Vereinigte Informationssysteme
Tierhaltung w.V.

INTERBULL OPEN MEETING
SEPTEMBER 7-8, 1995
PRUHONICE, PRAGUE, CZECH REPUBLIC

Analysis of Residuals to Evaluate the Fit of a Multiple Trait Model

F. REINHARDT
VIT, Heideweg 1, D-27283 Verden/Aller, Germany
E-mail: freinhar@rzv-server3.vit.de

INTRODUCTION

In the field of breeding value estimation the generation interval of new methods became shorter and shorter in the last decades. During the development of new and better evaluation systems we need statistical methods and parameters to compare reasonable models. After implementation of a new evaluation system checks of the actually used model are necessary time by time in order to recognize changes in precorrection factors over time periods, suboptimal definition of subclasses within fixed effects, changes in a priori assumed (co)variance components or other potential sources of biased estimates.

Much attention has been recently payed to check and to correct genetic trend resulting from IAM breeding value estimation in order to guarantee a correct ranking of old and young bulls in a given breeding population (Bonaiti et al, 1994). Reasons for an over- or underestimation of genetic trend are in most cases suboptimal precorrection or expansion factors. The theoretical background behind that methods is to test the goodness of evaluation systems by the stability of estimates with increasing information over time periods. Sigurdsson and Arnason (1995) showed in simulation studies that different estimated genetic trends are resulting from Repeatability and Multitrait models. The fact of correctly estimated genetic trends in national breeding value evaluation systems becomes even more important when the resulting sire proofs are used as base of a joint international ranking.

OBJECTIVE

The objectives of this report are to recall and to discuss the residual analysis, a well known and generally applicable method to check the goodness of fit of statistical models. In breeding value estimation we have to fit the models to very heterogeneous field data. This is a typical example to apply a residual analysis in order to get additional inferences about the quality of the used estimation procedure.

METHOD AND DATA

In Germany a Multi-Lactation Animal Model is used for breeding value estimation of milk production traits. All Holstein Friesian and Red and White dairy cattle are included into one joint evaluation system.

$$y = Xb + Zu + e$$

$$y_{ijm} = hys_{im} + a_{jm} + e_{ijm}$$

y_{ijm} yield in (part)lactation (m) of cow (j) in Herd*Year*Season (i)
hys_{im} fixed effect of Herd*Year*Season (i) having at least 6 contemporary comparisons, otherwise fixed effect of Region*Herdclass*Year*Season (i) in (part)lactation (m)
a_{jm} random additive genetic effect (BV) of animal (j) in (part)lactation (m)
e_{ijm} random residual error

The observerd yields (y) are precorrected for calving age and calving intervall and standardardized for within herd*year deviation depending on the average production level of the herd in that year. To get more (part)lactations compared directly within Herd*Year*Season classes only three calving seasons are defined.

During the last iteration round all relevant additional information beside breeding values is calculated - like average DYD or average mating niveau of sires . In the same step residuals for all animals having own performances can be calculated by a few additional program statements.

$$e = y - Xb + Zu$$

$$e_{ijm} > 0 \implies y_{ijm} > b_{im} + u_{jm}$$

$$e_{ijm} < 0 \implies y_{ijm} < b_{im} + u_{jm}$$

By definition the sum of residuals in a statistical model is zero, and this should hold likewise within subcells of fixed effects. We are assuming one set of additive genetic (co)variances and one set of residual (co)variances in the estimation model. Therefore we suppose that the ratio between genetic variation and residual variation is similar for all observed yields - over different genetic subpopulations and environmental subcells. This means that the variance ratio of estimated breeding values and residuals in all these subclasses should be similar.

Multi lactation models are considering selection effects between lactations (Henderson, 1984), if selection takes place on the analyzed trait. That means superior cows have a greater chance to survive to later lactations and the estimation model recognizes that the genetic competition in later lactations is higher as in the first part of the first lactation. This causes in average psitiv residuals in second and third lactation yields.

The observed residuals and the squared residuals can be analyzed by Least Squares technics. A simple calculation of average residuals within particular subclasses of effects is not possible, since data are not balanced and there is no random distribution of yields over all different fixed effects. In the least square model effects already considered in the breeding value estimation

may be defined, and further effects neglected in the original model. Mainly effects belonging to precorrection and prestandardization of observations are recommended to be cecked by an residual analysis. To detect a suboptimal modelling a smaller classification of groups should be defined within fixed effects. E.g. in the case of wide seasonal classes in the estimation model potential systematic differences in residuals between month can be analyzed. A reasonable model to analyse residuals out of the MME of the above mentioned breeding value estimation model is:

e = breed (HF, Red and White)
+ region
+ herdlevel (within herdclasses and within herds)
+ calving year
+ calving month
+ combination of (part)lactations
+ e'

Relevant statistic tests for significance of additional effects on residuals are the F-test and Student's t-test.

The large number of cows and (part)lactations included into the actual breeding value estimation, differences in production systems between regions, different breeds and variation in production levels of subpopulations within breed require a continuous checking of the statistical model. Is it appropriate to assume the same model for the whole dataset, and to suppose one set of a priori variance covariance components for all production levels and breeds? Especially when we put together the breeding value estimation for dairy cattle of the western part with the eastern part of Germany, we had to investigate such questions on the base of the results out of a first joint test evaluation.

RESULTS

By the above mentioned statistical model less than 1 % of the variation in residuals can be explained. Only the effect of calving month was significant ($p < 0.05$). The residuals in the months April and September differ significantly from zero. When we changed our model last year to the variable system of comparisons within **region*herdclass*year*season** or **herd*year*season**, we reduced the number of calving seasons in order to compare more cows directly within herds. April and September changed into the "summer" calving season. The residual analysis shows that we are making actually a compromise between best correction of calving seasons and comparing a lot of cows within herd. Direct within herd comparison are definitly preferable, since significant smaller residuals can be found in these environmental groups compared to region*herdclass*year*season groups .

There are no significant additional residual effects between different breeds, herdlevels, calving years and combinations of (part)lactations.

Using the absolute value of the residual, | e |, as dependent variable 2 % of the variation is caused by the residual model. Breed, herdlevel and calving year are significant effects. This means, there are different rest variances in a lot of groups of these effects. The residual variance increases systematically with higher herd production level and over calving years. In regions

with higher average production level variation in residuals is enlarged. Therefore it seems to be a scale effect. Using the same statistical model to analyze the variation of breeding values a similar tendency was found. The ratio between variation of breeding values and variation of residuals shows only a small increase over herdlevels but not over calving years. This ratio differs in a narrow range between breeds and regions. These results point out, that we should take into consideration different a priori (co)variance matrices (additive genetic and residual) in the MME.

CONCLUSION

The residual analysis is a generally applicable method, from which we get inferences about the goodnes of fit of a statistical model - the fixed and the random part. In both initially mentioned cases, the development of new methods and the checking of an actually running method in breeding value estimation, the residual analysis is a valuable method to find the weak points of a given model. The residual analysis is a useful tool to

- detect individual data errors,
- check precorrection factors,
- check fixed effects already included in the model,
- check fixed effects actually not considered in the model,
- check wether the used model is fitted equally over the total data base, and to
- check for heterogeneous variance covariance structure.

Further theoratical work and practical investigations are needed to clarify the expected average residuals of selected animals in an IAM and in a Multitrait model.

REFERENCES

Bonaiti, B., D. Boichard, A. Barbat, and S. Mattalia. 1994. Three methods to validate the estimation of genetic trend in dairy cattle. Proceeding of the Open Session of the INTERBULL Annual Meeting, Ottawa, August 5-6, 1994.

Henderson, C. R. 1984. Application of Linear Models in Animal Breeding. University of Guelph. ISBN 0-88955-030-1, Chapter 13.

Sidurdsson, A. and Th. Arnason. 1995. Predicting Genetic Trend by Uni- and Multitrait Models. Acta Agric. Scand. Sect. A, Animal Sci. 45, 1-10.