

POTENTIAL IMPROVEMENTS IN ANIMAL MODEL EVALUATION SYSTEMS

P.M. VANRADEN

Animal Improvement Programs Laboratory
Agriculture Research Service, United States Department of Agriculture
Beltsville, Maryland 20705, USA

SUMMARY

Animal model evaluation systems are replacing sire model evaluation systems as the basis of ranking animals. Mixed model techniques underlying these evaluations provide a powerful tool for further improvement of predictions. Improvements in the data to be analyzed, the models and algorithms used in the analysis, and the delivery of results to the industry all are possible. Inclusion of more data for more traits allows more animals to be compared and each animal's total genetic merit to be predicted more accurately. Evaluation models and variance assumptions should match actual distribution of records as closely as possible, but records themselves often can be adjusted or transformed to meet assumptions of a simpler model. Effects usually excluded from animal models that may deserve inclusion are inbreeding, heterosis, other nonadditive genetic effects, effects of individual genes, and genotype-by-environment interactions. Variances and covariances required in evaluations usually are not known with certainty and are difficult to estimate without selection bias from the large, unbalanced data sets available. Nonlinear mixed model equations offer improved predictions if data are not distributed normally. Enhanced models often demand enhanced algorithms or more powerful computers. Results delivered to the industry must be timely, well documented, and easily explained to those actually making selection decisions if they are to have maximum impact.

INTRODUCTION

Genetic evaluations help breeders choose genes for the next generation. Animal models predict merit of each animal's genes from the animal's own records and records of its relatives. Relatives have genes identical by descent (Wright, 1922), and information from all relatives is utilized by including an additive relationship matrix, which assumes that many genes each have small effects on the traits being evaluated. Each animal's genes are a random sample of its parents' genes (except for mutation, new gene insertion, or cytoplasmic inheritance), and genetic merits of animals covary in direct proportion to percentage of genes identical by descent.

Statistical foundations on which animal model evaluations are built are the mixed model (Henderson, 1974), best linear unbiased prediction (Henderson, 1973), and Bayes theorem (Gianola and Fernando, 1986). With these basic tools, animal breeders need only state their assumptions about means, variances, and distributions of records, and rankings of animals follow automatically provided that data are collected, programs are written, and sufficient computer resources are available. Because of the extreme size of the data sets and populations evaluated, compromises between accuracy and cost have been necessary. Fairly simple mixed models and approximate methods of the past have produced reasonable accuracy and substantial genetic gains. In the future, improved evaluation systems may yield higher accuracy and more progress with little extra cost.

IMPROVED DATA

Modern breeding programs rely on routine genetic evaluations computed from high quality data and parentage information for large numbers of animals. Improvements in quantity and quality of data collected allow more animals to be considered in selection decisions and each animal's evaluation to be more accurate, thus increasing genetic gain. In the future, centralized breeding

schemes; direct insertion of genes with large, known effects; or marketing of genetically identical animals could reduce dependence on large sets of field data. Even then, field testing of these improved varieties and discovery of offspring in the field that receive superior combinations of naturally occurring genes still may offer large benefits.

Data for genetic evaluations can be improved by inexpensive procedures such as applying transformations or adjustments to existing variables and combining existing data sets or by expensive procedures such as collecting data for additional traits or additional data for the same traits. Data can also be improved by redefining the way traits are measured. Extending lactation records to 305 days for all cows, including those dried off before 305 days, increased both repeatability and heritability (Norman *et al.*, 1985). Linear scoring (Thompson *et al.*, 1983) rather than descriptive, categorical scoring for conformation traits has allowed simpler models to produce more accurate evaluations.

Many organizations may collect data separately for the various traits that contribute to total merit, but joint analysis of these data may be required to account properly for effects of selection on correlated traits. Even when different traits reside in the same data base, single-trait rather than multitrait analysis often has been justified because of expense and complexity of multitrait evaluation or because unavailability of accurate estimates of covariances. Multitrait evaluation methods are well developed, and their use surely will increase as computer capacity allows.

International exchange of genetic material requires that foreign and domestic animals be compared fairly. Foreign animals often are selected based on data not included in the domestic data set. Inclusion of foreign breeding value estimates as an additional data source for domestic evaluations or combining data across countries are two solutions to this problem. Evaluations that do not combine data or use outside evaluations still could easily be made comparable by using standard units of expression and similar computing procedures and adjusting to an internationally recommended base.

Crossbreeding and resulting heterosis offer large short-term gains for many traits, and commercial populations often are crossbreeds. Long-term genetic gains possible from selection within purebred populations likely are inferior to genetic gains achievable by selecting within crossbred populations because one breed may not contain all the beneficial alleles available in other populations. Data from all breeds, including crossbred animals, could be evaluated as one data set to allow the best parents and combinations of parents to be chosen regardless of breed.

Incoming data should be monitored regularly for quality and agreement with the model. New technologies such as bovine somatotropin may generate responses large enough to invalidate previously well researched results such as 305-day projection factors and appropriate lactation measures. Performance enhancing products provide another tool useful to individual breeders to create high, misleading evaluations on their own prechosen animals. Nearly all genetic evaluation models assume that genetic and environmental effects are uncorrelated, but designed, randomized distribution of family members to many herds or regions may be the only way to ensure this.

IMPROVED MODELS

Accounting for means

Environmental effects are estimated from records of animals with the same or similar genetic merit performing in different environments. Large environmental effects such as age, season, and times milked daily often are not included in models; instead, adjustment factors are used to alter data. These adjustments often are multiplicative to help stabilize variance as well as mean. Including these factors directly in evaluation models might be preferred, but models then may have to account for unequal variances.

Genetic merit of an animal is predicted from its own and its relatives' records compared with records of animals with different genes performing in the same or similar environments. Similar environments usually are defined as contemporary groups or management groups that consist of animals of the same age in the same herd at the same time. Environment within a herd may change gradually across time, but contemporary grouping assumes changes occur abruptly only at time boundaries. Time series models may be more realistic but are more difficult and may not justify the added expense (Chauhan and Thompson, 1986). If group effects or group sizes are small, treatment of contemporary group effects as random rather than fixed may be justified (Chauhan, 1987).

Inbreeding affects many economically important traits, but inbreeding coefficients have not been included in models. Regressions on inbreeding coefficients are simple to include, but this requires computing an inbreeding coefficient for each animal with data. To obtain each animal's inbreeding coefficient, construct a list including the animal and its ancestors and then apply the tabular method (Henderson, 1976) to this short list. Simplicity, small memory requirement, and speed make this method of computing and including inbreeding coefficients attractive even for extremely large populations.

Crossbred populations have been more difficult to evaluate than purebreds because of larger nonadditive genetic effects among breeds than within breeds and difficulty of assigning crossbred animals to categories if various combinations of several different breeds are present. Including heterosis in the model and tracing pedigrees of all crossbred animals back to purebred ancestors should overcome these difficulties. The relationship matrix then would account properly for additive differences, and grouping would be required only for unknown parents (Westell *et al.*, 1988) within each pure breed.

Accounting for variances

Additive relationship matrices are easy to include in models because their inverses are extremely sparse and can be constructed from simple pedigree files (Henderson, 1976). An algorithm appropriate for noninbred populations is fast and often is applied to populations with inbreeding. Accounting for inbreeding while constructing inverses is nearly as fast once inbreeding coefficients are obtained by the method outlined above. Relationship matrices for populations with crossbred animals are needed, and these should assume higher genetic variation within crossbred populations than within purebreds if heterosis is modeled as a reduction of inbreeding.

Multitrait models allow data for one trait to act as additional data for all correlated traits. Information from correlated traits can be included with little expense if all traits are measured on each animal, models for each trait are identical, and no repeated records are present so that the only random effects are genetic and error (Thompson, 1979). Multitrait equations if some data are missing or if models differ for different traits are more difficult to set up and require more memory but are affordable for many models (Quaas and Pollak, 1980).

Genes may be expressed differently in different environments. Sire-by-herd or sire-by-region interactions have been studied and included in sire models (Bertrand *et al.*, 1987), and relationships among these interactions also can be included (Foulley and Henderson, 1989). For animal models, genotype-by-herd or genotype-by-region interactions are more appealing. To include these, give each animal an interaction effect for each herd or region in which it has progeny or records, and link interactions within each herd or region by the within-herd or within-region relationship matrix. Including interactions in evaluation models limits predicted merit and reliability of animals with data in one or few environments and provides an incentive to breeders to test genetic material in several environments.

Nonadditive genetic effects, major genes, genes located on the sex chromosomes or mitochondria, and imprinted genes all may influence performance but have not been included in routine genetic evaluation models. Recent discoveries of rapid inverses for dominance (Hoeschele and VanRaden, 1990), additive-by-additive (VanRaden and Hoeschele, 1990), and X-linked relationship matrices may make these effects nearly as easy to include in models as additive effects. Additive predictions that result then may be more accurate, and predictions of nonadditive effects would be useful in mating programs.

New reproductive techniques often create demands for better models. Widespread use of artificial insemination made sire models the method of choice because most relationships were through sires. Embryo transfer dramatically increased progeny numbers for some females and created a need for animal model techniques. In the future, cloning will require that one genetic effect be estimated per clone group (Kennedy and Schaeffer, 1989) rather than the current practice of treating identical animals as full sibs, and inclusion of nonadditive effects will become more imperative. If genes are inserted, they will act as mutations (Hoeschele, 1990), which also currently are not included in models.

Assumptions regarding variance parameters can be refined in several ways. First and later lactation performances often are assumed to depend on a single genetic effect but could be modeled with separate, correlated genetic effects. Heritabilities, repeatabilities, variances, and covariances may differ across time, from breed to breed, and from herd to herd but often are assigned a single value for an entire data set. Obtaining variance ratios and covariances needed to construct mixed model equations usually is more difficult than solving resulting equations. Estimation by restricted maximum likelihood (REML) often requires solving mixed model equations repeatedly, and this is the easy part computationally. The difficult part is obtaining traces of inverses and calculating likelihoods if number of animals is large, and this currently limits REML estimation with the animal model.

Accounting for distribution

If data are distributed nonnormally, observed traits often are nonlinear functions of breeding values. In such situations, best predictors of breeding values are nonlinear rather than linear functions of data. Nonlinear evaluation techniques are available for many nonnormally distributed traits, including binomial traits such as liveability (Foulley and Gianola, 1984), ordered categorical traits such as calving difficulty (Gianola and Foulley, 1983; Harville and Mee, 1984), hazard function traits such as longevity (Smith and Quaas, 1984; Ducrocq *et al.*, 1988), and production traits with a mixture distribution caused by an unknown major genotype (Hoeschele, 1988).

Estimates of unknown effects (such as breeding values) are those values of unknowns that maximize joint density of data and random effects. This joint density is the conditional density of data given random effects multiplied by the distribution function of random effects, which is almost always normal because, for example, polygenic inheritance is assumed. Estimates of unknown effects are obtained by solving mixed model equations. Equations are linear if conditional distribution of data is normal (Henderson, 1973) and nonlinear if distribution is nonnormal. Nonlinear mixed model equations must be solved repeatedly because elements of the coefficient matrix and right-hand sides are functions of the unknowns being estimated.

Equivalent estimates can be obtained in a Bayesian framework by maximizing joint conditional density of unknown effects given the data, where all unknown effects are treated as random. With normality, estimates of breeding values that maximize this joint density are also the mean of the conditional distribution of breeding values given the data, an optimum selection rule (Fernando and Gianola, 1986). With nonnormality, selection on joint modal rather than marginal mean values may give nearly maximum progress for sire models with many progeny and large contemporary groups; performance of the joint mode with animal models and for animals with less information is less well known and could be poor.

Production traits often are distributed normally or near normally, and linear models have been sufficient for these. Recent interest in health and fitness traits may stimulate more breeders to learn nonlinear methods. Those familiar with REML estimation likely have already used nonlinear mixed model equations, where elements of the coefficient matrix are functions of variance ratios, which in turn are functions of the random effects being estimated. Estimation of breeding values and variance components for nonnormally distributed traits is analogous to REML. Nonlinear techniques often are affordable and are used routinely for large data sets, including evaluation of calving ease of United States Holsteins (Djemali *et al.*, 1987). Further advances in nonlinear theory, computer algorithms, and ease of explanation could lead to increasing application of these techniques in the future.

IMPROVED ALGORITHMS

Large data sets and advanced statistical methods used in genetic evaluations often exceed capabilities of available software packages. Specialized algorithms for individual problems often have been required; however, when possible, programs should deal effectively with many different data sets and models. With availability of more generalized algorithms, programs, and packages (Groeneveld and Kovac, 1989), animal breeders soon may be able to spend less time writing and debugging their own programs and more time actually analyzing data and investigating genetic principles.

Genetic evaluation systems routinely reprocess large data sets as data for new animals arrive. Algorithms that allow rapid incorporation of new data without reprocessing all previous data might be preferred. Use of converged solutions from previous evaluations as starting values can save time when beginning iteration again, particularly if convergence is slow (Quaas and Pollak, 1980). Evaluation systems that minimize number of steps required to produce new evaluations from new data are simpler to use and maintain, and evaluations then may be easier to explain to the public.

Computer hardware evolves rapidly, and algorithms appropriate for today's machines may not make best use of future machines. Nevertheless, algorithms that reduce instructions and memory requirements by orders of magnitude are likely appropriate for many machines. Exact solutions to some problems still are not possible, and approximations are required (for example, in computing reliabilities for each animal). Nevertheless, rapid advances in vectorization and parallelization may allow today's approximations to be replaced by better approximations or exact answers in the future.

IMPROVED DELIVERY

For genetic evaluations to have greatest impact, they must be easily understood, accepted, and used by those making breeding decisions. The animal model's basic feature that information from all relatives contributes to each animal's evaluation can be explained easily (Wiggans and VanRaden, 1989). More complicated models may demand more creative thinking to translate matrix algebra formulas into common language. In free market economies, breeding goals cannot be dictated; instead, individual breeders must be convinced that the evaluation is accurate and the goal is appropriate for them.

Rankings for overall merit should be supplied in addition to predictions of individual traits for animals. Selection index theory is well established, but economic values may be known imprecisely, may differ across time or region, and may be nonlinear. Animals may be ranked differently for different markets, and individual matings should be recommended if economic values are nonlinear or if genes with nonadditive effects are present. Distribution of individual trait predictions gives breeders more flexibility in designing breeding programs, but release of traits without estimated economic values forces users to make their own estimates, use independent culling levels, correctively mate for the trait, or disregard the trait.

Evaluations must be available to breeders at the time when selections are made. National or international evaluations computed once or twice per year may be sufficient for many purposes but may not identify elite new animals quickly enough when genetic gains are rapid. Distributed calculations within herd, region, or country linked to national or international evaluations may provide faster service without the need to reprocess all data more regularly (Robinson and Chesnais, 1988). Alternatively, real time updating of all animals as new data arrive may be possible with centralized data storage on disk. Delivering evaluations on time may be more important than increasing their accuracy slightly (Akins, 1988, personal communication).

Reliabilities, standard errors of prediction, or other accuracy measures indicate how much information was included in an animal's evaluation. A simple but accurate method to compute these is to sum information provided by parents, own records, and progeny each expressed as progeny equivalents (Wiggans and VanRaden, 1989). A standard progeny contributes one progeny equivalent, and progeny equivalents from each source are a simple function of the reliability that source provides to this animal. Progeny equivalents also are an educational tool because they quantify exactly how much of an animal's total information is from its own records versus from parents and progeny. Users of genetic evaluations like to know what information was included in an evaluation and how it was calculated in addition to final rankings of animals. Providing accurate documentation in simple language and supplying natural subdivisions of breeding value such as parent average, yield deviation, and progeny contribution allow breeders to verify data sources and build trust in and understanding of the procedures.

CONCLUSIONS

Two basic techniques are useful for improving animal model evaluation systems: 1) improve the model where it differs from reality or 2) improve reality where it differs from the model. Models that account more fully for genetic factors, environmental factors, and interactions should rank animals more accurately for selection, indicate which animals have the highest combining abilities, and could specify which genotypes are appropriate for which environments. Reality sometimes can be recorded differently, transformed, or adjusted so that assumptions of a simpler model are met.

Because individual breeders have financial incentives to obtain high evaluations, independence of genetic and environmental effects might be guaranteed only if genetic material is tested across several environments in balanced or randomized designs. Improved statistical techniques can extract more information from data but cannot replace collection of large amounts of high quality data for all traits of economic importance.

REFERENCES

- BERTRAND, J.K., HOUGH, J.D. and BENYSHEK, L.L. 1987. *J. Anim. Sci.* 64: 77-82.
- CHAUHAN, V.P.S. 1987. *Livest. Prod. Sci.* 16: 117-130.
- CHAUHAN, V.P.S. and THOMPSON, R. 1986. *J. Anim. Breed. Genet.* 103: 321.
- DJEMALI, M., BERGER, P.J. and FREEMAN, A.E. 1987. *J. Dairy Sci.* 70: 2374-2384.
- DUCROQ, V., QUAAS, R.L., POLLAK, E.J. and CASELLA, G. 1988. *J. Dairy Sci.* 71: 3071-3079.
- FERNANDO, R.L. and GIANOLA, D. 1986. *Theor. Appl. Genet.* 72: 822.
- FOULLEY, J.L. and GIANOLA, D. 1984. *Genet. Sel. Evol.* 16: 285-306.
- FOULLEY, J.L. and HENDERSON, C.R. 1989. *J. Dairy Sci.* 72: 167-172.

- GIANOLA, D. and FOULLEY, J.L. 1983. *Genet. Sel. Evol.* 15: 201-223
- GIANOLA, D. and FERNANDO, R.L. 1986. *J. Anim. Sci.* 63: 217-244.
- GROENEVELD, E. and KOVAC, M. 1990. *J. Dairy Sci.* 73: (Feb.).
- HARVILLE, D.A. and MEE, R.W. 1984. *Biometrics* 40: 393-408.
- HENDERSON, C.R. 1973. *Proc. Anim. Breed. Genet. Symp. in Honor of Dr. J.L. Lush.* Amer. Soc. Anim. Sci. & Am. Dairy Sci. Assoc., Champaign, IL, pp. 10-41.
- HENDERSON, C.R. 1974. *J. Dairy Sci.* 57: 963-972.
- HENDERSON, C.R. 1976. *Biometrics* 32: 69-83.
- HOESCHELE, I. 1988. *Theor. Appl. Genet.* 76: 81.
- HOESCHELE, I. 1990. Potential gain from insertion of major genes into dairy cattle. *J. Dairy Sci.* 73: (in press).
- HOESCHELE, I. and VANRADEN, P.M. 1990. Rapid method to compute inverses of dominance relationship matrices for noninbred populations. *J. Dairy Sci.* (submitted).
- KENNEDY, B.W. and SCHAEFFER, L.R. 1989. *J. Anim. Sci.* 67: 1946-1955.
- NORMAN, H.D., DICKINSON, F.N. and WRIGHT, J.R. 1985. *J. Dairy Sci.* 68: 2646-2654.
- QUAAS, R.L. and POLLAK, E.J. 1980. *J. Anim. Sci.* 51: 1277-1287.
- ROBINSON, J.A.B. and CHESNAIS, J.P. 1988. *J. Dairy Sci.* 71(Suppl. 2): 70-78.
- SMITH, S.P. and QUAAS, R.L. 1984. *J. Dairy Sci.* 67: 2999-3007.
- THOMPSON, J.R., LEE, K.L., FREEMAN, A.E. and JOHNSON, L.P. 1983. *J. Dairy Sci.* 66: 325-331.
- THOMPSON, R. 1979. *Biometrics* 35: 339-353.
- VANRADEN, P.M. and HOESCHELE, I. 1990. Rapid method to compute inverses of additive-by-additive relationship matrices. *J. Dairy Sci.* (submitted).
- WESTELL, R.A., QUAAS, R.L. and VAN VLECK, L.D. 1988. *J. Dairy Sci.* 71: 1310-1318.
- WIGGANS, G.R. and VANRADEN, P.M. 1989. *Natl. Coop. Dairy Herd Impr. Progr. Handbook, Fact Sheet H-2*, Washington, DC, 8 pp.
- WRIGHT, S. 1922. *Amer. Naturalist* 56: 330-338.