boldsymbol{G}_{PP} & \boldsymbol{g}_P\\ \boldsymbol{g}_O' & \boldsymbol{g}_P' & 1+F\end{bmatrix}\sigma^2_A\right) \quad [7]$$

The "long regression" is represented as

$$a = \boldsymbol{b}_P'\boldsymbol{a}_P + \boldsymbol{b}_O'\boldsymbol{a}_O + \phi_G \quad [8]$$

In [8] the vectors of regression coefficients are $\boldsymbol{b}_P$ for the parents and $\boldsymbol{b}_O$ for the remaining non-independent BV from related individuals, whereas ϕ_G is the Mendelian residual of the "long regression" using genomic information, and is different from ϕ in [4].

Alternatively, the "short regression" [4] is written as

$$a = \boldsymbol{b}'\boldsymbol{a}_P + \phi \quad [9]$$

The 2×1 vector $\boldsymbol{b}$ has both elements equal to 0.5. We now prove that $\mathrm{Var}(\phi_G)$ in [9] is always smaller to $\mathrm{Var}(\phi)$ in [4], and is equal only if no genomic information is used.

First, the values of the regression coefficients in the "long regression" are obtained as in [2] by solving the following linear system

$$\begin{bmatrix}\boldsymbol{G}_{OO} & \boldsymbol{G}_{OP}\\ \boldsymbol{G}_{PO} & \boldsymbol{G}_{PP}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_O\\ \boldsymbol{b}_P\end{bmatrix} = \begin{bmatrix}\boldsymbol{g}_O\\ \boldsymbol{g}_P\end{bmatrix} \quad [10]$$

Using standard results on the inverse of a partitioned matrix (Greene, 2008, expression [A.74]), the inverse of the matrix in [10] is equal to

$$\begin{bmatrix}\boldsymbol{G}_{OO} & \boldsymbol{G}_{OP}\\ \boldsymbol{G}_{PO} & \boldsymbol{G}_{PP}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{G}^{-1}_{O|P} & -\boldsymbol{G}^{-1}_{OO}\boldsymbol{G}_{OP}\boldsymbol{G}^{-1}_{P|O}\\ -\boldsymbol{G}^{-1}_{P|O}\boldsymbol{G}_{PO}\boldsymbol{G}^{-1}_{OO} & \boldsymbol{G}^{-1}_{P|O}\end{bmatrix} \quad [11]$$

Conditional variances in [11] are equal to

$$\boldsymbol{G}_{O|P} = \boldsymbol{G}_{OO} - \boldsymbol{G}_{OP}\boldsymbol{G}^{-1}_{PP}\boldsymbol{G}_{PO} \quad \boldsymbol{G}_{P|O} = \boldsymbol{G}_{PP} - \boldsymbol{G}_{PO}\boldsymbol{G}^{-1}_{OO}\boldsymbol{G}_{OP} \quad [12]$$

Equating both regression models allows obtaining the following expression that relates both Mendelian residuals:

$$\phi_G = \phi - \left[\left(\boldsymbol{b} - \boldsymbol{b}_P\right)'\boldsymbol{a}_P + \boldsymbol{b}_O'\boldsymbol{a}_O\right] \quad [13]$$

Notice that, unless $\boldsymbol{G} = \boldsymbol{A}$, $\boldsymbol{b}_P$ will not be equal to $\boldsymbol{b}$: the regression coefficients resulting from genome IBD sharing from both parents of the individual will not necessarily be equal to 0.5, as the 50% of the genes inherited from a parent can be expressed as a linear combination of genes shared IBD by the individual with the parents' ancestors.

After a lot of algebra, it can be proved that

$$\mathrm{Var}(\phi_G) = \mathrm{Var}(\phi) - \boldsymbol{b}_O'\boldsymbol{G}_{O|P}\,\boldsymbol{b}_O \quad [14]$$

The quadratic form $\boldsymbol{b}_O'\boldsymbol{G}_{O|P}\boldsymbol{b}_O$ is always positive as long as $\boldsymbol{b}_O \neq 0$, which in turn can only happen if $\boldsymbol{G}_{PO} = \boldsymbol{0}$, i.e. if there is no shared IBD (genomic relationships reflecting common segregation) between the breeding values of parents and other individuals in the pedigree. Therefore, except for the latter case in which the variances will be equal, $\mathrm{Var}(\phi_G) < \mathrm{Var}(\phi)$.

Independence of the Mendelian residuals

An important property of the Mendelian residuals is that they are independent. In the block regression setting with a genomic covariance matrix such as in the "long regression" model [7]-[8], this property is reproduced as long as the BV of all other descendants from parents and ancestors, born up to the birthdate of the animal (we assume that individuals are ordered by date of birth into $\boldsymbol{a}$, as commonly used to build $\boldsymbol{A}^{-1}$ or $\boldsymbol{A}$) are included in the regression for animal i. To see this decompose $\boldsymbol{b}_O$ for the remaining non-independent BV from related individuals into $\boldsymbol{b}_O' = [\boldsymbol{b}_A'\ \boldsymbol{b}_C']$. The sub-indices A and C correspond to ancestors and relatives other than ancestors and parents (i.e. uncles, half and full sibs, cousins) of animal i. Observe that, up to the date i was born, its descendants are not yet in $\boldsymbol{G}$. This will confer to $\boldsymbol{G}$ a similar triangular Cholesky root free decomposition as found in $\boldsymbol{A}$. Without loss of generality we can write the vector of BV as $\boldsymbol{a}' = [\boldsymbol{a}_A'\ \boldsymbol{a}_P'\ \boldsymbol{a}_C'\ a_i]$, and system [10] is now equal to

$$\begin{bmatrix} \boldsymbol{G}_{AA} & \boldsymbol{G}_{OP} & \boldsymbol{G}_{AC} \\ \boldsymbol{G}_{PA} & \boldsymbol{G}_{PP} & \boldsymbol{G}_{PC} \\ \boldsymbol{G}_{CA} & \boldsymbol{G}_{CP} & \boldsymbol{G}_{CC} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_{O} \\ \boldsymbol{b}_{P} \\ \boldsymbol{b}_{C} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}_{A} \\ \boldsymbol{g}_{P} \\ \boldsymbol{g}_{C} \end{bmatrix} \quad [15]$$

Using formulae from the inverse of a positive definite matrix partitioned in three blocks (Hrone, 2006), we can proved that all matrices in the off-diagonal blocks of $\boldsymbol{G}^{-1}$ are not zero unless $\boldsymbol{G}_{AC} = \boldsymbol{0}$ and $\boldsymbol{G}_{PC} = \boldsymbol{0}$. This implies that ancestors and parents do not have other descendants than *i* (and its parents). Therefore, one should be careful to include all descendants of ancestors and parents that precede *i* in $\boldsymbol{G}$ in such a way that the Mendelian residual covariance matrix be diagonal.

The Genomic Covariance Matrix

Let $\boldsymbol{B}$ be a triangular matrix of order *q* that relates BV of individuals to ancestors, parents and contemporaries in $\boldsymbol{a}$, by associating the regression coefficients in $\boldsymbol{b}$ with a_i, coefficients that are calculated using genomic relationships. Matrix $\boldsymbol{B}$ enters into $\boldsymbol{G}$ in the following manner

$$\boldsymbol{a} = \boldsymbol{B}\,\boldsymbol{a} + \boldsymbol{\phi} \quad \text{or} \quad (\boldsymbol{I} - \boldsymbol{B})\,\boldsymbol{a} = \boldsymbol{\phi} \quad [16]$$

Elements of $\boldsymbol{B}$ are such that $-1 < \boldsymbol{B}_{ij} < 1$ as inbreeding will enlarge parent-offpring regressions but selfing or cloning will not be allowed. It can be verified that all eigenvalues of $\boldsymbol{B}$ are equal to zero and the largest eigenvalue (or spectral radius) of $\boldsymbol{B}$ ($\rho(\boldsymbol{B})$) is less than one. Then, Lemma 2.1 in Berman and Plemmons (1994) indicates that $\rho(\boldsymbol{B}) < 1$ if and only if $(\boldsymbol{I} - \boldsymbol{B})^{-1}$ exists and is equal to

$$(\boldsymbol{I} - \boldsymbol{B})^{-1} = \boldsymbol{I} + \sum_{k=1}^{K} \boldsymbol{B}^{k} \quad [17]$$

The value of *K* is the number of generations (or matrix powers) in the largest path between an ancestor and a descendant, and is similar to what happens with $\boldsymbol{A}$ (Quaas, 1988). The closer to 1 is the magnitude of any regression coefficient; the higher will be the value of *K*. This is a consequence of genomic data bringing information from all meiosis from a common ancestor in the base generation down to its last descendant, back and forth. This formulation allows taking the variance operator in [16] and obtaining

$$\begin{aligned} \mathrm{Var}(\boldsymbol{a}) &= \mathrm{Var}\left[(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\phi}\right] = (\boldsymbol{I} - \boldsymbol{B})^{-1}\,\mathrm{Var}(\boldsymbol{\phi})(\boldsymbol{I} - \boldsymbol{B}')^{-1} \\ &= (\boldsymbol{I} - \boldsymbol{B})^{-1}\,\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{B}')^{-1}\,\sigma^2_A = \boldsymbol{G}\,\sigma^2_A \end{aligned} \quad [18]$$

So that

$$\boldsymbol{G} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\,\boldsymbol{D}\,(\boldsymbol{I} - \boldsymbol{B}')^{-1} \quad [19]$$

The elements of $\boldsymbol{D}$ are $\boldsymbol{D}_{ii} = \mathrm{Var}(\phi_G) / \sigma^2_A$, and the buildup of $\boldsymbol{G}$ parallels the one for $\boldsymbol{A}$ as in Quaas (1988).

DISCUSSION

To relate genomic selection with classical prediction of BV from animal models using BLUP, an infinitesimal model using genomic information is needed. Some properties of such a model are: 1) Normality of BV using genomic information, while taking into account linkage disequilibrium; 2) regression properties of BV among generations; 3) reduction of Mendelian variance due to using genomic information; 4) independence of Mendelian residuals when using genomic information. Item 1) was dealt with by Dawson (1997) who found that only under extreme LD asymptotic normality was not attained. Our expression [8] characterizes 2), and we have given evidence that 3) and 4) hold. The reduction of Mendelian variance explains the increase in accuracy of BV prediction when using genomic data. Moreover, the increase in accuracy can be calculated with the following expression for prediction error variance (PEV; modified from Henderson, 1975):

$$\mathrm{Var}(\hat{\boldsymbol{a}} - \boldsymbol{a}) = \boldsymbol{C}^{aa} + \boldsymbol{C}^{aa}\,\boldsymbol{A}^{-1}(\boldsymbol{G} - \boldsymbol{A})\,\boldsymbol{A}^{-1}\,\boldsymbol{C}^{aa}\,\sigma^2_A$$

and $\boldsymbol{C}^{aa}$ is PEV under the animal model without genomic information. Finally, independence of the Mendelian residuals is useful to invert $\boldsymbol{G}$ in [19] avoiding direct inversion, i.e. by employing rules similar to the ones used to calculate $\boldsymbol{A}^{-1}$.

ACKNOWLEDGEMENT

The first author thanks the INP Toulouse, France, for the travel grant that allowed him to perform the current research.

LITERATURE CITED

Berman, A. and Plemmons, R. J. (1994). *Nonnegative Matrices in the Mathematical Sciences*. SIAM Press, Philadelphia, PA, 2nd edition.

Bulmer, M. G., (1985). *The mathematical theory of quantitative genetics*. Oxford University Press, Oxford, UK.

Dawson, K.J. (1997) Theor. Popul. Biol. 52, 137-154.

Donnelly, K. P. (1983). Theor. Popul. Biol. 23:34-63.

Foulley, J. L., and Chevalet C. (1981). Ann. Génét. Sél. Anim. 13:189 -196.

Greene, W. H. (2008). *Econometric analysis*. Prentice Hall, NJ, USA.

Guo, S. W. (1995). Amer. J. Human Genet. 56:1468-1476.

Hill, W. G. and Weir, B. S. (2011) Genetics Res. 93:47-64.

Henderson, C.R. 1975. J. Dairy Sci. 41:760-770.

Hrone, K. (2006) Acta Univ. Palacki. Olomuc., Fac. Rer. Nat., Mathematica 45:67–80.

Patry, C. & Ducrocq,V.(2011). J.Dairy Sci. 94:1011-1020.

Quaas, R. L. (1988). J. Dairy Sci. 71(Supp. 2):91–98.

Van Raden, P. (2008). J. Dairy Sci. 91:4414–4423.

Walsh, B. and Lynch, M. (2013). *Selection and Evolution of Quantitative Traits* I. *Foundations*. Chapter 22.

Wermuth, N. (1980). J. Amer. Statist. Assoc. 75:963-972.