Lecture 2
Linear Algebra and Linear Models

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.

Notes from a short course taught Jan-Feb 2012 at University of Uppsala

Much of quantitative-genetic analysis is based on models in which response variables are linear functions of two or more explanatory (or predictor) variables. For example, we routinely express 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. Such **linear models** form the backbone of parameter estimation in quantitative genetics. Linear (or matrix) algebra provides the necessary machinery for the analysis of linear models, and we start by reviewing some of its basic concepts.

ELEMENTARY MATRIX ALGEBRA

Basic Matrix Notation

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

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

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, **D** has three rows and two columns, and is thus a 3×2 matrix. An $r \times 1$ matrix, such as **a**, is a **column vector**, while a $1 \times c$ matrix, such as **b**, is a **row vector**. A matrix in which the number of rows equals the number of columns, such as **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

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

Alternatively, we could partition $\mathbf{C}$ into a single row vector whose elements are themselves column vectors,

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

or $\mathbf{C}$ could be written as a column vector whose elements are row vectors,

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

Addition and Subtraction

Addition and subtraction of matrices is straightforward. To form a new matrix $\mathbf{A}+\mathbf{B} = \mathbf{C}$, $\mathbf{A}$ and $\mathbf{B}$ 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 considerably 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

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{2.1}$$

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 (pre)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

$$L_{ij} = \mathbf{m_i} \cdot \mathbf{n_j} = \sum_{k=1}^{c} M_{ik} N_{kj} \tag{2.2a}$$

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

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1} \cdot \mathbf{n_1} & \mathbf{m_1} \cdot \mathbf{n_2} & \cdots & \mathbf{m_1} \cdot \mathbf{n_b} \\ \mathbf{m_2} \cdot \mathbf{n_1} & \mathbf{m_2} \cdot \mathbf{n_2} & \cdots & \mathbf{m_2} \cdot \mathbf{n_b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_r} \cdot \mathbf{n_1} & \mathbf{m_r} \cdot \mathbf{n_2} & \cdots & \mathbf{m_r} \cdot \mathbf{n_b} \end{pmatrix} \tag{2.2b}$$

As an example, suppose

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

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

$$\mathbf{m_1} = \begin{pmatrix} 3 & 1 & 2 \end{pmatrix}, \quad \mathbf{m_2} = \begin{pmatrix} 2 & 5 & 4 \end{pmatrix}, \quad \mathbf{m_3} = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix}$$

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 2.2b, 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 $\mathbf{N}$,

$$\begin{aligned} L_{11} = \mathbf{m_1} \cdot \mathbf{n_1} &= \begin{pmatrix} 3 & 1 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix} = \sum_{k=1}^{3} M_{1k} N_{k1} \\ &= M_{11}N_{11} + M_{12}N_{21} + M_{13}N_{31} = (3 \times 4) + (1 \times 1) + (2 \times 3) = 19 \end{aligned}$$

Computing the other elements gives

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

A second example is to consider the following system of equations, which arises in obtaining the least-squares solution for the β in $y = \mu + \beta_1 x_1 + \cdots \beta_n x_n$

$$\begin{array}{lllll} \sigma(y,z_1) = & \beta_1\sigma^2(z_1) & +\beta_2\sigma(z_1,z_2) & + \cdots & +\beta_n\sigma(z_1,z_n) \\ \sigma(y,z_2) = & \beta_1\sigma(z_1,z_2) & +\beta_2\sigma^2(z_2) & + \cdots & +\beta_n\sigma(z_2,z_n) \\ \quad\vdots & \quad\vdots & \quad\vdots & \ddots & \quad\vdots \\ \sigma(y,z_n) = & \beta_1\sigma(z_1,z_n) & +\beta_2\sigma(z_2,z_n) & + \cdots & +\beta_n\sigma^2(z_n) \end{array} \tag{2.3a}$$

The above set of equations (where the covariances are known and we wish to solve for the β_i) can be much more compactly written in matrix notation as

$$\begin{pmatrix} \sigma^2(z_1) & \sigma(z_1,z_2) & \dots & \sigma(z_1,z_n) \\ \sigma(z_1,z_2) & \sigma^2(z_2) & \dots & \sigma(z_2,z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(z_1,z_n) & \sigma(z_2,z_n) & \dots & \sigma^2(z_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sigma(y,z_1) \\ \sigma(y,z_2) \\ \vdots \\ \sigma(y,z_n) \end{pmatrix} \tag{2.3b}$$

Letting $\mathbf{V}$ be the **variance-covariance matrix**, where $V_{ij} = \sigma(x_i, x_j)$, $\boldsymbol{\beta}$ the vector of the β's and $\mathbf{c}$ the vector of variance matrices between y and the x_i, this becomes

$$\mathbf{V}\boldsymbol{\beta} = \mathbf{c} \tag{2.3c}$$

Dimensional Properties and Matrix Multiplication

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 above 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 $\mathbf{A}^T$ (the notation $\mathbf{A}'$ is also widely used), and is obtained simply by switching rows and columns of the original matrix. Thus $A^T_{ij} = A_{ji}$. 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

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{2.4a}$$

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

$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T \tag{2.4b}$$

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

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{2.5a}$$

Note that since the sum is a scaler, it equals its transpose, so that $\mathbf{a}^T\mathbf{b}^T = (\mathbf{a}^T\mathbf{b})^T = \mathbf{b}^T\mathbf{a}$. Likewise, the outer product yields the $n \times n$ matrix

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix} = \mathbf{ab}^T = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix} \tag{2.5b}$$

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

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \tag{2.6}$$

where

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad I_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{2.7}$$

is the **identity matrix**, 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

$$\mathbf{A} = \mathbf{I}\mathbf{A} = \mathbf{A}\mathbf{I} \tag{2.8}$$

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

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{2.9}$$

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 $\mathbf{A}^{-1}$,

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{2.10a}$$

When $\mathbf{A}$ is either singular or nonsquare, solutions for $\mathbf{x}$ can still be obtained using **generalized inverses** in place of $\mathbf{A}^{-1}$, but such solutions are not unique, applying instead to certain linear combinations of the elements of $\mathbf{x}$ (see Lynch and Walsh Appendix 3 for details.) Recalling Equation 2.3, the solution of the multiple regression equation can be expressed as

$$\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c} \tag{2.11}$$

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. Likewise, for any 2×2 matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \text{then} \qquad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{2.12}$$

If $ad = bc$, the **determinant** of $\mathbf{A}$, is zero the inverse does not exist, as division by zero is undefined. For any square matrices, the *inverse exists if and only if the determinant is nonzero.*

As an example of using inverses, 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 2.3, 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 2.12 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 2.11, 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

$$\beta_1 = \frac{1}{1-\rho_{12}^2} \left[\frac{\sigma(y, z_1)}{\sigma^2(z_1)} - \rho_{12}\, \frac{\sigma(y, z_2)}{\sigma(z_1)\sigma(z_2)} \right] \tag{2.13a}$$

and

$$\beta_2 = \frac{1}{1-\rho_{12}^2} \left[\frac{\sigma(y, z_2)}{\sigma^2(z_2)} - \rho_{12}\, \frac{\sigma(y, z_1)}{\sigma(z_1)\sigma(z_2)} \right] \tag{2.13b}$$

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,

$$\beta_1 = \frac{\sigma(y, z_1)}{\sigma^2(z_1)} \quad \text{and} \quad \beta_2 = \frac{\sigma(y, z_2)}{\sigma^2(z_2)} \tag{2.13c}$$

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

$$\mathbf{Y}_{m\times\ell} = \mathbf{A}_{m\times n}\mathbf{X}_{n\times k}\mathbf{B}_{k\times\ell} \tag{2.14}$$

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

$$E(\mathbf{X}+\mathbf{Z}) = E(\mathbf{X}) + E(\mathbf{Z}) \tag{2.15}$$

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 2.14 is

$$E(\mathbf{Y}) = E(\mathbf{AXB}) = \mathbf{A}E(\mathbf{X})\mathbf{B} \tag{2.16a}$$

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

$$E(\mathbf{y}) = E(\mathbf{Xb}) = E(\mathbf{X})\mathbf{b} \tag{2.16b}$$

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

$$E(y) = E(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T E(\mathbf{x}) \tag{2.16c}$$

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,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \tag{2.17}$$

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

$$\mathbf{b}^T\mathbf{A}\mathbf{a} = \left(\mathbf{b}^T\mathbf{A}\mathbf{a}\right)^T = \mathbf{a}^T\mathbf{A}^T\mathbf{b} \tag{2.18}$$

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 $\mathbf{x}$,

$$\begin{aligned}\sigma^2\left(\mathbf{c}^T\mathbf{x}\right) &= \sigma^2\left(\sum_{i=1}^n c_i x_i\right) = \sigma\left(\sum_{i=1}^n c_i\, x_i\,, \sum_{j=1}^n c_j\, x_j\right) \\ &= \sum_{i=1}^n\sum_{j=1}^n \sigma\left(c_i\, x_i, c_j\, x_j\right) = \sum_{i=1}^n\sum_{j=1}^n c_i\, c_j\, \sigma\left(x_i, x_j\right) \\ &= \mathbf{c}^T\mathbf{V}\,\mathbf{c} \end{aligned} \tag{2.19}$$

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

$$\sigma(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{x}) = \mathbf{a}^T\mathbf{V}\,\mathbf{b} \tag{2.20}$$

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 $\mathbf{x}$,

$$\begin{aligned}\boldsymbol{\sigma}(\mathbf{y}, \mathbf{z}) &= \boldsymbol{\sigma}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}) \\ &= E\left[(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})^T\right] \\ &= E\left[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{B}^T\right] \\ &= \mathbf{A}\mathbf{V}\,\mathbf{B}^T \end{aligned} \tag{2.21a}$$

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

$$\boldsymbol{\sigma}(\mathbf{y}, \mathbf{y}) = \mathbf{A}\mathbf{V}\,\mathbf{A}^T \tag{2.21b}$$

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

$$E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \tag{2.22}$$

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

THE MULTIVARIATE NORMAL DISTRIBUTION

Much of the theory for the evolution of quantitative traits is based on this distribution, which we hereafter denote as the MVN. To motivate the MVN, first 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,

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^n (2\pi)^{-1/2}\sigma_i^{-1}\exp\left(-\frac{(x_i-\mu_i)^2}{2\,\sigma_i^2}\right) \\ &= (2\pi)^{-n/2}\left(\prod_{i=1}^n \sigma_i\right)^{-1}\exp\left(-\sum_{i=1}^n\frac{(x_i-\mu_i)^2}{2\,\sigma_i^2}\right)\end{aligned} \tag{2.23}$$

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 2.23 can be rewritten as

$$p(\mathbf{x}) = (2\pi)^{-n/2} \, |\mathbf{V}|^{-1/2} \, \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{2.24}$$

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 2.24 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 the sum of a large number of underlying variables. 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

$$\text{for} \quad \mathbf{y} = \mathbf{x} + \mathbf{a}, \qquad \mathbf{y} \sim \text{MVN}_n(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V}) \tag{2.25a}$$

$$\text{for} \quad y = \mathbf{a}^T\mathbf{x} = \sum_{k=1}^{n} a_i x_i, \qquad y \sim \text{N}(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \mathbf{V}\, \mathbf{a}) \tag{2.25b}$$

$$\text{for} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \qquad \mathbf{y} \sim \text{MVN}_m\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}^T \mathbf{V} \mathbf{A}\right) \tag{2.25c}$$

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

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \qquad \text{and} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V_{x_1x_1}} & \mathbf{V_{x_1x_2}} \\ \mathbf{V}^T_{\mathbf{x_1x_2}} & \mathbf{V_{x_2x_2}} \end{pmatrix} \tag{2.26}$$

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

$$\boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) \tag{2.27}$$

and $(m \times m)$ covariance matrix

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V_{x_1x_1}} - \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}\mathbf{V}^T_{\mathbf{x_1x_2}} \tag{2.28}$$

4. *If* $\mathbf{x} \sim$ MVN, *the regression of any subset of* $\mathbf{x}$ *on another subset is linear and homoscedastic.* Rewriting Equation 2.27 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

$$\mathbf{x_1} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e} \tag{2.29a}$$

where

$$\mathbf{e} \sim \text{MVN}_m\left(\mathbf{0}, \mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}}\right) \tag{2.29b}$$

Example: The Regression of Offspring Value on Parental Value

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 (Lecture 3), giving the joint distribution as

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

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 2.29a, the regression of offspring value on parental values is linear and homoscedastic with

$$\begin{aligned} z_o &= \mu_o + \frac{h^2\sigma_z^2}{2}\,(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_s - \mu_s \\ z_d - \mu_d \end{pmatrix} + e \\ &= \mu_o + \frac{h^2}{2}\,(z_s - \mu_s) + \frac{h^2}{2}\,(z_d - \mu_d) + e \end{aligned} \tag{2.30a}$$

where, from Equations 2.28 and 2.29b, the residual error is normally distributed with mean zero and variance

$$\begin{aligned} \sigma_e^2 &= \sigma_z^2 - \frac{h^2\sigma_z^2}{2}\,(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{h^2\sigma_z^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_z^2 \left(1 - \frac{h^4}{2}\right) \end{aligned} \tag{2.30b}$$

Example: Regression of Offspring Breeding Value on Parental Breeding Values

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 the previous example, 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

$$A_o = \mu_o + \frac{A_s - \mu_s}{2} + \frac{A_d - \mu_d}{2} + e \tag{2.31a}$$

and

$$e \sim \text{N}(0, \sigma_A^2/2) \tag{2.31b}$$

OVERVIEW OF LINEAR MODELS

As mentioned, linear models form the backbone of most estimation procedures in quantitative genetics. 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.

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, e.g.,

$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x + e$$

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, here the within half-sib family variance). 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),

$$y_i = \sum_{k=1}^{p} \beta_k \, x_{ik} + e_i \tag{2.32a}$$

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.32b}$$

where the **design** or **incidence matrix** $\mathbf{X}$ 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.

Examples of GLMs

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.,

$$\sigma^2(e_i, e_j) = \begin{cases} \sigma_e^2 & \text{for } i = j \;\text{ (common variance)} \\ 0 & \text{for } i \neq j \;\text{ (uncorrelated)} \end{cases}$$

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

$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

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

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

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

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.33a}$$

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

$$\mathbf{V_b} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2 \tag{2.33b}$$

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 $\mathbf{V_b}$. Finally, we can estimate σ_e^2 by

$$\text{Var}(e) = \frac{\widehat{\mathbf{e}}^T \widehat{\mathbf{e}}}{n - \text{rank}(X)} \tag{2.33c}$$

Where the rank the number of independent columns of the design matrix $\mathbf{X}$. (This is also the number of non-zero eigenvalues of the matrix $\mathbf{X}^T\mathbf{X}$.) If $\mathbf{X}$ is of **full-rank**, then rank($\mathbf{X}$) = p, the lenght of the vector $\mathbf{b}$ (the number of estimated parameters in the linear model).

If the residuals follow a multivariate normal distribution with $\mathbf{e} \sim \text{MVN}(\mathbf{0},\, \sigma_e^2 \cdot \mathbf{I}\,)$, the OLS estimate is also the maximum-likelihood estimate. If $\mathbf{X}^T\mathbf{X}$ is singular, Equations 2.33a,b still hold when a generalized inverse is used, although only certain linear combinations of fixed factors can be estimated (see LW Appendix 3 for details).

Example: GLM Solution for Regression Through the Origin

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 2.33a,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 where the intercept value is not assumed to be equal to zero.

Example: Partial (or Multiple) Regression

Recall from Equations 2.3a-2.3c 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 $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The data are as follows: for the ith individual we observe y_i and the values of p predictor variables, $z_{i1}, \cdots, z_{ip}$. Since the regression satisfies $\overline{y} = \alpha + \beta_1\bar{z}_1 + \cdots + \beta_p\bar{z}_p$, subtracting the mean from each observation removes the intercept, with

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

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 ith 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

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

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 2.3c does indeed give the OLS estimates of the partial regression coefficients.

Polynomial Regressions and Interaction Effects

The general linear model is extremely flexible, covering many models that at first appear nonlinear. The key here is that linearity refers to the **parameters** to be estimated, not the data being observed. For example, consider the **quadratic regression** of y on a single predictor variable x. The *i*th observation is assumed to be of the form

$$y_i = \alpha + \beta_1\,x_i + \beta_2\,x_i^2 + e_i$$

While quadratic terms appear, these are in the *observed* variables, *not* the parameters to be estimated $(\alpha, \beta_1, \beta_2)$. Expressed in GLM form,

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

This sort of reasoning extends to polynomial regressions of any order and to just about any transformation of the observed variables, e.g., we could consider $\ln x$ or e^{-x}. Likewise the GLM can be used to account for **interaction effects**. For example, consider the model

$$y_i = \alpha + \beta_1\,x_{i1} + \beta_2\,x_{i2} + \beta_3\,x_{i1}\cdot x_{i2} + e_i$$

which in matrix form is

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

with

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}\cdot x_{12} \\ 1 & x_{21} & x_{22} & x_{21}\cdot x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}\cdot x_{n2} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

The β_j for this model have a slightly different interpretation from partial regression coefficients. Here, with x_1 held constant, a unit change in x_2 changes the mean value of y by $\beta_2 + \beta_3\cdot x_1$. Likewise with x_2 held constant, a unit change in x_1 changes the mean value by $\beta_1 + \beta_3\cdot x_2$.

Fixed vs. Random Effects

Linear models are based upon sets of variables that classify individuals into various groupings, often referred to as **factors** or **effects**. For example, suppose we have information on the sex of an individual, which diet it was raised on, and its age. These are the three factors for this analysis, and we can ask how much of the variation in a response variable is attributable to each factor individually and to interactions between the various factors (i.e., a sex-diet interaction not predicted by sex or diet alone).

There are two fundamentally different types of factors — **fixed** and **random.** The distinction between fixed and random effects is usually straightforward, but at times it can be extremely subtle. Consider the simple model in which a single factor takes on k discrete values. Then, $y_{ij} = \mu + \beta_j + e_{ij}$, where $1 \leq j \leq k$, so that y_{ij} is the ith observation at the jth value of the factor. Whether the factor is treated as fixed or random depends on how the k values of the factor are chosen. Under a random effects model, the k values are drawn at random from a probability distribution with mean zero and unknown variance. In this case, our interest is usually in estimating the variance of this distribution. Conversely, we may decide on a fixed set of factor values in advance (such as males versus females, or k distinct diets). These are **fixed effects,** as the factor values are assumed to be fixed in advance of the analysis, i.e., there is no variance associated with their choice. Because the distinction between fixed and random effects lies in how we treat the underlying sampling distribution of factor values, situations arise in which one investigator assumes fixed factor values, while another regards them as random. For example, when k diets are assayed, diet is a random effect if we regard the treatments as a random sample from some universe of possible diets. Conversely, if we are interested in these *k particular* diets, then diet is a fixed effect.

General linear models are used for three rather different classes of estimation problems: estimating fixed effects, estimating the variance of random effects, and predicting random effects. When dealing with random effects, the literature usually refers to **predicting**, rather than estimating their values, to distinguish this from the estimation of fixed effects. Estimation of variance components is most simply done using ANOVA methods, although the method of REML (restricted maximum likelihood) is a more advanced and flexible method of estimation of variance components. The ascertainment of values of random variables is of great concern in animal breeding, where the prediction of breeding values is a key issue. The most powerful approach here is BLUP (best linear unbiased prediction), which allows for **mixed** models with both fixed and random effects.

Example: Fixed vs. Random Effects in the Sire Model

Consider the sire model, $z_{ij} = \mu + s_i + e_{ij}$, that we used to motivate the concept of indicator variables. If five sires are measured, we could regard these individuals as a random sample from a larger population of interest. With this interpretation, the sire effects are random effects drawn from a distribution with mean 0 (a non-zero mean is absorbed by the population mean μ included in the model) and variance σ_s^2. Estimation of the sire variance is of interest as a means of estimating the additive genetic variance, since (ignoring complicating factors) $\sigma_s^2 = \sigma_A^2/4$ (Lecture 3). In addition to estimating the additive genetic variance of the population from which the five sires were drawn, we may also wish to predict the specific sire values $s_1, \cdots s_5$, as for example, when the individuals with the highest sire values are being sought for future breeding. Conversely, if these five males are all the sires in our breeding program and we do not plan to introduce any outside sires, we could regard this group of five as the entire population of sires. In this case, sires are fixed effects, as there is no population that we wish to draw inferences on other than these five sires.

GENERLIZED 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. 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

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \tag{2.34}$$

The covariance matrix for the GLS estimates is

$$\mathbf{V_b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\sigma_e^2 \tag{2.35}$$

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: Weighted Least Squares

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 2.34, the GLS estimates of α and β are

$$a = \overline{y}_w - b\overline{x}_w \tag{2.36a}$$

$$b = \frac{\overline{xy}_w - \overline{x}_w\,\overline{y}_w}{\overline{x^2}_w - \overline{x}_w^2} \tag{2.36b}$$

If all weights are equal ($w_i = c$), these expressions reduce to the standard (OLS) least-squares estimators. Applying Equation 2.35, the sampling variances and covariance for these estimates are

$$\sigma^2(a) = \frac{\sigma_e^2 \cdot \overline{x^2}_w}{w\,(\,\overline{x^2}_w - \overline{x}_w^2\,)} \tag{2.37a}$$

$$\sigma^2(b) = \frac{\sigma_e^2}{w\,(\,\overline{x^2}_w - \overline{x}_w^2\,)} \tag{2.37b}$$

$$\sigma(a, b) = \frac{-\sigma_e^2\,\overline{x}_w}{w\,(\,\overline{x^2}_w - \overline{x}_w^2\,)} \tag{2.37c}$$

MODEL GOODNESS-OF-FIT AND HYPOTHESIS TESTING

Chi-square and F-distributions

How does one decide if a particular linear model provides a reasonable fit to a data set or whether additional explanatory variables are warranted? A closely related issue is the ideal number of parameters in our final model. While we can always improve the fit of a model by adding more parameters, a point of diminishing returns can quickly be reached. We start by reviewing two very useful distributions, the **Chi-square** (χ^2) and **F distributions,** which play a key role in resolving these issues.

If $x_1, \cdots, x_k$ are independent unit normals, i.e., $x_i \sim \mathrm{N}(0,1)$, the sum of the squared values of these

$$X = \sum_{i=1}^{k} x_i^2$$

follows a Chi-square distribution with k degrees of freedom, and we write $X \sim \chi^2_k$.

The F-distribution, which is indexed by two sets of degrees of freedom, is defined as the scaled ratio of two χ^2 random variables,

$$F_{k,m} = \frac{\sum_{i=1}^{k}\, x_i^2\,/\,k}{\sum_{i=1}^{m}\, y_i^2\,/\,m}$$

where x_i and y_i are independent unit normals.

Sums of Squares

Hypothesis tests of linear models are often based on various sums of squares associated with the model being considered. Under appropriate normality assumptions, these sums of squares are χ^2-distributed and ratios of them are F-distributed. Such tests can be quite involved, especially if we are testing components of fit in a complex model. Here we consider the simplest case — the fit of the total model to the data.

The **total sum of squares** SS_T for a linear model can be written as the sum of two components, the **error** (or **residual**) **sum of squares** (SS_E) and the **model** (or **regression**) **sum of squares** (SS_M),

$$\mathrm{SS}_T = \mathrm{SS}_M + \mathrm{SS}_E$$

The total sum of squares measures the total variability in the data, while the model sum of squares measures the amount of variation accounted for by the linear model. The fraction of total variance explained by the linear model is given by the **coefficient of determination**,

$$r^2 = \frac{\mathrm{SS}_M}{\mathrm{SS}_T} = 1 - \frac{\mathrm{SS}_E}{\mathrm{SS}_T} \tag{2.38}$$

We first saw this (in simpler form) in Lecture 1 as a measure of the fraction of total variation explained by a simple (univariate) linear regression.

Sums of squares have different forms under OLS and GLS. Under OLS, the residuals are assumed to be independent with common variance σ_e^2. In this case, each observation/residual is weighted equally, with

$$\mathrm{SS}_T = \sum_{i=1}^n (\, y_i - \overline{y}\,)^2 = \sum_{i=1}^n y_i^2 - \overline{y}^{\,2} = \sum_{i=1}^n y_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n y_i\right)^2$$

which can be expressed as a quadratic form of the vector of observations $\mathbf{y}$,

$$\mathrm{SS}_T = \mathbf{y}^T\mathbf{y} - \frac{1}{n}\,\mathbf{y}^T\mathbf{J}\mathbf{y} = \mathbf{y}^T\left(\mathbf{I} - \frac{1}{n}\,\mathbf{J}\right)\mathbf{y} \tag{2.39a}$$

where each element in the matrix $\mathbf{J}$ is 1. The other sums of squares can also be written as quadratic products of $\mathbf{y}$. Consider the error sum of squares,

$$\mathrm{SS}_E = \sum_{i=1}^n (\, y_i - \widehat{y}_i\,)^2 = \sum_{i=1}^n \widehat{e}_i^{\,2}$$

Since $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and $\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$, we have

$$\mathrm{SS}_E = \widehat{\mathbf{e}}^T\widehat{\mathbf{e}} \quad \text{where} \quad \widehat{\mathbf{e}} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \tag{2.39b}$$

Expanding this expression and noting that $\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \mathbf{I}$, this simplifies to

$$\mathrm{SS}_E = \mathbf{y}^T\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y} \tag{2.39c}$$

Finally,

$$\mathrm{SS}_M = \mathrm{SS}_T - \mathrm{SS}_E = \mathbf{y}^T \left(\mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T - \frac{1}{n} \mathbf{J} \right) \mathbf{y} \tag{2.39d}$$

Since these sums of squares are functions of random data, they are also themselves random values. The **observed sums of squares** are distributed about their expected values, which can be obtained by using Equation 2.22. The resulting **expected sums of squares** are functions of the unknown means and variances. By equating observed and expected sums of squares, we can often estimate the unknown parameters. This is the basic methods-of-moments procedure for estimating variance components using ANOVA.

The matrix expressions for the sums of squares under generalized least squares (GLS) are slightly different, as we have to first correct for heteroscedasticity and/or the lack of independence among the residuals. Assuming the covariance matrix of residuals is $\sigma_e^2\,\mathbf{R}$, then recalling Equation 2.34, the total sum of squares for GLS becomes

$$\begin{aligned} \mathrm{SS}_T &= \mathbf{y}^T \mathbf{R}^{-1/2} \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{R}^{-1/2} \mathbf{y} \\ &= \mathbf{y}^T \left(\mathbf{R}^{-1} - \frac{1}{n} \mathbf{R}^{-1/2} \mathbf{J} \mathbf{R}^{-1/2} \right) \mathbf{y} \end{aligned} \tag{2.40a}$$

Likewise, the error sum of squares becomes

$$\begin{aligned} \mathrm{SS}_E &= \widehat{\mathbf{e}}^T \mathbf{R}^{-1} \widehat{\mathbf{e}} \\ &= \mathbf{y}^T \left(\mathbf{R}^{-1} - \mathbf{R}^{-1} \mathbf{X} \left(\mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{R}^{-1} \right) \mathbf{y} \end{aligned} \tag{2.40b}$$

and the model sum of squares is

$$\mathrm{SS}_M = \mathbf{y}^T \left(\mathbf{R}^{-1} \mathbf{X} \left(\mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{R}^{-1} - \frac{1}{n} \mathbf{R}^{-1/2} \mathbf{J} \mathbf{R}^{-1/2} \right) \mathbf{y} \tag{2.40c}$$

If the residuals are multivariate-normally distributed with

$$\mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \sigma_e^2\,\mathbf{I}) \quad \text{for OLS}; \qquad \mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \sigma_e^2\,\mathbf{R}) \quad \text{for GLS}$$

then SS_E/σ_e^2, the sum of squared unit normals, is χ^2-distributed. In particular, with n observations and p estimated parameters,

$$\frac{\mathrm{SS}_E}{\sigma_e^2} \sim \chi^2_{n-p} \tag{2.41}$$

as a degree of freedom is lost for each estimated model parameter.

Hypothesis Testing

We finally have all the necessary machinery and definitions in place for hypothesis testing of linear models. Suppose we have n observations and wish to compare two linear models, a **full** model fitting p parameters and a **reduced** model which uses only a subset ($q < p$) of the parameters in the full model. Do the additional $p - q$ fitted parameters provide a significant increase in the amount of variation accounted for by the model? Letting SS_{E_f} and SS_{E_r} denote the appropriate (OLS or GLS) error sums of squares for the full and reduced models, the test statistic for a significant improvement is

$$\left(\frac{\text{SS}_{E_r} - \text{SS}_{E_f}}{p-q} \right) \Bigg/ \left(\frac{\text{SS}_{E_f}}{n-p} \right) = \left(\frac{n-p}{p-q} \right) \left(\frac{\text{SS}_{E_r}}{\text{SS}_{E_f}} - 1 \right) \tag{2.42}$$

This is distributed as $F_{p-q,n-p}$, allowing F-tests to be used to evalutate if adding the $p-q$ parameters to the model significantly improves the fit. For example, we can ask if a particular linear model accounts for a significant fraction of the variation in y by considering that model versus the simplest reduced model $y_i = \mu + e_i$. It is easily seen that the least-squares solution for μ is the unweighted mean $\overline{y}$ for OLS and the weighted mean for GLS, giving $\text{SS}_{E_r} = \text{SS}_T$. Since the number of parameters in the reduced model is $q = 1$, the test for whether a particular linear model accounts for a significant amount of the variation is

$$\left(\frac{n-p}{p-1} \right) \left(\frac{\text{SS}_T}{\text{SS}_{E_f}} - 1 \right) = \left(\frac{n-p}{p-1} \right) \left(\frac{r^2}{1-r^2} \right) \tag{2.43}$$

where r^2 is the coefficient of determination for the full model (Equation 2.38). This test statistic follows an $F_{p-1,n-p}$ distribution.

Table 2.1 summarizes the features of OLS and GLS linear models.

Table 2.1. Summary of useful results for the general linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, under OLS and GLS assumptions for the distribution of residuals. We have discussed GLS above under the special case where $\mathbf{V} = \sigma_e^2\,\mathbf{R}$, where $\mathbf{R}$ is a matrix of constants.

	Ordinary Least Squares, OLS	**Generalized Least Squares, GLS**
Assumed Distribution of Residuals:	$\mathbf{e} \sim (\mathbf{0}, \sigma_e^2\,\mathbf{I})$	$\mathbf{e} \sim (\mathbf{0}, \mathbf{V})$
Least-squares estimator of $\boldsymbol{\beta}$:	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Covariance matrix for $\widehat{\boldsymbol{\beta}}$:	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$:	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Covariance matrix for predicted values, $\widehat{\mathbf{y}}$:	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

Lecture 2 Problems

1. Place the following system of equations in matrix form and solve

$$\begin{aligned} 5x_1 + 6x_2 &= 3 \\ 3x_1 - 4x_2 &= -6 \end{aligned}$$

2. Use the definition of the inverse to show that $(\mathbf{A}^{-1})^T$ = $(\mathbf{A}^T)^{-1}$

3. Consider a quadratic regression forced through the origin, $y = \beta_1 x + \beta_2 x^2 + e$. Write this model in GLM matrix form. If we have n pairs of data, (y_i, x_i) obtain expressions for the OLS solutions for β_1 and β_2. What are the sampling variances for these estimates?

Solutions to Lecture 2 Problems

1.

$$\begin{pmatrix} 5 & 6 \\ 3 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -6 \end{pmatrix}$$

The solution is given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 3 & -4 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ -6 \end{pmatrix}$$

where (Equation 2.12)

$$\begin{pmatrix} 5 & 6 \\ 3 & -4 \end{pmatrix}^{-1} = \frac{1}{5(-4) - 6 \cdot 3} \begin{pmatrix} -4 & -6 \\ -3 & 5 \end{pmatrix} = \frac{1}{-38} \begin{pmatrix} -4 & -6 \\ -3 & 5 \end{pmatrix}$$

giving

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{-38} \begin{pmatrix} -4 & -6 \\ -3 & 5 \end{pmatrix} \begin{pmatrix} 3 \\ -6 \end{pmatrix} = \begin{pmatrix} -12/19 \\ 39/38 \end{pmatrix}$$

2. Noting that an inverse $\mathbf{B}^{-1}$ for the matrix $\mathbf{B}$ satisfies $\mathbf{B}^{-1}\mathbf{B} = \mathbf{B}\mathbf{B}^{-1} = \mathbf{I}$, to show that $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$, we need to show that

$$\mathbf{A}^T(\mathbf{A}^{-1})^T = (\mathbf{A}^{-1})^T\mathbf{A}^T = \mathbf{I}$$

Taking the transpose of $\mathbf{A}^T(\mathbf{A}^{-1})^T$ gives

$$(\mathbf{A}^T(\mathbf{A}^{-1})^T)^T = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Note that $\mathbf{I} = \mathbf{I}^T$. Likewise,

$$[(\mathbf{A}^{-1})^T\mathbf{A}^T]^T = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

3.

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

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1 & x_1^2 \\ \vdots & \vdots \\ x_n & x_n^2 \end{pmatrix}$$

Hence

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} x_1 & \cdots & x_n \\ x_1^2 & \cdots & x_n^2 \end{pmatrix} \begin{pmatrix} x_1 & x_1^2 \\ \vdots & \vdots \\ x_n & x_n^2 \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^n x_i^2 & \sum\limits_{i=1}^n x_i^3 \\ \sum\limits_{i=1}^n x_i^3 & \sum\limits_{i=1}^n x_i^4 \end{pmatrix}$$

where

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{\sum x_i^4 \cdot \sum x_i^2 - \left(\sum x_i^3\right)^2} \begin{pmatrix} \sum\limits_{i=1}^n x_i^4 & -\sum\limits_{i=1}^n x_i^3 \\ -\sum\limits_{i=1}^n x_i^3 & \sum\limits_{i=1}^n x_i^2 \end{pmatrix}$$

Likewise,

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum x_i y_i \\ \\ \sum x_i^2 y_i \end{pmatrix}$$

Hence, the OLS estimator is

$$\widehat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{\sum x_i^4 \cdot \sum x_i^2 - \left(\sum x_i^3\right)^2} \begin{pmatrix} \sum\limits_{i=1}^n x_i^4 & -\sum\limits_{i=1}^n x_i^3 \\ -\sum\limits_{i=1}^n x_i^3 & \sum\limits_{i=1}^n x_i^2 \end{pmatrix} \begin{pmatrix} \sum x_i y_i \\ \\ \sum x_i^2 y_i \end{pmatrix}$$

hence,

$$\widehat{\beta_1} = \frac{\sum x_i^4 \cdot \sum x_i\, y_i - \sum x_i^3 \cdot \sum x_i^2\, y_i}{\sum x_i^4 \cdot \sum x_i^2 - \left(\sum x_i^3\right)^2}$$

$$\widehat{\beta_2} = \frac{-\sum x_i^3 \cdot \sum x_i\, y_i + \sum x_i^2 \cdot \sum x_i^2\, y_i}{\sum x_i^4 \cdot \sum x_i^2 - \left(\sum x_i^3\right)^2}$$

The associated sampling variances are

$$Var(\mathbf{b}) = \begin{pmatrix} \sigma^2(\beta_1) & \sigma(\beta_1, \beta_2) \\ \sigma(\beta_1, \beta_2) & \sigma^2(\beta_2) \end{pmatrix} = Var(e)(\mathbf{X}^T\mathbf{X})^{-1}$$

where

$$Var(e) = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y_i})^2$$

and $(\mathbf{X}^T\mathbf{X})^{-1}$ is as above.