Animal Breeding Methods - Course Notes

Instructor - L. R. Schaeffer

Overview of Animal Breeding
Fall 2008

1 Required Information

Successful animal breeding requires

1. the collection and storage of data on individually identified animals; and
2. complete pedigree information about the sire and dam of each animal.

Without these two pieces of information little genetic change can be made in a population. In dairy cattle, beef cattle, swine, sheep, and poultry, recording programs were established around 1900 in North America. Breed registry programs have been around for many years too. Selling purebred animals usually requires an official pedigree. Animal identification is important today for the ability to monitor animal movement for human health safety purposes. Animal recording and registrations are expensive programs to run, but are necessary to improve the breed or population. Much effort is needed to make sure as few errors as possible enter these databases. Much data are now collected and transmitted electronically to recording centres, and this has eliminated many errors, or has caught errors at the farm level which could be corrected on the spot. On farm computer systems have also helped in the collection of data.

The records and pedigrees need to be electronically stored for computer manipulation and data analyses. In the 1930's, Jay L. Lush began to show people how data could be used to identify genetically superior animals, mainly dairy bulls. The statistical methodology has been improved over the years, especially through the work of Charles R. Henderson from 1950 to 1989. Henderson's methods are chiefly used today in all countries. However, the models and methods are still being improved through the work of Gianola and Sorensen (the two Daniels), and others. Improvements are possible due to advances in computing power, i.e. more memory, more disk space, faster CPUs, and parallel processors.

Ideally, all animals within a herd should be recorded without any selection on which animals would be recorded. ICAR, the International Committee on Animal Recording, has put together guidelines for all recording programs in each species. This includes how animals should be weighed and measured, and at what ages, and so on. The guidelines are useful for species that tend to cross country boundaries, for example, dairy bull semen is sold from USA to many countries around the world. Thus, it is somewhat important to be able to compare cattle records between countries, and this is made possible when countries follow the similar data recording procedures.

2 What to do with information

Animal breeders analyze the data to estimate the breeding values of individual animals in a population using statistical linear models. Animals are ranked on the basis of the estimated breeding values (EBV), and the better animals are mated together, and the rest are culled (i.e. not allowed to mate). Animals are usually evaluated for several traits and these are weighted by their relative economic values allowing for the heritability of each trait and the genetic correlations among the traits.

An EBV incorporates data from (1) records on the animal, (2) information on the sire and dam of the animal, and (3) information on all progeny of that animal. At the same time effects due to the herd, the year of birth, the age of the animal, and many other factors need to be removed during the estimation process.

3 Mendelian Inheritance

Consider a single gene locus, call it locus A. The genotype of that locus describes the two alleles for an individual. Let the genotype of a male parent (usually referred to as the sire) be A_1A_2 and let the genotype of a female parent (referred to as the dam) be A_3A_4, where A_1, A_2, A_3, and A_4 are different alleles at that gene location. Each offspring inherits one of the two alleles from each parent with equal probability (i.e. 0.5). There are four possible genotypes of offspring described in the following table.

Table 1.1 Possible offspring genotypes.

		Female Gametes	
		A_3	A_4
		0.5	0.5
Male Gametes			
A_1	0.5	A_1A_3	A_1A_4
		0.25	0.25
A_2	0.5	A_2A_3	A_2A_4
		0.25	0.25

In this case, none of the offspring have the same genotype as one of the parents. Suppose the genotype of the dam is the same as that of the sire.

Table 1.2 Possible offspring genotypes.

		Female Gametes	
		A_1	A_2
		0.5	0.5
Male Gametes			
A_1	0.5	A_1A_1	A_1A_2
		0.25	0.25
A_2	0.5	A_2A_1	A_2A_2
		0.25	0.25

Half the progeny have the same genotype as the parents. If two copies of allele 2 was lethal, then one quarter of the progeny of this mating would be expected not to survive, and two thirds of the surviving offspring would be carriers of the lethal allele.

4 The Infinitesimal Model

Mendelian inheritance is assumed to occur at every locus in the genome. Molecular geneticists have estimated that there are between 30,000 and 60,000 gene loci in the genome. The number of alleles at each locus varies from 2 to 30 or more. Even if we assume that there are only two alleles at each locus, that gives 3 possible genotypes at each locus, and if we assume only 30,000 gene loci, and if we assume all of the gene loci are independent, then the number of possible genotypes (considering all loci simultaneously) would be 3^{30000} which is large enough to give the illusion of an infinite number of loci.

Genetic evaluation models typically assume that there are an infinite number of loci affecting each trait, called *quantitative trait loci* or QTL. The goal is to estimate the combined effect of all loci for each trait on each animal. Each locus, by itself, is assumed to have a relatively small effect. These assumptions give the *Infinitesimal Model*. Today there are challenges to the Infinitesimal model, and researchers are trying to find individual loci that have large effects on traits. These loci are called *major genes*. There could be up to 10 such loci for each trait.

The infinitesimal model is still the main tool for genetic evaluation, because the costs of genotyping many animals for major genes or markers close to the major gene are still pretty high. Also, the expression of one particular gene may be major, but it could be influenced by genes at other loci. The interactions of a major gene on loci affecting other traits are not completely understood either. Lastly, studies about the magnitude of effects of major genes require estimates of breeding values from usual genetic evaluation. Thus, the infinitesimal model will be needed for many years to come.

5 Types of Gene Action

Genetic evaluation is concerned with the **additive** influence of each allele. Suppose there is a gene locus that contributes to the growth of animals. Assume that allele 1, A_1, contributes 50 grams of weight at birth, A_2 contributes 30 grams, and A_3 contributes only 5 grams. The value of different genotypes can be determined as in the table below.

Table 1.3 Breeding values of Genotypes

Genotype	First Allele	+	Second Allele	=	Breeding Value
A_1A_1	50 g	+	50 g	=	100 g
A_1A_2	50 g	+	30 g	=	80 g
A_1A_3	50 g	+	5 g	=	55 g
A_2A_2	30 g	+	30 g	=	60 g
A_2A_3	30 g	+	5 g	=	35 g
A_3A_3	5 g	+	5 g	=	10 g

The breeding values are just the sum of the effects of each allele in the genotype. Another type of gene action is called *dominance*. Dominance occurs when there is an additional effect on a trait resulting from the particular combination of alleles. For example, suppose that when A_1 occurs with A_2 in a genotype there is an additional 10 grams of weight generated. Thus, instead of a genetic effect of 80 grams, the total genetic value is 90 grams.

Another type of gene action is called *epistasis*. This is an interaction between different loci caused by the particular genotype at one locus interacting with a particular genotype

at another locus. Perhaps when A_1A_2 occurs with B_4B_8 there is a loss of 7 grams in weight.

In genetic evaluation, only additive genetic effects are assumed to exist. The magnitudes of dominance and epistatic effects are assumed to be negligible. Also, genetic evaluation is based upon what is transmitted from parent to offspring which is only the additive genetic effect. With the infinitesimal model, the sum of additive effects at all loci are considered jointly in genetic evaluation.

6 Animal Identification

Animals must be uniquely identified. The birthdate and IDs of the sire and dam should also be known. To illustrate an ID system, the international standard system for dairy cattle is

```
HOCANF0036221749
```

where `HO` denotes a Holstein animal, `CAN` indicates Canada as the country of birth, `F` represents the code for a female animal, and the numeric part is the official registration number in Canada.

A problem with this ID is that it needs to be linked to a physical ID on the animal itself. Common physical IDs are tattoos, hot or freeze branding, ear notches, radio collars, fin clipping, ear tags, and pit tags(electronic chips). A disadvantage of physical IDs is that they are not permanent. Tags can be lost, or they are re-used after an animal is culled. DNA fingerprinting could be used to identify individuals uniquely, and have been used to identify full-sib groups of fish, but not individuals. DNA fingerprinting is also costly (at the moment).

Animal identification is a top priority for any genetic improvement program. Errors in identification can lower estimates of genetic variability and can result in biased genetic evaluations. The best investment for a genetic improvement strategy is in a top notch animal identification program. Besides being useful for genetic evaluation, animal identification is important for health and traceability concerning food safety for humans, which has become very important in recent years.

If possible, the ID system should be designed so that the numeric part of an offspring ID is a larger number than that of either parent. To compute inbreeding coefficients, for example, requires that animals be sorted chronologically. Animals that are measured for economic traits should be included in the pedigree. Animals that have not been measured for any traits and which also do not have any offspring may be removed from the pedigree files.

The number of generations that can be traced in the pedigree also has an effect on the estimation of breeding values, inbreeding coefficients, and amount of genetic variability remaining in the population. The more generations that are recorded, the better will be the analyses. However, going three generations back from the earliest recorded observation will likely be sufficient for most genetic analyses. There will always be a group of animals in a pedigree file that have unknown male and/or female parents, and these become the **base population** which are assumed to be animals that were randomly mating.

7 Data

Data refers to traits of economic importance to a livestock production system, and to all variables that could influence the expression of those traits. In dairy cattle, for example, the main trait of importance has been milk production. Successful selection for milk yield over the years has brought about correlated genetic responses of a detrimental nature in fertility and health traits. This has led to adding recording schemes for reproduction and health traits beyond the current milk recording schemes. While the future can never be known totally in advance, a recording scheme that attempts to define all of the traits that could influence productivity and profitability would likely best serve a genetic improvement program.

Traits are the observed and recorded variables associated with the productivity of an animal. Examples are milk yields, number of eggs, weight of calves, number of piglets born, weight of fleece produced, conformation of the animal, racing speed, jumping ability, behaviour, feed efficiency, reproductive efficiency, susceptibility to diseases, and others. Information should be recorded on all animals within a contemporary group, not just on selected individuals. Factors related to the observation should also be recorded, such as the age of the animal at the time of recording, the contemporary group, the location (herd, province, country), the month of the year (seasonal effects), the breed of the animal, the age of the dam, the track conditions, and who took the measurements. The factors affecting a trait are as important as the trait itself.

8 Breeding Objective

The *Breeding Objective* is a function of the traits that the owner(s) of the animals wish to change. The breeding objective includes all of the traits that need improvement, even if there are no records on some of the traits. Suppose five traits are identified as economically important. The breeding objective may be defined as

$$H = v_1T_1 \ + \ v_2T_2 \ + \ v_3T_3 \ + \ v_4T_4 \ + \ v_5T_5,$$

where v_1 to v_5 are relative economic values for each trait, and T_1 to T_5 are the true (unknown) breeding values for those traits. Suppose only the first 3 traits are recorded as well as another trait 6 which is correlated to trait 5. The breeding objective is approximated by an index, such as,

$$I = w_1 EBV_1 + w_2 EBV_2 + w_3 EBV_3 + w_6 EBV_6,$$

where w_1 to w_6 are economic weights, and EBV_i are estimated breeding values of the animal for traits 1, 2, 3, and 6. The selection index approach is a method for going from the breeding objective to the index.

9 Pathways of Selection

Selection emphasis can be applied differently to various ancestors. Generally, fewer males are needed for reproduction purposes than females. Thus, producers can be more strict about their requirements for males than for females. Only the very best (top 1%) sires and dams will be used to produce future sires in the species. The next females, however, will be offspring of sires in the top 25% of the species and of dams in the top 75% of the population. This is because nearly all females are kept for breeding purposes, while most males are culled or sent to market at an early age. These figures will vary depending on species, but there will always be these four pathways for selection:

- Sires of sires (top 1%),
- Dams of sires (top 1%),
- Sires of dams (top 25%),
- Dams of dams (top 75%).

10 Measurement of Genetic Change

Genetic change must be measured to determine the success of a breeding program. There are many different trends that could be monitored. Genetic change is estimated by averaging the EBVs of particular groups of animals. For example, the trend in all females that had offspring by year of birth of the offspring. A slightly different trend would be all females born in a particular year, even though some of them may never have progeny themselves. The trend in sires used for breeding could also be calculated. Genetic trend in each pathway of selection may be of interest. Thus, genetic trend must be carefully defined and interpretted.

11 Breeding Strategies

EBVs are used to determine which animals will be parents of the next generation. To optimize genetic improvement, the EBVs can be used to determine which male to breed to each female, such that the offspring have the highest possible average breeding value. Breeding strategies are concerned with the design of an efficient breeding program that maximizes genetic change under a certain set of conditions over the next few generations. What happens when conditions are changed or when restrictions are relaxed? Breeding strategies require an understanding of the biology of the species or the production system. These include

- Age at first breeding (for males and females);
- Length of gestation;
- Number of animals born per mating;
- Times in the lifecycle when traits are measured;
- Number of males and females needed for breeding;
- Generation interval; and
- Length of productivity of an animal.

Generation interval is the average age of an animal (sire or dam) when it can be replaced by one of its offpsring in the breeding program. Shortening the generation interval generally results in faster genetic change. Generation intervals depend largely on reproductive capacity of the species, but any technology that allows the breeding value of an animal to be estimated earlier in life will shorten the generation interval. Reproductive capacity of a species may be changed (with technology) to get more offspring per mating, to use fewer males or females, or to reduce the length of time to age at first breeding.

12 Mating Systems

A mating system is a set of rules for mating animals in a production system. In a beef production system, each cow that is bred is expected to produce a calf at weaning. Thus, any cow that does not become pregnant should be culled, and if the calf is born but does not survive to weaning, then the cow may be culled if it happens more than once. A mating system often refers to crossbreeding systems, or to linebreeding. Some producers have a short breeding season and will cull any females not pregnant at the end of that season. The mating system may allot females to breeding groups according to EBVs for

particular traits (or an index) which determines the males that will be used to service them. A mating system is a plan that should be followed as part of the overall breeding strategy.

13 Genome Selection

Due to the human genome mapping project, many livestock species are also having their genomes mapped. There are over 7 million single nucleotide polymorphisms (SNPs) in the human genome, and within a few years the same will be true for cattle and swine. There will be SNPs located very closely to each other all throughout the genome. Within a pair of adjacent SNPs could be none or 100 different genes that affect a particular trait of interest. By having 10,000 animals genotyped for 100,000 SNPs, the additive genetic contribution of each adjacent pair of SNPs can be estimated. The overall breeding value of the animal would be the sum of the estimated effects for all pairs of SNPs in the genome. This is a new area of genetics that is not yet implemented, and which needs considerable research. What is the best method to estimate the additive effects of adjacent pairs of SNPs? What are the effects of dominance and epistatic effects on the estimation of additive effects? The generation interval can be significantly reduced because animals can be genotyped at birth and the EBVs for many traits calculated immediately. How might breeding strategies be changed to maximize genetic change using genome selection? This strategy should help to locate major genes, if they exist, so that gene expression and gene function studies may be conducted.

Statistics Review

Fall 2008

1 Example Data Problem

Data on 10 Angus beef calves are given in Table 2.1. Each calf has a birthweight and all calves were weaned and weighed on the same day. Calves were born on different days, but were all weaned on the same day, and therefore, were different ages when they were weighed. Based on this data, how should the calves be ranked?

Table 2.1 Angus beef calf birth and weaning weight data.

Calf	BW(kg)	Weaning	
		Age(days)	Weight(kg)
1	30	180	180
2	28	192	198
3	36	204	200
4	31	210	224
5	29	195	205
6	35	200	199
7	25	208	212
8	40	216	222
9	32	198	209
10	34	205	195

To rank calves, beef producers commonly compute an adjusted 200-day weight. The assumption is made that growth in this stage of life is linear. The formula is

$$\text{Adj. 200-d Wt} = 200 \times (\text{Actual Wt} - \text{BW})/(\text{Age}) + \text{BW}.$$

Take the first calf, as an example,

$$\text{Adj. 200-d Wt} = 200\ d \times \frac{(180\ kg\ -\ 30\ kg)}{180\ d} + 30\ kg = 197\ kg.$$

Alternatively, producers may just compare the average daily gains (ADG) of the calves.

$$\text{ADG} = (\text{Actual Wt} - \text{BW})/(\text{Age}) = \frac{(180\ kg\ -\ 30\ kg)}{180\ d} = .833\ kg/d.$$

Table 2.2 Angus beef calf birth and weaning weight data with adjusted 200-d weight and average daily gain.

Calf	BW(kg)	Weaning		Adjusted	Ave. Daily
		Age(days)	Weight(kg)	200-d Wt.(kg)	Gain (kg/d)
1	30	180	180	197	.833
2	28	192	198	205	.885
3	36	204	200	197	.804
4	31	210	224	215	.919
5	29	195	205	210	.903
6	35	200	199	199	.820
7	25	208	212	205	.899
8	40	216	222	209	.843
9	32	198	209	211	.894
10	34	205	195	191	.785

2 Populations and Samples

A **population** refers to a group of animals that are part of the overall breeding structure in an industry. Examples are,

- Holstein dairy cattle in Canada that are on milk recording programs.
- Labrador retrievers in Ontario.
- Rainbow trout on the east coast of Canada.
- Racing pigeons of Quebec.

Populations have parameters that describe the means and variances of traits that are observed on that population. The population mean for a trait is designated by the Greek letter mu, μ. The population standard deviation for a trait is designated by a Greek sigma, σ. Population parameters need to be estimated for use in genetic evaluation. These are estimated from samples of animals from the population.

A **sample** is a subset of animals from the population. For example, the population of Holstein cows in Canada can be split into samples within each province. A sample might be cows in one herd.

3 Sample Means, Variances, Covariances, and Correlations

Let y_i be an observed trait value on an animal in the sample from the overall population, and let there be N such observations. The observation is composed of the population mean and a deviation (e_i) from that mean, i.e.,

$$y_i \;=\; \mu \;+\; e_i.$$

An unbiased estimator of the population mean is

$$\hat{\mu} \;=\; (\sum_{i=1}^{N} y_i)/N,$$

where $\sum$ is the summation symbol which means to add together the y_i's.

For the Angus calves in Table 2.2,

$$\begin{aligned}
\hat{\mu}_{BW} &= (30+28+36+31+29+35+25+40+32+34)/10 \\
&= 32, \\
\hat{\mu}_{200-d} &= (197+205+\cdots+191)/10 \;=\; 203.9, \quad \text{and} \\
\hat{\mu}_{ADG} &= (.833+.885+\cdots+.785)/10 \;=\; .8585.
\end{aligned}$$

Variance is an indicator of the range of possible values that y_i could have. For example, if the minimum value of y_i was 76 and the maximum value was 82, then the variance would be smaller than if the minimum was 25 and the maximum was 130. An estimator of the population variance is

$$\begin{aligned}
\hat{\sigma}^2 &= (\sum_{i=1}^{N} y_i^2 \;-\; (\sum_{i=1}^{N} y_i)^2/N)/(N-1), \\
&= \sum_{i=1}^{N} (y_i - \hat{\mu})^2/(N-1).
\end{aligned}$$

Using the data on the Angus calves,

$$\begin{aligned}
\hat{\sigma}^2_{BW} &= ((30-32)^2 \;+\; (28-32)^2 \;+\cdots+ (34-32)^2)/9, \\
&= 19.1111, \\
\hat{\sigma}^2_{200-d} &= ((197-203.9)^2 \;+\cdots+\; (191-203.9)^2)/9, \\
&= 58.3222, \\
\hat{\sigma}^2_{ADG} &= (7.390211 - 8.585(.8585))/9, \\
&= .00222.
\end{aligned}$$

Coefficient of Variation is a way to represent the degree of variation relative to the size of the mean,

$$CV = \frac{\hat{\sigma}}{\hat{\mu}} \times 100\%.$$

For birthweight, as an example,

$$CV = \frac{4.3716}{32} \times 100\% = 13.66\%,$$

while for 200-d weaning weight,

$$CV = \frac{7.6369}{203.9} \times 100\% = 3.75\%,$$

and for ADG is

$$CV = \frac{.04714}{.8585} \times 100\% = 5.49\%.$$

Larger values are better than small values of CV, in that there is a greater chance to make a change in the trait. Most traits of economic importance range from 5 to 20 %.

Covariance is used to measure how two traits vary together. Let y_i be one trait, like birthweight, and let w_i be a different trait, like adjusted 200-d weight, both measured on the same animal. An estimator of the population covariance is

$$\begin{aligned} \hat{\sigma}_{yw} &= (\sum_{i=1}^{N} y_i w_i \; - \; (\sum_{i=1}^{N} y_i)(\sum_{i=1}^{N} w_i)/N)/(N-1), \\ &= \sum_{i=1}^{N} (y_i - \hat{\mu}_y)(w_i - \hat{\mu}_w)/(N-1). \end{aligned}$$

Applied to the Angus calves, the covariances were

$$\begin{aligned} \hat{\sigma}_{BW-200d} &= (65193 - (320(2039))/10)/9 \; = \; -6.1111, \\ \hat{\sigma}_{BW-ADG} &= (273.583 - (320(8.585))/10)/9 \; = \; -.1263, \\ \hat{\sigma}_{200d-ADG} &= (1753.36 - (2039(8.585))/10)/9 \; = \; .3198. \end{aligned}$$

Covariances may be positive or negative. A positive covariance means that as one trait becomes larger in magnitude, so does the other trait. A negative covariance means that as one trait becomes larger the other trait becomes smaller. An easier way of looking at co-variation among traits is the **correlation coefficient**,

$$\hat{\rho} = \frac{\hat{\sigma}_{yw}}{(\hat{\sigma}_y^2 \; \hat{\sigma}_w^2)^{.5}},$$

so that

$$\begin{aligned}
\hat{\rho}_{BW-200d} &= \frac{-6.1111}{(19.1111 \times 58.3222)^{.5}} = -.183, \\
\hat{\rho}_{BW-ADG} &= \frac{-.1263}{(19.1111 \times .00222)^{.5}} = -.613, \\
\hat{\rho}_{200d-ADG} &= \frac{.3198}{(58.3222 \times .00222)^{.5}} = .889.
\end{aligned}$$

Correlation coefficients range between -1 and +1. Thus, weight at 200-days is highly correlated with average daily gain in this sample of animals, but BW and 200d weight are negatively correlated, but not too strongly.

4 Normal Distribution

Many quantitative traits of importance in livestock production follow the Normal Frequency Distribution. Every Normal distribution can be described entirely by its mean and variance as N(mean, variance). A Normal distribution with a mean of zero and a variance of one is known as the *standard Normal distribution* ($N(0,1)$). A more general formulation is

$$y_i \sim N(\mu, \sigma^2),$$

where y_i is the trait, μ is the mean of the population, and σ^2 is the variance of the observations.

Table 1. Some commonly used values for the standard Normal distribution.

z-values	Percentage Point	Confidence Interval	Selection Intensity
-3.00	99.9	99.8	.004
-2.50	99.4	98.8	.02
-2.00	97.7	95.4	.06
-1.50	93.3	86.6	.16
-1.00	84.1	68.2	.29
-0.50	69.2	38.4	.51
0.00	50.0	0.0	.80
0.50	30.8	38.4	1.14
1.00	15.9	68.2	1.53
1.50	6.7	86.6	1.94
2.00	2.3	95.4	2.37
2.50	0.6	98.8	2.83
3.00	0.1	99.8	3.41
-2.33	99.0	98.0	.03
-1.65	95.0	90.0	.11
-1.28	90.0	80.0	.20
-0.84	80.0	60.0	.35
-0.67	75.0	50.0	.43
-0.52	70.0	40.0	.50
-0.25	60.0	20.0	.64
0.00	50.0	0.0	.80
0.25	40.0	20.0	.97
0.52	30.0	40.0	1.16
0.67	25.0	50.0	1.26
0.84	20.0	60.0	1.40
1.28	10.0	80.0	1.75
1.65	5.0	90.0	2.06
1.75	4.0	92.0	2.15
1.88	3.0	94.0	2.27
2.05	2.0	96.0	2.42
2.33	1.0	98.0	2.67

A few points to remember:

- z-value ($z = (x_i - \mu)/\sigma$) is the trait value expressed as a difference from the population mean in standard deviation units.
- Percentage point (p) gives the portion of the population above the given z-value.

- Confidence interval gives the portion of the distribution within z-value units of the mean, i.e. between $-z$ and $+z$ on the horizontal axis.
- Selection intensity (i) is the average value (in standard deviation units and deviated from the mean) of the portion p of the population which lies above the z-value.
- The distribution is symmetric, with 50% above and 50% below the mean.
- Two-thirds (2/3) of the distribution, or about 67%, is within one standard deviation from the mean (i.e. between z-values of -1.0 and +1.0 on the standard Normal curve).
- About 95% of the distribution is within 2 standard deviations from the mean.

These rules apply to any trait that follows a Normal distribution, by first standardizing the distribution by converting the observations for the trait (y_i) to z-values. For example, if

$$y_i \sim N(\mu, \sigma^2),$$

then y_i can be converted into a z-value as follows:

$$z_i = \frac{y_i - \mu}{\sigma}.$$

The Normal distribution applies to the majority of traits recorded on livestock populations, but occasionally this is not the case. Examples of non-normality are as follows:

1. An animal either has a disease or does not have a disease (yes or no trait), which is a binomial distribution specified by p, a probability of having the disease.
2. Number of piglets born in a litter can be anywhere from 7 to 13 usually. The number born follows a Poisson distribution.
3. Calvings in cattle are categorized into 4 or 5 classes which range from Easy or Unassisted Calving, Assisted Calving, Difficult Calving, and Caesarian section. This is an example of a multinomial trait, i.e. more than two categories and probabilities associated with being in each.

In many cases, traits are assumed to follow a normal distribution even if they do not, and the results are almost as good as using the more appropriate distribution. Distributions other than normal are often more complicated computationally. In

practice, the first attempt should be to use the most appropriate distribution before making any simplification to a normal distribution. This course will only consider traits to follow a normal distribution.

Matrix Algebra
Fall 2008

1 Vectors and Matrices

Matrix algebra is a notation for representing arrays of items (usually data) and for theoretical derivation of methodology in a general manner. Only very basic matrix algebra is needed for this course.

A *vector* is a single column of numbers. Vectors are denoted by boldfaced small letters. Thus, examples of three vectors would be

$$\mathbf{y}_1 = \begin{pmatrix} 30 \\ 28 \\ 36 \\ 31 \\ 29 \\ 35 \\ 25 \\ 40 \\ 32 \\ 34 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 197 \\ 205 \\ 197 \\ 215 \\ 210 \\ 199 \\ 205 \\ 209 \\ 211 \\ 191 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} .833 \\ .885 \\ .804 \\ .919 \\ .903 \\ .820 \\ .899 \\ .843 \\ .894 \\ .785 \end{pmatrix}.$$

Row vectors are indicated by an apostrophe, e.g. $\mathbf{y}_1'$, which would have one row and 10 columns, i.e. $\mathbf{y}_1'$ is the transpose of $\mathbf{y}_1$.

A *matrix* is a two-dimensional array of numbers like a table, composed of rows and columns. The dimensions of a matrix are the number of rows and number of columns. The general designation of a matrix is a boldfaced upper case letter, and scalars are regular lower case letters, such as

$$\mathbf{M} = \{m_{ij}\},$$

where m_{ij} is the element in row i and column j. For example, the matrix containing $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$ would be

$$\mathbf{M} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 \end{pmatrix}$$

$$= \begin{pmatrix} 30 & 197 & .833 \\ 28 & 205 & .885 \\ 36 & 197 & .804 \\ 31 & 215 & .919 \\ 29 & 210 & .903 \\ 35 & 199 & .820 \\ 25 & 205 & .899 \\ 40 & 209 & .843 \\ 32 & 211 & .894 \\ 34 & 191 & .785 \end{pmatrix},$$

with 10 rows and 3 columns, and the element in the 6^{th} row and 2^{nd} column would be $m_{62} = 199$.

A special vector, **1**, has every element equal to 1, and a special matrix, **J**, has every element equal to 1. Concatenate **1** with **M** to get

$$\mathbf{W} = \begin{pmatrix} \mathbf{1} & \mathbf{Y} \end{pmatrix},$$

$$= \begin{pmatrix} 1 & 30 & 197 & .833 \\ 1 & 28 & 205 & .885 \\ 1 & 36 & 197 & .804 \\ 1 & 31 & 215 & .919 \\ 1 & 29 & 210 & .903 \\ 1 & 35 & 199 & .820 \\ 1 & 25 & 205 & .899 \\ 1 & 40 & 209 & .843 \\ 1 & 32 & 211 & .894 \\ 1 & 34 & 191 & .785 \end{pmatrix},$$

2 Addition of matrices

Matrices are *conformable for addition* if they have the same order. The resulting sum is a matrix having the same number of rows and columns as the two matrices to be added. Matrices that are not of the same order cannot be added together. If $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$, then

$$\mathbf{A} + \mathbf{B} = \{a_{ij} + b_{ij}\}.$$

For example, let

$$\mathbf{A} = \begin{pmatrix} 4 & 5 & 3 \\ 6 & 0 & 2 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 4 & 1 \end{pmatrix}$$

then

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 4+1 & 5+0 & 3+2 \\ 6+3 & 0+4 & 2+1 \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 5 & 5 \\ 9 & 4 & 3 \end{pmatrix} = \mathbf{B} + \mathbf{A}.$$

Subtraction is the addition of two matrices, one of which has all elements multiplied by a minus one (-1). That is,

$$\mathbf{A} + (-1)\mathbf{B} = \begin{pmatrix} 3 & 5 & 1 \\ 3 & -4 & 1 \end{pmatrix}.$$

3 Multiplication of Matrices

Two matrices are *conformable for multiplication* if the number of columns in the first matrix equals the number of rows in the second matrix. If $\mathbf{C}$ has order $p \times q$ and $\mathbf{D}$ has order $m \times n$, then the product $\mathbf{CD}$ exists only if $q = m$. The product matrix has order $p \times n$. In general, $\mathbf{CD}$ does not equal $\mathbf{DC}$, and most often the product $\mathbf{DC}$ may not even exist because $\mathbf{D}$ may not be conformable for multiplication with $\mathbf{C}$. Thus, the ordering of matrices in a product must be carefully and precisely written.

The computation of a product is defined as follows: let

$$\mathbf{C}_{p\times q} = \{c_{ij}\}$$

and

$$\mathbf{D}_{m\times n} = \{d_{ij}\}$$

and $q = m$, then

$$\mathbf{CD}_{p\times n} = \{\sum_{k=1}^{m} c_{ik}d_{kj}\}.$$

As an example, let

$$\mathbf{C} = \begin{pmatrix} 6 & 4 & -3 \\ 3 & 9 & -7 \\ 8 & 5 & -2 \end{pmatrix} \text{ and } \mathbf{D} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{pmatrix},$$

then

$$\mathbf{CD} = \begin{pmatrix} 6(1)+4(2)-3(3) & 6(1)+4(0)-3(-1) \\ 3(1)+9(2)-7(3) & 3(1)+9(0)-7(-1) \\ 8(1)+5(2)-2(3) & 8(1)+5(0)-2(-1) \end{pmatrix} = \begin{pmatrix} 5 & 9 \\ 0 & 10 \\ 12 & 10 \end{pmatrix}.$$

Let $\mathbf{y}$ be the vector of birthweights given earlier, with 10 rows and 1 column, then the product $\mathbf{y}'\mathbf{y}$ would have 1 row and 1 column, or a scalar quantity, and the result would be the sum of squares of the elements of $\mathbf{y}$.

$$\begin{aligned} \mathbf{y}'\mathbf{y} &= (30^2 + 28^2 + 36^2 + \cdots + 34^2), \\ &= 10,412. \end{aligned}$$

If $\mathbf{1}$ is a vector of 10 rows and 1 column with all elements equal to 1, then

$$\begin{aligned}\mathbf{1}'\mathbf{y} &= (1*30+1*28+1*36+\cdots+1*34),\\ &= \sum_{i=1}^{N} y_i,\\ &= 320, \quad \text{and}\\ \mathbf{1}'\mathbf{1} &= 10.\end{aligned}$$

4 Samples, Means, Variances, Covariances, and Correlations

Use the matrix $\mathbf{W}$ given earlier. First, multiply $\mathbf{W}'$ times $\mathbf{W}$. Note that $\mathbf{W}'$ has 4 rows and 10 columns, while $\mathbf{W}$ has ten rows and 4 columns. Thus, the resulting product will have 4 rows and 4 columns.

$$\begin{aligned}\mathbf{W} &= \begin{pmatrix}\mathbf{1} & \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3\end{pmatrix},\\ \mathbf{W}'\mathbf{W} &= \begin{pmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{y}_1 & \mathbf{1}'\mathbf{y}_2 & \mathbf{1}'\mathbf{y}_3\\ \mathbf{y}_1'\mathbf{1} & \mathbf{y}_1'\mathbf{y}_1 & \mathbf{y}_1'\mathbf{y}_2 & \mathbf{y}_1'\mathbf{y}_3\\ \mathbf{y}_2'\mathbf{1} & \mathbf{y}_2'\mathbf{y}_1 & \mathbf{y}_2'\mathbf{y}_2 & \mathbf{y}_2'\mathbf{y}_3\\ \mathbf{y}_3'\mathbf{1} & \mathbf{y}_3'\mathbf{y}_1 & \mathbf{y}_3'\mathbf{y}_2 & \mathbf{y}_3'\mathbf{y}_3\end{pmatrix},\\ &= \begin{pmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Y}\\ \mathbf{Y}'\mathbf{1} & \mathbf{Y}'\mathbf{Y}\end{pmatrix},\\ &= \begin{pmatrix}10 & 320 & 2039 & 8.585\\ 320 & 10,412 & 65,193 & 273.583\\ 2039 & 65,193 & 416,277 & 1753.36\\ 8.585 & 273.583 & 1753.36 & 7.390211\end{pmatrix}.\end{aligned}$$

The product, $\mathbf{W}'\mathbf{W}$, is therefore a matrix containing sums of the $\mathbf{y}$-vectors, sums of squares of those vectors, and sums of cross-products. For example,

$$1753.36 = 197(.833)+205(.885)+\cdots+191(.785),$$

is the sum of products of $\mathbf{y}_2$ elements with $\mathbf{y}_3$ elements summed together.

A matrix of variances and covariances, $\mathbf{V}$, can be obtained as follows:

$$\begin{aligned}\mathbf{V} &= (\mathbf{Y}'\mathbf{Y}-(\mathbf{Y}'\mathbf{1})(\mathbf{1}'\mathbf{Y})/N)/(N-1),\\ &= \frac{1}{9}\begin{pmatrix}10,412 & 65,193 & 273.583\\ 65,193 & 416,277 & 1753.36\\ 273.583 & 1753.36 & 7.390211\end{pmatrix}-\frac{1}{9}\begin{pmatrix}320\\ 2039\\ 8.585\end{pmatrix}\begin{pmatrix}32 & 203.9 & .8585\end{pmatrix},\\ &= \begin{pmatrix}19.1111 & -6.1111 & -.1263\\ -6.1111 & 58.3222 & .3198\\ -.1263 & .3198 & .00222\end{pmatrix}.\end{aligned}$$

The *diag()* function makes all of the off-diagonal elements of a matrix with the same number of rows as columns equal to zero. Thus,

$$\mathbf{D} = diag(\mathbf{V}) = \begin{pmatrix} 19.1111 & 0 & 0 \\ 0 & 58.3222 & 0 \\ 0 & 0 & .00222 \end{pmatrix}.$$

Now take the square root of the diagonals,

$$\mathbf{D}^{.5} = \begin{pmatrix} 4.3716 & 0 & 0 \\ 0 & 7.6369 & 0 \\ 0 & 0 & .04714 \end{pmatrix}.$$

The inverse of a diagonal matrix is created by dividing each diagonal element into 1, and is designated as

$$(\mathbf{D}^{.5})^{-1} = \mathbf{D}^{-.5} = \begin{pmatrix} .22875 & 0 & 0 \\ 0 & .13094 & 0 \\ 0 & 0 & 21.21321 \end{pmatrix}.$$

The correlation matrix, $\mathbf{C}$, is then

$$\begin{aligned} \mathbf{C} &= \mathbf{D}^{-.5}\mathbf{V}\mathbf{D}^{-.5}, \\ &= \begin{pmatrix} 1.000 & -.183 & -.613 \\ -.183 & 1.000 & .889 \\ -.613 & .889 & 1.000 \end{pmatrix}. \end{aligned}$$

5 Inversion of Matrices

The inverse of a square matrix (i.e. same number of rows and columns) is a matrix such that the product of the inverse with the original matrix gives an *Identity* matrix. An identity matrix is a diagonal matrix with all diagonals equal to 1, and all off-diagonal elements equal to 0. If $\mathbf{M}$ is the original matrix, then $\mathbf{M}^{-1}$ is the inverse and

$$\mathbf{M}\mathbf{M}^{-1} = \mathbf{M}^{-1}\mathbf{M} = \mathbf{I}.$$

Computing the inverse of a matrix is beyond the scope of this course, and so computers will be used to calculate them when they are needed. Inverses are needed to solve systems of equations. An example of a matrix and its inverse is shown below. Let a system of equations be

$$\mathbf{M}\mathbf{x} = \mathbf{r},$$

$$\mathbf{Mx} = \begin{pmatrix} 6 & -1 & 2 \\ 3 & 4 & -5 \\ 1 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$
$$\mathbf{r} = \begin{pmatrix} 81 \\ -51 \\ -11 \end{pmatrix},$$

then

$$\mathbf{M}^{-1} = \frac{-1}{57} \begin{pmatrix} -8 & -2 & -3 \\ 1 & -14 & 36 \\ -4 & -1 & 27 \end{pmatrix},$$

and the solutions are calculated by pre-multiplying the inverse times both sides of the equation,

$$\mathbf{M}^{-1}\mathbf{Mx} = \mathbf{M}^{-1}\mathbf{r},$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 9 \\ -7 \\ 10 \end{pmatrix}.$$

In genetic evaluation there is at least one equation for every animal in the pedigree file. There are other equations for factors that have a systematic effect on biological traits, such as age of the animal, temperature, management. There could be a million or more equations with a million or more unknowns to be estimated. Whatever the size, they can be simply represented as $\mathbf{Mx} = \mathbf{r}$.

Genetic Relationships

Fall 2008

1 Genomic Relationships

For every individual, there are a set of genes from the male parent and a set from the female parent. A set represents a random half of the alleles at each gene locus. Every progeny receives a different random half from the parent. The genomic relationship matrix, **G**, is just a large table consisting of the probabilities that alleles are in common between different kinds of relatives. Consider the pedigrees on the following five animals.

Example pedigree information on five animals.

Animal	Sire	Dam
A	-	-
B	-	-
C	A	B
D	A	C
E	D	B

Expand this table to identify the genomic pedigree structure. For any animal, X, let X_m and X_f represent the alleles inherited from the male and female parents, respectively.

Genomic pedigree structure of example pedigree.

Animal	Genome	Parent(m)	Parent(f)
A	A_m	-	-
A	A_f	-	-
B	B_m	-	-
B	B_f	-	-
C	C_m	A_m	A_f
C	C_f	B_m	B_f
D	D_m	A_m	A_f
D	D_f	C_m	C_f
E	E_m	D_m	D_f
E	E_f	B_m	B_f

The genomic relationship matrix will be of order 10. The diagonals of any genomic relationship matrix are always equal to 1. The probability of X_m having alleles in

common with X_m is always 1 because $X_m = X_m$.

		A		B		C		D		E	
		A_m	A_f	B_m	B_f	C_m	C_f	D_m	D_f	E_m	E_f
A	A_m	1	0	0	0						
	A_f	0	1	0	0						
B	B_m	0	0	1	0						
	B_f	0	0	0	1						
C	C_m					1					
	C_f						1				
D	D_m							1			
	D_f								1		
E	E_m									1	
	E_f										1

Because the parents of A and B are unknown, then they are assumed to be random individuals from a large random mating population and assumed to have no alleles identical by descent between them.

Let (A_m,C_m) indicate an element in the above table between the A_m male parent contribution of animal A and the C_m male parent contribution of animal C, then the value that goes into that location is the probability that A_m is related to the male and female parent contributions to animal C. Because C_m comes from animal A, then the probability that A_m and C_m share common alleles is

$$\begin{aligned}(A_m, C_m) &= 0.5 * [(A_m, A_m) + (A_m, A_f)] \\ &= 0.5 * [1 + 0] \\ &= 0.5\end{aligned}$$

Similarly, for the rest of the A_m row,

$$(A_m, C_f) = 0.5 * [(A_m, B_m) + (A_m, B_f)] = 0,$$

$$
\begin{aligned}
(\mathrm{A}_m, \mathrm{D}_m) &= 0.5 * [(\mathrm{A}_m, \mathrm{A}_m) + (\mathrm{A}_m, \mathrm{A}_f)] = 0.5, \\
(\mathrm{A}_m, \mathrm{D}_f) &= 0.5 * [(\mathrm{A}_m, \mathrm{C}_m) + (\mathrm{A}_m, \mathrm{C}_f)] \\
&= 0.5 * [0.5 + 0] = 0.25, \\
(\mathrm{A}_m, \mathrm{E}_m) &= 0.5 * [(\mathrm{A}_m, \mathrm{D}_m) + (\mathrm{A}_m, \mathrm{D}_f)] \\
&= 0.5 * [0.5 + 0.25] = 0.375, \\
(\mathrm{A}_m, \mathrm{E}_f) &= 0.5 * [(\mathrm{A}_m, \mathrm{B}_m) + (\mathrm{A}_m, \mathrm{B}_f)] = 0.
\end{aligned}
$$

The A_m column is equal to the A_m row, and therefore

$$
\begin{aligned}
(\mathrm{C}_m, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{C}_m) = 0.5, \\
(\mathrm{C}_f, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{C}_f) = 0, \\
(\mathrm{D}_m, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{D}_m) = 0.5, \\
(\mathrm{D}_f, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{D}_f) = 0.25, \\
(\mathrm{E}_m, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{E}_m) = 0.375, \\
(\mathrm{E}_f, \mathrm{A}_m) &= (\mathrm{A}_m, \mathrm{E}_f) = 0.
\end{aligned}
$$

This recursive method of calculating probabilities works as long as the animals are arranged chronologically, (parents come before progeny), and each row (column) should be completed before proceeding to the next row of the table. The complete table of genomic relationships is given below.

		A		B		C		D		E	
		A_m	A_f	B_m	B_f	C_m	C_f	D_m	D_f	E_m	E_f
A	A_m	1	0	0	0	.5	0	.5	.25	.375	0
	A_f	0	1	0	0	.5	0	.5	.25	.375	0
B	B_m	0	0	1	0	0	.5	0	.25	.125	.5
	B_f	0	0	0	1	0	.5	0	.25	.125	.5
C	C_m	.5	.5	0	0	1	0	.5	.5	.5	0
	C_f	0	0	.5	.5	0	1	0	.5	.25	.5
D	D_m	.5	.5	0	0	.5	0	1	.25	.625	0
	D_f	.25	.25	.25	.25	.5	.5	.25	1	.625	.25
E	E_m	.375	.375	.125	.125	.5	.25	.625	.625	1	.125
	E_f	0	0	.5	.5	0	.5	0	.25	.125	1

Notice the diagonal boxes for animals D and E. There is a probability of 0.25 that the alleles coming from male parent of D are shared with those of the female parent. Animal D is said to be **inbred**, and the **inbreeding coefficient** is 0.25. The female parent of D is animal C whose parent was animal A, which is the other parent of animal D. Thus, the alleles from animal A can occur in animal D from both sides of the pedigree with probability of 0.25. Inbreeding also means that 0.25 of the gene loci are expected to be homozygous (i.e. that same alleles).

For animal E, the probability is 0.125 that alleles are shared between the male and female contributions of its parents. Both animals A and B occur on both sides of the pedigree for animal E.

2 Additive Genetic Relationships

Both the additive and dominance relationship matrices may be obtained from the genomic relationship matrix. The additive relationship matrix gives the expected genetic variances or covariances between animals. The **additive relationship** between animals A and C, a_{AC}, for example, is given by

$$\begin{aligned} a_{AC} &= 0.5 * [(A_m, C_m) + (A_m, C_f) + (A_f, C_m) + (A_f, C_f)] \\ &= 0.5 * [0.5 + 0.0 + 0.5 + 0.0] = 0.5. \end{aligned}$$

Add the four numbers in each square of the table and divide by 2 (or multiply by 0.5). The **A** matrix is then

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & .5 & .75 & .375 \\ 0 & 1 & .5 & .25 & .625 \\ .5 & .5 & 1 & .75 & .625 \\ .75 & .25 & .75 & 1.25 & .75 \\ .375 & .625 & .625 & .75 & 1.125 \end{pmatrix}.$$

Note that values on the diagonals can go from 1 to 2, and off-diagonal elements range from 0 to 2. All of these elements are multiplied by the additive genetic variance, σ_a^2, to get the expected additive genetic variance or covariance between two individuals. Inbred individuals are expected to have a larger genetic variance between inbred individuals than between non-inbred individuals.

3 Dominance Genetic Relationships

Dominance effects occur from the specific combination of the male and female alleles of the parents. Two animals with the same parents could inherit the same specific combination of male and female alleles. This is measured by the dominance genetic relationship derived from the genomic relationship matrix. In general, the **dominance genetic relationship** between animals X and Y, d_{XY}, is given by

$$d_{XY} = (\mathrm{X}_m, \mathrm{Y}_m) * (\mathrm{X}_f, \mathrm{Y}_f) + (\mathrm{X}_m, \mathrm{Y}_f) * (\mathrm{X}_f, \mathrm{Y}_m).$$

For example, between animals D and E above, d_{DE} is

$$\begin{aligned} d_{DE} &= (\mathrm{D}_m, \mathrm{E}_m) * (\mathrm{D}_f, \mathrm{E}_f) + (\mathrm{D}_m, \mathrm{E}_f) * (\mathrm{D}_f, \mathrm{E}_m) \\ &= 0.625 * 0.25 + 0.625 * 0.0 \\ &= 0.15625 + 0.0 \\ &= 0.15625. \end{aligned}$$

The complete dominance relationship matrix is a matrix of expected dominance

genetic variances and covariances among animals.

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & .25 & 0 \\ 0 & 1 & 0 & 0 & .125 \\ 0 & 0 & 1 & .25 & .25 \\ .25 & 0 & .25 & 1.0625 & .15625 \\ 0 & .125 & .25 & .15625 & 1.015625 \end{pmatrix}.$$

4 Epistatic Genetic Relationships

All other possible gene interaction relationships can be computed from elements of **A** and **D**. For example, the relationship for the interaction of a additive genetic effect at locus A with an additive genetic effect at locus Z would be

$$a_{AZ} \times a_{AZ}.$$

The relationship for the interaction between the additive genetic effect at locus A with the dominance genetic effect at locus Z would be

$$a_{AZ} \times d_{AZ}.$$

The interactions can be as complex as desired, i.e. 3-way, 4-way, m-way interactions.

5 Meuwissen and Luo Algorithm for Inbreeding

The techniques of Meuwissen and Luo (1992) are used to find inbreeding coefficients. For each animal, a quantity, b_i, is computed, where

$$b_i \ = \ 0.5 \ - \ 0.25(F_s + F_d),$$

and F_s and F_d are the inbreeding coefficients of the sire and dam of animal i. If one of the parents is unknown, then $b_i = 0.75 - 0.25(F_p)$ where F_p is the inbreeding coefficient of the known parent. If both parents are unknown, then $b_i = 1$.

Animals must be ordered such that parents appear in the pedigree list before any of their progeny. Below is an example pedigree list, and the inbreeding coefficients and b_i values are given for all animals except the last two.

Example pedigree list.

Animal	Sire	Dam	F_i	b_i
A			0.0	1.0
B			0.0	1.0
C			0.0	1.0
D	A	B	0.0	0.5
E	A	C	0.0	0.5
F	D	E		0.5
G	A	F		

The mechanics of the algorithm to find the inbreeding coefficient are as follows:

1. Construct a table with three columns. The first column will contain animal IDs, the second column will contain one half to the power equal to the number of generations back in time, and the third column contains the b_i value of the animal. For animal F, the table begins as follows:

ID	t_i	b_i
F	1.0	0.5
D	0.5	0.5
E	0.5	0.5

   Note that $t_F = \frac{1}{2}^0 = 1$, and because D and E are the parents of F, then $t_D = t_E = \frac{1}{2}^1 = \frac{1}{2}$. The b_i values just come from the table above.

2. Now add the parents of animal D to the list.

ID	t_i	b_i
F	1.0	0.5
D	0.5	0.5
E	0.5	0.5
A	0.25	1.0
B	0.25	1.0

3. Add the parents of E to the list.

ID	t_i	b_i
F	1.0	0.5
D	0.5	0.5
E	0.5	0.5
A	0.25	1.0
B	0.25	1.0
A	0.25	1.0
C	0.25	1.0

4. The parents of A, B, and C are unknown so that no more animals can be added to the list.

5. Animal A appears twice in the list, and the two t_A values need to be added together. This should be done for any animal that appears more than once.

ID	t_i	b_i
F	1	0.5
D	$\frac{1}{2}$	0.5
E	$\frac{1}{2}$	0.5
A	$\frac{1}{2}$	1.0
B	$\frac{1}{4}$	1.0
C	$\frac{1}{4}$	1.0

6. The diagonal of **A** for animal F is calculated as

$$\begin{aligned} a_{FF} &= \sum_i t_i^2 b_i, \\ &= (1)^2(0.5) + (\frac{1}{2})^2(0.5) + (\frac{1}{2})^2(1) + (\frac{1}{4})^2(1) + (\frac{1}{4})^2(1), \\ &= \frac{18}{16}. \end{aligned}$$

In the additive genetic relationship matrix the diagonal is equal to 1 plus the inbreeding coefficient. Therefore,

$$F_F = a_{FF} - 1 \;=\; \frac{1}{8} \;=\; 0.125.$$

The same process is used for animal G. First, b_G is $(0.5 - 0.25(0.0 + 0.125)) = \frac{15}{32} = 0.46875$.

ID	t_i	b_i
G	1.0	0.46875
A	0.5	1.0
F	0.5	0.5
D	0.25	0.5
E	0.25	0.5
A	0.125	1.0
B	0.125	1.0
A	0.125	1.0
C	0.125	1.0

Animal A appears 3 times and the t_i values need to be added together, giving

ID	t_i	b_i
G	1.0	0.46875
A	0.75	1.0
F	0.5	0.5
D	0.25	0.5
E	0.25	0.5
B	0.125	1.0
C	0.125	1.0

Then

$$a_{GG} = 1(\frac{15}{32}) + (\frac{3}{4})^2(1) + (\frac{1}{2})^2(0.5) + 2(\frac{1}{4})^2(0.5) + 2(\frac{1}{8})^2(1),$$

or

$$a_{GG} = 1.25,$$

so that $F_G = 0.25$.

The complete table of inbreeding coefficients and b_i values is given below.

Example pedigree list.

Animal	Sire	Dam	F_i	b_i
A			0.0	1.0
B			0.0	1.0
C			0.0	1.0
D	A	B	0.0	0.5
E	A	C	0.0	0.5
F	D	E	0.125	0.5
G	A	F	0.25	0.46875

6 The Inverse

Let $\delta = b_i^{-1}$, then if both parents are known the following constants are added to the appropriate elements in the inverse matrix:

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

If one parent is unknown, then delete the appropriate row and column from the rules above, and if both parents are unknown then just add δ to the animal's diagonal element of the inverse.

Each animal in the pedigree is processed one at a time, but any order can be taken. Let's start with animal F as an example. The sire is animal D and the dam is animal E. In this case, $\delta = 2.0$. Following the rules and starting with an inverse matrix that is empty, the additions to the inverse matrix should appear as follows:

	A	B	C	D	E	F	G
A							
B							
C							
D				.5	.5	-1	
E				.5	.5	-1	
F				-1	-1	2	
G							

The contributions for each animal are accumulated into one matrix. Any elements that are empty have a zero in them. The least common denominator for this inverse is 30, so the complete inverse is

$$\mathbf{A}^{-1} = \frac{1}{30} \begin{pmatrix} 76 & 15 & 15 & -30 & -30 & 16 & -32 \\ 15 & 45 & 0 & -30 & 0 & 0 & 0 \\ 15 & 0 & 45 & 0 & -30 & 0 & 0 \\ -30 & -30 & 0 & 75 & 15 & -30 & 0 \\ -30 & 0 & -30 & 15 & 75 & -30 & 0 \\ 16 & 0 & 0 & -30 & -30 & 76 & -32 \\ -32 & 0 & 0 & 0 & 0 & -32 & 64 \end{pmatrix}.$$

Multiplying $\mathbf{A}^{-1}$ times $\mathbf{A}$ gives the expected $\mathbf{I}$, identity matrix. Using this algorithm the inverse for a relationship matrix for 4 million animals can be constructed in less than 10 minutes.

Writing Linear Models

Fall 2008

1 Introduction

A statistical model attempts to describe reality based upon variables that are observable. Statistical models are used to analyze all kinds of data. There are three parts to every model. Part 1 is an equation where the observation on a trait is described as being influenced by a list of factors (in an additive manner). The equation is written as

$$y_{ijkl} = \mu \ + \ A_i \ + \ B_j \ + \ C_k \ + \cdots + \ e_{ijkl},$$

where

y_{ijkl} is the observation on a trait of interest,

μ is the overall mean of the population,

A_i is the effect of factor A, level i, on the trait of interest,

B_j is the effect of factor B, level j, on the trait of interest,

C_k is the effect of factor C, level k, on the trait of interest, and

e_{ijkl} is a residual effect composed of all factors not observed.

The equation could contain any number of factors that influence the observed trait value. What are A, B, and C? Suppose y is the score of a dog at an obedience trial. Factor A could be the breed of dog, factor B could be the judge, and factor C could be the handler or trainer. Other factors such as the gender of the dog, the number of hours of training, number of previous obedience trials the dog may have participated, the conditions within the ring during the trial (noise and temperature conditions), and the number of competitors.

Part 2 of a model is an indication of which factors are fixed or random (see later). If a factor is random, then it is assumed to be a variable that is sampled from a population that has a particular mean and variance. The mean and variance should be specified. Determining whether a factor is fixed or random is not always easy, and takes experience in data analysis.

Part 3 of the model is a list of all implied or explicit assumptions or limitations about the first two parts. This part is often missing, but is important to be able

to judge the quality of the analysis. The best way to explain Part 3 is to give an example model.

2 Model for Weaning Weights of Beef Calves

Picture yourself as a beef calf and then try to think of the factors that would influence your growth and eventual weaning weight. For example,

$$y_{ijklm} = A_i + B_j + X_k + HYS_l + c_m + e_{ijklm},$$

where

y_{ijklm} is a weaning weight on a calf,

A_i is the age of the dam (in years), either 2, 3, 4, or 5 and greater,

B_j is a breed of calf effect,

X_k is a gender of calf effect (male or female),

HYS_l is a herd-year-season of birth effect, with three seasons per year (i.e. Nov-Feb, Mar-Jun, and Jul-Oct),

c_m is a calf additive genetic effect, and

e_{ijklm} is a residual effect.

The fixed factors are age of dam, breed of calf, and gender of calf. Herd-year-season effects, calf additive genetic effects, and residual effects are random. Instead of stating that the variance of calf additive genetic effects, for example, is 3000 kg^2, one could just say that the variance is 0.35 of the total variance, and herd-year-season effects comprise 0.15 of the total variance. The variance of residual effects is the remaining variation of 0.50 of the total. The means of the random effects are usually assumed to be zero. Calves could be related to each other because of a common sire, and/or related mothers. Thus, the analysis should take into account these relationships.

Part 3 of the model lists the assumptions and limitations of the data and model equation.

1. There are no interactions between age of dam, breed of calf, or gender of calf.

2. The weaning weights have been properly adjusted to a 200-d of age of calf weight.

3. There are no maternal effects on calf weaning weights.

4. Age of dam is known.

5. All calves in the same herd-year-season were raised and managed in the same manner.

A researcher would discuss the consequences of each assumption if it were not true. For example, if interactions among the fixed factors exist, then using this model might give biased estimates of age of dam, breed, and gender of calf, which might bias the estimates of calf additive genetic effects. However, So and So (1929) showed that interactions were negligible. (Note: this article would be considered to be too old to be used as a reference in 2006).

Maternal effects are known to exist for weaning weights. Thus, the model should be changed by adding a maternal genetic effect of the dam. Thus, the equation is revised, maternal genetic effects are another random factor, and the proportions of each to the total variance need to be revised. There is also a genetic correlation between calf additive genetic effects and the maternal genetic effects. (This is discussed more in the notes on Maternal Genetic Effects)

The last assumption may not be true in some herds, because owners sometimes separate male and female calves earlier than weaning. Also, some herds may be very large, and so there could be more than one management group within a herd-year-season. From the recorded data, this fact may not be obvious unless producers correctly fill in the management group codes.

For this course, students should be able to write an equation of the model (subscripts not necessary) in words, e.g.

$$\begin{aligned} \text{Wean. Wt.} &= \text{Age of dam} + \text{Breed} \\ &\quad + \text{Gender} + \text{HYS} \\ &\quad + \text{Calf} + \text{residual}. \end{aligned}$$

Then indicate the fixed and random factors, and the proportion of total variance for each random factor, and then a good attempt at Part 3.

3 Model Building

Developing an appropriate linear statistical model is best accomplished in discussions with other scientists. Full awareness of models that have been published in the literature for a particular species and trait is important. Model building, in the beginning, is a trial and error ordeal. The Analysis of Variance was created to allow factors in models to be tested for their significance. Factors that are significant should be in the model (for genetic evaluation). Sometimes factors that are not significant in your data, but which have consistently been important in previous studies, should also be included in the model. As more data accumulate, the model may need to be re-tested and refinements could be made. A genetic evaluation model will likely be used many times per year and over years. Therefore, scientists should be open towards making improvements to their models as new information becomes available.

4 Practice Models

Write a linear statistical model for one or more of the following cases. A similar case will be given on the mid-term exam.

Case 1. Body condition scores of cows during the lactation are assigned by the owner (from 1 to 5 in half increments, 1, 1.5, 2, 2.5,...), where 1 is very thin and lacking in condition, and 5 is very obese. A farmer has body condition scores on all cows every 30 days during the year. Write a model to analyze body condition scores.

Case 2. Beef bulls, at weaning, go to test stations for a 112 day growth test and the best bulls at the end of test are sold to beef producers in an auction. Growth, feed intake, and scrotal circumference are measured during the test period every 2 weeks. Write a model for either growth, feed intake, or scrotal circumference to evaluate the beef bulls. There are data from many test stations over the last 10 years. Several breeds and crossbreds are involved in the tests.

Case 3. Weight and length at two years of age in Atlantic cod are important growth traits. Fish are individually identified with pit tags. Fish are reared in tanks at a research facility with the capability of controlling water temperature and hours of daylight. Tanks differ somewhat in size and number of fish. Write a model for estimating the genetic variability in growth traits.

Case 4. Income from milk sales minus expenses for feed, breeding, and health problems from one calving to the next are available on many herds of dairy cows. Call the difference cow profit and write a model to analyze this trait for cows finishing their first lactation.

Case 5. A reproductive physiology study collected statistics on semen volume, sperm motility, and number of sperm per ejaculate on stallions from one year to ten years of age (on the same horses - a long term study) to see how semen characteristics change with age. Write a model to analyze one of these traits.

Case 6. Canadian Warmblood horses are raised for dressage and jumping. Mares can be sent to a central location for a brief training (breaking) period and are scored for a number of traits, such as gait and movement. Three experts score the horses as well as two riders, and the results are combined into a weighted average. Write a model for analyzing the combined averages on mares, from several test locations over several years.

Animal Models

Fall 2008

1 Introduction

An animal model is one in which there are one or more observations per animal, and all factors affecting those observations are described including an animal additive genetic effect. The animal additive genetic effects are random variables with an expected value of zero, and a covariance matrix that is equal to **A**, the additive genetic relationship matrix. Assumptions are that the trait of interest is influenced by an infinite number of loci each with a small, relatively equal effect, and that the population is randomly mating.

Animal models were first used in 1989, but the theory about these models was known since 1969. Animal models were not used before 1989 because computer power was not sufficient to handle so many equations. As computers became faster and had more memory, then the statistical models became more realistic, but also more complex.

2 Example Situation

Sheep are scanned at maturity by ultrasound(US) to determine the amount of fat surrounding the muscle. A model (equation) might be

$$\begin{aligned} \text{USFat} &= \text{YearMonth} + \text{FMG} + \text{b(Age at US)} \\ &\quad + \text{Animal} + \text{Residual} \end{aligned}$$

where

Year-month of birth is fixed,

FMG is a flock-year-management group effect (random),

Age at ultrasound is a covariate,

Animal additive genetic effects , and

Residual effects .

Fat thickness is in millimeters. Relationships among animals will be used. The purpose of the analysis is to estimate the variances, and afterwards to estimate the breeding values of the animals.

Animals are assumed to have only one US Fat measurement each, and that they have not been pre-selected on the basis of any other trait. The sex of the animal is assumed to not have any effect on the measurements. Within a FMG, all sheep are assumed to be treated and fed in the same manner.

3 Estimation of Variances

There are two methods of estimating variances that are used in animal breeding today. One is called Restricted Maximum Likelihood (or REML). REML has several different ways of being calculated. One is called Derivative Free REML (DFREML), and another is called Average Information REML (AIREML, ASREML). Other computational methods are too cumbersome or slow. Software is available for DFREML and ASREML from various sources (Denmark, Australia). To employ REML one needs to assume that the observations follow a normal distribution. Then the likelihood function can be written for the particular model. Both DFREML and ASREML try to maximize the log of the likelihood function, but in different ways. If both methods operate correctly, then both methods should give the same final answers. This does not always happen. The details of the methodology are too complex for this course.

The other method is known as the Bayesian method. Bayesian statisticians differ from traditional statisticians (known as Frequentists) because Bayesians assume that everything in a model is random. That means everything in the model comes from a population with a certain mean and variance. However, the Bayesians do not necessarily assume a normal distribution for everything. Even the variances that are to be estimated are assumed to be a random variable, and variances tend to have Chi-squared distributions. Fixed effects are assumed to have uniform distributions. Animal genetic effects and residual effects are usually assumed to have normal distributions.

The Bayesian methods indicate the distribution of every factor in the model equation, including the variances. Then the overall likelihood is the product of the likelihoods of all the factors in the model equation. This is the Joint Probability Function. The Bayesian method is to maximize the joint probability function. Usually this function is too complex to take derivates to find the maximum. To get around this problem Bayesians find the marginal probability functions of each fac-

tor assuming the parameters of all the other variables are known. Then a value is computed for an unknown parameter (based on its marginal probability function), and then a random amount is added or subtracted from that parameter depending on its expected variance (known as Gibbs sampling). Each unknown parameter in the joint probability function is treated this way, one at a time. One pass through all of the unknown parameters is one iteration or one sample. The Bayesians will perform as many iterations or samples as time permits - usually tens or hundreds of thousands of iterations. After some thousands of iterations, then the sample values of the unknown parameters begin to approximate samples from the joint probability function. The early samples are known as the 'burn-in' period. The averages of the sample values after the 'burn-in' period give an estimate of that parameter. The standard deviation of the sample values give the standard error of the estimates.

The Bayesian method is less limiting than REML because distributions other than normal can be utilized. The sampling process can take a long time, but software is easy to write for the Bayesian method. A good random number generator is needed for Gibbs sampling.

4 Comments

4.1 Examples

To illustrate either method of estimating variances is nearly impossible using small examples. Small examples tend not to give good results. If large examples are used, then too many pages of details need to be given. Thus, a good example is difficult to present.

4.2 Amount of Data

The estimation of variances requires data on at least a few thousand animals (2000 or more). The more animals that are included then the sharper will be the peak at the maximum of the likelihood function or joint probability function. With too few observations the peaks are less pronounced and find the maximum becomes more difficult. Success also depends on the model and the number of unknown parameters in the model.

4.3 Changes in Variances

Variance parameters tend to not change very much over time. This means that variances do not need to be re-estimated very often. Usually parameters need to be re-estimated every time the model is changed (adding or deleting factors to the model). Using estimates of variances that match the model is preferred. Of course, this will depend on the changes that were made.

4.4 Breed or Country Differences

Variance parameters may be specific to a breed. For example, the Holstein breed in dairy cattle generally has larger variances for milk production because Holsteins produce more milk than the other breeds. Charolais beef cattle grow more rapidly than Hereford or Angus. Variances may also be specific to a breed within a particular country. Holsteins in Canada have larger variances than Holsteins in South Africa or New Zealand.

4.5 Genetic Evaluation and Rankings of Animals

If heritability is estimated to be 0.30, then genetic evaluations that are calculated using either 0.20 or 0.40 would not greatly re-rank animals. By using 0.20 instead of 0.30, the estimated breeding values will have a smaller range in values, and using 0.40 the estimated breeding values will have a bigger range than those calculated using 0.30. Using the correct variance is important for measuring genetic trends, but not for ranking animals for selection.

5 Repeated Records on Animals

Often animals are observed more than once for a trait. However, animals get older between observations. An important question is whether the observation at one age is a different trait from the observation at a later age. Is the genetic correlation between the observations less than 1?

Assuming that the genetic correlation is not greatly less than 1, then there are **repeated** records on an animal. There are **permanent environmental** factors that are not genetic, but yet affect all observations on one animal. **Repeatability**,

r, is a number between 0 and 1 that reflects the degree of permanent environmental effects. Let σ_p^2 be the variance of permanent environmental effects, σ_a^2 is the additive genetic variance, σ_e^2 is the residual variance, and $\sigma_y^2 = \sigma_a^2 + \sigma_p^2 + \sigma_e^2$ then

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

and

$$h^2 = \frac{\sigma_a^2}{\sigma_y^2}.$$

For forming the mixed model equations, the ratios of residual variance to additive genetic variance, and of residual variance to permanent environmental variance are needed. Re-arranging the above formulas, then

$$\begin{aligned} r\sigma_y^2 &= \sigma_a^2 + \sigma_p^2, \\ h^2\sigma_y^2 &= \sigma_a^2, \\ \sigma_p^2 &= (r - h^2)\sigma_y^2, \\ \sigma_e^2 &= (1 - r)\sigma_y^2, \\ k_a &= \frac{\sigma_e^2}{\sigma_a^2} = \frac{(1-r)}{h^2}, \\ k_p &= \frac{\sigma_e^2}{\sigma_p^2} = \frac{(1-r)}{(r-h^2)}, \end{aligned}$$

Repeatability must always be greater than heritability.

5.1 Example on Horse Racing

Below are the time results of three training races on a mile and a quarter track. The races were held about 3 months apart and were always on the same track. The horses were all males at the same 'year' of age.

Time Results(seconds) for 3-yr-old Stallions.

Animal	Sire	Dam	Race 1 March	Race 2 June	Race 3 Sept
13	9	1	100	108	119
14	9	2	123	121	117
15	10	3	116		
16	10	4		112	133
17	11	5		117	
18	12	6	115		121
19	12	7	113	120	126
20	12	8			128

The best horse is the one with the lowest time. Animal 13 had the best time in races 1 and 2, but was beaten in race 3 by animal 14. Each horse did not necessarily compete in all three races. The rider of a horse was assumed to be the same for each race in which the horse competed. The linear statistical model for this example is

$$y_{ijk} = R_i + a_j + p_j + e_{ijk},$$

where y_{ijk} is the racing time of a horse in a particular race, R_i is the race effect (includes race conditions that day), a_j is the additive genetic effect of the animal, p_j is the permanent environmental effect of the animal, and e_{ijk} is the residual effect. Permanent environmental effects can only be estimated on animals that have raced, not on ancestors.

5.2 Results

For a heritability of 0.25 and a repeatability of 0.35, the variance ratios for the mixed model equations (MME) are

$$\begin{aligned} ka &= \frac{1-r}{h^2} = 2.6, \\ kp &= \frac{1-r}{r-h^2} = 6.5. \end{aligned}$$

In matrix notation the model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_p\mathbf{p} + \mathbf{e},$$

where $\mathbf{b}$ contains the race effects; $\mathbf{a}$ contains the animal additive genetic effects (for all 20 animals); $\mathbf{p}$ contains the permanent environmental effects for the 8 animals with records; and $\mathbf{e}$ are the residual effects. Previously, there was only one $\mathbf{X}$ and one $\mathbf{Z}$ matrices for the animal model. In a sense there is still only one $\mathbf{Z}$ matrix because

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_a & \mathbf{Z}_p \end{pmatrix}.$$

The random factors are $\mathbf{a}$ and $\mathbf{p}$, and the covariance matrix of those vectors is

$$\mathbf{G} = Var\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_p^2 \end{pmatrix}.$$

The inverse of this matrix times the residual variance is

$$\mathbf{G}^{-1}\sigma_e^2 = \begin{pmatrix} \mathbf{A}^{-1}\frac{1}{\sigma_a^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\frac{1}{\sigma_p^2} \end{pmatrix}\sigma_e^2,$$

which is

$$\mathbf{G}^{-1}\sigma_e^2 = \begin{pmatrix} \mathbf{A}^{-1}k_a & \mathbf{0} \\ \mathbf{0} & \mathbf{I}k_p \end{pmatrix}.$$

The mixed model equations are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_a & \mathbf{X}'\mathbf{Z}_p \\ \mathbf{Z}_a'\mathbf{X} & \mathbf{Z}_a'\mathbf{Z}_a + \mathbf{A}^{-1}k_a & \mathbf{Z}_a'\mathbf{Z}_p \\ \mathbf{Z}_p'\mathbf{X} & \mathbf{Z}_p'\mathbf{Z}_a & \mathbf{Z}_p'\mathbf{Z}_p + \mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_a'\mathbf{y} \\ \mathbf{Z}_p'\mathbf{y} \end{pmatrix}.$$

The solutions for race effects were 113.95 seconds for race 1, 115.90 seconds for race 2, and 124.13 seconds for race 3. The races became slower during the year, which may be due to the increased temperature during the summer months. There may be other explanations, like heavy rain before the September race.

The animal additive genetic effects and their standard errors of prediction and reliabilities are given in the table below, for the 8 animals that raced, and also their permanent environmental effects.

Estimates for Horses that Raced.

		Additive Genetic			Perm. Env.		
Horse	Sire	EBV	SEP	Rel	$\hat{\mathbf{p}}$	SEP	Rel
13	9	-3.76	2.82	0.35	-1.65	2.04	0.15
14	9	0.45	2.82	0.35	0.60	2.04	0.15
15	10	0.68	3.12	0.20	0.18	2.12	0.08
16	10	0.99	2.92	0.30	0.35	2.07	0.12
17	11	0.27	3.13	0.20	0.11	2.12	0.08
18	12	-0.14	2.95	0.29	-0.21	2.06	0.13
19	12	0.87	2.94	0.29	0.25	2.04	0.15
20	12	1.08	3.27	0.13	0.37	2.12	0.09

Animal 13 is genetically the fastest racer in the group, followed by animal 18, while the slowest was animal 20. Both animals 18 and 20 were sired by horse 12, but from different dams. There will always be good and poor progeny from each parent, but the average of a good sire should be better than the average of poorer sires. The solutions for the permanent environmental effects are similar in ranking to the EBVs, and they generally have a lower reliability than the EBVs. Why? Because σ_p^2 is often smaller than σ_a^2.

In an animal model where animals have only one record each and no progeny, the reliabilities of those EBVs can go no higher than the heritability of the trait. However, with the addition of progeny and repeated records per animal, then reliability can go as high as 100% (or close enough to it for all practical purposes).

This would require a large number of progeny and/or a large number of repeated records. Reliabilities account for 1) the number of records on the animal; 2) the number of progeny the animal has; 3) the number and type of relatives in the data; 4) the number of animals in each race; 5) the total number of observations in the data set; 6) the number of factors in the model; and 7) the variance parameters relative to the residual variance. Not all horses competed directly with all other horses. Animal 17 only raced in Race 2 and did not compete against animals 15, 18, or 20, but did compete against animal 19 which raced in all 3 races, thereby giving an indirect comparison of animal 17 to all other horses. All of this information is part of the MME.

Random Normal Deviates

-0.20	-1.86	0.74	0.09	0.14	1.27	-1.08	0.08
-0.32	2.38	0.30	0.13	-3.08	-0.58	-0.80	0.73
-0.91	-0.94	-0.02	1.42	-0.51	2.56	0.31	-0.72
-0.17	-0.77	-0.51	-0.24	-0.84	-0.55	0.03	-0.23
-0.57	-1.25	0.97	0.97	2.90	0.28	0.73	-0.52
0.41	-0.97	0.05	-0.66	0.99	1.05	-1.79	-0.68
-0.12	0.76	1.26	-1.05	0.60	-0.79	-1.96	0.03
1.16	0.40	-0.18	1.30	1.00	0.25	-0.38	0.25
-0.36	-0.97	-0.77	0.30	0.92	1.42	0.06	-0.65
-0.33	0.76	-1.23	0.40	-0.39	-0.99	0.37	0.46
-1.35	0.77	1.33	0.43	1.23	1.94	1.54	1.16
0.97	0.18	0.22	1.52	-0.07	-0.27	-1.04	2.39
0.85	0.25	-1.69	-0.42	1.00	0.99	0.98	1.54
-0.22	-0.17	0.02	-2.47	-0.48	0.62	1.38	0.56
0.23	0.62	-3.06	-0.78	1.48	-0.57	0.17	-0.36
-0.15	-0.94	-0.46	-0.47	1.39	0.17	2.07	0.07
-0.89	0.73	0.55	-1.22	1.13	-1.24	1.03	1.30
-0.46	0.02	-0.21	0.71	1.15	-0.04	-0.42	-0.36
-0.31	-0.43	0.93	0.31	0.86	-1.18	0.26	0.46
0.31	-1.04	-0.40	-0.00	-1.92	0.17	-0.68	-1.01
0.53	-0.66	0.06	0.64	1.30	-0.68	0.60	-0.59
1.64	0.10	-0.19	-0.94	-1.36	0.94	-1.69	-1.34
1.38	0.59	-0.37	-1.13	-0.35	0.57	1.33	0.37
-0.40	1.28	0.42	0.47	1.21	-1.45	0.70	-0.23
-1.70	-1.27	2.12	1.35	1.07	-0.10	0.79	1.14
-2.55	-1.36	-1.42	0.97	0.39	-0.67	-0.49	-0.04
0.13	-1.03	-1.05	3.01	0.40	0.29	0.44	-0.15
-0.40	0.31	-1.91	1.13	-0.10	-1.00	1.95	-0.51
-0.35	1.63	-0.10	-1.19	0.29	0.86	1.13	-0.08
-0.34	-2.69	-0.16	0.87	0.38	1.03	-0.31	-2.00
-0.38	-0.22	-0.28	1.75	1.00	1.33	0.87	-0.51
-0.17	-0.84	1.36	-0.73	0.08	1.23	0.04	0.19
-1.17	0.56	0.75	0.50	-0.90	-1.10	-0.73	1.01
-0.34	1.86	0.69	1.16	0.32	-1.57	0.24	-0.31
-0.57	-0.66	0.29	1.33	3.47	0.68	-0.95	1.21
1.64	1.19	-0.13	0.32	1.37	-1.38	-1.09	-1.58

Random Normal Deviates

0.22	0.20	-0.13	-0.16	0.44	0.21	0.74	0.45
-0.61	0.44	-0.72	-0.99	0.44	-0.49	-0.23	0.33
-1.27	-2.04	-0.99	-1.12	2.75	0.28	-0.22	0.10
-0.04	-0.06	-1.65	-0.18	0.01	-1.99	-0.49	1.37
-0.04	1.48	-0.54	-0.40	-0.71	-1.10	0.56	-0.03
0.19	0.88	-0.54	0.69	0.85	0.47	-1.19	-1.85
-1.72	1.24	-0.01	-0.53	-0.15	0.36	-0.17	0.83
0.97	-1.26	-2.39	0.60	-1.18	1.12	0.84	0.79
-0.71	-0.67	-0.52	-1.07	0.90	0.92	-0.94	-0.08
0.50	0.63	0.82	1.35	0.07	-0.02	-0.95	1.46
0.66	0.15	0.74	0.17	-1.26	0.62	-0.58	1.79
0.98	-0.26	0.06	-1.40	-0.57	0.72	0.96	0.43
-0.57	-0.24	0.69	0.82	-1.41	-0.77	-0.10	0.53
-0.32	0.33	-0.48	1.04	-0.58	0.87	-0.06	-0.20
0.99	-0.62	1.74	0.42	-1.10	-0.12	2.39	-0.45
-0.75	0.53	0.09	-0.98	-2.96	-1.12	-1.01	0.20
0.84	0.86	1.63	0.77	0.80	-0.67	0.07	0.10
0.77	0.26	0.08	0.68	0.64	-0.47	-0.81	0.75
-0.63	-0.23	0.72	1.32	-0.58	0.09	-2.38	0.91
0.00	0.49	1.25	0.71	-0.19	1.42	0.10	1.18
-1.07	-0.86	0.87	-0.66	0.12	0.27	0.59	-0.21
2.89	-0.65	0.56	0.69	-0.42	0.25	-0.39	-0.32
0.69	-0.97	0.31	-1.98	1.76	0.77	-0.56	0.71
1.00	1.19	0.85	0.80	0.59	1.00	-0.67	-1.70
0.25	-1.64	1.15	0.57	2.04	-1.53	-0.97	-1.38
0.55	-2.31	2.20	1.39	0.43	1.30	0.85	-1.24
0.75	3.20	-0.57	0.31	1.29	0.74	0.87	0.02
0.60	-0.90	1.55	2.52	0.82	-1.29	-1.28	0.39
1.16	-1.04	0.16	0.38	0.75	0.03	-0.30	-1.18
-0.62	-0.48	-0.87	-1.18	1.25	1.15	-0.20	1.57
-0.78	0.44	1.54	0.52	1.40	-0.42	-0.75	0.58
-0.63	-0.04	-1.04	-1.16	1.17	0.49	-0.41	-0.62
-0.15	0.65	0.16	-1.33	-1.49	0.26	-0.75	0.30
0.52	0.27	-0.13	-1.54	-0.63	-1.59	0.27	0.68
1.22	0.44	1.53	-1.07	0.17	-0.03	-0.12	-1.70
-0.66	0.02	2.14	-0.86	0.26	1.39	1.05	0.67

Random Normal Deviates

-0.47	-0.63	1.54	0.27	-2.11	-1.76	-0.43	-0.00
-1.14	0.46	1.26	0.82	-0.12	0.94	-0.48	0.28
0.06	0.12	0.26	0.54	0.35	0.10	-0.08	0.05
0.59	0.10	-0.73	-0.35	-0.34	0.08	-0.77	-1.87
0.69	-0.64	-0.55	1.34	1.32	-0.39	-2.27	2.07
0.46	-0.48	-0.68	-1.01	-0.14	-1.43	-0.93	1.39
-1.32	0.64	-0.29	-1.06	-1.62	-0.10	-1.37	0.08
-0.22	-1.31	0.46	0.65	2.15	-0.35	-1.11	-0.17
-1.19	0.40	-0.48	2.04	1.48	-1.58	-1.40	-0.74
0.87	2.50	0.64	-0.48	1.60	-0.46	0.85	-1.59
-0.11	-0.50	-0.68	1.79	-1.16	0.58	-0.78	1.95
-0.19	-0.40	-0.23	0.82	0.13	-1.73	0.18	1.76
1.29	0.79	1.06	-1.10	0.63	0.30	1.66	-0.21
1.15	0.01	-0.47	0.54	0.13	-1.33	1.09	-0.28
-0.21	-0.54	0.47	-0.08	0.46	-0.46	0.03	0.27
-0.16	-0.58	0.04	2.00	-0.90	1.65	-0.17	1.52
-1.48	1.90	-1.39	-1.14	-0.97	-0.14	-0.82	0.37
0.19	0.60	0.76	-1.73	0.10	0.11	-0.63	-2.00

Genetic Change

Fall 2008

1 Introduction

The main use of Estimated Breeding Values is to choose the parents of the next generation, and to optimize the matings of the selected parents. Another use is to monitor the success of selections on changing the genetic average of the population. Below are EBVs of 20 dairy goats from one herd in Ontario for lactation protein yields (Table 1). The EBVs were estimated using an animal model with repeated records from data on all herds in Ontario.

2 By Year of Birth

One trend that could be plotted would be the average EBV by year of birth. This would measure the trend in female goats that were born and then retained in the herd for producing milk. Female goats that were not retained were likely sold to other producers or culled because they were not needed. Results are shown in Table 2 by year of birth.

Trends within a herd are very erratic because the number of new female goats coming into a herd as replacements per year is small. Thus, the average of EBVs have a very large standard error. However, over a long period of time, a general trend should be observed.

Combining results across all goat herds in Ontario or in Canada can give a much better picture of the trend in the entire goat population. The standard error of those averages would be very small because they would be based on hundreds or thousands of animals.

Table 1
EBVs for Protein Yield of Dairy Goats.
'x' marks year in which goats had a lactation started.

Animal	Year of	Protein	Year of Production								
	Birth	EBV	91	92	93	94	95	96	97	98	99
1	90	+10	x								
2	90	-1	x	x							
3	91	+3	x	x							
4	91	+6	x	x	x						
5	91	-4	x								
6	91	+4		x	x	x	x				
7	92	+7		x	x	x					
8	92	-5			x						
9	92	+4			x	x	x	x	x		
10	93	+8				x	x				
11	93	+2				x	x	x			
12	93	+3					x	x	x	x	
13	94	-6					x	x			
14	94	+7						x	x	x	x
15	94	-8						x			
16	95	+5							x	x	x
17	95	0								x	x
18	96	+2								x	x
19	97	+1									x
20	97	+11									x

Table 2
Average EBVs of producing goats by year of birth.

Year of Birth	Number	Average EBV
90	2	4.50
91	4	2.25
92	3	2.00
93	3	4.33
94	3	-2.33
95	2	2.50
96	1	2.00
97	2	6.00

3 By Year of Production

The average EBV of female goats that began a lactation in each year of production would estimate the genetic average of live, active animals in each year. This would reflect the management policies of owners and somewhat the economic influences of the time. Economics may force owners to cull more stringently in one year than in others. This also accounts for the fact that some animals have longer productive lifetimes than others.

Table 3
Average EBVs by Year of Production.

Year of Kidding	Number	Average EBV
91	5	2.80
92	5	3.80
93	5	3.20
94	5	5.00
95	6	2.50
96	6	0.33
97	4	4.75
98	5	3.40
99	6	4.33

There is a trend upwards in these averages, and the averages are slighly higher than those in the previous table. This means that the better female goats are kept around to produce longer in the herd, and poor EBV goats are culled. There is about a 2 year difference between the two tables because one is based on year of birth and the other on year of kidding.

4 Trends in Males

The males in most species are more intensely selected than females. In dairy cattle, for example, about 75% of all female calves born are used as herd replacements, while only about 400 bull calves are chosen to be sires of the next generation. The average EBVs of bull calves chosen for breeding would be a useful statistic for measuring the change in the male side of the pedigree. The average EBVs of these animals should be much higher than the average of female calf replacements. The average of the males will take some years to be noticed in the population.

Another trend is the average EBV of males used to breed females in a given year. This is essentially a weighted average of male EBVs weighted by the number of matings they made in a year. Some males are more popular than others because of their EBVs for many traits, and therefore, they are chosen more frequently by producers.

5 Pathways of Selection

There are four basic pathways of selection in animal breeding, and each can have a different rate of genetic change.

1. Sires of males pathway (SM). This is the most stringent selection category. They represent the best 5% of all male animals that are chosen for breeding. They represent less than 0.1% of all male animals that are born, usually.
2. Sires of females pathway (SF). Males chosen for breeding to the general population of females.
3. Dams of males pathway (DM). Females chosen from which to obtain males for breeding.
4. Dams of females pathway (DF). Females chosen for breeding purposes to produce future female replacements.

Trends based on year of birth are probably more useful than trends based on year of production, although both might be of interest. Trends should also be calculated for each pathway of selection. These can be combined into one overall population trend if desired. The sire pathways are generally more accurately estimated because males have many more progeny than females. On the other hand, there are more females per year of birth than males in these pathways, and so the stability of the female trends is better.

Remember that the trends are a reflection of past selection and breeding decisions, and give an indication of how quickly the breeding goals are being achieved.

6 Biased Trends

6.1 Incorrect Heritability

The above trends assume that the correct heritabilities and other parameters of the model have been used to estimate the EBVs. If the heritability used in the MME is too high, then the range of EBVs becomes greater than it should be, which causes the estimates of average EBVs to be biased upwards. There is the appearance that there is more genetic change than actually exists.

On the other hand, if the heritability used in the MME is too low, then trends in average EBVs could be biased downwards. The solution is to use the best possible estimates of heritability. An experiment to test unbiasedness is to split the data into two sets. The first set has data up to time t, and the second set has data from time $t+1$ to the present. Using a value for heritability, estimate the EBV for all animals using the first data set only. Then combine the two data sets and re-estimate the breeding values. The regression of the predicted EBVs from the first data set on the EBVs from the combined data set should be 1 if the correct heritability has been used. If the regression is greater than 1, then the heritability was too high, or vice versa.

6.2 Wrong Model

Suppose a trait is significantly affected by the age of the animal, but the age effect was omitted from the animal model. Estimation of the EBVs could be biased by the age effects. That is, the age effects might end up in the EBVs if there is nothing to remove them in the model. Older animals might appear to be better genetically than young animals. Estimated genetic trends might be negative or close to zero. This is another reason to continuously update the model for genetic evaluation, to make sure that all necessary factors are in the model. The models should take into account phenotypic time trends.

7 Predicting Genetic Change

A breeding strategy describes the process (how and when) by which males and females are selected for the next generation of matings. The prediction of a future

progeny of sire X and dam Z is simply the average of the EBVs of the sire and dam;

$$EBV_{\text{progeny}} = 0.5 * (EBV_{\text{sire}} + EBV_{\text{dam}}).$$

The accuracy of that prediction depends on the accuracy of the EBVs of the sire and dam. If all of the matings were known in advance, then the EBV of each future progeny could be calculated and the average computed to give a prediction of genetic change. However, the mates for every mating are not known in advance for the next year or even the next 5 or 10 years, in most situations. Fortunately, there is a general equation that has been used by animal breeders for many years to predict future genetic change.

7.1 Formula

Assuming that the initial population is normally distributed, then the formula to predict genetic change is

$$\frac{\Delta \text{G}}{\text{year}} = \frac{r_{TI}\ i\ \sigma_a}{L},$$

where

ΔG is genetic change in a trait,

r_{TI} is the accuracy of selection (or reliability of the EBV),

i is the selection intensity,

σ_a is the additive genetic standard deviation of the trait, and

L is the generation interval in years.

7.1.1 Accuracy of Selection

The reliability of the EBVs is critical to genetic change. If EBVs are not very accurate then errors will be made in selecting animals for matings. This will limit or decrease the amount of genetic change that could be expected. Reliability of EBVs depends upon

- Heritability of the trait,
- The statistical linear model, and
- The quality and quantity of data.

7.1.2 Selection Intensity

Selection intensity, i, or selection differential is the difference in the mean of animals that have been selected versus the mean of all animals standardized to a variance of one. Table 13.1 (at the end of this chapter) contains i values by percentage of animals selected. The assumption is that truncation selection is applied to a normally distributed trait. For example, suppose the genetic standard deviation, σ_a, is 1000 kg of milk and the mean BV of all current animals is $\mu = +500$ kg, then if the top 8.3% of the animals have been selected the superiority of the mean of the selected animals, μ_s compared to the entire population would be

$$\begin{aligned} \mu_s &= \mu + (i)\sigma_a, \\ &= +500\text{kg} + (1.841)\ 1000\text{kg}, \\ &= 2341\text{kg} \end{aligned}$$

If the top 41% were chosen then the mean would be

$$\begin{aligned} \mu_s &= +500\text{kg} + (.948)\ 1000\text{kg}, \\ &= 1448\text{kg}. \end{aligned}$$

7.1.3 Genetic Standard Deviation

There is little that can be done, in the short term at least, to increase the genetic standard deviation of a trait. The genetic variance must be estimated, but this is usually not a problem.

7.1.4 Generation Intervals

Generation interval, L, is the average age of males or females when a future male or female replacement is born. The shortest generation interval (biologically) is the age of maturation plus the gestation length. This natural barrier can possibly be decreased through reproductive technology. For example, embryos can be removed from females before the female is mature (even while it is a fetus).

Often generation intervals are much longer than the minimum possible because producers want to have reliable EBVs before making breeding decisions. Thus, there is usually a balance between reliability of an EBV and the generation interval. Higher reliability means waiting for data on lots of progeny, or waiting until an animal has made a certain number of records itself. Decreasing the generation interval means choosing animals whose EBVs are usually much less reliable.

Obviously, the age at maturation and the gestation length of a species must be known in order to determine generation intervals. Some species have very long generation intervals (such as horses) while other species can have very short generation intervals (such as poultry).

7.2 Expansion of General Formula

Work by Dickerson and Hazel (1944) and Rendel and Robertson (1950) led to expansions of the genetic gain formula according to the four pathways of selection. Those pathways were

- Sires of males, SM,
- Sires of females, SF,
- Dams of males, DM, and
- Dams of females, DF.

The expanded formula is

$$\frac{\Delta \mathrm{G}}{\text{year}} = \frac{\Delta_{SM} + \Delta_{SF} + \Delta_{DM} + \Delta_{DF}}{L_{SM} + L_{SF} + L_{DM} + L_{DF}},$$

where each $\Delta_{ij} = r_{TI-ij} i_{ij} \sigma_a$. Thus, each pathway has a different reliability of EBVs (because sires are often based on many more progeny than dams), each pathway has a different selection intensity (because fewer sires are needed than dams), and each pathway has a different generation interval possible based on when animals can be replaced.

7.3 Example Predictions

Table 13.2 contains information related to dairy cattle selection programs for milk production where the genetic standard deviation is 1000 kg. Bulls are usually 9 to 11 years of age before a replacement is born. However, bulls will be 6 years of age when daughters are born that will be replacements for other females. Dams of males are also highly selected and have usually completed 3 lactations, so that they are at least 5 years of age when a replacement son is born. Dams of other females, however, only need one lactation record or less.

Reliabilities of EBVs differ. Sires of females need only a minimally reliable EBV which is .70 or higher, while sires of males must be much higher. Dams of males, because they have completed 3 lactations and might have one daughter, have a reliability as high as .50. Dams of females have a reliability close to the value of heritability .30.

Selection intensities also differ. Sires of males are the top 10 out of 400 bulls (2.5%) while sires of females are the top 40 out of 400 bulls tested per year (10%). Dams of males are the top 400 out of 100,000 (0.4%), and only 25% of females are culled per year leaving 75% to produce future females.

Table 13.2
Figures for Predicting Genetic Change in Dairy Cattle.

	Pathways of Selection			
	SM	SF	DM	DF
r_{TI}	.95	.70	.50	.35
i	2.336	1.755	2.975	.424
L years	9	6	5	3
Δ_G	2219.2	1228.5	1487.5	148.4

The total predicted response per year would be

$$\frac{\Delta_G}{\text{yr}} = \frac{2219.2 + 1228.5 + 1487.5 + 148.4}{9+6+5+3}$$
$$= \frac{5083.6}{23} = 221.0\text{kg/yr}$$

Note that the SM and DM pathways were the two largest contributors (about 73%) of the total predicted change per year. This indicates that male selection is very critical to genetic change in dairy cattle. Male selection is controlled by the artificial insemination industry.

7.3.1 Increasing Reliability

Molecular genetic technology may someday change the reliability of all EBVs to be 85% or better for all animals. Keeping the SS pathway at 95%, but improving all others to 85% changes the contributions of each pathway to

SM 2219.2 kg,

SF	1491.7 kg,
DM	2528.7 kg,
DF	360.4 kg,

gives $\Delta_G = 287$ kg/yr. This is 30% greater progress than the traditional selection program.

7.3.2 Decreasing Generation Intervals

The above calculations assumed that generation intervals would not be changed, but molecular genetics may be able to give us an 85% reliable EBV as soon as an animal is born. Dairy cattle are sexually mature at one year of age for bulls and 16 months for females. Suppose all generation intervals could be reduced to 2 years and accuracy of the SM pathway is now the same as the others at 85%, then the contributions of each pathway are

SM	1985.6 kg,
SF	1491.7 kg,
DM	2528.7 kg,
DF	360.4 kg,

giving $\Delta_G = 796$ kg/yr. This is a 260% increase in genetic change! Also notice that the DM pathway is relatively more important than the other three. Thus, selection of dams of males could become much more important in the future. Perhaps fewer sires will be needed and fewer dams of sires so that the selection intensities on those pathways can become more strict.

Table 13.1
Selection Differentials, i

	For .001 to .099 selected									
	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.00		3.400	3.200	3.033	2.975	2.900	2.850	2.800	2.738	2.706
.01	2.660	2.636	2.600	2.569	2.550	2.527	2.500	2.582	2.456	2.442
.02	2.420	2.400	2.386	2.370	2.363	2.336	2.323	2.311	2.293	2.283
.03	2.270	2.258	2.241	2.230	2.221	2.209	2.200	2.186	2.174	2.164
.04	2.153	2.146	2.136	2.126	2.116	2.107	2.098	2.087	2.079	2.071
.05	2.064	2.057	2.048	2.040	2.031	2.022	2.016	2.009	2.000	1.990
.06	1.985	1.977	1.971	1.965	1.958	1.951	1.944	1.937	1.931	1.925
.07	1.919	1.911	1.906	1.900	1.893	1.888	1.882	1.875	1.871	1.863
.08	1.858	1.852	1.846	1.841	1.837	1.834	1.826	1.820	1.815	1.810
.09	1.806	1.799	1.793	1.788	1.784	1.780	1.775	1.770	1.765	1.760
	For .01 to .99 selected									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.10	1.755	1.709	1.667	1.628	1.590	1.554	1.521	1.488	1.458	1.428
.20	1.400	1.372	1.346	1.320	1.295	1.271	1.248	1.225	1.202	1.180
.30	1.159	1.138	1.118	1.097	1.078	1.058	1.039	1.021	1.002	.984
.40	.966	.948	.931	.913	.896	.880	.863	.846	.830	.814
.50	.798	.782	.766	.751	.735	.720	.704	.689	.674	.659
.60	.644	.629	.614	.599	.585	.570	.555	.540	.526	.511
.70	.497	.482	.468	.453	.438	.424	.409	.394	.380	.365
.80	.350	.335	.320	.305	.290	.274	.259	.243	.227	.211
.90	.195	.179	.162	.144	.127	.109	.090	.070	.049	.027

Phantom Parent Groups
Fall 2008

In any pedigree file there are always animals that have unknown parents. The assumption is that these animals originated from a large, random mating population, such that they are unrelated to each other and non-inbred. In practice, parentage of animals is sometimes not recorded or is lost. This occurs most often when animals change ownership. In dairy cattle, for example, not all herds are enrolled on a milk recording program. Cows that move from a herd not on milk recording to a herd that is milk recorded often do not have known parents. This also occurs when animals change ownership over country borders (between Canada and the US, for example). However, with disease problems and their effects on humans becoming a high priority concern, the traceability of animals within and across countries is receiving increased attention. Eventually all livestock may need to have passports.

If an animal has unknown parents, is it safe to assume that this animal originates from the large, random mating base population? Suppose there are two animals with unknown parents, and one animal is known to have been born in 1970, and the other is known to have been born in 1980. If genetic trend is significant, then the genetic average of parents of animals born in 1970 will be different from the genetic average of parents of those born in 1980. The assumption that parents of both animals were from one population would not be valid. A way to handle this problem is to create **Phantom Parent Groups**.

1 Formation of Phantom Groups

Phantom Parent groups should be assigned according to the four pathways of selection, and by year of birth of the animal with the unknown parents. For example, suppose a female animal was born in 1970 with unknown parents. The male parent of this animal would be assigned to the Sire of Dam group-1970, and the female parent would be assigned to the Dam of Dam group-1970. Any female animal born in 1970 with unknown parents would have the male and female parents assigned to the same groups, SD-1970 for the male parent and DD-1970 for the female parent.

A male animal with unknown parents born in 1981 would have the male parent assigned to the Sire of Sire group-1981, and the female parent assigned to the Dam of Sire group-1981. Any male animal born in 1981 with unknown parents would have the male and female parents assigned to SS-1981 and DS-1981, respectively.

Depending upon the species, groupings might also include foreign versus domestic country of birth, or breed. The idea is to identify potentially different populations from which the parents might have been sampled. In genetic evaluation, there would be equations for each phantom parent group, and estimates of genetic differences between groups could be

estimated. There needs to be a sufficient number of animals within each phantom parent group in order to obtain an accurate estimate. Thus, if some phantom parent groups seem to be too small, then some groups could be combined

2 Example Problem

Table 1 contains information on dairy heifers (young female cows) for age at first service in days (age at which they are first inseminated artificially). The heritability of this trait is 0.12. Note the missing parent information on a few animals. Animals 15 and 18 have a known sire, but unknown dams. The first step is to create a pedigree file with all animals (1 through 20), and then to assign phantom parent group numbers to unknown parents. Six phantom parent groups were used for this example. Parents of animals 1, 6, and 8 were sires and therefore, were assigned to different groups than parents of animals 2, 3, 4, 5, and 7 which were females. Parents of animals 9, 10, 15, 18, and 19, were assigned to different groups. There were not enough animals in each year to make separate groups for parents of animals 18 and 19. The phantom groups are indicated by a 'P' in front of a number in Table 2.

Table 1
Age at First Service on Dairy Heifers.

Animal	Sire	Dam	Birth Year	Age at First Service(days)
9			2001	475
10			2001	498
11	1	2	2001	482
12	1	3	2001	500
13	6	4	2001	477
14	6	5	2001	503
15	6		2001	492
16	1	7	2002	513
17	6	2	2002	516
18			2002	487
19	8		2002	505
20	8	4	2002	494

Table 2
Pedigree File with Phantom Parent Groups Included.

Animal	Sire	Dam	b_i	Animal	Sire	Dam	b_i
1	P1	P2	1.00	11	1	2	0.50
2	P3	P4	1.00	12	1	3	0.50
3	P3	P4	1.00	13	6	4	0.50
4	P3	P4	1.00	14	6	5	0.50
5	P3	P4	1.00	15	6	P6	0.75
6	P1	P2	1.00	16	1	7	0.50
7	P3	P4	1.00	17	6	2	0.50
8	P1	P2	1.00	18	P5	P6	1.00
9	P5	P6	1.00	19	8	P6	0.75
10	P5	P6	1.00	20	8	4	0.50

2.1 A-Inverse

The formation of the inverse of $\mathbf{A}$ is similar to what has been given so far. One must first find the b_i values for all 'real' animals, which involves calculating the inbreeding coefficients of all animals. These are given in the above table for all animals. To include phantom parent groups in the animal model, think of the phantom groups as animals (with unknown parents). The b_i values for phantom groups will always be 1. The b_i value for animal 1 is also 1. The order of $\mathbf{A}^{-1}$ is therefore, $26 = 20$ animals plus 6 phantom groups.

Every animal in the pedigree file has both 'parents' listed. The rules are to add the

following quantities to the inverse matrix for each animal. Let i equal the animal, s equal the sire, d equal the dam, and $\delta_i = b_i^{-1}$, then the contributions to the inverse are

	i	s	d
i	δ_i	-0.5 δ_i	-0.5 δ_i
s	-0.5 δ_i	0.25 δ_i	0.25 δ_i
d	-0.5 δ_i	0.25 δ_i	0.25 δ_i

For the phantom groups, just add 1 to their diagonal elements. The inverse is given in the SAS IML program that follows, in a partitioned manner.

2.2 Results

When phantom parent groups are used, the formula for calculating the reliabilities of the EBVs no longer applies. A different procedure is needed for obtaining reliabilities, which is too complex for this course. The reliabilities could be approximated by re-running this example without phantom groups, and using the usual formula. The following table has the solutions for animals with records for the model with and the model without phantom parent groups, and the reliabilities for the later.

Table 10.3
EBVs for Age at First Service for Heifers
With and Without Phantom Parent Groups.

Animal	With		Without		
ID	EBV	SEP	EBV	SEP	Rel
9	-2.58	4.62	-1.71	3.90	0.10
10	0.18	4.62	1.05	3.90	0.10
11	0.40	4.28	0.01	3.89	0.11
12	1.73	4.30	1.33	3.90	0.10
13	-0.54	4.29	-0.94	3.90	0.10
14	2.12	4.30	1.71	3.91	0.10
15	0.53	4.53	0.69	3.91	0.10
16	1.66	4.30	1.25	3.90	0.10
17	1.81	4.29	1.41	3.90	0.11
18	-2.79	4.63	-1.94	3.92	0.10
19	-0.13	4.54	-0.01	3.93	0.09
20	-0.95	4.32	-1.38	3.92	0.10

The estimates of birth-year means were 489.31 for year 2001, and 503.08 for year 2002. With this trait, low values are seen to be better, or perhaps an average age of 500 days

might be best. The estimate of the residual variance was 121.99 for the model with phantom parent groups and was 124.55 for the model without. Thus, having phantom parent groups in the model reduced the residual variance. The SEP are larger in the model with phantom parent groups, because the EBV is actually an estimate of the phantom group effect plus the EBV. The error in estimating the phantom group effects is included in the SEP. The Rel should be relatively the same in both models. Note that the Rel values are all less than the heritability of the trait. This is because there was a need to estimate the Birth-Year means. If the Birth-Year means were known without error, then the Rel values would have been the same as heritability.

Maternal Genetic Effects

Fall 2008

1 Introduction

In all mammalian species, the female provides an environment for its offspring to survive and grow. Females vary in their ability to provide a good environment for their offspring, and this variability has a genetic basis. The offspring inherit directly an ability to grow (or survive) from both parents, and environmentally do better or poorer depending on their dam's maternal ability. Maternal ability is a genetic trait and is transmitted, as usual, from both parents, but maternal ability is only expressed by females when they have progeny (i.e. much like milk yield in dairy cows).

Direct and maternal genetic effects are genetically correlated. Estimates of this correlation differ widely in the literature. If the estimates are obtained from institutional herds (at universities or research stations), the estimated correlations are generally small and positive. If data are field recorded animals, then estimates tend to be zero to strongly negative. This is due to problems in data recording. In field data, the ability to follow a female calf from birth to first calving is often not possible. The identity of the calf is lost between birth and first calving for various reasons, usually related to management practices. These practices are not wrong or sloppy, but rather they are expedient to save time and effort. The cost is the loss of being able to tie a calf's birth weight and weaning weight with the same animal as a dam of the next generation. In a research setting, this connection is maintained for many generations, which leads to estimates of a direct-maternal correlation that are positive. Thus, the real situation is likely that the direct-maternal correlation is positive, but small. Good estimates are not likely from field data due to poor data structure.

2 Example Problem

Below are birthweights of beef calves in two contemporary groups. Dams provide good and poor pregnancy environments for calves too, based on how much they eat, what they eat, and exercise. Thus, maternal effects are important on birthweights as well as weaning weights.

Table 11.1
Example Birthweights of Beef Calves.

Animal	Sire	Dam	CG	Weight(lb)
8	1	5	1	76
9	2	6	1	44
10	1	7	1	55
11	3	8	2	73
12	3	10	2	59
13	4	7	2	52

The model is

$$y_{ijkl} = C_i + a_j + m_k + p_k + e_{ijkl},$$

where y_{ijkl} is a birthweight record on calf j from dam k, in contemporary group i; C_i is a contemporary group effect; a_j is the animal additive genetic effect (direct genetic); m_k is the dam's maternal genetic effect on the calf birthweight; p_k is the dam's permanent environmental effect on calf birthweight; and e_{ijkl} is the residual effect. Dams could have more than one calf over a period of years, (repeated records), and so would have a permanent environmental effect common to each calf birthweight.

In matrix notation this model would be written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{p} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of birthweights; $\mathbf{b}$ is a vector of contemporary group effects; $\mathbf{a}$ is the vector of additive genetic effects of the animals; $\mathbf{m}$ is the vector of maternal genetic (dam) effects, and $\mathbf{p}$ is a vector of maternal permanent environmental effects. Calves are assumed to be of the same sex and breed, and dams are assumed to be the same age within contemporary groups. Animals 8 and 10 appear as calves with a birthweight and as dams of calves, so that there is a connection between direct and maternal effects.

The expectations of the random vectors, $\mathbf{a}$, $\mathbf{m}$, $\mathbf{p}$, and $\mathbf{e}$ are all null vectors in a model without selection, and the variance-covariance structure is

$$Var\begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & \mathbf{0} & \mathbf{0} \\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix},$$

where σ_a^2 is the additive genetic variance, σ_m^2 is the maternal genetic variance, σ_{am} is the additive genetic by maternal genetic covariance, and σ_p^2 is the maternal permanent environmental variance. Let

$$\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} 49 & -7 \\ -7 & 26 \end{pmatrix}.$$

Let $\sigma_p^2 = 9$ and $\sigma_e^2 = 81$.

3 MME

Based on the matrix formulation of the model, the MME are represented as follows:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 & \mathbf{X}'\mathbf{Z}_3 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{-1}k_{11} & \mathbf{Z}_1'\mathbf{Z}_2 + \mathbf{A}^{-1}k_{12} & \mathbf{Z}_1'\mathbf{Z}_3 \\ \mathbf{Z}_2'\mathbf{X} & \mathbf{Z}_2'\mathbf{Z}_1 + \mathbf{A}^{-1}k_{12} & \mathbf{Z}_2'\mathbf{Z}_2 + \mathbf{A}^{-1}k_{22} & \mathbf{Z}_2'\mathbf{Z}_3 \\ \mathbf{Z}_3'\mathbf{X} & \mathbf{Z}_3'\mathbf{Z}_1 & \mathbf{Z}_3'\mathbf{Z}_2 & \mathbf{Z}_3'\mathbf{Z}_3 + \mathbf{I}k_{33} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \\ \mathbf{Z}_3'\mathbf{y} \end{pmatrix},$$

where

$$\begin{aligned} \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix} &= \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix}^{-1} \sigma_e^2, \\ &= \begin{pmatrix} 49 & -7 \\ -7 & 26 \end{pmatrix}^{-1} (81), \\ &= \begin{pmatrix} 1.7192 & .4628 \\ .4628 & 3.2400 \end{pmatrix}. \end{aligned}$$

Note that these numbers are not equal to

$$\begin{pmatrix} 81/49 & 81/(-7) \\ 81/(-7) & 81/26 \end{pmatrix}.$$

Finally, $k_{33} = \sigma_e^2/\sigma_p^2 = 81/9 = 9$.

The heritability of direct genetic effects is

$$h_d^2 = \frac{\sigma_a^2}{\sigma_y^2},$$

where

$$\sigma_y^2 = \sigma_a^2 + \sigma_m^2 + \sigma_p^2 + \sigma_e^2 + 0.5\sigma_{am} = 161.5,$$

so that

$$h_d^2 = 0.30.$$

The maternal heritability is

$$h_m^2 = \frac{\sigma_m^2}{\sigma_y^2} = 0.16.$$

The maternal repeatability is

$$r_m = \frac{\sigma_m^2 + \sigma_p^2}{\sigma_y^2} = 0.22.$$

The correlation between direct and maternal genetic effects is

$$\rho = \frac{\sigma_{am}}{\sigma_a \sigma_m} = -0.196.$$

4 Comments

Maternal effects can be expected in all mammals during the birth to weaning period. After weaning, maternal effects may still linger, but eventually disappear.

4.1 Poultry

Do maternal effects exist in other species? Take poultry as an example. There could be a maternal effect on chick development or hatchability due to the composition or amount of nutrients in the egg, or even due to the thickness of the shell of the egg. Generally, chicks are independent of the parents at birth, and in most commercial enterprises have little or no contact with the hen. Thus, the ability of a hen to find food or to protect her brood goes unexpressed. Most genetic analyses of poultry data ignore the possibility of maternal genetic effects.

4.2 Salmon

Suppose you are working with Pacific salmon, fish that must swim upstream in a river, lay eggs, then die before the eggs ever hatch. Maternal effects could be present due to the site which the female chooses to lay the eggs. Maybe it has a good flow of water all year round, or maybe the rocks in that area provide a good bed for the eggs, or maybe the spot is protected from predators that would eat the eggs. For some traits a maternal effect may be present. Because the female will die after spawning, there is no chance for her to spawn again and to have a repeated record. Thus, maternal permanent environmental effects can not be separated from maternal genetic effects and are usually ignored.

4.3 Data Structure

Estimation of the maternal genetic variance can be problematic. This is usually due to poor data structure. In dairy cattle, for example, a calf is not registered until the owner has decided whether to keep the animal as a replacement. Thus, the calf identification is 'unknown' from birth to registration. If that calf has a birthweight or calving ease record at birth, then commonly there is no connection between that record and any records on progeny of that calf when it becomes mature. There is an identification 'break' between the animal as a calf, and later when that animal is a mother. If there are enough breaks in the data, then the estimate of the maternal genetic variance can be biased downwards, and more importantly the estimate of the direct-maternal genetic correlation is biased downwards (such that it might become negative).

However, data structure is improving because of active animal identification programs that have been put into place in order to trace the movement of animals for health and consumer safety purposes. Data structure should be checked before estimating maternal genetic parameters. In some places, the estimate of the direct-maternal genetic correlation could not be trusted so that estimates of zero have been used, rather than attempting to estimate it.

4.4 Embryo Transfer

Embryo transfer has been popular in beef and dairy cattle. A fertilized embryo from a donor cow is implanted into the uterus of a recipient cow. The purpose is to produce more progeny from the donor cow, who is supposedly a superior animal. The recipient is regarded as an incubator for the embryo. The calf is 'born' from the recipient cow and receives the maternal environment provided by the recipient cow, but genetically the calf receives its maternal genetics from the donor cow and sire.

Often the identity of calves being born from recipient cows is not recorded. The calf is known to be produced by ET, but information about the recipient cow (age, breed) is usually unknown within the recording program because there is no attempt to retrieve that information. Thus, an ET calf in the recording program has its genetic sire and genetic dam identifications reported, and nothing about the recipient. If that information were known, the model for genetic evaluation could account for ET calves and for which cow provided the maternal environment for the calf.

Multiple Trait Models

Fall 2008

1 Introduction

Animals are commonly observed for more than one trait because many traits affect overall profitability of an animal. There are a few general categories of traits that apply to nearly all species. These are Production, Reproduction, Health, Behaviour, and Conformation. In dairy cattle, for example, production traits include milk, fat, and protein yields, and somatic cell scores, while in beef cattle, production includes growth and carcass composition. Reproduction is the ability to reproduce viable offspring without problems or delays in re-breeding, pregnancy, or parturition. Failure to become pregnant, difficulty with giving birth, or small litter size (in swine) are traits that cost producers money. Health traits relate to the ability of the animal to produce under stressful conditions. General immunity to fight off disease causing organisms is a useful trait for selection, but these traits often have low heritability. Behavioural traits, such as temperament, agressiveness towards progeny, desire to eat, and general ease of handling are traits that are not studied very much in livestock, but which contribute towards overall profitability. Conformation traits are important in some traits, such as horses or dairy cattle. Animals must have the correct body shapes to be able to jump hurdles, run fast, give more milk with fewer problems, and to win show competitions.

Multiple trait (MT) analyses make use of genetic and environmental correlations among traits in order to achieve greater reliabilities on EBVs. MT analyses are advantageous in the following situations.

- **Low Heritability Traits** When the difference between genetic and residual correlations is large (e.g. greater than .5 difference) or when one trait has a much higher heritability than the other trait, then the trait with the lower heritability tends to gain more in accuracy than the high heritability trait, although both traits benefit to some degree from the simultaneous analysis.
- **Culling** Traits that occur at different times in the life of the animal, such that animals may be culled on the basis of earlier traits and not be observed for traits that occur later in life can cause bias in EBVs of the later life traits. An MT analysis that includes all observations on an animal upon which culling decisions have been based, has been shown to partially account for the selection that has taken place, and therefore gives unbiased estimates of breeding values for all traits. Severe selection will tend to cause bias in most situations.

There are a couple of disadvantages to MT analyses.

- **Estimates of Correlations** An MT analysis relies on accurate genetic and residual correlations. If the parameter estimates are greatly different from the unknown true values, then an MT analysis could do as much harm as it might do good.
- **Computing Cost** MT analyses require more computing time and increased computer memory in order to analyze the data. Software programs are more complicated, more memory and disk storage are usually needed, and verification of results might be more complicated.

If culling bias is the main concern, then an MT model must be used regardless of the costs or no analysis should be done at all, except for the traits not affected by culling bias. More and more MT analyses are being conducted in animal breeding.

2 Models

MT situations may be simple or very complicated. A simple situation will be described. Consider two traits with a single observation per trait per animal. Table 1 contains data on body condition scores (1 to 10) and percentage fat in the tail of fat-tailed sheep at 120 days of age in Tunisia. Body condition is the degree of fatness in the body frame. A score of 1 is a very thin animal with bones sticking out and general unhealthy appearance. A score of 10 is a very fat animal, but perhaps prone to foot problems or back problems. A score of 5 is average and generally well-conditioned and healthy looking.

Table 1
Body Condition Scores and Percentage Fat in Fat-Tailed Sheep.

Animal	Sire	Dam	Group	Trait 1	Trait 2
1	0	0	1	2.0	39
2	0	0	2	2.5	38
3	0	0	3	9.5	53
4	0	0	1	4.5	45
5	0	0	2	5.5	63
6	1	3	3	2.5	64
7	1	4	2	8.5	35
8	1	5	3	8.0	41
9	2	3	1	9.0	27
10	2	4	1	7.5	32
11	2	5	2	3.0	46
12	6	10	3	7.0	67

A model should be specified separately for each trait. Usually, the same model is assumed for each trait, and this can greatly simplify the computational aspects, but such an assumption may be unrealistic in many situations. The same model will be assumed for both traits.

Let the model equation for trait t be

$$y_{tij} = G_{ti} + a_{tj} + e_{tij},$$

where G_{ti} is a group effect with 3 levels, a_{tj} is a random, animal additive genetic effect for trait t, and e_{tij} is a random residual environmental effect for trait t.

Because the two traits will be analyzed simultaneously, the variances and covariances need to be specified for the traits together. For example, the additive genetic variance-covariance (VCV) matrix could be written as

$$\begin{aligned} \mathbf{G} &= \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 15 \end{pmatrix}, \\ \mathbf{G}^{-1} &= \begin{pmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{pmatrix} = \frac{1}{11} \begin{pmatrix} 15 & -2 \\ -2 & 1 \end{pmatrix}, \end{aligned}$$

and the residual environmental VCV matrix as

$$\begin{aligned} \mathbf{R} &= \begin{pmatrix} e_{11} & e_{12} \\ e_{12} & e_{22} \end{pmatrix} = \begin{pmatrix} 10 & 5 \\ 5 & 100 \end{pmatrix}, \\ \mathbf{R}^{-1} &= \begin{pmatrix} e^{11} & e^{12} \\ e^{21} & e^{22} \end{pmatrix} = \frac{1}{975} \begin{pmatrix} 100 & -5 \\ -5 & 10 \end{pmatrix}. \end{aligned}$$

The genetic and residual correlations are, respectively,

$$\begin{aligned} \rho_g &= 2/(15)^{.5} = .516, \\ \rho_r &= 5/(1000)^{.5} = .158 \end{aligned}$$

with

$$h_1^2 = \frac{1}{11} = .0909,$$

and

$$h_2^2 = \frac{15}{115} = .1304.$$

For all data, then

$$Var \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}g_{11} & \mathbf{A}g_{12} \\ \mathbf{A}g_{12} & \mathbf{A}g_{22} \end{pmatrix}.$$

The structure of the residual VCV matrix over all observations can be written several ways depending on whether allowance is made for missing observations on either trait for some animals. If all animals were observed for both traits, then

$$Var \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}e_{11} & \mathbf{I}e_{12} \\ \mathbf{I}e_{12} & \mathbf{I}e_{22} \end{pmatrix}.$$

3 MME

Let the model for one trait, in matrix notation be

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

then the MME for one trait could be written as

$$\left[\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{pmatrix} \right] \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix},$$

or more simply as

$$(\mathbf{B} + \mathbf{H}^{-1})\hat{\mathbf{s}} = \mathbf{r}.$$

MT MME for two traits with animals observed for both traits, would be

$$\left[\begin{pmatrix} \mathbf{B}e^{11} & \mathbf{B}e^{12} \\ \mathbf{B}e^{21} & \mathbf{B}e^{22} \end{pmatrix} + \begin{pmatrix} \mathbf{H}^{-1}g^{11} & \mathbf{H}^{-1}g^{12} \\ \mathbf{H}^{-1}g^{21} & \mathbf{H}^{-1}g^{22} \end{pmatrix} \right] \begin{pmatrix} \mathbf{r}_1 e^{11} + \mathbf{r}_2 e^{12} \\ \mathbf{r}_1 e^{21} + \mathbf{r}_2 e^{22} \end{pmatrix},$$

Often observations for all traits are available on each animal. With two traits one of the two trait observations might be missing. This complicates the construction of MME, but they are still theoretically well-defined. Thus, an EBV could be calculated for an animal that has not been observed for a trait, through the genetic and residual correlations to other traits and through the relationship matrix. The models for each trait could also be different and this provides another layer of complexity to MT analyses.

4 Results

Both single trait and multiple trait analyses were conducted for this example with the results shown in Table 2.

Table 2
EBVs and Prediction Error Variances(VPE)
from Multiple and Single Trait Analyses.

	Multiple Trait EBVs				Single Trait EBVs			
Animal	Trait 1		Trait 2		Trait 1		Trait 2	
	EBV	VPE	EBV	VPE	EBV	VPE	EBV	VPE
1	-0.31	0.89	-0.72	12.88	-0.30	0.90	-0.41	12.99
2	-0.20	0.90	-1.44	13.11	-0.08	0.91	-1.41	13.20
3	0.18	0.92	0.14	13.55	0.21	0.94	-0.10	13.62
4	0.17	0.90	0.70	13.18	0.13	0.92	0.58	13.28
5	0.16	0.91	1.32	13.27	0.03	0.92	1.34	13.36
6	-0.15	0.94	0.36	13.94	-0.24	0.95	0.67	14.00
7	0.02	0.90	-0.51	13.16	0.09	0.92	-0.67	13.24
8	-0.12	0.91	-0.65	13.25	-0.07	0.92	-0.62	13.34
9	0.07	0.90	-1.02	13.16	0.22	0.92	-1.34	13.25
10	0.08	0.91	-0.17	13.31	0.12	0.92	-0.32	13.40
11	-0.10	0.94	-0.14	13.78	-0.11	0.95	-0.01	13.84
12	0.05	0.94	0.83	13.90	-0.05	0.95	0.93	13.96

Animal EBVs for both traits are highly correlated to each other between single and multiple trait analyses. There would be very little, if any, re-ranking of animals. Variances of prediction error from single trait analyses were slightly larger than those from the multiple trait analyses. In this example, there would be little advantage to multiple trait analyses. However, if observations were missing or if the difference between residual and genetic correlations was greater, then a multiple trait analysis would be more beneficial.

Non-Additive Genetic Effects

Fall 2008

1 Introduction

Genetic evaluation of animals is for the estimation of the additive genetic effects, i.e. those that are transmitted directly from parents to offspring. Additive genetic variation is usually greater than any non-additive variation, and additive genetic effects are easy to estimate using the animal model. Populations with high numbers of full-sibs (such as poultry, swine, or fish) could have significant non-additive genetic effects because full-sibs have a dominance genetic relationship of 0.25. Traits with low heritability could have substantial non-additive genetic variation too. Ignoring non-additive effects in the animal model could make the estimation of additive genetic effects less accurate.

Non-additive genetic effects are the interactions among alleles both within and across gene loci. A review of quantitative genetics is needed to explain these interactions.

2 Single Locus

Assume a single locus with 3 alleles, A_1, A_2, and A_3 with frequencies .4, .5, and .1, respectively. The possible genotypes, frequencies, and genotypic values are given in Table 1. Let the model for the genotypic values be given as

$$G_{ij} = \mu + a_i + a_j + d_{ij},$$

where

$$\begin{aligned}
\mu &= G_{..} = \sum_{i,j} f_{ij} G_{ij}, \\
a_i &= G_{i.} - G_{..}, \\
G_{i.} &= Pr(A_1)G_{11} + Pr(A_2)G_{12}, \\
a_j &= G_{.j} - G_{..} \\
G_{.j} &= Pr(A_1)G_{12} + Pr(A_2)G_{22}, \\
d_{ij} &= G_{ij} - a_i - a_j - \mu
\end{aligned}$$

Table 1. Example locus genotypes, frequencies, and values.

Genotype	Frequency, f_{ij}	Genetic Value, G_{ij}
A_1A_1	.16	5
A_1A_2	.40	3
A_1A_3	.08	1
A_2A_2	.25	4
A_2A_3	.10	2
A_3A_3	.01	0

Then

$$\begin{aligned}
\mu &= .16(5) + .40(3) + .08(1) + .25(4) + .10(2) + .01(0) \\
&= 3.28, \\
\sigma_G^2 &= .16(25) + .40(9) + .08(1) + .25(16) + .10(4) + .01(0) - \mu^2 \\
&= 12.08 - 10.7584 \\
&= 1.3216. \\
G_{1.} &= .4(5) + .5(3) + .1(1) = 3.6, \\
a_1 &= 0.32, \\
G_{2.} &= .4(3) + .5(4) + .1(2) = 3.4, \\
a_2 &= 0.12, \\
G_{3.} &= .4(1) + .5(2) + .1(0) = 1.4, \\
a_3 &= -1.88.
\end{aligned}$$

The dominance genetic effects are

$$\begin{aligned}
d_{ij} &= G_{ij} - a_i - a_j - \mu, \\
d_{11} &= 5 - 0.32 - 0.32 - 3.28 = 1.08, \\
d_{12} &= 3 - 0.32 - 0.12 - 3.28 = -0.72, \\
d_{13} &= 1 - 0.32 - (-1.88) - 3.28 = -0.72, \\
d_{22} &= 4 - 0.12 - 0.12 - 3.28 = 0.48, \\
d_{23} &= 2 - 0.12 + 1.88 - 3.28 = 0.48, \\
d_{33} &= 0 + 1.88 + 1.88 - 3.28 = 0.48.
\end{aligned}$$

Table 2. Additive and dominance effects added.

Genotype	Frequency, f_{ij}	Genetic Value, G_{ij}	$a_i + a_j$	d_{ij}
A_1A_1	.16	5	0.64	1.08
A_1A_2	.40	3	0.44	-0.72
A_1A_3	.08	1	-1.56	-0.72
A_2A_2	.25	4	0.24	0.48
A_2A_3	.10	2	-1.76	0.48
A_3A_3	.01	0	-3.76	0.48

The additive genetic variance is

$$\sigma^2_{10} = .16(0.64)^2 + \cdots + .01(-3.76)^2 = 0.8032,$$

and the dominance genetic variance is

$$\sigma^2_{01} + .16(1.08)^2 + \cdots + .01(0.48)^2 = 0.5184.$$

3 Two Loci

With two loci, each locus has its own additive and dominance genetic effects. In addition, there could be interactions between the two loci. In fact, there are three possible interactions. Assume just two alleles per locus, and let the two loci be A and B. An interaction means that there is an additional effect above or below that expected. Suppose the A_1 has effect of 5 and A_2 an effect of 2, and let B_1 have an effect of 4, and B_2 an effect of 9. Then the genotype A_1A_2 would be expected to be 7, but because of a dominance interaction maybe the value of that genotype is 10 (3 above the value of 7). Similarly, let the genotype B_1B_2 have a value of 11 (2 below the value of 13).

If the two genotypes occur in the same individual, then the expected value of an animal with genotype A_1A_2 B_1B_2 would be (10+11)=21. However, there could be interactions between A_1 with B_1, A_1 with B_2, A_2 with B_1, and A_2 with B_2. These would be called "additive by additive" interactions, and each could have a different value.

There could also be interactions between A_1 with B_1B_2, A_2 with B_1B_2, B_1 with A_1A_2, or B_2 with A_1A_2, each with a different value. This kind of interaction is called an "additive by dominance" gene interaction. Also, A_1 could interact with B_1B_1 or with B_2B_2, or all three genotypes at the B locus.

The last kind of interaction (among two loci) is called a "dominance by dominance" gene interaction, between A_1A_2 and B_1B_2. There could also be interactions between the other genotypes, A_1A_1 with B_2B_2, or A_1A_1 with B_1B_1, and so on.

In practice the variance of "additive by additive", "additive by dominance", and "dominance by dominance" interactions are estimated among all possible pairs of loci in the entire genome that affect the same trait. A special notation is used for these variances. The variance symbol, σ^2, is used with two subscripts. The first subscript indicates the degree of additive interaction, and the second subscript indicates the degree of dominance interaction. Thus, additive genetic variance of single loci is indicated by σ^2_{10}. The dominance genetic variance of single loci is denoted by σ^2_{01}. The others are

$$\begin{aligned}\sigma^2_{20} &= \text{additive by additive,}\\ \sigma^2_{11} &= \text{additive by dominance,}\\ \sigma^2_{02} &= \text{dominance by dominance.}\end{aligned}$$

With three loci, there are three way interactions which can be denoted as follows:

$$\begin{aligned}\sigma^2_{30} &= \text{add. by add. by add.}\\ \sigma^2_{03} &= \text{dom. by dom. by dom.}\\ \sigma^2_{21} &= \text{add. by add. by dom.}\\ \sigma^2_{12} &= \text{add. by dom. by dom.}\end{aligned}$$

plus the previously described interactions.

Given that there are approximately 30,000 total loci in the genome, the number of possible interactions can become very large. Estimating the gene interaction variances is a very complex problem which has not been attempted very frequently in the past. Usually, studies have not gone beyond interactions among two loci. The assumption is that variances of gene interactions for 3 or more loci are generally small and insignificant (because estimates of interactions for 2 loci have been small).

4 Genetic Variances and Covariances

The total genetic variance is the sum of all gene interaction variances. Let estimates of those variance be as follows (just an example):

$$\begin{aligned}\sigma^2_{10} &= 100,\\ \sigma^2_{01} &= 80,\\ \sigma^2_{11} &= 40,\end{aligned}$$

$$\begin{aligned} \sigma^2_{20} &= 60, \\ \sigma^2_{02} &= 20, and \\ \sigma^2_e &= 300. \end{aligned}$$

The total genetic variance would be

$$\sigma^2_G = (100 \ + \ 80 \ + \ 40 \ + \ 60 \ + \ 20) \ = \ 300.$$

The total phenotypic variance would be

$$\sigma^2_y = \sigma^2_G \ + \ \sigma^2_e \ = \ 600.$$

Heritability in the "broad" sense is the total genetic variance divided by the total phenotypic variance, which is 0.5 in this example.

Heritability in the "narrow" sense is the additive genetic variance divided by the total phenotypic variance, which is 0.1667 in this example.

The genetic covariance between two related individuals, X and Z, is given by the formula,

$$\sigma_{XZ} = \sum_i \sum_j (a_{XZ})^i \ (d_{XZ})^j \sigma^{ij}.$$

If X and Z have an additive relationship of 0.5, and a dominance relationship of 0.25, then the genetic covariance between them is

$$\begin{aligned} \sigma_{XZ} &= (0.5)^1 \ (0.25)^0 (100) \\ &\quad +(0.5)^0 \ (0.25)^1 (80) \\ &\quad +(0.5)^1 \ (0.25)^1 (40) \\ &\quad +(0.5)^2 \ (0.25)^0 (60) \\ &\quad +(0.5)^0 \ (0.25)^2 (20) \\ &= 91.25. \end{aligned}$$

Random Regression Models

Fall 2008

1 Introduction

All biological creatures grow over their lifetime. Traits that are measured at various times during that life are known as *longitudinal* data. Examples are body weights, body lengths, milk production, feed intake, fat deposition, and egg production. On a biological basis there could be different genes that turn on or turn off as an animal ages causing changes in physiology and performance. Also, an animal's age can be recorded in years, months, weeks, days, hours, minutes, or seconds, so that, in effect, there could be a continuum or continuous range of points in time when an animal could be observed for a trait. These traits have also been called *infinitely dimensional* traits.

Take body weight on gilts during a 60 day growth test, as an example.

Table 1 Pig weight data on performance test.

Animal	Days on Test					
	10	20	30	40	50	60
1	42	53	60	72	83	94
2	30	50	58	68	76	85
3	38	44	51	60	70	77
SD	1.6	3.7	3.9	5.0	5.3	5.6

The differences among the three animals increase with days on test as the gilts become heavier. As the mean weight increases, so also the standard deviation of weights increases. The weights over time could be modeled as a mean plus covariates of days on test and days on test squared.

2 Basic Structure of RRM

Random regression models (RRM) have been proposed for the analysis of longitudinal data. Such models have a basic structure that is similar in most applications. A simplified RRM for a single trait can be written as

$$y_{ijkn:t} = F_i + g(t)_j + r(a, x, m1)_k + r(pe, x, m2)_k + e_{ijkn:t},$$

where

$y_{ijkn:t}$ is the n^{th} observation on the k^{th} animal at time t belonging to the i^{th} fixed factor and the j^{th} group;

F_i is a fixed effect that is independent of the time scale for the observations, such as a cage effect, a location effect or a herd-test date effect;

$g(t)_j$ is a function or functions that account for the phenotypic trajectory of the average observations across all animals belonging to the j^{th} group;

$r(a, x, m1)_k = \sum_{\ell=0}^{m_1} a_{k\ell} x_{ijk:\ell}$ is the notation adopted for a random regression function. In this case, a denotes the additive genetic effects of the k^{th} animal, x is the vector of time covariates, and $m1$ is the order of the regression function. So that $x_{ijk:\ell}$ are the covariables related to time t, and $a_{k\ell}$ are the animal additive genetic regression coefficients to be estimated;

$r(pe, x, m2)_k = \sum_{\ell=0}^{m_2} p_{k\ell} x_{ijk:\ell}$ is a similar random regression function for the permanent environmental (pe) effects of the k^{th} animal; and

$e_{ijkn:t}$ is a random residual effect with mean null and with possibly different variances for each t or functions of t.

The function, $g(t)_j$, can be either linear or nonlinear in t. Such a function is necessary in a RRM to account for the phenotypic relationship between y and the time covariables (or other types of covariables that could be used in a RRM). In a test day model, $g(t)_j$ accounts for different lactation curve shapes for groups of animals defined by years of birth, parity number, and age and season of calving within parities, for example. With growth data, $g(t)_j$ accounts for the growth curve of males or females of breed X or breed Y from young or old dams.

The random regressions are intended to model the deviations around the phenotypic trajectories. Orthogonal polynomials of standardized units of time have been recommended as covariables (Kirkpatrick et al., 1990). Spline functions have also been suggested in some situations.

3 Example Data Analysis By RRM

Below are the data structure and pedigrees of four dairy cows. Given is the age at which they were observed for a trait during four visits to one herd.

Table 2. Example dairy cattle longitudinal data.

			Age, Obs. at Visit			
Cow	Sire	Dam	Visit 1	Visit 2	Visit 3	Visit 4
1	7	5	22,224	34,236	47,239	
2	7	6	30,244	42,247	55,241	66,244
3	8	5	28,224	40,242		
4	8	1		20,220	33,234	44,228

The model equation might be

$$\begin{aligned} y_{jik:t} &= V_j + b_0 + b_1(A) + b_2(A)^2 \\ &\quad +(a_{i0}z_0 + a_{i1}z_1 + a_{i2}z_2) \\ &\quad +(p_{i0}z_0 + p_{i1}z_1 + p_{i2}z_2) + e_{jik:t} \end{aligned}$$

where

V_j is a random contemporary group effect which is assumed to follow a normal distribution with mean 0 and variance, $\sigma_c^2 = 4$.

b_0, b_1, and b_2 are fixed regression coefficients on (A) = age and age squared which describes the general relationship between age and the observations,

a_{i0}, a_{i1}, and a_{i2} are random regression coefficients for animal i additive genetic effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix, $\mathbf{G}$,

p_{i0}, p_{i1}, and p_{i2} are random regression coefficients for animal i permanent environmental effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix, $\mathbf{P}$,

z_0, z_1, and z_2 are the Legendre polynomials based on standardized ages and derived as indicated earlier. The minimum age was set at 18 and the maximum age was set at 68 for calculating the Legendre polynomials.

and e_{jik} is a temporary residual error term assumed to follow a normal distribution with mean 0 and variance, $\sigma_e^2 = 9$. In this example, the residual variance is assumed to be constant across ages.

The model in matrix notation is

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

where

$$\mathbf{X} = \begin{pmatrix} 1 & 22 & 484 \\ 1 & 30 & 900 \\ 1 & 28 & 784 \\ 1 & 34 & 1156 \\ 1 & 42 & 1764 \\ 1 & 40 & 1600 \\ 1 & 20 & 400 \\ 1 & 47 & 2209 \\ 1 & 55 & 3025 \\ 1 & 33 & 1089 \\ 1 & 66 & 4356 \\ 1 & 44 & 1936 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 224 \\ 244 \\ 224 \\ 236 \\ 247 \\ 242 \\ 220 \\ 239 \\ 241 \\ 234 \\ 244 \\ 228 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\mathbf{z} = \begin{pmatrix} .7071 & -1.0288 & .8829 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & -.6369 & -.1493 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.7348 & .0632 & 0 & 0 & 0 \\ .7071 & -.4409 & -.4832 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & -.0490 & -.7868 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.1470 & -.7564 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -1.1268 & 1.2168 \\ .7071 & .1960 & -.7299 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & .5879 & -.2441 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.4899 & -.4111 \\ 0 & 0 & 0 & .7071 & 1.1268 & 1.2168 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & .0490 & -.7868 \end{pmatrix}.$$

In order to reduce rounding errors the covariates of age for the fixed regressions can be forced to have a mean of approximately zero by subtracting 38 from all ages and 1642 from all ages squared. Then

$$\mathbf{X} = \begin{pmatrix} 1 & -16 & -1158 \\ 1 & -8 & -742 \\ 1 & -10 & -858 \\ 1 & -4 & -486 \\ 1 & 4 & 122 \\ 1 & 2 & -42 \\ 1 & -18 & -1242 \\ 1 & 9 & 567 \\ 1 & 17 & 1383 \\ 1 & -5 & -553 \\ 1 & 28 & 2714 \\ 1 & 6 & 294 \end{pmatrix}.$$

The solutions for the animal additive genetic random regression coefficients are in Table 3.

Table 3. Solutions for animal effects.

Animal	a_0	a_1	a_2
1	-2.021529	.175532	-.002696
2	5.751601	-2.139115	.025848
3	-2.474456	2.554412	-.029269
4	-5.376687	-.370873	.002174
5	-1.886714	1.464975	-.016963
6	3.333268	-1.065525	.013047
7	1.503398	-1.081654	.012555
8	-2.948511	.681643	-.008633

Similarly, the solutions for the animal permanent environmental random regression coefficients can be given in tabular form.

Animal	p_0	p_1	p_2
1	-.296786	.246946	-.002521
2	3.968256	-.730659	.009430
3	-.834765	.925329	-.008164
4	-4.505439	-.441805	.001257

The problem is to rank the animals for selection purposes. If animals are ranked on the basis of a_0, then animal 2 would be the highest (if that was desirable). If ranked on the basis of a_1, then animal 3 would be the highest, and if ranked on the basis of a_2, then animal 2 would be the highest. To properly rank the animals, an EBV at different ages could be calculated, and then these could be combined with appropriate economic weights. Calculate EBVs for 24, 36, and 48 mo of age, and use economic weights of 2, 1, and .5, respectively, for the three EBVs. A Total Economic Value can be calculated as

$$\text{TEV} = 2 * \text{EBV}(24) + 1 * \text{EBV}(36) + .5 * \text{EBV}(48).$$

The Legendre polynomials for ages 24, 36, and 48 mo are given in the rows of the following matrix $\mathbf{L}$,

$$\mathbf{L} = \begin{pmatrix} .7071 & -.8328 & .3061 \\ .7071 & -.3429 & -.6046 \\ .7071 & .2449 & -.6957 \end{pmatrix}.$$

The results are shown in the following table.

Animal	EBV(24)	EBV(36)	EBV(48)	TEV
1	-1.58	-1.49	-1.38	-5.33
2	5.86	4.78	3.53	18.26
3	-3.89	-2.61	-1.10	-10.93
4	-3.49	-3.68	-3.89	-12.61
5	-2.56	-1.83	-.96	-7.43
6	3.25	2.71	2.09	10.25
7	1.97	1.43	.79	5.76
8	-2.66	-2.31	-1.91	-8.58

The animal with the highest TEV was animal 2. All animals ranked rather similarly at each age on their EBVs. Rankings of animals could change with age. Thus, the pattern of growth could be changed one that is desirable.

4 Application to Test Day Records

In a typical milk recording program, inspectors from the milk recording organization visit herds on a monthly basis in order to record the amount of milk given by each cow, and to take milk samples from each cow. The milk samples are sent to a testing laboratory and analyzed for protein and fat content, and somatic cell scores. The results from the lab are sent to the milk recording organization and merged with the milk yield information. Each visit is known as a "Test Day". For an entire lactation, a cow could be 'tested' 7 to 10 times.

In the past, a 305-day lactation yield would be calculated using the test day information and the Test Interval Method. If a cow gave 30 kg of milk on day 70 of lactation and 26 kg of milk on day 100, then the cow would receive credit for producing

$$(100 - 70) * (30 + 26)/2 \ = \ 30 * 28 \ = \ 840\text{kg}.$$

Special adjustment factors were needed for the first and last test days in lactation. By adding up the credits from 1 to 305 days would give the total lactation yield. TIM was replaced by MTP around 2000. MTP is essentially a mathematical function that accounts for the shape of the lactation curve, and handles milk, fat, protein, and somatic cell scores all at the same time. With the Test Interval Method regular herd visits were very important at roughly equal intervals. With MTP, the intervals between tests are not as important, and even the number of tests per lactation can be reduced. In both methods, a total yield over 305 days is calculated. The number 305 has been the standard lactation length since 1905 when milk recording began.

Two cows can produce equal 305-d lactation yields, but the manner in which they

produce this amount may be very different. Cow A could have a very high yield in the early part of lactation, but then test day yields could get smaller much more quickly by the end of the lactation. Cow B could have a lower peak yield, but may milk at that same level for more days than cow A before its test day yields start to decline. Cow B is said to be more persistent. Thus, if test day yields were analyzed instead of 305-d yields, the shapes of lactation curves could be evaluated. Random regression models were designed for this kind of problem.

Canada officially adopted the multiple-trait, random regression, test day model (CTDM) in February 1999 to replace a single-trait, repeated records, animal model. The changes were from a system

1. where lactation 305-d yields were considered as repeated measures of the same trait to a system where each lactation was considered to be a separate trait and the analyses were on test day yields;
2. where a standard lactation curve was assumed for each cow and lactation to a system where each lactation within a cowcould have a different shape of lactation;
3. that included 305-d yields from 1957 to a system that analyzed test day yields only from 1988 to the present; and
4. that could be computed easily in a few days to a system that required 2 weeks and a large amount of computer memory.

The advantages of the CTDM are

1. CTDM removes environmental effects from test day records more accurately;
2. CTDM models the shape of the lactation curve and the variability of yields around some general shapes;
3. CTDM provides more accurate genetic evaluations of cows in the range of 4 to 8% over evaluations based on 305-d yields.

On a given day, the k^{th} cow is at day t in its lactation, in parity n (limited to 1, 2, or 3), in herd-test date-parity subclass i, and calving within the j^{th} time period, region, age, and season subclass.

$$y_{nt:ijk} = \begin{pmatrix} \text{24-h milk yield, kg} \\ \text{24-h fat yield, kg} \\ \text{24-h protein yield, kg} \\ \text{somatic cell score} \end{pmatrix}$$

In some cases, one or more of these traits may be missing for some reason, but 24-h milk yield should always be present. Day t was limited to be between 5 and 305 days in lactation. Milk yields during the first 5 days are usually fed to calves or discarded.

5 Application to Growth Traits

The pattern of animal growth over time can be modeled by random regressions. In livestock species, growth is generally an economically important trait. Associated with growth are feed intake, feed efficiency, fat deposition, muscle development, bone length, degree of maturity, and body condition. Growth is slightly different from test day milk yields because body weights are cumulative over time. This would be analogous to accumulating daily milk yields of cows through the lactation rather than having individual test day yields on given days in the lactation. Accumulated weights have part-whole correlations from one weighing to the next, but will likely continue to be measured and analyzed as such.

One of the first applications of RRM to growth in pigs was made by Andersen and Pedersen (1996). In their study, pigs were weighed twice weekly from 30 kg live weight to 115 kg live weight. Machines monitored individual feed intake even though animals were in pens of twelve individuals. Thus, pigs started the test at different ages and consequently were weighed at different days on test. Weight and weight gains were modeled as a function of time, but were also modeled as a function of feed intake from which a measure of feed efficiency was derived. That is, the genetic merit for growth was a function of the amount of feed intake. Usually growth and feed intake are highly correlated both phenotypically and genetically, so that the genetic variation in growth remaining after accounting for feed intake would be reduced. The fixed curves of the model, $g(t)_j$, were a fourth order polynomial of days on test (not orthogonal polynomials), while the order of random regressions was 2. Growth rate was fairly linear between 30 to 115 kg, but did decrease between 30 and 50 days on test, and further decreased between 50 and 80 days on test, for both gilts and castrated males. Rather than model weights against feed intake, a multiple trait RRM model having both weight and feed intake traits against time on test would be a better way to examine feed efficiency without reducing the genetic variation in weight. A multiple trait RRM would simultaneously account for the changes in genetic and residual variation in each trait while allowing both traits and the relationship between those traits to vary together with time. The general concept would be not to model one trait against another if they are genetically correlated.

Maternal genetic effects of growth traits are known to be important in beef cattle.

Albuquerque and Meyer(2001) studied growth in Nelore cattle from birth to 630 days of age. The general RRM structure was augmented to include random regressions for maternal genetic effects and maternal permanent environmental effects. Let $r(ma, x, m3)$ and $r(mp, x, m4)$ denote the random regressions on maternal genetic of order $m3$ and maternal *pe* effects of order $m4$, respectively. However, Albuquerque and Meyer (2001) assumed zero correlations between direct and maternal genetic effects at all time points in order to simplify computations. Different orders of fit for the random regressions were applied to three different data sets. Using their notation, one of the favoured models was $k = 6\ 6\ 6\ 4$ which refers to the order (plus one) of the Legendre polynomials for direct genetic, maternal genetic, animal PE, and maternal PE effects, respectively, i.e. $k = (m1 + 1)\ (m3 + 1)\ (m2 + 1)\ (m4 + 1)$. Such a model has 77 (co)variances in addition to the residual variances to be estimated. Another favoured model based on Bayesian Information Content (BIC) was $k = 4\ 4\ 6\ 3$ with 51 parameters to be estimated. With either model, maternal genetic variance increased from birth to around 115 d of age and decreased thereafter, while direct genetic variance increased throughout from birth to 630 d of age and was generally much larger than the maternal genetic variance. Residual variances were small and increased only slightly with age. The effect of zero correlations between direct and maternal genetic effects was not examined, but perhaps may not be too important in these particular data.

Besides Andersen and Pedersen (1996), RRM have been applied to growth traits by Schnyder et al. (2001), Meyer (1999, 2000), Magnabosco et al. (2000), Schenkel et al. (2002), McKay et al. (2002), Veerkamp and Thompson (1999), and Uribe et al. (2000). The key issues in application of RRM to growth traits are the number of times individuals need to be measured, at what times in their lives, and what will be the upper age range. The costs of collecting these measurements would also play a role in determining how often and when to measure growth. A RRM provides some freedom in this regard, and animals are generally weighed at all ages from birth to maturity. Some animals could have many weights recorded while other animals may have only a few. Fixed growth curves should be estimated for each sex, within years of birth, within breeds or breed crosses, and within different parities of dams. Much work remains in applications to growth.

Besides animals, RRM could be applied to growth of plants, such as crops (which grow quickly) or trees (which grow slowly). RRM could be used to model growth of bacterial populations grown under certain conditions. Similar to growth would be a decay function such as the degradation of nutrients in the gut as they were digested in various parts of the gastrointestinal tract. Application of RRM to growth traits is in itself a growing area of research.

Economic Importance

Fall 2008

1 Introduction

The breeding objective in any livestock species is to improve the overall economic merit of the animals. Many traits contribute to the **Total Economic Value** of an animal. Suppose there are t traits of economic importance to a particular species, and let $\mathbf{g}$ be a vector of length t of true breeding values of an animal, then the **Aggregate Genotype**, H, is

$$H = \mathbf{v}'\mathbf{g},$$

where $\mathbf{v}$ is a vector of relative economic values of the t traits in $\mathbf{g}$. The Aggregate Genotype is approximated in practice by a **Selection Index**, I, as

$$I = \mathbf{w}'\hat{\mathbf{a}},$$

where $\hat{\mathbf{a}}$ are the EBVs on m traits for one animal and $\mathbf{w}$ are relative economic weights. Note that m could be more, less, or equal to t. The Aggregate Genotype could include more traits than those currently recorded on the species. Another difference between the Aggregate Genotype and Selection Index is that $\mathbf{g}$ are the true (unknown) breeding values on t traits and $\hat{\mathbf{a}}$ are the estimated breeding values on m traits, and lastly, $\mathbf{w}$ takes into account the reliabilities of the EBVs while $\mathbf{v}$ is based on perfect knowledge of the breeding values.

One problem with economic indexes is that economics can change over time and sometimes the change can be very rapid. For example, the discovery of BSE (bovine spongiform encephalitis or Mad cow disease) in Canada changed the value of beef cattle overnight from profitable to nothing. Most economic changes are not this drastic, and the relative economic importance of one trait to another stays 'constant' over time. For example, the value of reproductive performance to conformation traits does not fluctuate greatly over time. Genetic improvement is not instantaneous and does take some years to achieve, and hopefully relative economic values stay the same during this time.

An assumption of the Selection Index approach is that the value of traits is linear, as the trait EBV gets larger then the economic value also gets larger. However, for some traits, the added value above a particular EBV level actually remains constant or increases at a slower rate. Some traits have *intermediate optima*, such as birthweights of beef cattle. These are advanced issues that will not be covered in this course.

2 Aggregate Genotypes

The Aggregate Genotype contains all of the traits of economic importance in a species whether or not data are collected for all of these traits. The relative economic values may or may not be known for all of these traits. One must know the genetic variances and covariances among all t traits. These would be difficult to attain if some of the traits are not recorded in the population. The traits included are those that the breeder wishes to change for the better, or to not change while other traits are improved. Here, the breeder must define the longterm goals of the breeding program. The relative economic values may reflect true economic values or may also reflect the breeder's desired importance for one or more traits. The Aggregate Genotype is the plan that will be followed.

3 Selection Index

The Selection Index contains EBVs on traits that are readily available from the recording program for that species. The economic weights must be derived. The perfect way to estimate the economic weights would be to calculate the economic value of every animal, from an accountant's point of view. Animals would receive credit for producing offspring, but would lose money based on the amount of feed consumed, health costs, breeding costs, and particular management costs. Most animals that are kept should be net gainers in economic value.

Table 1 contains animals and their EBVs for two traits, one measured in centimeters and one measured in grams, plus a dollar value summarizing their accumulated costs and profits up to a fixed age. To determine the relative economic weights a linear regression model is applied,

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of dollar values of each animal, and $\mathbf{X}$ contains the EBVs of the traits, one column per trait, plus an overall mean (column of 1's).

Table 1
Animals, EBVs for two traits, and dollar value of animal.

Animal	Trait 1 EBV (cm)	Trait 2 EBV (g)	Dollar Value
1	+2	+119	61.80
2	+31	-72	357.70
3	-20	-124	302.80
4	+33	+76	184.70
5	-55	-61	146.70
6	-48	+71	4.40
7	+17	-73	326.70
8	+45	+61	230.30

The least squares equations to solve are

$$\begin{aligned}\hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \\ &= \begin{pmatrix} 8 & 5 & -3 \\ 5 & 10097 & 4445 \\ -3 & 4445 & 58309 \end{pmatrix}^{-1} \begin{pmatrix} 1615.10 \\ 18889.10 \\ -60347.30 \end{pmatrix}, \\ \begin{pmatrix} \hat{\mu} \\ \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} &= \begin{pmatrix} 200.00 \\ 2.30 \\ -1.20 \end{pmatrix}.\end{aligned}$$

The selection index equation would be

$$I = 2.30(EBV)_1 - 1.20(EBV)_2.$$

Ideally, this equation should be based on many animals. Using this equation, index values would be calculated for each animal as follows:

Animal	I -$
1	-138.20
2	+157.70
3	+102.80
4	-15.30
5	-53.30
6	-195.60
7	+126.70
8	+30.30

I values are called the Selection Index Criteria. Animals may be ranked based on the index values and the lower ranking animals may be removed from the breeding

population. Thus, both traits 1 and 2 will be improved simultaneously by selecting on the index values.

Table 14.2 contains the top four animals for each trait and for the index value. The last line gives the averages of the four animals. The same four animals were not necessarily selected in each group.

Table 2
Top 4 Ranking Animals by Trait and by Index.

By Trait 1 (cm)		By Trait 2 (g)		By Index ($)		
EBV	I	EBV	I	EBV_1	EBV_2	I
+45	+30.30	-124	+102.80	+31	-72	+157.70
+33	-15.30	-73	+126.70	+17	-73	+126.70
+31	+157.70	-72	+157.70	-20	-124	+102.80
+17	+126.70	-61	-53.30	+45	+61	+30.30
+31.5	+74.85	-82.5	+83.48	+18.2	-52.0	+104.38

If selection was only on the EBV for trait 1 versus the index, the average EBV for trait 1 would be +31.5 cm compared to +18.2 cm based on index selection, but the dollar value of the animals selected using only EBV for trait 1 would be $29.53 less than those selected on index value. Selection on EBV of trait 2 only, resulted in a better average EBV for trait 2 than index selection, but a loss of $20.90 in index value. Index selection will maximize the dollar value of animals selected, but will not result in the highest average EBVs for each trait. Therefore, genetic change in traits 1 and 2 will be slower with index selection than selecting on only one trait at a time, and genetic change in index value should be highest.

4 Relative Emphasis

In the previous example, by selecting on index values, was more weight put on trait 1 or trait 2? To answer this question the variances and covariances of the True BVs are needed. Variances of true BVs tend to be greater than variances of EBVs. For the example data in Table 14.1, assume that

$$Var\begin{pmatrix} BV_1 \\ BV_2 \end{pmatrix} = \begin{pmatrix} 1600 & 650 \\ 650 & 8836 \end{pmatrix}.$$

To determine the relative emphasis, the value of a one standard deviation change in traits must be compared. For trait 1, the value of one standard deviation change

is

$$w_1\sigma_{a_1} \;=\; 2.30(40) \;=\; 92.00\$,$$

and a one standard deviation change in trait 2 is

$$w_2\sigma_{a_2} \;=\; -1.20(94) \;=\; -112.80\$.$$

The relative emphasis of trait 1 to trait 2 in the index is

$$\text{RelativeEmphasis} = \frac{\mathrm{w_2}\sigma_{\mathrm{a_2}}}{\mathrm{w_1}\sigma_{\mathrm{a_1}}} = -1.226.$$

Thus, more emphasis is placed on trait 2 in this index.

If the emphasis is desired to be equal, then the weight on one trait needs to be adjusted so that the value of one standard deviation of each trait is equal. Thus, change w_1 to 2.82 \$/cm OR change w_2 to -0.98 \$/g.

If the emphasis on trait 2 was to be twice as large than on trait 1, then

$$w_2 = \frac{92.00(2)}{-94} \;=\; -1.96\$/g.$$

The index can therefore be manipulated to give the desired results. The regression equation, however, gives an indication of how each trait contributes to total economic value, at the present time and under the current financial situation. One could speculate on future changes in the economics of feed or products and change the dollar values of animals based on those projections. This would give weights on EBVs that are designed for that future financial scenario.

5 Custom Made Indexes

Some livestock industries develop selection index weights for producers to help the industry as a whole. However, each producer may have different feed costs and sales markets, so that an index specifically for that herd or flock would be better than a one-size-fits-all index. Some websites allow producers to enter their costs, prices received, and traits to be improved in order to design a custom-made index. That index should maximize the change in dollar value of the animals in that herd based on the producer's goals. This would be a desirable approach, in general.

6 Selection Practices

Selection index, if designed appropriately, will maximize the genetic change in dollar value of animals. Even so, other forms of selection are practiced in the industry.

Independent Culling Levels. A selection index is used, but added to this are minimum levels for each trait in the index (or for a few of them). For example, the minimum selection index dollar value may be +50, but if the EBV for trait 1 is negative, then that animal is culled regardless of the index dollar value. This changes the relative emphasis on the traits and is less efficient at maximizing the genetic change in dollar values. However, trait 1 may be a problem in that herd such that the producer can not afford to use animals with negative EBVs for that trait.

Tandem Selection. No selection index is used. The selection criterion changes from one year to the next. This year selection may be on EBVs for trait 1, and next year selection would be on EBVs for trait 2. This could result in no change in dollar value of the animals selected in the long term.

Phenotypic Selection. Producers often consider only the phenotypic values of animals and not what is transmitted to offspring. This is much less accurate than the selection index approach or the previous two methods unless the heritability of traits is very high. Residual effects can be very large with phenotypic records. EBVs are the best way to make genetic change.

Correlated Responses

Fall 2008

Application of the selection index method will cause two types of changes in the population, Direct and Indirect.

1. There will be a **direct** response in I values of animals.

$$\Delta I \;=\; r_{TI}\; i\; \sigma_I,$$

where (for two traits)

$$\begin{aligned}\sigma_I^2 &= Var(w_1 g_1 + w_2 g_2),\\ &= Var(\mathbf{w}'\mathbf{g}),\\ &= \mathbf{w}'\mathbf{G}\mathbf{w}.\end{aligned}$$

Assume the following values, where

$$\mathbf{G} = Var\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} 1600 & 650 \\ 650 & 8836 \end{pmatrix},$$

and

$$\mathbf{w}' = \begin{pmatrix} 2.30 & -1.20 \end{pmatrix},$$

then

$$\sigma_I^2 = 17,599.84\$^2,$$

or $\sigma_I = 132.66$ \$.

2. There will be **indirect** responses in any other traits that are genetically correlated with those in I. For trait k, the indirect response is

$$\begin{aligned}\Delta_c G_k &= b_{k.I}\Delta I,\\ &= b_{k.I}\; r_{TI}\; i\; \sigma_I,\\ &= \frac{\sigma_{G_k I}}{\sigma_I^2} r_{TI}\; i\; \sigma_I,\\ &= \frac{\sigma_{G_k I}}{\sigma_I} r_{TI}\; i.\end{aligned}$$

1 Traits Included in the Index

The covariance between the genetic value of a trait, k, and the index, I, is

$$Cov(G_k, I) = \sigma_{G_k I} \;=\; Cov(\mathbf{q}_k'\mathbf{g}, \mathbf{w}'\mathbf{g}),$$

where $\mathbf{q}_k'$ is a vector of all zeros except one 1 to designate trait k. For example, in a two trait index to indicate the first trait then

$$\mathbf{q}_k' = \begin{pmatrix} 1 & 0 \end{pmatrix}.$$

Then, for $k = 1$,

$$\begin{aligned} Cov(\mathbf{q}_1'\mathbf{g}, \mathbf{w}'\mathbf{g}) &= \mathbf{q}_1'\mathbf{G}\mathbf{w}, \\ &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1600 & 650 \\ 650 & 8836 \end{pmatrix} \begin{pmatrix} 2.30 \\ -1.20 \end{pmatrix}, \\ &= 2900\text{\$-cm}. \end{aligned}$$

Similarly, for $k = 2$,

$$Cov(\mathbf{q}_2'\mathbf{g}, \mathbf{w}'\mathbf{g}) = -9108.20\text{\$-g}.$$

Let $r_{TI} = 0.6$, $i = 1.5$, and σ_I was found to be 132.66 \$, then the direct response to selection would be

$$\Delta I = (0.6)(1.5)(132.66\$) = 119.39\$.$$

The indirect responses in each trait in the index would be

$$\begin{aligned} \Delta_c G_1 &= \frac{2900\text{\$-cm}}{132.66\$}(0.6)(1.5) = 19.67\text{cm}, \\ \Delta_c G_2 &= \frac{-9108.20\text{\$-g}}{132.66\$}(0.6)(1.5) = -61.79\text{g}. \end{aligned}$$

2 Traits Not Included in the Index

Suppose a third trait (measured in seconds, s) was recorded, but was not a part of the breeding objective or selection index. Assume that the covariances among the three traits were as shown below:

$$\mathbf{G} = \begin{pmatrix} 1600 & 650 & -100 \\ 650 & 8836 & -320 \\ -100 & -320 & 2704 \end{pmatrix}.$$

The index is still

$$I = 2.30g_1 - 1.20g_2,$$

and $\sigma_I = 132.66$ \$. The formula for correlated response is the same as before. The covariance between g_3 and I is

$$\begin{aligned} Cov(\mathbf{q}_3'\mathbf{g}, \mathbf{w}'\mathbf{g}) &= \mathbf{q}_3'\mathbf{G}\mathbf{w}, \\ &= \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1600 & 650 & -100 \\ 650 & 8836 & -320 \\ -100 & -320 & 2704 \end{pmatrix} \begin{pmatrix} 2.30 \\ -1.20 \\ 0.00 \end{pmatrix}, \\ &= 154\text{\$-s}, \end{aligned}$$

and

$$\Delta_c G_3 = \frac{154\text{\$-s}}{132.66\$}(0.6)(1.5) = 1.04\text{s}.$$

Trait 3 would have changed to be 1.04 seconds longer.

Remember that all traits that are genetically correlated to the traits in the index will have an indirect response associated with it. Thus, selection on an index could harm other traits not included in the index or not known to be correlated to traits in the index. If in doubt, assume that all other traits are correlated to the index and try to determine how much they might change and what direction, if the selection index was applied.

3 Efficiency of Alternative Indexes

Comparison of different indexes is often necessary. Suppose one alternative to the previous index was to select only on trait 2 alone. The alternative index could be written as

$$I_{a1} = \begin{pmatrix} 0.00 & -1.20 & 0.00 \end{pmatrix} \mathbf{g},$$

then

$$\begin{aligned} \sigma_{I_{a1}} &= 112.80\$, \\ \Delta I_{a1} &= (.6)(1.5)(112.80)\$ = 101.52\$, \\ \Delta_c G_1 &= \frac{-780.00\text{\$-g}}{112.80\$}(0.6)(1.5) = -6.22\text{cm}, \\ \Delta_c G_2 &= \frac{-10603.20\text{\$-g}}{112.80\$}(0.6)(1.5) = -84.60\text{g}, \\ \Delta_c G_3 &= \frac{384.00\text{\$-g}}{112.80\$}(0.6)(1.5) = 3.06\text{s}. \end{aligned}$$

Putting these results in a table, then

Table 15.1

Comparison of Correlated Responses from Alternative Indexes.

Trait	Original	Alternative
g_1 cm	19.67	-6.22
g_2 g	-61.79	-84.60
g_3 s	1.04	3.06
I \$	119.39	112.80

Selecting only on trait 2 would give 37% more response in trait 2 over the index on traits 1 and 2. Trait 1 would decrease in value, and trait 3 would increase more. The overall index dollars would be slightly less. The alternative index of selecting only on trait 2 would be 94% as efficient as the index on traits 1 and 2. Selecting on one trait gives the maximum response for that trait, but the indirect responses of other traits may not be in the desired direction and may give lower overall economic response.

4 Restricted Indexes

Suppose trait 3 is not to be changed. This means the correlated indirect response in trait 3 as a result of selecting on an index that includes traits 1 and 2 should be 0. This means that the covariance between trait 3 and the index must be 0. Trait 3 should be included in the index if it is not to change. The covariance between the index and trait 3 is

$$\begin{aligned} Cov(\mathbf{q}_3'\mathbf{g}, \mathbf{w}'\mathbf{g}) &= \mathbf{q}_3'\mathbf{G}\mathbf{w}, \\ &= \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1600 & 650 & -100 \\ 650 & 8836 & -320 \\ -100 & -320 & 2704 \end{pmatrix} \begin{pmatrix} 2.30 \\ -1.20 \\ w_3 \end{pmatrix}, \\ &= (154 + w_3(2704))\text{\$-s}, \end{aligned}$$

Equate this covariance to 0, then

$$w_3 = -0.05695\text{\$ / s.}$$

The new index would be

$$I_{a2} = \begin{pmatrix} 2.30 & -1.20 & -0.05695 \end{pmatrix} \mathbf{g}.$$

Determine the direct and indirect responses.

5 Desired Gains Index

Another way to determine the weights in a selection index is to define the desired responses in each trait after one generation of selection. Then determine the weights that would give those indirect responses. Suppose the desired gains (using the previous example) were 22 cm for trait 1, -65 g for trait 2, and 2.00 s for trait 3, then the appropriate index weights would be

$$\mathbf{w} = \mathbf{G}^{-1} \begin{pmatrix} 22 \\ -65 \\ 2.00 \end{pmatrix} * 132 = \begin{pmatrix} 2.28 \\ -1.14 \\ 0.0474 \end{pmatrix}.$$

The desired gains were not greatly different from the actual gains using the previous selection index, and therefore, the new weights are not greatly different. The value 132 is the desired standard deviation of the index, which is similar to standard deviation of the previous index (119).

Mating Systems

Fall 2008

After selecting the males and females that will be used to produce the next generation of animals, the next big decision is which males should be mated to which females. Mating decisions should consider

- the traits to be improved,
- traits not to be diminished,
- rate of inbreeding, and
- purpose of the mating.

The purpose of a mating may be to produce offspring for a potential market. For example, beef animals for slaughter need to be uniform in appearance (size, weight, fat thickness). A market for selling breeding stock, however, must produce genetically superior individuals with good maternal characteristics and good performance.

The rate of inbreeding is important. Increasing the degree of homozygosity of alleles at gene locations can be both good and bad. On the good side, a favourable allele may be fixed in the population by inbreeding. At the same time, an unfavourable allele for a different trait may be fixed. Inbreeding reduces the effects of dominance genetic effects.

The selection index, already discussed, includes the traits of economic importance and should have the appropriate weights on each trait. Even so, if a female has an EBV of -1.5 genetic standard deviations for one trait within the index, mating her to a male that also has a big negative EBV for the same trait may not be advisable. Try to pick a male that has the same overall index value, but who's EBV for this trait is average or above in the population.

1 Mating According to Index Values

There are two types of systems. **Random** mating is a mixing of the selected males and selected females in a random manner. Random mating can occur in herds of large numbers of animals because from a labour point of view this is an efficient strategy. With small groups of animals, owners prefer to make **Assortative** matings. There are **Positive Assortative** matings and **Negative Assortative** matings.

1.1 Positive Assortative Matings

This is mating "like to like". Suppose you have 8 females and 3 males for making matings. Rank the males from best to worst on index values, and rank the females from best to worst on their index values. Then decide the number of matings that each male will make. The best male could be mated to the top 4 females (maybe that is the limit), the second best male is mated to the next 3 females, and the third male is mated to the last female. Allowances must be made for females that do not become pregnant on the first mating. Perhaps the first two males are used for all first matings and the third male is used for all second and later matings.

Positive assortative mating tends to produce more genetic and phenotypic variability in the offspring generation compared to that produced by random mating. The usual normal distribution is 'flattened' slightly in the distribution of individuals. Positive assortative matings follow the goal of changing the mean of the population.

The purpose of positive assortative matings is to increase the probability of producing an outstanding extreme genetic individual. If the outstanding individual is a male, then it can be used more readily to spread its genes to many more individuals in the next generation. Positive assortative matings are the most likely matings in Thoroughbred race horses, for example.

1.2 Negative Assortative Matings

This mating system produces offspring that are closer to the mean of the population. Genetic and phenotypic variability is reduced. Negative assortative mating (or disassortative mating) is where the highest ranking males are mated to the lowest ranking females and vice versa. Rank the males from highest to lowest on index value and rank the females from lowest to highest on index value. Then mate the males to the females in the order of the two lists.

Negative assortative mating is a system to improve uniformity of the progeny. This is not a good system for making genetic change rapidly.

When looking at a single trait, negative assortative matings are also called **corrective matings**. A 'fault' in one parent is offset by a high EBV for the trait in the other parent.

2 Matings According to Relationships

2.1 Linebreeding

Some animal owners like to increase the number of animals in their herd (breeding program), that are related to a particular outstanding individual, usually a female. They are practicing **Linebreeding**. The daughters and granddaughters of an individual are kept for breeding purposes and may be used for a long time (increasing the generation interval). These females may also be highly related to a particular sire at the same time. Sons of this sire may be used for mating to the daughters and granddaughters. Obviously, these animals will be highly related, and most likely will be inbred, but 'the good alleles are being concentrated in the line'.

Linebreeding may be useful if the economic values of the animals in the line are enhanced because of their relatedness. Otherwise linebreeding is not a good strategy for maintaining genetic diversity and for making genetic change.

In some ways, Holstein dairy cattle in Canada, for example, can be considered a 'line' that is different from the 'lines' of Holstein dairy cattle in England or Finland. However, Holstein sires are now used world-wide and the existence of 'lines' is less evident than it used to be.

2.2 Deliberate Inbreeding

Mating of highly related individuals increases the homozygosity of alleles at gene loci. **Inbreeding Depression** is a decrease in performance of traits (generally with low heritability) which are thought to be influenced by non-additive genetic effects (i.e. dominance effects). Inbreeding is a way to 'fix' an allele in a population, so that all animals are homozygous for this allele, and therefore, all progeny receive this allele - which is hopefully beneficial to the population. In the process of 'fixing' an allele, other alleles may also become 'fixed' which may not be desirable.

Inbreeding depression commonly affects reproductive fitness, and once the level of inbreeding becomes too high then successful reproduction becomes more difficult to achieve. A population could actually breed itself to extinction if inbreeding levels are high. This is a major concern for species that only exist in zoos. There are few reproductive pairs of individuals, and these are most likely related to each other. Survival of offspring is another trait affected by inbreeding depression.

2.3 Inbreeding Avoidance

Matings can also be made with the intent of minimizing the average inbreeding coefficient in the progeny. Selection on BLUP EBVs from an animal model tends to automatically increase the probability of mating related individuals, and thus, inbreeding would increase rapidly. This is more of a problem with traits of low heritability because the EBVs would mostly be based on parent averages until an animal has a large number of offspring.

Outcrossing is a term given to matings within a breed that are as unrelated as possible. The purpose is to avoid inbreeding, but also to maximize heterozygosity of gene loci to capitalize on non-additive genetic effects.

Given a list of males and females to be used in matings, there are mating packages that will determine the matings that will minimize the average inbreeding coefficient of the offspring for a given desired level of genetic change. However, the owner must be prepared to follow the mating plan given by the program, without any deviance. If some females fail to conceive on the first mating and if semen from the same male is not available for a second mating, then the plan fails. Many owners also do not like to follow the mating suggestions of 'computer' programs, and make 'special cases' for specific females. Thus, the success of such programs is weakened by the number of 'special cases' that owners like to make.

3 Matings Between Breeds - Crossbreeding

Crossbreeding is the mating of individuals from different breeds within a species. The assumption is that each breed has been selected for several generations (within breed) and that the genes that have become 'fixed' or established in that breed are different from those that have become established in another breed. Thus, by mixing breeds, the favourable alleles of each breed are combined in some offspring. Heterozygosity should be at its maximum.

Heterosis, H, is defined as the superiority of crossbred offspring compared to the average of the two parental breeds.

$$H = 100 \times \frac{\text{Average of Crossbreds} - \text{Average of Parents}}{\text{Average of Parents}}.$$

Heterosis is also known as **Hybrid Vigor**.

Consider a single gene locus, say `A`, with three possible genotypes, i.e. (`AA, Aa,` and `aa`). In general, the genetic values of a single locus are denoted as

Genotype	Genetic Value
AA	s
Aa	t
aa	u

If gene action affecting a trait is entirely additive, then the genetic values of these genotypes would be such that `t = (s + u)/2`. The `AA` and `aa` genotypes would be the two parental breeds, and the `Aa` genotype would be the crossbred progeny. Heterosis would be zero because the average of the offspring would equal the average of the parental breeds. There would be no advantage to crossbreeding in this situation. Heterozygosity would be achieved in the progeny, but because the gene action is additive the heterozygotes would simply be half-way between the value of the two homozygote parent breeds.

Now assume that dominance gene action exists, and let `t = s = 2` and `u = 1`. The offspring average would be `t=2`, and the average of the parental breeds would be `(s+u)/2 = 1.5`, then heterosis would be

$$H = 100 \times \frac{2 - 1.5}{1.5} = 33.33\%.$$

Dominance gene action is the primary cause of heterosis. Overdominance is where the value of the heterozygote is superior to that of the best parental breed. This phenomenon also contributes to heterosis.

Lastly, there could be interactions between gene loci, **epistasis**, and this may also contribute to heterosis. However, the importance of this source of heterosis is considered to be low.

Within the entire genome, some gene loci will be acting in an entirely additive manner, and other gene loci will have dominance effects. Thus, you could get 100 % heterosis at some loci and 0 % heterosis at many other loci. The observed heterosis would be a combined average of the heterosis at every locus.

4 Crossbreeding Systems

A crossbreeding system is designed in order to take advantage of hybrid vigor in order to produce offspring that are consistent in performance. The breeds chosen must also complement each other.

4.1 Single Cross - Rotational System

In swine there are several breeds, some of which are Yorkshire (Y), Landrace (L), Hampshire (H), and Duroc (D). The Yorkshire and Landrace breeds are known for fast growth, while the Hampshire and Duroc are known for their meat quality. A single cross is a mating between two breeds. For example, females of the Duroc breed are mated to Yorkshire boars, and females of the Yorkshire breed are mated to Duroc boars. Purebred boars are always used, but the female replacements will be crossbred. A crossbred female who's sire was Yorkshire, would be mated to a Duroc boar, and a crossbred female who's sire was Duroc would be mated to a Yorkshire boar. This system requires two housing systems if natural matings are used, to make sure the female is mated to the correct breed of sire. Also, this system requires a source of superior purebred boars.

Only the first cross achieves all of the possible hybrid vigor. Offspring of crossbred females will be more than 50% of one breed, and so only a fraction of the heterosis will be expressed. How much? After about 7 generations of rotational matings, the equilibrium heterosis will be

$$\hat{H} = 100 \times \frac{2^n - 2}{2^n - 1},$$

where n is the number of breeds in the rotation. Thus, for a two-breed rotational crossing system,

$$\hat{H} = 100 \times \frac{4 - 2}{4 - 1} = 67\%.$$

For a three-breed rotation,

$$\hat{H} = 100 \times \frac{8 - 2}{8 - 1} = 86\%.$$

The following table illustrates the percentage of heterosis achieved in each cross up to generation 7.

Table 1. Breed Composition in Rotational Crossing System.

	Two breed rotation			Three breed rotation		
Gen.	Male	Progeny	$\hat{H}$	Male	Progeny	$\hat{H}$
0	Y	(50)Y +(50)D	100	Y	(50)Y +(50)D +(0)L	100
1	D	(25)Y +(75)D	50	L	(25)Y +(25)D +(50)L	100
2	Y	(63)Y +(37)D	75	D	(13)Y +(63)D +(25)L	75
3	D	(31)Y +(69)D	63	Y	(56)Y +(31)D +(13)L	88
4	Y	(66)Y +(34)D	69	L	(28)Y +(16)D +(56)L	88
5	D	(33)Y +(67)D	66	D	(14)Y +(58)D +(28)L	84
6	Y	(66)Y +(34)D	67	Y	(57)Y +(29)D +(14)L	86
7	D	(33)Y +(67)D	66	L	(29)Y +(14)D +(57)L	86

With four breeds a similar rotation could be established, but an additional twist would be to use crossbred boars. Using the swine breeds as an example, Yorkshire by Hampshire males could be mated to Landrace by Duroc females. More heterosis can be maintained with more breeds involved. One problem is getting a good estimate of the breeding values of crossbred animals. What would be a good statistical model for analyzing data from crossbred animals?

To simplify the rotational system, some breeders rotate breeds of sire from one breeding season to the next. Thus, in 1998 Yorkshire boars would be mated to all females. In the next year Duroc boars would be used on all females, and so on. This simplifies the practical breeding aspects, but may not optimize the utilization of heterosis.

4.2 Terminal Sire Systems

Breeds within a species have often been created by selection for a particular attribute. As already mentioned, Yorkshire and Landrace have been selected for growth while Hampshire and Duroc have been selected for meat quality traits. Some breeds have been selected for litter size and maternal characteristics of the sow.

In a terminal sire system, breeds that excel in the maternal characteristics are mated in a rotational system, and breeds that excel in performance traits are mated in a separate rotational system. There may be two or more breeds in each system. The crossbred females from the maternal rotational system are then mated to the best males from the performance rotational system. All offspring from this mating go to market and are not used for breeding purposes. A disadvantage is the need to maintain two rotational systems at one time, but heterosis is fully utilized.

4.3 Composite Breeds

A composite animal is a crossbred animal constructed from two or more breeds. Animals of the same genetic composition are mated to each other and selection is applied within this group. The crossbred animals become a composite breed. Offspring from composite animals may be more variable in performance and appearance because of segregation of alleles than either purebreds or F1 offspring (first cross). The more breeds that have gone into the composite, the more heterosis that is retained. A composite breed is created to have the 'good' qualities of each breed that has gone into it.

Dairy Cattle Breeding
Fall 2008

1 History

The first Holstein-Friesian in Canada was sold to Archibald Wright of Winnipeg in 1881. The Holstein makes up 95% of all dairy cattle in Canada. Milk recording programs began around 1905 with the aim of improving the milk production abilities of cows. Agriculture Canada was initially responsible for recording the performance of many species of livestock, and did so for many years. This included the computation of genetic evaluations for dairy bulls and cows. In dairy cattle, provincial recording programs started in the 1960's in Quebec and Ontario where the majority of dairy cattle are raised, but Agriculture Canada continued to compute genetic evaluations. In the early 1990's, Agriculture Canada decided to end its participation in animal performance recording and genetic evaluation. Each industry was given 3 years and about $3 million each to privatize the recording and genetic evaluation functions. The Canadian Dairy Network formed in 1995 with the mandate to compute genetic evaluations for all breeds and traits in Canada, and to participate in international evaluation programs. The annual budget for this activity is about $1 million, supported primarily by the artificial insemination (AI) industry. Milk recording was consolidated across provinces into one national program.

A milk quota system exists in Canada in which producers buy quota that allows them to produce milk, and in return they are guaranteed an income, and the amount of milk produced and transported is a fixed supply system.

The figures in the table below show that the numbers of herds and cows has been decreasing over the last 15 years, but the number of cows on milk recording has been more stable and actually increasing in the last four years. Average herd size is continually getting larger. So that the herds that are disappearing are the small farms. Note that the average herd size across Canada is greater than that in Quebec or Ontario. These trends are likely to continue.

Table 17.1. Numbers of cows and herds in Canada (all breeds).

Year	Quebec	Ontario	Canada
Herds			
1990	14,903	10,976	34,620
1995	11,782	8,509	25,700
2000	9,774	6,918	20,624
2004	8,054	5,641	16,970
Cows, '000			
1990	560.0	460.0	1,428.9
1995	507.0	419.0	1,274.0
2000	427.0	380.0	1,103.4
2005	407.0	354.8	1,065.0
Cows on Milk Recording, '000			
1990	105.0	193.6	416.8
1995	97.2	166.3	384.9
2000	111.4	154.5	386.3
2004	121.7	157.7	409.4
Average Herd Size			
1990	43.9	48.0	49.5
1995	46.3	50.5	53.3
2000	52.5	57.4	61.3
2004	60.9	67.3	72.4

2 Breeds

There are seven pure dairy breeds in Canada. AY=Ayrshire, BS=Brown Swiss, CA=Canadienne, GU=Guernsey, HO=Holstein, JE=Jersey, and MS=Milking Shorthorn. Crossbreeding is hardly used in dairy cattle. Breed associations are active and strong in promoting their breeds through shows, auctions, and classification. Table 17.2 contains information on each breed. Production averages are for 2004. The Holstein gives the most milk, and the Jersey has the highest fat and protein percentages. About 3-4% of animals registered are the result of embryo transfer, and this number is slowly increasing.

Table 17.2. Facts about dairy breeds in Canada.

Item	HO	AY	JE	BS	GU	MS	CA
Registrations							
1990	217,916	9,812	7,126	1,698	1,500	216	310
1995	202,102	8,812	6,565	1,775	1,005	277	209
2000	214,244	7,925	6,513	1,421	464	310	206
2004	232,754	7,217	6,245	1,450	292	210	194
Young bulls sampled							
1990	365	25	14	3	5	NA	1
1995	461	26	11	1	3	NA	1
2000	546	19	23	7	4	NA	0
2004	610	18	37	2	3	NA	0
Birth Wt., kg	44	33	30	41	30	40	30
Mature Wt., kg	680	540	450	630	555	555	450
Milk Yield, kg	9,658	7,323	6,291	8,048	6,435	6,595	5,776
Protein Yield, kg	307	243	236	279	221	213	203
Protein %	3.19	3.32	3.77	3.47	3.45	3.25	3.54
Fat Yield, kg	352	290	303	326	290	242	236
Fat %	3.67	3.97	4.85	4.07	4.54	3.69	4.12

3 Traits

There are many traits that affect the overall profitability of a dairy enterprise. However, the main source of income is through milk sales and somewhat through sale of animals for export or to other producers. The traits are related to the efficiency of production and reproduction.

Table 17.3. Genetic parameters for some traits in dairy cattle.

Trait	Heritability	Repeatability
Milk yield	.25-.45	.50
Fat yield	.25-.45	.50
Fat %	.40-.55	.60-.75
Protein yield	.20-.40	.55
Protein %	.40-.55	.60-.75
Lactose	.20	.50
Somatic Cell Score	.20	.50
Feed Intake	.30	
Final score	.25	
Mammary system	.20	
Feet and legs	.15	
Stature	.45	
Age at first service	.15	
Non Return rate	.03	
Longevity	.05	
Calving ease	.15	
Milking speed	.20	
Temperament	.12	

4 Industry Organization

The industry consists of the following main components.

- **Producers and consumers.** Producers raise and milk the cows.
- **Milk recording organizations.** Milk recording collects the records on cows and helps to provide management information to the producers. The national office is in Guelph.
- **Breed associations.** Breed associations register and identify animals and maintain pedigree records. Breed associations classify animals to ensure that certain standards are maintained in the appearance of animals in the breed. They assist producers in the sale of animals, transfers of ownership, finding markets, and representing the breed internationally. Today there is more concern about health of the animal and of the product that goes to consumers, and the breed associations have to lead in this area.
- **Artificial insemination organizations.** AI units select bulls for progeny testing, collect semen from bulls and store it in liquid nitrogen, and inseminate cows. They

also work to export Canadian genetics around the world. At one time there were five or six AI units across Canada, but the main units now are the Semex Alliance and Alta Genetics. There are other smaller units that represent US and European AI units.

- **Canadian Dairy Network.** Canadian Dairy Network (CDN) collects all data on cows (for all traits) into one large database for the purpose of genetic evaluation. CDN also represents the dairy industry internationally at INTERBULL and ICAR (International Committee on Animal Recording).

- **Dairy Farmers of Canada.** Lobbies for the dairy industry to federal and provincial governments. They are involved with the Canadian Milk Commission in the supply-management of the milk that is produced. They support research into dairy production through nutrition, food science, and genetics.

- **Veterinarians** are concerned with animal health, and universities are involved in research problems to produce healthy milk, from healthy cows, in a healthy environment.

- **Journals.** There are also many dairy oriented magazines such as the Holstein Journal (promoting the breed), and Hoard's Dairyman (tips and advice on farming).

- **Feed manufacturers** want to sell feeds to dairy producers, as well as pharmaceutical companies.

- **University researchers** provide research into the latest technologies that improve the life of the cow and the producer, as well as satisfying the consumer, with new milk products.

There is much more information available on the internet on each component.

5 Reproductive-Life Cycle

Timeline Months	Event
0	Calf is born Female calf is known as heifer calf Male calf is known as bull calf
12	Yearling heifer Decision is made to use calf as replacement
15	First breeding of heifer Gestation length is 9 months
24	Fresh heifer, first calving First parturition, first lactation
27+	Cow is re-bred
34	End of first lactation, 305 days Cow is dried off (rest)
36	Second calving Second lactation begins

The cycle continues for as long as the cow is kept in the herd. Some cows have had 17 lactations. The majority of cows have just 3 lactations. About 25-35% of cows are replaced each year.

The above cycle is for a typical ideal Holstein cow. Breeds differ slightly and individuals differ within a breed. Cows that take too long to re-breed or that are too old at first breeding are less profitable and should be culled.

Breedings can be either natural service (by a bull on the farm) or by artificial insemination (AI). Most dairy producers use AI for all first breedings. About 60% of all first AI breedings are successful. Producers may use AI for a second mating. After that, the cow is either culled or bred by natural service. Producers that use AI entirely may sometimes breed a cow up to 7 times, but the cow should be extremely valuable to spend this much time, effort, and money to get her pregnant.

Really valuable cows can be superovulated with hormones, the embryos collected and implanted into other less valuable cows. Usually 3 or more embryos can be collected per superovulation. Recipient cows can be cows that would otherwise be culled. There are companies that will also 'sex' the embryos to give you either female or male calves, but these are not totally efficient and the act of sexing the embryo lessens the chance of the embryo to survive. About 3-4% of cows registered were produced by ET. Nearly all bull calves have been produced by ET. Such animals usually have the letters ET at the end of their registered farm name. Putting a cow through the ET process delays the re-breeding of that cow, and subsequent calvings are usually at a later age. A producer may wait until a cow has completed at least two lactations before trying to get embryos. This is

so the EBV of the cow will be as accurate as possible, and so that the type conformation can be fully assessed.

6 Progeny Testing

Progeny testing has been the primary tool since 1950's for genetically improving dairy cattle. Through AI, a dairy bull can have many daughters and this provides a highly reliable EBV for the traits evaluated. Progeny testing is a very costly procedure. Bulls have to be kept until the EBV is available which is when the bull is roughly 6 years of age. AI units buy the bulls, feed them, and collect and store semen, and they pay technicians to travel the countryside to inseminate cows. The average cost to progeny test one bull is about $50,000. Note that there were 610 bulls progeny tested in the Holstein breed in 2004, which would be a cost of over $30 million.

The number of test matings depends on a formula. The general conception rate of AI semen, the percentage of matings to cows that are on milk recording, the 50-50 chance of producing a female calf, and the survival of female calves to maturity are factors in the formula. If the AI unit wants the bull to have 100 daughters with milk records, then at least 200 matings are needed because half will be males. Assume 50% of herds are on milk recording, then 400 matings are required, and if the conception rate is 63%, then 700 matings are needed. Of the female calves that are born in a milk recorded herd, the percentage that are kept for herd replacement has to be high. To achieve this, AI units offer money to producers for taking the first 50 heifers of an AI bull to complete their first lactation, and to get them classified by the breed association. The amount of money is not high, but is an incentive. Thus, AI units try to obtain 700 to 1000 test matings per young bull, and these are usually made within 2 to 4 weeks, depending on the popularity of the bull. In non-Holstein breeds, the testing period could be much longer.

Table 1. Progeny Testing Cycle of Events

Timeline Months	Event
-12	Identify dams of bulls Make contract matings to sires of bulls
0	Bull calf is born Calf is inspected If suitable, bull calf is bought
12	Bull calf produces semen Initial collections, once a week Test matings conducted, one month Enough matings to generate 50-100 daughters with records Bull "goes on the shelf", waiting period begins
24	Daughters are born
36	Daughters are inseminated
48	Daughters calve, first lactations
58-60	Daughters complete lactations Bull EBV available Some bulls culled immediately Some bulls "returned to active service" Active bulls get widespread use Second crop daughters are created
72	EBVs based on 1000's of daughters Decisions made on sires of bulls

AI units receive 'interim' EBVs on all of their bulls every 3 months based on the data that are available at the time. An EBV is not official, however, until there are a certain number of daughters and the EBV has a minimum reliability. Usually EBVs are fairly stable and do not change greatly, so that AI units can be fairly certain that some bulls will not do well and these can be culled before the EBV is official. Young bulls with very high, early EBVs can start to be collected again to be ready for the demand when the EBV (proof) becomes official.

7 International Scope

Dairy bull semen from Canada is exported to over 60 countries around the world. This export business is the main source of funds for an AI organization. The competition to sell semen is very intense. In the 1970's there was an interest in comparing bulls across countries. INTERBULL was created to address this problem. The first attempt was to run a trial comparison in Poland. Bulls from 14 countries were nominated for the trial and semen was sent 'free' to Poland to produce about 100 daughers per country. The

trial was conducted on large state farms in Poland. There were two years of matings. Several more years were needed until those daughters completed their first lactations, and another year before the data were analyzed. A similar trial was conducted in Bulgaria for Red and White cattle. The USA and Canada did well in the trials, but the results reflected the genetics that were available in the year when those bulls were test mated. By the end of the trials, genetic trends had taken each country further ahead at different rates, and so the comparisons were out of date. A less costly and less time consuming method was needed.

INTERBULL decided to take EBVs of bulls from different countries and to develop a way to compare countries from this information. Instantly it became clear that every country had different milk recording programs, different methods of genetic evaluation, and different criteria for EBVs to be official. Thus, efforts were made to 'standardize' by creating minimum standards for milk recording programs. A survey of genetic evaluation models was conducted and the results published. Tests were developed to determine the quality of EBVs that countries provided. This has benefitted all countries in that genetic evaluation methods have improved immensely over the years and all countries use very similar models now.

In 1993, a statistical model was proposed (called MACE) for making comparisons of bull EBVs from different countries. This method is still used by INTERBULL today, with some improvements, of course. MACE allows the EBV of a bull in the Netherlands, for example, to be expressed on the same scale as an EBV in Canada, and it combines the daughter information of that bull from every country in which it has daughters. Thus, young bulls could be progeny tested with daughters in several countries rather than just one. Some of Canada's young bulls are simultaneously test mated in more than one country.

As a result of international competition, AI units around the world are using the same sires of bulls to generate the next crop of young bulls. The number of effective sires of bulls, worldwide, is about 30. This leads to a rapid increase in inbreeding coefficients. Every country is progeny testing sons of the same bulls. The inbreeding rate in Canada goes up about 2.5% per year.

While statistically and scientifically the comparison of bulls in different countries can be achieved, there are often political hurdles to overcome. Thus, science sometimes has to wait for the politics to be settled. INTERBULL has a steering committee that governs what it does. The INTERBULL centre is in Uppsala, Sweden. Meetings are held in different countries every year. Between 100 and 200 people attend these meetings, and every country providing data to INTERBULL has at least one representative at the meetings. Visit the INTERBULL website on the internet, and look for MACE results on the CDN website.

8 Crossbreeding

Due to the increase in inbreeding coefficients in all breeds, producers are starting to consider the use of crossbreeding to avoid inbreeding depression. Agriculture Canada had a major research effort in crossbreeding during the 1970's and 1980's. The results were largely ignored because purebreeding was the 'in' thing in dairy breeding.

Norwegian Red cattle are being promoted as a breed to cross with Holsteins. Norwegian Red cattle have good production traits, few calving problems, and some resistance to diseases (at least better than Holsteins). Brown Swiss, Jersey, and Ayrshire could also be crossed to Holsteins. Producers, however, often feel that they are sacrificing too much milk yield when they use another breed. In the coming years, there will likely be an increase in the amount of crossbreeding in the dairy cattle industry. This will open up many new problems for breed associations, milk recording, and genetic evaluations. How do crossbred animals get identified so that pedigrees are stored? How do milk recording organizations treat crossbred animals in their systems? What models are needed for genetic evaluations including crossbred data? When will crossbred sires be progeny tested?

Genome Wide Selection
Fall 2008

1 The Genome

The cattle genome consists of 30 pairs of chromosome which are made of DNA. The are at least 3 billion base pairs within the DNA of those 30 chromosomes. Amino acids are coded by 3 bases, like TAA or TGC. A set of amino acids then codes for a protein or enzyme which influences activities within the body of an individual. Only about 5% of the genome actually codes for proteins and enzymes, with the remaining 95% seem to be redundant (as far as is known now). Thus, there are coding regions and non-coding regions in the genome.

From one individual to the next there are variations in the sequences of base pairs. Variations can be due to

1. A change in one base pair, where A changes to G, or G changes to C,
2. A few base pairs are missing between animals,
3. A few extra base pairs are added between two animals, or
4. The order of the base pairs can be inverted or moved to a different part of the chromosome.

Depending on the location of the variations in the genome, there could be different effects on the animal. Some variations (if they are in non-coding regions, for example) may not cause any change in the proteins and enzymes that are produced. Some variations may be in coding regions of the genome, but may still be harmless and result in no changes in functioning. Some variations could cause changes, such as in height of individuals or colour of the eyes or hair, which are also harmless. Finally, variations could be harmful and cause serious and even lethal changes in the individual due to an inability to produce the correct series of amino acids.

2 Single Nucleotide Polymorphism, SNP

The most abundant type of variation in human and cattle genomes is the single nucleotide polymorphism or SNP, where a single base pair has been changed. To be called a SNP, at least 1% of the population must have the different base change. To find SNP, one must start at one end of the genome and go through it base by base comparing between two individuals (Sequence Comparisons). SNPs are discovered by comparing individuals that

are greatly different in background - such as different breeds, or very high producers versus very low producers.

Millions of SNPs have been found in humans, and there are over 600,000 in cattle with more being discovered every day. Some of the same SNPs appear in both humans and cattle. In 2003, a company called Affymetrix (California) produced a 'chip' or 'panel' or 'array' of 10,000 SNP (from human studies). A DNA sample is put on the chip, and the genotypes of the animal for 10,000 SNP could be determined for a cost of about $350 per animal.

The Affymetrix chip, however, was designed for use with humans, and for the cattle genome, the 10,000 SNP did not fully cover the entire genome very well. In order for the SNP genotype estimates to be useful, the SNPs have to be situated about every 60,000 base pairs through out the genome. With 3 billion base pairs in total, that means a chip containing 50,000 SNPs would be needed, and it should be specifically made for cattle. This was the goal of a USDA-industry project started in 2006. The goal of the project was to discover Quantitative Trait Loci (i.e. genes) that had large, significant effects on various traits in cattle. Researchers went through all of the available known SNPs in cattle and deliberately chose which SNPs to be on the panel. The result was the Illumina 50K chip. DNA for the study was collected from semen samples from over 5,000 dairy and beef bulls from North America, including Canada.

3 Genome Wide Selection

For each SNP locus there are just 3 possible genotypes. In 2001, Meuwissen, Hayes, and Goddard published a paper that showed if the SNPs were evenly spread through the genome, then it was possible to estimate the effects of genotypes at each SNP locus on a trait of interest. The estimates could be put into a table as follows:

Genotype	Locus 1	Locus 2	Locus 3	$\cdots$	Locus n
11	0.10	3.60	10.97		-1.12
12	0.50	4.58	12.44	$\cdots$	-3.56
22	0.90	5.63	15.33		-5.87

There would be genotype estimates for every SNP locus. Thus, if a 50K chip was used, there would be 50,000 genotypes for one animal. A Genomic Estimated Breeding Value (GEBV), could be constructed from the table of genotype estimates. Suppose the genotypes of animal X were (11, 12, 22, $\cdots$, 12), then the animal's GEBV would be the sum of (0.10 + 4.58 + 15.33 + $\cdots$ -3.56) = 48.72, for example. Given the genotypes, sum the corresponding genotype estimates together for all SNP loci.

According the Meuwissen et al. (2001) the correlation between GEBV and an animal's true breeding value (TBV) would be as high as 0.85 or better. There estimate was based on simulation work in which many assumptions were made. In practice, so far, a correlation of 0.6 to 0.7 is probably the best that can be done. This is slightly more accurate than using a Parent Average EBV.

All animals with the same parents would receive the same Parent Average EBV as an estimate of their genetic merit. However, with a GEBV, each offspring would have a different GEBV because their genotypes would most likely be different. Thus, GEBV would allow the best offspring of a sire and dam to be chosen.

Since the early work of Meuwissen et al. (2001) others have proposed different methods of computing GEBV for individuals. As of August 2008 the best method has not yet been found.

An advantage of GEBV is that an animal can be genotyped at birth and a GEBV can be calculated with an acceptable accuracy. There is no need to wait until the animal is mature, or until the animal has some progeny, to select or cull that animal based on its genetic merit. The generation interval can be reduced. How this would work in dairy cattle was described by Schaeffer (2006), where genetic change could be doubled, and the cost of progeny testing could be reduced by two thirds or more. Also, fewer bulls would be needed.

Two countries have started to make use of GEBV. They are New Zealand and the Netherlands. France and Canada have been selecting bulls for progeny testing on the basis of genotypes for 14 or so markers (not 50,000). In 2009, the USDA will publish GEBV (combined with usual EBVs). Thus, the era of Genome Wide Selection is beginning. There will be significant changes in the dairy industry in the next few years because of this technology. The effect of GEBV on the increase in inbreeding will need to be monitored and controlled.

4 Future?

In humans, it is possible to have one's entire genome sequenced, so that the order of the 3 billion base pairs is known. With this information, the sequences of known genetic disorders can be "matched" to your genome to see if they are present or not. Thus, you will know which diseases you may incur in your life, and therefore, you might be able to alter your lifestyle to prevent the disease from occurring.

For livestock, the SNPs may help to discover all of the QTLs that affect economically important traits. Then chips having the QTLs rather than SNPs could be made. Accuracy of selection would be increased. GEBV will likely be used in all species of livestock with varying degrees of success.

If countries share their results on SNP genotype estimates, then genotype by environment interactions could be studied.

Everyone will be affected by this technology in the next few years.

REFERENCES

Meuwissen, T. H. E., B. J. Hayes, M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:1-6.

R Basics
Fall 2008

Double click on the R icon to get into the RGui screen and you are ready to use R.

1 Introduction

R is a programming language designed for the statistical analyses of data. The language has been developed over time by a number of different people. Anyone can contribute new packages to the language (for special analyses, for example) as long as it is well documented (there is a specific procedure for describing the package). The software is free and users agree not to create packages for sale from R. Other statistical packages, such as SAS and SPSS, are too expensive for many businesses and institutions, especially in countries outside of North America, and R is a logical alternative.

2 Object Oriented

R is an interactive language. Every command is an object and every object has some parameters that need to be given to it. Thus, the basic structure is

```
command( arg1, arg1, ... )
```

Every object has attributes, and there are commands to determine those attributes.

```
class( name of object )  = tells you the type of object

length( name of object ) = tells you the number of columns

mode( name of object )   = numeric, list, matrix, etc.
```

3 Data Frames

A Data Frame is a table of rows and columns very much like an Excel spreadsheet. Your data files need to be moved into the RGui when you need them. To do this,

```
mydframe = read.table("filepath/filename",
  header=FALSE,
  col.names=c("animal","date","gest", ...) )
```

The first row of a data frame usually contains the names of the columns. If the data that are read into R do not have column names in the first row, then `header=FALSE` and the user must specify the names of the columns, otherwise R gives them names of V1 to Vn where n is the number of columns.

If the file does contain column names in the first row, then

```
mydframe = read.table("filepath/filename",
  header=TRUE)
```

In R you can change the names of the columns with the `edit()` command.

```
xnew = edit(mydframe)
```

This gives you a spreadsheet like table. Go to the heading of each column and click on it. Then enter the name that you want. A short name of 3-5 letters that reminds you of the contents of a column is probably the most useful. The end result is saved in `xnew`.

To get a list of the names that you have used (if you forget),

```
xnewnames = names(xnew)
```

To display the names just type

```
xnewnames
```

4 summary()

The `summary()` command gives information about each column of a data frame or any vector, i.e. minimum and maximum values, mean, and standard deviation, unless the column contains characters.

If most of the columns are composed of character data, then you could do a summary of a single column of the data frame. Suppose the column of `xnew` is `bleeps`, then a summary of that column would be

```
summary(xnew$bleeps)
```

Notice that the $ is used and both the data frame name and the column name within the data frame are needed.

5 To View Selected Rows or Columns of a Data Frame

Sometimes the user likes to look at some of the data to see if it is correct. Below is an example of displaying rows 10 to 17, and a few particular columns.

```
xnew[10:17,c("animal","gest","date","yield")]
```

6 Changing a Value in a Data Frame

Suppose one of the variables in a data frame needs to be changed in one specific row. One could use the edit function, or if the row number were known, then

```
xnew$animal[irow] = newid
```

7 Histograms

Histograms are useful for visually determining the distribution shape of a data variable.

```
hist(xnew$yield)
```

Most statistical analyses assume that the observations follow a normal distribution. Generate a vector with 1000 random normal deviates, and do a histogram of that vector.

```
v = rnorm(1000)
hist(v)
```

8 Frequency Tables

One might have variables like gender and age which are categorical in nature. A frequency table of the two variables will show the number of observations in each subclass.

```
ftable(xnew$gender, xnew$age)
chisq.test(ftable(xnew$gender, xnew$age))
```

The Chi-square test can be used to test for an association between the two variables.

9 Matrix Algebra

Sometimes it is useful to perform example calculations in matrix algebra. Below are some of the simple operations.

9.1 Entering a Matrix

```
ww = matrix(data=c( 50, 6, 6.5, 6, 6, 0, 6.5, 0, 6.5),
   nrow=3,ncol=3 )
wy = matrix(data=c(251.7, 28.16, 31.09), nrow=3,
   ncol = 1 )

xt = t(x)  # to transpose the matrix x

gi = diag(c(0, 10, 10))
```

The last line creates a diagonal matrix with elements 0, 10, and 10 on the diagonals. If `G` is any square matrix, then `gd=diag(G)` gives a diagonal matrix using just the diagonals of `G`.

9.2 Generalized Inverses

A library needs to be loaded in order to get the generalized inverse function. The library call is needed only once in a session. Generalized inverses are used to get solutions to equations that are not of full rank. The function can also be used for matrices that are full rank too.

```
library(MASS)

cww = ginv(ww)

bhat = cww %*% wy
```

9.3 Trace

The trace of a square matrix is the sum of the diagonal elements.

```
k1 = sum(diag(cww %*% ww))
```

Traces are used to determine degrees of freedom for analysis of variance tables, and in estimation of variance components using the EM REML algorithm.

9.4 Cholesky Decomposition and Eigenvalues

A Cholesky decomposition is the factoring of a square, positive definite matrix into the product of a lower triangular matrix times its transpose. This is sometimes needed in the simulation of data for multiple trait models.

A canonical transformation is another decomposition of a square, positive definite matrix, e.g.

$$\mathbf{A} = \mathbf{UDU}',$$

where $\mathbf{D}$ is a diagonal matrix with the eigenvalues on the diagonal (all positive), and $\mathbf{U}$ is an orthogonal matrix such that $\mathbf{UU}' = \mathbf{I}$.

```
T = chol(A)

eee = eigen(A)
attributes(eee)
U = eee$vectors
D = diag(eee$values)
```

9.5 Block Function

Frequently matrices need to be combined as a direct sum. A function is needed that is not part of the R packages. Below is how to make a user function.

```
block = function( ... ){
  argv = list( ... ) # argv is a list of arguments
  argc = length(argv)
  i = 0
  for(a in argv) {
    m = as.matrix(a)
    if(i==0)
       rmat = m
    else
      {
        nr = dim(m)[1]  # number of rows of m
        nc = dim(m)[2]  # number of cols of m
        aa = cbind(matrix(0,nr,dim(rmat)[2]),m)
        rmat = cbind(rmat,matrix(0,dim(rmat)[1],nc))
  rmat = rbind(rmat,aa)
      }
    i = i + 1
   }
   rmat
 }

G = block(1, A, ww, wy)
```

10 Graphics

R has very good graphics capabilities to display data and results. Graphs created in R can be copied as .pdf files for incorporation into other documents. One of those was the histogram function. Below is an example of adding a title, making the columns blue, a black border, and restricting the number of bars to 5.

```
hist(xnew$yield, main='YIELD', col='blue',
    border='black', br=5)
```

More examples will be provided in the lab sessions.

Evolutionary Algorithms
Fall 2008

1 Introduction

Evolutionary algorithms (or EAs) are tools for solving complex problems. They were originally developed for engineering and chemistry problems. Much of the terminology around EAs involves genetic terminology, but the meanings are totally different from usual genetics, however, the ideas for them have come from genetics and from evolution.

EAs are for problems for which you can not *calculate* or *derive* an exact solution. The number of possible solutions is too large. An example is a fish research farm where after the fish are old enough (big enough) they need to be raised in a common pond. Fish are too difficult to identify individually. Suppose you have 25 full-sib families where the parents of each family have been genotyped for a set of 15 genetic markers. The progeny are now going to one of 5 large ponds. How do you assign families to ponds in order to maximize the probability of distinguishing individuals of each family through the genetic markers, and at the same time minimize the standard errors of estimated differences in growth and other traits among families, across ponds? One way is to enumerate all of the possible assignment of families to ponds and to compute the probabilities and standard errors of contrasts for each possibility. If the number of full-sib families was 150, the number of possible assignments becomes very large, and testing each one would take too long. In these cases, a solution has to be *found* by some other means.

Any problem where there are many components and a large number of combinations of those components exist, then an EA may be the only way to solve it in a reasonable amount of time. There are also Genetic Algorithms, Gene Expression, Genetic Programming, and Differential Evolution Algorithms which are now all under a common EA umbrella, although each is slightly different in how they work and/or the type of problems that they address.

2 EA Framework

One way to find the overall best solution to a problem is to compute all possible solutions and keep the one that is best. However, the number of possible solutions may be too large and would require years to go through each one. Additionally, you would know that a large percentage of them would not be acceptable, or maybe they are all acceptable to some degree. EAs were developed to go through the possible solutions without looking at all of them, but to find only those that give reasonably good solutions, in the least amount of time. EAs have a basic framework that is described in the following subsections.

2.1 Problem Representation

The problem has to be well defined. This includes the constraints to be imposed, the parameters that are known or needed, the equations that might describe the aspects of the problem (such as growth, or feed intake). For a given set of parameters and constraints there needs to be a way to compute a "phenotype" (which might be the profit or costs resulting from those input values, or could be the discrete allocation of animals to groups). In developing an EA, an understanding of the problem often improves, and sometimes the answers are unexpected, but correct.

2.2 Objective Function

Given the "phenotype" there needs to be a method of determining its value or "fitness" as a possible solution. Potential solutions can then be compared using the fitness values. Low fitness solutions are deleted and the higher fitness solutions are used to determine other possible solutions. The goal is to *maximize* fitness of the solutions (genotypes).

2.3 Optimization Engine

This is the algorithm by which new possible solutions (or "genotypes") are generated. Commonly, the algorithms involve "crossovers" or "recombination" in which two existing solutions are mixed. Another element is "mutation" in which a random component of an existing solution is changed to any of the other possible values for that component. The different EA algorithms utilize these processes in different ways, and in different relative frequencies. Differential Evolution involves a mixing of three different solutions into 1 new solution.

If the final best solution does not make any sense, biologically or practically, then the Problem Representation may need to be revised with additional or fewer constraints, parameters, etc. This would lead to another look at many possible solutions.

3 Example Problem

The following example is trivial, but simple enough to demonstrate the concepts of EAs. Suppose there are observations on 200 animals. The main variable, y, is the output of an animal. Output is determined by days on test, t, and amount of input, x. The formula that predicts output is

$$y \;=\; A\exp(-Bt) - \log(Cx) + \epsilon,$$

and the problem is to estimate A, B, and C which are the 3 'genes' of the genotype. The range of possible values for each can be specified. For A, it should be greater than zero and less than 500. For B and C, they should be greater than 0 and less than 1.

The data consist of values of y, t, and x for 200 animals. An example of the data are shown in the following table.

Example data for a few animals.

Animal	y	t	x
51	19	25	147
52	25	25	170
53	12	28	175
54	97	1	117
55	37	19	132
56	77	5	128

3.1 Parent Solutions

One has to decide on the parent population size, NP. Usually this is 10 to 20 possible solutions, but the user may need to try different values to see what is best for the given problem. For this example let $NP = 5$. Using a random uniform distribution variate, possible values of A, B, and C are generated, as shown in the next table.

Initial Parent Solutions.

Parent ID	A	B	C
1	200	0.5	0.8
2	19	0.4	0.1
3	48	0.3	0.6
4	120	0.2	0.7
5	31	0.1	0.05

3.2 Fitness Criterion

A fitness criterion needs to be constructed by which possible solution vectors can be ranked. In this example, the negative of sum of squares of differences between y and $\hat{y}$ can be used. The negative is used so that the fitness criterion can be maximized rather than minimized. Maximizing is a more positive attitude. So for each parent solution vector the fitness criterion is computed using the 200 animals in the data. The values are given in the following table.

Initial Parent Solutions.

Parent ID	A	B	C	Fitness
1	200	0.5	0.8	-380,011.0
2	19	0.4	0.1	-508,774.5
3	48	0.3	0.6	-448,562.4
4	120	0.2	0.7	-218,892.6
5	31	0.1	0.05	-329,392.7

Ranking the solutions on their fitness values, we get the order (4, 5, 1, 3, 2). The next step is to select the parents of the next set of solutions. Suppose that we take the top 3 solution vectors, (4, 5, 1).

3.3 Generating Progeny Solutions

Randomly choose one of the 3 selected parents as a 'template'. Then pick one of the other two selected parents as a 'mate'. Let those be vectors 1 and 4, respectively.

```
1  --------200-----0.5-----0.8-----
4  --------120-----0.2-----0.7-----
```

First, a decision needs to be made whether or not 'recombination' is going to occur. If the recombination percentage was set at 0.5, then generate a random uniform variate (between 0 and 1), and if that number is greater than 0.5, then a recombination is to be performed. Suppose the answer is yes. Then one needs to decide where the 'crossover' between 1 and 4 is going to occur. A random uniform variate can be used for that decision too. Suppose the answer is a break between the first and second genes. Then the new progeny solution is

```
6  --------200-----0.2-----0.7-----
```

The A allele came from parent 1, and the B and C alleles came from parent 4.

If the answer to the recombination query was no, then progeny 6 would be equivalent to parent 1. Parent 1 would be carried over to the next generation.

There could also be a mutation. Suppose parent 4 was chosen as the template (instead of 1 above), and suppose recombination answer was no, so that parent 4 alleles would be carried over. However, a mutation might occur at one loci. A random uniform variate is chosen, and if it is less than the mutation rate, then a mutation occurs. The mutation rate is generally low, say 0.10. If the answer is yes to mutation, then another random number is chosen to decide which loci is affected, and then the allele is replaced by another random value. Let the mutation occur in the B loci, and instead of being 0.2, a new value of 0.6 is given, then the new progeny solution is

```
7  --------120-----0.6-----0.7-----
```

NP new progeny solution vectors are generated from the 3 selected parents in the above manner. Their genotypes are shown below with their corresponding fitness values.

First Progeny Solutions.

Progeny ID	A	B	C	Fitness
6	200	0.2	0.7	-193,472.9
7	120	0.6	0.7	-449,234.2
8	31	0.1	0.8	-370,516.3
9	120	0.2	0.05	-189,019.7
10	200	0.1	0.8	-183,789.4

Notice that the average fitness value has gone up compared to the parent generation, and this is due to the selection of parents.

3.4 More Generations

After NP new solutions are generated, then the best 3 are selected to be parents of the next generation. This process is repeated for many thousands of generations. The solutions will evolve towards the best values of A, B, and C that satisfy the objective function, which is to fit the data with the proposed formula.

The values used to generate the data on 200 animals were $A = 104$, $B = 0.06$, and $C = 0.13$. The final solutions should be close to these values, but not exactly because there was some random residual variation (normal distribution) added to form y values with a mean of 0 and SD of 10. The x variable was also normally distributed with mean 300 and SD of 20. t were random numbers between 1 to 30.

4 Differential Evolution

The previous section described a genetic algorithm (GA) involving selection, recombination, and mutation. Usual GAs may take many thousands, if not, millions of generations to evolve to the final solution. Depending on the objective function and the number of data records, each generation could take a long time to compute. People that use EAs are generally in a hurry to find a solution. Thus, the Differential Evolution (DE) algorithm was developed, and with this algorithm solutions tend to evolve significantly more quickly.

Consider the example of the previous section, and the parent generation of solutions. The same fitness criterion is used.

Initial Parent Solutions.

Parent ID	A	B	C	Fitness
1	200	0.5	0.8	-380,011.0
2	19	0.4	0.1	-508,774.5
3	48	0.3	0.6	-448,562.4
4	120	0.2	0.7	-218,892.6
5	31	0.1	0.05	-329,392.7

In DE, in each generation, each parent solution is visited one at a time, in random order. Begin at parent 1. and randomly pick 3 out of the other 4 solutions, say 2, 4, and 5 for i, j, and k, respectively. Go through the three loci one at a time. For each loci, pick a random uniform variate. If the value is above 0.5 then the allele for a progeny is equal to the parent allele. If the value is below 0.5, then the progeny allele is set equal to

$$parent(i) \ + \ Factor * (parent(j) \ - \ parent(k)),$$

where i, j, and k are three random parent IDs (not equal to the parent ID being changed at the moment, so not equal to 1). For the A allele, for example, with $i = 2$, $j = 4$, and $k = 5$, then the new A allele would be

$$19 \ + \ 0.5 * (120 \ - \ 31) \ = \ 63.5.$$

The $Factor$ is usually equal to 0.5, but this number can be revised to get better mixing of solution vectors. A new set of i, j, and k are chosen for each loci.

Thus, for some alleles, the parent allele will carry through to the progeny, and for others, a new allele is generated from the existing solutions. Suppose the new progeny alleles are

```
6  --------63.5-----0.5-----0.1-----
```

Additionally, mutations can affect a loci with a certain percentage probability. The mutation can be a completely new random possibile value for that allele, or can be an average of a new random possible value with the value of the current best allele value (best is the most fit solution vector).

After the genotype of the new solution is set, then the fitness criterion is computed. If the fitness value of the progeny is greater than or equal to the fitness value of the parent, then the progeny genotype replaces the parent genotype in the solutions. Two solutions may have the same fitness value, but the genotypes could be different. Thus, if the progeny and parent have the same fitness value, replacing the parent with the progeny could introduce a new genotype to the set of parent solutions.

The process is repeated for the next parent ID, and again many generations are conducted. As can be seen, the DE algorithm provides a better mixing of alleles, and only progeny with better fitness are kept. Recombinations are replaced with the 'difference' function involving 3 different parent solutions. Thus, there has to be at least 4 parent solutions to run a DE algorithm. Ten to twenty parents are usually sufficient. This will depend on the number of loci in the genotypes. DE generally converges faster than GA towards the best solution, but a large number of generations may still be needed.

5 Global versus Local

EAs can easily converge towards a 'local' maximum rather than a 'global' maximum. The 'global' maximum, is the single, best solution possible. If you think of a mountain landscape, the global maximum is the mountain with the highest peak. All other mountain peaks are local maxima. If you start climbing one mountain, the EA algorithm may take you to the top, but once you are there you can see that there is another mountain with a higher peak. If the landscape is very mountainous, then a higher mutation rate may be needed to get away from a mountain with a local maximum. If the landscape is smooth with rolling hills, then maybe the mutation rate has to be lower. The user must be able to determine the type of landscape with which they are exploring, and adjust mutation rates accordingly.

Another method is to re-start the EA with very different initial parent solutions and see if the EA converge to the same or different maximum. The user must be aware and concerned about local versus global maxima.

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 1. Data, R, Matrix Algebra

1. **Data and R**

   (a) Enter the data in the table below into a data frame called "beef".

Calf	Breed	Sex	CE	BW(lbs)
1	AN	M	U	55
2	CH	M	E	68
3	HE	M	U	60
4	AN	M	U	52
5	SM	F	H	65
6	HE	F	E	64
7	CH	F	H	70
8	LM	F	E	61
9	SM	F	E	63
10	CH	M	C	75

   (b) Create design matrices for breed, sex, and CE.

   (c) Compute the mean BW by breed, sex, and CE.

   (d) Plot the data frame, i.e. plot(beef)

2. **Reading Outside Data in R**

   Go to (http://www.aps.uoguelph.ca/ lrs/ABMethods/DATA/). Copy the file "lab02.d" and store in the R subdirectory.

   Read the data into R, as follows:

```
zz = file.choose() #allows you to browse
trot = read.table(file=zz, header=FALSE, col.names=
 c("race","year","month","track","distance","condns",
 "driver","horse","age","sex","speed"))
```

   Answer the following questions:

   (a) How many records are in this data file?

   (b) How many horses are represented in the data?

   (c) How many drivers are represented in the data?

(d) What years were covered by the data?

(e) What was the age distribution by age? Can you represent it in a histogram?

(f) What was the mean and variance of speed?

3. **Matrix Algebra and R**

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 & 3 \\ -1 & 2 & 0 & -2 \\ 4 & 1 & -2 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 4 & 3 & -2 & 0 \\ -1 & 0 & 1 & 1 \\ 2 & -3 & -4 & 1 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 2 & -2 \\ -1 & 3 \\ 4 & 1 \\ 0 & 5 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} -2 & 12 \\ -4 & -2 \\ -1 & 3 \end{pmatrix}.$$

Perform the following operations in R, if they are conformable.

(a) $(\mathbf{A} * \mathbf{C}) - \mathbf{D}$

(b) $\mathbf{C}' * \mathbf{B}'$

(c) $\mathbf{A} - \mathbf{B}$

(d) $\mathbf{D} - \mathbf{C}$

(e) $(\mathbf{A} * \mathbf{B}')^{-1}$

(f) $\mathbf{D} * \mathbf{C}'$

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 2. Models, AOV and Dairy Facts

1. **Writing a Model**

   Daily feed intake (DFI) records of 653 barrows and gilts during a growing period from 27 to 108 kg live weight were available. Pigs were Yorkshire by Landrace crossbreds born from 1976 to 1982. Pigs are born in litters of various sizes, and litter size is known to affect growth. Age at start of the growing period was known. Feed intake will increase as the pigs grow. Write a model to analyze DFI and to genetically evaluate pigs for DFI.

2. **Analysis of Variance**

   Retrieve data for this question, as follows:

   - Go to (http://www.aps.uoguelph.ca/∼lrs/ABMethods/DATA).
   - Click on "dairy.d" and copy the file to your computer and save somewhere (in your R subdirectory).
   - Use the following R statements to read it.

```
zz = file.choose()
dairy = read.table(file=zz,header=FALSE,col.names=c("herd",
 "aid","birth","calve","parity","sexc","ease",
 "servs","dopen","milk"))
```

   The main model of analysis will be

```
milk = herd-year-season  + parity + ease + e
```

   where 'herd-year-season' is the herd, year, and season of calving; 'parity' is the number of calvings a cow has had; 'ease' is the calving ease category (unassisted, easy pull, hard pull, or caesarian); and 'milk' is the amount of milk given in that lactation over 305 days. Seasons will be defined four different ways.

   - Model 1: Each month (12) of calving will be a season.
   - Model 2: Every 2 months will be a season (6 of them).
   - Model 3: Every 3 months will be a season (4 of them).
   - Model 4: Every 4 months will be a season (3 of them).

   Do an analysis of variance on each model and determine which fits the data the best.

3. **Dairy Facts**

(a) List the seven (7) main dairy breeds in Canada.

(b) What functions do the following organizations perform?

i. Holstein Canada
ii. Canadian Dairy Network
iii. CanWest DHI
iv. ICAR
v. Interbull

(c) What are average values (in Holsteins) for

i. Age at first calving
ii. Gestation length
iii. Calving interval
iv. 305-d milk yield
v. 305-d fat yield
vi. 305-d protein yield

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 3. Genetic Relationships, Beef Facts

1. **The Tabular Method**

   (a) Calculate, by hand, the numerator additive relationship matrix for the following pedigrees:

Animal	Sire	Dam
A		
B		
C	A	B
D	A	C
E	D	B
F	A	B
G	E	F
H	A	C
J	G	H

   (b) Calculate, by hand, the b_i values for each animal in question 1, and write the inverse of the additive relationship matrix.

   (c) Animals C and F are full-sibs, such that $a_{CF} = 0.5$ and $d_{CF} = 0.25$. The variances of various gene interactions are $\sigma^2_{10} = 1600$, $\sigma^2_{01} = 1000$, $\sigma^2_{11} = 400$, $\sigma^2_{20} = 600$, and $\sigma^2_{02} = 200$, then calculate the genetic covariance between full-sibs.

2. **Using R routines for inbreeding coefficients.**

   Retrieve data for this question.

   Go to (http://www.aps.uoguelph.ca/~lrs/ABMethods/DATA). Click on "dped.d" and copy the file to your computer and save somewhere.

   Use the following R statements to read "dped.d".

```
zz = file.choose()
peds = read.table(file=zz,header=FALSE,col.names=c("anim",
  "sire","dam","birth","sex"))
```

   Calculate the inbreeding coefficients using the routines described in lecture. Average the inbreeding coefficients by year of birth, and plot the trend. Hand in a copy of this graph with your assignment.

3. **Facts about beef cattle.**

(a) List the five most numerous purebreds in Canada and give their relative numbers of animals born or registered per year.

(b) What are the traits of economic importance to beef cattle producers.

(c) How often and when are beef animals weighed?

(d) What is a station test?

(e) What organization computes genetic evaluations for beef cattle in Ontario?

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 4. Mixed Model Equations, Swine Facts

1. **Solving Mixed Model Equations**

   Below are data on seven animals from two contemporary groups with their pedigrees, giving a total of 12 animals. Apply the following animal model to this example.

   $$y_{ijk} = \mu + CG_j + a_k + e_{ijk},$$

   where μ is the overall mean, CG_j is a contemporary group effect (groups of animals raised under similar conditions), a_k is the animal additive genetic effect, and e_{ijk} is a residual effect.

Animal	Sire	Dam	b_i	CG_j	Obs.
1			1		
2			1		
3			1		
4			1		
5			1		
6	1	5	0.5	1	25
7	1	4	0.5	2	55
8	2	4	0.5	1	46
9	2	6	0.5	2	32
10	3	5	0.5	1	13
11	3	6	0.5	2	28
12	3	7	0.5	2	43

   Let the assumed variances be

   $$\begin{aligned} \sigma^2_{cg} &= 100, \\ \sigma^2_a &= 400, \\ \sigma^2_e &= 800. \end{aligned}$$

   (a) What is the heritability of the trait?

   (b) Construct $\mathbf{A}^{-1}$ using the Ainv function given in class.

   (c) Construct the mixed model equations - use the MME function.

   (d) Obtain a solution vector.

   (e) Estimate the residual variance.

   (f) Calculate the reliability and SEP of the animal EBVs.

(g) Define the genetic base as the average of all animals with records. Express all EBV relative to this genetic base.

(h) Using the same genetic base as in the previous question, express the EBV as relative EBVs (with a mean of 100).

2. **Swine Facts**

(a) What are the functions of the Canadian Centre for Swine Improvement?

(b) What breeds are important in the Canadian swine industry?

(c) What traits are of economic importance?

(d) How many chromosomes are there in swine?

(e) What is the gestation length of a sow?

(f) At what age or weight are pigs marketed?

(g) At what age are males and females sexually mature?

(h) What is the role of crossbreeding in swine?

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 5. Simulation of Data, Sheep Facts

1. **Simulation of Data**

   Simulate data according to a repeated records, animal model using the following specifications.

Number of animals(See Table)	15
Number of records per animal	1 to 3
Number of contemporary groups	3 (Means = 111, 87, 123)
Additive genetic variance	49
Permanent environmental variance	25
Residual variance	81

Pedigree Information. X indicates animal has a record in that Contemporary Group.

Animal	Sire	Dam	b_i	Contemp. Groups		
				1	2	3
1	-	-	1			
2	-	-	1			
3	-	-	1	X	X	X
4	-	-	1	X	X	
5	1	3	.5	X	X	
6	1	4	.5	X		
7	2	3	.5	X	X	X
8	2	4	.5	X		X
9	2	6	.5		X	X
10	2	6	.5		X	
11	1	7	.5		X	X
12	1	8	.5			X
13	5	8	.5			X
14	1	3	.5			X
15	2	4	.5			X

   (a) What are the heritability and repeatability of the trait.

   (b) Analyze the data with the appropriate model to obtain EBVs for each animal from MME.

   (c) Correlate the EBVs with the true breeding values.

   (d) Correlate the estimates of PE effects with their true values.

   (e) How do the estimates of CG effects compare to the values you used to simulate the data?

(f) Compare your correlation results to those of two other students.

2. **Sheep Facts**

(a) What are the physical differences between Suffolk, Dorset, Rideau Arcot, and Polypay breeds of sheep?

(b) Generation intervals are defined as the average age of the sires (and average age of the dam) when a replacement progeny is born. What is the average age of a ram when a male progeny that will replace the ram is born? Age of the ewe when that same male progeny is born?

(c) What is the number of chromosomes in sheep?

(d) What is the main sheep breed in New Zealand?

(e) How many sheep flocks are there in Ontario? What is the average number of ewes per flock?

(f) At what age and weight are lambs weaned? marketed?

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 6. Maternal Genetic, Random Regression, and Horse Facts

1. **Maternal Genetics Model**

Below are weaning weight (WW) data on calves of one beef breed.

Animal	Sire	Dam	Year	Age of Dam	WW(lbs)
8	1	3	2006	3	73
9	1	4	2006	3	98
10	2	5	2006	2	65
11	2	6	2006	3	87
12	2	4	2007	4	94
13	1	3	2007	4	71
14	2	5	2007	5	86
15	1	7	2007	4	79

Apply the following maternal effects model to the data.

$$y_{ijkl} = Y_i + A_j + a_k + m_l + p_l + e_{ijkl}$$

where

$$\begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} 55 & -10 \\ -10 & 25 \end{pmatrix},$$

$\sigma_p^2 = 11$, and $\sigma_e^2 = 220$.

(a) Set up $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{G}$.

(b) Construct the MME.

(c) Solve the MME.

(d) Rank the calves for direct weaning weight.

(e) Rank the dams for maternal genetic ability.

(f) Rank the dams for maternal most probably producing ability.

2. **Random Regression Model**

Below are milk yield data on goats at different days in milk. Assume the goats are not related.

Goat	HTD 1		HTD 2		HTD 3	
1	10	3.41	30	3.71	54	3.77
2	22	5.80	42	6.06	66	5.94
3	45	3.34	65	3.34	89	3.29
Goat	HTD 4		HTD 5			
1	77	3.73	97	3.67		
2	89	5.68	109	5.17		
3	112	3.23	132	2.94		

Let the model be

$$y_{im} = (b_0 + b_1 d + b_2 c) + (a_{i0} + a_{i1} d + a_{i2} c) + e_{im:t}$$

where d is days in milk, and c is $\exp^{-0.05d}$. Assume that

$$\mathbf{G} = \begin{pmatrix} 4 & -0.07 & 0.26 \\ -0.07 & .002 & -0.003 \\ 0.26 & -0.003 & 0.05 \end{pmatrix},$$

and $\sigma_e^2 = 2$.

(a) Set up $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{G}$.

(b) Construct the MME.

(c) Solve the MME.

(d) Rank the goats for yield at 25 days in milk.

(e) Rank the goats for yield at 100 days in milk.

(f) Which goat is most persistent from 25 to 100 days?

3. **Horse Facts**

(a) Distinguish between Thoroughbreds, Standard Breds, Quarter Horses, and Warmbloods.

(b) List traits and their approximate heritabilities that could be important in Warmbloods.

(c) Give the number of chromosomes, the average gestation length, and the average age at first breeding for stallions and mares.

(d) What countries are part of InterStallion?

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 7. Selection and Genetic Change

1. Simulate a trait for a population of 20,000 animals with a mean of 0 and a standard deviation of 10. Order the animals from highest to lowest, then calculate the mean and variance of the phenotypes for all animals, the top 90%, top 80%, etc. down to the top 10%, top 5%, and top 1%. Plot the results on a graph.

2. Repeat the previous question using a population of only 1,000 animals. How do the results compare? What is the effect of sample size?

3. **Swine Production Systems: Selection on Phenotypes** A swine breeder has a herd of 160 sows. Each sow is kept for four litters which begins when the sow is one year of age, and every 6 months thereafter. There are roughly 40 sows at each age (1, 1.5, 2, and 2.5 years). On average 10 piglets are weaned per litter.

   Production Cycle: Farrowing is continuous through the year with about 13 sows farrowing per month. Piglets are weaned at 4 weeks of age, and are transfered to a growing facility and raised to a market weight of approximately 100 kg.

   Sow Replacements: Sows are culled after their fourth litter. Approximately 67 new females are produced every month from the growing-finishing facility. Three or four of the fastest growing females are kept as replacements, with the restriction that each must be from a different litter.

   Boar Replacements: Boars can be used for breeding at 8 months of age. About 67 males are produced each month, of which 26 (2 from each litter) are performance tested for AGE at 100 kg, and the others are castrated and sent to the growing-finishing facility. Two boars per month are selected based on AGE at 100 kg. Boars are kept for only one year.

   Matings: Matings are random (boars to sows) except that a boar is never mated to a half-sib or full-sib female to avoid inbreeding.

   The Trait: AGE at 100 kg has a heritability of 0.32, and a genetic standard deviation of 6.0 days. The accuracy, (r_{TI}), of selecting animals based on their own growth record is equal to 0.30.

   Utilize the four pathways of selection formula, which is

$$\frac{\Delta \mathrm{G}}{\text{year}} = \frac{\Delta_{SM} + \Delta_{SF} + \Delta_{DM} + \Delta_{DF}}{L_{SM} + L_{SF} + L_{DM} + L_{DF}},$$

   where each $\Delta_{ij} = r_{TIij} i_{ij} \sigma_a$. Determine the expected genetic change under this system of selection.

4. **Selection ob EBVs** Suppose the swine breeder calculates EBVs for AGE at 100 kg, on all animals, and that EBVs are computed monthly. The average accuracy of evaluation of sows (with litters) increases to 0.50, and for boars (with matings) goes to 0.65. Female and male piglets have an increased accuracy of 0.40. The selection program is changed as follows:

   Sow Replacements: Dams and sires of replacement females must have an EBV above the average of the herd.

   Boar Replacements: Boars are selected based on their EBV for AGE at 100 kg.

   Determine the expected genetic change under this system of selection, and compare (graphically or in a table) to the results from the previous question.

5. What will be the expected change in inbreeding coefficients in this closed herd system?

Selection Differentials, i

	For .001 to .099 selected									
	.000	.001	.002	.003	.004	.005	.006	.007	.008	.009
.00		3.400	3.200	3.033	2.975	2.900	2.850	2.800	2.738	2.706
.01	2.660	2.636	2.600	2.569	2.550	2.527	2.500	2.582	2.456	2.442
.02	2.420	2.400	2.386	2.370	2.363	2.336	2.323	2.311	2.293	2.283
.03	2.270	2.258	2.241	2.230	2.221	2.209	2.200	2.186	2.174	2.164
.04	2.153	2.146	2.136	2.126	2.116	2.107	2.098	2.087	2.079	2.071
.05	2.064	2.057	2.048	2.040	2.031	2.022	2.016	2.009	2.000	1.990
.06	1.985	1.977	1.971	1.965	1.958	1.951	1.944	1.937	1.931	1.925
.07	1.919	1.911	1.906	1.900	1.893	1.888	1.882	1.875	1.871	1.863
.08	1.858	1.852	1.846	1.841	1.837	1.834	1.826	1.820	1.815	1.810
.09	1.806	1.799	1.793	1.788	1.784	1.780	1.775	1.770	1.765	1.760
	For .01 to .99 selected									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.10	1.755	1.709	1.667	1.628	1.590	1.554	1.521	1.488	1.458	1.428
.20	1.400	1.372	1.346	1.320	1.295	1.271	1.248	1.225	1.202	1.180
.30	1.159	1.138	1.118	1.097	1.078	1.058	1.039	1.021	1.002	.984
.40	.966	.948	.931	.913	.896	.880	.863	.846	.830	.814
.50	.798	.782	.766	.751	.735	.720	.704	.689	.674	.659
.60	.644	.629	.614	.599	.585	.570	.555	.540	.526	.511
.70	.497	.482	.468	.453	.438	.424	.409	.394	.380	.365
.80	.350	.335	.320	.305	.290	.274	.259	.243	.227	.211
.90	.195	.179	.162	.144	.127	.109	.090	.070	.049	.027

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 8. Correlated Responses

1. Estimated breeding values (EBVs) for 10,000 animals and 9 traits per animal have been prepared. Go to the ABMethods/DATA/ site to retrieve "sheep.RData". In R you would use "zz=file.choose()" to locate where you have stored the file, and then use "load(zz)" to bring it into your R session. This file contains the data called "tbv", the genetic covariance matrix for the 9 sheep traits, and the index weights for 4 different selection indices (shown below). The parameters in "VG" are for traits on Dorset sheep, lamb survival - direct and maternal, birth weights - direct and maternal, 50-day weights - direct and maternal, gain from 50 to 100 days, loin thickness, and fat thickness.

2. Apply each of the 4 indices (one at a time) to the 10,000 animals. Rank the animals on their index values. Compute the mean genetic values for all traits for the top 250 animals. Put the results into a table (on the next page).

Economic Weights on each Trait for 4 Indexes

Trait	Index A	Index B	Index C	Index D
Lamb Survival, Direct	0	0	2.85	0
Lamb Survival, Maternal	0	0	-8.33	0
Birthweight, Direct	0	-1.936	-8.94	0
Birthweight, Maternal	0	0	-33.84	0
50-d weight, Direct	1	1	1.75	1
50-d weight, Maternal	0	0	5.23	0
Gain 50-100d	0.61	0.61	0.19	0
Loin Thickness	0.686	0.686	-0.08	0
Fat Thickness	0	-2.626	-0.51	0

3. Which index gives the greatest favourable change in lamb survival, direct?

4. What are the relative emphases of traits in index C - compare everything to 50-d weight, direct effects?

5. Which index gives the greatest correlated response in maternal ability for 50-day weights?

6. Assume you are the owner of a sheep flock. List the nine traits in order of economic importance to you, from most to least important. Make an index (i.e. derive the weights) that will reflect your list. Apply your index and compare to the other four indices.

Means of Top 250 Animal EBVs

Trait	Index A—	Index B—	Index C—	Index D—	Index E—
Lamb Survival, Direct					
Lamb Survival, Maternal					
Birthweight, Direct					
Birthweight, Maternal					
50-d weight, Direct					
50-d weight, Maternal					
Gain 50-100d					
Loin Thickness					
Fat Thickness					

MBG*4030 - Animal Breeding Methods - Fall 2008

Lab 9. Crossbreeding in Beef Cattle

Below are values for conception rate (CR), calf survival (SV) (includes calving ease), and calf direct genetic weaning weight (WW) for five breeds of cattle. Assume that heterosis values for these traits are 0.10, 0.10, and 0.05, respectively.

Breed Parameters.

Breed	CR	SV	WW
A	0.60	0.75	244
L	0.64	0.85	232
C	0.70	0.80	226
H	0.56	0.80	240
S	0.60	0.70	250

The variable used to compare possible breed crosses is the expected kilograms of calf (EKC) per cow bred given by

$$EKC = CR \times SV \times WW.$$

The CR is that of the cow, and SV and WW are on the calf.

1. Compare all single crosses, and determine the best one.
2. Pick one cross as the maternal line for a two breed rotation. Pick another cross as the better growth breeds for a second two breed rotation. Determine the outcome for a terminal cross using males from the growth two breed rotation on females from the maternal two breed rotation.
3. Determine the best three breed rotational system and show the EKC per year of this system after it has settled to its average heterosis expression of 86%.
4. What is the amount of retained heterosis in a cross of $(\frac{1}{2}C \times (\frac{3}{8}A \times \frac{1}{8}S))$ males with $(\frac{1}{2}A \times (\frac{1}{4}L \times \frac{1}{4}H))$ females? Calculate the EKC for this cross.

MBG*4030
Animal Breeding Methods

Fall 2006

Midterm Exam

1. **(8 points):** A friend of yours, Sam, decides to raise dairy goats (Alpines) for milk production. After extensive reading and six months of experience, Sam has learned the following facts about dairy goats.

   - Goats are seasonal breeders, usually from August to mid-February.
   - Does kid between December and July, mostly February to May.
   - Gestation length is 5 months.
   - Two or three offspring are born per doe per gestation period.
   - Age at first breeding is between 7 to 12 months.
   - A mature doe gives about 900 kg of milk.

   Sam has a new deluxe computer with R installed, and wants you to design a genetic evaluation program for his does. Sam proposes the following model equation:

   $$\begin{aligned} y &= \text{First lactation milk yield (kg)}, \\ &= \mu + \text{Year of kidding effect}, \\ &\quad + \text{Age at kidding effect}, \\ &\quad + \text{Animal additive genetic effect}, \\ &\quad + \text{Residual effect}. \end{aligned}$$

   (a) What other factors(if any) would you add to this model?

   Length of lactation, breed of goat, interactions between lactation length, breed, and age at kidding. Maybe change Year of kidding to Year-Month of Kidding.

   (b) What factors (if any) would you take out?

   None to take out.

   (c) Complete Part 3 of the model(Assumptions and Limitations).

   - Breed effects are not important.
   - Lactation lengths are all very similar.
   - All animals are healthy.
   - No preferential treatment given to certain animals.
   - Pedigrees are known.
   - Heritability is about 0.35.

2. **(4 points):** Below are some data from Sam's dairy goat farm.

Doe	Year of Kidding	Age at Kidding (mo)	Age at Kidding Group	Milk Yield (kg)
1	2005	17	1	811
2	2005	18	2	792
3	2005	20	3	739
4	2006	16	1	832
5	2006	19	2	779
6	2006	18	2	873
7	2006	21	3	845
8	2006	19	2	903

Write the design matrices for Year of Kidding and Age at Kidding Group.

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ \vdots & & etc & & \end{pmatrix}.$$

3. **(2 points):** If

$$\mathbf{S} = \begin{pmatrix} 3 & -6 & 2 & -1 \\ -1 & 2 & -6 & 3 \end{pmatrix}, \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} 7 \\ 3 \\ -1 \\ 1 \end{pmatrix},$$

then calculate

$$\mathbf{Sw} = \begin{pmatrix} 0 \\ 8 \end{pmatrix}.$$

4. **(2 points):** The R^2 for fitting a particular model for the dairy goat data was 0.57. What does this R^2 value mean?

   Anything above 0.50 is a pretty good model, but could be made better possibly.

5. **(6 points):** The total phenotypic variance of a trait is 400. If the heritability is 0.25 and the repeatability is 0.35, then calculate the values of the additive genetic, permanent environmental, and residual variances.

$$h^2 = \frac{\sigma_a^2}{\sigma_y^2}, \text{ and } \quad r = \frac{\sigma_a^2 + \sigma_{pe}^2}{\sigma_y^2}.$$

$$\begin{aligned} \sigma_a^2 &= 100. \\ \sigma_a^2 + \sigma_{pe}^2 &= 140, \\ \sigma_{pe}^2 &= 40, \\ \sigma_e^2 &= 400 - 100 - 40 = 260. \end{aligned}$$

6. **(10 points):** Complete the calculations for the following genomic relationship matrix.

	Am	Af	Bm	Bf	A Cm	B Cf	C Dm	B Df
Am	1	0	0	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0
Af	0	1	0	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0
Bm	0	0	1	0	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
Bf	0	0	0	1	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
Cm	$\frac{1}{2}$	$\frac{1}{2}$	0	0	1	0	$\frac{1}{2}$	0
Cf	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{2}$	$\frac{1}{2}$
Dm	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{4}$
Df	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{4}$	1

Calculate $a_{CD} = \frac{3}{4}$ and $d_{CD} = \frac{1}{4}$.

What is the inbreeding level of animal D, and what does it mean?

The inbreeding coefficient is 0.25 and means that 25% of loci have alleles that are identical by descent.

7. **(2 points):** What is an additive by dominance gene interaction?

   This is the interaction of an allele at one gene locus with the pair of alleles at another gene locus.

8. **(4 points):** Give two important contributions of C. R. Henderson to animal breeding.

   BLUP and Mixed Model Equations, and calculation of A-inverse.

9. **(10 points):** Below are some R commands that we have used in the labs.

dim(dataframe)	tapply(ymilk,yrmnfac,mean)	as.numeric(ccg)
factor(yrmn)	c(1:m)	rep(0,m)
ls()	rm()	length()
ginv()		

   Which function would you use

   (a) to solve a system of equations? ginv(), or solve()

   (b) to force a variable to be numeric? as.numeric()

   (c) to prepare a variable for the lm() function? factor()

   (d) to remove a variable from your workspace? rm()

   (e) to find the number of items or records in an array or data frame? dim()

10. **(12 points):** Given the following pedigrees and b_i values for each animal, fill in the elements of $\mathbf{A}^{-1}$ for animals X and W. Remember that $\delta = 1/b_i$ and

    Pedigrees are

Animal	Sire	Dam	b_i
X	S	D	3/8
W	S	G	1/2

    The rules are

	animal	sire	dam
animal	δ	-0.5δ	-0.5δ
sire	-0.5δ	0.25δ	0.25δ
dam	-0.5δ	0.25δ	0.25δ

	—-S—-	—-D—-	—-G—-	–S—D– —-X—-	–S—G– —-W—-
S	$.5 + \frac{2}{3}$	$\frac{2}{3}$	.5	$-\frac{4}{3}$	-1
D	$\frac{2}{3}$	$\frac{2}{3}$		$-\frac{4}{3}$	
G	.5		.5		-1
X	$-\frac{4}{3}$	$-\frac{4}{3}$		$\frac{8}{3}$	
W	-1		-1		2

11. **(6 points):** Order the animals in the pedigrees below such that parents appear in the list before their progeny.

Animal	Sire	Dam	Work vector	New Order
909	101	202	3	101
101			4	202
202			4	909
606	1010	303	2	1010
1010			3	303
303			3	1111
404	101	1111	2	1212
1111			3	606
505	909	1212	2	404
1212			3	505
707	606	202	1	707
808	404	505	1	808

12. **Definitions (2 points each):**

    (a) Phantom groups: Needed to account for missing pedigrees. All missing parents are not necessarily from the same base generation.

    (b) BLUP: best linear unbiased prediction

    (c) Identical by descent (IBD): Alleles are the same because they descend from a common ancestor.

    (d) Parturition: Giving birth.

(e) Infinitesimal Model: Infinite number of loci affecting quantitative traits, all with equal and small effects. Only additive genetic effects considered, and the population is randomly mating.

(f) Floating Base System: Definition of the genetic base for comparison of EBVs is changed from one run to the next, or at least annually.

(g) Permanent Environmental Effects: Non-genetic effects associated with an animal that affects all of that animal's records for a trait.

MBG*4030
Animal Breeding Methods
Fall 2006
Final Exam

1. **Definition of Terms (2 points each):** Give one major difference between the following pairs of terms.

   (a) (Candidate Gene Approach) vs (QTL Detection)

   (b) (Daughter Design) vs (Granddaughter Design)

   (c) (MAS) vs (Genome Wide Selection)

   (d) (Aggregate Genotype) vs (Selection Index)

   (e) (Haplotype) vs (Genotype)

2. **(2 points)** The Lifetime Profit Index (LPI) used in Canada for ranking dairy bulls consists of three components (i.e. Production, Durability, and Health/Fertility). What components would you put into an LPI for swine?

3. **(2 points)** How do dominance genetic effects explain the existence of heterosis in cross bred animals?

4. **(2 points)** Which of the following two crosses has the most retained heterosis?

   Cross 1: ((A x B) x (C x D)) male with (A x D) female

   Cross 2: ((A x B) x C) male with B female

5. **(2 points)** What kind of model would you use to evaluate calving ease?

6. **(2 points)** Why is crossbreeding more popular with beef and swine breeders than with dairy cattle breeders, in Canada?

7. **(6 points)** List 3 facts about the Canadian Test Day Model.

   - Fact 1
   - Fact 2
   - Fact 3

8. **(2 points)** Give an advantage and a disadvantage of a terminal cross versus a 3 breed rotational cross system.

9. **(4 points)** Part A. Assume heterosis of 5%, and that breed A has a conception rate of 76% and breed B has a conception rate of 80%. Calculate the expected conception rate of an (A x B) crossbred female.

   Part B. Calculate the conception rate of a backcross progeny when the (A x B) female is bred back to a B male.

10. **(2 points)** A purebred sheep breeder must decide among three rams to breed to his ewes. Ram A has a Combined (polygenic and QTL) EBV of +6 kg for lamb weaning weight (WW), while Ram B also has a Combined EBV of +6 kg for WW. The difference between the rams is that ram A has only one favourable QTL allele while Ram B has two favourable QTL alleles. Ram C has a combined EBV of +8 and no favourable QTL alleles. Below is a table of additive genetic relationships of the 3 rams with three of his ewes. The ewes have been genotyped for the QTL (a Q indicates one favourable allele and QQ indicates two favourable alleles) and EBVs are given in the table. Recommend a ram for each ewe, and give a reasons for your choices.

Ewe	Genotype	EBV	Ram A	Ram B	Ram C	Recommend	Reason
1	Q	+1	.08	.08	.12		
2	QQ	-3	.25	0	.06		
3		-2	.06	.10	0		

11. **(2 points)** A three trait index is proposed.

$$I = b_1(EBV_1) \ + \ b_2(EBV_2) \ + \ b_3(EBV_3),$$

where $b_1 = 9$, $b_2 = 4$ and $b_3 = 6$.

$$\begin{pmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} & \sigma_{a_1a_3} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} & \sigma_{a_2a_3} \\ \sigma_{a_1a_3} & \sigma_{a_2a_3} & \sigma^2_{a_3} \end{pmatrix} = \begin{pmatrix} 16 & 5 & -1 \\ 5 & 81 & -2 \\ -1 & -2 & 144 \end{pmatrix}.$$

What are the relative weights of traits 2 and 3 compared to trait 1.

12. **(6 points)** Let

$$I = 1(BV_1) - 1(BV_2),$$

and

$$\begin{pmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} \\ \sigma_{a_1a_2} & \sigma^2_{a_2} \end{pmatrix} = \begin{pmatrix} 23 & -2 \\ -2 & 14 \end{pmatrix}.$$

(a) Trait 1 is in *kg* and trait 2 is in *cm*. Calculate the covariances of the traits with the index (keep the units straight).

(b) Calculate the variance of the index.

(c) If $\Delta I = 10.24$ \$, then calculate the correlated responses in traits 1 and 2 as a result of selection on this index.

13. **(2 points)** The Haldane mapping function gave the distance between the G and D loci as $m_{GD} = .30cM$. What would be the distance between G and D using the Kosambi mapping function? Let r_{GD} be the recombination rate between the G and D loci.

$$Haldane = -.5\log(1 - 2r_{GD}),$$

and

$$Kosambi = 0.25\log[(1 + 2r_{GD})/(1 - 2r_{GD})].$$

14. **(12 points)** The methods presented in this course were demonstrated primarily for dairy cattle, beef or swine. However, the methodology is applicable to any species of animals or plants. Companion animal genetics is being considered as a new undergraduate course. Briefly discuss what you would include in such a course involving dog breeding (for example) for the following areas.

    (a) Main breeds?

    (b) Relationships and Inbreeding?

    (c) Traits of importance?

    (d) Crossbreeding?

    (e) Genetic evaluation?

    (f) DNA testing?

15. **(2 points)** Give an example of a commercially available DNA test in cattle, and its purpose.

16. **(10 points)** Write the full name for each of the following abbreviations.

    - SNP
    - BIO
    - MACE
    - RFLP
    - QTL

MBG*4030 - Animal Breeding Methods

Fall 2007 - Midterm Exam

1. **(2 points):** The trace of a square matrix $tr()$ is equal to the sum of its diagonal elements. Calculate the trace of the following product.

$$tr\left(\begin{bmatrix} 10 & -1 & 5 \\ -2 & 0 & 4 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 3 & -3 \\ 2 & 2 \end{bmatrix}\right) = ANS = 27.$$

2. **(10 points):** Rabbits are grown for meat in many European countries, and in North America. The favourite species is the New Zealand White. A producer would like to assess the genetic merit of his breeding stock. You convinced the producer to keep weight records on the progeny at a specific age, and to keep good pedigree records for a number of years. Now it is time for you to analyze the data for the producer. Rabbits are grown separately in cages. The feed formulation has changed twice over the years that animals were recorded, but only at the start of a year and not in the middle. LIST the factors that you would put into the statistical model, and LIST the assumptions of your model.

List of Factors
Animal additive genetic
Contemporary Group(Year,Month of birth)
Sex of rabbit effect
Age at Weighing effect as a covariate
Maternal Genetic effects
Mat. Permanent Environ. effects

Assumptions:

- All rabbits are the same breed.
- Weighings are at roughly the same age.
- Effects of cages or hutches within a contemporary group are not important.
- No interactions between factors of the model.
- Pedigrees are assumed to be known.
- Heritability should be about 0.30.
- Effect of age at weighing assumed to be linear.

3. **(4 points):** Below are the test grades of different students, the hours of study for the test of each student, the hours of sleep the night before the test, and the colour of hair of each student. There are data on 378 students in one course, and below is a sample of seven students.

Grade	Study Hours	Sleep Hours	Hair Colour
53	0.5	12.5	Red
96	2.0	10.0	Brown
74	4.0	6.5	Black
69	3.5	9.0	Blonde
42	1.0	5.0	Brown
88	3.0	7.5	Brown
77	5.5	8.0	Black

The model for the analysis of these data is

$$Grade_{ij} = \mu + b_1(\text{Study,h}) + b_2(\text{Sleep,h}) + H_i + e_{ij},$$

where H_i is hair colour categories (Red, Brown, Black, and Blonde). Construct the $\mathbf{X}$ matrix for the sample data. (Hint: there should be 7 columns)

$$\mathbf{X} = \begin{pmatrix} 1 & 0.5 & 12.5 & 1 & 0 & 0 & 0 \\ 1 & 2.0 & 10.0 & 0 & 1 & 0 & 0 \\ 1 & 4.0 & 6.5 & 0 & 0 & 1 & 0 \\ \vdots & & etc & & & & \end{pmatrix}.$$

4. **(2 points):** Below is a list of pedigrees. Order the pedigrees chronologically.

Animal	Sire	Dam	Generation Numbers		New Order
Tom	Bart	Gerta	1		Chester
Harry	Starr	Lacy	1		Lacy
Chester			1	4	Starr
Mabel	Bart	Lucy	1		Bud
Starr			1	3	Lucy
Bud			1	3	Rose
Bart	Starr	Lucy	1	2	Bart
Gerta	Bud	Rose	1	2	Gerta
Lucy	Chester	Lacy	1	3	Fiona
Rose			1	3	Tom
Lacy			1	4	Harry
Dick	Chester	Fiona	1		Mabel
Fiona			1	2	Dick

5. **(2 points each):** Explain what each of the following R functions does.

   (a) sum(y) = adds together the elements in y

   (b) length(y) = gives number of elements in y

   (c) sum(diag(XX)) = adds the diagonals of XX, i.e. trace of XX.

   (d) tapply(weights,months,var) = calculates variances of weights by months.

   (e) source("Bills.R") = brings in the R code in Bills.R into your R session.

   (f) rbind(x,z) = row-wise binding of x and z.

   (g) sample(listA,nobs,replace=FALSE) = randomly select nobs items from listA without replacement.

6. **(6 points):** The following boxes are from a larger genomic relationship matrix.

					R	S
	R_m	R_f	S_m	S_f	T_m	T_f
A_m	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{4}$
A_f	0	$\frac{1}{2}$	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{1}{4}$	$\frac{1}{4}$

   (a) Fill in the missing coefficients between A and T.

   (b) Calculate a_{AS} and d_{AS}

$$a_{AS} = \frac{1}{2}(\frac{1}{8} + \frac{3}{8} + \frac{3}{16} + \frac{5}{16}).$$

$$d_{AS} = (\frac{1}{8}\frac{5}{16}) + (\frac{3}{16}\frac{3}{8}).$$

7. **(4 points):** An animal has an inbreeding coefficient of 0.15.

   (a) What is the meaning of an inbreeding coefficient?
   15% of the loci have two alleles that are identical by descent.

(b) What would be an expected inbreeding coefficient of a progeny of this animal if you bred the animal to another of its progeny (explain your calculations)?

Progeny of animal are related to this animal by at least (1.15/2) = .575. If animal is mated to one of its progeny, then the offspring of that mating would be inbred by 0.2875. If the animal is more highly related to its progeny, then the inbreeding coefficient will be greater.

8. **(2 points):** What are the differences between a genomic relationship matrix and an additive relationship matrix?

   Genomic relationship matrix is twice as large as the additive relationship matrix. All diagonals in genomic relationship matrix are equal to 1, but diagonals of additive relationship matrix can be between 1 and 2. Can easily get dominance genetic relationships from the genomic matrix, but not from the additive matrix.

9. **(10 points):** Given two animals, P and Q with the following parents and b_i values. Fill in the elements of $\mathbf{A}^{-1}$ for animals P and Q. The rules will be on the board. You do not need to add coefficients together, and please leave them in fraction form. Pedigrees are

Animal	Sire	Dam	b_i
P	A	B	7/16
Q	A	C	15/32

	—-A—-	—-B—-	—-C—-	–A—B– —-P—-	–A—C– —-Q—-
A	$\frac{4}{7}+\frac{8}{15}$	$\frac{4}{7}$	$\frac{8}{15}$	$-\frac{8}{7}$	$-\frac{16}{15}$
B	$\frac{4}{7}$	$\frac{4}{7}$		$-\frac{8}{7}$	
C	$\frac{8}{15}$		$\frac{8}{15}$		$-\frac{16}{15}$
P	$-\frac{8}{7}$	$-\frac{8}{7}$		$\frac{16}{7}$	
Q	$-\frac{16}{15}$		$-\frac{16}{15}$		$\frac{32}{15}$

10. **(2 points):** Why are phantom parent groups necessary?

    Are necessary to account for missing pedigree information, because all animals would not be traceable to the same base generation.

11. **(2 points):** Two traits are each measured more than once on the same animal during its life. One trait has $h^2 = 0.10$ with $r = 0.55$ and the other trait has $h^2 = 0.45$ with $r = 0.55$. Explain the differences that one might see between the two traits based on those parameter values?

    Trait 1 $r - h^2 = 0.45$, and Trait 2 is $r - h^2 = 0.10$. Thus, trait 1 would be more affected by permanent environmental effects than trait 2. Trait 2 would be more affected by additive genetic effects because $h^2 = 0.45$ versus $h^2 = 0.10$.

12. **(2 points):** What would be meant by a non-additive genetic variance that is represented by σ^2_{21} ?

    This would represent an additive by additive by dominance gene interaction.

13. **(4 points):** Maternal genetic effects are common in mammalian species. Explain the differences between a typical animal model, and a maternal genetics effect model?

    The dam provides an environmental effect to its progeny, and this ability is genetically transmitted to its progeny. There is a non zero covariance between direct and maternal genetic effects. There is usually a maternal permanent environmental effect too because females can have more than one offspring during their lifetime.

14. **(2 points):** Gibbs sampling is a tool that is used in Bayesian estimation of variances and covariances, and which was utilized in Lab 5. Why is a burn-in period needed with Gibbs Sampling? (Note we have not done Gibbs sampling in lab in Fall 2008)

    One needs samples from the joint posterior distribution, and during the early rounds of Gibbs sampling the samples are from the conditional posterior distributions. At burn-in, the samples have mixed and afterwards can be assumed to be samples from the joint posterior distribution.

15. **(2 points):** The Analysis of Variance is used to test the significance of factors in a model. Two models could be compared by looking at the R^2 value or the estimate of the residual variance from each model. How would you decide which model is better using these two measures?

    Models with lower residual variance are better. Models with higher R^2 are better.

16. **(2 points):** The course has emphasized simulation of data in order to better understand how data are explained by the different factors in the model. In a typical animal model, which factor(s) tend to have the largest contribution to each observation?

    The residual effects have usually the largest effect on the observations.

17. **(2 points):** How are SEP and reliability used to express the accuracy of an EBV?

    SEP gives a confidence interval-like figure, and reliability gives a percentage figure. Smaller SEP are good, and larger reliabilities are good.

18. **(2 points):** A nutritionist is creating biological models to explain the flow of nutrients through the stomach and intestines of the dairy cow. Blood samples are collected from cows every 15 minutes for 4 to 8 hours. Cholesterol, sugar, and other fatting acids can be monitored over this period. Data were collected on several cows on two different diets to compare the effects of dietary inputs on blood parameters. What kind of statistical model do you think might be useful for this data? Why?

    The data can be considered to be longitudinal data (taken over time following a curve of some sort). Thus, a random regression model might be considered.

19. **(6 points):** Suppose you have a herd of individuals that were all cloned from a single individual. That is, the entire herd is genetically identical, i.e. 100% of genes are identical by descent in all animals.

    (a) Would you expect all animals to have the same phenotype (e.g. same amount of milk yield)? Why?

    No, each animal's residual effect will be different, causing the observations to be different.

    (b) Would you expect all animals to have the same EBV? Why?

    Expect, yes. If the EBV methods are good, then EBVs should be the same, because all animals have the same genes.

    (c) What would be the advantages and disadvantages of having such a herd?

    Low variability among the animals. Feeding and management might be easier with all animals so similar. They could all be wiped out by one disease. Boring without variability.

MBG*4030 - Animal Breeding Methods
Fall 2007 - Final Exam

1 Problem Questions

Mick Dundee used his financial resources to purchase the "Now That's A Croc" crocodile farm that had been operating for a number of years in Darwin, Australia. There were 100 breeding females and 20 breeding males plus 1600 immature females and 320 immature males.

1. What would be the effective population size?

$$\frac{1}{N_e} = \frac{1}{4N_m} + \frac{1}{4N_f}$$

2. What would be the expected inbreeding rate per generation?

$$\Delta F = \frac{1}{2N_e}$$

3. How could Mick avoid inbreeding problems?

The farm would be open to tourists (to generate income), and would also supply skins to the luxury leather industry. The traits of economic interest, therefore, were body length at 2 years (the bigger they are, the better they scare tourists), tooth size at 2 years, skin quality, egg number, and egg mortality. Unfortunately, crocs take 16 years to reach maturity (breeding age), and skins are collected at 2-3 years of age. Number of eggs laid is about 40-50 per nest, but mortality is usually high and some females will eat their hatchlings (not very good maternal ability). Crocs can live to be 50 years old.

(True Note: A known expert, Dr. Grahame Webb, started a World Research Centre for studying all 22 species of crocodiles in 1990, and started a zoo in Darwin called Crocodylus Park.)

Mick decided to compute EBVs for the following traits: BL(body length at 2 yr), TZ(tooth size at 2 yr), SQ(skin quality at 2 yr), EN (egg number), and EM (egg mortality per cent).

4. There have not been enough studies on crocs to know what the heritability of these traits might be. What rough values would you suggest for each trait?

Assume the following additive genetic covariance matrix for the following traits (these are not actual estimates, but are simplified numbers for exam purposes):

$$\mathbf{G} = Var\begin{pmatrix} BL \\ TZ \\ SQ \\ EN \\ EM \end{pmatrix} = \begin{pmatrix} 25 & 2 & 5 & -1 & 10 \\ 2 & 16 & 3 & -2 & 4 \\ 5 & 3 & 36 & -3 & 9 \\ -1 & -2 & -3 & 49 & 10 \\ 10 & 4 & 9 & 10 & 36 \end{pmatrix}.$$

Mick decided to use the following index for selection of breeding animals:

$$I_1 = 10(BL) + 10(TZ).$$

5. What is the variance of his index?

$$Var(I_1) = w_1^2 Var(BL) + w_2^2 Var(TZ) + 2w_1 w_2 Cov(BL, TZ)$$

6. What would be the value of one genetic standard deviation change in BL and TZ?

7. What is the relative emphasis on BL compared to TZ?

8. If the accuracy of the index, (r_{TI}), is only 0.3, the intensity of selection, (i), is 0.5, and $L = 16$ years, then what is the expected genetic change in the index?

$$\Delta G = \frac{r_{TI}\ i\ \sigma_I}{L}.$$

9. What is the correlated response in TZ as a result of using this index? The covariances of the traits with the index is given by

$$\mathbf{Gw} = \mathbf{G}\begin{pmatrix} 10 \\ 10 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 270 \\ 180 \\ 80 \\ -30 \\ 140 \end{pmatrix}.$$

The correlated response is given by

$$\Delta_c G_i = \frac{Cov(G_i, I)}{\sigma_I}\ r_{TI}\ i.$$

10. What is the correlated response in EN as a result of using this index?

Below is a table for predicting the amount of genetic gain per year from the current breeding program. The Sire of Males and Sire of Females pathways are identical, and the Dam of Males and Dam of Females pathways are identical. About 20 males and 100 females are saved per year from roughly 4000 hatchlings, giving percentages of 1% for males and 5% for females, corresponding to selection intensity values of 2.65 and 2.06, respectively. The accuracy of selection is the same for all pathways and is 0.3. The generation interval is the age of the parent when the replacement progeny is born. If animals are not mature until 16 years, and if progeny are born 6 months later, then the generation interval is 16.5 years.

Pathway	Percentage	i	r_{TI}	L	$i \times r_{TI}$
SM	1%	2.65	0.3	16.5	0.795
SF	1%	2.65	0.3	16.5	0.795
DM	5%	2.06	0.3	16.5	0.618
DF	5%	2.06	0.3	16.5	0.618

11. What is the predicted genetic gain (in genetic standard deviation units)?

(Fiction) Another croc farm (called Krikey Krocs) raised a different 'breed' of croc than did Mick, and Mick began to consider crossbreeding. The crocs on the other farm had a better skin colour and better EN and EM than Mick's crocs. Apparently other breeders had thought of this possibility many years earlier, and below are the results of a crossbreeding experiment.

Cross	BL	TZ	SQ	EN	EM
A = Mick's	3.4 m	4.7 cm	74 pts	32	35%
B = Krikey's	2.8 m	3.9 cm	72 pts	48	25%
A x B	3.3 m	4.5 cm	75 pts	42	20%
B x A	3.0 m	4.3 cm	74 pts	45	22%

12. What is the heterosis for each trait by cross? Heterosis is the average of the cross minus the parent average divided by the parent average, all times 100%.

13. Should Mick buy males or females from Krikey for his crossbreeding exercise? Explain.

Research has found a marker for a major QTL that affects egg number. To genotype all of his animals Mick would have to spend $80,000.

14. Because of the negative correlations between EN with BL, TZ, and SQ, is selecting on QTL genotypes a good strategy for genetic improvement for Mick? Explain.

Twenty years later: A 300K SNP chip has been developed for crocs. SNP genotype effects for BL, TZ, SQ, EN, and EM are available, and the average accuracy of GEBVs for each of these traits is 0.80, plus the GEBV are available at birth.

15. Fill in the new values in the table below and predict genetic gain.

Pathway	Percentage	i	r_{TI}	L	$i \times r_{TI}$
SM	1%	2.65			
SF	1%	2.65			
DM	5%	2.06			
DF	5%	2.06			

16. Should Mick select breeding males and females from GEBV calculated at birth? Explain.

2 Multiple Choice Questions

Be careful, each question may have more than one correct answer. Circle the letter of the statements that are correct or apply.

17. The generation interval is defined as

1. X
2. L
3. the average age of the parent when a replacement progeny is born.
4. the average age of the progeny when the parent leaves the breeding population.
5. the interval between parturitions.

18. The breeding goal

1. includes all traits of economic importance.
2. is also known as the breeding net.
3. includes all possible traits.
4. is at the discretion of the breeder.
5. is also known as the aggregate genotype.

19. Examples of non-normally distributed traits in dairy cattle are

1. test day fat yield.
2. feet and legs score(1 to 9).
3. mastitis.
4. persistency.
5. temperament.

20. Retained heterosis is the percentage of heterosis

1. that is lost due to recombination.
2. that is saved by a parent and not transmitted to its progeny.
3. that remains after crossing animals that have alleles of one or more breeds in common in their genetic makeup.
4. that remains after accounting for inbreeding.
5. that is kept in reserve.

21. A multiple trait model should be used if

1. genetic and environmental covariance matrices are unknown.
2. traits are negatively correlated.
3. heritabilities of all traits are low.
4. culling might cause bias in one or more traits.
5. it has not been tried previously.

22. A random regression model should be used if

1. a trait has a trajectory, like growth.

2. a multiple trait model is not appropriate.
3. a Hazard's function can be used.
4. fixed regressions are not appropriate.
5. the genes affecting a trait change their output of proteins or enzymes with the age of the animal.

23. A threshold model should be used if

1. animals have reached a threshold age.
2. the trait has discrete, ordered categories.
3. the trait is a count variable like litter size or number of services to conception.
4. inbreeding has reached a dangerous threshold.
5. effective population size is low.

24. The Lifetime Profit Index (LPI) used in dairy cattle in Canada is composed of

1. a Production Component.
2. a Health/Fertility Component.
3. a Reproduction Component.
4. a Durability Component.
5. an Milk Quality Component.

25. The amount of retained heterosis in progeny from a cross of (A x C) males with ((A x B) x (C x D)) females would be

1. 0.25
2. 0.50
3. 0.6875
4. 0.75
5. 1.00

26. A random normal deviate has

1. a mean of 0 and variance of 1.
2. a mean of 1 and variance of 0.
3. a mean of 100 and variance of 100.
4. a kurtosis of 0.
5. a skewness of 0.

27. Variances and covariances are used in animal breeding for

1. making mixed model equations.
2. testing hypotheses.
3. giving heritability and repeatability values.
4. calculating heterosis.
5. computing correlated responses.

28. Selection index is

1. a practical representation of the Aggregate Genotype.

2. more useful than single trait selection.
3. only used in dairy cattle.
4. related to the Cost of Living Index.
5. a linear index not really useful for selection.

29. Single Nucleotide Polymorphisms (SNPs) are popular today because

1. there are millions of them spread across the genome.
2. they are co-dominant.
3. they are not inbred.
4. they are highly polymorphic.
5. they are easy to genotype.

30. A haplotype is

1. a model of a genotype.
2. a copy of a genotype.
3. a font size in R.
4. the set of alleles (one per locus) that happen to be on a particular chromosome and usually inherited as a single unit most of the time, except for recombination.
5. a combination of *haplo*id and geno*type*, the 'genotype' of a single chromosome.

31. Linkage Disequilibrium (LD) is

1. greater in humans than in cattle.
2. unequal recombination between loci on a chromosome.
3. not related to linkage.
4. created through random mating.
5. needed for QTL detection.

32. Selection intensity is

1. directly related to the amount of genetic change.
2. related to the heritability of the trait.
3. zero when 50% of the animals are selected.
4. related to the level of anxiety in animals.
5. the mean of individuals above a truncation point on a normal distribution with mean of zero and standard deviation of one.

33. Assume a two trait index,

$$I = b_1(EBV_1) \ + \ b_2(EBV_2)$$

and let $b_1 = 1$. If the additive genetic variance of trait 1 is $(40)^2$ and for trait 2 is $(60)^2$, then what should be the value of b_2 such that the emphasis on trait 2 is 3 (three) times greater than for trait 1?

1. 3
2. 2
3. 1
4. $\frac{1}{2}$

5. $\frac{1}{3}$

34. The Candidate Gene Approach is

1. popular with politicians.
2. an elective method.
3. where a known gene is assumed to be linked to a QTL for a trait that might be influenced by the gene.
4. a highly successful method.
5. costly because many animals need to be sequenced to find a mutation in the gene.

35. Permutation testing is

1. where observations are randomly shuffled with respect to marker genotypes.
2. where observations are randomly sampled with replacement within marker genotypes.
3. a method to derive confidence intervals on parameter estimates.
4. a method to test the significance of estimates of marker genotype differences.
5. the same as JackKnifing.

36. Bootstrapping is

1. where observations are randomly shuffled with respect to marker genotypes.
2. where observations are randomly sampled with replacement within marker genotypes.
3. a method to derive confidence intervals on parameter estimates.
4. a method to test the significance of estimates of marker genotype differences.
5. the same as JackKnifing.

37. Selection is defined as

1. natural, random culling of animals.
2. non-random pairing of mating individuals.
3. an act of God.
4. choosing sires for use on a group of females.
5. any action that changes the probability of an individual's chances to reproduce.

38. Effective population size, N_e,

1. is given by $\frac{1}{N_e} = \frac{1}{4N_m} + \frac{1}{4N_f}$.
2. equals the number of breeding males (N_m) in the population.
3. equals the number of breeding females (N_f) in the population.
4. determines the rate of inbreeding in a population.
5. is greater than the actual population size.

39. Henner Simianer's proposal for breed conservation was based upon

1. the rescue of breeds.
2. maintaining diversity between breeds.
3. the fact that breeds are not clearly defined.
4. maintaining a maximum number of alleles in a set of breeds.

5. genotyping animals with 2 microsatellite markers.

40. Recombination rate between two loci on one chromosome can be converted to a map distance in centiMorgans (cM) using

1. the Google map function.
2. the Hardy-Weinberg map function.
3. the Kosambi map function.
4. a Haldane map function.
5. a Chi-squared statistic.

41. Recombination rates are not additive because of

1. insertions and deletions in DNA.
2. parental types.
3. hot spots.
4. crossover interference.
5. recombinant types.

42. General combining ability is

1. the additive effect of the breed of sire.
2. the dominance effect of the breed of sire.
3. the additive effect of the breed of dam.
4. the dominance effect of the breed of dam.
5. the dominance effects of a particular breed of sire by breed of dam cross.

43. Specific combining ability is

1. the additive effect of the breed of sire.
2. the dominance effect of the breed of sire.
3. the additive effect of the breed of dam.
4. the dominance effect of the breed of dam.
5. the dominance effects of a particular breed of sire by breed of dam cross.

44. In choosing breeds for a crossbreeding program one should consider

1. Heterosis.
2. Linkage disequilibrium.
3. Climate.
4. How the breeds look in the barn.
5. Complimentarity.

45. Genetic variances can decrease over time due to

1. selection.
2. genetic drift.
3. heterosis.
4. linkage disequilibrium.
5. inbreeding.

MBG*4030 - Animal Breeding Methods

Fall 2008 - Midterm Exam
October 22, 2008

For multiple choice questions circle all answers that apply to a given question.

1. Four necessary pieces of information needed to make genetic change in a population are ANS=(a,b,e,f)

   (a) Records on the trait of interest.
   (b) Sire and dam information on all individuals.
   (c) Number of chromosomes in that species.
   (d) R software.
   (e) Knowledge about the production system.
   (f) Prior information about parameters.

2. Conformable for multiplication in matrix algebra means ANS=(none)

   (a) Two matrices conform to Geneva conventions.
   (b) Both matrices have the same number of rows and columns.
   (c) Both matrices are square.
   (d) The number of rows in the first matrix equals the number of columns in the second matrix.
   (e) Any two matrices are conformable.

3. A symmetric matrix is one which ANS=(d,e)

   (a) Has an even number of rows and columns.
   (b) The left half is the mirror image of the right half.
   (c) The top half is the mirror image of the lower half.
   (d) The lower left off-diagonals are the mirror image of the upper right off-diagonals.
   (e) The transpose equals the original matrix.

4. Estimates of the number of genes in a mammalian genome are ANS=(b)

   (a) Between 3,000 to 6,000.
   (b) Between 30,000 to 60,000.
   (c) Between 300,000 to 600,000.
   (d) Exactly 50,000.
   (e) Exactly 500,000.

5. The Infinitesimal Model ANS=(a,b,d,e,f)

   (a) Was put forward in 1909.
   (b) Assumes a random mating population.
   (c) Assumes an infinite number of effects in the model.
   (d) Assumes all loci have an equal and small effect.
   (e) Assumes an infinite number of loci.
   (f) Assumes only additive genetic effects.

6. EBV is ANS=(b,c,d,e,f)

   (a) short for expected breeding value.
   (b) used for culling and mating decisions.
   (c) used to measure genetic change.
   (d) obtained from statistical linear models.
   (e) short for estimated breeding value.
   (f) two times the ETA.

7. A linear model consists of three items, which are ANS=(a,c,e)

   (a) the equation of the model.
   (b) the analysis of variance table.
   (c) the expectations and variances of the random variables.
   (d) the Gibbs sampler.
   (e) the assumptions and limitations.

8. Different models can be compared for their fit of the data by ANS=(b,d)

   (a) comparing assumptions and limitations.
   (b) comparing their multiple R-squared values.
   (c) comparing their F-statistics.
   (d) comparing their estimated residual variances.
   (e) taking them for a test drive.

9. Holstein Canada ANS=(a,e)

   (a) is responsible for registrations of all dairy breeds.
   (b) is located in Guelph, Ontario.
   (c) is responsible for milk recording.
   (d) participates in AI sire selection.
   (e) is responsible for type classification in all dairy breeds.

10. The registrar's office has data files of all students (all curriculae), the courses they have taken, the grades they have received, the teachers of those courses, and the year and semester in which the course was taken. The data cover the last 20 years.

    (a) Write a linear model to analyze the final grades in each course, so that students could be ranked as well as the instructors. List the factors that you would put into the model initially. Students will have about 40 final grades each over their university years. Do not forget your assumptions and limitations.

    ANS:

Factors
Year-semester-course (contemporary group)
Instructor(s) of a course
Student (genetic + PE) - most important to have in model
Class size
Time of day when classes meet for a course
Student's program (Arts, Science, etc.)
Student' semester (1 to 8)

Assumptions and Limitations
No effect of age or sex of student (everyone equal).
No instructor by student interactions (preferential treatment).
Students work equally hard in all courses.
Variation in grades similar across courses.
Genetic relationships among students ignored.

    (b) Do you think the data will be sufficiently connected to make the analysis worthwhile? Why?

    ANS: I accepted any answer. Connectedness should not be a problem - everyone has to take electives and this will combine people from different programs.

(c) What do you think would be the repeatability of final grades?
ANS: I accepted any answer. My guess would be .4 to .6.

11. Below are example data on somatic cell scores (SCS) of dairy cows in one herd, and their days in milk (DIM), protein yield (Prot) on test day, and their age at calving (AGE). There are data on 58 cows.

SCS	DIM	Prot,kg	AGE,yr
3.53	12	1.25	2
3.96	63	1.00	2
3.74	49	1.65	3
3.69	113	0.90	3
4.42	197	0.85	5
3.88	88	0.98	4
3.77	36	1.54	6

The model for the analysis of these data is

$$SCS_{ij} = \mu + b_1(\text{DIM}) + b_2(\text{Prot}) + A_i + e_{ij},$$

where A_i is AGE (from 2 yr to 8 yr). Construct the $\mathbf{X}$ matrix for the sample data. (Hint: there should be 10 columns)

$$\mathbf{X} = \begin{pmatrix} 1 & 12 & 1.25 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 63 & 1.00 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 49 & 1.65 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 113 & .90 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 197 & .85 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 88 & 0.98 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 36 & 1.54 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \vdots & & & & & & & & & \end{pmatrix}.$$

12. Order the following pedigrees from oldest to youngest.

Animal	Sire	Dam	Generations	ANS
BF	DD	HE		GA
DD	GA	EC		FB
GA				EC
EC	GA	FB		DD
AG	BF	EC		HE
FB				BF
HE	DD	FB		AG

13. Give the R function(s) for performing the following tasks. (Note: 3 points for getting all but 1, then 2 points, but only 1 point if some were left blank.)

    (a) To compute the mean and variance of values in a vector.
    `mean(vector), var(vector)`

    (b) To determine the number of items in an array or vector.
    `length(vector) OR dim(vector)`

    (c) To bind two or more columns of numbers together.
    `cbind(vecA, vecB)`

    (d) To compute the inverse of a matrix.
    `ginv(matrix) OR inv(matrix)`

    (e) To list the names of variables in your workspace.
    `ls()`

    (f) To generate some random normal variates with variance 100.
    `rnorm(number, SD)`

    (g) To sort a vector of numbers.
    `order(vector) OR sort(vector) OR rank(vector)`

14. Complete the calculations for the following genomic relationship matrix. Animals R and S are unrelated to each other, and are the parents of T. Note that R is inbred.

	R_m	R_f	S_m	S_f	R T_m	S T_f
R_m	1	$\frac{1}{4}$	0	0	$\frac{5}{8}$	0
R_f	$\frac{1}{4}$	1	0	0	$\frac{5}{8}$	0
S_m	0	0	1	0	0	$\frac{1}{2}$
S_f	0	0	0	1	0	$\frac{1}{2}$
T_m	$\frac{5}{8}$	$\frac{5}{8}$	0	0	1	0
T_f	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	1

15. If one parent of an animal has an inbreeding coefficient of 0.20, and the other parent is unknown, then what is the b_i value of this animal? ANS=(c)

    (a) $b_i = 0.5 - 0.25 * (0.20 + 0.00) = 0.45.$

    (b) $b_i = 1$

    (c) $b_i = 0.75 - 0.25 * (0.20) = 0.70.$

    (d) $b_i = 0.5$

    (e) $b_i = 0.20$

16. The differences between a genomic relationship matrix(GEN) and a numerator additive relationship matrix(NA) are ANS=(a,b)

    (a) Inbreeding coefficients are in the off-diagonals of the GEN and on the diagonals in NA.

    (b) GEN allows the calculation of dominance genetic relationships.

    (c) NA has order equal to the number of animals while GEN has order equal to half the number of animals.

    (d) GEN is more accurate than NA.

    (e) NA can be inverted more easily using Henderson's rules than GEN.

17. Given two animals, P and Q with the following parents and b_i values. Fill in the elements of $\mathbf{A}^{-1}$ for animals P and Q. The rules will be on the board. You do not need to add coefficients together, and please leave them in fraction form.

Animal	Sire	Dam	b_i
P	A	B	1/8
Q	P	C	5/16

	—-A—-	—-B—-	—-C—-	–A—B– —-P—-	–P—C– —-Q—-
A	2	2		-4	
B	2	2		-4	
C			$\frac{4}{5}$	$\frac{4}{5}$	$-\frac{8}{5}$
P	-4	-4	$\frac{4}{5}$	$8+\frac{4}{5}$	$-\frac{8}{5}$
Q			$-\frac{8}{5}$	$-\frac{8}{5}$	$\frac{16}{5}$

18. Phantom parent groups are used when ANS=(c)

    (a) the pedigrees are unknown.

    (b) a medium says they are needed.

    (c) animals with unknown parents exist over many years, such that all unknown parents can not be assumed to be from the same base generation.

    (d) They are never used.

    (e) they are statistically significant.

19. Phantom parent groups are usually formed on the basis of year of birth and ANS=(c, but a,d worth 2 points)

    (a) Two pathways of selection.

    (b) Three pathways of selection.

    (c) Four pathways of selection.

    (d) Four pathways of selection and breed.

20. Two traits are each measured more than once on the same animal during its life. Trait 1 has $h^2 = 0.10$ with $r = 0.55$ and Trait 2 has $h^2 = 0.45$ with $r = 0.55$. The differences between the two traits are ANS(a,b,c,d)

(a) Trait 2 has smaller PE effects.

(b) The ratio of residual variance to PE variance in Trait 1 will be larger than for Trait 2.

(c) Selection on Trait 2 will be more effective than selection on Trait 1.

(d) More observations would be needed to evaluate animals for Trait 1 than for Trait 2.

(e) The mean of Trait 2 is greater than that of Trait 1.

21. Maternal genetic effects are common in mammalian species. Maternal genetic effects models ANS=(b,c,d,e)

(a) usually include animal PE effects because females generally have more than one record.

(b) usually include maternal PE effects because females generally have more than one progeny over their life.

(c) have a non-zero correlation between direct and maternal genetic effects.

(d) generally not used for traits observed after weaning.

(e) require special care when ET or cross fostering is used.

22. Methods of estimating variances from animal models in present day animal breeding research are ANS=(c,e)

(a) Henderson's Methods 1, 2, and 3.

(b) Fisher's ANOVA methods.

(c) Restricted Maximum Likelihood.

(d) Akaiki's method.

(e) Bayesian methodology.

23. Gibbs sampling is a computational tool that is used in Bayesian estimation of variances and covariances. ANS=(a,c,e)

    (a) Gibbs sampling is used because the joint posterior distribution is too complicated to maximize.

    (b) Gibbs sampling requires a short burn-in period before the conditional posterior distributions converge to the joint posterior distribution.

    (c) Estimates of standard errors are possible from the Gibbs samples.

    (d) This is the most accurate method of estimation of variances.

    (e) Gibbs sampling is used because it takes relatively little time to get results.

24. Relative Breeding Values ANS=(c)

    (a) Have an average value of 0 among all close relatives.

    (b) Have an average value of 100 among all close relatives.

    (c) Have an average value of 100 among all animals in the genetic base.

    (d) Have an average value of 0 among all animals in the genetic base.

    (e) Are not used in animal breeding.

25. Reliabilities ANS=(b,c)

    (a) are better than SEP to indicate the accuracy of EBVs.

    (b) go from 0 to 100%.

    (c) are derived from the inverse elements of the coefficient matrix of the mixed model equations.

    (d) are greater than heritabilities.

    (e) are smaller than heritability of the trait.

26. In a typical animal model, the factor having the largest influence on the observations is usually ANS=(e)

    (a) the additive genetic effect.

    (b) the contemporary group effect.

    (c) the maternal genetic effect.

    (d) the permanent environmental effect.

    (e) the residual effect.

27. A nutritionist is creating biological models to explain the flow of nutrients through the stomach and intestines of the dairy cow. Blood samples are collected from cows every 15 minutes for 4 to 8 hours. Cholesterol, sugar, and other fatting acids can be monitored over this period. Data were collected on several cows on two different diets to compare the effects of dietary inputs on blood parameters. What kind of statistical model would be useful for this data? ANS=(d and/or f)

    (a) A typical animal model.
    (b) A repeated records animal model.
    (c) A maternal genetic effects model.
    (d) A random regression model.
    (e) A non-additive genetic model.
    (f) A multiple trait model.

28. Suppose you have a herd of cows that were all cloned from a single individual. That is, the entire herd is genetically identical, i.e. 100% of genes are identical by descent in all animals. ANS=(b,c)

    (a) All animals would have exactly the same phenotype (e.g. same amount of milk yield).
    (b) All animals would have exactly the same EBV.
    (c) Animals would have different phenotypes because of different PE and residual effects.
    (d) Animals would have different EBVs because the phenotypes are all different.
    (e) All animals would look exactly the same.

29. The person that hated Karl Pearson for rejecting one of his papers was ANS=(b)

    (a) Everyone.
    (b) Sir R. A. Fisher
    (c) Jay L. Lush
    (d) Sewall Wright
    (e) C. R. Henderson

30. (**BONUS QUESTION**). Unscramble the four words below, then take the second letter of each word to spell the answer.

    SOVEIABRTSON LLLAEE MEPILLUT LASOPOPAA

    ANS = BLUP

    OBSERVATIONS ALLELE MULTIPLE APPALOOSA