This book is dedicated to Professor Dr G. K. Constantinescu, founder of the modern animal husbandry science in Romania, originator of the National Animal Husbandry Institute (1926) and initiator of the first laboratory of experimental genetics.

Preface

The idea of using the most valuable animals for breeding is ancient, being mentioned by VARRO, 2000 years B.C.; this idea was resumed, in one form or another starting with the 18th and 19th centuries (Andre). From its early moments, the genetic breeding of the dairy cattle focused on the identification of the best dairy cows in terms of genetics. The dependence of phenotype on genotype in cattle has been conceived since late 19th century, when H. BRANTH (1891), a Danish farmer said that "... the ability of a cow to produce more or less milk fat, from the feed it eats, depends on heredity". The author showed (1893) that the outer look of an animal cannot provide information on its ability to produce high or low-fat milk. This ability might be evaluated using the pedigree, but the best way is to make an evaluation using the progeny. Although BRANTH grasped the major role of the progeny for the genetic evaluation of the breeders he did no special investigations into this matter.

Daughter-Dam Comparison. BRANTH's ideas have been carried further by SEDELHOLM, who checked them in his own farm (1900). SEDELHOLM compared the daughters with their dams in terms of milk butterfat percentage, showing that the sire has a variable influence on daughter's performance. Historically, this was the first actual attempt to apply the progeny selection in dairy cattle. After 1900, the attempts to identify the best parental stock in dairy cattle entered a new stage, the stress falling on the genetic evaluation of the bulls. Thus, the Danish were the first to introduce, in 1902, the progeny testing of bulls using the records provided by the dairy cows breeding associations, while as of 1912 they started to use the daughter-dam comparison (J. Johanson, 1960). Subsequently, based on the work of Wright and Lush, several operational versions of the daughter-dam comparison method have been developed from 1925 to 1945. Within this context, the proposed indices can be classified into four categories: 1) indices that consider only daughter's average (Gifford, 1930); 2) indices that consider the actual daughter-dam difference (Pearl, Norton, Rice, Mount Hope); 3) indices that consider the number of daughters (Wright, 1932) and 4) indices that consider heritability and repeatability as genetic parameters (Lush, 1941).

In the USA, the daughter-dam comparison method, with its different versions, has been used officially for the genetic evaluation of the bulls from 1936 to 1962. Despite its several improved variants, the daughter-dam comparison model had to make place for a new method, which allows removing the environmental differences between the farms in which the candidates to selection performed.

Contemporary Comparison. The new proposal was called the Contemporary Comparison model and it was introduced by Robertson and Rendel, in 1954. Independently of the two, a variety of this method, the Herdmate Comparison, was presented in the USA by Henderson C.R., in the same year. The difference between the two methods is that the first method used just the records of the first lactation, while the second method considered all lactations of the surveyed animals. This method justified its applicability in the hypothesis of the genetic similarity of the herds.

Modified Contemporary Comparison. After two decades of its application, the situation of the dairy cattle populations changed very much due to the accumulated genetic progress. Under the new circumstances, a new variant was proposed, which allowed adjustment for the genetic merit of the contemporaries; the inclusion of the genetic merit of the ancestors; and the use of all lactations. The "Modified Contemporary Comparison" was used officially by USDA from 1974 to 1989, when it was replaced by the Repeatability Animal Model.

Cumulative difference method. Other countries, such as Israel and Germany, also tried to improve the Contemporary Comparison method. Thus, Bar-Anan and Sacks (1974), described a method which tried to correct the deficiencies of the Contemporary Comparison method. The new procedure was called the "Cumulative difference method", CD; the estimated breeding value of the sire consists of two parts: a) an estimate of the contemporary comparison; and b) an adjustment for the genetic level of the contemporaries of the sire's daughters. The practice of estimating the breeding value of the bulls showed that the bulls with a lower number of daughters are consistently disadvantaged against the other bulls. To remedy this deficiency of the method, L. Dempfle (1976), proposed a modification of the Bar-Anan and Sacks formula. Thus, the multiplication with the regression coefficient was to be done after considering the genetic level of the contemporaries of the sire's daughters.

Least Squares Method. Another method proposed for the calculation of the estimated breeding value of the bulls was the least squares method. Initially, the method was recommended by Robertson and Rendel (1954), but due to the low computing capacity at that time, this method was no longer used for genetic evaluation. Later, Henderson (1963), Searle (1964) and Cunningham (1965) improved this method. Although the least squares method was not widely used

for animal breeding purposes, it has historic importance because it allowed the advancement to a more elaborate method (BLUP).

Best Linear Unbiased Prediction. The father of "BLUP methodology" is Charles Roy Henderson, disciple of J. Lush. Henderson also formed a school (Cornell University, Ithaca, N.Y.), where his disciples (L.R. Schaeffer, B. W. Kennedy, S. Searle, D. Gianola, D. Sorensen, V. Ducrocq, C. Lin, etc.) studied and developed new methodologies for the prediction of the breeding value, taking further the work of Wright, Lush and Henderson. Theoretically, BLUP was established as far as in 1949, but its actual use in practice started only in 1970 due to technical reasons. Henderson said in 1973 that: "Theoretically, most of the calculation principles of BLUP methodology were already available, but the computing capacity was completely inadequate for its utilization. The contemporary comparison method was a technically possible compromise at that time"

Sire Models. Historically, the first type of model used for the genetic evaluation of the dairy cattle was the Sire Model, in the crossed two-factor variant, without genetic groups. The practical validity of the model relied on several working hypotheses among which: the group of sires is a random sample of the population, unselected, with no inbreeding; the sires are not related among them or with their mates; mating is random, one progeny being tested from each dam; the progeny of a sire are considered to be half-brothers, not related to the progeny of other sire; the dams are a random sample, representative for the original population. In time, due to the accumulation of genetic progress in the dairy cattle populations, part of these working hypotheses were invalidated and thus the breeding value of the candidates to selection was overestimated/underestimated. Thus, a first disagreement with the hypotheses refers to the situation in which the evaluated sires belong to the same homogenous population. As the artificial insemination (AI) techniques expanded, and as some outstanding bulls were being used intensely, the genetic differences between the subpopulations of the same breed became stronger. Within this context, the hypotheses that the candidates to selection come from a single homogenous population, with no genetic differences between farms, were invalidated. As genetic differences existed, the young bulls were consistently underevaluated compared to the older bulls. This difficulty was explained by an increasing trend of the genetic merit of the contemporaries within the same farm. Henderson raised again in 1966 the problem of the genetic groups and proposed a model which to eliminate the difficulties encountered by the classical methods of genetic evaluation of the bulls. The new model was called "NEAISC Model - the Northeast Artificial Insemination Sire Comparison", and it was introduced in the genetic evaluation of the sire in Northeast USA starting from 1972. Initially, the bulls were considered as being unrelated and the genetic groups have been defined as sets of sires from the same AI organization

(subpopulation) which started to be used for AI at the same time. Generally, the purpose of grouping was to account for the genetic trend and for the genetic differences between subpopulations. The introduction of the genetic groups in the biometric model to correct the estimates for the existing genetic differences was prompted by the fact that the relationships between the analysed sires were not accounted for, either because the genealogical information was not known, or due to the high cost of inversing the relationship coefficients matrix (A). Until 1974, the genetic evaluation methods used in the USA (Herdmate and Cornell Comparison) didn't use the relationships between the sires.The MCC included sire and maternal grandsire information using selection index equations in 1974 which added considerable accuracy to the genetic evaluations. A year later, Henderson C.R. discovered how to include these relationships in BLUP. The relationship information was transferred in both directions (to and from ancestors) by BLUP. However, the genetic group effects inherited between generations was included in the MCC but not in BLUP, until inherited groups were introduced by Westell et al. in 1988. Although the potential advantages of using sire relationships to calculate the estimated breeding value were known previously, their use was prevented by the prohibitive cost involved by the inversion of large matrices (A). After 1975, most researchers used the discoveries of Henderson, regarding the direct inversion of the relationship matrix without using the classical inversion methods. The use of relationships for the genetic evaluation of the bulls had several advantages among which: 1) higher accuracy of the genetic value prediction, particularly for the sire with lower progeny number or with no progeny; 2) it reduced the number of genetic groups necessary to account for the genetic trend and 3) allowed the earlier evaluation of the bulls by the possibility of considering the performance of their dams and of the paternal half-sisters. Some of the other hypotheses that were invalidated in time were: a) the mating is random and b) the dams are a random sample, representative for the original population. Because some farmers preferred to mate the best dams with the best sires, the hypothesis that the sires are randomly dispersed among farms was also invalidated. The breeding value of the bulls was corrected to account for the fact that the genetic level of the female mates is usually higher than the average population. Thus, Quaas et al. (1979) and Everett et al. (1979) developed the Sire-Maternal Grandsire model (Everett, R. W., and J.F. Keown, 1984). This model was used by the Cornell University to evaluate the bulls in North-eastern USA from 1979 to 1989, increasing the number of traits since 1982. Besides the hypotheses of the initial model, the new model also stipulated that the daughters of the maternal grandsires must be a random sample of the daughter population of all maternal grandsires. Later, it was shown that this hypothesis didn't verify in practice, because the daughters of a bull which calved subsequently, had already been selected after their first

calving, particularly if they were of second parity and following, which breached the assumption that they were a random sample.

Animal Models. Year 1989 was the start of a new era in the history of the genetic evaluation of the dairy cattle, by the introduction of the Animal Model. This became practically possible due to the progress in computer hardware. The BLUP methodology applied to the animal models rapidly became the reference method for the genetic evaluation of the animals. The notion of individual animal model was introduced explicitly by Quaas and Pollak (1980), although Henderson (1949) referred to this aspect in his paper on the estimation of the genetic and environmental trend. The animal model is the procedure that estimates the breeding value by describing the genetic effect of the progeny, not of the parents. Most previous models (contemporary comparison, herdmate comparison, and Sire Model) used only progeny performance for comparison. For these methods, dam evaluation was secondary and was done using the genetic evaluation of the sire. The animal model uses all sources of information: own performance, performance of the collateral ascendants and of the descendants. Compared to the method of the selection indices (BLP), BLUP methodology applied to an animal model has several advantages: it uses the information on all known relatives of an individual, thus enhancing the accuracy of prediction; it facilitates the genetic comparisons between the animals that performed in different environments and different periods of time; it facilitates the genetic comparisons between animals with different sources of information (different number of kin and different number of records for the same trait) - for instance, a cow with three lactations can be compared with a heifer; it allows genetic comparisons between animals that have been selected at different selection intensities; it allows an accurate measurement of the selection response. The implementation of the Repeatability animal model by USDA (in July 1989) was one of the greatest changes ever of the national methodologies of evaluation. VanRaden et al. (1989) have shown that the animal model was 3-5% better than MCC in terms of the EBV accuracy for dairy cattle. Although 1989 is cited as reference year for the practical implementation of the Repeatability animal model, this event took place earlier (1982) in Romania. Appendix B presents the "Romanian Animal Model, 1982", the contribution of the Romanian researcher Corneliu Drăgănescu to the development and implementation of a program for the genetic evaluation of the dairy cattle in Romania.

Multiple Across Country Evaluation. The intensification of semen export from USA and Canada to countries all over the world practically globalized dairy cattle breeding. This aspect complicated the process of sire selection. Thus, the importing countries were confronted with the difficulty of selecting the best bulls among the imported bulls vs. the local bulls. Thus, objective criteria had to be set for the identification of the best sire stock. This was no easy task

because each country had its own system of performance control and genetic evaluation, as well as different ways of expressing the breeding value of the bulls. The International Dairy Federation (IDF) proposed in 1981 to use regression equations for the conversion of the breeding value depending on the (importing/exporting) country. Thus, Goddard (1985) and Wilmink (1986) modified the regression equations to account for the accuracy of the breeding values in each country, but the procedure was shown to be very little efficient, the conversion being often limited to pairs of two countries. This prompted for additional work to enhance the working efficiency. The new method proposed by L.R. Schaeffer (1985) used the linear model which had the country where the genetic evaluation was done, the genetic group of the bull and the estimated breeding value as inputs. The model didn't account for the genotype-environment interaction, however. To deal with this aspect, L.R. Schaeffer (1994) also introduced the genetic correlations between the countries of origin of the bulls, which made it possible to obtain different classifications of the bull in different countries. The new method, known as MACE (Multiple Across Country Evaluation) was a crucial development in the field of the international genetic evaluation and it smoothed the way towards the establishment of INTERBULL (The International Bull Evaluation Service, Uppsala, Sweden). MACE was combining in an optimal manner the relationship information both at the national level and across countries.

Animal Model Multiple Traits. Even though the animal model for a single trait had become applicable in 1989, the variant for multiple traits at the individual level was not feasible, also due to the limitations of the computing capacity. However, there had been some limited applications, based on the Sire Model, for some breeds of meat cattle. One way of making applicable the multiple traits animal models was to use the procedure of data transformation. The best known and most used procedures of data transformation were the canonical transformation and the Cholesky transformation. Once transformed, the observation data, one may revert to the separate evaluation of the individual traits, which meant a consistent reduction of the required computing capacity. After the breeding values were computed on a transformed scale, they were ultimately reverted to their original scale using a retransformation procedure. Besides the traits of milk production (amount and quality of milk), the breeders associations have also been interested in the evaluation of the type traits because of the relations between the type traits and the production. The phenotypic expression of each type trait can be known when the primiparous cows are evaluated. In Canada, the Holstein cows are evaluated for 30 type traits, while in Europe they are evaluated for 16-20 type traits, with variations in individual countries. Because all traits are measured/observed on all primiparous cows and because their expression is influenced by the same environmental factors, the best way to ana-

lyse the animal model is the canonical transformation. Another variant may be the threshold models because some traits are evaluated in a subjective manner. However, a trend towards the simultaneous genetic evaluation for multiple traits can be noticed worldwide, because this makes it possible to use the information supplied by the genetic correlations between the traits. This means a higher accuracy of the obtained EBV, particularly for the traits with low inheritability, or when information lacks for some traits.

Test Day Models. Until 1998, the biometric model for the genetic evaluation of dairy cattle relied, according to ICAR norms, on the Lactation Model, which describes linearly in terms of genetic and environmental effects the production of milk/fat/protein per standard lactation, i.e. 305 days from the start of lactation. Thus, the intra-lactation genetic and environmental variation was set to be residual variance, which meant that, implicitly, the intra-lactation selection information was not used. This solution, accepted by INTERBULL regulations until not long ago, relied on the limitations due to the computing capacity. The explosive development of the computing capacity made possible the implementation of the Test-day model, which records the performance of the test-day, the sum of the productions of milk/fat/protein in the test-day. The linear description of the test-day performance in terms of genetic and environmental effects can be done in several ways. A dominant idea was that, same as the performance, the effects follow the lactation curve. Thus, the breakdown of performance shows regressions with constant coefficients for the fixed effects and regressions with random coefficients for the additive and environmental effects. Thus, the model allowed the construction of two categories of lactation curves: the first one, common for all contemporaries, and the second one particular for each individual cow. The first type of regression was called Fixed regression test day model, which affects similarly all the contemporary animals, while the second type of regression was called Random regression test day model because the values of the coefficients vary from one animal to the other. The major advantage of using the test-day model is the possibility to correct the records for the environmental factors whose impact changes during lactation, between two test-days. The stage of lactation is a basic element of such model; between the stage of lactation and the amount of milk there is a non-linear relation. Several types of functions are used to draw the lactation curve, such as Wood, 1967; Wilmink, 1987; Ali and Schaeffer, 1987; Legendre polynomials, Spline function. These mathematical functions have been incorporated by different authors in biometric models with the purpose to calculate the genetic value of the sires. Depending on the components of the model we have the Fixed regression test day and the Random regression test day (based on Ali and Schaeffer functions, Legendre polynomials, and the Spline function). Another biometric model used to analyse the records is the Autoregressive re-

peatability animal model developed by J. Carvalheira et al. (1998; 2002). The calculations related to these models are presented in detail, so that the reader may understand these categories of models. The most accurate results seem to be provided by the model that includes the Spline function.

Genetic Changes. The success of applying any breeding program is measured by the amount of genetic progress achieved by the populations along the successive generations. Knowing these achieved values is necessary in order to justify the usefulness of the genetic evaluations and of the new working methodologies. This aspect is included in a chapter dedicated to measuring the genetic progress in the dairy cattle populations. All the procedures used to measure the genetic progress in dairy cattle are presented in historic succession. These methods are closely related to the procedures used to estimate the breeding value.

Threshold Models. The traits that don't have normal (Gauss) distribution must be analysed using non-linear models. These traits have a discrete variation and follow the Poisson distribution, the analysed individuals falling into distinct classes. Such examples are the calving easiness, resistance to diseases, litter size, number of embryos and even some outer traits determined by subjective evaluation. Thus, for the calving easiness the animals can be classified at least into three distinct categories: 1. Unassisted calving, 2. Assisted calving, and 3. Dystocia. Although the Threshold models are best fitted for the analysis of the data with discrete variation, given the complexity of calculations, most genetic breeding programs worldwide used linear models. The classifications of sires using linear and non-linear models are very closely correlated (0.99).

Survival analysis. Survival analysis was initially used in other areas of activity (human medicine or reliability analysis). The first one to propose the use of this technique in animal breeding was S.P. Smith (1983), in his PhD dissertation "The extension of failure time analysis to problems of animal breeding". Cornell University, Ithaca, N.Y., USA. This method was subsequently reviewed and improved by Smith and Quass, 1984 and by Smith and Allaire, 1986. Ducrocq and Solkner (1994) developed a set of software for the analysis of survival data in cattle (The Survival Kit, a Fortran package for the analysis of survival data. In Proceedings of the 5th World Cong. on Genet. Appl. to Livest. Prod.). Survival analysis is defined as a set of analytical methods for data analysis when the output variable is the time left until the occurrence of a particular event, which may by the culling date, in the case of cows. The time from the first calving to the date of culling is called the average time of exploitation or the productive life. Usually, a full analysis of the actual productive life can be done only after all animals have been culled, which makes the final decision of identifying the best parents to be tardive, thus inefficient. One way to make an efficient analysis is to use the

information from live animals, so we can take into consideration the productive life achieved up to a specific point in life, before culling. These are called censored observations and they can be used to predict the potential moment of culling, one solution to estimating the productive life. The Weibull regression can be used to estimate the breeding value of the productive life, which a basic component of the survival analysis. Such working procedure has been available since 1996 within the Verden computer centre (Germany; The Survival Kit, 1998). Many countries currently use the algorithm developed by Ducrocq and Solkner (1994), for survival analysis. An alternative to that algorithm is the use of the Random regression test day because it is easier to apply than the first one, while it also allows re-ranking animals at different moments in time, according to Jamrozik et al. (2008).

Genomic analysis. One way to enhance the annual genetic progress is using information from genetic markers. The marker-assisted selection (MAS) has several advantages: 1) it can be applied for both sexes for the sex-limited traits; 2) it is much more efficient for the traits whose cost of testing is prohibitive, or which are very hard to test (resistance to diseases); 3) it can be applied very early for the reproduction traits or for the traits that are measure on the carcass. The genetic markers are useful for the identification of the chromosome parts that are associated to particular production traits. The use of information from the genetic markers correlated with the quantitative trait loci (QTL), next to the phenotypic information and to the genealogical data adds accuracy to the prediction of the breeding value, thus of the selection. This type of selection is highly efficient for the sex-limited traits and for the traits with low heritability, such as milk production in dairy cattle. It can also be used for the selection of young bulls before progeny testing, which means a shorter interval between generations and a higher annual genetic progress. To be useful, markers needed an LD of 30% or more. High LD means that an allele of the marker is on the same stretch of DNA as the favourable allele of the gene. Marker-assisted selection has its limits, because the intensity of the relation between marker and the QTL may decrease in time. A recommended alternative that has been used recently is the use of the huge variation of DNA following its sequencing. The most widespread form of genome variability is the Single Nucleotide Polymorphism (SNP). Consequently, the SNP have been increasingly used in the recent years for breeding value estimation in dairy cattle too, which is why we are now calling the process "Genomic selection". The basic requirement for selection efficiency is that the markers are in linkage disequilibrium with the QTL. The practical implementation of the genomic selection requires estimating SNP effects within a reference population and the prediction of the breeding value for the animals outside that population. Misztal et al. (2010) developed a one-step method

which uses the phenotypic, genealogic and genomic information to determine the genomic breeding value. In conclusion, the purpose of this book is to describe the evolution of the methods of genetic evaluation of the dairy cattle, starting with the daughter-dam comparison and ending with genomic selection, using several landmarks: description of the working method, showing the hypotheses and the statistic properties of the biometric models; numeric application for each method and, which is very important, justification of the transition from one working methodology to another.

In a book of this domain, it is impossible for the authors to balance the invaluable contribution of all those whose concepts and ideas are presented. We refer to basic and applied scientists in research institutes, universities. We owe a lot to the people in all these sectors with whom we communicated and whose scientific work we studied and commented.

At the local level, Romanian authors would like to acknowledge the encouragement, comments and suggestions we have had in this work from Professor Condrea Drăgănescu, Professor Popescu Ştefan Vifor. Also to Mihai Roman, Cristiana Grosu, Oana Bărbulescu, Dan Bărbulescu assisted us in the English version. We are grateful to Ms Doina Argesanu, the script supervision team, for the work to convert our digital script into a readable form.

Bucharest, 2013 Authors

Contents

Chapter 1

Genetic Evaluations

HORIA GROSU
P. A. OLTENACU
LARRY SCHAEFFER

1.1 Early History

The dependence of phenotype on genotype in dairy cattle has been considered since the late 19th century. Branth (cited by Bonnier, 1936), a Danish farmer, said (1891) that "the ability of a cow to produce more or less milk fat, from the feed it eats, depends on heredity". Also, the outer look of an animal cannot provide information on its ability to produce high fat or low fat milk. Such evaluation could be done using the pedigree, but a better approach is evaluation by progeny. Although Branth instinctively realised the role of the progeny for the genetic evaluation of a sire, he conducted no special investigations in this direction.

Branth's ideas have been further developed by Sedelholm (cited by Bonnier, 1936), who verified them on his own farm (1900). Practically, Sedelholm compared the daughters with their dams in terms of milk fat, proving that bulls have a variable influence on daughter records. Historically, this was the first real attempt to apply selection by progeny in cattle. After 1920, the research to identify the best animals in a dairy cattle population entered a new stage with the focus on the genetic evaluation of bulls. The scarcity of information regarding the role of genetics in the evolution of domestic animals is likely responsible for slow progress. Although Gregor Mendel had elaborated the laws of heredity as early as 1866, they were not understood for almost 40 years. Rediscovery of

Mendel's laws after 1900, should have provided a fresh and strong impulse to the field of genetic breeding. However, this happened almost a half century later due to a) the delayed acceptance of mathematics as an instrument of investigation in biology, and b) the failure to see clearly that genetic breeding does not refer to individuals, but rather to the population (Drăgănescu, 1979).

Quantitative genetics evolved as a science only after 1920 based on the works of R. A. Fisher (The Correlation Between Relatives on the Supposition of Mendelian Inheritance, 1918) and S. Wright (Mating Systems, 1921). Ten years later, the basics of population genetics had been laid out by the same two scientists: R. A. Fisher (The Genetical Theory of Natural Selection, 1930) and S. Wright (Evolution in Mendelian populations, 1931) and by the book of J. B. S. Haldane (cited by Drăgănescu, 1979) "A Mathematical Theory of Natural and Artificial Selection". These papers outlined the modern theory of evolutionism, population genetics and quantitative genetics, pioneering fields which had set the groundwork for the science of genetic breeding (J. L. Lush, 1945). The first books approaching genetic evaluation of bulls in a scientific manner were published after clarification of the principles governing the genetic evolution of the populations of domesticated animals (Wright, 1930, 1931, 1932; Lush, 1931, 1933, 1935, 1944, 1945). The works of these two giants of the science of animal breeding formed the basics for the development of new procedures of genetic evaluation of the sire by their disciples (L. N. Hazel; C. R. Henderson; L. D. Van Vleck).

1.1.1 Artificial Insemination

The first usage of artificial insemination (AI) in dairy cattle was around 1936. Before 1936, bulls were used through natural service and usually limited in the number of herds in which they had daughters (usually only 1 or 2). AI technology made it possible for dairy bulls to produce many thousands of progeny across many herds, if the bulls were genetically superior or popular with breeders. The need to assess the genetic merit of dairy bulls became more urgent with such profit potential from the sale of bull semen. Large progeny group size naturally led to greater accuracy in genetic evaluations of bulls. The need to make genetic evaluations as accurate as possible was driven by artificial insemination and competition among organizations and countries.

In Europe, organizations that computed genetic evaluations were the same companies that collected and sold bull semen, and sometimes were also concerned with breed registry. In North America, AI organizations were not part of genetic evaluation, which was undertaken by the United States Department of Agriculture, or Agriculture Canada. AI provided the impetus to improve genetic eval-

uation methods. AI was not immediately available in other species due to the inability to freeze semen in other species and afterwards have high conception rates.

1.1.2 Computers

Genetic evaluations are very suitable for computing machines, but advances in genetic evaluation procedures had to often wait for the advances in computer hardware. The modern day history of computing seems to begin in 1939, although the abacus was around since 2400 BC. The binary number system was described by Pingala (India) in 300 BC, and negative numbers were shown by the Chinese in 100 BC. Pascal in 1642 invented a mechanical calculator.

The first mechanical computer was designed by Charles Babbage in 1822 which by 1834 would have a program stored on punched cards. Funding was withdrawn when Babbage had trouble finding machinists who could make the parts that he needed within the exact tolerances that he required. Adding machines were invented often through the early 1900's. Vacuum tubes and circuit designs appeared by 1920. In 1928 IBM standardized the use of punch cards for storing data and programs. So while the basics of animal breeding and population genetics were being written, modern computers were also beginning to develop, but in the 1920's mostly adding machines and card counters were available for use, thus limiting the methods that could be used for genetic evaluation.

While AI was a driving force in genetic evaluations, developments in computer hardware were often lagging, and therefore limited the sophistication that could be applied to genetic evaluation, at a given point in time. For example, Henderson's Best Linear Unbiased Prediction procedure was known in 1949, but was not actually applied until 1970. The animal model was known in the 1960's, but could not be applied until 1989 due to insufficient computing power. Test-day models were discussed in the 1970's, but were not implemented until 2000. Genomics, through single nucleotide polymorphisms and gene sequences, is now challenging computer technology just to be able to store the information on thousands of animals, and to process the information efficiently. The connection between computer hardware and genetic theory will likely continue for many decades.

The aim of the book is to describe the evolution of genetic evaluation methods in dairy cattle by discussing their calculations, assumptions, and statistical properties, and the reasons for the advancements that were made at those points in time. Details of the theories and development of methods can be found in other texts.

1.2 References

BONNIER, G. 1936. Progeny Tests of Dairy Sires. Hereditas 22:145.

DRAGANESCU, C. 1979. Animal Breeding. Ed. Ceres, Bucuresti.

FISHER, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc., Edinburgh 52 part 2: 399-433.

FISHER, R. A. 1930. The Genetical Theory of Natural Selection. Oxford Clarendon Press. ABO-5438.

HALDANE, J. B. S. 1932. A mathematical theory of natural and artificial selection, Part IX. Proc. Cambridge Phil. Soc. 28: 244-248.

LUSH, J. L. 1931. The number of daughters necessary to prove a sire. Journal of Dairy Science 14: 209–220.

LUSH, J. L. 1933. The bull index problem in the light of modern genetics. Journal of Dairy Science 16: 501-522.

LUSH, J. L. 1935. Progeny test and individual performance as indicators of an animal's breeding value. Journal of Dairy Science, 18:1.

LUSH, J. L. 1944. The optimum emphasis on dams' records when proving dairy sires. Journal of Dairy Science 27: 937.

LUSH, J. L. 1945. Animal Breeding Plans, 3rd ed. Iowa State College Press, Ames. 443 p.

SEDERHOLM, G. 1900. Några iakttagelser om mjólkens fetthalt. - Landtmannen, p. 157-161.

WRIGHT, S. 1921, 1931, 1921, 1934. Systems of mating and other papers. Iowa State College Press.

WRIGHT, S. 1930. The genetical theory of natural selection (by R. A. Fisher). A review. J. Hered. 21:349-56.

WRIGHT, S. 1932. On the evaluation of dairy sires. Proc. Amer. Soc. Anim. Prod., p. 71-78.

Part I

SELECTION INDEX BASED METHODOLOGIES

Chapter 2

Daughter-Dam Comparisons

HORIA GROSU
PASCAL ANTON OLTENACU
LARRY SCHAEFFER

2.1 The Beginning

From the early 1900's to the 1960's genetic evaluation of dairy bulls was in the Era of Daughter Averages. Record keeping systems in North America began in 1905, and computing systems were limited to adding machines. The easiest quantities to calculate were averages, variances, and covariances. Rice (1933) listed criteria for a useful index (genetic evaluation):

1. The evaluations should be understandable by the users. Producers were familiar with lactation production figures, so that the evaluations should relate directly to lactation production.

2. The evaluations should include both dam and daughter records. This conclusion was not reached immediately, but generally all of the scientists concurred on this point by the 1930's.

3. The evaluations must clearly indicate how much the production level of the progeny would be improved. Producers were not used to looking at rankings of bulls, but only production levels.

Daughter-Dam Comparisons were the main interest in comparing dairy bulls. Many different formulas for combining daughter averages and dam averages

were proposed from 1925 to 1944. Some of the areas of concern regarding these proposals were as discussed in the following sections.

2.1.1 Merit of Dams

The main advantage of a daughter-dam comparison was the ability to account for the genetic worth of the dams as they contribute to a bull's daughter records. On the other hand, daughter records were not contemporary with dam records with often a gap of at least two and a half years between records, even within the same farm. If dam and daughter records are made in different herds, then the environmental effects could be even larger than for within herd comparisons.

2.1.2 Relatedness of Bulls and Dams

If bulls were related to each other, as some would be, the exact relationships were not considered, and another assumption was that bulls were not related to any of the dams to which they were mated. This could also have been violated in some instances, particularly when bulls were used within herds prior to artificial insemination. In general, the assumption was that bulls were unrelated to other bulls and to all mates.

2.1.3 Years

As years go by, hopefully genetic change has occurred, and therefore, the mates for younger bulls will have a higher genetic merit than the mates of bulls in earlier years. If a bull has daughters over a number of years, and if the mates are becoming genetically better, then the deviations of the daughters' averages compared to mates should decrease over time. Proposals did not allow for positive or negative genetic trends.

2.1.4 Dams and Daughters

Dams would most likely have more than one daughter, possibly sired by different bulls over their lifetime. Thus, the dams average phenotype would be used over and over in the indexes of different bulls. Given the computational limitations in those days, there was little that could be done to account for this situation.

2.1.5 Number of Daughter-Dam Pairs

The variability of means is a function of the number of observations going into the mean,

$$Var(mean) = \frac{\sigma^2}{n},$$

where σ^2 is the variance of individual observations and n is the number of observations in the mean. Thus, bulls with few daughter-dam pairs would have greater variation than bulls with many daughter-dam pairs. The result is that bulls with few pairs could rank higher than bulls with many pairs or could rank much lower. The likelihood of bulls with few pairs of daughter-dams to rank in the extremes was greater than for bulls with many pairs.

2.1.6 Other Considerations

Criteria like having evaluations that have a high correlation with the true breeding value of the bull, or evaluations that are unbiased or minimum variance were not considered in the proposals except for Bonnier (1936). These ideas did not occur until later with Lush, Henderson and others. At this time, probably no animal breeders knew about matrix algebra, least squares analyses or analysis of variance (1918). The application of mathematics to biology was not very popular.

2.2 Theoretical Concepts

2.2.1 A Daughter Record

Let the record of a bull's daughter be represented as

$$X_{ijkl} = \mu + .5\, s_i + .5\, d_j + m_k + p_k + e_{ijkl}$$

where

X_{ijkl} is the phenotypic observation on daughter k of sire i and dam$j$,

$.5s_i$ is an average half of the sire's true breeding value, which is the object to be estimated,

$.5d_j$ is an average half of the dam's true breeding value,

m_k is a Mendelian sampling effect of the k^{th} daughter, generated by the mixing of alleles from the sire and dam,

p_k is a permanent environmental effect common to all records on daughter k, and

e_{ijkl} is a residual error specific to this one record.

With respect to daughter k, all factors in the equation, s_i, d_j, m_k, and p_k are constants. The expected value of e_{ijk} is always zero. No other factors are assumed to affect the daughter record. In practice there are many factors that could contribute to X_{ijkl}, but the assumption was that adjustments could be made for most of these factors.

Thus,

$$E(X_{ijkl}) \;=\; \mu \;+\; .5\; s_i \;+\; .5\; d_j \;+\; m_k \;+\; p_k$$

2.2.2 Repeated Daughter Records

Consider the same daughter making n records, the average would be

$$\overline{X}_{ijk.} \;=\; \mu \;+\; .5\; s_i \;+\; .5\; d_j \;+\; m_k \;+\; p_k + \frac{1}{n}\sum_{l=1}^{n} e_{ijkl}$$

As n becomes larger, the average of the e_{ijkl} should approach zero.

2.2.3 Several Daughters

Consider the average of q daughters of sire i, the average would be

$$\overline{X}_i \;=\; \frac{1}{q}\; X_{i\ldots} \;=\; \mu \;+\; .5\; s_i \;+\; .5\; \overline{d} \;+\; \overline{m} \;+\; \overline{p} \;+\; \overline{e}_i$$

and with large q

$$(\overline{m} \;+\; \overline{p} \;+\; \overline{e}_i)$$

should approach zero. Thus,

$$2\; (\overline{X} \;-\; A) \;=\; s_i \;+\; \overline{d},$$

where A is the breed average estimate of μ. If the bull is mated randomly to dams, then $\overline{d}$ should also approach zero.

2.2.4 Dam Records

Similarly, the average of n dam records could be represented as

$$\overline{Y}_j \;=\; \mu \;+ d_j \;+\; p_j \;+\; \overline{e}_{j.}$$

where

$\overline{Y}_j$ is mean of n record of dam j,

μ is the breed mean,

d_j is assumed to be the dam's total additive genetic effect, or true breeding value,

p_j is the dam's permanent environmental effect, and

$\overline{e}_j$ is an average residual effect that should tend to zero as n increases.

Finally, the average of the dams or mates of the bull that produced the daughters of the bull would be

$$\overline{Y} \;=\; \mu \;+\; \overline{d} \;+\; \overline{p} \;+\; \overline{e},$$

where $\overline{p}$ and $\overline{e}$ are expected to tend towards zero over a large number of dams, but $\overline{d}$ is expected to differ between bulls because of differential usage of bulls on dams.

2.2.5 Bull Estimated Breeding Values

Combining the above results, then a bull's estimated breeding value would be given by

$$\begin{aligned} 2(\overline{X} \;-\; A) &= s_i \;+\; \overline{d} \\ (\overline{Y} \;-\; A) &= \overline{d} \\ s_i &= 2(\overline{X} - A) \;-\; (\overline{Y} - A) \\ &= -A + 2\,\overline{X} - \overline{Y} \\ &= -A + 2\,(\overline{X} \;-\; .5\,\overline{Y}) \end{aligned}$$

assuming that q is large enough so that $\overline{m}$, $\overline{p}$, and $\overline{e}_i$, in the daughter average go to zero, and $\overline{p}$ and $\overline{e}$ in the dams' average also goes to zero. The proposed methods in this chapter vary around this final formula by changing the assumptions or making new assumptions.

2.3 A General Expression

Graves (1925) was the first to use Daughter-Dam Comparisons under USA conditions (L. D. Van Vleck, 1985). Later on, the method was adopted officially (1935) by the US Department of Agriculture and used for about 30 years until it was replaced by the Herdmate Comparison Method (1962).

Several indexes were developed during this period for the genetic evaluation of dairy bulls, most of them being variations on the basic method (Daughter-Dam Comparison). As a rule, when evaluating the genetic merit of bulls, the average record of the daughters is expressed as a deviation from the average record of the dams. Before calculating the index, the records of both categories of females were corrected for age, length of lactation and times milked per day (305-2X-ME). If females had several previous lactations, their average was used in the calculations as well.

A number of different indexes were proposed over the years by Hansson, Yapp, Gifford, Pearl, Gowen, Turner, Goodale, Wright, Norton, Allen, Rice, and Lush. Among various indexes, those of Hansson-Yapp (recommended by Lush for intensive use, 1933) and Rice were used the most.

J. L. Lush (1944) showed that the proposed indexes were particular cases of a general expression:

$$M \;=\; a \;+\; c\,(\overline{X} \;-\; b\,\overline{Y}) \tag{2.1}$$

where

M is the estimated merit of a bull, non-observable,

$\overline{X}$ is the average record of a bull's daughters,

$\overline{Y}$ is the average record of the dams, and

a, b, and c are constants.

Some key assumptions about this general expression are that

- $\overline{X}$ and $\overline{Y}$ are adjusted perfectly for age and season at calving, lactation length, and number of times milked per day.
- The differences in environments between the years in which dams and daughters made records were negligible.

Because $\overline{X}$ and $\overline{Y}$ are the only random variables in the indexes, the manner in which these two components are combined determines the differences between the methods. If a and c are constants used on all bulls, then if b is the same between two methods, then bulls will rank identically for those two methods. The differences between bulls may vary, but the rankings will be identical. In a few proposed methods, c differs from bull to bull based on number of daughter-dam pairs, or some other parameter, and thus, rankings of bulls could be different.

The daughter average proposed by Gifford (1930) was the simplest index, where $a = 0$, $b = 0$, and $c = 1$, thus

$$M \ = \ \overline{X}.$$

However, an added assumption for this index is that the genetic level of dams is equal for all groups of bulls' daughters. This drawback quickly led to daughter-dam comparisons.

A table of the uniquely different methods are given in Table 2.1.

Note that the Hansson-Yapp, Rice, and Allen indexes all use $\overline{X} - 0.5\overline{Y}$, and thus, would rank bulls identically.

Similarly, the Graves and Pearl indexes use $\overline{X} - \overline{Y}$ although Pearl made calculations within lactations rather than across lactations. Wright (1932) also used $\overline{X} - 0.5\overline{Y}$, but the coefficient in front of this term differed among bulls depending on the number of daughter-dam pairs.

Turner (1925) concluded that dams or mates of bulls would have only a minor role in trait inheritance, and thus, daughter averages based on at least five daughters would be an accurate indicator of a bull's transmitting ability. If dams could be ignored, then all daughters' records could be used, even if the records of the dam were not available. More daughters would mean greater accuracy.

Meanwhile, Graves (1925) and others believed that dams did contribute to daughter averages, and good mates usually led to better daughter averages. The expected future progeny average was shown to be the average of the sire M and dam's record average.

Pearl (1930) argued that daughter first lactation records should be compared to their dam's first lactation records, then second lactation records, and so on. Then the average of those differences taken for M. The idea was that this would eliminate the need to require accurate age at calving adjustments across lactations.

Wright (1932) incorporated the number of daughter-dam pairs (q) per bull. His M varied between the breed average when q was 0, and $2(\overline{X} - .5\overline{Y})$ when q

Table 2.1: Proposed Daughter-Dam Indexes

Name	Year	Formula
Högström	1906	$M = -2\ A\ + 4\ (\overline{X}\ - .25\ \overline{Y})$
Pearl	1919	$M = 0\ + 1\ \sum(\overline{X}\ - 1\ \overline{Y})$ Sum within lactations
Graves	1925	$M = 0\ + 1\ (\overline{X}\ - 1\ \overline{Y})$
Hansson-Yapp	1925	$M = 0\ + 2\ (\overline{X}\ - .5\ \overline{Y})$
Turner	1925	$M = 0\ + \frac{100}{85}\ (\overline{X}\ - .15\ \overline{Y})$
Goodale (Mount Hope)	1927	If $\overline{X} > \overline{Y}$, $M = 0\ + 1.429\ (\overline{X}\ - \frac{.429}{1.429}\ \overline{Y})$ If $\overline{X} < \overline{Y}$, $M = 0\ + 3.333\ (\overline{X}\ - .7\ \overline{Y})$
Gifford	1930	$M = 0\ + 1\ \overline{X}$
Wright	1932	$M = \frac{2}{q+2}A\ + \frac{2q}{q+2}\ (\overline{X}\ - .5\ \overline{Y})$ q is no. of pairs
Rice	1933	$M = .5\ A\ + 1\ (\overline{X}\ - .5\ \overline{Y})$ A is breed average
Bonnier	1936	$M = 0\ + \frac{1}{1-\beta}\ (\overline{X}\ - \beta\ \overline{Y})$ $\beta = Cov(\overline{X}, \overline{Y})/Var(\overline{Y})$
Bonnier	1936	$M = 0\ + (1-\alpha)\ \overline{X}\ + \alpha\ \overline{Y}$ $\alpha = (\sigma^2_{\overline{X}} - \sigma_{\overline{X},\overline{Y}})/(\sigma^2_{\overline{X}} + \sigma^2_{\overline{Y}} - 2\sigma_{\overline{X},\overline{Y}})$
Lush	1941	$M = -F\ + 2\ (\overline{X}\ - .5\ D)$ $D =$Average of $\frac{nh^2}{1+(n-1)r}(\overline{Y} - F)$ n is number of records on dam F is the farm-year average
Allen	1944	$M = A\ + 2\ (\overline{X}\ - .5\ \overline{Y})$

was large. Thus, for two bulls with the same progeny average, the bull with the greater q would have the greater M.

Lush et al. (1941) was concerned about the number of records in the dam's averages. There could be 1 to 5 records per dam, and all records were being used, even if the daughter had only one record. His method involved heritability and repeatability as in traditional selection index formulas. The dam's record average was deviated from the farm average, and then this was weighted by

$$\frac{nh^2}{1 + (n-1)r}$$

where h^2 is heritability and r is repeatability, and n is the number of records of the dam. The average of all dams in the farm was then subtracted from each daughters' record average. Therefore, females without dams could be included.

Bonnier (1936) derived formulas that tried to have minimum variance in M. The result was a regression on daughter and dam record averages. Two different methods were derived in this manner. Goodale (1927) considered grouping dams according to their daughter averages. The methods in Table 2.1 show that two pieces of information, daughter average records and dam average records, could be manipulated in different ways, but which method among those was the best? All methods were biased by the environmental differences between dam and daughter records. The methods could also have been affected by the adjustment factors for age, lactation length, and number of times milked per day depending on how those factors were estimated.

2.4 Numerical Example

Assume a breed average production level (age corrected, lactation length of 305-d, and twice a day milkings) of 6500 kg with a variance of 60,000 kg^2. Let the heritability be 0.25 and the repeatability be 0.40. Below (Table 2.2), is the information on two bulls .

Table 2.2: Daughter-Dam Averages for Two Bulls

Bull	No. of pairs	No. of records per dam	Daughter Average, kg	Dam Average, kg
Babe	5	3.4	7800	6600
Jake	10	3.7	6200	6400

Based on simple daughter averages, then Babe would rank above Jake by 1600 kg. The results for each of the indexes is given in Table 2.3.

Every index gives different M for each bull, and the difference between Babe and Jake also varies. However, all agreed Babe was superior to Jake.

2.5 Simulation Study

To determine which index gave closer agreement to the true genetic merit of the bulls, a simulation study was undertaken. Assume a trait with heritability

Table 2.3: Comparisons of Two Bulls

Name	Formula	Babe	Jake
Högström	$M = -2\ A\ + 4\ (\overline{X}\ - .25\ \overline{Y})$	11600	5400
Graves	$M = 0\ + 1\ (\overline{X}\ - 1\ \overline{Y})$	1200	-200
Hansson-Yapp	$M = 0\ + 2\ (\overline{X}\ - .5\ \overline{Y})$	9000	6000
Turner	$M = 0\ + \frac{100}{85}\ (\overline{X}\ - .15\ \overline{Y})$	8012	6165
Goodale (Mount Hope)	If $\overline{X} > \overline{Y}$, $M = 0\ + 1.429\ (\overline{X}\ - \frac{.429}{1.429}\ \overline{Y})$ If $\overline{X} < \overline{Y}$, $M = 0\ + 3.333\ (\overline{X}\ - .7\ \overline{Y})$	8315	5732
Gifford	$M = 0\ + 1\ \overline{X}$	7800	6200
Wright	$M = \frac{2}{q+2}A\ + \frac{2q}{q+2}\ (\overline{X}\ - .5\ \overline{Y})$ q is no. of pairs	8286	6083
Rice	$M = .5\ A\ + 1\ (\overline{X}\ - .5\ \overline{Y})$ A is breed average	7750	6250
Bonnier	$M = 0\ + \frac{1}{1-\beta}\ (\overline{X}\ - \beta\ \overline{Y})$ for $\beta = 0.6$	9600	5900
Bonnier	$M = 0\ + (1-\alpha)\ \overline{X}\ + \alpha\ \overline{Y}$ $\alpha = 0.4$	7320	6280
Lush	$M = -F\ + 2\ (\overline{X}\ - .5\ D)$ $D =$Average of $\frac{nh^2}{1+(n-1)r}(\overline{Y} - F)$ n is number of records on dam $F = 6500$	9057	5944
Allen	$M = A\ + 2\ (\overline{X}\ - .5\ \overline{Y})$	15500	12500

of 0.25, and a repeatability of 0.40. The variance parameters, in kg^2 were

$$\begin{aligned}
\sigma_a^2 &= 15,000, \text{ genetic variance} \\
\sigma_p^2 &= 9,000, \text{ permanent environmental} \\
\sigma_h^2 &= 6,000, \text{ herd-yr variance} \\
\sigma_e^2 &= 30,000, \text{ residual variance}
\end{aligned}$$

with an overall trait mean of $6,500$ kg.

The base population had 300 bulls and 5,000 cows all assumed to be un-

related and non-inbred. The model to simulate a phenotypic record was

$$y_{ijk} = \mu + h_i + a_j + p_j + e_{ijk}$$

where

y_{ijk} is the phenotypic record,

μ is the overall mean of 6500 kg,

h_i is a herd-year contemporary group effect,

a_j is an animal additive genetic value, or true breeding value,

p_j is an animal permanent environmental effect, and

e_{ijk} is a residual effect.

Phenotypes were made for all animals (males and females) even though the trait simulated was 305-d lactation milk yield. The phenotypic records of males were used to select among bulls for the next generation of breeding males, but were not used for any other purpose.

Each generation, 5,000 females were mated to one of 300 bulls to produce one bull calf and one female calf per pregnancy.

The daughter was assigned to one of 200 herds within a generation. Daughters produced a lactation immediately, and could be chosen as a dam for the next generation. Thus, the generation interval was greatly shortened.

Male calves could be selected for breeding in the next generation. All dams and progeny were ranked on their latest phenotype and the top 5000 retained for breeding in the next generation. Similarly all bulls and male progeny were ranked on their "phenotype" and the top 300 kept for breeding the next generation. Selection intensity for females was 5 out of 10, and for males was 3 out of 53.

Pedigrees and inbreeding coefficients were computed for all animals. Six generations of breedings were conducted, and all resulting female phenotypes were used to calculate indexes of bulls.

Each method in Table 2.1 was applied to the simulated data and the M values of the bulls were correlated with the true breeding values of the bulls. Bulls had from 5 to 90 daughter-dam pairs.

Dams and daughters could have from 1 to 5 female progeny. The number of bulls in total ranged from 1072 to 1118. The results for 4 replicates are shown in Table 2.4.

Table 2.4: Comparison of Accuracies

Name	Formula	Replicates 1	2	3	4
ID Bulls		1117	1072	1099	1118
Högström	$M = -2\ A\ + 4\ (\overline{X}\ - .25\ \overline{Y})$	.88	.88	.87	.88
Graves	$M = 0\ + 1\ (\overline{X}\ - 1\ \overline{Y})$	.27	.26	.30	.32
Hansson-Yapp	$M = \ 0\ + 2\ (\overline{X}\ - .5\ \overline{Y})$	.81	.80	.81	.82
Turner	$M = \ 0\ + \frac{100}{85}\ (\overline{X}\ - .15\ \overline{Y})$	.89	.89	.88	.89
Goodale (Mount Hope)	If $\overline{X} > \overline{Y}$, $M = \ 0\ + 1.429\ (\overline{X}\ - \frac{.429}{1.429}\ \overline{Y})$ If $\overline{X} < \overline{Y}$, $M = \ 0\ + 3.333\ (\overline{X}\ - .7\ \overline{Y})$	.69	.68	.71	.72
Gifford	$M = 0\ + 1\ \overline{X}$	.89	.89	.88	.89
Wright	$M = \ \frac{2}{q+2}A\ + \frac{2q}{q+2}\ (\overline{X}\ - .5\ \overline{Y})$ q is no. of pairs	.81	.80	.81	.82
Rice	$M = \ .5\ A\ + 1\ (\overline{X}\ - .5\ \overline{Y})$ A is breed average	.81	.80	.81	.82
Bonnier	$M = \ 0\ + \frac{1}{1-\beta}\ (\overline{X}\ - \beta\ \overline{Y})$ for $\beta = 0.6$	.56 .78	.54 .79	.54 .82	.48 .88
Bonnier	$M = \ 0\ + (1-\alpha)\ \overline{X}\ + \alpha\ \overline{Y}$ $\alpha = 0.4$	.85 .42	.85 .44	.83 .51	.84 .65
Lush	$M = \ -F\ + 2\ (\overline{X}\ - .5\ D)$ $D =$Average of $\frac{nh^2}{1+(n-1)r}(\overline{Y} - F)$ n is number of records on dam $F = 6500$	.89	.88	.88	.89
Allen	$M = \ A\ + 2\ (\overline{X}\ - .5\ \overline{Y})$	.81	.80	.81	.82

2.5.1 Results of Simulation

The simple daughter average gave the highest correlation with the bulls' true breeding values. This likely occurred because the assumption of bulls having mates of equal genetic merit was generally true for all bulls.

Also, herd-year effects were small and random with respect to bulls' daughters. The simulation provided ideal conditions for each bull with no biases. Re-

cords were simulated without age at calving effects, without different lactation lengths, and all the same number of times milked per day. The only biases would be those caused by selection of animals to be parents.

The index developed in section 2.5 was equivalent to Rice or the Hansson-Yapp indices. The better methods were those that subtracted a fraction less than 0.5 of the dam average from the daughter average. Those that subtracted more than 0.5 of the dam average generally had lower accuracy.

Lush's index reduces the influence of the average of dams records, but under certain assumptions would give an index similar to that of section 2.5.

The simulations show that a simple daughter average was more accurate than any daughter-dam comparison. Comparisons made using real data do not have the advantage of knowing the true breeding values of the bulls. Splitting the progeny groups into two, based on the dam averages, and then correlating the resulting index values would give one measure of accuracy. The assumption then would be that bulls should rank similarly for either dam group.

Edwards (1932) did this comparison for five indices, and concluded that the best was the simple daughter average followed by Wright's index, and then the Hansson-Yapp index, similar to the comparisons in the simulation.

The belief, however, was that the adjustment for the dam's average was necessary, and so Daughter-Dam Comparisons persisted for many years. More progress might have been made using Daughter Averages.

2.6 References

ALLEN, N. 1944. A standard for evaluation of dairy sires proved in dairy herd improvement associations. Journal of Dairy Science, 27: 835.

BONNIER, G. 1936. Progeny tests of dairy sires. Hereditas. 22:145.

BONNIER, G. 1946. The sire index. Acta Agriculturae Suecana. 1:321.

EDWARDS, J. 1932. The progeny test as a method of evaluating the dairy sire. Journ. of Agr. Science, 22, p. 811-837.

GIFFORD, W. 1930. Data necessary to prove pure bred dairy sires. Guernsey Breeders J. Sept 1.

GOODALE, H. D. 1927. A sire breeding index with special reference to milk production. Amer. Nat., 671, p. 539-544.

GOODALE, H.D. 1927. Selecting a herd sire. Mt. Hope Farm Pub., Williamstown, Mass.

GOWEN, J. W. 1930. On Criteria for Breeding Capacity in Dairy Cattle. J. Anim. Sci., 47-49.

GRAVES, R. R. 1925. Improving dairy cattle by the continuous use of the proved sire. Journal of Dairy Science, 5: 391.

HANSSON, N. 1913. Kan man med fördel höja medelfetthalten i den av värа nötkrentursstammar och raser lamnade mjöilken? - Centralanst. för försöksväsendet pajordbruksomradet. Meddelande 78, p. 1-85.

HÖGSTRÖM, K. A . 1906. Komjölkens fetthalt, dess normala vaxlingar och arftlighet - Kungl. Landtbruksakademiens handlingar och tidskrift, p. 137-176.

LUSH, J. L. 1931. The number of daughters necessary to prove a sire. Journal of Dairy Science 14: 209–220.

LUSH, J. L. 1933. The bull index problem in the light of modern genetics. Journal of Dairy Science, 16: 501-522.

LUSH, J. L. 1944. The optimum emphasis on dams' records when proving dairy sires. Journal of Dairy Science, 27: 937.

LUSH, J. L. , H. NORTON, ARNOLD, FLOYD. 1941. Effects which selection of dams may have on sire indexes. Journal of Dairy Science, 24: 695-721.

NORTON, H. W., Jr. 1933. Unpublished data referred to by Lush in Journal of Dairy Science, 16: 501-522.

PEARL, R. , GOWEN, J. W. and MINER, J. R. 1919. Studies in milk secretion. Transmitting qualities of Jersey sires for milk yield, butterfat percentage and butterfat. Maine Agr. Exper. Stat. Bull. 281, 89, 165.

RICE, V. A. 1933. Which is the best index? Guernsey Breeders J. 43:238-239 and 261-262.

RICE, V. A. 1944. A new method for indexing dairy bulls. Journal of Dairy Science, 27: 921.

TURNER, C, W. 1925. A comparison of Guernesey Sires. Mo. Agr. Expt. Sta. Res. Bul. 79

WRIGHT, S. 1932. On the evaluation of dairy sires. Proc. Amer. Soc. Anim. Prod., p. 71-78.

YAPP, W. W. 1925. Transmitting ability of dairy sires. Proc. Amer. Soc. Anim. Prod., p. 90-92.

Chapter 3

Selection Index

HORIA GROSU
VALENTIN KREMER
LARRY SCHAEFFER

3.1 Jay L. Lush

Dr Jay L. Lush, the father of animal breeding, was born on January 3, 1896 in Shambough, Iowa. Influenced by Sewall Wright and Sir Ronald A. Fisher, he developed the selection index method during the 1920's and 30's. This work led to his book, "Animal Breeding Plans" first published in 1948, but used years earlier on his students. Over his career he advised 26 MSc students and 124 PhDs. In some way he influenced animal breeding around the globe, either directly or through his students. Iowa State became the center for animal breeding training and continues as such today. Dr Lush died on May 22, 1982. He was the person who changed animal breeding from an "art" into a "science". It is only fitting to acknowledge his contributions towards everything that follows in this book. His main advice to students was "just be productive".

3.2 Index Equation

The selection index for a single trait was used by Wright (1934) and Hazel (1942), and was given various names, like "index of combined selection" (Lush, 1947 and Osborne, 1952), or "breeding value index" (Rasch, 1974), or "family selection" (Gibson, 1995). The first selection indices for cattle were developed by

Harvey and Lush (1952), for the simultaneous improvement of production traits and body conformation. Legates and Lush (1954) developed an index for the intra-farm selection of cows for production using records from the cow and its relatives.

The selection index method was used by Lush (1944) when he determined the appropriate weight to put on the mate's information in the daughter-dam comparisons. The selection index method was also used to derive components of the contemporary comparisons which followed the daughter-dam comparisons. This chapter gives a brief overview of selection index methodology.

A simple selection index equation is essentially a regression equation.

$$M \;=\; b_1 y_1 \;+\; b_2 y_2 \;+ \ldots + b_q y_q$$

where M is the estimated breeding value of the candidate for selection; y_i can be individual records, means of progeny, means of half-sibs, mean of dam records, or any kind of phenotypic measure on relatives of the candidate for selection. All y_i should be records of animals that are genetically related to the candidate for selection. The selection index weights, b_i indicate the proportion of each phenotypic measure that goes into M. The process is to determine the appropriate weights, b_i, for the given information in the index equation. The y_i are assumed to be adjusted for age and season of calving, number of times milked per day, lactation length, and contemporaries. The y_i may also represent phenotypes of different traits. For example, M might be the milk production estimated breeding value of the bull, but some of the y_i may be progeny means for milk yield, fat yield, and protein yield.

The M for every candidate is assumed to be based upon the same types of y_i. Thus, for each candidate, there must be the same y_i available based upon the same number of progeny and records per progeny, and so on. Usually that is not the case and often M are based on different numbers of observations and progeny.

3.3 Estimating The Weights

Good references for the derivation of selection index equations are Bourdon (1997), Cameron (1997), Van Vleck (1993), and Henderson (1963), as well as Lush (1948). To estimate the b_i of the selection index equation, the phenotypic and genetic variances and covariances of the population need to be known. From these, the variances and covariances among the y_i are derived, and the covariances of

the y_i with the true breeding value of the selection candidate. Note that

$$\begin{aligned} M &= b_1 y_1 + b_2 y_2 + \cdots + b_q y_q \\ &= \sum_i b_i y_i \\ &= \mathbf{b}'\mathbf{y} \end{aligned}$$

so that

$$\begin{aligned} Var(M) &= Var(\mathbf{b}'\mathbf{y}) \\ &= \mathbf{b}' Var(\mathbf{y}) \mathbf{b} \\ &= \mathbf{b}'\mathbf{V}_y\mathbf{b} \end{aligned}$$

If M is the estimated breeding value, then let a represent the true breeding value. Then

$$\begin{aligned} Cov(a, M) &= \mathbf{b}'\mathbf{C}, \quad \text{and} \\ Var(a) &= V_a \end{aligned}$$

where $\mathbf{C}$ are genetic covariances of $\mathbf{y}$ with a. One property of b_i might be to maximize the correlation between true breeding value and the index. The correlation is

$$\begin{aligned} \rho(a, M) &= \frac{Cov(a, M)}{(V_a\ V_M)^{.5}} \\ &= \mathbf{b}'\mathbf{C} \times (V_a)^{-.5} \times (\mathbf{b}'\mathbf{V}_y\mathbf{b})^{-.5} \\ \ln(\rho(a, M)) &= \ln \mathbf{b}'\mathbf{C} - .5 \ln V_a - .5 \ln(\mathbf{b}'\mathbf{V}_y\mathbf{b}) \end{aligned}$$

To maximize the function, take derivatives with respect to the unknowns, which in this case is $\mathbf{b}$,

$$\frac{\partial(\ln(\rho(a, M)))}{\partial \mathbf{b}} = \frac{\mathbf{C}}{\mathbf{b}'\mathbf{C}} - \frac{\mathbf{V}_y\mathbf{b}}{\mathbf{b}'\mathbf{V}_y\mathbf{b}}$$

To solve, equate the derivative to 0, and solve for $\mathbf{b}$. To do this, we need to force

$$\mathbf{b}'\mathbf{C} = \mathbf{b}'\mathbf{V}_y\mathbf{b}$$

then

$$\mathbf{V}_y\mathbf{b} = \mathbf{C}$$

are the equations to solve giving

$$\mathbf{b} = \mathbf{V}_y^{-1}\mathbf{C}$$

Another property might be to minimize the variance of prediction errors, $(a - M)$,

$$\begin{aligned} Var(a - M) &= Var(a) + Var(M) - 2Cov(a, M) \\ &= V_a + \mathbf{b}'\mathbf{V}_y\mathbf{b} - 2\mathbf{b}'\mathbf{C} \end{aligned}$$

Take derivatives with respect to $\mathbf{b}$ and set to 0, then solve. The result is

$$\mathbf{b} = \mathbf{V}_y^{-1}\mathbf{C}.$$

Thus, two derivations give the same result, and the properties that the correlation of the index with true breeding value is maximized, and the variance of prediction error is minimized.

3.4 Variance of a Mean

Let a mean be denoted as

$$\overline{y} = (y_1 + y_2 + \ldots + y_n)/n$$

A crude method of obtaining the variance is to square the right hand side of the above equality, then replace squared terms with variances, and cross-products with covariances, giving

$$\begin{aligned} \sigma^2_{\overline{y}} &= (n\sigma^2_y + n(n-1)\sigma_{y,y'})/n^2 \\ &= (\sigma^2_y + (n-1)\sigma_{y,y'})/n \end{aligned}$$

The covariance among records depends on what the single observations are.

1. **Records on One Animal**

   $$\sigma_{y,y'} = r\ \sigma^2_y$$

   where r is repeatability of the trait. Then

   $$\sigma^2_{\overline{y}} = \frac{(1 + (n-1)r)}{n}\sigma^2_y$$

   for n being the number of records on the animal.

2. **Single Records on a Group of Animals**

$$\sigma_{y,y'} \;=\; a_{i,i'}\; h^2\sigma_y^2 \;+\; c_{i,i'}\sigma_y^2$$

where $a_{i,i'}$ is the additive genetic relationship among the animals in the group, such as a group of progeny of a bull, then $a_{i,i'} = 0.25$ for the case of half-sibs, or $a_{i,i'} = 0.5$ for a group of full-sibs from the same sire and dam. All members of the group have the same relationship to each other. Also, $c_{i,i'}$ is an environmental correlation common to records on members of the same group, assumed to be made in the same contemporary group, for example. Then

$$\sigma_{\overline{y}}^2 = \frac{1+(p-1)(a_{i,i'}h^2+c_{i,i'})}{p}\sigma_y^2$$

for p being the number of animals in the group.

3. **Mean of Means**
Suppose $Y = (\overline{y}_1 + \overline{y}_2 + \ldots + \overline{y}_p)/p$, where $\overline{y}_k$ is the average of n records on animal k, and there are n records in each average, then

$$\sigma_Y^2 = [\frac{1+(n-1)r}{n} + (p-1)(a_{i,i'}h^2+c_{i,i'})]\sigma_y^2/p$$

3.5 Covariances with True Breeding Values

The covariance of n records on a progeny of a bull with the true breeding value of the bull (T), is equal to

$$\sigma_{\overline{y},T} = a_{i\alpha}h^2\sigma_y^2$$

where $a_{i\alpha}$ is the additive genetic relationship of the animals in the group with the selection candidate. Heritability is assumed to be in the narrow sense, so that $h^2\sigma_y^2$ is an estimate of the additive genetic variance only.

3.6 Accuracy of Index

If $\mathbf{b} = \mathbf{V}_y^{-1}\mathbf{C}$ gives the weights for the selection index equation, then the accuracy of the index, or the correlation between true breeding value and index value is given by

$$r_{TI} = (\mathbf{C}'\mathbf{V}_y^{-1}\mathbf{C}/h^2)^{.5}$$

3.7 Example Index

Consider the situation where

y_1 is the average of n records on a cow, and

y_2 is the average of m records on the dams of the cow

Both y_1 and y_2 are assumed to be adjusted for age and season of calving, and other environmental effects, and are expressed as differences from their contemporaries. The candidate for selection is the cow, and we wish an index that combines y_1 and y_2 into an estimated breeding value for the cow. Assume the heritability of the trait is 0.25, and repeatability is 0.40.

Assuming the cow and dam are not inbred, then the additive relationship between daughter and dam is 0.5. The matrix $\mathbf{V}_y$ is

$$\begin{pmatrix} Var(y_1) & Cov(y_1, y_2) \\ Cov(y_1, y_2) & Var(y_2) \end{pmatrix},$$

and matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{pmatrix} Cov(y_1, T) \\ Cov(y_2, T) \end{pmatrix}$$

where

$$\begin{aligned} Var(y_1) &= (1 + (n-1)r)\sigma_y^2/n \\ Var(y_2) &= (1 + (m-1)r)\sigma_y^2/m \\ Cov(y_1, y_2) &= (0.5)h^2\sigma_y^2 \\ Cov(y_1, T) &= h^2\sigma_y^2 \\ Cov(y_2, T) &= (0.5)h^2\sigma_y^2 \end{aligned}$$

The variance, σ_y^2, is part of each of the above variances and covariances, and therefore, can be factored out, thus yielding the equations to solve as

$$\begin{pmatrix} (1+(n-1)r)/n & (0.5)h^2 \\ (0.5)h^2 & (1+(m-1)r)/m \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} h^2 \\ (0.5)h^2 \end{pmatrix}$$

Let $n = 1$ record on the cow, and $m = 3$ records on the dam, then the index becomes

$$M \;=\; 0.2299\; y_1 \;+\; 0.1604\; y_2.$$

The accuracy of the index is

$$r_{TI} = [(0.2299(0.25) + 0.1604(0.125))/0.25]^{.5}$$

or $r_{TI} = 0.5569$.

If $n = 2$ records on the cow, and $m = 3$, then the index changes to

$$M \;=\; 0.3323\; y_1 \;+\; 0.1391\; y_2,$$

and $r_{TI} = 0.6339$.

Suppose $h^2 = 0.32$ instead of 0.25, and $n = 1$, $m = 3$, then

$$M \;=\; 0.2896\; y_1 \;+\; 0.1894\; y_2,$$

and $r_{TI} = 0.62$. Thus, increasing number of records on the candidate and increasing the heritability both increase the weight on the candidate's own information, and also increases accuracy of the index.

3.8 Example Index 2

The candidate for selection is now the sire, and the index will be based upon the average of n records per progeny averaged over p total progeny.

$$M \;=\; b_1 Y$$

where

$$Var(Y) \;=\; \left[\frac{1+(n-1)r}{n} + (p-1)(a_{i,i'}h^2 + c_{i,i'})\right]\sigma_y^2/p$$

and

$$Cov(Y,T) \;=\; a_{i,\alpha}h^2\sigma_y^2$$

Thus,

$$b_1 \;=\; Cov(Y,T)/Var(Y).$$

Let $h^2 = 0.25$ and $r = 0.40$, and assume $n = 2$ records per daughter, and $p = 5$ daughters, and $c_{i,i'} = 0$.
Then

$$\begin{aligned} Var(Y) &= 0.19 \\ Cov(Y,T) &= 0.125 \\ b_1 &= 0.65789. \end{aligned}$$

The accuracy is $r_{TI} = [0.65789(0.125)/0.25]^{.5} = 0.5735$.

3.9 Two or More Traits

Hazel and Lush (1942) and Hazel (1943) developed the concept of aggregate genotype of an animal, which is a function of the additive genetic values of the selection candidate for the traits, weighted by their relative economic values, assuming a linear relation between phenotype and genotype.

If a_i is the true breeding value of an animal for trait i, and if w_i is the economic weight for trait i, then the aggregate genotype, H, is

$$H = \sum_i w_i \times a_i = \mathbf{w'a}$$

where $\mathbf{w}$ is the t by 1 vector of economic weights and $\mathbf{a}$ is the t by 1 vector of true breeding values, and t is the number of traits in the aggregate genotype.

Finding weights for a selection index for more than one trait comes by increasing $\mathbf{C}$ from a column vector (for one trait) to a t column matrix for t traits. In addition $\mathbf{C}$ is post multiplied by the relative economic values.

$$\mathbf{V}_y\mathbf{b} = \mathbf{Cw}$$

where b is the vector of weights on each trait.

$$\mathbf{b} = \mathbf{V}_y^{-1}\mathbf{Cw}$$

and

$$M = \mathbf{b'y} = \mathbf{w'C'V}_y^{-1}\mathbf{y}.$$

The aggregate genotype often contains more traits than are included in the selection index. Traits in the selection index may be "indicator" traits which

are highly correlated with one or more traits in the aggregate genotype.

3.10 Restricted Selection Index

Applying selection on one or more traits usually causes correlated responses in all traits that are genetically correlated to traits in the index.

Kempthorne and Nordskog (1959) proposed a restricted selection index where some traits are not to be changed. That meant deriving an index where the covariance between the index and the true breeding value for the trait not to be changed was required to be zero.

3.11 Desired Gains Index

Brascamp (1984) showed that the restricted selection index was a special case of a desired gains index (Pesak and Baker, 1969). The amount of gain desired in a trait or group of traits is pre-determined and then relative economic weights are derived which will give that amount of desired gain.

3.12 References

BRASCAMP, E. W. 1984. Selection indices with constraints. Anim. Breed. Abstr. 52:645-654.

BOURDON, R. M. 1997. Understanding Animal Breeding. Prentice Hall. Colorado State University Press.

CAMERON, N. D. 1997. Selection Indices and Prediction of Genetic Merit in Animal Breedind. CAB International.

CUNNINGHAM, E. P. 1969. The relative efficiencies of selection Indexes. Acta. Agr. Scand. 19:45-48.

FALCONER, D. S. T.F.C., MACKAY. 1997. Introduction in Quantitative Genetics. Fourth edition. Longman Group Ltd.

GIBSON, J. P. 1995. An introduction to the design and economics of animal breeding strategies. Guelph, Canada.

HARVEY, W. R. , J. L. LUSH. 1952. Genetic Correlation between Type and Production in Jersey Cattle. J. Dairy Sci., 35: 199-213.

HAZEL, L. N. 1943. The genetic basis for constructing selection indexes. Genetics 28:476-490.

HAZEL, L. N. , J.L. LUSH. 1942. The efficiency of three methods of selection. J. Heredity. 33:393.

HENDERSON, C. R. 1963. Selection index and expected genetic advance - Statistical Genetics and Plant Breeding.

JAMES, J. W. 1968. Index selection with restrictions. Biometrics 24:1015-1018.[Abstract].

KEMPTHORNE, O. , A.W. NORDSKOG. 1959. Restricted selection indices. Biometrics.

LEGATES, J. E. , J.L. LUSH. 1954. A Selection Index for Fat Production in Dairy Cattle Utilizing the Fat Yields of the Cow and Her Close Relatives. J. Dairy Sci., 37: 744-753.

LUSH, J. L. 1931. The Number of Daughters Necessary to Prove a Sire. J. Dairy Sci., 14:209-220.

LUSH, J. L. , 1944, The Optimum Emphasis on Dams' Records When Proving Dairy Sires. J. Dairy Sci., 27: 937.

LUSH, J. L. , 1947. Family Merit and Individual Merit as Bases for Selection. American Naturalist. 81: 256.

OSBORNE, R. 1957a. The use of sire and dam family averages in increasing the efficiency of selective breeding under a hierarhical mating system. Heredity, 11: 93-116.

OSBORNE, R. 1957b. Family selection in poultry: the use of sire and dam family averages in choosing male parents. Proc. Royal Soc. Edinburg B.66: 374-393.

PESEK, J. , R.J., BAKER. 1969. Desired improvement in relation to selection indices. Can. J. Plant Sci. 49:803-804.

SMITH, H. F. 1936. A discriminant function for plant selection. Annals for Eugenics 7:240.

STHAL, H. C. W. , RASCH, D., SILER, R., VACHAL, J., 1974. Genetica populaţiilor pentru zootehnişti. (Trad. A. FURTUNESCU). E. CERES, Bucureşti, 359 p.

VAN VLECK, L. D. 1993. *Selection index and introduction to mixed model methods.* ISBN 0-8493-8762-0. CRC Pres. Inc. USA.

WRIGHT, S. 1921, 1931, 1921, 1934. Systems of mating and other papers. Iowa State College Press.

Chapter 4

Contemporary Comparisons

H. DUANE NORMAN

4.1 Contemporary Comparison

The contemporary comparison sire evaluation contemporary comparison sire evaluation and the herdmate comparison, initiated in the early- to mid-1950s provided excellent opportunities to accelerate genetic improvement in production traits in the countries where they were introduced, well beyond the rather dismal pace underway at the time. These methods used fairly simple arithmetic calculations. Both procedures more accurately reflected the genetics contributed by the sire than the daughter-dam comparisons. The dilemma was illustrated quite well by A. H. Ward, Director of Herd Improvement of the New Zealand Dairy Board who stated at the Dairy Farmers' Conference, Massey College in 1949 " is there really anything in this breeding business - or are we just fooling ourselves? Can the difference in production between cows and herds be accounted for practically wholly by feeding and management, and high levels of butterfat production reached without worrying about breeding at all? " (New Zealand Dairy Board, 1950)

Robertson and Rendel are considered the originators of the Contemporary Comparison, and Henderson, Carter, and Godfrey are given credit for the Herdmate Comparison, both published in 1954. Searle (1964) took issue with these credits by pointing out that the New Zealand system started in 1950. Other versions of the contemporary comparison were being used in Sweden and Canada by 1956 (Robertson et al., 1956). There were several differences between the early evaluations among these groups. In any case, some of the fundamentals related to

these methods were simmering for several decades in different countries, even back to 1913 when German workers proposed the principle. The belief that success was achievable kept ideas emerging thereafter.

The distinction between the contemporary and herdmate comparison seemed obscure from the beginning and became even more so with time. The terms contemporaries and herdmates often referred to the other animals (cohorts) to which the sire's daughters are compared. To others, contemporaries meant animals performing at the same time or meant those animals of a similar age. Possibly adding to this confusion was that some procedures initiated in countries as a comparison among first lactation records later included all lactation records, and at that point could have been renamed. Nevertheless, if these had been the only differences between these procedures, then whichever method was more effective would have been determined largely by how well environmental effects were identified, estimated, and adjusted for prior to comparing the bull's daughters to their cohorts. Where contemporaries are thought of as animals of a similar age to the bull's daughters (i.e., those having the same or similar parity numbers), any deviation in yield caused by age differences between animals being compared is expected to be smaller for contemporaries than for herdmates.

The contemporary and herdmate comparisons developed and placed in operation had more differences than simply the cohorts to which the bull's daughters were compared, likely because the procedures were developed in different countries by independent groups. There were features that were advantageous to the different procedures, and if there had been more collaboration during the development there might have been modifications to produce a better procedure than any actually used. Some of the early methods that have been referred to as contemporary or herdmate comparisons will be described in this chapter and their advantages and disadvantages discussed. The main weakness of the daughter-dam comparison was that only a small fraction of the apparent superiority of high yielding cows reappears in their daughters. The variation in both dams' and daughters' yields caused considerable distortion in the information because the method was based on the difference in yields between dams and daughters which made the procedure rather ineffective. This was true even if one only compares dam and daughter yields in the same herd-year. Nevertheless, Miller and Corley (1965) showed this distortion was not due to the sires' mates (dams of daughters). They noted that the ranking of bulls was the same whether one chose to adjust or not for mates (correlations of 0.998). Nevertheless, even with the availability of these new procedures, it was usually a decade (or more) before much real progress was observed in the countries which adopted evaluations using these two methods. This was mostly due to a gradual phase-in to the new systems due to limitations at the AI centers (availability of facilities, etc) and the failure to sample enough

young bulls so that stringent selection could be carried out after the progeny tests were completed.

Without question, the comparisons, in their various versions, were welcome developments for genetic progress because they were more effective in reducing environmental effects from herd, year, and, to a lesser extent (depending on how modeled), season. They provided an opportunity for dairy producers to improve their herds genetically by searching for truly superior sires.

4.2 New Zealand

A wide-scale use of the comparison of daughters to the herd average was started in New Zealand in 1950 (Sire Survey and Merit Stud Register, 1955; Searle, 1964) to rank bulls on butterfat yield. When applying these comparisons, the primary goal was to neutralize the herd influence that is reflected in the Daughter Average (DA). The second purpose was to regress for the number of daughters because it was long known that when daughter numbers are limited, sire evaluations will have more variation than their true values. The New Zealand genetic evaluations were calculated by the Herd Improvement Department of the New Zealand Dairy Production and Marketing Board. Lactation records were excluded in the calculation of DA if age at calving exceeded 9 years, although they were still included and used in the herd averages. Those with < 100 days in milk and those between 100 and 200 days in milk which were ruled abnormal were excluded as well. No adjustment was made for lactation length if the record was < 305 days.

Age at calving was recorded only to the nearest year of age (e.g., 2 years, 3 years, etc.), so standardization of fat yield for age was done by adjusting 2- and 3- year old age groups to the mature cow production level, i.e., to those cows aged 4 to 9 years. Cows over 10 years also were factored. Thus, there were only 4 factors per breed. The additive factors tended to underestimate future yield in high producing herds and overestimate it in low producing herds. Therefore, after the 1956-57 calving season, the standardization was changed to use multiplicative factors. Because most cows in New Zealand calved in the spring, season was not considered an environmental source of concern and was ignored.

Sire proofs were done annually for pedigreed bulls and were called Preparatory when they had 6 to 9 daughters and were Official once they obtained 10 or more daughters. Each bull had their evaluations recalculated for 2 more years, and even longer if they still lacked 20 daughters. Evaluations were calculated for grade bulls (non-purebred) if requested by the owner, but these were not pub-

lished. In New Zealand it was assumed that some of the differences between herd averages were due to genetics so the bull was given credit for a portion of the difference between the herd and Breed Average (BA).

This was done by deriving Estimated True Daughter Level (ETDL) with an adjustment for the difference between Contemporary Average (CA) from breed average (BA):

$$\text{ETDL} = \text{DA - 0.9 (CA - BA)}$$

Simply stated, this gave a bull a credit of one extra pound for each 10 pounds of butterfat that the contemporaries produced above the BA, and penalized him the same if they were inferior to BA.

Each bull's daughter fat averages were combined across years by weighting by the number of records in each. They recognized that this caused an upward bias in the DA because it ignored that the later lactations of the bull's daughters were enhanced by culling the poor producers in their first lactations.

The explanatory material indicated that when a bull had a high positive difference from expected on his first crop of daughters, it was often because those daughters represented an above-average sample of all his possible daughters, i.e., this initial survey more often than not was an over-estimate of the true value of this bull for butterfat production.

Therefore, a final adjustment modified the ETDL to account for the number of daughters and lactations upon which the ETDL was based. Failing to adjust for this would have resulted in users selecting bulls too often that were overrated simply because they had been the recipient of favorable chance. By incorporating the adjustment for number of daughters, bulls take on the property (in theory) that their estimates have an equal probability of being above or below their true values. Bulls evaluated appropriately should continue having this property even as their sampling variance decreases with larger progeny numbers (i.e., as these estimates become closer to their true values).

The ETDL was multiplied by the calculated weighting factor. The New Zealand's bull ratings represented transmitting ability instead of breeding value, thus the rating was an estimate of the expected butterfat deviation of future daughters. The weighting factor used for New Zealand was $n/(n + 15)$ which corresponded to a heritabililty (h^2) of 0.25. This was based on assumptions that evaluations were derived from one year's data and that each daughter was in a different herd. The procedure lacked several refinements that were evident in later methods.

4.3 Great Britain

One of the more publicized contemporary comparisons was outlined by researchers in Great Britain, first presented by McArthur (1954) and further explained by Robertson and Rendel (1954). The British defined a contemporary lactation as one that ended in the same recording year in the same herd. The procedure had several differences from the early evaluations from New Zealand. Great Britain used only first parity records and applied no adjustment for age or month of calving so these effects remained unaccounted sources of variation. The restriction of first parity records limited the number of lactations entering the evaluation, but avoided the issue of bias in yield due to culling either daughter or contemporaries based on their performance.

The sire proofs were calculated from 305-day lactation records; milk and fat production after the 305^{th} day was excluded in the lactation records. Eliminated also were records from cows having non-normal lactations (e.g., cow was both suckled and milked, or had lost a mammary quarter by accident) and those from cows sold during lactation. Any lactation record of less than 200 days was rejected. Exclusion of records is always a concern, particularly if there is some incentive to change culling practices because it is known it can influence the bull evaluation that will be published.

Within each herd-year, they specified 2 groups of animals, the daughters of the bull being evaluated and the remaining heifers. At the time of initiation, most of the heifers had originated from natural service. Robertson et al. (1956) made a key point that having 5 daughters of a bull to compare to 1 contemporary was no more accurate than having 1 daughter of a bull with 5 contemporaries. This recognition was incorporated into the procedure they proposed.

The weight w_j given to the Daughter Contemporary Difference (DCD) of the lactation yields in each herd-recording year j was

$$w_j = (nD_j \times nC_j)/(nD_j + nC_j),$$

where nD_j and nC_j are the number of daughters and the number of contemporaries, respectively, in herd-year j. This method of weighting each comparison for accuracy was one of the primary advantages of the British contemporary comparison over the methodology used in most (if not all) herdmate comparisons. A sire's mean DCD across all his daughter groups was derived as

$$\text{Sire DCD} = \sum[w_j \times (DA_j - CA_j)]/\sum(w_j)$$

where $\sum$ indicates summation over j from 1 to number of herd-years and DA_j

and CA_j are the mean yield of daughter and contemporaries in herd-year j. An illustration of the calculation of sire's DCD in 3 herd-years are in Table 4.1.

Table 4.1: Example illustration of calculations

Herd-year	No. Dau.	No. Cont.	w_j	Yield (1000 gallons) DA_j	CA_j	DCD_j	$w_j \times \text{DCD}_j$
1	3	8	2.1818	8.0	7.0	1.0	2181.8
2	2	12	1.7143	10.0	9.0	1.0	1714.3
3	5	10	3.3333	9.0	10.0	-1.0	-3333.3
	10		7.2294				562.8

The sire's mean DCD across all herd-years is $562.8/7.2294 = 77.8$ gallons and was referred to as the Contemporary Comparison. Also, $\sum w_j$ was referred to as the total weight or W which is 7.2294. To derive the bull's estimate of breeding value, one then needed to regress the sire's DCD for the amount of information upon which it was based. The regression suggested by McArthur (1954) was based on a $h^2 = 0.30$ so was $2W/(W + 12.33)$ or in this case, 0.739. Transferring it into a formula for Estimated Breeding Value (EBV), it is:

$$\text{EBV} = 2W/(W + 12.33) \times \text{DCD} = 57.5 \text{ gallons} = (0.739 \times 77.8).$$

Robertson and Rendel (1954) reported this regression factor should use $h^2 = 0.17$.

The Estimated Sire Merit (ESM) was expressed on an actual yield basis (in contrast to a deviation), so was:

$$\text{ESM} = \text{BA} + \text{EBV},$$

which equals 1057.5 if BA was 1,000 gallons. The sire ratings published by the Milk Marketing Board (O'Connor, 1962) were taken a step further and expressed as Relative Breeding Value (RBV), i.e., as a percentage of the BA.

$$\text{RBV} = 100 \times (\text{ESM} \ / \ \text{BA})$$

which equals 105.8%. Robertson et al. (1956) suggested that a W of at least 15 to 20 was necessary to get reasonable accuracy for an evaluation.

Age adjustment factors to correct records prior to their use in any evaluation procedure will never fit all herds perfectly. However, any problem resulting

from imperfect age adjustment should always be less when daughters are compared to cohorts in the same parity, due to more uniformity in ages within the comparison group. In Great Britain, there was no age adjustments applied to the records for sire evaluation because only first lactations were used. When one considers the expected difference in milk or fat yield between cows first calving at 24 versus 34 months (typically 10%), there could have been a few large biases in their evaluations simply caused by age differences.

Robertson et al. (1956) gave the rationale for developing their original method in a British Society of Animal Production article. The primary objective was to examine and judge the effectiveness of 3 sire evaluation methods which seemed to have the best prospects for success at the time. These were the simple average yield of daughters, the comparison of daughters with their dams, and the comparison of daughters with their contemporaries in the same herd-years. They were interested in identifying the bulls transmitting the highest milk production. There seemed to be as much interest in the accuracy of natural-service sampled bulls as the AI bulls, since this was the first step for acquiring proven bulls at the time. More information was thought to be needed on natural service bulls to determine the value of information from the originating herd versus that coming from the daughters sold to other herds.

In the British evaluations, the assumption was made that all differences between herds at different levels of production were due to differences in management rather than to differences in breeding values. To support this assumption, Robertson et al. (1956) selected the daughter records of an AI bull and divided them into 4 groups according to the average production of the herds where they milked. Their results are shown in Table 4.2.

Table 4.2: First lactation daughters of an AI bull distributed according to herd production

Herd Ave.	No. of daus.	Daughter Average	Contemp. Average	Difference
< 800	38	814	668	+146
$800 - 900$	53	934	793	+141
$900 - 1000$	25	1,012	898	+114
> 1000	58	1,190	1,028	+162

There was a considerable difference in performance of both daughters and contemporaries, but the difference between the two was reasonably constant. The authors indicated they had considerable evidence this bull was typical of the

relationship generally observed. If the differences were approximately the same at all herd levels, then herd differences were mainly due to management. They indicated "this was not to say that genetic differences between herds do not exist, but that they are not large enough to interfere seriously with their method". This may have been true at that point in time.

Robertson et al. (1956) compared evaluations based on both DA and contemporary comparison, and noted that the contemporary comparison had much less variation between bulls than did DA due to removal of differences in management levels provided to the daughters. They concluded that their contemporary comparison had lived up to its early promise as a valuable method of evaluating sires for milk traits; simply using DA would be inaccurate, and would result in many good bulls failing to be considered as outstanding. They mentioned that since only first lactations were included, additional studies were needed to determine whether evaluation reports were missing bulls whose daughters matured slowly, and whether bulls which had most of their daughters in only one herd had distorted evaluations because they competed against only a small number of other bulls.

4.4 Cornell University

In 1954, Henderson, Carter, and Godfrey prepared an abstract for the proceedings of the American Society of Animal Science titled "Use of the Contemporary Herd Average in Appraising Progeny Tests of Dairy Bulls". They indicated that herds explained half of the variance in milk yield, and this appeared to be distorting evaluations for AI bulls with small numbers of progeny. They also proposed that a correction for the herd in which the bull's daughters appeared would reduce sampling error and bias. They supported incorporating an adjustment to the difference between daughter yield and herd average (same as done in New Zealand) for the quantity that the herd mean exceeded the population mean. This adjustment was derived from the intra-sire regression of daughter yield on contemporary herd mean (excluding the daughter in question). More details of the Cornell University evaluation system was provided by Henderson (1956), but a more comprehensive description of the procedure was given by Searle (1964) in his review comparing the sire evaluation systems used in New Zealand, Great Britain, and New York state. Research supporting genetic evaluation procedures advanced rapidly at Cornell University because of a number of favorable circumstances.

1. Dairy Herd Improvement (DHI) records were processed at Cornell University,

2. The New York State Artificial Breeders' Cooperative was located at the University, and

3. A capable research and extension staff in the College of Agriculture was dedicated to the cause.

The Cooperative provided financial support to the University and in exchange they received genetic evaluations on their bulls 3 times annually. This contributed to a general atmosphere at Cornell dedicated to determining how evaluations should be done effectively, perhaps more than any previous sustained efforts to date. Numerous research studies were carried out at Cornell University to support these commitments.

Most records from New York and 5 New England states were included in the Cornell evaluations; however, cows with lactations starting prior to 23 months and after 14 years of age were excluded. Also excluded were records initiated by an abortion, or those with less than 100 pounds of milk fat. Prior to the evaluation, records were pre-adjusted for lactation length, frequency of milking, and age at calving. Short records of less than 305 days were provided additional credit as long as the cessation of milking was not due to normal drying-off. For lactations longer than 305 days, the records were truncated at 305 days. Records made from milking more than twice-a-day were factored back to a twice-a-day basis using constants that differed by parity and lactation length.

Lactation records were standardized for age with multiplicative factors for the individual month up to 5 years and for individual years thereafter. Eight-year-old cows were considered mature in Brown Swiss and Guernsey and 6- and 7-year-olds in the other breeds. Accounting for age in months surely provided more accuracy than the approach used in New Zealand where one factor was used for each age in years or in Great Britain which had no age adjustment.

The Cornell sire evaluations used all lactations on a cow, and combined the lactations in an appropriate way. First the weighted mean across individual DA was calculated. The use of multiple records factored in the repeatability (r) of individual records and the number of lactations for each DA. The accuracy of the evaluation was determined by the sum of the weighted records instead of simply by the number of daughters. The weight given to daughter i was:

$$w_i = n_i r/[1 + (n_i - 1)r].$$

If daughter i had 4 records and the repeatability was 0.5, then

$$w_i = (4 \times 0.5)/[1 + (4 - 1)0.5] = 0.80$$

compared to a w_i of 0.50 from those daughters with a single record. Thus the contribution from a daughter with 4 records was equivalent to having 1.6 daughters (0.8/0.5), each with 1 record. The weights for 2 through 7 records per daughter compared to just one record were 1.33, 1.50, 1.60, 1.67, 1.71, and 1.75, respectively. The herdmate average for each sire was derived using the identical weights that combine the information across the daughter yields. The differences derived from using multiple records per cow in this manner are not biased by the effects of culling as long as all the cows' earlier records are present and age adjustments are appropriate.

The Cornell system made additional adjustments to the daughter and herdmate average yields that were not included in the methods from New Zealand or Great Britain. Each lactation record was designated as belonging to 1 of 3 seasons in each year depending on the month of calving. Having 3 seasons per year should have been effective in helping to eliminate most of the seasonal differences related to each herd, but sub-setting always reduces the number of cohorts to which each daughter is compared. These tradeoffs need to be evaluated for each situation. Herd-Year-Season (HYS) averages were calculated within each breed, as well as Year-Season (YS) averages from all records with the intention of estimating the True Production (T) of HYS.

This is an estimate of the conditional mean of T-of-HYS, assuming T-of-HYS and HYS Average are each from a bivariate normal distribution with a mean of YS.

$$\text{T-of-HYS} = \text{YS Avg.} + [(n/(n+1)) \times (\text{HYS Avg.} - \text{YS Avg.})].$$

This adjustment to HYS was effective in improving the cohorts' information thereby producing an improvement in each daughter-herdmate difference. However, because it was then assumed that each of these differences were of equal value, it failed to deliver the overall accuracy across the entire group of daughters that would have been achieved by individually weighting by the combination "number of daughters and contemporaries" as was done in the British method. For example, in the British method, if a daughter had one contemporary, it received about half the weight of a daughter having 20. In the Cornell herdmate comparison, both received near equal weight in producing the evaluation.

The Cornell procedure did not make the assumption that the British method did, that all differences between herd averages were due to management, but gave genetics credit for a portion of the difference. This adjustment to optimize was estimated to be 0.9 as it was in New Zealand's procedure. Applying this, a Herdmate Adjusted Daughter Average (HADA) was derived and used instead of the weighted DA by:

HADA = Weighted DA - 0.9 (Herdmate Avg. - BA).

This gave bulls extra credit when their daughters had herdmates with higher milk yield than BA, and penalized those when the herdmate yield was lower than BA. The BA was derived by summing all production records for each breed in the most recent 3-year period.

One advantage of the Cornell method over the British method was that there were more daughter records used on bulls plus more records used on cohort animals in the same herds. Assuming that standardization for age was done well, this should have increased the accuracy above any alternatives that used only first records. Cornell acknowledged that criticism was frequently directed toward sire selection based on progeny tests because so much emphasis is given to first lactation records which automatically places emphasis on early maturity. They countered by stating that their work (Hickman and Henderson, 1955) indicated that the daughters of different AI sires varied little with respect to increase from first to second lactation.

The regression for converting daughter-herdmate difference to the bulls transmitting ability was done by using the weight $n_D/(n_D + 12)$, where n_D is the effective daughters factoring in the additional information from multiple lactations per daughter. It assumes $h^2 = .31$. The Cornell herdmate comparison was produced regularly until 1972 at which time it was replaced by the Northeast AI Sire Comparison (Everett and Henderson, 1972).

4.5 USDA's Herdmate Comparison (1961)

A herdmate comparison was implemented in 1961 on a national basis in the United States for evaluating milk and fat yields (DHIA Proved Sire List, 1962) and was in many ways patterned after the procedures that were developed at Cornell University. However, the USDA procedure partitioned the daughters into 2 groups, those born as a result of matings through natural service and those resulting from AI service. Evaluations for the two groups were calculated independently. An initial evaluation was calculated for any bull that had 5 or more milking daughters, then a bull was re-summarized if the number of daughters increased by $\geq 50\%$.

Prior to entering milk records into the evaluation procedure, all lactations terminated by culling for low production or due to dairy sales were extended to 305 days with projection factors (Kendrick, 1953). Other lactations were credited as recorded if < 305 days, or truncated at 305 days if the cow milked longer. In addition, all lactations were standardized for age (to a mature basis) and factored

down if milked more than twice daily (Kendrick, 1953). Factors for reducing 305-d, age-corrected records to a twice-a-day milking basis were separate for number of days milked, milking frequency, and 3 age groupings and factors ranged from 0.74 to 0.99.

Having reliable evaluations on all AI bulls was a big step forward because getting accurate information for the first time on bulls already in service around the country, i.e., realizing who the outliers were, provided an opportunity for producers to try to acquire semen to enhance their genetics. For the first time, all national evaluations were derived by comparing the bulls' daughters with other cows experiencing the same management and feeding conditions, i.e., the daughters and their herdmates had an equal opportunity to produce. Because there were rather small genetic differences between herds at the time, the difference of the production of a sire's daughters and their herdmates did not differ markedly from one production level to another.

The USDA version of the herdmate comparison reflected a considerable effort to provide an optimum group of cohorts to which the bull's daughters were compared. The herd-year-seasons were based upon a 5-month moving average. The herdmate average was obtained by averaging all records of the daughters of other sires of the same breed, calving in the same herd, in the same month, and the preceding and subsequent 2 months. For example, if a cow calved in March, her herdmates calved in that same herd from January through May.

The Herd Average (HA) of each lactation of a daughter was adjusted using the number of herdmates (n_H) and the season averages. This estimated the true herdmate average as did the Cornell method although presented in a different way.

$$\text{Adjusted HA} = \text{Season Avg.} + [n_H/(n_H + 1)]\times \text{ (HA - Season Avg.)}$$

The season averages were calculated from the nationwide 305 day, 2X milking, age adjusted lactation yield (mature equivalent) of all DHI cows for the designated breed.

The DA was adjusted to consider the small genetic difference impacting herd averages, i.e., higher producing herds tend to have cows of a slightly higher genetic level. This increased the evaluation of sires used in high producing herds and decreased the evaluation of those used in low producing herds.

$$\text{Adjusted DA} = \text{DA - 0.9 (Adjusted HA - BA)}$$

where the BA used for each evaluation was computed from all DHIA cows of the breed calving in the latest 4 years. Eight breeds were evaluated, Ayrshire,

Brown Swiss, Guernseys, Holsteins, Jerseys, Milking Shorthorn, Red Dane and Red Poll. The procedure used multiple lactations and these were combined using the number of lactations and repeatability of the individual records. These later adjustments, including the incorporation of breed-season averages and BA, were incorporated into the herdmate comparison procedures, following the methodology used by C.R. Henderson and co-workers at Cornell University. To obtain the estimate of the sire's true transmitting ability from daughters scattered across all different levels of herd management, an additional calculation was required to regress for the amount of information (number of daughters and lactations). Its calculation was necessary in order to compare bulls having widely differing amounts of information. This regressed the Adjusted DA toward the BA with the formula:

$$\text{Predicted Avg.} = \text{BA} + [n_D/(n_D + 12)] \times \text{(Adj. DA - BA)}$$

where n_D is the effective daughters factoring in information from multiple lactations per daughter. The constant 12 was the same as used at Cornell at the time, assuming $h^2 = 0.31$. In hindsight, the heritability seems high, but it was certainly superior to not using a regression. In the August 1965 evaluations, the regression was changed to $[n_D/(n_D + 20)]$ (for $h^2 = 0.19$) and stated that recent research showed this revision more accurately reflects the true accuracy of AI sires as determined by the group regression of future daughters on earlier daughters. Producers observing the productivity of their own cows usually expected to see a higher relationship in their herd between the milking daughters of individual bulls and future daughters by the same bulls than what actually occurs.

The explanatory material (DHIA Proved Sire List, 1962) of the sire evaluation procedure indicated the greater the number of comparisons, the more reliable the information. "If a bull had < 9 natural service daughters or < 24 artificial insemination daughters, his published information should be considered preliminary and serve only as indicative and not conclusive evidence of the breeding value of the bull." If the natural service evaluation had from 17 to 25 unselected comparisons or the AI evaluation had 50 to 60 unselected comparisons, additional information would not greatly increase the accuracy of the evaluation. Today, this later suggestion was an overstatement; however, at the time, obtaining large numbers of daughters on each bull was a challenge.

4.6 USDA's Herdmate Comparison (1968)

Several revisions were made to USDA's national sire evaluation procedure in May 1968 (DHIA Proved Sire List, 1968, Plowman and McDaniel, 1968) follow-

ing discussions with industry groups. These changes helped unify the evaluation effort. As a consequence, all dairy cattle breed associations discontinued calculating their own evaluations which they had been doing using a fraction of the cows on DHIA test. Prior to that time, several sire evaluation lists were available to breeders, and multiple lists were a source of confusion for some producers in choosing sires. In 1968 obtaining a sire evaluation became easier. Evaluations were produced if a bull:

1. had ≥ 10 daughters with herdmates for the first time;
2. had semen marketed actively through an AI organization;
3. had enough additional information so that the evaluation might change appreciably; or
4. was the recipient of a special request for a new summary.

Continuing the practice used before, all production records entering the sire evaluation were pre-adjusted for a number of effects. These were primarily the same as before, i.e., extend records to 305 days if terminated by abortion or by sales from the herd (McDaniel et al., 1965), adjusting for age at calving to a mature basis (McDaniel et al., 1967), and factoring all cow lactations to twice-daily milking (Kendrick, 1953). However, changes were made to make these adjustments more precise; i.e., records were extended to 305 days with factors developed separately for individual breeds, parities, and traits (milk and fat), and standardization to 305 days differed for regions and seasons, as well as for each trait (Dairy Herd Improvement Letter, 1967).

Records were not used if they were coded as complete but < 180 days in milk, coded as incomplete but < 15 days in length, coded as initiated by abortion, or those missing ≥ 2 consecutive test periods. Lactation records were used when the lapse-time from calving date to start of the run ≥ 365 days. This avoided inclusion of an abnormally high incidence of incomplete records in early summaries that would otherwise bias the bull performance negatively; i.e., including all the culled daughter records in a bull's evaluation before the other daughters with more traditional length ($\geq$ 305-day) had the opportunity to be included.

Continuing in a similar approach as in the earlier version, albeit with more refinements by using regional information, the HA was adjusted for the number of herdmates by the following:

$$\text{Adjusted HA} = \text{Regional YS Avg} + [n_H/(n_H + 1)] \times \text{(HA - Reg YS Avg)}$$

where n_H is the number of herdmates. The designation of herdmate remained the same as in earlier USDA versions, i.e., a rolling 5-month herd-year-season of those by other sires in the same breed. The regional breed-year-season averages were for each 5-month rolling year-season from records with the same adjustments prior to entry into the evaluation described before. Three regional groups were defined for Ayrshires and Brown Swiss, 4 for Guernseys and Jerseys, and 14 for Holsteins.

The revised 1968 version of the USDA-DHIA herdmate comparison embedded an Adjusted DA as a subset of the formula as:

$$\text{Adj. DA} = \text{DA - Adj. HA} + 0.1\ (\text{Adj. HA - BA})$$

where the BA production was compiled from all DHIA cows of the breed calving in the same rolling 5-month year-season. The 1968 formula appears slightly different than in the 1961 version because in 1968 the evaluations were presented as a deviation instead of on an average milk and fat yield basis, i.e., with the BA included. Again, the purpose was to consider genetic difference in herd averages as higher producing herds tended to have cows of a slightly higher genetic level.

The final adjustment for the Herdmate Comparison (1968 version) was to regress the Adj. DA for the number of effective daughters. The new genetic indication of breeding merit was referred to as Predicted Difference (PD) instead of the earlier term Predicted Average, because the decision was made to present predictions of transmitting ability as deviations from a base zero. Predicted Difference was defined as the expected deviation for milk and fat yield of a bull's daughters from their herdmates in breed average herds. The formula was:

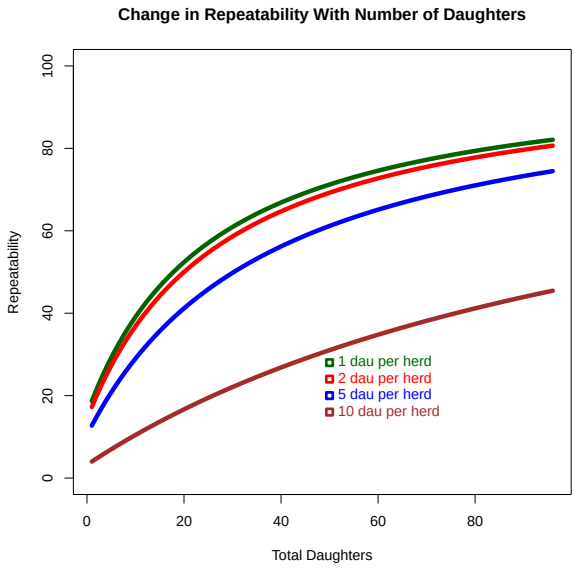
$$\text{PD} = (n_D h^2)/[4 + (n_D - 1)h^2 + 4\sum n_i(n_i - 1)c^2/n_D] \times \text{ Adj. DA}$$

where: n_i is the number of daughters in the i^{th} herd and $\sum n_i$ is the summation across all herds which equals n_D. The c^2 is the residual environmental correlation among half-sibs calving in the same herd due to common environment and/or genetic factors not accounted for by differences among bulls. The regression in the 1968 version included an additional adjustment to consider the distribution of daughters across herds. For purpose of comparing it with the 1961 version, if we temporarily ignore $4\sum n_i(n_i-1)c^2/n_D$, then $(n_D h^2)/[4+(n_D-1)h^2]$ reduces to $n_D/(n_D + 20)$ when $h^2 = 0.19$. The added adjustment to account for the distribution of daughters across herds was included because if a bull's daughters were in a single or few herds, the evaluation was not as accurate as having the same total number of daughters distributed across many herds. The c^2 adjustment

was quite severe in that it was set at 0.14, but it was effective for keeping bulls with daughters in only a few herds from having PD that were far more divergent than their true values. This formulation was from the research of Lush (1933) and Bereskin (1953), although re-estimated using national data. Lush showed that the upper limit of the correlation of estimated and actual breeding value were a function of the environmental correlation among half sisters in the same herd. He pointed out that if the value is 0.14, the maximum repeatability of a single-herd proof was near 0.25.

The consequence of this adjustment for distribution of daughters across herds upon bull Repeatability (and therefore likewise on PD) is illustrated in Figure 4.1. If a bull has 100 daughters in a single herd, his Repeatability was 26%. If an average herd has 12 cows, then if those 100 daughters were distributed 1, 2, 5, or 10 daughters per herd, the Repeatability was 82, 81, 74, and 45%, respectively.

Figure 4.1



Without this adjustment, bulls tested in a few herds (as the majority of bulls had been up to that time) would have had PD with more variation than that represented in their true transmitting ability. The concern was that the true merit of bulls with daughters in few herds having PD with large positive values would have been substantially lower than predicted by their published PD.

The new version contained many improvements beyond those shown in the basic formula. For example, $\sum w_j$ was substituted for n_D to provide effective daughters when some daughters had more than one record, i.e.,

$$w_j = n_j/[1 + (n_j - 1)r]$$

where: n_j was equal to the number of records on the j^{th} cow and r is the repeatability of individual records and was assumed to be 0.50. The accuracy of the bull evaluations was termed the Repeatability (R) of the PD and was simply the regression

$$(n_D h^2)/[4 + (n_D - 1)h^2 + 4\sum n_i(n_i - 1)c^2/n_D],$$

that portion multiplied by the Adj. DA in the PD formula, times 100. When a bull's R was low, the true breeding value of the bull was often substantially different from his PD. In contrast, as a bull's R approached 100%, his true breeding value represented quite accurately the PD. Another improvement of the 1968 version was weighting of the lactation records for their number of days in milk. Obviously the accuracy of the records is related to the length of lactation. The correlations in Table 4.3 were used as the weightings and these varied by the number of monthly tests included in each record and by parity.

Table 4.3: Phenotypic correlations between lactation records with various number of monthly tests and from the complete 305 day lactation

Parity	Number of monthly tests									
	1	2	3	4	5	6	7	8	9	10
1	.72	.83	.88	.92	.94	.96	.97	.98	.99	1.00
> 2	.60	.74	.82	.86	.91	.93	.96	.98	.99	1.00

The formula for this adjustment was:

Adj. Daughter Record = Adj. HA + ρ_i (Projected DA - Adj. HA)

where ρ_i was the phenotypic correlation between records with i months in milk and the 305 day records. For example, projected records of 2-yr old cows with 15 to 46 days in milk (1 monthly test) that were sold for dairy purposes or low production received 72% of the emphasis of others having complete 305-day records.

The herdmate comparison helped educate US dairy producers on a number of principles that continue to be beneficial to them in making genetic improvement to the present. Breeders learned that the predicted transmitting ability

(e.g., Predicted Average and PD) was the key to obtaining genetic improvement. They learned that Repeatability of the evaluation (R) should be only a secondary criterion, used primarily to determine the number of inseminations to obtain from AI sires, once they have been chosen based on their predicted transmitting abilities. This is because the Repeatability only indicates how sure we are that the PD really reflects his true transmitting ability.

Situations that resulted in inaccuracies with the use of the contemporary or herdmate comparisons (or any evaluation) was if daughters were fed or managed differently than the other cows in the herd. Obviously, this arises as a problem primarily when daughters are given better feed and care than most others in the herd, under the deliberate intent to make the bull look better than he would otherwise appear. This is primarily an issue when the majority of daughters are under the influence of management in only a few herds, such as natural service or syndicated bulls. Caution was also warranted if the cohorts were sired by other bulls that deviated substantially from what typically occurred in the population or if they were from only one or a few bulls.

4.7 References

BERESKIN, B. 1963. Effect of genetic and environmental variance on dairy sire evaluation. Ph. D. thesis. Iowa State University.

DAIRY HERD IMPROVEMENT LETTER. Feb. 1967.

DHIA Proved Sire List. 1962. Introductory material. DHIA Sire Summary List, May 1962, ARS-44-122. December 1962.

DHIA Proved Sire List. 1968. Introductory material. DHIA Sire Summary List, May 1968, ARS-44-202. July 1968.

EVERETT, R. W. , C. R. HENDERSON. 1972. The Northeast AI sire Comparison - why? Anim. Sci. Mimeo. Series No 19. Cornell University, Ithaca, NY.

HICKMAN, C. G. , C. R. HENDERSON. 1955. Components of the relationship between level of production and rate of maturity of dairy cattle. J. Dairy Sci. 38: 883-890.

HENDERSON, C. R. 1956. Cornell research on methods of selecting dairy sires. Proc. New Zealand Soc. Animal Prod. 16 Conf., 69-76.

HENDERSON, C. R. , H. W. CARTER, J. T. GODFREY. 1954. Use of the contemporary herd average in appraising progeny tests of dairy bulls. J. Animal Sci. 13:959.

KENDRICK, J. F. 1953. Standardizing Dairy Herd Improvement Association records in proving sires. USDA Bureau of Dairy Industry, Inf. 162.

LUSH, J. L. 1933. The bull index problem in light of modern genetics. J. Dairy Sci. 16:501-522.

McARTHUR, A. T. G. 1954. The assessment of progeny tests of dairy bulls made under farm conditions. Proc. Brit. Soc. Animal Prod., 75-82.

McDANIEL, B. T. , R. H. MILLER, E. L. CORLEY. 1965. DHIA factors for projecting incomplete records to 305 days. In Dairy Herd Improv. Ltr. 41, U.S. Dept. Agr. ARS-44-164, 21 pp.

McDANIEL, B. T. , R. H. MILLER, E. L. CORLEY, R. D. PLOWMAN. 1967. DHIA age adjustment factors for standardizing lactations to a mature basis. In Dairy Herd Improv. Ltr.,43, No. 1, U.S. Dept. Agr. ARS-44-188, 32 pp.

MILLER, R. H. , E. L. CORLEY. 1965. Usefulness of information on mates of sires in artificial insemination. J. Dairy Sci. 48:580-585.

NEW ZEALAND DAIRY BOARD. Sire Survey and Merit Register. 1949, 1950.

NEW ZEALAND DAIRY BOARD. Sire Survey and Merit Stud Register. 1955.

NEW ZEALAND DAIRY BOARD. 37th Annual Report. 1961.

O'CONNOR, L. K. 1962. The use of milk records in selection of male and female beeding stock. Proc. European Assoc. for Animal Prod.

PLOWMAN, R. D. , B.T. McDANIEL. 1968. Changes in USDA sire summary procedures. J. Dairy Sci. 51: 306-311.

ROBERTSON, A. , J. M. RENDEL. 1954. The performance of heifers got by A.I. J. Agr. Sci. 44: 184-192.

ROBERTSON, A. , A. STEWART, E. D. ASHTON. 1956. The progeny assessment of dairy sires for milk: The use of contemporary comparisons. Proc. Brit. Soc. of Animal Prod., 43-50.

SEARLE, S. R. 1964. Review of sire-proving methods in New Zealand, Great Britain, and New York State. J. Dairy Sci. 17: 402-413.

Chapter 5

USDA Modified Contemporary Comparisons

H. DUANE NORMAN
REX L. POWELL

5.1 Introduction

In 1974, USDA introduced a national sire and cow evaluation procedure for the United States incorporating several changes to address the shortcomings of its predecessor. The new procedure was tagged the Modified Contemporary Comparison (MCC) and replaced the USDA Herdmate Comparison which had been used for the US national evaluation in various versions since 1961. MCC evaluations remained in use in the United States for 15 years to estimate genetic merit of bulls and cows for yield of milk and fat and at a later time, protein. MCC had 3 primary advantages over most other evaluations in place at the time of implementation,

1. an adjustment for genetic merit of the contemporaries,
2. a highly effective method of weighting the information within across herds, and
3. the incorporation of genetic merit of ancestors.

In addition, it allowed the continued inclusion of all lactation records at a time when most other countries switched to using only first lactation records and continued doing so for several years.

Three concerns prior to converting to the MCC were

1. over evaluation of older bulls,
2. over evaluation of non-artificial insemination sampled bulls, and
3. misranking of bulls from different segments of the population (for example, from AI organizations or region).

If the contemporaries of an individual bull's daughter were sired by better than average bulls, the superiority of contemporaries' sires biased (in the opposite direction) differences of the daughter from contemporary average. Elimination of this bias by returning credit for the merit of the contemporaries sires delivered equitable.

The MCC model equation included nearly all the effects that were incorporated in 1989 into its successor, the USDA Animal Model, as well as more effects than modeled in most other evaluations even used today, the primary exception being the inclusion of genomic information. Evaluations produced by the MCC were shown in subsequent years to be highly accurate, and supported accelerated genetic progress.

MCC evaluations produced relationships between parent indexes and AI sons' performance that came close to being 100% as predictable as expected; in most prior studies using other evaluation methods, only about one-half to two-thirds the expected regressions were realized (Freeman, 1970; Van Vleck and Carter, 1972; Vinson and Freeman, 1972). Regardless of these somewhat disappointing regressions and correlations in previous studies, they were always positive, so inclusion of parent information could have been helpful, even though responses might not have been as accurate as one hoped.

The MCC was the first method used that incorporated relatives other than daughters into bull evaluations, which may have contributed to its improved predictions across generations. Around the same time MCC was implemented, many countries were converting their evaluation systems to use only first records because of computational limitations as well as from concern surrounding culling bias, but nearly all returned to using multiple records sometime later. When only first lactation records were used, it made it difficult to evaluate economic merit within herds, especially on cows. Using multiple records per cow placed more emphasis on performance throughout the entire productive life of the animals than did evaluations based only on first lactations.

Instead of each daughter's lactation yields being compared with the yields of herdmates of all ages as before, in MCC each was compared primarily with

the yields of the sire-identified contemporaries in the same parity grouping as the daughter (first lactation or later lactations). As in most evaluations, paternal half-sisters were excluded when calculating the contemporary or herdmate averages. Comparing daughters with contemporaries (where ages of animals compared are similar) instead of herdmates helped minimize unwanted variation due to age as age effects on yields in individual herds often differs from the age responses typical for the population. The MCC had the benefit of pre-adjustments using updated national factors for standardizing for age and month of calving which differed by breeds and regions (Miller, 1973; Norman et al., 1974).

All sire-identified herdmates not in the daughter's lactation group were averaged and included as one additional contemporary, i.e., combined with the contemporary average of the daughter's parity group. A correction to account for the average selection bias observed in lactation records due to culling across parities based on the results of Keown et al. (1976) was incorporated into the non-contemporaries averages. Combining the contemporaries yield plus a single average herdmates' yield from the other parity grouping enabled having a Modified Contemporary Average (MCA) for most lactations of nearly every daughter. This allowed daughters to contribute to the evaluation when they had no contemporaries from the same parity grouping, but had one or more herdmates from the other parity group; in these cases the comparison received approximately one-half the weight [similar to $(n_D n_C)/(n_D + n_C)$ of Robertson and Rendel (1954)], where n_D is the number of daughters and n_C is the number of contemporaries, of other comparisons that had dozens of contemporaries.

To be included as a contemporary or herdmate, a cow had to calve in the 5-month period around the daughter's calving month (from 2 months before to 2 months after), same as in the previous evaluation method. A rolling herd-year-season was used in MCC as monthly biases are typically smaller using rolling rather than using fixed herd-year-seasons, because calving dates of daughters and contemporaries are usually closer. Because herds are managed differently, all seasonal effects will never be removed by any genetic evaluation procedure when contemporaries span multiple months. As herd size increases, the consequence of reducing the number of months in the seasonal grouping should be periodically assessed.

5.2 Background Research for MCC

Henderson (1969) outlined some of the assumptions made in the herdmate comparison to the delegates at the National Association of Animal Breeders annual convention and stated "we need more research on whether there are really

problems serious enough to warrant changes in a successful, widely accepted (sire) program." Within a year, a major research effort was undertaken at USDA to answer many of the questions raised. McDaniel et al. (1973) made an attempt to document the advantages and disadvantages of comparing the daughters to contemporaries versus herdmates to determine which would be preferable in ranking sires in the US, based on data between 1966 and 1968. They characterized the distributions of contemporaries and herdmates calving in the same rolling 5-mo herd-year-season in 5 breeds. They found from 6 to 25% of progeny with first lactations did not have contemporaries in the same lactation and therefore up to 25% of the bull's daughters would not contribute to their sire's evaluation if contemporaries were restricted to "true" contemporaries. Similar values for second lactations were 10 to 29%. However, only 1 to 5% did not have herdmates of any age. For those with herdmates, the mean number of contemporaries ranged from 3 to 15 but was 5 to 10 for most groups. Number of herdmates ranged from 12 to 37, but most groups averaged over 20. Mean numbers for second lactation were slightly lower.

Biases from culling were also examined. Biases in milk yield against AI sired first lactations caused by comparing them to selected older cows were small in Ayrshires, Guernseys, and Holsteins (+7 to -10 kg) but were larger for Jerseys and Brown Swiss (-86 and -115 kg). Biases against non-AI sired first lactation cows were of similar magnitudes. Larger biases resulted when first lactations were compared to both their first and second lactation herdmates. They also showed sire summaries based on daughters' first lactations versus herdmates first lactations had a larger sampling variance (about 5 to 40%); these would not have been subject to bias as they were recorded prior to culling. Sampling variances were lower when first lactations were compared to herdmates of all ages, but biases resulting from older cows being the survivors of culling for yield were present. This study showed a contemporary comparison could be designed to use all cows with at least one herdmate of any age and the "average bias" observed from culling could be removed. If all the sires and dams of contemporaries associated with each daughter in a sire evaluation were a random sample of one single genetic population for each breed, complex sire evaluation techniques would not have been so important, although still would have been helpful because of reducing differences due to random variation. In such case, each cow could have been compared only with her contemporary average, and the common effects of herd environment (feeding and management) and genetics of contemporaries would have been relatively small, and each daughter's deviation would represent the genetics of her sire without bias.

The assumption of little genetic differences among contemporary averages was made for contemporary and herdmate comparisons. Although this assump-

tion was not completely true for either, its shortcomings did not prohibit an acceleration of genetic improvement. However, as the rate of genetic improvement gradually increased, more biases in the procedures resulted from discrepancies in genetic merit of contemporaries or herdmates with which a bull's daughters were compared (Miller and Corley, 1965; McDaniel et al., 1974).

No contemporary or herdmate comparsions previously accounted for the genetic merit of the herdmates' sires directly, but instead either assumed no differences or that a beneficial adjustment was made because of a correlation with herd yield. The effects that the actual genetic merit of the herdmates' sires had on daughter deviation from herdmate average and consequently upon sire summaries were examined by Norman et al. (1972) on a comprehensive basis throughout the United States. Lactation records from herds between 1966 and 1968 from 5 breeds on official Dairy Herd Improvement testing were used. Regressions within sire and year of calving of a) daughter yield, b) various herdmate averages, and c) daughter deviations from the various herdmate averages on average PD of the herdmates' sires were calculated.

As average PD for herdmates' sires in Artificial Insemination (AI) increased, so did first lactation daughters' yield (b = 0.35 to 0.73) and AI herdmate average (b = 1.34 to 1.78), but the daughters' deviation from AI herdmate average decreased (b = -0.72 to -1.14). Standard errors showed that these latter regressions did not differ from -1.00 in any of the 5 breeds, a highly desirable property if one considered using this as a means of adjusting for genetic merit of cohorts. Regression using AI contemporaries in both first and second lactation gave results similar to those expected. The results confirmed that differences between bulls in the genetic value of herdmates affected the herdmate comparison and therefore likely caused some misranking of bulls. These regressions provided convincing evidence that adjusting for the average genetic value of the herdmates' sires could significantly increase the accuracy of sire evaluations in use at the time.

In some contemporary or herdmate comparisons, the progeny of different bulls were compared with contemporaries or herdmates using the regression of herd mean (also contemporary or herdmate average) on daughter yield which adjusted for genetic differences in a non-specific manner. To determine the extent to which ignoring or adjusting based on herd level succeeded or failed to satisfy this issue, it was necessary to determine where these shortcuts were creating inequities and to document the extent of the problem. McDaniel et al. (1974) derived the genetic merit of sires of herdmates of bulls' progeny by geographical region, by AI organization, by age of bulls, and by calendar year. Average PD of herdmates' sires in Holsteins varied by 134 kg between the highest and lowest region and by 173 kg between the various AI organizations. Progeny of younger

bulls were compared to herdmates with higher transmitting abilities than were progeny of older bulls due to genetic improvements across years, but differences within the same year were small. These discrepancies were causing systematic errors in sire evaluations computed with the herdmate comparison procedures. Nevertheless, these biases were small compared to the variation among bulls across regions or AI organizations, but eliminating them was feasible by either modifying the contemporary or herdmate comparisons, i.e., using alternative sire evaluation procedures that adjust for genetic variation in herdmates.

5.3 The MCC Procedure

The MCC sire evaluation procedure used the following formula (Norman, 1976; Dickinson et al., 1976):

$$PD_{74} = R \times MCD + (1 - R) \times GA$$

where PD_{74} = Predicted Difference (PD), a measure of predicted transmitting ability set to a 1974 base; R= Repeatability, an indication of the accuracy of the progeny information, actually equivalent to "variation accounted for" in statistical terms; MCD = Modified Contemporary Deviation = DA - MCA + PDsmc (where DA is daughter average, MCA is modified contemporary average, and is the average genetic merit of the sires of modified contemporaries); and GA = group average MCD of bulls with similar pedigree indexes.

The statistical properties of the procedure were unknown. The same could have been said about the contemporary or herdmate comparison. Nevertheless, Quaas and Pollak (1981) reported that predicted transmitting ability would take on best linear unbiased prediction properties after iteration of the "average evaluation of contemporaries' sires". Their approach was remarkably similar to the major components of the MCC method except that in MCC, the sire of the daughter was included in GA. However, each genetic group in MCC usually represented over 100 bulls which meant there was virtually no difference from including or excluding them. Also, there was a very minor difference in the derivation of R.

In July 1983, GA was replaced with a prediction of GA called Ancestor Merit (AM). AM was shown to give a slightly better indication of daughter performance (Wiggans and Powell, 1984). The AM procedure was modified in July 1985 to include protein, to allow for genetic trend within the few bulls without a Pedigree Index (PI), and to eliminate the assumption that the specific trend estimated in the past would continue. The MCC was updated with a new genetic base in 1984 (10 years after initiation) so that dairy producers would be reminded

of the need to raise their selection standards when choosing service bulls. The 1984 base was set by making the weighted average PD of sires of first lactation cows sum to zero for each trait and it has been updated every 5 years since, continuing even after different evaluation methods were implemented.

In 1984, the MCC model was changed slightly to:

$$PD_{82} = R \times (DA - MCA + PDsmc) + (1\text{-}R) \times AM$$

where PD_{82} = PD under the 1982 genetic base and MCD = DA - MCA + PDsmc .

For young bulls, PD_{82} was equivalent to AM because no daughter information was available. As R in the MCC equation approached 1.0 (that is, 100%), the contribution from the pedigree was diluted to the point where PD was nearly the same as the MCD. The average R of contemporaries' sires averaged about 85% for most bulls.

5.4 Innovations of MCC

The term "modified" in MCC actually could have referred to several innovations:

1. the inclusion of a single non-contemporary in the contemporary average,
2. adjustment of each daughter-contemporary difference by the genetics of the contemporaries' sires,
3. inclusion of genetic grouping based on pedigree merit, or
4. the new weighting for increased accuracy.

For the contemporary comparison, the herdmate comparison, and the MCC, the genetic equality of dams of both daughters and contemporaries within herd-year of calving is assumed, and no attempt was ever made to adjust for differences. Norman et al. (1987) calculated the bias that assortative mating caused by documenting the effect of non-randomness of bulls' mates on daughter milk yield. First lactation records for 6 breeds were from cows with calving dates from 1967 to 1984. Correlations between sire PD and dam (mate) transmitting ability Cow Index (CI) for individual years ranged from -0.08 to 0.20. More often than not, the assortative mating was positive. However, the practice of assortative mating

only causes misranking in these procedures if it occurs within herd-years. Fortunately correlations across all records within herd-years indicated no assortative mating for milk yield (0.00 to 0.02) for any breed except Ayrshire (-0.07). The situation in Ayrshires was a result of a deliberate attempt to avoid inbreeding between relatives of Selwood Betty's Commander, an extremely high milk bull, who was used extensively for over 15 years in the breed. Additional details about his extensive use were given by Hudson and Van Vleck (1984). Within-sire regressions of daughter milk yield deviated from contemporary average (which had been adjusted for average PD of contemporaries' sires) on dam Cow Index (merit of mates) by breed were .84 to 1.08 for breed-regions. Expected regressions were 1.00.

Effect of merit of mates on MCC milk evaluations was determined by comparing evaluations from standardized yield with those from standardized yield minus dam's Cow Index. Correlations between evaluations for 4233 Ayrshire, 5275 Brown Swiss, 13,742 Guernsey, 32,572 Holstein, and 13,688 Jersey each rounded to 1.00; average absolute differences in evaluations were 9 to 16 kg, and maximum differences were 49 to 118 kg. Adding a correction to the MCC to account for nonrandomness of mates would have done virtually nothing to increase the accuracy except for about 15 of the 70,000 bulls examined who would have been improved slightly. One bull was biased badly, but he was a progeny test bull who died during his waiting period, and a release of the limited quantity of remaining semen (second round) was only used on elite cows for purposes of obtaining sons for AI service.

Including pedigree information was one of the most valuable improvements in the MCC and yet was one of its most controversial aspects. Previous sire summaries had disregarded ancestor information that could have been valuable in predicting breeding value. The earlier practice was to value the pedigree information when choosing the young animals, but then discard the pedigree information as soon as the first daughters started milking, which seemed irrational. Ironically, information from relatives had been included in cow evaluations for many years. The inclusion of genetic grouping or ancestor merit in bull evaluation made use of pedigree information similar to its use in selection index procedures.

The relative contribution from pedigree information varied inversely with the amount of progeny information available. Pedigree information was weighted with progeny information according to the value of each source. Most bulls had AI sires and maternal grandsires with high R so the young bull's pedigree information was usually equivalent to the information from nine milking daughters, each in a different herd. Initially, pedigree information aided in the choice of bulls to progeny test, but also increased the accuracy of sire summaries, especially when

the bull had low R. Norman et al. (1976) showed that pedigree information was a valuable addition to the evaluations, even for bulls with moderate to high R. Correlation of PD by MCC with future daughter information (those calving later) was 0.13 higher than the herdmate comparison when a bull had ≤ 10 daughters, but was 0.04 higher when they had ≥ 100 daughters. The genetic grouping of the bulls accounted for 49 to 73% of the change from the herdmate comparison to the MCC, while the genetic merit of the herdmate sires accounted for only 26 to 39%.

Powell et al. (1977) estimated the regression of bull's daughter yields on pedigree index based on the MCC procedure and these averaged near 1.0, with corresponding correlations close to the expected value. Before MCC, ironically, research showed pedigree information consistently was less effective than theory indicated it should be (usually ranging from about one-half to two-thirds), probably because evaluations were calculated from procedures without a fixed base, without multiple-population grouping, and usually with no accounting for genetic merit of herdmates' sires. The MCC sire evaluation was not the first method to incorporate grouping, as these had been assigned by stud-year in the Northeast AI Sire Comparison (Everett and Henderson, 1972), but MCC was the first method to group directly on pedigree merit, which proved to be considerably more effective than the Cornell grouping. The evidence for this conclusion was determined by examining the differences in group means; stated another way, if the estimated group means are similar in magnitude, there is little gained from the grouping strategy.

The daughter and contemporary information were weighted according to days in milk, number of daughters and contemporaries, and number of and average Repeatability of contemporaries' sires. This weighting was more accurate than if all records were assumed to be equal in length and to have an infinite number of herdmates. Contemporaries with complete records received more weight in calculating contemporary averages than did those with in-progress or short terminal records (Wiggans and Dickinson, 1985). The statistical procedures for weighting daughter and contemporary information in the calculation of MCC Sire Summary calculations were quite extensive (Norman, 1976; Dickinson et al., 1976). Each cow's records (contemporaries as well as daughters) were weighted by the inverse of their expected variances while forming linear functions of the information. Differences between daughter and contemporary averages also were combined within and across herds by the inverse of their respective variances. Finally, R of the sire was calculated with the formula that accounts for the cumulative variation (within and between herds).

The MCC procedure combined daughter information by weighting for daughter distribution across herds in the presence of residual environmental correlations. This recognizes that a bull's daughters in the same herd were more alike than they were expected to be just from having a common sire. This technique limited the influence of daughter information from any single herd and thus produced more reliable combined information from all herds. Repeatability increased faster with new daughters in new herds than with additional daughters or records in a herd that already had daughters. Thus, the influence of a single herd having a high percentage of a bull's daughters was markedly limited. MCC was the first sire evaluation procedures in use with this feature incorporated. In most countries, release of evaluations was delayed by requiring a bull to have daughters in a large number of herds. This delay in release of genetic information kept the absence of this feature that accounted for residual environmental correlations from causing problems, but slowed the wider use of a number of good bulls.

The previous USDA herdmate comparison procedure also restricted the information coming from an individual herd, but in a less desirable way. It gave equal weight to each daughter with the same number of records regardless of how many daughters were in each herd, and then used the distribution to derive the appropriate R and PD for that method. In that earlier procedure, the R was correct for that average derived, but by adding daughters in a herd that already had many, in certain situations the bull's R actually decreased. In contrast, in the MCC where the information was weighted for the residual environmental correlation in deriving the MCD, adding daughters never resulted in a reduction of R, but of course would not increase R much either if there already were a large number of daughters in the herd adding new ones.

The weighting procedure, as well as the use of pedigree information were two of the reasons that the MCC summaries for bulls with daughters in only a few herds have greater accuracy than did such evaluations using earlier methods. This accuracy was documented for the first 192 bulls that entered AI service based on natural service daughters in a few herds after the MCC implementation in 1974 (Norman et al., 1985). Estimated transmitting abilities before entering AI were compared with those estimated after each bull had hundreds of daughters in a large number of herds. Average PD milk decreased by 0.9 kg; average PD fat remained the same. Any losses incurred by use of individual bulls with evaluations that declined were compensated for by use of other bulls with evaluations that increased. Results examined later (unpublished) showed that additional bulls entering AI with limited herds did not hold up as well as the first 192, but the restriction on R from the weighting forced dairy producers to sample these bulls in many more herds, and eventually there were few that entered AI using this sampling method.

5.5 The Genetic Base

A stepwise genetic base is used in the MCC. A stepwise genetic base (that is, a fixed base for a specified number of years) is a compromise between a fixed and a moving base. Maintaining a fixed (or constant) base over a long period minimizes any problems with comparing bulls over time. A fixed base permits the appropriate adjustment for sires of contemporaries because all bulls are evaluated to the same base within breed. Therefore, sire summaries produced at different times with the same base are directly comparable regardless of the evaluation date. When the base was changed in January 1984, all bulls and cows with summaries that had been released were reevaluated. The weighted average PD of sires of first-lactation cows calving in 1982 was defined as zero for the updated MCC genetic base for each trait and breed. Having all estimates of genetic merit with the same base increases the accuracy of PD's, Cow Indexes and pedigree evaluations. Comparability of these genetic tools increases genetic gain because bulls and cows are selected on the basis of progeny performance as well as pedigree potential. Unnecessary problems are encountered with a moving base when evaluations from different runs are compared.

5.6 Calculation of Ancestor Merit

To compute AM, bulls were assigned to groups based on breed, birth year and pedigree information available (sire and maternal grandsire (MGS), sire only or none). Bulls with only MGS information were included with the group of bulls with no pedigree information available.

Bulls without pedigree information were grouped by bull's birth year. In addition, Holstein bulls with sire information only (no MGS information) were grouped separately. For other breeds, an MGS evaluation was estimated from average Cow Indexes of dams of contemporaries, which were the adjustments to Cow Indexes for genetic merit of contemporaries' dams. This estimate was combined with the known sire evaluation so that these bulls would be included with those that had both sire and MGS pedigree information. Holstein and Jersey bulls with both sire and MGS information (including Jersey bulls with estimated MGS evaluations) were categorized further by the type of sampling program. The sampling programs were natural service versus AI.

Bulls without pedigree information available were grouped by average year of daughter birth for the bull's first evaluation with ≥ 5 daughters. For bulls with ≤ 5 daughters, birth year from the bull's latest evaluation was used. If the birth date was known for a bull with no pedigree information, average birth year of

daughters was constrained to be no more than 3 years after the bull's birth year. Difference between MCD and pedigree index was computed individually for all bulls. Then means of differences weighted by R were computed for each breed, yield trait (milk, fat or protein), birth year and pedigree category. These means were smoothed by regression over 9 consecutive years with the estimate for the middle year retained. Means for the most recent years were calculated from regression coefficients from the last complete set of 9 years. After calculation of means, AM was calculated by:

$$\text{AM} = \text{mean} + \text{bull's Pedigree Index.}$$

5.7 Ranking Percentiles

Genetic progress occurred at an impressive rate (Council on Dairy Cattle Breeding, 2013). Average PD for milk of active AI bulls increased by more than 45 kg per year. A weakness of evaluations from a fixed genetic base is that they do not reflect how each bull compares with the current average bull; i.e., an evaluation by itself does not indicate whether a specific bull is above or below average for the current "bull battery". For example, in the early 1970's, bulls that were +400 kg for PD milk were some of the best bulls available. However, before the base changed in 1984, a bull with +400 kg for PD milk was a candidate for culling.

Dollar percentiles based on economic pricing for milk, fat, and protein (PD$) were added to USDA sire summaries to indicate how bulls compare with active AI (marketed) bulls at any time. Percentiles provided information about the ranking of each bull for PD$ relative to the PD$ of all active AI bulls of that breed. Specifically, a bull's percentile showed the percentage of all active Al bulls that the bull exceeded for PD$. Information for active AI bulls for each breed was sorted by PD$ from high to low. Bulls in the top 1% were in percentile 99; this means that their PD$ was better than the PD$ of 99% of all active AI bulls. Bulls in the bottom 1% were in percentile 0, their PD$ exceeded less than 1% of all active AI bulls.

Competition within and across breeds was intense. Breeders were encouraged to not use bulls with percentiles lower than that of the average active AI bulls, i.e., percentile 50. Dairy producers were encouraged to discontinue services to bulls below percentile 50 and instead shift these services to young AI bulls with high pedigree predictions. Research showed that the daughters of the young bulls were more productive, but also when these were sampled more extensively, it made it easier to identify the top AI bulls for siring the next generation. Use

of percentile rankings helped breeders put active AI bulls in their proper perspective both before and after updates of the genetic base. Eventually percentiles were shifted from PD$ to the index Lifetime Net Income that considered many additional traits.

5.8 Choosing Among Published Genetic Evaluations

Historically, genetic information has often been provided from many sources. Many dairy breeders are still faced with this issue today. One example that faced breeders in several countries in the last decade was whether when buying semen to use the domestic evaluation lists, calculated only from daughters in their own country, or instead to select bulls from the list provided by Interbull. Even in the US, evaluations based on regional data have been published (and still are), with the implication that they were more useful for the dairy producers in the region. Nevertheless, different evaluations use different records, different edits, in addition to different procedures, so it seemed worthwhile to examine the merits of relying on different bull lists, which had seldom been done.

National and regional evaluations published in the United States were compared for their ability to predict standardized milk yield of subsequent daughters (Norman et al., 2005). This was repeated 2 ways. First, there was a comparison in each year between the national (USDA MCC) and a regional list (from Cornell University), both of which had been produced for 14 years, and second, a comparison between national US evaluations and others derived from 4 regional subsets (California, North Central, Northeast, and Southeast) for the same time period calculated using the same procedure (the USDA Animal Model). This later comparison addressed whether there was value from having separate evaluations when there are regional differences in management of the daughters (e.g., large California herds versus smaller Midwestern US herds).

In the first study, correlations between evaluations and first-, second-, and third-parity yields of future daughters were calculated within herd-year-month group. Mean correlations with predicted yield of future daughters across the United States were higher for national Holstein evaluations (0.109, 0.111, and 0.082 for first, second, and third parities, respectively) than for Northeast evaluations (0.098, 0.085, and 0.061); corresponding correlations for predicting only future Northeast daughters yields were similar, meaning the national evaluation worked better even in the Northeast region.

Bull evaluations based on the first 5 parities of daughters that first calved through 1991 in one of the following, either California, North Central, Northeast,

or Southeast regions, as well as from the entire United States were compared with standardized milk yields of daughters that calved later. Correlations with first-, second-, and third-parity yields of future daughters were higher (from 0.001 to 0.011) for national than for regional evaluations. National evaluations were better predictors of future-daughter yield, especially for California and the Southeast. Evaluations based on only first parity were slightly better than those based on the first 5 parities in predicting first-parity yield for 3 of 4 regions but were far less useful in predicting second- or third-parity yield regardless of region.

Regional evaluations included fewer bulls because of limited numbers of daughters in each region. The top 100 bulls for genetic merit for milk yield based on regional rankings were inferior to the top 100 bulls based on national ranking by 25 to 173 kg. Increased reliance on any actual or proposed regional rather than national evaluations would reduce current US genetic gains. These studies make it clear that even if one chooses to select their service bulls from a publication list that might be based on more desirable statistical properties, there is no guarantee that this decision will produce future daughters that are more profitable in his/her own herd. The value of the ranking is not only dependent on the sophistication of the evaluation procedure itself, but is highly dependent on the data edits, standardization for any important effects not included in the evaluation model, what data are available to use, and a number of other reasons.

Effectiveness of various genetic evaluations since 1960 (Council on Dairy Cattle Breeding, 2013) and their widespread acceptance and use have been extremely successful in aiding US production efficiency and have allowed milk to remain affordable to consumers by keeping prices low in relation to the "cost of living" index. Sometimes the acceptance of new genetic discoveries took longer than one felt it should have, but when the guidelines were sound and as more producers eventually adopted them, the adopters became more efficient than the non-adopters, and eventually became the majority.

5.9 MCC Cow Indexes

Cow evaluations were called CI in the MCC (Powell et al., 1976) as they had been in previous evaluations (Miller, 1968). Information for the MCC CI was a by product of the MCC bull evaluation process. The equation exhibited some similarities to the PD equation in that it has 2 parts and combined direct information with pedigree information.

$$\text{CI} = \tfrac{1}{2}[w(\text{MCD}) + (1-w)(\text{Sire's PD})]$$

where w is a weighting factor derived from the amount of information available for the cow and her sire, MCD is the cow's average MCD, and the $\frac{1}{2}$ is necessary to go from a breeding value to transmitting ability.

In January 1981, the dam's CI was included (Powell, 1978). Since the dam's CI includes information from her parents and the sire's PD includes AM, this new CI contains much of the information that would be included from a full pedigree system.

$$\text{CI} = \tfrac{1}{2}[w(\text{MCD}) + (1-w)(\text{Sire's PD} + \text{Dam's CI})]$$

The Repeatability for the CI was a function of the amount of information for the MCD (largely number and length of lactations and number of modified contemporaries) and the Reliabilities for the parents' evaluations. For a given amount of information on the cow, the higher the total R on parents, the lower w becomes. For a given total parental R, the more information on the cow, the higher w becomes. Later it was recognized that the CI even with the same stated base were not quite comparable across time due to an unaccounted for improvement over time in the merit of the dams of contemporaries (Powell, 1984). Thus, an addition was made to the cow part of the equation to add an adjustment for dams of contemporaries (ADC) according to the birth year of the cow. This small adjustment resulted in making the base stationary which continued as each new base was defined.

5.10 References

Council on Dairy Cattle Breeding, 2013. Genetic trend. CDCB website.

DICKINSON, F. N. , H. D. NORMAN, R. L. POWELL, L. G. WAITE, B. T. McDANIEL. 1976. Procedures used to calculate the USDA-DHIA Modified Contemporary Comparison. The USDA-DHIA Modified Contemporary, USDA Prod. Res. Rpt. No. 165, p. 18-34.

EVERETT, R. W. , C. R. HENDERSON. 1972. The Northeast A.I. Sire Comparison, why? Animal Science Mimeo Series, No. 19. Dept of Animal Science, Cornell University, Ithaca, NY.

FREEMAN, M. G. 1970. What has been realized from pedigree selection of dairy bulls? J. Dairy Sci. (Supplement, ADSA meeting, Univ. of FL, Gainesville).

HENDERSON, C. R. 1969. A new sire evaluation method. Proc. 22nd Ann. Conv. National Assoc. Animal Breeders, Milwaukee, WI.

HUDSON, G. F. S. , L. D. VAN VLECK. 1984. Effects of inbreeding on milk and fat production, stayability, and calving interval of registered Ayrshire cattle in the northeastern United States. J. Dairy Sci. 67: 171-179.

KEOWN, J. F. , H. D. NORMAN, R. L. POWELL. 1976. Effects of selection bias on sire evaluation procedures. J. Dairy Sci. 59(10): 1808-1816.

McDANIEL, B. T. , H. D. NORMAN, F. N. DICKINSON. 1973. Herdmate versus contemporaries for evaluating progeny tests of dairy bulls. J. Dairy Sci. 56 (12): 1545-1558.

McDANIEL, B. T. , H. D. NORMAN, F. N. DICKINSON. 1974. Variation in genetic merit of sires of herdmates of first lactation cows. J. Dairy Sci. 57 (10):1234-1244.

MILLER, P. D. 1973. A recent study of age adjustment. J. Dairy Sci. 56:952-958.

MILLER, R. H. 1968. A cow index in selecting dams of bulls. Jersey J. 15:17.

MILLER, R. H. , E. L. CORLEY. 1965. Usefulness of information on mates of sires in artificial insemination. J. Dairy Sci. 48:580-585.

NORMAN, H. D. 1976. Theoretical background for the USDA-DHIA Modified Contemporary Comparison Sire Summary procedure. The USDA-DHIA Modified Contemporary Comparison, USDA Prod. Res. Rpt. No. 165, p. 8-17.

NORMAN, H. D. , B. T. McDANIEL, F. N. DICKINSON. 1972. Regression of daughter and herdmate milk yield on genetic value of the herdmates' sires. J. Dairy Sci. 55:1735.

NORMAN, H. D. , P. D. MILLER, B. T. McDANIEL, F. N. DICKINSON, C. R. HENDERSON. 1974. USDA-DHIA factors for standardizing 305-day lactation records for age and month of calving. USDA ARS-NE-40.

NORMAN, H. D. , R. L. POWELL, F. N. DICKINSON. 1976. Modified contemporary and herdmate comparisons in sire summary. J. Dairy Sci. 59:2155.

NORMAN, H. D. , R. L. POWELL, J. R. WRIGHT. 1985. Changes in evaluation for natural-service sampled bulls brought into artificial insemination service. J. Dairy Sci. 68:1513.

NORMAN, H. D. , R. L. POWELL, J. R. WRIGHT. 1987. Influence of genetic differences in merit of mates on sire evaluation. J. Dairy Sci. 70 (1): 141-157.

NORMAN, H. D. , P. M. VanRADEN, R. L. POWELL, J. R. WRIGHT, W. R. VerBOORT. 2005. Effectiveness of national and regional sire evaluations in predicting future-daughter milk yield. J. Dairy Sci. 88: 812-826.

POWELL, R. L. 1978. A procedure for including the dam and maternal grandsire in USDA-DHIA Cow Indexes. J. Dairy Sci. 61:794-800.

POWELL, R. L. 1984. Genetic base for cow evaluation. J. Dairy Sci. 67:1359-1363.

POWELL, R. L. , H. D. NORMAN, F. N. DICKINSON. 1977. Relationships between bulls' pedigree indexes and daughter performance in the modified contemporary comparison. J. Dairy Sci. 60:961.

POWELL, R. L. , H. D. NORMAN, F. N. DICKINSON. 1976. The USDA-DHIA Modified Contemporary Comparison Cow Index. The USDA-DHIA Modified Contemporary, USDA Prod. Res. Rpt. No. 165, p. 35-40.

QUAAS, R. L. , E. J. POLLAK. 1981. Modified BLUP equations for sire models with groups. J. Dairy Sci. 64:1868.

ROBERTSON, A. , J. M. RENDEL. 1954. The performance of heifers got by artificial insemination. J. Agric. Sci.44:184.

VAN VLECK, L. D. , H. W. CARTER. 1972. Comparison of estimated daughter superiority from pedigree records with daughter evaluation. J. Dairy Sci. 55(2):214-217.

VINSON, W. E. , A. E. FREEMAN. 1972. Selection of Holstein bulls for future use in artificial insemination. J. Dairy Sci. 55:1621-1630.

WIGGANS, G. R. , F. N. DICKINSON. 1985. Standardization of NCDHIP dairy cattle lactation records. National Cooperative Dairy Herd Improvement Program Handbook. Chapter G-2.

WIGGANS, G. R. , R. L. POWELL. 1984. Increasing pedigree contribution to dairy sire evaluation. J. Dairy Sci. 67:893.

Chapter 6

Cumulative Differences

HORIA GROSU
LARRY SCHAEFFER
SORIN LUNGU

6.1 Introduction

After almost two decades of the contemporary comparison method, dairy cattle populations changed due to genetic progress. The assumptions upon which the contemporary comparisons were made became less valid than they were initially because:

1. The tested bulls were no longer a random sample of bulls;
2. Sires originated from several populations of the same breed, and
3. The distribution of sires by farm was no longer random.

Genetic differences between farms caused unwanted changes in the evaluations of bulls. As the bulls were ageing, their newer daughters had contemporaries from younger bulls. Because the young bulls were sons of selected sires, their daughters were expected to have higher genetic potential than the daughters from older bulls. Under these circumstances, the comparison of young bulls with old bulls systematically made the older bulls seem undervalued, because the differences between the two categories of daughters decreased in magnitude. Bar-Anan and Sacks (1974) described a method which attempted to correct the deficiencies of the contemporary comparison method by adjusting for the genetic level of the

sires of the contemporaries. The Cumulative Difference method (CDM), defined the breeding value of a sire consisting of

1. the comparison of daughter averages with contemporaries' averages and
2. the adjustment for the genetic level of the sires of the contemporaries.

6.2 CDM Calculations

Table 6.1 contains data from 3 herds or contemporary groups for sires A through I. Interest is in evaluating sire A. The first number is the number of daughters, and the number in parentheses is the average of those daughters.

Table 6.1: Data to illustrate calculation of the cumulative difference method

Sire	Previous	Herds		
	CD	1	2	3
A		2(5800)	3(6000)	1(6200)
B	-60	1(5500)	1(5400)	1(5700)
C	+60	1(6000)	1(6100)	
D	+150	1(6500)		1(6400)
E	-100		1(5500)	
F	+90		2(6100)	
G	+40		1(5900)	
H	+350			1(7100)
I	-40			1(5400)

Theory gives that the daughter average, $(\mathrm{DA})_{ik}$, of sire i in contemporary group k has expectation equal to

$$\mathrm{DA}_{ik} \;=\; \frac{1}{2}S_i + \frac{1}{2}\overline{M}_{ik} + H_k + \overline{e}_{ik}$$

where

S_i is the sire true breeding value, $\overline{M}_{ik}$ is the average true breeding value of the dams of those daughters (mates of sire i in herd k), H_k is the herd environmental effect, and $\overline{e}_{ik}$ is the average of the residual effects of those daughters.

Similarly, the contemporary average, (CA_{-ik}), of daughters of all sires

except sire i in contemporary group k has expectation

$$\text{CA}_{-ik} \;=\; \frac{1}{2}\overline{S}_{-i} + \frac{1}{2}\overline{M}_{-ik} + H_k + \overline{e}_{-ik}$$

where the terms have a similar definition to those in the daughter average.

Taking the difference, (DIFF), daughter average minus contemporary group average gives

$$\text{DIFF} \;=\; \text{DA}_{ik} - \text{CA}_{-ik} \;=\; \frac{1}{2}(S_i - \overline{S}_{-i}) + (\overline{e}_{ik} - \overline{e}_{-ik})$$

where $(\overline{M}_{ik} - \overline{M}_{-ik})$ is assumed to be zero in the absence of selective matings. The differences are weighted by a factor using the number of daughters and number of contemporaries, w_{ik},

$$w_{ik} = \frac{n_{ik} \cdot n_{-ik}}{(n_{ik} + n_{-ik})}$$

The numerical results for the example data are shown in Table 6.2 .

Table 6.2: Example Calculations for Sire A

Item	Herds		
	1	2	3
DA	5800	6000	6200
CA	6000	5850	6150
DIFF=DA-CA	-200	+150	+50
daughters	2	3	1
contemporaries	3	6	4
w_k	1.2	2.0	0.8
w_k(DIFF)	-240	+300	+40
Previous ave. CD	+50	+20	+100
w_k (CD)	+60	+40	+80

Accumulating the differences and dividing by the sum of the weights gives

$$C_i \;=\; \frac{(-240 + 300 + 40)}{(1.2 + 2.0 + 0.8)} \;=\; +25.$$

Some countries ranked bulls based on C_i values, but other countries preferred to adjust for the number of effective progeny and for heritability of the trait using the following:

$$CC_i \;=\; \frac{\sum w_{ik} \cdot h^2}{(4 + (\sum w_{ik} - 1) \cdot h^2)} \times C_i,$$

which gives

$$CC_A \;=\; \frac{4 \cdot 0.25}{(4 + (3) \cdot 0.25)}(+25) \;=\; +5.26$$

if $h^2 = 0.25$. To adjust for the genetic merit of the contemporaries, the weighted average of the previous cumulative difference, CD, of the sires of the contemporaries, are needed as shown in Table 6.2.

Let A_i be the weighted average of the contemporary sires' CD values.

$$A_i \;=\; \frac{(+60 + 40 + 80)}{(1.2 + 2.0 + 0.8)} \;=\; 180/4 \;=\; +45.$$

Finally, the CD_A for sire A is

$$CD_A \;=\; CC_i \;+\; A_i \;=\; 5.26 + 45 \;=\; +50.26.$$

Unfortunately, sires with lower numbers of daughters were systematically disadvantaged compared to other bulls, and thus, Dempfle(1976) proposed using the regression to account for number of effective progeny and heritability AFTER adding A_i to C_i, rather than BEFORE. Then

$$CA_i = C_i + A_i \;=\; +25 + 45 \;=\; +70$$

followed by

$$CD_i = \frac{\sum w_{ik} \cdot h^2}{(4 + (\sum w_{ik} - 1) \cdot h^2)} \times CA_i \;=\; 0.2104 \times +70 \;=\; +14.73.$$

The results can be seen to be different.

The method of cumulative differences should be iterated until the estimates of CD_i stabilize (i.e. stop changing). The resulting CD of a sire is an estimated transmitting ability. Estimated breeding values are equal to twice the estimated

transmitting ability.

6.3 References

BAR-ANAN, R. , J.M., SACKS. 1974. Sire evaluation and estimation genetic gain in Israeli dairy herds. Anim.Prod. 18: 59-66.

DEMPFLE, L. 1976. A note on the properties of the cumulative difference method for sire evaluation. Anim.Prod. 23: 121-124.

Chapter 7

Regressed Least Squares

HORIA GROSU
LARRY SCHAEFFER
SORIN LUNGU

7.1 Introduction

Robertson and Rendel (1954) first proposed the use of least squares for calculating breeding values of bulls. The idea was further developed by Searle (1964) and Cunningham (1965). Henderson (1952, 1963) developed several procedures for calculation of estimates from weighted least squares for fixed effects.

Regressed least squares involves solving a set of least squares equations for a model with contemporary groups and sires, and afterwards regressing the solutions for sires towards the population mean based on number of effective daughters and heritability. Henderson(1978) discussed the deficiencies of regressed least squares in detail, one of which is that the estimators are not unique, but depend on the restrictions used to obtain a solution to the least squares equations.

The model for an individual animal is

$$y_{ijkl} \;=\; H_i \;+\; \frac{1}{2}S_j \;+\; \frac{1}{2}M_k \;+\; \epsilon_{ijkl}$$

where y_{ijkl} is the observation on daughter l of sire j and dam k making a record in herd i; H_i is the herd (or contemporary group) effect; S_j is the sire true breeding value; M_k is the dam true breeding value; and ϵ_{ijkl} is the residual effect. The phenotype is assumed to be adjusted for age and season of calving, lactation length, and number of times milked per day. Dams are assumed to be random

and of equal genetic quality for all cows with records. Each dam is assumed to have only one progeny in the data. Each daughter is assumed to have only one record. Sires and herds are random factors, but least squares treats these factors as fixed effects in the calculations. Now let $\frac{1}{2}S_j$ be equal to s_j, which is the transmitting ability of the sire. Dams are assumed to have only one progeny each so that it can be combined with the residual term, giving

$$e_{ijkl} = \frac{1}{2}M_k \ + \ \epsilon_{ijkl}.$$

Note that σ^2_ϵ contains $0.5\sigma^2_A$, one half the additive genetic variance. The final model is (eliminating the k subscript)

$$y_{ijl} \ = \ H_i \ + \ s_j \ + \ e_{ijl}.$$

The daughter average, DA, in a particular herd is

$$DA_j \ = \ H_i \ + \ .5S_j \ + \ .5\overline{M}_j \ + \ \bar{\epsilon}_j.$$

Assume there are n progeny in the herd, then

$$Var(DA_j) = 0.25\sigma^2_S + \frac{0.25}{n} \cdot \sigma^2_M + \frac{\sigma^2_\epsilon}{n}$$

If $\sigma^2_S \ = \ \sigma^2_M \ = \ \sigma^2_A$, where σ^2_A is the additive genetic variance, then

$$Var(DA_j) = 0.25 \cdot \sigma^2_A \ + \ \frac{\sigma^2_e}{n}$$

where

$$\sigma^2_e \ = \ 0.25 \cdot \sigma^2_A + \sigma^2_\epsilon$$

also,

$$\begin{aligned} \sigma^2_e &= (1 - 0.25 \cdot h^2)\sigma^2_y \\ \sigma^2_A &= h^2\sigma^2_y \end{aligned}$$

then

$$Var(DA_j) \;=\; (0.25 \cdot h^2 + \frac{1 - 0.25 \cdot h^2}{n})\sigma_y^2.$$

The covariance of the daughter average with the sire's true breeding value is

$$\begin{aligned} Cov(S_j, DA_j) &= 0.5 \cdot \sigma_A^2 \\ &= 0.5 \cdot h^2 \cdot \sigma_y^2. \end{aligned}$$

Let the model be written in matrix notation as

$$\mathbf{y} \;=\; \mathbf{Xh} + \mathbf{Zs} + \mathbf{e}$$

where $\mathbf{h}$ are the herd effects; $\mathbf{s}$ are the sire transmitting abilities; $\mathbf{y}$ are the daughter records; $\mathbf{X}$ relates observations to the herds in which they were made, and $\mathbf{Z}$ relates observations to the sires of the cows; and $\mathbf{e}$ are the residual effects. The variability of the residual effects was commonly assumed to be the same for each herd or contemporary group.

The ordinary least squares equations are written as

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{h}} \\ \widehat{\mathbf{s}} \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix}$$

The order, or size, of the equations is equal to the number of herds plus the number of sires, and so the number of unknowns for which to solve could be many thousands.

7.2 Absorption

Often there were many more contemporary groups than there were sires, and one technique for reducing the size of the equations was to "absorb" contemporary group equations into sire equations. The reduced equations would be

$$\mathbf{Z'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})\mathbf{Z}\,\widehat{\mathbf{s}} \;=\; \mathbf{Z'}(\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})\mathbf{y}$$

and let

$$\mathbf{S} \;=\; (\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})$$

then the equations are written simply as

$$\mathbf{Z'SZ}\,\widehat{\mathbf{s}} \;=\; \mathbf{Z'Sy}.$$

If the absorption is done properly, then the sum of elements in each column of $\mathbf{Z'SZ}$ should be zero, and the sum of the elements of $\mathbf{Z'Sy}$ should be zero. This means that $\mathbf{Z'SZ}$ has a zero determinant, which implies that the matrix does not have an unique inverse, which means that there are an infinite possibility of solution vectors for $\widehat{\mathbf{s}}$. In order to apply regressions to the LS solutions, the restriction to the equations would be to force the sum of the sire solutions to be zero. Thus, the equations are now

$$\begin{pmatrix} \mathbf{Z'SZ} & \mathbf{k} \\ \mathbf{k'} & 0 \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{s}} \\ x \end{pmatrix} = \begin{pmatrix} \mathbf{Z'Sy} \\ 0 \end{pmatrix}$$

where $\mathbf{k}$ is a column vector with all elements equal to 1 with length equal to the number of sires.

Also, the matrix $\mathbf{Z'SZ}$ is of order equal to the number of sires, and almost all elements of this matrix are non-zero. In 1963 with the computer hardware of that day, a direct inverse of this matrix was not possible and could be subject to large rounding errors. Solving the equations was not a trivial exercise in the 1960's. One problem would have been storing $\mathbf{Z'SZ}$ in memory so that it could be inverted.

7.3 Solutions and Estimated Breeding Values

The LS solutions were

$$\begin{pmatrix} \widehat{\mathbf{s}} \\ x \end{pmatrix} = \begin{pmatrix} \mathbf{Z'SZ} & \mathbf{k} \\ \mathbf{k'} & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Z'Sy} \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{C} & \mathbf{c}_k \\ \mathbf{c'}_k & 0 \end{pmatrix} \begin{pmatrix} \mathbf{Z'Sy} \\ 0 \end{pmatrix}$$

The diagonals of $\mathbf{C}$ times σ_e^2 gives the variance of the sire solutions.

Estimated breeding value for sire j is given by

$$\text{EBV}_j \;=\; b_j \cdot \widehat{s}_j$$

where

$$b_j = \frac{Cov(S_j, \widehat{s}_j)}{Var(\widehat{s}_j)}$$

and S_j is the true breeding value of the sire.

From earlier, if $\widehat{s}_j$ is replaced by DA_j, then the regression of sire true breeding value on daughter average (subtracting the average of the contemporaries) would be

$$\begin{aligned} b_j &= \frac{Cov(S_j, DA_j)}{Var(DA_j)} \\ &= \frac{0.5 \cdot h^2 \cdot \sigma_y^2}{(0.25h^2 + \frac{1-0.25h^2}{n})}\sigma_y^2 \\ &= \frac{0.5h^2}{(0.25h^2 + \frac{1-0.25h^2}{n})} \\ &= \frac{0.5h^2 \cdot n}{(0.25h^2 \cdot n \ + \ (1 - 0.25h^2))} \\ &= \frac{2nh^2}{(nh^2 + (4 - h^2))} \\ &= \frac{2n}{n \ + \ \frac{(4-h^2)}{h^2}} \\ &= \frac{2n}{n \ + \ k} \end{aligned}$$

for $k = \frac{(4-h^2)}{h^2}$.

If c_{jj} is a diagonal element of $\mathbf{C}$, and if we replace DA_j by $\widehat{s}_j$, then

$$Var(\widehat{s}_j) \ = \ 0.25 \cdot h^2 \cdot \sigma_y^2 + c_{jj} \cdot (1 - 0.25h^2)\sigma_y^2$$

thus,

$$b_j = \frac{2c_{jj}^{-1}}{c_{jj}^{-1} + k},$$

and

$$\text{EBV}_j \ = \ b_j \cdot \widehat{s}_j.$$

7.4 Numerical Example

Table 7.1 contains example data on 3 bulls and two contemporary groups. The first number is the number of daughters, and the second number in parentheses is the sum total of the daughter records. The resulting LS equations are of order 6 including the restriction to force the sire solutions to add to zero.

Table 7.1: Example Data for Least Squares Method. CG = Contemporary Group

Sire	Herd 1	Herd 2	Sire totals
1	2(9,100)	2(8,000)	4(17,100)
2	5(20,200)	3(13,100)	8(33,300)
3	1(4,500)	5(19,600)	6(24,100)
CG Totals	8(33,800)	10(40,700)	

7.4.1 LS Equations

The full equations are

$$\begin{pmatrix} 8 & 0 & 2 & 5 & 1 & 0 \\ 0 & 10 & 2 & 3 & 5 & 0 \\ 2 & 2 & 4 & 0 & 0 & 1 \\ 5 & 3 & 0 & 8 & 0 & 1 \\ 1 & 5 & 0 & 0 & 6 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \widehat{h}_1 \\ \widehat{h}_2 \\ \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \\ x \end{pmatrix} = \begin{pmatrix} 33,800 \\ 40,700 \\ 17,100 \\ 33,300 \\ 24,100 \\ 0 \end{pmatrix}$$

The equations with the restriction are full rank, and therefore can be inverted. However, the contemporary group equations will be absorbed into the sire and restriction equations (with no effect on the restriction equation). The result is

$$\begin{pmatrix} 3.100 & -1.850 & -1.250 & 1 \\ -1.850 & 3.975 & -2.125 & 1 \\ -1.250 & -2.125 & 3.375 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \\ x \end{pmatrix} = \begin{pmatrix} +510 \\ -35 \\ -475 \\ 0 \end{pmatrix}$$

The effective number of daughters of the sires are the diagonals of the above matrix, namely, 3.100, 3.975, and 3.375 for sires 1, 2, and 3, respectively.

7.4.2 Solutions and EBV

The solutions are

$$\begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \\ x \end{pmatrix} = \begin{pmatrix} 0.1448 & -0.0612 & -0.0836 & 0.3333 \\ -0.0612 & 0.1120 & -0.0509 & 0.3333 \\ -0.0836 & -0.0509 & 0.1345 & 0.3333 \\ 0.3333 & 0.3333 & 0.3333 & 0.0000 \end{pmatrix} \begin{pmatrix} +510 \\ -35 \\ -475 \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \\ x \end{pmatrix} = \begin{pmatrix} 115.7303 \\ -10.9551 \\ -104.7753 \\ 0.0000 \end{pmatrix}.$$

The residual variance would be estimated by taking the total sum of squares of all records minus the reduction due to fitting the model and then dividing by the number of records minus the rank of $(X \quad Z)$.

$$\sigma_e^2 = (310,150,000 \; - \; 308,563,174)/(18 \; - \; 4) \; = \; 113,345.$$

The variances of the estimators of the sire solutions would be

$$\begin{aligned} \text{Sire1} &= 0.1448 \times 113,345 = 16,412.356 \\ \text{Sire2} &= 0.1120 \times 113,345 = 12,694.64 \\ \text{Sire3} &= 0.1345 \times 113,345 = 15,244.9025 \end{aligned}$$

The square root of the above give the standard errors, 128.1, 112.7, and 123.5, respectively. Note that the standard errors are larger than the corresponding sire solutions.

If heritability is assumed to be 0.25, then the EBV for sire 1, $c_{jj}^{-1} = 6.9061$ and $k = 15$ for $h^2 = 0.25$, would be

$$\begin{aligned} \text{EBV}_1 &= \frac{2 \cdot 6.9061}{6.9061 + 15} \times 115.7303 \\ &= 0.63052 \times 115.7303 \\ &= +72.9701 \end{aligned}$$

Likewise for sires 2 and 3, giving

$$\begin{aligned} \text{EBV}_2 &= \frac{2 \cdot 8.9286}{8.9286 + 15} \times -10.9551 \\ &= -8.1754 \\ \text{EBV}_3 &= \frac{2 \cdot 7.4349}{7.4349 + 15} \times -104.7753 \\ &= -69.4452. \end{aligned}$$

7.4.3 Comparison to CDM

Below is Table 7.2 showing information for sire 1, similar to Table 6.2 (page 75), in order to show the linkage between the Cumulative Difference Method and Least Squares.

Table 7.2: Sire 1 Information For CDM

Item	Herds	
	1	2
DA	4550	4000
CA	$4116\frac{2}{3}$	$4087\frac{1}{2}$
DIFF	$433\frac{1}{3}$	$-87\frac{1}{2}$
daughters	2	2
contemporaries	6	8
w_k	1.5	1.6

Notice that

$$\begin{aligned}\sum_{k=1}^{2} w_k(\mathrm{DA}_k - \mathrm{CA}_k) &= 1.5(433.33333) + 1.6(-87.5) \\ &= 510\end{aligned}$$

which is the value in $\mathbf{Z'Sy}$ for sire 1. Also,

$$\sum_{k=1}^{2} w_k = (1.5 + 1.6) = 3.1$$

which is the diagonal of $\mathbf{Z'SZ}$ for sire 1. Next, the CDM gives the sire solution as

$$C_1 = \frac{510}{3.1} = 164.5161$$

This step should be followed with the correction for the sires' C_i of the contemporaries, and the process should be iterated until the C_i stabilize. The correction for sire 1 would be

$$A_i = (1.85 \times C_2 + 1.25 \times C_3)/3.1$$

The new C_1 would be $C_1 + A_1$. After each iteration, the C_i should be forced to sum to zero, by subtracting the mean of the C_i from each C_i. This will give the

solutions to the least squares equations when the solutions stabilize.

After these solutions stabilize, then the last step would be to regress the C_i for numbers of effective daughters and heritability of the trait.

$$\text{EBV}_1 = \frac{2 \cdot 3.1}{3.1 + 15} C_1.$$

The CDM, however, adjusts the C_i for the sires of contemporaries using regressed values rather than other C_i values, and therefore, the EBV from CDM should be different from regressed least squares EBVs.

The least squares equations appear to be a more direct and easier calculation strategy than CDM.

7.5 Summary

Regressed least squares had a number of deficiencies as pointed out by Henderson(1978), but by modifying the procedure the results could be made equivalent to solutions from mixed model equations, although much more difficult to calculate. However, the least squares solutions at least accounted for the level of competition within each contemporary group, and for the unequal distribution of sires' daughters among contemporary groups. The contemporary comparison methods were simplified versions of the least squares method to make calculations practical.

7.6 References

CUNNINGHAM, E. P. 1965. The evaluation of sires from progeny test data. Anim. Prod. 7:231.

HENDERSON, C. R. 1952. Specific and general combining ability. Heterosis. Iowa State College Press, Ames.

HENDERSON, C. R. 1963. Selection index and expected genetic advance. Statistical Genetics and Plant Breeding.

HENDERSON. C. R. 1978. Undesirable properties of regressed least squares prediction of breeding vales. J. Dairy Sci. 61:114-120.

SEARLE, S. R. 1964. Review of sire proving methods in New Zealand, Great Britain and New York State. J.Dairy Sci., 47, 402-412.

ROBERTSON, A. , J.M. RENDEL. 1954. The performance of heifers got by artificial insemination. J. Agric. Sci. Camb. 44: 184-192.

Part II

LINEAR MODEL BASED METHODOLOGIES

Chapter 8

Linear Models

HORIA GROSU
PASCAL A. OLTENACU
LARRY SCHAEFFER

8.1 Charles R. Henderson

Charles R. Henderson was a graduate of Iowa State University under L. N. Hazel and J. L. Lush. He was one of the more statistically minded students, and during his thesis studies he modified least squares equations and found that the resulting solutions were equivalent to Lush's selection index, when the means were known. However, it was not until he met Shayle Searle that he became converted to matrix algebra and with Searle's help proved that mixed model solutions were equivalent to Best Linear Unbiased Prediction estimates. Henderson advocated the linear model approach for many years in his graduate course. Paul Miller (1970) was his first student to apply mixed models to the estimation of age and month of calving adjustment factors for dairy milk production.

Henderson was a dominant force at scientific meetings in the USA, and not many people outside of Cornell University could understand the subtle points of linear models, estimability, and mixed model equations because matrix algebra was not commonly taught. Even for Cornellians, one had to sit through Henderson's course more than once, and after that you became used to his speech and his emphasis on different points. He was always very precise about the assumptions and conditions relating to his comments, and was usually always right. Once Henderson had spoken, the answer was definitive and no further questions

remained.

Cornell University started with a Sire Model for genetic evaluation in 1972, and in 1976 added genetic relationships among the sires. The programs were all written by Henderson, in Fortran, and ran on an IBM 360 machine with 128K of memory. Data were stored on large reels of magnetic tape, and later on 5 or 6 tiered disk drives, as large as a hat box, that gave fast access to data, and made it easier to sort files than on magnetic tapes. All data had to be entered on punched cards.

Henderson was forced to retire in 1976, but he became more productive after he retired, and seemed to be trying to solve all current problems in animal breeding. By 1989, the year he died, the animal model had come into vogue. He had taught the animal model since the 1960's but he called it the individual cow model. The cow model was dubbed the animal model by Quaas and Pollak (1980). Nearly every country in the world now uses a linear model approach (animal model) and applies mixed model equations. The mixed model equations are also used with genomic data today. It is appropriate to acknowledge this huge contribution of C. R. Henderson to animal genetic evaluations, in all species of livestock.

8.2 Best Linear Prediction

A mixed linear model, in matrix notation, is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

where

$\mathbf{y}$ is an $N \times 1$ vector of phenotypic observations on the traits and animals of interest,

$\mathbf{b}$ is a $p \times 1$ vector of fixed effects (environmental variables that can be identified) that influence the phenotypic observations,

$\mathbf{u}$ is a $q \times 1$ vector of random effects (like contemporary groups, litters, and permanent environmental effects) that influence the phenotypic observations,

$\mathbf{e}$ is an $N \times 1$ vector of residual effects, and

$\mathbf{X}, \mathbf{Z}$ are matrices that relate elements of $\mathbf{b}$ and $\mathbf{u}$ to $\mathbf{y}$. Thus, $\mathbf{X}$ has order $N \times p$, and $\mathbf{Z}$ has order $N \times q$.

Random factors in a model follow a distribution function, usually a normal distribution, which has a mean and variance structure. To find the Best Linear Predictor of $\mathbf{u}$ the distribution, means, and variance structures must be known. That is,

$$\begin{aligned} Var(\mathbf{y}) &= \mathbf{V} \\ Var(\mathbf{u}) &= \mathbf{G} \\ Cov(\mathbf{u}, \mathbf{y}) &= \mathbf{GZ}' \\ E(\mathbf{y}) &= \mathbf{Xb} \\ E(\mathbf{u}) &= \mathbf{0} \end{aligned}$$

The Best Linear Prediction (BLP), is

$$\widehat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

which does not depend on the distribution functions of $\mathbf{y}$ or $\mathbf{u}$. This formula is exactly the formula used in the selection index method. This formulation has some serious deficiencies.

1. The means of the random variables must be known. Note that $\mathbf{Xb}$ includes the effects of age at calving, herd-year-season of calving, genetic groups, number of times milked, lactation lengths, and possibly other factors. Practically, the means for all of these effects are not known perfectly. The means depend on the sample of data that is available.

2. In some cases the variances and covariances are unknown, such as when analyzing a new breed or new trait for the first time.

3. Inversion of $\mathbf{V}$ with very large numbers of observations is impossible to calculate.

8.3 Best Linear Unbiased Prediction

The model is the same as in the previous section. Now assume that $E(\mathbf{y})$ is not known. Logically, we need to replace $\mathbf{Xb}$ with an estimate of $\mathbf{Xb}$. The variance and covariance structures and parameters are still required to be known. The predictor of $\mathbf{u}$ is to be a linear function of $\mathbf{y}$, say $\mathbf{L}'\mathbf{y}$, and we must require that

$$E(\mathbf{u}) = E(\mathbf{L}'\mathbf{y}).$$

By minimizing

$$E(\mathbf{L}'\mathbf{y} - \mathbf{u})^2$$

subject to requiring the predictor to be unbiased, then the resulting Best Linear Unbiased Prediction (BLUP) is

$$\widehat{\mathbf{u}} \;=\; \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} \;-\; \mathbf{X}\widehat{\mathbf{b}})$$

where

$$\widehat{\mathbf{b}} \;=\; (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\; \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Thus, $\widehat{\mathbf{b}}$ is the Generalized Least Squares (GLS), of $\mathbf{b}$. In selection index, often raw means for some fixed factors were used, but from this development, the better alternative would have been to use GLS estimates of those fixed factors.

Note again, that the predictor involves the inversion of $\mathbf{V}$. Henderson frequently knew where developments were headed, so he would try ideas out on empirical data. In doing this he found that modifying least squares equations gave selection index results, or BLUP solutions. Thus, he deduced that there must be a mathematical proof or derivation to go from

$$\widehat{\mathbf{u}} \;=\; \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} \;-\; \mathbf{X}\widehat{\mathbf{b}})$$

to his modified least squares equations. To prove this, Henderson needed to show that

$$\mathbf{V}^{-1} \;=\; \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1}$$

where

$$\mathbf{T} \;=\; (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \;+\; \mathbf{G}^{-1})^{-1}.$$

You multiply $\mathbf{V}\mathbf{V}^{-1}$ and show that the result is an identity matrix.

The story Henderson gave was that he left the problem on a piece of paper at Shayle Searle's desk while he went for a coffee break (sometime in 1967 while Henderson was in New Zealand).

When Henderson returned to his office, there was a proof given on the same paper. So, even though Henderson published his modified least squares equations in 1950, it was not until 1967 that he proved the solutions were equivalent to BLUP.

8.4 Mixed Model Equations

The Mixed Model Equations (MME) of Henderson (1973) are

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

Thus, in order to apply these equations we need to know how to construct $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}^{-1}$, $\mathbf{R}^{-1}$, and $\mathbf{y}$. The MME are general for any linear model with fixed and random effects. The linear model could be a sire model, an animal model, a test-day model, or genomics model. Each model might have a different $\mathbf{X}$ matrix, for example, but we know where it goes in MME, and the same for the other components of MME. Hence in the following chapters we will show how to create these components for different models. Once the parts are known, then the next step is to set up and solve the MME.

The solutions to the MME are

$$\begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-} \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

Let

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xz} \\ \mathbf{C}_{zx} & \mathbf{C}_{zz} \end{pmatrix},$$

then Henderson (1973) has shown

$$\begin{aligned} Cov(\widehat{\mathbf{b}}, \widehat{\mathbf{u}}) &= \mathbf{0} \\ Cov(\widehat{\mathbf{b}}, \mathbf{u}) &= \mathbf{C}_{xz} \\ Cov(\widehat{\mathbf{b}}, \widehat{\mathbf{u}}-\mathbf{u}) &= \mathbf{C}_{xz} \\ Var(\widehat{\mathbf{u}}) &= \mathbf{G}-\mathbf{C}_{zz} \\ Var(\widehat{\mathbf{u}}-\mathbf{u}) &= \mathbf{C}_{zz} \end{aligned}$$

The diagonals of $\mathbf{C}_{zz}$ give the variances of prediction error for the predictors of $\mathbf{u}$. Because the size of the MME are often too large to obtain an inverse of the coefficient matrix, then the diagonals of $\mathbf{C}_{zz}$ are approximated, and various methods have been proposed. The solutions to MME are computed using iteration techniques, of which there are several possible variations (Schaeffer and

Kennedy, 1986). Sometimes the models allow simplifications to be made which make the calculations easier to complete.

The residual variance can be estimated using

$$\widehat{\sigma}_e^2 \;=\; (\mathbf{y}'\mathbf{y} \;-\; \widehat{\mathbf{b}}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \;-\; \widehat{\mathbf{u}}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y})/(N - r(\mathbf{X}))$$

where $r(\mathbf{X})$ is the rank of the matrix $\mathbf{X}$ which is the number of linearly independent columns, also known as a degrees of freedom for fitting the fixed effects of the model.

8.5 Linear Models

The readers should be clear that BLUP and MME are just tools that are used to derive predictors that have certain desirable properties. The more important aspect in genetic evaluation is the description of the linear model. What factors are being considered? What parameters are being used? Are all important factors included in the model? These are the questions that need to be properly answered.

A person might tell you that BLUP was used, but when they disclose their model you discover that one or two important factors have been omitted, and thus, the sire EBV could be severely biased. So that any method could have been used to get predictors, and they all might give biased EBV, due to a poor model. The critical point is finding the most useful model. Being able to write a good model requires experience, and trial and error. Model comparisons are a common task in animal breeding.

A complete linear model consists of

- an equation giving a list of all of the factors to be considered in the analysis,
- the distributions of the random variables in the model including the expectations and variance structure of each factor, and
- the assumptions and limitations imposed by the data and computer hardware.

Only if all of these parts are provided can the quality of the model be judged. There is a saying in statistics that all models are wrong, and that is definitely true. However, when we write a linear model we are trying to provide the best approximation to our vision of the correct model. Another way to put it, is that the true underlying model is totally unknown. The best model may not be

linear and may be a complex amalgamation of several simple processes. Animal breeders have, so far, worked primarily with linear models because they are easy to understand, and they perform very adequately in the majority of situations.

8.5.1 Fixed or Random Factors

A difficult part of modelling is determining which factors are fixed and which are random. When a factor is random we usually think of a population of levels of that factor with an overall mean (usually zero) and with a certain variance. The levels that appear in our data are assumed to be a random sample of levels from the overall population. There is no process that limits which samples appear in our data. Sires have always been considered to be a random factor. Even though sires are selected for progeny testing, usually the sire and dam have been selected to produce the young bull, but Mendelian sampling has produced a random progeny from that mating. Thus, the young bull is still a random entity, in that sense.

Herd-year-seasons or contemporary group effects are also a random factor. The animals within a contemporary group arrive there by chance. The environment that exists in that contemporary group is a random event due to a combination of location, weather, management, and composition of animals within the group. That particular combination will never exist again. The contemporary group has a definite genetic level by virtue of the genotypes of the animals within the group. In the contemporary comparison method, however, the assumption was that the genetic level was fairly equal across contemporary groups. In an animal model we can account for both the environmental and genetic levels of the contemporary groups because we know which animals are in each contemporary group. In a sire model, we need the assumption that the genetic levels of contemporary groups are equal, or at least the effect is not very large.

When sire models were first implemented, Henderson made contemporary groups as fixed factors. According to his selection bias theory, any association between genetic level of the herd-year-season and the sires used in that herd could be removed by treating herd-year-season effects as fixed. Thus, from that point on, everyone believed Henderson was correct and contemporary groups have always been fixed factors. However, no one has ever demonstrated the existence of the association, nor how large of an effect it was. Secondly, Henderson's selection bias theory has been disclaimed by several scientists in recent years and therefore, treating contemporary groups as fixed does not necessarily remove bias. Lastly, sire models are no longer used, and it has not been shown that any bias exists with animal models. Thus, the best course of action, until proven otherwise, is to

let contemporary groups remain random, as they were originally classified.

If contemporary groups are random, then models need a fixed factor like year-month of calving to account for time trends. This will be discussed further in the chapter on animal models. The key point of this discussion is that determining factors to be fixed or random is not always simple.

8.6 Relationship Matrices

Besides BLUP and MME, Henderson (1953) developed three methods of estimating variance components (known as Henderson's Methods 1, 2, and 3) during his PhD thesis project.

These methods were unbiased procedures that were relatively easy to calculate in those days, but which often yielded negative estimates of variances. However, his methods were used extensively until the 1970's, when they were replaced by likelihood methods which kept estimates of variances within their allowable parameter space. Henderson's three methods provided heritability and repeatability estimates for use in genetic evaluation methods. His variance estimation methods indirectly impacted on genetic evaluation methods.

The third most important discovery of Henderson's career, after MME and the methods for estimating variance components, was the discovery of a method to invert a matrix of additive genetic relationships using only a pedigree list and a simple set of rules (Henderson, 1976). This discovery made it possible to account for genetic relationships among individuals within and across herds in both sire and animal models. Without this discovery, animal models could have been delayed another ten years. Although Henderson's paper only talked about matrices for non-inbred animals, several subsequent papers were written on the fast calculation of inbreeding coefficients, such as Meuwissen and Luo (1992). This enabled inbred animals to be included in analyses.

Henderson wanted to include relationships among bulls in the Northeast AI Sire Comparison at Cornell University. He inverted many small example relationship matrices and eventually noticed a pattern. Once he verified the pattern with other examples, then it was not long (a day) before he proved the results mathematically.

Today everyone uses relationship matrices in genetic evaluation and they use MME to obtain the evaluations without referring to the original Henderson publications. His discoveries have become commonplace and accepted while the name Henderson fades into ancient history.

8.6.1 Sire-MGS Relationships

Below (Table 8.1), are a few sire-MGS pedigrees.

Table 8.1: Example pedigrees

Bull	Sire	MGS
1	-	-
2	-	-
3	-	-
4	1	2
5	2	3
6	4	-
7	-	1

Notice that bulls can have both parents unknown, both parents known, or either sire or MGS unknown. Following Henderson (1975), let

$\delta = 1$ if both parents are unknown
$\delta = 16/11$ if both parents are known
$\delta = 4/3$ if the MGS is unknown, and
$\delta = 16/15$ if the sire is unknown.

The inverse of the relationship matrix, commonly denoted by $\mathbf{A}^{-1}$, begins as a null matrix. Then each animal in the pedigree list is processed, one at a time, to add numbers to the appropriate locations in $\mathbf{A}^{-1}$. The rules for adding to $\mathbf{A}^{-1}$ are

	sire	MGS	bull
sire	.25 δ	.125 δ	$-.5$ δ
MGS	.125 δ	.0625 δ	$-.25$ δ
bull	$-.5$ δ	$-.25$ δ	δ

Bulls 1, 2, and 3 have $\delta = 1$ and both parents are unknown, so that δ is simply added to the diagonal elements for the rows corresponding to bulls 1, 2, and 3. Bulls 4 and 5 have $\delta = 16/11$ because both parents are known. For bull 4, add δ to the diagonal for bull 4, $.25\delta$ to the diagonal for bull 1, and $.0625\delta$ to the diagonal for bull 2. Then add $.125\delta$ to the off-diagonals between bulls 1 and 2, subtract $.5\delta$ to the off-diagonals between bulls 4 and 1, and subtract $.25\delta$ to the off-diagonals between bulls 4 and 2. The results to this point are

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{15}{11} & \frac{2}{11} & 0 & -\frac{8}{11} & 0 & 0 & 0 \\ \frac{2}{11} & \frac{12}{11} & 0 & -\frac{4}{11} & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -\frac{8}{11} & -\frac{4}{11} & 0 & \frac{16}{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

For bull 6, the MGS is unknown, so additions are made to the bull diagonal, bull 4 diagonal, and a subtraction to the off-diagonal between bulls 6 and 4. For bull 7, the sire is unknown, so additions are made to the diagonals for bull 7 and bull 1, and to their off-diagonals. This completes $\mathbf{A}^{-1}$, as shown below.

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{136}{165} & \frac{2}{11} & 0 & -\frac{8}{11} & 0 & 0 & -\frac{4}{15} \\ \frac{2}{11} & \frac{16}{11} & \frac{2}{11} & -\frac{4}{11} & -\frac{8}{11} & 0 & 0 \\ 0 & \frac{2}{11} & \frac{12}{11} & 0 & -\frac{4}{11} & 0 & 0 \\ -\frac{8}{11} & -\frac{4}{11} & 0 & \frac{59}{33} & 0 & -\frac{2}{3} & 0 \\ 0 & -\frac{8}{11} & -\frac{4}{11} & 0 & \frac{16}{11} & 0 & 0 \\ 0 & 0 & 0 & -\frac{2}{3} & 0 & \frac{4}{3} & 0 \\ -\frac{4}{15} & 0 & 0 & 0 & 0 & 0 & \frac{16}{15} \end{pmatrix}.$$

8.6.2 Sire-Dam Relationships

For animal models, with the usual animal, sire, and dam pedigree lists, we need to know the inbreeding coefficient of every animal. To calculate inbreeding coefficients, animals must be sorted chronologically so that the an animal's parents appear in the pedigree list prior to the animal itself. One should not rely 100% on birthdates to sort animals, because errors are known to exist in pedigree lists. Meuwissen and Luo (1992) give a good algorithm for computing inbreeding coefficients.

In Table 8.2, is a partial pedigree list for a few animals with their inbreeding coefficients.

Now we must determine the fraction of Mendelian sampling variance is remaining in each animal. That quantity is given by

$$f_i \;=\; 0.50 \;-\; 0.25 \times (F_{sire} + F_{dam})$$

where F_{sire} and F_{dam} are the inbreeding coefficients of the sire and dam of the animal. For animals A through E the parents are all non-inbred, so that the

Table 8.2: Example pedigrees with inbreeding coefficients, F_i

Animal	Sire	Dam	F_i
A	X	Q	0
B	X	Q	0
C	X	W	0
D	A	B	1/4
E	B	C	1/8
F	D	E	1/4

fraction of Mendelian sampling variance in their genome is one half of the additive genetic variance.

The parents of animal F, however, are both inbred, and therefore,

$$f_F \;=\; 0.5 - 0.25 \times (0.25 + 0.125) \;=\; 0.40625 = 13/32.$$

The value of δ for animal F is

$$\delta \;=\; 32/13 = 1/f_F.$$

Because sires and dams can have many combinations of inbreeding coefficients, the number of different values for f_F can be very large, and therefore, there are also many different possible values for δ. However, is an animal has both parents unknown, then $\delta = 1$, and if either one of the two parents is unknown, then $\delta = 4/3$, assuming that the unknown animals are not related.

The rules for adding to $\mathbf{A}^{-1}$ are similar to those for the sire-MGS inverse matrix, but simpler, i.e.

	sire	dam	animal
sire	.25 δ	.25 δ	$-.5\ \delta$
dam	.25 δ	.25 δ	$-.5\ \delta$
animal	$-.5\ \delta$	$-.5\ \delta$	δ

The animals in the complete pedigree list are processed one animal at a time until $\mathbf{A}^{-1}$ is completed.

8.7 References

HENDERSON, C. R. 1953. Estimation of variance and covariance components. Biometrics 9:226.

HENDERSON, C. R. 1973. Sire evaluation and genetic trends. Proc. of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush. ASAS, ADSA, PSA. p. 10-41.

HENDERSON, C. R. 1975. Inverse of a matrix of relationships due to sires and maternal grandsires. J. Dairy Sci. 58:1917-1921.

HENDERSON, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. Biometrics 32:69.

MEUWISSEN, T.H.E. , Z. LUO. 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24:305-313.

MILLER, P. D. , W. E. LENTZ, C. R. HENDERSON. 1970. Joint influence of month and age of calving on milk yield of Holstein cows in the northeastern United States. J. Dairy Sci. 53:351-357.

QUAAS, R. L. , E. J. POLLAK. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. J. An. Sci. 51:1277.

SCHAEFFER, L. R. , B. W. KENNEDY. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575-579.

Chapter 9

Sire Models

HORIA GROSU
PASCAL A. OLTENACU
LARRY SCHAEFFER

9.1 Northeast AI Sire Comparison - 1972

The sire model implemented in the first Northeast AI Sire Comparison was

$$y_{ijkl} = H_i + g_j + s_{jk} + e_{ijkl},$$

where

y_{ijkl} was a first lactation production yield of the l^{th} daughter of the k^{th} sire in the i^{th} herd-year-season of calving,

H_i was a fixed, herd-year-season of calving effect,

g_j was a fixed, genetic group of sire effect,

s_{jk} was a random, sire transmitting ability, and

e_{ijkl} was a random residual effect (including dam and Mendelian sampling effects as well as temporary environment).

The expected values of s_{jk} and e_{ijkl} were all zero, and $Var(s_{jk}) = 0.25 \cdot h^2 \cdot \sigma_y^2$ and $Var(e_{ijkl}) = \sigma_e^2 = (1 - 0.25h^2)\sigma_y^2$. The residual variance was assumed constant for all daughters and all herd-year-seasons. The ratio $\sigma_e^2/\sigma_s^2 = 15$,

assuming heritability was 0.25. Sires were assumed to be unrelated, and therefore, if **s** represents the vector of all sire effects, then

$$Var(\mathbf{s}) \;=\; \mathbf{I}\sigma_s^2.$$

9.1.1 Genetic Groups

Genetic groups of sires were based on their year of birth, and on the AI organization that entered them into progeny testing. Natural service bulls were not evaluated. The idea was that different AI studs applied different selection strategies when entering young bulls, and thus, those populations could have different genetic means over time. Because the data came from the Northeast USA, most of the bulls were from Eastern Breeders in Ithaca, NY, or natural service bulls. Bulls from other AI studs were not completely represented, but were only bulls that dairy producers in the Northeast USA could afford to buy from other AI studs.

9.1.2 Data

Only first lactation production records were used which gave fewer records per sire upon which to base EBVs. The belief was that culling of cows after first lactation would introduce biases into genetic evaluations of bulls. Some years later the model was changed to include multiple records of cows. Production records were adjusted for age and month of calving and only records from herds that milked 2 times per day were included. If a bull had daughters in the herd from which he was born, then those daughters were excluded from the data. The assumption was that the herd owner may have provided preferential treatment to those daughters to make the bull look better than he might actually be. This edit often reduced the size of herd-year-season groups, which affected the accuracy of bull genetic evaluations.

9.1.3 Dams

Sires were assumed to be randomly mated to dams (same genetic level, on average), and dams were assumed to have one progeny only in the data. This was obviously not true, but the assumption was necessary with this model. The genetic level of the contemporaries was considered by having Contemporary Group (CG) in the model, so that the assumptions about dams were not as critical.

9.1.4 Herd-year-seasons

Henderson (1975) published a paper in Biometrics on his selection bias theories, and the paper was the foundation for the treatment of herd-year-season effects as a fixed factor in the Northeast AI Sire Comparison. Subsequently, this theory was criticized by Robin Thompson (1979), Daniel Gianola (1988), and Richard Quaas (1980). In retrospect, it may have been better to let HYS effects be random as they obviously are. However, this one little assumption has carried over into all genetic evaluation methods in almost all livestock species, when for many situations HYS should be a random factor. One problem with fixed HYS effects is that any HYS where all the cows are from the same sire, then none of that information is used in genetic evaluation. If you "absorb" that HYS equation, then all zeros are created and nothing is added to the sire equations. Also, if a HYS has only one cow in it, then that information is also lost. For many European countries, HYS size was much smaller than in the USA, and consequently much data were not being used in genetic evaluations. However, if HYS had been random, then those records would have been utilized.

9.1.5 Random Samples

The sire model implicitly assumes that sires are mated randomly to dams. From matings to those dams, the daughters represented in the data are assumed to be a random sample of all possible daughters. Within HYS each cow is expected to receive the same level of treatment and care. That means that no cow is to receive special treatment (like extra feed or being kept in its own stall). For the most part these assumptions are met, but there can be situations where preferential treatment does exist. Also, due to different price structures for proven versus unproven bulls, the randomness of mates may not be a valid assumption.

9.1.6 Numerical Example

The following table (Table 9.1)contains the subclass numbers of cows per herd, year-season, and sire subclasses.

Thus, there are 9 herd-year-season subclasses and 6 sires. Let sires 1, 2, and 3 belong to genetic group 1, and sires 4, 5, and 6 belong to genetic group 2. The records are not shown, but there are 91 first lactation records in this small example. The total sum of squares was $4,049,535,091$.

In matrix notation, the model is

$$\mathbf{y} = \mathbf{X}_h\mathbf{h} + \mathbf{X}_g\mathbf{g} + \mathbf{Z}_s\mathbf{s} + \mathbf{e}.$$

Table 9.1: Subclass numbers for example calculations

Herd	Year-Season	Sires					
		1	2	3	4	5	6
1	1	2	1	0	0	4	3
	2	1	3	5	0	1	2
	3	0	2	3	0	2	2
2	1	1	2	0	5	1	2
	2	1	1	2	3	1	0
	3	1	3	2	2	0	2
3	1	0	1	3	2	0	3
	2	0	4	1	2	1	5
	3	0	1	4	2	0	2

Thus,

$$\mathbf{X} = (\mathbf{X}_h \quad \mathbf{X}_g) \tag{9.1}$$

$$\mathbf{Z} = \mathbf{Z}_s \tag{9.2}$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{h} \\ \mathbf{g} \end{pmatrix} \tag{9.3}$$

$$\mathbf{u} = \mathbf{s} \tag{9.4}$$

$$\mathbf{G} = \mathbf{I}\sigma_s^2 \tag{9.5}$$

$$\mathbf{R} = \mathbf{I}\sigma_e^2 \tag{9.6}$$

The MME are

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

where

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} =$$

$$\frac{1}{\sigma_e^2} \begin{pmatrix} \mathbf{X}'_h\mathbf{X}_h & \mathbf{X}'_h\mathbf{X}_g & \mathbf{X}'_h\mathbf{Z}_s \\ \mathbf{X}'_g\mathbf{X}_h & \mathbf{X}'_g\mathbf{X}_g & \mathbf{X}'_g\mathbf{Z}_s \\ \mathbf{Z}'_s\mathbf{X}_h & \mathbf{Z}'_s\mathbf{X}_g & \mathbf{Z}'_s\mathbf{Z}_s + \mathbf{I}\frac{1}{\sigma_s^2} \end{pmatrix}$$

$$\begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{h}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{s}} \end{pmatrix},$$

and

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} = \frac{1}{\sigma_e^2} \begin{pmatrix} \mathbf{X}'_h\mathbf{y} \\ \mathbf{X}'_g\mathbf{y} \\ \mathbf{Z}'_s\mathbf{y} \end{pmatrix}.$$

Multiply both sides of the MME by σ_e^2, and the resulting equations are

$$\begin{pmatrix} \mathbf{X}'_h\mathbf{X}_h & \mathbf{X}'_h\mathbf{X}_g & \mathbf{X}'_h\mathbf{Z}_s \\ \mathbf{X}'_g\mathbf{X}_h & \mathbf{X}'_g\mathbf{X}_g & \mathbf{X}'_g\mathbf{Z}_s \\ \mathbf{Z}'_s\mathbf{X}_h & \mathbf{Z}'_s\mathbf{X}_g & \mathbf{Z}'_s\mathbf{Z}_s + \mathbf{I}k \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{h}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{s}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_h\mathbf{y} \\ \mathbf{X}'_g\mathbf{y} \\ \mathbf{Z}'_s\mathbf{y} \end{pmatrix}.$$

The order of these equations, for the example data, are $(9 + 2 + 6)$ or 17 and the rank of the equations is 16.

Matrix $\mathbf{X}_h$ was 91 rows by 9 columns, one for each of the herd-year-seasons. Each row contained one 1 corresponding to the herd-year-season in which the record was made.

Matrix $\mathbf{X}_g$ was 91 rows by 2 columns, one for each genetic group. Each row had a one and a zero, with the one in the column of the group to which the sire of the cow was a member.

Matrix $\mathbf{Z}_s$ was 91 rows by 6 columns, one for each sire. Each row had a one corresponding to the sire of the cow making the record.

The 91 records were ordered herd-year-seasons within sires (i.e. by columns of Table 9.4). Thus, the first 6 observations in $\mathbf{y}$ were progeny of sire 1, followed by the 18 of sire 2, and so forth. The order of records in $\mathbf{y}$ must be maintained in $\mathbf{X}_h$, $\mathbf{X}_g$, and $\mathbf{Z}_s$. The mixed model equations were

$$\left(\begin{array}{cccccccccc|cc|cccccc}
10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 7 & 2 & 1 & 0 & 0 & 4 & 3 \\
0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9 & 3 & 1 & 3 & 5 & 0 & 1 & 2 \\
0 & 0 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 4 & 0 & 2 & 3 & 0 & 2 & 2 \\
0 & 0 & 0 & 11 & 0 & 0 & 0 & 0 & 0 & 3 & 8 & 1 & 2 & 0 & 5 & 1 & 2 \\
0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 & 4 & 4 & 1 & 1 & 2 & 3 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 & 6 & 4 & 1 & 3 & 2 & 2 & 0 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 & 4 & 5 & 0 & 1 & 3 & 2 & 0 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 13 & 0 & 5 & 8 & 0 & 4 & 1 & 2 & 1 & 5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9 & 5 & 4 & 0 & 1 & 4 & 2 & 0 & 2 \\
\hline
3 & 9 & 5 & 3 & 4 & 6 & 4 & 5 & 5 & 44 & 0 & 6 & 18 & 20 & 0 & 0 & 0 \\
7 & 3 & 4 & 8 & 4 & 4 & 5 & 8 & 4 & 0 & 47 & 0 & 0 & 0 & 16 & 10 & 21 \\
\hline
2 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 6 & 0 & 21 & 0 & 0 & 0 & 0 & 0 \\
1 & 3 & 2 & 2 & 1 & 3 & 1 & 4 & 1 & 18 & 0 & 0 & 33 & 0 & 0 & 0 & 0 \\
0 & 5 & 3 & 0 & 2 & 2 & 3 & 1 & 4 & 20 & 0 & 0 & 0 & 35 & 0 & 0 & 0 \\
0 & 0 & 0 & 5 & 3 & 2 & 2 & 2 & 2 & 0 & 16 & 0 & 0 & 0 & 31 & 0 & 0 \\
4 & 1 & 2 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 10 & 0 & 0 & 0 & 0 & 25 & 0 \\
3 & 2 & 2 & 2 & 0 & 2 & 3 & 5 & 2 & 0 & 21 & 0 & 0 & 0 & 0 & 0 & 36
\end{array}\right)
\left(\begin{array}{c}
\widehat{H}_1 \\ \widehat{H}_2 \\ \widehat{H}_3 \\ \widehat{H}_4 \\ \widehat{H}_5 \\ \widehat{H}_6 \\ \widehat{H}_7 \\ \widehat{H}_8 \\ \widehat{H}_9 \\ \widehat{g}_1 \\ \widehat{g}_2 \\ \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \\ \widehat{s}_4 \\ \widehat{s}_5 \\ \widehat{s}_6
\end{array}\right)$$

$$= \begin{pmatrix} 56,484 \\ 80,694 \\ 63,317 \\ 68,581 \\ 53,426 \\ 71,054 \\ 58,639 \\ 86,029 \\ 65,623 \\ \hline 290,395 \\ 313,452 \\ \hline 32,149 \\ 113,554 \\ 144,692 \\ 112,748 \\ 57,792 \\ 142,912 \end{pmatrix}$$

There are an infinite number of solution vectors possible, but we will use the restriction that $\widehat{g}_2 = 0$. The solutions are (Table 9.2)

Table 9.2: Solutions to MME for example data to sire model

HYS	estimate	Group	estimate	Sire	estimate
1	5877	1	-298	1	-235
2	6862	2	0	2	-102
3	7153			3	336
4	6281			4	187
5	6748			5	-264
6	7218			6	78
7	6480				
8	6699				
9	7260				

Estimated transmitting abilities are calculated as the sum of genetic group and sire solutions.

Sires 1, 2, and 3 are in group 1, and sires 4, 5, and 6 are in group 2, hence

Sire 1	-298	-235	=	-533
Sire 2	-298	-102	=	-400
Sire 3	-298	+336	=	+38
Sire 4	0.0	+187	=	+187
Sire 5	0.0	-264	=	-264
Sire 6	0.0	+78	=	+78

Accuracy of the estimates is based on the standard errors of prediction. For sire 1, as an example,

$$\begin{aligned} \text{ETA}_1 &= \widehat{g}_1 + \widehat{s}_1 \\ Var(\text{ETA}_1 - \text{TA}_1) &= Var(\widehat{g}_1) + Var(\widehat{s}_1 - s_1) + 2Cov(\widehat{g}_1, \widehat{s}_1 - s_1) \\ &= (c_{g1} + c_{s1} + 2c_{g1,s1}) \times \widehat{\sigma}_e^2 \end{aligned}$$

where c_{g1} is the diagonal inverse element for the group 1 solution, c_{s1} is the diagonal inverse element for the sire 1 solution, and $c_{g1,s1}$ is the off-diagonal element between group 1 and sire 1. The residual variance is given by

$$\widehat{\sigma}_e^2 = (\mathbf{y}'\mathbf{y} - \widehat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \widehat{\mathbf{u}}'\mathbf{Z}'\mathbf{y})/(N - r(\mathbf{X}))$$

For this example, $\widehat{\sigma}_e^2 = 174,940.9$. For sire 1, $c_{g1} = 0.09604$, $c_{s1} = 0.05252$, and $c_{g1,s1} = -0.01446$, so that

$$Var(\text{ETA}_1 - (TA)_1) = (.11965 \times 174,940.9).$$

The SEP, standard error of prediction is the square root of the variance of prediction error, and is ±144.68 for sire 1. The results for the six sires are shown below (Table 9.3).

Table 9.3: ETA and SEP for 6 example sires

Sire	ETA	SEP
1	-533	145
2	-400	124
3	38	125
4	187	90
5	-264	93
6	78	88

9.2 Sire-MGS Relationships

In 1975, Henderson discovered a method for writing the inverse of a relationship matrix constructed from sire and MGS (maternal grandsire) relationships. Some details of this method were given in the previous chapter. In the previous example, suppose that sires 1, 2, and 5 were half-sibs, sires 4 and 6 were full-sibs, and sire 3 was unrelated to the others. Then a possible relationship matrix would be as follows:

$$\mathbf{A} = \begin{pmatrix} 1 & .25 & 0 & 0 & .25 & 0 \\ .25 & 1 & 0 & 0 & .25 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & .50 \\ .25 & .25 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & .50 & 0 & 1 \end{pmatrix}.$$

Then, for $k = 15$,

$$\mathbf{A}^{-1}k = \begin{pmatrix} 16\frac{2}{3} & -3\frac{1}{3} & 0 & 0 & -3\frac{1}{3} & 0 \\ -3\frac{1}{3} & 16\frac{2}{3} & 0 & 0 & -3\frac{1}{3} & 0 \\ 0 & 0 & 15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & -10 \\ -3\frac{1}{3} & -3\frac{1}{3} & 0 & 0 & 16\frac{2}{3} & 0 \\ 0 & 0 & 0 & -10 & 0 & 20 \end{pmatrix}.$$

This matrix is used in place of $\mathbf{I}k$ in the MME. Everything else in the MME is the same as before. However, this one little change, accounting for relationships among bulls, has an effect of the solutions and ETAs. The new results are given in Table 9.4.

Notice that the HYS estimates are very similar to those from the previous model. The estimate of the difference between genetic group 1 and genetic group 2, however, is smaller by 88 kg. This happened in real life in the Northeast AI Sire Comparison as well only much more drastic. Many genetic group differences became almost zero. After some contemplation, the estimated differences were now a reflection of differences in selection differentials among the dams of AI bulls. All AI studs were buying bulls from the same herds and types of cows so that there were small differences remaining. Additive genetic relationships were accounting for many of the prior differences between AI studs and years of birth.

Note also that the solutions for sires 1, 2, and 5 are more similar than they were in the previous model, due to the fact that they are now considered to be

Table 9.4: Solutions to MME for example data to sire model using sire-MGS relationships

HYS	estimate	Group	estimate	Sire	estimate
1	5874	1	-211	1	-308
2	6832	2	0	2	-187
3	7131			3	305
4	6263			4	209
5	6729			5	-304
6	7190			6	133
7	6438				
8	6673				
9	7218				

half-sibs. Also sires 4 and 6 are more similar. The residual variance is given by

$$\widehat{\sigma}_e^2 \;=\; (\mathbf{y}'\mathbf{y} - \widehat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \widehat{\mathbf{u}}'\mathbf{Z}'\mathbf{y})/(N - r(\mathbf{X}))$$

For this example, $\widehat{\sigma}_e^2 = 163,762.9$. Notice that the estimate is smaller than for the model without sire-MGS relationships. This indicates that sire-MGS relationships are helping to provide more accurate sire solutions, and therefore, more of the variation is being explained with relationships present. Presumably, the more complete the relationships are, then the solutions should be more accurate.

After this model was implemented, then concerns about errors in pedigrees rose in priority, and many errors were found. Parentage checks were used by breed associations to verify parents for registration of progeny, but these checks were random and sporadic, except for young bulls going into AI service in which all had to be tested. Thus, errors were present in all pedigrees with estimates from 5 to 12% of all registrations.

Sire ETAs are created as before, as the sum of genetic group and sire solutions, and standard errors of prediction are calculated in the same way using the inverse elements of the coefficient matrix of the MME.

At first glance the SEP in the Table 9.5, appear larger than in the previous section. However, the SEP in each table are calculated assuming that the model used in the analysis was the true, correct model. In the first case, that means bulls were not related was the true state of nature. If not true, then the SEP would need to be re-calculated under the true state of nature, which would make them larger than the values in the Table 9.5. Similarly, when bulls were assumed to be related, then SEP were calculated assuming that this was the true state of

Table 9.5: ETA and SEP for 6 example sires, using sire-MGS relationships

Sire	ETA	SEP
1	-519	147
2	-398	130
3	94	130
4	209	96
5	-304	93
6	133	94

nature.

9.3 Random HYS

If Henderson had left HYS effects as a random factor in the model, then another factor to account for time trends would have been needed. The model would be

$$y_{ijklm} = YS_i + HYS_{ij} + g_k + s_{kl} + e_{ijklm},$$

where

y_{ijklm} was a first lactation production yield of the m^{th} daughter of the l^{th} sire in the ij^{th} herd-year-season of calving,

YS_i was a fixed, year-season of calving effect,

HYS_{ij} was a random, herd within year-season of calving effect,

g_k was a fixed, genetic group of sire effect,

s_{kl} was a random, sire transmitting ability, and

e_{ijklm} was a random residual effect (including dam and Mendelian sampling effects as well as temporary environment).

The expected values of HYS_{ij}, s_{kl} and e_{ijklm} were all zero, and

$$\begin{aligned} Var(HYS_{ij}) &= \sigma_e^2/6.5, \\ Var(s_{kl}) &= \sigma_e^2/15.0, \\ Var(e_{ijklm}) &= \sigma_e^2. \end{aligned}$$

The residual variance was assumed constant for all daughters and all herd-year-seasons. Sires were assumed to be related, as in the previous section. The same **A** matrix applies to this section.

In matrix notation, the model is

$$\mathbf{y} = \mathbf{X}_{ys}\mathbf{ys} + \mathbf{X}_g\mathbf{g} + \mathbf{Z}_h\mathbf{h} + \mathbf{Z}_s\mathbf{s} + \mathbf{e}.$$

Thus,

$$\begin{aligned}
\mathbf{X} &= (\mathbf{X}_{ys} \quad \mathbf{X}_g) \\
\mathbf{Z} &= (\mathbf{Z}_h \quad \mathbf{Z}_s) \\
\mathbf{b} &= \begin{pmatrix} \mathbf{ys} \\ \mathbf{g} \end{pmatrix} \\
\mathbf{u} &= \begin{pmatrix} \mathbf{h} \\ \mathbf{s} \end{pmatrix} \\
\mathbf{G} &= \begin{pmatrix} \mathbf{I}\sigma_h^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma_s^2 \end{pmatrix} \\
\mathbf{R} &= \mathbf{I}\sigma_e^2
\end{aligned}$$

The MME are, after multiply both sides of the MME by σ_e^2,

$$\begin{pmatrix}
\mathbf{X}'_{ys}\mathbf{X}_{ys} & \mathbf{X}'_{ys}\mathbf{X}_g & \mathbf{X}'_{ys}\mathbf{Z}_h & \mathbf{X}'_{ys}\mathbf{Z}_s \\
\mathbf{X}'_g\mathbf{X}_{ys} & \mathbf{X}'_g\mathbf{X}_g & \mathbf{X}'_g\mathbf{Z}_h & \mathbf{X}'_g\mathbf{Z}_s \\
\mathbf{Z}'_h\mathbf{X}_{ys} & \mathbf{Z}'_h\mathbf{X}_g & \mathbf{Z}'_h\mathbf{Z}_h + \mathbf{I}(6.5) & \mathbf{Z}'_h\mathbf{Z}_s \\
\mathbf{Z}'_s\mathbf{X}_{ys} & \mathbf{Z}'_s\mathbf{X}_g & \mathbf{Z}'_s\mathbf{Z}_h & \mathbf{Z}'_s\mathbf{Z}_s + \mathbf{A}^{-1}(15)
\end{pmatrix}
\begin{pmatrix} \widehat{\mathbf{ys}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{h}} \\ \widehat{\mathbf{s}} \end{pmatrix}
=
\begin{pmatrix} \mathbf{X}'_{ys}\mathbf{y} \\ \mathbf{X}'_g\mathbf{y} \\ \mathbf{Z}'_h\mathbf{y} \\ \mathbf{Z}'_s\mathbf{y} \end{pmatrix}.$$

The order of these equations, for the example data, are $(3+2+9+6)$ or 20 and the rank of the equations is 19.

Table 9.6: Solutions to MME for example data to sire model using sire-MGS relationships and random HYS effects

YS	estimate	Group	estimate	HYS	estimate	Sire	estimate
1	6181	1	-193	1	-184	1	-323
2	6732	2	0	2	55	2	-193
3	7165			3	-26	3	315
				4	45	4	226
				5	-9	5	-323
				6	6	6	141
				7	139		
				8	-46		
				9	20		

ETA are formed by adding genetic group and sire solutions, as before (Table 9.6). Note that the genetic group difference is smaller yet compared to the previous two sections. The residual variance estimate was $160,443.6$ which is smaller than the previous sections. Thus, having HYS as random and adding the YS fixed effect to account for time trends, the resulting model accounts for more variation. The ETA and their SEP (Table 9.7), assuming this model is the true state of nature, are

Table 9.7: ETA and SEP for 6 example sires, using sire-MGS relationships, and random HYS effects

Sire	ETA	SEP
1	-516	147
2	-386	130
3	122	128
4	226	95
5	-323	92
6	141	93

9.4 Maternal Grandsire Model

An assumption of the Sire Model was that sires were randomly mated to dams. However, the differential prices of proven versus unproven bull semen often meant that high priced bulls were mated to higher quality dams. Quaas et al. (1979) proposed the Maternal Grandsire Model to partially account for this problem. Computing restrictions still kept scientists from using an Animal Model in 1979. Animal models were not feasible for another 10 years.

The model would be

$$y_{ijkhmn} = HYS_i + g_j + s_{jk} + \frac{1}{2}(g_h + s_{hm}) + e_{ijkhmn},$$

where

y_{ijkhmn} was a first lactation production yield of the n^{th} daughter of the jk^{th} sire in the i^{th} herd-year-season of calving,

HYS_i was a fixed, herd within year-season of calving effect,

g_j was a fixed, genetic group of sire effect,

s_{jk} was a random, sire transmitting ability,

g_h was a fixed, genetic group of sire effect (for the maternal grandsire),

s_{hm} was a random, sire transmitting ability (for the maternal grandsire), and

e_{ijkhmn} was a random residual effect (including dam and Mendelian sampling effects as well as temporary environment).

The expected values of (s_{jk}, s_{hm}) and e_{ijkhmn} were all zero, and

$$\begin{aligned} Var(s_{kl}) &= \sigma_e^2/15.0, \\ Var(e_{ijklm}) &= \sigma_e^2. \end{aligned}$$

Assume again that the additive relationships were

$$\mathbf{A} = \begin{pmatrix} 1 & .25 & 0 & 0 & .25 & 0 \\ .25 & 1 & 0 & 0 & .25 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & .50 \\ .25 & .25 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & .50 & 0 & 1 \end{pmatrix}.$$

The residual variance was assumed constant for all daughters and all herd-year-seasons.

9.4.1 Assumptions

Besides the assumptions for the usual Sire Model, other assumptions were

- The sire of the dams (MGS) were assumed known for all cows with lactation records.
- The daughters of the MGS which were dams (mates) to the sires were assumed to be a random sample of daughters of that MGS.
- Dams were still assumed to have only one progeny each.

By accounting for the MGS, some genetic level of the dam was assumed. Unfortunately, the daughters of a bull which later became dams of other cows were slightly selected, especially if the dam was beginning her second or later lactation. The poorer daughters would have been culled after their first lactations. Depending on the bull, the number of daughters culled would vary.

When MGS were not known, then the model reverted back to the Sire Model for those cows, but the residual variance was assumed to be larger than for the MGS model. The variances for the MGS model were

$$\sigma_y^2 = \sigma_s^2 + \frac{1}{4}\sigma_s^2 + \sigma_E^2$$

and for cows with unknown MGS,

$$\begin{aligned} \sigma_y^2 &= \sigma_s^2 + (\frac{1}{4}\sigma_s^2 + \sigma_E^2) \\ &= \sigma_s^2 + \sigma_e^2 \end{aligned}$$

If we let $\sigma^2_E = 1$, then $\sigma^2_e = 1.01667$. The MME require the ratio $k = \sigma^2_E/\sigma^2_s$ or 15.25.

In matrix notation, the model is

$$\mathbf{y} = \mathbf{X}_h\mathbf{h} + \mathbf{X}_g\mathbf{g} + \mathbf{Z}_s\mathbf{s} + \mathbf{e}.$$

Thus,

$$\begin{aligned}
\mathbf{X} &= (\mathbf{X}_h \quad \mathbf{X}_g) \\
\mathbf{Z} &= \mathbf{Z}_s \\
\mathbf{b} &= \begin{pmatrix} \mathbf{h} \\ \mathbf{g} \end{pmatrix} \\
\mathbf{u} &= \mathbf{s} \\
\mathbf{G} &= \mathbf{A}\sigma^2_s \\
\mathbf{R} &= \mathbf{I}\sigma^2_E
\end{aligned}$$

In the MGS model, $\mathbf{X}_g$ and $\mathbf{Z}_s$ now contain two non-zero elements per row rather than one. Suppose there are five genetic groups and the sire of a cow belongs to group 4, and the MGS belongs to group 2, then the row of $\mathbf{X}_g$ for this cow would appear as

$$\begin{pmatrix} 0 & .5 & 0 & 1 & 0 \end{pmatrix}$$

Similarly, the row of $\mathbf{Z}_s$ would have a 1 in the location of the sire and .5 in the column for the MGS. Usually the sire and MGS would belong to different genetic groups, and usually the sire and MGS would be two distinct individuals, otherwise the 1 and .5 would be combined for the same genetic group column or the same sire column.

In general, the MGS model increased the number of sires that needed to be evaluated. Some of the MGS were natural service bulls, and many were older sires the pre-dated the first sires with progeny in the data.

9.4.2 Numerical Example

The previous example data do not have any MGS identified, and therefore can not be used to illustrate the MGS model. Instead, consider the small example in the table 9.8.

Table 9.8: MGS Model Example Data

Cow	Sire Group	Sire	MGS Group	MGS	HYS	Yield,kg
1	2	4	1	1	1	6062
2	2	4	1	2	1	6516
3	2	5	1	2	1	5325
4	2	6	1	3	2	7535
5	1	1	1	3	2	6183
6	1	2	-	-	2	7223
7	2	6	2	4	2	7466

The components of the model are

$$\mathbf{y} = \begin{pmatrix} 6062 \\ 6516 \\ 5325 \\ 7535 \\ 6183 \\ 7223 \\ 7466 \end{pmatrix}, \quad \mathbf{X}_h = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{X}_g = \begin{pmatrix} .5 & 1 \\ .5 & 1 \\ .5 & 1 \\ .5 & 1 \\ 1.5 & 0 \\ 1 & 0 \\ 0 & 1.5 \end{pmatrix},$$

and

$$\mathbf{Z}_s = \begin{pmatrix} .5 & 0 & 0 & 1 & 0 & 0 \\ 0 & .5 & 0 & 1 & 0 & 0 \\ 0 & .5 & 0 & 0 & 1 & 0 \\ 0 & 0 & .5 & 0 & 0 & 1 \\ 1 & 0 & .5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .5 & 0 & 1 \end{pmatrix}.$$

The residual matrix is scaled, and is a diagonal matrix,

$$\mathbf{R} = diag(\ 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1.01667 \quad 1\ )$$

The one element is different from one because the MGS for cow 6 was unknown.

The resulting MME were of order 10 = (2 HYS + 2 genetic groups + 6

sires). The solutions were (Table 9.9)

Table 9.9: Solutions to MGS MME

HYS	estimate	Genetic groups	estimate	Sire	estimate
1	6363	1	-801	1	-22
2	7695	2	0	2	4
				3	-2
				4	30
				5	-40
				6	14

Thus, the evaluation of a bull consists of both his daughters and his granddaughters. The ETA would be constructed in the same manner by adding the genetic group solutions to the sire solutions. Sires 1, 2, and 3 belong to genetic group 1, and therefore, their ETA would be -823, -797, and -803, respectively. However, the estimate of the group difference has a very high standard error because group 1 sires had only 2 daughter records. Sires 4, 5, and 6 have ETA equal to the sire solutions because the solution for genetic group 2 was restricted to be zero.

When the MGS was applied to real data at Cornell University, the change in sire ETA was much greater than anticipated. This was apparently due to the invalid assumption that dams were not a random group of daughters of their sires. Those that became dams were the result of culling. Thus, MGS effects were inflated above their complete progeny level. This would cause the sire solutions to be adjusted upwards because the genetic level of the dams would be over-rated. The MGS model was a failure for production traits, and should only be used if daughters of MGS are random samples of all daughters of each MGS.

9.5 All Lactations

Another criticism of the Northeast AI Sire Comparison was that it did not include later lactations of cows. Thus, sires were not being estimated as accurately, and sires were being chosen for early first lactation yields rather than lifetime production. Ufford et al. (1979) modified the sire model so that all lactations of each cow could be included. To do so, cows had to be nested within sires (which they always were) and nested within herds (which was not always

true). Cows which changed herds, only the records in the herd in which the first lactation was made were kept. Cows were assumed to be unrelated between herds, which was also not true. These assumptions allowed the cow equations to be absorbed into the HYS and sire equations. The model equation was

$$y_{ijkhmn} = HYS_{ij} + g_k + s_{kh} + sH_{ikh} + c_{ikhm} + e_{ijkhmn}$$

where

y_{ijkhm} is the n^{th} record of the m^{th} cow of the jk^{th} sire in the i^{th} herd and j^{th} year-season;

H_{ij} is the j^{th} year-season within the i^{th} herd, fixed;

g_k is the k^{th} genetic group for sires, fixed;

s_{kh} is a sire transmitting ability, random;

sH_{ijkh} is a sire by herd interaction (common environment effect among daughters of the same sire in one herd), random;

c_{ikhm} is a cow within sire effect (includes dam's contribution to that daughter plus permanent environmental effect), nexted within herd and within sire, random;

e_{ijkhmn} is a residual effect peculiar to each lactation record, random.

In matrix notation, the model is

$$\mathbf{y} = \mathbf{X}_h\mathbf{h} + \mathbf{X}_g\mathbf{g} + \mathbf{Z}_s\mathbf{s} + \mathbf{Z}_{sh}\mathbf{sh} + \mathbf{Z}_c\mathbf{c} + \mathbf{e}.$$

Thus,

$$\begin{aligned}
\mathbf{X} &= (\mathbf{X}_h \quad \mathbf{X}_g) \\
\mathbf{Z} &= (\mathbf{Z}_s \quad \mathbf{Z}_{sh} \quad \mathbf{Z}_c) \\
\mathbf{b} &= \begin{pmatrix} \mathbf{h} \\ \mathbf{g} \end{pmatrix} \\
\mathbf{u} &= \begin{pmatrix} \mathbf{s} \\ \mathbf{sh} \\ \mathbf{c} \end{pmatrix} \\
\mathbf{G} &= \begin{pmatrix} \mathbf{A}\sigma_s^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_{sh}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_c^2 \end{pmatrix} \\
\mathbf{R} &= \mathbf{I}\sigma_e^2
\end{aligned}$$

The residual variance was assumed constant across lactation records and across herds. The assumed variance ratios were

$$\begin{aligned} k_s &= \sigma_e^2/\sigma_s^2 = 8.33 \\ k_{sh} &= \sigma_e^2/\sigma_{sh}^2 = 3.57 \\ k_c &= \sigma_e^2/\sigma_c^2 = 1.67 \end{aligned}$$

The equations for cows were absorbed into the other equations, then sire by herd interactions were absorbed in HYS, groups, and sire equations. Then HYS equations were absorbed into groups and sire equations. Thus, the solutions for cows, sire by herd interactions, and HYS were never obtained explicitly. These were considered to be nuisance parameters. One needed to account for cows, number of records and distribution of records, and needed to account for sire by herd interactions, and for the level of herdmates (contemporaries) within herd-year-seasons, but the actual values of those effects were forfeited. This was the only computational strategy feasible at that point in time with the computer hardware available, and even this approach took a long time to compute.

No numerical example is given for this model.

9.6 References

GIANOLA, D. , Im, S., Fernando, R.L. 1988. Prediction of breeding values under Henderson's selection model: A revisitation. J. Dairy Sci. 71:2790-2798.

HENDERSON, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-448.

HENDERSON, C. R. 1975. Inverse of a matrix of relationships due to sires and maternal grandsires. J. Dairy Sci. 58:1917-1921.

QUAAS, R. L. , R. W. EVERETT, A. C. McCLINTOCK. 1979. Maternal grandsire model fo dairy sire evaluation. J. Dairy Sci. 62:1648-1654.

QUAAS, R. L. , 1980. Personam Communication

THOMPSON, R. 1979. Sire evaluation. Biometrics 35:339.

UFFORD, G. R. , C. R. HENDERSON, L. D. VAN VLECK. 1979. Computing algorithms for sire evaluation with all lactation records and natural service sires. J. Dairy Sci. 62:511-513.

UFFORD, G. R. , C. R. HENDERSON, J. F. KEOWN, L. D. VAN VLECK. 1979. Accuracy of first lactation versus all lactation sire evaluations by best linear unbiased prediction. J. Dairy Sci. 62:603-612.

Chapter 10

Animal Models

HORIA GROSU
PASCAL A. OLTENACU
LARRY SCHAEFFER

10.1 Microcomputers

The first official microcomputer, the Datapoint 2200, was introduced in 1970. People were beginning to think about having their own computing power under their control, but the cost of the machines was too great. The first generation of microcomputers lasted from 1971 to 1976, and mostly their usage was limited to games. Monitors were very small, and the computers were used mostly for games because of the small amount of Random Access Memory (RAM). In essence they were just large calculators.

The second generation of microcomputers started in 1977 and machines now had BASIC as a programming language. In 1979 Apple II was made, followed by Atari and Commodore in 1980. The Microsoft Disk Operating System (MS-DOS) became available in 1980 and was sold with IBM PC-like machines. The advances in microprocessors was rapid and each new machine had more RAM. By 1989 the Macintosh SE/30 had a 386 processor, 4 MB of RAM, and an 80 MB hard disk. The cost of personal computing was coming down, but was still high. People wanted one, but were afraid that it would be obsolete in a year.

By 1989, however, computer hardware was faster and had more memory than ever before. The possibility of using animal models for genetic evaluations was finally here.

10.2 Basic Animal Model

In a sire model, the focus was on the sire, and data were on progeny of that sire. By assuming sires were randomly mated to dams, dams were unrelated to each other, and that dams had only one progeny each, then sire models gave a good prediction of the sire. However, the assumptions about random matings became less valid over time. Herds contained cow families of which the producers were proud, and therefore assuming dams were unrelated was also invalid. Dams had several female progeny and this could not be ignored. Sires were related through a sire-MGS relationship matrix which did not account for inbreeding.

With an animal model, the focus is on the cow, and data are on the cows. The animal effect (or cow effect) could be written as

$$a_i = .5\ a_s\ +\ .5\ a_d\ +\ M_i$$

where a_s is the breeding value of the sire, a_d is the breeding value of the dam, and M_i is the Mendelian sampling effect, which accounts for the specific alleles inherited by the animal from its parents. Animals are related through sires and dams, and inbreeding can be included using the algorithm of Meuwissen and Luo (1992). The additive genetic relationship matrix, $\mathbf{A}$, accounts for every specific mating of sires to dams, and it accounts for multiple progeny (male or female) of each parent, and it can account for increases in inbreeding coefficients.

In writing an animal model we must consider the environmental and genetic factors that would influence the cow during the making of its record(s). In a dairy cattle context, a basic animal model for single milk production records per cow would be

$$y_{ijklm} = AMG_i + YM_j + HYS_k + f(g_l)_m + a_m + e_{ijklm}$$

where

y_{ijklm} is a first lactation 305-d yield record of cow m, adjusted for lactation length and number of times milked per day;

AMG_i is a fixed age-month(AM) of calving subclass within a five-year period by region of country group(G), recognizing that differences between ages can change over time due to genetics and nutrition (environment), and differ between regions;

YM_j is a fixed year-month of calving, which can be partitioned according to regions of the country to account for drastic environmental differences;

HYS_k is a random herd-year-season of calving effect, or contemporary group;

$f(g_l)_m$ is a function of phantom group effects for animal m;

a_m is a random additive genetic value of an animal; and

e_{ijklm} is a random residual effect.

10.2.1 Lactation Records

In the past, records were pre-adjusted for age and month of calving, lactation length, and number of times milked per day. Now it was possible to include these factors in the model and to estimate them simultaneously with the genetic values. Instead of re-estimating adjustment factors every 5 to 10 years, they were re-estimated every time genetic evaluations were calculated.

Adjustment factors could be either additive (adding or subtracting values from 305-d yields) or multiplicative (multiplying 305-d yields by constants to increase or decrease them). Because production generally continued to increase over time, 305-d yields were larger and larger values resulting in more variation. Thus, multiplicative factors were thought to be better than additive. Including age effects in the model implied additive adjustments were being applied. Thus, AMG_i was put into the models to allow age-month differences to change with time and remain additive. By making further groups to account for differences in regions of the country, the age-month adjustments could be more accurate than using one set of adjustments for all years and regions of the country.

Multiplicative adjustment factors were still needed for extending incomplete lactations to a 305-d basis. Only records beyond 90 to 120-d in milk were extended (depending on the country). The goal was to include every possible daughter, especially for young sires, so that young bulls got their first proof (ETA or EBV) as soon as possible. Multiplicative factors were also used for adjusting for number of times milked per day. Extension factors also had to be re-estimated every 5 to 10 years.

In North America, Dairy Herd Improvement Associations offered different recording services to producers. The best program was where a supervisor visited each herd about once a month to weigh the morning and evening milk yields and to take samples for fat and protein content analyses at a central lab. Another program was an alternating AM-PM program in which the supervisor only weighed and collected samples from one milking per day, alternating between morning and evening milkings between herd visits. Estimating 24-h milk weights then required adjustment factors too. Finally there was an Owner-Sampler program, in which producers recorded their own weights and sent in milk samples on a regular basis. Owner-Sampler records were considered to be the least accurate milk recording program, and for many years a debate carried on about whether Owner-Sampler records should or should not be included in genetic evaluations. Eventually, they

were allowed into genetic evaluation, but with larger residual variances and hence, less weight in the MME.

Milk recording programs were carried out by governments, by breed associations, by private cooperatives, or by organizations that did progeny testing, AI, milk recording, and genetic evaluations - depending on the country. No matter what the organization structure was in a country, cooperation between breed registration (pedigrees), milk recording, AI, genetic evaluation and government had to exist. Otherwise the breeding program in a country could suffer. Students of Lush that returned to their homelands after graduate training made sure that their respective dairy industries collaborated.

10.2.2 YM and HYS

For the majority of implementations of an animal model for genetic evaluation, the HYS effects have been treated as fixed effects, as a carry-over practice from sire models. No one seemed to critically question this aspect of the model. Henderson originally made HYS fixed to avoid an association between sire true breeding values and the true levels of each HYS. However, in an animal model, that association is not so critical, and therefore, HYS should have been a random factor. If HYS are random, then a fixed factor that accounts for phenotypic trends, such as Year-Month of calving effect, needs to be in the model. If YM effects are omitted when HYS are random, then estimated breeding values can be severely biased.

For a large country, like the USA, YM effects should be separated according to regions of the country (e.g. Southeast, Southwest, West Coast, West, Midwest, East Coast, and Northeast) which clearly delineate environmentally different areas. In Europe countries may have mountainous regions or regions close to the Mediterranean, or areas below sea level.

The HYS effects are assumed to be samples from a large population with mean zero and common variance, σ^2_h, and are independent of each other and independent of levels of other random factors in the model.

10.2.3 Animal Effects

The animal additive genetic effect (individual cow effect, to Henderson) is based upon an infinitesimal genetic model. That is, there are assumed to be an infinitely large number of genes that affect a given trait, like milk production, and each of these genes have a small and fairly equal sized effect on the overall trait. The a_m element of the model accounts for the sum of all of these gene effects over the entire genome. All interactions among gene loci, and all dominance effects are ignored in the animal model. At least, the non-additive effects are assumed

to be inconsequential. Only the genetics that are passed from parent to offspring are considered.

The animal model requires that complete pedigrees are known for all animals with records. A base population is assumed, in which all the individuals were non-selected, non-inbred, and randomly mating. The mean of the base population is zero, and the additive genetic variance is σ_a^2. Thus, the covariance structure is $\mathbf{A}\sigma_a^2$. Fortunately, we do not require having $\mathbf{A}$ created and stored anywhere. The $\mathbf{A}$ matrix is usually 95% full of non-zero numbers, and for anything over 50,000 animals the capacity to store this matrix in memory in a computer becomes stretched to the limit. On the other hand, $\mathbf{A}^{-1}$, is fairly sparse, and the elements of the matrix do not need to be stored, but can be generated when needed using Henderson's rules (1976). After advances in computer hardware, Henderson's method of inverting $\mathbf{A}$ was a key factor in making the implementation of animal models possible.

10.2.4 Phantom Parent Groups

The base population is assumed to be large with individuals mating randomly, and therefore, they would be non-inbred. All other animals have parents and can be traced through the pedigree to this base population. In real life, however, herds go on and off milk recording and in the process create gaps in pedigrees within the herd. Also, cows are transferred from non-milk recorded herds to milk recorded herds, and cows appear "out of nowhere" with unknown parents. Cows are also exchanged between countries (USA and Canada, for example) and their parents can not be traced to the base population in the importing country.

Because animals with unknown parents can not be traced back to the base population, then the unknown parents have to be assigned to phantom parent groups (Robinson 1986, Westell et al. 1988). Phantom parent groups are based upon whether the animal is a male or female, and if the parent is a sire or dam, which gives four pathways of selection. There are sires of cows (SC), sires of bulls (SB), dams of cows (DC), and dams of bulls (DB). Phantom groups are also based upon the year of birth of the animal. Thus, a cow born in 1992 with an unknown dam, the unknown dam is assigned to DC-92, i.e. dams of cows pathway for progeny born in 1992.

Thus, every animal has either a real parent ID or an unknown parent replaced by a phantom parent group ID. The only 'animals' with unknown parents are the phantom parent group IDs. If g_l represents one of the phantom parent groups, then every animal has a function that shows what fraction of each phantom parent group makes up part of the animal's true breeding value. Suppose we have three phantom parent groups, g_1, g_2, and g_3, and animal A has an unknown sire assigned to group 1, and an unknown dam assigned to group 2. The

function is then

$$(0.5\ g_1\ +\ 0.5\ g_2\ +\ 0\ g_3).$$

Animal B might have function

$$(0.5\ g_1\ +\ 0\ g_2\ +\ 0.5\ g_3).$$

Then if animal C is a progeny of A and B, then its function is the average of the two parent functions, giving

$$(0.5\ g_1\ +\ 0.25\ g_2\ +\ 0.25\ g_3).$$

After many generations these functions can become very complicated. Fortunately we do not need to specify them for each animal, as will be shown later through Quaas (1988).

The animal's estimated breeding value is then the sum of $f(\widehat{g}_l)_m$ and $\widehat{a}_m$. Most implementations of animal models utilize phantom parent groups.

10.2.5 Relationship Matrix Inverse

Looking at Henderson's rules for writing $\mathbf{A}^{-1}$, every animal has its parents (or phantom parent group) identified, and there are non-zero coefficients between the animal, its sire, and its dam. A row of $\mathbf{A}^{-1}$ can be written as

$$b_i a_i - .5 b_i a_s - .5 b_i a_d + \sum_j (-.5 b_j a_j + .25 b_j a_i + .25 b_j a_m)$$

where $b_i = 1/\delta_i$, δ_i is the amount of Mendelian sampling variance remaining, a_s is the additive genetic effect of the sire (or sire phantom group) of animal i, a_d is the additive genetic effect of the dam (or dam phantom group) of animal i, a_m is the additive genetic effect of the mate of animal i that produced progeny j, and a_j is the additive genetic effect of progeny j from animals i and m.

The expression can be re-arranged as

$$(b_i + \sum_j .25 b_j) a_i - .5 b_i (a_s + a_d) - \sum_j .5 b_j (a_j - .5 a_m)$$

which can be seen as

- a part for the animal, $(b_i + \sum_j .25 b_j) a_i$,
- a second part for the parent average, $-.5 b_i (a_s + a_d)$, and
- a third part for the sum of contributions from its progeny, adjusted for the mates, $-\sum_j .5 b_j (a_j - .5 a_m)$.

Progeny groups are assumed to be random samples of progeny. There has been no pre-selection of progeny before or after birth to be included as part of the progeny group. Until genomics entered the picture, this assumption has usually been close to true.

The animal model, through the relationship matrix, analyzed by BLUP through MME, provides estimated breeding values for all cows, all sires, and all dams, simultaneously, and accounting for all identified additive relationships.

10.2.6 Residual Effects

The residual effects are assumed to be sampled from one population with a mean of zero and variance of σ_e^2. During the 1990's this assumption was questioned (Meuwissen et al. 1996), and the consensus was that there was a separate population of residual effects for each herd, with zero means, but with different variances. Thus, the concept of heterogeneous variances arose. This also raised the question that possibly the additive genetic variance was also different between herds, and therefore, heritability varied amongst herds.

Many countries now have adjustments for heterogeneous residual variances among herds for the production traits. Meuwissen et al. (1996) provided a sound theoretical approach for solving the problem, although most countries opted for simpler methods.

10.3 BLUP and MME

Let the animal model be written in matrix notation as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za} + \mathbf{Wh} + \mathbf{e},$$

where

y is the vector of single records per animal,

b is the vector of age-month-region group effects, and a vector for year-month-region effects,

a is the vector of animal additive genetic effects,

h is the vector of herd-year-season effects,

Z is the matrix that relates animals to their observations,

g is the vector of phantom parent genetic group effects, and

Q is the matrix of functions of the phantom groups that relate to each animal,

$\mathbf{X}, \mathbf{W}$ are design matrices relating fixed effects and herd-year-season effects to the observation vector, and

$\mathbf{e}$ is the vector of residual effects.

The expectations of the random vectors are null vectors, and the variances are

$$Var \begin{pmatrix} \mathbf{a} \\ \mathbf{h} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_h^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix}.$$

The *Estimated Breeding Value*, EBV, of an animal is equal to

$$\text{Vector of EBVs} = \mathbf{Q}\widehat{\mathbf{g}} + \widehat{\mathbf{a}}.$$

Quaas and Pollak (1981) showed that the MME with phantom parent grouping simplify significantly. The MME are

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'ZQ} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Q'Z'X} & \mathbf{Q'Z'ZQ} & \mathbf{Q'Z'Z} & \mathbf{Q'Z'W} \\ \mathbf{Z'X} & \mathbf{Z'ZQ} & \mathbf{Z'Z} + \mathbf{A^{-1}}\alpha & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'ZQ} & \mathbf{W'Z} & \mathbf{W'W} + \mathbf{I}\alpha_h \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Q'Z'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{pmatrix}.$$

Notice that $\mathbf{Q'}$ times the third row subtracted from the second row gives

$$\mathbf{Q'A^{-1}}\widehat{\mathbf{a}}\alpha = \mathbf{0}.$$

Quaas and Pollak (1981) showed that $\mathbf{Q}\widehat{\mathbf{g}} + \widehat{\mathbf{a}}$ can be computed directly. Note that

$$\begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Q} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Q} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \mathbf{Q}\widehat{\mathbf{g}} + \widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix}.$$

Substituting this equality into the left hand side (LHS) of the MME gives

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{0} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Q'Z'X} & \mathbf{0} & \mathbf{Q'Z'Z} & \mathbf{Q'Z'W} \\ \mathbf{Z'X} & -\mathbf{A^{-1}Q}\alpha & \mathbf{Z'Z} + \mathbf{A^{-1}}\alpha & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{0} & \mathbf{W'Z} & \mathbf{W'W} + \mathbf{I}\alpha_h \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \mathbf{Q}\widehat{\mathbf{g}} + \widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Q'Z'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{pmatrix}.$$

To make the equations symmetric again, both sides of the above equations must be premultiplied by

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{Q}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

This gives the following system of equations as

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{0} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}\alpha & -\mathbf{Q}'\mathbf{A}^{-1}\alpha & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & -\mathbf{A}^{-1}\mathbf{Q}\alpha & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\alpha & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{0} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W}+\mathbf{I}\alpha_h \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{g}} \\ \mathbf{Q}\widehat{\mathbf{g}}+\widehat{\mathbf{a}} \\ \widehat{\mathbf{h}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{0} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{pmatrix}. \quad (10.1)$$

Quaas (1988) examined the structure of $\mathbf{Q}$ and the inverse of $\mathbf{A}$ under phantom parent grouping and noticed that $\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$ and $-\mathbf{Q}'\mathbf{A}^{-1}$ had properties that followed the rules of Henderson (1976) for forming the elements of the inverse of $\mathbf{A}$. Thus, the elements of $\mathbf{A}^{-1}$ and $\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$ and $-\mathbf{Q}'\mathbf{A}^{-1}$ can be created by a simple modification of Henderson's rules. Use δ_i as computed earlier, (i.e. $\delta_i = B_{ii}^{-1}$ for B_{ii} being the fraction of Mendelian sampling variation remaining), and let i refer to the individual animal, let s and d refer to either the parent or the phantom parent group if either is missing, then the rules are

Constant to Add	Location in Matrix
δ_i	(i,i)
$-\delta_i/2$	(i,s),(s,i),(i,d), and (d,i)
$\delta_i/4$	(s,s),(d,d),(s,d), and (d,s)

Thus, $\mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q}$ and $\mathbf{Q}'\mathbf{A}^{-1}$ can be created directly without explicitly forming $\mathbf{Q}$ and without performing the multiplications times $\mathbf{A}^{-1}$. In essence, phantom groups can almost be treated in the same manner as a real animal. The first step would be to use the algorithm of Meuwissen and Luo (1992) to obtain the inbreeding coefficients, F_i, of all the animals, and a B_{ii} value for each animal, where

$$B_{ii} \;=\; 0.5 - 0.25(F_s + F_d)$$

for the case when both s, sire and d, dam are known, and $B_{ii} = 4/3$ if one parent is unknown, and $B_{ii} = 1$ if both parents are unknown. The second step would be to add in the phantom group identification, and let their $B_{ii} = 1$. Then construct the relationship matrix inverse using Henderson's rules and the B_{ii} values of each animal or phantom group.

10.4 Partitioning EBVs

Sometimes partitioning solutions to MME can be informative and helpful to understanding MME. Take the equations for the animal EBVs from 10.1 which are

$$\mathbf{Z'X}\widehat{\mathbf{b}} - \mathbf{A^{-1}Q}\alpha\widehat{\mathbf{g}} + (\mathbf{Z'Z} + \mathbf{A^{-1}}\alpha)\widehat{\mathbf{Qg} + \mathbf{a}} + \mathbf{Z'W}\widehat{\mathbf{h}} \;=\; \mathbf{Z'y}$$

Re-arrange these to give

$$(\mathbf{Z'Z} + \mathbf{A^{-1}}\alpha)\widehat{\mathbf{Qg} + \mathbf{a}} \;=\; \mathbf{Z'y} - \mathbf{Z'X}\widehat{\mathbf{b}} - \mathbf{Z'W}\widehat{\mathbf{h}} + \mathbf{A^{-1}Q}\alpha\widehat{\mathbf{g}}$$

If we consider just one animal, i, having a record, then

$$\begin{aligned}(\mathbf{Z'Z} + \mathbf{A^{-1}}\alpha)\widehat{\mathbf{ebv}} - \mathbf{A^{-1}Q}\alpha\widehat{\mathbf{g}} \;&=\; (1 + b_i\alpha + \sum_j .25b_j\alpha)\widehat{ebv}_i \\ &\quad -.5b_i\alpha(\widehat{ebv}_s + \widehat{ebv}_d) \\ &\quad -\sum_j .5b_j\alpha(\widehat{ebv}_j - .5\widehat{ebv}_m)\end{aligned}$$

and

$$\mathbf{Z'y} - \mathbf{Z'X}\widehat{\mathbf{b}} - \mathbf{Z'W}\widehat{\mathbf{h}} \;=\; (y_{ijklm} - \widehat{AMG}_i - \widehat{YM}_j - \widehat{HYS}_k).$$

Combining results and expressing things in terms of the animal's EBV,

$$\begin{aligned}\widehat{ebv}_i \;&=\; w_1(y_{ijklm} - \widehat{AMG}_i - \widehat{YM}_j - \widehat{HYS}_k) \\ &\quad +w_2(\alpha.5b_i(\widehat{ebv}_s + \widehat{ebv}_d)) \\ &\quad +w_3(\sum_j .25b_j\alpha(\widehat{ebv}_j - .5\widehat{ebv}_m))\end{aligned}$$

where

$$\begin{aligned}D_i \;&=\; (1 + b_i\alpha + \sum_j .25b_j\alpha) \\ w_1 \;&=\; 1/D_i \\ w_2 \;&=\; (b_i\alpha)/D_i \\ w_3 \;&=\; (\sum_j .25b_j\alpha)/D_i\end{aligned}$$

This shows that an EBV can be partitioned into a part due to the record on the animal (if present), a part due to the parent average EBV, and a part due to the sum of its progeny deviated from one-half their dam's EBV. The weights, w_1, w_2, and w_3 depend on the amount of information and α in each part.

Suppose an animal has its own record, both parents are known, and it has 2 progeny, with $\alpha = 1.5$, then

$$\begin{aligned} D_i &= (1 + 2(1.5) + .25(1.5)(2 + 2) \\ &= 5.5 \text{ so that} \\ w_1 &= 0.1818 \\ w_2 &= 0.5454 \\ w_3 &= 0.2727 \end{aligned}$$

In this case the parent average plays a large part in this animal's EBV.

Progeny are more important than the animal's own record. If the animal had 10 progeny (all with both parents known and non-inbred) then

$$\begin{aligned} D_i &= (1 + 2(1.5) + .25(1.5)(2 * 10) \\ &= 11.5 \\ w_1 &= 0.0870 \\ w_2 &= 0.2609 \\ w_3 &= 0.6522 \end{aligned}$$

Now the progeny are carrying more weight in the animal's EBV than its parents, and the contribution of one record diminishes very quickly.

A higher heritability (lower α) also shifts the weights, for example let $\alpha = 1.1$, then

$$\begin{aligned} D_i &= (1 + 2(1.1) + .25(1.1)(2 * 10) \\ &= 8.7 \\ w_1 &= 0.1149 \\ w_2 &= 0.2529 \\ w_3 &= 0.6322 \end{aligned}$$

which results in more weight back on the record, and a little less on the parent average and on the progeny, but progeny are still the major contributor.

The animal model MME produce the appropriate weights on each source of information. Partitioning is not usually calculated, but is used to improve

understanding.

10.5 Accuracies

The residual variance is estimated, as usual, as

$$\sigma_e^2 = (\mathbf{y}'\mathbf{y} - \widehat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \widehat{(\mathbf{Qg} + \mathbf{a})}'\mathbf{Z}'\mathbf{y} - \widehat{\mathbf{h}}'\mathbf{W}'\mathbf{y})/(N - r(\mathbf{X})).$$

Then if the MME coefficient matrix can be inverted, so that c_{ii} represents the diagonal of the inverse for animal i, then the variance of prediction error Prediction Error Variance (PEV) is given by

$$PEV = c_{ii} \times \sigma_e^2.$$

Accuracy is given by the squared correlation between the EBV and the true breeding value,

$$r^2_{a,\widehat{a}} = Cov(a, \widehat{a})/Var(a)$$

where

$$Cov(a, \widehat{a}) = Var(a) - PEV \;=\; [(1 + F_i) - c_{ii}\alpha]\sigma_a^2$$

Then

$$r^2_{a,\widehat{a}} = [(1 + F_i) - c_{ii}\alpha]/[1 + F_i].$$

If the animal is not inbred, then

$$r^2_{a,\widehat{a}} = 1 \;-\; c_{ii}\alpha.$$

Harris and Johnson (1998), Meyer (1987), and others have developed approximate methods to calculate accuracies from animal models. Meyer (1987) uses an "absorption"-like strategy, but not everything gets "absorbed" and so the approximation does not work well on animals with lots of progeny or animals with few progeny. A selection index procedure incorporating

- the number of records on the animal,
- the number of progeny of the animal,
- the number of progeny of the sire of the animal,
- the number of records on the dam of the animal, and
- the number of progeny of the dam of the animal,

seems to be a good approximation of $r^2_{a,\widehat{a}}$. When using an animal model, there are always too many animals being evaluated and thus, the inverse of the coefficient

matrix of the MME is never practical to compute. Thus, one must find a good approximation of either c_{ii} or $r^2_{a,\widehat{a}}$. Unfortunately, there are many approximations from which to choose.

10.6 Reduced Animal Model

In the early days of animal models, not everyone had adequate computing power to set up and solve MME. Thus, the search for computational tricks to reduce the number of equations to be solved was in play. Only animals that were possible candidates for becoming parents of the next generation needed to be evaluated. Eliminating other animals from the data would cause bias by creating a data set with only selected animals, and so the trick was finding a way to use their records without estimating their breeding value. Pollak and Quaas (1980) came up with the reduced animal model or RAM to cover this situation. Animals that would never have any progeny had their true breeding value in the model replaced with

$$a_i = .5a_s + .5a_d + M_i$$

and the M_i was combined with the residual variance to give a larger residual variance for that animal's record. The resulting MME only had equations for animals that either were parents or would eventually be parents. This trick was especially useful for species of animals with high reproductive rates like pigs, poultry, rabbits, fish, beef cattle, or sheep. In dairy cattle, the reproductive rate was not great enough to make a great savings on the size of MME to be solved. This did not matter because main memory in computers, and disk storage capacity grew so quickly during the 1990's that such a trick was not necessary for very long. Consider a very simple animal model with periods as a fixed factor and one observation per animal, as in the table 10.1. There are no phantom groups or random contemporary groups in this example. The purpose is to present an example of the reduced animal model.

10.6.1 Usual Animal Model Analysis

Assume that the ratio of residual to additive genetic variances is $\alpha = 2$. The MME for this data would be of order 11 (nine animals and two periods). The

Table 10.1: Example Data For Reduced Animal Model

Animal	Sire	Dam	Period	Observation
5	1	3	2	250
6	1	3	2	198
7	2	4	2	245
8	2	4	2	260
9	2	4	2	235
4	-	-	1	255
3	-	-	1	200
2	-	-	1	225

left hand sides and right hand sides of the MME are:

$$\begin{pmatrix} 3 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 4 & 0 & 2 & 0 & -2 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 6 & 0 & 3 & 0 & 0 & -2 & -2 & -2 \\ 1 & 0 & 2 & 0 & 5 & 0 & -2 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 3 & 0 & 6 & 0 & 0 & -2 & -2 & -2 \\ 0 & 1 & -2 & 0 & -2 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 5 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 5 \end{pmatrix}, \begin{pmatrix} 680 \\ 1188 \\ 0 \\ 225 \\ 200 \\ 255 \\ 250 \\ 198 \\ 245 \\ 260 \\ 235 \end{pmatrix}$$

and the solutions to these equations are:

$$\begin{pmatrix} \widehat{b}_1 \\ \widehat{b}_2 \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \\ \widehat{a}_6 \\ \widehat{a}_7 \\ \widehat{a}_8 \\ \widehat{a}_9 \end{pmatrix} = \begin{pmatrix} 225.8641 \\ 236.3366 \\ -2.4078 \\ 1.3172 \\ -10.2265 \\ 11.3172 \\ -2.3210 \\ -12.7210 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{pmatrix}.$$

10.6.2 Reduced AM

RAM results in fewer equations to be solved, but the solutions from RAM are exactly the same as from the usual MME. In a typical animal model with **a** as

the vector of additive genetic values of animals, there will be animals that have had progeny, and there will be other animals that have not yet had progeny (and some may never have progeny). Denote animals with progeny as $\mathbf{a_p}$, and those without progeny as $\mathbf{a_o}$, so that

$$\mathbf{a}' = \begin{pmatrix} \mathbf{a}'_\mathbf{p} & \mathbf{a}'_\mathbf{o} \end{pmatrix}.$$

In terms of the example data,

$$\begin{aligned} \mathbf{a}'_\mathbf{p} &= \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix}, \\ \mathbf{a}'_\mathbf{o} &= \begin{pmatrix} a_5 & a_6 & a_7 & a_8 & a_9 \end{pmatrix}. \end{aligned}$$

For any individual, i, the additive genetic value may be written as

$$a_i = .5(a_s + a_d) + M_i.$$

Therefore,

$$\mathbf{a}_o = \mathbf{T}\mathbf{a}_p + \mathbf{m},$$

where $\mathbf{T}$ is a matrix that indicates the parents of each animal in $\mathbf{a}_o$, and $\mathbf{m}$ is the vector of Mendelian sampling effects. Then

$$\begin{aligned} \mathbf{a} &= \begin{pmatrix} \mathbf{a}_p \\ \mathbf{a}_o \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} \\ \mathbf{T} \end{pmatrix} \mathbf{a}_p + \begin{pmatrix} \mathbf{0} \\ \mathbf{m} \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} Var(\mathbf{a}) &= \mathbf{A}\sigma_a^2 \\ &= \begin{pmatrix} \mathbf{I} \\ \mathbf{T} \end{pmatrix} \mathbf{A}_{pp} \begin{pmatrix} \mathbf{I} & \mathbf{T}' \end{pmatrix} \sigma_a^2 + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \sigma_a^2 \end{aligned}$$

where $\mathbf{B}$ is a diagonal matrix with diagonal elements equal to

$$b_i = 0.5 - .25(F_s + F_d),$$

when both parents are known, which is assumed to be true for these particular animals.

$$Var(\mathbf{a}_p) = \mathbf{A}_{pp}\sigma_a^2.$$

The animal model can now be written as

$$\begin{pmatrix} \mathbf{y}_p \\ \mathbf{y}_o \end{pmatrix} = \begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_o \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{Z}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_o \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{T} \end{pmatrix} \mathbf{a}_p + \begin{pmatrix} \mathbf{e}_p \\ \mathbf{e}_o + \mathbf{Z}_o\mathbf{m} \end{pmatrix}.$$

Note that the residual vector has two different types of residuals and that the additive genetic values of animals without progeny have been replaced with $\mathbf{T}\mathbf{a}_p$. Because every individual has only one record, then $\mathbf{Z}_o = \mathbf{I}$, but $\mathbf{Z}_p$ may have fewer rows than there are elements of $\mathbf{a}_p$ because not all parents may have observations themselves. In the example data, animal 1 does not have an observation, therefore,

$$\mathbf{Z}_p = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Consequently,

$$\begin{aligned} \mathbf{R} &= Var \begin{pmatrix} \mathbf{e}_p \\ \mathbf{e}_o + \mathbf{m} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 + \mathbf{B}\sigma_a^2 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_o \end{pmatrix} \sigma_e^2 \end{aligned}$$

The mixed model equations for the reduced animal model are

$$\begin{pmatrix} \mathbf{X}_p'\mathbf{X}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{X}_p'\mathbf{Z}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{T} \\ \mathbf{Z}_p'\mathbf{X}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{Z}_p'\mathbf{Z}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{T} + \mathbf{A}_{pp}^{-1}\alpha \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_p \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}_p'\mathbf{y}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{y}_o \\ \mathbf{Z}_p'\mathbf{y}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{y}_o \end{pmatrix}.$$

Solutions for $\widehat{\mathbf{a}}_o$ are derived from the following formulas.

$$\widehat{\mathbf{a}}_o = \mathbf{T}\widehat{\mathbf{a}}_p + \widehat{\mathbf{m}},$$

where

$$\widehat{\mathbf{m}} = (\mathbf{Z}_o'\mathbf{Z}_o + \mathbf{D}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\widehat{\mathbf{b}} - \mathbf{T}\widehat{\mathbf{a}}_p).$$

Using the example data,

$$\mathbf{T} = \begin{pmatrix} .5 & 0 & .5 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \end{pmatrix},$$

and

$$\mathbf{B} = \text{diag} \begin{pmatrix} .5 & .5 & .5 & .5 & .5 \end{pmatrix},$$

then the MME with $\alpha = 2$ are

$$\begin{pmatrix} 3 & 0 & 0 & 1 & 1 & 1 \\ 0 & 4 & .8 & 1.2 & .8 & 1.2 \\ 0 & .8 & 2.4 & 0 & .4 & 0 \\ 1 & 1.2 & 0 & 3.6 & 0 & .6 \\ 1 & .8 & .4 & 0 & 3.4 & 0 \\ 1 & 1.2 & 0 & .6 & 0 & 3.6 \end{pmatrix} \begin{pmatrix} \widehat{b}_1 \\ \widehat{b}_2 \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \end{pmatrix} = \begin{pmatrix} 680. \\ 950.4 \\ 179.2 \\ 521. \\ 379.2 \\ 551. \end{pmatrix}$$

The solutions are as before, i.e.

$\widehat{b}_1$ = 225.8641

$\widehat{b}_2$ = 236.3366

$\widehat{a}_1$ = -2.4078

$\widehat{a}_2$ = 1.3172

$\widehat{a}_3$ = -10.2265

$\widehat{a}_4$ = 11.3172

10.6.3 Backsolving for Omitted Animals

To compute $\widehat{\mathbf{a}}_o$, first calculate $\widehat{\mathbf{m}}$ as:

$$(\mathbf{I}+\mathbf{B}^{-1}\alpha) = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

$$\mathbf{y}_o = \begin{pmatrix} 250 \\ 198 \\ 245 \\ 260 \\ 235 \end{pmatrix}$$

$$\mathbf{X}_o\widehat{\mathbf{b}} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 225.8641 \\ 236.3366 \end{pmatrix}$$

$$\mathbf{T}\widehat{\mathbf{a}}_p = \begin{pmatrix} -6.3172 \\ -6.3172 \\ 6.3172 \\ 6.3172 \\ 6.3172 \end{pmatrix}$$

$$\begin{aligned} \widehat{\mathbf{m}} &= (\mathbf{I}+\mathbf{B}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\widehat{\mathbf{b}} - \mathbf{T}\widehat{\mathbf{a}}_p) \\ &= \begin{pmatrix} 3.9961 \\ -6.4039 \\ .4692 \\ 3.4692 \\ -1.5308 \end{pmatrix} \end{aligned}$$

and

$$\mathbf{T}\widehat{\mathbf{a}}_p + \widehat{\mathbf{m}} = \begin{pmatrix} -2.3211 \\ -12.7211 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{pmatrix}.$$

10.7 Repeated Records Model

Cows can have several lactation records. The question is whether the successive measures are records on the same genetic trait, or whether the successive measures are different genetic traits, but with high genetic correlation. Assuming the successive lactation records are repeated measures of the same trait, then a repeatability animal model can be employed. Such a model was adopted by the USDA in 1990 (Wiggans and VanRaden, 1991) and discussed by Wiggans (1988). A repeatability model is similar to the single record animal model, but with the addition of a permanent environmental effect for each animal making a lactation record. The model is

$$y_{ijklmn} = PAMG_i + PYM_j + HYS_k + f(g_l)_m + a_m + p_m + e_{ijklmn}$$

where

y_{ijklm} is the n^{th} lactation 305-d yield record of cow m, adjusted for lactation length and number of times milked per day;

$PAMG_i$ is a fixed parity-age-month(AM) of calving subclass within a five-year period by region of country group(G), recognizing that differences between ages can change over time due to genetics and nutrition (environment), and differ between regions;

PYM_j is a fixed parity-year-month (AM) of calving, which can be partitioned according to regions of the country to account for drastic environmental differences;

HYS_k is a random herd-year-season of calving effect, or contemporary group, where contemporary groups could be split by first versus later lactations;

$f(g_l)_m$ is a function of phantom group effects for animal m;

a_m is a random additive genetic value of an animal;

p_m is a random permanent environmental (PE) value of an animal; and

e_{ijklm} is a random residual effect.

A necessary requirement for this model is that every cow should have their first lactation record included. This is so that culling of cows after first lactation can be taken into account.

Age-month of calving effects are different depending on parity number. The effects for a 36 month old cow calving for the first time are different from those for a 36 month old cow calving for the second time. Similarly, trends for year-months of calving will differ by parity number because yields tend to go up in later parities. Contemporary groups should be split by first lactation cows versus all others. This is because first lactation cows have lower yields, but also because heifers are generally raised separately because they are not being milked during their pregnancy. Once a cow calves, she is kept with the older cows.

Permanent environmental effects are non-genetic effects that each cow encounters during their lives that affect every lactation record they make. In comparison to an athlete, for example, the permanent environmental effects is the training the athlete followed prior to becoming competitive. Did the athlete have a good coach, did the athlete carry out the training appropriately, did the training have an adverse effect, or did the training allow the genetic potential to come forward? Permanent environmental effects can be either good or bad, but they have an effect on every performance. Traits are said to have a *repeatability* which is

$$r = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_y^2}$$

where σ_a^2 is the additive genetic variance, σ_p^2 is the permanent environmental variance from cow to cow, and σ_y^2 is the phenotypic variance of the trait. Repeatabilities go from 0 to 1, but because of the definition should always be greater than the heritability of the trait.

If $\mathbf{h}$ is the vector of herd-year-season effects, $\mathbf{a}$ is the vector of animal additive genetic effects, $\mathbf{p}$ is the vector of permanent environmental effects of cows that made records, and $\mathbf{e}$ is the vector of residual effects, then the assumed covariance matrix is

$$Var \begin{pmatrix} \mathbf{h} \\ \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{I}\sigma_h^2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix}.$$

The design matrix for the animal additive genetic effects, $\mathbf{Z}_a$, and the matrix for permanent environmental effects, $\mathbf{Z}_p$, are not identity matrices. A column for an animal has n ones corresponding to the records of that cow.

10.7.1 Numerical Example

Table 10.2: Example for Repeated Records Model

Animal	Sire	Dam	Year 1 y_{1jk}	Year 2 y_{2jk}	Year 3 y_{3jk}
1	-	-			
2	-	-			
3	-	-			
4	-	-			
5	-	-			
6	-	-			
7	1	2	39	51	62
8	3	4	48	72	
9	5	6	71		96
10	1	4		56	47
11	3	6			86
12	1	2		46	

None of the animals are inbred (Table 10.2), so that the inverse of the additive genetic relationship matrix is

$$\mathbf{A}^{-1} = \frac{1}{2}\begin{pmatrix} 5 & 2 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & -2 & 0 & -2 \\ 2 & 4 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 4 & 1 & 0 & 1 & 0 & -2 & 0 & 0 & -2 & 0 \\ 1 & 0 & 1 & 4 & 0 & 0 & 0 & -2 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 4 & 0 & 0 & -2 & 0 & -2 & 0 \\ -2 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & -2 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & -2 & 0 & 0 & 4 & 0 & 0 & 0 \\ -2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 \\ -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Let

$$\mathbf{W} = \left[\begin{array}{cccc} \mathbf{X} & \left(\begin{array}{cc} \mathbf{0} & \mathbf{Z} \end{array} \right) & \mathbf{Z} \end{array} \right],$$

then

$$\mathbf{W}'\mathbf{W} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}, \quad \mathbf{W}'\mathbf{y} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{0} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{00}k_a & \mathbf{A}^{0r}k_a & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{r0}k_a & \mathbf{A}^{rr}k_a & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}k_p \end{pmatrix},$$

where $\mathbf{A}^{ij}$ are corresponding elements of the inverse of the additive genetic relationship matrix (given earlier) partitioned according to animals without and with records. In this example, each submatrix is of order 6. Also,

$$k_a = \sigma_e^2/\sigma_a^2 = 1.33333, \quad \text{and} k_p = \sigma_e^2/\sigma_p^2 = 3.$$

MME are therefore,

$$\begin{aligned} (\mathbf{W}'\mathbf{W} + \Sigma)\beta &= \mathbf{W}'\mathbf{y} \\ (\mathbf{W}'\mathbf{W} + \Sigma)\beta &= \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{0} & \mathbf{A}^{00}k_a & \mathbf{A}^{0r}k_a & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{A}^{r0}k_a & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{rr}k_a & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_0 \\ \widehat{\mathbf{a}}_r \\ \widehat{\mathbf{p}} \end{pmatrix}. \end{aligned}$$

Let a generalized inverse of the coefficient matrix be represented as

$$(\mathbf{W}'\mathbf{W} + \Sigma)^- = \begin{pmatrix} - & - & - \\ - & \mathbf{C}_{aa} & - \\ - & - & \mathbf{C}_{pp} \end{pmatrix},$$

where $\mathbf{C}_{aa}$ is of order 12 in this case, and $\mathbf{C}_{pp}$ is of order 6. The full HMME are too large to present here as a whole, so parts of the matrix are given as follows.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{pmatrix} 158 \\ 225 \\ 291 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\mathbf{Z}'\mathbf{Z} = diag(3 \;\; 2 \;\; 2 \;\; 2 \;\; 1 \;\; 1),$$

and

$$\mathbf{Z}'\mathbf{y} = \begin{pmatrix} 152 \\ 120 \\ 167 \\ 103 \\ 86 \\ 46 \end{pmatrix}.$$

The solutions for animals are given in the table 10.3. Solutions for year effects were

$$\begin{aligned} \widehat{t}_1 &= 50.0858, \\ \widehat{t}_2 &= 63.9612, \\ \widehat{t}_3 &= 72.0582. \end{aligned}$$

Table 10.3: Solutions for Example Data

Animal	$\widehat{\mathbf{a}}$	$\widehat{\mathbf{p}}$
1	-7.9356	
2	-4.4473	
3	2.8573	
4	-2.6039	
5	5.0783	
6	7.0512	
7	-8.0551	-1.6566
8	1.0111	0.7861
9	11.1430	4.5140
10	-8.7580	-3.1007
11	6.9271	1.7537
12	-8.7750	-2.2965

10.7.2 Cumulative Permanent Environments

Schaeffer (2011) presented a case for permanent environmental effects being cumulative in nature. Animals experience new environmental influences every day of their lives. Thus, there would be permanent environmental effects that would influence the first record of a cow, and during that lactation the cow would be influenced by new effects which add on to those already present. Hence, each record would have a different PE effect, but which would include the PE effect on the previous record. This can be accommodated by allowing the design matrix for permanent environmental effects to have more than one non zero element in a row. For example, suppose a cow has four records, then the appropriate design matrix for this cow would be

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \quad \mathbf{Z}_p = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Thus, there would be 4 PE effects to estimate for that animal. The last row of $\mathbf{Z}_p$ says that y_4 has 4 PE effects affecting it, the same three that affected the previous record, y_3, plus a new one.

Questions that remain unanswered are whether each PE effect is coming from the same population of PE effects, or different populations. If we assume different populations, then the covariance matrix of the PE effects would be:

$$Var \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} \sigma^2_{p1} & \sigma^2_{p1} & \sigma^2_{p1} & \sigma^2_{p1} \\ \sigma^2_{p1} & (\sigma^2_{p1} + \sigma^2_{p2}) & (\sigma^2_{p1} + \sigma^2_{p2}) & (\sigma^2_{p1} + \sigma^2_{p2}) \\ \sigma^2_{p1} & (\sigma^2_{p1} + \sigma^2_{p2}) & (\sigma^2_{p1} + \sigma^2_{p2} + \sigma^2_{p3}) & (\sigma^2_{p1} + \sigma^2_{p2} + \sigma^2_{p3}) \\ \sigma^2_{p1} & (\sigma^2_{p1} + \sigma^2_{p2}) & (\sigma^2_{p1} + \sigma^2_{p2} + \sigma^2_{p3}) & (\sigma^2_{p1} + \sigma^2_{p2} + \sigma^2_{p3} + \sigma^2_{p4}) \end{pmatrix}.$$

Possibly the size of these variances decrease as more records are made. Or the variance of PE effects are the same, and therefore, PE variance increases with record number.

The concept of cumulative PE effects is relatively new, and has not been implemented into any genetic evaluation models. If such effects exist, then perhaps each record should be considered a different trait, and then the PE effects can be combined with the temporary environmental effects in a multiple trait analysis.

10.8 Heterogeneous Variances

The variability of production records in some herds was found to be much greater or much lower than the average herd. The consequence of unequal herd variances was that more animals would be selected from herds with greater variation. This was especially important in the selection of bull dams (Hill, 1984). An attempt to solve the problem was to group herds by level of production and to estimate variances within groups. However, a high herd average does not necessarily mean high variance. In fact, Winkelman and Schaeffer (1988) found no relationship between herd means and herd variances. Also, within-herd residual and sire variances seemed to vary together so that heritability seemed to be constant across herds. Usually, comparisons of herds were made using phenotypic variances (Hill, 1984). Phenotypic variances are the sum of genetic, contemporary group, and residual variances in most situations. Thus, variation could be in one component or may be due to variation in all three components. The problem was the estimation of these variances with enough accuracy to know if the differences are real or are only due to random variability. Herds are not very large and cover many years of data collection, so that estimation of within herd variances for genetic merit and residual effects is problematic. Gianola et al. (1992) and Weigel and Gianola (1992) proposed a Bayes method to estimate within herd variances. Thus, within herd estimates were weighted towards an overall residual and sire variances based on herd size using priors and hyperparameters known as degrees of belief. Meuwissen et al. (1996) also proposed a method for simultaneously estimating herd variances and breeding values using an autoregressive structure within herds over years. The estimated autocorrelation was 0.984 so that herd-year-season variances were similar within a herd, but between herds were more different. Because heritability seemed to be constant, a simple approach was to calculate phenotypic or residual variances within each contemporary group, call it S_k, and let the average residual variance be V, then the sample variance was regressed towards the average variance using

$$S_k^* = \left[\frac{n_k}{n_k + \gamma} (S_k - V) \right] + V,$$

where γ had to be chosen appropriately, and n_k was the number of records in that herd-year-season. The smaller is n_k, then the closer S_k^* should be to V. If n_k is large, then the closer S_k^* should be to S_k.

10.8.1 Numerical Example

Assume a simple model,

$$y_{ijk} = Y_i + HY_{ij} + a_k + e_{ijk}$$

where

y_{ijk} is a production record on a cow,

Y_i is a fixed year effect,

HY_{ij} is a random herd-year effect,

a_k is a random animal additive genetic effect, and

e_{ijk} is a random residual effect assumed to have come from a different population within each herd-year.

Animals are assumed to be traced to the same base population through the pedigrees, thus, there is no need to have phantom genetic groups in the model.

Table 10.4: Heterogeneous variances example data

Year	HY	cow	sire	dam	protein,kg
1	1	14	1	5	230
	1	15	1	6	310
	1	16	2	7	260
	1	17	2	8	250
1	2	18	1	9	280
	2	19	3	10	320
	2	20	3	11	340
	2	21	4	12	270
	2	22	4	13	240
2	3	23	2	5	250
	3	24	3	6	290
	3	25	4	7	220
2	4	26	1	8	290
	4	27	2	9	310
	4	28	3	10	300
	4	29	4	11	210
	4	30	1	12	220
	4	31	2	13	320

The ratio of the average residual variance to additive genetic variance is $\alpha = 1.4$. The example data are presented in the table 10.4.

To get starting values for the residual variances, use the phenotypic variances for each HY. The overall phenotypic variance of all 18 yields was 1574.183. Below are the phenotypic variances for the 4 HY, their regressed values, and their values relative to the overall phenotypic variance (Table 10.5). The regression towards the overall variance used $\gamma = 10$.

Table 10.5: Within herd-year phenotypic variances

HY	n_k	mean	S_k	S_k^*	S_k^*/V
1	4	262.5	1158.33	1455.37	0.924523
2	5	290.0	1600.00	1582.79	1.005467
3	3	253.3	1233.33	1495.52	0.950033
4	6	275.0	2270.00	1835.11	1.165757

Instead of $\mathbf{R}$ being an identity matrix times a scalar constant, now it is a block diagonal matrix,

$$\mathbf{R} = \begin{pmatrix} \mathbf{I}_4(0.9245) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_5(1.0055) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_3(0.9500) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_6(1.1658) \end{pmatrix} \sigma_e^2$$

The ratio of σ_e^2/σ_a^2 is 1.4, and σ_e^2 is an average residual variance for all records. $\mathbf{R}^{-1}$ goes into the MME, so that HY with larger variance are given less weight in the equations. Also, the ratio $\sigma_e^2/\sigma_h^2 = 4$ was assumed. After the MME are constructed and solved, then calculate the residuals of each observation,

$$\widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}} - \mathbf{W}\widehat{\mathbf{h}} - \mathbf{Z}\widehat{\mathbf{a}})$$

The overall variance of $\widehat{\mathbf{e}}$ was 503.2015. Now compute the variances of the residuals for each HY, regress those towards the overall variance, and then express them relative to the overall variance (Table 10.6). HY 2 and 3 have smaller residual variances compared to their relative phenotypic variances, while HY 4 has a much larger residual variance. The relationship of residual variances to phenotypic means is not very strong.

The new relative values go into $\mathbf{R}$ and the process should be iterated a few more times, although the changes to the diagonals in $\mathbf{R}$ will not be very large and should settle after 5 iterations.

Table 10.6: Within herd-year residual variances

HY	n_k	S_k^*	S_k^*/V
1	4	494.48	0.9812
2	5	483.33	0.9354
3	3	424.09	0.8347
4	6	662.62	1.3482

Below (Table 10.7) is a comparison of sire EBVs when heterogeneous residual variances are used or if homogeneity is assumed. In this example there was

Table 10.7: Sire EBVs for two models

Sire	HOM Var	HET Var
1	-5.61	-4.82
2	8.04	5.84
3	23.53	23.93
4	-25.96	-24.95

no re-ranking of sires or dams(not shown), but the cows with records in HY 4 had lower EBVs after heterogeneous variance adjustments, as would be expected. The other cows had only minor changes to their EBVs because the relative values in **R** were closer to one for those HY.

10.9 References

GIANOLA, D. , J. L. FOULLEY, R. L. FERNANDO, C. R. HENDERSON, K. A. WEIGEL. 1992. Estimation of heterogeneous variances using empirical Bayes methods: Theoretical considerations. Journal of Dairy Science, 75:2805-2823.

HARRIS, B. , D. JOHNSON. 1998. Approximate reliability of genetic evaluations under an animal model. Journal of Dairy Science, 81:2723-2728.

HENDERSON, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. Biometrics 32:69.

HILL, W. G. 1984. On selection among groups with heterogeneous variance. Anim. Prod. 39:473.

MEUWISSEN, T. H. E. , Z. LUO. 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24:305-313.

MEUWISSEN, T. H. E. , G. DeJONG, B. ENGEL. 1996. Joint estimation of breeding values and heterogeneous variances of large data files. Journal of Dairy Science, 79:310-316.

QUAAS, R. L. and POLLAK E.J., 1980. Mixed model methodology for form and ranch beef cattle testing programs. J.Anim.Sci. 51:1277.

QUAAS, R. L. , E. J. POLLAK. 1981. Modified equations for sire models with groups. Journal of Dairy Science, 64:1868-1872.

QUAAS, R. L. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science, 71:1338-1345.

ROBINSON, G. K. 1986. Group effects and computing strategies for models for estimating breeding values. Journal of Dairy Science, 69:3106-3111.

SCHAEFFER, L. R. 2011. Cumulative permanent environmental effects in a repeated records animal model. J. Anim. Breed. Genet. 128(2):95-9.

VanRADEN, P. M. , and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. Journal of Dairy Science, 74:2737 2746.

WEIGEL, K. A. , D. GIANOLA. 1992. Estimation of heterogeneous within-herd variance components using empirical Bayes methods: A simulation study. Journal of Dairy Science, 75:2824-2833.

WESTELL, R. A. , R. L. QUAAS, L. D. VAN VLECK. 1988. Genetic groups in an animal model. Journal of Dairy Science, 71:1310-1318.

WIGGANS, G. R. , I. MISZTAL, L. D. VAN VLECK. 1988. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. Journal of Dairy Science, 71:54-69.

WINKELMAN, A. , L. R. SCHAEFFER. 1988. Effect of heterogeneity of variance in dairy sire evaluation. Journal of Dairy Science, 71:3033.

Chapter 11

International Models

GEORGIOS BANOS
LARRY SCHAEFFER

11.1 The Holstein-Friesian

The Holstein-Friesian breed of dairy cattle originated in the Netherlands (Holland). The first established herd in the United States was 1869, and the first herd in Ontario, Canada was 1881 (Lewington, 1983). The United States placed much emphasis on improved milk production and were successful at it, such that in the 1960's US cattle were being exported back to the Netherlands and into Europe and elsewhere in the world. Canadian cattle also started to be exported around the globe.

Although the Holstein breed composes 90% or more of the cattle in North America, many red breeds are the majority in other countries, like Sweden and Norway. However, the Holstein breeds drove the need for international sire comparisons. Due to increased movement of Holstein genetics during the 1970's from North America to other parts of the world, it became important to be able to compare bulls from the US and Canada to bulls from the importing countries. One problem was that each country had its own system of milk recording and genetic evaluations, and more importantly, each country had different standards and methods of expressing EBVs of bulls. Importers of bull semen were facing the challenge of selecting sires from several exporting countries. Producers understood their own country's EBV system, but did not know or trust the EBV system in other countries, and therefore, they did not know how to rank foreign bulls compared to their own. By the same token, it was important for the semen exporter to make sure their bulls were ranked highly, by some means, in

the importing countries or that producers in the importing country knew how to interpret foreign EBVs.

11.1.1 International Friesian Strain Comparison Trial

The Food and Agriculture Organization (FAO) of the United Nations organized a country comparison trial in Poland in the mid 1970's, the most extensive cattle experiments of all time. During the first three years, 80,000 doses of semen from ten different countries were used on 30,000 Polish Black and White cows located on 70 state farms. The countries involved were Canada, the United States, the United Kingdom, Germany, Denmark, Israel, the Netherlands, Sweden, New Zealand, and Poland. Each country was to choose a random sample of young sires in their progeny test programs that were eligible for exportation to Poland. Thus, it would take almost 5 years before those daughters completed their first lactations so that comparisons between countries could be made. As well as production traits (milk, fat, protein), beef performance, meat quality, body size, feed conversion, health, and reproduction were carefully recorded. At the insistence of Canada and the United States, type classification traits were also scored. Enough traits were measured in the trial such that every country was superior for at least one trait, and none were at the bottom for all traits. The organizers on the Polish side were Jasiorowski, Stolzman, and Reklewski (1988).

While the idea looked reasonable on paper, the notion of a fair comparison was not possible. The samples of bulls were not totally random. However, Poland benefitted greatly by receiving all of this free superior genetic material from several countries. The final results were published in 1988, but by that time each country had improved genetically by different amounts, so that the results no longer applied to the current generation of bulls.

A similar trial for eight red and white breeds was set up in Bulgaria, but the results were never published. Importers needed a faster method of getting good comparisons among individual bulls from different countries.

11.2 Conversion Methods

Banos (2010) described the history of the formation of Interbull which began with debates over conversion methods. In 1981 the International Dairy Federation (IDF) sanctioned the use of equations to convert a sire's genetic merit assessed in one country to the genetic base and scale of another country. The approach was to use a simple regression,

$$\text{EBV}_I = \ a \ + \ b(\text{EBV}_E) + e$$

where EBV_I is the EBV in the Importing country, and EBV_E is the EBV in the Exporting country. Thus, there had to be a number of bulls that had enough daughters in both countries in order to obtain estimates of a, the intercept, and b, the slope. Then those a and b values were applied to all EBV_E to convert them to the Importing country mean and scale. The debate became who was responsible for calculating a and b, the Importing, the Exporting, or some other country. The results varied tremendously depending on which bulls were included in the analyses and who was doing the calculations. Goddard (1985) and Wilmink et al. (1986) modified the regression approach to account for the accuracies of the EBVs from each country. Regressions were used for more than 2 decades.

Another source of bias was the proofs of foreign bulls in an importing country. If a bull was being imported to another country, the semen was probably worth much more than semen of domestic bulls in the importing country. The result was that only the more prosperous herds could afford to buy the semen, and this semen was undoubtedly used to breed the superior animals of the herd. Also, the daughters of those bulls could be preferentially treated once they were in the herd. Therefore, the EBV of the foreign bull in the importing country could be significantly biased. Biased EBVs should not be used in deriving a and b values.

The biases could be noticed by trying to derive reciprocal equations to convert back and forth between countries. For example, if

$$\begin{aligned} \text{EBV}_I &= a_E + b_E \cdot \text{EBV}_E \\ \text{EBV}_E &= a_I + b_I \cdot \text{EBV}_I \end{aligned}$$

when you try to convert EBV_I back to EBV_E, you get

$$\text{EBV}_E = a_I + b_I \cdot (a_E + b_E \cdot \text{EBV}_E)$$

which only works if $a_I = -b_I \cdot a_E$ and if $b_I \times b_E = 1$, and this seldom happened.

Another problem with the conversion method was that bulls from the exporting country, after conversion, would rank exactly the same in the importing country as they did in their home country. Thus, there was no allowance for the possibility of a genotype by environment interaction. EBV of foreign bulls in importing countries often did not rank the same as in their home countries.

Finally, there were sometimes importing countries that did not have any bulls from country X proven in their country, and therefore, there was no data from which to calculate a and b values. However, some manipulations were done sometimes involving three countries. For example, bulls from country A were used in country B, but not in country C, and bulls from country B were used in country

C. Then the conversion equation of EBV from country A to country B and the conversion equation from country B to country C could give a conversion equation from country A to country C, indirectly. These types of conversion equations were always less useful than those having actual data between the two countries, and were discouraged.

Conversion equations were always limited to a comparison of bulls in two countries. The methods did not allow good comparisons between 20 or more countries at the same time.

11.3 Linear Model

Schaeffer (1985) proposed an alternative method to conversion for the comparison of bull EBV across country. The method was based on a linear model fitting the effects of country of evaluation, genetic groups, and sire to national EBV of bulls from different countries. This approach allowed the inclusion of all bulls with a national genetic evaluation (not just those with EBV in more than one country). Bulls were linked through their pedigree, making it possible to express the genetic merit of all bulls, independently of the country of origin, on the scale and base of each country separately. However, all bull rankings were initially identical as the model assumed a genetic correlation of unity among different countries, implying no genotype by environment interactions, and constituting an important limitation of the method. Rozzi et al. (1990) applied this model to a few countries.

Schaeffer (1994) extended the model to account for a genetic correlation among countries of less than unity, thereby allowing bulls to be ranked differently in each country, depending on the locally prevailing conditions. The new method was termed Multiple Across Country Evaluation (MACE) and required estimates of genetic parameters for the participating countries. Different methods for estimating the genetic parameters were proposed. MACE was a crucial development in international genetic evaluations that paved the way for the Interbull services.

11.3.1 The Model

The first step in MACE is to convert sire proofs (EBV or ETA) into de-regressed proofs, one country at a time, using an estimate of the effective number of daughters in each country. An overall pedigree file is used for bulls from all countries, based on sire, maternal grandsire, and maternal granddam, for which a set of rules similar to those of Henderson were derived. The model for a sire de-regressed proof, y_{kji}, for bull i in country k is

$$y_{kji} = \mu_k + g_{kj} + s_{kji} + e_{kji}$$

for μ_k being the country mean proof, g_{kj} is a genetic group (defined over bulls in all countries in MACE), and s_{kji} is the sire true proof, and e_{kji} is the residual which has variance equal to

$$Var(e_{kji}) = \sigma^2_{e_k}/d_{ki}$$

where d_{ki} is the effective number of daughters of bull i in country k, and $\sigma^2_{e_k}$ is the residual variance of country k. The variance of sire true proofs is $\mathbf{A}\sigma^2_s$, where $\mathbf{A}$ is based on all animals going into MACE. Either the ratio of residual to sire variance is known, or it has to be estimated.

Usually y_{kji} is known and the $g_{kj} + s_{kji}$ have to be predicted, but in de-regression, we want to find y_{kji} such that they give us the sire proofs that have been provided. We also need to determine the country mean and group effects. So we must work iteratively. Construct MME, except for the right hand sides. Then put in the sire proofs for the bulls from one country and these can generate y_{kji}, which are then used to estimate μ_k, and then all other animals are solved based on elements of $\mathbf{A}^{-1}$. Using the full set of solutions, then the entire process is repeated.

1. For sires with proofs,

$$\mathbf{Z'Dy} = \mathbf{Z'DX}\mu_k + \mathbf{Z'DZQg} + (\mathbf{Z'DZ} + \mathbf{A}^{ww}\alpha)\mathbf{s}_w + \mathbf{A}^{wo}\mathbf{s}_o$$

   where $\mathbf{X}$ is a column of ones, $\mathbf{Z}$ is an identity matrix, $\mathbf{A}^{ww}$ are the inverse elements of the bulls with proofs in country k, $\mathbf{s}_w$, $\mathbf{A}^{wo}$ are the inverse elements of bulls with proofs in country k with all other relatives in all countries NOT having proofs in country k, $\mathbf{s}_o$, and $\mathbf{D}$ is a diagonal matrix with elements equal to the number of effective daughters in country k for each bull with a proof in country k.

2. To estimate μ_k, then

$$\mu_k = \sum_i \mathbf{Z'Dy} / \sum_i \mathbf{Z'DZ} = \mathbf{X'Dy}/\mathbf{X'DX}$$

   where $\mathbf{X'DX}$ is a scalar equal to the total effective number of daughters of all bulls in country k.

3. Solve for $\mathbf{s}_o$ as

$$\mathbf{A}^{oo}\mathbf{s}_o = \mathbf{A}^{ow}\mathbf{s}_w.$$

The rules for making $\mathbf{A}^{-1}$ can be summarized in the following table. The value of x is $16/(m+11)$ where $m = 0$ if both sire and MGS are known, $m = 1$

if the sire is known and MGS is unknown, $m = 4$ if sire is unknown and MGS known, and $m = 5$ if both sire and MGS are unknown. The MGD is assumed unknown in all cases. Inbreeding is ignored.

	Bull	Sire	MGS	MGD
Bull	x	-.5x	-.25x	-.25x
Sire	-.5x	.25 x	.125 x	.125x
MGS	-.25x	.125x	.0625x	.0625x
MGD	-.25x	.125x	.0625x	.0625x

The deregressed proofs are then

$$\mathbf{y}_k = (\mathbf{Z}'\mathbf{D}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}\mathbf{y}.$$

Once all countries have had the sire proofs de-regressed, then they go into MACE, which is a multiple trait version of the equations that were just used. Correlations among countries either have to be assumed or estimated. There have been several proposed methods for estimating the correlations.

11.3.2 Numerical Example of MACE

The following example (Table 11.1) is taken exactly from Schaeffer (1994). Assume the following pedigrees and proofs of bulls from two countries. Phantom groups are indicated with the letter P in front of the group number. This example has 6 phantom groups.

Bull 1 has daughters in both countries. Sire 10 has a son in both

Bull 2 has daughters in country A but a son in country B.

Bulls 9 and 11 appear only in their respective countries as MGS.

The relationship matrix inverse contains both animals and phantom groups.

The part for the real animals is

$$\mathbf{A}^{-1} = \frac{1}{176}\begin{pmatrix}
256 & 0 & 0 & 0 & 0 & -128 & -64 & 0 & 0 & 0 & 0 \\
0 & 320 & 0 & 0 & -128 & 32 & 0 & -128 & -64 & 0 & 0 \\
0 & 0 & 256 & 0 & 0 & 0 & 0 & -64 & 0 & -128 & 0 \\
0 & 0 & 0 & 256 & 0 & 0 & 0 & 0 & 0 & -128 & -64 \\
0 & -128 & 0 & 0 & 256 & -64 & 0 & 0 & 0 & 0 & 0 \\
-128 & 32 & 0 & 0 & -64 & 256 & 32 & 0 & 0 & 0 & 0 \\
-64 & 0 & 0 & 0 & 0 & 32 & 192 & 0 & 0 & 0 & 0 \\
0 & -128 & -64 & 0 & 0 & 0 & 0 & 256 & 32 & 32 & 0 \\
0 & -64 & 0 & 0 & 0 & 0 & 0 & 32 & 192 & 0 & 0 \\
0 & 0 & -128 & -128 & 0 & 0 & 0 & 32 & 0 & 304 & 32 \\
0 & 0 & 0 & -64 & 0 & 0 & 0 & 0 & 0 & 32 & 192
\end{pmatrix}$$

Table 11.1: Example Data for MACE d is effective number of daughters DRP is de-regressed proof

Bull	Sire	MGS	MGD	m	Country A		Country B	
					d_A	DRP_A	d_B	DRP_B
1	6	7	P5	0	10	+56	100	+9
2	8	9	P5	0	20	-23		
3	10	8	P5	0	50	+8		
4	10	11	P6	0			40	+3
5	2	6	P6	0			20	-11
6	P1	P2	P6	5				
7	P1	P2	P6	5				
8	P1	P2	P6	5				
9	P3	P4	P6	5				
10	P3	P4	P6	5				
11	P3	P4	P6	5				

and the coefficients between animals and phantom groups are in the Table 11.2. All values multiplied by 176.

The assumed heritabilities were 0.36 for country A, and 0.235 for country B.

The assumed genetic correlation between countries was 0.89.

The resulting solutions from MACE are in the Table 11.3.

To compare bull 4 in country A, add 10.50 and 4.30 to give 14.80, which is then comparable to other proofs in country A. Bull 4's proof in country B would be 1.21 plus 1.12 or 3.33 which is similar to its within country proof.

Bull 1 had progeny in both countries, from MACE its proof in country A should be 41.63 (compared to 56 based on 10 effective daughters) and in country B should be 8.23 (compared to 9 based on 100 effective daughters). Thus, MACE combines information on a bull from different countries, and incorporates relationships to bulls within and across countries. MACE can handle any number of countries.

Suppose the daughters of bull 1 in country A were suspected of being from highly selected dams, and preferentially treated daughters, then bull 1's proof in country A could be omitted from the analysis. Sons of bull 1 that enter into country A's progeny testing programs should receive daughters that are not from highly selected dams nor preferentially treated daughters, and therefore, their proofs in country A should be nearly unbiased. MACE relies on each country doing a good job at national genetic evaluations. If a country has a bias in their evaluations, then that could affect MACE results, for bulls in all countries. Hence

Table 11.2: Bull by phantom group coefficients of the relationship matrix inverse

ID	P1	P2	P3	P4	P5	P6
1	0	0	0	0	-64	0
2	0	0	0	0	-64	32
3	0	0	0	0	-64	0
4	0	0	0	0	0	-64
5	0	0	0	0	0	-64
6	-88	-44	0	0	32	-28
7	-88	-44	0	0	16	-44
8	-88	-44	0	0	48	-44
9	0	0	-88	-44	16	-44
10	0	0	-88	-44	32	-12
11	0	0	-88	-44	0	-28
P1	308	66	0	0	0	66
P2	66	209	0	0	0	33
P3	0	0	308	66	0	66
P4	0	0	66	209	0	33
P5	0	0	0	0	224	0
P6	0	0	0	0	0	224

the need for data validation prior to putting data into MACE.

11.4 The International Bull Evaluation Service

Interbull was formed in 1983 as a joint venture of the International Committee for Animal Recording (ICAR), the European Association for Animal Production, and the International Dairy Federation. The mandate of Interbull was to monitor, support and promote developments in the international genetic evaluation of dairy sires. In 1991 the Interbull Centre was established in Uppsala, Sweden, as the headquarters of the organization. For the first 3 years research on international genetic evaluations was collaborative in nature exemplified by different joint projects with the European Community and a consortium of Nordic countries. At the same time MACE was developed allowing a large-scale implementation of international genetic evaluations. Year 1994 marked the onset of routine genetic evaluation services, using national genetic evaluations from 4 different countries (Denmark, Finland, Norway and Sweden) and two breeds (Ayrshire and Holstein) as input. In February 1995 the number of countries in the service increased to 10, including most of the major semen exporters.

Table 11.3: Solutions to multiple country analysis (MACE)

Item	Country A	Country B
Mean	10.50	1.21
bull 1	31.13	7.02
bull 2	-26.54	-5.95
bull 3	-2.41	-0.48
bull 4	4.30	1.12
bull 5	-29.22	-7.07
bull 6	10.94	2.32
bull 7	9.00	2.01
bull 8	-12.88	-2.92
bull 9	-8.30	-1.84
bull 10	1.47	0.44
bull 11	0.05	0.07
P1	2.59	0.55
P2	1.29	0.28
P3	-0.98	-0.15
P4	-0.49	-0.08
P5	1.56	0.39
P6	-3.98	-0.99

11.5 From Past to Present

There were a number of technical issues with MACE and the Interbull Services that needed attention before there would be worldwide acceptance of international genetic evaluations. Scientists from around the world joined forces and conducted valuable research aimed at the technical issues.

11.5.1 Data Validation

The quality of International Genetic Evaluation (IGE) was the main issue. Quality was dependent on the National Genetic Evaluation (NGE) that were the input into MACE. Any problems in NGE would permeate into MACE and reduce the effectiveness of IGE. Boichard et al. (1995) developed three methods to validate the genetic trend in NGE. Genetic trend was found to be overestimated in a number of countries, resulting in considerable bias in IGE. All participating countries had to pass the 3 tests in order to have their NGE included in MACE. Thus, the countries that had overestimates of genetic trend, then had to figure out how to correct their national genetic evaluation models and methods. Klei et

al. (2002) proposed a method to improve error detection in NGE, based on the consistency of EBV and their reliabilities in consecutive evaluation runs.

11.5.2 Dependent Variables

Initially MACE used NGE as the observations in the model, but because some bull EBVs would be less than 99% accurate, the EBV had to be de-regressed so that all fixed and random effects that are in the MACE model are removed from the national EBV prior to MACE. This avoids double regressing the EBVs and compromising the variance of the dependent variable. The de-regression process is based on work by Sigurdsson and Banos (1995) and Jairath et al. (1998).

11.5.3 Genetic Correlations

Genetic variances and covariances constitute an important feature of IGE. The genetic correlations were originally meant to describe genotype by environment (GxE) interaction, that daughters of a bull may fare differently in different countries and environments. However, correlation estimates also reflect differences in data, genetic evaluation models, and trait definitions across countries. The first few IGE used a genetic correlation of 0.995 due to a lack of appropriate estimates. Thus, the bulls ranked almost identically in all countries.

Sigurdsson et al. (1996) proposed the first method for estimating the correlations and it was implemented by Interbull. Klei and Weigel (1998) proposed an alternative method that was used for conformation traits, and by January 2004, all traits analyzed by Interbull used this method. The method was computationally demanding and not all countries could be included in the estimation at one time. Subsets of data had to be run in sequential steps and then the results combined to produce the final matrix, with the need to insure the matrix was positive definite, Jorjani et al. (2003). By 2010, the average genetic correlation between the USA, the United Kingdom and New Zealand decreased to 0.84.

11.5.4 Time Edits

Every country uses a different time frame of data in the NGE. Canada goes back to 1957, for example, while some countries may not have data prior to 1970. Simulation work by Weigel and Banos (1997) showed that a certain time frame for data inclusion should be defined in all countries to ensure that the genetic parameters are relevant for the most recent populations of bulls. A sliding time window was originally adopted, but as of January 2004, the time window has stopped sliding and is fixed to the beginning of 1986 and 1981 for all countries.

11.5.5 Reliability

Reliabilities of IGE are based on the principle of Information Source, Harris and Johnson (1998). Reliability is based on the number of daughter records in each country, incorporating information from granddaughters, combined across country and finally includes parent information. This is an approximation because an inverse of the MACE MME is not possible. MACE is solved by an iterative algorithm, currently based on an approach by Klei (1998). The notion of Effective Daughter Contribution was developed (Fikse and Banos, 2001), which considered contemporary group structure, correlation between repeated records and reliability of dams of daughters, which gives a weighting factor in MACE.

11.5.6 Practicality

The final check was whether MACE results were reliable and meaningful within each country. Rex Powell of USDA made several studies showing the weaknesses of MACE in the early years, and the strengths of MACE in the later years. Fabiola Canevesi of Italy and others also made reports to Interbull meetings that led to changes in procedures which helped improve MACE rankings of bulls. Bert Klei of Holstein USA revised the software package so that MACE would run more efficiently, as did Peter Sullivan of Canadian Dairy Network. Interbull and MACE have truly been based on international collaboration and has been on-going for over 30 years.

11.6 Current Status

IGE now encompasses 30 countries, 6 breeds, 114,400 bulls and 38 traits with the number of bulls increasing each year. Some countries have been grouped into regional areas, such as Denmark-Finland-Sweden, or Germany-Austria, or the Netherlands-Luxemburg-Belgium. The genetic correlations among countries in a group is unity.

Multiple trait MACE (Schaeffer, 2001) is now being explored to accommodate more than one trait per country at a time. Zero residual covariances are assumed across countries which makes the MME easier to create.

Genomics and the availability of SNP arrays are now offering interesting possibilities for better IGE. International cooperation is imperative because each country may not be able to afford to genotype their bulls. Sharing genotypes will improve IGE calculations. Methodologies need to be developed.

11.7 Interbull Workshops

The dairy industries of different countries have collaborated very well over the years and the amount of disagreement between countries has been greatly lessened, although each is pursuing different selection strategies within their national programs in order to gain an edge in the export markets. No other species has had this amount of international cooperation.

Part of the reason for this has been the organization of annual (or more frequent) workshops sponsored by Interbull. In 1994, for example, there were about 30 participants in the workshop at Arhus, Denmark when MACE was first proposed to the world. Gradually over time the number of participants in each workshop has grown, to where it is nearly close to 200. There are more presentations wanting to be made, than can be accommodated at times. Many topics are discussed at each meeting where research on IGE is the main theme. There are some presentations on changes to NGE from some countries from time to time. Or countries may find a problem, they think, in the MACE results and try to illustrate it. The workshops have been extremely successful and useful to Interbull.

11.8 References

BANOS, G. 2010. Past, present and future of international genetic evaluations of dairy bulls. Proceedings of 9^{th} WCGALP, Leipzig, Germany.

BOICHARD, D. , B. BONAITI, A. BARBAT. 1995. Journal of Dairy Science, 78:431-437.

FIKSE, F. , G. BANOS. 2001. Journal of Dairy Science, 84:1759-1767.

GODDARD, M. E. 1985. A method of comparing sires evaluated in different countries. Livest. Prod. Sci. 13:321-331.

HARRIS, B. , D. JOHNSON. 1998. Interbull Bulletin 17:31-36.

JAIRATH, L. , J. C. M. DEKKERS, L. R. SCHAEFFER. 1998. Journal of Dairy Science, 81:550-562.

JASIOROWSKI, H. A. , M. STOLZMAN, Z. REKLEWSKI. 1988. The International Friesian Strain Comparison Trial: A World Perspective. FAO, Rome.

JORJANI, H. , L. KLEI, U. EMMANUELSON. 2003. Journal of Dairy Science, 86:677-679.

KLEI, L. 1998. Interbull Bulletin 17:3-7.

KLEI, L. , K. A. WEIGEL. 1998. Interbull Bulletin 17:8-14.

ROZZI, P. , L. R. SCHAEFFER, E. B. BURNSIDE, W. SCHLOTE. 1990. International evaluation of Holstein-Friesian dairy sires from three countries. Livest. Prod. Sci. 24:15.

SCHAEFFER, L. R. 1985. Model for international evaluation of dairy sires. Livest. Prod. Sci. 12:105-115.

SCHAEFFER, L. R. 1994. Multiple-country comparison of dairy sires. Journal of Dairy Science, 77:2671-2678.

SCHAEFFER, L. R. 2001. Multiple trait international bull comparisons. Livest. Prod. Sci. 69:145-153.

SIGURDSSON, A. , G. BANOS. 1995. Acta. Agric. Scand. 45:207-219.

SIGURDSSON, A. , G. BANOS, J. PHILIPSSON. 1996. Acta. Agric. Scand. 46:129-136.

WEIGEL, K. A. , G. BANOS. 1997. Journal of Dairy Science, 80:3425-3430.

WILMINK, J. B. M. , A. MEIJERING, B. ENGEL. 1986. Conversion of breeding values for foreign populations. Livest. Prod. Sci. 14:223-229.

Chapter 12

Multiple Traits

LARRY SCHAEFFER

12.1 Multiple Lactation Records

The Repeated Records Animal Model was suitable to analyze multiple lactation records per cow if the genetic correlation between lactation records was unity. However, if the genetic correlation was less than unity, then each lactation record would be essentially a different trait. First lactation cows were known to have lower production levels than in later lactations. By assuming each lactation is a different trait, then there can be a different model for each trait. There could be different factors affecting each lactation, but you could also have specific effects on each lactation rather than an amalgamation of one effect on all lactations. Finally, if permanent environmental effects change or accumulate from one lactation to the next, then a multiple trait model would allow different Permanent Environmental (PE) effects combined with the residual variance for each trait.

A multiple trait model would yield separate EBVs per animal for each trait, whether the cow was observed for each trait or not. These would be estimated through the genetic correlations and through the additive genetic relationships among animals. In 1990, however, when the animal model was introduced, the computing power of the time did not allow including more than one trait at a time. There had been multiple trait sire models for beef cattle by 1990, but the number of equations to solve were much less than the number of cows in a single trait animal model. Henderson (1976) described multiple trait models and applications. Pollak and Quass (1976) and Pollak et al. (1984) demonstrated the ability of multi-trait models to account for the selection on growth records in the beef cattle. Finally, Tier and Meyer (2004) provided an approximation method

to obtain accuracies of multi-trait EBV.

12.1.1 Canonical Transformation

In the 1990's multiple trait models could be made practical if the data could be transformed. The canonical transformation could be applied if every animal was observed for every trait (i.e. no missing data), if the model was the same for each trait (i.e. same factors and levels of factors), and if there was only residual and additive genetic effects as the only random variables in the model (i.e. all other factors had to be fixed). Then there was a linear transformation of the data that would yield diagonal covariance matrices for the genetic and residual effects for the transformed variables. With diagonal covariances, then each transformed trait could be analyzed as a single trait. Once the EBVs were calculated for the transformed data, they could be reverse transformed back to their original scales. The advantage of the canonical transformation was to change one m trait problem into m single trait problems. However, the assumptions to apply a canonical transformation were too restrictive because most models would have more than two random factors, and observations would be missing on some of the traits, and models could be different for some traits. The application of transformations were limited.

Assume the following covariance matrices for residual and sire genetic effects.

$$\mathbf{R} = \begin{pmatrix} 120 & 30 & -10 \\ 30 & 90 & -10 \\ -10 & -10 & 60 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} 40 & 15 & 3 \\ 15 & 20 & 4 \\ 3 & 4 & 10 \end{pmatrix},$$

To obtain the canonical transformation matrix, follow these steps:

1. Determine the eigenvalues and eigenvectors or $\mathbf{R}$.

$$\begin{aligned} \mathbf{R} &= \mathbf{UDU}' \\ \mathbf{U} &= \begin{pmatrix} .8343918 & .5506134 & .02480492 \\ .5249279 & -.8075789 & .26882522 \\ -.1680507 & .2112848 & .96286952 \end{pmatrix} \\ \mathbf{D} &= diag(\ 140.88748 \quad 72.16205 \quad 56.95047\ ) \end{aligned}$$

2. Determine the transformation to make the transformed residual matrix equal an identity matrix.

$$\begin{aligned}
\mathbf{T} &= \mathbf{U}'\mathbf{D}^{-.5} \\
&= \begin{pmatrix} 9.903900 & 4.677364 & .1871917 \\ 6.230686 & -6.860241 & 2.0287038 \\ -1.994695 & 1.794827 & 7.2663463 \end{pmatrix} \\
\mathbf{T}^{-1} &= \begin{pmatrix} .070296518 & .044422455 & -.01415807 \\ .064816505 & -.09506716 & .02487217 \\ .003286921 & .03562225 & .12759063 \end{pmatrix} \\
\mathbf{R}^* &= \mathbf{T}^{-1}\mathbf{R}\mathbf{T}'^{-1} \\
&= \mathbf{I}
\end{aligned}$$

3. Apply the transformation to $\mathbf{G}$ and compute the eigenvectors of the transformed matrix.

$$\begin{aligned}
\mathbf{H} &= \mathbf{T}^{-1}\mathbf{G}\mathbf{T}'^{-1} \\
&= \begin{pmatrix} .32106906 & .04968076 & .10974890 \\ .04968076 & .16089013 & -.01744568 \\ .10974890 & -.01744568 & .23099418 \end{pmatrix} \\
&= \mathbf{P}\mathbf{M}\mathbf{P}' \\
\mathbf{P} &= \begin{pmatrix} .8358324 & -.3019151 & .4585101 \\ .1354278 & -.6959903 & -.7051644 \\ .5320183 & .6514943 & -.5408435 \end{pmatrix}
\end{aligned}$$

4. The final transformation matrix is calculated as

$$\begin{aligned}
\mathbf{Q} &= \mathbf{P}^{-1}\mathbf{T}^{-1} \\
&= \begin{pmatrix} .06928290 & .04304126 & .05941517 \\ -.06419453 & .07602146 & .07008831 \\ -.01525304 & .06804932 & -.09303716 \end{pmatrix} \\
\mathbf{Q}\mathbf{R}\mathbf{Q}' &= \mathbf{I} \\
\mathbf{Q}\mathbf{G}\mathbf{Q}' &= \begin{pmatrix} .3989753 & 0 & 0 \\ 0 & .1987716 & 0 \\ 0 & 0 & .1152064 \end{pmatrix}
\end{aligned}$$

The data are transformed by converting each set of m observations on animal i

$$\mathbf{z}_i = \mathbf{Q}\mathbf{y}_i.$$

A canonical transformation for 3 or more covariance matrices is not possible such that the resulting transformed covariance matrices are all diagonal. However, Misztal et al. (1995) found an approximate method of diagonalization for more random factors.

12.1.2 Cholesky Transformation

Another transformation is the Cholesky decomposition of $\mathbf{R}$, so that the transformed residual covariance matrix is diagonal. This simplifies the creation of MME, but all of the random factor covariance matrices are not diagonal after transformation. The Cholesky decomposition also works if some traits are missing (i.e. not observed), provided that if trait k is missing then all traits from 1 to $k-1$ must be present, and all traits k to m must be missing. This is possible if the traits are observed over time, like lactations in dairy cows. Using the previous $\mathbf{R}$ matrix, then the Cholesky decomposition is a lower triangular matrix.

$$\begin{aligned}
\mathbf{R} &= \mathbf{L}\mathbf{L}' \\
\mathbf{L} &= \begin{pmatrix} 10.95445 & 0 & 0 \\ 2.738613 & 9.082951 & 0 \\ -.9128709 & -.8257228 & 7.6475387 \end{pmatrix} \\
\mathbf{L}^{-1} &= \begin{pmatrix} .0912871 & 0 & 0 \\ -.02752409 & .11009638 & 0 \\ .00792491 & .01188737 & .13076102 \end{pmatrix} \\
\mathbf{L}^{-1}\mathbf{R}\mathbf{L}'^{-1} &= \mathbf{I}
\end{aligned}$$

The transformed data are

$$\mathbf{z} = \mathbf{L}^{-1}\mathbf{y},$$

and the transformed genetic covariance matrix is

$$\mathbf{G}^* = \mathbf{L}^{-1}\mathbf{G}\mathbf{L}'^{-1}.$$

Multiple trait equations are easier to set up if the residual covariances are zero.

12.1.3 Scale Transformation

In 2012, no one worries about using transformations because there are more than enough of computer memory, storage, and speed, to handle any multiple trait models with very large numbers of traits. However, if the traits are greatly different in scale, such as milk yield measured in thousands of kilograms and a second trait measured in tenths of a cm, there can be rounding error problems and difficulty in estimating covariances between the traits. Thus, it is sometimes useful to divide the traits by their overall means, convert trait t_i as

$$z_i = (t_i - \bar{t})/\bar{t}.$$

The transformed variables have a mean of zero and the range of their observations is similar for each trait. After estimating variances and covariances for the z_i then multiply by the trait means to return to the original scales. Correlation estimates do not change between scales.

12.2 Numerical Example

Consider a herd of cows with one to three lactation records per cow for a trait. The model for this simple example is

$$y_{tij} = \mu_t + HYS_{ti} + a_{tj} + e_{tij}$$

where

t is the trait number, 1, 2, or 3,

y_{tij} is a lactation yield for trait t,

HYS_{ti} is a random herd-year-season effect,

a_{tj} is a random animal additive genetic effect for trait t, and

e_{tij} is a random residual effect.

The assumed additive genetic covariance matrix is

$$\mathbf{G} = \begin{pmatrix} 36 & 30 & 30 \\ 30 & 40 & 34 \\ 30 & 34 & 45 \end{pmatrix},$$

and

$$\mathbf{R} = \begin{pmatrix} 85 & 0 & 0 \\ 0 & 88 & 0 \\ 0 & 0 & 90 \end{pmatrix},$$

and

$$Var(\mathbf{h}) = \mathbf{I}(14).$$

The residual covariance matrix is diagonal because there are no common environmental effects on records made in different years. Similarly, herd-year-season effects are also uncorrelated because they occur at different times. The genetic covariance matrix, however, shows that the variability of second and third records are greater than for first records, and because they are made on the same animal there is a genetic correlation between records of about .7 to .8. The data are shown in the Table 12.1.

Table 12.1: Multiple Lactation Model

Cow	Sire	Dam	HYS	t_1	HYS	t_2	HYS	t_3
10	1	5	1	53	3	64	5	68
11	2	6	1	62	3	73	5	70
12	3	7	1	74	3	81		
13	4	8	1	46				
14	1	9	2	58	3	66	5	69
15	2	5	2	65	4	76		
16	3	6	2	37				
17	3	8	2	49	4	61	5	65
18	4	7	2	51	4	62	5	59

The MME are going to be of order (3 + 5 + 18*3)=62. There will be 3 equations for each animal, 3 equations for overall means, and 5 equations for HYS effects. The solutions to the MME were

$$\mu = \begin{pmatrix} 55.163 \\ 68.000 \\ 66.660 \end{pmatrix}$$

and for HYS were

$$\begin{pmatrix} \widehat{h}_1 \\ \widehat{h}_2 \\ \widehat{h}_3 \\ \widehat{h}_4 \\ \widehat{h}_5 \end{pmatrix} = \begin{pmatrix} 1.1614 \\ -1.1614 \\ 0.5046 \\ -0.5046 \\ 0.0000 \end{pmatrix},$$

and for the animals are shown in the Table 12.2.

Table 12.2: Multiple trait animal solutions for example data

	Multi-trait MME			Single trait MME		
Animal	Lact 1	Lact 2	Lact 3	Lact 1	Lact 2	Lact 3
1	-0.39	-0.69	-0.27	-.03	-1.39	0.73
2	3.37	3.39	3.24	2.54	1.74	0.65
3	-0.06	0.26	0.18	-0.49	0.74	-0.19
4	-2.92	-2.95	-3.14	-2.01	-1.09	-1.19
5	1.17	1.10	1.18	0.94	0.37	0.27
6	-1.57	-1.08	-1.09	-1.85	0.45	0.65
7	2.34	2.07	1.63	2.36	0.89	-1.19
8	-2.49	-2.40	-2.23	-2.11	-1.25	-0.19
9	0.55	0.32	0.51	0.66	-0.47	0.46
10	-0.55	-0.81	-0.33	-0.23	-1.43	0.78
11	2.16	2.43	2.34	1.25	1.55	1.29
12	5.08	5.07	4.56	3.83	2.80	-0.69
13	-4.04	-3.79	-3.80	-3.53	-1.17	-0.69
14	0.63	0.13	0.63	0.97	-1.40	1.05
15	4.38	4.36	4.18	3.37	2.34	0.46
16	-3.64	-2.77	-2.81	-3.92	0.60	0.23
17	-2.44	-2.37	-2.15	-1.93	-1.50	-0.38
18	-1.89	-2.29	-2.79	-0.37	-1.19	-2.38

The last three columns of Table 12.2 are corresponding solutions if each lactation was analyzed separately as a single trait. The solutions for lactation 1 are similar between multiple trait and single trait analyses, but single trait solutions for lactations 2 and 3 are much more different and much lower in variability between high and low animals. Some of the solutions are opposite in sign compared to their multiple trait values. The reason for the bigger differences is due to culling biases that are not considered during single trait analyses. Cows that make lactation 2 and 3 records have been selected based on their previous lactations. When lactations 2 and 3 are analyzed with lactation 1 in the multiple trait model, then the differences in production between cows that have later records versus those that do not is present in the data. Multiple trait analyses, to some extent, can account for culling biases, provided that the genetic covariances are well estimated. Such covariances should be estimated using cows that have been allowed to have all three lactations without any being culled (even if they should have been culled). Such data are not usually available.

The multiple trait EBVs combine information from all lactations and due to the assumed genetic correlations of .7 to .8, the solutions are very similar for

each lactation, i.e. animals rank the same. The accuracies of multiple trait EBVs are higher than for single trait EBVs due to the combined use of data from all lactations contributing to each individual lactation EBV.

12.3 Economic Traits

This book has concentrated on genetic evaluation for milk production, but milk production can included milk yield, fat and protein percentages, fat and protein yields, and somatic cell scores. New traits are frequently being studied, and if possible, are included into milk recording programs, or breed association recording programs.

Breed associations have been very concerned with body conformation or type classification of cows nearly as long as milk production. In Canada, the Holstein breed association, for many years, had about 30 main conformation traits that were subjectively scored by trained classifiers, and an additional 60 defective characteristics. The defective characteristics were similar to genetic mutations that were contrary to the breed standards, and which rarely occurred. Genetic evaluations for conformation traits have been conducted along with production traits, assuming the traits are normally distributed, but as single traits. However, conformation traits are collected on every first lactation heifer, and each heifer is scored for all of the traits, thus, conformation traits could have been evaluated using multiple trait models applying the canonical transformation. Conformation traits, because they were subjective, categorical traits, should have been analyzed by threshold models which will be discussed in a later chapter.

A list of traits evaluated in dairy cattle are given in the Table 12.3.

There are Multiple Trait (MT), systems for calving and reproduction from the cow perspective, MT systems for disease traits, and MT systems for production traits. An MT system for conformation traits would be beneficial, but will likely not be implemented because there are too many conformation traits.

12.3.1 Numerical Example

Suppose we have milk, fat, and protein 305-d first lactation yields on cows. Protein was not always observed for each cow, but milk and fat were always available on each cow. The model for each trait was assumed to be the same. Correlations between HYS and residual effects were non-zero because the records for each trait were made during the same time period. The data are shown in the Table 12.4.

The model for this example is

$$y_{tij} = \mu_t + HYS_{ti} + a_{tj} + e_{tij}$$

Table 12.3: Economically Important Traits

Area	Specific
Calving	Calf vitality/mortality
	Calf size
	Calving ease
	Stillbirths
Reproduction	Non-return rate
	Age at first breeding
	Days open
	Gestation length
	Number of services to conception
Workability	Milking speed
	Temperament
	Leakage
	Udder
	Likability
Locomotion	Rear legs-side view
	Rear legs-rear view
	Feet-legs score
Conformation	Stature
	Body depth
	Chest width
	Rump angle
Disease	Mastitis treatments
	Other treatments
	Ketosis
Longevity	Survival
	Herd life

Table 12.4: Data for MT example on cows

Cow	Sire	Dam	HYS	Milk,kg	Fat,kg	Protein,kg
10	1	5	1	5386	364	316
11	2	6	1	6213	373	270
12	3	7	1	7428	405	344
13	4	8	1	4639	321	
14	1	9	2	5873	366	269
15	2	5	2	6507	346	297
16	3	6	2	4988	333	276
17	3	8	2	5149	351	
18	4	7	2	6651	384	317

where

t is the trait number, 1, 2, or 3,

y_{tij} is a lactation yield for trait t,

HYS_{ti} is a random herd-year-season of first lactation effect,

a_{tj} is a random animal additive genetic effect for trait t, and

e_{tij} is a random residual effect.

Assume the following covariance matrices. The genetic covariance matrix is

$$\mathbf{G} = \begin{pmatrix} 330400 & 8000 & 4800 \\ 8000 & 268 & 160 \\ 4800 & 160 & 260 \end{pmatrix},$$

and the HYS covariance matrix is

$$\mathbf{H} = \begin{pmatrix} 70800 & 1714 & 1028 \\ 1714 & 57 & 34 \\ 1028 & 34 & 56 \end{pmatrix},$$

and the residual covariance matrix is

$$\mathbf{R} = \begin{pmatrix} 424800 & 10286 & 6172 \\ 10286 & 345 & 206 \\ 6172 & 206 & 334 \end{pmatrix}.$$

For cows with missing protein yields, the appropriate residual covariance matrix

is

$$\mathbf{R}_{-p} = \begin{pmatrix} 424800 & 10286 \\ 10286 & 345 \end{pmatrix}.$$

Covariance matrices should always be tested to ensure they are positive definite. That means, the eigenvalues need to be calculated and should all be positive.

The order of the MT MME was 63, 3 equations for each of the mean, 2 HYS, and 18 animals. The solutions to the MME are given in the Table 12.5.

Table 12.5: MT solutions to MME

	Milk, kg	Fat, kg	Protein, kg
μ	5873	361	294
HYS 1	14.69	1.73	1.84
HYS 2	-14.69	-1.73	-1.84
Sire 1	-111	2	-1
Sire 2	227	0	-4
Sire 3	-4	2	6
Sire 4	-112	-4	-1
Dam 5	19	-3	6
Dam 6	-144	-3	-9
Dam 7	523	15	15
Dam 8	-416	-10	-6
Dam 9	17	2	-6
Cow 10	-173	0	7
Cow 11	121	2	-12
Cow 12	618	18	21
Cow 13	-540	-17	-9
Cow 14	-30	3	-9
Cow 15	270	-4	2
Cow 16	-297	-8	-6
Cow 17	-350	-5	-1
Cow 18	370	11	12

Notice that cows 13 and 17 obtained EBVs for protein yields even though their own protein observations were missing. The variances of prediction error should be smaller than those for single trait analyses, if the same parameters are used for the variances. MT analyses assume that the covariance matrices are accurately estimated. Errors in these matrices can result in biased EBV with larger true prediction error variances.

Genetic evaluations are shifting more towards MT analyses in order to utilize genetic correlations and information from other traits to achieve more accurate EBVs. Schaeffer (1984) showed the advantages of MT analyses over single traits and the conditions that make MT analyses favored over single traits. Traits with low heritability gain accuracy in EBV if they are analyzed with traits having higher heritabilities. Also if the difference between genetic and residual correlations is large, then MT analyses are much better than single trait. Lastly, if there is selection on a trait such that another trait is observed only on selected animals, then EBV for the second trait can be more accurate as opposed to ignoring information from the first trait. MT analyses do not completely remove the bias of selection, but reduce the effects of it on the second trait EBVs. This depends on the degree of selection, or severity of culling.

MT analyses are helpful for traits that are recorded on only a few animals rather the entire population. These observations do not occur due to selection or culling on other traits, but are usually limited due to their cost of recording. An example would be ultrasound backfat measures in sheep, where only a few herds find it beneficial to take ultrasound measurements.

12.4 References

HENDERSON, C. R. , R. L. QUAAS. 1976. Multiple trait evaluation using relatives' records. J. Anim. Sci. 43:1188-1197.

MISZTAL, I. , K. A. WEIGEL, T. J. LAWLOR. 1995. Approximation of estimates of (co)variance components with multiple-trait restricted maximum likelihood by multiple diagonalization for more than one random effect. J. Dairy Sci. 78:1862-1872.

POLLAK, E. J. , R. L. QUAAS. 1981. Monte Carlo study of within-herd multiple trait evaluation of beef cattle growth traits. J. Anim. Sci. 52:248-256.

POLLAK, E. J. , R. L. QUAAS. 1981. Monte Carlo study of genetic evaluations using sequentially selected records. J. Anim. Sci. 52:257-264.

POLLAK, E. J. , J. Van der WERF, R. L. QUAAS. 1984. Selection bias and multiple trait evaluation. J. Dairy Sci. 67:1590-1595.

SCHAEFFER, L. R. 1984. Sire and cow evaluation under multiple trait models. J. Dairy Sci. 67:1567-1580.

TIER, B. , K. MEYER. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J. Anim. Breed. Genet. 121:77-89.

Chapter 13

Test Day Models

RAPHAEL MRODE
JULIO CARVALHEIRA
LARRY SCHAEFFER

13.1 Test Day Records

The basic unit of information gathered by milk recording organizations around the world is a Test-Day (TD) record. A TD record is a measure of the amount of milk produced by a cow within 24 h on a given day during her lactation, and includes the fat and protein percentage in the milk. TD yields have been gathered since 1905 in Michigan started by Helmer Rabild of Denmark. He was hired in 1908 by the US Department of Agriculture to create the national milk recording program.

For every cow, there are 7 to 10 TD records per lactation stored in their milk recording agency. The number of tests depends on the frequency of testing and the lactation length of the cow. Each TD record is an estimate of the yield given in 24 h, even though the cow may have been milked 2 or 3 times in a day, or 5 times over 2 days. Today there are many herds with robotic milking systems, so that cows decide how often they are milked per day, although they are forced to go through at least twice in order to receive feed.

In Canada, the federal government had two recording plans for many years. One was to collect the 24 h yields every day during the lactation (expensive in time and cost), or one TD every 30 days. With the first plan, a 305-d yield was just the sum of all TD yields from day 1 to day 305. Due to the cost, this plan was dropped by 1970.

In the second plan, assumptions had to be made in order to calculate a

305-d yield. The Test Interval Method (TIM) computed the average yield between two consecutive tests and multiplied the average by the number of days interval between the two tests. Problems existed for the first and last tests, and special procedures and formula were derived for these two situations. Unfortunately, these methods and formula had to be updated and modified often. During a Dairy Technical Committee meeting in Ottawa in 1975, Dr. John Moxley of Macdonald College of McGill University in Montreal, stated that he believed genetic evaluations would be easier if scientists analyzed TD records themselves, rather than trying to find appropriate methods to adjust first and last test days. However, in 1975 computers still had limited memory and speed, and genetic evaluations had just started to use Sire Models in Canada.

Australians had been using test day records since 1985 to produce indexes on cows. Robert W. Everett of Cornell University had a plan for analyzing TD records within herds that he developed while in Australia, and which he patented in 1993 (See Rothschild and Newman, 2002), and implemented at Cornell University. In 1990, the animal model was implemented in Canada for genetic evaluation, personal computers were becoming more common and inexpensive, and it was obvious that computer speed and memory were going to keep improving rapidly. Now seemed to be the right time to begin thinking about the analysis of TD records rather than 305-d yields. In 1991, Schaeffer and others started thinking about lactation curves or trajectories. Early work by Ptak et al. (1993) considered analyses of TD records where the shape of the curve was assumed the same for all cows, just the height of the curve was different from cow to cow. Thinking about this, Jack Dekkers said, "if only you could have random regressions". Upon checking Henderson (1984), there was a small paragraph on "random regressions", and Henderson, Jr. (1982) had a paper on random regressions in Biometrics. Thus, the Random Regression Model (RRM), or Test Day Model was developed. Random regressions meant that each cow could have its own curve, its own shape. Obviously, cows differed in persistency of milk production, and producers fed cows according to persistency and level of production. Selecting cows to change the shape of the lactation curve was now possible.

RRM test day models also allowed changes to be made in milk recording programs. Cows did not need to be tested at regular 30-d intervals. Early work showed that at least 4 test day records gave the same level of accuracy in an EBV as one 305-d lactation record based on 8 or more TD records. A variety of testing programs were offered to producers resulting in lower costs to the producers due to less frequent testing. Intervals between TD now vary considerably, but the ability of the RRM to account for Days In Milk (DIM) during the lactation made the intervals less critical. However, producers could choose the level of accuracy they wanted in 305-d yields.

13.2 Lactation Curves

Phenotypic lactation curves are shown in the following three figures (Figure 13.1, Figure 13.2, Figure 13.3) for milk, fat, and protein yields, respectively. Over the years there have been more than 20 proposed curve functions to model the yields in these figures. Functions had from 3 to 5 parameters and were both linear and non-linear functions. Functions could also be classified as having parameters which had biological interpretations or those which were basically regression functions. The more parameters there were, the better was the fit to actual data. The trade-off was the number of TD records per animal, when estimating a curve for each cow.

Figure 13.1

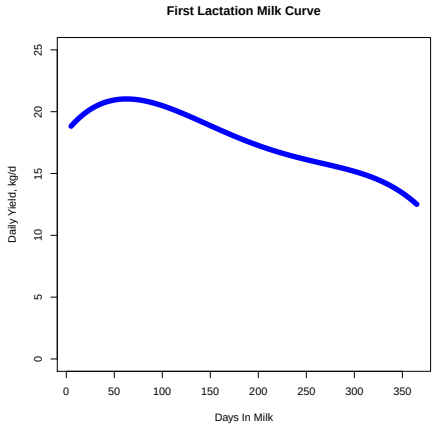


Figure 13.2

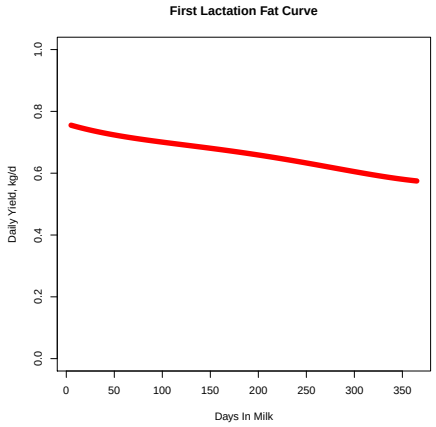
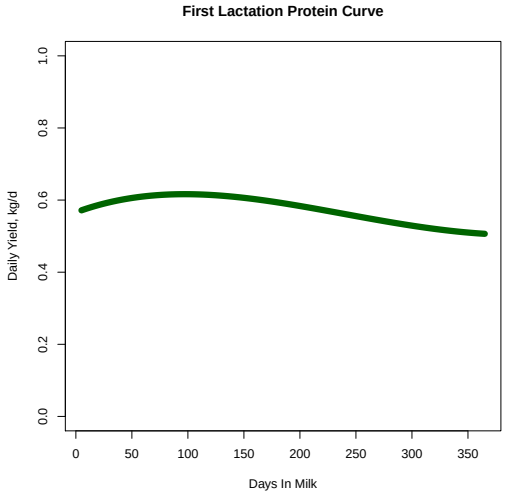


Figure 13.3



Although the milk production curve in Figure 13.1 represents average first lactation cows, there is great variability in curve shape from cow to cow. % Most first lactation cows have their peak milk yield by 45 days in milk, but there are documented cases where cows have not peaked until 100 days, and on the opposite end of the spectrum cows that peaked at 10 days and decreased yield thereafter.

Different curve functions may be appropriate for different cows, and one function does not fit all animals. However, the exceptions are assumed to be in the minority, and the majority of cows adhere to the curve function.

There have been many proposed functions to model the shape of the lactation curve, mostly in phenotypic terms. Three of the more popular ones are mentioned here.

13.2.1 Wood's Function, 1967

Wood (1967) proposed the following non-linear model for describing lactation curves in cattle.

$$y_t \;=\; a\; t^b\; \exp^{-ct},$$

where

y_t was yield in week t of lactation,

a was peak yield,

t^b was the slope of yields up to the peak, and

$\exp^{ct}$ was the decline in production after the peak,

a,b,c were parameters to be estimated for an individual or a group of cows.

Typically, the equation would be transformed by taking natural logarithms, and then solving by simple regression.

$$\ln(y_t) \;=\; \ln(a) \;+\; b\;\ln(t) \;-\; c\;t.$$

This function was widely used after it was introduced, and is still used in 2012. There were only three parameters and those parameters corresponded to biological interpretations. The biological interpretations fit the majority of cows, but lactation curves, in practice, were not smooth curves. There were dips or waves in lactation curves that did not follow the Wood function.

The parameter estimates in the Wood function tend to have high correlations among them.

13.2.2 Wilmink's Function, 1987

Another popular curve function was given by Wilmink (1987).

$$y_t \;=\; a \;+\; b\;t \;+\; c\;\exp(-.05\;t)$$

which was a linear function of days in milk (or weeks) and had three unknown parameters. The biological interpretation was not as clear as with the Wood function, but a transformation of the data was not needed to estimate the parameters.

The $-.05$ in the exponential function was derived through trial and error. Those who have attempted to estimate this value usually find estimates close to $-.05$, and the minor discrepancies from $-.05$ do not cause a significant change in the lactation curve shape. The parameter estimates in the Wilmink model also tend to have high correlations among them.

13.2.3 Ali-Schaeffer Function, 1987

Ali and Schaeffer (1987) used the function

$$y_t \;=\; \mu + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e_t,$$

where

$$
\begin{aligned}
y_t &= \text{daily yield of cow on DIM } t \\
t &= \text{days in milk} \\
X_1 &= (t/305) \\
X_2 &= (t/305)^2 \\
X_3 &= \ln(305/t) \\
X_4 &= (\ln(305/t))^2 \\
\mu, b_1, b_2, b_3, b_4 &= \text{unknown regressions} \\
e_t &= \text{residual for day } t
\end{aligned}
$$

X_1 and X_2 were chosen because they increased as t became larger, and X_3 and X_4 decreased as t increased. A problem in estimation is that these variables have high correlations (positive or negative) among themselves, which makes it difficult to obtain stable estimates for a particular animal. Let θ_i represent a vector of the five parameters in the above equation for cow i, then the next step was to estimate genetic parameters of the p^{th} element of θ_i using a model like

$$\theta_{pijkm} = (AM)_{pj} + (YM)_{pk} + HYS_{pm} + a_{pi} + e_{pijkm},$$

where AM are fixed age-month of calving group effects, YM are fixed year-month of calving effects, HYS are random, herd-year-season effects, a are random animal additive genetic effects, and e are random residual effects. The observations in this model have a degree of error of estimation associated with them that could differ between animals (but which was usually ignored). The elements could be analyzed singly or as a multiple trait problem. Thus, there would be a heritability estimate for μ, b_1 through b_4 and possibly genetic and residual correlations.

In estimating the curve parameters for a cow, the information on all other animals was generally ignored. Two cows which belonged to the same age-month of calving group were not considered in estimating the parameters for each cow, but such information might have led to parameter estimates with smaller standard errors. A more accurate approach would be to analyze TD records of all cows simultaneously, accounting for the fact that there are many cows in the same AM group, in the same YM group, and in the same HYS. A model might be

$$
\begin{array}{rcll}
y_{tijkm} & = & (\mu_j + b_{1j}X_1 + b_{2j}X_2 + b_{3j}X_3 + b_{4j}X_4) & (AM)_j \text{ effects} \\
 & & +(\mu_k + b_{1k}X_1 + b_{2k}X_2 + b_{3k}X_3 + b_{4k}X_4) & (YM)_k \text{ effects} \\
 & & +HTD_m & \text{herd-test date-parity} \\
 & & +(\mu_i + b_{1i}X_1 + b_{2i}X_2 + b_{3i}X_3 + b_{4i}X_4) & a_i \text{ effects} \\
 & & +(\mu_i + b_{1i}X_1 + b_{2i}X_2 + b_{3i}X_3 + b_{4i}X_4) & p_i, \text{ PE effects} \\
 & & +e_{tijkm} & \text{residual effects}
\end{array}
$$

PE effects are needed because we have more than one TD record per animal in a lactation. The same curve function is used in all factors, in this model, but this does not need to be the case. For the AM effects, for example, rather than fitting a curve function, the lactation could be split up into 7-day periods from day 5 to day 365. Yields in the first five days are not recorded as they are used to feed calves and not measured. The means for each 7 day period would be estimated within age and months of calving. These means may not necessarily be very smooth, but should be a better fit to the data than the curve function. In some studies the number of covariates for the genetic and PE effects were allowed to differ.

Notice the HTD effect, which is a particular day in which the supervisor visits the herd to weigh the milk and take samples. Cows in first lactation on that test day are all influenced by the environment at that point in time, but the cows can be in very different stages of lactation on that day, and likely are in countries where cows calve all year round. In countries with seasonal calvings, cows in a HTD will be closer together in stage of lactation.

13.3 Example Data

The following data (Table 13.1), on 5 first lactation cows will be used throughout this chapter to illustrate the methods.

The assumed pedigrees were

Cow	Sire	Dam
8	1	3
9	2	4
10	1	5
11	2	6
12	1	7

A preliminary study was made to determine the milk yields of cows in first lactation every 30 days from day 5 to 305. The resulting phenotypic covariance matrix is given in the table Table 13.2. The corresponding table of phenotypic correlations are in Table 13.3.

Table 13.1: TD milk yields, MY, kg on five cows

HTD	Cow 8		Cow 9		Cow 10		Cow 11		Cow 12	
	DIM	MY	DIM	MY	DIM	MY	DIM	MY	DIM	MY
1	4	17.0	6	23.0	106	23.0				
2	38	18.6	40	21.0	140	16.8				
3	73	24.0	74	18.0	174	11.0				
4	106	20.0	108	17.0	208	13.0	7	22.8		
5	140	20.0	142	16.2	242	17.0	41	22.4		
6	174	15.6	176	14.0	276	13.0	75	21.4	11	10.4
7	201	16.0	203	14.2	303	12.6	102	18.8	38	12.3
8	242	13.0	244	13.4			143	18.3	79	13.2
9	276	8.2	278	11.8			177	16.2	113	11.6
10	303	8.0	305	11.4			204	15.0	140	8.4

Table 13.2: (Co)variances for milk yields on specific days in first lactation

	Days in Milk										
	5	35	65	95	125	155	185	215	245	275	305
5	22	15	10	8	7	6	6	5	5	5	4
35	15	17	12	12	11	10	9	9	8	8	7
65	10	12	16	13	13	13	12	11	10	10	9
95	8	12	13	16	14	13	13	12	12	11	10
125	7	11	13	14	16	14	13	13	13	12	11
155	6	10	13	13	14	16	13	13	13	12	11
185	6	9	12	13	13	13	16	14	14	13	12
215	5	9	11	12	13	13	14	16	14	14	12
245	5	8	10	12	13	13	14	14	16	14	13
275	5	8	10	11	12	12	13	14	14	17	13
305	4	7	9	10	11	11	12	12	13	13	18

Let the matrix represented in Table 13.2 be called **V**, a phenotypic covariance matrix.

13.4 Covariance Functions

Kirkpatrik et all (1990; 1994), proposed the use of covariance functions for longitudinal data of this kind. A covariance function (CF) is a way to model

Table 13.3: Correlations for milk yields on specific days in first lactation

	Days in Milk										
	5	35	65	95	125	155	185	215	245	275	305
5	1.00	0.78	0.53	0.43	0.37	0.32	0.32	0.27	0.27	0.26	0.20
35	0.78	1.00	0.73	0.73	0.67	0.61	0.55	0.55	0.49	0.47	0.40
65	0.53	0.73	1.00	0.81	0.81	0.81	0.75	0.69	0.62	0.61	0.53
95	0.43	0.73	0.81	1.00	0.88	0.81	0.81	0.75	0.75	0.67	0.59
125	0.37	0.67	0.81	0.88	1.00	0.88	0.81	0.81	0.81	0.73	0.65
155	0.32	0.61	0.81	0.81	0.88	1.00	0.81	0.81	0.81	0.73	0.65
185	0.32	0.55	0.75	0.81	0.81	0.81	1.00	0.88	0.88	0.79	0.71
215	0.27	0.55	0.69	0.75	0.81	0.81	0.88	1.00	0.88	0.85	0.71
245	0.27	0.49	0.62	0.75	0.81	0.81	0.88	0.88	1.00	0.85	0.77
275	0.26	0.47	0.61	0.67	0.73	0.73	0.79	0.85	0.85	1.00	0.74
305	0.20	0.40	0.53	0.59	0.65	0.65	0.71	0.71	0.77	0.74	1.00

the variances and covariances of a longitudinal trait. Orthogonal polynomials are used in this model and the user must decide the order of fit that is best. Legendre polynomials, founded in 1797, are the easiest to apply.

13.4.1 Legendre Polynomials

The Legendre polynomials are defined by a recursive formula. The first two are pre-defined to be

$$\begin{aligned} P_0(x) &= 1, \text{ and} \\ P_1(x) &= x, \end{aligned}$$

then the $n+1$ polynomial is described by the following recursive equation:

$$P_{n+1}(x) = \frac{1}{n+1}\left((2n+1)xP_n(x) - nP_{n-1}(x)\right).$$

These quantities are "normalized" using

$$\phi_n(x) = \left(\frac{2n+1}{2}\right)^{.5} P_n(x).$$

This gives the following series of polynomials,

$$\begin{aligned}
\phi_0(x) &= \left(\frac{1}{2}\right)^{.5} P_0(x) = .7071 \\
\phi_1(x) &= \left(\frac{3}{2}\right)^{.5} P_1(x) \\
&= 1.2247x \\
P_2(x) &= \frac{1}{2}(3xP_1(x) - 1P_0(x)) \\
\phi_2(x) &= \left(\frac{5}{2}\right)^{.5} (\frac{3}{2}x^2 - \frac{1}{2}) \\
&= -.7906 + 2.3717x^2,
\end{aligned}$$

and so on. The first six can be put into a matrix, Λ, as

$$\Lambda' = \begin{pmatrix}
.7071 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.2247 & 0 & 0 & 0 & 0 \\
-.7906 & 0 & 2.3717 & 0 & 0 & 0 \\
0 & -2.8062 & 0 & 4.6771 & 0 & 0 \\
.7955 & 0 & -7.9550 & 0 & 9.2808 & 0 \\
0 & 4.3973 & 0 & -20.5206 & 0 & 18.4685
\end{pmatrix}.$$

Legendre polynomials are defined within the range of values from -1 to +1. Thus, days in milk have to be standardized (converted) to the interval between -1 to +1. The formula is

$$q_\ell = -1 + 2\left(\frac{t_\ell - t_{min}}{t_{max} - t_{min}}\right).$$

The minimum time for the matrix, $\mathbf{V}$, in Table 13.2 is 5 days, and the maximum time is 305 days. For the days in milk given in Table 13.2, let

$$\mathbf{x}' = \begin{pmatrix} -1 & -.8 & -.6 & -.4 & -.2 & 0 & .2 & .4 & .6 & .8 & 1 \end{pmatrix}.$$

Now define $\mathbf{M}$ as a matrix containing the polynomials of standardized time values, raised to different powers.

$$\mathbf{M} = \begin{pmatrix} \mathbf{x}^0 & \mathbf{x}^1 & \mathbf{x}^2 & \mathbf{x}^3 & \mathbf{x}^4 & \mathbf{x}^5 & \mathbf{x}^6 & \mathbf{x}^7 & \mathbf{x}^8 & \mathbf{x}^9 & \mathbf{x}^{10} \end{pmatrix}.$$

Let

$$\begin{aligned} \Phi &= \mathbf{M}\Lambda, \text{ then} \\ \mathbf{V} &= \Phi\mathbf{H}\Phi' \\ &= \mathbf{M}(\Lambda\mathbf{H}\Lambda')\mathbf{M}' \\ &= \mathbf{M}\mathbf{T}\mathbf{M}'. \end{aligned}$$

Note that Φ, $\mathbf{M}$, and Λ are matrices defined by the Legendre polynomial functions and by the standardized time values and do not depend on the data or values in the matrix $\mathbf{V}$. Therefore, it is possible to estimate either $\mathbf{H}$ or $\mathbf{T}$,

$$\mathbf{H} = \Phi^{-1}\mathbf{V}\Phi^{-T},$$

and

$$\mathbf{T} = \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-T}$$

$\mathbf{H}$ can be used to calculate the covariance between any two days in milk between 5 and 305 days. To compute the covariance between days t_1 and t_2, calculate the Legendre polynomial covariates as in calculating a row of Φ using the standardized time values for days t_1 and t_2. Then the Legendre polynomials are stored in $\mathbf{L}$, and the variances and covariance for those two DIM are

$$\mathbf{LHL}'$$

The matrix $\mathbf{H}$ is order 11 by 11 or less, but it can be used to calculate variances and covariances between any two DIM from 5 to 305 days.

Legendre polynomials were chosen because they are orthogonal, which means that the covariance matrix of polynomials over the entire range from -1 to +1 are very close to zero. Looking at the correlations in Table 13.3 shows that, phenotypically, the elements in $\mathbf{V}$ are highly correlated to each other, but the elements in $\mathbf{H}$ will have much lower correlations. The correlation issue is related to problems in estimating covariance matrices and breeding values.

13.4.2 Order of Fit

The matrix $\mathbf{V}$ is order 11, and therefore, a full order of fit is one with 11 polynomials which go from 0 to 10 in powers of the vector of standardized time variables. A full order fit explains all of the variation in the elements of $\mathbf{V}$,

without error. Thus, if you use $\mathbf{H}$ to predict $\mathbf{V}$ you predict $\mathbf{V}$ perfectly.

Reduced orders of fit try to find an $\mathbf{H}$ of smaller dimension which can predict $\mathbf{V}$ with similar accuracy, but there will be errors. The problem is to find the smallest order for $\mathbf{H}$ which predicts $\mathbf{V}$ with a low error. That is, find Φ^* such that it is rectangular and $\mathbf{H}^*$ has a smaller order, $m < 11$, but still

$$\mathbf{V}^* = \Phi^* \mathbf{H}^* \Phi'^*.$$

To determine $\mathbf{H}^*$, first pre-multiply $\mathbf{V}$ by Φ'^* and post-multiply by Φ^* as

$$\begin{aligned} \Phi'^* \mathbf{V} \Phi^* &= \Phi'^* (\Phi^* \mathbf{H}^* \Phi'^*) \Phi^* \\ &= (\Phi'^* \Phi^*) \mathbf{H}^* (\Phi'^* \Phi^*). \end{aligned}$$

Now pre- and post- multiply by the inverse of $(\Phi'^* \Phi^*) = \mathbf{P}$ to determine $\mathbf{H}^*$,

$$\mathbf{H}^* = \mathbf{P}^{-1} \Phi^{*'} \mathbf{V} \Phi^* \mathbf{P}^{-1}.$$

Φ^* is the first m columns of Φ, and $(\Phi'^* \Phi^*)$ has order m, is symmetric, and has an inverse.

Having obtained $\mathbf{H}^*$, then calculate

$$\mathbf{V}^* = \Phi^* \mathbf{H}^* \Phi'^*,$$

and let

$$\mathbf{E} = \mathbf{V}^* - \mathbf{V},$$

then let $\mathbf{e}$ be a vector of the upper triangular portion of $\mathbf{E}$, so that

$$SS_m = \mathbf{e}'\mathbf{e}$$

with degrees of freedom equal to $m(m+1)/2$. In Table 13.4 there are sums of squares of errors for m decreasing from 11 to 2, for matrix $\mathbf{V}$ (Table 13.2).

The SS_m become larger as m decreases, meaning that the fit of the model is becoming poorer. An $F-$test can be constructed by taking the difference in SS_m from SS_{10} divided by the difference in degrees of freedom as the numerator, and that divided by an estimate of the residual variance. The residual variance is

$$\sigma_E^2 = SS_{10}/(66-55) = 3.4956/11 = 0.3178.$$

To test the significance of order of fit 5, then

$$F_{40,11} = [(37.5334 - 3.495584)/(55-15)]/0.3178 = 2.68$$

Table 13.4: Sums of Squares of Errors for Predicting Elements of Covariance Matrix, **V**

Order of fit	degrees of freedom	SS_m
11	66	0.0
10	55	3.495584
9	45	8.089266
8	36	15.83321
7	28	23.96126
6	21	30.34071
5	15	37.53340
4	10	55.06217
3	6	111.1478
2	3	377.9466

compared to the table value of 2.53 at the 0.95 confidence level means it is just barely significantly different from an order 10 fit.

The matrix **H** from an order 5 fitting was

$$\mathbf{H}_5 = \begin{pmatrix} 24.0199 & 0.8687 & -1.6743 & 0.5643 & -0.4771 \\ 0.8687 & 3.4882 & -0.7436 & 0.0340 & -0.0899 \\ -1.6743 & -0.7436 & 1.4325 & -0.3654 & 0.0552 \\ 0.5643 & 0.0340 & -0.3654 & 0.6815 & 0.0110 \\ -0.4771 & -0.0899 & 0.0552 & 0.0110 & 0.2726 \end{pmatrix}.$$

To get the phenotypic variances and covariances between days 43 and 81, for example, the standardized values would be

$$\begin{aligned} q_{43} &= -.7466667 \\ q_{81} &= -.4933333 \end{aligned}$$

then the Legendre Polynomials for those two days are

$$\mathbf{L}' = \begin{pmatrix} 0.7071068 & -0.9144762 & 0.5316843 & 0.1483803 & -0.7548403 \\ 0.7071068 & -0.6042075 & -0.2133483 & 0.8228542 & -0.5908374 \end{pmatrix}.$$

Finally,

$$\mathbf{L}'\mathbf{H}_5\mathbf{L} = \begin{pmatrix} 14.23285 & 13.32235 \\ 13.32235 & 14.56514 \end{pmatrix}.$$

Thus, covariances can be calculated for any pair of days between 5 to 305 days. Technically, the covariances can not be calculated between or with days that are outside of the range of t_{min} or t_{max}, because their standardized time values would be less than -1 or greater than +1. Another matrix of phenotypic covariances would need to be obtained that covers 5 to 365 days in milk, or 400 DIM, depending how far cows give milk in one lactation.

The standard of 305 days for an official lactation record was established back in the 1930's, but today cows produce a lot more milk and for a longer period of time, such that the standard lactation length should probably be extended. Test day models often allow TD records up to 365 days in milk to be included in genetic evaluation, but the official standard length is still 305 days.

13.5 Fixed Regression Model

Ptak and Schaeffer (1993) analyzed test day records with a model similar to the following:

$$\begin{aligned} y_{tijk} &= (b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4) \\ &\quad + HTD_j + a_i + p_i + e_{tijk} \end{aligned}$$

where

y_{tijk} is a 24 h TD milk yield in first lactation,

HTD_j is a random herd-test date effect,

t is days in milk,

$X_1 = (t/305)$,

$X_2 = (t/305)^2$,

$X_3 = \ln(305/t)$

$X_4 = (\ln(305/t))^2$

b_0, b_1, b_2, b_3, b_4 are the fixed, overall mean regressions,

a_i are the animal additive genetic effects,

p_i are the animal permanent environmental effects, and

e_{tijk} is a random residual effect.

In Ptak and Schaeffer (1993) the model had fixed regressions on days in milk variables for eight age-season groups within first lactations. Of the four test day models they compared, two had herd-year-season of calving effects, and two had herd-test-day effects. Also, two models included covariances among the residual effects between TD records on the same cow, and two models assumed residual covariances were zero. The residual variances, however, were allowed to vary with days in milk, using a quadratic regression on days.

$$\sigma^2_{e_t} = 9.7 - 0.072\ t + 0.0002\ t^2$$

The covariance matrices were

$$\begin{aligned} Var(\mathbf{h}) &= \mathbf{I}\sigma^2_h \\ Var(\mathbf{a}) &= \mathbf{A}\sigma^2_a \\ Var(\mathbf{p}) &= \mathbf{I}\sigma^2_p \\ Var(\mathbf{e}) &= \mathbf{I}\sigma^2_e \end{aligned}$$

and $\sigma^2_a = 4$, $\sigma^2_p = 1.6$, $\sigma^2_h = 8$, and $\sigma^2_e = 5$.

13.5.1 Solutions and EBV

Data from Table 13.1 were analyzed with the fixed regression model. The order of the MME was 32, (5 fixed regressions, 10 HTD effects, 12 animal genetic, and 5 PE effects). The solutions were

$$\widehat{\mathbf{b}} = \begin{pmatrix} 33.3101 & -34.7340 & 13.0129 & -5.3289 & 0.3188 \end{pmatrix}$$

for the overall mean. To estimate the average yield on day 65, the covariates for day 65 were 1, 0.213, 0.045, 1.546, and 2.390, respectively, thus

$$\begin{aligned} \widehat{y}_{65} &= 33.3101 - 34.7340(0.213) + 13.0129(0.045) \\ &\quad -5.3289(1.546) + 0.3188(2.390) \\ &= 19.02 \text{ kg} \end{aligned}$$

The HTD solutions were

$$\begin{pmatrix} 3.136 & -0.116 & -0.299 & 0.624 & 1.489 \\ -0.590 & -0.379 & -0.258 & -1.566 & -2.042 \end{pmatrix}$$

The animal additive genetic and the animal PE solutions are given in Table 13.5. Also, the EBV for the animals are obtained by multiplying the animal genetic solutions times 301 (the number of days from 5 to 305). The EBV are also in Table 13.6.

Table 13.5: Animal genetic and animal PE solutions from fixed regression test-day model

Animal	Genetic	PE	EBV
1	-0.979	0	-295
2	0.979	0	295
3	0.491	0	148
4	0.075	0	23
5	0.398	0	120
6	0.904	0	272
7	-1.868	0	-562
8	0.247	0.393	74
9	0.602	0.060	181
10	0.108	0.319	33
11	1.845	0.723	555
12	-3.291	-1.494	-991

In this model there is only one additive genetic value per animal which represents differences in the heights of the lactation curves for each animal. The shape of the curve was assumed to be the same for all cows, given by the overall mean parameters.

13.6 Autoregressive Model

Following Carvalheira et al. (2002), the fixed regression model was modified by changing the PE effects into separate short term environmental (STE) effects within each cow which have an autocorrelation amongst themselves. The model can be written as

$$\begin{aligned} y_{tijkm} &= (b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4) \\ &\quad + HTD_j + a_i + p_{ik} + e_{tijkm} \end{aligned}$$

where

y_{tijkm} is a 24 h TD milk yield in first lactation,

HTD_j is a random herd-test date effect,

t is days in milk,

$X_1 = (t/305)$,

$X_2 = (t/305)^2$,

$X_3 = \ln(305/t)$

$X_4 = (\ln(305/t))^2$

b_0, b_1, b_2, b_3, b_4 are the fixed, overall mean regressions,

a_i are the animal additive genetic effects,

p_{ik} are the animal permanent environmental effects, and

e_{tijkm} is a random residual effect.

If $\mathbf{p}$ is the vector of short term environmental (STE) effects for all five cows, then it has length 39 (one for each record), with expected value of null, and covariance matrix that looks like

$$\mathbf{P} = \begin{pmatrix} \mathbf{F}_8 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_9 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{F}_{12} \end{pmatrix} \sigma_p^2,$$

for cows 8 to 12, and where

$$\mathbf{F}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-2} & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-3} & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-4} & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & \rho & 1 \end{pmatrix},$$

for a cow with n TD records. The inverse of an autocorrelation matrix has a resulting tri-diagonal format. That is,

$$\mathbf{F}_i^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

As in the fixed regression model, the covariance matrices were

$$\begin{aligned} Var(\mathbf{h}) &= \mathbf{I}\sigma_h^2 \\ Var(\mathbf{a}) &= \mathbf{A}\sigma_a^2 \\ Var(\mathbf{p}) &= \mathbf{P}\sigma_p^2 \\ Var(\mathbf{e}) &= \mathbf{I}\sigma_e^2 \end{aligned}$$

and $\sigma_a^2 = 4$, $\sigma_p^2 = 1.6$, $\sigma_h^2 = 8$, and $\sigma_e^2 = 5$. We will assume that $\rho = 0.80$.

13.6.1 Solutions and EBV

Using data from Table 13.1, the MME in this example has order 66, (5 covariates for the overall mean, 10 HTD effects, 12 animal genetic effects, and 39 STE effects).

$$\widehat{\mathbf{b}} = \begin{pmatrix} 33.729 & -34.741 & 12.899 & -5.788 & 0.409 \end{pmatrix},$$

for the overall mean curve, and the HTD solutions were

$$\begin{pmatrix} 3.145 & 0.020 & -0.203 & 0.629 & 1.460 \\ -0.638 & -0.418 & -0.274 & -1.606 & -2.115 \end{pmatrix}$$

The short term environmental effects for each animal are given in Table 13.7.

The EBV for the fixed regression and autoregressive models are very similar in this example. However, there is still only one animal additive genetic effect being estimated per animal, which relates to the relative heights of the lactation curves. The shape is assumed to be the same for each cow. The STE effects may be an improvement over the fixed regression model, but depends on the validity of assuming an autoregressive model. If one compares the correlations in the phenotypic covariance matrix with the autoregressive values, as in Table 13.8, then it

Table 13.6: Animal genetic and animal PE solutions from fixed regression and autoregressive test-day models

	Fixed Regression		Autoregressive	
Animal	Genetic	EBV	Genetic	EBV
1	-0.979	-295	-1.108	-334
2	0.979	295	1.108	334
3	0.491	148	0.439	132
4	0.075	23	0.086	26
5	0.398	120	0.437	132
6	0.904	272	1.022	308
7	-1.868	-562	-1.985	-597
8	0.247	74	0.105	32
9	0.602	181	0.684	206
10	0.108	33	0.102	31
11	1.845	555	2.087	628
12	-3.291	-991	-3.531	-1063

Table 13.7: Short term environmental effects on five cows

HTD	Cow 8	Cow 9	Cow 10	Cow 11	Cow 12
1	0.116	0.341	0.212		
2	0.467	0.107	-0.021		
3	1.030	-0.245	-0.322		
4	1.018	-0.478	-0.125	0.677	
5	0.956	-0.545	0.280	0.487	
6	0.763	-0.415	0.383	0.460	-1.256
7	0.553	-0.227	0.379	0.334	-1.289
8	0.130	-0.069		0.355	-1.130
9	-0.296	0.066		0.394	-0.972
10	-0.408	0.145		0.389	-0.924

appears that the autocorrelation of 0.80 is almost suitable to describe $\mathbf{V}$, but in general, the autocorrelations are much smaller than the actual correlations.

The autoregression model requires estimation of one short term environmental effect per TD record within an animal. This example assumed that σ_p^2 was the same for each monthly interval. This variance could possibly be different for each monthly interval, but the autocorrelation structure might still exist. Finally,

Table 13.8: Comparison of correlations for milk yields on specific days in first lactation, actual (above diagonals) versus autocorrelation (0.80) (below diagonals)

	Days in Milk										
	5	35	65	95	125	155	185	215	245	275	305
5	1	0.78	0.53	0.43	0.37	0.32	0.32	0.27	0.27	0.26	0.20
35	0.80	1	0.73	0.73	0.67	0.61	0.55	0.55	0.49	0.47	0.40
65	0.64	0.80	1	0.81	0.81	0.81	0.75	0.69	0.62	0.61	0.53
95	0.51	0.64	0.80	1	0.88	0.81	0.81	0.75	0.75	0.67	0.59
125	0.41	0.51	0.64	0.80	1	0.88	0.81	0.81	0.81	0.73	0.65
155	0.33	0.41	0.51	0.64	0.80	1	0.81	0.81	0.81	0.73	0.65
185	0.26	0.33	0.41	0.51	0.64	0.80	1	0.88	0.88	0.79	0.71
215	0.21	0.26	0.33	0.41	0.51	0.64	0.80	1	0.88	0.85	0.71
245	0.17	0.21	0.26	0.33	0.41	0.51	0.64	0.80	1	0.85	0.77
275	0.13	0.17	0.21	0.26	0.33	0.41	0.51	0.64	0.80	1	0.74
305	0.11	0.13	0.17	0.21	0.26	0.33	0.41	0.51	0.64	0.80	1

the autocorrelation works for fixed intervals between TD records, but the intervals between TD within a herd are never exactly equal. How to correctly write the $\mathbf{F}_i$ for the i^{th} cow is not clear.

13.6.2 Multiple Lactations

Harville (1979) suggested that the cow's permanent environmental effect might be a first-order autoregressive process from one lactation to the next. This is similar to the cumulative PE model described in Chapter 10, except that the PE effects are now correlated between lactations rather than independent and cumulative. The PE effects between lactations can be viewed as long term environmental effects (LTE), and then STE effects within each lactation and within cow. Both of these factors could have autocorrelation structures, with different correlation values. Thus, there would be a PE estimate for each lactation on the cow, and then a STE estimate for each TD record within a lactation. The number of equations could become very large.

13.7 Ali-Schaeffer RRM

The Ali and Schaeffer (1987) covariates will be used to analyze the data in Table 13.1. The cows were all in one herd, and there were 10 test dates. To simplify matters, the cows' records are assumed to be perfectly adjusted for age and month of calving, and are all in the same year. The model might be

$$\begin{aligned} y_{tijk} &= (b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4) \\ &\quad +HTD_j \\ &\quad +(a_{0i} + a_{1i}X_1 + a_{2i}X_2 + a_{3i}X_3 + a_{4i}X_4) \\ &\quad +(p_{0i} + p_{1i}X_1 + p_{2i}X_2 + p_{3i}X_3 + p_{4i}X_4) \\ &\quad +e_{tijk} \end{aligned}$$

where

y_{tijk} is a 24 h TD milk yield in first lactation,

HTD_j is a random herd-test date effect,

t is days in milk,

$X_1 = (t/305)$,

$X_2 = (t/305)^2$,

$X_3 = \ln(305/t)$

$X_4 = (\ln(305/t))^2$

b_0, b_1, b_2, b_3, b_4 are the fixed, overall mean regressions,

$a_{0i}, a_{1i}, a_{2i}, a_{3i}, a_{4i}$ are the animal additive genetic random regressions,

$p_{0i}, p_{1i}, p_{2i}, p_{3i}, p_{4i}$ are the animal permanent environmental random regressions , and e_{tijk} is a random residual effect.

Let the HTD effects have covariance matrix $\mathbf{I}\sigma_h^2$, and within an animal,

$$Var \begin{pmatrix} a_{0i} \\ a_{1i} \\ a_{2i} \\ a_{3i} \\ a_{4i} \end{pmatrix} = \mathbf{G}$$

where $\mathbf{G}$ has order 5, and across animals is

$$Var(\mathbf{a}) = \mathbf{A} \bigotimes \mathbf{G}$$

where $\bigotimes$ is the direct product of two matrices, each element of $\mathbf{A}$ times the entire matrix $\mathbf{G}$. Similarly,

$$Var(\mathbf{p}) = \mathbf{I} \bigotimes \mathbf{P}$$

for the permanent environmental effects. The residual variance will be assumed to be constant throughout the lactation. In practice, however, the residual variance

changes through the lactation, such that there could be a different variance for each DIM. We will assume that the residual variances are independent from day to day.

The initial problem is how to find a $\mathbf{G}$ and $\mathbf{P}$ matrix when such an analysis has not been attempted previously. Start with the matrix $\mathbf{V}$ given in Table 13.2. Assuming heritability is 0.25, then a matrix of additive genetic variances and covariances for those 11 days in milk would be $\mathbf{V}_g = 0.25 \times \mathbf{V}$. Instead of making Φ as with Legendre polynomials, make a matrix of order 11 by 5 with the appropriate covariates from the above model, as shown in the following table (Table 13.9).

Table 13.9: Covariates for Variance Model

DIM	X_0	X_1	X_2	X_3	X_4
5	1	0.01639344	0.0002687450	4.1108739	16.89928393
35	1	0.11475410	0.0131685031	2.1649637	4.68706789
65	1	0.21311475	0.0454178984	1.5459245	2.38988258
95	1	0.31147541	0.0970169309	1.1664349	1.36057034
125	1	0.40983607	0.1679656006	0.8919980	0.79566050
155	1	0.50819672	0.2582639076	0.6768867	0.45817555
185	1	0.60655738	0.3679118517	0.4999560	0.24995595
215	1	0.70491803	0.4969094329	0.3496737	0.12227173
245	1	0.80327869	0.6452566514	0.2190536	0.04798446
275	1	0.90163934	0.8129535071	0.1035407	0.01072067
305	1	1.00000000	1.0000000000	0.0000000	0.00000000

If the above matrix is $\mathbf{B}$, then

$$
\begin{aligned}
\mathbf{V}_g &= 0.25 \times \mathbf{V} \\
\mathbf{V}_g &= \mathbf{BGB}' \\
\mathbf{G} &= (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{V}_g\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \\
&= \begin{pmatrix}
783.40 & -1251.74 & 483.64 & -431.55 & 59.15 \\
-1251.74 & 2032.57 & -801.16 & 688.39 & -94.25 \\
483.64 & -801.16 & 326.74 & -264.17 & 36.06 \\
-431.55 & 688.39 & -264.17 & 239.86 & -32.98 \\
59.15 & -94.25 & 36.06 & -32.98 & 4.55
\end{pmatrix}
\end{aligned}
$$

Now compute the correlations among the elements of $\mathbf{G}$ giving

$$\mathbf{C}_g = \begin{pmatrix} 1 & -0.9920 & 0.9559 & -0.9955 & 0.9904 \\ -0.9920 & 1 & -0.9831 & 0.9859 & -0.9799 \\ 0.9559 & -0.9831 & 1 & -0.9437 & 0.9351 \\ -0.9955 & 0.9859 & -0.9437 & 1 & -0.9981 \\ 0.9904 & -0.9799 & 0.9351 & -0.9981 & 1 \end{pmatrix}.$$

Notice the many correlations that are close to -1 or +1. Such a matrix can lead to estimation problems in the MME due to the high dependency between covariates. For example, an iterative solution program could take a very long time to converge to a solution.

If repeatability is .35, and using the same $\mathbf{B}$ as for finding $\mathbf{G}$, then

$$\begin{aligned} \mathbf{V}_p &= 0.10 \times \mathbf{V} \\ \mathbf{V}_p &= \mathbf{BPB}' \\ \mathbf{P} &= (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{V}_p\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1} \\ &= \begin{pmatrix} 313.36 & -500.70 & 193.46 & -172.62 & 23.66 \\ -500.70 & 813.03 & -320.46 & 275.36 & -37.70 \\ 193.46 & -320.46 & 130.69 & -105.67 & 14.43 \\ -172.62 & 275.36 & -105.67 & 95.94 & -13.19 \\ 23.66 & -37.70 & 14.43 & -13.19 & 1.82 \end{pmatrix} \end{aligned}$$

The resulting correlation matrix is the same as that for $\mathbf{G}$ due to that both $\mathbf{G}$ and $\mathbf{P}$ were derived from multiples of $\mathbf{V}$.

13.7.1 MME and Solutions

The $\mathbf{X}$ matrix for this model has 39 rows (because there are 39 TD records) and 5 columns for the five Ali and Schaeffer covariates. The design matrix for HTD effects is 39 by 10, containing zeros and one 1 in each row for the appropriate HTD number. The design matrix for animal genetic effects has 39 rows and 60 columns (12 animals times 5 covariates each). The animal PE design matrix has 39 rows and 25 columns (only 5 animals with records and 5 covariates each). In total there are 100 equations in MME. The assumptions were that σ_h^2 was 8, and σ_e^2 was 5. Additive genetic relationship matrix inverse was used. The resulting solutions were

$$\widehat{\mathbf{b}} = \begin{pmatrix} 42.547 & -48.853 & 18.412 & -10.869 & 1.148 \end{pmatrix}$$

for the overall mean. Thus, an estimate of an average yield on day 65 (using covariates from Table 13.7.1) would be

$$\begin{aligned}\widehat{y}_{65} &= 42.547 - 48.853(0.213) + 18.412(0.045) \\ &\quad -10.869(1.546) + 1.148(2.390) \\ &= 18.91\text{kg}\end{aligned}$$

The HTD solutions were (2.998 ; 0.088 ; -0.234 ; 0.438 ; 1.526 ; -0.325 ; -0.307 ; -0.283 ; -1.686 ; -2.215) for HTD 1 to 10, respectively.

The animal additive genetic solutions are given in Table 13.10, and the solutions for animal PE effects are in Table 13.11.

Table 13.10: Animal Genetic Random Regression Solutions

Animal	a_{0i}	a_{1i}	a_{2i}	a_{3i}	a_{4i}
1	10.255	-16.813	5.828	-6.290	0.843
2	-10.255	16.813	-5.828	6.290	-0.843
3	4.519	-5.166	0.616	-2.185	0.261
4	-5.362	7.480	-1.998	3.110	-0.414
5	3.228	-5.316	2.483	-1.587	0.210
6	-4.893	9.334	-3.831	3.180	-0.429
7	2.509	-6.332	2.729	-2.517	0.372
8	11.906	-16.155	3.838	-6.422	0.813
9	-13.171	19.626	-5.910	7.810	-1.043
10	9.970	-16.381	6.638	-5.526	0.736
11	-12.467	22.407	-8.660	7.914	-1.064
12	8.891	-17.904	7.008	-6.921	0.979

13.7.2 EBV for 305-d Yields

The regression coefficients can not be used, easily, to determine the best cows and sires. The solutions need to be converted to a 305-d basis.

To do that, we need: $\sum_{i=5}^{305} X_0 = 301.0000$, $\sum_{i=5}^{305} X_1 = 152.9672$, $\sum_{i=5}^{305} X_2 = 102.1669$, $\sum_{i=5}^{305} X_3 = 281.5174$, $\sum_{i=5}^{305} X_4$=482.9813.

Table 13.11: Animal PE Random Regression Solutions

Animal	p_{0i}	p_{1i}	p_{2i}	p_{3i}	p_{4i}
8	3.615	-4.133	0.493	-1.748	0.209
9	-4.290	5.984	-1.598	2.488	-0.331
10	2.582	-4.253	1.986	-1.270	0.168
11	-3.915	7.467	-3.065	2.544	-0.343
12	2.007	-5.065	2.184	-2.014	0.298

For sire 1, for example, his 305-d milk EBV is

$$\begin{aligned} EBV_1 &= 301(10.255) - 152.9672(16.813) + 102.1669(5.828) \\ &\quad -281.5174(6.290) + 482.9813(0.843) \\ &= -253.06kg \text{ kg} \end{aligned}$$

The same constants are used with the regression solutions of each animal. The resulting EBV are given in table 13.12:

Table 13.12: Animal EBVs from different models

Animal	Fixed Reg.	Autoreg.	A&S
1	-295	-334	-253
2	295	334	253
3	148	132	144
4	23	26	1
5	120	132	67
6	272	308	252
7	-562	-597	-464
8	74	32	89
9	181	206	129
10	33	31	-27
11	555	628	504
12	-991	-1063	-822

13.8 Legendre Polynomial RRM

Legendre polynomials are used for the covariates in this test day model. The model equation is identical to the first test day model.

Table 13.13: Covariates for Variance Model

DIM	X_0	X_1	X_2	X_3	X_4
5	0.7071	-1.2247	1.5811	-1.8708	2.1213
35	0.7071	-0.9798	0.7273	-0.1497	-0.4943
65	0.7071	-0.7348	0.0632	0.6735	-0.8655
95	0.7071	-0.4899	-0.4111	0.8232	-0.2397
125	0.7071	-0.2449	-0.6957	0.5238	0.4921
155	0.7071	0.0000	-0.7906	0.0000	0.7955
185	0.7071	0.2449	-0.6957	-0.5238	0.4921
215	0.7071	0.4899	-0.4111	-0.8232	-0.2397
245	0.7071	0.7348	0.0632	-0.6735	-0.8655
275	0.7071	0.9798	0.7273	0.1497	-0.4943
305	0.7071	1.2247	1.5811	1.8708	2.1213

If the above matrix is Φ, then assuming $h^2 = 0.25$ as before, and $\mathbf{V}_g = 0.25 \times \mathbf{V}$, then

$$\mathbf{V}_g = \mathbf{\Phi G \Phi'}$$

$$\begin{aligned} \mathbf{G} &= (\mathbf{\Phi'\Phi})^{-1}\mathbf{\Phi'V}_g\mathbf{\Phi}(\mathbf{\Phi'\Phi})^{-1} \\ &= \begin{pmatrix} 6.0050 & 0.2172 & -0.4186 & 0.1411 & -0.1193 \\ 0.2172 & 0.8720 & -0.1859 & 0.0085 & -0.0225 \\ -0.4186 & -0.1859 & 0.3581 & -0.0914 & 0.0138 \\ 0.1411 & 0.0085 & -0.0914 & 0.1704 & 0.0028 \\ -0.1193 & -0.0225 & 0.0138 & 0.0028 & 0.0682 \end{pmatrix}, \end{aligned}$$

and the correlations among the elements of **G** are

$$\mathbf{C}_g = \begin{pmatrix} 1.00 & 0.09 & -0.29 & 0.14 & -0.19 \\ 0.09 & 1.00 & -0.33 & 0.02 & -0.09 \\ -0.29 & -0.33 & 1.00 & -0.37 & 0.09 \\ 0.14 & 0.02 & -0.37 & 1.00 & 0.03 \\ -0.19 & -0.09 & 0.09 & 0.03 & 1.00 \end{pmatrix}.$$

With Legendre Polynomials, which are orthogonal over the entire range (-1 to +1), the correlations are much smaller than with the Ali and Schaeffer function. This leads to better estimation of EBV and covariance components from this analysis. Convergence to solutions for EBV is not delayed due to a high dependency between covariates.

Repeatability is 0.35, then $\mathbf{V}_p = 0.10 \times \mathbf{V}$, and using the same $\mathbf{\Phi}$ as for finding **G**, then

$$\mathbf{P} = \begin{pmatrix} 2.4020 & 0.0869 & -0.1674 & 0.0564 & -0.0477 \\ 0.0869 & 0.3488 & -0.0744 & 0.0034 & -0.0090 \\ -0.1674 & -0.0744 & 0.1432 & -0.0365 & 0.0055 \\ 0.0564 & 0.0034 & -0.0365 & 0.0681 & 0.0011 \\ -0.0477 & -0.0090 & 0.0055 & 0.0011 & 0.0273 \end{pmatrix}$$

The resulting correlation matrix is the same as that for **G** because both **G** and **P** were derived from multiples of **V**.

13.8.1 MME and Solutions

The **X** matrix consists of order 5 Legendre Polynomials (LP) for the DIM on which the TD records were made. In practice, one would construct a table of LP for days 5 through 305, then if a TD record occurs on day 14, just pick out the row for day 14. (NOTE: In the example data, the first record is made on day 4, this was changed to day 5 to fit the Legendre polynomials for days 5 to 305). The design matrix for additive genetic and PE effects use the same LP as in **X**. The dimensions of the matrices are the same as in the previous test day model.

Again, the assumed variances were $\sigma^2_h = 8$, and $\sigma^2_e = 5$. The additive genetic relationship matrix inverse was used. The resulting solutions were

$$\widehat{\mathbf{b}} = \begin{pmatrix} 22.212 & -3.239 & -0.254 & 0.820 & -0.620 \end{pmatrix}$$

for the overall mean. An estimate of an average yield on day 65 (using LP covariates from Table 13.13) would be

$$\begin{aligned}\widehat{y}_{65} &= 22.212(0.7071) + 3.239(0.7348) - 0.254(0.0632) \\ &\quad +0.820(0.6735) + 0.620(0.8655) = 19.16 \text{ kg}\end{aligned}$$

The HTD solutions were

$$(3.069 \quad 0.221 \quad -0.341 \quad 0.411 \quad 1.504$$

$$-0.460 \quad -0.248 \quad -0.432 \quad -1.779 \quad -1.945)$$

for HTD 1 to 10, respectively.

The animal additive genetic solutions are given in Table 13.14, and the solutions for animal PE effects are in Table 13.15.

Table 13.14: Animal Genetic Random Regression Solutions

Animal	a_{0i}	a_{1i}	a_{2i}	a_{3i}	a_{4i}
1	-1.168	0.163	-0.239	0.148	-0.036
2	1.168	-0.163	0.239	-0.148	0.036
3	0.679	-0.158	-0.302	0.127	-0.042
4	-0.017	-0.055	0.204	-0.081	0.023
5	0.324	0.030	0.040	0.042	-0.019
6	1.185	-0.107	0.035	-0.067	0.013
7	-2.170	0.291	0.023	-0.021	0.025
8	0.434	-0.156	-0.573	0.264	-0.081
9	0.559	-0.164	0.425	-0.195	0.053
10	-0.098	0.126	-0.059	0.137	-0.047
11	2.361	-0.242	0.172	-0.175	0.037
12	-3.839	0.518	-0.086	0.043	0.020

13.8.2 EBV for 305-d Yields

The solutions need to be converted to a 305-d basis. To do that we need the sum of the LP covariates from 5 to 305 days, as shown below:

$\sum_{i=5}^{305} X_0 = 212.8391$, $\sum_{i=5}^{305} X_1 = 0.0000$, $\sum_{i=5}^{305} X_2 = 1.5864$,

$\sum_{i=5}^{305} X_3 = 0.0000$, $\sum_{i=5}^{305} X_4 = 2.1449$.

Table 13.15: Animal PE Random Regression Solutions

Animal	p_{0i}	p_{1i}	p_{2i}	p_{3i}	p_{4i}
8	0.543	-0.126	-0.242	0.101	-0.034
9	-0.014	-0.044	0.163	-0.064	0.019
10	0.259	0.024	0.032	0.034	-0.015
11	0.948	-0.086	0.028	-0.054	0.010
12	-1.736	0.233	0.018	-0.017	0.020

For sire 1, for example, his 305-d milk EBV is

$$\begin{aligned} EBV_1 &= 212.8391(-1.168) + 1.5864(-0.239) + 2.1449(-0.036) \\ &= -249 \text{ kg} \end{aligned}$$

The same constants are used with the regression solutions of each animal. The resulting EBV are

Table 13.16: Animal EBVs from different models

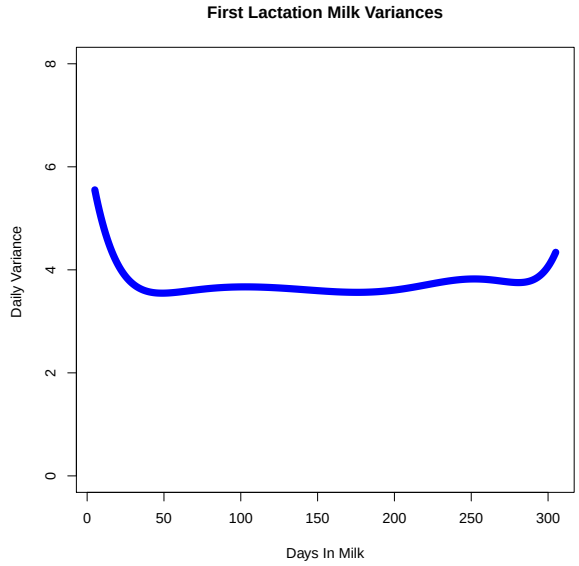
Animal	Fixed Reg.	Autoreg.	A&S	LP
1	-295	-334	-253	-249
2	295	334	253	249
3	148	132	144	144
4	23	26	1	-3
5	120	132	67	69
6	272	308	252	252
7	-562	-597	-464	-462
8	74	32	89	91
9	181	206	129	120
10	33	31	-27	-21
11	555	628	504	503
12	-991	-1063	-822	-817

13.9 Spline Function RRM

Take the **G** matrix from the LP test day model in the previous section. From that we can estimate the genetic variance for every day in the lactation from day 5 to day 305 using the LP covariates for each day. If we plot those variances

we obtain the following figure (Figure 13.4).

Figure 13.4



Notice the higher genetic variances at the beginning and at the tail of lactation. These are not the true genetic variances, but are known as artifacts created by the polynomial nature of the model. What is more important is the **G** matrix which describes the variation in the genetic regression coefficients among animals. However, some researchers are offended by the artifacts as in the above figure.

Spline functions (segmented polynomials) have been suggested to replace LPs. The lactation is divided into sections by locations known as "knots". The production between any two knots is assumed to be changing linearly. The challenge is to determine the number of knots and where they should be placed throughout the lactation (Jamrozik et al. 2010). The more knots used the better the fit of the models, but more unknowns give more parameters to estimate per animal. Their study looked at 4 to 7 knots. In this example, we will use 5 knots, and the locations founded in their study. Namely, $T_1 = 7$, $T_2 = 54$, $T_3 = 111$, $T_4 = 246$, and $T_5 = 302$,

Let t be a particular days in milk between 4 and 305, and T_i represent the five knots, for this example, then the covariates, x_i for the spline function are determined as follows:

- If $t < T_1$, then $x_1 = t/T_1$ and other x_i for $i > 1$ are zero.

- If $t > T_5$, then $x_5 = T_5/t$ and other x_i for $i < 5$ are zero.
- If $T_i < t < T_{i+1}$, then

$$x_i = (t - T_i)/(T_{i+1} - T_i)$$

$$x_{i+1} = (T_{i+1} - t)/(T_{i+1} - T_i)$$

and the other covariates are zero.

Thus, there are at most two non-zero covariates per days in milk, which gives spline functions some computational advantages. An example of the covariates are given in the following table (13.17), for the 11 days represented in $\mathbf{V}$.

Table 13.17: Spline Function Covariates for Particular DIM

Day	X_0	X_1	X_2	X_3	X_4
5	0.7143	0	0	0	0
35	0.5957	0.4043	0	0	0
65	0	0.1930	0.8070	0	0
95	0	0.7193	0.2807	0	0
125	0	0	0.1037	0.8963	0
155	0	0	0.3259	0.6741	0
185	0	0	0.5481	0.4519	0
215	0	0	0.7704	0.2296	0
245	0	0	0.9926	0.0074	0
275	0	0	0	0.5179	0.4821
305	0	0	0	0	0.9902

If we let $\mathbf{X}$ represent the table of covariates above (Table 13.17), in matrix notation, then note that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 0.8651 & 0.2408 & 0 & 0 & 0 \\ 0.2408 & 0.7181 & 0.3576 & 0 & 0 \\ 0 & 0.3576 & 2.7262 & 0.7446 & 0 \\ 0 & 0 & 0.7446 & 1.7829 & 0.2497 \\ 0 & 0 & 0 & 0.2497 & 1.2129 \end{pmatrix}$$

which is a tri-diagonal matrix, which can be used to advantage in setting up and solving MME.

The derivation of the appropriate $\mathbf{G}$ and $\mathbf{P}$ follows as before. Assuming

$h^2 = 0.25$ as before, and $\mathbf{V}_g = 0.25 \times \mathbf{V}$, then

$$\begin{aligned} \mathbf{V}_g &= \mathbf{XGX}' \\ \mathbf{G} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_g\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \begin{pmatrix} 8.3229 & 2.3771 & 2.1918 & 2.0917 & 1.4394 \\ 2.3771 & 4.9380 & 2.6510 & 3.4222 & 1.9740 \\ 2.1918 & 2.6510 & 3.4241 & 3.2226 & 3.0439 \\ 2.0917 & 3.4222 & 3.2226 & 3.8411 & 2.7010 \\ 1.4394 & 1.9740 & 3.0439 & 2.7010 & 4.5052 \end{pmatrix}, \end{aligned}$$

and $V_p = 0.10 \times \mathbf{V}$ assuming repeatability of 0.35. Thus,

$$\mathbf{P} = \begin{pmatrix} 3.3292 & 0.9508 & 0.8767 & 0.8367 & 0.5757 \\ 0.9508 & 1.9752 & 1.0604 & 1.3689 & 0.7896 \\ 0.8767 & 1.0604 & 1.3696 & 1.2890 & 1.2176 \\ 0.8367 & 1.3689 & 1.2890 & 1.5364 & 1.0804 \\ 0.5757 & 0.7896 & 1.2176 & 1.0804 & 1.8021 \end{pmatrix}.$$

13.9.1 Solutions and EBV

The solutions for the overall means represent the average yield at the location of the knots.

$$\widehat{\mu} = (18.67 \;\; 17.79 \;\; 16.12 \;\; 14.88 \;\; 12.98\ ).$$

The solutions for herd-testdays were

$$(4.99 \;\; 1.14 \;\; 1.32 \;\; -0.01 \;\; 1.87$$

$$-0.63 \;\; -0.86 \;\; -1.44 \;\; -2.37 \;\; -2.96)$$

The additive genetic solutions are shown in Table 13.18. Note that the solutions are similar, but not the same, for each knot, meaning that their production was consistently above or below the overall means.

The animal PE solutions are in Table 13.19.

The sum of the spline function covariates over days 4 to 305 give

$$(26.1429 \;\; 52 \;\; 96 \;\; 95.5 \;\; 31.48028),$$

which represent the average number of days between the knots. These numbers are multiplied times the animal genetic solutions to give a 305-d EBV, as shown in the last column of Table 13.20

Table 13.18: Animal Genetic Solutions for Spline Function RRM

Animal	a_{0i}	a_{1i}	a_{2i}	a_{3i}	a_{4i}
1	-1.396	-0.719	-0.800	-0.777	-0.853
2	1.396	0.719	0.800	0.777	0.853
3	0.323	0.911	0.377	0.568	-0.209
4	0.374	-0.297	-0.257	-0.332	-0.037
5	-0.262	-0.160	-0.398	-0.386	-0.109
6	1.022	1.016	1.057	1.109	0.889
7	-1.457	-1.470	-0.779	-0.959	-0.534
8	-0.213	1.006	0.165	0.464	-0.740
9	1.259	-0.085	0.014	-0.110	0.371
10	-1.091	-0.600	-0.996	-0.967	-0.590
11	2.230	1.884	1.986	2.051	1.760
12	-2.883	-2.565	-1.569	-1.827	-1.228

Table 13.19: Animal PE Solutions for Spline Function RRM

Animal	p_{0i}	p_{1i}	p_{2i}	p_{3i}	p_{4i}
8	0.258	0.729	0.301	0.455	-0.167
9	0.299	-0.237	-0.206	-0.266	-0.029
10	-0.210	-0.128	-0.318	-0.309	-0.087
11	0.817	0.813	0.846	0.887	0.711
12	-1.165	-1.176	-0.623	-0.767	-0.427

The EBVs from the spline function RRM are very different from the previous RRM that utilize smooth functions. This suggests that possibly more knots are required, which increases the number of parameters to estimate per animal. Five knots are not enough to capture the curves in the lactation curve shape. Jamrozik et al. (2010) suggest that at least 7 knots are needed, but that more knots significantly increase the computing cost.

13.10 Multiple Trait RRM

Most applications of RRM have been multiple trait systems. In Canada, there are the first two lactations and third and later lactations, and within each of those includes milk, fat, and protein yields, and somatic cell scores giving 12 traits. Each trait has 5 LP covariates, giving 60 parameters for the additive

Table 13.20: Animal EBVs from different models

Animal	Fixed Reg.	Autoreg.	A&S	LP	Spline
1	-295	-334	-253	-249	-252
2	295	334	253	249	252
3	148	132	144	144	140
4	23	26	1	-3	-63
5	120	132	67	69	-94
6	272	308	252	252	315
7	-562	-597	-464	-462	-298
8	74	32	89	91	84
9	181	206	129	120	31
10	33	31	-27	-21	-266
11	555	628	504	503	598
12	-991	-1063	-822	-817	-572

genetic effects, and 60 for the animal permanent environmental effects.

Clearly first lactations have a lactation shape that does not peak as high as later lactations, and first lactation is usually more persistent, more slow to decline in yield than later lactations. Lactations 2 and 3 were kept separate because it was not known how different those two lactations could be when the research began. Combining later lactations with third lactations seemed appropriate because cows average 3.5 lactations in Canada, so that there would be many fewer lactations after third lactation. Also, the shape of later lactations were similar to those of lactation three. Lactations greater than 5 were omitted from genetic evaluations because production levels in those lactations decreases.

If lactations 4 and 5 were also separated, then there would be too many parameters per animal to estimate, and many of those would be based on the assumed genetic correlations between lactations.

Genetic parameters were estimated from subsets of the data in which cows had at least 7 TD records per lactation, in the first three lactations. Bayesian methods were used employing Gibbs Sampling procedures. Many months of computing were spent to obtain the estimated covariance matrices.

13.11 Lifetime Production RRM

Jensen (2001) discussed a lifetime production RRM, in which TD milk yields in any lactation, for example, were modelled with covariates across lactations and within lactations. The across lactation covariates would account for intervals between lactations, for days not pregnant, and gestation lengths. The

results would give EBV for milk yield for each lactation, but also an EBV for milk yield over the cow's lifetime.

The goal of a lifetime production RRM would be to include all lactations on every cow. After accounting for the first five lactations, additional lactations would only add another 20% of TD records. Thus, is it worthwhile to build a more complex model in order to add in more TD records on animals which are too old to matter?

13.12 References

ALI, T. E. , L. R. SCHAEFFER. 1987. Accounting for covariances among test day milk yields in dairy cows. Can. J. Anim. Sci. 67:637-644.

CARVALHEIRA, J. G. V. , E. J. POLLAK, R. L. QUAAS, R. W. BLAKE. 2002. An autoregressive repeatability animal model for test-day records in multiple lactations. J. Dairy Sci. 85:2040-2045.

HARVILLE, D. A. 1979. Recursive estimation using mixed linear model with autoregressive random effects. In "Variance Components and Animal Breeding." Proceedings of Conf. in Honor of C. R. Henderson, Cornell Univ. page 157.

HENDERSON, C. R. 1984. Applications of Linear Models in Animal Breeding University of Guelph, Guelph, on Third Edition edited by L.R. Schaeffer.

HENDERSON, Jr, C. R. 1982. Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. Biometrics 38:623-640.

JAMROZIK, J. , J. BOHMANOVA, L. R. SCHAEFFER. 2010. Selection of locations of knots for linear splines in random regression test-day models. J. Anim. Breed. Genet. 127:87-92.

JENSEN, J. 2001. Genetic evaluation of dairy cattle using test-day models. J. Dairy Sci. 84:2803-2812.

KIRKPATRICK, M. , D. LOFSVOLD, M. BULMER. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. Genetics 124:979-993.

KIRKPATRIK, M. , Thompson, R., and Hill, W.G. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. Genetical Research 64:57-69.

PTAK, EWA , L. R. SCHAEFFER. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. Livest. Prod. Sci. 34:23-34.

ROTHSCHILD, M. F. , S. NEWMAN. 2002. *Intellectual Property Rights in Animal Breeding and Genetics.* p.233-246.

WILMINK, J. B. M. 1987. Adjustment of test day milk, fat, and protein yield for age, season and stage of lactation. Livest. Prod. Sci. 16:335-348.

WOOD, P. D. P. 1967. Algebraic model of the lactation curve in cattle. Nature 216:164-165.

Chapter 14

Genetic Change

HORIA GROSU
P. A. OLTENACU

14.1 Introduction

The success or failure of a breeding program is measured by an estimate of genetic trend in the population. Estimates are needed to justify the usefulness of genetic evaluations, and therefore, the justification of new methodologies. The average performance of a population can change over years due to genetic and environmental causes. The change due to the environment can be caused by the changes in feeding and management technologies, and in the health status of the herd. Individual performance also changes due to ageing. The task of separating genetic and environmental causes became a concern early in dairy cattle breeding.

Prior to 1950, there were no procedures that could separate the genetic change, (g), and the environmental change, (t), from the total change. After 1950, the average performance of a selected population could be compared to an unselected control population. Such procedures were particularly suitable for small species, such as poultry and pigs (Goodwin at al., 1955; Gowe at al., 1959; Dickerson, 1960; cited by Smith, 1962). For larger species, the use of control populations was economically impractical. The maintenance of control populations was too expensive, although some were attempted in Minnesota, USA and elsewhere in university research herds. Instead, the advantages provided by artificial insemination (AI) were used for these species. In cattle, the measurement of genetic trend was done by processing field records. The methodology consisted in the comparison, under similar environmental conditions, of contemporary daughters obtained with frozen semen from the sires which were active several generations

back, with the daughters from the younger sires currently used for breeding. The difference between the average performances of the two categories of daughters reflected half of the genetic gain.

Comparisons could also be done between AI and Natural Service (non-AI) (NS) bulls (Van Vleck and Henderson, 1961). The first attempts to measure the effect of selection in cattle were those of Lorthscher (1937) and Nelson(1943; cited by Rendel and Robertson, 1950), who quantified the genetic change of the population by considering dam performance in successive years. The result may measure the change due to the environment which, subtracted from the total change, gives an estimate of the genetic trend for that particular trait. The performance of the same cows will differ, however, in successive years, due to the genetic factor and to other factors with systematic effects, such as environmental (feeding) or biological (cow age) factors. This presumes that before any comparison is done between successive years, cows' performance records must be standardized to a mature equivalent basis, i.e., to a standard age. This implies the use of correction factors. This approach was criticised, however, at that time, because the correction factors for age were mistaken for age effects and because the errors associated with the estimation of the correction factors would bias the measure of the genetic gain (Rendel and Robertson, 1950). The conclusion was that the estimates of genetic gain were biased because improper correction factors were used. Elston (1959, cited by Smith, 1962) estimated the effects of selection by measuring the change in the average sire effect in relation with time.

14.2 Comparison to Non-AI Sired Daughters

Van Vleck and Henderson (1961) proposed the use of Least Squares (LS) to measure genetic gain in milk yield achieved from 1951 to 1958 in New York state. For that, they used field records from both AI and non-AI cows which completed their first lactations in the same farm-year-season. The results showed that during the considered period, the genetic gain was 7.71 kg for fat yield and 181 kg for milk yield for natural service cows, and was 11.4 kg fat and 232.2 kg milk yield for AI cows.

According to Van Vleck and Henderson (1961), the first lactation performance of a cow with records adjusted to a mature equivalent basis can be described by the following biometric model:

$$y_{ijkp} = \mu + h_i + s_j + m_k + e_{ijkp},$$

where

y_{ijkp} is first lactation performance of a daughter, in farm i, year-season k, by cow

p, daughter of sire j,

μ is the population average,

h_i is the farm effect,

s_j is the sire effect,

m_k is the year-season effect, and

e_{ijkp} is the residual effect.

The first step in the procedure was to calculate the mean differences of an AI sired daughter average from the average of non-AI sired daughters within farm-year-seasons. Let

$\overline{y}_{1ijk}$ be the average of n_{1ijk} AI sired daughters of sire j, in farm-year-season ik,

$\overline{y}_{2i.k}$ be the average of $n_{2i.k}$ non AI sired daughters in farm-year-season ik, then

d_{ijk} is $\overline{y}_{1ijk} - \overline{y}_{2i.k}$, and

w_{ijk} is $(n_{1ijk}n_{2i.k})/(n_{1ijk} + n_{2i.k})$.

Below is a table (14.1) of data for one sire with progeny in 3 farms all within year-season 1.

Table 14.1: Example data for one sire, one year-season, 3 farms

Farm	No. daus.	Dau. Ave.	Non-AI daus.	Non-AI Ave.	w_{ijk}	d_{ijk}
1	2	15	3	5	1.20	5
2	1	18	3	8	0.75	10
3	1	12	2	7	0.67	5
Totals	4		8		2.62	

The weighted difference for this year-season and sire would be

$$d_{.jk} = \frac{\sum_i (w_{ijk} d_{ijk})}{\sum_i w_{ijk}}$$

or

$$d_{.jk} = \frac{1.20(5) + 0.75(10) + 0.67(5)}{1.20 + 0.75 + 0.67} = \frac{16.85}{2.62} = 6.43$$

This is computed for every AI sire and year-season. Suppose the results for 3 AI sires and 4 year-seasons are as shown in Table 14.2.

Table 14.2: Example weights and differences for 3 sires in 4 year-seasons

	Year-Seasons				Sire Total
Sire	1	2	3	4	$\sum_k(w_{.jk}d_{.jk})$
1	6.43 (2.62)	7.00 (3.00)	7.50 (2.50)	0 (0)	56.60
2	0 (0)	5.25 (1.75)	6.66 (1.33)	8.00 (3.33)	44.69
3	4.00 (1.00)	6.00 (2.00)	0 (0)	5.00 (1.33)	22.65
Season Totals $\sum_j(w_{.jk}d_{.jk})$	20.85	42.19	27.61	33.29	

where 6.43 and 2.62 have the follwing significances: $d_{.jk} = 2.62$ and $w_{.jk} = 6.43$

Below we illustrate how the numbers 2.62 and 6.43 in the above table were calculated, for sire 1, in year-season 1. Let's consider that this sire has 4 daughter spread in three herds (Table 14.3).

Table 14.3: Example to estimate actual number of daughters, for sire 1, in year-season 1

Herd	Number of daughters (A.I.)	Number of contemporaries (non-A.I.)	Actual number of daughters ($w_{.jk}$)	Difference between the daughters and the contemporaries d_{ijk}=y_{1ijk}-y_{2ijk}
1	2	3	1.2	d_{111}=15-10=5
2	1	3	0.75	d_{211}=18-8=10
3	1	2	0.67	d_{311}=12-7=5
Total	4	8	$w_{.jk}$=2.62	

The weighted average difference of sire 1, for season 1, determined from its daughters and contemporaries grouped in the three farms are 6.43 lb.

The next step is to construct LS equations for AI sires and year-seasons, as follows:

$$\begin{pmatrix} \mathbf{D} & \mathbf{C} \\ \mathbf{C}' & \mathbf{W} \end{pmatrix} \begin{pmatrix} \mathbf{s} \\ \mathbf{m} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{pmatrix},$$

where (using numbers from Table 14.2)

$$\mathbf{D} = \begin{pmatrix} (2.62+3+2.5) & 0 & 0 \\ 0 & (1.75+1.33+3.33) & 0 \\ 0 & 0 & (1+2+1.33) \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} -2.62 & -3.00 & -2.50 & 0.00 \\ 0.00 & -1.75 & -1.33 & -3.33 \\ -1.00 & -2.00 & 0.00 & -1.33 \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} (2.62+1) & 0 & 0 & 0 \\ 0 & (3+1.75+2) & 0 & 0 \\ 0 & 0 & (2.5+1.33) & 0 \\ 0 & 0 & 0 & (3.33+1.33) \end{pmatrix},$$

$$\mathbf{r}_1 = \begin{pmatrix} 56.60 \\ 44.69 \\ 22.65 \end{pmatrix},$$

and

$$\mathbf{r}_2 = \begin{pmatrix} 20.85 \\ 42.19 \\ 27.61 \\ 33.29 \end{pmatrix}.$$

These equations have a dependency, therefore, one restriction is necessary, which can be imposed by eliminating one equation, or by adding an equation that would force the sum of the year-season solutions to be zero.

Van Vleck and Henderson (1961) describe how to absorb the sire equations into the year-season equations, thus,

$$\mathbf{S} = \mathbf{W} - \mathbf{C}'\mathbf{D}^{-1}\mathbf{C},$$

and

$$\mathbf{t}_2 = \mathbf{r}_2 - \mathbf{C}'\mathbf{D}^{-1}\mathbf{r}_1,$$

then a solution is

$$\widehat{\mathbf{m}} \;=\; \mathbf{S}^{-}\mathbf{t}_2$$

using a generalized inverse of $\mathbf{S}$. One possible solution vector is

$$\widehat{\mathbf{m}} = \begin{pmatrix} 80.996 \\ 70.464 \\ 75.535 \\ 0.000 \end{pmatrix}.$$

Then solutions for each of the sires is obtained by

$$\widehat{\mathbf{s}} = \mathbf{D}^{-1}(\mathbf{r}_1 - \mathbf{C}\widehat{\mathbf{m}}) = \begin{pmatrix} 82.40 \\ 41.88 \\ 56.48 \end{pmatrix}.$$

Thus, the change in non-AI sire merit from year-season 1 to year-season 4 was $\widehat{m}_4 - \widehat{m}_1 = -80.996$. The solutions for AI sires have the non-AI sire averages removed, and thus, a weighted average of the AI sire solutions within year-seasons gives the averages per year-season, as shown below (Table 14.4).

Table 14.4: Trend in AI sires over year-seasons

Year-Season		
1	[2.62(82.40)+1(56.48)]/3.62=	75.24
2	[3(82.40)+1.75(41.88)+2(56.48)]/6.75=	64.21
3	[2.5(82.40)+1.33(41.88)]/3.83=	68.33
4	[3.33(41.88)+1.33(56.48)]/4.67=	45.95

The amount of genetic change was in a negative direction for this trait.

Some of the assumptions of the previous method were that the non-AI sires were random samples of all non-AI sires in each farm-year-season subclass. All sires were randomly mated to dams within year-seasons. This method would have become less reliable over time as the percentage of cows in a farm-year-season that were from non-AI sires became less. Today a very large percentage of cows are from AI sires, and there could be herds without any non-AI sired daughters.

14.3 Regressions of Performance on Time

Most studies conducted on dairy cattle relied on production data from commercial farms. In this case, one of the most used methods was proposed by Smith (1962), who showed that genetic trend can be measured by within-sire regression of performance on time or from differences in the means with time. Several modifications have been proposed to the method of Smith (1962) to reduce biases in genetic trend estimates (Burnside, et. al., 1967; Everett, R. W., et. al., 1967; Harville, D. A., and C. R. Henderson. 1967; Powell, R. L., and A. E. Freeman. 1974; Rothschild, M. F., and C. R. Henderson. 1979).

Smith (1962) showed that when genetic trend is measured using field re-

cords, sires have progeny with a wide distribution over time (years-seasons) and over space (different farms), which ensures genotype continuity and allows quantifying genetic change. Thus, if in a year, the total change in the population is $(g+t)$, then, for a random bull, considering that the mates are a random sample, the change in the successive groups of progeny will be $(0.5*g+t)$. Assuming that genetic change in the population is g, for a particular bull, the genetic change in its progeny during one year is $0.5g$, assuming that the genetic value of the bull is a constant. Thus, the difference $[(g+t)-(0.5*g+t)]$ measures half of the genetic gain achieved in that year.

On the basis of these principles, Smith (1962) proposed two approaches to measure annual genetic gain:

1. Measuring the effect of selection by regression of the performance on time:

$$2(b_{PT} - b_{ST})$$

where b_{PT} is the linear regression of the average population performance on time, while b_{ST} is the within sire regression of progeny performance on time. In order to eliminate the influence of the annual environmental fluctuations, Smith proposed that the within sire regression on time is calculated from the difference between the average of the populations and those of the sire families.

2. Measuring the effect of selection by the differences between averages in time:

$$\frac{2[(\overline{X}_{Ty} - \overline{X}_{Sy}) - (\overline{X}_{T0} - \overline{X}_{S0})]}{y},$$

where $\overline{X}_{Ti}$ is the population mean in year i and $\overline{X}_{Si}$ is a sire family mean in year i, and y is a given number of years.

The procedure of Smith (1962) has been applied in pigs, in order to quantify the genetic change over time for growth and carcass quality traits. Most subsequent studies conducted on dairy cattle used Smith's regression methodology or variants of it. These approaches are based on the following expectations of regressions:

$$E(b_{PT}) = g + t \quad (14.1)$$

$$E(b_{PT/S}) = 0.5g + t \quad (14.2)$$

$$E(b_{(P-\overline{P})T/S}) = -0.5g \quad (14.3)$$

$$E(b_{PT/SD}) = t \quad (14.4)$$

The total trend b_{PT} is measured by the regression of the performance (P) on time

(T). The expected value of the within sire regression ($b_{PT/S}$) is only ($0.5g + t$), because the sire is common for all progeny, and thus its genetic merit is a constant, and only the dams contribute to the genetic gain.

The expected value of the within sire regression on time, calculated on the basis of the differences between the individual performances and the population average ($P - \overline{P}$), and $b_{(P-\overline{P})T/S}$, respectively.

Finally, the regression of within sire and within dam performance on time includes only the environment component (t), because the progeny have the same known parents and thus their parent average breeding values are expected to be the same. From the first 3 equations we may obtain the formulas quantifying genetic trend (g), as follows:

$$\widehat{g} = 2(b_{PT} - b_{PT/S}) \tag{14.5}$$

$$\widehat{g} = b_{PT} - b_{PT/SD} \tag{14.6}$$

$$\widehat{g} = -2(b_{(P-\overline{P})T/S}) \tag{14.7}$$

These relations are valid provided that we have

1. random mating of sires to dams;
2. no culling of dams;
3. no differential mating of dams according to age or genetic ability, and
4. no maternal effects for the production trait.

If deviations from these assumptions occur, the formulas must be corrected in order to produce estimates with higher accuracy.

Everett et al. (1967) made a first attempt to modify these equations. Their purpose was to test the method of Smith (1962), with slight changes, in order to determine the accuracy of measuring genetic trend. The authors introduced corrections for the effect of culling and age of dam effect.

If the bulls are mated to older dams, the estimates of g will be biased.

$$E(b_{PT}) = t + g(1 - 0.5\, b_{DAT}) \tag{14.8}$$

$$E(b_{PT/S}) = t + g(1 - b_{DAT/S})/2 \tag{14.9}$$

where b_{DAT} is the regression of age of dam on time, and $b_{DAT/S}$ is the regression within sire of age of dam on time. The difference

$$b_{DAT/S} - b_{DAT}$$

is a measure of the within herd non-random assignment of cows to sires. This is also true when bulls are mated to younger dams.

With these modifications, Everett et al. (1967) quantified genetic trend using the following equation:

$$2\ (b_{PT}\ -\ b_{PT/S})\ =\ g(1\ -\ b_{DAT}\ +\ b_{DAT/S}) \tag{14.10}$$

$$\widehat{g}\ =\ \frac{2(b_{PT}\ -\ b_{PT/S})}{1\ -\ b_{DAT}\ +\ b_{DAT/S}} \tag{14.11}$$

In a study of 1,556 first lactation cows, Everett et al. (1967) presented estimates for milk production of

$$\begin{aligned} b_{DAT} &= 0.012 \\ b_{DAT/S} &= 0.533 \\ b_{PT} &= 59.3\ kg \\ b_{PT/S} &= -47.7\ kg \\ \widehat{g} &= \frac{2(59.3-(-47.7))}{1-0.012+0.533} \\ &= 140.70\ kg \end{aligned}$$

When the effect of culling on dams is significant, a correction is needed for the regression of progeny performance over time within sire as follows:

$$b_{PT/S}\ =\ (t\ +\ g(1\ -\ b_{DAT/S})/2) - \Delta C$$

where ΔC shows the additive genetic superiority due to dams culled over time. The value is given by the difference between the regression of daughters' performances (total trend) and the within sire regression of dams' production on time:

$$\Delta C\ =\ 0.5(b_{PT}\ -\ b_{DPT/S})$$

where $b_{DPT/S}$ was estimated by Everett et al. (1967) to be 170.2 kg, and thus

$$\Delta C\ =\ 0.5(59.3-170.2)\ =\ -55.45\text{kg}.$$

The 0.5 appears in the formula because dams contribute half their genes to the progeny. Therefore, the new value of $b_{PT/S}$ is

$$b_{PT/S}\ =\ -47.7-\Delta C\ =\ +7.75\text{kg}.$$

Finally, the estimate of annual genetic trend is

$$\widehat{g} \;=\; \frac{2(59.3 - 7.75)}{1 - 0.012 + 0.533} \;=\; 67.78\text{kg}.$$

The effects of dam age and culling are very important for the measurement of genetic gain. However, ΔC was a phenotypic trend, and actually its value should be regressed by heritability and added rather than subtracted, to the within sire regression (Powell and Freeman, 1974). For example,

$$b_{PT/S} \;=\; (t \;+\; g(1 \;-\; b_{DAT/S})/2) + \Delta D$$

where ΔD is one half the genetic merit of the dams of the sire's progeny deviated from the population mean. The following was proposed by Harville and Henderson (1967),

$$\Delta D \;=\; 0.5\; h^2\; b_{(DP-\overline{P})T/HS}$$

where HS refers to within sire-herd subclasses. In their study, Harville and Henderson (1967) used three methods to measure genetic change:

1. Method 1: is the original variant of Smith (1962) using the estimation of within farm regression and within farm by sire regression;
2. Method 2: uses within-sire by dam subclass regressions and
3. Method 3: uses the principles of the contemporary comparison method.

One of the three estimators of the genetic trend can be calculated with the following equation:

$$\widehat{g} \;=\; \frac{2(b_{PT/H} - b_{PT/HS} + \Delta D)}{1 + b_{DAT/HS}}$$

where the regressions are calculated within herds. In this situation $\widehat{g}$ estimates the within herd trend rather than the gross trend (Powell and Freeman, 1974). During the period 1956 to 1962, the total (phenotypic) trend in the cattle population of New York, USA was 176 kg milk and 6.4 kg fat. The genetic trend was -12 kg (Method 1); 68 kg (Method 2) and 58 kg (Method 3), for milk yield, and was -0.1 kg; 3 kg; and 1.6 kg for milk fat for Methods 1, 2, and 3, respectively.

14.4 Using Relatives Other Than Progeny

In previous approaches, the genetic and environmental trends were quantified from progeny performance. There also have been approaches which replaced progeny by collateral relatives (sisters and half-sisters). Thus, Burnside and Legates (1967) studied the use of first lactation performance of the full sisters and

paternal half-sisters to measure the genetic and environmental trends in a population of dairy cattle. The milk yield and the milk fat were analysed for 1953 to 1961. Least squares constants for year of calving were obtained. They used an estimator similar to (14.6), with some corrections for use of full sisters. Thus, when these relatives were used, their performances might be higher than expected when they are the result of directed mating. This would introduce a bias, and the estimator, $b_{PT/SD}$, might be negative, even with no environmental trend in the population. To remove the bias, the authors proposed to subtract 78 kg milk from the production of each first full sister (78 kg was the amount of the difference of 99 kg that was supposedly due to environmental causes). The authors used two variants of model (14.6) to estimate genetic trend of the population:

$$\begin{aligned} \Delta G1 &= b_{(\Delta G+\Delta E)} - b_{(\Delta E)} \\ &= 2(b_{(\Delta G+\Delta E)} - b_{(\Delta G/2+\Delta E)}) \end{aligned}$$

where $b_{(\Delta G+\Delta E)}$; $b_{(\Delta E)}$ and $b_{(\Delta G/2+\Delta E)}$ are the weighted regressions on year of calving. For this population, the total annual genetic gain was 63 kg milk and 0.007% for milk fat. The performances of full sisters were analysed to obtain solutions for the effect of years, corrected for parental genetic effects and for selection. The weighted regressions of the coefficients showed an annual genetic trend of 45 kg milk and 0.018% for milk fat.

The second estimate of genetic gain was obtained by comparing the total trend with half of the genetic trend plus the environmental trend, by analysing the performance of the half-sisters corrected for the genetic effects of sires. The results showed an annual gain of 55 and 45 kg milk and 0.016% for the milk fat. For the milk yield, the two estimates (45 and 55 kg) represent 0.75% and 0.92%, respectively, from the population average (6005 kg ME). The results were similar to the rate of the annual genetic gain estimated by Rendel and Robertson (1950), when progeny testing was not used.

Acharya and Lush (1969) measured the genetic gain in a cattle population from India, for three traits: age at first calving; milk yield at first calving and first calving interval, using two methods.

1. Two times the difference between the total regression and the within sire regression on time, $2(b_{PT} - b_{ST})$;

2. Two times the within sire regression of progeny performance on time, each performance being expressed as deviation from the average of the contemporaries, $-2b_{(S-P)T}$.

The results using the first method were -2.90 months; 30.7 kg milk and

-0.66 months, for each of the three traits. For the second method, the results were -1.48 months; 10.3 kg milk and -0.29 months. As can be seen, the estimates obtained with the first method were larger than estimates from the second method. The authors considered the results produced by the second method to be more credible, because this method eliminates annual environmental fluctuations.

Miller et al. (1969) estimated the genetic merit of a sire by direct and indirect comparisons between all sires. By taking into consideration the sire of each contemporary, the results were free from bias due to genetic trend.

The genetic trend estimated by the regression between the genetic merit of the sire and time for the investigated population, showed an annual genetic gain of 48 kg milk for the daughters of the proven bulls and of 18 kg for the sampled sires.

Hargrove and Legates (1971) used the method suggested by Smith (1962), to measure the annual genetic gain, which avoids the annual fluctuations in yield due to the environmental factors. Thus, the genetic trend was estimated as two times the regression within sire calculated from the difference between the population average and the average of the sire families.

The within sire regression was calculated using the contemporary comparison because it gives a clear estimate of the genetic trend compared to the Herdmate Comparison. Thus, according to the following equation (notations of the authors), the estimate of the trend is half the genetic gain due to the sire:

$$\begin{aligned} b_{(P-S)T} &= (\overline{D} - \overline{C})_2 - (\overline{D} - \overline{C})_1 \\ &= (\overline{D}_2 - \overline{D}_1) - (\overline{C}_2 - \overline{C}_1) \\ &= .5\Delta G_{dD} + .5\Delta G_{sD} - .5\Delta G_{dC} - .5\Delta G_{sC} \\ &= -.5\Delta G_{sC} \end{aligned}$$

where 1 and 2 are the years when the daughters (D) and the contemporaries (C) are compared, and d for dams and s for sires, and ΔG is annual genetic trend. Because the genetic value of the bulls is a constant, then $\Delta G_{sD} = 0$ and the dams of the daughters are assumed to have he same genetic value as the dams of contemporaries, hence $(\Delta G_{dD} - \Delta G_{dC}) = 0$.

If the herdmate comparison was used instead of contemporary comparisons, the within sire regression would be more complex, but $\Delta G_{sD} = 0$. Let H

represent Herdmate Method, then

$$\begin{aligned} b_{P-S}T &= (\overline{D} - \overline{H})_2 - (\overline{D} - \overline{H})_1 \\ &= (\overline{D}_2 - \overline{D}_1) - (\overline{H}_2 - \overline{H}_1) \\ &= .5\Delta G_{dD} + .5\Delta G_{sD} - .5\Delta G_{dH} - .5\Delta G_{sH} \\ &= .5\Delta G_{dD} - .5\Delta G_{dH} - .5\Delta G_{sH} \end{aligned}$$

with other terms as defined above. The authors estimated a genetic trend of 53 kg milk and 1.8 kg fat for the Holstein population and 25 kg milk and 0.9 kg fat for the Jersey breed. The total phenotypic trends were 133 kg milk for Holstein and 68 kg for Jersey.

14.5 Regression within sire, within farm

Hickman (1971) proposed the regression within sires, within farms, on time, of the difference between daughter average performance and contemporary cows' average performance.

Let n_1 equal the number of daughters, n_2 equals the number of contemporaries, d_{ijk} is the difference between daughter and contemporary averages within farm j for sire k and year-season i, where t_{ijk} is the time variable (i.e. number of months in a season).

Also, $w_{ijk} = (n_1 n_2)/(n_1 + n_2)$.

The regression was calculated as

$$b_{dt} = \frac{\sum_j \sum_k \left[\sum_i (w_{ijk} d_{ijk} t_{ijk}) - \frac{(\sum_i (w_{ijk} d_{ijk}))(\sum_i (t_{ijk}))}{\sum_i (w_{ijk})} \right]}{\sum_j \sum_k \left[\sum_i (w_{ijk} t_{ijk}^2) - \frac{(\sum_i (w_{ijk} t_{ijk}))^2}{\sum_i (w_{ijk})} \right]}$$

14.5.1 Example Data

The time variable is $t_{ijk} = 6$ months for all calculations in this example. Data are shown in Table 14.5, and calculated values are given in Table 14.6.

Table 14.5: Example Data for Within Sire and Herd Regression

Herd	Sire	Year	Season	Dau. Ave.	Cont. Ave.	Diff.	No. Daus.	No. Cont.
1	1	1	1	5000	4000	1000	3	4
			2	5500	4800	700	5	3
1	1	2	1	5700	5900	-200	2	8
			2	4200	4500	-300	4	2
1	2	1	1	5250	4100	1150	3	5
			2	4800	5300	-500	5	4
1	2	2	1	4150	5200	-1050	4	5
			2	5200	5800	-600	7	3
1	3	1	1	5700	4200	1500	5	7
			2	4350	4800	-450	8	3
1	3	2	1	7100	6200	900	4	5
			2	5800	7200	-1400	9	3
2	1	1	1	4300	3800	500	8	3
			2	5700	4900	800	4	5
2	1	2	1	5235	4800	435	7	4
			2	4800	5100	-300	3	8
2	3	1	1	4800	5000	-200	5	3
			2	5400	4700	700	4	8
2	3	2	1	5900	5200	700	7	4
			2	4800	5100	-300	3	5

Table 14.6: Intermediate Quantities for Within Sire and Herd Regression

Herd	Sire	Year	Season	**w**	**wd**	**wdt**	$\mathbf{wt}^2$	$(\mathbf{wt})^2$
1	1	1	1	1.71	1714.29	10,285.71	61.71	105.80
			2	1.88	1312.50	7,875.00	67.50	126.56
		Total		3.59	3026.79	18,160.71	129.21	232.36
1	1	2	1	1.60	-320.00	-1,920.00	57.60	92.16
			2	1.33	-400.0	-2,400.00	48.00	64.00
		Total		2.93	-720.00	-4,320.00	105.60	156.16
1	2	1	1	1.88	2156.25	12,937.50	67.50	126.56
			2	2.22	-1111.11	-6,666.67	80.00	177.78
		Total		4.10	1045.14	6,270.83	147.50	304.34
1	2	2	1	2.22	-2333.33	-14,000.00	80.00	177.78
			2	2.10	-1260.00	-7,560.00	75.60	158.76
		Total		4.32	-3593.33	-21,560.00	155.60	336.54
1	3	1	1	2.92	4375.00	26,250.00	105.00	306.25
			2	2.18	-981.82	-5,890.91	78.55	171.37
		Total		5.10	3393.18	20,359.09	183.55	477.62
1	3	2	1	2.22	2000.00	12,000.00	80.00	177.78
			2	2.25	-3150.00	-18,900.00	81.00	182.25
		Total		4.47	-1150.00	-6,900.00	161.00	360.03
2	1	1	1	2.18	1090.91	6,545.45	78.55	171.37
			2	2.22	1777.78	10,666.67	80.00	177.78
		Total		4.40	2868.69	17,212.12	158.55	349.15
2	1	2	1	2.55	1107.27	6,643.64	91.64	233.26
			2	2.18	-654.55	-3,927.27	78.55	171.37
		Total		4.73	452.74	2,716.36	170.18	404.63
2	3	1	1	1.88	-375.00	-2,250.00	67.50	126.56
			2	2.67	1866.67	11,200.00	96.00	256.00
		Total		4.54	1491.67	8.950.00	163.50	382.56
2	3	2	1	2.55	1781.82	10,690.91	91.64	233.26
			2	1.88	-562.50	-3,375.00	67.50	126.56
		Total		4.42	1219.32	7,315.91	159.14	359.82

For the first herd-sire-year subclass, then

$$\begin{aligned} \text{NUM} &= \sum_i (w_{ijk} d_{ijk} t_{ijk}) - \frac{\left(\sum_i (w_{ijk} d_{ijk})\right)\left(\sum_i (t_{ijk})\right)}{\sum_i (w_{ijk})} \\ &= \left(18,160.71 - \frac{(3026.79)(12)}{3.59}\right) \\ &= 8041.31, \end{aligned}$$

and

$$\begin{aligned} \text{DEN} &= \sum_i (w_{ijk} t_{ijk}^2) - \frac{\left(\sum_i (w_{ijk} t_{ijk})\right)^2}{\sum_i (w_{ijk})} \\ &= \left(129.21 - \frac{232.36}{3.59}\right) \\ &= 64.48. \end{aligned}$$

The other herd-sire-year subclasses are in Table 14.7.

Table 14.7: Herd-sire-year subclass Totals

Herd	Sire	Year	NUM	DEN
1	1	1	8,041.31	64.48
1	1	2	-1,374.55	52.36
1	2	1	3,209.82	73.22
1	2	2	-11,583.65	77.74
1	3	1	12,372.76	89.87
1	3	2	-3,814.29	80.50
2	1	1	9,395.61	79.27
2	1	2	1,567.13	84.59
2	3	1	5,008.72	79.27
2	3	2	4,005.88	77.74
Totals			26,828.74	759.04

The regression is

$$b_{dt} = \frac{26828.74}{759.04} = +35.35kg,$$

and the genetic trend is 2 times b_{dt} or +70.69 kg.

14.6 Powell and Freeman Review

Powell and Freeman (1974) reviewed five estimators of genetic trend, as follows:

Method 1

$$\widehat{g}_1 = \frac{2(b_{PT/H} - b_{PT/HS} + \Delta D_1)}{1 + b_{DAT/HS} - b_{DAT/H}}$$

where $\Delta D_1 = 0.5h^2(b_{DPT/HS} - b_{DPT/H})$.

Method 2

$$\widehat{g}_2 = b_{PT/H} - b_{PT/HSD}$$

Method 3

$$\widehat{g}_3 = \frac{-2(b_{(P-\overline{P})T/HS} - \Delta D_2)}{1 + b_{DAT/HS} - b_{DAT/H}}$$

where $\Delta D_2 = 0.5h^2 b_{(DP-\overline{P})T/HS}$ as described by Harville and Henderson (1967).

Method 4

$$4\overline{S}_i = \sum_j (n_{ij}\widehat{S}_j) / \sum_j n_{ij}$$

where $\widehat{S}_j$ is an estimate of one half the transmitting ability of sire j and n_{ij} is the number of daughters of the sire calving in year-season i. The transmitting abilities are obtained from a linear model analysis that accounts for genetic trends as presented in either Powell and Freeman (1974) or in Schaeffer (1973).

Powell et al. (1974) applied this method to the Holstein population from 1960 to 1975 in the USA. For milk yield the trend was 18 kg per year. From 1968 to 1975 the trend was 38 kg per year using average estimated breeding values of sires. When using the average EBV of cows, the values were 8 and 21 kg of milk, respectively.

Method 5

$$E(b_{(P-\overline{P})T/HSD}) = t - (t+g) \;=\; -g.$$

Trend is estimated by the regression within full sister families as deviations from the adjusted herdmate average.

Method 3 gave estimates of 82 kg milk and 1.5 kg fat. Methods which excluded progeny performance overestimated genetic change, while Method 5 gave estimates with very high standard errors.

Schaeffer et al.(1974) estimated the genetic change in Ontario Holsteins as twice the average sire transmitting abilities by year from 1958 to 1972. Sire effects were obtained from a sire model without genetic relationships among bulls, similar to the Northeast AI Sire Comparison Method used at Cornell University (See Chapter 9 for details). The estimates were 41.8 kg milk and 1.26 kg fat yields.

By moving from a sire model to a cow model, Slanger et al. (1976) and Hintz et al. (1978) used estimated breeding values of cows averaged by year of first calving. Another estimator was to obtain a weighted average of the sires' estimated transmitting abilities within years of calving. Both sets of averages could be regressed on years to obtain a single number. For Holsteins, the cow trend for AI sired cows was 17.9 kg while that for non-AI sired cows was 26.1 kg. The trend using sire estimated transmitting abilities was 17.9. These results were lower than the theoretically possible trend, if selection were entirely on milk production, predicted by Rendel and Robertson (1950) or Van Vleck (1976). This is due to the emphasis placed on many other traits by dairy producers, such as conformation, fertility, disease, calving ease, and milking speed. The estimated trends also depend on the assumed value of heritability. If the applied value is larger than the actual heritability, then trends can be overestimated, and likewise if the applied value was lower than the actual heritability, then trends would be underestimated.

Lee et al. (1985) estimated trends in the US Holstein population from 1960 to 1979 using transmitting abilities of sires, dams, maternal grandsires and maternal granddams. One method used

$$\widehat{g} \;=\; \widehat{\beta}_S \;+\; \widehat{\beta}_D,$$

where $\widehat{\beta}_S$ is the regression of average sire PD(Predicted Differences, which are estimated transmitting abilities) on birth year of the progeny, and $\widehat{\beta}_D$ is the regression of average dam CI (Cow Index) on birth year of the progeny, each regression estimated separately. The second method used

$$\widehat{g} \;=\; 2\,\widehat{\beta}_{ETA},$$

where $\widehat{\beta}_{ETA}$ is the regression of average cow CI on birth year of the cow, or the regression of average bull PD on birth year of the bull. Both methods used on dam populations, produced similar results of 54.77 kg for the first method and 51.55 kg milk for the second method, but results were different for the sire population.

Van Vleck et al. (1986) estimated genetic change for Holsteins in the northeast USA using an animal model that considered all genetic relationships among individuals in the population. Until this study, genetic groups had only been applied to sires, but now genetic groups had to be considered for cows also. Phantom parent genetic groups were not yet known. Their estimates of trend from 1970 to 1980 were 39.5 for registered cows, and 38.1 kg for non-registered cows.

Boichard et al. (1995) proposed three methods for estimating genetic trends.

1. Evaluations from repeatability animal model compared to evaluations from a first lactation only animal model. The trends should be similar.

2. Daughter yield deviations (DYD) within sire by calving year compared over years to see that there is little change (regression of zero on years).

3. Variance of successive official evaluation runs could detect systematic trends due to information from new daughters.

The first two methods produce very reliable estimates of genetic trends, but the model must be correct and the data should be accurate. The third method, although having a lower precision of results, was easier to apply. These three methods are employed by Interbull to validate national estimates of genetic trend before a country's EBV are used in international comparisons.

Bonaiti (1993) found that a bias of 100 kg in age correction factors could result in a 40 kg bias in annual genetic trends. Banos (1992) combined data from the USA and Canada, and genetic evaluations for the combined data were run in Canada and the USA. The resulting genetic trends were different, presumably due to the genetic evaluation methods. Bonaiti (1993) also found inconsistent genetic trends between USA and France, and Boichard et al. (1995) showed large differences between estimated and realized genetic change between the USA, the Netherlands, Germany, Italy, and France.

A summary of estimates (from the literature) of genetic trends in milk production for Holsteins only, are presented in Table 14.8 (References are at the end of the chapter).

Table 14.8: Estimates of Genetic Trend For Milk Yield (kg) in Holsteins

AI or non-AI	Period	Location	Estimate	First author	Year
AI		New York	33	Elston,R.C	1959
AI	1951-1959	New York	29	Van Vleck,L.D	1961
non-AI	1951-1959	New York	23	Van Vleck,L.D	1961
both		One herd	3	Gaalaas, R.F	1961
both		Texas	81	Qureshi,A.W	1963
AI	1956-1962	New York	47	Harville,D.A	1967
AI		One herd	51	Branton, C	1967
AI	1953-1961	North Carolina	45	Burnside,E.B	1967
AI	1955-1965	One herd	0	Burnside,E.B	1968
AI		New York	48	Miller,P.D	1969
both		SE USA	53	Hargrove,G.L	1971
both	1958-1967	Florida	33	Verde,O.G	1972
both	1957-1969	Midwest USA	66	Powell,R.L	1974
AI	1958-1972	Ontario	42	Schaeffer,L.R	1975
AI	1966-1972	Quebec	46	Kennedy,B.W	1975
both		Wisconsin	43	Olson,K.L	1976
AI	1957-1975	NE USA	46	Everett,R.W	1976
AI	1957-1976	NE USA	39	Ufford,G.R	1977
AI	1961-1974	NE USA	17.9	Hintz,R.L	1978
both	1958-1975	Canada	38.2	Batra,T.R	1979
both	1959-1975	Florida	43	Moya,J	1985
both	1960-1979	USA	54	Lee,K.L	1985
both	1960-1980	NE USA	39	Van Vleck,L.D	1986
both	1956-1971	North Carolina	120	Legates,J.E	1988
both	1955-1981	NE USA	35	Van Tassell,C.P	1991

14.7 Animal Models

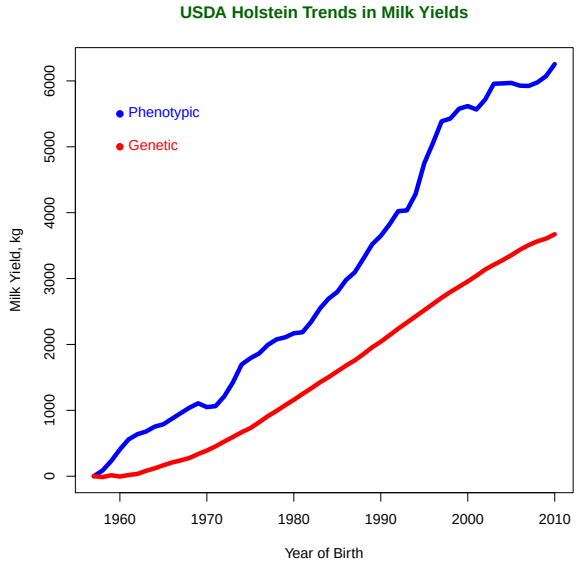
The easiest measures of genetic change can be obtained from EBV calculated in an animal model through MME. However, the precise definition needs to be stated, which means that the cows included in the estimates need to be clearly stated. For example, you could calculate the average EBV of cows by their year of birth, which means each cow's EBV is used only once. Alternatively, you could have the average EBV of cows by year of calving which implies a cow's EBV is

used each year in which it calves, and thus, gives an estimate of the genetic merit of the milking population of cows in a given year. Another estimate could be derived by a weighted average of the sire EBV of each cow by the cow's year of birth. The agreement among these estimates would be interesting to study. The trend in AI sires, by year of sire birth, and the trend in dams of cows, by their year of birth, would also provide relevant information to indicate which pathways of selection are providing most of the genetic change.

With a Test Day model (TDM) cows receive EBV for first, second, and third and later parities, which are genetically correlated traits, although highly correlated. Trends should be similar for each trait.

In the USA, data in their animal model date back to 1957. Both the phenotypic and genetic trends are given in Figure 14.1.

Figure 14.1



Courtesy of USDA AIPL Website

Data were available from the USDA AIPL website. The average EBVs of cows by year of birth increase smoothly and almost evenly over the entire period from 1957 to 2010. The phenotypic trend is not as smooth due to environmental changes and demands of the consumers for milk products. There is a slight leveling of phenotypic and genetic trends in the last few years. Some producers believe cows are giving enough milk, so that it is time to concentrate on efficient producers of milk. Cows that do not need a lot of feed, do not have reproduction problems, or do not have health problems should be favoured. Indexes that combine several

traits are being used more by producers than just milk yield evaluations.

Note that the genetic trend has been 3629 kg over 53 years or 68 kg per year, compared to a phenotypic trend of 6350 kg or 120 kg per year. Genetic change has been more than half of the phenotypic change. Nutrition, technology, and improved management practices account for the remainder. Genetic evaluation methods have been critical to genetic change.

14.8 References

ACHARYA, R. M. , J. L. LUSH. 1968. Genetic progress through selection in a closed herd of Indian cattle. J. Dairy Sci. 51:1059.

BANOS, G. , G. R. WIGGANS, J.A.B. ROBINSON. 1992. Comparison of methods to rank bulls across countries. J. Dairy Sci. 75:2560.

BONAITI, B. , D. BOICHARD, A. BARBAT, S. MATTALIA. 1993. Problems arising with genetic trend estimation in dairy cattle. Proc. INTERBULL Mtg., Aarhus, Denmark, August 19-20. 1998. INTERBULL. Bull. 8.

BRANDTON, C. , EVANS, D.L., STEELS, J.R. and FARTHING, B.R. 1967. Estimated genetic progress in milk and fat yield in Louisiana Holstein Herd. J. Dairy Sci. 50:974.

BOICHARD, D. , B. BONAITI, A. BARBAT, S. MATTALIA. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. J. Dairy Sci. 78:431-437.

BURNSIDE, E. B. , J. E. LEGATES. 1967. Estimation of genetic trends in dairy cattle populations. J. Dairy Sci. 50:1448.

BURNSIDE, E. B. , RENNIE, J. C. and BOWMAN, G. H. 1968. Genetic trends and selection in a dairy cattle herd. Can. J. Anim. Sci. 48:243.

DICKERSON, G. E. 1960. In "Techniques and Procedures in Animal Production Research". American Society of Animal Production, Beltsville, Maryland.

ELSTON, R. C. 1959. The estimation of genetic gain in milk yield due to sire selection over a period of time. Ph.D. Thesis. Cornell University, Ithaca, New York.

EVERETT, R. W. , C. E. MEADOWS, J. L. GILL. 1967. Estimation of genetic trends in simulated data. J. Dairy Sci. 50:550.

EVERETT, R.W. , KEOWN, J.F. and CLAPP, E.E. 1976. Production and stayability trends in dairy cattle. J. Dairy Sci. 59:1532.

FREEMAN, A. E. , G.L. LINDBERG. 1993. Challenges to dairy cattle management: Genetic considerations. J. Dairy Sci. 76:3143-3159.

GAALAAS, R. F. , and PLOWMAN, R. D. 1961. Effectiveness of statistical adjustments for yearly fluctuations in production. J. Dairy Sci. 44:1188.

GOODWIN, K. , G. E. DICKERSON, W. F. LAMOUREUX. 1955. A technique for measuring genetic progress in poultry breeding experiments. Poultry Sci. 34:1197.

BIOMETRICAL GENETICS. Edited by O. KEMPTHORNE, Pergamon Press, London.

GOWE, R. S. , A. ROBERTSON, B. D. H. LATTER. 1959. Environment and poultry breeding problems: 5. The design of poultry control strains. Poultry Sci. 38:462.

HARGROVE, G. L. , J. E. LEGATES. 1971. Biases in dairy sire evaluation attributable to genetic trend and female selection. J. Dairy Sci. 54:1041.

HARVILLE, D. A. , C. R. HENDERSON. 1967. Environmental and genetic trends in production and their effects on sire evaluation. J. Dairy Sci. 50:870.

HICKMAN, C. G. 1971. Response to selection of breeding stock for milk solids production. J. Dairy Sci. 54: 191-198.

HINTZ, R. L. , R.W. EVERETT, L.D. VAN VLECK. 1978. Estimation of genetic trends from cow and sire evaluations. J. Dairy Sci. 61:607.

KENNEDY, B.W. , and MOXLEY J.E. , 1975. Genetic trends among artificially breed Holsteins in Quebec. J. Dairy Sci. 58:1871.

LEE, K. L. , A. E. FREEMAN, L. P. JOHNSON. 1985. Estimation of genetic change in the registered Holstein cattle population. J. Dairy Sci. 68:2629-2638.

LEE, K. L. , FREEMAN, A. E. , JOHNSON L. P. 1985. Estimation of Genetic Change in the Registered Holstein Cattle Population. J Dairy Sci 68:2629-2638.

LEGATES, J. E. , R. M. MYERS. 1988. Measuring genetic change in a dairy herd using a control population. J. Dairy Sci. 71:1025-1033.

LORTHSCHER, H. 1937. Variationsstatistische Untersuchungen an Leistungserhebungen in einer British-Friesian Herde. Z. ZuchtBiol. 39:257.

MILLER, P. D. , W. E. LENTZ, C. R. HENDERSON. 1969. Comparison of contemporary daughters of young and progeny tested dairy sires. J. Dairy Sci. 52:926.

MOYA, J. , WILCOX, C. J. , BACHMAN, K. C. and MARTIN, F. G. 1985. Genetic trends in milk yield and composition in a subtropical dairy herd. Brazil J. Genet. VIII:509-521.

NELSON, R. H. 1943. Measuring the amount of genetic change in a herd average. J. Anim. Sci. 2:358 (Abstract).

OLSON, K. E. , and. JENSEN, E. L. 1976. Estimation of genetic trend in Wisconsin Holsteins. ADSA 71st Ann. Meeting, p. 74.

POWELL, R. L. , A. E. FREEMAN. 1974. Estimators of sire merit. J. Dairy Sci. 57:1228.

POWELL, R. L. , A. E. FREEMAN. 1974. Genetic trend estimators. J. Dairy Sci. 57:1067.

QURESHI, A. W. 1963. Genetic trend in milk and milk fat production in Texas Dairy Herd Improvement Association cows. J. Dairy. Sci. 46:629.

RENDEL, J. M. , A. ROBERTSON. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. J. Genet. 50:1.

ROTHSCHILD, M.F , and HENDERSON, C.R. 1979. Maximum likelihood estimates of parameters of first and second lactation milk records. J. Dairy Sci. 62:990-995.

SCHAEFFER, L. R. , R. W. EVERETT, C. R. HENDERSON. 1973. Lactation records adjusted for days open in sire evaluation. J. Dairy Sci. 56:602.

SCHAEFFER, L. R. , M. G. FREEMAN, E. B. BURNSIDE. 1974. Evaluation of Ontario Holstein dairy sires for milk and fat production. J. Dairy Sci. 58:109.

SCHAEFFER, L. R. , M. G. FREEMAN, and E. B. BURNSIDE.1975. Evaluation of Ontario Holstein dairy sires for milk and fat production. J. Dairy Sci. 58:109.

SLANGER, W. D. , E. L. JENSEN, R. W. EVERETT, C. R. HENDERSON. 1976. Programming cow evaluation. J. Dairy Sci. 59:1589.

SMITH, C. 1962. Estimation of genetic change in farm livestock using field records. Anim. Prod. 4:239.

UFFORD, G. R. 1977. Dairy sire evaluation using all lactation records in best linear unbiased production procedures. Ph.D. Thesis, Cornell niversity, Ithaca, NY.

VERDE, O. G. , WILCOX, C. J., MARTIN, F. G. and REAVES, C. W. 1972. Genetic trends in milk production in Florida Dairy Herd Improvement Association herds. J. Dairy Sci. 55:1010.

VAN TASSELL, C. P. , Van VLECK, L.D. 1991. Estimates of genetic selection differentials and generation intervals for four paths of selection. J. Dairy Sci. 74:1078-086.

VAN VLECK, L. D. , C. R. HENDERSON. 1961. Measurement of genetic trend. J. Dairy Sci. 44:1705.

VAN VLECK, L. D. , 1976. Theoretical and actual genetic progress in dairy cattle. In "International Conference on Quantitative Genetics". Iowa State University Press, Ames, Iowa.

VAN VLECK, L. D. , R. A. WESTELL, J. C. SCHNEIDER. 1986. Genetic change in milk yield estimated from simultaneous genetic evaluation of bulls and cows. J. Dairy Sci. 69:2963.

Chapter 15

Threshold Models

LARRY SCHAEFFER
JANUSZ JAMROZIK

15.1 Categorical Data

There are many traits recorded for dairy cattle which are subjective assessments of the animal. A person decides to which category out of m categories an animal should be assigned. Examples are calving ease, severity of claw disorders, and stature. Some traits categorize themselves, such as diseased or not diseased. When $m = 2$, then the trait is an "all-or-none" or binary trait. Most disease traits are binary in nature.

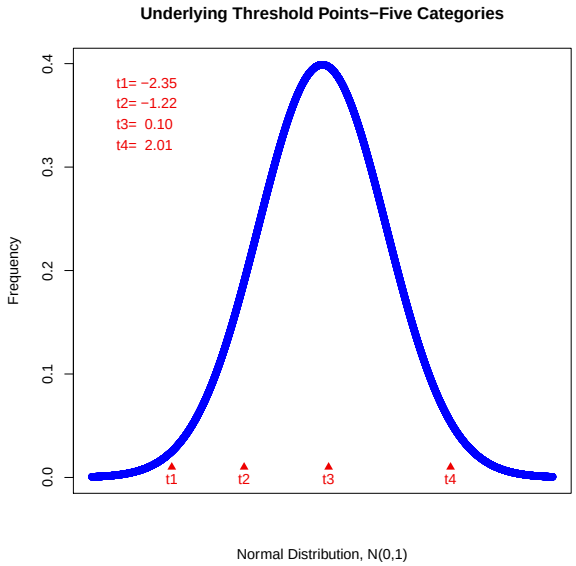
Categories are arranged in a sequence from one extreme expression to the opposite extreme expression. Calving ease, for example, can range from completely unassisted calving, to very difficult calving, even to caesarian section. There could be 3, 4, or 5 categories of calving difficulties.

Categorical traits may be inherited in a polygenic manner. The underlying susceptibility to a disease trait, or to calving ease may actually be continuous and may follow a normal distribution. The underlying continuous scale is known as the *liability* scale. On the liability scale are one or more threshold points $(t_1 < t_2 <,...,< t_{m-1})$ as shown in Figure 15.1.

If the liability value, λ, of an animal is between t_i and t_{i+1}, then the animal belongs to category $i + 1$, and if the liability is below t_1, then the animal belongs to category 1 and if the liability is above t_{m-1} then the animal belongs to category m. The liability scale is only conceptual and cannot be observed.

Threshold models were proposed by Gianola (1982), Gianola and Foulley (1983) and Harville and Mee (1984) as the most theoretically acceptable method

Figure 15.1



of analysis for categorical data. However, Meijering and Gianola (1985) showed that for some situations a linear model may perform just as well in terms of correctly ranking dairy sires. The threshold models assume a continuous normally distributed underlying liability scale for the trait. The thresholds define the categories that are observed. The solution to a threshold model is non-linear in computational complexity, and there must be back and forth calculations of thresholds and effects in the model until convergence of the system of equations stabilizes.

There are various quantities which need to be computed repeatedly in a threshold model analysis, and these are based on normal distribution functions.

1. $\Phi(\lambda)$ is known as the cumulative distribution function of the normal distribution with mean 0 and variance 1. This function gives the area under the normal curve up to the value of λ, for λ going from minus infinity to plus infinity (the range for the normal distribution). For example, if $\lambda = .4568$, then $\Phi(\lambda) = .6761$, or if $\lambda = -.4568$, then $\Phi(\lambda) = .3239$. Let Φ_k represent the value up to and including category k for $k = 1$ to m. Therefore, $\Phi_m = 1$.

2. $\phi(\lambda)$ is a function that gives the height of the normal curve at the value λ, for a normal distribution with mean zero and variance 1. That is,

$$\phi(\lambda) = (2\pi)^{-.5} \exp -.5\lambda^2.$$

For example, if $\lambda = 1.0929$, then $\phi(\lambda) = .21955$.

3. $P(k)$ is the probability that λ from a $N(0,1)$ distribution is between two threshold points, or is in category k. That is,

$$P(k) = \Phi_k - \Phi_{k-1}.$$

If $k = 1$, then $\Phi_{k-1} = 0$.

15.2 Example Data

Consider calving scores of calves from first lactation heifers in two herds within the same year-season (Table 15.1). There are $m = 3$ categories, i.e., 1 is unassisted calving, 2 is assistance required, and 3 is a very difficult calving. Calving ease could be considered a trait of the calf being born, or it could be considered a trait of the cow that is giving birth. In this example, the trait is observed on the calf. Maternal effects are ignored in this example because each dam has only one calf.

Table 15.1: Calving scores of calves from first lactation heifers

Year Season	Herd	Calf ID	Sire ID	Dam ID	Sex of Calf	Calving Score
1	1	15	1	4	F	1
1	1	16	1	5	F	2
1	1	17	2	6	M	1
1	1	18	3	7	M	2
1	1	19	3	8	M	3
1	2	20	1	9	M	1
1	2	21	2	10	F	2
1	2	22	2	11	F	1
1	2	23	2	12	F	3
1	2	24	3	13	F	1
1	2	25	3	14	M	3

15.3 Linear Model

The most common analysis of categorical data is the use of a linear model on the observed scores, ignoring the fact that categorical data are non-normally distributed. This is the easiest approach because no distribution is assumed.

However, hypothesis tests about fixed effects in the model could be biased because of non-normality of the data. Analysis of category number assumes the 'distances' between categories is the same. For example, to go from unassisted to assisted calvings assumes the same degree of difficulty as to go from assisted to very difficult. Usually the percentages of observations in each category indicate that it is easier to be in the unassisted category and there are typically fewer observations in the assisted and very difficult categories.

Let the model be

$$y_{ijk} = S_i + H_j + a_k + e_{ijk}$$

where

y_{ijk} is an observed score (a number from 1 to m) on calf k, of sex i, in herd j,

S_i is a sex of calf effect,

H_i is a herd-year-season effect (in this example there is only one year-season, but two herds),

a_k is a calf additive genetic effect, and

e_{ijk} is a residual error effect.

Calves are assumed related through sires and dams, and let $\sigma_e^2 = 2\ \sigma_h^2$, and $\sigma_e^2 = 6\ \sigma_a^2$. The usual MME are formed and solved. The solutions for animal additive genetic effects are used to rank animals (negative values are better than positive values because the categories were numbered from easiest to most difficult calvings).

The solution for female calves was 1.66 while for male calves was 1.99, which indicates that male calves had more difficult births than female calves. The difference between herd-year-season 1 and 2 was 0.08 in favour of herd-year-season 1. Thus, there were more difficult births in herd-year-season 2. Below is a table of the additive genetic solutions (Table 15.2), for the three sires and animals 15 and 25.

The sires were ranked in order to their ID with sire 1 having the easiest calvings. These solutions can only be used to rank the animals, but to predict the calving ease of a future progeny is not simple. There would need to be separate predictions for a future male or female calf. For a future female progeny of sire 1, the prediction would be $1.66 - 0.08 = 1.58$, or somewhere between category 1 and category 2.

Table 15.2: Additive genetic values from linear model analysis of category numbers

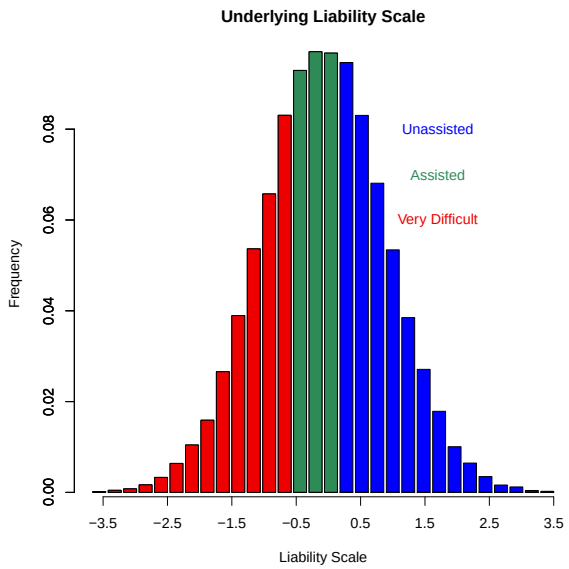
Animal	Solution
1	-0.08
2	-0.00
3	0.09
15	-0.11
25	0.15

15.4 Use of Scores

Snell (1964) proposed changing category numbers to scores ranging from 0 to 100, where the scores were obtained assuming an underlying exponential distribution. The purpose of the scores was to provide homogeneous residual variances over observations and approximately normal distribution of residuals. In the example data, there were 5 observations in category 1 (45.45%), and 3 each in categories 2 (27.27%) and 3(27.27%). The scoring process is known as "normalizing" .

A simple normalizing method (not Snell's) is to determine the averages of the means of liabilities within each category.

Figure 15.2



The mean of values on the x-axis for Unassisted calvings (blue area) was

0.8725, the mean for Assisted calvings (green area) was -0.2379, and the mean for Very Difficult calvings (red area) was -1.215. The distance between means of the categories is not equal, but depends on the percentage of observations in each category and the location of truncation points. The means are used in the linear model analysis rather than the category numbers. The linear model was exactly the same as in the previous section. In this analysis, positive solutions are favourable and negative solutions are not good.

Female calves had a mean liability of 0.16, while male calves had a mean liability of -0.17. The difference in liabilities between herd-year-season 1 and 2 was 0.07, in favour of herd-year-season 1.

Animal genetic values from linear model analysis of normalized scores

Animal	Solution
1	0.09
2	0.00
3	-0.09
15	0.11
25	-0.15

The genetic solutions are nearly identical to the linear model analysis of category numbers, but in this case the solutions are interpretted as differences in the underlying normal scale. Thus, the results could be converted to probabilities. Suppose we wanted to predict the probability of female progeny of sires 1 and 3 to be unassisted. For sire 1, locate $0.16 + 0.09$ on the x-axis of the normal density function and find the probability up to that point, $\Phi(0.25) = 0.60$ or 60%, and for sire 3, locate $0.16 - 0.09$ giving $\Phi(0.07) = 0.53$ or 53%. Thus, sire 1 would be expected to have 7% more unassisted births than sire 3.

15.5 Separate Traits

Quaas and Van Vleck (1980) used a model where each category was a different trait, and then a multiple trait model was applied to the data. Thus, trait i was a binary trait with 1 if the animal belonged to that category, or 0 if it did not. If traits 1 and 2 were 0, then trait 3 had to be a 1. The multiple trait equations would estimate the probability of having progeny in each category. With m categories there would be $m - 1$ traits, with the m^{th} trait determined by subtraction from the other traits.

In the above example, the frequencies of observations in the three categories were 0.4545, 0.2727, and 0.2727. Let $\mathbf{y}_i$ represent the vector of observations

for animal i of length m, where $\mathbf{y}_i$ would have one 1 and $(m-1)$ zeros. The phenotypic covariance matrix of $\mathbf{y}$ over all animals would be

$$Var(\mathbf{y}) = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_m \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \cdots & -\pi_2\pi_m \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_m\pi_1 & -\pi_2\pi_m & \cdots & \pi_m(1-\pi_m) \end{pmatrix},$$

where π_i is the frequency of observations in category i. Numerically, for the example data

$$\mathbf{V} = Var(\mathbf{y}) = \begin{pmatrix} 0.2480 & -0.1240 & -0.1240 \\ -0.1240 & 0.1984 & -0.0744 \\ -0.1240 & -0.0744 & 0.1984 \end{pmatrix}.$$

Note that the rows add to zero, and thus, the matrix is not positive definite. Assuming heritability of each category is the same and equal to 0.10, and herd-year-season effects are 0.3 of the total variance, then

$$\mathbf{G} = \begin{pmatrix} 0.02480 & -0.01240 & -0.01240 \\ -0.01240 & 0.01984 & -0.00744 \\ -0.01240 & -0.00744 & 0.01984 \end{pmatrix},$$

$$\mathbf{H} = \begin{pmatrix} 0.0744 & -0.0372 & -0.0372 \\ -0.0372 & 0.05952 & -0.02232 \\ -0.0372 & -0.02232 & 0.05952 \end{pmatrix},$$

and $\mathbf{R} = \mathbf{V} - \mathbf{G} - \mathbf{H}$,

$$\mathbf{R} = \begin{pmatrix} 0.1488 & -0.0744 & -0.0744 \\ -0.0744 & 0.11904 & -0.04464 \\ -0.0744 & -0.04464 & 0.11904 \end{pmatrix}.$$

Because the covariance matrices are singular, analyze only the first two categories and omit category m from the multiple trait analysis. The model is the same as in the previous two sections. The solutions are given in Table 15.3.

To predict the probability of sire 1 having a female calf in category 1, add 0.49 to 0.04 to get 0.53, or 53%, and for sire 3 it would be 0.44 or 44%, a difference of 9%. For a female progeny in category 3, the corresponding probabilities would be 0.10 for sire 1 and 0.18 for sire 3.

There could be predictions for each category, but in practice you would need to predict one quantity for an average of female and male calves because the sex of the calf would be equal probability male or female.

Table 15.3: Solutions from multiple trait analysis of categories as binary traits

	Cat. 1	Cat. 2	Cat. 3
Female	0.49	0.37	0.14
Male	0.42	0.18	0.40
HYS 1	-0.03	0.10	-0.07
HYS 2	0.03	-0.10	0.07
Sire 1	0.04	-0.00	-0.04
Sire 2	0.00	-0.01	0.01
Sire 3	-0.05	0.01	0.04
Calf 15	0.08	-0.05	-0.03
Calf 25	-0.07	-0.01	0.08

15.6 Threshold Model

In a threshold model, the underlying liabilities are modelled, and these give rise to the phenotypes which are either 1, 2, or 3.

$$\lambda_{ijkl} = f(t)_i + S_j + H_k + a_l + e_{ijkl}$$

where

λ_{ijkl} is the unknown, underlying liability value for calf l, of sex j, in herd k,

$f(t)_i$ is a function of the thresholds and probabilities of the calving score of the calf belonging to category i,

S_j is a sex of calf effect,

H_k is a herd within year-season effect,

a_l is a calf additive genetic effect, and

e_{ijkl} is a residual error effect.

Because the liability values are unknown, a scale has to be imposed on the liabilities. A convention is to set the residual variance to be 1. The heritability of calving scores will be 0.10 in this example, but is often lower, and sometimes greater than 0.10. The herd-year-season variation will be assumed 0.30 of the total. The relationship matrix will be utilized.

Let the model be written in matrix notation as

$$\lambda = \mathbf{Ft} + \mathbf{Xb} + \mathbf{Zu} + \mathbf{e},$$

where

λ is the vector of unobserved liabilities of each animal,

$\mathbf{t}$ is the vector of $m-1$ thresholds,

$\mathbf{b}$ is the vector of fixed effects in the model,

$\mathbf{u}$ is the vector of random effects, including random animal additive genetic effects,

$\mathbf{e}$ is the vector of random residuals, assumed to have mean 0 and variance of 1,

$\mathbf{F}$ is a matrix of probabilities of an animal being in the various categories resulting in a function of the unknown thresholds, and

$\mathbf{X}, \mathbf{Z}$ are the usual design matrices of a linear model.

A non-linear system of equations were derived separately by Harville and Mee (1984) and by Gianola and Foulley (1984) for a sire model application, but which can be extended to an animal model, where the animal is the smallest subclass with only one observation in it. The equations can be written as follows:

$$\begin{pmatrix} \mathbf{Q} & \mathbf{L}'\mathbf{X} & \mathbf{L}'\mathbf{Z} \\ \mathbf{X}'\mathbf{L} & \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{L} & \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \Delta\mathbf{t} \\ \Delta\mathbf{b} \\ \Delta\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{p} \\ \mathbf{X}'\mathbf{v} \\ \mathbf{Z}'\mathbf{v} - \mathbf{G}^{-1}\mathbf{u} \end{pmatrix}.$$

The equations must be solved iteratively. Note that $\Delta\mathbf{b}$, for example, is the change in solutions for $\mathbf{b}$ between iterations. The calculations for $\mathbf{Q}$, $\mathbf{L}$, $\mathbf{W}$, $\mathbf{p}$, and $\mathbf{v}$ need to be described. The values of these matrices and vectors change with each iteration of the non-linear system. The amount of change each iteration decreases to zero.

15.6.1 Calculations

The process is begun by choosing starting values for $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{t}$. Let $\mathbf{b} = \mathbf{0}$ and $\mathbf{u} = \mathbf{0}$, then starting values for $\mathbf{t}$ can be obtained from the data by knowing the fraction of animals in each category. For the example, let the threshold values be $t_1 = 0.3904$, and $t_2 = 0.9563$ for categories arranged from left to right. Category 1 is Unassisted calvings, 2 is Assisted calvings and 3 is very difficult calvings. The categories could be ordered in the opposite direction too, if desired.

The following calculations are performed for each calving score for animals 15 to 25, those for $l = 15$:

1. $f_{li} = t_i - \mathbf{x}_l'\mathbf{b} - \mathbf{z}_l'\mathbf{u}$ for $i = 1$ to $(m-1)$.

$$\begin{aligned} f_{15,1} &= t_1 - S_F - H_1 - a_1 \\ &= .3904 - 0 - 0 - 0 \\ &= .3904, \text{ and} \\ f_{15,2} &= t_2 - S_F - H_1 - a_1 \\ &= 0.9563. \end{aligned}$$

2. For $i = 0$ to m, $\Phi_{li} = \Phi(f_{li})$.

$$\begin{aligned} \Phi_{15,0} &= 0 \\ \Phi_{15,1} &= \Phi(.3904) \\ &= .6519, \\ \Phi_{15,2} &= \Phi(0.9563) = .8305, \text{ and} \\ \Phi_{15,3} &= 1. \end{aligned}$$

3. For $i = 1$ to m, $P_{li} = \Phi_{li} - \Phi_{l(i-1)}$.

$$\begin{aligned} P_{15,1} &= .6519 - 0 = .6519, \\ P_{15,2} &= .8305 - .6519 = .1787, \\ P_{15,3} &= 1 - .8305 = .1695. \end{aligned}$$

4. For $i = 0$ to m, $\phi_{li} = \phi(f_{li})$.

$$\begin{aligned} \phi_{15,0} &= 0.0 \\ \phi_{15,1} &= \phi(.3904) = .36967, \\ \phi_{15,2} &= \phi(.9563) = .25254, \\ \phi_{15,3} &= 0.0. \end{aligned}$$

5. The matrix $\mathbf{W}$ is a diagonal matrix of order equal to the number of observations, N. The elements of $\mathbf{W}$ provide weighting factors for each animal with a record. The j^{th} diagonal is given by

$$w_{jj} = \sum_{k=1}^{m} (\phi_{j(k-1)} - \phi_{jk})^2 / P_{jk}.$$

For the first observation,

$$\begin{aligned} w_{15,15} &= \left(\frac{(\phi_{15,0} - \phi_{15,1})^2}{P_{15,1}} + \frac{(\phi_{15,1} - \phi_{15,2})^2}{P_{15,2}} + \frac{(\phi_{15,2} - \phi_{15,3})^2}{P_{15,3}} \right) \\ &= 1 \left(\frac{(-.36967)^2}{.6519} + \frac{(.36967 - .25254)^2}{.1787} + \frac{(.25254)^2}{.1695} \right) \\ &= .66277. \end{aligned}$$

For all observations, **W** has diagonals equal to 0.66277, in this first iteration with all solutions to sex effects, herd-year-seasons, and animals equal to zero.

6. The vector **v** is used as the phenotypes, which are unknown. For the j^{th} observation,

$$v_j = \sum_{k=1}^{m} n_{jk}(\phi_{j(k-1)} - \phi_{jk})/P_{jk}.$$

Then for animal 15,

$$\begin{aligned} v_{15} &= 1(\phi_{15,0} - \phi_{15,1})/P_{15,1} \\ &\quad +0(\phi_{15,1} - \phi_{15,2})/P_{15,2} \\ &\quad +0(\phi_{15,2} - \phi_{15,3})/P_{15,3}, \\ &= -.5671. \end{aligned}$$

The complete vector **v** is

$$\begin{array}{c} (\ -.5671 \quad .6556 \quad -.5671 \quad .6556 \quad 1.4902 \quad -.5671 \\ .6556 \quad -.5671 \quad 1.4902 \quad -0.5671 \quad 1.4902\ )' \end{array}$$

7. The matrix **L** is of order $N \times (m-1)$ and the jk^{th} element is

$$\ell_{jk} = -\phi_{jk}[(\phi_{jk} - \phi_{j(k-1)})/P_{jk} - (\phi_{j(k+1)} - \phi_{jk})/P_{j(k+1)}].$$

For the example data,

$$\mathbf{L} = \begin{pmatrix} -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \\ -0.4520 & -0.2108 \end{pmatrix}.$$

8. The matrix $\mathbf{Q}$ is a tri-diagonal matrix of order $(m-1) \times (m-1)$, however, this is only noticeable when m is greater than 3. The diagonals of $\mathbf{Q}$ are given by

$$q_{kk} = \sum_{j=1}^{N} \phi_{jk}^2 (P_{jk} + P_{j(k+1)}) / P_{jk} P_{j(k+1)},$$

and the off-diagonals are

$$q_{k(k+1)} = -\sum_{j=1}^{N} \phi_{jk} \phi_{j(k+1)} / P_{j(k+1)}.$$

For the example data,

$$\mathbf{Q} = \begin{pmatrix} 10.7198 & -6.0599 \\ -6.0599 & 8.0664 \end{pmatrix}.$$

9. The elements of vector $\mathbf{p}$ are given by

$$p_k = \sum_{j=1}^{N} \phi_{jk} [\frac{n_{jk}}{P_{jk}} - \frac{n_{j,k+1}}{P_{j,k+1}}],$$

for k equal to the category in which the animal was assigned. Hence,

$$\mathbf{p}' = \begin{pmatrix} -3.3720 & -0.2302 \end{pmatrix}.$$

Put these elements into the MME given earlier along with $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{G}^{-1}$, then

$$\Delta\mathbf{t} = \begin{pmatrix} -0.061146 \\ 0.061146 \end{pmatrix}$$

which are added to the current values of $\mathbf{t}$ giving

$$\mathbf{t} = \begin{pmatrix} 0.329254 \\ 1.017446 \end{pmatrix}$$

$$\Delta\mathbf{b} = \begin{pmatrix} 0.252511585 \\ 0.719326847 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0.252511585 \\ 0.719326847 \end{pmatrix}$$

The new solutions for HYS effects were

$$\mathbf{h} = \begin{pmatrix} -0.03156714 \\ 0.03156714 \end{pmatrix}.$$

There were also new solutions for the animal additive genetic effects, but there are 25 animals, too many to show these early results.

Once the new solutions are available, then the entire process is repeated using the current solutions. The new results for the next iteration are as follows:

$$\mathbf{p} = \begin{pmatrix} -0.5120 \\ 0.8925 \end{pmatrix},$$

$$\mathbf{v} = \begin{pmatrix} -0.6541 \\ 0.4217 \\ -0.9543 \\ -0.0782 \\ 0.9068 \\ -0.9661 \\ 0.3261 \\ -0.7132 \\ 1.2053 \\ -0.7424 \\ 0.8680 \end{pmatrix},$$

$$diag(\mathbf{W}) = (0.7062 \;\; 0.7346 \;\; 0.7715 \;\; 0.7725$$
$$0.7674 \;\; 0.7722 \;\; 0.7503 \;\; 0.7270 \;\; 0.7617$$
$$0.7357 \;\; 0.7624 \;),$$

$$\mathbf{Q} = \begin{pmatrix} 10.054433 & -4.992588 \\ -4.992588 & 9.122229 \end{pmatrix},$$

and

$$\mathbf{L} = \begin{pmatrix} -0.4700668 & -0.2361804 \\ -0.4608371 & -0.2738132 \\ -0.4103202 & -0.3611655 \\ -0.3676418 & -0.4048709 \\ -0.3433935 & -0.4239606 \\ -0.4067816 & -0.3654070 \\ -0.4496203 & -0.3006381 \\ -0.4644020 & -0.2626043 \\ -0.4353492 & -0.3263621 \\ -0.4602620 & -0.2754458 \\ -0.3281838 & -0.4342064 \end{pmatrix}.$$

Construct the MME again, and solve for the changes in the solution vectors. Repeat until the values of the changes are all equal to zero.

After five more iterations, the nearly converged solutions were

$$\mathbf{t} = \begin{pmatrix} 0.2759 \\ 1.0708 \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} 0.1965 \\ 0.6648 \end{pmatrix},$$

$$\mathbf{h} = \begin{pmatrix} -0.0504 \\ 0.0504 \end{pmatrix},$$

and the animal EBVs are in the Table 15.4. Keep in mind that category 1 was Unassisted calvings, so that negative values are good in this analysis.

The EBV can be converted to probabilities. Suppose we want to predict the probability Pr that a female progeny of sire 1 would be in category 1.

$$\begin{aligned} Pr &= \Phi(t_1 - \mathbf{x}'\widehat{\mathbf{b}} - \mathbf{z}'\widehat{\mathbf{a}}) - \Phi(t_0 - \mathbf{x}'\widehat{\mathbf{b}} - \mathbf{z}'\widehat{\mathbf{a}}) \\ &= \Phi(0.2527) \\ &= 0.60 \end{aligned}$$

For sire 3, the same probability would be

$$\begin{aligned} Pr &= \Phi(t_1 - \mathbf{x}'\widehat{\mathbf{b}} - \mathbf{z}'\widehat{\mathbf{a}}) - \Phi(t_0 - \mathbf{x}'\widehat{\mathbf{b}} - \mathbf{z}'\widehat{\mathbf{a}}) \\ &= \Phi(-0.2190) \\ &= 0.41 \end{aligned}$$

Table 15.4: EBV from Threshold Animal Model (TAM) Example

Animal	EBV	Animal	EBV
1	-0.0943	15	-0.1274
2	-0.0030	16	0.0137
3	0.0974	17	-0.1193
4	-0.0535	18	0.0498
5	0.0406	19	0.1721
6	-0.0785	20	-0.1693
7	0.0007	21	0.0438
8	0.0823	22	-0.0918
9	-0.0814	23	0.1567
10	0.0302	24	-0.0452
11	-0.0602	25	0.1641
12	0.1055		
13	-0.0626		
14	0.0770		

Thus, sire 1 would have 19% more Unassisted calvings than sire 3.

15.7 ECP

There can be computational problems with categorical traits. The Extreme Category Problem (EPC) is one case. This problem occurs when most of the observations are in one category. In a herd-year-season, for example, all of the animals may be scored in category 1. Deriving some of the probabilities in the threshold model, when all observations are in one category have led to problems. There are Bayesian methods for using prior information to reduce the ECP problems.

The ECP is more common in disease or binary traits where the frequency of a disease is very low, so that many subclasses have all observations in one category or the other.

15.8 Multiple Traits

Analyses of categorical traits with continuous traits is becoming more common. Simianer and Schaeffer (1989) provide details of an analysis using a threshold model for one binary trait, and a regular mixed model for one continuous trait. This method was applied to disease and yield traits in Norwegian dairy

cattle by Simianer et al. (1991). There have also been analyses of calving difficulty with birth weights, fertility traits with production traits, and many others.

15.9 Comments

As mentioned earlier, most routine applications of genetic evaluation to categorical traits have utilized linear models rather than threshold models. Comparisons of rankings of sires from the two models has shown very high correlations between the two rankings (over 0.99). Given that the threshold model is slightly more difficult to carry out, routine linear model packages are more readily at hand.

The more categories there are, such as 9 or 18 for conformation traits, the more normally distributed are the observations. Threshold models are more suitable for traits with just 3 or 4 categories. Even binary traits are served well by linear models (Meijering and Gianola, 1985).

However, the threshold model provides estimates of heritability on the underlying liability scale, which is normally distributed. These heritability estimates are generally larger than those from a linear model analysis. There are conversion formulas to change heritability from the linear model to what it would likely be from the threshold model, based on category frequencies. In most cases the converted value is the same as the heritability estimate from the threshold model.

Now there are Bayesian methods for threshold models, which are supposedly easier to implement than a threshold model. See Sorensen and Gianola (2002, p. 605).

There are other kinds of discrete data, such as number born, or number of embryos produced. These are count data which have a Poisson distribution. Bayesian methods can be utilized for these data too. Today's animal breeders have to know and understand Bayesian methodology, which means better analyses for non-normally distributed data.

15.10 References

GIANOLA, D. 1982. Theory and analysis of threshold characters. J. Anim. Sci. 54:1079-1096.

GIANOLA, D. , J. L. FOULLEY. 1983. Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15:201.

HARVILLE, D. A. , R. W. MEE. 1984. A mixed model procedure for analyzing ordered categorical data. Biometrics 40:393-408.

MEIJERING, A. , D. GIANOLA. 1985. Linear versus nonlinear methods of sire evaluation for categorical traits: a simulation study. Genet. Sel. Evol. 17:115.

QUAAS, R. L. , L. D. VAN VLECK. 1980. Categorical trait sire evaluation by best linear unbiased prediction of future progeny category frequencies. Biometics 36:117-122.

SIMIANER, H. , L. R. SCHAEFFER. 1989. Estimation of covariance components between one continuous and one binary trait. Genet. Sel. Evol. 21:303-315.

SIMIANER, H. , H. SOLBU, L. R. SCHAEFFER. 1991. Estimated genetic correlations between disease and yield traits in dairy cattle. J. Dairy Sci. 74:4358-4365.

SNELL, E. J. 1964. A scaling procedure for ordered categorical data. Biometrics 20:592.

SORENSEN, D. , D. GIANOLA. 2002. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Chapter 16

Survival

LARRY SCHAEFFER

16.1 Definitions

If a dairy cow was allowed to live a natural life, it could live to be 10 to 12 years of age, on average. Cows have been known to reach 17 years of age or more. However, in most dairy herds, cows only live through about 3 and a half lactations, on average, which corresponds to 5.5 years of age. This implies that owners are deciding when cows will leave the herd, which is a reflection of the productivity and profitability of the cow relative to other cows in the herd, and relative to possible younger replacements. Thus, there are many reasons why cows leave the herd, and the owner is the primary deciding factor, except for involuntary causes. Every cow reaches a point where it is no longer profitable to keep her in the herd.

In a country where herds are governed by a quota system of production, cows may remain longer in the herd, or may quickly disappear if the herd might go over quota in production. Thus, there are outside financial considerations that affect the owners' decisions, and which are totally independent of the cows.

At the same time, dairy producers want cows that will stay in the herd for three or more lactations because that could reduce the cost of raising or buying a replacement animal. Cows should have "longevity" or "stayability". The problem is how to define longevity, how to analyze that definition of longevity, and how much importance to give to this trait in an overall profit index.

Regardless of the definition of longevity, the heritability estimates of the trait range from 0.02 to 0.10. In the past, selection for increased production also increased herdlife, but today herdlife encompasses the functionality of the cow. A

functional cow has good fertility and reproductive ability, has good conformation (feet and legs, udder), and is virtually disease free. Thus, a functional cow is less costly to maintain. See Vollema (1998) and Interbull Bulletin 21 (1999) for different types of survival analyses that have been applied.

16.1.1 Censored Data

A basic description of survival data and terminology is given by Collett (2003). Usually the date of the last test date is the cull date for an animal, and this would be an uncensored record, if true. A cow's record is said to be censored when it has not yet died or been culled. All current, in-milk cows are censored data. When analyzing survival there are three possible situations.

1. Censored data are removed from the analysis,
2. Censored data are included in the analysis, but are NOT properly taken into account, or
3. Censored data are included in the analysis and are properly taken into account.

Cows can change to herds that are not on milk recording, and thus, their actual cull date is not recorded. Thus, reasons for disposal from herds are important to determine if records are censored. The analysis of survival or herdlife should include censored data because dairy producers are most interested in the young progeny tested bulls more than the older proven bulls.

16.1.2 Indirect Herdlife

Indirect herdlife is estimated through a multiple trait analysis of indicator traits that are correlated with herdlife, and the results combined in an index. The usual indicator traits are conformation traits including feet and legs, udder, rump, and stature (Jairath et al. 1998). The indicator traits are usually available long before cows are culled, and therefore, give an early indication of herdlife, although not very accurately. The other definitions that follow would be measures of Direct Herdlife.

16.1.3 Length of Productive Life

The age of the cow at the time it is culled is the observation, measured in days, months, years, or number of completed lactations. These observations are known for animals that have been culled, but do not include censored cows. This trait is not normally distributed and would have most observations clustered around 5.5 years, but there would be a long thin tail above 5.5 years.

For censored cows, a predicted length of productive life is usually made based on probabilities estimated from past data. Thus, if a cow has lived to time t, then the probability that it will live to the next time $t + 1$ is used as the observation.

16.1.4 Stayability

Stayability refers to a fixed time period, such as survival to 60 months of age, yes or no. Sires are ranked on the percentage of daughters that live to 60 months. Or it may be stayability to the completion of first lactation, second lactation, or third. The analyses are single or multiple trait (binary traits) (Jairath et al. 1998).

In Canada, the stayability evaluations have been expanded to a five trait analysis. First lactation is split into survival up to 120 DIM and survival up to 240 DIM, then lactations 2, 3, and 4. The results are combined with Indirect Herdlife evaluations using the MACE procedure described in Chapter 11.

16.1.5 Survival

A non-linear approach is taken where time to failure is modelled. Censored data can be included. A survivor function is derived and from this a hazard function is created, which is influenced by time dependent variables, and time independent variables. This approach will be described in more detail in a later section.

16.1.6 Functional Herdlife

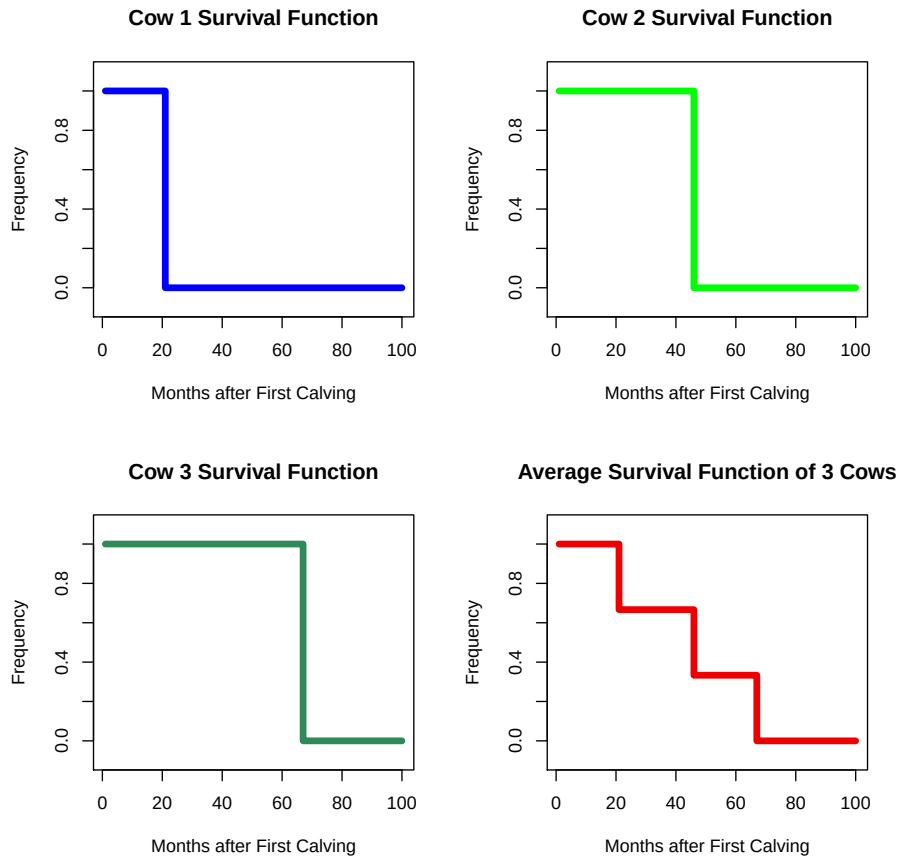
Whatever measure of herdlife is analyzed, there is a need to remove the effects of production on culling. Production is a primary trait because it contributes to the income of a dairy enterprise. What is left are the effects of functional or secondary traits which contribute to the costs of keeping a cow. Thus, Uncorrected Herdlife includes culling on primary AND secondary traits. Functional Herdlife has culling on production removed through the analysis. Functional Herdlife includes culling reasons for conformation, reproduction, and health, and natural longevity.

16.2 Survival Functions

Most cows have been culled by the time they reach 100 months after first calving. A survival function goes from 1 for an animal that is alive to 0 when the animal is dead or culled. The vertical line from 1 to 0 indicates the moment in the

test period when the animal's function changes, i.e. when the animal is removed from production. The survival function for one cow is a one-step function. Figure 16.1 has survival functions for 3 cows, one has died at 20 months after first calving, one at 45 months, and one at 66 months. The fourth graph in Figure 16.1 is the average step function for the three cows combined.

Figure 16.1



As you accumulate more and more cows and average them together, you obtain the survival function for the population, over the years covered by those cows, as in Figure 16.2. This survival function is almost a smooth curve. The values on the curve give the expected probability of an animal being alive in x months after first calving. By the time a cow reaches 100 months, it has a pretty high probability of being culled in the next month.

Because there is a curve over time, it is natural to think of an analysis that involves a random regression model. This was the approach taken by Veerkamp

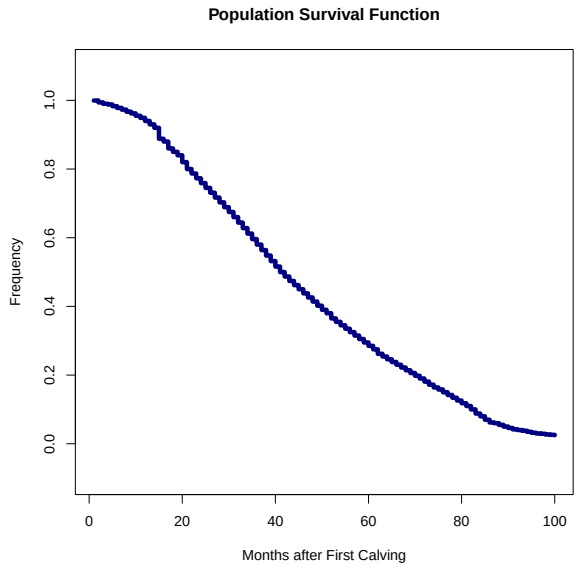
et al. (1999) and Galbraith (2003).

The survival function in Figure 16.2, for this example, is

$$S_t = \frac{n - d_t}{n}$$

where t is the month in which an animal was last alive, n is the total number of live animals that had the opportunity to live for 100 months, and d_t is the number that have died up to and including period t.

Figure 16.2



Eventually d_t comes closer to n. The population survival function can be modelled by a Weibull function,

$$S_t = \exp^{-(\lambda t)^\rho}$$

where $\lambda = 0.01941$ and $\rho = 1.746501$ give the best fit for Figure 16.2.

For an individual cow, the survival function is a one step function that can be represented by a vector, $\mathbf{s}$. Suppose a cow was last alive at 20 months after first calving, then the first 20 elements of $\mathbf{s}$ are 1, and the remaining 80 elements are 0.

For a cow that has lived for 34 months and is still milking in the herd, the first 34 elements of its $\mathbf{s}$ are equal to 1, and the remaining 66 elements are unknown, or not specified yet, because the animal has not had the opportunity to live through those next 66 months. This is a censored survival function.

16.3 Random Regression Analysis

A population survival function is shaped similar to a lactation curve, and so using Legendre polynomials of order 4 (5 covariates) may be appropriate for fitting the general shape. However, because the scale goes from 1 down to 0, at the beginning of the curve many animals are alive, so that the variation in the first months after calving is very small. In general, the variance is the frequency times one minus the frequency, which has the greatest value when the frequency is 0.5. Then the variance becomes smaller again until the end when most animals are dead. Thus, a quadratic shape for the variances seems appropriate. Legendre Polynomials of order 2 (3 covariates) will be used to model the random animal additive genetic, and permanent environmental effects.

Cows calve for the first time in different year-month subclasses, and at different ages and seasons, and therefore, there would be separate fixed curves for year-month of first calving, and for age-seasons of first calving. There could be different survival functions within each herd-year-season of first calving, for HYS as a random factor. To account for production, cows should be divided into production level groups within herd, from low to high. Animal additive genetic effects and animal PE effects would also be modelled with random regressions. The observation for a cow is its survival function. For culled cows, the survival function has 100 data points, and for censored cows, the survival function has less than 100 data points.

16.3.1 Example Data

In Table 16.1 are the culling times of cows in one year-month of first calving and one age-season of first calving, from two HYS. There were just 3 production levels identified across HYS, in this case. Culling times are number of months after first calving.

A model is

$$y_{tijk} = \sum_{m=1}^{5} b_{im} z_{tm} + \sum_{\ell=1}^{3} h_{j\ell} z_{t\ell} + \sum_{n=1}^{3} a_{kn} z_{tn} + \sum_{f=1}^{3} pe_{kf} z_{tf} + e_{tijk}$$

where

y_{tijk} is either 1 or 0, depending on alive or dead at time t, for animal k in HYS j and production level i.

b_{im} are fixed regression coefficients for production level i of order 4, (5 covariates).

$h_{j\ell}$ are random regression coefficients for HYS j of order 2, (3 covariates).

Table 16.1: Example survival data on cows. Production levels are High, Medium, and Low

HYS	Cow	Sire	Dam	Prod. Level	Cull Time
1	17	1	4	L=low	36
1	18	1	5	M=medium	42
1	19	1	6	M	46
1	20	2	7	M	50
1	21	2	8	M	30
1	22	2	9	H=high	60
2	23	1	10	L	20∗
2	24	1	11	M	25
2	25	2	12	M	58
2	26	2	13	H	49∗
2	27	3	14	L	40∗
2	28	3	15	M	51
2	29	3	16	H	68

∗ indicates censored records

a_{kn} are random regression coefficients for animal additive genetic effects of order 2.

pe_{kf} are random regression coefficients for animal permanent environmental effects of order 2.

z_{t-} are Legendre polynomials dependent on time t, and

e_{tijk} are residual effects for each observation, where the variance depends on time t.

Every uncensored animal has 100 data points, and every censored animal has number of observations equal to its censored last month alive. Thus, in total for this example there were 1109 observations, on 13 animals.

The covariance matrices for the random factors are of dimension 3.

$$\mathbf{G} = \begin{pmatrix} 0.246 & -0.172 & 0.332 \\ -0.172 & 0.246 & -0.246 \\ 0.332 & -0.246 & 0.451 \end{pmatrix},$$

$$\mathbf{P} = \begin{pmatrix} 0.197 & -0.138 & 0.266 \\ -0.138 & 0.197 & -0.197 \\ 0.266 & -0.197 & 0.361 \end{pmatrix},$$

and

$$\mathbf{H} = \begin{pmatrix} 0.560 & -0.046 & -0.023 \\ -0.046 & 0.883 & 0.009 \\ -0.023 & 0.009 & 0.044 \end{pmatrix}.$$

The residual variance was allowed to vary from 0.01 at months 1 and 100, to 0.25 in the middle months around 40 to 55 months, based on phenotypic population variances in each month after calving.

There were 147 equations in the MME for this example, 5 for each of three production levels, 3 for each of two HYS, and 3 for each of 29 animal additive genetic effects, and for 13 animal PE effects.

16.3.2 Production Level Solutions

The solutions for the production levels are in Table 16.2.

Table 16.2: Production Level Regression Coefficients

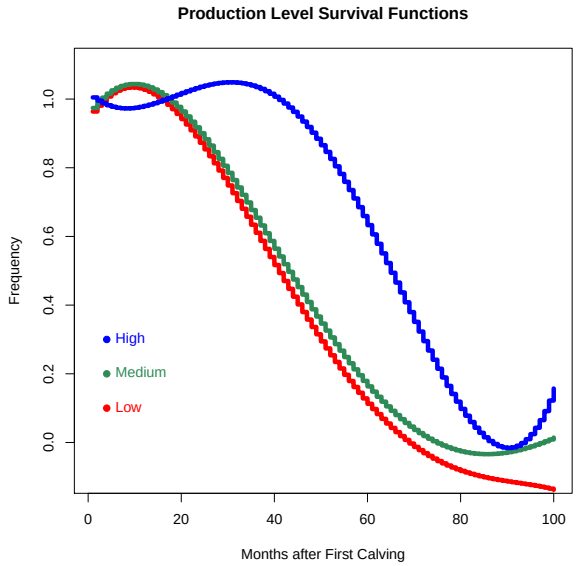
Level	b_0	b_1	b_2	b_3	b_4
Low	0.538	-0.599	0.078	0.098	-0.043
Med.	0.607	-0.563	0.085	0.112	-0.033
High	0.918	-0.524	-0.168	0.117	0.093

The interpretation of the above numbers is difficult to visualize. Thus, it is better to use the regressions to calculate expected frequencies for each month after calving and then plot the production level survival functions, as shown in Figure 16.3.

Note that the production level survival functions go above 1 and below 0, which are outside the natural range. This could be due to the small number of observations and the fact that there were no cows below 20 months or above 68 months, thus the fit on the two ends is not good. The reason could also be that the results on the ends are artifacts of the Legendre polynomial functions. Alternatively, production levels could be modelled by 100 classes per level, if there were enough observations to fill all of the classes.

The results show that the high production level had greater survival than either medium or low. Galbraith (2003) found this to be true in Ayrshire and Jersey cow populations in Canada. Bigger differences in survival functions were found when conformation score levels were used in place of production levels.

Figure 16.3



16.3.3 HYS Solutions

The HYS solutions are in Table 16.3.

Table 16.3: HYS Regression Coefficients

HYS	h_0	h_1	h_2
1	-0.0413	0.0059	0.0079
2	0.0413	-0.0059	-0.0079

Based on the plots in Figure 16.4, HYS 2 had the greater survival function. One could rank HYS on the basis of h_0. Positive h_0 lead to greater survival, but plots should always be made because the sign and magnitude on the other two coefficients could have an impact.

16.3.4 Sire Breeding Values

The solutions just for sires 1, 2, and 3 are given in Table 16.4, and a plot of their survival function is in Figure 16.5. Neither the Table nor the Figure help to determine the better sire.

Figure 16.4

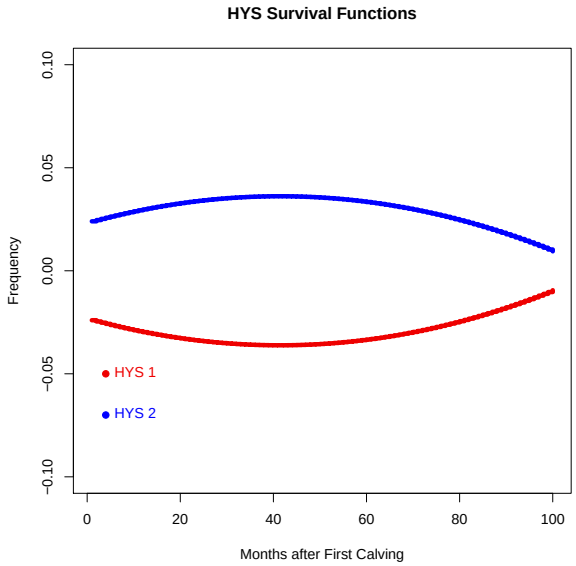


Table 16.4: Sire Regression Coefficients

Animal	a_0	a_1	a_2
1	0.0092	0.0255	0.0226
2	0.0085	0.0039	0.0044
3	-0.0177	-0.0294	-0.0269

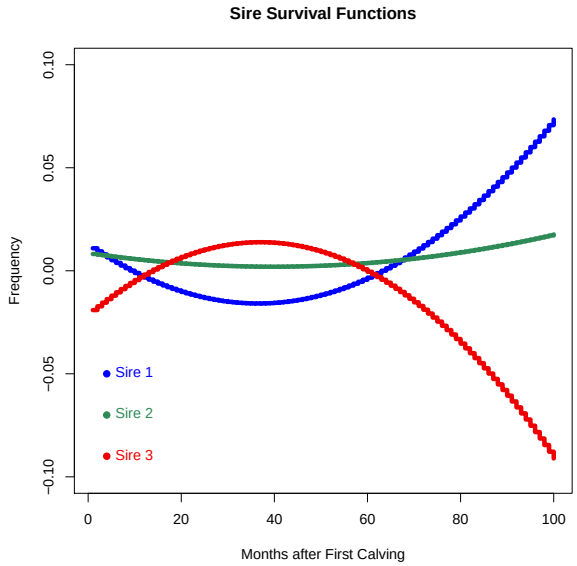
The sire EBV for each month after calving could be added to the population survival function, and then plotted as in Figure 16.6. Now the differences among sires are not very large, at least through 40 to 80 months, but the differences are greater at the end of the 100 month period. This is due to the small number of cows in the analysis, and also due to the heritability of the trait being very low.

To rank sires for survival, one could look at one point, say 50 months after first calving, and compute the value at that time for each sire, from their survival functions (Figure 16.6). Thus sire 1 would have the lowest survival at 50 months of 0.378, then sire 2 at 0.392, and sire 3 at 0.399.

Other options for ranking sires would be

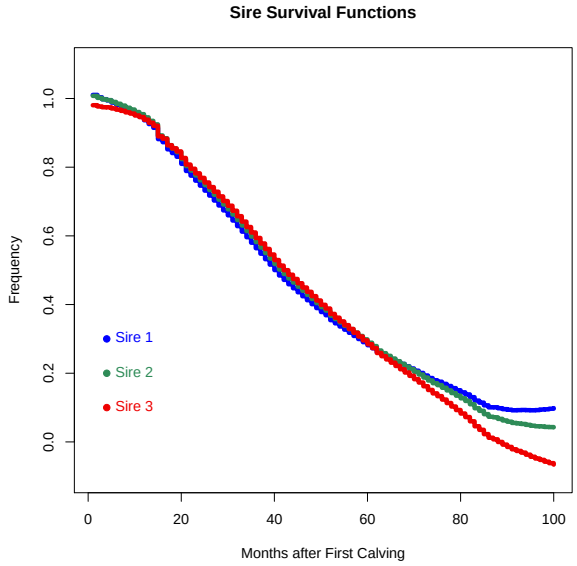
- Use a_0, the intercept parameter, to indicate the overall level. Sire 1 would be the best sire in this case.
- Do a principal component analysis of $\mathbf{G}$ to form functions of the animal

Figure 16.5



solutions for ranking purposes.

Figure 16.6



Cows also have solutions, but these are less reliable than for sires. Predictions of survival for censored cows may be of interest. Obviously predictions for 100 months would be subject to more error, than predictions for 50 months

or earlier. Below (Table 16.5), are the three censored cows in the example, and their predicted survival EBV for 50 months after first calving. Differences in probabilities are minor, but tend to follow their sires' predictions.

Table 16.5: Censored Cow EBVs for 50 months

Cow	Current Age	EBV 50 mo
23	20	0.383
26	49	0.390
27	40	0.402

Solutions for animal PE effects are not shown.

The RRM would be similar to a Multiple Trait (MT) analysis of survival (stayability), if the MT analysis considered each month after calving as a trait (i.e. 100 traits). Thus, the MT analysis is just a much smaller RRM. On the other hand, the RRM for animal additive genetic effects considers just 3 covariates per animal to determine the entire shape of the animal's survival function while the MT analysis looks at a few specific points in time. The RRM uses more information in the analysis, and includes animal permanent environmental effects (which incorporate other correlated effects along the survival curve of each animal).

16.4 Proportional Hazard Model

In 1984, Smith and Quaas applied a failure time analysis to productive lifespan of bull progeny groups, which was further extended by Smith and Allaire (1986). Ducrocq et al. (1988a,b) gave justification for using a Weibull model and how to estimate variances in survival analyses. In 1994, Ducrocq and Solkner presented their software package to analyze survival data, and called it the "Survival Kit". A good explanation of the methods in the programs is given by Kachman (1998).

The survival and hazard functions were described as following a Weibull model (Figure 16.2). The two functions for the population are

$$\begin{aligned} S(t) &= \exp^{-(\lambda t)^{\rho}} \text{ and} \\ h(t) &= \rho\lambda(\lambda\rho)^{\rho-1} \end{aligned}$$

where λ is a scale parameter, and ρ is a shape parameter known as the Weibull

modulus. For Figure 16.2, ρ was found to be equal to 1.7465, which is greater than one and implies that the failure time or mortality is increasing as animals age. If $\rho = 1$, then mortality is random with respect to time and constant, and if $\rho < 1$, then mortality decreases with time, so that most failures are early in life rather than later.

Individual functions are based on risk factors, η_i, which either increase or decrease survival from the population curves for individual i. The risk factors can be modelled as a linear model,

$$\eta_i = \mathbf{x}_i'\mathbf{b} + \mathbf{z}_i'\mathbf{u}$$

which can be a sire or animal model, and where $\mathbf{b}$ are fixed factors affecting survival, like production level, sex of animal, and year-months of calving, and $\mathbf{u}$ are random factors such as herd-year-seasons, and animal additive genetic effects.

The survival function and hazard function can be re-written to include the risk factors as

$$S(t, \eta_i) = \exp^{-t^\rho \exp^{\eta_i}}$$

for $\eta = \rho \ln(\lambda)$, and

$$\begin{aligned} h(t, \eta_i) &= \rho\ t^{\rho-1}\ \exp^{\eta_i} \\ &= h_0(t)\ \exp^{\eta_i} \end{aligned}$$

where $h_0(t)$ is the baseline hazard function where animals have no risks, $\eta = 0$.

Using a parametric approach, the joint likelihood of survival times and η_i is maximized. Assuming that ρ is known, the computational steps are

1. With current estimates of $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{u}}$, for each animal, calculate

$$\eta_i = \mathbf{x}_i'\widehat{\mathbf{b}} + \mathbf{z}_i'\widehat{\mathbf{u}}.$$

2. Let

$$\begin{aligned} q_i &= \exp^{\rho \ln(t)} \\ r_{ii} &= q_i \exp^{\eta_i} \\ Y_i &= w_i - q_i \exp^{\eta_i} + r_{ii}\eta_i \end{aligned}$$

where $w_i = 1$ if the record is uncensored, and $w_i = 0$ if the record is censored.

3. Obtain new estimates of $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{u}}$ from mixed model-like equations,

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}\mathbf{X} & \mathbf{X}'\mathbf{R}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}\mathbf{X} & \mathbf{Z}\mathbf{R}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix},$$

where $\mathbf{X}$ contains the vectors $\mathbf{x}_i'$, $\mathbf{Z}$ contains the vectors $\mathbf{z}_i'$, $\mathbf{R}$ is a diagonal matrix with diagonals equal to r_{ii}, and $\mathbf{y}$ is a vector of Y_i.

The steps are iterated until convergence is reached.

The example data were analyzed, where the risk factors were modelled as

$$\eta_i = b_j + h_k + a_i,$$

for

b_j is the fixed level of production effect (one of three groups),

h_k is a random herd-year-season effect (one of two), and

a_i is a random animal additive genetic effect.

The solutions, upon convergence, were

$$\widehat{\mathbf{b}}' = \begin{pmatrix} -7.1943 & -6.6202 & -7.5019 \end{pmatrix},$$

for the production level effects,

$$\widehat{\mathbf{h}}' = \begin{pmatrix} 0.1153 & -0.1153 \end{pmatrix},$$

for the herd-year-season effects, and animal additive genetic solutions are in Table 16.6.

The solutions can be manipulated in different ways. Suppose we want to predict the percentage of live daughters of the sires at 60 months after first calving in low production herds and an average herd-year-season. For sire 1,

$$\begin{aligned} \eta_1 &= \widehat{b}_1 + \widehat{a}_1 \\ &= -7.1943 + 0.0445 = -7.1498 \\ S(60, \eta_1) &= \exp^{-t^\rho \exp^{\eta_1}} \\ &= 0.3675 \end{aligned}$$

For sires 2 and 3, the results were 0.3938 and 0.3902, respectively. The random regression results (for month 50) were 0.378, 0.392, and 0.399, for sires 1, 2, and 3, respectively.

Table 16.6: Animal Genetic solutions from Survival Analysis

Animal	$\widehat{a}_i$	Animal	$\widehat{a}_i$
1	0.0445	17	0.0664
2	-0.0273	18	0.0187
3	-0.0172	19	0.0051
4	0.0294	20	-0.0412
5	-0.0024	21	0.0213
6	-0.0114	22	0.0036
7	-0.0184	23	0.0117
8	0.0233	24	0.0761
9	0.0115	25	-0.0441
10	-0.0070	26	-0.0486
11	0.0359	27	-0.0422
12	-0.0203	28	-0.0186
13	-0.0233	29	0.0091
14	-0.0224		
15	-0.0066		
16	0.0118		

In the random regression model, sires could rank differently at month 20 compared to month 80, but in the PH model, sires would rank the same at any month from 1 to 100 months after first calving. This is because there is only one parameter per animal being estimated, which applies to the entire time scale.

16.5 Comments

The model for the risk factors could be more complex.

Ducrocq and Solkner (1998) included the following fixed and random factors:

- year-season of first calving effects with two seasons per year,
- lactation number and days in milk within lactation, with 6 lactations and days in milk changes at days 30, 60, 150, and 240, and date dried off,
- ten classes of production levels within herds for milk yield, and five for fat and protein contents,
- herd size classes (4) and herd size variation changes (5), to know if herds are increasing or decreasing in size, culling would be greater in herds that are decreasing in size,

- herd-year-season random effects, with two seasons per year, and
- animal additive genetic effects.

In Austria, random herd-year-season effects were used due to small herd sizes, and also variation in herd size was omitted, but replaced with the average age of herdmates (9 groups). Cows sold for dairy purposes were treated as censored records. In France, the estimates of heritability from the survival analysis were 0.16 to 0.22 depending on breed.

The Survival Kit is used extensively around the world by many countries to evaluate length of productive life. The random regression model may be more easily applied than the survival kit, and allows animals to re-rank over time. A comparison was given by Jamrozik et al. (2008).

16.6 References

COLLETT, D. 2003. *Modelling Survival Data in Medical Research.* Chapman & Hall/CRC, London. Second Edition.

DUCROCQ, V. , R. L. QUAAS, E. J. POLLAK, G. CASELLA. 1988. Length of productive life of dairy cows. 1. Justification of a Weibull model. J. Dairy Sci. 71:3061-3070.

DUCROCQ, V. , R. L. QUAAS, E. J. POLLAK, G. CASELLA. 1988. Length of productive life of dairy cows. 2. Variance component estimation and sire evaluation. J. Dairy Sci. 71:3071-3079.

DUCROCQ, V. , J. SOLKNER. 1994. The Survival Kit, a Fortran package for the analysis of survival data. In Proc. 5^{th} World Cong. on Genet. Appl. To Livest. Prod.

DUCROCQ, V. , J. SOLKNER. 1998. Implementation of a routine breeding value evaluation for longevity of dairy cows using survival analysis techniques. In Proc. 6^{th} World Cong. on Genet. Appl. To Livest. Prod. p. 359-362.

GALBRAITH, F. 2003. Random regression models to evaluate sires for daughter survival. Master's Thesis, University of Guelph, Ontario, Canada, August.

INTERBULL Bulletin. 1999. Proceedings of International Workshop on EU Concerted Action: Genetic Improvement of Funtional Traits in Cattle. Jouy-en-Josas, France. Bulletin No. 21.

JAIRATH, L. , J. C. M. DEKKERS, L. R. SCHAEFFER, Z. LIU, E. B. BURNSIDE, B. KOLSTAD. 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81:550-562.

JAMROZIK, J. , J. FATEHI, L. R. SCHAEFFER. 2008. Comparison of models for genetic evaluation of survival traits in dairy cattle: a simulation study. J. Anim. Breed. Genet. 125:75-83.

KACHMAN, S. D. 1998. Applications in survival analysis. Presented at 1998 ADSA/ASAS meeting in Denver, Colorado. J. Anim. Sci. 77:Suppl. 2.

SMITH, S. P. , and R. L. QUAAS. 1984. Productive Lifespan of Bull Progeny Groups: Failure Time Analysis. J. Dairy Sci. 67:2999-3007.

SMITH, S. P. , and F. R. ALLAIRE. 1986. Analysis of failure times measured on dairy cows: theoretical considerations in animal breeding. J. Dairy Sci. 69:217-227.

VEERKAMP, R. F. , S. BROTHERSTONE, T. H. E. MEUWISSEN. 1999. Survival analysis using random regression models. INTERBULL Bulletin 21:36-40.

VOLLEMA, A. R. 1998. Selection for longevity in dairy cattle. Doctoral thesis, Animal Breeding and Genetics Group, Wageningen Agricultural University, The Netherlands.

Part III

YEARS 2001-present

Chapter 17

Genomics Era

LARRY SCHAEFFER
RAPHAEL MRODE

17.1 Infinitesimal Model

All of the previous chapters, and all of the years of genetic evaluation of dairy cattle have assumed that the *Infinitesimal Model* holds true. This model was described by Fisher (1918). The assumptions of this model are that

- there are an infinite number of genes (loci) that influence milk production, and
- each of those loci contribute an infinitesimally small amount to the overall total genetic merit of an animal.

Genetic evaluation methods were designed to estimate the genetic merit due to the sum total of the effects of the infinite number of loci. One reason for using this model is that science could not yet determine the correct number of loci that were involved, and whether each loci had the same size of effect.

With the discovery of Deoxyribonucleic Acid by Watson and Crick (1953) and the identification of four base pairs making up DNA, scientists have been trying to find loci with large effects (major genes) and trying to determine the actual number of loci. In the 1970's molecular geneticists promised the discovery of some major genes, such that when found there would be a genetic test to discover the genotype of individuals for that gene. That is, which combination of alleles did an individual have. Animals with favourable genotypes would be kept for breeding and others would be culled. There would be no need to calculate genetic evaluations any longer.

For many years the molecular geneticists used markers, which were fairly large segments of DNA (100 to 1000 base pairs in length), but the markers were not very close to the actual genes. Sometimes the marker genotypes were not consistently determined, and so selection on marker genotypes was not very successful. Markers were generated by finding enzymes that cut DNA at specific locations. Microsatellites were one of the smaller segments of DNA that could be used as a marker. However, there were not a lot of microsatellites to be found.

The success of a Marker Assisted Selection (MAS) scheme depended on how close the marker was to the actual gene locus. The term Linkage Disequilibrium (LD) was used to indicate the usefulness of a marker. High LD meant that the marker was close to the gene. To be useful, markers needed an LD of 30% or more. High LD means that an allele of the marker is on the same stretch of DNA as the favourable allele of the gene. Recombinations between the marker allele and the gene allele would be few over hundreds of thousands of meiosis events.

17.2 Single Nucleotide Polymorphisms

The Human Genome Project (HGP) (1984) was started with the goal of sequencing all of the DNA, which was thought to be 5 billion base pairs in length. The first rough draft of the sequence was available in June 2000, and the final version was published in April 2003. Scientists knew the base pair sequences that indicated the start of a gene, and the estimate of the number of genes dropped continually during the project and finally settled at around 25,000 genes. Thus, the number of genes was not infinite and could be counted. The genes, however, only accounted for 5% of the genome or less. What was the purpose of the other 95% of the genome? At the time it was called "junk" DNA, but now scientists are finding that this DNA does have a purpose.

During the project scientists discovered Single Nucleotide Polymorphism(s) (SNP) by comparing DNA sequences of different individuals. A SNP was where a single base pair difference was detected for one individual compared to the majority of individuals. SNPs were found to be everywhere in the genome, millions of them. By 2004 there was a SNP panel developed such that a dairy bull or cow could be genotyped for 5000 SNP at one time. The SNP that were chosen tried to cover the entire genome evenly. It was not known if those SNP were close to any genes, and indeed, the location of some of the 5000 SNPs in the genome was not known. The gaps between any two SNPs could have contained hundreds of genes. So the goal was to find more SNP and to develop panels that could test a million SNP at one time, and to ensure that the SNP were evenly spaced throughout the genome.

Dairy bulls in Canada were initially genotyped with a 5000 SNP panel,

known as the 5K chip, and the EBV of the bulls for many traits were collected. In 2004, the cost of genotyping a bull was over $300 US, and to genotype 400 bulls cost more than $120,000. First, DNA samples had to be collected from the bulls and sent to a lab for processing. The second step was to extract the DNA from the sample, and then to run it on the 5K chip. A reader then scanned the chip results and based on the colour of the reaction, the genotype for each SNP could be determined. The results were then returned to the person who requested the genotyping. All of this took 1 to 3 months.

After receiving the genotype results, (5000 pieces of information per bull), the genotypes had to be verified, which meant comparing son genotypes to their sires' genotypes to detect inconsistencies, or impossible genotypes. Sometimes genotypes could not be determined by the scanner because the reaction of the DNA on the chip was not conclusive. Also, checks were made to determine the frequencies of the genotypes for each SNP. In some cases the genotypes for a SNP were identical for all sampled bulls, and therefore, that SNP marker was not informative.

The locations of the SNPs had to be known, and therefore, comparisons to available maps had to be made. The order of some SNPs was not confirmed, and some SNPs were allocated to the wrong chromosomes. In the end, there were usually much fewer than 5000 SNP genotypes available.

17.3 Example Data

Below, in the Table 17.1, are records on fat yields of cows in one herd-year-season, with 4 different age groups within first lactation.

Table 17.1: Fat yield data on first lactation cows from one herd-year-season

Animal	Sire	Dam	Animal	Sire	Dam	Age	Fat Yield
1	0	0	11	9	7	4	296
2	0	0	12	9	2	1	357
3	0	0	13	8	3	4	387
4	0	0	14	1	4	1	303
5	0	0	15	1	6	2	363
6	0	0	16	1	10	4	301
7	0	0	17	8	4	1	297
8	0	0	18	8	5	4	338
9	0	0	19	8	3	3	376
10	0	0	20	1	6	2	318

The relationship matrix can be partitioned into

$$\mathbf{A} = \begin{pmatrix} \mathbf{I} & \mathbf{A}_{12} \\ \mathbf{A'}_{12} & \mathbf{A}_{22} \end{pmatrix},$$

where

$$\mathbf{A}_{12} = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{A}_{22} = \frac{1}{4}\begin{pmatrix} 4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 4 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 4 & 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 1 & 4 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 4 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 1 & 1 & 4 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

Note that the relationships are either one-half or one-quarter, and that none of the cows are inbred.

A typical animal model with age groups and animal additive genetic effects was applied to the data. A variance ratio for residual to additive genetic variances of 3.08 was used. The EBV from the analysis are given in the Table 17.2.

Now animals 7 through 14 have been genotyped with a 50K SNP chip. Animals 8 and 9 are sires, and the others are females from the herd. In the Table 17.3, are the results for 10 SNP. Typically there are more SNPs than number of genotyped animals.

Table 17.2: EBV for animals in Example data

Animal	EBV	Animal	EBV
1	-5.10	11	-6.49
2	4.83	12	7.44
3	7.15	13	13.08
4	-4.83	14	-6.62
5	0.75	15	0.60
6	0.00	16	-7.71
7	-4.45	17	-3.24
8	4.72	18	3.48
9	0.38	19	5.93
10	-3.44	20	-5.69

Table 17.3: SNP genotypes for 10 markers: 1=AA, 2=Aa, 3=aa genotypes

Animal	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
7	1	2	2	1	2	3	2	2	2	1
8	2	2	2	3	2	3	1	2	1	1
9	1	1	2	3	2	2	2	2	1	1
10	3	3	3	1	2	2	2	2	2	2
11	1	2	2	2	2	3	3	2	2	1
12	1	1	2	3	3	2	3	2	1	2
13	3	2	2	3	2	3	1	2	2	1
14	1	2	2	3	2	1	1	1	1	2

17.4 Association Studies

An association study is where SNP markers are used, one marker at a time, to determine its effect on the trait of interest. To illustrate we will use the small example of the previous section. The observations are the EBV of the genotyped animals, and the assumption is that these EBV have high accuracy (0.99). In practice, bull EBVs based on many thousands of daughters are used as the observations. We will assume the EBV in this example are highly accurate, just to illustrate the methods.

Markers are examined before the regression models are applied. The allele frequencies are estimated for each marker and if the frequency of the less frequent allele (minor allele frequency) is less than 0.10, then that marker may not be

analyzed. Markers 8, 9, and 10 did not have any genotype 3 animals in the example for instance.

The EBV are regressed on the marker genotypes, in this case marker 1.

$$\begin{pmatrix} -4.45 \\ 4.72 \\ 0.38 \\ -3.44 \\ -6.49 \\ 7.44 \\ 13.08 \\ -6.62 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 \\ 2 \\ 1 \\ 3 \\ 1 \\ 1 \\ 3 \\ 1 \end{pmatrix} b + \mathbf{e}$$

The least squares result is $\widehat{\mu} = -5.26$, and $\widehat{b} = 3.59$. The estimate of the residual variance was 47.98, and the resulting F-value to test the significance of $\widehat{b}$ was 1.58. The F-value had to be above 5.99 to be significant at the 0.05 level.

The results for all 10 markers are given in the next Table (17.4).

Table 17.4: Association Tests of 10 SNP markers

Marker	$\widehat{b}$	σ_e^2	F-stat	Sign.
1	3.59	47.98	1.58	0.26
2	-3.72	54.01	0.74	0.43
3	-4.59	57.55	0.32	0.60
4	4.28	42.66	2.53	0.17
5	7.84	51.65	1.04	0.35
6	3.03	54.69	0.65	0.46
7	-1.98	57.44	0.33	0.59
8	8.22	50.76	1.17	0.33
9	-1.80	59.54	0.11	0.76
10	-2.32	58.94	0.17	0.70

Marker 4 was the most significant of the ten examined, but remember there were 50K SNP genotypes on each animal, all of which need to be tested. The lack of significance is due to the inaccurate EBV that were used as well as the small number of animals.

Because so many markers are being tested at one time, and because EBV for several traits may be analyzed, the significance tests must use higher critical values to be conservative in reporting significant results. With 50,000 SNP, for example, and a 0.05 significance level, 2500 SNP could be significant simply due

to chance. The Bonferoni correction has been widely used.

After significant markers are discovered, then one must check if those markers are near to possible major genes that could have biological significance on the trait of interest. If yes, then more studies using additional SNPs close to that gene could be used to find out more about the gene and its functions. If the marker is highly significant, then a test for that marker alone could be developed and offered to producers, so that they may increase the frequency of the favourable allele for that marker in their herds.

One such important gene has been found in dairy cattle, known as the DGAT1 gene located on chromosome 14. In New Zealand, this gene was found to have a 6 kg effect on milk fat yields, with a corresponding decrease in milk protein and milk volume (Spelman, 2002). This gene was patented, but later found that the majority of cattle have the favourable allele, such that testing and selecting for it is not of great significance.

There have been relatively few genes with large effects found in dairy cattle that could be exploited by MAS. The infinitesimal model seems to have been a very realistic model for dairy cattle genetic evaluations.

17.5 Genome Wide Selection

In 2001, Meuwissen et al. showed through simulation that if you had thousands of SNP markers spread evenly through the genome then you could simultaneously estimate the small effects of each SNP genotype on the overall trait. Thus, if you genotyped animals with the 10K chip, then you would estimate genetic effects for those 10,000 SNP from EBV of a group of genotyped bulls. Then you could apply those estimates to a new group of unproven bulls that are genotyped, but which do not have any progeny with records, and obtain an estimate of their BV. They showed that the accuracy of the Genomic Estimated Breeding Value (gEBV) could be as high as 0.81 (in their simulations). That accuracy is much greater than the accuracy of a simple parent average prediction of that bull's EBV, which is usually less than 0.40. Also, you could genotype the young bull at birth and obtain a gEBV immediately, and thereby save 5 years of progeny testing a bull that may not be suitable. Schaeffer (2006) showed how this strategy could double genetic progress in dairy cattle.

In reality, of course, the cost of genotyping enough bulls was great. The first studies in Canada involved less than 500 proven bulls. The bulls had to be split into two groups. One group was used to estimate the effects of 10K SNP, and the other group was used to validate the accuracy of predictions (gEBV). Instead of achieving an accuracy of 0.81 as in Meuwissen et al., (2001), the validation accuracy was 0.50 to 0.60 The gain in genetic progress, however, was not com-

ing from increased accuracy of the gEBV over the parent average, but from the fact that selection decisions on young bulls were being made 5 years earlier than current progeny test schemes. Genetic progress was due to significantly reduced generation intervals.

The USDA started a study around 2006 where several thousand dairy bulls of various breeds and from different countries were to be genotyped. By that time, everyone was using a 10K chip. The project involved developing a new 50K chip. Over 600,000 SNP were screened in order to pick 50K that were suitable.

Also, by this time, the cost of genotyping was coming down to about $100 US per animal for 50K chips. More than 3000 bulls had been genotyped when studies began and shortly thereafter, gEBV were calculated in the United States and Canada and made official.

17.5.1 Least Squares Estimation of SNP Effects

As in Meuwissen et al. (2001), and similarly in Xu (2001), one could estimate all of the SNP effects simultaneously using least squares equations. The model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Sb} + \mathbf{e},$$

where

$\mathbf{y}$ is the vector of EBV from daughters on animals that have been genotyped,

μ is an overall mean effect,

$\mathbf{b}$ is a vector of fixed SNP effects to be estimated where $\mathbf{S}$ is an N by m matrix, with N equal to the number of genotyped animals, and m equal to the number of SNP markers, and the elements are equal to -1, 0, and 1 (genotypes minus 2), so that -1 is genotype AA, 0 is Aa, and 1 is aa,

$\mathbf{e}$ is the residual error.

The residual variance is assumed to be $\mathbf{I}\sigma_e^2$ because the EBV in $\mathbf{y}$ are assumed to have very high accuracy.

If EBV vary in accuracy, then the residual variances could be varied to reflect the different accuracies. Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{S} \end{pmatrix}.$$

One problem with a least squares approach is that m is usually much greater than N, more SNPs than genotyped animals. Hence the LS equations are less than full rank and can not be inverted. The solutions are also not unique.

Using ten SNPs and eight genotyped animals from the example data, the LS equations are

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 8 & -3 & -1 & 1 & 3 & 1 & 3 & -1 & -1 & -4 & -5 \\ -3 & 7 & 3 & 1 & -2 & -1 & 0 & -2 & 1 & 3 & 2 \\ -1 & 3 & 3 & 1 & -3 & -1 & 0 & -1 & 0 & 2 & 1 \\ 1 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & -2 & -3 & -1 & 7 & 1 & 0 & -2 & -1 & -4 & -2 \\ 1 & -1 & -1 & 0 & 1 & 1 & 0 & 1 & 0 & -1 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 1 & 0 & -4 \\ -1 & -2 & -1 & 0 & -2 & 1 & 0 & 5 & 1 & 1 & 1 \\ -1 & 1 & 0 & 0 & -1 & 0 & 1 & 1 & 1 & 1 & 0 \\ -4 & 3 & 2 & 0 & -4 & -1 & 0 & 1 & 1 & 4 & 2 \\ -5 & 2 & 1 & 0 & -2 & 0 & -4 & 1 & 0 & 2 & 5 \end{pmatrix},$$

and

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 4.61 \\ 19.38 \\ -11.26 \\ -3.44 \\ 26.89 \\ 7.44 \\ 13.47 \\ -10.23 \\ 6.62 \\ -5.92 \\ -7.23 \end{pmatrix}.$$

A Moore-Penrose inverse of the coefficient matrix was used to obtain a solution (because the order of the equations was small in this case) in Table 17.5.

Note that the solutions in Table 17.5, do not agree with the solutions from the association study where each SNP was analyzed separately. In fact, the SNP solutions in Table 17.5 may have no relationship at all to the actual size of SNP effects in reality. The SNP genotypes are merely covariates that may or may not help to explain variation in EBVs.

Because there were more SNPs than data, the above model fits the data perfectly. That is, $\mathbf{y} = \mathbf{1}\widehat{\mu} + \mathbf{S}\widehat{\mathbf{b}}$. The validation test is where the above solutions are applied to animals not included in the analysis.

Suppose animal 20 was genotyped, shown in Table 17.6, then multiplying its genotype times the SNP effect solutions and adding up the results and the overall mean, the genomic EBV (gEBV) for animal 20 would be -9.26. This can

Table 17.5: SNP effect solutions and overall mean

	Solution
mean	1.929518
Marker 1	5.864284
2	-7.635916
3	-2.344765
4	1.255575
5	8.271216
6	2.817283
7	-3.292362
8	1.917075
9	2.497527
10	2.079857

be compared to its actual EBV from the animal model analysis of -6.62.

Table 17.6: SNP effect solutions and overall mean

	Solution	Animal 20 Genotypes	Accumulation
mean	1.929518		1.929518
Marker 1	5.864284	-1	-3.934766
2	-7.635916	1	-11.570682
3	-2.344765	0	-11.570682
4	1.255575	1	-10.315107
5	8.271216	0	-10.315107
6	2.817283	-1	-13.132390
7	-3.292362	-1	-9.840028
8	1.917075	-1	-11.757103
9	2.497527	1	-9.259576
10	2.079857	0	-9.259576

The validation data set would need to include a large number of animals that have accurate EBV and genotypes, but which were not included in the estimation of the SNP effects. The correlation of the gEBV and the animal's actual EBV would then be an indication of the accuracy of the gEBV.

17.5.2 Using BLUP

Meuwissen et al. (2001) also considered a BLUP analysis using a variance ratio added to the diagonals of the SNP equations in the LS system. Suppose we add 1 to the diagonals. This would remove the dependencies among the columns of **S**. An assumption is that the SNPs are members of a population of SNP effects with an overall mean of zero and variance equal to that of the residual variance. Obviously, the correct variance would need to be estimated. The results using a ratio of 1 are given in the next two tables.

Another analysis was performed using a variance ratio of 10, and those results are also given in the next two Tables (17.7, 17.8).

Table 17.7: BLUP Estimates of SNP Effects

	Solution ratio=1	Solution ratio=10
μ	0.0787	-0.0634
Marker 1	4.2626	1.4627
2	-3.1732	-0.7886
3	-0.8258	-0.2317
4	2.9590	1.5048
5	3.2227	0.6289
6	2.2377	0.8251
7	-0.7794	-0.4013
8	1.8796	0.5752
9	0.1762	-0.1536
10	0.5173	-0.1779

Using these estimates to predict animal 20, the predictions were -7.56 for a ratio equal to 1, and -1.96 for a ratio equal to 10. The smaller ratio seems to produce gEBV that agree better with the animal model EBV than the ratio of 10.

17.5.3 BLUP with Unequal Ratios

Meuwissen et al. (2001) also proposed allowing the ratio for each SNP to differ depending on the magnitude of the SNP effects (squared). The larger is the SNP effect, the larger would be the variance for that SNP, and the smaller would be the ratio. They described a Bayesian method of estimating the correct ratio. Subsequently there have been several revised proposals from various researchers.

Table 17.8: BLUP gEBV of genotyped animals

Animal	gEBV ratio=1	gEBV ratio=10
7	-5.42	-2.03
8	5.36	3.00
9	1.25	1.10
10	-2.62	-1.13
11	-3.24	-0.92
12	4.22	1.15
13	9.80	4.31
14	-4.74	-0.87

In Table 17.9, the LS SNP effect estimates were squared and then divided into the variance of the EBV, which was 52. While not the best procedure, these ratios were used on the diagonals of the SNP equations.

Table 17.9: SNP effect solutions and overall mean from LS analysis

Marker	Solution	Squared	Ratio
1	5.864284	34.39	1.51
2	-7.635916	58.31	0.89
3	-2.344765	5.50	9.45
4	1.255575	1.58	32.91
5	8.271216	68.41	0.76
6	2.817283	7.94	6.55
7	-3.292362	10.84	4.80
8	1.917075	3.68	14.13
9	2.497527	6.24	8.33
10	2.079857	4.33	12.01

The results from the LS, BLUP with ratio equal to 1, and BLUP with variable ratios are shown in Table 17.10 for comparisons. Using a different ratio for each SNP based on the magnitude of its effect made many of the SNP effect estimates much smaller than in the LS analysis, and a few (SNPs 1, 2, and 5) still had relatively large effects. Gianola (2010) questioned whether SNPs should be considered as random effects, and what kind of distribution SNP effects follow. Ideally, there should be more animals genotyped than there are SNP markers to estimate, but this will not happen for a long time. We already have many times

more genotyped than SNP markers, albeit females.

Table 17.10: SNP effect solutions and overall mean

	LS	BLUP ratio=1	BLUP varied
mean	1.93	0.08	0.49
Marker 1	5.86	4.26	4.71
2	-7.64	-3.17	-5.18
3	-2.34	-0.83	-0.30
4	1.26	2.96	0.32
5	8.27	3.22	3.94
6	2.82	2.23	0.95
7	-3.29	-0.78	-0.89
8	1.92	1.88	0.18
9	2.50	0.18	-0.11
10	2.08	0.52	-0.20

The gEBV from the BLUP analysis with variable ratios for the SNP effects are shown in the Table 17.11. A prediction for animal 20 was -9.45.

Table 17.11: gEBV for genotyped animals from BLUP with variable variance ratios

Animal	gEBV
7	-3.38
8	2.96
9	1.60
10	-0.60
11	-3.95
12	4.45
13	7.56
14	-4.04

A problem with the direct approach of estimating SNP effects becomes obvious when there are 600,000 SNP effects to estimate at one time and around 100,000 animals with phenotypes or EBVs. Solving for 600,000 effects simultaneously could give too many dependencies among the SNP markers. If BLUP with

variable variance ratios was used, there would be many variances to estimate, and this could take a long time. With so much information available, one begins to ask if all of the SNPs are necessary for accurate gEBV.

17.5.4 Relationships Among Animals

The matrix $\mathbf{S}$ was defined earlier to be a matrix with N rows and m columns containing the SNP genotypes (expressed as -1, 0, and 1). Early on it was noted that a matrix of relationships based upon the SNP genotypes could be created, as

$$\mathbf{G} \;=\; \mathbf{SS}'/(\sum 2p_iq_i),$$

where p_i is the frequency of one allele at marker i, and $q_i = (1 - p_i)$. Thus, in an animal model $\mathbf{A}$ could be replaced by $\mathbf{G}$. There were two immediate drawbacks.

1. Not all animals with phenotypes have been genotyped, and therefore, $\mathbf{G}$ could not be calculated for all animals with data.

2. The matrix $\mathbf{G}$ did not have an easy inverse like Henderson's method for $\mathbf{A}^{-1}$.

The procedure became a Two-Step procedure. First step, use the EBVs on the genotyped animals, and construct the relationships among them using the SNP genotypes.

$$Var(\mathbf{a}) \;=\; \mathbf{G}\sigma_a^2$$

Construct MME using the model

$$\mathbf{y} \;=\; \mathbf{1}\mu \;+\; \mathbf{Ia} \;+\; \mathbf{e},$$

where

$\mathbf{y}$ is the vector of highly accurate EBV on the genotyped animals,

μ is the overall mean,

$\mathbf{a}$ are the animal additive genetic values, and

$\mathbf{e}$ is the vector of residual effects.

The MME are

$$\begin{pmatrix} \mathbf{1'1} & \mathbf{1'} \\ \mathbf{1} & \mathbf{I} + \mathbf{G}^{-1}k \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{1'y} \\ \mathbf{y} \end{pmatrix},$$

and k is the residual to additive genetic variances, which will be much smaller than the same ratio when $\mathbf{y}$ is a vector of phenotypes (i.e. single records) rather than EBVs. The solution to the MME give the genomic EBV, gEBV directly.

For the example data and genotyped animals,

$$\mathbf{G} = \frac{256}{1134}\begin{pmatrix} 4 & 1 & 1 & 0 & 3 & 0 & 0 & -1 \\ 1 & 5 & 3 & -1 & 1 & 1 & 4 & 2 \\ 1 & 3 & 5 & -3 & 2 & 4 & 1 & 3 \\ 0 & -1 & -3 & 4 & -1 & -3 & 0 & -2 \\ 3 & 1 & 2 & -1 & 4 & 2 & 0 & -1 \\ 0 & 1 & 4 & -3 & 2 & 6 & -1 & 2 \\ 0 & 4 & 1 & 0 & 0 & -1 & 5 & 0 \\ -1 & 2 & 3 & -2 & -1 & 2 & 0 & 6 \end{pmatrix}$$

where $\sum 2p_iq_i = 1134/256$, and the frequencies of the 10 SNP markers were

$$\frac{1}{16}\begin{pmatrix} 11 & 9 & 7 & 5 & 7 & 5 & 9 & 9 & 12 & 13 \end{pmatrix}.$$

Note that there are some negative relationships in **G** which would not normally occur in **A**. Also, some diagonal elements are greater than 1 which indicates inbreeding, however, the homozygosity picked up in **G** includes the result of identity by chance, while in **A** the inbreeding is due to identity by descent.

Solving the MME give the following results (Table 17.12), which are the gEBV. The assumed variance ratio was 1. The overall mean was estimated to be -0.24.

Table 17.12: gEBV using genomic relationship matrix

Animal	gEBV
7	-3.22
8	4.37
9	1.53
10	-1.59
11	-1.43
12	2.17
13	6.64
14	-1.89

The second step of the procedure is to incorporate the gEBV into the EBV of all other non-genotyped animals. Different methods are followed in various countries for this step. One problem is the genetic base of gEBV versus EBV, and so the base must be made equal for both results. Similarly, the variances of gEBV and EBV need to be made the same.

17.5.5 One-Step Method

To avoid the ad hoc nature of the second step in the Two-Step Method, Misztal et al. (2010) proposed a one step procedure in which all animals are evaluated simultaneously in one set of equations. Ducrocq and Legarra (2011) described a feasible strategy for applying the One-Step Method, and that will be shown here. The original observations are analyzed in this method and not EBVs. Let the model be

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

with the usual definitions of the vectors and matrices. Partition $\mathbf{a}$ into

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix},$$

where subscript 1 indicates animals not genotyped, and 2 denoting genotyped animals, and correspondingly partition $\mathbf{Z}$ and $\mathbf{A}$, the design matrix and relationship matrix, respectively, as

$$\begin{aligned} \mathbf{Z} &= \begin{pmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \end{pmatrix}, \\ \mathbf{A} &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}. \end{aligned}$$

The usual MME (ignoring genotype relationships) are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}'_1\mathbf{X} & \mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{A}^{11}k & \mathbf{A}^{12}k \\ \mathbf{Z}'_2\mathbf{X} & \mathbf{A}^{21}k & \mathbf{Z}'_2\mathbf{Z}_2 + \mathbf{A}^{22}k \end{pmatrix} \begin{pmatrix} \mathbf{b}^* \\ \mathbf{a}_1^* \\ \mathbf{a}_2^* \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_1\mathbf{y} \\ \mathbf{Z}'_2\mathbf{y} \end{pmatrix}.$$

Misztal et al. (2010) show that the correct equations including $\mathbf{G}$ are as follows:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}'_1\mathbf{X} & \mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{A}^{11}k & \mathbf{A}^{12}k \\ \mathbf{Z}'_2\mathbf{X} & \mathbf{A}^{21}k & \mathbf{Z}'_2\mathbf{Z}_2 + \mathbf{A}^{22}k + \mathbf{G}^{-1}k - \mathbf{A}_{22}^{-1}k \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_1 \\ \widehat{\mathbf{a}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_1\mathbf{y} \\ \mathbf{Z}'_2\mathbf{y} \end{pmatrix}.$$

Ducrocq and Legarra (2011) assume the differences between solutions to the above equations are

$$\begin{pmatrix} \widehat{\mathbf{a}}_1 \\ \widehat{\mathbf{a}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^* + \mathbf{d}_1 \\ \mathbf{a}_2^* + \mathbf{d}_2 \end{pmatrix},$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are differences due to genotype relationships. They showed that

$$\mathbf{d}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{d}_2$$

then substitution of this equality into the above equations (page 294), gives an equivalent set of equations below.

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 & \mathbf{0} \\ \mathbf{Z}'_1\mathbf{X} & \mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{A}^{11}k & \mathbf{A}^{12}k & \mathbf{0} \\ \mathbf{Z}'_2\mathbf{X} & \mathbf{A}^{21}k & \mathbf{Z}'_2\mathbf{Z}_2 + \mathbf{A}^{22}k & -\mathbf{A}_{22}^{-1}k \\ \mathbf{0} & \mathbf{0} & -\mathbf{A}_{22}^{-1}k & (\mathbf{A}_{22}^{-1} + [\mathbf{G} - \mathbf{A}_{22}]^{-1})k \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_1 \\ \widehat{\mathbf{a}}_2 \\ \mathbf{d}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_1\mathbf{y} \\ \mathbf{Z}'_2\mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Solving the equations can be done in two pieces in an iterative manner. Start with $\mathbf{d}_2 = \mathbf{0}$.

1. Solve the following equations

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 \\ \mathbf{Z}'_1\mathbf{X} & \mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{A}^{11}k & \mathbf{A}^{12}k \\ \mathbf{Z}'_2\mathbf{X} & \mathbf{A}^{21}k & \mathbf{Z}'_2\mathbf{Z}_2 + \mathbf{A}^{22}k \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_1 \\ \widehat{\mathbf{a}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'_1\mathbf{y} \\ \mathbf{Z}'_2\mathbf{y} + \mathbf{A}_{22}^{-1}k\ \mathbf{d}_2 \end{pmatrix}.$$

2. Solve for $\mathbf{d}_2$ using

$$(\mathbf{A}_{22}^{-1} + [\mathbf{G} - \mathbf{A}_{22}]^{-1})\mathbf{d}_2 \ = \ \mathbf{A}_{22}^{-1}\mathbf{a}_2$$

Repeat parts 1 and 2 until convergence is achieved. Now all animals' EBV are influenced by the SNP genotypes.

Tables 17.13 and 17.14 contain results for the animal model using **A** only, and for the animal model incorporating **G** into **A**. The two models give close agreement for the age group solutions and similar rankings of animals.

Table 17.13: Age group solutions from animal models

Age Group	With **A**	With **G**
1	319.81	318.44
2	343.05	342.01
3	370.07	369.20
4	329.91	328.42

17.5.6 Not All SNPs

In Israel, Weller et al. (2012) do not believe that more SNP markers are better for accuracy of gEBV. They considered a random subset, a subset of evenly spaced markers, and a subset of SNPs with large estimated effects. The subsets were used in an animal model where the animal effects were used to pick up the remaining polygenic effects after the SNPs.

Table 17.14: EBV for animals in Example data

Animal	With **A**	With **G**	Animal	With **A**	With **G**
1	-5.10	-3.01	11	-6.49	-2.14
2	4.83	2.04	12	7.44	5.14
3	7.15	5.37	13	13.08	12.17
4	-4.83	-3.04	14	-6.62	-2.70
5	0.75	0.71	15	0.60	1.64
6	0.00	0.00	16	-7.71	-6.92
7	-4.45	-5.57	17	-3.24	-0.76
8	4.72	8.23	18	3.48	5.18
9	0.38	4.16	19	5.93	6.80
10	-3.44	-4.16	20	-5.69	-4.65

For this method to work properly, we need a better segregation analysis procedure (Kerr and Kinghorn, 1998) so that marker genotypes can be determined for all animals that have data records, and all bulls of interest should have genotypes known. Then all animals can be included and all available genotypes.

If an animal is homozygous for a marker, say AA, then each parent would have contributed an A allele, and each progeny would receive an A allele. If it is known that an animal has an A allele, then its genotype can only be either AA or Aa, with probability equal to the frequency of the A or a alleles, respectively. If the genotype of the other parent is known, then this can give a better estimate of the frequencies of the two alleles. Table 17.15 contains the animals with their known genotypes for marker 1, and then next to that the probabilities of having the other genotypes, if they were not originally genotyped. The frequency of the A allele was $p = 11/16$, and $q = 5/16$ is the frequency of the a allele.

The probabilities in Table 17.15 can go into the analysis for animals 11 through 20. The probabilities might be refined by utilizing the phenotypic observations as in Kerr and Kinghorn (1996) in a Bayesian segregation analysis. Another refinement might be to also use 2 or 3 flanking markers on either side of the markers of interest and to use marker haplotypes (i.e. which alleles are in phase) to help determine probabilities of genotypes for all ungenotyped animals.

Markers 1, 4, and 7 were chosen randomly to illustrate the model analysis. The probabilities for each marker genotype for animals 11 through 20 are shown in Table 17.16 (derived separately for each marker).

The model for fat yield is

$$y_{ij} = A_i + b_1 z_{1ij} + b_4 z_{4ij} + b_7 z_{7ij} + a_j + e_{ij},$$

Table 17.15: Segregation Analysis for Marker 1

Animal	Sire	Dam	Known Genotype	Possible Genotypes AA	Aa	aa
1				p	q	0
2				p	q	0
3				0	p	q
4				p	q	0
5				p^2	$2pq$	q^2
6				p^2	$2pq$	q^2
7			AA	1	0	0
8			Aa	0	1	0
9			AA	1	0	0
10			aa	0	0	1
11	9	7	AA	1	0	0
12	9	2	AA	1	0	0
13	8	3	aa	0	0	1
14	1	4	AA	1	0	0
15	1	6		$p^2 + pq/2$	$(3pq + q^2)/2$	$q^2/2$
16	1	10		p	q	0
17	8	4		$(2p + q)/4$	0.5	$q/4$
18	8	5		$p/2$	0.5	$q/2$
19	8	3		$p/4$	0.5	$(p + 2q)/4$
20	1	6		$p^2 + pq/2$	$(3pq + q^2)/2$	$q^2/2$

where

y_{ij} is fat yield on cow j at age i,

A_i is an age group effect,

z_{kij} is the genotype covariate for marker k, animal j, in age group i, and in this case $k = 1$, 2, or 3,

b_k are regression coefficients on the genotype covariates for marker k,

a_j is the animal polygenic effect after accounting for SNP markers 1, 4, and 7,

e_{ij} is the residual effect of the ij^{th} record.

Post-multiply the frequencies by $(-1 \;\; 0 \;\; 1)$ to obtain a covariate for each

Table 17.16: Probabilities of genotypes for markers 1, 4, and 7

Animal	Marker 1			Marker 4			Marker 7		
	AA	Aa	aa	BB	Bb	bb	CC	Cc	cc
11	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000
12	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
13	0.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000	0.000
14	1.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	0.000
15	0.580	0.371	0.049	0.049	0.371	0.580	0.439	0.465	0.096
16	0.688	0.312	0.000	0.156	0.844	0.000	0.391	0.500	0.109
17	0.422	0.500	0.078	0.000	0.156	0.844	0.781	0.219	0.000
18	0.344	0.500	0.166	0.000	0.312	0.688	0.562	0.438	0.000
19	0.172	0.500	0.328	0.000	0.156	0.844	0.781	0.219	0.000
20	0.580	0.371	0.049	0.049	0.371	0.580	0.439	0.465	0.096

marker for every animal with a record, then

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 1 & -1.000 & 0.000 & 1.000 \\ 1 & 0 & 0 & 0 & -1.000 & 1.000 & 1.000 \\ 0 & 0 & 0 & 1 & 1.000 & 1.000 & -1.000 \\ 1 & 0 & 0 & 0 & -1.000 & 1.000 & -1.000 \\ 0 & 1 & 0 & 0 & -0.531 & 0.531 & -0.343 \\ 0 & 0 & 0 & 1 & -0.688 & -0.156 & -0.282 \\ 1 & 0 & 0 & 0 & -0.344 & 0.844 & -0.781 \\ 0 & 0 & 0 & 1 & -0.178 & 0.688 & -0.562 \\ 0 & 0 & 1 & 0 & 0.156 & 0.844 & -0.781 \\ 0 & 1 & 0 & 0 & -0.531 & 0.531 & -0.343 \end{pmatrix}$$

where the first four columns are for the age effects (see Table 17.1), and the last three columns are for markers 1, 4, and 7. Thus, the SNP genotypes or covariates are fixed effects in the model.

Construct the MME and solve. The solutions are in the Table 17.17.

In the original animal model, without SNP genotypes, the estimate of the residual variance was

$$\sigma_e^2 = (\mathbf{y}'\mathbf{y} - \widehat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \widehat{\mathbf{a}}'\mathbf{Z}'\mathbf{y})/(N - r(\mathbf{X})) \;=\; 1114.64,$$

Accounting for markers 1, 4, and 7, then

$$\sigma_e^2 \;=\; 561.18.$$

Table 17.17: Solutions using Actual and Predicted Genotypes for a subset of SNP markers

	$\widehat{b}$	Animal	$\widehat{a}_k$	Animal	$\widehat{a}_k$
Age 1	282.18	1	2.45	11	-3.43
Age 2	325.05	2	1.66	12	2.26
Age 3	334.59	3	0.52	13	-0.21
Age 4	315.60	4	-1.66	14	0.57
SNP 1	24.63	5	-0.66	15	4.37
SNP 4	65.35	6	0.00	16	4.63
SNP 7	21.59	7	-2.13	17	-3.67
		8	-1.98	18	-1.98
		9	-0.47	19	-0.73
		10	2.27	20	-1.92

Thus, accounting for SNP genotypes has lowered the residual variance, which means the estimated breeding values must be more accurate.

The gEBV for animal 9, for example, where his marker covariates would have been -1, 1, and 0, for markers 1, 4, and 7, respectively, would be calculated as

$$-1(24.63) + 1(65.35) + 0(21.59) \; - \; 0.47 \; = \; 40.25,$$

and for animal 20 would be

$$-0.531(24.63) + 0.531(65.35) - 0.343(21.59) \; - \; 1.92 \; = \; 12.30.$$

In this small example the gEBV are very large due to the estimated regression coefficients being very large. More data and especially more genotypes are needed.

The question becomes how many SNP markers need to be included, and how to determine the best set of markers, out of 50K or 700K. An analysis of including one marker at a time, as in the association studies, but with the above model, will be needed. This could take a long time to try 50K or 700K markers. The number of markers should likely be 50 to 200, but this is a guess.

The best set of markers could differ depending on the trait being analyzed. In total, perhaps 5000 markers need to be saved to include in genetic evaluations of all traits.

The problem of more SNPs to estimate than data is resolved because all animals with records are given marker genotypes based on probabilities, and a limited number of markers are used per trait. All data records are utilized simultaneously.

The SNP genotype probabilities can be stored and revised as more anim-

als are genotyped and as more animals are added to the pedigrees, rather than starting over from nothing each genetic evaluation run. The process would be to use genotyped animals to determine which alleles likely came from each parent. In some cases, the exact genotype of a parent for a marker may be possible to derive.

After all parents have been assigned probabilities for each genotype, then all ungenotyped progeny can be processed. If both parents are homozygous, then the genotype of the progeny may be determined exactly. Use of phenotypic records in a Bayesian approach may be helpful, but should be tested. A study is needed to compare the methods in this section in terms of accuracy of estimating true breeding values. True breeding values would be simulated using genotypes of 100K markers. Different percentages of genotyped animals should also be compared, and whether it is more important to genotype parents or progeny.

17.6 Imputation

Other approaches to estimating genotype probabilities have been given by Van Arendonk et. al (1989) and by Fernando et. al. (1993). There are many sizes of SNP panels with the latest being 700K for dairy cattle. With 700K SNPs, the genome is thoroughly and densely covered with markers and the gaps between SNPs are small. Estimation of the effects of 700K SNPs is a problem because by comparison there are very few genotyped animals from which to estimate all of these effects. Instead the SNPs are used to obtain a better genetic relationship matrix among animals. Because animals have been genotyped with different sizes of SNP panels, the process of imputation has been studied heavily and several methods developed. Imputation is where, for example, the genotypes from a 50K chip are used to 'guess' the genotypes for a 700K chip using pedigrees and population knowledge about LD between SNPs. There needs to be enough animals genotyped with the 700K panel, and from that their 50K genotypes are also known (same SNPs on both chips), so that algorithms to extend from 50K to 700K can be worked out. The strategy would be to genotype relatively few animals with the 700K chip (maybe bulls and bull dams), and all other animals would be genotyped with the cheaper 50K chip, and then imputed by computer to 700K. The accuracy of imputation is fairly high (95% in most cases). Thus, money for genotyping can be used to genotype more animals at 50K rather than paying much more for 700K genotypes. Eventually, costs of genotyping may become low for everyone to be able to use 700K chips or larger.

17.7 Genome Sequencing

SNPs may not have a long life in genomics. The cost of sequencing the entire genome is decreasing. A person can have their genome sequenced for a cost of only $1000 US. Nanotechnology has developed a tube through which DNA passes, one base pair at a time, allowing 'reading' of the sequence. What can we do with a complete 3.5 billion base pair sequencing of DNA per cow or bull? That is a lot of information to store and to process, per animal. Sequence data will need to be condensed into a smaller number of useful pieces of information. Supposedly we should be able to identify the 25,000 genes and their alleles. Genes will likely have more than two alleles each, probably dozens of alleles.

Genes are known to interact with each other. Traits are affected by hundreds or thousands of genes, and some genes can affect several traits. These pathways need to be understood before selection on individual genes should be made. Genomics is new and valuable to genetic improvement, but there is still much that needs to be learned about the genome.

17.8 References

DUCROCQ, V. , A. LEGARRA. 2011. Interbull Meeting. Stavangar, Norway, August 2011.

FERNANDO, R. L. , C. STRICKER, R. C. ELSTON. 1993. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. Theor. Appl. Genet. 87:89-93.

FISHER, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans. Royal Soc. Edinburgh. 52:399-433.

GIANOLA, D. 2010. Human Genome Project. 1984

KERR, R. , B. KINGHORN. 1996. An efficient algorithm for segregation analysis in large populations. Livest. Prod. Sci. 113:457-469.

MEUWISSEN, T. H. E. , B. J. HAYES, M. E. GODDARD. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

MISZTAL, I. , I. AGUILAR, A. LEGARRA, S. TSURUTA, D. L. JOHNSON, T. J. LAWLOR. 2010. A unified approach to utilize phenotypic, full pedigree and genomic information for genetic evaluation. 9^{th} World Congress on Genetics Applied to Livestock Production. Leipzig, Germany. Paper 0050.

SCHAEFFER, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:1-6.

SPELMAN, R. J. , C. A. FORD, P. McELHINNEY, G. C. GREGORY, R. G. SNELL. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85:3514-3517.

VAN ARENDONK, J. A. M. , C. SMITH, B. W. KENNEDY. 1989. Method to estimate genotype probabilities at individual loci in farm livestock. Theor. Appl. Genet. 78:735-740.

WATSON, J. D. , F. H. C. CRICK. 1953. A structure for deoxyribose nucleic acid. Nature 171:737-738.

WATSON, J. D. , F. H. C. CRICK. 1953. Genetical implications of the structure of deoxyribose nucleic acid. Nature 171:964-967.

WELLER, J. I. , G. GLICK, A. SHIRAK, E. EZRA, Y. ZERON, M. RON. 2012. Predictive ability of selected subsets of SNPs for moderately sized dairy cattle populations.

XU, S. 2001. Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.

Part IV

APPENDICES

Appendix A

Relationship Among Methods

HORIA GROSU
SORIN LUNGU

This appendix attempts to illustrate the subtle differences between three methods of sire evaluation, and how the methods are connected.

A.1 Contemporary Comparison

This is the method of Roberston and Rendel (1954). Consider the case of only first lactations and the model

$$y_{ijk} \;=\; h_i \;+\; s_j \;+\; e_{ijk},$$

where

y_{ijk} represents the heifer yield of daughter k of sire j in herd-year-season i,

h_i is the fixed herd-year-season effect, (i.e. contemporary group),

s_j is the sire effect, and

e_{ijk} is the residual effect.

Let n_{ij} represent the number of daughters of sire j in HYS i. In matrix notation,

$$\mathbf{y} \;=\; \mathbf{Xh} \;+\; \mathbf{Zs} \;+\; \mathbf{e}.$$

The least squares estimates for sires and HYS are given by

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{h}} \\ \widehat{\mathbf{s}} \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix}.$$

The equation for the i^{th} HYS is

$$n_{i.}\widehat{h}_i \; + \; \sum_j (n_{ij}\widehat{s}_j) \; = \; y_{i..}$$

and for the j^{th} sire is

$$\sum_i \widehat{h}_i \; + \; n_{.j}\widehat{s}_j \; = \; y_{.j.}$$

When the number of HYS are large, then those equations may be absorbed into the sire equations, to give

$$\mathbf{Z'SZ\hat{s}} \; = \; \mathbf{Z'Sy}.$$

The equation for the j^{th} sire is

$$\left[n_{.j} - \sum_i \left(\frac{n_{ij}^2}{n_{i.}} \right) \right] \widehat{s}_j \; - \sum_i \sum_{j' \neq j} \left(\frac{n_{ij} n_{ij'}}{n_{i.}} \right) \widehat{s}_{j'} = \; y_{.j.} - \sum_i (n_{ij} \overline{y}_{i..})$$

Let $n_{ij'}$ represent all of the contemporaries of daughters of sire j, then

$$\begin{aligned} n_{i.} &= n_{ij} + n_{ij'} \\ w_j &= n_{.j} - \sum_i \left(\frac{n_{ij}^2}{n_{i.}} \right) \\ w_{ij} &= n_{ij} \left(1 - \frac{n_{ij}}{n_{i.}} \right) \\ &= \frac{n_{ij} n_{i.} - n_{ij}^2}{n_{i.}} \\ w_{ij} &= \frac{n_{ij} n_{ij'}}{n_{ij} + n_{ij'}} \end{aligned}$$

Using the above results, then Thompson (1976) showed that the equation for the j^{th} sire could be written as

$$w_j \widehat{s}_j \; = \; \sum_i \left[w_{ij} \left(\overline{y}_{ij} - \overline{y}_{ij'} \right) \right] + \overline{A}_{j'},$$

where

$\overline{y}_{ij}$ is the average yield of daughters of sire j in HYS i,

$\overline{y}_{ij'}$ is the average yield of daughters of all other sires, except j, in HYS i, and

$\overline{A}_{j'}$ is a measure of the genetic merit of the contemporaries of daughters of sire j, where

$$\overline{A}_{j'} = \sum_i \sum_{j' \neq j} \left(\frac{n_{ij} n_{ij'}}{n_{i.}} \right) \widehat{s}_{j'}$$

Solving the above equation would be difficult if there are many bulls and HYS, so Robertson and Rendel (1954) suggested ignoring the last term, $\overline{A}_{j'}$, assuming that the average genetic merit of contemporaries was equal among sires. Then C_j is used to approximate $\widehat{s}_j$, so that

$$w_j C_j = \sum_i \left[w_{ij} \left(\overline{y}_{ij} - \overline{y}_{ij'} \right) \right]$$

Consequently, C_j is an average of weighted deviations of daughter averages from their contemporary averages, and called the Contemporary Comparison. The solution for C_j is then regressed using

$$\widehat{g}_j = \left(\frac{w_j}{w_j + k} \right) C_j$$

A.1.1 Numerical Example

Consider the data from Chapter 7, duplicated below.

Table A.1: Example Data for Least Squares Method. CG = Contemporary Group

Sire	Herd 1	Herd 2	Sire totals
1	2(9,100)	2(8,000)	4(17,100)
2	5(20,200)	3(13,100)	8(33,300)
3	1(4,500)	5(19,600)	6(24,100)
CG Totals	8(33,800)	10(40,700)	

After setting up the LS equations and aborbing the HYS equations, the resulting sire equations are

$$\begin{pmatrix} 3.100 & -1.850 & -1.250 \\ -1.850 & 3.975 & -2.125 \\ -1.250 & -2.125 & 3.375 \end{pmatrix} \begin{pmatrix} \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \end{pmatrix} = \begin{pmatrix} +510 \\ -35 \\ -475 \end{pmatrix}$$

The off-diagonals of the coefficient matrix are ignored, as suggested by

Robertson and Rendel (1954), then

$$\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} = \begin{pmatrix} 3.100 & 0 & 0 \\ 0 & 3.975 & 0 \\ 0 & 0 & 3.375 \end{pmatrix}^{-1} \begin{pmatrix} +510 \\ -35 \\ -475 \end{pmatrix} = \begin{pmatrix} +164.5161 \\ -8.8050 \\ -140.7407 \end{pmatrix}.$$

If $k = 15$ is assumed, then

$$\begin{aligned} CC_1 &= \left(\frac{3.100}{3.100+15}\right) \times 164.5161 = 28.18 \\ CC_2 &= \left(\frac{3.975}{3.975+15}\right) \times -8.8050 = -1.85 \\ CC_3 &= \left(\frac{3.375}{3.375+15}\right) \times -140.7407 = -25.85 \end{aligned}$$

Bar Anan and Sacks (1974) give an example of the biases encountered as a bull ages, because the contemporaries of the later daughters will likely be daughters of progressively younger bulls with potentially greater genetic merit.

A.2 Cumulative Difference Method

Bar Anan and Sacks (1974) described a method to correct for the deficiencies of the contemporary comparison. The Cumulative Difference Method gave sire estimated breeding values that consisted of two parts.

1. An estimate of the comparison of daughters to contemporaries, and
2. An adjustment for the genetic level of the contemporaries.

The method was to calculate

$$CD_j = CC_j + \overline{A}_{j'}$$

where $\overline{A}_{j'}$ is the average genetic deviation of the sires of the contemporaries. Let

$$\mathbf{W} = diag(\mathbf{Z'SZ}) = \begin{pmatrix} 3.100 & 0 & 0 \\ 0 & 3.975 & 0 \\ 0 & 0 & 3.375 \end{pmatrix}$$

$$\begin{aligned}\mathbf{W}-\mathbf{Z}'\mathbf{S}\mathbf{Z} &= \begin{pmatrix} 0 & 1.850 & 1.250 \\ 1.850 & 0 & 2.125 \\ 1.250 & 2.125 & 0 \end{pmatrix} \\ &= \mathbf{Q}\end{aligned}$$

then

$$\begin{pmatrix} \overline{A}_{1'} \\ \overline{A}_{2'} \\ \overline{A}_{3'} \end{pmatrix} = \mathbf{W}^{-1}\mathbf{Q} \begin{pmatrix} CC_1 \\ CC_2 \\ CC_3 \end{pmatrix}$$

or

$$\begin{pmatrix} \overline{A}_{1'} \\ \overline{A}_{2'} \\ \overline{A}_{3'} \end{pmatrix} = \begin{pmatrix} -11.5200 \\ -0.7056 \\ +9.2745 \end{pmatrix}$$

which shows that sire 1 had contemporaries whose sires were of greater genetic merit than those of sires 2 or 3, thus, a downward adjustment on CC_1

Finally,

$$\begin{pmatrix} CD_1 \\ CD_2 \\ CD_3 \end{pmatrix} = \begin{pmatrix} 28.1800 \\ -1.8455 \\ -25.8503 \end{pmatrix} + \begin{pmatrix} -11.5200 \\ -0.7056 \\ +9.2745 \end{pmatrix} = \begin{pmatrix} +16.6525 \\ -2.5502 \\ -16.5759 \end{pmatrix}.$$

Thus, the Cumulative Difference Method is the Contemporary Comparison Method plus an adjustment for the genetic level of the contemporaries' sires. The off-diagonals in $\mathbf{Z}'\mathbf{S}\mathbf{Z}$ are not ignored in the Cumulative Difference Method.

A.3 Modified Cumulative Difference Method

Dempfle (1976) found that sires with lower numbers of daughters were disadvantaged in the Cumulative Difference Method, and therefore, suggested the following order of calculations. First,

$$CA_j = C_j + \overline{A}_{j'}$$

where

$$\begin{pmatrix} \overline{A}_{1'} \\ \overline{A}_{2'} \\ \overline{A}_{3'} \end{pmatrix} = \mathbf{W}^{-1}\mathbf{Q} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}$$

or

$$\begin{pmatrix} \overline{A}_{1'} \\ \overline{A}_{2'} \\ \overline{A}_{3'} \end{pmatrix} = \begin{pmatrix} -62.0049 \\ +1.3285 \\ +55.3280 \end{pmatrix}$$

Then

$$\begin{pmatrix} CA_1 \\ CA_2 \\ CA_3 \end{pmatrix} = \begin{pmatrix} +164.5161 \\ -8.8050 \\ -140.7407 \end{pmatrix} + \begin{pmatrix} -62.0049 \\ +1.3285 \\ +55.3280 \end{pmatrix} = \begin{pmatrix} +102.5112 \\ -7.4765 \\ -85.3527 \end{pmatrix}.$$

Now the result is regressed to obtain the sire evaluations, (assuming $k = 15$ again),

$$\begin{pmatrix} CD_1 \\ CD_2 \\ CD_3 \end{pmatrix} = \begin{pmatrix} \frac{3.100}{3.100+15}(102.5112) \\ \frac{3.975}{3.975+15}(-7.4765) \\ \frac{3.375}{3.375+15}(-85.3527) \end{pmatrix} = \begin{pmatrix} +17.5572 \\ -1.5662 \\ -15.6770 \end{pmatrix}.$$

The procedure is iterated using CD_j instead of C_j to derive new $\overline{A}_{j'}$, then new CA_j followed by new CD_j. After several iterations,

$$\begin{pmatrix} CD_1 \\ CD_2 \\ CD_3 \end{pmatrix} = \begin{pmatrix} +26.2945 \\ -2.0015 \\ -24.2931 \end{pmatrix}.$$

The Mixed Model Equations (MME) for this model, after absorbing HYS, would be

$$(\mathbf{Z}'\mathbf{S}\mathbf{Z} + \mathbf{I}k)\widehat{\mathbf{s}} = \mathbf{Z}'\mathbf{S}\mathbf{y}$$

and the solutions are

$$\begin{pmatrix} \widehat{s}_1 \\ \widehat{s}_2 \\ \widehat{s}_3 \end{pmatrix} = \begin{pmatrix} +26.2945 \\ -2.0015 \\ -24.2931 \end{pmatrix}.$$

Thus, the modified cumulative difference method and MME would be identical for a model where sires are assumed to be unrelated.

A.4 Thompson Approach

From the previous section, at the first iteration step,

$$CD_j = b_j \times CA_j$$

where $b_j = \frac{w_j}{w_j+k}$. Thompson (1976) showed that this could be re-arranged as

$$\begin{aligned} CD_j &= b_j(C_j + \overline{A}_{j'}) \\ &= b_j \times C_j + b_j \times \overline{A}_{j'} \\ &= CC_j + b_j \times \overline{A}_{j'} \end{aligned}$$

Then $\overline{A}_{j'}$ are updated using the new CD_j, as above, but CC_j is kept constant through the iterations. The results converge to the same as the modified cumulative difference method and to the MME sire model solutions from the previous section.

Let

$$\begin{aligned} \mathbf{Z'SZ} &= \mathbf{W} + (\mathbf{Z'SZ} - \mathbf{W}) \\ &= \mathbf{W} - \mathbf{Q} \end{aligned}$$

where $\mathbf{W}$ are the diagonals of $\mathbf{Z'SZ}$, and $\mathbf{Q}$ are the off-diagonals. Then the MME for a sire model can be written as

$$\begin{aligned} (\mathbf{W} - \mathbf{Q} + \mathbf{I}k)\widehat{\mathbf{s}} &= \mathbf{Z'Sy} \\ (\mathbf{W} + \mathbf{I}k)\widehat{\mathbf{s}} &= \mathbf{Z'Sy} + \mathbf{Q}\widehat{\mathbf{s}} \end{aligned}$$

where the coefficient matrix on the left is diagonal, thus for the j^{th} sire

$$(w_j + k)\widehat{s}_j = \sum_i \left(w_{ij}(\overline{y}_{ij} - \overline{y}_{ij'})\right) + \overline{A}_{j'}$$

where

$$\overline{A}_{j'} = (\mathbf{Q}\widehat{\mathbf{s}})_j$$

which is the j^{th} element of $\mathbf{Q}\widehat{\mathbf{s}}$. Also,

$$\begin{aligned} \widehat{\mathbf{s}} &= (\mathbf{W} + \mathbf{I}k)^{-1}\mathbf{Z'Sy} + (\mathbf{W} + \mathbf{I}k)^{-1}\mathbf{Q}\widehat{\mathbf{s}} \text{ or} \\ \widehat{s}_j &= CC_j + b_j\overline{A}_{j'}. \end{aligned}$$

A.5 Summary

The following comments can be made.

1. The Contemporary Comparison of Robertson and Rendel (1954) calculate C_j which are weighted deviations of bull's daughter records from contemporary records.

$$C_j = \frac{\sum_i \left(w_{ij}(\overline{y}_{ij} - \overline{y}_{ij'})\right)}{\sum_i w_{ij}}$$

The final evaluation is

$$CC_j = \frac{w_j}{w_j + k} C_j.$$

The off-diagonals of $\mathbf{Z'SZ}$ are ignored, which means that the genetic level of the contemporaries are ignored.

2. The Cumulative Difference Method of Bar Anan and Sacks (1974) takes CC_j and adds an adjustment for the genetic level of the contemporaries, which uses the off-diagonals of $\mathbf{Z'SZ}$.

3. The Modified Cumulative Difference Method of Dempfle(1976) or of Thompson (1976) also account for the genetic level of the contemporaries, but in a different order of calculations, and iteratively. Thus,

$$CD_j = b_j(C_j + \overline{A}_{j'}), or$$

or

$$CD_j = CC_j + b_j \overline{A}_{j'}.$$

Iterating to update $\overline{A}_{j'}$ leads to CD_j which are identical to sire solutions from corresponding MME.

4. If genetic relationships among sires are to be considered, then the Modified Cumulative Difference method would not be identical to the MME solution.

A.6 References

BAR-ANAN, R., and J.M., SACKS, 1974. Sire evaluation and estimation genetic gain in Israeli dairy herds. Anim.Prod. 18: 59-66;

DEMPFLE, L. , 1976. A note on the properties of the cumulative difference method for sire evaluation. Anim.Prod. 23: 121-124;

ROBERTSON, A. , and J.M. RENDEL, 1954. The performance of heifers got by artificial insemination. J. agric. Sci. Camb. 44: 184-192;

THOMPSON, R. , 1976. Relationship between the Cumulative Difference and Best Linear Unbiased Predictor Methods of Evaluating BULLS, Anim. Prod. 23: 15-24.

Appendix B

Romanian Animal Model, 1982

CORNELIU DRĂGĂNESCU

B.1 Animal Model, 1982

An animal model-like procedure for Romania was described in 1982, but which does not utilize the additive genetic relationship matrix. The calculations involve coefficients that look like Henderson's rules for computing the inverse of **A**. Multiple lactation records per cow, assumed to be adjusted for parity, age, and season of calving are modelled as

$$y_{ijk} = h_i + a_j + p_j + e_{ijk},$$

where

y_{ijk} is lactation k on cow j in herd-year-season i,

h_i is a fixed herd-year-season effect,

a_j is a random animal additive genetic effect,

p_j is a random animal permanent environmental effect, and

e_{ijk} is a random residual effect.

Also,

$$\begin{aligned} E(a_j) &= 0 \\ E(p_j) &= 0 \\ E(e_{ijk}) &= 0 \end{aligned}$$

and the variances are

$$\sigma_y^2 \;=\; \sigma_a^2 \;+\; \sigma_p^2 \;+\; \sigma_e^2,$$

for phenotype, additive, PE, and residual variances, respectively. Let

$$k_p \;=\; \frac{\sigma_e^2}{\sigma_p^2} \;=\; \frac{1-r}{r-h^2},$$

and

$$k_a \;=\; \frac{\sigma_e^2}{\sigma_a^2} \;=\; \frac{1-r}{h^2}.$$

Instead of additive genetic relationships, each animal is assigned to one of four groups on the basis of available parentage information, and coefficients are assigned for weighting the information, as shown in the following table (Table B.1). In actual fact, these constants correspond to $\mathbf{A}^{-1}$ times 6. The iteration strategy uses the constants exactly as in using Henderson's rules to create $\mathbf{A}^{-1}$.

Table B.1: Grouping Information

Sire Known	Dam Known	K	M	T
No	No	6	0	0
Yes	No	8	4	0
No	Yes	8	0	4
Yes	Yes	12	6	6

An iteration on data strategy has been used. Start with all solutions being zero, then the steps are as follows:

1. Solve for herd-year-season solutions,

$$\hat{h}_i \;=\; \frac{\sum_j \sum_k (y_{ijk} - \hat{a}_j - \hat{p}_j)}{n_{i.}}$$

2. Solve for animal PE effects,

$$\hat{p}_j \;=\; \frac{\sum_i \sum_k (y_{ijk} - \hat{h}_i - \hat{a}_j)}{n_{.j} + k_p}$$

3. Solve for animal additive genetic effects, in parts,

   - Data part, D and its weight, L,

$$D \;=\; \sum_i \sum_k (y_{ijk} - \hat{h}_i),$$

     and

$$L \;=\; \frac{6\ h^2}{n_{.j}(r - h^2) + (1 - r)}.$$

   - Parent part, PA, and

$$PA \;=\; M\hat{a}_s \;+\; T\hat{a}_d$$

     where s and d indicate sire and dam of animal j.

   - Progeny part, O, and its weight W,

$$O \;=\; 4\ \sum_m (\hat{a}_m) \;+\; 6\ \sum_\ell (\hat{a}_\ell - \hat{a}_{d\ell}/2),$$

     where m denote progeny of animal j when the mate is unknown, and ℓ and $d\ell$ indicates mate of animal j that co-produced progeny ℓ for progeny when both parents are known, and

$$W = \sum_m (2) + \sum_\ell (3)$$

     for progeny without mate known, and with mate known.

   Combining the pieces gives

$$\hat{a}_j \;=\; \frac{(L \times D) + PA + O}{n_{.j}L + K + W}.$$

   When j refers to a bull or dam without records, then $L = 0$.

The coefficients in B.1 take the place of Henderson's rules for inverting the additive genetic relationship matrix. The method ignores inbreeding. During the iteration process it is necessary to combine animal solutions from 3 separate rounds of

iteration. Let $a^{(i)}$ be the solution for an animal from round i, then at the end of each iteration calculate

$$a^{next} = (-a^{(i-1)} + 2a^{(i)} + 5a^{(i+1)})/6$$

This method, described in 1982, was not implemented in Romania until 1996-7.

B.1.1 Example

Below are repeated records on cows in 3 herd-year-seasons (Table B.2). Assume that $h^2 = 0.267$ and $r = 0.398$, then $k_a = 2.25$ and $k_p = 4.59$.

Table B.2: Example Protein Yield Data To Illustrate Methods

Cow	Sire	Dam	HYS 1	HYS 2	HYS 3
12	1	5	284	285	299
13	1	6	301	371	320
14	1	7	329	324	371
15	2	8	285	309	
16	2	9	258	302	
17	2	10	306		
18	2	11	300		
19	3	5		327	352
20	3	8		323	340
21	4	7			334
22	4	10			338

Table B.3 contains the solutions to the above example and also, the solutions to the usual animal model MME using **A**.

The solutions from the two models are similar and highly correlated except for the animals that did not have records. Their solutions tended to differ more than other solutions.

B.2 Animal model - 1982

Biometric model:

$y_{ijk} = m_i + a_j + p_j + e_{ijk}$,

y_{ijk} = is a performance in lactation k on cow j in herd-year-season i;

m_i = is fixed hys effect i, i =1 .. F ;

Table B.3: Results of Drăgănescu (1982) and usual animal model analysis

	Drăgănescu	Usual MME		Drăgănescu	Usual MME
h_1	296.68	296.98	a_{14}	9.83	9.40
h_2	321.18	321.43	a_{15}	-5.01	-5.43
h_3	333.23	333.53	a_{16}	-11.09	-11.06
a_1	4.71	1.48	a_{17}	-0.20	0.07
a_2	-7.38	-5.27	a_{18}	-1.63	-1.42
a_3	0.65	3.64	a_{19}	2.33	2.46
a_4	3.89	0.15	a_{20}	0.79	1.80
a_5	-1.78	-4.63	a_{21}	3.16	2.47
a_6	5.14	3.17	a_{22}	1.77	1.66
a_7	7.75	5.53	p_{12}	-7.88	-7.44
a_8	-3.00	-1.41	p_{13}	2.90	3.11
a_9	-8.91	-5.62	p_{14}	5.72	5.78
a_{10}	0.20	2.15	p_{15}	-2.10	-2.05
a_{11}	-2.60	0.81	p_{16}	-5.42	-5.51
a_{12}	-7.77	-9.16	p_{17}	1.70	1.60
a_{13}	6.30	5.49	p_{18}	0.89	0.79
			p_{19}	3.02	2.90
			p_{20}	1.06	0.67
			p_{21}	-0.43	-0.36
			p_{22}	0.54	0.50

a_j = is a random additive genetic value of the cow j;

p_j = is a random nonadditive (dominance, epistatic and permanent environmental effect) on cow j;

e_{ijk} = is random residual effect. All variables are uncorrelated, $E(a_j)=E(p_j)=E(e_{ijk})=0$, $var(a_j)=V_A$, $var(p_j)=V_p$, $var(e_{ij})=V_e$

$h^2 = V_A/V_y$, $V_y = var\ (y) = V_A + V_p + V_e$

$r = (V_A+V_p\ /V_y$,

U = the set of animals (cows and bulls) to be evaluated;

X = the set of animals (cows and bulls) with dam unknown and sire unknown;

Y = the set of animals (cows and bulls) with dam known and sire unknown;

Z = the set of animals (cows and bulls) with dam unknown and sire known;

W = the set of animals (cows and bulls) with dam known and sire known;

D'(u) = the set of offspring (cows and bulls) of a cow (u=v) or of a bull (u=t), with the unknown partner;
D"(u) = the set of offspring (cows and bulls)
of a cow (u=v) or of a bull (u=t), with the known partner;
N_u' = number of offspring in the set D_u' ;
$N_u"$ = number of offspring in the set $D_u"$;
M(d) = dam of d;
T(d) = sire of d.

Multiplying with V_y , the WLS function to minimize is as follows:
$f(m, a, p) = \Sigma_{ijk}(y_{ijk} - m_i - a_j - p_j)^2 / (1 - r) + \Sigma_j p_j^2/(r - h^2) + \Sigma_{u\in X}(a_u^2) / (h^2) + \Sigma_{u\in Y} (a_u - a_{M(u)} /2)^2 / (3h^2/4) + \Sigma_{u\in Z}(a_u - a_{T(u)} /2)^2 / (3h^2/4) + \Sigma_{u\in W}(a_u - a_{M(u)} /2 - a_{T(u)} /2)^2 / (h^2/2)$

Further, the notations for predicted or estimated values will be the same with those of true ones.

Effect of hys i (m_i)

The derivative of f(m, a, p) with respect to m_i is set to zero:
$(-2)\Sigma_{jk}(y_{ijk} - m_i - a_j - p_j)/(1 - r) = 0$
$\Sigma_{jk}(y_{ijk} - m_i - a_j - p_j)= 0$
$\Sigma_{jk} m_i = \Sigma_{jk} (y_{ijk} - a_j - p_j)$
$m_i = \Sigma_{jk}(y_{ijk} - a_j - p_j)/ \Sigma_j n_{ij} = [y_{i..} - \Sigma_j n_{ij}(a_j + p_j)] / n_{i.}$ **(1)**

The animal PE effect on cow j(p_j)

The derivative of f(m, a, p) with respect to p_j is set to zero:
$(-2)\Sigma_{ik}(y_{ijk} - m_i - a_j - p_j)/(1 - r) + [2/(r - h^2)]p_j = 0$
$(-1)\Sigma_{ik}(y_{ijk} - m_i - a_j - p_j) / (1 - r) + [1/(r - h^2)]p_j = 0$
$[n_{.j} + (1 - r)/(r - h^2)]p_j = (y_{.j.} - \Sigma_i n_{ij}m_i) - n_{.j} a_j$
$p_j = [y_{.j.} - \Sigma_i n_{ij}m_i - n_{.j} a_j] / [n_{.j} + (1 - r)/(r - h^2)]$ **(2)**

The proof and result is based on cows with known lactations.

Additive genetic effect of the cow j=v

Cows without lactations

Let us denote:
α=1 if the cow v has the dam and the sire unknown, and
α=0 in others situations;

β=1 if the cow v has the dam known and the sire unknown, and β=0 in others situations;

γ=1 if the cow v has the dam unknown and the sire known, and γ=0 in others situations;

δ=1 if the cow v has the dam known and the sire known, and δ=0 in others situations.

The derivative of f(m, a, p) with respect to a_v is set to zero:

$$\alpha(2/h^2)a_v + \beta 8/(3h^2)(a_v - a_{M(v)} /2) + \gamma 8/(3h^2)(a_v - a_{T(v)} /2) +$$
$$+ \delta 4/ h^2(a_v - a_{M(v)} /2 - a_{T(v)} /2) +$$
$$(2)(-1/2)/(3h^2/4)\Sigma_{d\in D'(v)}(a_d - a_v/2) +$$
$$(2)(-1/2)/(h^2/2)\Sigma_{d\in D''(v)}(a_d - a_v /2 - a_{T(d)} /2) = 0$$

$$[2\alpha + 8\beta/3 + 8\gamma/3 + 4\delta]a_v + [(2/3)\Sigma_{d\in D'(v)}a_v + \Sigma_{d\in D''(d)} a_v]=$$
$$= \beta(4/3)a_{M(v)} + \gamma(4/3)a_{T(v)} + \delta(4)(a_{M(v)} /2 + a_{T(v)} /2) +$$
$$+ (4/3) \Sigma_{d\in D'(v)} a_d + 2[\Sigma_{d\in D''(u)} (a_d - a_{T(d)} /2)]$$

$$(6\alpha + 8\beta + 8\gamma + 12\delta + 2N_v' + 3N_v'')a_v =$$
$$(4\beta + 6\delta)a_{M(v)} + (4\gamma + 6\delta)a_{T(v)} + 4\Sigma_{d\in D'(v)} a_d +$$
$$6\Sigma_{d\in D''(v)} (a_d - a_{T(d)} /2) \qquad \textbf{(3a)}$$

id est

α=1 => $(6 + 2N_v' + 3N_v'')a_v = 4\Sigma_{d\in D'(v)} a_d + 6\Sigma_{d\in D''(v)} (a_d - a_{T(d)} /2)$

β=1 => $(8 + 2N_v' + 3N_v'')a_v = 4a_{M(v)} + 4\Sigma_{d\in D'(v)} a_d + 6\Sigma_{d\in D''(v)} (a_d - a_{T(d)} /2)$

γ=1 => $(8 + 2N_v' + 3N_v'')a_v = 4a_{T(v)} + 4\Sigma_{d\in D'(v)} a_d + 6\Sigma_{d\in D''(v)} (a_d - a_{T(d)} /2)$

δ=0 => $(12 + 2N_v' + 3N_v'')a_v = 6a_{M(v)} + 6a_{T(v)} + 4\Sigma_{d\in D'(v)} a_d + 6\Sigma_{d\in D''(v)} (a_d - a_{T(d)} /2)$

Cows with lactations

The derivative of f(m, a, p) with respect to a_v is set to zero:

$$[-2/(1 - r]\Sigma_{ik} (y_{ivk} - m_i - a_v - p_v) +$$
$$\alpha(2/h^2)a_v + \beta(8)/(3h^2)(a_v - a_{M(v)} /2) + \gamma(8)/(3h^2)(a_v - a_{T(v)} /2) +$$
$$\delta(4/ h^2)(a_v - a_{M(v)} /2 - a_{T(v)} /2)$$
$$+ (2)(-1/2)/(3h^2/4)\Sigma_{d\in D'(v)}(a_d - a_v/2) +$$
$$(2)(-1/2)/(h^2/2)\Sigma_{d\in D''(v)}(a_d - a_v /2 - a_{T(d)} /2) =0$$
$$[-h^2/(1 - r]\Sigma_{ik} (y_{ivk} - m_i - a_v - p_v) +$$
$$\alpha a_v + \beta(4/3)(a_v - a_{M(v)} /2) + \gamma(4/3)(a_v - a_{T(v)} /2) +$$
$$\delta 2(a_v - a_{M(v)} /2 - a_{T(v)} /2) + (2)(-1/2)/(3/2)\Sigma_{d\in D'(v)}(a_d - a_v/2) +$$
$$(2)(-1/2)\Sigma_{d\in D''(v)}(a_d - a_v /2 - a_{T(d)} /2) =0$$

$$[h^2/(1 - r)]\Sigma_{ik} (a_v + p_v) +[\alpha + 4\beta/3 + 4\gamma/3 + 2\delta)] a_v +$$
$$+ (1/3)N_v'a_v + (1/2)N_v''a_v = [h^2/(1 - r)]\Sigma_{ik} (y_{ivk} - m_i) +$$

$+ \quad (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

$[h^2/(1 - r)][n_{.v}\, a_v + n_{.v}\, p_v] + [\alpha + 4\beta/3 + 4\gamma/3 + 2\delta) + (1/3)N_v' + (1/2)N_v'']a_v = [h^2/(1 - r)][y_{.v.} - \Sigma_i n_{iv}m_i] + (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

Because $p_v = [y_{.v.} - \Sigma_i n_{iv}m_i - n_{.v}\, a_v] / [n_{.v} + (1 - r)/(r - h^2)]$,
it results
$[h^2/(1 - r)]\{ n_{.v}\, a_v + n_{.v}\, [y_{.v.} - \Sigma_i n_{iv}m_i - n_{.v}\, a_v] / [n_{.v} + (1 - r)/(r - h^2)] \} + [\alpha + 4\beta/3 + 4\gamma/3 + 2\delta) + (1/3)N_v' + (1/2)N_v'']a_v =$
$[h^2/(1 - r)][y_{.v.} - \Sigma_i n_{iv}m_i] +$
$+ (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$
$[h^2/(1 - r)]\{ n_{.v}\, a_v + n_{.v}\, [- n_{.v}\, a_v] / [n_{.v} + (1 - r)/(r - h^2)] \} +$
$+ [\alpha + 4\beta/3 + 4\gamma/3 + 2\delta) + (1/3)N_v' + (1/2)N_v'']a_v =$
$= [h^2/(1 - r)][y_{.v.} - \Sigma_i n_{iv}m_i] - n_{.v}\, [y_{.v.} - \Sigma_i n_{iv}m_i] /$
$[n_{.v} + (1 - r)/(r - h^2)] + (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

$[h^2/(1 - r)]\{ (1 - r) / [n_{.v}(r - h^2)] + (1 - r)]\} n_{.v}\, a_v +$
$+ [\alpha + 4\beta/3 + 4\gamma/3 + 2\delta) + (1/3)N_v' + (1/2)N_v'']a_v =$
$= [h^2/(1 - r)]\{ (1 - r) / [n_{.v}(r - h^2) + (1 - r)]\}[y_{.v.} - \Sigma_i n_{iv}m_i] +$
$+ (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

$\{ h^2/ [n_{.v}(r - h^2)] + (1 - r)]\} n_{.v}\, a_v +$
$+ [\alpha + 4\beta/3 + 4\gamma/3 + 2\delta) + (1/3)N_v' + (1/2)N_v'']a_v =$
$= \{h^2 / [n_{.v}(r - h^2) + (1 - r)]\}[y_{.v.} - \Sigma_i n_{iv}m_i] +$
$+ (2\beta/3 + \delta)a_{M(v)} + (2\gamma/3 + \delta)a_{T(v)} + (2/3)\Sigma_{d\in D'(v)}(a_d) + \Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

Let us denote $L_v = 6\, h^2 / [\, n_{.v}\, (\, r - h^2\,) + 1 - r\,]$, if the cow has known lactations, and $L_v = 0$, if the cow has no known lactation. We multiply the equation by 6:

$[n_{.v}\, L_v + 6\alpha + 8\beta + 8\gamma + 12\delta + 2N_v' + 3N_v'']\, a_v =$
$= L_v\, [y_{.v.} - \Sigma_i n_{iv}m_i] + [4\beta + 6\delta]a_{M(v)} +$
$+ \quad (4\beta + 6\delta)a_{M(v)} + (4\gamma + 6\delta)a_{T(v)} + 4\Sigma_{d\in D'(v)}(a_d) + 6\Sigma_{d\in D''(v)}(a_d - a_{T(d)}/2)$

(3b)

id est:
α=1 =>

$(n_{.v}\, L_v + 6 + 2N_v' + 3N_v'')a_v = L_v\, [y_{.v.} - \Sigma_i\, n_{iv} m_i\,] + 4\Sigma_{d \in D'(v)}(a_d) + 6\Sigma_{d \in D''(v)}(\, a_d - a_{T(d)}\, /2)$

$\beta=1 \Rightarrow$

$(n_{.v}\, L_v + 8 + 2N_v' + 3N_v'')a_v = L_v\, [y_{.v.} - \Sigma_i\, n_{iv} m_i\,] + + 4a_{M(v)} + 4\Sigma_{d \in D'(v)}\, a_d + 6\Sigma_{d \in D''(v)}\, (a_d - a_{T(d)}\, /2)$

$\gamma=1 \Rightarrow$

$(n_{.v}\, L_v + 8 + 2N_v' + 3N_v'')a_v = L_v\, [y_{.v.} - \Sigma_i\, n_{iv} m_i\,] + + 4a_{T(v)} + 4\Sigma_{d \in D'(v)}\, a_d + 6\Sigma_{d \in D''(v)}\, (a_d - a_{T(d)}\, /2)$

$\delta=1 \Rightarrow$

$(n_{.v}\, L_v + 12 + 2N_v' + 3N_v'')a_v = L_v\, [y_{.v.} - \Sigma_i\, n_{iv} m_i\,] + + 6a_{M(v)} + 6a_{T(v)} + 4\Sigma_{d \in D'(v)}\, a_d + 6\Sigma_{d \in D''(v)}\, (a_d - a_{T(d)}\, /2)$

Additive genetic effect of the bull j=t

By analogy with the formula proved for cows without known lactation, for bulls there is the following formula:

$(6\alpha + 8\beta + 8\gamma + 12\delta + 2N_t' + 3N_t'')a_t = (4\beta + 6\delta)a_{M(t)} + (4\gamma + 6\delta)a_{T(t)} + 4\Sigma_{d \in D'(t)}\, a_d + 6\Sigma_{d \in D''(t)}\, (a_d - a_{M(d)}\, /2)$

id est:

$\alpha=1 \Rightarrow$

$(\, 6 + 2N_t' + 3N_t'')a_t = 4\Sigma_{d \in D'(t)}\, a_d + 6\Sigma_{d \in D''(t)}\, (a_d - a_{M(d)}\, /2)$

$\beta=1 \Rightarrow$

$(\, 8 + 2N_t' + 3N_t'')a_t = 4a_{M(t)} + 4\Sigma_{d \in D'(t)}\, a_d + 6\Sigma_{d \in D''(t)}\, (a_d - a_{M(d)}\, /2)$

$\gamma=1 \Rightarrow$

$(\, 8 + 2N_t' + 3N_t'')a_t = 4a_{T(t)} + 4\Sigma_{d \in D'(t)}\, a_d + 6\Sigma_{d \in D''(t)}\, (a_d - a_{M(d)}\, /2)$

$\delta=1 \Rightarrow$

$(12 + 2N_t' + 3N_t'')a_t = 6a_{M(t)} + 6a_{T(t)} + 4\Sigma_{d \in D'(t)}\, a_d + 6\Sigma_{d \in D''(t)}\, (a_d - a_{M(d)}\, /2)$

B.3 References

DRĂGĂNESCU,C., 1982. Evaluation simultanée des taureaux et des vaches pour les caracteres de la production laitiere. XXXIII ejegodnaia conferentia Evropeiskcoi Assotiatii po jivotnovodstvu, 982-Leningrad, SSSR.

List of Tables

Index