Approximating Reliabilities of Estimated Direct Genomic Values

Z. Liu, F. Seefried, F. Reinhardt and R. Reents
***vit** w.V., Heideweg 1, 27283 Verden, Germany*

Abstract

Genomic evaluation usually combines estimated direct genomic values (DGV) with conventional EBVs using their associated reliabilities. Matrix inversion is required to obtain reliabilities of DGV (VanRaden, 2008), and the dimension of inverted matrices corresponds to the number of genotyped animals in genomic reference population. As more animals get genotyped, direct inversion of those matrices becomes increasingly less feasible. The objective of this study was to develop an approximation method for the reliabilities of DGV for genotyped candidates. A total of 5025 German Holstein reference bulls were considered to derive a prediction formula for DGV reliabilities of 5344 genotyped Holstein candidates. Four out of 12 predictor variables were selected for the final prediction equation, which was consistent across all evaluated traits. The four predictor variables were reasonably highly correlated with the response variable, DGV reliability. The correlations ranged from 0.61 to 0.72. Predicted DGV reliabilities had an average correlation of 0.91 with observed ones for all trait groups, except female fertility traits. No systematic bias was observed via residual analysis. The developed prediction formulae were proven to be accurate for using the German national reference population. However, they were found to be inapplicable to the case of using EuroGenomics reference population for predicting German candidates' reliabilities, because both reference populations differed in size and structure significantly. Consequently, new prediction formulae must be derived when the genomic reference population changes from German national to EuroGenomics training sets.

1. Introduction

Genomic evaluation usually includes the steps of estimating DGV and combining them with conventional evaluation. Estimated DGV, the sum of all SNP effect estimates, are associated with certain reliabilities, which are calculated by inverting matrices and genomic relationship matrix (VanRaden, 2008; Liu *et al.,* 2009) among other statistical procedures. Despite desirable properties of those reliabilities of DGV estimates, the direct matrix inversion approach is only feasible as long as the number of genotyped animals does not exceed a certain threshold, because computing requirements for matrix inversion increases quickly with the number of genotyped animals. As genomic reference population for German Holsteins was switched from German national to a joint four European countries (EuroGenomics) reference population, the number of genotyped reference bulls reached 17,054 in January 2010. Due to the much bigger reference population, inverting matrices of this dimension became more and more difficult. The objective of this study was to develop a statistical method for approximating reliabilities of DGV for routine genomic evaluation without matrix inversion.

2. Materials and Methods

Genomic data and evaluation results were obtained from January 2010 national genomic evaluation for German Holsteins. Among 10,487 genotyped animals, there were 5025 Holstein bulls included in German national genomic reference population and 5344 genotyped Holstein candidates without phenotypes. Those candidates were born from 2005 to 2009. Their reliabilities of estimated DGV, calculated by matrix inversion (Reinhardt *et al.*, 2009), were used as response variable in this study. A total of 12 predictor variables were developed to derive a prediction formula for approximating reliabilities of estimated DGV.

For a given candidate *i*, an average genomic relationship coefficient, $\bar{g}_i$, was calculated as:

$$\bar{g}_i = (\sum_{j=1}^{n} g_{ij}) / n \qquad [1]$$

where g_{ij} is genomic relationship coefficient (VanRaden, 2008) between the candidate *i* and a reference bull *j* ($j = 1, \cdots, n$), and *n* is the number of genotyped bulls in the reference population. Squared value of the average genomic relationship, $\bar{g}_i^2$, was considered as predictor variable, too. The third predictor variable was the maximum value of genomic relationship of candidate *i* with all reference bulls:

$$g_i^{\max} = \max(g_{i1,} \cdots, g_{in}). \qquad [2]$$

And the fourth investigated predictor variable was defined as sum of squared genomic relationship of candidate *i* with all reference bulls:

$$g_i^2 = (\sum_{j=1}^{n} g_{ij}^2) / n. \qquad [3]$$

Reference bulls may differ in their reliabilities of phenotypic daughter information (REL_j^{DAU}) or combined genomic information (REL_j^G). In order to account for the impact of individual difference in the reliabilities of reference bulls on candidate's DGV reliabilities, the following three predictor variables were studied as well:

$$pg_i^2 = (\sum_{j=1}^{n} g_{ij}^2 * \frac{g_{jj}(1 - REL_j^{DAU})}{\lambda}) / n \qquad [4]$$

$$gg_i^2 = (\sum_{j=1}^{n} g_{ij}^2 * \frac{g_{jj}(1 - REL_j^{G})}{\lambda}) / n \qquad [5]$$

$$dg_i^2 = (\sum_{j=1}^{n} g_{ij}^2 * \frac{g_{jj}(1 - (REL_j^{G} - REL_j^{DAU}))}{\lambda}) / n \qquad [6]$$

where pg_i^2, gg_i^2, and dg_i^2 represent predictor variables for considering individual reference bull in phenotypic daughter reliability, combined genomic reliability, and their difference, respectively; g_{jj} is diagonal element of genomic relationship matrix (VanRaden, 2008) for reference bull *j*, λ is the ratio of residual to sire variance. Five additional predictors were developed and investigated. Due to their poorer association with the response variable, they were no longer considered in further analyses,.

Model fitting was examined for all combinations of the remaining predictor variables using multiple regression method. Optimal subset regression was determined by jointly considering goodness of fit of each model (R^2 value) and the number of fitted predictor variables.

3. Results

3.1. Correlations of predictor variables with DGV reliabilities of candidates

Table 1 shows simple correlations of the candidates' reliabilities of DGV (response variable) with the seven predictor variables. Because those correlations did not differ much between traits, only correlations of protein yield are given here. It can be seen in Table 1 that average genomic relationship is reasonably well correlated with the response variable, 0.66. The four quadratic predictor variables are almost equally highly correlated with DGV reliabilities of the candidates.

Table 1. Correlations of candidates' reliabilities of DGV with predictor variables for protein.

Predictor variable	Correlation
Average genomic relationship $\bar{g}_i$	0.66
Squared average value $\bar{g}_i^2$	0.64
Max. genomic relationship $g_i^{\max}$	0.61
Sum of squared relationship g_i^2	0.72
Daughter reliability pg_i^2	0.71
Genomic reliability gg_i^2	0.71
Genomic-daughter reliability dg_i^2	0.72

3.2. Optimal subset regressions for predicting DGV reliabilities

All combinations of the seven predictor variables were considered in the multiple regression analysis for predicting reliabilities of the candidates' DGV. Goodness of fit of the multiple regressions was measured by R^2 value of the prediction models. With additional consideration of the number of fitted predictor variables, an optimal subset prediction formula was found for predicting DGV reliabilities of the candidates:

$$REL_i = b_0 + b_1\bar{g}_i + b_2\bar{g}_i^2 + b_3 g_i^{\max} + b_4 g_i^2 \qquad [7]$$

where REL_i denotes reliability of DGV of candidate *i*, and b's are multiple regression coefficients. The regression equation [7] was identified as the optimal prediction formula consistently across all the genomically evaluated traits (Reinhardt *et al.*, 2009). Although all of the predictor variables were positively correlated with the response variable, regression coefficient for the squared genomic relationship value, b_2, was negative. The relationship between DGV reliabilities and genomic relationship of candidates to reference animals was clearly non-linear. It is interesting to note that reliabilities of individual reference bulls were no longer important, which was possibly caused by fitting the predictor variable of the maximum genomic relationship $g_i^{\max}$.

All of the four predictor variables were highly correlated among themselves, except $g_i^{\max}$. Table 2 shows mean, standard deviation, minimum and maximum values of the response and selected predictor variables. Because of similarity across traits, only non-return rate cow is given here as an example. It can be seen that the DGV reliabilities varied considerably among candidates. Average genomic relationship of candidates with German national reference population had a mean of 0.059 and standard deviation of 0.010. Maximum genomic relationship of the candidates had an average of 0.42, indicating that not all of the candidates had a genotyped sire or fullsib in the reference population.

Table 2. Descriptive statistics of the response and selected predictor variables for non-return rate cow trait.

Variable	Mean	Std	Min	Max
DGV reliability REL_i	.53	.05	.27	.69
Mean relationship value $\bar{g}_i$	.059	.010	.006	.085
Squared average relationship $\bar{g}_i^2$	.0036	.0011	.0000	.0073
Maximum relationship $g_i^{\max}$	.42	.12	.11	.69
Sum of squared relationship g_i^2	.0050	.0014	.0005	.0095

3.3. Accuracy of the reliability prediction formulae

Table 3 summaries averaged accuracy of the genomic reliability prediction formula [7] for all traits in each of seven trait groups. Across all trait groups, predicted genomic reliability was highly correlated with its true value, above 0.9, except the female fertility trait group. The high correlation and low MSE indicated a high level of goodness of fit of the prediction equation [7]. The poorer fit of the fertility trait group may be attributed to their very low heritability values. In one test run, non-Holstein candidates were added to the analysis, which resulted in higher R^2 value of the prediction model but much unfavourable MSE value. This suggests that R^2 value alone may not be enough to make model selection accurately.

Residuals of predicted genomic reliabilities of the candidates were distributed sysmetrically around 0. Additionally, no systematic biases were observed with respect to the predictor variables.

4. Discussion

A prediction equation was developed for approximating reliabilities of DGV for genotyped candidates using German national genotyping population. Consistent prediction formula comprising four predictor variables was obtained across all the trait groups. Reasonably high goodness of fit was achieved for all the traits. No systematic bias was found via residual analysis. Interestingly, individual phenotypic or genomic reliabilities of reference bulls no longer had a major impact on the reliabilities of the candidates, once the four predictor variables had already been included in the reliability prediction formula.

Table 3. Accuracy of the genomic reliability prediction formula, averaged for each group.

Trait group	Correlation with predicted reliability	MSE (x1000)
Milk production	.907	.36
Udder health	.908	.40
Longevity	.906	.50
Female fertility	.863	.74
Calving	.906	.46
Workability	.907	.58
Conformation	.910	.44

The approximated DGV reliabilities should be adjusted to the level of realised reliabilities via a validation study. The prediction formulae need to be validated using a different set of candidates. It is import to keep in mind that the same reference population must be kept in such cross-validation studies.

The reliability prediction formulae derived from the German national reference population were applied to the EuroGenomics data. Poorer results were obtained, due to changes in the level of average as well as maximum genomic relationship of the German candidates to EuroGenomics reference bulls. This finding suggests a new derivation would be required, if genomic reference population changes in structure and size significantly.

5. References

Ducrocq, V. & Liu, Z. 2009. Combining genomic and classical information in national BLUP evaluations. *Interbull Bulletin 40,* 172-177.

Liu, Z., Seefried, F., Reinhardt, F. & Reents, R. 2009. Dairy cattle genetic evaluation using genomic information. *Interbull Bulletin 39,* 23-28.

Reinhardt, F., Liu, Z., Seefried, F. & Thaller, G. 2009. Implementation of genomic evaluation in German Holsteins. *Interbull Bulletin 40,* 219-226.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91,* 4414-4423.