An Overview of Validation Issues in National Genetic Evaluation Systems (N-GES)

Hossein Jorjani
[1]*Interbull Centre, Box 7023, S - 750 07, Uppsala, Sweden,*

Abstract

Current attempts at validation of national genetic evaluation systems lack a satisfactory level of transparency. Further, those internationally agreed validation methods that are in effect rely heavily on validation of the genetic theory underlying national evaluation model. The present paper is an attempt to distinguish among different levels of validation, to briefly review some trend in validation studies, and a start point for a house-keeping check-list of validation.

Introduction

Validation of national genetic evaluation systems (N-GES) is a seemingly difficult issue, partly because the issue at hand is a complicated one. However, before one gets into the technicalities of validation an urgent question is: "What is the target of validation"? There is a need to distinguish three different targets:

a) Genetic theory underlying evaluation model;
b) Evaluation model used in N-GES; and
c) N-GES.

Validation of the genetic theory underlying N-GES as such is a legitimate scientific question and it is of interest to all quantitative geneticists and animals breeders. However, this is a question of primarily academic interest that needs to be addressed by utilization of the data accrued through appropriately designed experiments and analyzed accordingly. There is also a long tradition on this in our scientific community (see for example: Clayton & Robertson, 1957). To test, say, the infinitesimal model, and its associated assumptions such as constancy of variances, is of little interest unless it has a strong bearing on published values of EBV/PTA of animals. Let's be clear about the fact that in the Interbull community we are committed to the industry, especially farmers. With these points in mind, it seems that validation of genetic theory underlying genetic evaluation model is a too narrow target and a slightly misguided one.

In contrast to the (a) above, validation of N-GES in its entirety is a too wide of a target to be focused on by the Interbull Community alone. It is worth mentioning that "GES is meant to include all aspects from population structure and data collection to publication of results. Each and every statistical treatment of the data that has a genetic-breeding motivation or justification is an integrated part of GES" (Jorjani et al., 2001). This is a target more suitable to be handled by Interbull's parent organization, ICAR and its working groups, especially the working group on "Quality Assurance". There are also organizational aspects related to the validation of N-GES that need to be addressed by governmental agencies, breeding organizations, and so on.

Based on the preceding arguments, we, at the Interbull Centre and the Interbull community, need to concentrate on the second target, i.e. validation of evaluation model used in N-GES.
This target, as alluded to in this paper, encompasses previous validation / verification methods and tools.

Validation of evaluation model

Before embarking on a discussion of validation of evaluation models may I point to the fact that animal breeders commonly have multiple roles to play: the roles of a statistician, a geneticist, a market planner, and an extension specialist, not to mention other minor roles. This strange combination has had some unfortunate consequences. Our

role as geneticists has prevented us from exploring a variety of models, our role as extension specialists has forced us to adopt models that are easy to explain, and so on. What seems to be accommodating for the others, and therefore suffering, is our role as statisticians, with the final result that statistics role has been reduced to an under-exploited tool. Let me bore the reader with a long quote from Oscar Kempthorne (1976):

"*It is interesting that one can obtain some results without invoking Mendelism at all, but merely use purely statistical ideas of correlation and regression. One can go further, I believe. The whole area of selection can be approximated by purely statistical ideas of correlation and regression. The ideas of Mendelism merge with these ideas, as Fisher showed (more or less), and the fact that the theory does not need Mendelism in some respects, and one can almost say, does not use Mendelism is, I think, a reason for it having a moderate degree of robustness in relation to assumptions. Apart from a difficulty I shall mention later, one could proceed as follows.*

Let there be a population; let rules of forming mating couples be defined in terms of metric traits of individuals and/or in terms of relationship; let there be selection of individuals on the basis of metric traits or metric traits of related individuals; and finally let the offspring be measured. Then without an atom of formal Mendelism and with a large data set, the joint distribution of offspring and parents can be determined. One can examine this distribution and determine a prediction equation, which one can then apply for a few generations. The only flies in the ointment for this proposal are that every covariance have to be determined from data and not inferred from, say, a coefficient of relationship and heritability, and large data sets would be needed to control sampling error.

So one could have a completely empirical selection procedure and a purely empirical process of obtaining a prediction of the result of continued selection. I suggest that this type of thinking should not be dismissed as a cranky idea. The reason that some predictions of the results of selection theory seem to work is that they are based on a process rather close to what I have sketched."

Having this in mind there are **at least three approaches** to validation of evaluation models:

1- The old statistical way
2- The old heuristic way
3- The right way

Validation of evaluation model: The old statistical way

In the days before our current highly sophisticated computers became available a statistician would start with a careful consideration of the problem, study design and protocol for data collection. Then, our imaginary and thorough statistician would think of a model for evaluation. Consequently, based on Fisher's likelihood theory the model structure is assumed to be known (i.e. to be correct and true) and that only the parameters in that structural model are to be estimated. Finally, one needs to determine the precision of the estimated parameters.

Thorough validation of the evaluation method then would involve several steps:

1- Assess conformity to the protocol for data collection by:

   1.1 Check the data for logical inconsistencies by checking uniqueness of ID's, cross-checking of the life history milestones (dates for birth, insemination, calving) in parents and offspring, and so on.
   1.2 Explore the data exhaustively by tables, graphs, plots of various descriptive statistics (including 1^{st} to 4^{th} moments), properly cross-classified for all of the explanatory variables considered in the model.
   1.3 Examine conformity of the data to the assumptions of the model, such as distributional properties (e.g. normality) or homogeneity of variances.

2- Investigate the fit of the model to the data by a variety of statistical tools that, depending on the statistician's school of thought, are available for this purpose. Examples are RMSE, R^2, deviance, or formal χ^2 goodness-of-fit, etc.

3- Check the results against theoretical expectations that one has from the biology/genetics of the issue under examination or the statistical properties and consequences of the analysis model. Equivalent to the checks for data one has to:

   3.1 Explore the results exhaustively by tables, graphs, plots of various descriptive statistics, properly cross-classified for all of the explanatory variables considered in the model.
   3.2 Examine the solutions for main effect(s) and residuals in the same exhaustive way described in 1.2 and 3.1 above.

What has been described above is believed to be in common practice in all national genetic evaluation centers around the world. An examination of the fields for "criteria for inclusion of the data" and "method of validation" in the countries' fact sheets in the latest Interbull survey (Jorjani, 2000; IBB 24) prove that extensive validatory measures are employed by different countries. However, unfortunately there is no transparent record of what is performed in one country available to the rest of the world. More importantly, it is not known how and when the results lead to changes of the evaluation models.

In order to increase transparency, and also to provide a better opportunity for a mutual learning among evaluation centers, we need to set up a check-list of what verification / validation tools should be employed and try to keep it as updated as possible.

Validation of evaluation model: The old heuristic way

Presently, the most widely used methods of validation (Boichard et al., 1995) result from a heuristic utilization of the knowledge of genetic structure of dairy cattle populations, statistical properties of the evaluation models and the genetic theory underlying these models. The problem with the methods currently sanctioned by Interbull (Boichard et al., 1995) and many others proposed validation / verification methods (e.g. Thompson, 2001; Klei et al., 2002) is that they focus heavily on examination of means and variances of estimated breeding values and Mendelian sampling terms. It is very easy to see that national evaluation models are capable of yielding more information than just breeding values and Mendelian sampling terms. An example of other quantities is daughter yield deviations (DYD) that is used in Interbull Validation Method II (Boichard et al., 1995) and the method proposed by Liu et al. (2003). However, national evaluation models are way too under-utilized. Maybe an example helps in shedding some light on this matter.

The example to be discussed is taken from evaluation model in the USA for two reasons. First, through the paper by VanRaden & Wiggans (1991) we probably know more about this model than any other country's model. Second, the paper by VanRaden & Wiggans (1991) has been very influential, both because of its educational value and also because of its extension to other situations (e.g. Mrode & Swanson, 1999, 2002, VanRaden, 2001; and Liu et al., 2003). The evaluation model described by VanRaden & Wiggans (1991) is as follows:

$$y = Mm + Za + ZA_g g + Pp + Cc + e$$

where

y = Standardized milk, fat, or protein yield
m = Vector of effects for management group
a = Vector of effects for random portion of additive genetic merit
g = Vector of effects for unknown-parent group
p = Vector of effects for permanent environment
c = Vector of effects for herd-sire interaction
M = Incidence matrix for management group
Z = Incidence matrix for random portion of additive genetic merit
ZA_g = Incidence matrix for unknown-parent group (A_g = Related animals to the unknown-ancestor group)
P = Incidence matrix for permanent environment
C = Incidence matrix for herd-sire interaction
e = Error

Based on the above model and the usual assumptions about variances, the MME are as follows.

$$\begin{bmatrix} M'R^{-1}M & & & & \\ Z'R^{-1}M & Z'R^{-1}Z + A^{-1}k_a & & & symmetric \\ 0 & -A_g'A^{-1}k_a & A_g'A^{-1}A_gk_a & & \\ P'R^{-1}M & P'R^{-1}Z & 0 & P'R^{-1}P + Ik_p & \\ C'R^{-1}M & C'R^{-1}Z & 0 & C'R^{-1}P & C'R^{-1}C + Ik_c \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{u} \\ \hat{g} \\ \hat{p} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} M'R^{-1}y \\ Z'R^{-1}y \\ 0 \\ P'R^{-1}y \\ C'R^{-1}y \end{bmatrix}$$

VanRaden & Wiggans (1991) working on the second row of this system of equations (Equation for u) manage to derive an array of useful measurements, one of which is DYD, that are being used in many areas, from understanding of the results to the validation of evaluation models.

An intuitive question is that would it be possible to utilize this and other parts of the same equation system to arrive at other measurements? Whether this turns out to be labeled with the derogatory name of "unsuccessful fishing expedition" or the flattering name of "new heuristic validation methods" remains to be seen. But, without being exhaustive, let's examine some other possibilities.

Equation for m (management groups)

Can we examine the vector of management group solutions:

- To identify an appropriate distribution to model management groups?
- To identify cases of preferential treatment, very bad management, outbreak of disease, sub-clinical disease, etc. by studying outliers?
- To study the correlation of management group effects & their sizes to better understand heterogeneous variances?
- If management group solutions prove to be heteroscedastic, can we assume that pre-adjustments have not been effective?

Equation for g (unknown parent groups)

Is it so unrealistic to assume that:

- Solutions should be constant across evaluations?
- Elements of this must be significantly different from each other?
- There should be a correspondence between this and the "a" value in a conventional conversion equation?
- For countries with deep pedigrees it should be possible to treat some animals as unknown and compare their solutions as members of the unknown parent groups with their EBV as known animals?

Equation for p (permanent environment effects)

Permanent environment (PE) effect, as its name suggests, is an environmental effect. However, it must be remembered that because it is specific to the animal, it is partly environmental and partly genetic, in the sense that it is related to the genotype of the animal at the time of conception, but in contrast to the additive genetic value, it is not heritable / transmittable. In a way, as it is customary in simulation studies, it should be treated just like breeding value of the animal, that is, to acknowledge that it takes a random value, constant for the whole lifetime of the animal. Therefore, in a limited way, and for specific purposes, all the assumptions imposed on breeding values can be imposed on permanent environment effect.

Further, a comparison of some new attempts to model quantitative genetic variation (San Cristobal-Gaudy et al., 1998; Hill, 2002; and Sorensen & Waagepetersen, 2003) suggests that what we used to define as permanent environment effect may indeed be related to a genetic ability to influence residual variance. In this sense, there might be a confounding and correspondence between PE and the slope of a norm of reaction curve, which undoubtedly is under genetic control (Kolmodin et al., 2002, 2003).

Conditional on the discussion above, estimates of the PE must remain constant across evaluations, be constant over lifetime, follow the same distributional properties of BV/EBV. Alternatively and conservatively, consecutive estimates of PE must have high auto-correlation among themselves and the regression of new estimate on old estimate should be equal to unity.

Equation for c (sire-herd interaction)

Inclusion of a sire-herd interaction term in the model actually suggests that we explicitly have singled out one environmental descriptor (explanatory variable) as having interaction with a major random effect, and implicitly that such

interactions for other environmental descriptors are negligible. Studies conducted by the Wisconsin group (e.g. Weigel & Rekaya, 2000; Zwald et al., 2001) and at the Interbull Centre (Fikse et al., 2001) on Holstein and Guernsey breeds, respectively, suggest that even other environmental descriptors may have non-trivial effects.

The purpose of the above questions and arguments is to point to the fact that a) Evaluation model are modeling more than just breeding values, and b) We may be surprised to see other validation tools emerge if we look beyond breeding values.

Validation of evaluation model: The right way

We have seen in the past how the choice of national evaluation models has been influenced by seemingly small, yet ingenious methodological achievements (e.g. Henderson, 1976), by technological progress made in computer performance (e.g. adoption of iterative methods and even random regression models), and by market forces.

In my opinion we are on the verge of a new change of paradigm in choice of evaluation models in N-GES. The change that we are about to experience is a move away from single model evaluations to multi-model evaluations. The prerequisites for this change have been accumulated gradually during the past two decades or so, and now there are both methodological and technological means to extend the application of multi-model evaluations from academic research to real life situations.

Multi-model evaluation is a logical extension of Fisherian method of data analysis alluded to before. In the same way that estimation of parameters has been considered as an optimization problem, one can extend the optimization principle to cover even the process of model selection. Different steps of a multi-model evaluation can be summarized as:

a) Formulation of a set of models;
b) Model selection;
c) Estimation of parameters; and
d) Obtaining measurement of precision (including the uncertainty about the model).

As an analogy one can compare the first two steps of this process to a classical multiple regression analysis, in which a set of a priori explanatory variables are examined to arrive at a smaller set that provide optimum levels of bias and variance.

Multi-model evaluation is still an area of active research in both frequentist and Bayesian schools of thought (see for example Burnham & Anderson, 2002; and Sorensen & Gianola, 2002, respectively). Despite their differences, both frequentist and Bayesian statisticians advocate multi-model approachs and not surprisingly they arrive at a large number of similar recommendations.

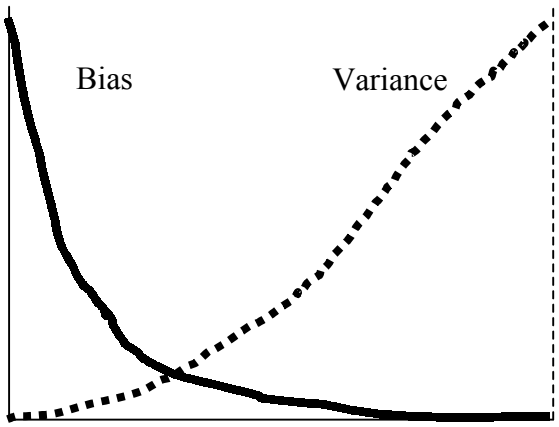
Formulation of a set of models

In a frequentist setting formulation of a set of models can start with a global model that includes all relevant effects that our scientific understanding of the subject deems relevant. One should investigate the fit of the global model to the data and proceed with the analysis of the data, and formulation of more parsimonious models, only if the global model provides an acceptable fit to the data and narrow confidence intervals.

Alternatively, in a Bayesian setting one can start with a very simple model and build up more complex models based on the examination of, say, posterior predictive distributions.

Model selection

Obviously as the number of parameters in the model increases, i.e. the model gets more complex, the fit is improved, bias decreases and variance increases. The following schematic diagram is used to ease the discussions.



Conceptually it may be possible to find a model that lies at the intersection of the two curves shown in this figure, but this happens very seldom. A model lying to the left of the intersection is an underfitted model and a model to the right of the intersection an overfitted model. The balance between these two is a dilemma to an animal breeder because of the multiple roles that we have. As extension specialists we would like to see as parsimonious a model as possible. Then we would be able to explain the results very easily to farmers. However, for prediction of "future" data we would like to have a model with low prediction error variance.

In any case, modern statisticians, in both frequentist and Bayesian camps, as judged by their advocacy of different "information criterion=IC" prefer a more parsimonious model (principle of parsimony being defined (Box & Jenkins, 1976) as a model with the smallest possible number of parameter for adequate representation of the data.).

There are a number of methods available for selection of a model, especially from a Bayesian perspective (by using the Bayes Factor, and all the IC methods that might or might not rely on it). However, from a frequentist point of view there is no theoretical foundation for the notion of hypothesis testing with a fixed α level for model selection (Burnham & Anderson, 2002).

Cross validation has widely been studied and suggested as a basis for model selection (e.g. Thompson, 2001). For this purpose, the data are divided into two parts. The first part is used for model fitting; and the second (which may contain only one data point) is used for validation.

Model averaging

As noted by Box (1976) all models are wrong, but some are useful. Therefore, if none of the models is "true", then the question arises that why should we choose a model, rather than average a small number of more plausible models.

In Bayesian model averaging (BMA) uncertainties pertinent to the relative plausibility of each model are taken into account. It can be shown that, with the exception of the very unlikely event of having the "true" model among the models, BMA has better predictive performance than any of the models alone (Sorensen and Gianola, 2002).

It is worth mentioning that the number of all possible (conjectural) models to be considered in BMA may be very large. However, one can use the available scientific knowledge to weed out some of the possible (but purely speculative) combinations of model building elements (for more statistical aspects of how to reduce the number of models see, for example, Hoeting et al., 1999).

Conclusions

In order to bring our validation tool box into order we need to think of a comprehensive check-list for all "purely statistical" aspects of data analysis. For the genetically motivated "heuristic" validation method we need to expand our horizon and even consider other things that are modeled in the national genetic evaluation models. And finally we need to be more open minded to the newly developed computing intensive mulimodel evaluations and their imbedded validation mechanisms.

Acknowledgement

I would like to acknowledge fruitful discussions with my team-mates in Interbull Centre and with Professor Daniel Gianola, university of Wisconsin. On a personal level, I should say that after each round of discussion with Daniel and his refusal of my ideas, the only thing that I could think of was that "The longest distance between any two points, is the shortcut".

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci. 78,* 431-437.

Box, G.E.P. 1976. Science and statistics. *Journal of the American Statistical Association 71,* 791-799.

Box, G.E.P. & Jenkins, G.M. 1970. *Time series analysis: forecasting and control.* Holden-Day, London.

Burnham, K.P. & Anderson, D.R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach.* Springer-Verlag, New York.

Clayton, G.A., Morris, J.A. & Robertson, A. 1957. An experimental check on quantitative genetical theory. I. Short term response to selection. *Journal of Genetics 55,* 131-151.

Clayton, G.A. & Robertson, A. 1957. An experimental check on quantitative genetical theory. II. The long-term effects of selection. *Journal of Genetics 55,* 152-170.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics 32,* 69-74.

Hill, W.G. 2002. Direct effect of selection on phenotypic variability of quantitative traits. *7th WCGALP*. Communication No. 19-02.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. 1999. *Bayesian Model Averaging: A tutorial. Statistical Science 14,* 382-417. (The original article contains a number of typographical errors. For a correct version see Jennifer Hoeting's web site at: www.stat.colostate.edu/~jah)

Jorjani, H. 2000. National Genetic Evaluation Programmes for Dairy Production Traits Practiced in Interbull Member Countries 1999-2000. *Interbull Bulletin 24,* 111 pp.

Jorjani, H., Philipsson, J. & Mocquot, J.-C. 2001. Interbull Guidelines for national and international genetic evaluation systems in dairy cattle with focus on production traits. *Interbull Bulletin 28,* 30 pp.

Kempthorne, O. 1977. Status of quantitative genetics theory. In: Pollak, E., Kempthorne, O. & Bailey, T.B. (eds). *Proceedings of the International Conference on Quantitative Genetics*. Ames, Iowa, Iowa State University Press. pp 719-760.

Klei, L., Mark, T., Fikse, F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Interbull Bulletin 29,* 178-182.

Mrode, R.A. & Swanson, G.J.T. 1999. Simplified equations for evaluations of bulls in the Interbull international evaluation system. *Livest. Prod. Sci. 61,* 43-52.

Mrode, R.A. & Swanson, G.J.T. 1999. The calculation of cow and daughter yield deviations and partitioning of genetic evaluations when using a random regression model. *7th WCGALP*. Communication No. 01-04.

Thompson, R. 2001. Statistical validation of genetic models. *Livest. Prod. Sci. 72,* 129-134.

SanCristobal-Gaudy, M., Elsen, J.-M., Bodin, L. & Chevalet, C. 1998. Prediction of the response to a selection for canalization of a continuous trait in animal breeding. *Genet. Sel. Evol. 30,* 423-451.

Sorensen, D. & Gianola, D. 2002. *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer-Verlag, New York.

VanRaden, P.M. 2001. Methods to combine estimated breeding values obtained from separate sources. *J. Dairy Sci. 84 (E. Suppl.),* E47-E55.

VanRaden, P.M. & Wiggans, G. 1991. Derivation, calculation, and use of national animal model information. *J. Dairy Sci. 74,* 2737-2746.

Weigel, K.A. & Rekaya, R. 2000. A Multiple-Trait Herd Cluster Model for International Dairy Sire Evaluation. *J. Dairy Sci. 83,* 815-821.

Zwald, N.R., Weigel, K.A., Fikse, W.F. & Rekaya, R. 2001. Characterization of Dairy Production Systems in Countries that Participate in the International Bull Evaluation Service. *J. Dairy Sci. 84,* 2530-2534.

Zwald, N.R., Weigel, K.A., Fikse, W.F. & Rekaya, R. 2003. Identification of Factors That Cause Genotype by Environment Interaction Between Herds of Holstein Cattle in Seventeen Countries. *J. Dairy Sci. 86,* 1009-1018.