Lecture 5
Basic Designs for Estimation of Genetic Parameters

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.
Notes from a short course taught Jan-Feb 2012 at University of Uppsala

Heritability

The reason for our focus, indeed obsession, on the heritability is that it determines the degree of resemblance between parents and offspring, which in turn determines the response to selection. In particular, the slope of a midparent-offspring regression is just $h^2 = V_A/V_P$.

Why h^2 instead of h?

Students often ask why we use h^2 rather than h to refer to heritability. You can blame Sewall Wright for this, as he used h to denote the correlation between breeding and phenotypic values within an individual. Recalling that the square of the correlation is the total fraction of variation accounted by a variable leads to the universal use of h^2. Indeed, whenever we speak of heritability, we always refer to h^2 and never to h. To see that h is indeed the correlation between breeding and phenotypic values within an individual, note from the definition of a correlation that

$$r_{Ap} = \frac{\sigma(A,P)}{\sigma_A\,\sigma_P} = \frac{\sigma(A,A+D+E)}{\sigma_A\,\sigma_P} = \frac{\sigma(A,A)}{\sigma_A\,\sigma_P} = \frac{\sigma_A^2}{\sigma_A\,\sigma_P} = \frac{\sigma_A}{\sigma_P} = h \tag{5.1}$$

Heritabilities are Functions of a Population

As heritability is a function of both the genetic and environmental variances, it is strictly a property of a population. Different populations, even if closely related, can have very different heritabilities. Since heritability is a measure of the *standing genetic variation* of a population, a zero heritability DOES NOT mean that a trait is not genetically determined. For example, an inbred line may show consist features that are clearly the result of genetic differences relative to other lines. However, since there is no variation *within* this hypothetical inbred population, h^2 is zero.

Increasing the Heritability

Note that h^2 decreases as the phenotype variance σ_P^2 increases. Hence, if one can reduce the environmental variance (for example, by more careful measurements of a trait or by using a more uniform environment) one increases the heritability. One note of caution, however. The heritability is not only a genetic property of a population, but also of the distribution (or *universe*) of environmental values that the population experiences. Thus, a heritability measured in a laboratory population may be rather different from the same population measured in a natural setting due to a wider range of environments. This is not a serious problem for breeders and experimentalists, *provided* that genotype-environment interactions are small. As the universe of environments change, when significant G x E is present, this can change the genotypic values, and hence any appropriate genetic variance. This issue is of special concern to plant breeders, where even slightly different growing regions may have subtle, but consistent, differences in their distribution of environmental values.

Heritability and the Prediction of Breeding Values

As mentioned, h^2 is the proportion of the total variance attributable to differences in breeding values. Further, h^2 is the slope of the regression predicting breeding value given an individual's phenotypic

value, as

$$A = \frac{\sigma(P,A)}{\sigma_P^2}(P-\mu_p) + e = h^2(P-\mu_p) + e \tag{5.2a}$$

This follows from the definition of a regression slope and the fact that the regression must pass through the mean of both A and P (0 and μ_p, respectively). The error e in predicting breeding value A from phenotypic value P has mean zero and variance

$$\sigma_e^2 = (1-h^2)\sigma_A^2 \tag{5.2b}$$

Hence, the larger the heritability, the tighter the distribution of true breeding values around the value $h^2(P-\mu_p)$ predicted by an individuals' phenotype.

Since heritability is a function of genetic variances, as allele frequencies change (for example, by selection and/or drift), the heritabilities also change. The slope of the parent-offspring regression changes during the course of selection, and as a result, our prediction of the response to selection using some initial estimate of heritability from an unselected population is good for only a few generations.

Heritability Values and Population Divergence

While the heritability of a population provides a measure of its genetic potential to response to a generation of selection, the magnitude of h^2 only provides information on the potential over a few generations. As allele frequencies change, so does heritability. A population showing a high h^2 value may have heritability erode to zero very quickly, while another population with a much smaller h^2 value may actually have heritability increase during selection as rare favorable alleles become more frequent. Hence, *heritability is a completely unreliable predictor for long-term response*, although it is generally a good to excellent predictor of short-term response.

Likewise, measuring heritability values in two populations that show a difference in their means provides no information on whether the underlying difference is genetic — h^2 is only a measure of the *current variation* in each population, it provides no information on the past history of either population. Thus, high estimated h^2 values in two divergent populations does not imply that the divergence is genetic (it could be strictly environmental). Likewise, low estimates of h^2 does not imply that an observed difference between two populations is environmental — both population could have exhausted genetic variation during selection for their divergence. In short, variances within populations and means between populations are not comparable.

Broad-Sense Heritability H^2 and Plant Breeding

There are actually two measures of heritability. We have seen the **narrow-sense** heritability $h^2 = \sigma_A^2/\sigma_z^2$, the fraction of total variance due to variance in breeding values. However, when examining offspring that are *clones* (exact genetic copies) of their parent, the more appropriate measure is the **broad-sense heritability**,

$$H^2 = \frac{\sigma_G^2}{\sigma_z^2} \tag{5.3}$$

H^2 is the fraction of phenotypic variance that is due to variation in genotypic values. Note that $H^2 \geq h^2$. Just as h^2 is the slope of the parent-offspring regression when offspring are sexually produced, H^2 is the parent-offspring regression slope with clones.

One interesting (and subtle) complication with measuring H^2 is that when we examine clones, we typically do not measure individuals, but instead measure a **plot** or a **block** of individuals. This

can result in inconsistent measures of H^2 even for otherwise identical populations. To see this, consider the value of the ℓ-th individual in plot k for genotype i in macroenvironment j,

$$z_{ijk\ell} = G_i + E_j + GE_{ij} + p_{ijk} + e_{ijk\ell} \tag{5.4a}$$

If we set our unit of measurement as the average over all plots, then for individuals measured in e macro-environments, each with p replicate plots that each consist of n individuals, the phenotypic variance becomes

$$\sigma^2(z_i) = \sigma^2_G + \sigma^2_E + \frac{\sigma^2_{GE}}{e} + \frac{\sigma^2_p}{e \cdot r} + \frac{\sigma^2_e}{e \cdot r \cdot n} \tag{5.4b}$$

Thus, the value of the phenotypic variance, and hence the value of H^2, depends on our choice of e, r, and n, the experimental design parameters.

Estimation: One-way ANOVA and the Simple Full-sib Design

We now turn to common designs for estimating heritability. Perhaps the simplest sib-based design is to examine N full-sib families, each with n offspring. The traditional approach to analyzing such data is the **one-way analysis of variance**, based on the linear model

$$z_{ij} = \mu + f_i + w_{ij} \tag{5.5}$$

where z_{ij} is the phenotype of the *j*th offspring of the *i*th family, f_i is the effect of the *i*th family and w_{ij} is the residual error resulting from segregation, dominance, and environmental contributions. We further assume that the w_{ij} are uncorrelated with each other and have common variance σ^2_w, the **within-family variance** (we will also use w_{FS} and w_{HS} to distinguish between the within-family variance for full- and half-sibs respectively, but for now it is clear that we are simply dealing with a full-sib family). The variance among family effects (the **between-family**, or **among family, variance**) is denoted by σ^2_f.

A basic assumption of linear models underlying ANOVA is that the random factors are uncorrelated with each other. This leads to a key feature:

- *The analysis of variance partitions the total phenotypic variance into the sum of the variances from each of the contributing factors.*

For example, for the full-sib model, the critical assumption is that the residual within-family deviations are uncorrelated with the family effects, i.e., $\sigma(f_i, w_{ij}) = 0$. Thus, the total phenotypic variance equals the variance due to differences among family means plus the residual variance of individual family members about their mean,

$$\sigma^2_z = \sigma^2_f + \sigma^2_{w(FS)} \tag{5.6}$$

The second ANOVA relationship that proves to be very useful is that

- *The phenotypic covariance between members of the same group equals the variance among groups.*

To see this for full-sibs, note that members of the same group (full sibs) share family effects, but have independent residual deviations, so

$$\begin{aligned} \text{Cov(Full Sibs)} &= \sigma(\, z_{ij}, z_{ik}\,) \\ &= \sigma[\,(\mu + f_i + w_{ij}), (\mu + f_i + w_{ik})\,] \\ &= \sigma(\, f_i, f_i\,) + \sigma(\, f_i, w_{ik}\,) + \sigma(\, w_{ij}, f_i\,) + \sigma(\, w_{ij}, w_{ik}\,) \\ &= \sigma^2_f \end{aligned} \tag{5.7}$$

The identity Cov(within) = Var(between) allow us to relate an estimated variance component (e.g., the between-family variance σ_f^2) with the causal variance components (e.g., σ_A^2) that are our real interest. For example, the variance among family effects equals the covariance between full sibs,

$$\sigma_f^2 = \sigma_A^2/2 + \sigma_D^2/4 + \sigma_{Ec}^2 \tag{5.8a}$$

where E_c is the common (or shared) family environmental effects (such as shared maternal effects) Likewise, since $\sigma_P^2 = \sigma_f^2 + \sigma_{w(FS)}^2$, the within-group variance $\sigma_{w(FS)}^2$ (i.e., the variance of full-sib values about their family mean) is

$$\begin{aligned}\sigma_{w(FS)}^2 &= \sigma_P^2 - (\,\sigma_A^2/2 + \sigma_D^2/4 + \sigma_{Ec}^2\,) \\ &= \sigma_A^2 + \sigma_D^2 + \sigma_E^2 - (\,\sigma_A^2/2 + \sigma_D^2/4 + \sigma_{Ec}^2\,) \\ &= (1/2)\sigma_A^2 + (3/4)\sigma_D^2 + \sigma_E^2 - \sigma_{Ec}^2\end{aligned} \tag{5.8b}$$

The ANOVA table for a balanced full-sib design becomes:

Table 5.1 ANOVA for a balanced full-sib design for N families each with n sibs.

Factor	df	SS	MS	E(MS)
Among-families	$N-1$	$\text{SS}_f = n\sum_{i=1}^{N}(\overline{z}_i - \overline{z})^2$	$\text{SS}_f/(N-1)$	$\sigma_{w(FS)}^2 + n\sigma_f^2$
Within-families	$T-N$	$\text{SS}_w = \sum_{i=1}^{N}\sum_{j=1}^{n}(z_{ij} - \overline{z}_i)^2$	$\text{SS}_w/(T-N)$	$\sigma_{w(FS)}^2$

Note: The total sample size is $T = Nn$. Degrees of freedom are denoted by df, observed sums of squares by SS, and expected mean squares by E(MS).

Estimating Variances and Variance Components

Unbiased estimators of σ_f^2, $\sigma_{w(FS)}^2$, and σ_z^2 follow from the expected mean squares

$$\text{Var}(f) = \frac{\text{MS}_f - \text{MS}_w}{n} \tag{5.9a}$$

$$\text{Var}(w) = \text{MS}_w \tag{5.9b}$$

$$\text{Var}(z) = \text{Var}(f) + \text{Var}(w) \tag{5.9c}$$

Recalling Equation 5.4,

$$2\sigma_f^2 = \sigma_A^2 + \sigma_D^2/2 + 2\sigma_{E_c}^2$$

so that $2\sigma_f^2$ provides an *upper bound* on σ_A^2.

Standard errors for the variance estimators given by Equation 5.9a-5.9c follow (under the assumptions of normality and balanced design) since the observed mean squares extracted from an analysis of variance are distributed independently with expected sampling variance

$$\sigma^2(\text{MS}_x) \simeq \frac{2(\text{MS}_x)^2}{\text{df}_x + 2} \tag{5.10}$$

Since Equations 5.9a–5.9c are linear functions of the observed mean squares, the rules for obtaining variances and covariances of linear functions (Lecture 3) can be used in conjunction with Equation 5.10 to obtain the large-sample approximations

$$\text{Var}[\,\text{Var}(w(FS))\,] = \text{Var}(\text{MS}_w) \simeq \frac{2(\text{MS}_w)^2}{T-N+2} \tag{5.11a}$$

$$\text{Var}[\,\text{Var}(f)\,] = \text{Var}\left[\frac{\text{MS}_f - \text{MS}_w}{n}\right] \simeq \frac{2}{n^2}\left(\frac{(\text{MS}_f)^2}{N+1} + \frac{(\text{MS}_w)^2}{T-N+2}\right) \tag{5.11b}$$

Estimating Heritability

Since the intraclass correlation for full-sibs is given by

$$t_{\text{FS}} = \frac{\text{Var}(f)}{\text{Var}(z)} = \frac{1}{2}h^2 + \frac{\sigma_D^2/4 + \sigma_{E_c}^2}{\sigma_z^2}, \tag{5.12a}$$

an upper bound for the estimate of heritability is given by

$$h^2 \simeq 2t_{\text{FS}} \tag{5.12b}$$

This has a (large-sample) standard error of

$$\text{SE}(h^2) \simeq 2(1-t_{\text{FS}})[1+(n-1)t_{\text{FS}}]\sqrt{2/[Nn(n-1)]} \tag{5.12c}$$

Worked Example of a Full-sib Design

Table 5.2. Suppose $N = 10$ full-sib families each with $n = 5$ offspring are measured.

Factor	df	SS	MS	E(MS)
Among-families	9	$\text{SS}_f = 405$	45	$\sigma_w^2 + 5\sigma_f^2$
Within-families	40	$\text{SS}_w = 800$	20	σ_w^2

$$\text{Var}(f) = \frac{\text{MS}_f - \text{MS}_w}{n} = \frac{45-20}{5} = 5, \quad \text{Var}(w) = \text{MS}_w = 20, \quad \text{Var}(z) = \text{Var}(f) + \text{Var}(w) = 25$$

Hence, an upper bound for the additive variance is $Var(A) = 2Var(f) = 10$. Likewise, the estimated heritability (assuming dominance and shared environmental effects can be ignored) is

$$2t_{FS} = 2\,\frac{5}{25} = 0.4, \quad \text{with} \quad \text{SE}(h^2) \simeq 2(1-0.4)[1+(5-1)0.4]\sqrt{2/[50(5-1)]} = 0.312$$

illustrating the (usually) large standard errors on heritability estimates.

Estimation: The Nested Full-sib, Half-sib Analysis

The simple full-sib design suffers in that one cannot obtain a clean estimate of σ_A^2. A more efficient design is the **nested full-sib, half-sib design**, wherein each male (or **sire**) is mated to several unrelated females (or **dams**), generating a series of full-sib families nested within half-sibs.

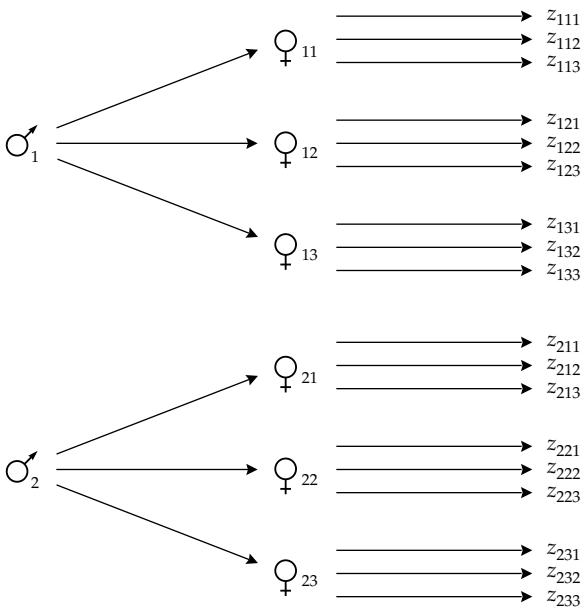


Figure 18.3 A nested full-sib, half-sib mating design. Each male is mated to several unique (unrelated) females, from each of which several offspring are assayed.

The linear model for this **nested design** is

$$z_{ijk} = \mu + s_i + d_{ij} + w_{ijk} \tag{5.13a}$$

where z_{ijk} is the phenotype of the *k*th offspring from the family of the *i*th sire and *j*th dam, s_i is the effect of the *i*th sire, d_{ij} is the effect of the *j*th dam mated to the *i*th sire, and w_{ijk} is the residual deviation (the within-full-sib family deviations). As usual, under the assumption that individuals are random members of the same population, the s_i, d_{ij}, and w_{ijk} are defined to be independent random variables with expectations equal to zero. It then follows that the total phenotypic variance is

$$\sigma_z^2 = \sigma_s^2 + \sigma_d^2 + \sigma_w^2 \tag{5.13b}$$

where σ_s^2 is the variance among sires, σ_d^2 the variance among dams within sires, and σ_w^2 the variance within full-sib families.

To relate the observable components of variance to covariances between relatives, first note that the total phenotypic variance can be partitioned into two components, the variance within- and among- full-sib families. Since the variance among groups is equivalent to the covariance of members within groups, the variance among full-sib families equals the phenotypic covariance of full sibs, $\sigma(\text{FS})$. Thus, the variance within full-sib families (the residual variance in the model) is simply

$$\sigma_w^2 = \sigma_z^2 - \sigma(\text{FS}) \tag{5.14a}$$

Similarly, the variance among sires is equivalent to the covariance of individuals with the same father but different mothers, i.e., the covariance of paternal half sibs,

$$\sigma_s^2 = \sigma(\text{PHS}) \tag{5.14b}$$

Since the three components of variance must sum to σ_z^2, the dam variance is found to be

$$\begin{aligned} \sigma_d^2 &= \sigma_z^2 - \sigma_s^2 - \sigma_w^2 \\ &= \sigma(\text{FS}) - \sigma(\text{PHS}) \end{aligned} \tag{5.14c}$$

Recalling the genetic/environmental interpretations of $\sigma(\text{PHS})$ and $\sigma(\text{FS})$ gives

$$\sigma_s^2 \simeq \frac{\sigma_A^2}{4} \tag{5.15a}$$

$$\sigma_d^2 \simeq \frac{\sigma_A^2}{4} + \frac{\sigma_D^2}{4} + \sigma_{E_c}^2 \tag{5.15b}$$

$$\sigma_w^2 \simeq \frac{\sigma_A^2}{2} + \frac{3\sigma_D^2}{4} + \sigma_{E_s}^2 \tag{5.15c}$$

where $\sigma_{E_c}^2$ is the component of variance due to common family environmental effects, and $\sigma_{E_s}^2$ the remaining environmental variation. An obvious problem with this set of equations is that they are overdetermined — there are four causal sources of variance ($\sigma_A^2, \sigma_D^2, \sigma_{E_c}^2, \sigma_{E_c}^2$) but only three observable variance components ($\sigma_s^2, \sigma_d^2, \sigma_w^2$). We will deal with this in the worked example below.

The variance-component estimators are given by,

$$\text{Var}(s) = \frac{\text{MS}_s - \text{MS}_d}{Mn} \tag{5.16a}$$

$$\text{Var}(d) = \frac{\text{MS}_d - \text{MS}_w}{n} \tag{5.16b}$$

$$\text{Var}(e) = \text{MS}_w \tag{5.16c}$$

while the intraclass correlations for paternal half sibs and full sibs are

$$t_{\text{PHS}} = \frac{\text{Cov(PHS)}}{\text{Var}(z)} = \frac{\text{Var}(s)}{\text{Var}(z)} \tag{5.16d}$$

$$t_{\text{FS}} = \frac{\text{Cov(FS)}}{\text{Var}(z)} = \frac{\text{Var}(s) + \text{Var}(d)}{\text{Var}(z)} \tag{5.16e}$$

$4t_{\text{PHS}}$ provides the best estimate of h^2 since it is not inflated by dominance and/or common environmental effects. If, however, $\text{Var}(s)$ and $\text{Var}(d)$ are found to be approximately equal, then dominance and maternal effects can be ruled out as significant causal sources of covariance.

Table 5.3: Summary of a (balanced) nested analysis of variance involving N sires, M dams per sire and and n offspring per dam. $T = MNn$ is the total number of sibs in the design.

Factor	df	Sums of Squares	MS	E(MS)
Sires	$N-1$	$Mn\sum_{i=1}^{N}\sum_{j=1}^{M_i}(\overline{z}_i - \overline{z})^2$	SS_s/df_s	$\sigma_w^2 + n\sigma_d^2 + Mn\sigma_s^2$
Dams (sires)	$N(M-1)$	$n\sum_{i=1}^{N}\sum_{j=1}^{M}(\overline{z}_{ij} - \overline{z}_i)^2$	SS_d/df_d	$\sigma_w^2 + n\sigma_d^2$
Sibs (dams)	$T-NM$	$\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{n}(z_{ijk} - \overline{z}_{ij})^2$	SS_w/df_e	σ_w^2

Worked Example of a Nested Design

Table 5.4:. Suppose $N = 10$ sires are each crossed to $M = 3$ dams and $n = 10$ offspring are measured in each full-sib family, with resulting ANOVA table

Factor	df	SS	MS	E(MS)
Sires	9	$SS_s = 4,230$	470	$\sigma_w^2 + 10\sigma_d^2 + 30\sigma_s^2$
Dams (sires)	20	$SS_d = 3,400$	170	$\sigma_w^2 + 10\sigma_d^2$
Within Dams	270	$SS_w = 5,400$	20	σ_w^2

$$\begin{aligned}
\sigma_w^2 &= MS_w = 20 \\
\sigma_d^2 &= \frac{MS_d - MS_w}{n} = \frac{170 - 20}{10} = 15 \\
\sigma_s^2 &= \frac{MS_s - MS_d}{Nn} = \frac{470 - 170}{30} = 10 \\
\sigma_P^2 &= \sigma_s^2 + \sigma_d^2 + \sigma_w^2 = 45
\end{aligned}$$

Hence

$$\sigma_A^2 = 4\sigma_s^2 = 40$$

and

$$h^2 = \frac{\sigma_A^2}{\sigma_z^2} = \frac{40}{45} = 0.89$$

Likewise, since

$$\sigma_d^2 = 15 = (1/4)\sigma_A^2 + (1/4)\sigma_D^2 + \sigma_{E_c}^2 = 10 + (1/4)\sigma_D^2 + \sigma_{E_c}^2$$

we are left with the estimate of the linear combination

$$\sigma_D^2 + 4\sigma_{E_c}^2 = 20$$

Hence, if $\sigma_D^2 = 0$, then $\sigma_{E_c}^2 = 5$, while if $\sigma_{E_c}^2 = 0$, then $\sigma_D^2 = 20$. These represent the extreme values of these two variance components consistent with the ANOVA.

Estimation: Parent-Offspring Regressions

In some sense the simplest design is the parent-offspring regression, the regression of offspring phenotype z_o given the phenotypic value of one of its parents, z_p. Here the linear model is

$$z_{o_i} = \alpha + b_{o|p} z_{p_i} + e_i = \mu + b_{o|p}(z_{p_i} - \mu) + e_i$$

The alternative formulation follows since the regression passes through the mean of both variables (offspring and parental phenotypes).

The expected regression slope $b_{o|p}$ is

$$E(b_{o|p}) = \frac{\sigma(z_o, z_p)}{\sigma^2(z_p)} \simeq \frac{(\sigma_A^2/2) + \sigma(E_o, E_p)}{\sigma_z^2} \tag{5.17}$$

For males, it is generally expected that the covariance between parent and offspring environmental values is zero and the regression slope is $h^2/2$. This is not necessarily the case for females, as one can imagine how a larger female could better provision her offspring, leading to larger offspring, creating a positive environmental covariance. For this reason, single parent-offspring regressions usually

involve the fathers, although if the regression slopes for father-offspring and mother-offspring are the same, we can rule out shared mother-offspring environmental values. Thus a simple (possibly biased) estimate of $h^2 = \sigma_A^2/\sigma_z^2$ is twice the (single) parent-offspring regression, $2b_{|o|p}$.

Greater precision is possible when both parents can be measured, in which case one can regress offspring phenotypes on the mean phenotypes of their parents (also known as the **midparent values**). The linear model is now

$$z_{oi} = \mu + b_{o|MP}\left(\frac{z_{mi} + z_{fi}}{2} - \mu\right) + e_i$$

where z_{mi} and z_{fi} refer to the phenotypes of mothers and fathers. The slope $b_{o|MP}$ is a direct estimate of the heritability. To see this, note that

$$\begin{aligned} b_{o\|MP} &= \frac{\text{Cov}[z_o, (z_m + z_f)/2]}{\text{Var}[(z_m + z_f)/2]} \\ &= \frac{[\text{Cov}(z_o, z_m) + \text{Cov}(z_o, z_f)]/2}{[\text{Var}(z) + \text{Var}(z)]/4} \\ &= \frac{2\text{Cov}(z_o, z_p)}{\text{Var}(z)} = 2b_{o|p} \end{aligned} \tag{5.18}$$

What happens when multiple (n) offspring are measured in each family? The expected phenotypic covariance of a parent i and the average of its $j = 1, \cdots, n$ offspring may be written $\sigma[(\sum_{j=1}^{n} z_{oij}/n), z_p]$. Since all n of the covariance terms have the same expected value, this reduces to $n\sigma(z_o, z_p)/n = \sigma(z_o, z_p)$, the same as the expectation for single offspring. Thus, provided family sizes are equal, the interpretation of a single parent-offspring regression is the same whether individual offspring data or the progeny means are used in the analysis.

The sampling variance of the regression of a single parent on its (n) offspring is approximately

$$\text{Var}(b_{o|p}) \simeq \frac{n(t - b_{o|p}^2) + (1 - t)}{Nn} \tag{5.19a}$$

where N is the number of parent-offspring pairs and t is the covariance between sibs. Since sibs in a single-parent regression can potentially be either full- or half-sibs,

$$t = \begin{cases} t_{HS} = h^2/4 & \text{for half-sibs} \\ t_{FS} = h^2/2 + \dfrac{\sigma_D^2 + \sigma_{E_c}^2}{\sigma_z^2} & \text{for full sibs} \end{cases}$$

Since h^2 is estimated as $2b_{o|p}$,

$$\text{Var}(h^2) = \text{Var}(2b_{o|p}) = 4\text{Var}(b_{o|p})$$

Likewise, the sampling variance for a midparent-offspring regression with N parent-offspring pairs and n offspring per set of parents is approximately

$$\text{Var}(b_{o|MP}) \simeq \frac{2[n(t_{FS} - b_{o|MP}^2/2) + (1 - t_{FS})])}{Nn} \tag{5.19b}$$

Since h^2 is estimated by $b_{o|MP}$, $\text{Var}(h^2)$ is given by Equation 5.19b.

Natural Population Parent-Laboratory Offspring Regressions

Riska et al. (1989) have shown that a lower bound, h^2_{min}, to the heritability in the field can be estimated by regressing the phenotypes of lab-reared progeny on their field-reared parents. Let the regression coefficient involving wild midparents and lab-reared offspring be $b'_{o\overline{p}}$, the phenotypic variance of the natural population be $\mathrm{Var}_n(z)$, and the additive genetic variance in the laboratory environment (obtained either from the covariance of lab-reared sibs or of lab-reared parents and offspring) be $\mathrm{Var}_l(A)$. Then,

$$h^2_{min} = (b'_{o\overline{p}})^2 \frac{\mathrm{Var}_n(z)}{\mathrm{Var}_l(A)} = \left[\frac{\mathrm{Cov}_{l,n}(A)}{\mathrm{Var}_n(z)}\right]^2 \frac{\mathrm{Var}_n(z)}{\mathrm{Var}_l(A)} \tag{5.20a}$$

where $\mathrm{Cov}_{l,n}(A)$ is the additive genetic covariance between the trait as expressed in the wild and in the lab. To see that this provides a lower bound, define

$$\gamma = \frac{\mathrm{Cov}_{l,n}(A)}{\sqrt{\mathrm{Var}_n(A)\mathrm{Var}_l(A)}} \tag{5.20b}$$

to be the additive genetic correlation between environments . The expected value of h^2_{min} is then $\gamma^2 h^2_n$, which is necessarily $\leq h^2_n$, the heritability in the wild. h^2_{min} is an unbiased estimate of h^2_n only if the genetic correlation across environments is equal to one.

Estimation of σ^2_A and Breeding Values in General Pedigrees

The above simple designs (balanced sib families, only a single type of relative) do not really describe the sort of data that is typically gathered, which often includes highly unbalanced designs (for example, big differences in the number of sibs across families), collections of several different types of relatives, and the possibility of a number of confounding fixed factors (such as differences in the mean value for sexes, location effects, etc.). The general way to handle estimation over a complex pedigree is the methods of BLUP and REML, with BLUP predicting the breeding values, and REML used to estimate the variances. As an introduction to these topics (which are covered in gory detail in Chapters 26 and 27 of Lynch and Walsh), we first start with some comments on the general mixed model.

The General Mixed Model

Consider a column vector $\mathbf{y}$ containing the phenotypic values for a trait measured in n individuals. We assume that these observations are described adequately by a linear model with a $p \times 1$ vector of fixed effects ($\boldsymbol{\beta}$) and a $q \times 1$ vector of random effects ($\mathbf{u}$). The first element of the vector $\boldsymbol{\beta}$ is typically the population mean, and other factors included may be gender, location, year of birth, experimental treatment, and so on. The elements of the vector $\mathbf{u}$ of random effects are usually genetic effects such as additive genetic values. In matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{5.21}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are respectively $n \times p$ and $n \times q$ **incidence matrices** ($\mathbf{X}$ is also called the **design matrix**), and $\mathbf{e}$ is the $n \times 1$ column vector of residual deviations assumed to be distributed independently of the random genetic effects. Usually, all of the elements of the incidence matrices are equal to 0 or 1, depending upon whether the relevant effect contributes to the individual's phenotype. Because this model jointly accounts for fixed and random effects, it is generally referred to as a **mixed model**.

Example 1. Suppose that three sires are chosen at random from a population, and each mated to a randomly chosen dam. Two offspring from each mating are evaluated, some in environment 1

and some in environment 2. Let y_{ijk} denote the phenotypic value of the kth offspring of sire i in environment j. The model is then

$$y_{ijk} = \beta_j + u_i + e_{ijk}$$

This model has three random effects (u_1, u_2, u_3), which measure the contribution from each sire, and two fixed effects (β_1, β_2), which describe the influence of the two environments. The model assumes an absence of sire $\times$ environment interaction.

As noted above, a total of six offspring were measured. One offspring of sire 1 was assigned to environment 1 and had phenotypic value $y_{1,1,1} = 9$, while the second offspring was assigned to environment 2 and had phenotypic value $y_{1,2,1} = 12$. The two offspring of sire 2 were both assigned to environment 1 and had values of $y_{2,1,1} = 11$ and $y_{2,1,2} = 6$. One offspring of sire 3 was assigned to environment 1 and had phenotypic value $y_{3,1,1} = 7$, while the second offspring was assigned to environment 2 and had phenotypic value $y_{3,2,1} = 14$. The resulting vector of observations can be written as

$$\mathbf{y} = \begin{pmatrix} y_{1,1,1} \\ y_{1,2,1} \\ y_{2,1,1} \\ y_{2,1,2} \\ y_{3,1,1} \\ y_{3,2,1} \end{pmatrix} = \begin{pmatrix} 9 \\ 12 \\ 11 \\ 6 \\ 7 \\ 14 \end{pmatrix}$$

giving the mixed model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where the incidence matrices for fixed and random effects and the vectors of these effects are respectively

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Now consider the means and variances of the component vectors of the mixed model. Since $E(\mathbf{u}) = E(\mathbf{e}) = \mathbf{0}$ by definition, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Denote the $(n \times n)$ covariance matrix for the vector $\mathbf{e}$ of residual errors by $\mathbf{R}$ and the $(q \times q)$ covariance matrix for the vector $\mathbf{u}$ of random genetic effects by $\mathbf{G}$. Excluding the difference among individuals due to fixed effects, the assumption that $\mathbf{u}$ and $\mathbf{e}$ are uncorrelated gives the covariance matrix for the vector of observations $\mathbf{y}$ as

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \tag{5.22}$$

The first term accounts for the contribution from random genetic effects, while the second accounts for the variance due to residual effects. We will generally assume that residual errors have constant variance and are uncorrelated, so that $\mathbf{R}$ is a diagonal matrix, with $\mathbf{R} = \sigma_E^2 \mathbf{I}$.

We are now in a position to contrast the mixed model and the general linear model. Under the general linear model (Lecture 2),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^* \qquad \text{where} \quad \mathbf{e}^* \sim (\mathbf{0}, \mathbf{V}) \quad \text{implying} \ \ \mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where the notation $\sim (a, b)$ means that the random variable has mean a and variance b. On the other hand, the mixed model partitions the vector of residual effects into two components, with $\mathbf{e}^* = \mathbf{Z}\mathbf{u} + \mathbf{e}$, giving

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad \text{where} \quad \mathbf{u} \sim (\mathbf{0}, \mathbf{G}) \ \text{ and } \mathbf{e} \sim (\mathbf{0}, \mathbf{R})$$

$$\text{implying } \mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) = (\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})$$

When analyzed in the appropriate way, both formulations yield the same estimate of the vector of fixed effects $\boldsymbol{\beta}$, while the mixed-model formulation further allows estimates of the vector of random effects $\mathbf{u}$.

For the mixed model, we observe $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, while $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{R}$, and $\mathbf{G}$ are generally unknown. Thus, mixed-model analysis involves two complementary estimation issues: (1) estimation of the vectors of fixed and random effects, $\boldsymbol{\beta}$ and $\mathbf{u}$, and (2) estimation of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$. These covariance matrices are generally assumed to be functions of a few unknown variance components.

Estimating Fixed Effects and Predicting Random Effects

The primary goal of a quantitative-genetic analysis is often solely to estimate variance components. However, there are also numerous situations in which inferences about fixed effects (such as the effect of a particular environment or year) and/or random effects (such as the breeding value of a particular individual) are the central motivation. Inferences about fixed effects have come to be called **estimates,** whereas those that concern random effects are known as **predictions.** Procedures for obtaining such estimators and predictors have been developed using a variety of approaches, such as likelihood theory (Lecture 1, LW Appendix 4). The most widely used procedures are BLUE and BLUP, referring respectively to **best linear unbiased estimator** and **best linear unbiased predictor.** They are *best* in the sense that they minimize the sampling variance, *linear* in the sense that they are linear functions of the observed phenotypes $\mathbf{y}$, and *unbiased* in the sense that $E[\,\text{BLUE}(\boldsymbol{\beta})\,] = \boldsymbol{\beta}$ and $E[\,\text{BLUP}(\mathbf{u})\,] = \mathbf{u}$.

For the mixed model given by Equation 5.21, the BLUE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{5.23a}$$

with $\mathbf{V}$ as given by Equation 5.22. Notice that this is just the generalized least-squares (GLS) estimator discussed in Lecture 2. Henderson (1963) showed that the BLUP of $\mathbf{u}$ is

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{5.23b}$$

which is equivalent to the conditional expectation of $\mathbf{u}$ given $\mathbf{y}$ under the assumption of multivariate normality. As noted above, the practical application of both of these expressions requires that the variance components be known. Thus, prior to a BLUP analysis, the variance components need to be estimated by ANOVA or REML.

Example 2. What are the BLUP values for the sire effects (u_1, u_2, u_3) in Example 1? In order to proceed, we require the covariance matrices for sire effects and errors. We assume that the residual variances within both environments are the same (σ^2_E), so $\mathbf{R} = \sigma^2_E\,\mathbf{I}$, where $\mathbf{I}$ is the 6×6 identity matrix. Assuming that all three sires are unrelated and drawn from the same population, $\mathbf{G} = \sigma^2_S\,\mathbf{I}$, where $\mathbf{I}$ is the 3×3 identity matrix and σ^2_S is the variance of sire effects. Assuming only additive genetic variance, the sire effects (breeding values) are half the sires' additive genetic values. Thus, since the sires are sampled randomly from an outbred base population, $\sigma^2_S = \sigma^2_A/4$, where σ^2_A is the additive genetic variance. Assuming that $\sigma^2_A = 8$ and $\sigma^2_E = 6$, the covariance matrix $\mathbf{V}$ for the vector of observations $\mathbf{y}$ is given by $\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$, or

$$\mathbf{V} = \frac{8}{4}\begin{pmatrix}1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix} + 6\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}$$

$$= \begin{pmatrix} 8 & 2 & 0 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & 0 & 2 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{pmatrix} \quad \text{giving} \quad \mathbf{V}^{-1} = \frac{1}{30} \cdot \begin{pmatrix} 4 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{pmatrix}$$

Using this result, a few simple matrix calculations give

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \end{pmatrix}$$

and

$$\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \frac{1}{18}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Note that the solution of Equations 5.23a and 5.23b requires the inverse of the covariance matrix $\mathbf{V}$. In the preceding example, $\mathbf{V}^{-1}$ was not particularly difficult to obtain. However, when $\mathbf{y}$ contains many thousands of observations, as is commonly the case in cattle breeding, the computation of $\mathbf{V}^{-1}$ can be quite difficult. As a way around this problem, Henderson (1950, 1963, 1973, 1984) offered a more compact method for jointly obtaining $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ in the form of his **mixed-model equations** (MME),

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{5.24}$$

While these expressions may look considerably more complicated than Equations 5.23a and 5.23b, $\mathbf{R}^{-1}$ and $\mathbf{G}^{-1}$ are trivial to obtain if $\mathbf{R}$ and $\mathbf{G}$ are diagonal, and hence the submatrices in Equation 5.24 are much easier to compute than $\mathbf{V}^{-1}$. A second advantage of Equation 5.24 can be seen by considering the dimensionality of the matrix on the left. Recalling that $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ respectively, $\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}$ is $p \times p$, $\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}$ is $p \times q$, and $\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}$ is $q \times q$. Thus, the matrix that needs to be inverted to obtain the solution for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ is of order $(p+q) \times (p+q)$, which is usually considerably less than the dimensionality of $\mathbf{V}$ (an $n \times n$ matrix).

Example 3. Using the values from Examples 1 and 2, we find that

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} = \frac{1}{6}\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} = \left(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\right)^T = \frac{1}{6}\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{G}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} = \frac{5}{6}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 33 \\ 26 \end{pmatrix}, \quad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 21 \\ 17 \\ 21 \end{pmatrix}$$

Thus, after factoring out $1/6$ from both sides, the mixed-model equations for these data become

$$\begin{pmatrix} 4 & 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 1 & 5 & 0 & 0 \\ 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix}$$

Taking the inverse gives the solution

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \frac{1}{270} \begin{pmatrix} 100 & 25 & -25 & -40 & -25 \\ 25 & 175 & -40 & -10 & -40 \\ -25 & -40 & 67 & 10 & 13 \\ -40 & -10 & 10 & 70 & 10 \\ -25 & -40 & 13 & 10 & 67 \end{pmatrix} \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix} = \frac{1}{18} \begin{pmatrix} 148 \\ 235 \\ -1 \\ 2 \\ -1 \end{pmatrix}$$

which is identical to the results obtained in Example 2.

The Animal Model

The basic BLUP model used in most analysis is the animal model (or individual model if you perfer), in which we consider individuals as the unit of analysis. Assuming only a single fixed factor (the population mean) under the simplest animal model, the observation for individual *i* is expressed as

$$y_i = \mu + a_i + e_i \tag{5.25}$$

where a_i is the additive genetic value of individual *i*. With k individuals, the model can be expressed as in Equation 5.23a with

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \mu, \qquad \mathbf{u} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

The matrix **G** describing the covariances among the random effects (here the breeding values) follows from standard results for the covariances between relatives. The additive genetic covariance between two relatives *i* and *j* is given by $2\Theta_{ij}\sigma_A^2$, i.e., by twice the coefficient of coancestry times the additive genetic variance in the base population. Hence, under the animal model, $\mathbf{G} = \sigma_A^2\,\mathbf{A}$, where the **additive genetic** (or **numerator**) **relationship matrix A** has elements $A_{ij} = 2\Theta_{ij}$.

The covariance matrix **R** for the vector of residual errors requires a little more care. The standard assumption is that $\mathbf{R} = \sigma_E^2\,\mathbf{I}$, so that the residual error for each observation has the same variance σ_E^2 and is uncorrelated with all other residual errors. There are many ways in which this assumption can fail. For example, if the character displays any dominance and *i* and *j* are full sibs, $\sigma(e_i, e_j) = \sigma_D^2/4$. Shared environmental effects can also cause correlations between residual effects.

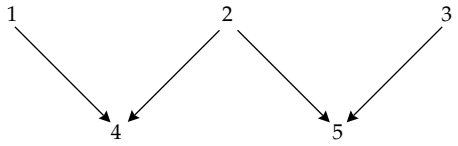
Since $\mathbf{G}^{-1} = \sigma_A^{-2}\,\mathbf{A}^{-1}$, the mixed-model equations (Equation 5.25) for the animal model reduce to

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix} \tag{5.26a}$$

where $\lambda = \sigma_E^2/\sigma_A^2 = (1-h^2)/h^2$ under the assumption of additive gene action. Since the only fixed factor is the mean μ (so that $\boldsymbol{\beta} = \mu$ and $\mathbf{X} = \mathbf{1}$, a vector of ones) and each individual has only a single observation (so that $\mathbf{Z} = \mathbf{I}$), with n individuals, Equation 5.26a reduces to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \sum^n y_i \\ \mathbf{y} \end{pmatrix} \tag{5.26b}$$

Example 5. Consider the pedigree of individuals given in the figure below, where each individual has a single measurement and the only fixed factor is the mean.



With the vector of observations,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

we can use Equation 5.26b with $\widehat{\mathbf{u}}^T = (\widehat{a}_1,\, \widehat{a}_2,\, \widehat{a}_3,\, \widehat{a}_4,\, \widehat{a}_5)$. Assuming that individuals 1, 2, and 3 are unrelated and not inbred, the relationship matrix becomes

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/4 \\ 0 & 1/2 & 1/2 & 1/4 & 1 \end{pmatrix}$$

Suppose it is known that $\sigma_E^2 = \sigma_A^2$, so that $\lambda = 1$. Then,

$$\mathbf{I} + \lambda\, \mathbf{A}^{-1} = \begin{pmatrix} 5/2 & 1/2 & 0 & -1 & 0 \\ 1/2 & 3 & 1/2 & -1 & -1 \\ 0 & 1/2 & 5/2 & 0 & -1 \\ -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & 0 & 3 \end{pmatrix}$$

Since $n = 5$ and $\sum y_i = 41$, Equation 5.26b gives the mixed-model equations for these data as

$$\begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 5/2 & 1/2 & 0 & -1 & 0 \\ 1 & 1/2 & 3 & 1/2 & -1 & -1 \\ 1 & 0 & 1/2 & 5/2 & 0 & -1 \\ 1 & -1 & -1 & 0 & 3 & 0 \\ 1 & 0 & -1 & -1 & 0 & 3 \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \end{pmatrix} = \begin{pmatrix} 41 \\ 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

the solutions of which are

$$\widehat{\mu} = \frac{440}{53} \simeq 8.302, \qquad \begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \end{pmatrix} = \begin{pmatrix} -662/689 \\ 4/53 \\ 610/689 \\ -732/689 \\ 381/689 \end{pmatrix} \simeq \begin{pmatrix} -0.961 \\ 0.076 \\ 0.885 \\ -1.062 \\ 0.553 \end{pmatrix}$$

Note that the average breeding value in the base population (individuals 1, 2, and 3) is zero (as expected for a random sample of the population). This is no longer the case once we leave the base population, unless all base-population individuals contribute equally to progeny production.

ANOVA vs. REML Variance Estimates

Typically in analysis of variance (ANOVA), variance components were estimated by equating observed mean squares to expressions describing their expected values, these being functions of the

variance components. ANOVA has the nice feature that the estimators for the variance components are unbiased regardless of whether the data are normally distributed, but it also has two significant limitations. First, field observations often yield records on a variety of relatives, such as offspring, parents, or sibs, that cannot be analyzed jointly with ANOVA. Second, ANOVA estimates of variance components require that sample sizes be well balanced, with the number of observations for each set of conditions being essentially equal. In field situations, individuals are often lost, and even the most carefully crafted balanced design can quickly collapse into an extremely unbalanced one. Although modifications to the ANOVA sums of squares have been proposed to account for unbalanced data (Henderson 1953, Searle et al. 1992), their sampling properties are poorly understood.

Unlike ANOVA estimators, maximum likelihood (ML) and restricted maximum likelihood (REML) estimators do not place any special demands on the design or balance of data. Such estimates are ideal for the unbalanced designs that arise in quantitative genetics, as they can be obtained readily for any arbitrary pedigree of individuals. Since many aspects of ML and REML estimation are quite difficult technically, the detailed mathematics can obscure the general power and flexibility of the methods.

ML Versus REML Variance Estimates

Although algebraically tedious, maximum likelihood (ML) is conceptually very simple. It was introduced to variance component-estimation by Hartley and Rao (1967). For a specified model, such as Equation 5.23a, and a specified form for the joint distribution of the elements of $\mathbf{y}$, ML estimates the parameters of the distribution that maximize the likelihood of the observed data. This distribution is almost always assumed to be multivariate normal. An advantage of ML estimators is their efficiency — they simultaneously utilize all of the available data and account for any nonindependence.

One drawback with variance-component estimation via the usual maximum likelihood approach is that all fixed effects are assumed to be known without error. This is rarely true in practice, and as a consequence, ML estimators yield biased estimates of variance components. Most notably, estimates of the residual variance tend to be downwardly biased. This bias occurs because the observed deviations of individual phenotypic values from an estimated population mean tend to be smaller than their deviations from the true (parametric) mean. Such bias can become quite large when a model contains numerous fixed effects, particularly when sample sizes are small.

Unlike ML estimators, restricted maximum likelihood (REML) estimators maximize only the portion of the likelihood that does not depend on the fixed effects. In this sense, REML is a *restricted* version of ML. The elimination of bias by REML is analogous to the removal of bias that arises in the estimate of a variance component when the mean squared deviation is divided by the degrees of freedom instead of by the sample size. REML does not always eliminate all of the bias in parameter estimation, since many methods for obtaining REML estimates cannot return negative estimates of a variance component. However, this source of bias also exists with ML, so REML is clearly the preferred method for analyzing large data sets with complex structure. In the ideal case of a completely balanced design, REML yields estimates of variance components that are identical to those obtained by classical analysis of variance.

As an example of the differences between ML and REML estimates of a variance, consider the simplest case of n values from a normal with unknown mean and variance. The ML estimate of the variance is just

$$\frac{1}{n}\sum_{i+1}^{n}(x-\overline{x})^2$$

while the REML estimate is

$$\frac{1}{n-1}\sum_{i+1}^{n}(x-\overline{x})^2$$

Lecture 5 Problems

1. Consider a simple full-sib design. Suppose $\sigma^2_A = 30$, $\sigma^2_D = 10$, $\sigma^2_{Ec} = 5$ and $\sigma^2_E - \sigma^2_{Ec} = 5$.

   a: What is σ^2_f? σ^2_w?

   b: Now consider a nested full-sib design. What are the sire (σ^2_s) and dam (σ^2_d) variances?

2. Recalling Equation 5.7b, what is the variance of our estimate of σ^2_A for the worked full-sib problem?

3. Create your own ANOVA! Consider a strictly half-sib analysis, wherein each of N sires are mated to n dams, each of which leaves exactly one offspring (an example of this is beef or dairy cattle). Under this model, the ijth observation is the jth offspring from sire i and is the sum of a sire effect s_i and a within (half-sib) family deviation from the sire effect.

   a: What is the linear model for this design?

   b: Express the sire σ^2_s and within-family $\sigma^2_{w(HS)}$ variance in terms of the genetic and environmental variance components.

   c: What would the resulting ANOVA table look like? (i.e., what are the required sums of squares, the associated degrees of freedom, the mean squares, and the expected value of the mean squares expressed in terms of the genetic and environmental variance components).

4. Make your own animal model! You have 7 measured individuals, $y_1, \cdots, y_7$. individual 1 is the father and individual 2 the mother to individuals 3 (male) and 4 (female). Individual five is related to 1-4, and is the mother of individuals 6 (male) and 7 (female). Add a fixed effect due to sex, with the mean of males being β_m, while the mean for females is β_f. Assume residuals for the fixed factors are independent and homoscedatic with variance σ^2_e (note that this assumes no dominance nor shared maternal effects among sibs). Write out the animal model (in matrix form, including the covariance matrices) for this case.

Solution to Lecture 5 Problems

1: a: $\sigma^2_f = \sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{E_c} = 30/2 + 10/4 + 5 = 22.5$.
$\sigma^2_w = \sigma^2_z - \sigma^2_f$. Here $\sigma^2_z = 30 + 10 + 5 + 5 = 50$, giving $\sigma^2_w = 50 - 22.5 = 27.5$

b: The within-sire variance $\sigma^2_s = \sigma^2_A/4 = 30/4 = 7.5$, while (Equation 5.10a) $\sigma^2_d = Cov(FS) - \sigma^2_s = \sigma^2_f - \sigma^2_s = 22.5 - 7.5 = 15$.

2:

$$\text{Var}[\,\text{Var}(A)\,] = \text{Var}[\,2\text{Var}(f)\,] \simeq 2^2\frac{2}{n^2}\left(\frac{(\text{MS}_f)^2}{N+1} + \frac{(\text{MS}_w)^2}{T-N+2}\right)$$

$$= 4\cdot\frac{2}{5^2}\left(\frac{(45)^2}{11} + \frac{(20)^2}{42}\right) = 61.96$$

Giving an standard error ($\sqrt{Var}$) of 7.87.

3:

a: $z_{ij} = s_i + w_{ij}$

b: $\sigma^2_s = \text{Cov(half-sibs)} = \sigma^2_A/4$, $\quad \sigma^2_w = \sigma^2_z - \sigma^2_s = (3/4)\sigma^2_A + \sigma^2_D + \sigma^2_E$

c:

Factor	df	SS	MS	E(MS)
Between-sires	$N-1$	$\text{SS}_s = n\sum_{i=1}^{N}(\overline{z}_i - \overline{z})^2$	$\text{SS}_s/(N-1)$	$\sigma^2_w + n\sigma^2_s$
Within-sire families	$T-N$	$\text{SS}_w = \sum_{i=1}^{N}\sum_{j=1}^{n}(z_{ij} - \overline{z}_i)^2$	$\text{SS}_w/(T-N)$	σ^2_w

where $T = Nn$

4: Letting A_i be the breeding value of individual i, the animal model becomes

$$\begin{pmatrix} y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\\ y_7 \end{pmatrix} = \begin{pmatrix} 0 & 1\\ 1 & 0\\ 0 & 1\\ 1 & 0\\ 1 & 0\\ 0 & 1\\ 1 & 0 \end{pmatrix}\begin{pmatrix} \beta_f\\ \beta_m \end{pmatrix} + \begin{pmatrix} A_1\\ A_2\\ A_3\\ A_4\\ A_5\\ A_7\\ A_6 \end{pmatrix} + \begin{pmatrix} e_1\\ e_2\\ e_3\\ e_4\\ e_5\\ e_6\\ e_7 \end{pmatrix}$$

$$\mathbf{V} = \sigma^2_A\mathbf{A} + \sigma^2_e\mathbf{I}, \quad \text{where} \quad \mathbf{A} = \begin{pmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 0 & 0\\ 0 & 1 & 1/2 & 1/2 & 0 & 0 & 0\\ 1/2 & 1/2 & 1 & 1/2 & 0 & 0 & 0\\ 1/2 & 1/2 & 1/2 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1/2 & 1 & 1/4\\ 0 & 0 & 0 & 0 & 1/2 & 1/4 & 1 \end{pmatrix}$$