8

Introduction to Matrix Algebra and Linear Models

We have already encountered several examples of models in which response variables are linear functions of two or more explanatory (or predictor) variables. For example, we have been routinely expressing an individual's phenotypic value as the sum of genotypic and environmental values. A more complicated example is the use of linear regression to decompose an individual's genotypic value into average effects of individual alleles and residual contributions due to interactions between alleles (Chapters 4 and 5). Such **linear models** form the backbone of parameter estimation in quantitative genetics (Chapters 17–27).

This chapter provides a more formal introduction to the general features of linear models, which will be used extensively throughout the rest of this volume, most notably in Chapters 9, 26, and 27. We start by introducing multiple regression, wherein two or more variables are used to make predictions about a response variable. A review of elementary matrix algebra then follows, starting with matrix notation and building up to matrix multiplication and solutions of simultaneous equations using matrix inversion. We next use these results to develop tools for statistical analysis, considering the expectations and covariance matrices of transformed random vectors. After introducing the multivariate normal distribution, which is by far the most important distribution in quantitative-genetics theory, we discuss parameter estimation via both ordinary and generalized least squares. Those with strong statistical backgrounds will find little new in this chapter, other than perhaps some immediate contact with quantitative genetics in the examples. Additional material on matrix algebra and linear models is given in Appendix 3.

MULTIPLE REGRESSION

As a point of departure, consider the multiple regression

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e \tag{8.1}$$

where y is the **response variable**, and the z_i are the **predictor** (or **explanatory**) **variables** used to predict the value of the response variable. This multivariate equation is similar to the expression for a simple linear regression, Equation 3.12a, except that y is now a function of n predictor variables, rather than of one. The variables $y, z_1, \ldots, z_n$ represent observed measures, whereas α and $\beta_1, \ldots, \beta_n$ are constants to be estimated. As in the case of simple linear regression, e (the **residual error**) is the deviation between the observed and fitted value of y. Recall that the use of a linear model involves no assumptions regarding the true form of relationship between y and $z_1, \ldots, z_n$. It simply gives the best linear approximation. Many statistical techniques, including path analysis (Appendix 2) and analysis of variance (Chapter 17), are based on versions of Equation 8.1.

The terms $\beta_1, \ldots, \beta_n$ are known as **partial regression coefficients**. Notice that when all but the *i*th predictor variable are held constant in Equation 8.1, the formula reduces to a univariate model similar to Equation 3.12a but with slope β_i. This partial regression coefficient often differs from the simple regression coefficient, β_i. Suppose, for example, that a simple regression of y on z_1 has a slope of zero. This might lead to the suggestion that there is no relationship between z_1 and y. However, it is conceivable that z_1 actually has a strong positive effect on y that is obscured by positive correlations of z_1 with other variables that have negative influences on y. A multiple regression that included the appropriate variables would clarify this situation by yielding a positive β_1.

Since it is usually impossible for biologists to evaluate partial regression coefficients by empirically imposing constancy on all extraneous variables, we require a more indirect approach to the problem. From Chapter 3, the covariance of y and a predictor variable is

$$\begin{aligned} \sigma(y, z_i) &= \sigma\left[(\alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e), z_i\right] \\ &= \beta_1 \sigma(z_1, z_i) + \beta_2 \sigma(z_2, z_i) + \cdots + \beta_n \sigma(z_n, z_i) + \sigma(e, z_i) \end{aligned} \tag{8.2}$$

The term $\sigma(\alpha, z_i)$ has dropped out because the covariance of z_i with a constant (α) is zero. By applying Equation 8.2 to each predictor variable, we obtain a set of n equations in n unknowns ($\beta_1, \ldots, \beta_n$),

$$\begin{array}{ccccccc} \sigma(y, z_1) = & \beta_1 \sigma^2(z_1) & + \beta_2 \sigma(z_1, z_2) + & \cdots & + \beta_n \sigma(z_1, z_n) & + \sigma(z_1, e) \\ \sigma(y, z_2) = & \beta_1 \sigma(z_1, z_2) + & \beta_2 \sigma^2(z_2) & + \cdots & + \beta_n \sigma(z_2, z_n) & + \sigma(z_2, e) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma(y, z_n) = & \beta_1 \sigma(z_1, z_n) + & \beta_2 \sigma(z_2, z_n) + & \cdots & + \beta_n \sigma^2(z_n) & + \sigma(z_n, e) \end{array} \tag{8.3}$$

As in univariate regression, our task is to find the set of constants (α and the partial regression coefficients β_i) that gives the best linear fit of the conditional expectation of y given $z_1, \cdots, z_n$. Again, the criterion we choose for "best" relies on the **least-squares** approach, which minimizes the squared differences between observed and expected values. Thus, our task is to find that set of $\alpha, \beta_1, \cdots, \beta_n$ giving $\widehat{y} = \alpha + \sum \beta_i z_i$ such that $E[(y - \widehat{y})^2 | z_1, \cdots, z_n]$ is minimized. Taking derivatives of this expectation with respect to α and the β_i and setting each equal to zero, it can be shown that the set of equations given by Equation 8.3 is, in fact, the least-squares solution to Equation 8.1. If the appropriate variances and covariances are known, the β_i can be obtained exactly. If these are unknown, as is usually the case, the least-squares estimates b_i are obtained from Equation 8.3 by substituting the observed (estimated) variances and covariances for their (unknown) population values.

The properties of least-squares multiple regression are analogous to those for simple regression. First, the procedure yields a solution such that the average deviation of y from $\widehat{y}$, $E(e)$, is zero. Hence $E(\,y\,) = E(\,\widehat{y}\,)$, implying

$$\bar{y} = a + b_1\bar{z}_1 + \cdots + b_n\bar{z}_n$$

Thus, once the fitted values $b_1, \ldots, b_n$ are obtained from Equation 8.3, the intercept is defined by $a = \bar{y} - b_1\bar{z}_1 - \cdots - b_n\bar{z}_n$. Second, least-squares analysis gives a solution in which the residual errors are uncorrelated with the predictor variables. Thus, the terms $\sigma(e, z_i)$ can be dropped from Equation 8.3. Third, the partial regression coefficients are entirely defined by variances and covariances. However, unlike simple regression coefficients, which depend on only a single variance and covariance, each partial regression coefficient is a function of the variances and covariances of all measured variables. Notice that if $n = 1$, then $\sigma(y, z_1) = \beta_1\sigma^2(z_1)$, and we recover the univariate solution $\beta_1 = \sigma(y, z_1)/\sigma^2(z_1)$.

A simple pattern exists in each of the n equations in 8.3. The *i*th equation defines the covariance of y and z_i as the sum of two types of quantities: a single term, which is the product of the *i*th partial regression coefficient and the variance of z_i, and a set of $(n-1)$ terms, each of which is the product of a partial regression coefficient and the covariance of z_i with the corresponding predictor variable. This general pattern suggests an alternative way of writing Equation 8.3,

$$\begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) & \ldots & \sigma(z_1, z_n) \\ \sigma(z_1, z_2) & \sigma^2(z_2) & \ldots & \sigma(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(z_1, z_n) & \sigma(z_2, z_n) & \ldots & \sigma^2(z_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \\ \vdots \\ \sigma(y, z_n) \end{pmatrix} \tag{8.4}$$

The table of variances and covariances on the left is referred to as a **matrix**, while the columns of partial regression coefficients and of covariances involving y are called **vectors**. If these matrices and vectors are abbreviated as $\mathbf{V}$, $\boldsymbol{\beta}$, and $\mathbf{c}$, Equation 8.4 can be written even more compactly as

$$\mathbf{V}\boldsymbol{\beta} = \mathbf{c} \tag{8.5}$$

The standard procedure of denoting matrices as bold capital letters and vectors as bold lowercase letters is adhered to in this book. Notice that $\mathbf{V}$, which is generally called a **covariance matrix**, is symmetrical about the main diagonal. As we shall see shortly, the *i*th equation in 8.3 can be recovered from Equation 8.4 by multiplying the elements in $\boldsymbol{\beta}$ by the corresponding elements in the ith horizontal row of the matrix $\mathbf{V}$. Although a great deal of notational simplicity has been gained by condensing the system of Equations 8.3 to matrix form, this does not alter the fact that the solution of a large system of simultaneous equations is a tedious task if performed by hand. Today, such solutions are rapidly accomplished on computers. Before considering matrix methods in more detail, we present an application of Equation 8.1 to quantitative genetics.

An Application to Multivariate Selection

Karl Pearson developed the technique of multiple regression in 1896, although some of the fundamentals can be traced to his predecessors (Pearson 1920, Stigler 1986). Pearson is perhaps best known as one of the founders of statistical methodology, but his intense interest in evolution may have been the primary motivating force underlying many of his theoretical endeavors. Almost all of his major papers, including the one of 1896, contain rigorous analyses of data gathered by his contemporaries on matters such as resemblance between relatives, natural selection, correlation between characters, and assortative mating (recall the assortative mating example in Chapter 7). The foresight of these studies is remarkable considering that they were performed prior to the existence of a genetic interpretation for the expression and inheritance of polygenic traits.

Pearson's (1896, 1903) invention of multiple regression developed out of the need for a technique to resolve the observed directional selection on a character into its direct and various indirect components. In Chapter 3 we defined the selection differential S (the within-generation change in the mean phenotype due to selection) as a measure of the total directional selection on a character. However, S cannot be considered to be a measure of the direct forces of selection on a character unless that character is uncorrelated with all other selected traits. An unselected character can appear to be under selection if other characters with which it is correlated are under directional selection. Alternatively, a character under strong directional selection may exhibit a negligible selection differential if the indirect effects of selection on correlated traits are sufficiently compensatory.

Because he did not employ matrix notation, some of the mathematics in Pearson's papers can be rather difficult to follow. Lande and Arnold (1983) did a great service by extending this work and rephrasing it in matrix notation. Suppose that a large number of individuals in a population have been measured for n characters and for fitness. Individual fitness can then be approximated by the linear model

$$w = \alpha + \beta_1 z_1 + \cdots + \beta_n z_n + e$$

where w is relative fitness (observed fitness divided by the mean fitness in the population), and $z_1, \ldots, z_n$ are the phenotypic measures of the n characters. Recall from Chapter 3 that the selection differential for the ith trait is defined as the covariance between phenotype and relative fitness, $S_i = \sigma(z_i, w)$. Thus, we have

$$\begin{aligned} S_i = \sigma(z_i, w) &= \sigma(z_i, \alpha + \beta_1 z_1 + \cdots + \beta_n z_n + e) \\ &= \beta_1 \sigma(z_i, z_1) + \cdots + \beta_n \sigma(z_i, z_n) + \sigma(z_i, e) \end{aligned}$$

Note that this expression is of the same form as Equation 8.3, so that by taking the β_i to be the partial regression coefficients we have $\sigma(z_i, e) = 0$. Note also that the selection differential of any trait may be partitioned into a component estimating the **direct selection** on the character and the sum of components from **indirect selection** on all correlated characters,

$$S_i = \beta_i \sigma^2(z_i) + \sum_{j \neq i}^{n} \beta_j \sigma(z_i, z_j)$$

It is important to realize that the labels "direct" and "indirect" apply strictly to the specific set of characters included in the analysis; the partial regression coefficients are subject to change if a new analysis includes additional correlated characters that are under selection.

Example 1. A morphological analysis of a pentatomid bug (*Euschistus variolarius*) population performed by Lande and Arnold (1983) provides a good example of the insight that can be gained from a multivariate approach. The bugs were collected along the shore of Lake Michigan after a storm. Of the 94 individuals that were recovered, 39 were alive. All individuals were measured for four characters: head and thorax width, and scutellum and forewing length. The data were then logarithmically transformed to more closely approximate normality (Chapter 11). All surviving bugs were assumed to have equal fitness ($W = 1$), and all dead bugs to have zero fitness ($W = 0$). Hence, mean fitness is the fraction p of individuals that survived, giving **relative fitnesses**, $w = W/\overline{W}$, as

$$w = \begin{cases} 1/p & \text{if the individual survived} \\ 0 & \text{if the individual did not survive} \end{cases}$$

The selection differential for each of the characters is simply the difference between the mean phenotype of the 39 survivors and the mean of the entire sample. These are reported in units of phenotypic standard deviations in the following table, along with the partial regression coefficients of relative fitness on the four morphological characters. Here * and ** indicate significance at the 5% and 1% levels. All of the phenotypic correlations are highly significant.

Character	Selection Differential	Partial Regression Coef. of Fitness	Phenotypic Correlations			
z_i	S_i	b_i	H	T	S	F
Head (H)	–0.11	–0.7	1.00	0.72	0.50	0.60
Thorax (T)	–0.06	11.6**		1.00	0.59	0.71
Scutellum (S)	–0.28*	–2.8			1.00	0.62
Forewing (F)	–0.43**	–16.6**				1.00

The estimates of the partial regression coefficients nicely illustrate two points discussed earlier. First, despite the strong directional selection operating directly on thorax size, the selection differential for thorax size is negligible. This lack of apparent selection results because the positive correlation between thorax width and wing length is coupled with negative forces of selection on the latter character. Second, there is a significant negative selection differential on scutellum length even though there is no significant direct selection on the character. The negative selection differential is largely an indirect consequence of the strong selection for smaller wing length.

ELEMENTARY MATRIX ALGEBRA

The solutions of systems of linear equations generally involve the use of matrices and vectors of variables. For those with little familiarity with such constructs and their manipulations, the next few pages provide an overview of the basic tools of matrix algebra.

Basic Notation

A matrix is simply a rectangular array of numbers. Some examples are:

$$\mathbf{a} = \begin{pmatrix} 12 \\ 13 \\ 47 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 2 & 0 & 5 & 21 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} 0 & 1 \\ 3 & 4 \\ 2 & 9 \end{pmatrix}$$

A matrix with r rows and c columns is said to have **dimensionality** $r \times c$ (a useful mnemonic for remembering this is *r*ailroad *c*ar). In the examples above, $\mathbf{D}$ has three rows and two columns, and is thus a 3×2 matrix. An $r \times 1$ matrix, such as $\mathbf{a}$, is a **column vector**, while a $1 \times c$ matrix, such as $\mathbf{b}$, is a **row vector**. A matrix in which the number of rows equals the number of columns, such as $\mathbf{C}$, is called a **square matrix**. Numbers are also matrices (of dimensionality 1×1) and are often referred to as **scalars**.

A matrix is completely specified by the **elements** that comprise it, with M_{ij} denoting the element in the *i*th row and *j*th column of matrix **M**. Using the sample matrices above, $C_{23} = 4$ is the element in the second row and third column of **C**. Likewise, $C_{32} = 1$ is the element in the third row and second column. Two matrices are equal if and only if all of their corresponding elements are equal.

Partitioned Matrices

It is often useful to work with **partitioned matrices** wherein each element in a matrix is itself a matrix. There are several ways to partition a matrix. For example, we could write the matrix **C** above as

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & \vdots & 1 & 2 \\ \cdots & \cdots & \cdots & \cdots \\ 2 & \vdots & 5 & 4 \\ 1 & \vdots & 1 & 2 \end{pmatrix} = \begin{pmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{B} \end{pmatrix}$$

where

$$\mathbf{a} = (\,3\,), \quad \mathbf{b} = (\,1 \quad 2\,), \quad \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 5 & 4 \\ 1 & 2 \end{pmatrix}$$

Alternatively, we could partition **C** into a single row vector whose elements are themselves column vectors,

$$\mathbf{C} = (\,\mathbf{c_1} \quad \mathbf{c_2} \quad \mathbf{c_3}\,) \quad \text{where} \quad \mathbf{c_1} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c_2} = \begin{pmatrix} 1 \\ 5 \\ 1 \end{pmatrix}, \quad \mathbf{c_3} = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix}$$

or **C** could be written as a column vector whose elements are row vectors,

$$\mathbf{C} = \begin{pmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \mathbf{b_3} \end{pmatrix} \quad \text{where} \quad \mathbf{b_1} = (\,3 \quad 1 \quad 2\,), \quad \mathbf{b_2} = (\,2 \quad 5 \quad 4\,), \quad \mathbf{b_3} = (\,1 \quad 1 \quad 2\,)$$

Addition and Subtraction

Addition and subtraction of matrices is straightforward. To form a new matrix $\mathbf{A} + \mathbf{B} = \mathbf{C}$, **A** and **B** must have the same dimensions. One then simply adds the corresponding elements, $C_{ij} = A_{ij} + B_{ij}$. Subtraction is defined similarly. For example, if

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

then

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} 4 & 2 \\ 3 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}$$

Multiplication

Multiplying a matrix by a scalar is also straightforward. If $\mathbf{M} = a\mathbf{N}$, where a is a scalar, then $M_{ij} = aN_{ij}$. Each element of $\mathbf{N}$ is simply multiplied by the scalar. For example,

$$(-2)\begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 0 \\ -6 & -2 \end{pmatrix}$$

Matrix multiplication is a little more involved. We start by considering the **dot product** of two vectors, as this forms the basic operation of matrix multiplication. Letting $\mathbf{a}$ and $\mathbf{b}$ be two n-dimensional vectors (the first a column vector, the second a row vector), their dot product $\mathbf{a} \cdot \mathbf{b}$ is a scalar given by

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

For example, for the two vectors

$$\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 4 & 5 & 7 & 9 \end{pmatrix}$$

the dot product is $\mathbf{a} \cdot \mathbf{b} = (1 \times 4) + (2 \times 5) + (3 \times 7) + (4 \times 9) = 71$. Note that the dot product is not defined if the vectors have different lengths.

Now consider the matrix $\mathbf{L} = \mathbf{MN}$ produced by multiplying the $r \times c$ matrix $\mathbf{M}$ by the $c \times b$ matrix $\mathbf{N}$. Partitioning $\mathbf{M}$ as a column vector of r row vectors,

$$\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_r} \end{pmatrix} \qquad \text{where} \qquad \mathbf{m_i} = \begin{pmatrix} M_{i1} & M_{i2} & \cdots & M_{ic} \end{pmatrix}$$

and $\mathbf{N}$ as a row vector of b column vectors,

$$\mathbf{N} = \begin{pmatrix} \mathbf{n_1} & \mathbf{n_2} & \cdots & \mathbf{n_b} \end{pmatrix} \qquad \text{where} \qquad \mathbf{n_j} = \begin{pmatrix} N_{1j} \\ N_{2j} \\ \vdots \\ N_{cj} \end{pmatrix}$$

the *ij*th element of $\mathbf{L}$ is given by the dot product

$$L_{ij} = \mathbf{m_i} \cdot \mathbf{n_j} = \sum_{k=1}^{c} M_{ik} N_{kj} \tag{8.6a}$$

Hence the resulting matrix $\mathbf{L}$ is of dimension $r \times b$ with

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1} \cdot \mathbf{n_1} & \mathbf{m_1} \cdot \mathbf{n_2} & \cdots & \mathbf{m_1} \cdot \mathbf{n_b} \\ \mathbf{m_2} \cdot \mathbf{n_1} & \mathbf{m_2} \cdot \mathbf{n_2} & \cdots & \mathbf{m_2} \cdot \mathbf{n_b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_r} \cdot \mathbf{n_1} & \mathbf{m_r} \cdot \mathbf{n_2} & \cdots & \mathbf{m_r} \cdot \mathbf{n_b} \end{pmatrix} \tag{8.6b}$$

Note that using this definition, the matrix product given by Equation 8.4 recovers the set of equations given by Equation 8.3.

Example 2. Compute the product $\mathbf{L} = \mathbf{MN}$ where

$$\mathbf{M} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{N} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 3 \\ 3 & 2 & 2 \end{pmatrix}$$

Writing $\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \mathbf{m_3} \end{pmatrix}$ and $\mathbf{N} = (\, \mathbf{n_1} \quad \mathbf{n_2} \quad \mathbf{n_3} \,)$, we have

$$\mathbf{m_1} = (\, 3 \quad 1 \quad 2 \,), \quad \mathbf{m_2} = (\, 2 \quad 5 \quad 4 \,), \quad \mathbf{m_3} = (\, 1 \quad 1 \quad 2 \,)$$

and

$$\mathbf{n_1} = \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix}, \quad \mathbf{n_2} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{n_3} = \begin{pmatrix} 0 \\ 3 \\ 2 \end{pmatrix}$$

The resulting matrix $\mathbf{L}$ is 3×3. Applying Equation 8.6b, the element in the first row and first column of $\mathbf{L}$ is the dot product of the first row vector of $\mathbf{M}$ with the first column vector of $\mathbf{N}$,

$$L_{11} = \mathbf{m_1} \cdot \mathbf{n_1} = (\, 3 \quad 1 \quad 2 \,) \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix} = \sum_{k=1}^{3} M_{1k} N_{k1}$$

$$= M_{11}N_{11} + M_{12}N_{21} + M_{13}N_{31} = (3 \times 4) + (1 \times 1) + (2 \times 3) = 19$$

Computing the other elements gives

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1} \cdot \mathbf{n_1} & \mathbf{m_1} \cdot \mathbf{n_2} & \mathbf{m_1} \cdot \mathbf{n_3} \\ \mathbf{m_2} \cdot \mathbf{n_1} & \mathbf{m_2} \cdot \mathbf{n_2} & \mathbf{m_2} \cdot \mathbf{n_3} \\ \mathbf{m_3} \cdot \mathbf{n_1} & \mathbf{m_3} \cdot \mathbf{n_2} & \mathbf{m_3} \cdot \mathbf{n_3} \end{pmatrix} = \begin{pmatrix} 19 & 8 & 7 \\ 25 & 15 & 23 \\ 11 & 6 & 7 \end{pmatrix}$$

Certain dimensional properties must be satisfied when two matrices are to be multiplied. Specifically, since the dot product is defined only for vectors of the same length, for the matrix product $\mathbf{MN}$ to be defined, the number of columns in $\mathbf{M}$ must equal the number of rows in $\mathbf{N}$. Thus, while

$$\begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 12 \\ 10 \end{pmatrix}, \qquad \begin{pmatrix} 4 \\ 3 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \quad \text{is undefined.}$$

Writing $\mathbf{M}_{r\times c}\mathbf{N}_{c\times b} = \mathbf{L}_{r\times b}$ shows that the inner indices must match, while the outer indices (r and b) give the number of rows and columns of the resulting matrix. The order in which matrices are multiplied is critical. In general, $\mathbf{AB}$ is not equal to $\mathbf{BA}$. For example, when the order of the matrices in Example 2 is reversed,

$$\mathbf{NM} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 3 \\ 3 & 2 & 2 \end{pmatrix} \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 14 & 9 & 12 \\ 8 & 9 & 12 \\ 15 & 15 & 18 \end{pmatrix}$$

Since order is important in matrix multiplication, it has specific terminology. For the product $\mathbf{AB}$, we say that matrix $\mathbf{B}$ is **premultiplied** by the matrix $\mathbf{A}$, or that matrix $\mathbf{A}$ is **postmultiplied** by the matrix $\mathbf{B}$.

Transposition

Another useful matrix operation is **transposition**. The transpose of a matrix $\mathbf{A}$ is written $\mathbf{A}^T$ (while not used in this book, the notation $\mathbf{A}'$ is also widely used), and is obtained simply by switching rows and columns of the original matrix. For example,

$$\begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix}^T = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 4 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 7 & 4 & 5 \end{pmatrix}^T = \begin{pmatrix} 7 \\ 4 \\ 5 \end{pmatrix}$$

A useful identity for transposition is that

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{8.7a}$$

which holds for any number of matrices, e.g.,

$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T \tag{8.7b}$$

Vectors of statistics are generally written as column vectors and we follow this convention by using lowercase bold letters, e.g., $\mathbf{a}$, for a column vector and $\mathbf{a}^T$

for the corresponding row vector. With this convention, we distinguish between two vector products, the **inner product** (the dot product) which yields a scalar and the **outer product** which yields a matrix. For the two n-dimensional column vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

the inner product is given by

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{8.8a}$$

while the outer product yields the $n \times n$ matrix

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix} = \mathbf{a}\mathbf{b}^T = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix} \tag{8.8b}$$

Inverses and Solutions to Systems of Equations

While matrix multiplication provides a compact way of writing systems of equations, we also need a compact notation for expressing the solutions of such systems. Such solutions utilize the **inverse** of a matrix, an operation analogous to scalar division. The essential utility of matrix inversion can be noted by first considering the solution of the simple scalar equation $ax = b$ for x. Multiplying both sides by a^{-1}, we have $(a^{-1}a)x = 1 \cdot x = x = a^{-1}b$. Now consider a square matrix $\mathbf{A}$. The **inverse of** $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, satisfies $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$, where $\mathbf{I}$, the **identity matrix**, is a square matrix with diagonal elements equal to one and all other elements equal to zero. The identity matrix serves the role that 1 plays in scalar multiplication. Just as $1 \times a = a \times 1 = a$ in scalar multiplication, for any matrix $\mathbf{A} = \mathbf{IA} = \mathbf{AI}$. A matrix is called **nonsingular** if its inverse exists. Conditions under which this occurs are discussed in the next section. A useful property of inverses is that if the matrix product $\mathbf{AB}$ is a square matrix (where $\mathbf{A}$ and $\mathbf{B}$ are square), then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{8.9}$$

The fundamental relationship between the inverse of a matrix and the solution of systems of linear equations can be seen as follows. For a square nonsingular matrix $\mathbf{A}$, the unique solution for $\mathbf{x}$ in the matrix equation $\mathbf{Ax} = \mathbf{c}$ is obtained by premultiplying by $\mathbf{A}^{-1}$,

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{Ax} = \mathbf{A}^{-1}\mathbf{c} \tag{8.10a}$$

When $\mathbf{A}$ is either singular or nonsquare, solutions for $\mathbf{x}$ can still be obtained using **generalized inverses** in place of $\mathbf{A}^{-1}$ (Appendix 3), but such solutions are not unique, applying instead to certain linear combinations of the elements of $\mathbf{x}$. (See Appendix 3 for details.) Recalling Equation 8.5, the solution of the multiple regression equation can be expressed as

$$\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c} \tag{8.10b}$$

Likewise, for the Pearson-Lande-Arnold regression giving the best linear predictor of fitness,

$$\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{s} \tag{8.10c}$$

where $\mathbf{P}$ is the covariance matrix for phenotypic measures $z_1, \ldots, z_n$, and $\mathbf{s}$ is the vector of selection differentials for the n characters.

Before developing the formal method for inverting a matrix, we consider two extreme (but very useful) cases that lead to simple expressions for the inverse. First, if the matrix is **diagonal** (all off-diagonal elements are zero), then the matrix inverse is also diagonal, with $\mathbf{A}_{ii}^{-1} = 1/A_{ii}$. For example,

$$\text{for} \quad \mathbf{A} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \qquad \text{then} \qquad \mathbf{A}^{-1} = \begin{pmatrix} a^{-1} & 0 & 0 \\ 0 & b^{-1} & 0 \\ 0 & 0 & c^{-1} \end{pmatrix}$$

Note that if any of the diagonal elements of $\mathbf{A}$ are zero, $\mathbf{A}^{-1}$ is not defined, as $1/0$ is undefined. Second, for any 2×2 matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{then} \qquad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{8.11}$$

To check this result, note that

$$\mathbf{A}\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$= \frac{1}{ad - bc} \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix} = \mathbf{I}$$

If $ad = bc$, the inverse does not exist, as division by zero is undefined.

Example 3. Consider the multiple regression of y on two predictor variables, z_1 and z_2, so that $y = \alpha + \beta_1 z_1 + \beta_2 z_2 + e$. In the notation of Equation 8.5, we have

$$\mathbf{c} = \begin{pmatrix} \sigma(y, z_1) \\ \\ \sigma(y, z_2) \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) \\ \\ \sigma(z_1, z_2) & \sigma^2(z_2) \end{pmatrix}$$

Recalling that $\sigma(z_1, z_2) = \rho_{12}\,\sigma(z_1)\sigma(z_2)$, Equation 8.11 gives

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix}$$

The inverse exists provided both characters have nonzero variance and are not completely correlated ($|\rho_{12}| \neq 1$). Recalling Equation 8.10b, the partial regression coefficients are given by $\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c}$, or

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix} \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \end{pmatrix}$$

Again using $\sigma(z_1, z_2) = \rho_{12}\,\sigma(z_1)\sigma(z_2)$, this equation reduces to

$$\beta_1 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y, z_1)}{\sigma^2(z_1)} - \rho_{12}\,\frac{\sigma(y, z_2)}{\sigma(z_1)\sigma(z_2)}\right]$$

and

$$\beta_2 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y, z_2)}{\sigma^2(z_2)} - \rho_{12}\,\frac{\sigma(y, z_1)}{\sigma(z_1)\sigma(z_2)}\right]$$

Note that only when the predictor variables are uncorrelated ($\rho_{12} = 0$), do the partial regression coefficients β_1 and β_2 reduce to the univariate regression slopes,

$$\beta_1 = \frac{\sigma(y, z_1)}{\sigma^2(z_1)} \quad \text{and} \quad \beta_2 = \frac{\sigma(y, z_2)}{\sigma^2(z_2)}$$

Determinants and Minors

For a 2×2 matrix, the quantity

$$|\mathbf{A}| = A_{11}A_{22} - A_{12}A_{21} \tag{8.12a}$$

is called the **determinant**, which more generally is denoted by det($\mathbf{A}$) or $|\mathbf{A}|$. As with the 2- dimensional case, $\mathbf{A}^{-1}$ exists for a square matrix $\mathbf{A}$ (of any dimensionality) if and only if det($\mathbf{A}$) $\neq 0$. For square matrices with dimensionality greater than two, the determinant is obtained recursively from the general expression

$$|\mathbf{A}| = \sum_{j=1}^{n} A_{ij}(-1)^{i+j}|\mathbf{A}_{ij}| \tag{8.12b}$$

where i is any fixed row of the matrix $\mathbf{A}$ and $\mathbf{A}_{ij}$ is a submatrix obtained by deleting the ith row and jth column from $\mathbf{A}$. Such a submatrix is known as a **minor.** In words, each of the n quantities in this equation is the product of three components: the element in the row around which one is working, -1 to the $(i+j)$th power, and the determinant of the ijth minor. In applying Equation 8.12b, one starts with the original $n \times n$ matrix and works down until the minors are reduced to 2×2 matrices whose determinants are scalars of the form $A_{11}A_{22} - A_{12}A_{21}$. A useful result is that the determinant of a diagonal matrix is the product of the diagonal elements of that matrix, so that if

$$A_{ij} = \begin{cases} a_i & i = j \\ 0 & i \neq j \end{cases} \qquad \text{then} \qquad |\,\mathbf{A}\,| = \prod_{i=1}^{n} a_i$$

The next section shows how determinants are used in the computation of a matrix inverse.

Example 4. Compute the determinant of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

Letting $i = 1$, i.e., using the elements in the first row of $\mathbf{A}$,

$$|\mathbf{A}| = 1 \cdot (-1)^{1+1} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} + 1 \cdot (-1)^{1+2} \begin{vmatrix} 1 & 2 \\ 1 & 1 \end{vmatrix} + 1 \cdot (-1)^{1+3} \begin{vmatrix} 1 & 3 \\ 1 & 2 \end{vmatrix}$$

Using Equation 8.12a to obtain the determinants of the 2×2 matrices, this simplifies to

$$|\mathbf{A}| = [1 \times (3 - 4)] - [1 \times (1 - 2)] + [1 \times (2 - 3)] = -1$$

The same answer is obtained regardless of which row is used, and expanding around a column, instead of a row, produces the same result. Thus, in order to reduce the number of computations required to obtain a determinant, it is useful to expand using the row or column that contains the most zeros.

Computing Inverses

The general solution of a matrix inverse is

$$A_{ij}^{-1} = \left[\frac{(-1)^{i+j} |\mathbf{A}_{ij}|}{|\mathbf{A}|} \right]^T \tag{8.13}$$

where A^{-1}_{ij} denotes the ijth element of $\mathbf{A}^{-1}$, and $\mathbf{A}_{ij}$ denotes the ijth minor of $\mathbf{A}$. It can be seen from Equation 8.13 that a matrix can only be inverted if it has a nonzero determinant. Thus, a matrix is singular if its determinant is zero. This occurs whenever a matrix contains a row (or column) that can be written as a weighted sum of any other rows (or columns). In the context of our linear model, Equation 8.4, this happens if one of the n equations can be written as a combination of the others, a situation that is equivalent to there being n unknowns but less than n independent equations.

Example 5. Compute the inverse of

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

First, find the determinants of the minors,

$$|\mathbf{A}_{11}| = \begin{vmatrix} 5 & 4 \\ 1 & 2 \end{vmatrix} = 6 \qquad |\mathbf{A}_{23}| = \begin{vmatrix} 3 & 1 \\ 1 & 1 \end{vmatrix} = 2$$

$$|\mathbf{A}_{12}| = \begin{vmatrix} 2 & 4 \\ 1 & 2 \end{vmatrix} = 0 \qquad |\mathbf{A}_{31}| = \begin{vmatrix} 1 & 2 \\ 5 & 4 \end{vmatrix} = -6$$

$$|\mathbf{A}_{13}| = \begin{vmatrix} 2 & 5 \\ 1 & 1 \end{vmatrix} = -3 \qquad |\mathbf{A}_{32}| = \begin{vmatrix} 3 & 2 \\ 2 & 4 \end{vmatrix} = 8$$

$$|\mathbf{A}_{21}| = \begin{vmatrix} 1 & 2 \\ 1 & 2 \end{vmatrix} = 0 \qquad |\mathbf{A}_{33}| = \begin{vmatrix} 3 & 1 \\ 2 & 5 \end{vmatrix} = 13$$

$$|\mathbf{A}_{22}| = \begin{vmatrix} 3 & 2 \\ 1 & 2 \end{vmatrix} = 4$$

Using Equation 8.12b and expanding using the first row of $\mathbf{A}$ gives

$$|\mathbf{A}| = 3|\mathbf{A}_{11}| - |\mathbf{A}_{12}| + 2|\mathbf{A}_{13}| = 12$$

Returning to the matrix in brackets in Equation 8.13, we obtain

$$\frac{1}{12}\begin{pmatrix} 1\times 6 & -1\times 0 & 1\times -3 \\ -1\times 0 & 1\times 4 & -1\times 2 \\ 1\times -6 & -1\times 8 & 1\times 13 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 6 & 0 & -3 \\ 0 & 4 & -2 \\ -6 & -8 & 13 \end{pmatrix}$$

and then taking the transpose,

$$\mathbf{A}^{-1} = \frac{1}{12}\begin{pmatrix} 6 & 0 & -6 \\ 0 & 4 & -8 \\ -3 & -2 & 13 \end{pmatrix}$$

To verify that this is indeed the inverse of $\mathbf{A}$, multiply $\mathbf{A}^{-1}$ by $\mathbf{A}$,

$$\frac{1}{12}\begin{pmatrix} 6 & 0 & -6 \\ 0 & 4 & -8 \\ -3 & -2 & 13 \end{pmatrix}\begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 12 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

EXPECTATIONS OF RANDOM VECTORS AND MATRICES

Matrix algebra provides a powerful approach for analyzing linear combinations of random variables. Let $\mathbf{x}$ be a column vector containing n random variables, $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$. We may wish to construct a new univariate (scalar) random variable y by taking some linear combination of the elements of $\mathbf{x}$,

$$y = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T \mathbf{x}$$

where $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$ is a column vector of constants. Likewise, we can construct a new k-dimensional vector $\mathbf{y}$ by premultiplying $\mathbf{x}$ by a $k \times n$ matrix $\mathbf{A}$ of constants, $\mathbf{y} = \mathbf{A}\mathbf{x}$. More generally, an $(n \times k)$ matrix $\mathbf{X}$ of random variables can be transformed into a new $m \times \ell$ dimensional matrix $\mathbf{Y}$ of elements consisting of linear combinations of the elements of $\mathbf{X}$ by

$$\mathbf{Y}_{m\times\ell} = \mathbf{A}_{m\times n}\mathbf{X}_{n\times k}\mathbf{B}_{k\times\ell} \tag{8.14}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are constants with dimensions as subscripted.

If $\mathbf{X}$ is a matrix whose elements are random variables, then the expected value of $\mathbf{X}$ is a matrix $E(\mathbf{X})$ containing the expected value of each element of $\mathbf{X}$. If $\mathbf{X}$ and $\mathbf{Z}$ are matrices of the same dimension, then

$$E(\mathbf{X} + \mathbf{Z}) = E(\mathbf{X}) + E(\mathbf{Z}) \tag{8.15}$$

This easily follows since the ijth element of $E(\mathbf{X} + \mathbf{Z})$ is $E(x_{ij} + z_{ij}) = E(x_{ij}) + E(z_{ij})$. Similarly, the expectation of $\mathbf{Y}$ as defined in Equation 8.14 is

$$E(\mathbf{Y}) = E(\mathbf{A}\mathbf{X}\mathbf{B}) = \mathbf{A}E(\mathbf{X})\mathbf{B} \tag{8.16a}$$

For example, for $\mathbf{y} = \mathbf{X}\mathbf{b}$ where $\mathbf{b}$ is an $n \times 1$ column vector,

$$E(\mathbf{y}) = E(\mathbf{X}\mathbf{b}) = E(\mathbf{X})\mathbf{b} \tag{8.16b}$$

Likewise, for $y = \mathbf{a}^T\mathbf{x} = \sum_i^n a_i x_i$,

$$E(y) = E(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T E(\mathbf{x}) \tag{8.16c}$$

COVARIANCE MATRICES OF TRANSFORMED VECTORS

To develop expressions for variances and covariances of linear combinations of random variables, we must first introduce the concept of quadratic forms. Consider an $n \times n$ square matrix $\mathbf{A}$ and an $n \times 1$ column vector $\mathbf{x}$. From the rules of matrix multiplication,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j \tag{8.17}$$

Expressions of this form are called **quadratic forms** (or **quadratic products**) and yield a scalar. A generalization of a quadratic form is the **bilinear form**, $\mathbf{b}^T\mathbf{A}\mathbf{a}$, where $\mathbf{b}$ and $\mathbf{a}$ are, respectively, $n \times 1$ and $m \times 1$ column vectors and $\mathbf{A}$ is an $n \times m$ matrix. Indexing the matrices and vectors in this expression by their dimensions, $\mathbf{b}^T_{1\times n}\mathbf{A}_{n\times m}\mathbf{a}_{m\times 1}$, shows that the resulting matrix product is a 1×1 matrix — in other words, a scalar. As scalars, bilinear forms equal their transposes, giving the useful identity

$$\mathbf{b}^T\mathbf{A}\mathbf{a} = \left(\mathbf{b}^T\mathbf{A}\mathbf{a}\right)^T = \mathbf{a}^T\mathbf{A}^T\mathbf{b} \tag{8.18}$$

Again let $\mathbf{x}$ be a column vector of n random variables. A compact way to express the n variances and $n(n-1)/2$ covariances associated with the elements of $\mathbf{x}$ is the matrix $\mathbf{V}$, where $V_{ij} = \sigma(x_i, x_j)$ is the covariance between the random variables x_i and x_j. We will generally refer to $\mathbf{V}$ as a **covariance matrix**, noting that the diagonal elements represent the variances and off-diagonal elements the covariances. The $\mathbf{V}$ matrix is symmetric, as

$$V_{ij} = \sigma(x_i, x_j) = \sigma(x_j, x_i) = V_{ji}$$

Now consider a univariate random variable $y = \sum c_k\, x_k$ generated from a linear combination of the elements of $\mathbf{x}$. In matrix notation, $y = \mathbf{c}^T\mathbf{x}$, where $\mathbf{c}$ is a column vector of constants. The variance of y can be expressed as a quadratic form involving the covariance matrix $\mathbf{V}$ for the elements of $\mathbf{x}$,

$$\begin{aligned}\sigma^2\left(\mathbf{c}^T\mathbf{x}\right) &= \sigma^2\left(\sum_{i=1}^{n} c_i x_i\right) = \sigma\left(\sum_{i=1}^{n} c_i\, x_i\, , \sum_{j=1}^{n} c_j\, x_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma\left(c_i\, x_i, c_j\, x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i\, c_j\, \sigma\left(x_i, x_j\right) \\ &= \mathbf{c}^T\mathbf{V}\,\mathbf{c}\end{aligned} \tag{8.19}$$

Likewise, the covariance between two univariate random variables created from different linear combinations of $\mathbf{x}$ is given by the bilinear form

$$\sigma(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{x}) = \mathbf{a}^T\mathbf{V}\,\mathbf{b} \tag{8.20}$$

If we transform $\mathbf{x}$ to two new vectors $\mathbf{y}_{\ell\times 1} = \mathbf{A}_{\ell\times n}\mathbf{x}_{n\times 1}$ and $\mathbf{z}_{m\times 1} = \mathbf{B}_{m\times n}\mathbf{x}_{n\times 1}$, then instead of a single covariance we have an $\ell \times m$ dimensional covariance matrix, denoted $\boldsymbol{\sigma}(\mathbf{y},\mathbf{z})$. Letting $\boldsymbol{\mu}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\mathbf{z}} = \mathbf{B}\boldsymbol{\mu}$, with $E(\mathbf{x}) = \boldsymbol{\mu}$, then $\boldsymbol{\sigma}(\mathbf{y},\mathbf{z})$ can be expressed in terms of $\mathbf{V}$, the covariance matrix of $\mathbf{x}$,

$$\begin{aligned}\boldsymbol{\sigma}(\mathbf{y},\mathbf{z}) &= \boldsymbol{\sigma}(\mathbf{A}\mathbf{x},\mathbf{B}\mathbf{x}) \\ &= E\left[(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})^T\right] \\ &= E\left[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{B}^T\right] \\ &= \mathbf{A}\mathbf{V}\,\mathbf{B}^T \end{aligned} \tag{8.21a}$$

In particular, the covariance matrix for $\mathbf{y} = \mathbf{A}\mathbf{x}$ is

$$\boldsymbol{\sigma}(\mathbf{y},\mathbf{y}) = \mathbf{A}\mathbf{V}\,\mathbf{A}^T \tag{8.21b}$$

so that the covariance between y_i and y_j is given by the ijth element of the matrix product $\mathbf{A}\mathbf{V}\mathbf{A}^T$.

Finally, note that if $\mathbf{x}$ is a vector of random variables with expected value $\boldsymbol{\mu}$, then the expected value of the scalar quadratic product $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is

$$E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \tag{8.22}$$

where $\mathbf{V}$ is the covariance matrix for the elements of $\mathbf{x}$, and the **trace** of a square matrix, $\text{tr}(\mathbf{M}) = \sum M_{ii}$, is the sum of its diagonal values (Searle 1971).

THE MULTIVARIATE NORMAL DISTRIBUTION

As we have seen above, matrix notation provides a compact way to express vectors of random variables. We now consider the most commonly assumed distribution for such vectors, the multivariate analog of the normal distribution discussed in Chapter 2. Much of the theory for the evolution of quantitative traits is based on this distribution, which we hereafter denote as the MVN.

Consider the probability density function for n independent normal random variables, where x_i is normally distributed with mean μ_i and variance σ_i^2. In this case, because the variables are independent, the joint probability density function is simply the product of each univariate density,

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^{n} (2\pi)^{-1/2}\sigma_i^{-1}\exp\left(-\frac{(x_i-\mu_i)^2}{2\,\sigma_i^2}\right) \\ &= (2\pi)^{-n/2}\left(\prod_{i=1}^{n}\sigma_i\right)^{-1}\exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{2\,\sigma_i^2}\right) \end{aligned} \tag{8.23}$$

We can express this equation more compactly in matrix form by defining the matrices

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n^2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Since $\mathbf{V}$ is diagonal, its determinant is simply the product of the diagonal elements

$$|\mathbf{V}| = \prod_{i=1}^{n} \sigma_i^2$$

Likewise, using quadratic products, note that

$$\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{x} - \boldsymbol{\mu})$$

Putting these together, Equation 8.23 can be rewritten as

$$p(\mathbf{x}) = (2\pi)^{-n/2} \, |\mathbf{V}|^{-1/2} \, \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{8.24}$$

We will also write this density as $p(\mathbf{x}, \boldsymbol{\mu}, \mathbf{V})$ when we wish to stress that it is a function of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$.

More generally, when the elements of $\mathbf{x}$ are correlated, Equation 8.24 gives the probability density function for a vector of multivariate normally distributed random variables, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. We denote this by

$$\mathbf{x} \sim \text{MVN}_n(\boldsymbol{\mu}, \mathbf{V})$$

where the subscript indicating the dimensionality of $\mathbf{x}$ is usually omitted. The multivariate normal distribution is also referred to as the **Gaussian distribution**.

Properties of the MVN

As in the case of its univariate counterpart, the MVN is expected to arise naturally when the quantities of interest result from a large number of underlying variables. Since this condition seems (at least at first glance) to describe many biological systems, the MVN is a natural starting point in biometrical analysis. Further details on the wide variety of applications of the MVN to multivariate statistics can be found in the introductory texts by Morrison (1976) and Johnson and Wichern (1988) and in the more advanced treatment by Anderson (1984). The MVN has a number of useful properties, which we summarize below.

1. *If* $\mathbf{x} \sim$ MVN, *then the distribution of any subset of the variables in* $\mathbf{x}$ *is also MVN*. For example, each x_i is normally distributed and each pair (x_i, x_j) is bivariate normally distributed.

2. *If* $\mathbf{x} \sim$ MVN, *then any linear combination of the elements of* $\mathbf{x}$ *is also MVN*. Specifically, if $\mathbf{x} \sim \text{MVN}_n(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{a}$ is a vector of constants, and $\mathbf{A}$ is a matrix of constants, then

$$\text{for} \quad \mathbf{y} = \mathbf{x} + \mathbf{a}, \qquad \mathbf{y} \sim \text{MVN}_n(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V}) \tag{8.25a}$$

$$\text{for} \quad y = \mathbf{a}^T\mathbf{x} = \sum_{k=1}^{n} a_i x_i, \qquad y \sim \text{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\mathbf{V}\,\mathbf{a}) \tag{8.25b}$$

$$\text{for} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \qquad \mathbf{y} \sim \text{MVN}_m\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}^T\mathbf{V}\mathbf{A}\right) \tag{8.25c}$$

3. *Conditional distributions associated with the MVN are also multivariate normal*. Consider the partitioning of $\mathbf{x}$ into two components, an $(m \times 1)$ column vector $\mathbf{x_1}$ and an $[(n-m)\times 1]$ column vector $\mathbf{x_2}$ of the remaining variables, e.g.,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix}$$

The mean vector and covariance matrix can be partitioned similarly as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \qquad \text{and} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V_{x_1x_1}} & \mathbf{V_{x_1x_2}} \\ \mathbf{V^T_{x_1x_2}} & \mathbf{V_{x_2x_2}} \end{pmatrix} \tag{8.26}$$

where the $m \times m$ and $(n-m) \times (n-m)$ matrices $\mathbf{V_{x_1x_1}}$ and $\mathbf{V_{x_2x_2}}$ are, respectively, the covariance matrices for $\mathbf{x_1}$ and $\mathbf{x_2}$, while the $m \times (n-m)$ matrix $\mathbf{V_{x_1x_2}}$ is the matrix of covariances between the elements of $\mathbf{x_1}$ and $\mathbf{x_2}$. If we condition on $\mathbf{x_2}$, the resulting conditional random variable $\mathbf{x_1}|\mathbf{x_2}$ is MVN with $(m \times 1)$ mean vector

$$\boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V^{-1}_{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) \tag{8.27}$$

and $(m \times m)$ covariance matrix

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V_{x_1x_1}} - \mathbf{V_{x_1x_2}}\mathbf{V^{-1}_{x_2x_2}}\mathbf{V^T_{x_1x_2}} \tag{8.28}$$

A proof can be found in most multivariate statistics texts, e.g., Morrison (1976).

4. *If* $\mathbf{x} \sim$ MVN, *the regression of any subset of* $\mathbf{x}$ *on another subset is linear and homoscedastic*. Rewriting Equation 8.27 in terms of a regression of the predicted value of the vector $\mathbf{x_1}$ given an observed value of the vector $\mathbf{x_2}$, we have

$$\mathbf{x_1} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V^{-1}_{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e} \tag{8.29a}$$

where

$$\mathbf{e} \sim \text{MVN}_m\left(\mathbf{0}, \mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}}\right) \tag{8.29b}$$

Example 6. Consider the regression of the phenotypic value of an offspring (z_o) on that of its parents (z_s and z_d for sire and dam, respectively). Assume that the joint distribution of z_o, z_s, and z_d is multivariate normal. For the simplest case of noninbred and unrelated parents, no epistasis or genotype-environment correlation, the covariance matrix can be obtained from the theory of correlation between relatives (Chapter 7), giving the joint distribution as

$$\begin{pmatrix} z_o \\ z_s \\ z_d \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_z^2 \begin{pmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{pmatrix} \right]$$

Let

$$\mathbf{x_1} = (z_o), \quad \mathbf{x_2} = \begin{pmatrix} z_s \\ z_d \end{pmatrix}$$

giving

$$\mathbf{V}_{\mathbf{x_1},\mathbf{x_1}} = \sigma_z^2, \quad \mathbf{V}_{\mathbf{x_1},\mathbf{x_2}} = \frac{h^2\sigma_z^2}{2} (1 \quad 1), \quad \mathbf{V}_{\mathbf{x_2},\mathbf{x_2}} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

From Equation 8.29a, the regression of offspring value on parental values is linear and homoscedastic with

$$\begin{aligned} z_o &= \mu_o + \frac{h^2\sigma_z^2}{2} (1 \quad 1) \sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_s - \mu_s \\ z_d - \mu_d \end{pmatrix} + e \\ &= \mu_o + \frac{h^2}{2} (z_s - \mu_s) + \frac{h^2}{2} (z_d - \mu_d) + e \end{aligned} \tag{8.30a}$$

where, from Equations 8.28 and 8.29b, the residual error is normally distributed with mean zero and variance

$$\begin{aligned} \sigma_e^2 &= \sigma_z^2 - \frac{h^2\sigma_z^2}{2} (1 \quad 1) \sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{h^2\sigma_z^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_z^2 \left(1 - \frac{h^4}{2} \right) \end{aligned} \tag{8.30b}$$

Example 7. The previous example dealt with the prediction of the phenotypic value of an offspring given parental phenotypic values. The same approach can be used to predict an offspring's additive genetic value A_o given knowledge of the parental values (A_s, A_d). Again assuming that the joint distribution is multivariate normal and that the parents are unrelated and noninbred, the joint distribution can be written as

$$\begin{pmatrix} A_o \\ A_s \\ A_d \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_A^2 \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \right]$$

Proceeding in the same fashion as in Example 6, the conditional distribution of offspring additive genetic values, given the parental values, is normal, so that the regression of offspring additive genetic value on parental value is linear and homoscedastic with

$$A_o = \mu_o + \frac{A_s - \mu_s}{2} + \frac{A_d - \mu_d}{2} + e \tag{8.31a}$$

and

$$e \sim \text{N}(0, \sigma_A^2/2) \tag{8.31b}$$

OVERVIEW OF LINEAR MODELS

Linear models form the backbone of most estimation procedures in quantitative genetics and will be extensively used throughout the rest of this book. They are generally structured such that a vector of observations of one variable (y) is modeled as a linear combination of other variables observed along with y. The remainder of this chapter introduces some of the basic tools and key concepts underlying the use of linear models. Advanced topics are examined in detail in Chapters 26 and 27, and further comments are given in Appendix 3.

In multiple regression, the commonest type of linear model, the predictor variables $x_1, \cdots, x_n$ represent observed values for n traits of interest. More generally, some or all of the predictor variables could be **indicator variables**, with values of 0 or 1 indicating whether an observation belongs in a particular category or grouping of interest. As an example, consider the half-sib design wherein each of p unrelated sires is mated at random to a number of unrelated dams and a single offspring is measured from each cross. The simplest model for this design is

$$y_{ij} = \mu + s_i + e_{ij}$$

where y_{ij} is the phenotype of the *j*th offspring from sire *i*, μ is the population mean, s_i is the **sire effect**, and e_{ij} is the residual error (the "noise" remaining in the data after the sire effect is removed). Although this is clearly a linear model, it differs significantly from the regression model described above in that while there are parameters to estimate (the sire effects s_i), the only measured values are the y_{ij}. Nevertheless, we can express this model in a form that is essentially identical to the standard regression model by using p indicator (i.e., zero or one) variables to classify the sires of the offspring. The resulting linear model becomes

$$y_{ij} = \mu + \sum_{k=1}^{p} s_k\, x_{ik} + e_{ij}$$

where

$$x_{ik} = \begin{cases} 1 & \text{if sire } k = i \\ 0 & \text{otherwise} \end{cases}$$

By the judicious use of indicator variables, an extremely wide class of problems can be handled by linear models. Models containing only indicator variables are usually termed ANOVA (**analysis of variance**) models, while regression usually refers to models in which predictor variables can take on a continuous range of values. Both procedures are special cases of the **general linear model** (GLM), wherein each observation (y) is assumed to be a linear function of p observed and/or indicator variables plus a residual error (e),

$$y_i = \sum_{k=1}^{p} \beta_k\, x_{ik} + e_i \tag{8.32a}$$

where $x_{i1}, \cdots, x_{ip}$ are the values of the p predictor variables for the *i*th individual. For a vector of n observations, the GLM can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{8.32b}$$

where the **design** or **incidence matrix X** is $n \times p$, and $\mathbf{e}$ is the vector of residual errors. It is important to note that $\mathbf{y}$ and $\mathbf{X}$ contain the observed values, while $\boldsymbol{\beta}$ is a vector of parameters (usually called **factors** or **effects**) to be estimated.

Example 8. Suppose that three different sires used in the above half-sib design have two, one, and three offspring, respectively. This can be expressed in GLM form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ s_1 \\ s_2 \\ s_3 \end{pmatrix}, \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{31} \\ e_{32} \\ e_{33} \end{pmatrix}$$

Likewise, the multiple regression

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j \, x_{ij} + e_i$$

can be written in GLM form with

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

Ordinary Least Squares

Estimates of the vector $\boldsymbol{\beta}$ for the general linear model are usually obtained by the method of least-squares, which uses the observations $\mathbf{y}$ and $\mathbf{X}$ and makes special assumptions about the covariance structure of the vector of residual errors $\mathbf{e}$. The method of **ordinary least squares** assumes that the residual errors are homoscedastic and uncorrelated, i.e., $\sigma^2(e_i) = \sigma_e^2$ for all i, and $\sigma(e_i, e_j) = 0$ for $i \neq j$.

Let $\mathbf{b}$ be an estimate of $\boldsymbol{\beta}$, and denote the vector of y values predicted from this estimate by $\widehat{\mathbf{y}} = \mathbf{Xb}$, so that the resulting vector of residual errors is

$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{Xb}$$

The ordinary least-squares (OLS) estimate of $\boldsymbol{\beta}$ is the $\mathbf{b}$ vector that minimizes the residual sum of squares,

$$\sum_{i=1}^{n} \widehat{e}_i^{\,2} = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{Xb})^T (\mathbf{y} - \mathbf{Xb})$$

Taking derivatives, it can be shown that our desired estimate satisfies

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8.33a}$$

Under the assumption that the residual errors are uncorrelated and homoscedastic (i.e., the covariance matrix of the residuals is $\sigma_e^2 \cdot \mathbf{I}$), the covariance matrix of the elements of $\mathbf{b}$ is

$$\mathbf{V_b} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2 \tag{8.33b}$$

Hence, the OLS estimator of β_i is the ith element of the column vector $\mathbf{b}$, while the variance of this estimator is the ith diagonal element of the matrix $\mathbf{V_b}$. Likewise, the covariance of this estimator with the OLS estimator for β_j is the ijth element of $\mathbf{V_b}$.

If the residuals follow a multivariate normal distribution with $\mathbf{e} \sim$ MVN($\mathbf{0}$, $\sigma_e^2 \cdot \mathbf{I}$), the OLS estimate is also the maximum-likelihood estimate. If $\mathbf{X}^T\mathbf{X}$ is singular, Equations 8.33a,b still hold when a generalized inverse is used, although only certain linear combinations of fixed factors can be estimated (see Appendix 3 for details).

Example 9. Consider a univariate regression where the predictor and response variable both have expected mean zero, so that the regression passes through the origin. The appropriate model becomes

$$y_i = \beta\, x_i + e_i$$

With observations on n individuals, this relationship can be written in GLM form with $\boldsymbol{\beta} = \beta$ and design matrix $\mathbf{X} = (x_1,\, x_2,\, \cdots x_n)^T$, implying

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^n x_i^2 \qquad \text{and} \qquad \mathbf{X}^T\mathbf{y} = \sum_{i=1}^n x_i\, y_i$$

Applying Equations 8.33a,b gives the OLS estimate of β and its sample variance (assuming the covariance matrix of $\mathbf{e}$ is $\mathbf{I} \cdot \sigma_e^2$) as

$$b = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \frac{\sum x_i\, y_i}{\sum x_i^2}, \qquad \sigma^2(b) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma_e^2 = \frac{\sigma_e^2}{\sum x_i^2}$$

This estimate of β differs from the standard univariate regression slope (Equation 3.14b) where the intercept value is not assumed to be equal to zero.

Example 10. Recall from Equation 8.10b that the vector of partial regression coefficients for a multivariate regression is defined to be $\mathbf{b} = \mathbf{V}^{-1}\,\mathbf{c}$ (where $\mathbf{V}$ is the estimated covariance matrix, and $\mathbf{c}$ is the vector of estimated covariances between $\mathbf{y}$ and $\mathbf{z}$). Here we show that this expression is equivalent to the OLS estimator $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Using the notation from Example 8, for the ith individual we observe y_i and the values of p predictor variables, $z_{i1}, \cdots, z_{ip}$. Since the regression satisfies $\bar{y} = \alpha + \beta_1 \bar{z}_1 + \cdots + \beta_p \bar{z}_p$, subtracting the mean from each observation removes the intercept, with

$$y_i^* = (y_i - \bar{y}) = \beta_1(z_{i1} - \bar{z}_1) + \cdots + \beta_p(z_{ip} - \bar{z}_p) + e_i$$

For n observations, the resulting linear model $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ has

$$\mathbf{y}^* = \begin{pmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} (z_{11} - \bar{z}_1) & \cdots & (z_{1p} - \bar{z}_p) \\ \vdots & \ddots & \vdots \\ (z_{n1} - \bar{z}_1) & \cdots & (z_{np} - \bar{z}_p) \end{pmatrix}$$

where z_{ij} is the value of character j in the *i*th individual. Partitioning the design matrix $\mathbf{X}$ into p column vectors corresponding to the n observations on each of the p predictor variables gives

$$\mathbf{X} = (\, \mathbf{x}_1, \quad \cdots, \quad \mathbf{x}_p \,) \quad \text{where} \quad \mathbf{x}_j = \begin{pmatrix} z_{1j} - \bar{z}_j \\ z_{2j} - \bar{z}_j \\ \vdots \\ z_{nj} - \bar{z}_j \end{pmatrix}$$

giving the *j*th element of the vector $\mathbf{X}^T\mathbf{y}^*$ as

$$\left(\mathbf{X}^T\mathbf{y}^*\right)_j = \mathbf{x}_j^T\mathbf{y}^* = \sum_{i=1}^{n}(y_i - \overline{y})(z_{ij} - \bar{z}_j) = (n-1)\text{Cov}(y, z_j)$$

and implying $\mathbf{X}^T\mathbf{y}^* = (n-1)\,\mathbf{c}$. Likewise, the *jk*th element of $\mathbf{X}^T\mathbf{X}$ is

$$\mathbf{x}_j^T\mathbf{x}_k = \sum_{i=1}^{n}(z_{ij} - \bar{z}_j)(z_{ik} - \bar{z}_k) = (n-1)\text{Cov}(z_j, z_k)$$

implying $\mathbf{X}^T\mathbf{X} = (n-1)\mathbf{V}$. Putting these results together gives

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}^* = \mathbf{V}^{-1}\mathbf{c}$$

showing that Equation 8.10b does indeed give the OLS estimates of the partial regression coefficients.

Generalized Least Squares

Under OLS, the unweighted sum of squared residuals is minimized. However, if some residuals are inherently more variable than others (have a higher variance), less weight should be assigned to the more variable data. Correlations between residuals can also influence the weight that should be assigned to each individual, as the data are not independent. Thus, if the residual errors are heteroscedastic and/or correlated, ordinary least-squares estimates of regression parameters and standard errors of these estimates are potentially biased.

A more general approach to regression analysis expresses the covariance matrix of the vector of residuals as $\sigma_e^2\,\mathbf{R}$, with $\sigma(e_i, e_j) = R_{ij}\sigma_e^2$. Lack of independence between residuals is indicated by the presence of nonzero off-diagonal elements in $\mathbf{R}$, while heteroscedasticity is indicated by differences in the diagonal elements of $\mathbf{R}$. **Generalized** (or **weighted**) **least squares** (GLS) takes these complications into account. As shown in Appendix 3, if the linear model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{with } \mathbf{e} \sim (0, \mathbf{R}\,\sigma_e^2)$$

the GLS estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \tag{8.34}$$

(Aitken 1935). The covariance matrix for the GLS estimates is

$$\mathbf{V}_{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\sigma_e^2 \tag{8.35}$$

If residuals are independent and homoscedastic, $\mathbf{R} = \mathbf{I}$, and GLS estimates are the same as OLS estimates. If $\mathbf{e} \sim$ MVN($\mathbf{0}$, $\mathbf{R}\,\sigma_e^2$), the GLS estimate of $\boldsymbol{\beta}$ is also the maximum-likelihood estimate.

Example 11. A common situation requiring weighted least-squares analysis occurs when residuals are independent but heteroscedastic with $\sigma^2(e_i) = \sigma_e^2/w_i$, where w_i are known positive constants. For example, if each observation y_i is the mean of n_i independent observations (each with uncorrelated residuals with variance σ_e^2), then $\sigma^2(e_i) = \sigma_e^2/n_i$, and hence $w_i = n_i$. Here

$$\mathbf{R} = \text{Diag}(w_1^{-1}, w_2^{-1}, \ldots, w_n^{-1})$$

where Diag denotes a diagonal matrix, giving

$$\mathbf{R}^{-1} = \text{Diag}(w_1, w_2, \ldots, w_n)$$

With this residual variance structure, consider the weighted least-squares estimate for the simple univariate regression model $y = \alpha + \beta\,x + e$. In GLM form,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Define the following weighted means and cross products,

$$w = \sum_{i=1}^{n} w_i, \quad \overline{x}_w = \sum_{i=1}^{n} \frac{w_i x_i}{w}, \quad \overline{x^2}_w = \sum_{i=1}^{n} \frac{w_i x_i^2}{w}$$

$$\overline{y}_w = \sum_{i=1}^{n} \frac{w_i y_i}{w}, \quad \overline{xy}_w = \sum_{i=1}^{n} \frac{w_i x_i \, y_i}{w}$$

With these definitions, matrix multiplication and a little simplification give

$$\mathbf{X}^T \mathbf{R}^{-1} \mathbf{y} = w \begin{pmatrix} \overline{y}_w \\ \\ \overline{xy}_w \end{pmatrix} \qquad \text{and} \qquad \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} = w \begin{pmatrix} 1 & \overline{x}_w \\ \\ \overline{x}_w & \overline{x^2}_w \end{pmatrix}$$

Applying Equation 8.34, the GLS estimates of α and β are

$$a = \overline{y}_w - b\overline{x}_w \tag{8.36a}$$

$$b = \frac{\overline{xy}_w - \overline{x}_w \, \overline{y}_w}{\overline{x^2}_w - \overline{x}_w^2} \tag{8.36b}$$

If all weights are equal ($w_i = c$), these expressions reduce to the standard (OLS) least-squares estimators given by Equation 3.14. Applying Equation 8.35, the sampling variances and covariance for these estimates are

$$\sigma^2(a) = \frac{\sigma_e^2 \cdot \overline{x^2}_w}{w \, (\, \overline{x^2}_w - \overline{x}_w^2 \,)} \tag{8.37a}$$

$$\sigma^2(b) = \frac{\sigma_e^2}{w \, (\, \overline{x^2}_w - \overline{x}_w^2 \,)} \tag{8.37b}$$

$$\sigma(a, b) = \frac{-\sigma_e^2 \, \overline{x}_w}{w \, (\, \overline{x^2}_w - \overline{x}_w^2 \,)} \tag{8.37c}$$