Genetic Evaluation

2006-06-14

Erling Strandberg and Birgitta Malmfors

Dept of Animal Breeding and Genetics
Swedish University of Agricultural Sciences, Uppsala, Sweden

A major principle of animal breeding is to select those animals to become parents that will improve the genetic level in the next generation the most. For this purpose we need to identify animals with the best genes. In most situations this amounts to finding animals with the best additive genetic effects (or breeding values). The breeding value is not a measure of how good an animal is in itself, but rather of the effect its genes will have in the population.

For quantitative traits we are unable to observe the genotype, we can only measure the phenotypic value, which is influenced both by genotype and environment. Therefore, we need a way to infer the breeding value from the phenotypic value in such a way that we maximize the probability of choosing the correct animals to become parents. This is the objective of the genetic evaluation.

Contents

A general approach: regression of breeding value on phenotypic value, $b_{A/P}$ 1
The regression of breeding value on a phenotypic value = heritability 3
The $b_{A/P}$ accounts for amount of information available 4
Some comments on terminology 4
Different sources of information are used in genetic evaluation 5
Complex situations need flexible methods for calculating E(A|P) 7
Two specific approaches: selection index theory and mixed linear models........... 8
Selection index theory 10
Finding the selection index weights... 11
Example: One trait in breeding goal – recorded on the individual................ 12
Selection index theory can also handle more complicated situations 13
Adjustment of phenotypic records 15
Finding the precision of the selection index 16
Special case: one and the same trait in index and in breeding goal 20
Special types of selection indexes.. 21
Mixed linear models (BLUP) 22
Animal model with unrelated animals 22
Solving the mixed model equations.. 25
Solution by inversion 26
Iterative methods.. 27
Animal model with relationship matrix 27
Example of animal model with relationships.. 28
Precision of predicted breeding values 31
Other mixed linear models.. 32
Reduced animal model... 32
Genetic groups 32
Sire models... 33
Models with maternal effects 33
Models including non-additive genetic components..................................... 34
Multiple trait models.. 34
Repeated records model 35
BLUP breeding values are useful for ranking and selection............................. 36
References 36

Appendices 1-6 are in a separate document.

Genetic Evaluation. Compendium, version 2006-06-14

A general approach: regression of breeding value on phenotypic value, $b_{A/P}$

We will now try to derive a general approach for predicting breeding values for any type of situation. Even though the procedure is general we will use a simple example to describe it.

Assume a simple genetic model:

$$P = A + E \qquad [1]$$

where *P* is the phenotypic value of an individual as a deviation from the mean, *A* is the (true) breeding value, and *E* is the environmental deviation. We cannot exactly know the values of *A* and *E*, only their sum. The phenotypic value can be thought of as a black box, and we need a light to see what's in it. We need a way to calculate the breeding value *A*, given that we know the phenotypic value.

Let's have a look at some data to see if we can come up with a solution to the problem. In Figure 1 the breeding values (*y*-axis) of 200 individuals are plotted against their phenotypic values expressed as deviations from the mean (*x*-axis). (This cannot be done using real data, so we have simulated the individuals in the computer according to eq. [1].)

We can observe the following:

- The good news is that in general, an increased phenotypic value is associated with an increased breeding value. The cluster of points move from the lower left towards the upper right of the graph.
- The not-so-good news is that if we choose a certain phenotypic deviation (say +2, as shown by the vertical line), we can see that all individuals with (approximately) that value do not have the same breeding value.
- For this group, we can also see that their breeding values lie between approximately 0.2 and 1.0 with an average of about 0.6, i.e. the average breeding value is smaller than the phenotypic deviations.

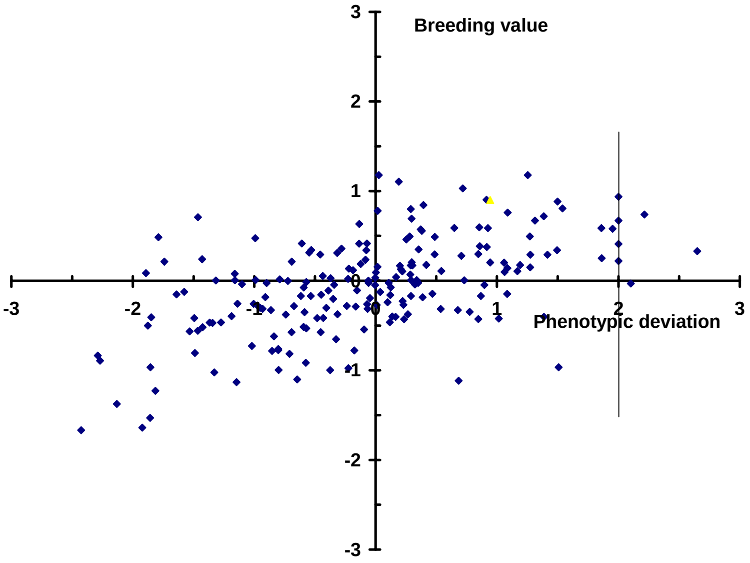


Figure 1. Plot of breeding values and phenotypic deviations for 200 individuals.

So, in conclusion, we are fairly safe in assuming that increased phenotypic value is associated with increased breeding value, but the increase in breeding value seems to be smaller than the increase in phenotypic value.

One way of quantifying the relationship between breeding values and phenotypic deviations in Figure 1 is by a linear regression equation. So, we estimate the regression of breeding values on phenotypic deviations, $b_{A/P}$ and put the regression line into the graph (Figure 2). The slope of this regression is 0.3 and it passes through the origin.

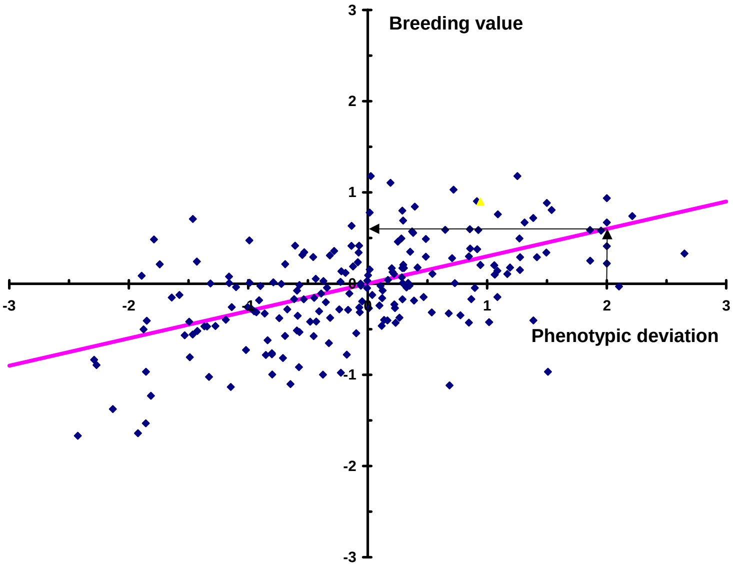


Figure 2. Plot of breeding values and phenotypic deviations for 200 individuals, together with the regression line corresponding to $b_{A/P}$, assuming a h^2 of 0.3

Using the regression line we can see the following:

- The expected breeding value can be calculated for any individual with a certain phenotypic deviation. For example, in the example a phenotypic deviation of +2 results in an expected breeding value of +0.6.
- We can also note that the expected breeding values are closer to zero (the mean) than are the phenotypic deviations. We say that **the expected breeding values are regressed towards the mean**.
- We can also see that we are usually not so lucky as to hit the target exactly, there is always some prediction error, i.e. there is a difference between predicted and true breeding value.

What we have done using the line in Figure 2, can be expressed in statistical or mathematical terms, as that we calculated ***the expected value of the breeding value, given the phenotypic value, E(A|P).***

E(A|P) was obtained by multiplying ***the regression coefficient of breeding value on phenotypic value*** (**$b_{A/P}$**) by the phenotypic deviation. We usually call E(A|P) a ***predicted breeding value***, and it is often denoted with a “hat”: $\hat{A}$.

The calculations we have done can be summarized in an equation:

$$\mathbf{E}(\boldsymbol{A}|\boldsymbol{P}) = \hat{\boldsymbol{A}} = \boldsymbol{b}_{\boldsymbol{A/P}}\,\boldsymbol{P} \qquad [2]$$

This function is valid irrespective of if the phenotypic records are on the individual itself or on its relatives. The b-value will, however be different for different sources of information. How to calculate the b-values will be dealt with later in this chapter.

The regression of breeding value on a phenotypic value = heritability

In Figures 1-2 we had both true breeding values and phenotypic values and were able to estimate $b_{A/P}$ from the data. That is not possible in real life. However, we can calculate $b_{A/P}$ from our knowledge of genetic parameters. Let us try to calculate this regression coefficient from its definition in our simple example where the trait is measured once on the individual itself:

$$b_{A/P} = \frac{\text{cov}(A,P)}{\sigma_P^2} = \frac{\text{cov}(A,A+E)}{\sigma_P^2} = \frac{\text{cov}(A,A)+\text{cov}(A,E)}{\sigma_P^2} = \frac{\sigma_A^2}{\sigma_P^2} = h^2 \quad [3]$$

assuming that there is no covariance between genotype and environment,
i.e. cov(A,E)=0, and that cov(A,A) $=\sigma_A^2$.

What we see from equation [3] is that ***the regression of breeding value on phenotypic deviation is equal to the heritability, h^2***. So, if we know h^2 we can use that to predict the breeding value of an individual if we know its phenotypic value.

In the example in Figures 1-2, the heritability was 0.3. In Figure 3 a corresponding graph is drawn for a trait with heritability equal to 0.9. (This is an unrealistically high value but we use it to give a clearer picture).

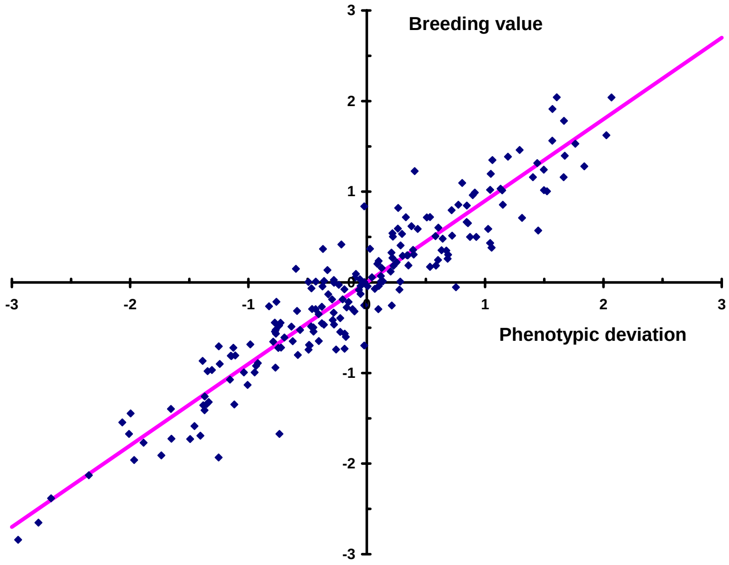


Figure 3. Plot of breeding values and phenotypic deviations for 200 individuals, together with the regression line corresponding to $b_{A/P}$, assuming a h^2 of 0.9.

If we compare the graph in Figure 3 with that in Figure 2 we see that:

- The slope of the regression line is much steeper.
- The points lie much closer to the regression line indicating that our prediction of breeding values is much more precise, the prediction error is much lower.
- The phenotypic deviation gives much more information about the breeding value when the heritability (h^2) is high.
- If you were to select, say, the top 10% of the animals based on their phenotypic values, the average breeding value of those selected would be much higher in Figure 3 than in Figure 2.

The $b_{A/P}$ accounts for amount of information available

Having come this far one might again raise a serious objection, perhaps articulated as: *So what?* If we rank the candidates of selection on their phenotypic deviations (*P*) or on their predicted breeding values (h^2P), the ranking would still be the same. So what have we gained by calculating predicted breeding values?

The answer is: nothing - so far. However, this is because we have the simplest of examples where all individuals have exactly the same information. Let's do an exercise in thinking that will hopefully make you see the need for accounting for the amount of information available on an individual.

Assume that we have two pigs, A and B, and that their growth records are identical: a daily gain of +30 g above the average of the population. So it would seem that we cannot distinguish between them. However, say now that we also know that these growth records are based on 1 and 10 measurements, respectively. Which one of the pigs would you believe to have the highest breeding value?

Intuitively, we would say that we believe that the phenotypic deviation of pig B is a better representation of its breeding value, than is the phenotypic deviation of pig A. By measuring the trait 10 times on individual B we have excluded much of the temporary variation that affects the single measure of pig A.

Intuition could help us in this situation as both A and B had an identical phenotypic deviation. But what if A had grown 90 g better than the average? And what if we have other individuals with varying phenotypic values and number of observations as well? Then we clearly cannot get by on intuition alone, we need a more scientific approach.

Let's put the intuition into statistical terms. By increasing the number of observations, the denominator in equation [3] will be decreased, while the numerator will remain unchanged. The reason is that with repeated measurements the environmental variation is reduced and the phenotypic variance is calculated as

$$\sigma_P^2 = \sigma_A^2 + \frac{\sigma_E^2}{n}$$

where *n* is the number of measurements (if we have no other permanent effect than genetic). Compare with $\sigma_P^2 = \sigma_A^2 + \sigma_E^2$ when we had one measurement.

If the phenotypic variation is reduced the heritability will be increased (because $h^2 = \sigma_A^2 / \sigma_P^2$). The heritability of an average is thus higher than the heritability of a single measurement of the same trait. If h^2 of a single measure was 0.3, then the heritability based on 10 measurements becomes 0.81. So to get the predicted breeding value of pigs A and B in the example above, we use equation [2] and multiply +30 by 0.3 and 0.81, respectively, and get $\hat{A}_A$ = +9 and $\hat{A}_B$ = +24. So pig B is better, just as we thought!

Some comments on terminology

Before going on, we would like to make a note on terminology, because it may often be quite bewildering for the new student of animal breeding. Up till now we have been a bit "sloppy" in our use of terminology. We have used the symbol *P* both for a single phenotypic measure and for an average of *n* observations, as we referred to equation [3] in both cases. This may be confusing. Therefore, in the following when we talk about a phenotypic measure that is used in prediction of breeding values, we will call it *X* instead, and specify exactly what the variable *X* stands for. Therefore, we will write $b_{A/X}$ instead of $b_{A/P}$. When *X=P* (i.e. a single measurement), $b_{A/X} = h^2$ and when $X = \overline{P}_n$, i.e. an average of n observations,

$$b_{A/X} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_E^2}{n}}$$, and so on.

Also, we have sometimes talked about phenotypic values and phenotypic deviations interchangeably. As pointed out earlier the phenotypic values are normally expressed as deviations from the mean, and in the future *X* will always be a deviation.

Different sources of information are used in genetic evaluation

So far we have only given examples where the traits have been measured on the individuals themselves, i.e. on the individuals to be genetically evaluated. In practice we often also use records from relatives, such as progenies, half-sibs, full-sibs, parents and grandparents (Figure 4).

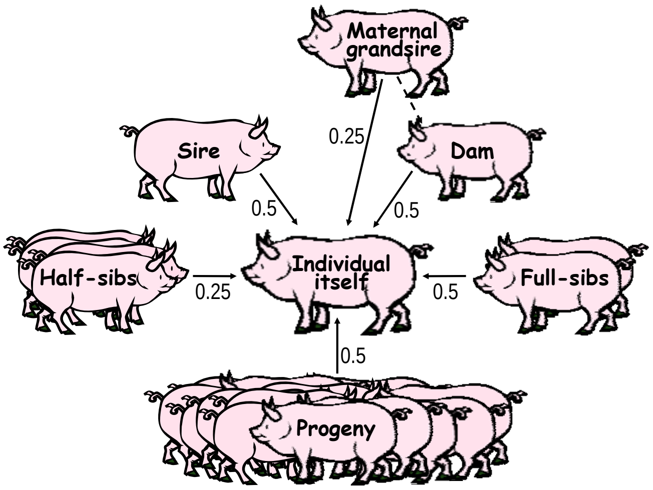


Figure 4. Examples of sources of information used in genetic evaluation. The figures show the additive genetic relationship ($a_{i\alpha}$) between the various sources (i) and the individual itself, i.e. the candidate to be evaluated, the proband (α)

Several factors influence which sources of information to use when predicting breeding values for a trait: what information is available, the heritability of the trait, and how and on what individuals the trait can be measured. In genetic evaluation in practice it is common to combine information from several sources.

- Information on the ***individual itself*** (α)***, i.e. the candidate to be evaluated*** for selection, is commonly used, when the trait in question can be measured on the individual (directly or indirectly). Sometimes this is not possible, e.g., traits that are sex-limited (e.g. milk production, female fertility) cannot be measured in male animals. Traits like carcass composition and meat quality cannot be measured on live animals, unless an indirect method can be used (e.g. ultra-sonic measurement of carcass composition). Use of records on the candidate itself is called *performance testing*.

  For performance testing to be efficient, heritability should be at least moderately high. As we have seen already, the phenotypic deviation comes closer to the true breeding value as the heritability increases (e.g., compare Figure 2 and 3).

- Using phenotypic records on ***progenies*** is generally the most accurate source of information for genetic evaluation. The average phenotypic value of a progeny group gives a good indication of the additive genetic effect (i.e. the breeding value) of the candidate. The value of the information increases with the size of the progeny group.

  Progeny testing is useful also when the heritability is low, and can be used even for traits with a heritability below 0.1, assuming the candidate has a large number of progenies (~100-150). The disadvantage is that it takes time before results on progenies are available. Progeny testing is first of all used for genetic evaluation of male animals, as they usually get many more progenies than females, especially when AI is practised.

- Phenotypic records on the candidate's ***sibs, half sibs and full sibs,*** are often used in addition to other information, or to give supplementary information, for example on traits that cannot be measured on the candidate itself. The accuracy of sib testing depends on the number of sibs that have records. Full sibs are usually raised in the same herd, they have a common environmental effect. This may cause a bias when they are used for prediction of breeding values, unless we are able to adjust for it.

- Information on ***pedigree*** (parents, grandparents) is generally available even before the candidate is born, and can thus give very early information. However, the genes from each locus of the parents are transmitted at random, so information based on pedigree alone is not very accurate, but can be valuable as additional information. The additive genetic relationship, and thus the proportion of common genes between the candidate and the pedigree, is halved for every generation backwards. If the breeding values of the parents are well known there is little to gain in using information on grandparents (actually, if the parents true breeding values were known, there would be *nothing* gained in using grandparental information).

As already mentioned, all information available is usually utilized when an animal's breeding value is predicted. The weight given to a specific source of information depends on the additive genetic relationship with the candidate, the heritability and the amount of information, i.e. the number of progenies or sibs, etc. In the coming pages we will show how breeding values can be calculated when different types of information is available.

Complex situations need flexible methods for calculating E(*A*|*P*)

To get the predicted breeding value or E(*A*|*P*), we need to get a $b_{A/P}$ that we can multiply by the phenotypic deviation (as E(*A*|*P*) = $b_{A/P}$ *P).* We have only shown two simple examples so far, when the phenotypic deviation is based either on a single measure or on several measures on the individual itself. There are innumerable other situations, some of which will be described in the rest of the chapter or in the appendix. Just to mention some, we progeny test bulls to get their breeding value for sex-limited traits like milk production, we use indicator traits, e.g. measures on live animals, when we want to improve the carcass characteristics, and we may want to combine different traits and different sources of information.

Not only $b_{A/P}$, but also the phenotypic value (P) depends on the practical situation. As you probably remember, a phenotypic record may be influenced by several factors, illustrated in Figure 5.

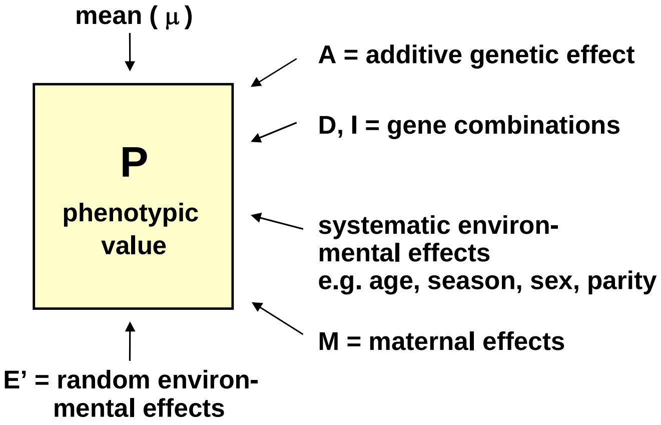

mean (μ)
A = additive genetic effect
D, I = gene combinations
P
phenotypic value
systematic environmental effects
e.g. age, season, sex, parity
M = maternal effects
E' = random environmental effects

Figure 5. Various factors influencing the phenotypic value.

To get a proper prediction of an animal's breeding value (the additive genetic effect) we must adjust for the influence of systematic environmental effects and whenever applicable also for maternal effects and dominance. It is important to adjust for systematic effects, as those are common for groups of animals, and may be interpreted as contributing to genetic differences if not accounted for. For instance, say that you want to compare the growth rate of two lambs born in the same herd. One of the lambs is single-born and the other is a triplet. The single-born lamb will have a better chance to grow fast and its genetic capacity will likely be overestimated if no adjustment is done for litter size at birth. Moreover, using breeding values for growth rate that are not adjusted for litter size may result in reduced fertility, as lambs from small litter sizes will have an advantage in growth rate.

A flexible machinery is needed to handle all these possible situations in a practical manner.

Two specific approaches: selection index theory and mixed linear models

There are two commonly used methods for prediction of breeding values, selection index theory and mixed linear models (often called BLUP, which stands for Best Linear Unbiased Prediction. To be strictly correct, BLUP is not a method but a property of the predictor.). The selection index theory was developed in the 1940's and was the first method used in practice. The principle for using mixed linear model methodology for the prediction of breeding values was chiefly developed by C.R. Hendersson in the late 1940's and the method has been used since the mid-1970's, increasingly so with improved computing resources.

The actual procedures involved in the two methods are very different and it may be difficult to see any similarities between them. Therefore, in this introduction some characteristics of the methods will be mentioned briefly. We will come back to these characteristics later as well. We will also try to point out the similarities with the general approach outlined previously.

First of all, remember that ***what we are looking for is E(A|P), i.e. the breeding value!*** This is true, regardless of whether selection index theory or mixed model methods are used.

In our previous general approach we used the knowledge of h^2 to calculate $b_{A/P}$. This is still true: ***we need to know the genetic parameters*** for both of these methods to work.

The two methods for prediction of breeding values, selection index theory and mixed model methodology (BLUP) can both handle adjustment for systematic (often fixed) "environmental" effects, but it is done in different ways. In our general approach we stressed that we dealt with phenotypic values as deviations from the mean. We also assumed (but never actually said it) that the individuals were influenced by the same systematic environmental effects, or that the records for these effects were adjusted before deviating the value from the mean. This is the procedure ***for the selection index method: we need to adjust the records beforehand and deviate them from the mean.*** This means that the adjustment factors are estimated on previous ("historical") data. In statistical terms a breeding value predicted through selection index theory is **BLP** (Best Linear Prediction), but it is not guaranteed to be unbiased.

As a contrast, ***with the mixed linear models (BLUP) the adjustment is done automatically*** in the method itself. Genetic and environmental effects, as well as the mean, are estimated simultaneously. This is a big advantage, as it guarantees that the predicted breeding values are ***unbiased***. That means that we predict what we expect to do.

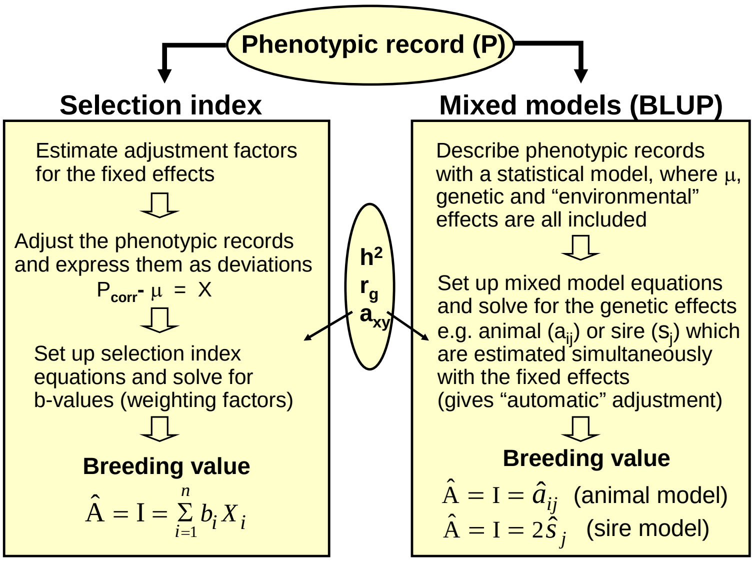
What we will now say may seem strange at first: The selection index theory is a method to predict breeding values, but it is very seldomly used for that purpose any more. So why learn about it at all? The main reason is ***that selection index theory provides a simple way to calculate the precision (accuracy) of selection before you set up a breeding program.*** This is very useful for comparing alternative strategies. Mixed linear models, on the other hand, are typically only used when you actually have real data (phenotypic records), and want to get predicted breeding values. You can calculate the precision with this method as well, but this is usually not done until one has the actual data.

For selection index theory, you don't need data, you only need to know the expected structure of the data or, expressed differently, which sources of information that are planned to be used in the genetic evaluation. Some examples, one would typically use selection index theory to compare the expected precision in selecting pigs on 1 or 10 measures of growth rate, respectively, or to compare how the precision would change if heritability increases from 0.3 to 0.9. One could also compare the precision of using performance testing and progeny testing at a given heritability and progeny group size.

We will come back to the exact definition of precision later, but you already saw the difference in precision in predicting breeding values from Figure 2 vs. Figure 3, there was much less prediction error in Figure 3 and thus a higher precision.

Another reason for learning about ***selection index theory is that it provides a very useful framework when you want to improve several traits at the same time***, by ensuring that you put the correct relative weighting on all traits in the selection criterion.

The two types of methods are summarized in Figure 6.



Phenotypic record (P)
Selection index
Mixed models (BLUP)
Estimate adjustment factors for the fixed effects
Adjust the phenotypic records and express them as deviations
$P_{corr} - \mu = X$
Set up selection index equations and solve for b-values (weighting factors)
Breeding value
$\hat{A} = I = \sum_{i=1}^{n} b_i X_i$
h^2
r_g
a_{xy}
Describe phenotypic records with a statistical model, where μ, genetic and "environmental" effects are all included
Set up mixed model equations and solve for the genetic effects e.g. animal (a_{ij}) or sire (s_j) which are estimated simultaneously with the fixed effects (gives "automatic" adjustment)
Breeding value
$\hat{A} = I = \hat{a}_{ij}$ (animal model)
$\hat{A} = I = 2\hat{s}_j$ (sire model)

Figure 6. Schematic description of the working procedure for selection index and mixed model methods.

Selection index theory

We will describe selection index theory starting from a very general and broad description, but going almost directly to a very simple example.

One of the first steps in setting up a breeding program is to **determine the breeding goal**, i.e. to decide what traits should be genetically improved, as well as to determine their relative importance. In selection index theory this is done by defining a linear breeding goal function, usually called *T* (short for ***T***rue breeding value):

Breeding goal *(or true breeding value)* $= T = v_1A_1 + v_2A_2 + \ldots + v_mA_m =$ **v'a** [4]

where v_i expresses the relative importance of the breeding value A_i, i.e. the relative importance of each one of the traits in the breeding goal. The weights v_i are usually called **economic weights**, but they may be based on other factors than purely economical. The procedures of calculating economic weights will not be dealt with in this chapter, here we just assume that we know them. The last term of the equation (**v'a**) describes the equation in matrix language, and **v'** is a row vector of economic weights and **a** is a column vector of true breeding values.

In some literature the breeding goal is called "aggregate genotype" because it gives a good description of eq. [4]. When several traits are included in the breeding goal we often want to predict a value combining all the traits, i.e. the "aggregate genotype" of the individual. If only one trait is included in the breeding goal ($m=1$) then the breeding goal function will be $T = A$, i.e. the true breeding value of the individual with regard to the specific trait.

The breeding goal *T* itself is unobservable, because it contains the *true* breeding values, so *T* needs to be estimated by some other function. We call this estimator (or predictor) **the index**, and it contains phenotypic information, i.e. information that we really *can* observe:

Index *(or predicted breeding value)* $= \hat{T} = I = b_1X_1 + b_2X_2 + .. + b_nX_n =$ **b'x** [5]

where b_i is a so-called ***selection index weight*** (sometimes just called b-value), for the phenotypic measure X_i. As mentioned earlier, all the *X*'s are pre-adjusted for systematic environmental effects and deviated from the mean. Just as in [4] the last part (**b'x**) is the expression in matrix language, where **b'** is a row vector of index weights and **x** is a column vector of phenotypic deviations.

As an example, X_1 could be birth weight of the individual, X_2 growth from birth to weaning of the individual, X_3 birth weight average of full sibs, X_4 growth from birth to weaning average of full sibs, et cetera. So, we may utilize phenotypic observations of several traits, measured on one or several sources of information.

Please note that the number of breeding goal traits (*m*) is not necessarily the same as the number of selection index traits (*n*). Also, trait 1 in the goal, need not be identical to trait 1 in the index (we may use an indirect measure of it), and so on. But we will show that clearly later.

So, we know how to define what traits are included in the breeding goal, their economic weights and what phenotypic measures are available on the selection candidates or their relatives. But we do not know the index weights, the b-values. So, the million-dollar question is: how do we get the index weights?

Finding the selection index weights

Selection index theory is based on linear regression using least squares. We try to predict *T* with *I*:

$$T = I + e \qquad [6]$$

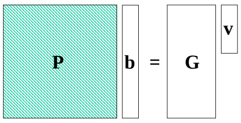
$$\mathbf{v'a} = \mathbf{b'x} + e \qquad [7]$$

where *e* =(*T*-*I*), the residual, often called "error". We use least squares which means that we want the residual or error sum of squares to be as small as possible, i.e. **E(*T*-*I*)^2 is minimized**. This is the same as requiring that the **correlation between the breeding goal (*T*) and the index (*I*) be maximized**. We call this correlation r_{TI}. In words, this means that we find the b-values (the vector **b**) such that the index predicts the true breeding value in the best possible way. For this reason, the selection index is also sometimes called **B**est **L**inear **P**redictor or **BLP**.

Derivation of selection index equations: Appendix 1

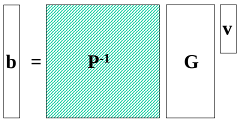
The proof of how the b-values are chosen is given in appendix 1. Here we will just show the end result of that proof. As we may have several traits in both *T* and *I* it becomes very convenient to describe the resulting equation system using matrix algebra, also because computer programs often deal with matrices. If you feel that you would need a brief review of matrix algebra, please have a look in the chapter "Statistical concepts" before continuing here.

Now, back to finding the selection index weights. The system of equations from which the b-values that minimize E(*T*-*I*)^2 can be found is shown in appendix 1 to be:



$$\mathbf{Pb = Gv} \qquad [8]$$

with solution:



$$\mathbf{b = P^{-1}Gv} \qquad [9]$$

where

P = var(**x**), is a square (*n*x*n*) matrix of variances (diagonal) and covariances (off-diagonal elements) among the phenotypic measures X_i (*i*=1..*n*), and $\mathbf{P^{-1}}$ is its inverse,
b is the wanted (*n*x1) vector of index weights (the b-values),
G = cov(**x, a**), is a (*n*x*m*) matrix of additive genetic covariances between the *n* index traits and the *m* breeding goal traits, and
v is a (*m*x1) vector of economic weights.
(Please note that the matrix **P** is something different than the symbol *P* for the phenotypic value that we used e.g. in eq. [1])

If we write out equation [8] specifying the matrix elements it becomes:

$$\begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdot & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdot & \sigma_{X_2X_n} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdot & \sigma^2_{X_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ b_n \end{bmatrix} = \begin{bmatrix} \sigma_{X_1A_1} & \sigma_{X_1A_2} & \cdot & \sigma_{X_1A_m} \\ \sigma_{X_2A_1} & \sigma_{X_2A_2} & \cdot & \sigma_{X_2A_m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{X_nA_1} & \sigma_{X_nA_2} & \cdot & \sigma_{X_nA_m} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ v_m \end{bmatrix} \qquad [10]$$

$$\mathbf{P} \qquad \mathbf{b} \quad = \qquad \mathbf{G} \qquad \mathbf{v}$$

and if we carry out the matrix multiplication on both sides in [10] we get:

$$\begin{aligned} \sigma^2_{X_1} b_1 + \sigma_{X_1X_2} b_2 + ... + \sigma_{X_1X_n} b_n &= \sigma_{X_1A_1} v_1 + \sigma_{X_1A_2} v_2 + ... + \sigma_{X_1A_m} v_m \\ \sigma_{X_1X_2} b_1 + \sigma^2_{X_2} b_2 + ... + \sigma_{X_2X_n} b_n &= \sigma_{X_2A_1} v_1 + \sigma_{X_2A_2} v_2 + ... + \sigma_{X_2A_m} v_m \\ &\cdot \\ \sigma_{X_1X_n} b_1 + \sigma_{X_2X_n} b_2 + ... + \sigma^2_{X\,n} b_n &= \sigma_{X_nA_1} v_1 + \sigma_{X_nA_2} v_2 + ... + \sigma_{X_nA_m} v_m \end{aligned} \qquad [11]$$

Looking at this mess, it is easy to understand why one rather would use matrix algebra and write **Pb=Gv**!

The equation systems [8]-[11] may seem a bit difficult to understand right now so we will directly apply these equations in the same simple example that we used in our general approach previously. What is easy to see, though, is that we will get solutions for the b-values if we find out what figures to put into the other matrices/vectors.

Example: One trait in breeding goal – recorded on the individual

This is the same example that we had in our general approach previously, i.e. a breeding value to be predicted from the phenotypic performance of the individual itself. This means that the breeding goal *T* in this case consists of only one trait (say, growth of the pig), and therefore there is no need for economic weights, or one can say that all weight is given to this trait, i.e. v_1 = 1. We also have only one phenotypic observation, *X* (growth), and thus **P** contains only one value, the phenotypic variance of the trait. Similarly, **G** only contains one value, which should be the additive genetic covariance between the index trait and the goal trait.

Thus in this example we have the situation that T = A_1 and I = b_1 $X_{1.}$ So, if we insert this in [8] or [10] we get:

$$\left[\sigma^2_{X_1}\right]\left[b_1\right] = \left[\sigma_{X_1A_1}\right]\left[v_1\right] \qquad [12]$$

where v_1 =1. We have only one measurement and thus $\sigma^2_{X_1}$ will be the phenotypic variance for this trait. Because the index trait and the goal trait are one and the same and measured on the same individual (additive relationship is 1), the covariance $\sigma_{X_1A_1}$ becomes identical to the additive genetic variance, $\sigma^2_{A_1}$. Inserting this in the equation above gives

$$\begin{aligned} \sigma_{X_1A_1} &= \sigma_{(A_1+E_1)A_1} = \sigma^2_{A_1} + \sigma_{E_1A_1} = \sigma^2_{A_1} \\ &= \sigma^2_{P_1} b_1 = \sigma^2_{A_1} \end{aligned} \qquad [13]$$

and if we solve for the index weight we get:

$$b_1 = \frac{\sigma^2_{A_1}}{\sigma^2_{P_1}} = h_1^2 \qquad [14]$$

and if we put that into the index equation [5] we get:

$$I = h_1^2 X_1 \qquad [15]$$

Now, we are done! Using selection index theory gave **exactly the same result as the general approach** used previously. And we see that the index weight we get by using the equation **Pb=Gv** is the regression of breeding value on phenotype, or the heritability of the trait.

Selection index theory can also handle more complicated situations

The selection index equations given in [8] can be used for any situation. What will differ is the calculation of the variances and covariances in **P** and **G**. For example, say that the trait in the breeding goal is measured on *p* progenies instead of on the individual itself. Then the variance element σ_X^2 in **P** will be the phenotypic variance of a mean instead of just the variance of the trait. For calculation of **G** we need to account for the additive genetic relationship between the proband α (the individual for which we want to get the breeding value) and the information source. This relationship between α and the progeny is 0.5. Another example, if the trait measured is not the same as the trait in the breeding goal then the covariance between the traits will be encountered in the element on the right hand side.

In summary we can say state that selection index theory can handle situations with:

- One or several traits included in the breeding goal and /or in the index
- Trait(s) recorded on one or several sources of information
- Traits in index and in breeding goal being the same or different.

What you need to do is to define which traits will be in the breeding goal (will affect how **G** is set up). You also need to define which traits will be in the index and the type of individuals (information source) each trait is recorded on (will affect the construction of both **P** and **G**). If there are several traits in the breeding goal you also need to enter values for their relative importance ("economic weights").

In the next page you find formulas to calculate the diagonal and off-diagonal elements of **P**, as well as the elements of **G**. Having these formulas you can predict breeding values for almost any situation with regard to traits and sources of information.

Selection index equations in general

Normal equations:

$$\begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdot & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdot & \sigma_{X_2X_n} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdot & \sigma^2_{X_n} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ b_n \end{bmatrix} = \begin{bmatrix} \sigma_{X_1A_1} & \sigma_{X_1A_2} & \cdot & \sigma_{X_1A_m} \\ \sigma_{X_2A_1} & \sigma_{X_2A_2} & \cdot & \sigma_{X_2A_m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{X_nA_1} & \sigma_{X_nA_2} & \cdot & \sigma_{X_nA_m} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ v_m \end{bmatrix}$$

P b = G v

- Diagonal elements ($\sigma^2_{X_i}$) of **P**: $$\frac{\frac{1+(n-1)r}{n}+(p-1)(a_{ii}h^2+c^2)}{p}\sigma^2_P$$

- Off-diagonal elements ($\sigma_{X_iX_j}$) of **P**:
  - same info source - different traits: $$\frac{\sigma_{P\tilde{P}}+(p-1)a_{ii}\sigma_{A\tilde{A}}}{p}$$
  - different info sources - same trait $$a_{ij}\sigma^2_A + c^2\sigma^2_P$$
  - different info sources - different traits: $$a_{ij}\sigma_{A\tilde{A}}+\sigma_{C\tilde{C}}$$

- Elements ($\sigma_{X_iA_j}$) of ***G***:

  trait in I & T is: • the same $a_{i\alpha}h^2\sigma^2_P = a_{i\alpha}\sigma^2_A$ • different $a_{i\alpha}\sigma_{A\tilde{A}}$

where

n	number of observations per individual in information source *i*
p	number of individuals in information source *i*
a_{ii}	relationship between individuals in information source *i*
a_{ij}	relationship between individuals in two different information sources
$a_{i\alpha}$	relationship between the candidate for evaluation (α) and the individuals in information source *i*
r	repeatability of the trait with repeated measures
h^2	heritability of the trait
c^2	influence of common environment (within or between info sources) on a trait
σ^2_P	phenotypic variance for the trait
σ^2_A	additive genetic variance for the trait ($\sigma^2_A = h^2\sigma^2_P$)
$\sigma_{P\tilde{P}}$	phenotypic co-variance between two traits ($\sigma_{P\tilde{P}} = r_p\sqrt{\sigma^2_P\sigma^2_{\tilde{P}}}$)
$\sigma_{A\tilde{A}}$	additive genetic co-variance between two traits ($\sigma_{A\tilde{A}} = r_g\sqrt{\sigma^2_A\sigma^2_{\tilde{A}}}$)
$\sigma_{c\tilde{c}}$	effect of common environment (between info sources) on two traits

Now, a suggestion that might be of help to you when you need to calculate the elements of the matrices in the selection index equations.

- Write a "symbol" for each information source and trait both to the left and on top of the empty **P** matrix (e.g., trait 1 measured on the individual itself you may call "α-1", trait 2 measured on 50 progenies "50P-2", etc)
- Write a "symbol" for each trait in the breeding goal (traits to be improved in individuals) on top of the **G** matrix (e.g.: trait 1, i.e. same trait as in the index "α-1", trait 3, say, a trait not included in the index "α-3", etc).

Make symbols that *you* find informative. Indicating systematically what to include in **P** and **G** will make it easy for you to choose the appropriate formula for every element to be calculated, and to include the information required.

As you could see above, the selection index theory does consider a number of factors, such as heritabilities, phenotypic variances and economic weights of traits, genetic and phenotypic correlations between traits, family size and type, and influence of common environment. It is important to remember, though, that the selection index theory *per se* does not take into account any systematic effects that may have influenced the phenotypic records. Any adjustment required needs to be done in a separate step. As pointed out earlier, the phenotypic records (X_i) that are included in the index (I) must be the adjusted values (expressed as deviations).

Adjustment of phenotypic records

The phenotypic records often need to be adjusted for systematic (fixed) effects, such as age, parity, litter size, days open, sex, herd, year, season, management, etc. Several of those effects fluctuate very little over time, so accurate estimates of their effect may be obtained from previous ("historical") sets of data. Effects of factors like herd, year, season, and management fluctuate more and are therefore best estimated directly from the data to be used in the genetic evaluations, as is done in the BLUP procedure. The option available when selection index theory is used for genetic evaluation is adjustment of fixed effects based on previous data sets. This is sometimes used also in BLUP, e.g., for factors where the effects do not fluctuate very much from year to year.

Phenotypic records are adjusted in order to be comparable. In statistical terms this means that they after the adjustment should have the same mean and variance. If this is not the case some of the predicted breeding values may be overestimated, while others are underestimated. It is thus important that the adjustment is done as correctly as possible. The main procedures for adjustment of phenotypic data are:

- Additive adjustment (affects the level only)
- Multiplicative adjustment (affects both level and variation)
- Deviation from a mean of comparable individuals (affects level only)

Additive and multiplication adjustments are both based on using adjustment factors based on estimates from a statistical analysis (usually least squares analysis). It may for example be the effect of male versus female animals. In additive adjustment the adjustment factor is added to one of the sexes to make the phenotypic records comparable to those of the other sex. In multiplicative adjustment the same is achieved through multiplication with the adjustment factor. Multiplicative adjustment is preferable when the variation is related to the level of the trait (higher level, higher variation). The principle for additive and multiplicative adjustment, respectively, is illustrated in Table 1. We have analyzed milk production records with a model containing effects of mean, calving month and possibly other factors as well. We arbitrarily chose the month August to be the reference point.

Table 1. Example of additive and multiplicative adjustment factors for effect of calving month on 305-day milk production in dairy cattle.

			Adjustment factor	
Model			Additive	Multiplicative
term	Estimate	$\hat{\mu}+\hat{a}_i$	$K_\mathrm{i}=\hat{a}_8-\hat{a}_i$	$M_\mathrm{i}=(\hat{\mu}+\hat{a}_8)/(\hat{\mu}+\hat{a}_i)$
μ	7987			
a_1	-72	7915	-30	0.996
a_2	-132	7855	30	1.004
a_3	-311	7676	209	1.027
a_4	-347	7640	245	1.032
a_5	-407	7580	305	1.040
a_6	-302	7685	200	1.026
a_7	-287	7700	185	1.024
a_8	**-102**	**7885**	**0**	**1.000**
a_9	-20	7967	-82	0.990
a_{10}	41	8028	-143	0.982
a_{11}	33	8020	-135	0.983
a_{12}	0	7987	102	0.987

The adjusted records would become:

$$y_i{*} = y_i + K_i \text{ or}$$
$$y_i{*} = y_i M_i$$

where y_i is the unadjusted record of a cow calving in month *i*.

As pointed out previously, when selection index theory is used for prediction of breeding values the phenotypic observations are expressed as deviations from a mean (often the current herd mean). This is in fact also an adjustment, which may adjust for systematic effects that fluctuate from year to year (e.g. herd, year, season).

Finding the precision of the selection index

As mentioned initially, one of the main uses for selection index theory is to find the precision of the selection criterion. This is actually the measure that was one of our starting points in deriving the selection index equations, i.e. the correlation between the true and the predicted breeding value ($\boldsymbol{r_{TI}}$), which is often called the ***accuracy*** of the selection criterion. The accuracy in prediction of breeding values has an important impact on the genetic improvement that can be expected as a result of selection, and is thus a useful measure.

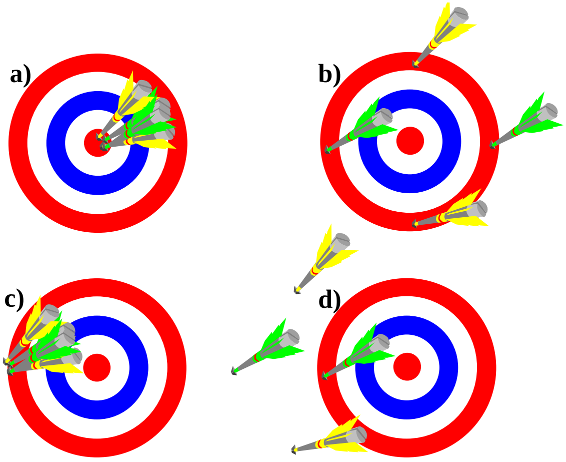
Rather unfortunately, animal breeders have traditionally called r_{TI} accuracy. In statistical literature what r_{TI} measures, is called precision. Accuracy on the other hand, is related to bias. In the graph in the next page we try to illustrate the distinction between these two measures of goodness of fit.

The panel a) describes an ideal situation, where we both hit the target without bias and have very good precision (the darts are very close to each other).

Panel b) shows a situation with no bias but with rather bad precision, the arrows are spread fairly wide around the board.

Panel c) shows a case with high precision but large bias. We may have had a large material at our hands, but somehow the data was flawed, or we failed to account for an important environmental factor.

Panel d) finally shows the worst situation of all: we have both bad precision and large bias.



There is a measure that combines both of these aspects, precision (defined as prediction error variance, PEV) and accuracy, the mean squared error which is defined as:

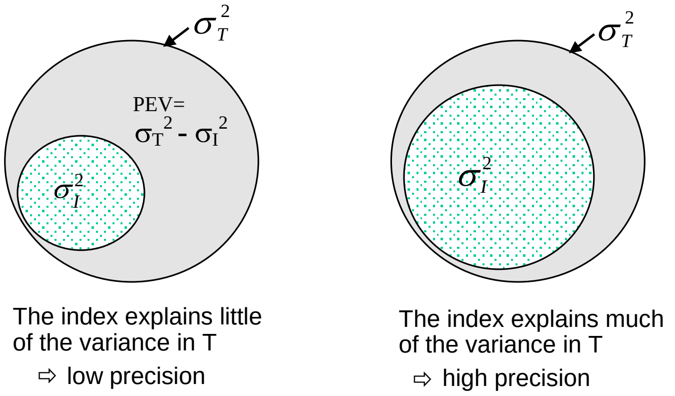
$$\text{MSE} = \text{PEV} + (\text{bias})^2$$

where the bias is squared because it can take both positive and negative values, none of which are favourable.

So, whenever you hear the word “accuracy” please make sure that the person who uses the word means the same thing as you do (whatever that is).

Let us try to explain precision in an intuitive way first. As you know by now, we try to predict *T* by using the index, *I*. On another level, one can say that the variance of *I* should explain as much of the variance in *T* as possible, or reversely, the variance **not** explained by *I* should be as **small** as possible. If we use the analogy of circles for variances, this can be described as in Figure 7.

Figure 7. Schematic description using Venn diagrams of how much variation in *T* that is explained by variation in *I*. The outer, big circle reflects σ_T^2 and the inner circle σ_I^2. The dark grey surface is the part of the variation in T that is not explained by the index, the prediction error variance, PEV.



σ_T^2
PEV=
$\sigma_T^2 - \sigma_I^2$
σ_I^2
The index explains little of the variance in T
⇨ low precision
σ_T^2
σ_I^2
The index explains much of the variance in T
⇨ high precision

Looking at the circles, one can see that the larger the variation is in the index (*I*), the higher will the precision be in our genetic evaluation and the lower will the ***prediction error variance***, PEV, be. At maximum the index explains all the variation in the true breeding value (*T*), and the circle for *I* is as big as that for *T* and falls fully within the circle for *T*.

Actually, the above description is not unique for the selection index method. In ordinary statistical analysis we often use the R^2-value (coefficient of determination) when we want to find out how much of the variation in *y* (corresponds to our *T*) that can be explained by a certain model (corresponds to our *I*). The R^2-value, which is in fact a squared correlation, can be obtained as the amount of variation explained by the model divided with the total variation. So, for our situation with *T* and *I* we define:

$$r_{TI}^2 = \frac{\sigma_I^2}{\sigma_T^2} \qquad [16]$$

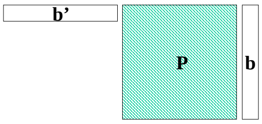
This value is often called ***reliability***, at least in international dairy cattle applications. From [16] we get the accuracy as:

$$r_{TI} = \sqrt{\frac{\sigma_I^2}{\sigma_T^2}} = \frac{\sigma_I}{\sigma_T} \qquad [17]$$

So, we need to know how to calculate the variance of the index (σ_I^2) and of the breeding goal (σ_T^2), in order to get r_{TI}.

The variance of the index is equal to:

$$\sigma_I^2 = Var(b_1 X_1 + b_2 X_2 + ... + b_n X_n)$$

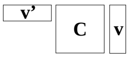


or in matrix terms

$$\mathbf{b´Pb} \qquad [18]$$

and the variance of the breeding goal is:

$$\sigma_T^2 = Var(b_1 A_1 + b_2 A_2 + ... + b_m A_m)$$



$$\mathbf{v’Cv} \qquad [19]$$

Derivation of variance of *I* and *T*: Appendix 2

where **C** is a *mxm* matrix containing the additive genetic variances (diagonal) and covariances (off-diagonal elements) among the breeding goal traits. Then equation [16] becomes:

$$r_{TI}^2 = \frac{\mathbf{b'Pb}}{\mathbf{v'Cv}} \qquad [20]$$

Even though this equation involves matrices and vectors, the end result is a scalar.

Let us examine the value of r_{TI}^2 in our simple example of phenotypic information on the individual itself. Equation [20] would become:

$$r_{TI}^2 = \frac{h^2 \sigma_P^2 h^2}{1 \times \sigma_A^2 \times 1} = \frac{\frac{\sigma_A^2}{\sigma_P^2} \sigma_P^2 \frac{\sigma_A^2}{\sigma_P^2}}{\sigma_A^2} = \frac{\sigma_A^2}{\sigma_P^2} = h^2$$

and thus $r_{TI} = \sqrt{h^2} = h$ [21]

So, in this simple example, r_{TI}^2 becomes identical to the heritability. We alluded to that precision was related to the level of the *h^2* when we compared Figures 2 and 3. We saw in the graph that prediction was more precise when *h^2* was higher, and that prediction error was smaller. Now, we have also proven it, mathematically.

As seen from [16] the r_{TI}^2 value is directly proportional to the variance of the index (σ_I^2). This means that also this variance, or more commonly the **standard deviation of the index** ($\sigma_I = r_{TI} \times \sigma_T$) can be used as a measure of the precision of the predicted breeding value, as long as you have the same breeding goal (and thus σ_T is constant). A high σ_I thus indicates a high precision.

We may also be interested in the **prediction error variance**, i.e. the variance unexplained by the index (the dark grey parts of the circles in Figure 7). This variance is:

$$PEV = \sigma_T^2 - \sigma_I^2 = \sigma_T^2 - r_{TI}^2 \sigma_T^2 = (1 - r_{TI}^2)\sigma_T^2 \qquad [22]$$

and the **standard error of the index, *SE*(*I*), is the square root of *PEV***. A low standard error means an index with a high precision. Just as for an ordinary estimate, the standard error can be used to calculate a confidence interval for a given index.

As you have seen, there are several measures than can be used to indicate the precision of a breeding value and you need to be careful with the interpretation of these. Be especially aware of the difference between the *standard deviation* of the index and then *standard error* of the index! Remember:

Measure	**should be**
Accuracy, r_{TI}	HIGH
Standard deviation of index, σ_I	HIGH
Prediction error variance, *PEV*	LOW
Standard error of the index	LOW

The efficiency of the index also depends on the parameters that are entered into it. For example, we generally assume that the genetic and phenotypic variances and covariances are known without error, but this may not be true in practice. It has been shown that errors in the estimates of the heritability and repeatability of a trait have comparably little effect on the usefulness of the index, while errors in the genetic and phenotypic correlations can be more serious.

Special case: one and the same trait in index and in breeding goal

For the situation when there is ***only one trait in the breeding goal and this trait is measured on one or several sources of information*** the r_{TI} value can simply be calculated as:

$$r_{TI} = \sqrt{\sum_{i=1}^{n} b_i \times a_{i\alpha}} \qquad [23]$$

where $a_{i\alpha}$ is the additive genetic relationship between the proband α and the information source for index trait *i*, and the sum is over all information sources.

Using this formula for our simple example with phenotypic information on the individual itself gives $r_{TI} = \sqrt{h^2 \times 1} = h$. If the trait is measured on the individual itself and on a number of half-sibs as well, then $r_{TI} = \sqrt{b_1 \times 1 + b_2 \times 0.25}$, etc. To calculate the r_{TI} value you thus first need to use selection index theory to calculate the b-values, as shown previously.

Another common situation is when you have progeny testing for the breeding goal trait, e.g. milk yield in dairy cattle. Then the r_{TI} for *p* progenies can be calculated as:

$$r_{TI} = \sqrt{\frac{p}{p+\lambda}} \qquad [24]$$

where $\lambda = \dfrac{\sigma_e^2}{\sigma_s^2} = \dfrac{4-h^2}{h^2}$.

For heritabilities of 5% and 30% and 100 progenies, r_{TI} would be 0.75 and 0.94, respectively.

Special types of selection indexes

Selection indexes can occur in different variants. For example, the index can be based on ***sub-indexes*** (one for each trait in the breeding goal) that are weighed together with their respective economic weights, i.e. $I = v_1I_1 + v_2I_2 + \ldots + v_mI_m$. This means that the economic weights can easily be modified, without the need to recalculate the index equations.

Another useful variant is ***restricted selection index*** or ***desired gains selection index***. By using this we can for example restrict one of the breeding goal traits to remain unchanged, while at the same time the expected genetic change in the other traits is maximized. Maintained adult weight in ewes, while maximizing growth rate of their lambs might be an example.

Mixed linear models (BLUP)

As mentioned previously, whereas selection index theory nowadays is mainly used to get the precision of selection when planning a breeding program, mixed linear models are used to get the predicted breeding values once the breeding program is in place and we have the performance records. The breeding values we get are **B**est **L**inear **U**nbiased **P**redictions, and the procedure of using mixed linear models in genetic evaluation is therefore often called **BLUP**. The *Best* means that the procedure maximizes the correlation between true and predicted breeding values (r_{TI}) or minimizes the prediction error variance; the *Linear* means that the predictors are linear (additive) functions of the observations; the *Unbiased* means that the predicted breeding values is equal to the expected value of the true breeding value; and the *Prediction* comes from the fact that we are dealing with random variables that we predict the outcome of (e.g., future offspring of a sire).

Having a statistical model as the starting point for genetic evaluation means that the analysis can be applied to many different practical situations, and that we get a proper adjustment for systematic "environmental" effects, as those are estimated jointly with the genetic effects. Animals can then be compared across groups, e.g. herds, age groups et cetera, which gives a wider scope for selection. No longer do we have to work with pre-adjusted phenotypic records deviated from the mean (X_i's), we are able to use the actual phenotypic records directly. Just as with selection index theory, though, BLUP too requires that the genetic parameters are known.

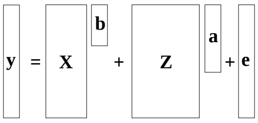
There are many different possible models that one can use in the mixed linear model framework. We will start by describing what we think is the most natural model when one can assume a purely additive genetic model– **the animal model.**

Animal model with unrelated animals

We assume that the phenotypic observations (in the following called *y*, instead of *P*, to conform better with statistical and animal breeding literature) can be described by the following model:

$$y = mean + systematic\ environmental\ effects + animal + residual$$

or expressed in matrix terms:



$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [25]$$

where

y is a column vector of phenotypic observations

b is a column vector of fixed effects (mean + systematic environmental effects)

a is a column vector of animals' breeding values, ~IND($0, \sigma_a^2$) where $\sigma_a^2 = \sigma_A^2$, the additive genetic variance,

X, Z are incidence matrices relating fixed effects and breeding values, respectively, to the observations, and

e is a vector of residuals, ~IND($0, \sigma_e^2$), where $\sigma_e^2 = \sigma_E^2$, the environmental variance

Unless you are used to reading about linear models expressed in matrix algebra, equation [25] may not mean very much to you right now. Therefore, we will

Table 2. Example data.

Growth	Sex	Animal
200	1	1
250	2	2
270	1	3

explain with the help of a small example. We will assume that we have observations on growth of pigs, and that we only have a mean and a fixed effect of sex in the model, apart from the breeding value effect. We also assume that the pigs are unrelated. The data is given in Table 2.

If we forget about the matrix language for a while we can write the model as:

$$y_{ij} = \mu + k_i + a_{ij} + e_{ij}$$

indicating that each single phenotypic observation is influenced by the mean (μ), the sex i (k_i) of the animal, its breeding value (a_{ij}) and a random error (e_{ij}) specific for each animal.

More specifically the model for the three observation becomes:

$$\begin{aligned} 200 &= \mu + k_1 + a_{11} + e_{11} \\ 250 &= \mu + k_2 + a_{21} + e_{21} \\ 270 &= \mu + k_1 + a_{12} + e_{12} \end{aligned} \qquad [26]$$

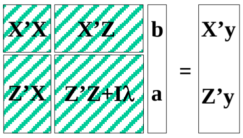
Now, putting this into matrix language (as in [25]) it becomes:

$$\begin{bmatrix} 200 \\ 250 \\ 270 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ k_1 \\ k_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{21} \\ e_{12} \end{bmatrix} \qquad [27]$$

$$\mathbf{y} \quad = \quad \mathbf{X} \qquad \mathbf{b} \; + \qquad \mathbf{Z} \qquad \mathbf{a} \; + \; \mathbf{e}$$

If you carry out the matrix multiplications you will see that you end up with the same result as in [26].

We want to solve this equation system to get solutions for the fixed effects and for the breeding values at the same time. These solutions can be found by minimizing the prediction error variance. Doing this minimization leads to the following equation system, called Henderson's **M**ixed **M**odel **E**quations (**MME**):



$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [28]$$

where $\lambda = \dfrac{\sigma_e^2}{\sigma_a^2} = \dfrac{\sigma_E^2}{\sigma_A^2} = \dfrac{\sigma_P^2 - \sigma_A^2}{\sigma_A^2} = \dfrac{\dfrac{\sigma_P^2 - \sigma_A^2}{\sigma_P^2}}{\dfrac{\sigma_A^2}{\sigma_P^2}} = \dfrac{1-h^2}{h^2}$ [29]

and **I** is an identity matrix, i.e. a a matrix with 1's on the diagonal and 0's everywhere else. This is almost the same equation system as for ordinary least squares, except for the variance ratio λ added to the diagonal of **Z'Z**, the random part of the equation system. As you can see this addition means that genetic parameters, in this case h^2, needs to be known. The matrix on the left hand side is often called the ***coefficient matrix***.

Derivation of MME, Appendix 3

The proof of the MME will not be given here, because it is long and complicated; it is given only in appendix 3. So, for now, we hope you will be satisfied by knowing that the solution is found by minimizing the prediction error variance, while also ensuring unbiasedness of the estimates and predictions, by estimating all effects simultaneously. Therefore, the solutions to the MME are called Best Linear Unbiased Predictors or **BLUP**. This is discussed more in appendix 3.

Now back to our small example. Let's assume we have h^2 = 0.3, which gives $\lambda = \frac{1-0.3}{0.3} = 2.33$. The MME resulting from the three observations thus will be (to facilitate for you to see where the figures in the coefficient matrix belong, we have indicated the effects included in the model outside this matrix):

$$\begin{array}{c} \\ \mu \\ k_1 \\ k_2 \\ a_1 \\ a_2 \\ a_3 \end{array}
\begin{array}{c}
\begin{array}{cccccc} \mu & k_1 & k_2 & a_1 & a_2 & a_3 \end{array} \\
\begin{bmatrix} 3 & 2 & 1 & 1 & 1 & 1 \\ 2 & 2 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1+2.33 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1+2.33 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1+2.33 \end{bmatrix}
\end{array}
\begin{array}{c} \\ \begin{bmatrix} \mu \\ k_1 \\ k_2 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \end{array}
\begin{array}{c} \\ = \end{array}
\begin{array}{c} \\ \begin{bmatrix} 720 \\ 470 \\ 250 \\ 200 \\ 250 \\ 270 \end{bmatrix} \end{array} \qquad [30]$$

Using symbols we can see that the coefficient matrix and the vector on the right hand side contain the following:

$$\begin{array}{c} \\ \mu \\ k_1 \\ k_2 \\ a_1 \\ a_2 \\ a_3 \end{array}
\begin{array}{c}
\begin{array}{cccccc} \mu & k_1 & k_2 & a_1 & a_2 & a_3 \end{array} \\
\begin{bmatrix} N_{tot} & n_{1.} & n_{2.} & 1 & 1 & 1 \\ n_{1.} & n_{1.} & 0 & 1 & 0 & 1 \\ n_{2.} & 0 & n_{2.} & 0 & 1 & 0 \\ 1 & 1 & 0 & 1+\lambda & 0 & 0 \\ 1 & 0 & 1 & 0 & 1+\lambda & 0 \\ 1 & 1 & 0 & 0 & 0 & 1+\lambda \end{bmatrix}
\end{array}
\begin{array}{c} \\ \begin{bmatrix} \mu \\ k_1 \\ k_2 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} \end{array}
\begin{array}{c} \\ = \end{array}
\begin{array}{c} \\ \begin{bmatrix} \sum\sum y \\ \sum y_1 \\ \sum y_2 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} \end{array} \qquad [31]$$

What you can see is that the coefficient matrix consists of numbers of observations at different levels of specification, e.g. total number of observations, the number of observations for each one of the two sexes ($n_{1.}$, $n_{2.}$), as well as for each animal. In the random part of the coefficient matrix you recognize the addition of λ, and you can also see that the other elements in this section are zeros. This is, however, only the case when the animals are unrelated.

To be even more precise we can say that the entry in any given cell (row-column position) in **X'X, X'Z**, **Z'X** or **Z'Z** is the number of observations that have both the row-effect and the column-effect. For instance, in equation [30] in our example the position (row 1, col. 1) shows that there are 3 observations that have μ in them (this is always equal to the total number of observations). Similarly, position (row 1, col. 2) indicates that there are 2 observations that have both μ and k_1 in them, and so on. The 0 in position (row 2, col. 3) shows that no observations have both k_1 and k_2 (no animal could be of both sexes; not the kind of animals we work with, anyway).

For the right hand side, the values are sums of all observations pertaining to the effect in question. For instance, the first value (720) is the sum of all observations (because they all contain the mean), the second value is the sum of all observations with sex 1, and so on.

If you want to, you can carry out the matrix multiplications described in [28] (e.g. multiply **X'** by **X**) to see that you actually get the numbers given in [30]. (If you do that, you can answer the following somewhat tricky question: I put to you that the numbers in the coefficient matrix of MME are not really the number of observations (as suggested in [31]) but they are still equal to the number of observations! Can you explain that?)

Now, let us study the equation system [30] in more detail. We look at the line pertaining to animal 1 (line 4) and write it out:

$1\mu + 1k_1 + 0k_2 + 3.33a_1 + 0a_2 + 0a_3 = 200$, which simplifies to:

$$\mu + k_1 + 3.33a_1 = 200 \qquad [32]$$

If we solve for a_1 we can see how the breeding value of animal 1 is calculated in the BLUP procedure, i.e.:

$$a_1 = (200-k_1-\mu)/3.33 = 0.3\ (200-k_1-\mu) \qquad [33]$$

To actually get the predicted breeding value we need to solve the equation system in [30], so that we get simultaneous solutions also for μ and k_1. How to solve such an equation system and get the solutions is discussed in the next section of this chapter.

The important thing to see from eq. [33] is that the phenotypic value (200) is adjusted for the fixed effect of sex 1 and deviated from the mean, and then multiplied by the heritability, 0.3.

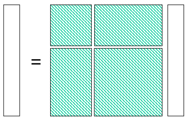
So, as you may have noticed already, ***we get basically the same solution for the breeding value when we use the BLUP procedure, as we had in our general approach, as well as when we used selection index theory.*** We multiply the phenotypic deviation by the regression of breeding value on phenotype, i.e. in our simple example, the heritability. The only difference is that in BLUP we adjust for the fixed effects directly when we solve the equation system, whereas in the other procedures we had to pre-adjust for these effects.

Solving the mixed model equations

So far we have only set up the MME ([28], [30-31]) and looked at the equations for an animal ([32-33]) but we haven't solved the equation system. There are several different ways to get the solution from the MME.

Solution by inversion

The perhaps most natural way is to invert the left hand side coefficient matrix and multiply that inverse by the right hand side:



$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{I}\lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad [34]$$

(assuming the MME of [28]). Note, that we have now put hats on the **b** and **u** vectors, they now contain the solutions, the estimates and predictions.

In order to calculate the inverse we need to make sure the matrix is of full rank. This means in practice that we have to make sure that no columns (or rows) are additive combinations of other columns (rows). In our simple example in [30], the column for μ is identical to the sum of the two columns for k_1 and k_2. So, we cannot estimate both μ *and* k_1 and k_2, we need to ***reparameterize*** (see chapter on “Statistical concepts” p 15 and 17). The simplest way is to delete the row and column pertaining to μ and then take the inverse of the remaining coefficient matrix:

$$\begin{array}{c|cccccc|} & \mu & k_1 & k_2 & a_1 & a_2 & a_3 \\ \mu & \cancel{3} & \cancel{2} & \cancel{1} & \cancel{1} & \cancel{1} & \cancel{1} \\ k_1 & \cancel{2} & 2 & 0 & 1 & 0 & 1 \\ k_2 & \cancel{1} & 0 & 1 & 0 & 1 & 0 \\ a_1 & \cancel{1} & 1 & 0 & 1+2.33 & 0 & 0 \\ a_2 & \cancel{1} & 0 & 1 & 0 & 1+2.33 & 0 \\ a_3 & \cancel{1} & 1 & 0 & 0 & 0 & 1+2.33 \end{array} \begin{bmatrix} \cancel{\mu} \\ k_1 \\ k_2 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \cancel{720} \\ 470 \\ 250 \\ 200 \\ 250 \\ 270 \end{bmatrix} \qquad [35]$$

$$\begin{bmatrix} \hat{k}_1 \\ \hat{k}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{bmatrix} = \begin{bmatrix} 0.714 & 0 & -0.214 & 0 & -0.214 \\ 0 & 1.428 & 0 & -0.428 & 0 \\ -0.214 & 0 & 0.364 & 0 & 0.0643 \\ 0 & -0.428 & 0 & 0.428 & 0 \\ -0.214 & 0 & 0.0643 & 0 & 0.364 \end{bmatrix} \begin{bmatrix} 470 \\ 250 \\ 200 \\ 250 \\ 270 \end{bmatrix} = \begin{bmatrix} 235 \\ 250 \\ -10.5 \\ 0 \\ 10.5 \end{bmatrix} [36]$$

So, the predicted breeding values for the three animals are –10.5, 0, and 10.5, respectively. Note that the three breeding values add to zero (our assumption was that **a**~IND(0, σ_A^2), i.e. has expectation zero).

In practice, predicted breeding values are rarely used directly as they are from the solution of the MME. Often they are set relative to the breeding values of some group of animals that is not the base population, e.g., the average of the last three years of bulls. Sometimes the chosen average is set to 100. If we apply that in our example, the predicted breeding values of the three animals would be 89.5, 100, and 110.5. Another often used transformation is to standardize the variance, e.g., that one genetic standard deviation corresponds to, say, 10 units of the presented breeding values. If both these transformations are used the predicted breeding values all have positive values and are distributed around 100, with extremes at around 70 and 130.

We can now check that equation [33] actually is true (remember that after reparameterization our new k_1 includes the μ:

$$\hat{a}_1 = 0.3\ (200\text{-}235) = \text{-}10.5$$

The method of inversion can, however, often be very difficult to apply in reality, especially for an animal model with many animals and perhaps also many fixed effect, such as herd effects. The matrix just becomes too large to invert. For small examples, however, this method works fine.

Iterative methods

Iterative solution methods, Appendix 4

There are several methods that use iterative procedures, i.e. they do not give the correct answer right away, but they need to be repeated until the solutions do not change any more, which is called convergence. Such methods are needed when the equation system becomes so large that a solution by inversion is not possible. We give a brief outline of such a method in appendix 4.

Animal model with relationship matrix

The previous example is interesting because it is simple, and it shows that the estimate of the breeding value of an animal from MME is built on the same principles as in our general approach and as in selection index theory. However, **the main benefit of the animal model is that you can use information from all relatives of an individual** when solving for the breeding value.

To do this we need to amend the MME a bit. Our assumption about the breeding values is no longer that they are ~IND($0, \sigma_A^2$), now we assume they are ~ND($0, \mathbf{A}\sigma_A^2$), where the matrix **A,** the so-called *relationship matrix,* contains the additive genetic relationships among the animals. Now, the MME become:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [37]$$

Instead of just adding the variance ratio λ to the diagonal elements of the random part, we add a whole matrix, $\mathbf{A}^{-1}$, the inverse of the relationship matrix, multiplied by λ, to **Z'Z.**

Setting up $\mathbf{A}^{-1}$ directly, Appendix 5

Note that **A** (and thus $\mathbf{A}^{-1}$) is of the same size as **Z'Z**. This also means that if you want to predict breeding values for 10 000 animals, you need to set up a 10 000 x 10 000 **A** matrix first, by use of the tabulation method, and then invert this matrix to get $\mathbf{A}^{-1}$, before it can be inserted into the equation system. There is, however, a trick to get the $\mathbf{A}^{-1}$ directly, without having to invert **A**. This very efficient procedure works animal by animal, just as the setting up of the MME does. The procedure is given in appendix 5.

Example of animal model with relationships

We will use a very simple example to describe what the MME looks like when you have related animals. We assume the pedigree structure as in Figure 8.

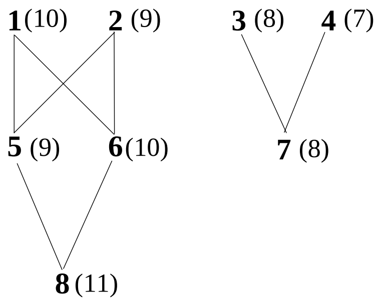


Figure 8. Description of pedigree structure of 8 animals (phenotypic values within parantheses) (from Kennedy *et al.*, 1988).

The model is simply:

$$y_{ij} = \mu + a_i + e_i$$

but with the assumption that **a** ~ND(0,$\mathbf{A}\sigma_a^2$). The ordinary least squares equation system (i.e. before adding $\mathbf{A}^{-1}\lambda$ to **Z’Z** in [26]) becomes:

$$\begin{array}{c} \\ \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{array}\begin{array}{c} \begin{array}{ccccccccc} \mu & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \end{array} \\ \begin{bmatrix} 8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} \\ \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix} \end{array} \begin{array}{c} \\ = \end{array} \begin{array}{c} \\ \begin{bmatrix} 72 \\ 10 \\ 9 \\ 8 \\ 7 \\ 9 \\ 10 \\ 8 \\ 11 \end{bmatrix} \end{array} \qquad [38]$$

The relationship matrix **A** is:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0.5 \\ 0 & 1 & 0 & 0 & 0.5 & 0.5 & 0 & 0.5 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0.5 & 0 & 0 & 1 & 0.5 & 0 & 0.75 \\ 0.5 & 0.5 & 0 & 0 & 0.5 & 1 & 0 & 0.75 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 1 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0.75 & 0.75 & 0 & 1.25 \end{bmatrix} \qquad [39]$$

and its inverse becomes:

$$\mathbf{A}^{-1} = \begin{bmatrix} 2 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 2 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1.5 & 0.5 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0.5 & 1.5 & 0 & 0 & -1 & 0 \\ -1 & -1 & 0 & 0 & 2.5 & 0.5 & 0 & -1 \\ -1 & -1 & 0 & 0 & 0.5 & 2.5 & 0 & -1 \\ 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix} \qquad \text{[40]}$$

The MME (after adding in $\mathbf{A}^{-1}\lambda$ to the oLS, where in this situation we assume h^2 to be 0.5 so λ is 1):

$$\begin{array}{c|ccccccccc|}
 & \mu & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\
\mu & 8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
a_1 & 1 & 3 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\
a_2 & 1 & 1 & 3 & 0 & 0 & -1 & -1 & 0 & 0 \\
a_3 & 1 & 0 & 0 & 2.5 & 0.5 & 0 & 0 & -1 & 0 \\
a_4 & 1 & 0 & 0 & 0.5 & 2.5 & 0 & 0 & -1 & 0 \\
a_5 & 1 & -1 & -1 & 0 & 0 & 3.5 & 0.5 & 0 & -1 \\
a_6 & 1 & -1 & -1 & 0 & 0 & 0.5 & 3.5 & 0 & -1 \\
a_7 & 1 & 0 & 0 & -1 & -1 & 0 & 0 & 3 & 0 \\
a_8 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 3
\end{array}
\begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix} = \begin{bmatrix} 72 \\ 10 \\ 9 \\ 8 \\ 7 \\ 9 \\ 10 \\ 8 \\ 11 \end{bmatrix} \qquad \text{[41]}$$

Now, let’s return to finding the actual solutions to the equation system in [39].The inverse of the left hand side matrix becomes:

$$\begin{array}{c|ccccccccc|}
 & \mu & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\
\mu & 0.4118 & -0.2647 & -0.2647 & -0.2353 & -0.2353 & -0.3235 & -0.3235 & -0.2941 & -0.3529 \\
a_1 & -0.2647 & 0.5987 & 0.0987 & 0.1513 & 0.1513 & 0.3151 & 0.3151 & 0.1891 & 0.2983 \\
a_2 & -0.2647 & 0.0987 & 0.5987 & 0.1513 & 0.1513 & 0.3151 & 0.3151 & 0.1891 & 0.2983 \\
a_3 & -0.2353 & 0.1513 & 0.1513 & 0.5987 & 0.0987 & 0.1849 & 0.1849 & 0.3109 & 0.2017 \\
a_4 & -0.2353 & 0.1513 & 0.1513 & 0.0987 & 0.5987 & 0.1849 & 0.1849 & 0.3109 & 0.2017 \\
a_5 & -0.3235 & 0.3151 & 0.3151 & 0.1849 & 0.1849 & 0.6352 & 0.3018 & 0.2311 & 0.4202 \\
a_6 & -0.3235 & 0.3151 & 0.3151 & 0.1849 & 0.1849 & 0.3018 & 0.6352 & 0.2311 & 0.4202 \\
a_7 & -0.2941 & 0.1891 & 0.1891 & 0.3109 & 0.3109 & 0.2311 & 0.2311 & 0.6387 & 0.2521 \\
a_8 & -0.3529 & 0.2983 & 0.2983 & 0.2017 & 0.2017 & 0.4202 & 0.4202 & 0.2521 & 0.7311
\end{array} \qquad \text{[42]}$$

The solution vector (after multiplying the inverse by the right hand side) becomes:

$$\begin{bmatrix}\hat{\mu}\\ \hat{a}_1\\ \hat{a}_2\\ \hat{a}_3\\ \hat{a}_4\\ \hat{a}_5\\ \hat{a}_6\\ \hat{a}_7\\ \hat{a}_8\end{bmatrix}=\begin{bmatrix}8.7059\\ 0.8676\\ 0.3676\\ -0.3676\\ -0.8676\\ 0.6716\\ 1.0049\\ -0.6471\\ 1.3235\end{bmatrix} \qquad [43]$$

Note that here the sum of ***all*** breeding values is not zero, only the sum of the breeding values of the individuals from the so-called base population, i.e. animals that have unknown parents (animals 1 to 4). It seems that there is some kind of selection going on (or random drift) because the average breeding values in generation 1 (animals 5-7) is above the average of the base population. This is because the two best animals in the base population (1 and 2) contributed two offspring to the next generation whereas animals 3 and 4 only contributed one.

Let's have a look at two of the equations in equation [41]. First we write out the equation pertaining to animal 1, an animal with unknown parents but with offspring:

$$\hat{\mu}+3\hat{a}_1+\hat{a}_2-\hat{a}_5-\hat{a}_6=10 \qquad [44]$$

and solve for the breeding value of animal 1:

$$\hat{a}_1=\frac{1}{3}[(10-\hat{\mu})+(\hat{a}_5-0.5\hat{a}_2)+(\hat{a}_6-0.5\hat{a}_2)] \qquad [45]$$

We can note that:

- the phenotypic value is adjusted for the fixed effects, in this case only the mean.
- the breeding value of the offspring is adjusted for that part that comes from the other parent (the mate of animal 1).
- everything is multiplied by a regression factor, in this case the within-family heritability, defined as $(0.5k\sigma_A^2/(0.5k\sigma_A^2+\sigma_E^2)=0.5h^2/(0.5h^2+(1-h^2))$, where $k=(1-\bar{F})$, and $\bar{F}$ is the average inbreeding of the parents of animal 1. In this example k=1 and h^2 is 0.5, so the within-family heritability becomes 1/3.

Let’s also have a look at the equation for animal 8, an animal with known parents:

$$\hat{\mu} - \hat{a}_5 - \hat{a}_6 + 3\hat{a}_8 = 11$$

$$\begin{aligned}\hat{a}_8 &= \frac{1}{3}(11 - \hat{\mu} + \hat{a}_5 + \hat{a}_6) = \frac{1}{3}(11 - \hat{\mu} + \frac{3\hat{a}_5}{2} - \frac{\hat{a}_5}{2} + \frac{3\hat{a}_6}{2} - \frac{\hat{a}_6}{2}) \\ &= \frac{\hat{a}_5 + \hat{a}_6}{2} + \frac{1}{3}(11 - \hat{\mu} - \frac{\hat{a}_5 + \hat{a}_6}{2})\end{aligned} \qquad [46]$$

The first term is the expected breeding value based on the breeding values of both parents. The second term is an estimate of the Mendelian sampling term. It is based on the adjusted phenotypic observation deviated from the parent average breeding value, again multiplied by the within-family heritability.

Precision of predicted breeding values

The precision of predicted breeding values can be calculated also from mixed linear models. Let’s describe the inverse of the coefficient matrix in equation [37] in abbreviated form as:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\lambda \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix} \qquad [47]$$

where $\mathbf{C}^{11}$ corresponds to the fixed effects part (**X’X)** and $\mathbf{C}^{22}$ to the random part (**Z’Z+A^{-1}**λ). Now, prediction error variance (*PEV*) for the breeding values can be calculated as:

$$PEV = \mathbf{C^{22}}\sigma_e^2 \qquad [48]$$

or expressed more simply, the PEV for animal *i* is:

$$PEV_i = c^{ii}\sigma_e^2 \qquad [49]$$

i.e. the diagonal element of the inverse matrix corresponding to animal *i*, multiplied by the residual variance. If we want the accuracy, r_{TI}, we can use the relationship from [22] and combine it with [49]:

$$PEV_i = (1 - r_{TI}^2)\sigma_a^2 = c^{ii}\sigma_e^2 \qquad [50]$$

If we solve for r_{TI}^2 we get:

$$r_{TI}^2 = 1 - c^{ii}\lambda$$

and

$$r_{TI} = \sqrt{1 - c^{ii}\lambda} \qquad [51]$$

where λ is as before, $\dfrac{\sigma_E^2}{\sigma_A^2} = \dfrac{1-h^2}{h^2}$.

In the example shown in Figure 8, if we get the inverse elements from equation [42] we get the following accuracies (assuming that λ=1):

	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$	$\hat{a}_6$	$\hat{a}_7$	$\hat{a}_8$
Inverse element	.5987	.5987	.5987	.5987	.6352	.6352	.6387	.7311
r_{TI}	.6335	.6335	.6335	.6335	.6040	.6040	.6011	.5186

Other mixed linear models

So far we have dealt with the animal model, where every individual animal is included in the analysis for prediction of breeding values. You have seen how the data is adjusted for systematic environmental effects, and how information on relatives is utilized by including additive genetic relationships between the individuals. We have, however, just considered one single trait, with only a single observation per animal. Moreover, dominance and maternal effects have been assumed to be of no importance.

Using mixed linear models for prediction of breeding values means that various situations can be handled. There are several types of models that can be used for different purposes. We will briefly comment on some of those in the following. For more thorough discussion and illustrations of the use of different models, see for example: (Mrode, 1996) “Linear Models for the Prediction on Animal Breeding Values” or (Van Vleck, 1993) “Selection Index and Introduction to Mixed Model Methods”.

Reduced animal model

An alternative to predict breeding values for animals that have progeny records is to use a *reduced animal model (RAM)*. Only equations for animals that are parents are then included, which reduces the number of equations to be solved and thus makes the computing faster than when the full animal model is used. Breeding values of individual progenies can thereafter be predicted by back-solving from the predicted parental breeding values.

Genetic groups

Quite often animals with unknown parents are included in the genetic evaluation analyses (like animals 1, 2, 3 and 4 in our example on animal model including relationships). Such animals are generally called base population animals and the breeding values of animals in subsequent generations are usually expressed relative to those base animals. It may happen that the base animals come from populations with different genetic means, their origin may be from different breeds, different year batches or sires from different countries. If so, this must be accounted for in the model, or the predicted breeding values will be biased. A way to handle it is to include genetic group as a fixed effect in the model and assign a group-identification to each animal lacking pedigree.

Sire models

The sire model can be used when observations are solely on progenies. The breeding value of the sire is then predicted from the phenotypic average of the progeny group. The number of equations to solve is thus much lower than when the individual animal model is used. The accuracy of the predictions depends on the number of progenies.

In model terms the sire model includes:

Progeny observation = mean + systematic environmental effects + sire + residual

Sire model equations and prediction of breeding values, see Appendix 6

How to set up the sire model equations and predict the breeding values, is shown in appendix 6. The progenies only have half of their genes from the sire. An estimate based on their phenotypic value thus reflects one half of his breeding value. The breeding value for a certain sire is therefore predicted as:

$I = 2\hat{s}_j$ where $\hat{s}_j$ is the predicted effect of sire j.

The sire model basically assumes that all progeny of a sire are from different dams, and that all dams are equally good. If some sires are used on better dams than others the quality of the dams need to be corrected for. One way to do this, at least partially, is to also include the father of the dam in the model, which actually means that the sire is evaluated both through his own daughters and through his granddaughters. The model used is sometimes called the ***sire-maternal grandsire model*** (see appendix 6). However, be aware that in this model we still assume that there are only additive effects, i.e. the genes of the male act in the same way whether they occur in the father or in the maternal grandsire. This is not the case in the “sire-maternal grandsire model” used when one assumes maternal effects (see below).

If the dams have several offspring with the same sire the dam can be included in the model instead of the maternal grandsire. The model then used is a ***sire-dam model***.

Models with maternal effects

Some traits may be influenced by *maternal effects*, which means that the mother has an impact on the performance of her offspring that depends on her ability to provide a suitable environment for them (mothering ability). This may the case for traits like early growth, survival and weaning weight of pigs, lambs and beef cattle, as well as behaviour characteristics in many species. These maternal effects are strictly environmental for the offspring, but with regard to the mother the mothering ability can be partly genetic and partly environmental. The mothering ability can be considered to be an unobservable phenotype of the mother, the effect of which is only seen in the offspring.

The mixed model methodology can handle this situation by extending the animal model to include:

Phenotypic observation = mean + systematic environmental effects
+ breeding value of ***animal*** (often called ***direct genetic effect***)
+ breeding value of the ***dam*** in providing a suitable environment (***maternal genetic effect***)
+ permanent environmental effect of the dam (maternal non-genetic effect)
+ residual

Note that the genes for mothering ability exist in both males and females, however, the genes are only expressed in females, and not until they get progeny of their own. The assumption is also that there can be a correlation between the direct genetic effect and the maternal genetic effect. For instance, if a negative direct-maternal correlation for growth exists, that means that animals that have good genes for their own growth have bad genes for supporting the growth of their offspring.

Another way of estimating maternal genetic effects is to apply a ***sire-maternal grandsire model***, where the effect of the maternal grandsire now includes both the direct effect of genes, and the maternal effect of the genes.

Phenotypic observation = mean + systematic environmental effects
+ effect of sire (direct effect of genes)
+ effect of maternal grandsire (direct effect and maternal effect)
+ residual

Note the difference between this model and the sire-maternal grandsire model for purely additive direct effects described in appendix 6.

Models including non-additive genetic components

The models that we have discussed so far have all exclusively dealt with additive genetic effects. For some traits, however, the contribution of non-additive genetic effects, like dominance, might be significant. Theoretically, the Hendersson mixed model equations (MME) can handle a situation where we have both additive and dominance genetic effects (by also including dominance relationships), and thus predict both the additive genetic effect and the dominance effect of each individual. In practice, however, the application of such models has been limited due to lack of reliable genetic parameters, and also because dominance effects are often highly confounded with effects due to common environment between close relatives.

Multiple trait models

Genetic evaluation of an animal mostly includes several traits, and as you could see in the section on selection index theory, we often want to predict a breeding value (index) that combines those traits. This is possible to achieve also when the mixed model methodology is used for the genetic evaluation. Moreover, including several traits in the same analysis means that the genetic and phenotypic correlations between traits are considered, and that correlated traits thus add information to each other. This usually increases the accuracy of the evaluations.

In multiple-trait BLUP the mixed model equations are extended in proportion to the number of traits included. If we consider a multi-trait analysis including two traits the mixed models in matrix form for each trait will be:

Trait 1: $$\mathbf{y_1} = \mathbf{X_1 b_1} + \mathbf{Z_1 a_1} + \mathbf{e_1}$$

Trait 2: $$\mathbf{y_2} = \mathbf{X_2 b_2} + \mathbf{Z_2 a_2} + \mathbf{e_2}$$

The model for a two-trait analysis can thus be written as:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

and the Hendersson Mixed Model Equations (MME) [37], will be expanded accordingly, i.e. the number of equations will be doubled compared to single-trait analysis. For each trait included in the analysis we will also need to incorporate the respective variances for random effects (σ_u^2), residuals (σ_e^2), as well as the covariances between the traits.

The number of equations to be solved can be very large when multi-trait BLUP is used, especially if many traits are included. This may result in computational problems and a high computational cost. However, with increased computer capacity and computational "tricks" such problems are becoming less important. Iterative methods, such as iteration on data, have made it possible to solve extremely large equation systems.

Another problem with muilti-trait analysis can be lack of reliable estimates of genetic and phenotypic correlations between the traits, but this problem is the same whether selection index theory or the BLUP procedure is used.

What we get from the multi-trait mixed model analysis are predictions of breeding values for each trait included in the analysis (which are not the same as those we would have got in separate analyses ignoring correlations between traits). To get a composite breeding value (Total Merit Index) we weigh the multi-trait BLUP values by their respective economic weights:

$$I = v_1 I_1 + v_2 I_2 + \ldots\ldots\ldots + v_m I_m$$

where I_1 to I_m are predicted breeding values from a multiple trait model. This is actually the same procedure as we use when calculating a selection index based on sub-indexes.

Repeated records model

Sometimes the same trait is measured repeatedly on the same animal. Milk yield in successive lactations and litter size in successive pregnancies are some examples. Repeated records on the same animal show resemblance not only for genetic reasons, but also because they often are influenced by permanent environmental factors. The mixed model with repeated records may thus look like:

$$\mathbf{y = Xb + Za + Zp + e}$$

where **p** is a vector of permanent environmental effects (specific for each animal), and the other elements are the same as in an ordinary single-trait analysis. The assumption in this model is that, say, milk yield in first and second lactation is genetically the same trait. If this is not (at least approximately) true one should use the multiple trait animal model instead.

BLUP breeding values are useful for ranking and selection

BLUP breeding values, especially from the animal model including relationships, are useful tools in selection. Selection on BLUP breeding values maximizes the probability for correct ranking of breeding animals and selection on them maximizes genetic gain from one generation to another. There are many factors that contribute to this:

- The animal model makes full use of information from ***all*** relatives, which increases accuracy (precision).
- The breeding values are adjusted for systematic environmental effects in an optimal way. This means that animals can also be compared across herds, age classes etc, assuming the data is connected.
- The procedure is flexible, various practical situations can be handled.
- Non-random mating can be accounted for.
- Several traits can be included
- Bias due to culling (e.g., between 1st and 2nd lactations) and selection (over generations) is accounted for, assuming that also non-selected animals' data are included in the analysis.

It should, however, be noted that the genetic evaluation is based on phenotypic observations, and that regardless of how splendid the BLUP procedure may be, it ***cannot compensate for bad data***. So a good recording is necessary for a reliable genetic evaluation and subsequent genetic gain. It should also not be forgotten that BLUP (as well as selection index) assumes that ***the genetic parameters used are the true ones***. In practice that means that they should be close to the true parameters.

Something that should be noticed is the potential ***risk for increased inbreeding*** when selection is based on breeding values including information on all relatives. The probability that several family members are selected jointly is increased, which may result in increased inbreeding. To avoid this, and to optimize long-term selection response, selection on BLUP breeding values might be combined with some restriction on average relationship of the selected animals.

A useful side effect of BLUP genetic evaluation is that it gives ***estimates of the realized genetic trend***. This is achieved by comparing BLUP breeding values of animals born in different years, assuming there are connections between years through successive time overlapping or through relationships. We already did a simple estimation of genetic trend in the small animal model example starting on page 29.

References

Kennedy, B.W., Schaeffer, L.R. and Sorensen, D.A. 1988. Genetic properties of animal models. *J. Dairy Sci.* 71 (Suppl 2):17-26.

Mrode, R.A. 1996. Linear Models for the Prediction on Animal Breeding Values. CAB International, Wallingford, UK.

Van Vleck, L.D. 1993. Selection Index and Introduction to Mixed Model Methods. CRC Press, Boca Raton, FL, USA.