The use of a genetic relationship matrix biases the best linear unbiased prediction

BONGSONG KIM

RiceTec, Inc, Alvin, TX 77511, USA
E-mail: bkim@ricetec.com.

Received 12 December 2019; revised 20 April 2020; accepted 22 April 2020

Abstract. The best linear unbiased prediction (BLUP), derived from the linear mixed model (LMM), has been popularly used to estimate animal and plant breeding values (BVs) for a few decades. Conventional BLUP has a constraint that BVs are estimated from the assumed covariance among unknown BVs, namely conventional BLUP assumes that its covariance matrix is a $\lambda K$, in which $\lambda$ is a coefficient that leads to the minimum mean square error of the LMM, and $K$ is a genetic relationship matrix. The uncertainty regarding the use of $\lambda K$ in conventional BLUP was recognized by past studies, but it has not been sufficiently investigated. This study was motivated to answer the following question: is it indeed reasonable to use a $\lambda K$ in conventional BLUP? The mathematical investigation concluded: (i) the use of a $\lambda K$ in conventional BLUP biases the estimated BVs, and (ii) the objective BLUP, mathematically derived from the LMM, has the same representation as the least squares.

Keywords. best linear unbiased prediction; breeding values; genetic relationship matrix; numerator relationship matrix; least squares.

Introduction

A breeding value (BV) refers to the combining ability of an entity as a parent. Estimating the BVs is purposeful in seed and livestock industries, aiming to save time and resources in increasing genetic gains by maximizing selection accuracies and efficiencies. The best linear unbiased prediction (BLUP), derived from the linear mixed model (LMM), has been popularly used to estimate BVs for a few decades (Piepho 1994; Panter and Allen 1995a, b; Meuwissen et al. 2001; Choi et al. 2017). Conventional BLUP has a constraint that the estimated breeding values (EBVs) are derived from the assumed covariance among unknown BVs. Conventional BLUP uses a matrix representing pairwise genetic variations for an assumed covariance matrix (Henderson 1975). Pedigrees or genomic fingerprints are typically required for calculating the assumed covariance matrix (Emik and Terrill 1949; VanRaden 2008; Kim et al. 2016; Kim and Beavis 2017).

Previous studies reported that EBVs obtained by conventional BLUP have a high correlation with empirically observed combining abilities, based on field tests and computer simulations (Belonsky and Kennedy 1988; Piepho 1994; Panter and Allen 1995a, b; Bauer et al. 2006; Nielsen et al. 2011; Choi et al. 2017; Manzanilla-Pech et al. 2017). However, no studies clarified two uncertainties regarding conventional BLUP. First, there are no clues that the assumed covariance matrix is objective. If the assumed and objective covariance matrices are not equal, conventional BLUP is biased. Second, BVs are not physically measurable but conceptual. In fact, the combining ability for an entity can vary according to its gametes, mating partners and environments.

In this study, I investigated the aforementioned uncertainties from a mathematical perspective. As a result, this study discovered the following facts: (i) the objective covariance matrix can be estimated; (ii) conventional BLUP is biased in that the assumed and objective covariance matrices are different; and (iii) the unbiased BLUP, mathematically derived from the LMM, is equal to the least squares.
Materials and methods

Naïve BLUP

The LMM is denoted as

\[ y = X\beta + Zu + \epsilon \]  

(1)

where \( y \) is the phenotypic variable; \( X \) is the design matrix for fixed effect; \( Z \) is the design matrix for random effect; \( \beta \) is the unknown fixed-effect variable; \( u \) is the unknown random-effect variable; and \( \epsilon \) is the error-term variable. Equation (1) assumes \( \epsilon \sim N(0, \sigma_\epsilon^2) \). The \( u \) represents a variable for BVs and can be calculated as follows:

\[ u = (Z'Z)^{-1}Z'(y - X\beta - \epsilon) \]  

(2)

The \( \text{var}(u) \) can be calculated as follows:

\[
\text{var}(u) = (Z'Z)^{-1}Z'\text{var}(y - X\beta - \epsilon)Z(Z'Z)^{-1} \\
= (Z'Z)^{-1}Z'(\text{var}(y) - 1\sigma_\epsilon^2)Z(Z'Z)^{-1} \\
= (Z'Z)^{-1}Z'\text{var}(y)Z(Z'Z)^{-1} - \sigma_\epsilon^2(Z'Z)^{-1} 
\]  

(3)

The \( \text{var}(y) \) can be calculated as follows:

\[ \text{var}(y) = Z\text{var}(u)Z' + 1\sigma_\epsilon^2 \]  

(4)

Let us substitute equation (4) for equation (3):

\[
\text{var}(u) = (Z'Z)^{-1}Z'\text{var}(y)Z(Z'Z)^{-1} \\
+ \sigma_\epsilon^2(Z'Z)^{-1}Z'\text{var}(y)Z(Z'Z)^{-1} - \sigma_\epsilon^2(Z'Z)^{-1} 
\]  

(5)

Equation (5) demonstrates that \( \sigma_\epsilon^2(Z'Z)^{-1} \) in equation (3) is negligible because it will be cancelled by \( \text{var}(y) \). For the sake of simplicity, therefore, equation (3) can omit \( \sigma_\epsilon^2(Z'Z)^{-1} \) so that:

\[ \text{var}(u) = (Z'Z)^{-1}Z'\text{var}(y)Z(Z'Z)^{-1} \]  

(6)

Because \( \text{var}(y) = \frac{1}{n-1}(y - \bar{y})(y - \bar{y})' \), equation (6) can be rewritten as:

\[ \text{var}(u) = \frac{1}{n-1}Z'(y - \bar{y})(y - \bar{y})Z(Z'Z)^{-1} \]  

(7)

where \( n \) is the number of values in \( u \); and \( \bar{y} \) is a vector containing the arithmetic averages of \( y \).

Equation (7) is in the same structure as \( \text{var}(u) = \frac{1}{n-1}(u - \bar{u})(u - \bar{u})' \). Therefore,

\[ \hat{u} = (Z'Z)^{-1}Z'y \]  

(8)

Hereafter, equation (8) is called naïve BLUP. This is the unbiased solution of the LMM and has the same representation as the least squares. Every value of \( \hat{u} \) in equation (8) represents the arithmetic average of multiple phenotypic observations for each entity. Given the \( \hat{u} \), the \( \hat{\beta} \) for equation (1) can be calculated using the following equation:

\[ \hat{\beta} = (X'X)^{-1}X'(y - Z\hat{u}) \]  

(9)

Conventional BLUP

Henderson et al. (1959) derived the following formula from the LMM:

\[
\begin{pmatrix}
X'X & X'Z \\
Z'Z + \text{var}(u)^{-1} & \text{var}(u)
\end{pmatrix}
\begin{pmatrix}
\hat{\beta} \\
\hat{u}
\end{pmatrix}
= \begin{pmatrix}
X'y \\
Z'y
\end{pmatrix}
\]  

(10)

In equation (10), \( \hat{u} \) can be obtained using the row operation (Kim et al. 2019) as follows:

\[
\hat{u} = (Z'Z + \text{var}(u)^{-1} - Z'X(X'X)^{-1}X'Z)^{-1} \\
\quad (Z'y - Z'X(X'X)^{-1}X'y)
\]  

(11)

Henderson (1975) assumed that \( \text{var}(u) = \lambda K \), where \( \lambda \) is a coefficient that leads to the minimum mean square error (MMSE) of the LMM, and \( K \) is a genetic relationship matrix (Robinson 1991). Therefore, equation (11) can be rewritten as:

\[
\hat{u} = (Z'Z + (\lambda K)^{-1} - Z'X(X'X)^{-1}X'Z)^{-1} \\
\quad (Z'y - Z'X(X'X)^{-1}X'y)
\]  

(12)

Equation (12) is a mathematical representation of conventional BLUP. Because \( \lambda = \frac{\sigma_\epsilon^2}{\sigma_u^2} \), conventional BLUP can be fitted by estimating \( \sigma_\epsilon^2 \) and \( \sigma_u^2 \), in which the former and the latter are the variance components for \( \epsilon \) and \( u \), respectively. In this study, equation 12 was fitted using the expectation-maximization (EM) algorithm (Kim et al. 2019). Given the \( \hat{u} \), the \( \hat{\beta} \) can be calculated using equation 9.

Rice dataset

To compare between two sets of EBVs, obtained by naïve BLUP and conventional BLUP, an open-access rice data set (Spindel et al. 2015) comprising phenotypic data and single-nucleotide polymorphism (SNP) data was used, which was downloaded from http://ricediversity.org/data/index.cfm. The phenotypic trait used in this study was grain yield (kg/ha), which was recorded over four years (2009, 2010, 2011, 2012), two seasons (DS: dry season, WS: wet season) per year and three replications (1, 2, 3) per season. The three variables were reduced to two by averaging the three replications within each season. Consequently, the grain yield observations were split into eight environments: 2009/DS, 2009/WS, 2010/DS, 2010/WS, 2011/DS, 2011/WS, 2012/DS and 2012/WS. Meanwhile, the SNP data consisted of 108,024 SNPs, which was used to calculate a genetic relationship matrix (K) using Numericware i (Kim and Beavis 2019).
2017). The common entities between the SNP data and phenotypic data were 107, which was subjected to statistical analyses.

**Condition of conventional BLUP being unbiased**

The following proposition was established to judge whether conventional BLUP is biased.

**Proposition 1.** Conventional BLUP is unbiased, only if 

\[
\lambda K = \frac{1}{n-1}(Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}.
\]

**Proof.** Conventional BLUP assumes that \( \text{var}(u) = \lambda K \). According to the process from equations (1) through (7), the objective \( \text{var}(u) \) is defined as \( \text{var}(u) = \frac{1}{n-1}(Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1} \). Therefore, if conventional BLUP is unbiased, it must be satisfied that 

\[
\lambda K = \frac{1}{n-1}(Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}.
\]

**Results**

**Comparison between naïve BLUP and conventional BLUP**

Figure 1 shows the correlation between two sets of EBVs, obtained by naïve BLUP and conventional BLUP. Its Pearson correlation coefficient \( (\rho) \) is 0.9526. Table 1 summarizes the two sets of EBVs. Figure 1 and table 1 show that the two sets of EBVs deviate from each other. Naïve BLUP is equal to the least squares, suggesting that its resulting EBVs are statistically unbiased. The deviation between the two sets of EBVs indicates that conventional BLUP is biased due to the assumption that \( \text{var}(u) = \lambda K \).

**Comparison between the objective and assumed covariance matrices**

The objective and assumed covariance matrices were computed using the same rice dataset. The submatrices of the former and the latter are shown in tables 2 and 3, respectively. The objective covariance matrix was calculated by equation (7). The assumed covariance matrix was calculated by \( \lambda K \). The \( \lambda \) is defined as \( \frac{\sigma_e^2}{\sigma_u^2} \), in which \( \sigma_e^2 \) and \( \sigma_u^2 \) are the variance components for \( e \) and \( u \), respectively. In this study, the resulting estimates for \( \sigma_e^2 \) and \( \sigma_u^2 \) were 80019.630 and 86076.655, respectively. It led to \( \lambda = 1.076 \). Therefore, table 3 represents a submatrix of 1.076*\( K \). The comparison between tables 2 and 3 shows that the two covariance matrices are different. Figure 2 shows a low correlation \( (\rho = 0.099) \) between the two covariance matrices. The discrepancy between the two covariance matrices led to the conclusion that 

\[
\lambda K \neq \frac{1}{n-1}(Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}.
\]

According to Proposition 1, therefore, conventional BLUP is biased.

**Discussion**

Naïve BLUP is based on the objective covariance matrix defined as equation (7), which implies that \( y \) influences the direction of \( u \). The fact that \( u \) depends on \( y \) is crucial because the phenotypic variable \( (y) \) must impact the BV variable \( (u) \). Naïve BLUP is equal to the least squares, i.e. each naïve BLUP estimate represents the arithmetic average of multiple phenotypic observations per entity. This means that naïve BLUP estimates are unbiased because the arithmetic average indicates the point where the variance among phenotypic

---

**Table 1.** Summaries of the naïve BLUP estimates and conventional BLUP estimates.

|                | Min.  | 1st Qu. | Median | Mean   | 3rd Qu. | Max.  |
|----------------|-------|---------|--------|--------|---------|-------|
| Conventional BLUP | -253.58 | 30.46   | 118.45 | 113.18 | 201.40  | 381.81|
| Naïve BLUP       | 2445  | 2855    | 3005   | 2986   | 3103    | 3390  |
Table 2. The objective covariance submatrix calculated by equation (7).

|    | A1257 | A1258 | A1259 | A1260 | A1261 | A1262 | A1263 | A1264 | A1265 | A1267 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A1257 | 24.60617 | -124.532 | 37.15311 | -54.4295 | -19.9907 | -180.25 | -56.2965 | -94.8407 | -23.1525 | -61.8573 |
| A1258 | -124.532 | 630.2552 | -188.032 | 275.4677 | 101.1728 | 912.2462 | 284.9165 | 479.9887 | 117.1748 | 313.0598 |
| A1259 | 37.15311 | -188.032 | 56.09787 | -82.1837 | 44.21993 | 16.24093 | 45.73671 | 77.051 | 18.80968 | 50.25445 |
| A1260 | -54.4295 | 275.4677 | -82.1837 | 120.3996 | 44.21993 | 146.4399 | 85.0027 | 143.201 | -34.9582 | -93.399 |
| A1261 | -19.9907 | 101.1728 | -30.1841 | 44.21993 | 16.24093 | 85.0027 | 143.201 | -34.9582 | 136.8301 |
| A1262 | -180.25 | 912.2462 | -272.162 | 398.7185 | 124.5294 | 209.7902 | 51.21398 | 136.8301 |
| A1263 | -56.2965 | 284.9165 | -143.201 | 51.21398 | 136.8301 |

Table 3. The assumed covariance submatrix calculated by $\lambda K$.

|    | A1257 | A1258 | A1259 | A1260 | A1261 | A1262 | A1263 | A1264 | A1265 | A1267 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A1257 | 2.151388 | 1.851162 | 1.850721 | 1.800379 | 1.812319 | 1.731394 | 1.827389 | 1.873655 | 1.873053 | 1.885466 |
| A1258 | 1.851162 | 2.151388 | 1.849775 | 1.811587 | 1.79897 | 1.664529 | 1.797334 | 1.865082 | 1.826185 | 1.817923 |
| A1259 | 1.850721 | 1.849775 | 2.151388 | 1.722724 | 1.768743 | 1.662496 | 1.772594 | 1.926719 | 1.78477 | 1.801519 |
| A1260 | 1.800379 | 1.811587 | 1.722724 | 2.151388 | 1.795947 | 1.659366 | 1.8301 | 1.814675 | 1.924923 | 1.874623 |
| A1261 | 1.812319 | 1.79897 | 1.768743 | 1.795947 | 2.151388 | 1.801766 | 1.742151 | 1.79795 | 1.796958 | 1.796079 |
| A1262 | 1.731394 | 1.664529 | 1.662496 | 1.659366 | 1.801766 | 2.151388 | 1.742833 | 1.705642 | 1.719659 | 1.715926 |
| A1263 | 1.827389 | 1.797334 | 1.772594 | 1.8301 | 1.742151 | 1.724833 | 2.151388 | 1.800562 | 1.835156 | 1.846408 |
| A1264 | 1.873655 | 1.865082 | 1.926719 | 1.814675 | 1.792795 | 1.705642 | 1.800562 | 2.151388 | 1.8696 | 1.837092 |
| A1265 | 1.873053 | 1.826185 | 1.78477 | 1.924923 | 1.796958 | 1.719659 | 1.835156 | 1.8696 | 2.151388 | 1.876742 |
| A1267 | 1.885466 | 1.817923 | 1.801519 | 1.874623 | 1.796079 | 1.876742 | 2.151388 | 1.8696 | 1.837092 | 1.876742 | 2.151388 |
Conventional BLUP has two fatal flaws regarding its constraint, $\text{var}(u) = \lambda K$. First, conventional BLUP ignores the fact that the direction of $u$ must depend on $y$. According to the constraint, $\lambda$ just multiplies $K$. This means that the direction of $u$ is solely determined by $K$. Second, the constraint per se is unrealistic. In order for the constraint to be true, it must be satisfied that $\lambda K = \frac{1}{\pi^2} (Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}$. However, figure 2 and the comparison between tables 2 and 3 demonstrate that $\lambda K \neq \frac{1}{\pi^2} (Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}$. This means that conventional BLUP violates the constraint. According to Proposition 1, therefore, conventional BLUP is biased.

Conclusion

The uncertainty regarding the use of $\lambda K$ in conventional BLUP was recognized by past studies (Blasco 2001; Postma 2006; Hadfield et al. 2010). Blasco (2001) argued that conventional BLUP may not be entitled to being called best and unbiased. Nevertheless, the use of $\lambda K$ in conventional BLUP has been considered as the best practice to overcome the unknowable $\text{var}(u)$. Notably, this study derived naïve BLUP from the linear mixed model, and two facts were found therefrom. First, the objective covariance matrix can be defined as $\text{var}(u) = \frac{1}{\pi^2} (Z'Z)^{-1}Z'(y - \bar{y})(y - \bar{y})'Z(Z'Z)^{-1}$. Second, naïve BLUP is equal to the least squares. Considering that each naïve BLUP estimate represents the arithmetic average of multiple phenotypic observations per entity, the application of naïve BLUP is suited for autogamous species. This is because multiple entities being genetically identical can be made in autogamous species. The naïve BLUP estimate for each entity indicates a genetic performance that implies the adaptability to environments. From this perspective, naïve BLUP estimates can be EBVs. However, the use of naïve BLUP estimates is at a breeder’s discretion. This is because the combining ability for a parent must have a large variability due to unknown complex polygenic effects such as hybrid vigour or inbreeding depression.

References

Bauer A. M., Reetz T. C. and Léon J. 2006 Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. *Crop Sci.* 46, 2685–2691.

Belonsky G. M. and Kennedy B. W. 1988 Selection on individual phenotype and best linear unbiased predictor of breeding value in a closed swine herd. *J. Anim. Sci.* 66, 1124–1131.

Blasco A. 2001 The Bayesian controversy in animal breeding. *J. Anim. Sci.* 79, 2023–2046.

Choi T., Lim D., Park B., Sharma A., Kim J.-J., Kim S. et al. 2017 Accuracy of genomic breeding value prediction for intramuscular fat using different genomic relationship matrices in Hanwoo (Korean cattle). *Asian-Australas J. Anim. Sci.* 30, 907–911.

Emik L. O. and Terrill C. E. 1949 Systematic procedures for calculating inbreeding coefficients. *J. Hered.* 40, 51–55.

Hadfield J. D., Wilson A. J., Garant, D., Sheldon B. C. and Kruuk L. E. B. 2010 The misuse of BLUP in ecology and evolution. *Am. Nat.* 175, 116–125.

Henderson C. R. 1975 Best Linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447.

Henderson C. R., Kemphorne O., Searle S. R. and von Krosigk C.M. 1959 The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15, 192–218.

Kim B. and Beavis W. D. 2017 Numericware i: Identical by State Matrix Calculator. *Bioinformatics* 33, 11769-31688663 (online).

Kim B., Beavis W. D. and Léon J. 2016 Numericware N: Numerator Relationship Matrix Calculator. *J. Hered.* 107, 686–690.

Kim B., Dai X., Zhang W., Zhuang Z., Sanchez D. L., Lübberstedt T. et al. 2019 GWASpro: a high-performance genome-wide association analysis server. *Bioinformatics* 35, 2512–2514.

Manzannilla-Pech C. I. V., Veerkamp R. F., de Haas Y., Calus M. P. L. and Ten Napel J. 2017 Accuracies of breeding values for dry matter intake using nongenotyped animals and predictor traits in different lactations. *J. Dairy Sci.* 100, 9103–9114.

Meuwissen T. H. E., Hayes B. J. and Goddard M. E. 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Nielsen H. M., Sonesson A. K. and Meuwissen T. H. E. 2011 Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. *J. Anim. Sci.* 89, 630–638.

Panter D. and Allen F. L. 1995a Using best linear unbiased predictions to enhance breeding for yield in soybean. II: Selection of superior crosses from a limited number of yield trials. *Crop Sci.* 35, 405–410.
Corresponding editor: H. A. Ranganath