

Lecture 9

Analysis of Short-term Selection Experiments

Bruce Walsh. Aug 2003. Nordic Summer Course

Variance in Response

While the breeders' equation predicts the *expected* response, it is important to realize that there is considerable *variation* in response among otherwise identical replicate lines. To examine the various sources contributing to the variance in response, write the population mean in generation t as

$$\bar{z}_t = \mu + g_t + d_t + e_t \quad (9.1)$$

where g_t is the mean breeding value and d_t the mean environmental deviation in generation t . Under this model, a series of lines initiated simultaneously have expected value

$$E(\bar{z}_t) = \mu + E(g_t) + d_t \quad (9.2)$$

The variance of \bar{z}_t around its expected value $E(g_t)$ in this particular generation is $\sigma_g^2(t) + \sigma_e^2(t)$. However, we usually don't know the particular value of d_t , which is assumed to be uncorrelated between generations ($\sigma(d_t, d_{t'}) = 0$), with mean 0 and variance σ_d^2 . Thus, the expected value of \bar{z}_t measured at a random time (e.g., the value of d is chosen at random) is $\mu + E(g_t)$ and the variance of \bar{z}_t about this mean value is inflated by σ_d^2 to give

$$\sigma_{\bar{z}}^2(t) = \sigma_g^2(t) + \sigma_e^2(t) + \sigma_d^2 \quad (9.3)$$

If M_0 individuals are initially sampled to form each line, then

$$\sigma_g^2(t) = \left(\frac{1}{M_0} + 2f_t \right) \sigma_A^2 = \left(\frac{1}{M_0} + 2f_t \right) h^2 \sigma_z^2 \quad (9.4)$$

where f_t is the amount of inbreeding at generation t . The M_0 term accounts for variation in mean breeding value between lines in the founding generation while the f_t term accounts for variation generated by subsequent drift.

The variance of e_t is more involved, with its exact form depending on the distribution of family sizes. To be conservative, we will use the upper bound

$$\sigma_e^2(t) = \frac{\sigma_z^2}{M_t} \quad (9.5)$$

Equation 9.3 thus becomes

$$\sigma_{\bar{z}}^2(t) = \left(\frac{1}{M_0} + 2f_t \right) h^2 \sigma_z^2 + \sigma_d^2 + \sigma_z^2/M_t \quad (9.6)$$

Since drift variance accumulates each generation (via f_t increasing each generation) while the other terms do not, drift is expected to dominate, usually after a few generations. If population size remains constant,

$$2f_t = 2 \left[1 - \left(1 - \frac{1}{2N_e} \right)^t \right] \simeq t/N_e \quad \text{for } t/N_e \ll 1 \quad (9.7)$$

If different numbers of males (N_m) and females (N_f) are sampled and/or N varies over time,

$$2f_t \simeq \sum_{k=0}^{t-1} \left[\frac{1}{4N_m(k)} + \frac{1}{4N_f(k)} \right] \quad t > 0 \quad (9.8)$$

Equation 9.6 describes the divergence *between* lines due to drift. Drift also introduces a positive correlation between the means at different generations *within* a line. If the errors in estimating the mean breeding values in generations t and t' (g_t and $g_{t'}$) from the phenotypic means (\bar{z}_t and $\bar{z}_{t'}$) are uncorrelated and if between-generation environmental effects are uncorrelated, then

$$\sigma(g_t, g_{t'}) = \sigma(\bar{z}_t, \bar{z}_{t'}) = \left(\frac{1}{M_0} + 2f_t \right) \sigma_A^2 = \left(\frac{1}{M_0} + 2f_t \right) h^2 \sigma_z^2 \quad \text{for } t < t' \quad (9.9)$$

The assumption that σ_A^2 and σ_z^2 remain constant is a major one and can be violated in at least three ways. First, changes in the underlying allele frequencies can change the variance. If major alleles are segregating, large changes in the variance can occur within a few generations. Major alleles initially at low frequencies can result in a substantial increase in the between-line variance over that predicted by Equation 3.9, as these alleles are lost in some lines and increase in frequency in others. The net result being that different lines from the same base population can start with very different values of additive variance, increasing the variance in response. Second, directional selection generates negative gametic-phase disequilibrium, reducing the additive genetic variance within a line. Third (as mentioned above), inbreeding due to finite population size reduces additive genetic variance within a line by fixing alleles. In the absence of mutational input the expected additive genetic variance within a line in generation t is

$$\sigma_A^2(t) = \left(1 - \frac{1}{2N_e} \right)^t \sigma_A^2(0) \simeq \left(1 - \frac{t}{2N_e} \right) \sigma_A^2(0)$$

where $\sigma_A^2(0)$ is the variance in the base population. Provided $t/2N_e \ll 1$, the error introduced by ignoring the reduction in the within-line variance due to inbreeding is small. A final complication that we have ignored is that drift generates between-line variance in σ_A^2 itself, further inflating the between-line variance in means.

Realized Heritabilities

The breeders' equation immediately suggests that heritability can be estimated as the ratio of observed response to observed selection differential,

$$\hat{h}_r^2 = \frac{R}{S} \quad (9.10)$$

Estimates of heritability based on the response to selection are referred to as **realized heritabilities**, and we denote these estimates by \hat{h}_r^2 . While one can use this approach to estimate h^2 , any complication in predicting response using the breeders' equation will usually make \hat{h}_r^2 a biased estimator of h^2 . Turning this point around, however, suggests that one test for the success of the breeders' equation is to compare how close realized heritabilities are to heritabilities estimated from resemblance between relatives in the unselected base population. If the breeders' equation generally provides an accurate model of selection response, we expect these two different estimates to be similar (i.e., within sampling error).

Estimators for Several Generations of Selection

While the estimator given by Equation 9.10 is unambiguous for a single generation of selection, two different estimates have been proposed when several generations of selection are considered. Both are based on the **cumulative selection response** $R_C(t)$ and **cumulative selection differential** $S_C(t)$, which are defined by

$$S_C(t) = \sum_{i=1}^t S_i \quad (9.11a)$$

and

$$R_C(t) = \sum_{i=1}^t R_i \quad (9.11b)$$

where S_i and R_i are the (directional) selection differential and single-generation response for generation i ($\bar{z}_i^* - \bar{z}_i$ and $\bar{z}_{i+1} - \bar{z}_i$, respectively). Perhaps the most common multigeneration estimator of realized heritability is to use the slope of the unweighted (ordinary) least-squares (OLS) regression of cumulative response on cumulative selection differential,

$$R_C(t) = b_C S_C(t) + e_t \quad \text{for } t = 1, \dots, T \quad (9.12)$$

with $\hat{h}_r^2 = b_C$. Since the expected response is zero if there is no selection, the regression line is constrained to pass through the origin and hence lacks an intercept term. Recalling Lecture 5, the OLS estimator for the slope is $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Since the design matrix \mathbf{X} for the regression given by Equation 9.12 is just the vector of cumulative selection differentials \mathbf{S} and \mathbf{y} is the vector of cumulative responses \mathbf{R} , it follows that the OLS estimate of the slope is given by

$$\hat{b}_C(\text{OLS}) = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{R} = \frac{\sum_{i=1}^T S_C(i) \cdot R_C(i)}{\sum_{i=1}^T S_C^2(i)} \quad (9.13)$$

We will refer to this as the **OLS regression estimator** of the realized heritability.

An alternate estimator for multigenerational data is to simply consider the ratio of total response to total differential,

$$\hat{b}_T = \frac{R_C(T)}{S_C(T)} \quad (9.14)$$

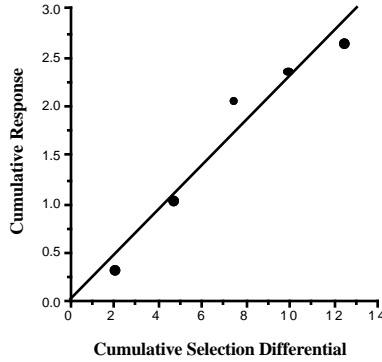
with $\hat{h}_r^2 = \hat{b}_T$. We refer to this as the **ratio estimator** of the realized heritability.

Example 9.1 Consider the following data from Mackay (1985), who performed a divergent selection experiment on abdominal bristle number in replicate lines of *Drosophila melanogaster*. Fifty males and fifty females were measured in each line, with ten of each sex selected to form the next generation. Her data for the High (up-selected) line from replicate pair 2 for the first five generations of selection are

t	\bar{z}	\bar{z}^*	$S(t)$	$R(t)$	$S_C(t)$	$R_C(t)$
1	18.02	20.10	20.10–18.02 = 2.08	18.34–18.02 = 0.32	2.08	0.32
2	18.34	21.00	21.00–18.34 = 2.66	19.05–18.34 = 0.71	4.74	1.03
3	19.05	21.75	21.75–19.05 = 2.70	20.07–19.05 = 1.02	8.44	2.05
4	20.07	22.55	22.55–20.07 = 2.48	20.36–20.07 = 0.29	9.92	2.34
5	20.36	22.95	22.95–20.36 = 2.59	20.65–20.36 = 0.29	12.51	2.63
6	20.65					

The ratio estimate of the realized heritability (Equation 9.14) is

$$\hat{h}_r^2 = \frac{2.63}{12.51} = 0.2102$$



The regression, forced through the origin, of cumulative response (R_C) on cumulative selection differential (S_C) is plotted above. Equation 9.13 gives the OLS regression estimator of the realized heritability as

$$\hat{h}_r^2 = \hat{b}_C(OLS) = \frac{\sum_{i=1}^5 S_C(i) \cdot R_C(i)}{\sum_{i=1}^5 S_C^2(i)} = \frac{78.96}{350.45} = 0.2245$$

Weighted Least-Squares Estimates of Realized Heritability

Ordinary least-squares regression assumes that the residuals are homoscedastic and uncorrelated, so that the covariance matrix for the vector of residuals \mathbf{e} is $\text{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$. However, genetic drift causes the covariance structure of the regression given by Equation 9.12 to depart significantly from this simple form. In particular, the residual variance increases with time (Equation 9.6) and residuals from different generations are correlated (Equation 9.9). Thus, $\text{Var}(\mathbf{e}) = \mathbf{V}$ and generalized least-squares (GLS) must be used to account for this covariance structure. The GLS estimator of the regression slope is given by

$$\hat{b}_C(\text{GLS}) = (\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{V}^{-1} \mathbf{R} \quad (9.15)$$

where \mathbf{V} is the variance-covariance matrix associated with selection response,

$$V_{ij} = \sigma_e(i, j) = \sigma [R_C(i), R_C(j)]$$

The elements of \mathbf{V} are the variances V_{ii} given by Equation 9.6 and covariances V_{ij} given by Equation 9.9. Even though OLS assumes an incorrect residual structure, it still provides an unbiased estimate of b_C . However, OLS *significantly underestimates the standard error* of the OLS estimator, and it is this reason that GLS estimators are greatly preferred and should be used when ever possible.

Example 9.2. To compute the GLS regression using the data from Example 9.1, we need to variance-covariance matrix of the residuals, which we obtain by using Equations 9.6 and 9.9. From Example 9.1, $M =$

100, $N = 20$, while the estimated phenotypic variance is $\text{Var}(z) = 3.293$ (Mackay, personal communication). Assuming that both initial sampling and between-generation environmental effects can be ignored (i.e., $M_0 \gg 1$ and $\sigma_d^2 \simeq 0$), then from Equations 9.6 and 9.7, the variance associated with the response in generation i is

$$\sigma^2 [R_C(i)] = \left(\frac{i}{N}\right) h^2 \sigma_z^2 + \frac{\sigma_z^2}{M} = i \cdot h^2 \cdot 0.1647 + 0.03292$$

Similarly, Equations 9.9 and 9.7 give for the covariance between generations as

$$\sigma [R_C(i), R_C(j)] = \left(\frac{i}{N}\right) h^2 \sigma_z^2 = i \cdot h^2 \cdot 0.1647 \quad \text{for } i < j$$

The resulting covariance matrix becomes

$$\mathbf{V} = 0.1647 \cdot \begin{pmatrix} h^2 + 0.2 & h^2 & h^2 & h^2 & h^2 \\ h^2 & 2h^2 + 0.2 & 2h^2 & 2h^2 & 2h^2 \\ h^2 & 2h^2 & 3h^2 + 0.2 & 3h^2 & 3h^2 \\ h^2 & 2h^2 & 3h^2 & 4h^2 + 0.2 & 4h^2 \\ h^2 & 2h^2 & 3h^2 & 4h^2 & 5h^2 + 0.2 \end{pmatrix}$$

Since the matrix \mathbf{V} is a function of the unknown heritability, estimation is an iterative process, starting with some initial estimate of h^2 , updating \mathbf{V} in subsequent iterations with the current estimate until convergence. For a starting value, we will use the ratio estimate $h^2 = 0.21$. Applying Equation 9.15 gives a first estimate as

$$\hat{h}_r^2 = \hat{b}_C(GLS)^{(1)} = (\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{V}^{-1} \mathbf{R} = 0.222197$$

Substituting this new value into \mathbf{V} gives upon a second iteration $\hat{b}_C(GLS)^{(2)} = 0.222135$, which remains unchanged upon subsequent iteration.

Standard Errors for Realized Heritability Estimates

The final piece of statistical machinery necessary for assessing the success of the breeders' equation is computation of the standard errors for realized heritability estimates. Consider first the realized heritability estimated from the unweighted (i.e., OLS) regression (Equation 9.13). The variance for an OLS estimator,

$$\begin{aligned} \text{Var} [\hat{b}_C(\text{OLS})] &= \sigma_e^2 (\mathbf{X}^T \mathbf{X})^{-1} = \sigma_e^2 (\mathbf{S}^T \mathbf{S})^{-1} \\ &= \sigma_e^2 / \sum_{i=1}^T S_C^2(i) \end{aligned} \quad (9.16a)$$

The residual variance σ_e^2 can be estimated from the residual sums of squares divided by the degrees of freedom. Here

$$\hat{\sigma}_e^2 = \frac{1}{T-1} \sum_{i=1}^T \hat{e}_i^2 = \frac{1}{T-1} \sum_{i=1}^T \left(R_C(i) - \hat{h}_r^2 S_C(i) \right)^2 \quad (9.16b)$$

As mentioned, because the OLS estimator assumes residuals are uncorrelated and have equal variances (both of which are incorrect), it significantly underestimates the correct variance (see Example 9.3). The GLS regression estimator (Equation 9.15) avoids these problems by properly accounting for the variance structure. From standard GLS theory,

$$\text{Var} [\hat{b}_C(\text{GLS})] = (\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} \quad (9.17)$$

As above, Equations 9.6 and 9.9 are used to obtain the elements of \mathbf{V} , with \widehat{h}_r^2 used in place of h^2 .

Finally, consider the variance for the estimator b_T , the ratio of total response to total selection (Equation 9.14). Since $\text{Var}(y/c) = \text{Var}(y)/c^2$ for a constant c , it immediately follows that

$$\text{Var}(\widehat{b}_T) = \frac{\text{Var}[R_C(T)]}{S_C^2(T)} \simeq \frac{(T/N)\widehat{h}_r^2\sigma_z^2 + \sigma_z^2/M}{S_C^2(T)} \quad (9.18)$$

To obtain the variance in response in Generation T , we again use Equation 9.6, assuming that initial sampling can be ignored $M_0 \gg 1$ and no significant between-generation environmental variance ($\sigma_d^2 = 0$).

Example 9.3. Using the data from Examples 9.1 and 9.2, we compare the standard errors associated with the three different realized heritability estimates (total response, weighted and unweighted regressions). Consider the unweighted regression estimator $\widehat{b}_C(\text{OLS})$ first. The residual sums of squares is

$$\sum_{i=1}^T \left(R_C(i) - \widehat{h}_r^2 S_C(i) \right)^2 = 0.091$$

giving an estimated residual variance of $\widehat{\sigma}_e^2 = 0.091/4 = 0.0228$. Equation 9.16a gives

$$\text{Var} \left[\widehat{b}_C(\text{OLS}) \right] = \sigma_e^2 / \sum_{i=1}^T S_C^2(i) = \frac{0.0228}{350.45} = 0.0000649$$

Taking the square root gives the standard error as 0.0081. Turning to the estimate \widehat{b}_T based on the total response to total selection, Equation 9.18 gives

$$\text{Var}(\widehat{b}_T) = \frac{(5/20) \cdot 0.21 \cdot 3.292 + 0.03292}{12.51^2} = 0.00132$$

for a standard error of 0.0363. Finally, substituting the GLS estimate of $\widehat{h}_r^2 = 0.222135$ in \mathbf{V} (Example 9.2), Equation 9.17 gives variance of this estimate as

$$(\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} = \frac{1}{790.4} = 0.00126554$$

for a standard error of 0.0356.

In summary, the three approaches give extremely similar estimates,

Unweighted least-squares regression, $\widehat{b}_C(\text{OLS})$	$\widehat{h}_r^2 = 0.2245 \pm 0.0081$
Total response/total differential, \widehat{b}_T	$\widehat{h}_r^2 = 0.2102 \pm 0.0363$
Weighted least-squares regression, $\widehat{b}_C(\text{GLS})$	$\widehat{h}_r^2 = 0.2221 \pm 0.0356$

Note that the difference in the standard error between $\widehat{b}_C(\text{GLS})$ and \widehat{b}_T is very small, with \widehat{b}_T considerably more straightforward to compute. Also note that the unweighted regression badly underestimates the standard error.

Experimental Evaluation of the Breeders' Equation

Just how good is the general agreement between the estimated heritability and realized heritability? An extensive review was offered by Sheridan (1988), and the highlights are summarized in Tables 9.1 and 9.2.

Table 9.1. Tests of significance between estimated and realized heritabilities. Only those experiments with estimated standard errors for both heritability estimates are included. After Sheridan (1988).

Species	Significant Differences	NS Differences	Total
<i>Drosophila</i>	14 (23%)	47 (77%)	61
<i>Tribolium</i>	7 (27%)	19 (73%)	26
Mice and Rats	6 (18%)	28 (82%)	34
Poultry and Quail	5 (45%)	6 (55%)	11
Swine and Sheep	8 (53%)	7 (47%)	15
Summary over all groups			
Laboratory Species	27 (21%)	104 (79%)	131
Commerical Species	11 (37%)	19 (63%)	30
All Species	38 (25%)	113 (75%)	151

Sheridan also looked at the goodness-of-fit as a function of the duration of the experiment (Table 9.2). Surprisingly, longer experiments tended to have a better fit. While this is contrary to expectations, this could also be design artifact. First, longer experiments tend to have smaller standard errors, as the SE scales as the inverse of the total selection differential. Second, in many cases longer experiments may employ larger population sizes than experiments of shorter duration, reducing the effect of drift.

Table 9.2. Agreement between realized and estimated base-population heritability as a function of the duration of the experiment. After Sheridan (1988).

Generations	Percent absolute disagreement (relative to \hat{h}_r^2)			n
	0–10%	10–30%	> 30 %	
1 - 5	18%	27%	55%	44
6 - 10	24%	17%	59%	98
10 - 15	52%	10%	38%	90

Asymmetric Selection Response

A common design is to perform a **divergent selection experiment**, wherein replicate lines are selected in opposite directions. Many such experiments show different amounts of response in the up versus down direction, a phenomenon referred to as an **asymmetric selection response**. This is in sharp contrast with the expectation from the breeders' equation, which predicts that the absolute magnitude of response should depend only the absolute value of S .

There are a variety of possible explanations for asymmetric responses (Table 9.3). It may simply be an artifact of the experimental design and/or analysis. In particular, the prediction of equal positive and negative slopes holds only for plots of cumulative response versus cumulative selection differentials ($R_C(t)$ vs. $S_C(t)$). Asymmetry in response based on differences in slope of cumulative response versus generations of selection ($R_C(t)$ vs. t) can thus be very misleading as the different lines may have experienced different amounts of selection.

Table 9.3. Possible explanations for asymmetric response (including reversed response).

Design Defects	Drift
	Scale effects
	Different effective selection differentials
	Undetected environmental trends
	Transient effects from previous selection in the base population
	Undetected selection on correlated characters
Nonlinear parent-offspring regressions	Major genes with dominance
	Genotype-environment interactions
	Departures from normality
Other sources	Genetic asymmetries
	Inbreeding depression
	Maternal effects

Even if the amount of artificial selection is the same in both directions, there may be major differences in natural selection — e.g., up-selected lines may experience lower fertility and thus have a lower effective selection differential. Using effective selection differentials allows for some correction, but the investigator often lacks the data (e.g., fertilities for each parent) necessarily to compute them.

Even though lines may have quite different values of \hat{h}_r^2 , there is still the issue of whether these differences are significant. As we have seen, genetic drift can generate considerable variation between replicate lines and it is important to distinguish between a real difference in response versus the expected variation in two realizations of a process with the same (absolute) expected value. Directional trends in environmental change can also produce asymmetry, accentuating the response in the direction of the trend and retarding response in the opposite direction. Finally, differences in response could simply be scale effects (Figure 9.1). For example, if the genetic variance increases with the mean, heritability can increase with the mean, giving a faster response in the upwardly selected lines.

Although spurious asymmetric responses can result from defects in design and analysis, true asymmetric responses can be generated by a variety of genetic situations. For example, if the parent-offspring regression is nonlinear, S is not sufficient to predict response and it is not surprising that

asymmetric responses can be generated in such cases. Failure to detect departures from linearity in the base population does not rule out nonlinearity as an explanation for an observed asymmetric response. The range of variation in the base population may not be sufficient to detect departures at the extreme ends of the initial range of phenotypes. As the selected lines diverge, differences in the tails of the initial phenotypic distribution can become quite important.

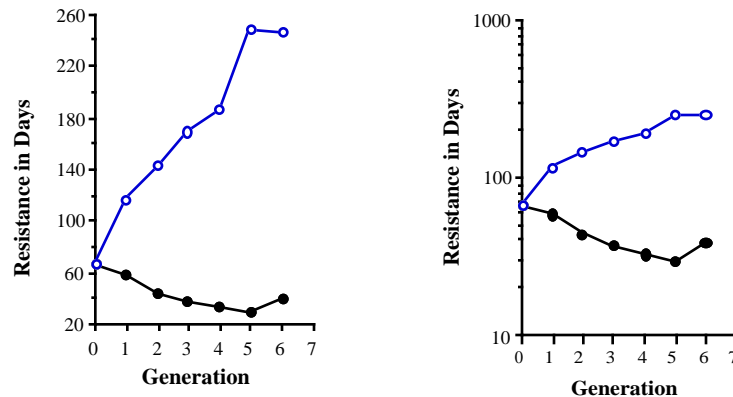


Figure 9.1. An example of a scale effect generating an asymmetric selection response. **Left:** Selection for resistance to dental caries in albino rats (*Rattus norvegicus*), response measured as the expected number of days to develop caries on a standard diet. **Right:** Same data on a log scale.

Characters displaying inbreeding depression show asymmetric selection response, accentuating the response in one direction and retarding it in the other. A simple test for inbreeding as an explanation of asymmetric response is to see whether the mean of an unselected control population changes in the direction of greater response when inbred. The effects of inbreeding depression can be corrected for by inbreeding a control population to the same level as the divergent selection lines, using the contrast between selected and control lines to estimate response. However, this is not necessarily as straightforward as it appears, as selection generally increases the amount of inbreeding within a line by decreasing the effective population size. Thus, simply keeping the control line at the same size as the selected lines underestimates the amount of inbreeding, especially when selection is intense.

When major alleles are segregating, asymmetries in response arise as a consequence of allele frequencies changing in different directions in the selected populations, as an increase in an allele from its initial frequency results in a different additive variance from that produced by an equal decrease in allele frequency. Falconer (1954) refers to this feature as **genetic asymmetry**. This is most easily seen by considering LW Figure 4.8, which shows additive genetic variation as a function of allele frequency for a single diallelic locus. If alleles are completely additive, σ_A^2 as a function of allele frequency is symmetric about $p = 1/2$. Suppose that the frequency of an allele that increases the character value is initially below $1/2$. In upwardly-selected lines, this allele increases in frequency (ignoring drift), resulting in an increase in σ_A^2 and h^2 as the allele frequency increases to $1/2$. Conversely, in downwardly-selected lines, σ_A^2 always decreases. If dominance is present, σ_A^2 is no longer a symmetric function of allelic frequencies (LW Figure 4.8) and asymmetric changes in the contribution to σ_A^2 from a single locus are almost always expected.

Control Populations and Experimental Designs

One complication with estimating realized heritabilities is distinguishing between genetic versus environmental trends. For example, Newman et al. (1973) found that 60 percent of the increase in

yearling weight in a selected line of Shorthorn cattle was due to environmental, rather than genetic, improvement. Using a large unselected **control population** reared under the same environmental conditions as the selected line(s) allows for correction of between-generation environmental change. The use of control populations is not a fool-proof approach for removing environmental trends, as the control and selected lines can develop different genotype-environment interactions. Likewise, extensive genetic drift in the control population can also result in biases. When full pedigree data are available, the powerful mixed-model analysis (discussed below) can also be used to remove genetic and environmental trends, even in the absence of control populations (although they are certainly preferred).

Basic Theory of Control Populations

Assuming no genotype-environment interaction, the true mean of a line in generation t can be decomposed as $\mu + g_t + d_t$, where μ is the base population mean, g_t the change in mean breeding value due to selection and drift, and d_t the change in mean due to environmental change. Under this model, observed means \bar{z} for a selected (s) and control (c) population reared in a common environment can be decomposed as

$$\bar{z}_{s,t} = \mu + g_{s,t} + d_t + e_{s,t} \quad (9.19a)$$

$$\bar{z}_{c,t} = \mu + g_{c,t} + d_t + e_{c,t} \quad (9.19b)$$

where e_t is the error in estimating the mean breeding value ($\mu + g_t$) from the observed mean corrected for the change in the environment ($\bar{z}_t - d_t$) and has expected value 0.

Assuming that the breeders' equation holds, the expected total response at generation t is $E[R_C(t)] = E(g_{s,t}) = h^2 S_C(t)$. Under drift alone, there is no expected directional change in the mean breeding value of the control population, $E(g_{c,t}) = 0$. Assuming no genotype-environment interactions (i.e., $d_{s,t} = d_{c,t}$), the contrast between the selected and control populations has expected value

$$E(\bar{z}_{s,t} - \bar{z}_{c,t}) = E(g_{s,t}) - E(g_{c,t}) = h^2 S_C(t) \quad (9.20a)$$

If the control population is small, then $g_{c,t}$ can drift significantly away from zero, resulting in an over (or under) estimation of the true selection response. Hence, if the goal is simply to remove an environmental trend, control populations should be kept as large as possible so that $g_{c,t}$ is close to zero. However, control populations are also used to attempt to correct for any effects of inbreeding depression on the trait. In these cases, significant drift has likely occurred, and the use of a control population introduces additional uncertainty in the estimate of the response (although this may be more than compensated for the accounting for inbreeding depression and/or environmental trends).

If genotype-environmental interactions are present, then

$$E(\bar{z}_{s,t} - \bar{z}_{c,t}) = h^2 S_C(t) + (d_{s,t} - d_{c,t}) \quad (9.20b)$$

resulting in $\bar{z}_{s,t} - \bar{z}_{c,t}$ being a potentially biased estimator of $h^2 S_C(t)$. If the environmental trends are positively correlated between populations, then the use of a control will still improve the estimate of the genetic trend. However, if the environmental values are negatively correlated between populations, the use of a control can lead to a more inaccurate estimate than simply using a selected population without a control. Mixed-model analysis may be able to provide some insights (as they estimate the $d_{x,t}$ s), but they require extensive information (full pedigrees) and that the model assumptions hold.

If a control population is used, then the responses and differentials are estimated by

$$R_t = (\bar{z}_{s,t} - \bar{z}_{c,t}) - (\bar{z}_{s,t-1} - \bar{z}_{c,t-1}) \quad (9.21a)$$

$$R_C(t) = \bar{z}_{s,t} - \bar{z}_{c,t} \quad (9.21b)$$

$$S_t = \bar{z}_{s,t-1}^* - \bar{z}_{s,t-1} \quad (9.21c)$$

where \bar{z}^* denotes the mean of the selected individuals. No correction is necessary for the selection differential S_t , as we have assumed the environment stays constant within a generation.

Divergent Selection Designs

A related approach to comparing a selected and a control line is the **divergent** (or **bidirectional**) selection design, wherein one compares lines selected in opposite directions (typically denoted by the up and down, or high and low, lines). Again assuming no significant genotype \times environment interactions between lines, the basic statistical model for this design is

$$\bar{z}_{u,t} = \mu + g_{u,t} + d_t + e_{u,t} \quad (9.22a)$$

$$\bar{z}_{d,t} = \mu + g_{d,t} + d_t + e_{d,t} \quad (9.22b)$$

where u and d refer to the upwardly- and downwardly-selected lines. With this design, the responses and differentials are estimated by

$$R_t = (\bar{z}_{u,t} - \bar{z}_{u,t-1}) - (\bar{z}_{d,t} - \bar{z}_{d,t-1}) \quad (9.23a)$$

$$R_C(t) = \bar{z}_{u,t} - \bar{z}_{d,t} \quad (9.23b)$$

$$S_t = (\bar{z}_{u,t-1}^* - \bar{z}_{u,t-1}) - (\bar{z}_{d,t-1}^* - \bar{z}_{d,t-1}) \quad (9.23c)$$

Again, the expected response (using Equations 9.23a and 9.23c for the response and selection differential) is just $R = h^2 S$. In addition to previous concerns about genotype-environment interactions, asymmetric response to selection also complicates the interpretation of results with divergent selection and can result in a biased estimate of the realized heritability.

Variance in Response

For unidirectional selection plus a control population,

$$\begin{aligned} R_C(t) = \bar{z}_{s,t} - \bar{z}_{c,t} &= (\mu + g_{s,t} + d_t + e_{s,t}) - (\mu + g_{c,t} + d_t + e_{c,t}) \\ &= g_{s,t} - g_{c,t} + e_{s,t} - e_{c,t} \end{aligned} \quad (9.24)$$

Similarly, for divergent selection,

$$R_C(t) = \bar{z}_{su,t} - \bar{z}_{sd,t} = g_{su,t} - g_{sd,t} + e_{su,t} - e_{sd,t} \quad (9.25)$$

Since each term in Equations 9.24 and 9.25 is independent, the variance in response becomes

$$\begin{aligned} \sigma^2 [R_C(t)] &= (2f_t + B_0) 2f_t h^2 \sigma_z^2 + B_t \sigma_z^2 \\ &\simeq (tA + B_0) h^2 \sigma_z^2 + B_t \sigma_z^2 \end{aligned} \quad (9.26a)$$

and the covariance between generations in the same line

$$\begin{aligned} \sigma [R_C(t), R_C(t')] &= (2f_t + B_0) h^2 \sigma_z^2 \\ &\simeq (tA + B_0) \sigma_z^2 h^2 \quad \text{for } t < t' \end{aligned} \quad (9.26b)$$

where the coefficients A and B_t are given in Table 9.4. Recall (Equation 9.3) that for unidirectional selection without a control, the variance in response has an additional term, σ_d^2 accounting for the between-generation environmental variation

Table 9.4. Coefficients for the variances and covariances in response (Equations 9.26a and 9.26b). M_x individuals are sampled, of which N_x are allowed to reproduce. The subscripts s and c refer to the selected and control populations, u and d to the up- and down-selected lines, respectively.

Selection in a single direction without a control line. Equation 9.26a has an extra term, σ_d^2 , accounting for the between-generation' variation in environmental effects.

$$f_t = f_{s,t}, \quad A = \frac{1}{N_s}, \quad B_t = \frac{1}{M_{s,t}} \quad \text{for } t \geq 0$$

Selection in a single direction with a control line

$$f_t = f_{s,t} + f_{c,t}, \quad A = \frac{1}{N_s} + \frac{1}{N_c}, \quad B_t = \frac{1}{M_{s,t}} + \frac{1}{M_{c,t}} \quad \text{for } t \geq 0$$

Divergent Selection Without a Control Line

$$f_t = f_{u,t} + f_{d,t}, \quad A = \frac{1}{N_u} + \frac{1}{N_d}, \quad B_t = \frac{1}{M_{u,t}} + \frac{1}{M_{d,t}} \quad \text{for } t \geq 0$$

Control Populations and Variance in Response

When does using a control population reduce the variance in response? Assuming $M = M_s = M_c$ and $N = N_s = N_c$, subtracting the expected variance in response using a unidirectionally-selected population adjusted using a control from the response estimated without a control gives

$$\left(\frac{t}{N} + \frac{1}{M_0} \right) h^2 \sigma_z^2 + \frac{1}{M} \sigma_z^2 - \sigma_d^2 \quad (9.27a)$$

Assuming the between-line drift variance dominates (terms involving M are ignored), the condition for the variance in response with a control to be larger than the response without one is approximately

$$\frac{t\sigma_z^2 h^2}{N} > \sigma_d^2 \quad (9.27b)$$

Hence, regardless of the value of σ_d^2 , if sufficient generations are used, the optimal design (in terms of giving the smallest expected variance in response) is not to use a control. However, this approach runs the risk of an undetected directional environmental trend compromising the estimated heritability. Further, as the number of generations becomes large, the key assumption of short-term response (essentially no changes in the genetic and phenotypic variances) becomes untenable.

Optimal Experimental Design

As Equation 9.27b illustrates, when to use different designs is not entirely clear-cut. What in general can we say? The coefficient of variation in response,

$$CV[R_C(t)] = \frac{\sigma[R_C(t)]}{E[R_C(t)]}$$

is especially useful in comparing efficiencies of different designs, as it is independent of σ_z^2 and further provides an appropriate measure of comparing efficiencies when the expected response

differs between designs. Table 9.5 gives expressions for the CV under some simplifying assumptions. Note that the coefficient of variation is a function of tN , the total number of adults selected during the course of the experiment (provided drift variance dominates error variance). A short experiment with many selected adults per generation thus gives the same expected CV as a long experiment with few adults per generation (provided the total numbers are the same). However, if the error variance is nontrivial relative to the drift variance (as would be expected if h^2 small), increasing the duration of the experiment results in some improvement in precision.

Table 9.5. Coefficients of variation for various designs, assuming the pure-drift approximation and further that $\sigma^2 [R_C(t)] \simeq t Ah^2 \sigma_z^2$. This assumes that the selection experiment is sufficiently long that the between-line drift dominates (i.e., $tA \gg B_0$ and $tAh^2 \gg B_t$). For all designs, we assume that the absolute selection intensity on all selected lines is \bar{i} .

Selection in a single direction with a control line

$$E [R_C(t)] = t h^2 \bar{i} \sigma_z, \quad CV [R_C(t)] \simeq \frac{1}{h\bar{i}} \sqrt{\frac{2}{Nt}}$$

Selection in a Single Direction Without a Control Line (Assuming $\sigma_d^2 = 0$)

$$E [R_C(t)] = t h^2 \bar{i} \sigma_z, \quad CV [R_C(t)] \simeq \frac{1}{h\bar{i}} \sqrt{\frac{1}{Nt}}$$

Divergent Selection Without a Control Line

$$E [R_C(t)] = 2t h^2 \bar{i} \sigma_z, \quad CV [R_C(t)] \simeq \frac{1}{h\bar{i}} \sqrt{\frac{1}{2Nt}}$$

As an example of using CV, consider unidirectional selection without a control population versus divergent selection. Which is more efficient if the same total number of adults are selected (e.g., N under unidirectional, $N_d = N_u = N/2$ under divergent selection)? If there is no between-generation environmental variance both designs are equally efficient, while divergent selection is more efficient if $\sigma_d^2 > 0$.

Example 9.4. Suppose we plan to select the upper 5% of a population for a normally distributed character with $h^2 = 0.25$. What value of Nt is needed for the expected CV of response to be no greater than 0.01 if no control population is used? For this amount of selection, $E(\bar{i}) = 2.06$ if the population is large, and slightly less in small populations (for simplicity assume the large population value). Further assuming that drift variance dominates error variances (including σ_d^2), applying the expressions from Table 9.5 gives

$$0.01 = \frac{1}{0.5 \times 2.06} \times \sqrt{\frac{1}{Nt}}$$

Solving, gives $Nt \simeq 9426$. Hence, during the entire course of the experiment using a total of at least 9426 selected parents gives an approximate expected CV less than 1%. If the desired CV is 0.05 or 0.10, $Nt \simeq 377$ and $Nt \simeq 94$, respectively.

Nicholas' Criterion

An alternative criterion for choosing Nt was suggested by Nicholas (1980). Often the investigator is interested in ensuring that (at least) a certain response will occur with a preset probability. To a reasonable approximation, the expected mean value in any given replicate line after t generations of selection is normally distributed, with mean $E[R_C(t)]$ and variance $\sigma^2 [R_C(t)]$. Consider the probability that the observed response is at least β of the expected response,

$$\begin{aligned}\Pr(R_C(t) > \beta E[R_C(t)]) &= \Pr\left(\frac{[R_C(t) - E[R_C(t)]]}{\sigma [R_C(t)]} > \frac{(\beta - 1)E[R_C(t)]}{\sigma [R_C(t)]}\right) \\ &= \Pr\left(U > \frac{\beta - 1}{\text{CV}[R_C(t)]}\right)\end{aligned}\quad (9.28)$$

where U is a unit normal random variable. Note that the probability that the observed response exceeds the expected response ($\beta = 1$) is one half (as $\Pr[U > 0] = 1/2$).

Example 9.5. Again suppose that $\bar{i} = 2.06$, $h^2 = 0.25$, and the design is unidirectional selection without a control population. What value of Nt is required in order for a 95% probability that the observed response is at least 90% of its expected response? Here, $\beta = 0.9$ and (from normal tables) $\Pr[U > -1.65] = 0.95$. Hence,

$$\frac{\beta - 1}{\text{CV}[R_C(t)]} = \frac{-0.1}{\text{CV}[R_C(t)]} = -1.65$$

Rearranging gives

$$\text{CV}[R_C(t)] = \frac{1}{0.5 \times 2.06} \sqrt{\frac{1}{Nt}} = \frac{0.1}{1.65}$$

implying that $Nt \simeq 257$.

Mixed-model Estimation

When we have access to the complete pedigree of all individuals in the selection experiment, we can apply the more powerful approaches of mixed-model (i.e. BLUP and REML) estimation. While OLS and GLS essentially uses only the means, mixed-model methods makes full use of the covariance structure of not just between means but also between all individuals as well.

The basic building block of mixed-model analysis of selection experiments is the **animal model**, which estimates the breeding (or additive genetic) values of all individuals measuring during the course of experiment. To apply the animal model to selection experiments, first vectorize the observations from the entire experiment by letting y_{ij} denote the j measured individual from generation i , where $0 \leq i \leq t$ (generation 0 representing the unselected base population) and $1 \leq j \leq n_i$. Let the vector \mathbf{y} denote all measured individuals from the entire experiment,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_t \end{pmatrix}, \quad \text{where} \quad \mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{in_i} \end{pmatrix}$$

The vector \mathbf{y}_i includes the values for all measured individuals from generation i , including those culled as well as those allowed to reproduce. The simplest animal model for these data is

$$y_{ij} = \mu + a_{ij} + e_{ij} \quad (9.29a)$$

where μ is an overall mean, a_{ij} the breeding value of the j th measured individual from generation i , and e_{ij} the deviation between breeding and phenotypic values. With exactly one record per individual, the only fixed effect is the mean, giving $\beta = (\mu)$ and $\mathbf{X} = \mathbf{1}$ (a vector of ones), giving the model as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \mathbf{e} \quad (9.29b)$$

Here \mathbf{a} is the vector of breeding values for all individuals measured during the course of the experiment, with $\text{Var}(\mathbf{a}) = \sigma_A^2 \mathbf{A}$. The relationship matrix \mathbf{A} is the key to mixed-model analysis, as it includes all the pedigree information. The diagonal elements of \mathbf{A} describe the amount of inbreeding, with $A_{ii} = (1 + f_i)$, while the off-diagonal elements $A_{ij} = 2\Theta_{ij}$ (twice the coefficient of coancestry) describe the relatedness of individuals i and j . The simple animal model assumes that all genetic variance is additive, so that there is no (genetic) covariance between residuals. In this case, it is generally assumed that $\text{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$.

For Equation 9.29b, the mixed-model equations simplify to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda \mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \mathbf{y} \end{pmatrix} \quad (9.29c)$$

where n is the total number of individuals in the experiment, $\lambda = \sigma_e^2/\sigma_A^2 = (1 - h^2)/h^2$, $\hat{\mathbf{a}}$ is an n -dimensional vector of the predicted breeding values of all measured individuals, and $\mathbf{1}$ is a vector of ones.

Under a mixed-model analysis, response is measured by the change in the mean breeding value of a selected population over time. The estimated mean breeding value in generation k is simply given by

$$\hat{\mathbf{a}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \hat{a}_{kj} \quad (9.30a)$$

Total response at generation t is estimated by $\bar{a}_t - \bar{a}_0 = \bar{a}_t$, as the predicted mean breeding value from generation 0 (the unselected base population) is zero by construction. In matrix notation, the vector $\bar{\mathbf{a}}$ of mean breeding values is estimated by

$$\hat{\bar{\mathbf{a}}} = \mathbf{K}^T \hat{\mathbf{a}} \quad (9.30b)$$

where the i th row of the matrix \mathbf{K} consists of $1/n_j$ when the column corresponds to an individual from generation j , otherwise all elements in that row are zero. Thus, for t generations of data (corresponding to $t - 1$ generations of selection, as the analysis includes the unselected base population, generation 0), \mathbf{K} is $n \times t$, with $\mathbf{K}^T \mathbf{1}_n = \mathbf{1}_t$ (a $t \times 1$ vector of ones).

Example 9.6. Suppose in the base population (unrelated and non-inbred) individuals 1-4 have trait values 3, 6, 5, and 2, respectively. The two largest individuals are allowed to reproduce and their resulting offspring (individuals 5-8) have values 4, 5, 6, 5. Assuming the only fixed effect is the mean, the resulting animal

model is $\mathbf{y} = \mathbf{1}\beta + \mathbf{a} + \mathbf{e}$, where

$$\mathbf{y} = \begin{pmatrix} 3 \\ 6 \\ 5 \\ 2 \\ 4 \\ 5 \\ 6 \\ 5 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{pmatrix}, \quad \beta = (\mu)$$

What is the relationship matrix \mathbf{A} ? Since individuals 2 and 3 are the parents, and all offspring are full-sibs, all related individuals have values of $1/2$ as $2\theta_{ij} = 1/2$ for both parent-offspring and full-sibs. The resulting numerator relationship matrix becomes

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

Suppose the heritability of the trait is $h^2 = 0.3$. Here

$$\hat{\mu} = (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} \mathbf{1}^T \mathbf{V}^{-1} \mathbf{y} = 4.22$$

The BLUP estimate of the genetic values as

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.366 \\ 0.666 \\ 0.366 \\ -0.666 \\ 0.386 \\ 0.562 \\ 0.739 \\ 0.562 \end{pmatrix}. \quad \text{Here } \mathbf{K} = \frac{1}{4} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{implying } \mathbf{K}^T \hat{\mathbf{a}} = \begin{pmatrix} 0 \\ 0.562 \end{pmatrix}$$

Hence, the estimated response (for $h^2 = 0.3$) is 0.562. Under a least squares analysis, $\bar{y}_0 = 4$ and $\bar{y}_1 = 5$ for an estimated response of 1. The estimated response for different assumed heritabilities are as follows:

h^2	Estimated response	h^2	Estimated response
0.0	0	0.6	0.940
0.1	0.211	0.7	1.026
0.2	0.398	0.8	1.083
0.4	0.707	0.9	1.095
0.5	0.833	1.0	1

Note that a REML/BLUP analysis of a selection experiment has a very different character from a LS analysis. In the latter, one estimates the realized heritability from a suitable regression

of phenotypic means on selection differentials. With a strictly BLUP analysis, one starts with an assumed base-population heritability and then computes a genetic trend by plotting the mean estimated breeding values. Response is underestimated if the assumed heritability is less than the true value, while it is overestimated if the heritability is overestimated. The BLUP response curves of mean breeding values are much smoother than the curves for phenotypic means. This is because BLUP compares an individual's estimated breeding value with an index based on information from its relatives. Individual breeding values are regressed towards the value predicted by the index, smoothing out excessive fluctuations. In a REML/BLUP analysis, one first estimates the additive genetic variance in the base population (via REML), using this value in the subsequent BLUP analysis.

One caveat that must be stressed is that the estimated mean breeding values should not themselves be used to estimate heritabilities. For example, Blair and Pollak (1984) regressed the BLUP-estimated mean breeding values on cumulative selection differentials to obtain a realized heritability estimate. The problem with this approach is that the heritability (or additive variance) used in the BLUP analysis very strongly influences the results. Hence, realized heritability estimates obtained from regressions based on BLUP estimates of mean breeding values depend on the assumed heritability, not the actual population heritability (Thompson 1986). The correct estimate of heritability in a mixed-model analysis should be based on REML (or other) estimates of the base-population variance components.

The Relationship Matrix Accounts for Drift and Disequilibrium

Selection changes the additive variance by generating gametic-phase disequilibrium even in the absence of allele frequency changes. Additionally, with a finite number of loci selection also changes allele frequencies, further changing the genetic variances. While gametic-phase disequilibrium changes in the genetic variances are generally restricted to the first few generations of selection, allele frequency changes become increasingly more important as selection proceeds. Thus, the additive variance in a particular generation after selection is likely different from the variance in the unselected base population. A least-squares analysis does not account for these changes, but rather assumes that the realized heritability is the same in each generation. Given that the reduction in h^2 from disequilibrium reaches its equilibrium value in only a few generations of directional selection, the LS assumption of a constant h^2 may not induce too serious of an error, *provided* the reduction is corrected for.

Under the infinitesimal model (in particular, assuming no significant selection-induced changes in allele frequencies), a mixed-model analysis fully accounts for the effects of gametic-phase disequilibrium as well as genetic drift. This occurs because under the infinitesimal model, even in the face of selection and drift, the variance-covariance matrix of the vector of breeding values remains the product of the base-population additive genetic variance and the numerator relationship matrix, $\text{Var}(\mathbf{a}) = \sigma_A^2 \mathbf{A}$.

This independence of the covariance relationships of \mathbf{a} from selection and drift (given \mathbf{A}) follows as a consequence of the behavior of the residual in regression of the breeding value of an individual A_i on the breeding values of its sire (A_{m_i}) and dam (A_{f_i}). Recall (lecture 3) that the regression of offspring breeding values on parental breeding values is

$$A_i = \frac{1}{2}A_{f_i} + \frac{1}{2}A_{m_i} + s_i \quad (9.31)$$

where the **segregation residual** s (also referred to as **Mendelian sampling**) results from segregation of alleles at heterozygous loci in the parents. Under the infinitesimal model, s is independent of parental breeding values and has mean zero and variance $(1 - \bar{f}_i) \sigma_A^2 / 2$. Here \bar{f}_i is the mean inbreeding of the parents of individual i and σ_A^2 the base-population (before selection) additive variance. More generally, provided the vector \mathbf{s} of Mendelian sampling residuals remains multivariate

normal, then $\mathbf{s} \sim \text{MVN}(\mathbf{0}, (\sigma_A^2/2) \mathbf{F})$. The matrix \mathbf{F} is diagonal with i th element $(1 - \bar{f}_i)$, one minus the average inbreeding in the parents. If k and j are the parents of offspring i , then

$$F_{ii} = (1 - \bar{f}_i) = \left(1 - \frac{f_k + f_j}{2}\right) = \left(2 - \frac{A_{kk} + A_{jj}}{2}\right) \quad (9.32)$$

where $A_{kk} = (1 + f_k)$ denotes the k th diagonal element of the relationship matrix \mathbf{A} . The effects of drift on the additive variance enter through the inbreeding coefficients f . Thus, the distribution of the Mendelian sampling terms \mathbf{s} is unaffected by the breeding values of the parents (and hence by selection and assortative mating). When we have the complete pedigree of all individuals in the selection experiment, along with all their records, we can express any breeding value as a linear function of the base population breeding values (the coefficients following from \mathbf{A}) and Mendelian sampling terms (\mathbf{s}) not affected by selection. In particular, we can express \mathbf{a} as a linear function of the Mendelian sampling terms, $\mathbf{a} = \mathbf{T}\mathbf{s}$. The resulting covariance matrix is $\text{Var}(\mathbf{a}) = \mathbf{T} \text{Var}(\mathbf{s}) \mathbf{T}^T$, where $\text{Var}(\mathbf{s}) = (\sigma_A^2/2)\mathbf{F}$ is independent of selection and assortative mating, while \mathbf{F} accounts for the reduction in additive variance from genetic drift.

To show this, we follow Sorensen and Kennedy (1984a). Ordering individuals so that parents proceed their offspring, \mathbf{A} can be written as a function of a diagonal matrix, \mathbf{D} , and an upper triangular matrix \mathbf{T} ,

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^T \quad (9.33)$$

\mathbf{T} traces the passage of genes from one generation to the next, while \mathbf{D} is the variance in offspring breeding value, conditioned on the parental breeding values (the Mendelian sampling variance). To see this last point, consider the transformation $\mathbf{g} = \mathbf{T}^{-1}\mathbf{a}$ of the vector of breeding values. From Equation 9.33, the covariance matrix for \mathbf{g} becomes

$$\text{Var}(\mathbf{g}) = \mathbf{T}^{-1} \text{Var}(\mathbf{a})(\mathbf{T}^T)^{-1} = \sigma_A^2 \mathbf{T}^{-1} \mathbf{T}\mathbf{D}\mathbf{T}^T (\mathbf{T}^T)^{-1} = \sigma_A^2 \mathbf{D} \quad (9.34)$$

which follows using the general matrix relationship $(\mathbf{B}^T)^{-1} = (\mathbf{B}^{-1})^T$ (the inverse of transpose equals the transpose of the inverse). Sorensen and Kennedy show that the i th element of \mathbf{g} equals

$$g_i = A_i - \frac{1}{2}A_{f_i} - \frac{1}{2}A_{m_i} = s_i \quad (9.35)$$

which is simply the segregation residual. Hence, $\mathbf{g} = \mathbf{s}$, and since $\text{Var}(\mathbf{s}) = (\sigma_A^2/2)\mathbf{F}$, implies $\mathbf{D} = \mathbf{F}/2$. Writing $\mathbf{a} = \mathbf{T}\mathbf{T}^{-1}\mathbf{a} = \mathbf{T}\mathbf{g} \equiv \mathbf{T}\mathbf{s}$, shows that the vector of breeding values \mathbf{a} is a simple vector transformation of the vector of Mendelian sampling residuals \mathbf{s} .

Thus, the variance of the vector of breeding values is a linear transformation of the variance of the Mendelian sampling residuals. Under the infinitesimal model, the distribution of these residuals is unaffected by selection. However, if the infinitesimal model does not hold, then residual values may indeed vary with parental breeding values, in which case selection can certainly influence the distribution of residuals. If, however, the change in allele frequencies is small over the course of the experiment, the bias may not be too serious. Likewise, another key element is that the distribution of residuals does not significantly deviate from normality.

Model Validation

Given the sensitivity of a mixed-model analysis to the validity of the assumptions (in particular, the infinitesimal model), some form of model validation is required to apply MM methods with confidence. One approach is to test the infinitesimal model prediction that estimates of the base population σ_A^2 should remain stable as additional generations of selection are considered. If the infinitesimal model holds, \mathbf{A} completely accounts for changes in the additive variance in these later

generations of selection. On the other hand, if σ_A^2 is changing in ways not predictable from the infinitesimal model, using data from additional generations of selection may result in dramatically different estimates of the base-population additive variance. Likewise, if the same base population is used to form both the control and selected lines, the estimated the base population additive variance has the same expected value in both lines. Departures from either of these two predictions indicates failure of the model.

Example 9.7. One of the first REML/BLUP analyses of a selection experiment was by Meyer and Hill (1991), who examined the response to selection for adjusted food intake (AFI) in mice. AFI is defined as food intake between 4 and 6 weeks corrected for 4-week weight. Meyer and Hill had three replicates, each consisting of a high, low, and control lines, for a total of almost 11,000 mice over the course of the experiment. Within-family selection on AFI was followed for 23 generations. Meyer and Hill included a number of fixed effects in their model, as well as adding a random effect to control for common litter effects.

As a check of the validity of the assumptions (in particular, the infinitesimal model), Meyer and Hill compared variance estimates based on data from generations 5 -7 with estimates based on generations 14-23. In both cases, the full pedigree structure was incorporated into the numerator relationship matrix (i.e., the relationships among individuals in generation 5 were used to generate the submatrix of **A** associated with this generation, and similarly for future generations). While incorporation of the complete pedigree information reduces the bias in estimates of the base-population additive variance, it does not completely reduce the bias if the records for all individuals from previous generations (back to the base population) are ignored (van der Werf and de Boer 1990). Even with this caveat in mind, Meyer and Hill observed a dramatic decline in the estimated additive variances (from 7.2 based on generations 5-7 to 2.5 based on generations 14-23). Under the infinitesimal model, both estimates should be for the base population variance. This large decrease suggested that the infinitesimal model may not be appropriate for this trait. It is interesting to note that this decrease occurred even as the total variance increased dramatically (from 23.88 to 33.93). This increase resulted mainly from an increase in the environmental variance (from 12.9 to 25.5), although there was also a slight increase in the litter-effects variance (from 4.78 to 5.96).

Several other REML/BLUP analyses of selection experiments in mice also found differences in estimates of base population additive variance when comparing data from early generations versus data from later generations. Beniwal et al. (1992a,b) observed decreases in the additive variance (in body weight, litter size, and lean mass), while Heath et al. (1995) observed an increase in the additive variance in body weight.

Lecture 9 Problems

1. Consider the following data for four generations of selection with $N = 200$ individuals chosen (after selection) to form the next generation. The phenotypic variance is $\sigma_z^2 = 9$ and assume M is sufficiently large that we can ignore terms of order $1/M$.

Generation	Mean	
	Before selection	After selection
1	10.00	11.48
2	10.81	12.53
3	11.64	12.90
4	12.21	13.60
5	12.90	

- Compute the vectors of the cumulative differential and cumulative response.
- Compute the ratio estimator for realized heritability and its standard error.
- Compute the OLS regression estimator of realized heritability and its standard error.
- Compute the GLS regression estimator of realized heritability and its standard error.

Lecture 9 Solutions, page 1 of 2

a:

$$\mathbf{R} = \begin{pmatrix} 10.81 - 10.00 \\ 11.64 - 10.80 \\ 12.90 - 11.64 \\ 13.60 - 12.9 \end{pmatrix} = \begin{pmatrix} 0.81 \\ 0.83 \\ 0.58 \\ 0.68 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 11.48 - 10.00 \\ 12.53 - 10.81 \\ 12.90 - 11.64 \\ 13.60 - 12.21 \end{pmatrix} = \begin{pmatrix} 1.48 \\ 1.72 \\ 1.26 \\ 1.39 \end{pmatrix}$$

The cumulative differential and response thus become

$$\mathbf{R}_C = \begin{pmatrix} 0.81 \\ 1.64 \\ 2.21 \\ 2.90 \end{pmatrix}, \quad \mathbf{S}_C = \begin{pmatrix} 1.48 \\ 3.20 \\ 4.47 \\ 5.86 \end{pmatrix}$$

b:

$$h_r^2 = R(4)/S(4) = 2.90/5.86 = 0.49$$

The standard error is obtained by first noting that

$$\sigma^2 [R_C(4)] = \left(\frac{4}{N}\right) h^2 \sigma_z^2 = \frac{4}{200} * 0.49 * 9 = 0.0882$$

$$\sigma^2(h_r^2) = \sigma^2 \left(\frac{\sigma^2 [R_C(4)]}{S_C(4)} \right) = \frac{\sigma^2 [R_C(4)]}{S_C(4)^2} = \frac{0.0882}{5.86^2} = 0.00257$$

giving a standard error of $\sqrt{0.00257} = 0.051$

c:

$$\hat{h}_r^2 = \hat{b}_C(OLS) = \frac{\sum_{i=1}^4 S_C(i) \cdot R_C(i)}{\sum_{i=1}^4 S_C^2(i)} = \frac{33.3}{66.7} = 0.50$$

Turning to the standard error,

$$\text{Var} [\hat{b}_C(OLS)] = \sigma_e^2 / \sum_{i=1}^4 S_C^2(i) = \frac{\sigma_e^2}{66.7} = \frac{0.0023}{66.7} = 0.000035$$

giving a standard error of 0.006. We estimated the variance for the residuals from

$$\sigma_e^2 = \frac{1}{4-1} \sum_{i=1}^4 e_i^2 = \frac{1}{3} \sum_{i=1}^4 \left(R_C(i) - \hat{h}_r^2 S_C(i) \right)^2 = \frac{0.0070}{3} = 0.0023$$

d: For the GLS solution, note that the elements in the covariance matrix \mathbf{V} are

$$V_{ii} = \sigma^2 [R_C(i)] = \left(\frac{i}{N}\right) h^2 \sigma_z^2 = i * h^2 * 9/200 = ih^2 0.045$$

and

$$V_{ij} = \sigma [R_C(i), R_C(j)] = \left(\frac{i}{N}\right) h^2 \sigma_z^2 = ih^2 0.045 \quad \text{for } i < j$$

Lecture 9 Solutions, page 2 of 2

giving

$$\mathbf{V} = 0.045 \cdot \begin{pmatrix} h^2 & h^2 & h^2 & h^2 \\ h^2 & 2h^2 & 2h^2 & 2h^2 \\ h^2 & 2h^2 & 3h^2 & 3h^2 \\ h^2 & 2h^2 & 3h^2 & 4h^2 \end{pmatrix} = 0.045h^2 \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

$$\hat{h}_r^2 = \hat{b}_C(GLS) = (\mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{S}_C)^{-1} \mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{R}_C$$

Notice that the $0.045h^2$ term on \mathbf{V} cancels.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$$

Hence

$$\frac{\mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{S}_C}{0.045h^2} = (1.48 \quad 3.20 \quad 4.47 \quad 5.86) \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1.48 \\ 3.20 \\ 4.47 \\ 5.86 \end{pmatrix} = 8.6938$$

while

$$\frac{\mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{R}_C}{0.045h^2} = (1.48 \quad 3.20 \quad 4.47 \quad 5.86) \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 0.81 \\ 0.83 \\ 0.58 \\ 0.68 \end{pmatrix} = 4.3094$$

thus

$$\hat{h}_r^2 = \hat{b}_C(GLS) = (\mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{S}_C)^{-1} \mathbf{S}_C^T \mathbf{V}^{-1} \mathbf{R}_C = \frac{4.3094}{8.6938} = 0.4957$$

The resulting standard error is

$$(\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} = \frac{0.045h^2}{8.6938} = \frac{0.045 * 0.4957}{8.6938} = 0.00256572$$

for a standard error of $\sqrt{0.00256572} = 0.050653$