Model building pipeline to maximize the accuracy of breeding values prediction

J. Jenko[1*] and Ø. Nordbø[2]

[1] Geno SA, Storhamargata 44, 2317 Hamar, Norway; [2] Norsvin R&D, Storhamargata 44, 2317 Hamar, Norway; [*]janez.jenko@geno.no

Abstract

To maximize the accuracy of selection, models used for the prediction of breeding values need to correct phenotypes for all known important environmental effects. We have developed a pipeline where a fixed part of the model is built automatically, and the proposed models are validated in terms of accuracy and bias of their predictions. After the best model for a single trait is selected, genetic correlations are estimated between all the combination of traits and the optimal multitrait models are proposed. We have tested the model building pipeline on the development of single step genomic predictions for a conformation data set with repeated records on the Norwegian Red dairy cattle breed. The comparison between the single step genomic predictions using the old multitrait models and the new multitrait models averaged over nineteen traits showed a relative increase of 26% (from 0.62 to 0.78) in the accuracy of predictions.

Introduction

Accuracy of selection is among selection intensity, genetic variation, and generation interval one of the four components influencing genetic improvement of a population (Falconer, 1996). For a trait with a single record per individual the accuracy of prediction is a square root of heritability. When multiple measurements are taken on the animal for a single trait, the breeding value is predicted from the mean of these records. Repeated records enable estimation of resemblance between records due to the permanent environmental factors (Mrode, 2014). This is decreasing the variance within the individual which leads to an increase in the accuracy of prediction. The gain in accuracy of prediction is increasing with more records, but the marginal benefit in precision is decreasing.

Before variance component estimation and prediction of breeding values, phenotypes have to be corrected for all known important effects. The decision whether the effect is included into the model is often based on F-test where the simple and alternative models are compared. Other criteria like coefficient of determination and adjusted coefficient of determination can be used as well. Finally, models can be validated for their prediction accuracy.

Conformation traits have been part of the breeding goal from the very beginning of the breeding program for Norwegian Red dairy cattle in the nineteen sixties. The importance of conformation traits in breeding goal has increased from around 10% in the beginning to almost 30% in the 2020. Back in the history phenotypes for exterior traits in Norwegian Red dairy cattle were only collected on the first lactating cows. However, since 2014, repeated recording of conformation traits on cows in the first five lactations were collected and the starting point of the current work was to utilize the repeated records into the routine genomic evaluations. To increase the accuracy of predictions furthermore, new multitrait models will be built to utilize information from genetically correlated traits.

Materials & Methods

Data. The data comprised of twenty-four traits collected between 1988 and 2021. There were 646,291 records collected on 585,306 cows, from which 16,924 were genotyped. Only cows

in the first five lactations were recorded. At the time of recording the majority were in the first lactation (545,518) followed by the cows in the second (59,827), third (32,156), fourth (6006), and fifth (2784) lactation. The data were recorded on 22,036 herds. There were seven traits that were standardized since they have changed the scale of recording from 1-3 to 1-9 in 2014. To keep the phenotypic variance constant, phenotypes recorded in 2014 or later were standardized for heterogenous variance due to the lactation number. Except stature and supernumerary teats all the traits were also standardized for heterogenous variance due to the classifier and the year of taking the scoring.

Model development and estimation of variance components. To build the fixed part of the model and estimate the variance components, only data recorded after 2011 from herds with at least 100 recordings for stature were used. Furthermore, only data recorded between 14 and 305 days in lactation were kept. This gave 68,227 records collected from 440 herds. Fixed part of the model was built with an inductive approach. Here, the effects were included in the fixed part of the model only if they have passed the inclusion criterion for a specific scenario. Three different inclusion criteria were set for the four optimization parameters. The four optimization parameters were i) R square or the proportion of the variance explained ii) Adjusted R square, which in comparison to R square penalizes for addition of extra variables iii) P value that compares the reference and alternative model using the F-test and iv) Gain by loss which is the relative change in residual sum of squares between the reference and alternative model per change in degrees of freedom between the reference and alternative model. Each of the optimization criteria was tested with three inclusion criteria. R square and Adjusted R square with 0.1, 0.01, and 0.001, P value with 0.01, $1*10^{-5}$, and $1*10^{-10}$, and Gain by loss with 0.01, 0.001, and $1*10^{-4}$. If the tested effect passed the inclusion criterion for the optimization parameter and its effect was the biggest among all the effects, it was included in the model. By default, each model had already included the fixed effects of herd-five-years and year-month of calving. The inclusion of the other twelve candidate effects (number of days since calving, barn type, recording version, feeding system, milking system, classifier, lactation number, time since milking within time from calving, heterozygosity, year of recording, hours since milking, and age at calving) was tested in a sequential order so that the most important effect was included first. If a linear covariate was included in the model, the squared covariate was also tested. The same was done for cubic covariate if the squared covariate was included in the model. Once none of the effect was passing the inclusion criterion, the model building was finished. Finally, two and three-way interactions were tested and included in the model if passing the inclusion criterion for the optimization parameter. With this approach twelve models were developed for each of the twenty-four traits and variance components were estimated using DMU 6 (Madsen and Jensen, 2013).

Cross-validation of alternative models. The proposed models were then tested for the accuracy of prediction using a cross-validation approach. Here 269,114 records collected on 226,485 cows after the year 2005 were used. First estimated breeding values (EBV) using a single step genomic approach on this complete data set were calculated with the DMU 6 (Legarra et al., 2014; Madsen and Jensen, 2013). Then, phenotypes of 6841 cows that were genotyped and were not included in the variance component estimation were masked and EBV using the single step genomic approach were calculated again. Finally, yield deviations (YD) were estimated from the complete data set and correlated with the EBV from the data set with masked phenotypes. This was done on the animals with masked phenotypes only. The model with the highest accuracy and the lowest bias was then selected for the construction of multitrait models.

Estimation of variance-covariance parameters and construction of multitrait models. The variance-covariance parameters were estimated on the bivariate models across all the twenty-four traits. Traits with high genetic correlations were then combined into six multitrait models. Some traits were included in more than one multitrait model to increase the accuracy of prediction for the traits in a particular multitrait model.

Cross-validation of final multitrait models. The accuracy of predictions of the newly developed multitrait models were finally compared with the accuracy of predictions from the old multitrait models. This was done with masking of the phenotypes from the youngest 5000 genotyped cows. The YD calculated from the complete data set and EBV from the masked data set were calculated using the MiX99 software (MiX99 Development Team, 2019). The obtained YD and EBV were correlated (Cor) and corrected for the heritability to estimate the accuracy of prediction (Acc) using the following equation:

$$Acc=Cor(YD,EBV)/(n /(n +(\sigma_e/\sigma_a)))^{0.5} \quad (1)$$

where n is the mean number of observations for each animal with masked phenotype, σ_e is the error variance and σ_a is the additive variance. Error and additive variances estimated from the new models were used when estimating the accuracy for both old and new models.

Results

The heritability and prediction accuracy from old and new multitrait models are presented in Table 1. Out of twenty-four traits breeding values for nineteen traits were predicted with both old and new models.

Table 1. Prediction accuracy for the conformation traits of Norwegian Red dairy cattle with previous and current multitrait models (old/new multitrait model)

Trait	Previous heritability	Current heritability	Accuracy previous model	Accuracy current model
Stature (1./1.)	0.41	0.58	0.72	0.89
Rump angle (1./2.)	0.19	0.28	0.64	0.87
Body depth (1./3.)	0.14	0.19	0.53	0.77
Chest width (1./3.)	0.11	0.15	0.53	0.74
Rump width (1./3.)	0.22	0.30	0.61	0.80
Rear legs rear view (2./1.)	0.08	0.14	0.45	0.60
Rear legs side view (2./1.)	0.09	0.17	0.62	0.70
Foot angle (2./1.)	0.09	0.14	0.64	0.71
Udder balance (3./4.)	0.11	0.17	0.68	0.85
Rear udder width (3./4.)	0.13	0.17	0.71	0.81
Rear udder height (3./4.)	0.11	0.16	0.70	0.83
Rear teat placement (4./5.)	0.22	0.26	0.69	0.83
Fore udder attachment (4./4.)	0.12	0.17	0.51	0.63
Front teat placement (4./5.)	0.18	0.27	0.70	0.79
Central ligament (5./5.)	0.10	0.14	0.54	0.64
Supernumerary teat (5./6.)	0.21	0.17	0.56	0.95
Teat thickness (5./6.)	0.20	0.29	0.64	0.82
Teat length (5./6.)	0.27	0.41	0.71	0.86
Udder depth (5./4.)	0.25	0.31	0.68	0.83

Most of the traits have moderate heritability spanning from 0.14 to 0.31. Only teat length and stature have high heritability of 0.41 and 0.58, respectively. The prediction accuracy has increased for all the traits when the new multitrait models were used. Average accuracy when the old multitrait models were used for predictions was 0.62 where a minimum accuracy of 0.45 was achieved for rear legs rear view and a maximum accuracy of 0.72 was achieved for stature. Average accuracy when the new multitrait models were used for predictions was 0.79 where a minimum accuracy of 0.60 was achieved for rear legs rear view and a maximum accuracy of 0.95 was achieved for supernumerary teat. This means that on average the new multitrait models increase the prediction accuracy for 26% (from 0.62 to 0.79). Correlations between the predicted breeding values from old and new models were from 0.78 to 0.97.

Discussion

Three main reasons for the increase in the accuracy of predictions with the new models exist. The first one is the inclusion of phenotypes from multiple lactations instead of only the first one. As mentioned earlier this enables estimation of resemblance between records due to the permanent environmental factors and decreases the variance within the individual which leads to an increase in the accuracy of predictions (Mrode, 2014). The second reason is a better correction of phenotypes for all known environmental effects. Old models were all correcting traits for the same fixed effects. Model development pipeline corrected phenotypes only for the important known effects and different fixed effects were included in the different models. Using the cross-validations we have compared the proposed models and selected the model that is maximising prediction accuracy and minimising bias. The third reason is the optimal construction of multitrait models. The old models were built so that trait describing the same part of the cow exterior were combined into a multitrait model. The new multitrait models were constructed so that traits with high genetic correlations were combined into a single multitrait model.

The new multitrait models for the prediction of breeding values for exterior traits were sent for Interbull test in August 2021. After their approval they were successfully implemented in the routine genetic evaluations for the Norwegian Red dairy cattle in December 2021. Given the improved accuracy of genomic predictions with new, updated models, we expect to see an increased genetic gain in the coming years.

Acknowledgement

We want to thank the Research Council of Norway for funding this research through the project 309611, “Large scale single step genomic selection in practice” and under the BIONÆR program, project number 282252.

References

Falconer D.S., and Mackay T.F.C. (1996) Introduction to quantitative genetics. Fourth edition. Longman, Harlow, UK.

Legarra A., Christensen O.F., Aguilar I., and Misztal I. (2014) Livestock Science 166:54-65. https://doi.org/10.1016/j.livsci.2014.04.029

Madsen P., and Jensen J. (2013) DMU Ver. 6, rel. 5.2. Available at: https://www.researchgate.net/publication/291444592_A_user's_guide_to_DMU.ls

MiX99 Development Team (2019) MiX99: A software package for solving large mixed model equations. Release XI/2019. Available at: http://www.luke.fi/mix99

Mrode R.A. (2014) Linear models for the prediction of animals breeding values. 3rd edition. CABI, Wallingford, UK.