Quantitative Methods Applied to Animal Breeding

Guilherme J. M. Rosa
Department of Animal and Dairy Sciences,
Department of Biostatistics & Medical
Informatics, University of Wisconsin-Madison,
Madison, WI, USA

Article Outline

Glossary

Bayesian inference Statistical inference approach based on the combination of prior information and evidence (i.e., observations) for estimation or hypothesis testing. In Bayesian analysis the prior information is updated with the experimental data to generate the posterior distribution of unknowns, such as model parameters. The name "Bayesian" comes from the use of the Bayes' theorem in the updating process.

Breeding value A measure of the genetic merit of an individual for breeding purposes.

Genetic correlation The correlation between traits that is caused by genetic as opposed to environmental factors. Genetic correlations can be caused by pleiotropy (genes that affect multiple traits simultaneously) or by linkage disequilibrium between genes affecting the different traits.

Genomic selection Genomic selection is a form of marker-assisted selection in which genetic markers covering the whole genome are used such that all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker.

Heritability (narrow sense) The fraction of the phenotypic variance that is due to additive genetic effects.

Infinitesimal genetic model A genetic model that assumes that a trait is influenced by a very large (effectively infinite) number of loci, each with infinitesimal effect.

Linkage disequilibrium Non-random association of alleles at two or more loci, leading to combinations of alleles (haplotypes) that are more or less frequent in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies.

Mixed models A mixed-effects model (or simply mixed model) is a statistical model containing both fixed and random effects. Such models are useful in a wide variety of disciplines in the physical, biological, and social sciences, especially for the analysis of data with repeated measurements on each statistical unit or with measurements taken on clusters of related statistical units.

Population genetics The study of allele frequency distribution and change under the influence of the four main evolutionary processes: selection, genetic drift, mutation, and migration.

Quantitative genetics The study of complex traits (e.g., production and reproductive traits, disease resistance) and their underlying genetic mechanisms. It is effectively an extension of simple Mendelian inheritance in that the combined effect of the many underlying genes results in a continuous distribution of phenotypic values or of some underlying scale or liability thereof.

R. A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*,
https://doi.org/10.1007/978-1-4939-2493-6_334-3

Definition of the Subject

The term *Animal Breeding* refers to the human-guided genetic improvement of phenotypic traits in domestic animals such as livestock and companion species [1]. The genetic improvement of production and reproductive traits, as well as of disease resistance traits, is essential for the sustainability of animal agriculture operations, not only in terms of their economic viability, but also increased animal welfare and decreased environmental impact of production. Animal breeding is based on principles of *Quantitative Genetics* [2–4] and aims to increase the frequency of favorable alleles and allelic combinations in the population, which is achieved through selection of superior individuals and specific mating systems strategies. Selection methods and mating strategies are developed by combining principles of quantitative and population genetics with sophisticated statistical methods and computational algorithms for integrating phenotypic, pedigree, and genomic information, along with the utilization of reproductive technologies that allow for larger progeny cohorts from superior animals as well as shorter generation intervals.

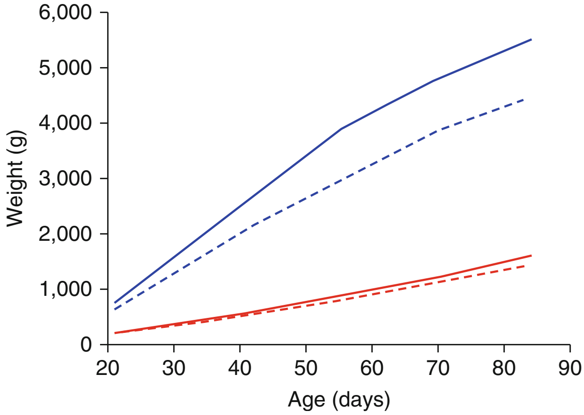
Through selection and mating of superior animals the frequency of favorable alleles is increased, so the overall additive genetic merit of a population is increased through successive generations [5]. Selection can be regarded as the most important tool for the improvement of lines or breeds within a specific species in terms of additive genetic effects. Such lines or breeds can be then inter-mated such that non-additive genetic effects such as dominance and epistasis can be exploited through specific inter- and intra-locus allelic combinations [1–4].

The theoretical foundations of population and quantitative genetics can be traced back to the work of R. A. Fisher, J. B. S. Haldane, and S. Wright. The rational animal breeding has its origins in the work of J. L. Lush, who made substantial contributions to animal genetics and biometrics research and is generally referred to as the father of modern scientific animal breeding [1].

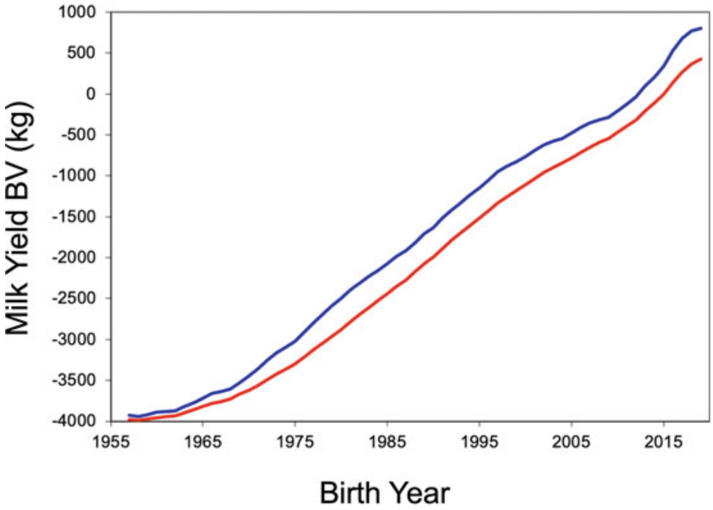
More recent theoretical developments in population and quantitative genetics have been fostered by researchers such as C. C. Cockerham, C. W. Cotterman, J. F. Crow, W. J. Ewens, W. G. Hill, M. Kimura, G. Malécot, T. Nagylaki, and B. S. Weir, among others. A landmark in the area of animal breeding and genetics is the development of mixed model methodology, first proposed by C. R. Henderson, which has been used extensively in many applications in the field, ranging from breeding value prediction under the infinitesimal assumption to gene mapping and segregation analysis. Most recently, Bayesian methods, Monte Carlo, and re-sampling techniques have been employed to fit and evaluate complex models in different contexts, including nonlinear systems, generalized models, survival analysis, and situations in which the number of parameters or covariates surpasses the number of observations, such as in association analysis and whole-genome marker-assisted selection using high density panels of single nucleotide polymorphism (SNP) markers.

Introduction

Since domestication, artificial selection has greatly changed the shape, size, and production and reproduction performance of livestock and companion animal species. For example, there is an incredible diversity of canine breeds –and between dogs and their wolf ancestor – from differences in overall appearance to behavior and their ability to perform specific tasks. Although to a lesser degree, the same can be observed in many other companion animal species, such as cats and horses. With livestock species, tremendous genetic changes have been accomplished as well, markedly in the last 60 years or so. For example, Fig. 1 depicts the average growth curves of broilers from selected and control populations. These results refer to a population of birds selected for over 40 years for increased growth rate, and another population kept without artificial selection, with both groups derived from the same base population, starting in 1957 [6]. In the experiment presented in Fig. 1, the two groups of birds



Quantitative Methods Applied to Animal Breeding, Fig. 1 Average growth curves of commercial broilers. Blue and red lines represent birds with "2001" and "1957" genetics, respectively. Solid and dashed lines represent birds fed diets typical of 2001 or 1957, respectively. (Adapted from Ref. [6])



Quantitative Methods Applied to Animal Breeding, Fig. 2 Genetic trend for milk yield in the US Holstein or Red & White populations. Males and females average breeding values are in blue and red, respectively; genetic base refers to cows born in year 2015. (Source: Council on Dairy Cattle Breeding – CDCB; https://www.uscdcb.com/)

were fed diets typical of 1957 and 2001, such that the interaction between genetics and feed, as well as the genetic contribution to the phenotypic differences observed, could be assessed. It is seen that the 2001 genetics group presented an average body weight of about 4 kg at 56 days of age, while its 1957 counterpart weighed only 800 g or so. Moreover, it is shown that 85–90% of this fivefold improvement is accounted for by genetics, with the remaining 10–15% to nutrition.

Similar levels of genetic improvement can also be observed in many other species, such as swine, beef and dairy cattle, and some species of fish. For example, as illustrated in Fig. 2, the average breeding value for milk yield in the US Holstein or Red & White populations has increased over 4400 kg in the last 60 years.

Such genetic improvements have been accomplished mostly through the selection and breeding of superior animals, which can be chosen using specific statistical methods such as those discussed in the subsequent sections. In this chapter, the discussion will focus on methods developed for normally distributed (Gaussian) traits, under the infinitesimal assumption, i.e., that traits are affected by a large (virtually infinite) number of genes of small effects [2–4], although this assumption is somewhat alleviated in marker assisted selection, which is discussed later.

Principles of Selection

Basic Genetic Model for Quantitative Traits

The basic genetic model can be expressed as [2, 3, 7]:

$$y_i = \mu + g_i + e_i \tag{1}$$

where y_i is the phenotypic value of animal i (i.e., the animal's performance for a specific trait); μ is the population mean (average performance of the animals); g_i is the genotypic value of the animal, expressed as a deviation from the mean; and e_i is a term representing environmental factors affecting the animal's performance, also expressed as a deviation from the mean. Hence, it is assumed that $E[g_i] = 0$ and $E[e_i] = 0$, such that $E[y_i] = \mu$, where $E[.]$ represents the expectation function. Moreover, the variance of y_i is given by $\mathrm{Var}[y_i] = \sigma_y^2 = \sigma_g^2 + \sigma_e^2$, where $\sigma_g^2 = \mathrm{Var}[g_i]$ and $\sigma_e^2 = \mathrm{Var}[e_i]$ are the genetic and environmental variances, respectively. Normally distributed traits, i.e., phenotypic traits with a bell-shaped distribution, are generally represented as $y_i \sim N\left(\mu, \sigma_y^2\right)$. Such distribution has a probability density function that can be described as [2, 4]:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left\{-\frac{1}{2\sigma_y^2}(y_i - \mu)^2\right\},$$

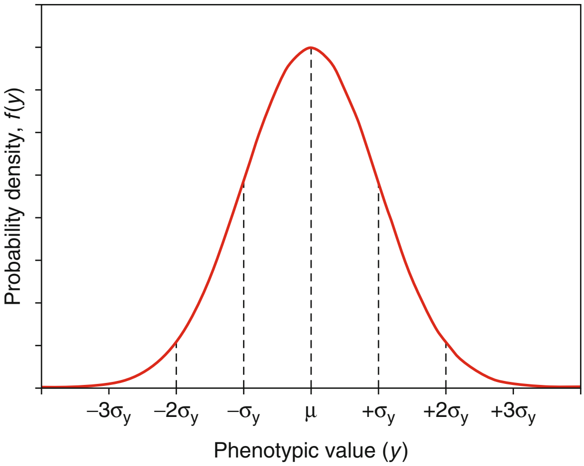
for $-\infty < y_i < \infty$, $-\infty < \mu < \infty$, and $\sigma_y^2 > 0$, which can be represented as in Fig. 3. To simplify the notation used throughout the text, it is noted that either random variables or their realizations will be represented with lower case letters. However, the context should make it clear to the reader when a letter represents one or the other.

The genetic component g_i of Model (1) can be partitioned into additive (a_i) and non-additive (c_i) genetics effects, i.e., $g_i = a_i + c_i$, where a_i is also called "breeding value," and c_i refers to the "gene combination value," which encompasses interaction effects between alleles within each locus (i.e., dominance effects) or between alleles in different loci (i.e., epistatic effects).

Hence, Model (1) can be expressed also as:

$$y_i = \mu + a_i + c_i + e_i, \tag{2}$$

where $a_i \sim N\left(0, \sigma_a^2\right)$, $c_i \sim N\left(0, \sigma_c^2\right)$, and $e_i \sim N\left(0, \sigma_e^2\right)$, with all these terms assumed independent from each other. The phenotypic variance can be then expressed as $\mathrm{Var}[y_i] = \sigma_y^2 = \sigma_a^2 + \sigma_c^2 + \sigma_e^2$, from which two important definitions are derived. The first one is called broad sense heritability (H^2), expressed as $\mathrm{H}^2 = \sigma_g^2/\sigma_y^2$, where $\sigma_g^2 = \sigma_a^2 + \sigma_c^2$, which represents the proportion of the phenotypic variance that is due to genetic effects. The second, called narrow sense heritability (h^2), refers to the specific contribution of additive genetic effects to the phenotypic variance, i.e., $h^2 = \sigma_a^2/\sigma_y^2$. These two quantities,



Quantitative Methods Applied to Animal Breeding, Fig. 3 Probability density function of a normally distributed trait with mean $\mu = E[y_i]$ and variance $\sigma_y^2 = \mathrm{Var}[y_i]$, i.e., $y_i \sim N\left(\mu, \sigma_y^2\right)$

particularly the narrow sense heritability, will be further discussed and used in the next sections.

The breeding value of an individual (a_i) is equal to the sum of the additive effects of individual alleles within and across loci, and it is sometimes called "additive genetic deviation" or "additive genetic effect". Because individual alleles, and therefore independent allele effects, are passed from parent to offspring, the breeding value of an individual is important for predicting its progeny's performance, and so it is central to selection of superior animals [1, 3]. The gene combination value (c_i) is the difference between the genetic merit (g_i) of an animal and its breeding value, i.e., $c_i = g_i - a_i$, so it is often called "non-additive genetic deviation". Because the component c_i involves interactions between alleles (both within and between loci), and only a single allele (as opposed to a pair of alleles) in each locus is transmitted from parents to offspring, non-additive effects are not transmitted in a predictable manner. Hence, while average breeding value in a population can be changed over generations with the selection of superior animals, the gene combination value should be explored through specific mating systems. In this Chapter, the discussion will focus on selection approaches and the genetic improvement of a population in terms of additive genetics effects only. For a discussion on mating systems, such as inbreeding and outbreeding strategies, see, for example, [1, 7, 8]. Additional discussion on inbreeding depression and heterosis (or hybrid vigor) can be found in [3, 4].

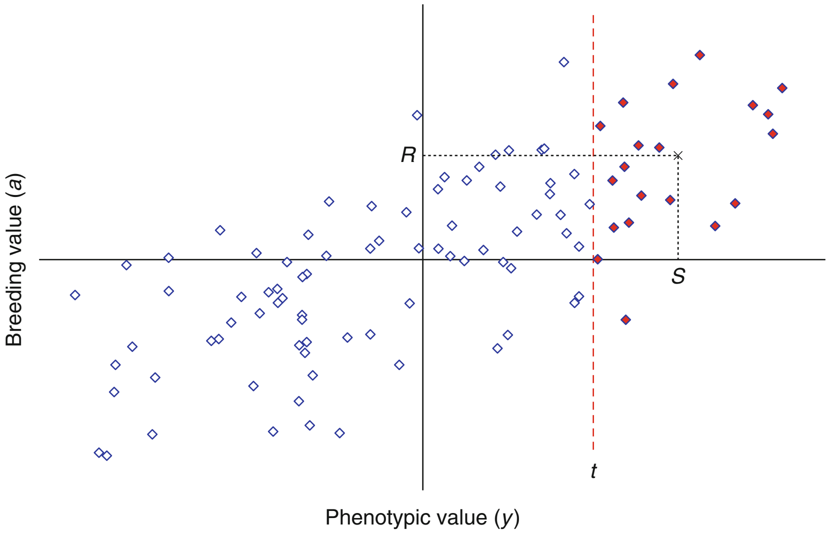
As discussed previously, the breeding value of an individual is equal to the sum of its independent allele effects. Because a parent passes a random sample of half of its alleles to its progeny, an animal's breeding value is twice what is often called "transmitting ability" or "expected progeny difference" [1, 5]. The expected breeding value of an offspring (a_o) is then equal to the average of its parents' breeding values (the same as the sum of its parents' transmitting abilities), i.e., $E[a_o|a_s, a_d] = \frac{a_s + a_d}{2}$, where a_s and a_d represent the (realized) breeding values of the offspring's sire and dam, respectively. However, there will be variability in terms of breeding values within a full-sib family because of the random sampling of parents' alleles that each offspring receives, the so-called Mendelian sampling [4].

The breeding value of an individual can be expressed as a function of its parents' breeding values as $a_o = 0.5a_s + 0.5a_d + \delta$, where δ refers to the Mendelian sampling component. It is interesting to notice that the variance of breeding values in a specific generation is equal to $\mathrm{Var}[a_o] = 0.25\mathrm{Var}[a_s] + 0.25\mathrm{Var}[a_d] + \mathrm{Var}[\delta]$. Assuming the same additive genetic variance across generations and for both sexes (i.e., $\mathrm{Var}[a_o] = \mathrm{Var}[a_s] = \mathrm{Var}[a_d] = \sigma_a^2$), it is shown that the Mendelian sampling variance is equal to half the additive genetic variance, i.e., $\mathrm{Var}[\delta] = \sigma_a^2/2$.

Phenotypic Selection

The most traditional approach of genetic improvement of livestock (and more generally any domestic animal or plant species) is based on selection of animals with the best performance, or "phenotypic selection" [1–4]. Accordingly, given a group of animals supposedly reared in similar environmental conditions, only those with the highest performance are allowed to breed to produce the next generation. As discussed previously (Model 2), the performance of each animal is a combination of its breeding value and all other non-additive genetics effects and environmental factors, such that a superior performance does not always represent superior breeding value. Nonetheless, whenever $\sigma_a^2 > 0$, there will be a positive correlation between performance and breeding value, and the phenotypic selection will result in genetic progress. Moreover, higher values of such a correlation will increase the genetic response, i.e., the effectiveness of phenotypic selection.

To illustrate this concept, consider Fig. 4, in which a scatter plot of breeding values and phenotypes (centered on zero, i.e., $y_i - \mu$) for a few fictitious animals is presented. As indicated before, in this chapter the discussion will be focused on selection approaches and the genetic improvement of a population in terms of additive genetics effects only, such that Model (2) can be conveniently re-expressed as:



Quantitative Methods Applied to Animal Breeding, Fig. 4 Scatter plot of breeding values versus phenotypic values. Each dot represents a specific animal and those colored in red are selected animals with performance (i.e., phenotypic value) above a specified threshold (*t*). *S* and *R* represent the average phenotypic and breeding values of the selected (top) animals, respectively

$$y_i = \mu + a_i + \varepsilon_i, \tag{3}$$

where $\varepsilon_i = c_i + e_i$ represents all non-additive genetic and environmental effects affecting the phenotypic value y_i, assumed $\varepsilon_i \sim N\left(0, \sigma^2_\varepsilon\right)$.

Assuming that each effect in Model (3) is independent from each other, the covariance between phenotype and breeding value is given by:

$$\mathrm{Cov}[y_i, a_i] = \mathrm{Cov}[\mu + a_i + \varepsilon_i, a_i] = \mathrm{Var}[a_i] \\ = \sigma^2_a,$$

such that the correlation between phenotype and breeding value is:

$$r_{y_i,a_i} = \frac{\mathrm{Cov}[y_i, a_i]}{\sqrt{\mathrm{Var}[y_i]\mathrm{Var}[a_i]}} = \frac{\sigma^2_a}{\sigma_y \sigma_a} = \frac{\sigma_a}{\sigma_y} = \sqrt{h^2},$$

i.e., the square root of the (narrow sense) heritability.

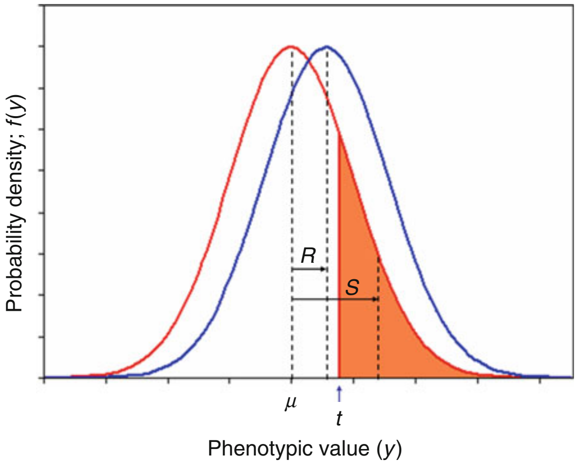
In practice, the breeding values of animals are unknown, so what phenotypic selection essentially does is to predict (or estimate) the animals' breeding values based on their own performance. The prediction is based on the regression of breeding values on phenotypes, and the regression coefficient (slope) is given by:

$$b_{a_i \cdot y_i} = \frac{\mathrm{Cov}[y_i, a_i]}{\mathrm{Var}[y_i]} = \frac{\sigma^2_a}{\sigma^2_y} = h^2.$$

This means that an animal's estimated breeding value (EBV) based solely on its performance (and with a single measurement only) can be expressed as:

$$\hat{a}_i = h^2 \times (y_i - \mu).$$

The correlation between such EBV (which is a linear transformation of y_i) and the true breeding value (a_i) is $r_{\hat{a}_i,a_i} = \frac{\mathrm{Cov}[\hat{a}_i, a_i]}{\sqrt{\mathrm{Var}[\hat{a}_i]\mathrm{Var}[a_i]}} = \frac{h^2\sigma^2_a}{\sqrt{h^4\sigma^2_y\sigma^2_a}} = h,$ which is generally referred to as 'prediction accuracy' in the animal breeding literature [5]. In this case, the square of the accuracy, which is often called "prediction reliability," is equal to the heritability of the trait. The prediction accuracy (and consequently the reliability) can be increased by using additional sources of information on an animal (such as repeated measurements of the trait or performance of progeny and other



Quantitative Methods Applied to Animal Breeding, Fig. 5 Probability density of the distribution of phenotypic values in the candidates-for-selection (red) and the progeny (blue) populations. The candidates-for-selection group represents the parental population (or generation 0), from which the top performing animals (above the threshold *t*) are selected and mated to produce the next generation or progeny (generation 1). The difference between the phenotypic average of the selected animals and that of the generation 0 is called selection differential (represented by *S*), and the difference between the phenotypic mean of the progeny and that of the generation 0 is called genetic progress, or genetic response (represented by *R*)

relatives) when estimating its breeding value. Computing EBVs using multiple sources of information can be accomplished using selection indexes and mixed model methodology, which will be discussed later in this chapter.

As indicated in Fig. 4, the selected animals (i.e., the best performing animals) will have an average phenotypic value equal to *S* and an average breeding value equal to *R*. The expected average breeding value (and also the expected phenotypic performance) of the progeny of the selected animals is also *R*, as illustrated in Fig. 5, and the ratio R/S is equal to the heritability (h^2) of the trait under selection. The genetic progress after one generation of selection is then given by:

$$R = h^2 \times S,$$

where $R = \mu_P - \mu$ and $S = \mu_S - \mu$, with μ_P, μ_S, and μ representing the average phenotypic performance of the progeny (generation 1), of the selected animals, and of the selection candidate (generation 0) populations, respectively.

The selection differential (*S*) can also be expressed as $S = i\sigma_y$, where $i = \frac{\mu_S - \mu}{\sigma_y}$ is called "selection intensity," and represents the selection differential in terms of phenotypic standard deviations. In addition, as *R* represents the genetic progress expected in a single generation of selection, the genetic improvement per unit of time is then given by $R^* = R/L$, where *L* is the generation interval. Hence, the expected genetic progress of single-trait phenotypic selection is given by [1, 3]:

$$R^* = \frac{h^2 \times i \times \sigma_y}{L},$$

which, given that $\sigma_y = \sigma_a/h$, can be expressed also as:

$$R^* = \frac{h \times i \times \sigma_a}{L}.$$

This equation is a special form of the so-called breeder's equation (or "key equation"), for the case of phenotypic selection. In its general form, the breeder's equation is expressed as [5]:

$$R^* = \frac{\text{accuracy} \times \text{intensity} \times \text{variation}}{\text{generation interval}},$$

meaning that the genetic progress per unit of time is proportional to the accuracy of breeding values prediction, to the selection intensity, and to the genetic variation, and inversely proportional to the generation interval.

Hence, to increase the genetic progress in a population (e.g., breed or line) through selection, animal breeders (and similarly plant breeders) work to improve the four components of the equation above. As the genetic variability is a natural characteristic of a population and cannot be easily changed, genetic progress is generally incremented either by improving prediction accuracy (e.g., by using specific statistical techniques to combine different sources of information regarding the animals' genetic merit), or by increasing the selection intensity, or by shortening the generation interval, which can be accomplished using molecular genetics techniques (e.g., the use of marker-assisted selection) and biotechnology approaches (e.g., artificial insemination).

It is important to mention that the breeder's equation discussed here can be extended for more complex scenarios, such as when males and females contribute differently to some components of the Eq. (5). For example, prediction accuracies and selection intensity are generally higher for males if artificial insemination is used. Another important issue to mention here is that selection not only shifts the mean of the breeding values in a population but it also changes the genetic variance (and heritability). A primary cause of the change in genetic variance is due to the fact that selected parents represent one tail of the phenotypic distribution, therefore their phenotypic variance is smaller than that of the whole candidates-for-selection population. This leads to a reduction in both the phenotypic and additive genetic variances in the progeny population, which is known as the "Bulmer effect" [2]. In addition, as selection modifies allele frequencies toward the fixation of favorable alleles, selection in one direction over many generations is also expected to reduce the genetic variation. Additional discussion on effects of selection on variance and other short- and long-term consequences of artificial selection can be found, for example, in [2, 3].

In the remainder of this chapter, specific statistical techniques (such as the selection index, BLUP, and genomic selection) for the improvement of accuracy, intensity, and generation interval, and consequently the increase of genetic progress from artificial selection will be discussed.

Correlated Response and Indirect Selection

If two traits x and y are genetically correlated, direct selection on one of the traits (say y) will also cause a genetic change in the other trait (trait x), which is called "correlated response" [3]. Correlated response to selection ($R_{x \cdot y}$), that is, genetic change in trait x as a consequence of direct selection on trait y, can be predicted by:

$$R_{x \cdot y} = b_{x \cdot y} R_y,$$

where R_y is the genetic progress of trait y through direct selection on itself, and $b_{x \cdot y}$ is the genetic regression coefficient, given by:

$$b_{x \cdot y} = \frac{\mathrm{Cov}\left(a_x, a_y\right)}{\sigma^2_{a_y}},$$

where Cov(a_x, a_y) is the genetic covariance between traits x and y.

The genetic correlation between two traits x and y is given by:

$$\rho_{a_x, a_y} = \frac{Cov\left(a_x, a_y\right)}{\sigma_{a_x} \sigma_{a_y}},$$

such that $\mathrm{Cov}\left(a_x, a_y\right) = \rho_{a_x, a_y} \sigma_{a_x} \sigma_{a_y}$, and the genetic regression can be expressed as:

$$b_{x \cdot y} = \frac{\rho_{a_x, a_y} \sigma_{a_x} \sigma_{a_y}}{\sigma^2_{a_y}} = \rho_{a_x, a_y} \frac{\sigma_{a_x}}{\sigma_{a_y}}.$$

Using this term, and recalling the selection response formula discussed before, given by $R_y = h_y i_y \sigma_{a_y}$, the correlated response can then be expressed as:

$$R_{x\cdot y} = \rho_{a_x,a_y} \frac{\sigma_{a_x}}{\sigma_{a_y}} h_y i_y \sigma_{a_y} = \rho_{a_x,a_y} \sigma_{a_x} h_y i_y,$$

or, given that $\sigma_{a_x} = h_x \sigma_{y_x}$, it can be finally written as:

$$R_{x\cdot y} = \rho_{a_x,a_y} h_x h_y i_y \sigma_{y_x}.$$

Such an equation can be used either to monitor potential genetic changes in correlated traits when performing direct selection on a specific trait of economic importance or, alternatively, to explore indirect selection strategies using indicator traits [5]. The latter use may be of interest when a trait of economic importance is difficult or expensive to measure, or it is expressed later in an animal's life, so it may be advantageous to select on a correlated trait, which would be the indicator trait. To assess the effectiveness of indirect selection relative to direct selection, one may look at the ratio of expected genetic progress per unit of time in each scenario, i.e.:

$$\frac{R_{x\cdot y}}{R_x} = \frac{\rho_{a_x,a_y} \sigma_{a_x} h_y i_y / L_y}{h_x i_x \sigma_{a_x} / L_x} = \frac{\rho_{a_x,a_y} h_y i_y L_x}{h_x i_x L_y}$$
$$= \rho_{a_x,a_y} \frac{h_y}{h_x} \frac{i_y}{i_x} \frac{L_x}{L_y}.$$

So, it can be seen that this ratio can be higher than 1 (meaning that the indirect selection is more effective than the direct selection) depending on the genetic correlation between the economic and indicator traits, the ratios of their heritabilities, and their potential selection intensities and generation intervals.

Selection Index

In section "Phenotypic Selection", selection based on a single measurement on each animal was discussed. However, it is not always possible to observe the phenotype for all animals, such as traits that are expressed in only one sex or that require the sacrifice of animals to be measured, etc. In addition, even when it is possible to measure the phenotypic trait in each animal, information from relatives can be used to obtain earlier or more reliable predictions of breeding values. In this section, the prediction of breeding values using different sources of information (e.g., multiple measurements of the trait in each animal and progeny performance) will be discussed, and a methodology (the selection index) that combines multiple sources of information into a single prediction for each animal will be presented.

When multiple measurements of the same trait are recorded (e.g., milk yield in multiple lactations), breeding values can be predicted using the average of observations $(\bar{y}_i)$ from each animal as $\hat{a}_i = b_{a_i \cdot \bar{y}_i}(\bar{y}_i - \mu)$. However, to derive the genetic regression of breeding value on average phenotypic value, Model (3) must be expanded to include an additional term, which is discussed next.

It can be empirically shown that the covariance (or resemblance) between repeated measurements on the same animal is larger than σ_a^2, which is what would be expected under the assumptions of Model (2). This additional source of covariance between records for the same animal refers to environmental factors that affect all records similarly, the so-called permanent environmental effects [1, 4]. Under these circumstances, the Model (2) can be extended to:

$$y_{ij} = \mu + a_i + c_i + p_i + e_{ij} \tag{4}$$

where y_{ij} represents the observation j ($j = 1, \ldots, n_i$) on animal i, with n_i being the total number of records on animal i; μ, $a_i \sim N\left(0, \sigma_a^2\right)$ and $c_i \sim N\left(0, \sigma_c^2\right)$ are as defined previously; p_i refers to the permanent environmental effects affecting records on animal i, assumed $p_i \sim N\left(0, \sigma_p^2\right)$; and $e_{ij} \sim N\left(0, \sigma_e^2\right)$ represents residual effects (temporary environmental effects) associated with observation y_{ij}. In addition, it is assumed that all random terms in Model (4) are independent from each other, i.e., Cov(a_i, c_i) = Cov(a_i, p_i) = Cov(c_i, p_i) = 0 and $\mathrm{Cov}\left(e_{ij}, e_{ij'}\right) = 0$, for any i, j, and j' $(j \neq j')$.

Under these settings, the average phenotypic value of an animal is given by $\bar{y}_i = \mu + a_i + c_i + p_i + \bar{e}_i$, where $\bar{e}_i = \frac{1}{n_i} \sum_{i=1}^{n_i} e_{ij}$, such that its variance is given by $\mathrm{Var}[\bar{y}_i] = \sigma_a^2 + \sigma_c^2 + \sigma_p^2 +$

σ_e^2/n_i, and the covariance between a_i and $\overline{y}_i$ is $\mathrm{Cov}(a_i, \overline{y}_i) = \sigma_a^2$. In this case, the regression of breeding values on phenotypic means is given by:

$$b_{a_i \cdot \overline{y}_i} = \frac{\mathrm{Cov}[a_i, \overline{y}_i]}{\mathrm{Var}[\overline{y}_i]} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_c^2 + \sigma_p^2 + \sigma_e^2/n_i}.$$

An important definition related to repeated measurements refers to repeatability (r), which is given by the intraclass correlation, i.e., the ratio of the within-individual (or between repeated measurements) to the phenotypic variances [1, 4]:

$$r = \frac{\sigma_a^2 + \sigma_c^2 + \sigma_p^2}{\sigma_y^2} = \frac{\sigma_a^2 + \sigma_c^2 + \sigma_p^2}{\sigma_a^2 + \sigma_c^2 + \sigma_p^2 + \sigma_e^2},$$

and measures the correlation between records on the same animal.

Noting that $r = 1 - \frac{\sigma_e^2}{\sigma_a^2 + \sigma_c^2 + \sigma_p^2 + \sigma_e^2}$, the variance of the average phenotypic value of an animal can be expressed as a function of the repeatability as $\mathrm{Var}[\overline{y}_i] = [r + (1 - r)/n_i]\sigma_y^2$, such that the genetic regression becomes:

$$b_{a_i \cdot \overline{y}_i} = \frac{\sigma_a^2}{[r + (1 - r)/n_i]\sigma_y^2} = \frac{n_i h^2}{1 + (n_i - 1)r}.$$

The prediction accuracy in this case, i.e., the correlation between an animal's estimated breeding value using repeated records and its true breeding value is given by:

$$\begin{aligned} r_{\hat{a}_i, a_i} = r_{\overline{y}_i, a_i} = & \frac{\mathrm{Cov}(\overline{y}_i, a_i)}{\sqrt{\mathrm{Var}(\overline{y}_i)\mathrm{Var}(a_i)}} \\ = & \frac{\sigma_a^2}{\sqrt{r + (1 - r)/n_i}\sigma_y \sigma_a} \\ = & h\sqrt{\frac{n_i}{1 + (n_i - 1)r}} = \sqrt{b_{a_i \cdot \overline{y}_i}}. \end{aligned}$$

Hence, it can be seen that compared with single record phenotypic selection, there is a gain in accuracy when predictions are based on repeated records, and that the gain will depend on the values of r and n_i; higher gain in accuracy is obtained when r is low and when n_i is high.

Another alternative to predict breeding values is to use progeny performance, which is often employed for predicting breeding values of males for traits for which records can be obtained only on females, such as milk yield. For example, let $\overline{y}_i$ be the average of single records on n_i progeny of sire i, and assume that the sire was mated to a random sample of females not related to him. In this case, each progeny record can be expressed as:

$$y_{ij} = \mu + \frac{1}{2}a_i + \frac{1}{2}d_{ij} + \delta_{ij} + \varepsilon_{ij},$$

where a_i is the breeding value of a specific sire i; d_{ij} is the breeding value of dam j ($j = 1, \ldots, n_i$) mated with sire i; and δ_{ij} and ε_{ij} refer to the Mendelian sampling and residual (non-additive genetic and environmental) components associated with the observation y_{ij}. Using this notation, the following model can be used to describe the progeny average of sire i:

$$\overline{y}_i = \mu + \frac{1}{2}a_i + \frac{1}{2}\overline{d}_i + \overline{\delta}_i + \overline{\varepsilon}_i \tag{5}$$

where $\overline{y}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} y_{ij}$, $\overline{d}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} d_{ij}$, $\overline{\delta}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} \delta_{ij}$, and $\overline{\varepsilon}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} \varepsilon_{ij}$.

Given that $E\left[\overline{d}_i\right] = 0$ and $E\left[\overline{\delta}_i\right] = 0$, the breeding value of sire i can be then predicted by $\hat{a}_i = b_{a_i \cdot \overline{y}_i}(\overline{y}_i - \mu)$, where $b_{a_i \cdot \overline{y}_i} = \mathrm{Cov}[a_i, \overline{y}_i]/\mathrm{Var}[\overline{y}_i]$. It is shown that:

$$\mathrm{Cov}(a_i, \overline{y}_i) = \mathrm{Cov}(a_i, a_i/2) = \sigma_a^2/2$$

and

$$\begin{aligned} \mathrm{Var}[\overline{y}_i] = & \mathrm{Var}\left[\frac{1}{2}a_i + \frac{1}{2}\overline{d}_i + \overline{\delta}_i + \overline{\varepsilon}_i\right] \\ = & \frac{1}{4}\sigma_a^2 + \frac{1}{4}\frac{\sigma_a^2}{n_i} + \frac{\sigma_a^2}{2n_i} + \frac{\sigma_\varepsilon^2}{n_i} \\ = & \frac{(n_i + 3)\sigma_a^2 + 4\sigma_\varepsilon^2}{4n_i} \\ = & \frac{(n_i + 3)h^2 + 4\left(1 - h^2\right)}{4n_i}\sigma_y^2 \\ = & \left[k + \frac{1 - k}{n_i}\right]\sigma_y^2, \end{aligned}$$

where $k = h^2/4$ is the intraclass correlation between half-sibs, such that the genetic regression coefficient is given by:

$$\begin{aligned} b_{a_i \cdot \bar{y}_i} = & \frac{\sigma_a^2/2}{[k + (1-k)/n_i]\sigma_y^2} \\ = & \frac{h^2\sigma_y^2/2}{\left[h^2/4 + \left(1 - h^2/4\right)/n_i\right]\sigma_y^2} \\ = & \frac{2n_i h^2}{4 + (n_i - 1)h^2}, \end{aligned}$$

and the prediction accuracy, by:

$$\begin{aligned} r_{a_i,\bar{y}_i} = & \frac{\mathrm{Cov}[a_i, \bar{y}_i]}{\sqrt{\mathrm{Var}[a_i]\mathrm{Var}[\bar{y}_i]}} = \frac{h^2\sigma_y^2/2}{\sqrt{h^2\sigma_y^2[k + (1-k)/n_i]\sigma_y^2}} \\ = & \sqrt{\frac{n_i h^2/4}{1 + (n_i - 1)k}} \\ = & \sqrt{\frac{n_i h^2}{4 + (n_i - 1)h^2}} = \sqrt{b_{a_i \cdot \bar{y}_i}/2}, \end{aligned}$$

which approaches unity (one) as the number of progeny records increases.

Up to this point, it has been discussed how breeding values can be predicted using different sources of information, such as an animal's own performance (either a single record or multiple measurements) or progeny performance. Other sources of information that could also be used are the performance of parents, sibling, or other kinds of relatives. However, in practice, what generally happens is that multiple sources of information are available simultaneously, so that the question becomes how to best combine all the information available in order to maximize prediction accuracy. Here, a classical approach will be discussed, the "selection index," and later on in this chapter a more general and modern alternative, based on mixed model methodology, will be presented.

Consider, for example, that there are three sources of information available on animal i (represented here as y_{i1}, y_{i2}, and y_{i3}, and expressed as deviations from their means). The goal is to predict the animal's breeding value with a linear combination of such information, i.e.:

$$\hat{a}_i = b_{i1}y_{i1} + b_{i2}y_{i2} + b_{i3}y_{i3},$$

so that the prediction accuracy (i.e., correlation between predicted and true breeding value) is maximized.

Maximization of $r_{\hat{a}_i,a_i}$ is equivalent to the maximization of $\log\left(r_{\hat{a}_i,a_i}\right)$, which is generally easier to accomplish. The log correlation can be expressed as (here, to simplify the notation, the index i indicating the animal is dropped):

$$\begin{aligned} \log\left(r_{\hat{a},a}\right) = & \log\left[\frac{\sigma_{\hat{a},a}}{\sqrt{\sigma_{\hat{a}}^2\sigma_a^2}}\right] \\ = & \log\left(\sigma_{\hat{a},a}\right) - \frac{1}{2}\sigma_{\hat{a}}^2 - \frac{1}{2}\sigma_a^2, \end{aligned}$$

where the covariance between $\hat{a}$ and a, and the variance of $\hat{a}$ are given respectively by:

$$\sigma_{\hat{a},a} = b_1\sigma_{y_1,a} + b_2\sigma_{y_2,a} + b_3\sigma_{y_3,a}$$

and

$$\begin{aligned} \sigma_{\hat{a}}^2 = & b_1^2\sigma_{y_1}^2 + 2b_1b_2\sigma_{y_1,y_2} + 2b_1b_3\sigma_{y_1,y_3} + b_2^2\sigma_{y_2}^2 \\ & + 2b_2b_3\sigma_{y_2,y_3} + b_3^2\sigma_{y_3}^2. \end{aligned}$$

Substituting these expressions into $\log\left(r_{\hat{a},a}\right)$, taking the partial derivatives of $\log\left(r_{\hat{a},a}\right)$ with respect to each of the regression coefficients b_j ($j = 1, 2, 3$), and setting them to zero, gives the following set of equations:

$$\begin{cases} \dfrac{\partial \log\left(r_{\hat{a},a}\right)}{\partial b_1} = \dfrac{\sigma_{y_1,a}}{\sigma_{\hat{a},a}} - \dfrac{b_1\sigma_{y_1}^2 + b_2\sigma_{y_1,y_2} + b_3\sigma_{y_1,y_3}}{\sigma_{\hat{a}}^2} \\ \dfrac{\partial \log\left(r_{\hat{a},a}\right)}{\partial b_2} = \dfrac{\sigma_{y_2,a}}{\sigma_{\hat{a},a}} - \dfrac{b_1\sigma_{y_1,y_2} + b_2\sigma_{y_2}^2 + b_3\sigma_{y_2,y_3}}{\sigma_{\hat{a}}^2} \\ \dfrac{\partial \log\left(r_{\hat{a},a}\right)}{\partial b_3} = \dfrac{\sigma_{y_3,a}}{\sigma_{\hat{a},a}} - \dfrac{b_1\sigma_{y_1,y_3} + b_2\sigma_{y_2,y_3} + b_3\sigma_{y_3}^2}{\sigma_{\hat{a}}^2} \end{cases}$$

which can be rearranged as:

$$\begin{cases} b_1\sigma^2_{y_1} + b_2\sigma_{y_1,y_2} + b_3\sigma_{y_1,y_3} = k\sigma_{y_1,a} \\ b_1\sigma_{y_1,y_2} + b_2\sigma^2_{y_2} + b_3\sigma_{y_2,y_3} = k\sigma_{y_2,a} \\ b_1\sigma_{y_1,y_3} + b_2\sigma_{y_2,y_3} + b_3\sigma^2_{y_3} = k\sigma_{y_3,a} \end{cases}$$

where $k = \sigma^2_{\hat{a}}/\sigma_{\hat{a},a}$.

Extending the system for any number m of components (i.e., sources of information), these equations can be expressed in matrix notation as:

$$\mathbf{Pb} = k\mathbf{c},$$

where $\mathbf{P} = \begin{bmatrix} \sigma^2_{y_1} & \sigma_{y_1,y_2} & \cdots & \sigma_{y_1,y_m} \\ \sigma_{y_1,y_2} & \sigma^2_{y_2} & \cdots & \sigma_{y_2,y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_1,y_m} & \sigma_{y_2,y_m} & \cdots & \sigma^2_{y_m} \end{bmatrix}$ is the variance-covariance matrix of the vector $\mathbf{y} = [y_1, y_2, \ldots, y_m]'$, $\mathbf{b} = [b_1, b_2, \ldots, b_m]'$ is the vector of regression coefficients (weights) of each source of information, and $\mathbf{c} = \left[\sigma_{y_1,a}, \sigma_{y_2,a}, \ldots, \sigma_{y_m,a}\right]'$ is the vector of covariances between each piece of information and the breeding value of the animal, such that the weights $\mathbf{b}$ of the index $\hat{a} = \mathbf{b}'\mathbf{y}$ are given by $\mathbf{b} = k\mathbf{P}^{-1}\mathbf{c}$.

It should be noted that the constant k does not change the relative size of the regression coefficients $\mathbf{b}$ or the value of $r_{\hat{a},a}$, so it can be set to 1. In fact, if instead of maximizing $r_{\hat{a},a}$, the average square prediction error $E[\hat{a} - a]^2$ is minimized, then $\sigma^2_{\hat{a}} = \sigma_{\hat{a},a}$ and the system (usually called selection index equations) becomes:

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{c}.$$

The correlation between the index and the true breeding value is given by $r_{\hat{a},a} = \sigma_{\hat{a},a}/\sqrt{\sigma^2_{\hat{a}}\sigma^2_a} = \sqrt{\sigma_{\hat{a},a}/\sigma^2_a} = \sqrt{\frac{1}{\sigma^2_a}\sum_{j=1}^{m} b_j\sigma_{y_j,a}}$.

Multiple-Trait Selection

Usually more than one trait is considered in a selection program, as multiple traits may be economically (or societally) important in a production system (e.g., [9]). There are many strategies for multi-trait selection, including the tandem approach (which selects rotationally one trait at a time) and the independent culling levels strategy (which sets minimum performance levels for each of the traits of interest), but they are generally sub-optimal.

Here, the selection of a combination of multiple traits evaluated in economic terms will be discussed. Such a combination of traits is generally called “aggregate breeding value” or “breeding objective”, and can be expressed as [3]:

$$T = \mathbf{w}'\mathbf{a} = w_1a_1 + w_2a_2 + \cdots + w_ka_k,$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_k)'$ is the vector of economic weights (expressed as net economic value per unit of trait) for k traits of linear economic value, and $\mathbf{a} = (a_1, a_2, \ldots, a_k)'$ is a vector of breeding values relative to the k traits defining T. Here again, to simplify the notation, the subscript indexing the animal is suppressed.

Suppose records are available for m traits, which may or may not be included in the k traits describing the breeding objective. The goal then is to predict T based on the m traits observed, using the so-called economic selection index. The theory of selection index was introduced in the previous subsection as a means of combining multiple sources of information to predict breeding values for a specific trait. Here, similar methodology will be considered, but it will be used instead to combine information from multiple traits to predict an overall economic merit for each animal. i.e.:

$$\hat{T} = I = \mathbf{v}'\mathbf{y} = v_1y_1 + v_2y_2 + \cdots + v_my_m,$$

where $\hat{T}$ is the predicted overall economic merit of an animal, $\mathbf{v} = (v_1, v_2, \ldots, v_m)'$ is the vector of weighting factors, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)'$ is the vector of phenotypic measurements.

An alternative for determining the weights $\mathbf{v} = (v_1, v_2, \ldots, v_m)'$ is to first predict separately the breeding values a_j, $j = 1, 2, \ldots, k$, for each trait involved in the breeding objective, using information from all traits with measurements, $\mathbf{y} = (y_1, y_2, \ldots, y_m)'$. Afterward, such predictions are substituted for the true breeding values in the breeding objective equation, and then coefficients are grouped accordingly.

The breeding values a_j for each trait can be predicted by $\hat{a}_j = b_{j1}y_1 + b_{j2}y_2 + \cdots + b_{jm}y_m$, in which the weights are obtained as usual, to maximize $r_{\hat{a}_j, a_j}$ or minimize $E\left[\hat{a}_j - a_j\right]^2$. The equations that define the weights for the prediction of a_j are then given by:

$$\left\{\begin{array}{llll} b_{j1}\sigma^2_{y_1} + & b_{j2}\sigma_{y_1,y_2} + & \cdots + b_{jm}\sigma_{y_1,y_m} & = \sigma_{y_1,a_j} \\ b_{j1}\sigma_{y_1,y_2} + & b_{j2}\sigma^2_{y_2} + & \cdots + b_{jm}\sigma_{y_2,y_m} & = \sigma_{y_2,a_j} \\ \vdots & \vdots & \vdots & \vdots \\ b_{j1}\sigma_{y_1,y_m} + & b_{j2}\sigma_{y_2,y_m} + & \cdots + b_{jm}\sigma^2_{y_m} & = \sigma_{y_m,a_j} \end{array}\right.$$

This procedure is repeated for all k traits in the breeding objective, and the predictions $\hat{\mathbf{a}} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k)'$ are then substituted for the true values $\mathbf{a} = (a_1, a_2, \ldots, a_k)'$ in the aggregate breeding value, i.e.:

$$\hat{T} = w_1\hat{a}_1 + w_2\hat{a}_2 + \cdots + w_k\hat{a}_k.$$

This overall index estimating T can be rewritten as $I = v_1y_1 + v_2y_2 + \cdots + v_my_m$, by using appropriate multiplications and grouping of coefficients, with each coefficient v_i given by $v_i = w_1b_{1i} + w_2b_{2i} + \cdots + w_kb_{ki}$, with $i = 1, 2, \ldots, m$.

Another way of deriving the weights $\mathbf{v} = (v_1, v_2, \ldots, v_m)'$ defining the economic selection index $I = \mathbf{v}'\mathbf{y}$ is to maximize the correlation $r_{T,\,I}$, which will generate the following equations:

$$\left\{\begin{array}{llll} v_1\sigma^2_{y_1} + & v_2\sigma_{y_1,y_2} + & \cdots + v_m\sigma_{y_1,y_m} & = \sigma_{y_1,T} \\ v_1\sigma_{y_1,y_2} + & v_2\sigma^2_{y_2} + & \cdots + v_m\sigma_{y_2,y_m} & = \sigma_{y_2,T} \\ \vdots & \vdots & \vdots & \vdots \\ v_1\sigma_{y_1,y_m} + & v_2\sigma_{y_2,y_m} + & \cdots + v_m\sigma^2_{y_m} & = \sigma_{y_m,T} \end{array}\right.$$

where $\sigma_{y_i,T}$ is the covariance between each measured trait i ($i = 1, 2, \ldots, m$) and the linear function $T = \mathbf{w}'\mathbf{a}$, i.e., the aggregate breeding value. It can be shown that both approaches for determining the weights $\mathbf{v} = (v_1, v_2, \ldots, v_m)'$ are equivalent.

Mixed Model Methodology

Introduction

Many statistical methods for analysis of genetic data are specific (or more appropriate) for phenotypic measurements obtained from planned experimental designs with balanced data sets. While such situations may be possible within laboratory or greenhouse experimental settings, data from natural populations and agricultural species are generally highly unbalanced and fragmented by numerous kinds of relationships. Culling of data to accommodate conventional statistical techniques (such as those discussed to this point) may introduce bias and/or lead to a substantial loss of information. The mixed model methodology, on the other hand, allows efficient estimation of genetic parameters (such as variance components and heritability) and breeding values while accommodating extended pedigrees, unequal family sizes, overlapping generations, sex-limited traits, assortative mating, and natural or artificial selection.

The single trait prediction methods discussed in the previous section use only a single source of information or, when multiple sources of information are available, they require them to be split into independent subgroups, i.e., specific groups of relatives such as half-sibs, full-sibs, progeny, etc. However, in practice the data may be extremely complex due to the intricate pedigree structure commonly found in livestock species, e.g., beef and dairy cattle populations. Other drawbacks of the selection index include an inability to account for genetic trend over time, and that the phenotypes must be pre-adjusted for environmental effects, which can be done, for example, using the average of contemporary groups of animals. However, contemporary group effects can be inferred only under the unrealistic assumption that they are genetically equal. Hence, a selection index can be reliably applied only to individual animals within same herd and born in same year.

In view of such limitations, linear mixed models (models including both fixed and random effects) and best linear unbiased prediction (BLUP) of breeding values were developed [10–12]. The BLUP methodology uses performance information from all known relatives to estimate breeding values, and it can be applied to whole herds or large populations using data from many years, and can also accommodate genetic differences between contemporary groups. Presently, mixed models are widely used in many fields of science as a flexible tool for the analysis of data where responses are clustered around some random effects, such that there is a natural dependence between observations in the same cluster [13]. Examples of applications of mixed models in genetics and genomics include gene mapping and association analysis (e.g., [14, 15]), and gene expression assays using microarrays [16, 17] or RT-PCR [18], to name a few.

In some applications of mixed models, the central objective is the estimation and hypothesis testing regarding fixed effects (e.g., treatment effects in an experimental study), in which case the random effects (e.g., block effects) are nuisance effects. In animal breeding, however, the main goal is the prediction of realized values of random effects (e.g., breeding values of animals), and the fixed effects are generally environmental factors that must be considered to adjust the observed phenotypic values. A third application or goal of mixed models is the estimation of variance components, such as genetic and environmental variances, or functions of them, such as heritability and repeatability.

In this section, some basics regarding mixed models are briefly reviewed, with some emphasis toward the prediction of random effects, and subsequently some specific applications of the mixed model methodology in animal breeding and genetics are presented.

A linear mixed effects model is defined as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{6}$$

where $\mathbf{y}$ is the vector of responses (observations), $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are known design or incidence matrices relating $\mathbf{y}$ to the vectors $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively, and $\boldsymbol{\varepsilon}$ is a vector of residual terms. Generally, it is assumed that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are independent from each other and normally distributed with zero-mean vectors and variance-covariance matrices $\mathbf{G}$ and $\boldsymbol{\Sigma}$, respectively.

As mentioned before, in animal breeding a central goal refers to the prediction of random effects (breeding values). In linear (Gaussian) models as in (6), such predictions are given by the conditional expectation of $\mathbf{u}$ given the data, i.e., $E[\mathbf{u}|\ \mathbf{y}]$. Given the model specifications above, the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{V} & \mathbf{ZG} \\ \mathbf{GZ}' & \mathbf{G} \end{bmatrix} \right),$$

where $\mathbf{V} = \mathbf{ZGZ}' + \boldsymbol{\Sigma}$.

From the properties of multivariate normal distributions, $E[\mathbf{u}|\ \mathbf{y}]$ is given by:

$$E[\mathbf{u}|\mathbf{y}] = E[\mathbf{u}] + \mathrm{Cov}[\mathbf{u}, \mathbf{y}']\mathrm{Var}^{-1}[\mathbf{y}](\mathbf{y} - E[\mathbf{y}]),$$

so that in this case:

$$\begin{aligned} E[\mathbf{u}|\mathbf{y}] &= \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{GZ}'(\mathbf{ZGZ}' + \boldsymbol{\Sigma})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

This expression, however, depends on the fixed effects values $\boldsymbol{\beta}$, which also need to be inferred from the data. The fixed effects are then typically replaced by their estimates, such that predictions are made based on the following expression:

$$\hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right).$$

To estimate the fixed effects $\boldsymbol{\beta}$, all random effects in Model (6) can be combined into a single vector $\boldsymbol{\xi} = \mathbf{Zu} + \boldsymbol{\varepsilon}$, such that the following fixed effects model is obtained $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$. It is shown that the expectation of the $\boldsymbol{\xi}$ term is $E[\boldsymbol{\xi}] = E[\mathbf{Zu} + \boldsymbol{\varepsilon}] = \mathbf{Z}E[\mathbf{u}] + E[\boldsymbol{\varepsilon}] = \mathbf{0}$, and that its variance is $\mathrm{Var}[\boldsymbol{\xi}] = \mathrm{Var}[\mathbf{Zu} + \boldsymbol{\varepsilon}] = \mathbf{Z}\mathrm{Var}[\mathbf{u}]\mathbf{Z}' + \mathrm{Var}[\boldsymbol{\varepsilon}] = \mathbf{ZGZ}' + \boldsymbol{\Sigma} = \mathbf{V}$. Under these settings, the distribution of $\mathbf{y}$ is multivariate normal with mean vector $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}$, i.e., $\mathbf{y}$~MVN ($\mathbf{X}\boldsymbol{\beta}$, $\mathbf{V}$), and the maximum likelihood estimator of $\boldsymbol{\beta}$ can be shown to be:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

which is distributed as $\hat{\boldsymbol{\beta}} \sim \mathrm{MVN}\left(\boldsymbol{\beta}, \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\right)$. If the design matrix $\mathbf{X}$ is not full column rank, a generalized inverse of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ must be used to obtain a solution $\boldsymbol{\beta}^0 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ of the system, from which estimable functions $\boldsymbol{\theta} = \mathbf{L}\boldsymbol{\beta}$ are estimated as $\hat{\boldsymbol{\theta}} = \mathbf{L}\boldsymbol{\beta}^0$.

The solutions $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ discussed before require $\mathbf{V}^{-1}$. As $\mathbf{V}$ can be of huge dimensions, especially in animal breeding applications, its inverse is generally computationally demanding if not unfeasible. However, Henderson [19] presented the mixed model equations (MME) to estimate $\boldsymbol{\beta}$ and $\mathbf{u}$ simultaneously, without the need for computing $\mathbf{V}^{-1}$. The MME were derived by maximizing (for $\boldsymbol{\beta}$ and $\mathbf{u}$) the joint density of $\mathbf{y}$ and $\mathbf{u}$, expressed as:

$$p(\mathbf{y},\mathbf{u}) \propto |\boldsymbol{\Sigma}|^{-1/2}|\mathbf{G}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\right.$$

$$\left.\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\right\}.$$

The logarithm of this function is:

$$\begin{aligned}\ell = \log\left[p(\mathbf{y},\mathbf{u})\right] &\propto |\,\boldsymbol{\Sigma}\,| + |\,\mathbf{G}\,| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})'\boldsymbol{\Sigma}^{-1} \\ &\quad (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u} \\ &= |\,\boldsymbol{\Sigma}\,| + |\,\mathbf{G}\,| + \mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - 2\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{u} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{u} \\ &\quad + \mathbf{u}'\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\mathbf{u} + \mathbf{u}'\mathbf{G}^{-1}\mathbf{u}\end{aligned}$$

The derivatives regarding $\boldsymbol{\beta}$ and $\mathbf{u}$ are:

$$\begin{bmatrix}\dfrac{\partial\ell}{\partial\boldsymbol{\beta}} \\ \dfrac{\partial\ell}{\partial\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y} - \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} - \mathbf{G}^{-1}\hat{\mathbf{u}}\end{bmatrix}.$$

Equating them to zero gives the following system:

$$\begin{bmatrix}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} + \mathbf{G}^{-1}\hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\end{bmatrix},$$

which can be expressed as:

$$\begin{bmatrix}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\end{bmatrix},$$

known as the mixed model equations (MME).

Using the second part of the MME,

$$\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \left(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)\hat{\mathbf{u}} = \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{y},$$

such that:

$$\hat{\mathbf{u}} = \left(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right).$$

It can be shown that this expression is equivalent to $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \boldsymbol{\Sigma})^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$ and, more importantly, that $\hat{\mathbf{u}}$ is the best linear unbiased predictor (BLUP) of $\mathbf{u}$. Using this result into the first part of the MME,

$$\begin{aligned}&\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ &\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1} \\ &\qquad\qquad \mathbf{Z}'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}\end{aligned}$$

$$\hat{\boldsymbol{\beta}} = \left\{\mathbf{X}'\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\right]\mathbf{X}\right\}^{-1}$$

$$\mathbf{X}'\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\right)^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\right]\mathbf{y}.$$

Similarly, it is shown that this expression is equivalent to $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, which is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

It is important to note that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ require knowledge of $\mathbf{G}$ and $\boldsymbol{\Sigma}$, or at least some function of them. As these matrices are rarely known, the practical approach is to replace $\mathbf{G}$ and $\boldsymbol{\Sigma}$ by some sort of point estimates $\hat{\mathbf{G}}$ and $\hat{\boldsymbol{\Sigma}}$ into the MME.

Many methods have been proposed to estimate variance components in mixed effects models. The simplest is the analysis of variance (ANOVA) method, which works well for simple models (such as a one-way structure) or balanced data (such as data from designed experiments with no missing data), but they are not indicated for more complex models and data structures such as those generally found in the animal breeding context.

Alternative methods proposed for estimating variance components in more complex scenarios include the expected mean squares approach of Henderson [20] and the minimum norm quadratic unbiased estimation [21]. However, maximum likelihood-based methods are currently the most popular (see, for example, [22], especially the restricted (or residual) maximum likelihood (REML) approach [23], which attempts to correct for the well-known bias in the classical maximum likelihood (ML) estimation of variance components. Additional literature on variance component estimation and mixed model methodology can be found, for example, in [24–28].

The Animal Model

The advent of mixed effect models has undoubtedly revolutionized the animal breeding field, and today they are widely used in the genetic improvement of many livestock and companion animal species. In this sub-section some of the applications of mixed models for the genetic evaluation of populations using phenotypic and pedigree information will be presented. In the following section, applications incorporating molecular maker information will be discussed as well.

As a first application of mixed models in animal breeding, the so-called animal model is considered here, for the specific situation of a single trait and a single phenotypic observation (including missing values) per animal. The animal model can be described as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of observations (phenotypic scores), $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects (e.g., herd-year-season effects in cattle evaluations), and $\boldsymbol{\varepsilon}$ represents residual effects, assumed $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ as before. In most applications of animal models, however, residuals are assumed independent across animals, such that the residual covariance structure can be expressed as $\mathbf{R} = \mathbf{I}\sigma^2_\varepsilon$, where $\mathbf{I}$ is an identity matrix of appropriate order, and σ^2_ε is the residual variance. In the case of animal models, the random effects $\mathbf{u}$ represent the breeding values, i.e., $\mathbf{u} = \mathbf{a}$, assumed to be $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G})$. The vector $\mathbf{a}$, of dimension $(q \times 1)$, may include breeding values of all animals with record or in the pedigree file, such that q is generally bigger than n.

The matrix $\mathbf{G}$, which in this case describes the covariances among the breeding values, follows from standard results for the covariances between relatives. It can be shown that the additive genetic covariance between two relatives i and i' is given by $2\theta_{ii'}\sigma^2_a$, where $\theta_{ii'}$ is the coefficient of co-ancestry between individuals i and i', and σ^2_a is the additive genetic variance in the base population [29]. Hence, under the animal model, $\mathbf{G} = \mathbf{A}\sigma^2_a$, where $\mathbf{A}$ is the "additive genetic (or numerator) relationship matrix," having elements given by $a_{ii'} = 2\theta_{ii'}$.

As mentioned earlier, in animal breeding the usual main interest is prediction of breeding values – for selection of superior individuals– and on estimation of variance components. The fixed effects are, in some sense, nuisance factors with no central interest in terms of inferences, but which need to be taken into account (i.e., they need to be corrected for when inferring breeding values).

Because under the animal model $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2}$ and $\mathbf{R}^{-1} = \mathbf{I}\sigma_\varepsilon^{-2}$, the mixed model equations reduce to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

where $\lambda = \frac{\sigma^2_\varepsilon}{\sigma^2_a} = \frac{1-h^2}{h^2}$, such that:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}.$$

It is worth mentioning that $\mathbf{A}^{-1}$ can be obtained directly from the pedigree, without setting up $\mathbf{A}$ [30, 31], which is computationally very convenient.

Conditionally on the variance components ratio λ, the BLUP of the breeding values are then given by $\hat{\mathbf{a}} = \left(\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1}\right)^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$, which are the EBVs. Alternatively, some breeders' associations express their results as predicted transmitting abilities (PTA) or expected progeny differences (EPD), which are equal to half the EBVs, representing the portion of an animal's breeding values that is passed to its offspring.

The amount of information contained in an animal's genetic evaluation depends on the availability of its own record, and of phenotypic information from its relatives (including how many and how closely related). As a measure of amount of information in livestock genetic evaluations, EBVs are typically reported with their associated accuracies, i.e., the correlation between true and estimated breeding values, $r_i = r_{\hat{a}_i, a_i}$. Instead of accuracy, some livestock species genetic evaluations use reliability, which is the squared accuracy (r_i^2).

A model-derived calculation of r_i requires the diagonal elements of the inverse of the MME coefficient matrix, represented as:

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta a} \\ \mathbf{C}^{a\beta} & \mathbf{C}^{aa} \end{bmatrix}.$$

It is shown that the prediction error variance (PEV) of EBV $\hat{a}_i$ is given by:

$$\text{PEV} = \text{Var}(\hat{a}_i - a_i) = c_i^{aa}\sigma_\varepsilon^2,$$

where c_i^{aa} is the i-th diagonal element of $\mathbf{C}^{aa}$, relative to animal i. The PEV can be interpreted as the fraction of additive genetic variance not accounted for by the prediction. Therefore, PEV can also be expressed as:

$$\text{PEV} = \left(1 - r_i^2\right)\sigma_a^2,$$

such that $c_i^{aa}\sigma_\varepsilon^2 = \left(1 - r_i^2\right)\sigma_a^2$, from which the reliability is obtained as $r_i^2 = 1 - c_i^{aa}\sigma_\varepsilon^2/\sigma_a^2 = 1 - \lambda c_i^{aa}$.

Extensions and Variations of the Animal Model

The animal model discussed above can be extended also to multiple (correlated) traits [32, 33]. For instance, consider as an example the analysis of k traits, in which the model for each trait is expressed as:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_j + \boldsymbol{\varepsilon}_j,$$

where $\mathbf{y}_j$, $\mathbf{X}_j$, $\boldsymbol{\beta}_j$, $\mathbf{Z}_j$, $\mathbf{a}_j$, and $\boldsymbol{\varepsilon}_j$ are defined as before, but here have an additional index to indicate the trait ($j = 1, 2, \ldots, k$).

For a joint analysis of the k traits, the single trait models can be combined as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = [\mathbf{y}_1{}' \ \mathbf{y}_2{}' \ \ldots \ \mathbf{y}_k{}']'$, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1{}' \ \boldsymbol{\beta}_2{}' \ \ldots \ \boldsymbol{\beta}_k{}']'$, $\mathbf{a} = [\mathbf{a}_1{}' \ \mathbf{a}_2{}' \ \ldots \ \mathbf{a}_k{}']'$, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1{}' \ \boldsymbol{\varepsilon}_2{}' \ \ldots \ \boldsymbol{\varepsilon}_k{}']'$, and the design matrices in this case are:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_k \end{bmatrix} \text{ and } \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_k \end{bmatrix}.$$

It is assumed that $\text{Var}\begin{bmatrix} \mathbf{a} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \otimes \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \otimes \mathbf{I} \end{bmatrix}$, where

$$\mathbf{G} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} & \cdots & \sigma_{a_1a_k} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} & \cdots & \sigma_{a_2a_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_1a_k} & \sigma_{a_2a_k} & \cdots & \sigma^2_{a_k} \end{bmatrix} \text{ and } \mathbf{\Sigma} = \begin{bmatrix} \sigma^2_{\varepsilon_1} & \sigma_{\varepsilon_1\varepsilon_2} & \cdots & \sigma_{\varepsilon_1\varepsilon_k} \\ \sigma_{\varepsilon_1\varepsilon_2} & \sigma^2_{\varepsilon_2} & \cdots & \sigma_{\varepsilon_2\varepsilon_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\varepsilon_1\varepsilon_k} & \sigma_{\varepsilon_2\varepsilon_k} & \cdots & \sigma^2_{\varepsilon_k} \end{bmatrix}$$

are the genetic and residual variance-covariance matrices, respectively, $\mathbf{A}$ and $\mathbf{I}$ are the numerator relationship matrix and an identity matrix, and $\otimes$ represents the direct (Kronecker) product.

The MME for multi-trait analyses are of the same form as before, i.e.:

$$\begin{bmatrix} \mathbf{X}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{X} & \mathbf{X}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{Z} \\ \mathbf{Z}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{X} & \mathbf{Z}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{Z}+\mathbf{G}^{-1}\otimes\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{y} \\ \mathbf{Z}'\left(\mathbf{\Sigma}^{-1}\otimes\mathbf{I}\right)\mathbf{y} \end{bmatrix},$$

from which the BLUEs and BLUPs of $\boldsymbol{\beta}$ and $\mathbf{a}$ can be obtained, respectively.

The dimensionality of such multi-trait MME, however, can become a hurdle for solving it when more than two or three traits are considered. An alternative for the analysis of multiple traits is to use a canonical transformation of the traits [34–36], which consists of transforming the vectors of correlated traits into a new vector of uncorrelated variables. In such case, each transformed variable can be analyzed independently using standard single trait models, and subsequently the estimated breeding values are transformed back to the original scale of measurement.

Some other interesting applications of mixed models in animal breeding involve multiple random effects, as in the cases of repeated measurements of the same trait or traits with maternal effects. For the analysis of repeated measurements, as discussed in subsection "Selection Index" (Model 4), environmental effects can be partitioned into permanent and temporary effects. In this case, the mixed model, usually called "repeatability model," can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Za} + \mathbf{Wp} + \boldsymbol{\varepsilon},$$

where all terms are as previously defined for a single trait animal model, and $\mathbf{p}$ is the vector of permanent environmental effects, with each level pertaining to a common effect to all observations of each animal, and $\mathbf{W}$ is a known incidence matrix relating $\mathbf{y}$ to the vector $\mathbf{p}$.

It is often assumed that $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{A}\sigma^2_a\right)$, $\mathbf{p} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_p\right)$, and $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2_\varepsilon\right)$, which are independent from each other. Under these assumptions, the MME becomes:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{W} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\lambda_a\mathbf{A}^{-1} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W}+\lambda_p\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix},$$

where $\lambda_a = \sigma^2_\varepsilon/\sigma^2_a$ and $\lambda_p = \sigma^2_\varepsilon/\sigma^2_p$.

There are some traits of interest in livestock, such as weaning weight in beef cattle, in which progeny performance is affected by the dam's ability to affect the calf's environment, such as in the form of nourishment through her milk production, the quantity and quality of which is in part genetically determined. In some cases, there can be also a paternally provided environmental component. In such cases, parents contribute to the performance of their progeny not only through the genes passed to the progeny (the "direct genetic effects") but also through their ability to provide a suitable environment (the "indirect genetic effects").

Here, maternally influenced traits are considered, for which the mixed model can be written as [37]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Za} + \mathbf{Km} + \mathbf{Wp} + \boldsymbol{\varepsilon},$$

where all terms are as before, except that the model now includes a vector **m** of random maternal genetic effects, and a vector **p** of random maternal permanent environmental effects, with **K** and **W** as their respective incidence matrices. It is assumed that $\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_a^2\right)$, $\mathbf{m} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_m^2\right)$, $\mathbf{p} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_p^2\right)$, and $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right)$, and quite often a covariance structure between direct and maternal additive genetic effects is considered, assumed equal to $\mathbf{A}\sigma_{a,\,m}$.

Some other variations of the animal model, which are computationally convenient, include the "sire model" and the "reduced animal model" [38]. In the sire models, only sires are evaluated, using progeny records under the assumption of randomly selected mates. In the reduced animal model, instead of having equations set up for every animal (i.e., parents and progeny), it allows equations to be set up only for parents in the MME, making the dimensions of the system greatly reduced. The breeding values of the parents are estimated directly from the MME, and the progeny breeding values are then inferred by back solving from the predicted parental breeding values.

As a final note regarding the use of mixed models in animal breeding, it is important to mention that solving the MME does not necessary require the inversion of the coefficient matrix **C**. More computationally convenient alternatives for solving high dimensional systems of linear equations include methods based on iteration on the MME, such as the Jacobi or Gauss-Seidel iteration [39], and the "iteration on the data" strategy [40], which is the commonly used methodology in national genetic evaluations involving millions of records.

Marker-Assisted Selection

Introduction

The advent of molecular markers has created opportunities for a better understanding of genetic inheritance and for developing novel strategies for genetic improvement in agriculture. Molecular markers are used, for example, to study quantitative trait loci (QTL), which are defined as chromosomic regions contributing to variation in phenotypic traits. The location and effects of QTL can be inferred by combining information from marker genotypes and phenotypic scores of individuals, and by exploring genetic linkage [41–44] and linkage disequilibrium [45, 46] information between marker loci and QTL, such as in experimental or mapping populations (e.g., backcross or F_2, or granddaughter designs) or in complex pedigrees in outbred populations. Information on markers associated with QTL can be used to enhance prediction of genetic merit of animals [47]. This is especially useful for low heritability traits, traits that are expensive or difficult to measure, or traits expressed in only one sex [48].

Classical Approaches with Few Markers

The application of molecular information for genetic improvement of animals and plants, or marker-assisted selection (MAS), requires that candidate-for-selection individuals are genotyped for specific markers. For MAS purposes, there are three types of genetic markers, and for each type there are specific statistical approaches for incorporating their information into selection programs [48]. A first type of marker refers to situations in which the functional polymorphism itself can be genotyped. These markers are called "direct markers," as they indicate exactly the genotype an animal has at specific causative loci.

A second type of marker refers to those that are in population-wide linkage disequilibrium (LD) with the causative or functional mutations. In such cases, although the marker genotype of an animal does not unambiguously indicate the genotype at a specific functional locus, it still provides information regarding how likely an animal carries a specific allele or genotype at such a locus. Finally, a third kind of molecular marker refers to those loci that are in population linkage equilibrium with the functional mutations, which are often called "indirect markers". In such cases, although the marker information on a single animal in a population does not provide any information regarding the genetic merit of that animal, it still can useful in exploring family (pedigree) structure when genotyped animals are related to each other.

While direct markers are the simplest and most efficient in MAS programs, their identification is much more difficult and generally involves a pre-screening step using QTL mapping methods to identify promising chromosomic regions, followed by fine mapping (often using functional and positional candidate gene strategies), followed by validation (using some strategy such as a knock-out approach). On the other extreme, indirect markers are extensively available for most livestock species, but their use in MAS is more complex and the results are generally modest.

Statistical models to incorporate direct and/or LD markers in the genetic evaluations of animals are relatively straightforward. For example, a marker can be included into an animal model context with the following specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}^{*} + \mathbf{M}\mathbf{g} + \boldsymbol{\varepsilon},$$

where all terms are as defined before, except that $\mathbf{a}^{*} \sim N\left(\mathbf{0}, \mathbf{A}\sigma^2_{a^{*}}\right)$ represents now the random additive (non-marker) polygenic effects, and $\mathbf{g}$ and $\mathbf{M}$ are the (fixed) QTL effects and an incidence matrix, respectively. In the case of direct markers, the matrix $\mathbf{M}$ represents the marker genotypes and is obtained directly from the genotyping of animals. In the case of LD markers, the incidence matrix $\mathbf{M}$ will represent genotype probabilities at each QTL locus, which can be derived using segregation analysis. The overall genetic merit of the animals are then given by the sum of their $\mathbf{a}^{*}$ and $\mathbf{g}$ components. Other strategies for combining the infinitesimal and the QTL components to increase long-term genetic gain have also been proposed (e.g., [49–51]); a review of MAS strategies can be found, for example, in [48]).

In the case of indirect markers, however, the within-family LD between QTL and linked markers must be explored. One approach is to determine the marker effects or the marker-QTL linkage phases separately for each family. Alternatively, more general MAS models have been proposed to incorporate marker data in genetic evaluations for complex pedigrees [14, 52], which can be represented as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}^{*} + \mathbf{M}\mathbf{q} + \boldsymbol{\varepsilon},$$

where the terms are as before, but here the QTL effects $\mathbf{q}$ are assumed random and normally distributed, such that:

$$\begin{bmatrix} \mathbf{a}^{*} \\ \mathbf{q} \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{A}\sigma^2_{a^{*}} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\lambda}\sigma^2_q \end{bmatrix}\right),$$

where $\mathbf{G}_{\lambda}$ is the gametic relationship matrix for the QTL, and σ^2_q is the additive variance of the QTL allelic effects. The gametic relationship matrix gives the probabilities of identity between each of the two alleles in each individual, and it can be derived based on the QTL position λ and the marker information.

Genomic Selection

As most quantitative traits are influenced by many genes, tracking a small number of them using molecular markers (as in the MAS approaches discussed above) will explain only a small fraction of the total genetic variance. Moreover, individual genes are likely to have small effects and so a large amount of data is needed to accurately estimate their effects [53]. Genome-wide Marker-Assisted Selection (GWMAS), or simply Genomic Selection (GS), on the other hand, makes use of a very dense set of markers covering the entire genome, which potentially explain all genetic variance. In addition, given the LD between the dense markers and the QTL, estimated marker effects pertain across the population [54].

Meuwissen et al. [55] were the first to propose GS and suggested a model that can be described as:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{j=1}^{p} \mathbf{m}_j q_j + \boldsymbol{\varepsilon},$$

where **y** is a vector of phenotypic observations; μ is an intercept and **1** is a vector of ones; q_j represents the genetic effect captured by each of a large number ($j = 1, 2, \ldots, p$) of biallelic markers (e.g., SNP loci); and $\mathbf{m}_j$ represents the vector of genotypes for each genetic marker (coded for example as 0, 1 and 2), which present different levels of LD with QTL affecting the phenotypic trait of interest (**y**). Here, it is assumed that the QTL affecting the trait act additively, and that q_j refers to per-allele effects; non-additive effects as well as effects relative to non-marked QTL are lumped together into the residual term of the model.

Fitting such GS model using standard regression approaches is not trivial, as the number p of markers (and so the number of genetic effects to be estimated) may easily exceed the number n of individuals available. The "large p small n paradigm" is central in many applications of genomic technologies, including expression profiling and association analysis, and various statistical strategies have been proposed in the literature to overcome this problem, such as dimension-reduction techniques, stepwise fitting procedures, ridge regression [56], and least absolute selection and shrinkage operator – LASSO [57].

Alternatively, GS regression models can be implemented using some sort of hierarchical Bayesian modeling, given its flexibility and good statistical properties. Within this approach, the genetic effects $\mathbf{q}_j$ are assumed random and distributed according to some pre-specified distribution [55]. For example, $\mathbf{q}_j$ may be assumed normally distributed with mean 0 and variance σ_j^2, and the hierarchy can be extended by assuming a prior distribution for the variances σ_j^2 [55, 58–60]. Alternative distributions can be adopted for $\mathbf{q}_j$, such as double exponential or mixture distributions including a mass point at zero. It is interesting to notice the connection between the ridge regression approach and a Bayesian model with normal priors with common variances $\sigma_j^2 = \sigma_0^2$, as well as the LASSO methodology and a Bayesian model with double exponential priors for the genetic effects [61].

Another approach to fit a GS model is to use the genetic marker information to build a genomic relationship matrix **G** describing genetic relatedness among individuals, and replace **G** for the pedigree-based relationship matrix **A** used in the animal model discussed previously. This approach is called GBLUP, and the genomic relationship matrix **G** is generally computed as $\mathbf{G} = c^{-1} \times \mathbf{MM}'$, where **M** is an $(n \times p)$ matrix of genotypes with each column (i.e., each marker) centered on zero, and $c = 2\sum_{j=1}^{p} p_j\left(1 - p_j\right)$, in which p_i represents the frequency of a reference allele in each marker [62]. Moreover, some other recent methods aim to combine all available phenotypic, pedigree, and genomic information for prediction of genetic merit of animals [63].

The potential of GS to accelerate genetic progress has been demonstrated through many simulation studies (e.g., [55, 64, 65]), and more recently confirmed with real data applications. The first use of GS using thousands of markers in livestock has been in dairy cattle [66, 67], followed by some breeds of beef cattle and more recently in poultry and pigs. Table 1 shows some early results with dairy cattle obtained by the USDA over 10 years ago. Since then GS has been implemented in commercial breeding programs across various livestock and crop species.

Future Directions

As shown here, the mixed model methodology is extremely flexible and can be used in a wide variety of applications in quantitative genetics and genomics. Other extensions of the methods discussed here include models with non-additive genetic effects (e.g., [68, 69]), mixed models for the analysis of non-Gaussian traits such as binary and categorical (e.g., [70, 71]) or counting data (e.g., [72]), robust models [73, 74], survival traits [75], nonlinear models to study, e.g., growth curves (e.g., [76, 77]), among others. However, such models can get extremely complex and asymptotic statistical methods are generally required. Alternatively, Bayesian analysis employing Markov Chain Monte Carlo (MCMC) methods can be used, given their

Quantitative Methods Applied to Animal Breeding, Table 1 Comparison of April 2010 genomic and traditional evaluations for bulls with an AI status of active or foreign

	Average reliability (%)		
Trait	Genomic	Traditional	Difference
Net merit	87	81	+6
Milk yield	93	91	+2
Fat yield	93	91	+2
Protein yield	93	91	+2
Productive life	81	71	+9
Somatic cell score	88	83	+5
Daughter pregnancy rate	79	69	+10
Final score	89	85	+4
Sire calving ease	90	84	+6
Daughter calving ease	80	67	+13

Source: AIPL – USDA; http://www.aipl.arsusda.gov/

exceptional flexibility and the possibility of incorporating prior information regarding the model parameters [78]. Bayesian analysis has been increasingly used in genetics and animal breeding; for a review the reader can refer, for example, to [79–81]. A comprehensive treatment of Bayesian MCMC approaches in animal breeding is presented in [24]. As discussed earlier, Bayesian hierarchical modeling has been extensively used also in genomic selection [55, 82–84]. In addition, non-parametric and semi-parametric methods, and machine learning techniques based on artificial intelligence have been used for the analysis of high-density marker panels in the context of animal breeding, such as in [85–91].

As indicated in the beginning of this chapter, the genetic improvement observed in many livestock and companion animal species over the years is truly remarkable. Most of this genetic progress has been accomplished through selection, using the methods discussed here. Two technological and methodological developments, however, must be mentioned as turning points in the genetic trends observed in some species; these are the advent of artificial insemination and the mixed models. Seemingly, the development of high-density SNP panels and, more recently, next generation sequencing technologies and their application in genomic selection strategies promise to be the next turning point. Another new area of research and interest to commercial breeding programs refers to the genetic improvement of novel (and hard to measure) traits, such as feed intake and feed efficiency, methane emission, product quality (e.g., meat and milk fatty acids, and milk protein profile), and even animal behavior traits. Such phenotypic traits are generally measured or monitored using high-throughput phenotyping (HTP) techniques based on digital sensor technologies [92, 93], such as image analysis and computer vision [94], and infrared spectroscopy [95].

This new era of animal breeding and genetics demands a multidisciplinary approach for both the development and the deployment of modern tools and techniques for efficient genetic improvement of livestock populations to enhance its sustainability, especially in terms of customers' requirements regarding animal welfare standards and product quality, and also environmental stewardship of animal production. As such, an efficient and contemporary breeding program requires nowadays not only expertise on population and quantitative genetics, and traditional statistical and computational methods, but also on biosystems engineering and on modern data mining techniques [61, 96, 97] suitable for large databases including multiple sources of information (phenotypic, genomic, and environmental variables) and data structures (tabular data, images, text, etc.). It is indeed a very exciting time to work in animal breeding!

Bibliography

Primary Literature

1. Lush JL (1994) The genetics of populations. Prepared for publication by A. B. Chapman and R. R. Shrode, with an addendum by J. F. Crow. Special Report 94, College of Agriculture, Iowa State University, Ames, IA
2. Bulmer MG (1985) The mathematical theory of quantitative genetics. Clarendon, Oxford
3. Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow
4. Lynch M, Walsh B (1998) Genetic analysis of quantitative traits. Sinauer Associates, Sunderland
5. Hill WG (1969) On the theory of artificial selection in finite populations. Genet Res 13:143–163
6. Havenstein B, Ferket PR, Qureshi MA (2003) Growth, livability, and feed conversion of 1957 versus 2001 broilers when fed representative 1957 and 2001 broiler diets. Poult Sci 82:1509–1518
7. Bourdon RM (2000) Understanding animal breeding, 2nd edn. Prentice Hall, Upper Saddle River
8. Crow J, Kimura M (1970) An introduction to populations genetics theory. Haraper and Row, New York
9. Shook GE (2006) Major advances in determining appropriate selection goals. J Dairy Sci:1349–1361
10. Henderson CR (1949) Estimation of changes in herd environment. J Dairy Sci 32:709
11. Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447
12. Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph
13. Gianola D, Rosa GJM (2015) One hundred years of statistical developments in animal breeding. Book Ser Annu Rev Anim Biosci 3:19–56
14. Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467–477
15. Yu J et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208
16. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637
17. Rosa GJM, Steibel JP, Tempelman RJ (2005) Reassessing design and analysis of two-color microarray experiments using mixed effects models. Comp Funct Genomics 6:123–131
18. Steibel JP, Poletto R, Coussens PM, Rosa GJM (2009) A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data. Genomics 94:146–152
19. Henderson CR (1950) Estimation of genetic parameters. Ann Math Stat 21:309
20. Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226
21. Rao CR (1971) Estimation of variance and covariance components MINQUE theory. J Multivar Anal 1: 257–275
22. Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72(358):320–338
23. Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545–554
24. Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York
25. Littell RC, Miliken GA, Stroup WW, Wolfinger RD (2006) SAS system for mixed models, 2nd edn. SAS Institute Inc., Cary
26. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-plus. Springer, New York
27. Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
28. Verbeke G, Molenberghs G (1997) Linear mixed models in practice: a SAS-oriented approach. Lecture notes in statistics 126. Springer, New York
29. Wright S (1921) Systems of mating. I. The biometric relations between parents and offspring. Genetics 6: 111–123
30. Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83
31. Quaas RL (1976) Computing the diagonal elements of a large numerator relationship matrix. Biometrics 32: 949–953
32. Henderson CR, Quaas RL (1976) Multiple trait evaluation using relatives' records. J Anim Sci 43: 1188–1197
33. Schaeffer LR (1984) Sire and cow evaluation under multiple trait models. J Dairy Sci 67:1567–1580
34. Thompson R (1977) Estimation of quantitative genetic parameters. In: Pollak E, Kempthorne O, Bailey TB (eds) Proceedings of the international conference on quantitative genetics. Iowa State University Press, Ames, pp 639–657
35. Meyer K (1985) Maximum-likelihood estimation of variance-components for a multivariate mixed model with equal design matrices. Biometrics 41(153):1985
36. Ducrocq V, Besbes B (1993) Solution of multiple trait animal models with missing data on some traits. J Anim Breed Genet 110:81–92
37. Quaas RL, Pollak EJ (1981) Modified equations for sire models with groups. J Dairy Sci 64:1868–1872
38. Quaas RL, Pollak EJ (1980) Mixed model methodology for farm and ranch beef cattle testing programs. J Anim Sci 51:1277–1287
39. Misztal I, Gianola D (1988) Indirect solution of mixed model equations. J Dairy Sci 77(Suppl. 2):99–106
40. Schaeffer LR, Kennedy BW (1986) Computing solutions to mixed model equations. In: 3rd world congr genet appl livest prod, vol XII, pp 382–393

41. Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199
42. Haley CS, Knott SA (1992) A simple regression method to for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324
43. Haley CS, Knott SA, Elsen J-M (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195–1207
44. Pérez-Enciso M, Misztal I (2004) Qxpak: a versatile mixed model application for genetical genomics and QTL analyses. Bioinformatics 20(16):2792–2798
45. Meuwissen THE, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155: 421–430
46. Pérez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163:1497–1510
47. Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756
48. Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3(1):22–32
49. Dekkers JCM, van Arendonk JAM (1998) Optimizing selection for quantitative traits with information on an identified locus in outbred populations. Genet Res 71(3):257–275
50. Manfredi E, Barbieri M, Fournet F, Elsen JM (1998) A dynamic deterministic model to evaluate breeding strategies under mixed inheritance. Genet Selet Evol 30:127–148
51. Chakraborty R, Moreau L, Dekkers JCM (2002) A method to optimize selection on multiple identified quantitative trait loci. Genet Sel Evol 34(2):145–170
52. Goddard ME (1992) A mixed model for analyses of data on multiple genetic-markers. Theor Appl Genet 83:878–886
53. Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124(6):323–330
54. Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–223
55. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
56. Whittaker JC, Thompson R, Visscher PM (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252
57. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 58:267–288
58. Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. Genetics 163:347–365
59. Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163(2): 789–801
60. ter Braak CJF, Boer MP, Bink MCAM (2005) Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170(3):1435–1438
61. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and predictions. Springer
62. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423
63. Misztal I, Legarra A, Aguilar I (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci 92:4648–4655
64. Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124:362–368
65. Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124:342–355
66. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor J, Schenkel FS (2009) Reliability of genomic predictions for North American dairy bulls. J Dairy Sci 92:16–24
67. Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, GJM R, Gianola D (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J Dairy Sci 92:5248–5257
68. Henderson CR (1985) Best linear unbiased prediction of non-additive genetic merits in non-inbred populations. J Anim Sci 60:111–117
69. Hoeschele I, VanRaden PM (1991) Rapid inverse of dominance relationship matrices for noninbred populations by including sire and dam subclass effects. J Dairy Sci 74:557–569
70. Gianola D (1982) Theory and analysis of threshold characters. J Anim Sci 54:1079–1096
71. Gianola D, Foulley JL (1983) Sire evaluation for ordered categorical-data with a threshold-model. Genet Sel Evol 15(2):201–223
72. Tempelman RJ, Gianola D (1996) A mixed effects model for overdispersed count data in animal breeding. Biometrics 52:265–279
73. Strandén I, Gianola D (1998) Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study. Genet Sel Evol 31:25–42
74. Rosa GJM, Padovani CR, Gianola D (2003) Robust linear mixed models with normal/independent distributions and Bayesian MCMC implementation. Biom J 45(5):573–590
75. Ducrocq V, Casella G (1996) A Bayesian analysis of mixed survival models. Genet Sel Evol 28(6):505–529

76. Varona L (1997) Multiple trait genetic analysis of underlying biological variables of production functions. Livest Prod Sci 47:201–209
77. Forni S, Piles M, Blasco A et al (2009) Comparison of different nonlinear functions to describe Nelore cattle growth. J Anim Sci 87(2):496–506
78. Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. J Anim Sci 63:217–244
79. Shoemaker JS, Painter IS, Weir BS (1999) Bayesian statistics in genetics – a guide for the uninitiated. Trends Genet 15:354–358
80. Blasco A (2001) The Bayesian controversy in animal breeding. J Anim Sci 79(8):2023–2046
81. Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. Nat Rev Genet 5:251–261
82. Yi N, Xu S (2008) Bayesian Lasso for quantitative trait loci mapping. Genetics 179:1045–1055
83. Gianola D, de los Campos G, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183(1):347–363
84. De los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes J (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182:375–385
85. Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semi-parametric procedures. Genetics 173:1761–1776
86. Gianola D, van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303
87. Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S (2007) Machine learning procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J Anim Breed Genet 124(6):377–389
88. González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendano S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178(4):2305–2313
89. De los Campos G, Gianola D, Rosa GJM (2009) The linear model of quantitative genetics is a reproducing kernel Hilbert spaces regression. J Anim Sci 87:1883–1887
90. Gianola D, Okut H, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet 12:87
91. Okut H, Gianola D, Rosa GJM, Weigel KA (2011) Prediction of body mass index in mice using dense molecular markers and a regularized neural network. Genet Res 93:189–201
92. Koltes JE, Cole JB, Clemmens R et al (2019) A vision for development and utilization of high-throughput phenotyping and big data analytics in livestock. Front Genet 10:1197
93. Silva FF, Morota G, Rosa GJM (2021) High-throughput phenotyping in the genomic improvement of livestock. Front Genet 12:707343. https://doi.org/10.3389/fgene.2021.707343
94. Fernandes AFA, Dórea JRR, Rosa GJM (2020) Image analysis and computer vision applications in animal sciences: an overview. Front Vet Sci 7:551269
95. Bresolin T, Dórea JRR (2020) Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems. Front Genet 11:923. https://doi.org/10.3389/fgene.2020.00923
96. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
97. Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York

Books and Reviews

Chapman AB (1980) General and quantitative genetics. World animal science series. Elsevier, Amsterdam

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman & Hall, London

Gondro C, van der Werf J, Hayes B (2013) Genome-wide association studies. Springer, New York

Lange K (2002) Mathematical and statistical methods for genetic analysis, 2nd edn. Springer, New York

Liu BH (1998) Statistical genomics. CRC Press, Boca Raton

Mrode R (2005) Linear models for the prediction of animal breeding values, 2nd edn. CAB Int, New York

Ott J (1991) Analysis of human genetic linkage. Johns Hopkins

Sham P (1998) Statistics in human genetics. Arnold

Van Vleck LD (1993) Selection index and introduction to mixed model methods for genetic improvement of animals. CRC Press, Boca Raton