Simultaneous Deregression of Cow and Bull Breeding Values

M. P. L. Calus[1], G. Schopen[2], R. F. Veerkamp[1]
[1]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, The Netherlands
[2]CRV BV, The Netherlands

Abstract

The next step to increase the accuracy of genomic prediction is to extend reference populations with cows next to daughter proven bulls. Cows typically have estimated breeding values (EBV) with considerably lower reliabilities compared to bulls. This suggests that commonly used (approximate) deregression procedures for bulls may not be appropriate for cows. The objective of this study was to test an alternative approach to simultaneously de-regress EBV of cows and bulls, and to derive appropriate weights for those de-regressed EBV. First, the appropriate weights of the de-regressed EBV were derived, and then the de-regressed EBV were computed using those weights. The analyses showed that the methods were well able to accurately de-regress EBV and compute their weights, both for bulls and cows. Despite observed discrepancies between intermediate results and simulated values, final EBV and reliabilities correlated very well with original values.

Key words: Deregression, reference population, cows

Introduction

The size of the reference population is known to be a key parameter that influences the accuracy of genomic prediction. This fact has motivated breeding companies to genotype as much as possible bulls with highly reliable estimated breeding values (EBV), and to exchange genotypes between organizations and countries (e.g. Lund, *et al*., 2011). The next step to further increase the accuracy of genomic prediction is to start genotyping cows and add them to the reference population.

For genomic prediction, EBV of the bulls included in the reference population are commonly used as phenotypes, after being "deregressed". The deregression procedure involves adjusting the scale of the EBV, such that all have the same variance, and removing information from relatives (e.g. Jairath, *et al*., 1998). Most deregression procedures are approximate by nature, that perform quite well when de-regressing highly reliable EBV of bulls. These approximate procedures, however, are likely not able to properly deal with EBV of cows, that typically have much lower reliabilities. The objective of this study was to test an alternative approach to simultaneously de-regress EBV of cows and bulls, and to derive appropriate weights for those de-regressed EBV (also termed "de-regressed proofs"; DRP) to be used in genomic prediction with a combined cow and bull reference population.

Deregression of EBV

The procedure used to compute DRP has been termed "matrix deregression", because it involves calculations involving matrices based on the mixed model equations of the original model. Consider that we want to deregress a set of EBV of genotyped animals (hereafter referred to as "reference animals"), and that these reference animals are a subset of a larger dataset used in an evaluation system to obtain the EBV. In this case, the deregression procedure should correct within the subset all EBV for information of relatives within this subset. For instance, the EBV of a sire that is based on 100 daughters of which 10 are genotyped, should be corrected for the 10 genotyped daughters, such that the DRP still includes the information of the 90 daughters that are not genotyped. The matrix deregression procedure involves setting up the following mixed model equations for a pedigree based model that only includes the reference animals in the subset (i.e. the genotyped animals in our case):

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad [1]$$

where **1** is a vector or ones, **Z** is an incidence matrix for random additive genetic effects (note that those are diagonal in this case and can therefore be omitted), $\mathbf{R}^{-1}$ is a diagonal matrix with EDC (effective daughter contribution) values, $\mathbf{A}^{-1}$ is the inverse relationship matrix between reference animals (note that this matrix is obtained by first setting up **A** for the reference animals using the full pedigree, and then inverting it), $\hat{\mu}$ is a mean, $\hat{\mathbf{a}}$ is the vector of EBV that needs to be deregressed, and **y** is a vector of DRP that need to be computed. Note that fixed effects can be omitted from the above (e.g. Harris and Johnson, 2010). However, not accounting for the mean may lead to inconsistent results for traits with a low heritability (Jairath, *et al*., 1998, Rogers, *et al*., 1996). Therefore, we performed matrix deregression that accounts for the mean. Based on equation [1], this was done as follows:

1. Initialize $\hat{\mu}$ by assigning the mean of $\hat{\mathbf{a}}$
2. Compute:

$$\begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{R}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \end{bmatrix} \quad [2]$$

3. Compute $\hat{\mu} = (\mathbf{1}'\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{R}^{-1}\mathbf{y}$
4. Iterate 2 and 3 until convergence
5. Compute DRP as $\mathbf{R}\mathbf{R}^{-1}\mathbf{y}$

Note that this procedure is similar to the procedure described by Jairath *et al*. (1998), apart we use the inverse of **A** including only reference animals (computed using the full pedigree), while Jairath *et al*. (1998) use an $\mathbf{A}^{-1}$ that includes the whole pedigree, and therefore can be composed directly using Henderson's rules.

Deregression of EDC

The procedure to compute DRP requires "deregressed EDC" (dEDC) as input. Here we developed a new procedure to "deregress" EDC, that adjusts an animals' EDC (EDC_i) for information from its parents, offspring and sibs. This is achieved by iteratively computing the EDC contribution of those relatives in the reference population (EDC_{rel_k}), and then the dEDC of animal *k* is computed as

$$dEDC_k = EDC_k - EDC_{rel_k} \quad [3]$$

The procedure starts with initializing dEDC values by setting them equal to EDC values for all reference animals (i.e. $dEDC_k = EDC_k$). Thereafter, the following steps are taken for each reference animal:

1. Set up part of the coefficient matrix of the mixed model equations of the model that computes EBV from the DRP, considering all reference animals, using current dEDC values
2. For each reference animal, obtain inverse of the coefficient matrix that is created in step 1, where the dEDC value of the animal is set to zero, to compute EDC_{rel_k}
3. For each reference animal, compute updated value of $dEDC_k$ using [3]
4. Repeat steps 1-3 until convergence

Step 1 requires an inverse relationship matrix for all reference animals. Therefore, the full numerator relationship matrix was computed, and inverted, for all reference animals, using the full pedigree. In step 2, for each reference animal the inverse of the coefficient matrix was required that used a dEDC value of zero for this particular animal. Realizing that the only difference between the two coefficient matrices is a change of one diagonal value, the procedure outlined by Hager (1989) can be used. Consider that we have an $n \times n$ square invertible matrix **LHS** (the "full" coefficient matrix) and a matrix **LHSR** that is identical to **LHS** except for the elements in one column, being the diagonal element for reference animal *k* in our case, then the element $lhsr_{kk}^{-1}$ of $\mathbf{LHSR}^{-1}$ can be expressed in terms of the elements lhs_{ij}^{-1} of $\mathbf{LHS}^{-1}$ as:

$$lhsr_{kk}^{-1} = \frac{lhs_{kk}^{-1}}{\sum_{i=1}^{n} lhs_{ki}^{-1} lhsr_{ik}} \quad [4]$$

This value can then be used to compute the reliability of reference animal k (REL_k) as (Mrode, 2005):

$$REL_k = 1 - lhsr_{kk}^{-1} \times \lambda \quad [5]$$

where $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$. This is then used to compute EDC_k as:

$$EDC_k = \frac{\lambda \times REL_k}{1 - REL_k} \quad [6]$$

Validation of deregression procedure

To validate the deregression procedure, phenotypic data were simulated through an existing pedigree of (supposedly genotyped) 13,720 cows and 1532 bulls. Cows received 1 to 5 records with equal probability. For the bulls, 50-200 additional daughters with one record each and unknown dams were simulated. True EDC were computed based on simulated own records for cows, and based on simulated daughter records for bulls.

The following analyses were performed on this data:

1. EBV and EDC were estimated from the phenotypic data
2. Obtained EBV and EDC were deregressed
3. The DRP and dEDC obtained in step 2 were used to estimate EBV

Steps 1 and 3 were performed with a straightforward BLUP animal model. Step 2 was performed with the procedures that were outlined previously. In all three steps, the simulated variance components were used.

The formal validation involved comparison of the DRP and dEDC values obtained in step 2 to the simulated values. The EBV obtained from step 3 were also compared to those from step 1, since this is a straightforward test that may be performed for real data, where the true breeding values and EDC are not known.

Results

Breeding values were estimated from the simulated data that included 748,308 records, and 673,740 animals in the pedigree. In this analysis, simulated variance components were used.

The first step of the deregression procedure involved estimating dEDC. The algorithm converged after 5 iterations, while using only 1-2 iterations gave very similar results (not shown). The dEDC values were expected to be close to the true values, which was indeed the case for many animals, but for several bulls (Figure 1) and cows (Figure 2), the dEDC values were substantially larger than the true values.

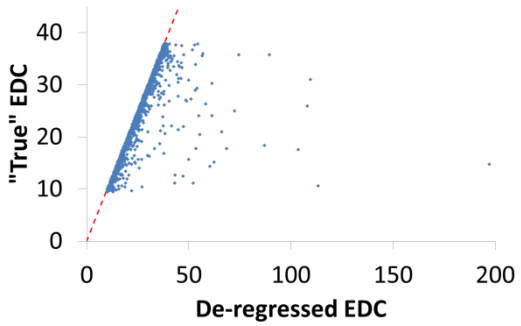

Figure 1. Simulated (true) versus de-regressed EDC for bulls.

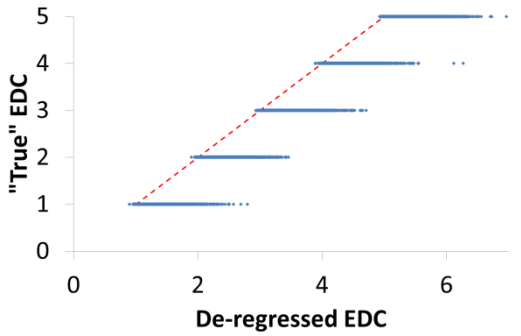

Figure 2. Simulated (true) versus de-regressed EDC for bulls.

Using the dEDC values, the DRP values were computed with formula [1]. In the next step, DRP and dEDC values were used in a straightforward BLUP model. The (final) EBV obtained from this BLUP model, were very close to the original breeding values (Figure 3), as expected.

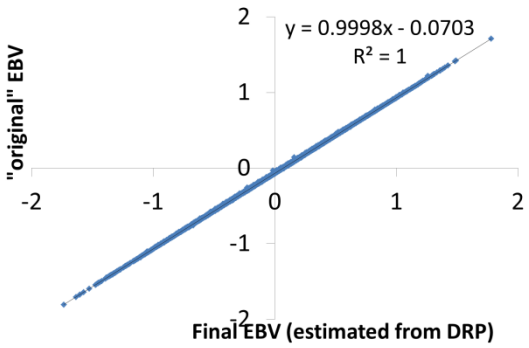

Figure 3. Original versus final EBV.

The same comparison was made between the reliabilities of original and final EBV. Those values were also very close (Figure 4), and more so for the bulls (reliability values > 0.81) than for the cows (reliability values < 0.81).

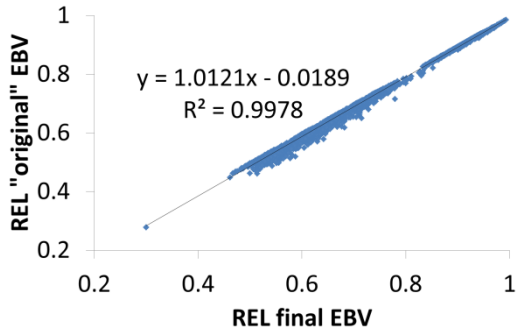

Figure 4. Reliabilities of original versus final EBV.

Discussion

Our objective was to test an alternative approach to compute dEDC and DRP values, that was expected to yield more optimal results for reference animals with low reliability EBVs (e.g. cows), compared to commonly used approximate deregression methods. Our results confirmed that final EBV and their reliabilities were close to the original values, as expected.

During the Interbull meeting, two important comments were made. The first one was that deregression procedures should include all EBV that were estimated in the initial "full" genetic evaluation. The second one expanded on this to state that at least all the ancestors of reference animals are included in the deregression procedure. The latter was in fact achieved, because although the A matrix used included only animals with EBV, the whole pedigree was used to compute the relationships. An alternative approach is to directly build $\mathbf{A}^{-1}$ including all animals and their ancestors, and include that in equations [1] and [2], as done by Jairath *et al*. (1998). This avoids inversion of A, which may become problematic if the number of reference animals becomes very large.

The method used to compute dEDC, was theoretically expected to yield exact results, and indeed the final results obtained were very close to the expectations. Intermediate results, i.e. comparing dEDC to "true" EDC values, did show discrepancies, especially for bulls with many (grand)sons in the data. It is currently unclear why the dEDC values for those animals are overestimated.

One drawback of the approach taken to compute dEDC, is that it involves inversion of potentially large matrices. Nevertheless, it is expected that approximate methods will be developed that avoid matrix inversion. Our validation approach and our method can serve as a reference to investigate how accurate such methods may be.

Conclusions

The presented methods can be used to derive phenotypes and weights for genomic prediction using a combined cow and bull reference population. The analyses showed

that the methods were well able to accurately de-regress EDC and breeding values. Despite observed discrepancies between intermediate results and simulated values, final EBV and reliabilities correlated very well with original values.

Acknowledgments

CRV is acknowledged for financial support.

References

Hager, W.W. 1989. Updating the inverse of a matrix. *SIAM Review. 31:2,* 221-239.

Harris, B.L. & Johnson, D.L. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci. 93:3,* 1243-1252.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B. & Kolstad, B. 1998. Genetic evaluation for herd life in Canada. *J. Dairy Sci. 81:2,* 550-562.

Lund, M., de Roos, S., de Vries, A., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F. & Su, G. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol. 43:1,* 43.

Mrode, R. 2005. *Linear Models for the Prediction of Animal Breeding Values.* 2nd edition ed. CABI Publishing.

Rogers, G., Banos, G., Nielsen, U.S. & Philipsson, J. 1996. Genetic correlations among somatic cell scores, productive life, and type traits from the United States and udder health measures from Denmark and Sweden. *Interbull Bulletin 14,* 34-38.