15

Mixed-Model and Bayesian Analysis of Short-Term Selection Experiments

Unnecessarily complex analysis should not be used as a foil to disguise lower quality datasets: estimates of genetic parameters are only as good as the data on which they are based — Kruuk (2004)

Draft version 13 September 2008

A least-squares analysis of a selection experiment (Chapter 14) distills the data down to the means and variances of the trait within each generation and the between-generation covariances in the means — any additional information is simply ignored. In some selection experiments we have access to much more information — in the extreme trait measurements (or **records**) of all individuals throughout the course of the selection experiment and the pedigree of how these individuals are related to each other. When such data is available, a LS analysis ignores this structure (and hence information). **Mixed-model** (MM) **analysis** (LW Chapters 26, 27), on the other hand, fully considers the covariances between *all* observations. By virtue of using this additional information, a MM analysis is potentially far more powerful than a LS one (Sorensen and Kennedy 1983, 1984a, 1984b; Kennedy and Sorensen 1988; Kennedy 1990). For example, under infinitesimal-model assumptions, incorporation of the full covariance structure accounts for any within- and between-line changes due to genetic drift, as well as changes in σ^2_A from gametic-phase disequilibrium due to selection and/or assortative mating.

There are two general approaches for a mixed-model analysis. The first is the **two-step approach**, wherein one first uses **REML** (LW Chapter 27) or some other method to estimate the appropriate variance components, and then uses **BLUP** to estimate breeding values (LW Chapter 26). With BLUPs for individual breeding values in hand, one can estimate the mean breeding value for any particular generation and hence (in theory) directly follow *genetic*, as opposed to *phenotypic* change, allowing for the separation of genetic versus environmental change. While straight-forward, the two-step approach does account for the uncertainty in BLUP estimates that results from using estimates, as opposed to the true values, of the variances. Further, the sampling distribution of this two-step process is currently not obtainable. In contrast, **Bayesian approaches** simultaneously estimate variance components and predict of random effects, returning (given a specified prior) a posterior distribution for random variables of interest given the data. Bayesian analysis provides an exact account of the uncertainty in the estimation of the unknown parameters, especially after the confounding effects of nuisance parameters are removed.

We start this chapter with a (brief) review of mixed-models and then consider various applications of the **animal model** to the analysis of selection experiments. We then introduce the basics of Bayesian analysis, and conclude by re-analyzing many of the same mixed models under a Bayesian framework.

MIXED MODEL vs. LEAST SQUARES ANALYSIS

Figure 15.1 illustrates the result of a mixed-model analysis of selection response. Note that instead of measuring response from the *observed phenotypic means* (the LS approach),

response is measured from the *estimated mean breeding values* obtained from BLUP. Further, instead of estimating a realized heritability, a mixed-model analysis estimates the additive genetic variance in the base population.

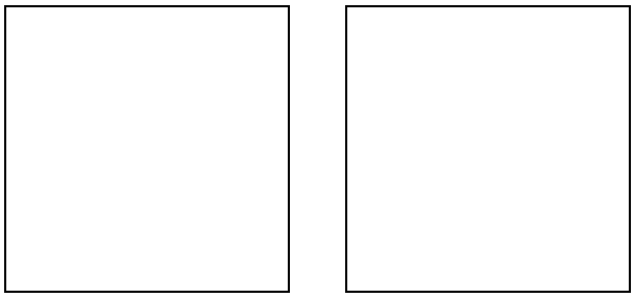


Figure 15.1. Results from high and low selection on 6-week weight in mice. **Left:** Observed mean phenotypic values in the up and down-selected lines. <M/H Fig 1.e>. **Right:** Estimated mean breeding values for both selected populations and the control (middle curve). <M/H Fig 2.3>. See Examples 15.2 and 15.6 for more details on this experiment. After Meyer and Hill (1991).

Mixed-models readily allow records to be adjusted for any number of **fixed effects**. For example, one might correct for trait differences between sexes, between individuals from different size litters, age effects, known environmental factors, etc. Such adjustments for fixed-factors result in more accurate prediction of an individual's genetic value (and a more accurate estimate of the population's genetic response). While any number of fixed factors are readily incorporated into a mixed-model analysis, it can be more problematic to adjust for fixed-factors in a LS analysis. For example, a LS analysis cannot separate genetic from environmental trends when only a single line is considered. By contrast, under a properly-formatted MM analysis, one can separate phenotypic changes into genetic and environmental components without using a control population. As we will see, this is made possible by using the full covariance structure associated with the complete pedigree of all measured individuals in the experiment.

Another advantage of mixed model analysis is its great flexibility in handling any number of selection designs within the same framework. For example, a MM analysis allows for overlapping generations (e.g., when a parent contributes offspring over several different years of selection), while LS analysis of response in overlapping generations can be difficult to formulate correctly. Finally, a properly-designed mixed model analysis can account for assortative mating, drift, and selection-induced gametic phase disequilibrium (provided we can assume the infinitesimal model).

For all the power of mixed-model analysis, there are tradeoffs relative to a simpler LS analysis. First, a MM analysis requires far greater record keeping (e.g., following all individuals and their relatives) and is more computationally demanding. Second, a MM analysis can be rather model-sensitive, in particular the infinitesimal model assumption is critical. If selection-induced changes in allele frequencies are significant during the course of the experiment, the assumptions of a MM analysis can be violated

BASICS OF MIXED-MODEL ANALYSIS

While we provide a brief overview of the analysis of mixed models here, we strongly

encourage the reader to review LW Chapters 26 and 27 before proceeding. These chapters provide many worked examples to give the reader a feel for mixed models, as well as considering advanced topics in MM analysis in far great detail than we do here.

Mixed models are so named because they consider both fixed and random effects. Recall that fixed effects are unknown constants while random effects have values that are drawn from some underlying distribution (LW Chapters 8, 26). Hence, any particular value for a random effect represents just one possible realization from this underlying distribution. Typically, statisticians speak of *estimating* fixed effects and *predicting* the realized values of random effects. Both LS and MM analyses estimate the fixed effects in a model, while MM analysis also predicts the values of the random effects by using the covariances between observations (after adjusting for fixed effects). Note that under a Bayesian analysis, every effect is assumed to be random, and this distinction between fixed vs. random effects disappears.

The standard mixed model for a vector $\mathbf{y}$ of n observations is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e} \tag{15.1}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed effects, $\mathbf{a}$ is a $p \times 1$ vector of random effects (in our case, these are the breeding values of the individuals in our experiment), $\mathbf{e}$ the vector of residuals (which are also random effects), and $\mathbf{X}$ and $\mathbf{Z}$ are $n \times k$ and $n \times p$ **incident matrices** associated with the fixed and random effects. If each individual is measured exactly once, then $\mathbf{Z} = \mathbf{I}_{n \times n}$, an identity matrix of dimension n. In the absence of the vector of random effects $\mathbf{a}$, Equation 15.1 reduces to a least squares model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with ordinary (unweighted) least squares (OLS) used to estimate $\boldsymbol{\beta}$ if the residuals are uncorrelated and homoscedastic, giving the covariance matrix for the vector of residuals as $\mathbf{Var}(\mathbf{e}) = \sigma^2\,\mathbf{I}$. More generally, if the covariance structure of the residuals is more complex, $\mathbf{Var}(\mathbf{e}) = \mathbf{V}$ where the only constraint on the matrix $\mathbf{V}$ is that it is symmetric and positive definite, then general (weighted) least squares (GLS) is used (Equation 15.3a).

An important feature about a mixed-model is that the variances are estimated *after* the variation introduced by fixed effects is removed. Wilson (2008) notes that this implies different models (even using the same data) with different fixed effects give different residual variances, and hence potentially different heritabilities, as $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_e^2)$.

In order to solve Equation 15.1, we need to specify the covariance structure for the vectors of random effects. It is generally assumed that the residuals are uncorrelated and homoscedastic, $\mathbf{Var}(\mathbf{e}) = \sigma^2\,\mathbf{I}$. The covariance of $\mathbf{a}$ (in our case, the vector of breeding values) has a more complicated structure governed by the details of the pedigree, $\mathbf{Var}(\mathbf{a}) = \sigma_A^2\,\mathbf{A}$. Here $\mathbf{A}$ is a matrix of known constants (the **numerator relationship matrix**) whose elements, as detailed below, are given by the pedigree structure. The resulting $n \times n$ covariance matrix $\mathbf{V}$ for the vector of observations $\mathbf{y}$ becomes

$$\mathbf{V} = \sigma_A^2\,\mathbf{Z}\mathbf{A}\mathbf{Z}^T + \sigma_e^2\,\mathbf{I} \tag{15.2a}$$

The covariance matrix $\mathbf{V}$ is thus a function of the variance components σ_A^2 and σ_e^2. Since we will assume throughout that $\mathbf{a}$ is a vector of breeding values, so that σ_A^2 is the additive genetic variance, we can alternatively express $\mathbf{V}$ as a function of the heritability (h^2) and phenotypic variance ($\sigma_z^2 = \sigma_A^2 + \sigma_e^2$) of the trait of interest (after removal of fixed effects),

$$\mathbf{V} = \sigma_z^2\left(h^2\mathbf{Z}\mathbf{A}\mathbf{Z}^T + (1-h^2)\mathbf{I}\right) \tag{15.2b}$$

as $\sigma_e^2 = \sigma_z^2 - \sigma_A^2 = \sigma_z^2(1-h^2)$. Since our focus is generally on breeding values versus everything else, any dominance variance gets swept into the residual variance. Lumping

the dominance into the residual results in residues within families being correlated, and we correct for this later.

Estimation of the vector of fixed-effects $\boldsymbol{\beta}$ follows from GLS using the covariance matrix $\mathbf{V}$ (LW Chapter 8),

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{15.3a}$$

Equation 15.3a is called the **best linear unbiased estimator** (**BLUE**) of the vector $\boldsymbol{\beta}$ of fixed effects. **Estimability** of the fixed effects can be an issue, as the structure of the data (indicated by the column rank of $\mathbf{X}$) may not allow for unique estimates of all fixed effects. In such cases, generalized inverses can be used to obtain unique estimates of certain linear combinations of the fixed effects (LW Appendix 2). If $\mathbf{X}$ has column rank $\ell \leq q$, then exactly ℓ combinations of fixed effects can be estimated, see LW Chapter 26 and LW Appendix 2 for further details. Finally, note that although the BLUEs are a function of $\mathbf{V}$ (and hence σ_A^2 and σ_e^2), applying Equation 15.2b shows that the phenotypic variance σ_z^2 in $\mathbf{V}$ cancels out in Equation 15.3a ($\mathbf{V}^{-1}$ scales as $1/\sigma_z^2$, while the inverse of $\mathbf{V}^{-1}$ scales as σ_z^2), leaving the BLUE estimate as a function of just the heritability.

The **best linear unbiased predictor** (**BLUP**) of the vector of random effects is given by

$$\widehat{\mathbf{a}} = \mathbf{Var}(\mathbf{a})\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \sigma_A^2\mathbf{A}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{15.3b}$$

These BLUPs for breeding values are often called **PBV** or **EBVs**, for **predicted** or **estimated breeding values**. Equation 15.3b is the regression of $\mathbf{a}$ on the vector of observations $\mathbf{y}$ adjusted for fixed effects ($\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$) and then suitably scaled. Even if the number of random effects exceeds the number of actual observations (i.e., $p > n$), Equation 15.3b still provides unique estimates of each (provided $\mathbf{V}^{-1}$ exists). This occurs because the covariance structure of the vector $\mathbf{a}$ is incorporated in the model. As with the BLUEs, BLUPs are just functions of h^2 as $\sigma_A^2 = h^2\sigma_z^2$, while $\mathbf{V}^{-1}$ scales as $1/\sigma_z^2$, leaving $\widehat{\mathbf{a}}$ as only a function of the heritability. An alternative expression for the vector of PBVs is

$$\widehat{\mathbf{a}} = \left(\mathbf{Z}^T\mathbf{M}\mathbf{Z} + \lambda\mathbf{A}^{-1}\right)^{-1}\mathbf{Z}^T\mathbf{M}\mathbf{y} \tag{15.3c}$$

where $\lambda = \sigma_e^2/\sigma_A^2 = (1-h^2)/h^2$ and

$$\mathbf{M} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{15.3d}$$

is the absorption matrix for the fixed effects (Kenndey and Trus 1993).

In practice, Equations 15.3a-c are often not used, as they require inversion of the potentially very large matrix $\mathbf{V}$. Instead, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{a}}$ are obtained without computing an inverse by numerically solving (for example, by Gaussian elimination) **Henderson's mixed model equations**

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\,\mathbf{A}^{-1} \end{pmatrix}\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix} \tag{15.4}$$

The careful reader might wonder why we worried so much about avoiding taking the inverse of $\mathbf{V}$, but that the mixed-model equations contain $\mathbf{A}^{-1}$, which as first blush looks just as complicated to invert at $\mathbf{V}$. However, it turns out that $\mathbf{A}$ is very easy to invert (Henderson 1976, Quaas 1976), indeed $\mathbf{A}^{-1}$ is often directly computed from the pedigree in place of $\mathbf{A}$. The variance-covariance matrices for $\widehat{\mathbf{a}}$ and $\widehat{\boldsymbol{\beta}}$ also follow from the mixed-model equations. Denote the inverse of the matrix in Equation 15.4 by

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\,\mathbf{A}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{15.5a}$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $k \times k$, $k \times p$, and $p \times p$ submatrices. Using this notation, Henderson (1975) showed that the covariance matrix for the BLUE vector of fixed effects $\boldsymbol{\beta}$ is given by

$$\mathbf{Var}(\widehat{\boldsymbol{\beta}}) = \sigma_e^2\,\mathbf{C}_{11} \tag{15.5b}$$

The **prediction error variances** (**PEVs**), the variances (and covariances) among the vector of prediction errors ($\widehat{\mathbf{a}} - \mathbf{a}$) are given by

$$\mathbf{Var}(\,\widehat{\mathbf{a}} - \mathbf{a}\,) = \sigma_e^2\,\mathbf{C}_{22} \tag{15.5c}$$

following Equation 15.3c, we can also expressthis as

$$\mathbf{Var}(\,\widehat{\mathbf{a}} - \mathbf{a}\,) = \sigma_e^2\left(\mathbf{Z}^T\mathbf{M}\mathbf{Z} + \lambda\mathbf{A}^{-1}\right)^{-1} \tag{15.5d}$$

The prediction error variances for *individual* breeding values are not a serious concern here, as we measure response by taking the average breeding values over all individuals within each generation. However, as we will see in Chapter 16, they can be important when we try to disentangle selection in natural populations. Finally, the covariances between estimated fixed effects and prediction errors is

$$\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{a}} - \mathbf{a}\,) = \sigma_e^2\,\mathbf{C}_{12} \tag{15.5d}$$

These covariance matrix expressions assume that the variance is known without error. When BLUP estimates are obtained using an *estimated* variance, this additional source of error is *not* accounted from by Equations 15.5b–d.

One concern is that the above BLUPs and BLUEs may be biased by selection. Henderson (1975) showed that these are unbiased if (i) selection decisions are based on linear combinations of data (such as truncation selection based on individual phenotypes or a linear index based on the phenotypes of an individual and its relatives) and (ii) that if selection is based on records that are adjusted for fixed effects, the model uses estimates of these fixed effects in an unbiased fashion when selection is absent. Hence, Henderson's conditions for BLUPs and BLUEs being unbiased by selection hold under many reasonable forms of artificial selection.

REML Estimation of Unknown Variance Components

The variance components (σ_A^2 and σ_e^2), or at a minimum the heritability $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_e^2)$, must be specified to obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{a}}$. These variances are generally unknown, but can be estimated using **restricted maximum likelihood** (**REML**). REML is closely related to BLUP, with (roughly speaking) REML estimates obtained from iterating and updating BLUP estimates until suitable convergence. REML maximizes that part of the likelihood function that is not influenced by fixed effects, so the restricted likelihood refers to that part of the likelihood function unaffected by fixed effects (Patterson and Thompson 1971). Harville (1977) coined the term restricted ML, but Thompson (2008) notes that REML maximizes a *residual* likelihood, and hence prefers the term **residual maximum likelihood**. One advantage of REML estimates (over other variance estimation procedures) is that they are unbiased by the estimates of fixed effects (Patterson and Thompson 1971). We refer the

reader to the extensive discussion of REML variance estimation in LW Chapter 27 for further details, and to Hofer (1998), Thompson and Mäntysaari (2004), Thompson et al. (2005), and Thompson (2008) for a review of more recent developments, including computational issues. For the much of remainder of this chapter, we assume we have already obtained REML estimates of the variances proceeding the BLUP analysis (relaxing this assumption when discussing Bayesian Mixed-model analysis). This two-stage approach of BLUP using estimated variance components (in place of their true values) is often called **empirical BLUP** or **REMBL/BLUP**.

The covariance matrix for the REML estimates can be approximated by using the best quadratic fit of the restricted likelihood surface, centered at the REML estimates (Smith and Graser 1986, Graser et al. 1987). If $\boldsymbol{\sigma} = (\sigma_A^2, \sigma_e^2)^T$ is a vector of assumed variances, one computes the restricted likelihood $L(\boldsymbol{\sigma})$ for a grid of values close to the REML solution and then fits the best quadratic surface to the data,

$$L(\boldsymbol{\sigma}) = \mathbf{b}_0 + \boldsymbol{\sigma}^T \mathbf{b}_1 + \boldsymbol{\sigma}^T \mathbf{S} \boldsymbol{\sigma} \tag{15.6a}$$

where the vectors $\mathbf{b}_i$ and the symmetric matrix $\mathbf{S}$ are fitted using the data. (Chapter 28 discusses fitting the best quadratic surfaces in the context of fitness surface estimation.) The approximate covariance matrix for the vector of REMLs, $\widehat{\boldsymbol{\sigma}}$, is given by

$$\mathbf{Var}(\widehat{\boldsymbol{\sigma}}) \simeq (-2\mathbf{S})^{-1} \tag{15.6b}$$

The rationale for this approach is that the inverse of the matrix of second-order partial derivatives of the likelihood surface at the likelihood estimate (for large samples) approaches the covariance matrix of these estimates (LW Appendix 4). Equation 15.6a is a (second-order) multidimensional Taylor series (Appendix 6), with $2\mathbf{S}$ corresponding to the matrix of second-order partial derivatives (the **Hessian matrix**) of the likelihood function. Even though Equation 15.6 allows a (large-sample) estimate of the uncertainty in our estimate of the variance, this cannot easily be incorporated to provide a measure of how much additional uncertainty this introduces into BLUP estimates. As is discussed at the end of this chapter, Bayesian approaches, on the other hand, are exact (given a specified prior) for any sample size and fully incorporate the uncertainty in the variance estimates into the uncertainty in the BLUP estimates.

REML Can (Often) Return Variance Estimates Unbiased by Selection

As with BLUPs and BLUEs, REML estimates of variance are often unbiased by selection. In particular, if the base population consists of unselected and noninbred individuals and phenotypic data are available for all selected and unselected individuals, then under the infinitesimal model REML yields essentially unbiased estimates of the additive genetic variance in the base population (Henderson 1949, Henderson et al. 1959, Curnow 1961, Thompson 1973, Rothschild et al. 1979, Sorensen and Kennedy 1984b, Gianola and Fernando 1986, Gianola et al. 1988, Juga and Thompson 1989, Gianola et al. 1989, Fernando and Gianola 1990). We will show shortly that this arises because (assuming the infinitesimal model) the relationship matrix $\mathbf{A}$ fully accounts for drift and linkage-disequilibrium generated by selection and/or assortative mating. Thus, if the base-population is non-inbred and in linkage-equilibrium ($d(0) = 0$, see Chapter 13), then $\mathbf{A}$ in conjunction with the records of all individuals is sufficient to account for the linkage disequilibrium that is subsequently generated by selection and/or assortative mating. Simulations by van der Werf and de Boer (1990) show that if the model includes the pedigree information for all individuals ($\mathbf{A}$) but is missing records for some individuals (e.g., Example 15.4), then REML does not necessarily yield unbiased estimates. Simulations by Jeyaruban and Gibson (1996) showed that the bias may vary with heritability, finding (in their model) that bias increased with heritability.

When the base population consists of previously selected individuals, REML provides no protection from biased estimates of the additive genetic variance in the population prior to selection, even if the entire pedigree of individuals back to the base population is included (van der Werf 1990, van der Werf and de Boer 1990, van der Werf and Thompson 1992). This arises because $d \neq 0$ in the base population and hence (without knowledge of the actual base-population d value), $\mathbf{A}$ does not fully account for the dynamics of d (as it does when we start with $d = 0$). Finally, if selection acts on a suite of unmeasured characters that are correlated with characters included in the model, REML can generate biased estimates of the variances and covariances of the measured characters (Schaeffer and Song 1978).

ANIMAL-MODEL ANALYSIS OF SELECTION EXERIMENTS

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. We examine the simplest version of the animal model first, considering various elaborations in later sections (also see LW Chapter 26). While this model has its origin in the animal breeding literature, it has very widespread applicability. We trust that plant scientists will not be not greatly offended, as the "animal" model can be used to analyze plant selection experiments as well!

The Basic Animal Model

To apply the animal model to a selection experiment, first vectorize the observations from the entire experiment by letting y_{ij} denote the jth 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,

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

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

$$y_{ij} = \mu + a_{ij} + e_{ij} \tag{15.7a}$$

where μ is an overall mean, a_{ij} the breeding value of the jth measured individual from generation i, and e_{ij} the deviation between breeding and phenotypic values. With exactly one record per individual, $\mathbf{Z} = \mathbf{I}$. In this simple model the only fixed effect is the mean, giving $\boldsymbol{\beta} = (\mu)$ and $\mathbf{X} = \mathbf{1}$ (a vector of ones), reducing Equation 15.1 to

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \mathbf{e} \tag{15.7b}$$

Here $\mathbf{a}$ is the vector of breeding values for all individuals measured during the course of the experiment, with $\mathbf{Var}(\mathbf{a}) = \sigma_A^2\mathbf{A}$. The relationship matrix $\mathbf{A}$ is the key to mixed-model analysis, as it includes all the pedigree information. Because of the information provided by $\mathbf{A}$, breeding values can be estimated for individuals without records, *provided* they have measured relatives in the analysis (see Example 15.4). 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, see LW Chapters 7, 26) describe the relatedness of

individuals i and j. Recursive methods for obtain the elements of $\mathbf{A}$ (and $\mathbf{A}^{-1}$) given a pedigree are discussed in LW Chapter 26. The simple animal model assumes that all genetic variance is additive, so that there is no (genetic) covariance between residuals (we relax this assumption below). In this case, it is generally assumed that $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$, and the mixed-model equations (Equation 15.4) simplify to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} n\,\overline{y} \\ \mathbf{y} \end{pmatrix} \tag{15.7c}$$

where n is the total number of individuals in the experiment, $\lambda = \sigma_e^2/\sigma_A^2 = (1-h^2)/h^2$, $\widehat{\mathbf{a}}$ is the n-dimensional vector of the predicted breeding values of all measured individuals, and $\mathbf{1}$ is a vector of ones. Likewise, the covariance matrices for the fixed effects $\mathbf{C}_{11}$, predictor errors for the BLUPs of breeding values $\mathbf{C}_{22}$, and the covariances $\mathbf{C}_{12}$ between these estimators are given by

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda\,\mathbf{A}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{15.7d}$$

Response is Measured by Change in Mean Breeding Values

Under a mixed-model analysis, response is measured by the change in the mean breeding value of a selected population over time. Mixed-models easily allow for overlapping generations by simply predicting breeding values at discrete time points (say every year) instead of each generation. Hence, in what follows, one can easily replace "generation" by "year" or some other time measure.

The estimated mean breeding value in generation k is simply given by the average of individual breeding values for that generation,

$$\widehat{\overline{\mathbf{a}}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \widehat{a}_{kj} \tag{15.8a}$$

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

$$\widehat{\overline{\mathbf{a}}} = \mathbf{K}^T \widehat{\mathbf{a}} \tag{15.8b}$$

where the ith 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).

From Equation 15.5d, and recalling that $\mathbf{Var}(\mathbf{Bx}) = \mathbf{B}\,\mathbf{Var}(\mathbf{x})\mathbf{B}^T$ (LW Equation 8.21b), the sampling covariance matrix for the vector of estimated genotypic means becomes

$$\mathbf{Var}\left(\widehat{\overline{\mathbf{a}}}\right) = \sigma_e^2\,\mathbf{K}^T\mathbf{C}_{22}\mathbf{K} \tag{15.8c}$$

where the $n \times n$ matrix $\mathbf{C}_{22}$ is the solution to Equation 15.7d (under the simple animal model) or more generally by Equation 15.5b. Again, these expressions assume that the

residual variance is known without error, and hence using an estimate adds an additional source of uncertainty not accounted for by Equation 15.8c.

Example 15.1. As an example of how one uses a mixed-model analysis in a selection experiment, consider the following very simple situation. From a base population of unrelated and non-inbred individuals, four (indexed by 1-4) are measured and have trait values of 3, 6, 5, and 2, respectively. The two largest individuals are mated, and their resulting offspring (individuals 5-8) have values 4, 5, 6, 5. Now suppose we have either a REML-based estimate of the heritability (not advisable here given the every small sample size) or have previous knowledge of its value, so that our goal is to estimate the size of the genetic response. Assuming the only fixed effect is the mean, the resulting animal model is $\mathbf{y} = \mathbf{1}\boldsymbol{\beta} + \mathbf{a} + \mathbf{e}$, where

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

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

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

For example, individuals 2 and 5 are parent and offspring, so $A_{5,2} = A_{2,5} = 1/2$, as cov(P,O) = $\sigma_A^2/2$. Similarly, individuals 7 and 8 are full-sibs, so that $A_{7,8} = A_{8,7} = 1/2$. Note that the relationship matrix for the founders (base population members) is given by the 4×4 identity submatrix in the upper left of $\mathbf{A}$. This identity matrix implies that non-inbred (diagonal elements are one) and unrelated individuals (off-diagonal elements zero) formed the base population.

Turning to the covariance matrix of the residuals, we make the standard assumption that $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\mathbf{I}$, i.e., all residuals are uncorrelated. However, if there is dominance, the residuals among full-subs are inflated by $\sigma_D^2/4$. Likewise, if there are common family effects (e.g., maternal effects or other shared environmental effects) the residuals are inflated by σ_c^2, where c the is common family effect. We ignore these possible complications for now, which are easily accounted for by adding additional random effects to the model (see below).

Suppose (from REML or prior knowledge) that the heritability of the trait is $h^2 = 0.3$. Applying Equation 15.3a gives

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

where we have computed $\mathbf{V}$ using Equation 15.2b scaled to remove the phenotypic variance σ_z^2, e.g., $\mathbf{V} = 0.3\,\mathbf{A} + 0.7\,\mathbf{I}$. Substituting into Equation 15.3b gives the BLUPs for the genetic values and resulting genetic means as

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

Note that (by construction) the mean breeding value in the base population is zero. Hence, the estimated response (for $h^2 = 0.3$) is 0.562. 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

A Bayesian analysis removes this dependency of the responses on h^2 by computing a weighted average of response over all possible h^2 values (weighted by their likelihood), resulting in a marginal posterior distribution for the response that is independent of the heritability.

Turning to a more standard analysis, the selection differential is the mean of the selected parents minus the mean of all parents, $S = 5.5 - 4 = 1.5$. Likewise, the response is the mean of the offspring minus the mean in the previous generation, $5 - 4 = 1$, giving a realized heritablity of $R/S = 1.0/1.5 = 0.67$.

Note that a mixed-model 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 using the mean estimated breeding values (Figure 15.1). Response is underestimated if the assumed heritability is less than the true value, while it is overestimated if the heritability is overestimated (Sorensen and Kennedy 1984a). As Figure 15.1 shows, 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 (Sorensen and Kennedy 1986). In a two-stage **REML/BLUP analysis**, one first estimates the additive genetic variance in the base population (via REML), using this value in the subsequent BLUP analysis. Kackar and Harville (1981) and Gianola et al. (1988) show that using the REML estimates does not result in biased values for BLUPs, but that the resulting predictors may not be "best" (there may be other linear predictors with smaller errors).

One caveat that must be stressed is that the estimated mean breeding values should not themselves be used to estimate heritabilities under a standard MM analysis. 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 instead be based on REML (or other) estimates of the base-population variance components. However, under a Bayesian analysis the (marginal posterior) estimate of mean breeding value in any particular generation is independent of the heritability, and thus a regression of the mean breeding value on generations can indeed return an unbiased estimate of realized heritability.

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 base population additive variance has the same expected value in each line. Departures from either of these two predictions indicates potential failure of the model.

Example 15.2. 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 (Figure 15.1). 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 high, low, and control lines, for a total of almost 11,000 mice over the course of the experiment. Within-family selection (Chapter 17) 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 (i.e., family) effects (see Example 15.6 for details).

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 $\mathbf{A}$ 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 complete 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 15.2 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 versus late 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. In contrast, Martinez et al. (2000) found no changes in variance estimates over 20 generations of selection for body composition (fat pad to body weight ratio) in mice. These authors examined REML estimates of the additive variance (and heritability) using various subsets of the full twenty-generation data. Consistent estimates of the additive variance were obtained using all records and the complete pedigree from generations 0-20, using records from generations 9–20 but with pedigree information from generation 0, and finally looking at the phenotypic data in blocks of three generations. They conclude that the selection response, while resulting in a roughly four-fold change in mean, was still well-fit by the infinitesimal model.

Separating Genetic and Environmental Trends

The observed improvement in a trait over time (such as milk yield) could be do entirely to improvement in the environment (better husbandry and nutrition), entirely from genetic changes (response from selected breeding), or (most likely) a combination of both. Thus, it is critical to partition an observed phenotypic change into genetic and environmental components. For example, Southwood and Kennedy (1991) showed that the improvement in several litter-size related traits in pigs over a ten year period in Quebec was entirely due to environmental, rather than genetic, changes.

In a least-squares analysis, any underlying environmental trend is assumed to be removed by contrasting selected and control populations (or contrasting populations selected in opposite directions). The rational is that the kth individual from population j in generation t can be written as

$$y_{tjk} = \mu + d_t + a_{tjk} + e_{tjk} \tag{15.9}$$

where d_t is the environmental trend. If the common environmental value is the same in both selected and control populations, then the difference in phenotypic means in generation t is

$$\overline{z}_{s,t} - \overline{z}_{c,t} = (\,\overline{a}_{s,t} - \overline{a}_{c,t}\,) + (e_{s,t} - e_{c,t}) \tag{15.10}$$

The residuals e have expected value zero, hence the contrast provides an estimate of $\overline{a}_{s,t}$, *provided* there is no significant drift in the mean breeding value of the control population (so that $\overline{a}_{c,t} \simeq 0$). However, if genotype-environment interactions are present, the environment value for generation t can differ between populations, in which case Equation 15.9 has an additional term $(d_{s,t} - d_{c,t})$. Hence, even when a control population is used, a least-squares analysis can still give biased results if there is significant drift in the mean of the control population ($|\,\overline{a}_{c,t}\,| >> 0$) and/or significant $G \times E$.

A mixed-model analysis estimates the mean *breeding value*, rather than the *phenotypic mean*, of the population. Hence a MM analysis allows for the separation of the genetic change from any environmental change (Henderson et al. 1959, Blair and Pollak 1984, Sorensen and Kennedy 1984a). This occurs because **A** tracks the flow of genes through the population, allowing for estimates of breeding values independent of environmental effects. Of course, this is strictly dependent on the model assumptions holding, but if they do, a mixed-model analysis does not require a control population. Common environment effects are incorporated into the basic animal model by simply adding a fixed effect d_i for common environmental effect in generation i,

$$y_{ij} = \mu + d_i + a_{ij} + e_{ij} \tag{15.11}$$

The vector of fixed effects now becomes

$$\boldsymbol{\beta} = (\mu, d_1, d_2, \cdots, d_T)^T$$

and the corresponding incident (or design) matrix $\mathbf{X}$ has ones in the columns (two through $T+1$) corresponding to the generation in which the individual was scored, viz.,

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Example 15.3. To examine the potential bias from $G \times E$ and drift in the control population, Blair and Pollak (1984) examined a seven-generation selection experiment on 14-month greasy fleece weight in sheep. The model they assumed was that the mth individual in generation t with fixed sex effect (male/female) i, fixed age of dam effect (mature/immature) j, and fixed rearing rank effect (single/twin) k had a phenotypic value of

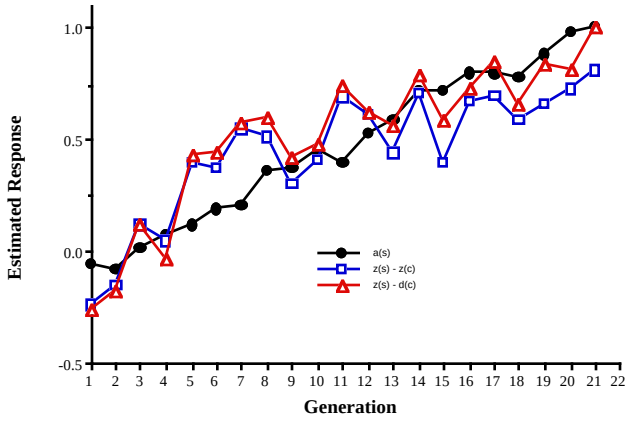
$$y_{tmijk} = sx_i + b_j + r_k + d_t + a_{tm} + e_{tmijk}$$

In matrix form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$, where the vector $\boldsymbol{\beta}$ contains the fixed-effects for sex (sx), dam age (b), and rearing rank (r) in addition to the effects for years (d). Both the selected and control line were subjected to a BLUP analysis using this model, and three different estimates of selection response were considered:

(i) $\widehat{\overline{z}}_{st} - \widehat{\overline{z}}_{ct}$, the estimated phenotypic means following adjustment for the fixed-effects (sx, b, and r), obtained by $\widehat{\overline{z}}_{xt} = \widehat{\overline{a}}_{xt} + \widehat{d}_{xt}$. This is an unbiased estimate of the response if there is no significant drift in the control population ($\overline{a}_{ct} \simeq 0$) and no $G \times E$, so that $d_{ct} = d_{st}$.

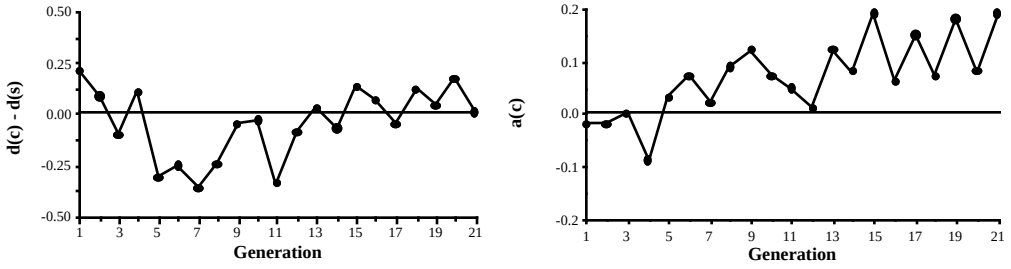
(ii) $\widehat{\overline{z}}_{st} - \widehat{d}_{ct}$, the (fixed-effects adjusted) phenotypic mean of the selected population minus the common environmental effect, as estimated from the control population.

(iii) $\widehat{\overline{a}}_{st}$, the BLUP estimate of the mean breeding value in the selected population.



Estimate **(i)** mimics that used in a least-squares analysis, and Blair and Pollak show it is independent of the assumed heritability. Estimates **(ii)** and **(iii)** are highly dependent on the assumed (or estimated) heritabilties in the control and selected populations. As the above figure shows, all three estimates show a positive genetic trend (following a reversed response over the first few generations). The estimated response using only the predicted mean breeding value is smoother (as expected) than the other two estimates.

The potential biases in a least-squares analysis of this data using the contrast between the control and selected phenotypic means are seen in the two graphs below. The left graph plots the difference in the estimated common environmental effects between selected and control populations ($\widehat{d}_{ct} - \widehat{d}_{st}$). Ignoring the inherent variance in estimating the $\widehat{d}$'s, the difference (via a paired t-test) is not significantly different from zero. The right graph plots the predicted mean breeding value of the control population, which is assumed to be zero under the least-squares analysis. As the figure shows, there is a slight, but positive, trend. When the control mean is subtracted off, the net result is that the LS analysis slightly underestimates the true response. Thus, there is no evidence of error being introduced by $G \times E$ differences between the control and selected lines, but error is introduced by the mean breeding value of the control population departing significantly from zero.



Validation That a Trend is Indeed Genetic

The estimated environmental (d) and additive-genetic (a) effects are highly-dependent on the estimated (or assumed) base-population heritability h^2. Hence, using BLUP to separate genetic from environmental values is highly dependent on the heritability used being close to its true value. Other departures from mixed-model assumptions (e.g., the infinitesimal model, and that BLUP and REML estimates are unaffected by selection) can also result an incorrect assignment of the relative importance of environmental versus genetic values. Thus, one must be cautious when relying on a BLUP analysis to separate genetic from environmental effects, and some sort of validation of the trend is critical.

We have already discussed validation of the general animal model, namely by looking at the consistency of the estimated additive variance using different subsets of the data. Similar validation that a trend is indeed genetic can be performed by again looking for consistency across the analysis. For example, if a control population is used, a BLUP analysis can estimate the amount of drift in the mean breeding value from its expected value of zero and estimates of the selection and control environmental effects can be compared (e.g., Example 15.3). If these are reasonably consistent, then a joint analysis (assuming the same environmental values in both populations) may yield more precise estimates of the generational environmental values. Likewise, if the estimates are significantly different, the possibility of either $G \times E$ and/or different environmental values in selected versus controlled lines needs to be seriously considered. Thus, even with a control population, there is still much to be gained by subjecting each to a mixed-model (e.g., BLUP) analysis.

Even in the absence of a control population, one can still attempt trend validation. For

example, Boichard et al (1995) examined several different methods to attempt validation of estimated genetic trends in dairy cattle, all of which involving comparing predicted trend values using different subsets of the data. The authors were interested in comparing the performance of a model (AM90) used for French Holsteins from 1990 to 1992 with a more recent model (AM93) used since 1993. For a variety of reasons, there was concern that the older AM90 model yielded biased estimates of the genetic trend. One check was based on the fact that milk yield data is in the form of multiple records per individual so that a repeatability model is appropriate (Example 15.5). The authors compared estimates of EBVs based only on the first lactations with estimates based on using all lacations. They found that the trend estimated from the first versus all lactations agree well under the AM93 model, but differ dramatically under the AM90 model. Two other measures of consistency (for example, looking at the stability of estimated breeding values of individuals as more information is added) also showed that the newer AM93 model seem relatively robust, while the older AM90 model seemed to produce biased estimates of the trend.

Replicate Lines

It is straightforward to jointly analyze multiple lines simultaneously. For k lines, write the total vector of observations as $\mathbf{y}^T = (\mathbf{y}_{l1}^T, \mathbf{y}_{l2}^T, \cdots \mathbf{y}_{lk}^T)$ where $\mathbf{y}_{li}$ is the vector of total observations from line i. If the generational environmental effects are assumed to be the same in each line, the model for the ith individual in generation t from line k is

$$y_{kti} = \mu + d_t + a_{kti} + e_{kti} \tag{15.12a}$$

Alternatively, if the environmental effects are potentially different in each population, then

$$y_{kti} = \mu + d_{kt} + a_{kti} + e_{kti} \tag{15.12b}$$

The power of combining multiple lines arises when effects can be assumed to be the same across lines. In this case, the effective sample size for estimating each effect is increased, and (presumably) the resulting sampling variance decreased, improving the precision of the estimates. The assumed covariance matrix for the vector of joint breeding values can also take several forms. If the founding members for each line are drawn from the same base population (but otherwise unrelated), then the covariance matrix for the vector of breeding values $\mathbf{a}$ has block-diagonal form, with the ith block corresponding to $\sigma_A^2\mathbf{A}(i)$, the numerator relationship matrix for line i times the base-population additive variance. If the founding members of at least some different lines are related, then $\mathbf{A}$ is more complex, reflecting these relationships. Further modifications for joint analysis have been proposed by Visscher and Thompson (1990), and extended by Beniwal et al. (1992a,b) and Heath et al. (1995) by allowing the additive variance to change. For example, one might assume that the additive variance remains constant for the first few generations of selection, after which it assumes a different value. This is a logical, but still ad-hoc, approach towards dealing with potential departures from the infinitesimal model.

Estimating the Additive Variance at Generation t

Even under the infinitesimal model, the additive variance changes over time. While REML provides an estimate of the base-population additive variance (which is unbiased provided the model assumptions hold), it does not immediately provide estimates of the actual additive variance in any particular generation of selection. The most straightforward approach is to use the parent-offspring regression for each generation of selection to estimate the additive genetic variance in the parents. With parents from generation t and offspring in generation $t+1$, the regression estimates the heritability of the parents, $h_A^2(t)$ (Robertson

1977). The drawback to this approach is the typically small sample size associated with each generation (resulting in large standard errors for each heritability/variance estimate). Ideally, one would like to be able to combine information across generations in such a way as to improve the variance estimates.

Sorenson and Kennedy (1984b) suggest one approach for combining information to estimate the variance in generation t is to use a mixed-model analysis treating generation t as the base population. In particular, one considers only the data from generation t onward (say to generation T) and the relationship matrix is adjusted to assume that generation t is the base population. The resulting covariance matrix for the breeding values becomes

$$\mathbf{Var}\begin{pmatrix}\mathbf{a}_t\\ \mathbf{a}_{t+1}\\ \vdots\\ \mathbf{a}_T\end{pmatrix} = \sigma^2_A(t)\begin{pmatrix}\mathbf{I} & \mathbf{A}_{t,t+1} & \cdots & \mathbf{A}_{t,T}\\ \mathbf{A}_{t,t+1} & \mathbf{A}_{t+1,t+1} & \cdots & \mathbf{A}_{t+1,T}\\ \vdots & \ddots & \vdots & \vdots\\ \mathbf{A}_{t,T} & \mathbf{A}_{t+1,T} & \cdots & \mathbf{A}_{T,T}\end{pmatrix} \tag{15.13}$$

where $\mathbf{a}_k$ is the vector of breeding values in generation k. $\mathbf{A}_{jk}$ is the relationship matrix of associations between individuals in generations j and k. By assuming $\mathbf{Var}(\mathbf{a}_t) = \mathbf{I}$, we are assuming that all individuals in generation t are unrelated and noninbred, as this is now our assumed base population. All measured individuals from generation t (including those not leaving offspring) are included in the base population. While this approach seems logical, it is still somewhat ad-hoc and is not exact. Simulation studies by van der Werf and de Boer (1990) show that Sorenson and Kennedy's approach tends to overestimate the true variance.

Another potential candidate would be to use the variance among the predicted breeding values within a generation, viz.,

$$\text{Var}(A_t) = \frac{1}{n-1}\sum_{i=1}^{n_t}\left(\widehat{a}_{ti} - \widehat{\overline{a}}_t\right)^2 \tag{15.14}$$

While again this is a reasonable suggestion, there are complications. First, there is a level of uncertainty in that the breeding values are all estimated. Second is that the assumed genetic variance used to obtain the BLUP estimates has a strong influence on the values of the estimated $\widehat{a}_{ti}$. However, the use of Equation 15.14 in a *Bayesian* framework (wherein the uncertainty in variance estimates is naturally incorporated into the analysis) avoids many of these problems (Sorensen et al. 2001), although this approach is computationally intense.

THE RELATIONSHIP MATRIX ACCOUNTS FOR DRIFT AND DISEQUILIBRIUM

As reviewed in Chapter 13, selection changes the additive variance by generating gametic-phase disequilibrium even in the absence of allele frequency change. Additionally, with a finite number of loci, selection also changes allele frequencies, further changing the genetic variance. While gametic-phase disequilibrium changes in the genetic variances are generally restricted to the first few generations of selection (Chapter 13), allele frequency changes become increasingly more important as selection proceeds (Chapters 24 - 26). 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* one corrects for this reduction (e.g., Equation 14.21).

Under the infinitesimal model assumptions, a MM analysis fully accounts for the effects of gametic-phase disequilibrium as well as genetic drift (Sorensen and Kennedy 1983, 1984a), *provided* the base population consists of unrelated, non-inbred individuals in linkage equilibrium. 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, $\mathbf{Var}(\mathbf{a}) = \sigma_A^2\mathbf{A}$.

This conditional 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 parents, A_{m_i} and A_{f_i}. Recall from Equation 13.8a that

$$A_i = \frac{1}{2}A_{f_i} + \frac{1}{2}A_{m_i} + s_i \tag{15.15}$$

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 $(\,1 - \overline{f}_i\,)\,\sigma_A^2/2$. Here $\overline{f}_i$ is the average 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 $\mathbf{F}$ a diagonal with ith element $(1 - \overline{f}_i)$, one minus the average inbreeding of the parents of i. If k and j are the parents of offspring i, then

$$F_{ii} = (\,1 - \overline{f}_i\,) = \left(1 - \frac{f_k + f_j}{2}\right) = \left(2 - \frac{A_{kk} + A_{jj}}{2}\right) \tag{15.16}$$

where $A_{kk} = (1 + f_k)$ denotes the kth 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). Drift effects the variance of the mendelian sampling but this is fully accounted for by the matrix $\mathbf{F}$. 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 $\mathbf{Var}(\mathbf{a}) = \mathbf{T}\,\mathbf{Var}(\mathbf{s})\,\mathbf{T}^T$, where $\mathbf{Var}(\mathbf{s}) = (\sigma_A^2/2)\mathbf{F}$ is independent of selection and assortative mating, with $\mathbf{F}$ accounting 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, Henderson (1976) and Thompson (1977) show that $\mathbf{A}$ can be written as a function of a diagonal matrix, $\mathbf{D}$, and an upper triangular matrix $\mathbf{T}$,

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^T \tag{15.17}$$

$\mathbf{T}$ 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 15.17, the covariance matrix for $\mathbf{g}$ becomes

$$\mathbf{Var}(\mathbf{g}) = \mathbf{T}^{-1}\mathbf{Var}(\mathbf{a})(\mathbf{T}^{\,T})^{-1} = \sigma_A^2\,\mathbf{T}^{-1}\mathbf{T}\mathbf{D}\mathbf{T}^{\,T}(\mathbf{T}^{\,T})^{-1} = \sigma_A^2\,\mathbf{D} \tag{15.18}$$

which follows using the 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 ith element of $\mathbf{g}$ equals

$$g_i = A_i - \frac{1}{2}A_{f_i} - \frac{1}{2}A_{m_i} = s_i \tag{15.19}$$

which is simply the segregation residual. Hence, $\mathbf{g} = \mathbf{s}$, and since $\mathbf{Var}(\mathbf{s}) = (\sigma_A^2/2)\mathbf{F}$, 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, while $\mathbf{F}$ accouns for drift.

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. Provided that the change in allele frequencies is small over the course of the experiment, this bias may not be too serious. Another key assumption from the infinitesimal model is that the distribution of residuals does not significantly deviate from normality. Chapter 24 examines this rather technical issue in some detail.

MODIFICATIONS OF THE BASIC ANIMAL MODEL

The strength of a mixed-model analysis is its flexibility. For example, one can (often) predict breeding values for individuals with no records, as the following example shows.

Example 15.4. Consider the simple case of an unmeasured parent with two measured offspring, both of which have a single record. We index the parent by 0, and the offspring by 1 and 2. Assuming a single fixed effect (the mean μ), the resulting mixed model becomes

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

To complete the mixed-model, we need the relationship matix $\mathbf{A}$, whose structure depends on the relationship between the two sibs. Assuming no inbreeding, these are either full- or half-sibs, with resulting $\mathbf{A}$ matrices

$$\text{full-sibs:} \quad \mathbf{A} = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}, \qquad \text{half-sibs:} \quad \mathbf{A} = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/4 \\ 1/2 & 1/4 & 1 \end{pmatrix}$$

Thus, we can estimate the breeding value a_0 for the unmeasured parent. Provided unmeasured individuals have measured relatives, $\mathbf{A}$ allows us to estimate their breeding values.

Connections between relatives are often referred to as **links**, and the amount of links (or **connectiveness**) in a relationship matrix is one measure of its precision in estimating breeding values (i.e., related to the preidction error variance, see Kennedy and Trus 1993). The breeding value of an unmeasured individual with few measured relatives will have a much less precision than that for an individual with a large number of measured relatives. Despite this flexibility in predicting breeding values for individuals with missing records, it is important to stress that simple inclusion of all pedigree relationships appears to be *not* sufficient to yield unbiased

BLUP/REML estimates when selection occurs (van der Werf 1990, van der Werf and de Boer 1990, van der Werf and Thompson 1992).

In many cases, it is prudent to modify the simple animal model by considering additional fixed and random effects. For example, genetic and environmental effects can be separated without a control population by adding fixed effects to account for environmental trends. Likewise, it is often reasonable to include additional random effects, such as maternal/litter effects. Another modification of the basic model occurs when the phenotypic scores (records) of parents are unknown. Example 15.4 showed how we can estimate these as random effects, but likely with bias (when selection occurs). An alternative is to treat these breeding values of unmeasured parents as fixed, rather than random, effects. We deal with each of these modification in turn.

Models with Additional Random Effects

We have assumed that residuals are uncorrelated and homoscedastic, giving their covariance matrix as $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$. When additional random effects are present, but ignored by the model, they are subsumed into the residuals, potentially introducing correlations and heteroscedasticity. For example, if sibs share a common maternal environment, this introduces correlations between sibs beyond those accounted for by $\mathbf{A}$. If the model only includes $\mathbf{a}$ and $\mathbf{e}$, this additional covariance appears between the residuals, and the true covariance matrix for $\mathbf{e}$ is no longer the assumed diagonal, leading to biased estimates of the BLUEs and BLUPs. By suitably incorporating additional random effects, we can develope a new model where the residuals again have the simple covariance structure $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$.

Suppose there is a second vector $\mathbf{u}$ of m random effects in addition to the vector $\mathbf{a}$ of p breeding values and vector of residuals $\mathbf{e}$. Equation 15.1 becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{u} + \mathbf{e} \tag{15.20a}$$

where $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are $n \times q$, $n \times p$ and $n \times m$ incident matrices. The covariance structure assumed is $\mathbf{Var}(\mathbf{a}) = \sigma_A^2\,\mathbf{A}$, $\mathbf{Var}(\mathbf{u}) = \sigma_u^2\,\mathbf{I}$, and $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$, giving the covariance matrix for $\mathbf{y}$ as

$$\mathbf{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}^T\sigma_A^2 + \mathbf{W}\mathbf{W}^T\sigma_u^2 + \mathbf{I}\,\sigma_e^2 \tag{15.20b}$$

If we had incorrectly assumed the true model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$, the (true) covariance matrix for the residuals is $\mathbf{Var}(\mathbf{e}) = \mathbf{W}\mathbf{W}^T\sigma_u^2 + \mathbf{I}\,\sigma_e^2$, showing how the additional random effects alters the covariance matrix. The resulting mixed-model equations for 15.20a become

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{W} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T\mathbf{W} + \lambda_u\mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{pmatrix} \tag{15.21a}$$

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_u^2} \tag{15.21b}$$

Additional vectors of random effects can be incorporated in a similar manner, see LW Chapters 26 and 27 for details. The mixed-model equation again form the basis for iterative REML estimates of the unknown variance components (σ_A^2, σ_u^2, and σ_e^2), as discussed in detail in LW Chapter 27.

Example 15.5. Often the same trait is measured multiple times in the same individual, for example, the sizes of different litters from a sigle female. When multiple records are present for at least some individuals, a **repeatability model** should be used (LW Chapter 26). Repeated measures from the same individual have three components: a genetic value a_k, a common (permanent) environmental value p_k that is the same in each measurement, and the residual environmental value e varying between each measurement, giving the ith measurement of kth individual as $a_k + p_k + e_{ki}$. The **repeatability** of the trait is $r = (\sigma^2_A + \sigma^2_p)/\sigma^2_z$, giving the variance of the residuals as $\sigma^2_e = (1-r)\sigma^2_z$ and the variance of permanent environmental effects as $\sigma^2_p = (r - h^2)\sigma^2_z$. We remind the reader that "environmental "effects can also include non-additive genetic components, as these are not passed along to offspring.

The repeatability model was used by Estany et al. (1989) to examined the selection response for litter size in rabbits. Their model assumed two groups of fixed effects, d_t the year-season (environmental) effect which had 22 levels in this experiment and the reproductive state l_i of the doe (l has three levels: l_1 for primiparious does, l_2 for lactating does, and l_2 for non-primiparious and non-lactating does). Since only two of these l_x factors are estimable, l_1 was assigned a value zero. Their model had three random effects, a_k and p_k for the additive genetic and permanent environmental effect of the kth doe, and the residual e, giving the overall model as

$$y_{tk\ell i} = \mu + l_i + d_t + a_k + p_k + e_{tk\ell i}$$

where $y_{tk\ell i}$ denotes the litter size for the ℓth litter of doe k in reproductive state i in season-year t. In matrix form, the mixed-model becomes

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

where $\mathbf{a}$ and $\mathbf{p}$ are $n \times 1$ vectors corresponding to the n does, $\mathbf{Var}(\mathbf{a}) = \sigma^2_A\,\mathbf{A}$, $\mathbf{Var}(\mathbf{p}) = \sigma^2_p\,\mathbf{I}$, and $\mathbf{Var}(\mathbf{e}) = \sigma^2_e\,\mathbf{I}$. $\mathbf{X}$ and $\mathbf{Z}$ are incident matrices, and the vector of fixed effects is

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ l_1 \\ l_2 \\ d_1 \\ \vdots \\ d_{22} \end{pmatrix}$$

The mixed-model equations are given by Equation 15.21a with

$$\lambda_A = \frac{\sigma^2_e}{\sigma^2_A} = \frac{1-r}{h^2} \quad \text{and} \quad \lambda_u = \frac{\sigma^2_e}{\sigma^2_p} = \frac{1-r}{r-h^2}$$

The careful reader might notice that the two vectors of random effects, the breeding values $\mathbf{a}$ and permanent environment effects $\mathbf{p}$, enter the model as $\mathbf{Za}$ and $\mathbf{Zp}$, respectively. Why then do we simply not combine these, e.g., $\mathbf{Zu}$ where $\mathbf{u} = \mathbf{a}+\mathbf{p}$? The reason we cannot do this (and indeed the reason we can estimate $\mathbf{a}$ and $\mathbf{p}$ separately!) is that $\mathbf{a}$ and $\mathbf{p}$ have *different covariance structures*, $\sigma^2_A\,\mathbf{A}$ versus $\sigma^2_p\,\mathbf{I}$. Thus, we assume that permanent environment effects are uncorrelated across individuals and are homoscedastic. On the other hand, breeding values generate covariances in relatives. Again, the critical importance of the covariance matrix to a mixed model analysis is apparent.

Common Family and Material Effects

Random effects are frequently included to account for any common family environmental effect when sibs are present. For example, if two sibs (i and j) share a common environmental value c, then $\sigma(e_i, e_j) = \sigma_c^2$. Hence, there are off-diagonal elements in the covariance matrix of residuals and we no longer have the standard assumption $\mathbf{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$.

Example 15.6. Meyer and Hill (1991), examining the response to selection on adjusted food intake (AFI) in mice (Example 15.2), formulated a model incorporating shared family values c as random effects. In addition, their model accounts for fixed-effects due to generations (d, 22 levels), lines (ln, 3 levels), sex (sx, male/female), and litter size (lt, 7 levels for litters of size 6 to 12 individuals). Under their model, the observed value for AFI from the kth individual from generation t, line ℓ and full-sib family i is given by

$$y_{t\ell ik} = \mu + d_t + ln_\ell + sx_j + lt_m + a_{t\ell ik} + c_{t\ell i} + e_{t\ell ik}$$

where this individual has sex j and experienced litter size m. In matrix form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{c} + \mathbf{e}$. The vector of fixed effects $\boldsymbol{\beta}$ contains the d, ln, sx, and lt values, while the random effects are the vector of common family effects $\mathbf{c}$, the vector of additive genetic values $\mathbf{a}$, and the vector of residuals $\mathbf{e}$. The incident matrix $\mathbf{W}$ has as its *ij*-th element 1 if individual i is from family j, else the element is zero. Note that Meyer and Hill have two model variables to account for litter effects — a fixed effect lt common to all litters of the same size and a random effect $c_{t\ell i}$ that varies between families but is the same for all individuals from a particular family. The resulting REML estimate for heritability was 0.15, while the fraction of the total variation (after removal of the fixed effects) accounted for by random family effects was estimated to be $c^2 = \sigma_c^2/\sigma_z^2 = 0.22$ ($\sigma_z^2 = \sigma_A^2 + \sigma_c^2 + \sigma_e^2$ is the trait variance following the removal of fixed effects). Hence, a larger fraction of the resemblance between sibs (the intraclass correlation, $t = h^2/2 + c^2$) is due to shared family environments, rather than due to shared genes. One caveat is that the model assumes no dominance variance. If present, sibs share dominance variance ($\sigma_D^2/4$), and under this model this is incorporated into σ_c^2.

Sib correlations (beyond those accounted for by their correlations in breeding value for the focal trait) can araise for (at least) three reasons. The first is dominance, which we will address shortly. The second and third are common family effects, which can potentially have two components: shared family and/or maternal environmental effects and shared maternal effects with a genetic component (i.e., material performance itself has a genetic component). The distinction between these two components of maternal effects (both of which may be included in a model) was foreshadowed in Example 15.5, in that they result in different covariance structures.

In Example 15.6, the common environment c was assumed to be uncorrelated across sibships, implying (assuming equal variances across families) that its covariance structure is $\mathbf{Var}(\mathbf{c}) = \sigma_c^2\,\mathbf{I}$. Hence, there is no correlation across different sibships and thus no shared information. Now suppose that a mother has several sibships. In this case, we can treat the common family environment in two distinct ways. In the first, each particular litter has a unique common family effect that is uncorrrelated across litters from the same mother. The second is to consider a repeatability model, wherein the contribution from the mother to a particular litter has two components. Considering litter k from mother i, the common family effect can be written as $c_{ik} + cp_k$, where (in terms of the repeatability model) cp_i is the permanent environmental effect of this mother (a common effect shared by all her litters), while c_{ik} is the unique environment shared by all members of her kth litter. In this case,

there are two additional random effects added to the model, and we also need to estimate the variances σ^2_{cp} and σ^2_c, responding to the contributions from cp_k and c_{ik}, respectively. We assume each litter has two or more individuals, otherwise c_{ik} and cp_k cannot be estimated seperately.

If we assume all of the maternal contribution is environmental, namely that a mother does not pass along any of her performance genetically to daughters, then the covariance structure for these two vectors of random effects would be $\mathbf{Var}(\mathbf{c}) = \sigma^2_c \mathbf{I}$ and $\mathbf{Var}(\mathbf{cp}) = \sigma^2_{cp} \mathbf{I}$. However, maternal performance could also have a genetic component, and thus female relatives have correlated maternal performances. In this case, we would add a third random effect ma (the breeding value of this maternal effect) to the model. This vector of breeding values for maternal performance has covariance matrix $\mathbf{Var}(\mathbf{ma}) = \sigma^2_A(ma)\mathbf{A}$. Distinguishing between maternal versus direct effects requires that there are paternal, as well as maternal, links in the pedigree (Kruuk 2004).

A further complication is that if we treat maternal performance as a genetic trait, our trait is determined by direct and maternal contributions, and the breeding values for these two can be correlated. This now makes the problem a multiple trait one (LW Chapter 27). We discuss maternal effects in greater detail in **Chapters xx,xx**.

From an operational standpoint, if maternal effects are suspected, at a minimum a common family effect should be included, and in the form of a repeatability model if the female has multiple litters (each of which results in several sibs). More generally, if there are many links between female relatives (with litters) in our data set, then one should seriously consider a genetic maternal effects model. Failure to do so may result in contributions from shared genetic maternal performance being regarded as breeding values for the direct trait, giving a biased picture of the nature of selection response (Milner et al. 2000, Kruuk 2004).

Treating Certain Breeding Values as Fixed Effects

How should one proceed if the base-population has itself been under selection? If this is known or suspected to be the case, Graser et al. (1987) suggest that the base population breeding values be treated as fixed, rather than random, effects. The motivation for this suggestion is that if parents are selected, they are not a random sample from the base population. Since REML estimates are unbiased by fixed-effects, any bias in the variance of the initial sample is ignored by treating the original parental breeding values as fixed. Simulation studies, however, show that even if initial bias is reduced by treating the parents as fixed, selection on the resulting offspring (or future generations) introduces additional bias (van der Werf 1990). Despite this reservation, we briefly review the approach here as parents whose records are missing are also often treated as fixed, which requires some modifications of the mixed-model equations. Let $\mathbf{a_b}$ be the vector of breeding values for the base population and $\mathbf{a_r}$ breeding values of the remaining individuals that descent from the base population. Following Graser et al. (1987), we can express the dependence of $\mathbf{a_r}$ on the base population breeding values $\mathbf{a_b}$ as follows

$$\begin{pmatrix} \mathbf{a_b} \\ \mathbf{a_r} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_1 & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{a_b} \\ \mathbf{a_r} \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{s} \end{pmatrix} \tag{15.22}$$

where $\mathbf{s}$ is the vector of segregational residuals (Equation 15.15) and $\mathbf{P}_1$ and $\mathbf{P}_2$ are matrices with values of $1/2$ in the parents' column in each row. Here, $\mathbf{a_r}$ is a random effect because it is a function of a fixed effect ($\mathbf{a_b}$) and a random effect ($\mathbf{s}$). The resulting mixed-model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a_b} + \mathbf{Z}_2\mathbf{a_r} + \mathbf{e} \tag{15.23}$$

Graser et al. show that the resulting mixed-model equations can be written as

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_1 & \mathbf{X}^T\mathbf{Z}_2 \\ \mathbf{Z}_1^T\mathbf{X} & \mathbf{Z}_1^T\mathbf{Z}_1 + \lambda\mathbf{Q}^T\mathbf{G}^{-1}\mathbf{Q} & -\lambda\mathbf{Q}^T\mathbf{G}^{-1} \\ \mathbf{Z}_2^T\mathbf{X} & -\lambda\mathbf{G}^{-1}\mathbf{Q} & \mathbf{Z}_2^T\mathbf{Z}_2 + \lambda\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}}_\mathbf{b} \\ \widehat{\mathbf{a}}_\mathbf{r} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_1^T\mathbf{y} \\ \mathbf{Z}_2^T\mathbf{y} \end{pmatrix} \tag{15.24a}$$

where $\lambda = (2\sigma_e^2/\sigma_A^2)$ and

$$\mathbf{Q} = (\mathbf{I} - \mathbf{P}_2)^{-1}\mathbf{P}_1 \quad \text{and} \quad \mathbf{G} = (\mathbf{I} - \mathbf{P}_2)^{-1}\mathbf{F}\left[(\mathbf{I} - \mathbf{P}_2)^{-1}\right]^T \tag{15.24b}$$

and the elements of the diagonal matrix $\mathbf{F}$ are given by Equation 15.14.

Dominance

Up to this point, we have been assuming all genetic variation is additive, so that we need only consider the vector $\mathbf{a}$ of breeding values and the numerator relationship matrix $\mathbf{A}$. When nonadditive genetic variance is present, it creates additional genetic correlations between relatives beyond those accounted for $\mathbf{A}$. The simplest setting is when dominance is present. As we have seen, the covariance among (noninbred) full sibs is inflated by $\sigma_D^2/4$. However, sibs can also have their covariance inflated because of common environmental effects. Separating the contribution of dominance from common family environment is nontrivial and requires specific types of links in the pedigree. If the goal is simply to reduce the bias in estimation of breeding values when dominance is potentially present, an animal model adding an additional random factor for common sib effects (e.g., Example 15.6) will often be satisfactory. This model simply estimates the common sib variance σ_c^2, which may include contributions from both dominance and shared family environments.

While dominance inflates the covariance between full sibs, it can also do so for other sets of relatives as well. To treat this general problem, letting the vector $\mathbf{d}$ denote the dominance effects, the mixed model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{d} + \mathbf{e} \tag{15.25a}$$

The overall genetic merit of an individual is estimated by $\widehat{\mathbf{g}} = \widehat{\mathbf{a}} + \widehat{\mathbf{d}}$. Turning to the covariance structure of this model, as before $\mathbf{Var}(\mathbf{a}) = \sigma_A^2\,\mathbf{A}$ and $\mathbf{Var}(\mathbf{e}) = \sigma_E^2\,\mathbf{I}$, while the covariance matrix for dominance effects is $\mathbf{Var}(\mathbf{d}) = \sigma_D^2\,\mathbf{D}$, giving

$$\mathbf{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}^T\sigma_A^2 + \mathbf{Z}\mathbf{D}\mathbf{Z}^T\sigma_D^2 + \mathbf{I}\,\sigma_e^2 \tag{15.25b}$$

Equation 15.25b shows the error in the residual variance structure if one (incorrectly) assumes dominance is absent, where we would (incorrectly) use $\sigma_e^2\,\mathbf{I}$ instead of $\mathbf{Z}\mathbf{D}\mathbf{Z}^T\sigma_D^2 + \mathbf{I}\,\sigma_e^2$.

The elements of the **dominance genetic relationship matrix D** are obtained as follows. The covariance between dominance effects for (non-inbred) individuals *i* and *j* is the product of the dominance genetic variance and the coefficient of fraternity, $\sigma_D^2\Delta_{ij}$. From LW Equation 7.7, the latter is given by

$$\Delta_{ij} = \Theta_{gk}\,\Theta_{hl} + \Theta_{gl}\,\Theta_{hk} \tag{15.26a}$$

where *i*'s parents are indexed by *g* and *h* and *j*'s by *k* and *l*, and (as above), Θ is the coefficient of coancestry. Recalling that the elements of the numerator relationship matrix $\mathbf{A}$ are $2\Theta_{ij}$, the off-diagonal elements of $\mathbf{D}$ can be computed from the elements of $\mathbf{A}$ by

$$D_{ij} = \frac{A_{gk}\,A_{hl} + A_{gl}\,A_{hk}}{4} \tag{15.26b}$$

whereas the diagonal elements are all $D_{ii} = 1$. Note that $\mathbf{D}$ is expected to be considerably more **sparse** (most off-diagonal elements zero) than $\mathbf{A}$, and hence may not contribute information for most individuals.

The mixed-model equations for Equation 15.25a become

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{Z} + \lambda_D\mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{d}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix} \tag{15.27}$$

where $\lambda_A = \sigma_e^2/\sigma_A^2$, and $\lambda_D = \sigma_e^2/\sigma_D^2$. Hoeschele and Van Raden (1991) present a quick method for computing $\mathbf{D}^{-1}$ for a non-inbred population.

Epistatic terms can be included in the mixed model equations in a similar fashion, see LW Chapter 26 for details and LW Chapter 27 for modifications of the REML equations to estimate nonadditive variances.

While the above treatment suggests that dominance is easily incorporated, this is misleading, as we have made the tact assumption of no selection and (much more importantly) no inbreeding. Assuming the infinitesimal model, selection does not significantly change the dominance variance (Chapter 13, Bulmer 1971b) and hence the estimates of $\mathbf{a}$ and $\mathbf{d}$ are not likely to be biased by selection. Inbreeding (which occurs in all selection experiments), on the other hand, introduces major complications. First, there may be inbreeding depression. In some situations, this can be dealt with by including inbreeding depression as a covariate, for example by using a model such as

$$y_{ti} = \mu + If(t) + a_{ti} + d_{ti} + e_{ti} \tag{15.28}$$

where $f(t)$ is the inbreeding in generation t and I is the inbreeding depression under complete inbreeding. Recall (LW Chapter 10) that with only dominance, inbreeding depression is a linear function of the inbreeding f, while with epistasis it is a nonlinear function of f. Thus, if there is significant epistasis, Equation 15.28 may not properly correct for inbreeding depression, especially at high values of f (those approaching one). One can also include a control population with known levels of inbreeding to provide an independent estimation of I. In a LS analysis, inbreeding depression is typically assumed to be corrected for by estimating response by subtracting off the mean of such a control population. This assumes that the control and selected lines have the same level of inbreeding, when in fact the selected population is typically far more inbred (Chapter 26). If the levels of inbreeding are very similar in both populations, the use of a control can account for nonlinear inbreeding depression.

A second, and far more serious, complication is that the covariance between inbred relatives with dominance is no longer a function of just σ_A^2 and σ_D^2. As discussed in Chapter 3, these covariances depend upon four other quadratic components as well (σ_{DI}^2, σ_{ADI}, $\imath^*$, $\imath^2 - \imath^*$), see Equations 3.14-3.15. Recall that these additional components fully account for inbreeding depression. While one could formulate a mixed model incorporating all six quadratic components (using the covariances given by Equation 3.14-3.16), the resulting model is extremely complex and numerically very demanding. A start in this direction has been developed by Smith and Mäki-Tanila (1990), who should be consulted for more details. A second approach is to use Equation 15.27, with a $\mathbf{D}$ matrix that approximates the elements under inbreeding (Smith and Mäki-Tanila 1990). One could also combine the use of a modified $\mathbf{D}$ with a covariate for inbreeding depression (Equation 15.28), but this is still a largely ad-hoc approach to a complex problem.

Some guidance is potentially offered by simulations by de Boer and van Arendonk (1992), who examined the consequences of ignoring a cofactor inbreeding depression and the full covariance structure under inbreeding. When a standard dominance model (no accounting for inbreeding) was used in populations with inbreeding, estimates of both additive and dominance effects were biased. However, when a simple cofactor for inbreeding depression was included, but the full covariance structure under inbreeding ignored, effects estimates were unbiased, at least up to the level of inbreeding used in the simulations ($f = 0.35$). Thus, with low to modest levels of inbreeding, simple inclusion of a factor for inbreeding depression included in the model given by Equation 15.25 may be relatively robust.

The situation with epistasis is even more complex than with simple dominance. The good news is that the weighting of the nonadditive variance component terms for the covariances distant relatives is very small (e.g., LW Equation 7.12), so that even if nonadditive components are significant, their *actual* contribution to the covariance of most relatives, especially those separated by more than one generation, are very minor.

BAYESIAN MIXED MODEL ANALYSIS

As mentioned throughout this chapter, a standard mixed-model analysis does not fully account for the uncertainty introduced into estimates of the breeding values by using estimates of the variances (instead of their true values). While there are large-sample approximations for the sample variance of a REML variance estimator, it is never fully clear what constitutes "large". Further, quantities of interest, such as the heritability, are often functions of the estimated quantities. The sample variance and sampling distribution of such functions are very complicated (if they are obtainable at all). **Bayesian approaches** offer solutions to both these issues. While Bayesian statistics (as opposed to more standard, or **frequentist**, statistics) is often touted for its ability to incorporate prior information, we view their main utility as providing a complete description of the uncertainty of an estimate.

Frequentists assume the true value of a parameter is (typically) a constant, and the samples are variable. Statistics (such as confidence intervals) are computed by conceptually drawing an infinite number of samples, in which case (for example), the true value is in 95% of all such constructed confidence intervals. In contrast, a bayesian assumes that the sample is fixed, while the parameter is random. Their interest is in how that data changes the probability distribution for possible locations of the parameter. Thus, the term "Bayesian mixed models" is (formally) inappropriate, as *all* terms in a Bayesian analysis are assumed to be random and hence never "mixed". However, we use this term to emphasize that much of the basic foundations of a mixed-model analysis of selection experiments remain unchanged (such as the model formulation). What does change is how we analyze such data.

Appendix 2 introduces some of the basic ideas in a Bayesian analysis (beyond our short introduction here). Computational issues are extremely important, and covered in Appendix 3. Indeed, the recent explosion in the application of Bayesian approaches largely follows from relatively new computational approaches (such as MCMC methods) that allow for very complex distributions to be handled through straight-forward (but computationally intensive) procedures.

Introduction to Bayesian Statistics

While very deep (and very subtle) differences in philosophy separate hard-core Bayesians from hard-core frequentists (Efron 1986, Glymour 1981), our treatment here of Bayesian methods is motivated simply by their use as a powerful statistical tool. Their introduction

into quantitative genetics can be largely credited to the influential paper of Gianola and Fredando (1986), which reviews Bayesian appilications to animal breeding. Blasco (2001) provides a very nice overview of Bayesian vs. frequentist approaches in quantitative genetics and is highly recommended, while a very complete treatment of applications to quantitative genetics is provided by Sorensen and Gianola (2002).

The foundation of Bayesian Statistics is **Bayes' theorem** (Appendix 2; LW Chapter 13). The continuous, vector-valued version of this theorem is

$$p(\boldsymbol{\Theta} \,|\, \mathbf{y}) = \frac{p(\mathbf{y} \,|\, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{\int p(\mathbf{y}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta}} = \frac{p(\mathbf{y} \,|\, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{y})} \tag{15.29a}$$

where $\boldsymbol{\Theta}^T = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(k)})$ is a vector of k (potentially) continuous variables. Here $p(\boldsymbol{\Theta})$ is our prior belief (**prior** for short) about the distribution of the unknown values $\boldsymbol{\Theta}$, while $p(\mathbf{y} \,|\, \boldsymbol{\Theta})$ is just a standard likelihood function for the probability of the observed vector of data $\mathbf{y}$ given that the unknown parameters have specified value $\boldsymbol{\Theta}$. The product of these two, normalized by $p(\mathbf{y})$ to form a proper probability distribution, is our posterior belief (**posterior**) for the distribution of the unknown parameters given both the data $\mathbf{y}$ and the prior information/brief $p(\boldsymbol{\Theta})$. Since $p(\mathbf{y})$, the probablity of the data vector $\mathbf{y}$ is a constant, it is typically ignored, and the posterior is often written as

$$p(\boldsymbol{\Theta} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta}) \tag{15.29b}$$

Example 15.7. As an example of a Bayesian analysis, consider the simple case of n observations from a normal with unknown mean μ and known variance σ^2. The details for this analysis (and more realistic cases, such as both the mean and variance unknown) are fully detailed in Appendix 2. Assuming the data $\mathbf{y} = (y_1, \cdots y_n)^T$ are independent, the resulting likelihood function (which corresponds to $p(\mathbf{y} \,|\, \mu)$ is

$$p(\mathbf{y} \,|\, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \right)$$

Suppose we assume a Gaussian prior for the location of the mean, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, so that

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

The mean and variance of the prior, μ_0 and σ_0^2, are referred to as **hyperparameters**. Here, μ_0 specifies a prior location for the mean, while σ^2 specifies our uncertainty in this prior location – the larger the variance, the greater our uncertainty. In the limit (as $\sigma_o^2 \to \infty$) this corresponds to a $p(\mu) = c$, a constant, which is **uniform** or **flat** or **improper** prior, where all values of μ are assumed to be (a priori) equally likely. (Keyner 1921 called this is *principle of indifference* — all possible events are equally probable.) A little algebra (Appendix 2) yields

$$p(\mu \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \mu)\, p(\mu) = \exp\left(-\frac{(\mu - \mu_*)^2}{2\sigma_*^2} \right)$$

where the expressions for μ_* and σ_*^2 are given in Appendix 2. Thus, the posterior density function for μ is a normal with mean μ_* and variance σ_*^2, e.g.,

$$\mu \,|\, (\mathbf{y}, \sigma^2) \sim \mathrm{N}\left(\mu_*, \sigma_*^2\right)$$

Here, the prior **conjugated** with the likelihood function – the product of the prior and likelihood returned a distribution in the same family as the prior. The use of such **conjugate priors** (for a given likelihood) is a key concept in Bayesian analysis and we explore it in detail in Appendix 2. For example, with normally-distributed data and an unknown variance, using a **scaled inverse chi-square** (χ^2) prior for the variance also conjugates the likelihood, with the posterior distribution for the variance also following a scaled inverse χ^2 distribution (Appendix 2). Further note that while μ is normally a fixed effect in a standard (frequentist) analysis, in a Bayesian analysis it is a *random* effect. There are no fixed effects in a Bayesian analysis, as everything is treated as random, namely being drawn from a distribution.

A Bayesian analysis returns a *distribution*, rather than a **point estimate**. Thus, often several summary statistics are reported for the posterior such as its mean, medium (50% value) and mode (maximal value). Note that the mode is what a likelihood analysis returns as the MLE. More generally, when nice analytic expressions (like ours above) are not available, one can simply plot the distribution via a histogram from values generated through MCMC methods (as discussed below).

What is relative importance of the prior information versus the actual data? The mean of the posterior distribution is given by

$$\mu_* = \mu_0 \frac{\sigma_*^2}{\sigma_0^2} + \overline{x} \frac{\sigma_*^2}{\sigma^2/n}$$

With a very diffuse prior on μ (i.e., $\sigma_0^2 >> \sigma^2$), $\sigma_*^2 \to \sigma^2/n$ and $\mu_* \to \overline{x}$. Likewise, as we collect enough data (i.e., large n), $\sigma_*^2 \to \sigma^2/n$ and again $\mu_* \to \overline{x}$. Thus, with very weak prior information, the mean of the posterior distribution is close to the sample mean. Likewise, even with a very strong prior belief about the location of the mean, as our sample size gets very large, the mean of the posterior is still close to the sample mean.

The dependence of the posterior on the prior (which can easily be assessed by trying different priors) provides an indication of how much information on the unknown parameter values is contained in the data. If the posterior is highly dependent on the prior, then the data likely has little signal, while if the posterior is largely unaffected by the shape of the assumed prior, the data are highly informative. Such explorations of the effects of a prior under a careful Bayesian analysis offers some protection from incorrect conclusions based on weak likelihoods.

Example 15.8. Sorensen et al (1994) examined the effects of different priors for the variance components in a simulation study using the analysis model. The data were analyzed in two sets: the entire data set (ALL) and a partial subset (PART). The simulated heritability was 0.5. Both uniform and scaled inverse χ^2 priors for the additive and residual variance were used. The uniform spreads belief evenly over all possible values, while the inverse χ^2 places more weight on specific values. The mean and variance of the marginal posterior distribution for h^2 in these four cases were as follows:

Data Set	Prior	$E[h^2]$	$\sigma^2[h^2]$
PART	Uniform	0.737	2.26×10^{-2}
PART	Inverse χ^2	0.501	0.29×10^{-2}
ALL	Uniform	0.550	1.63×10^{-2}
ALL	Inverse χ^2	0.529	0.24×10^{-2}

Note the disparity of the estimates under the two priors in the partial data case, and their agreement in the full data case. In the partial data case, the effect of the prior had a strong influence, indicating a weak signal (likelihood) for h^2 in this particular data set. With the full data set, the signal greatly increases, mitigating the effects of the prior. Note (as expected) for both priors that the posterior variance decreased under the larger sample size (c.f., PART versus ALL). Also note that the posterior variance was smaller under the assumed inverse χ^2 priors. Thus, the choice of a prior not only influences the mean of the estimate, but also its variance as well. In this case, while the different priors gave the same mean heritability in the full data set, their variances differed by an order of magnitude.

This is an example of a **sensitivity analysis** using different priors to look at the stability of the posterior. With complex posteriors, one can observe broad stability for many of the variables (insensitivity to changes in the priors), but extreme dependence in the others. Thus, the use of different priors provides one means to explore the amount of signal along the different directions (variables) of the likelihood function.

Often, only a subset of the unknown parameters are of concern the us, the rest being **nuisance parameters** that we wish to remove (or at least ignore). Write the vector of unknown parameters as $\boldsymbol{\Theta}^T = (\boldsymbol{\Theta}_1^T, \boldsymbol{\Theta}_n^T)$, where $\boldsymbol{\Theta}_n$ is the vector of nuisance parameters. We would like to condition out the dependence of the parameters of interest from the effects of the nuisance parameters, generating a **marginal posterior distribution** that does not depend on the nuisance parameters. Integrating over $\boldsymbol{\Theta}_n$ gives the desired marginal as

$$\begin{aligned} p(\boldsymbol{\Theta}_1 \,|\, \mathbf{y}) &= \int p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_n \,|\, \mathbf{y})\, d\boldsymbol{\Theta}_n \\ &= \int p(\boldsymbol{\Theta}_1 \,|\, \boldsymbol{\Theta}_n, \mathbf{y})\, p(\boldsymbol{\Theta}_n \,|\, \mathbf{y}) d\boldsymbol{\Theta}_n \\ &= E_{\boldsymbol{\Theta}_n}\left[\, p(\boldsymbol{\Theta}_1 \,|\, \boldsymbol{\Theta}_n, \mathbf{y})\,\right] \end{aligned} \tag{15.30}$$

This marginal probability calculation illustrates both the strength, and weakness, of a Bayesian analysis. The strength is that obtaining such a marginal is very powerful for inference, the weakness is that the integration to obtain this marginal can be horrendous (at best!)

Example 15.9. In the context of analyzing a selection response experiment, the vector of breeding values $\mathbf{a}$ is of interest, while the q fixed effects ($\boldsymbol{\beta}$) and variances (σ_A^2, σ_e^2) are often regarded as nuisance parameters. In this case, Equation 15.30 gives the marginal distribution of the breeding values $\mathbf{a}$ given the data as

$$p(\mathbf{a} \,|\, \mathbf{y}) = \int p(\mathbf{a}, \boldsymbol{\beta}, \sigma_A^2, \sigma_e^2 \,|\, \mathbf{y})\, d\boldsymbol{\beta}\; d\sigma_A^2\; d\sigma_e^2$$

The integration is over the $q + 2$ dimensional space given by the q possible values for the $\boldsymbol{\beta}$ and the two variances. This conditioning removes any dependencies of estimates of the response on estimates of the variance components (and fixed effects). Uncertainties introduced by estimating these nuisance parameters are automatically accommodated when considering the marginal distribution. While this multidimensional integral is complex, the Gibbs sampler (below) can be use to approximate drawings from this marginal distribution.

With the marginal density $p(\mathbf{a} \mid \mathbf{y})$ in hand, one can obtain estimates of the response to selection $\mathbf{K}^T\mathbf{a}$ that are *independent* of the assumed (or estimated) additive variance σ_A^2. The error due to estimation of the additive variance from the data is directly incorporated when the marginal is computed as we integrate over possible values of σ_A^2 and their support given the data. This independence of the estimate of response from the estimate of additive variance and the subsequent incorporation of the error in estimating σ_A^2 in the estimate of the response are two very compelling reasons for a Bayesian analysis of response.

This example hints at a key feature noted by Gianola and Fernando (1986). If the data on which selections was based are included in the analysis, then (by integrating over all nuisance parameters) the Bayesian approach accounts for any potential bias by selection (*provided* that the model assumptions, such as multivariate normality, hold).

Computiting Posteriors and Marginals: MCMC and The Gibbs Sampler

Historically, the widespread implementation of Bayesian approaches was limited by the difficulty in obtaining posterior distributions, which typically requires the integration of high dimensional functions (e.g., Equations 15.29 and 15.30). **Markov Chain Monte Carlo** (**MCMC**) approaches (Appendix 3) provide a solution to this problem, offering a straightforward (although computationally demanding) procedure for generating random draws from very complex distributions.

Simulating random vectors drawn from some complex **target distribution** can be a very difficult task. The idea behind MCMC approaches is to successively draw samples from far simpler distributions in such a way that the distribution of the samples converges to the target distribution. MCMC approaches are so-named because one uses the previous sample value to randomly generate the next sample value, generating a **Markov chain** (Appendix 3). While there are a wide range of MCMC methods, two of the most commonly encountered in the quantitative-genetics literature are the **Metropolis-Hastings algorithm** (Metropolis and Ulam 1949, Metropolis et al. 1953, Hastings 1970), and the **Gibbs sampler** (Geman and Geman 1984). Under Metropolis-Hastings, one simulates draws from a complex target distribution by first drawing a random variable from a specified (and simpler) distribution and then using a second probability distribution to decide whether to keep that realization or reject it (details in Appendix 2). The strength of Metropolis-Hastings is that is can be applied to a very wide range of problems, such as priors that do not conjugate (and hence do not have a simple form). Its weakness is that the candidate value can end up being rejected with very high probability, making the sampler very inefficient (requiring very long runs to produce a reasonably-size trimmed sequence with low correlation between elements), especially when one is dealing with vectors of random variables. The Gibbs sampler is a special case of Metropolis-Hastings sampling wherein the random value is always accepted. The key to the Gibbs sampler is that one only considers univariate conditional distributions — the distribution when all of the random variables but one are assigned fixed values. Typically, one uses conjugate priors to form a Gibbs sampler (see below).

To introduce the Gibbs sampler, consider a bivariate random variable (x, y), and suppose we wish to compute one or both marginals, $p(x)$ and $p(y)$. The idea behind the sampler is that it is far easier to consider a sequence of conditional distributions, $p(x \mid y)$ and $p(y \mid x)$, than it is to obtain the marginal by integration of the joint density $p(x, y)$, e.g., $p(x) = \int p(x, y)dy$. The sampler starts with some initial value y_0 for y and obtains x_0 by generating a random variable from the conditional distribution $p(x \mid y = y_0)$. The sampler then uses x_0 to generate a new value of y_1, drawing from the conditional distribution based

on the value x_0, $p(y \mid x = x_0)$. The sampler proceeds as follows

$$x_i \sim p(x \mid y = y_{i-1}) \tag{15.31a}$$

$$y_i \sim p(y \mid x = x_i) \tag{15.31b}$$

Repeating this process k times, generates a **Gibbs sequence** of length k, where a subset of points (x_j, y_j) for $1 \leq j \leq m < k$ are taken as our simulated draws from the full joint distribution. To obtain the desired total of m sample points, one samples the chain (i) after a sufficient **burn-in** to removal the effects of the initial starting values and (ii) at set time points (say every n samples) following the burn-in (**trimming** the sequence). For example, Wang et al (1994) in their Gibbs sampler for an animal model generated a total of 1,205,000 sample vectors. The first 5,000 were discarded (corresponding to the burn-in), and then every tenth subsequent iteration was saved (to reduce correlations between sample vectors) to yield a total sampler of 120,000 vectors. The burn-in period, and sampling interval following the burn-in can be delicate, and careful analysis of the resulting sequence using conversion diagnostic tools is critical (Appendix 3).

When more than two variables are involved, the sampler is extended in the obvious fashion. For example, if there are four variables, (w, x, y, z), the sampler becomes

$$\begin{aligned} w_i &\sim p(w \mid x = x_{i-1}, y = y_{i-1}, z = z_{i-1}) \\ x_i &\sim p(x \mid w = w_i, y = y_{i-1}, z = z_{i-1}) \\ y_i &\sim p(y \mid w = w_i, x = x_i, z = z_{i-1}) \\ z_i &\sim p(z \mid w = w_i, x = x_i, y = y_i) \end{aligned}$$

Any feature of interest for the marginals can be computed from the m realizations of the Gibbs sequence. For example, the expectation of any function f of the random variable x is approximated by

$$E[f(x)]_m = \frac{1}{m} \sum_{i=1}^{m} f(x_i) \tag{15.32a}$$

which is simply the average of the function evaluated over the points in the sampler. This is the **Monte-Carlo** (MC) **estimate** of $f(x)$, as $E[f(x)]_m \to E[f(x)]$ as $m \to \infty$. Likewise, the MC estimate for any function of n variables $(\theta^{(1)}, \cdots, \theta^{(n)})$ is given by

$$E[f(\theta^{(1)}, \cdots, \theta^{(n)})]_m = \frac{1}{m} \sum_{i=1}^{m} f(\theta_i^{(1)}, \cdots, \theta_i^{(n)}) \tag{15.32b}$$

Example 15.10. Suppose we are interested in the distribution of breeding values in a particular generation (measured by four individuals in the analysis), as well as in the base population heritability. A Gibbs sampler has been generated and the realizations at three different iterations (say 145, 365, and 679) after a sufficiently burn-in period are as follows:

factor	Sample 145	Sample 365	Sample 679
$a(1)$	1.5	1.8	2.2
$a(2)$	2.1	3.4	1.4
$a(3)$	3.1	2.9	4.4
$a(4)$	3.3	4.3	3.6
σ_A^2	0.55	0.64	0.46
σ_e^2	1.10	0.98	1.20

Here $a(1)$ through $a(4)$ correspond to the value of the four estimated breeding values for our generation of interest in that particular iteration of the sampler, and σ_A^2 and σ_e^2 are similarly the realizations for the variances in that iteration. Using these realizations, the values of the mean breeding value

$$\overline{a}_i = \frac{1}{4}\sum_{j=1}^{4} a(j)_i,$$

the variance in breeding values

$$\text{Var}(a)_i = \frac{1}{4}\sum_{j=1}^{4} \left[\, a(j)_i - \overline{a}_i \,\right]^2,$$

and the base-population heritability

$$h_i^2 = \sigma_{A,i}^2/(\sigma_{A,i}^2 + \sigma_{e,i}^2)$$

for these three realizations are simply

	Sample 145	Sample 365	Sample 679
$\overline{a}$	2.5	3.1	2.9
$\text{Var}(a)$	0.54	0.82	1.37
h^2	0.33	0.40	0.28

Thus, the sampler has returned three values for each of the quantities of interest. Of course, a full sampler consist of thousands to tens of thousands of such realization, allowing us to empirically generate the full distribution of any of these functions.

Density Estimation Under MCMC

Suppose $x_1, \cdots, x_m$ represent the sequence of values from a (burned-in and trimmed) Gibbs sampler generated for variable of interest x. As shown above (Equation 15.32, Example 15.10), one could directly use this sequence to compute any summary statistic of interest for the marginal posterior of x (such as a mean, variance, or skewness). In essence, we are simply estimating the conditional distributions through a histogram, assigning a probability mass of $1/m$ to each value x_i. Given a long enough sequence, this simple approach works just fine.

An alternative approach (to provide for both better smoothing between points and treatment of extreme values) is to is to replace the point values with functions. Two methods commonly appear in the quantitative-genetics literature. The first is the **Gaussian Kernel estimator** (Silverman 1986), taking the marginal density for x as the weighted sum of normals,

$$p(x) = \frac{1}{m}\sum_{i=1}^{m} \varphi(x, x_i, w^2), \quad \text{where} \quad \varphi(x, x_i, w^2) = \frac{1}{\sqrt{2\pi\, w^2}} \exp\left(\frac{(x - x_i)^2}{2w^2}\right) \qquad (15.33a)$$

here w, the **window width**, is a user-defined smoothing constant. (For ease of presentation, we have denote the marginal posterior for x as $p(x)$, while formally it should be $p(x \mid \mathbf{y}, \Theta_p)$, as it is conditional on the data $\mathbf{y}$ and the hyperparameter values Θ_p chosen for the priors.) The gaussian kernel replaces putting all of the probability mass at the value x_i with *centering* the probability mass at x_i, but spreading out the possible values (w serving as the variance about this value – the smaller w, the more the mass is concentrated tightly around x_i).

An alternative approach exploits that fact that, with a Gibbs sampler, we already have the univariate conditional distributions, and thus it should be more powerful to simply average over these conditionals,

$$p(x) = \frac{1}{m}\sum_{i=1}^{m} p(x \mid \Theta = \Theta_i) \tag{15.33b}$$

where now Θ_i are the values for the other variables in the sampler at (trimmed) iteration i. Note the distinction in the values from the sampler used by Equation 15.33a versus 15.33b. In the Gaussian kernel estimator the specific realizations $x_1, \cdots, x_m$ for the focal variable are used, while in the average conditional estimator uses (by conditioning) the values of the sampler *excluding* the value of the focal variable x_i.

Example 15.11. Suppose we wish to use a Gibbs sequence to estimate the marginal posterior estimate for the additive genetic variance, σ_A^2 in an animal model. Suppose our (suitably burned-in and trimmed) Gibbs sequence is of length m, and let $x_1, \cdots, x_m$ denote the realizations of σ_A^2 in this sequence (suppose the first two are 3.2 and 4.5). Then the Gaussian kernel estimator of the marginal posterior for the additive variance is simply

$$\begin{aligned} p\left(\sigma_A^2\right) &= \frac{1}{m}\sum_{i=1}^{m} \varphi(\sigma_A^2, x_i, w^2) \\ &= \frac{1}{m}\Big(\varphi(\sigma_A^2, 3.2, w^2) + \varphi(\sigma_A^2, 4.5, w^2) + \cdots + \varphi(\sigma_A^2, x_m, w^2)\Big) \end{aligned}$$

where the smoothing parameter w is set by the user. Using this marginal, means, variances, and support intervals are found in the standard way via integration. For example, to finds an interval (a, b) such that 95% of the probability mass for σ_A^2 in the marginal is within this region, we simply solve

$$\int_a^b p\left(\sigma_A^2\right)\, d\sigma_A^2 = 0.95$$

As an aside, note that there are infinitely many such intervals, and typically one takes the smallest such interval as our 95% Bayesian credible interval.

Alternatively, if we have the univariate conditional distribution for σ_A^2 (which we have from constructing the Gibbs sampler), then we can use Equation 15.33b to take the average of these to estimate the marginal posterior. For a standard animal model , Wang et al (1994a) show that the univariate distribution for the additive variance σ_A^2, conditional on the values for the fixed effects, breeding values, and environmental variance, follows an scaled inverse χ^2 distribution, with the parameters of the posterior a function of the assumed prior hyperparameters and the value of $\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}$. Thus, as the realization of $\mathbf{a}$ changes in different iterations of the sampler, so does the distribution parameters. The posterior distribution has $\nu + p$ degrees of freedom and scaling parameter

$$S = \frac{\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a} + vS_o^2}{v + p}$$

where ν and S_o are the hyperparameters of the prior and q the number of breeding values to estimate. The resulting estimate of the marginal for σ_A^2 thus becomes

$$p\left(\sigma_A^2\right) = \frac{1}{m}\sum_{i=1}^{m} \phi(\sigma_A^2, \nu + q, S_i)$$

where $\phi(x, n, a)$ denotes the density for a random variable x distributed as an inverse χ^2 with n degrees of freedom and scaling factor a (see Appendix 2 for details). Here S_i is obtained by using the realizations of the vector $\mathbf{a}$ for the i value from the gibbs sequence. Note that, unlike the Gaussian kernel estimator, we *do not* use the sampler values for σ^2_A. Rather, we use the sampler values for the *other* variables (in this case, the vector of breeding values) when constructing the marginal posterior.

Bayesian Analysis of the Animal Model

The use of Bayesian approaches for the analysis of selection experiments was first suggested by Sorensen and Johansson (1992). Starting with the standard animal model

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

Wang et al. (1993, 1994a,b), Sorensen et al. (1994) and Jensen et a.l (1994) developed Gibbs-samplers for this model and its extensions. As before, the conditional distribution of the data given the vectors of fixed effects ($\boldsymbol{\beta}$), breeding values ($\mathbf{a}$) and the environmental variance is multivariate normal,

$$\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{a}, \sigma^2_e \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma^2_e) \tag{15.34a}$$

The infinitesimal model is assumed, with the distribution of breeding values given the relationship matrix $\mathbf{A}$ and additive genetic variance σ^2_A is also multivariate normal,

$$\mathbf{a} \mid \mathbf{A}, \sigma^2_A \sim \text{MVN}(\mathbf{0}, \mathbf{A}\sigma^2_A) \tag{15.34b}$$

Sorensen et al. (1994) assumed a uniform prior for $\boldsymbol{\beta}$, a normal prior for $\mathbf{a}$, and both uniform and inverse χ^2 priors for the variances. For example, the prior for $\mathbf{a}$ is the product of $p(\mathbf{a} \mid \sigma^2_A, \mathbf{A}) \cdot p(\sigma^2_A)$, where the first distribution is a multivariate normal (Equation 15.34c) and the second either a uniform or an inverse χ^2. These choices are conjugate priors given the multivariate normal, and the result is that analytic expressions can be obtained for the $p+q+2$ univariate conditional distributions for each factor in the model (p breeding values a_i, q fixed effects β_j, and the variances σ^2_A and σ^2_e). Using these univariate conditionals, the Gibbs sampler can be constructed. The outline for the sampler is as follows:

1. Set initial values for $\mathbf{a}$, $\boldsymbol{\beta}$, σ^2_A, σ^2_e.

2. Using the current values of $\mathbf{a}$, $\boldsymbol{\beta}$, σ^2_A σ^2_e and the conditional distributions (see Sorensen et al. for exact expressions) :

(i) update the fixed effects by sequentially drawing (for $j = 1, \cdots, q$) from the conditionals (which are univariate normals)

$$\beta_{j,i} \sim p(\beta_j \mid \beta_{1,i}, \cdot\cdot, \beta_{j-1,i}, \beta_{j+1,i-1}, \cdot\cdot, \beta_{q,i-1}, \mathbf{a}_{i-1}, \sigma^2_{A,i-1}, \sigma^2_{e,i-1})$$

where $\beta_{j,i}$ is the value of β_j during the ith iteration of the sample. Hence, for factor j, we take the values for $\mathbf{a}$ and the variances from the last iteration ($i-1$), the values of β_1 to β_{j-1} from the current iteration (i), and the values of β_{j+1} to β_q from the last iteration ($i-1$). These values are inserted to give the parameters (here, the conditional mean and variance) for the univariate normal that corresponds to the conditional distribution for β_j and a random variable is drawn from this distribution to give $\beta_{j,i}$

(ii) Update the breeding values by sequentially drawing (for $i = 1, \cdots, p$) from the conditionals (again univariate normals)

$$a_{j,i} \sim p(a_j \mid \boldsymbol{\beta}_i, a_{1,i}, \cdot\cdot, a_{j-1,i}, a_{j+1,i-1}, \cdot\cdot, a_{p,i-1}, \sigma^2_{A,i-1}, \sigma^2_{e,i-1})$$

(iii) Update the additive variance, drawing from the conditional (a scaled inverse χ^2 distribution)

$$\sigma^2_{A,i} \sim p(\sigma^2_A \,|\, \boldsymbol{\beta}_i, \mathbf{a}_i, \sigma^2_{e,i-1})$$

(iv) Update the error variance, drawing from the conditional (again, a scaled inverse χ^2 distribution)

$$\sigma^2_{e,i} \sim p(\sigma^2_e \,|\, \boldsymbol{\beta}_i, \mathbf{a}_i, \sigma^2_{A,i})$$

3. Using the updated values, repeat **(2)** until k samples are obtained, from which m are extracted (followng the burn-in and trimming) for the Gibbs sampler chain.

Note that the Bayesian analysis of the animal model makes most of the standard animal-model assumptions, in particular that the infinitesimal model holds (so that covariance matrix for breeding values is $\sigma^2_A \mathbf{A}$). Thus, both MM and Bayesian analysis are potentially biased by changes in allele frequencies. A Bayesian analysis has all the advantages of a MM analysis (over a LS analysis) and, in addition, the posterior marginals correctly give the distribution of any parameter of interest independent of the values assumed for other parameters. The uncertainly introduced by estimating these additional parameters is reflected in the posterior marginal. Thus, the Bayesian approach gives the correct distribution (assuming the model assumptions hold) for the estimated response independent of the additive genetic variance. By contrast, a MM analysis is highly dependent on the assumed (or estimated) additive variance, and the standard error of a REML/BLUP estimate for the response (Equation 15.8c) does not account for the variance in REML estimation of σ^2_A.

Application: Estimating Response in Pig Litter Size Components

Blasco et al. (1998) used the method of Sorensen et al. (1994) to estimate the response to selection on ovulation rate and prenatal survival in French Large White pigs. Three lines were followed, one selecting on each trait and a control line. The relevant selection and control lines were jointly analyzed to estimate response. Ovulation rate was examined using the standard animal model,

$$\mathbf{y} \,|\boldsymbol{\beta}, \mathbf{a}, \sigma^2_e \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}, \mathbf{I}\sigma^2_e), \quad \mathbf{a} \,|\, \mathbf{A}, \sigma^2_A \sim N(\mathbf{0}, \mathbf{A}\,\sigma^2_A)$$

Prenatal survival (as a function of the mother) was examined using the repeatability model,

$$\mathbf{y} \,|\boldsymbol{\beta}, \mathbf{a}, \mathbf{c}, \sigma^2_e \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{c}, \mathbf{I}\,\sigma^2_e)$$

$$\mathbf{a} \,|\, \mathbf{A}, \sigma^2_A \sim N(\mathbf{0}, \mathbf{A}\,\sigma^2_A), \quad \mathbf{c} \,|\, \sigma^2_c \sim N(\mathbf{0}, \mathbf{I}\,\sigma^2_c)$$

Among the fixed effects in $\boldsymbol{\beta}$ are terms for the **parity** of the mother (1st parity = 1st litter, 2nd parity = second litter and so on). The marginal posterior distribution for breeding values (and hence for the response via $\mathbf{K}^T\mathbf{a}$) was obtained by using the Gibbs sampler approach of Sorensen et al. (1994). For each trait, two independent chains of length 100,000 were computed, with the first 10,000 samples discarded (to remove burn-in effects) and sampling at every 30 iterations thereafter, generating a (trimmed) sampler of length 3000. The authors obtained these burn-in and resampling values after several initial runs and using the diagnostics suggested by Raftery and Lewis (1992) for level of precision and Geyer (1992) for autocorrelation between samples. A uniform prior was taken for the fixed effects, while different priors used for the variances (discussed below).

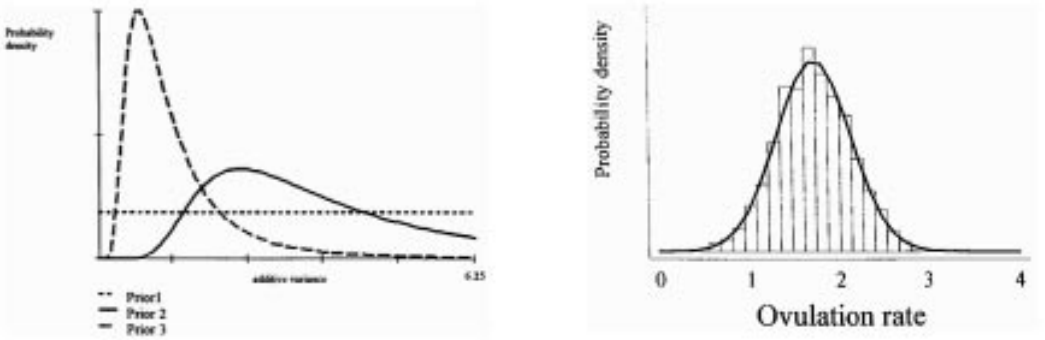


Figure 15.2. Analysis of ovulation rate at puberty in French Large White pigs. **Left:** Assumed priors for σ_A^2 (see text for details). **Right:** The Bayesian estimate of response is given by the posterior density for the mean breeding value in ovulation rate in the last generation of selection, presented as a histogram. This distribution is approximately normal (solid curve). After Blasco et al. (1998).

Consider the results for ovulation rate at puberty first. Figure 15.2 shows the three priors assumed for the additive variance in this trait based on prior information. The phenotypic variance of this trait is 6.25, setting an upper limit on σ_A^2. Prior one is a uniform distribution that weights all values in the parameter space equally. Priors two and three (scaled inverted χ^2 distributions) reflect additional information. Published heritabilities for this trait in pigs and rabbits ranges from 0.1 to 0.6, and prior two assumes a broad distribution around the approximate medium value ($\sigma_A^2 = 0.4 \cdot 6.25 = 2.5$). A study specifically in French Large Whites gave an estimate of $h^2 = 0.11 \pm 0.02$ and the tight distribution around this value is reflected in prior three. Using the approach for Sorensen et al. (1994), Blasco et al. obtain Monte-Carlo estimates of the (base population) heritability under these three priors of $h^2 = 0.39 \pm 0.07$, 0.39 ± 0.06, and 0.32 ± 0.06. Table 15.1 shows the estimated response during each of the four generations of selection, comparing these with the least squares (differences between generation means) and mixed model (REML/BLUP) estimates. Note that the three different priors give very consistent estimates of response, implying that the data contain sufficient information to overpower most of the signal coming from the assumed prior. The Bayesian and MM analysis give very similar results, while the LS results give a slightly different estimates of response.

Table 15.1 Estimated response to selection for ovulation rate at puberty and prenatal survival in French Large White pigs. Bayesian analysis with three different priors (Figure 15.2 for ovulation, Figure 15.3 for prenatal survival) were used to obtain Monte-Carlo estimates of the mean response and their associated standard deviations (the later incorporating the additional error from estimating the additive variance and other parameters). For comparison, least squares (LS) estimates ($\overline{z}_{i+1} - \overline{z}_i$) and mixed-model (REML/BLUP) estimates are also included. After Blasco et al. (1998).

Ovulation Rate at Puberty

Method	Gen 1	Gen 2	Gen 3	Gen 4
Bayesian, Prior 1	0.30 ± 0.31	0.51 ± 0.35	1.03 ± 0.39	1.58 ± 0.43
Bayesian, Prior 2	0.31 ± 0.30	0.51 ± 0.34	1.05 ± 0.38	1.55 ± 0.42
Bayesian, Prior 3	0.31 ± 0.31	0.51 ± 0.35	1.01 ± 0.35	1.53 ± 0.38
LS	-0.09	0.35	1.98	1.87
REML/BLUP	0.27	0.45	1.00	1.54

Prenatal Survival

Method	Gen 1	Gen 2	Gen 3	Gen 4
Bayesian, Prior 1	-0.53 $\pm$ 1.44	1.23 $\pm$ 1.61	2.83 $\pm$ 1.94	2.89 $\pm$ 2.12
Bayesian, Prior 2	-0.64 $\pm$ 1.70	1.50 $\pm$ 1.87	3.46 $\pm$ 2.05	3.49 $\pm$ 2.30
Bayesian, Prior 3	-0.46 $\pm$ 1.45	1.22 $\pm$ 1.60	2.84 $\pm$ 1.82	2.90 $\pm$ 2.01
LS	-5.71	2.11	4.13	-2.82
REML/BLUP	-0.54	1.49	3.26	3.42

While the results for ovulation rate are consistent across the three priors and with the MM analysis, the results are more problematic for prenatal survival (Table 15.1). Figure 15.3 shows the assumed different priors. As with ovulation rate, prior one is the uninformative prior, weighting all potential additive variances equally. Prior two (as with prior 2 for ovulation rate) assumes a broad distribution around the mean heritability ($h^2 \simeq 0.2$) for a number of studies, while prior three uses the estimate of $h^2 = 0.03 \pm 0.03$ found using French Large Whites. The three priors give Monte Carlo estimates of heritability (and its standard deviation) of $h^2 = 0.12 \pm 0.06$, 0.16 ± 0.04, and 0.11 ± 0.04. Likewise, these priors give Monte Carlo estimates of the repeatability of 0.23 ± 0.05, 0.23 ± 0.04, and 0.19 ± 0.04. As Table 15.1 shows, the standard deviations for the Monte Carlo estimates of mean response are very large, but that the three priors and the MM analysis give consistent results, while the LS results are quite different. Clearly, the information in the experiment is sufficiently small that the posterior is strongly influenced by the prior.

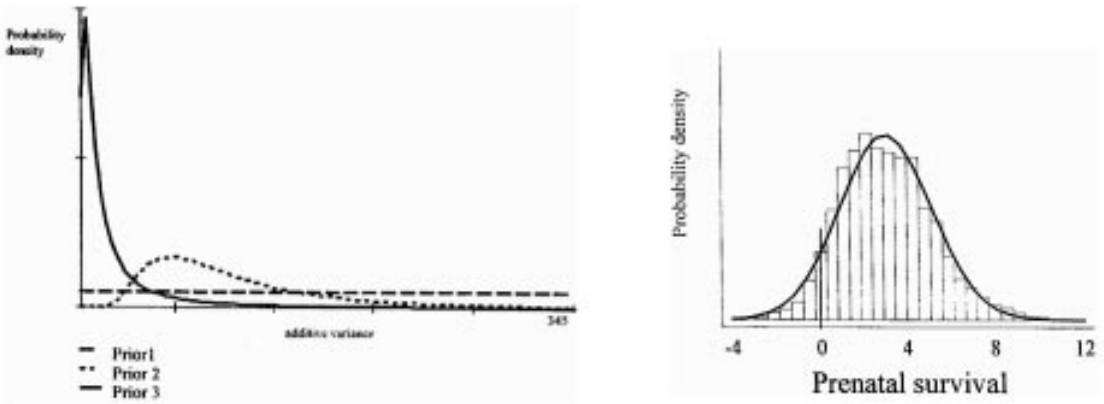


Figure 15.3. Analysis of prenatal survival in French Large White pigs. **Left:** Assumed priors for the additive variance (see text for details). **Right:** Posterior distribution of mean breeding values (at generation four) in prenatal survival, which deviates from the best-fitting normal. After Blasco et al. (1998).

Additional early examples of Bayesian analysis of selection experiments are Rodriguez et al. (1996) who looked at lean growth in pigs and Su et al. (1997) who examined body weight in chickens.

LS, MM, OR BAYES?

So just what analysis should an investigator use on a selection experiment? Obviously, in the absence of any pedigree information, a least-squares analysis is the only option,

although this could also be placed in a Bayesian framework. With the pedigree in hand (either observed or inferred, see Chapter 16), a mixed-model analysis is much more powerful and is strongly preferred over LS, unless there is strong evidence that model assumptions are violated. If a mixed-model approach is appropriate and chosen, should the analysis be a standard or bayesian? As mentioned, the bayesian approach does a much better job of treating uncertainty, but this comes at a high computational cost, especially when one does a proper analysis using several different priors to assess sensitivity. Perhaps the best advice is that offered by Blasco (2001)

> " The choice of one school or the other should be related to whether these are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable scientists feel with the way they convey their results. "

Blasco's last point is especially important: It is far more important for an invesigator to use a method with which they are comfortable, in the sense of knowing its limitations and having some intution into the approach than to simply use a method because it is new and trendy.

Literature Cited

Beniwal, B. K., I. M. Hastings, R. Thompson, and W.G. Hill. 1992a. Estimation of changes in genetic parameters in selected lines of mice using REML with an animal model. 1. Lean mass. *Heredity* 69: 352–360. [15]

Beniwal, B. K., I. M. Hastings, R. Thompson, and W.G. Hill. 1992b. Estimation of changes in genetic parameters in selected lines of mice using REML with an animal model. 2. Body weight, body compostion and litter size. *Heredity* 69: 361–371. [15]

Blasco, A. 2001. The Bayesian controversy in animal breeding. *J. ANim. Sci.* 79: 2023–2046. 15]

Blair, H. T., and E. J. Pollak. 1984. Estimation of genetic trend in a selected population with and without the use of a control population. *J. Animal Sci.* 58: 878–886. [15]

Boichard, D., B. Bonaiti, A. Barbat, and S. Mattalia. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Diary Sci.* 78: 431–437. [15]

Blasco, A., D. Sorensen, and J. P. Bidanel. 1998. Bayesian inference of genetic parameters and selection response for litter size components in pigs. *Genetics* 149: 301–306. [15]

Bulmer, M. G. 1971. The effect of selection on genetic variability. *Amer. Nat.* 105: 201–211. [15]

Curnow, R. N. 1961. The estimation of repeatability and heritability from records subject to culling. *Biometry* 17: 553–566. [15]

de Boer, I. J. M., and J. A. M. van Arendonk. 1992. Prediction of additive and dominance effects in selection or unselected populations with inbreeding. *Theor. Appl. Genet.* 84: 451–459. [15]

Efron, B. 1986. Why isn't everyone a bayesian? *American Statistician* 40: 1-11. [15]

Estany, J., M. Baselga, A. Blasco, and J. Camacho. 1989. Mixed model methodology for the estimation of genetic response to selection in litter size of rabbits. *Livestock Production Science* 21: 67–75. [15]

Fernando, R. L., and D. Gianola. 1990. Statistical inferences in populations undergoing selection or nonrandom mating. *In* D. Gianola and K. Hammond (eds.) *Statistical Methods for Genetic Improvement of Livestock,* pp. 437–453. Springer-Verlag, NY. [15]

Geman, S. and D. Geman. 1984. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741. [15] J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), pp. 169-193. Oxford University Press. [15]

Geyer, C. J. 1992. Practical Markov chain Monte Carlo (with discussion). *Stat. Sci.* 7: 473–511. [15]

Gianola, D., and R. L. Fernando. 1986. Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63: 217–244. [15]

Gianola, D., R. L. Fernando, S. Im, and J. L. Foulley. 1989. Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems. *Genome* 31: 768–777. [15]

Gianola, D., S. Im, and R. L. Fernando. 1988. Prediction of breeding values under Henderson's selection model: a revisitation. *J. Dairy Sci.* 71: 2790–2798. [15]

Glymour, C. 1981. Why I am not a Bayesian, in *The philosophy of science*, *ed.* by D. Papineau. Oxford University Press.

Graser, H.-U., S. P. Smith, and B. Tier. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *J. Anim. Sci.* 64: 1362–1370. [15]

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* 57: 97–109. [15]

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72: 320–338. [15]

Heath, S. C., G. Bulfield, R. Thompson, and P. D. Keightley. 1995. Rates of change of genetic parameters of body weight in selected mouse lines. *Genet. Res. Camb.* 66: 19–25. [15]

Henderson, C. R. 1949. Estimates of changes in herd environment. *J. Dairy Sci.* 32: 706. [15]

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31: 423–447. [15]

Henderson, C. R. 1976. A simple method for the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32: 69–83. [15]

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. von Krosigk. 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15: 192–218. [15]

Hoeschele, I, and P. M. Van Raden. 1991. Rapid inversion of dominance relationship matrices for noninbred populations by including sire by dam subclass effects. *J. Dairy Sci.* 74: 557–569. [15]

Hofer, A. 1998. Variance component estimation in animal breeding: a review. *J. Anim. Breed. Genet.* 115: 247–265. [15]

Jensen, C. S. Wang, D. A. Sorensen, and D. Gianola. 1994. Marginal inferences of variance and covariance components for traits influenced by maternal and direct genetic effects using the Gibbs sampler. *Acta Agric. Scxand.* 44: 193–201 [15]

Jeyaruban, M. G., and J. P. Gibson. 1996. Estimation of additive genetic variance in commercial layer poultry and simulated populations under selection. *Theor. Appl. Genet.* 92: 483–491. [15]

Juga, J., and R. Thompson. 1989. Estimation of variance components in populations selected over multiple generations. *Acta Agric. Scand.* 39: 79–89. [15]

Kackar, R. N., and D. A. Harville. 1981. Unbiasedness of two-stage estimation and prediction for mixed linear models. *Comm. Stat. Theor. Meth.* A10: 1249–1261. [15]

Kennedy, B. W. 1990. Use of mixed model methodology in analysis of designed experiments. *In* D. Gianola and K. Hammond (eds.), *Statistical methods for genetic improvement of livestock,* pp. 77–97. Springer-Verlag, NY. [15]

Kennedy, B. W., and D. A. Sorensen. 1988. Properties of mixed-model methods for prediction of genetic merit. *In* B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong (eds.), *Proceedings of the second international conference on quantitative genetics,* pp. 91 – 103. Sinauer Assoc., Sunderland, MA. [15]

Kennedy, B. W., and D. Trus. 1993. Considerations on genetic connectedness between management units under an animal model. *J. Anim. Sci.* 71: 2341–2352. [15]

Keyes, J. M. 1921. *A treatise on probability*. Macmillan, London. [15]

Kruuk, L. E. B. 2004. Estimating genetic parameters in natural populations using the 'animal model'. *Phil. Trans. R. Soc. Lond. B* 359: 873–890. [15]

ADD Maki-Tanila, A., and B. W. Kennedy. 1986. Mixed model methodology under genetic models with a small number of additive and non-additive loci. *In, Proc. Proc. Third World Cong. on Genetics Applied to Livestock Production* XII: 443–448.. Ames, Iowa. [15]

Martinez, V., L. Bünger, and W. G. Hill. 2000. Analysis of response to 20 generations of selection for body composition in mice: fit to the infinitesimal model assumptions. *Genet. Sel. Evol* 32: 3–21. [15]

Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. *J. Amer. Statist. Assoc.* 44: 335–341. [15]

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A.Teller, and H. Teller. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21: 1087–1091. [15]

Meyer, K., and W. G. Hill. 1991. Mixed model analysis of a selection experiment for food intake in mice. *Genet. Res. Camb.* 57: 71–81. [15]

Milner, J. M., J. M. Pemberton, S. Bortherstone, and S. D. Albon. 2000. Estimating variance components and heritabilities in the wild: a case study using the 'animal model' approach. *J. Evol BIol.* 13: 804–813. [15]

Patterson, H. D., and R. Thompson. 1971. Recovery of interblock information when block sizes are unequal. *Biometrika* 58: 545–554. [15]

Quaas, R. L. 1976. Computing the diagonal elments and inverse of a large numerator relationship matrix. *Biometrics* 32: 949–953. [15]

Raftery, A. E., and S. Lewis. 1992 . How many iterations in the Gibbs sampler? *In, Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), pp. 763–773. Oxford University Press. [15]

Robertson, A. 1977. The effect of selection on the estimation of genetic parameters. *Z. Tier Zuchtungsbiol.* 94: 131–135. [15]

Rodriguez, M. C., M. Toro and L. Silió. 1996. Selection on lean growth in a nucleus of Landrace pigs: an analysis using Gibss sampling. *Anim. Sci.* 63: 243–253. [15]

Rothschild, M. F., C. R. Henderson, and R. L. Quaas. 1979. Effects of selection on variances and covariances of simulated first and second lactations. *J. Dairy Sci.* 62: 996–1002. [15]

Schaeffer, L. R. and H. Song. 1978. Selection bias and REML variance-covariance component estimation. *J. Dairy Sci.* 61: 91–92. [15]

Silverman, B. W. 1986. *Density estimation for statistics and data analysis.* Chapman and Hall, London. [15]

Smith, S. P., and H.-U. Graser. 1986. Estimating variance components in a class of mixed models by restricted maximum likelihood. *J. Dairy Sci.* 69: 1156–1165. [15]

Smith, S. P., and A. Mäki-Tanila. 1990. Genotypic covariance matrices and their inverses for models allowing dominance and inbreeding. *Genet. Sel. Evol.* 22: 65-91. [15]

Southwood, O. I., and B. W. Kennedy. 1991. Genetic and environmental trends for litter size in swine. *J. Anim. Sci.* 69: 3177–3182. [15]

Sorensen, D., R. Fernando, and D. Gianola. 2001. Inferring the trajectory of genetic variance in the course of artificial selection. *Genet. Res. Camb.* 77: 83–94. [15]

Sorensen, D. and D. Gianola. 2002. *Likelihood, bayesian and MCMC methods in quantitative genetics.* Springer

Sorensen, D. A., and B. W. Kennedy. 1984a. Estimation of response to selection using least-squares and mixed model methodology. *J. Anim. Sci.* 58: 1097–1106 [15]

Sorensen, D. A., and B. W. Kennedy. 1983. The use of the relationship matrix to account for genetic drift variances in the analysis of selection experiments. *Theor. Appl. Genet.* 66: 217–220. [15]

Sorensen, D. A., and B. W. Kennedy. 1984b. Estimation of genetic variances from unselected and selected populations. *J. Anim. Sci.* 59: 1213–1223. [15]

Sorensen, D. A., and B. W. Kennedy. 1986. Analysis of selection experiments using mixed model methodology. *J. Anim. Sci.* 63: 245–258. [15]

Sorensen, D. A., C. S. Wang, J. Jensen, and D. Gianola. 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genet. Sel. Evol.* 26: 333–360. [15]

Su, G., P. Sørensen, and D. Sorensen. 1997. Inferences about variance components and selection response for body weight in chickens. *Genet. Sel. Evol* 29: 413–425. [15]

Thompson, R. 1973. The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics* 29: 527–550. [15]

Thompson, R. 1977. The estimation of heritability with unbalanced data. II. Data available on more than two generations. *Biometrics* 33: 496–504. [15]

Thompson, R. 1986. Estimation of realized heritability in a selected population using mixed model methods. *Genet. Sel. Evol.* 18: 475–484. [15]

Thompson, R. 2008. Estimation of quantitative genetic parameters. *Proc. R. Soc. B* 275: 679–686. [15]

Thompson, R., S. Brotherstone, and I. M. White. 2005. Estimation of quantitative genetic parameters. *Phil. Trans. R. Soc. b* 360: 1469–1477. [15]

Thompson, R., and E. Mäntysaari. 2004. Prospects for statistical methods in animal breeding. *J. Ind. Soc. Agri. Stat.* 57: 15–25. [15]

van der Werf, J. H. J. 1990. A note on the use of conditional models to estimate additive genetic variance in selected populations. *In* W. G. Hill, R. Thompson, and J. A. Woolliams (eds.), *Proc. 4th World Congr. Genet. Appl. Livestock Prod.*, Vol. 13, pp. 476–479. Edinburgh. [15]

van der Werf, J. H. J., and I. J. M. de Boer. 1990. Estimation of additive genetic variance when base populations are selected. *J. Anim. Sci.* 68: 3124–3132. [15]

van der Werf, J. H. J., and R. Thompson. 1992. Variance decomposition in the estimation of genetic variance with selected data. *J. Anim. Sci.* 70: 2975–2985. [15]

Visscher, P., and R. Thompson. 1990. REML estimates of parameters for fat yield in pedigree herds in the U.K. using an individual animal model: male and female heritability estimates. *Proc. 4th World Congr. Genet. Appl. Livest. Prod. Edinburgh* 14: 484–487. [15]

Wang, C. S., J. J. Rutledge, and D. Gianola. 1993. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genet. Sel. Evol.* 21: 41–62. [15]

Wang, C. S., D. Gianola, D. A. Sorensen, J. Jensen, A. Christensen, and J. J. Rutledge. 1994a. Response to selection for litter size in Danish landrace pigs: a bayesian analysis. *Theor. Appl. Genet.* 88: 229–230. [15]

Wang, C. S., J. J. Rutledge, and D. Gianola. 1994b. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet. Sel. EVol.* 26: 91–115. [15]

Wilson, A. 2008. Why h^2 does not always equal VA/VP. *J. Evol. Bio.* **xxx: xxx-xxx.** [15]