Validation of Genomic Evaluation Models

H. Jorjani[*]

Introduction

Validation of animal breeding genetic evaluation models, especially national genetic evaluation models, is a research area that has received less attention than it deserves. The reason, at least partly, is because the usual genetic evaluation model uses a mix of some statistical and some genetic theories. As Kempthorne (1976) noted, an animal (or plant) improvement plan can be conducted "*without an atom of formal Mendelism*". A purely statistically constructed model can use the rich repertoire of validation tools from statistical literature for validation of the model (e.g. Burnham and Anderson, 2004). However, animal breeders, with their strong background in quantitative genetics, have the habit of incorporating some (quantitative) genetic model elements in the evaluation model as well. One example is the incorporation of the infinitesimal model in the genetic evaluation models. Use of data from single nucleotide polymorphism (SNP) does not change the importance of such genetic models.

Validation of a genetic evaluation model not only should consider statistical issues, it must also consider quantitative genetics aspects of model building and their validation. The purpose of this study is to establish a connection between statistical and quantitative genetic aspects of genetic evaluation models.

Method

Validation under no-selection scenario. One suggestion for validation of genomic evaluation models (Mäntysaari, VanRaden and Liu, 2010) is comparison of the following two Models:

[1] $DYD = b_0 + b_1 EBV_r + e$
[2] $DYD = b_0 + b_1 GEBV_r + e$

where *DYD* is the daughter yield deviation for a bull calculated from a full data set, EBV_r and $GEBV_r$ are the conventional breeding value and genomically enhanced breeding value of the same bull from a reduced data set (when the bull has no daughter yet), b_0 and b_1 are the intercept and the slope of the linear regression and *e* is the residual error. The rationale behind the suggestion is the assumption that $GEBV_r$ is based on more data than EBV_r, *i.e.* the SNP genotype data, and therefore Model [2] must have a higher coefficient of determination, R^2.

[*] Interbull Centre, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-75007, Uppsala, Sweden

It is easy to see that EBV_r, $GEBV_r$ and *DYD* are predictions of an animal's genetic merit (expressed as EBV or PTA) based on three different sources of information, *i.e.* animal's parents (*P*), animal's own genotype (*G*), and animal's offspring (*O*). These three pieces of information represent the progression of acquiring more data during the lifetime of an animal. Before conception the only available information is that of parents. Later, at some point of time after the conception, the information from genotype becomes available. Finally, the information from animal's offspring becomes available. In other words, an animal's predicted genetic merit (*PGM*) can be calculated from part of the data (coming from parents and/or genotype) or from the full data. Re-writing Models [1] and [2], we obtain

[3] $PGM_{P+G+O} = b_0 + b_1 PGM_P + e$
[4] $PGM_{P+G+O} = b_0 + b_1 PGM_{P+G} + e$

However, there are two other intermediate Models

[5] $PGM_{P+O} = b_0 + b_1 PGM_P + e$
[6] $PGM_{P+G} = b_0 + b_1 PGM_P + e$

By examination of Models [3] and [4] it is easier to see that we are indeed dealing with comparison of predictions at different times, *i.e.* validation of the model through comparing the predictions from part data and full data. This is commonly known as *Method R*, as suggested by Reverter et al. (1994a and b). It is interesting to ascertain if the expected value of b_1, $Exp(b_1)$, equals to unity or not, and also how it is affected by selection. These two points have been discussed before (e.g. Reverter *et al.*, 1994a,b and Thompson, 2001). Here, rather than statistical reasons, the emphasis is more on the quantitative genetic aspects.

From Model [5] it can be seen that the regression of full data (PGM_{P+O}) on part data (PGM_P) equals the $Cov(P,(P+O)) / Var(P) = Cov(P,P) / Var(P) = \sigma^2_a / \sigma^2_a = 1.0$. The strange, nonetheless the obvious, point is the fact that $Cov(P,O) = 0.0$, because this term depends on the random segregation of alleles from parent to offspring, *i.e.* the Mendelian sampling terms, which is independent of parents' PGM and is distributed as

[7] $\gamma \sim N(0, \sigma^2_a /2 * (1- (F_s + F_d)/2))$.

Thompson (2001) used the following argument to prove that $Exp(b_1) = 1.0$. Consider the following model

[8] $y = X\beta + u + e$

where $Var(u) = A \sigma^2_a$. For distinct generations the *A* matrix can be partitioned into TDT' with *T* relating animals of different generations to each other. Then $T\gamma = u$ and $Var(\gamma) = D$. Elements of the diagonal matrix of *D* are shown in Equation [7]. Then, following Reverter (1994a), predictions based on full data, $u'_1 = (u'_{10}, u'_{11})$, and predictions based on part data, $u'_0 = (u'_{00}, u'_{01})$, have an expected regression of 1.0, *i.e.* $Cov(u_1, u_0) / Var(u_0) = 1.0$, or $E(u'_{new} A^{-1} u_{old}) / E(u'_{old} A^{-1} u_{old}) = 1.0$.

The same argument can be used to prove that $Exp(b_1)$ in Model [6] is also equal to unity. The regression of full data (PGM_{P+G}) on part data (PGM_P) equals the $Cov(P,(P+G)) / Var(P) = Cov(P,P) / Var(P) = \sigma^2_a / \sigma^2_a = 1.0$. The reason is that $Cov(P,G) = 0.0$, again because this term depends on the random segregation of alleles from parent to offspring, *i.e.* the Mendelian sampling terms, which is independent of parents' *PGM*. To appreciate this point, consider a group of full sib animals from a pair of parents. The parental genotype and therefore the parents' *PGM* is the same for all of them, however, they have all different genotypes.

This type of argument can be use to prove that the $Exp(b_1)$ in Models [1] to [6], or any other combination of *P*, *G* and *O* to estimate *PGM*, is equal to 1.0.

Validation under selection scenario. Now, that it has been established that the expected regression of full data on part data, as in the $Exp(b_1)$ validation Models [1] to [6] is 1.0, let's examine the effect of selection on regression.

More than 50 years ago Henderson and his colleagues (Henderson *et al.*, 1959) showed that selection on the independent variable, *i.e.* part data in the right hand side of Models [1] to [6], has no effect on the regression equation. However, selection on the dependent variable, *i.e.* full data in the left hand side of Models [1] to [6], leads to bias. Therefore, significant departures of estimated values of b_0 from 1.0 indicate bias in the genetic evaluation process.

Discussion

Any statistical analysis model needs to be validated by rigorous statistical tools and modified until a model that provides the best fit to the data in found. However, the degree of determination, R^2, alone is not enough for validation. An over complicated (over parameterized) model may have high R^2 value, but it would probably have little predictive value. To increase the value of validation tests, there is also a need to check the conformity of the model results to the quantitative genetic assumptions that were employed in the model construction. One such assumption in animal breeding genetic evaluation is the segregation of a large number of loci. As mentioned above, selection on dependent variable leads to bias. Please notice that "*selection*" in this connect means "*non-random pattern of missing data*", and should not be confused with phenotypic selection. In this respect, the nature of bias is prevention of individuals with certain genotypes from having offspring data, and hence being excluded from the full data, which it its turn means excluding certain genotypes from full data. This is equivalent to violating random segregation of alleles. In other words, "*non-random pattern of missing data*" of this sort creates a covariance between two sources of data, *i.e. G* and *O* leading to $Cov(G, O) \neq 0.0$.

Many concepts in quantitative genetics theory have a dependence on the random segregation of allele, and re-creation of Mendelian sampling variance in each generation. If random segregation of alleles is distorted, then many quantitative genetic concepts, e.g. breeding value, lose their meaning. Therefore, it is imperative that all genotyped animals are included in the evaluation.

Conclusion

Some general methods of validation of genetic and genomic evaluation models are presented. In these models, a parameter of interest is the comparison of predictions based on full data and part data. Based on statistical and quantitative genetic arguments, the regression of these two predictions is expected to be equal to 1.0. Departure from the expectation of 1.0 indicates bias. In genomic evaluation, a possible source of bias is likely to be non-random pattern of missing (genotypic) data, which has effect equivalent to distortion of random segregation of alleles.

References

Burnham, K.P. and Anderson, D.R. (2002). Model Selection and Multimodel Inference. Springer-Verlag, new York.

Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.M. (1959) *Biometrics*, 15: 192-218.

Kempthorne, O. (1976). Proceedings of the International Conference on Quantitative Genetics. Ames, Iowa, Iowa State University Press. pp 719-760.

Mäntysaari, E., VanRaden, P. and Liu, Z. (2010) Personal Communication.

Reverter, A., Golden, B., Bourdon, R.M. and Brinks, J.S. (1994a) J. Animal Science. 72: 34-37.

Reverter, A., Golden, B., Bourdon, R.M. and Brinks, J.S. (1994b) J. Animal Science. 72: 2247-2253.

Thompson, R. (2001*). Livesock Prod. Sci*., 72:129–134.