Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 284-286

ROUTINE DIAGNOSIS OF GENETIC EVALUATIONS AND AUTO STANDARDISATION TO NON-LINEAR COST FUNCTION

X. Zhang, F. Hely, K. Stachowicz and P. Amer

AbacusBio, Level 5, 333 Princes Street Dunedin 9016, New Zealand

SUMMARY

Routine genetic or genomic evaluations are often conducted multiple times a year. When there is a new distribution of estimated breeding values, non-linear cost functions that have been developed will often need to be adjusted or recalculated. The purpose of this paper is to present routine methods to 1) diagnose and display whether the genetic evaluations have deviated from the past course, and 2) standardise the genetic evaluations of a new population so that the previously developed non-linear cost function can be applied to the standardised evaluations. Automation of this routine could expedite the decision on whether the new data still suits the already-developed non-linear cost function and avoid potential errors created by manually recalibrating the non-linear cost functions incorrectly.

INTRODUCTION

The index value, or cost function, of a measured trait with linear economic weights is simply the product of its economic value in trait units and its estimated breeding value (EBV; Falconer and Mackay 1996; Hazel 1943). For traits with non-linear economic weights, the shape of the non-linear cost function (NLCF) and its position relative to the range of estimated breeding values adds a complexity into the index formulation (Quinton *et al* 2019). For example, a cost, C($/trait unit) for a trait, could be connected to EBV via a conditional NLCF as below

$$C = \begin{cases} \beta_0 + \beta_1(g - T), & g \leq T \\ \beta_0 + \beta_1(g - T) + \beta_2(g - T)^\gamma, & g > T \end{cases} \quad (1)$$

where β_0, β_1 and β_2 are intercept, linear and nonlinear (curve) coefficients, respectively, g_i is the EBV, T is the threshold in g scale where the economic value changes from linear to non-linear, and γ is the exponent. Values of C, g and T are known. The NLCF is only nonlinear when $g > T$. β_0, β_1, β_2 and γ determine the shape of the NLCF. β_0 and β_1 are easy to solve by taking two pairs of g and C observations into the linear part, i.e. $g \leq T$. For β_2 and γ, given two sets of observations of g and C, we can now form two equations

$$C_1 = \beta_0 + \beta_1(g_1 - T) + \beta_2(g_1 - T)^\gamma$$

and

$$C_2 = \beta_0 + \beta_1(g_2 - T) + \beta_2(g_2 - T)^\gamma$$

Since the value of β_0 merely shifts the curve up or down but does not change the shape, we can simplify the above by assuming the cost is 0 at T, then $\beta_0 = 0$, and the above becomes

$$C_1 - \beta_1(g_1 - T) = \beta_2(g_1 - T)^\gamma$$

and

$$C_2 - \beta_1(g_2 - T) = \beta_2(g_2 - T)^\gamma$$

We can solve for γ by taking the ratio of the above two equations

$$\frac{C_1 - \beta_1(g_1 - T)}{C_2 - \beta_1(g_2 - T)} = \left(\frac{g_1 - T}{g_2 - T}\right)^\gamma$$

Then taking the log of both sides and rearranging we have

$$\gamma = \frac{log\big(C_1 - \beta_1(g_1 - T)\big) - log\big(C_1 - \beta_1(g_2 - T)\big)}{log(g_1 - T) - log(g_2 - T)}$$

Then we can solve β_2 using either of the initial equations, e.g.

$$\beta_2 = \frac{C_1 - \beta_1(g_1 - T)}{(g_1 - T)^\gamma}$$

When the distribution of estimated breeding values changes due to a new evaluation method, a new genetic base, pre-selection, or the trait is evaluated in a different environment for a new cohort of candidates, the original coefficients including T_i may not hold anymore. Thus, the NLCF needs to be adjusted or recalculated (Burndon 1990; Quinton *et al.* 2019). This is required to ensure that the index still penalises or incentivises the right individuals in the right degree based on where they are on the estimated breeding value scale.

A set of economic weights is usually calculated to last for multiple years unless there is a major change in the genetic/genomic evaluation results described above, when an update of the index may be triggered (Cole and VanRaden 2018). However, there is often a lack of communication between genetic evaluation providers and those developing or maintaining the index regarding whether and how the evaluation results have changed between runs, sometimes due to data security or IP protection. It would be useful to make a diagnosis tool as part of the evaluation pipeline, allowing index experts to detect major changes in the evaluation so that they can decide the next appropriate step, e.g. to change the NLCF, standardise the EBV, or to understand more about the evaluation results before making a decision. The purpose of this paper is to present a routine method to diagnose changes in EBV and to standardise these when the latest EBV doesn't deviate from the past distribution.

ROUTINE DESCRIPTION

The routine will consist of two parts. The diagnostics part and the auto-standardisation part.

Diagnosis. The first diagnostic is to verify whether the evaluation distribution changed. The EBV of every trait in each evaluation run will go through the routine for a series of descriptive analyses including summary statistics and tests for a normal distribution. The results will be recorded in a report, including the number of animals, mean, median, SD, min, max, outlier values, histograms, results from Shapiro-Wilk or Kolmogorov-Smirnov test of normality, and ranking of animals.

The second diagnostic is to check if there is a genetic base shift. When we have accumulated multiple evaluation results, then for a set of reference animals that have EBV across all evaluations, we can run correlation, pairwise t-test (when EBV are normally distributed), pairwise F-test and ANOVA, or Kruskal-Wallis H test (when using genotypes) to check if the latest EBV have deviated from the original EBV for which the NLCF was created. If genetic progress in the EBV is expected, then we can run a linear regression on the EBV means across evaluations, leaving out the latest mean, then compare the observed latest mean with the predicted mean, to see if the observation is within the expected range, say, within 2 SD of the predicted mean. The formula to calculate the standard deviation of a predicted value, $\hat{y}_i$, as a linear regression is:

$$SD(\hat{y}_i) = s\sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$ is the standard error of $\hat{y}$, n is the number of observations, x_i is the i-th predictor, and $\bar{x}$ is the mean of all predictors. Then we repeat the analysis for EBV for every trait. Lastly, we can check the rank correlation of animals overlapped across evaluations, especially base or proven animals.

The third diagnostic is to check if the EBV threshold, T, has changed. The relationship between the non-linear cost changing point and the phenotype threshold usually stays stable, whereas the corresponding T to the phenotype threshold can change across evaluations. When the phenotype

threshold is available to both the initial and the latest datasets, we can run a linear regression between phenotype and EBV of the initial data and apply the same coefficients to the latest data to estimate T in the latest dataset. If phenotypes of at least one dataset are not available, we can still run a linear regression between the EBV values of the 2 sets of data to estimate new T from this regression. In addition, we can check the proportion of individuals with EBV beyond the threshold where the cost is non-linear. The threshold will be recorded from the latest NLCF, along with other parameters.

If there has been a distribution change, genetic base shift or a change of threshold is detected, the routine programme will show a warning and flag the out-of-range EBV, so that index experts can decide if they want to communicate with evaluation experts about the cause and impact of the changes, and if they need to re-design the NLCF. These diagnostics can run every time a new set of EBV is available, and the results will be recorded for the comparison.

Auto standardisation. When an index review is needed, commonly every 3-5 years, the historical diagnosis results can be used to help with decision making. If the vector of the latest set of EBV, **g**, has passed the aforementioned diagnostics, e.g. their distribution has not changed significantly, or the evaluation expert has confirmed they are homogenous, then we can simply standardise them to the initial scale, $\mathbf{g}_0$, for which the NLCF was created.

$$\mathbf{g}_{new} = \frac{(\mathbf{g} - \mu_{\mathbf{g}}) \cdot \sigma_{\mathbf{g}_0}}{\sigma_{\mathbf{g}}} + \mu_{\mathbf{g}_0} \quad (2)$$

where g_{new} is the vector of standardised latest EBV; $\mu_{\mathbf{g}}$ and $\sigma_{\mathbf{g}}$ are the mean and standard deviation of the latest EBV, and $\mu_{\mathbf{g}_0}$and $\sigma_{\mathbf{g}_0}$are the mean and standard deviation of EBV at the initial evaluation where the NLCF was made. The standardisation formula parameters will be recorded for future reference.

Auto fitting of NLCF. Finally, after diagnosis and auto standardisation of EBV, we fit $\mathbf{g}_{new}$ into equation (1) to obtain the corresponding new cost for **g** and record the new NLCF.

CONCLUSION

We have described a routine diagnostics pipeline to first validate new trait EBV against historical trait EBV, then use an automation pipeline to perform a series of calculations, ultimately generating new non-linear cost values. By streamlining and simplifying the selection index review process, this pipeline provides timely warnings to index experts when data deviates from expectations, aiding them in making informed decisions about necessary adjustments to the NLCF.

REFERENCES

Burndon R.D. (1990) *Theor. Appl. Genet.***79**: 65.

Cole J.B. and VanRaden P.M. (2018) *J. Dairy. Sci.* **101**: 3686.

Falconer D.S. and Mackay T.F.C. (1996) 'Introduction to Quantitative Genetics' 4th Edition Longman Group Ltd.

Hazel L.N. (1943) *Genetics* **28**: 476.

Quinton C.D., Amer P.R., Byrne T.J., Archer J.A., Santos B. and Hely F. (2019) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **23**: 47.