J. Dairy Sci. 99:1–17
http://dx.doi.org/10.3168/jds.2016-11028

Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights

M. P. L. Calus,[1] J. Vandenplas, J. ten Napel, and R. F. Veerkamp
Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 6700 AH Wageningen, the Netherlands

ABSTRACT

Training of genomic prediction in dairy cattle may use deregressed proofs (DRP) as phenotypes. In this case, DRP should be estimated breeding values (EBV) corrected for information of relatives included in the data used for genomic prediction, and adjusted for regression to the mean (i.e., their reliability). Deregression is especially important when combining animals with EBV with low reliability, as commonly the case for cows, and high reliability. The objective of this paper, therefore, was to compare the performance of different deregression procedures for data that include both cow and bull EBV, and to develop and test procedures to obtain the appropriate deregressed weights for the DRP. Considered DRP were EBV: without any adjustment, adjusted for information of parents and regression to the mean, or adjusted for information of all relatives and regression to the mean. Considered deregressed weights were weights of initial EBV: without any adjustment, adjusted for information of parents, or adjusted for information of all relatives. The procedures were compared using simulated data based on an existing pedigree with 1,532 bulls and 13,720 cows that were considered to be included in the data used for genomic prediction. For each cow, 1 to 5 records were simulated. For each bull, an additional 50 to 200 daughters with 1 record each were simulated to generate a source of data that was not used for genomic prediction. The simulated trait had either a heritability of 0.05 or 0.3. The validation involved 3 steps: (1) computation of initial EBV and weights, (2) deregression of those EBV and weights, (3) using deregressed EBV and weights to compute final EBV, (4) comparison of the initial and final EBV and weights. The methods developed to compute appropriate weights for the DRP were either very precise and computationally somewhat demanding for larger data sets, or were less precise but computationally trivial due their approximate nature. Adjusting DRP for all relatives, known as matrix deregression, yields by definition final EBV that are identical to the original EBV. Matrix deregression is therefore preferred over other approaches that only correct for information of parents or not performing any deregression at all. It is important to use appropriate weights for the DRP, properly corrected for information of relatives, especially when individual reliabilities of final EBV are computed based on the prediction error variance of the model.

Key words: deregression, deregressed proofs, reference population

INTRODUCTION

Several applications on dairy cattle data use pseudo-phenotypes as response variables, such as international genetic evaluations performed by Interbull (Schaeffer, 1994), integration of foreign information into a national genetic evaluation (e.g., Banos et al., 1992), QTL mapping (e.g., Rodriguez-Zas et al., 2002; Boichard et al., 2003), and genomic prediction (Meuwissen et al., 2001). The pseudo-phenotypes usually summarize the phenotypic performance of female progeny of bulls. They can also summarize the phenotypic performance of individual dairy cows, for instance across lactations. Commonly used pseudo-phenotypes for cows and bulls are so-called yield deviations (**YD**) and daughter yield deviations (VanRaden and Wiggans, 1991). Those YD are derived from a model that uses raw phenotypic data as input and adjusts for fixed effects such as herd-year-season. For many applications, YD cannot be derived, because the raw phenotypic data are not available. In addition, the analysis of certain traits requires complicated models, such as survival analysis using a proportional hazards model (Ducrocq, 1994), which makes derivation of YD difficult. In those cases where YD are not available, information of EBV, commonly derived from national genetic evaluations, are used to compute pseudo-phenotypes (e.g., Rozzi et al., 1990; Banos et al., 1993; Bonaiti and Boichard, 1995; Jairath et al., 1998).

Breeding values are commonly estimated using BLUP (Henderson, 1949). One of the features of BLUP

Received February 12, 2016.
Accepted April 6, 2016.
[1] Corresponding author: mario.calus@wur.nl

is that it regresses EBV to the mean to avoid bias in the prediction of phenotypes. This feature makes the variance of EBV proportional to their reliability, and the higher the reliability, the higher the variance of the EBV. For the aforementioned applications that use pseudo-phenotypes as response variables, it is required that the pseudo-phenotypes give the same EBV and reliabilities as obtained from the underlying genetic evaluation. This is commonly achieved by deregressing the EBV. The most simple form of deregression involves dividing each EBV by its reliability (Goddard, 1985; Garrick et al., 2009). The resulting variables are commonly termed deregressed proofs (**DRP**).

Apart from the difference in variance of EBV with different reliabilities, an additional issue is that EBV include information of related animals whose EBV will also be used in the subsequent analyses. Nevertheless, the subsequent analyses assume that the pseudo-phenotypes derived from the EBV are measured on the animal itself, and this information is then propagated toward relatives in the same subset. To avoid issues due to double counting of the same information, information that comes from relatives that are included with their data in the subsequent analyses should be removed from the pseudo-phenotypes. Sophisticated deregression procedures enable the simultaneously adjusting of EBV for differences in reliabilities as well as to account for information of relatives (e.g., Sigurdsson and Banos, 1995; Jairath et al., 1998). For each DRP value, a weight, such as the number of effective daughter contributions (**EDC**; Fikse and Banos, 2001), is commonly computed that reflects its reliability. When DRP are corrected for information of relatives, this also means that appropriate weights of the DRP are on average expected to be lower than the corresponding weights of the original EBV because the DRP contain less information than the EBV.

Whether or not deregression of EBV and computation of appropriate weights has a (large) effect on the results in subsequent analyses depends on differences between reliabilities of EBV across animals due to different amounts of information per individual. In dairy cattle, typically a few influential bulls have many progeny in the population, whereas a large number of bulls each have a limited number of progeny. Nevertheless, bulls with proofs based on daughter performance generally have reliabilities that are greater than 0.5. In these cases, the amount of "own" information is relatively large, and the effect of accounting (or not) for information of relatives in the deregression procedure is expected to have relatively little effect (Sigurdsson and Banos, 1995; Guo et al., 2010). When animals with substantially lower reliabilities are included, such as cows, the effect of deregression is expected to be much more important. For example, Vandenplas et al. (2014b) estimated that >85% of the total information of animals with a reliability lower than 0.50 for milk traits was due to information from relatives in the data, whereas this percentage was lower than 7% for dairy bulls having daughters with records. In the context of genomic prediction, deregression of EBV with low reliability that are heavily influenced by information of relatives is expected to become more important in the near future. This follows directly from the observation that reference populations, which currently mainly include genotyped bulls with proofs based on daughter information (VanRaden et al., 2009; Lund et al., 2011), will be further expanded by adding genotyped cows, both for conventional (Mc Hugh et al., 2011; Lourenco et al., 2014; Koivula et al., 2016) and new breeding goal traits (Buch et al., 2012; Calus et al., 2013; Egger-Danner et al., 2014).

The objective of this paper, therefore, was to compare the performance of different deregression procedures for data that includes both cow and bull EBV, and to develop and test procedures to obtain the appropriate deregressed weights that should be used with those DRP. In this paper, we refer to a situation in which EBV of animals included in a reference population for genomic prediction need to be deregressed, to be used as pseudo-phenotypes in a subsequent genomic prediction model, but the described methods are also applicable for other situations.

MATERIALS AND METHODS

Deregression of EBV

Deregression is a procedure that computes pseudo-phenotypes, here referred to as DRP, that will yield a set of currently known EBV in a BLUP evaluation. The procedure used to compute DRP will be referred to as matrix deregression. Consider that we want to deregress a set of EBV of animals included in a reference population, and that this set of animals is a subset of a larger data set 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 and also included in the reference population, should be corrected for those 10 genotyped daughters, such that the DRP still includes the information of the 90 daughters that are not genotyped and therefore not included in the reference population. The same applies for the weight given to the DRP, as this should reflect only the information of the 90 daughters not included in the reference population. The matrix