De-Regressing MACE Versus Domestic EBV for Genomics

P.G. Sullivan, J. Jamrozik and G.J. Kistemaker
Canadian Dairy Network, 660 Speedvale Ave W, Suite 102, Guelph, Ontario N1K 1E5, Canada
Email: sullivan@cdn.ca

Abstract
De-regressed EBV are commonly used phenotypes in national genomic evaluation systems. This study compared REML estimates of variance for de-regressed MACE proofs of foreign sires on the Canadian scale, versus de-regressed national EBV of domestic sires. Variances for nearly all traits were higher for foreign than domestic sires. Ratios of SD for foreign relative to domestic sires, based on December 2010 Holstein data were; 1.05 and 1.07 for protein and fat yields, 1.17, 1.04 and 1.24 for mammary system, feet & legs and conformation, 1.50 for cow survival and 1.20 for cow non-return rate. Based on current data from December 2014, some of the more extreme ratios of SD were closer but still higher than 1. After applying a variance adjustment to de-regressed MACE proofs of foreign sires, genomic validation results improved for all traits. Slopes of regression, of the 2014 Canadian LPI index of trait EBV, on the 2010 genomic-enhanced parent averages (GPA), increased from 0.93 to 0.97, and biases of over-prediction for top young genomic bulls were accordingly reduced.

Key words: de-regression, genomic evaluation.

Introduction

De-regression is a computational technique to transform genetic evaluations derived from a complicated model, into pseudo-observations suitable for re-analysis under a simpler model. The simpler model can be modified in different ways, for example to include and combine information from multiple countries (Schaeffer *et al.*, 1996), or to switch from pedigree-based to genotype-based covariances among animals (Nejati-Javaremi *et al.*, 1997; VanRaden, 2008).

For most of the dairy cattle traits evaluated in Canada, multiple-trait traditional evaluations (EBV) are transformed into the de-regressed pseudo-observations (dEBV) used as input for single-trait genomic evaluation models. Effects of individual SNP on dEBV are estimated, then summed across SNP to get total direct genomic value (DGV) of an animal.

In its simplest form, de-regression is applying a variance adjustment to the difference between EBV of an animal and its parents (VanRaden *et al.*, 2009). If this variance scaling in de-regression is too high, an upward bias can be expected in the variance of resulting DGV. Initial (Kistemaker and Sullivan, 2010) and ongoing genomic validation test results for Canada have consistently detected inflated variances of DGV for many traits, and while effective measures have been taken to reduce the problem in genomic evaluations officially published in Canada, it is still unclear why the variance of DGV was inflated in the first place.

A potential explanation is that the dEBV derived from correlated traits have a different covariance structure among animals than is assumed under the single-trait genomic evaluation model. Information from related animals accumulates to a lesser degree (e.g. with maximum reliability equal to the genetic correlation squared) when all information is from a correlated trait. The single-trait genomic model does not limit accumulation of dEBV information from correlated traits of relatives, as it should, causing upward bias in both the reliability and variance of DGVs, even if methods are used to account for correlated information in the animal's own dEBV (Liu *et al.*, 2004; Sullivan *et al.*, 2006). The expected degree of bias is a function of the relative amount of information accumulated from correlated traits, the amount from dEBV of related animals, and the genetic correlations among all traits involved.

In the Canadian genomic evaluation system, dEBV of both domestic and foreign sires are used as input data. The MACE proofs of foreign sires are expressed on the same scale as Canadian EBV of domestic sires, but are based on only foreign, correlated-trait data. Thus, the expected degree of bias in variance of dEBV is higher for foreign sires. The goals of the present study were to estimate genetic variances for foreign versus domestic sires, and by re-scaling the foreign sires' dEBV, remove the relative difference in bias between domestic and foreign sires.

Materials and Methods

Holstein data for all traits officially evaluated for both domestic (EBV) and foreign (MACE) sires, and published in December 2014, were used for the present study. After removing four years of most recent phenotypic data, all traditional and genomic evaluation systems were re-run on the reduced data sets. The genomic evaluations were re-run a second time, after estimating variances and applying a variance adjustment, to the difference between dEBV and parent average, for foreign sires.

Proof De-regression

Since December 2012, the de-regressed proofs (dEBV) used for genomic evaluation at CDN have been computed with the approximation formula presented by VanRaden *et al.* (2009) for national evaluation, and by VanRaden and Sullivan (2010) for international evaluation:

$$dEBV = PA + (EBV - PA) / Rel(EBV|PA),$$

where PA is the parent average and Rel(EBV|PA) is the reliability of EBV, adjusted by removing the contribution of PA. For genomic evaluation, only EBV of sires were included in the phenotypic data used for de-regression, and the EBV of cows were not used. The dam's contribution to PA was determined solely from the EBV of male ancestors of the dam.

The term EBV-PA is an estimate of Mendelian Sampling (M), and while Rel(EBV-PA) is different than Rel(EBV|PA), these two reliabilities are very closely related for proven sires. The variance of dEBV is strongly influenced by the variance of estimated M and the estimates of reliability of M. The reliability is a direct function of prediction error variance (PEV(M)), which is nearly always approximated.

Variance Estimation

Sullivan (1999) showed that the EM-REML equation for estimating a (co)variance can be simplified to an average involving two terms; estimates and prediction error (co)variances (PEV) of Mendelian sampling (M). The formula is general for variances (i=j) and covariances (i≠j) between traits i and j, where for any group of q animals:

$$\hat{\sigma}_{ij} = \frac{1}{q}\sum_{k=1}^{q} \delta_k (\hat{M}_{ik}\hat{M}_{jk} + PEV(\hat{M}_{ik}, \hat{M}_{jk}))$$

The term δ is a function of known ancestry of the animal (e.g. δ =2 when sire and dam are known). The REML estimates of variance are thus determined solely by the variance and prediction error variance of M, as was basically also the case for variance of dEBV described above. If the estimates of Rel(M), or similarly Rel(EBV|PA), are inaccurate, then the de-regression of EBV will scale the dEBV incorrectly, hence biasing the variance of de-regressed proofs (V(dEBV)). Estimates of σ_i^2 from the dEBV can be used to detect when Rel(EBV|PA) is the incorrect term for scaling M, and to adjust the dEBV formula accordingly. This approach corrects both for errors in individual reliability approximations as well as for errors due to ignoring effects of correlated data on the prediction error covariance between animal and parents.

The REML methods used in the present study are also used by Interbull for sire variance estimation in routine MACE services,

and to estimate sire and cow variances by year for trend validation tests (Fikse *et al.*, 2003; Tyrisevä *et al.*, 2012).

Results and Discussion

Slopes of prediction, of current EBV on DGV estimated from 4-years truncated data, have an expected value of 1 in the absence of genomic pre-selection. These slopes have been much lower than 1 for DGV of many traits, however, so a reliability-weighted average of DGV and PA (called a genomic PA or GPA) is officially published for genomic selection of young bulls in Canada (Kistemaker and Sullivan, 2010). The GPA are highly correlated with, but have lower variance than DGV, and with GPA the slopes of prediction are much closer to 1.

The GPA have relatively less bias in variance, especially for traits affecting the LPI selection index used in Canada. This advantage for GPA over DGV was apparent both before (old) and after (new) adding a variance adjustment to the dEBV of foreign sires (Table 1). Slopes of prediction for DGV improved after adding variance adjustments, but continued to be lower than 1 for nearly all traits. For the LPI index, slopes of prediction improved for both domestic and foreign bulls. The best slopes of prediction for LPI were with GPA and the variance adjustment to dEBV of foreign sires. Prediction model R^2 values were also slightly higher, but in practical terms essentially the same as before. Thus in Canada, using a weighted average of DGV and PA is still preferred for genomic selection of young bulls.

Variance adjustments were generally larger for fertility and survival traits, which had and continue to have relatively low slopes of prediction. The scaling of DGV for these traits has been improved but is still to some degree incorrect. Effects of correlated data on M for domestic sires was not yet considered, nor the effects on covariances between M of related animals across the population. These factors are of interest but may be difficult to address.

Consistent with improved slopes of prediction, there were also reductions in over-prediction of top bulls, as expected. When selected by LPI using DGV of traits, the over-prediction for top young bulls dropped from .68 to .50, and using GPA from .45 to .29, standard units of LPI (Table 2). For domestic proven sires, over-prediction was very small in all cases. For foreign sires, the percentage reduction in bias, relative to young bulls, was higher for DGV (50% vs. 19%) and lower for GPA (16% vs. 36%). Variance adjusting de-regressed MACE proofs had a similar effect as averaging DGV and MACE results for foreign sires. In general, the most desirable patterns of lowest over-prediction of LPI plus greatest consistency among the different types of bulls, was with the current approach in Canada; blending DGV with traditional evaluations for all animals, to publish GEBV for domestic sires, GMACE for foreign sires and GPA for young bulls.

Computing time to estimate variances was very small, because reliabilities are used to approximate variance of M on an individual animal basis, thus avoiding the inversion of mixed-model equations that is usually required for REML. The evaluation systems at CDN have been updated, and since April 2015, now include a variance estimation and adjustment to dEBV of foreign sires with every evaluation.

The variance estimates and adjustments are expected to remain stable over time, and were very similar for all traits between December 2014 and April 2015 (Table 1). Any future changes to the estimates would be considered as important updates, to maintain consistency between the adjustments being used and the input data being adjusted. The estimated adjustment for truncated data (December 2010) were consistent but more variable than the estimates for current data. This was expected, partly because the truncated data sets were smaller, but more importantly because we did not have an optimal set of truncated-data MACE proofs. For proper alignment of MACE with the truncated-data national EBVs, all countries would need to submit 4-years

truncated national EBV to Interbull for a special truncated-data MACE evaluation run. For the present study, MACE proofs officially published 4 years ago were used instead, and these do not reflect any of the changes to national models at CDN, or by any other country over the past four years.

Conclusions

Approximations of reliability and distributional properties of MACE proofs are different than for domestic EBVs. Adding a variance adjustment, to the de-regressed MACE proofs used as input data, reduced the problem of inflated variances in genomic evaluation results, especially for the more problematic traits. Computational requirements are minor, which has allowed CDN to include a variance estimation and adjustment with every evaluation run, since the first official implementation in April 2015. Re-estimating variance adjustment factors with every run ensures an ongoing consistency between the variance adjustments being applied and the input data being adjusted.

References

Fikse, W.F. & Banos, G. 2001. Weighting factors of sire daughter information in international genetic evaluations. *J. Dairy Science 84,* 1759-1767.

Fikse, W.F., Klei, L., Liu, Z. & Sullivan, P.G. 2003. Procedure for validation of trends in genetic variance. *Interbull Bulletin 31,* 30-36.

Kistemaker, G.J. & Sullivan, P.G. 2010. Experiences in validating genomic evaluations. *Interbull Bulletin 40,* 235-239.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. Derivation and Calculation of Approximate Reliabilities and Daughter Yield-Deviations of a Random Regression Test-Day Model for Genetic Evaluation of Dairy Cattle. *J. Dairy Science 87,* 1896-1907.

Nejati-Javaremi, A., Smith, C. & Gibson, J. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. *J. Animal Science 75,* 1738-1745.

Schaeffer, L.R., Reents, R. & Jamrozik, J. 1996. Factors Influencing International Comparisons of Dairy Sires. *J. Dairy Science 79,* 1108–1116.

Sullivan, P.G. 1999. REML estimation of heterogeneous sire (co)variances for MACE. *Interbull Bulletin 22,* 146-148.

Sullivan, P.G., Liu, Z., Jakobsen, J.H. & Fikse, W.F. 2006. More on weighting factors for complicated models. *Interbull Bulletin 35,* 112-116.

Tyrisevä, A., Mäntysaari, E.A., Jakobsen, J., Aamand, G.P., Dürr, J., Fikse, W.F. & Lindauer, M.H. 2012. Validation of consistency of Mendelian sampling variance in national evaluation models. *Interbull Bulletin 46,* 97-102.

VanRaden, P. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Science 91,* 4414–4423.

VanRaden, P., Van Tasssell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J. & Schenkel, F. 2009. Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Science 92,* 16-24.

VanRaden, P.M. & Sullivan P.G. 2010. International genomic evaluation methods for dairy cattle. *Genetics Selection Evolution 42,* 7.

Table 1. Slopes of prediction of current (Dec 2014) EBV from genomic evaluations using truncated (Dec 2010) data, and REML estimates of variance for de-regressed proofs of foreign relative to domestic sires from different data sets.

		Slope 2010 DGV		Slope 2010 GPA		$\hat{\sigma}_{foreign}/\hat{\sigma}_{domesti}$		
Trait[z] (2014 EBV)		**Old**	**New**	**Old**	**New**	**Dec 2010**	**Dec 2014**	**Apr 2015**
Milk	MIL	0.91	0.94	1.06	1.08	1.05	1.02	0.98
Fat**	FAT	0.87	0.90	1.04	1.07	1.07	1.04	1.03
Protein**	PRO	0.85	0.88	0.99	1.02	1.05	1.02	0.96
Stature	STA	0.94	0.98	1.18	1.22	1.08	1.07	1.08
Chest Width	CWI	0.75	0.77	0.95	0.99	1.06	1.04	1.03
Body Depth	BDE	0.91	0.95	1.15	1.18	1.09	1.04	1.04
Angularity	ANG	0.77	0.79	0.94	0.96	1.01	1.03	1.03
Rump Angle	RAN	0.89	0.91	1.10	1.12	1.03	1.02	1.02
Pin Width	RWI	0.97	0.99	1.17	1.20	1.08	1.04	1.04
Rear Legs Side View	RLS	0.81	0.82	1.01	1.02	1.03	1.02	1.02
Rear Legs Rear View	RLR	0.76	0.81	0.96	1.01	1.10	1.06	1.03
Foot Angle	FAN	0.61	0.66	0.76	0.82	1.12	1.04	1.05
Fore Attachment	FUA	0.84	0.89	1.03	1.08	1.11	1.09	1.08
Rear Attach. Height	RUH	0.86	0.92	1.05	1.12	1.16	1.12	1.10
Median Suspensory	USU	0.79	0.86	0.94	1.00	1.12	1.08	1.05
Udder Depth	UDE	0.95	0.99	1.21	1.25	1.07	1.06	1.05
Fore Teat Placement	FTP	0.77	0.83	0.93	0.97	1.11	1.05	1.06
Teat Length	FTL	1.01	1.02	1.22	1.22	1.01	0.98	0.98
Rear Teat Placement	RTP	0.84	0.87	1.03	1.06	1.06	0.99	1.00
Conformation	OCS	0.74	0.82	0.90	0.97	1.24	1.26	1.25
Mammary System**	OUS	0.78	0.84	0.95	1.01	1.17	1.18	1.16
Feet & Legs**	OFL	0.76	0.79	0.92	0.96	1.04	0.99	0.99
Somatic Cell Score*	SCS	0.80	0.81	0.94	0.96	1.03	1.01	1.03
Direct Herd Life*	DLO	0.63	0.71	0.75	0.83	1.50	1.19	1.20
Sire CE in Heifers	DCE	0.77	0.77	0.90	0.92	0.96	0.92	0.91
CE in Heifers	MCE	0.71	0.75	0.85	0.90	1.17	1.08	1.08
Sire CSV in Heifers	DSB	0.69	0.72	0.82	0.87	0.90	0.83	0.84
CSV in Heifers	MSB	0.81	0.90	0.96	1.03	1.43	1.15	1.16
NRR in Heifers	HCO	0.60	0.62	0.78	0.80	1.19	1.03	1.02
NRR in Cows*	CC1	0.69	0.75	0.88	0.94	1.20	1.06	1.06
CTFS	CRC	0.55	0.60	0.69	0.76	1.15	1.02	1.01
FSTC in Cows*	CC2	0.71	0.75	0.89	0.94	1.13	1.06	1.05
Days Open	INT	0.73	0.76	0.90	0.94	1.09	1.02	1.02
Milking Speed	MSP	0.68	0.69	0.86	0.87	1.10	1.09	1.11
Milking Temperament	TEM	0.61	0.61	0.73	0.74	1.08	1.08	1.07
LPI domestic sires (EBV)		0.78	0.83	0.93	0.97			
LPI foreign sires (MACE)		0.75	0.80	0.86	0.90			

[z]Trait definition in Canada followed by the Interbull 3-letter trait code. Traits are divided by the Interbull trait groupings: production, conformation, udder health, longevity, calving, fertility and workability. CE is calving ease, CSV is calf survival, NRR is non-return rate, CTFS is calving to first service, FSTC is first service to conception.

**Very important and *important traits contributing to the 2014 LPI formula.

Table 2. Average over-prediction of December 2014 LPI index of trait EBV, for genomic evaluations from December 2010 data, before and after adding a variance adjustment to the dEBV of foreign bulls for all traits[z].

	No Adjustment		Variance Adjustment	
Bull and Proof Type	**All bulls**	**Top 100**	**All bulls**	**Top 100**
2010 DGV	**Average Difference (2010 DGV – 2014 EBV)**			
Domestic (DGV)	-0.02	0.12	0.00	0.06
Foreign (DGV)	0.23	0.30	0.10	0.15
Young (DGV)	0.29	0.68	0.21	0.50
2010 GEBV	**Average Difference (2010 GEBV – 2014 EBV)**			
Domestic (GEBV)	0.00	0.09	0.00	0.03
Foreign (GMACE)	0.23	0.32	0.21	0.27
Young (GPA)	0.37	0.45	0.31	0.29

[z]Results are presented in standardized units of LPI.