$$\mathbf{X}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 3 & 8 \\ 1 & 5 & 4 & 6 \end{pmatrix}$$

The transpose of a column vector is a row vector. The transpose of a product $(\mathbf{AB})'$ is the product of the transposes, in opposite order, so $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Suppose that $\mathbf{a}$ is an $r \times 1$ vector with elements $a_1, \ldots, a_r$. Then the product $\mathbf{a}'\mathbf{a}$ will be a 1×1 matrix or scalar, given by

$$\mathbf{a}'\mathbf{a} = a_1^2 + a_2^2 + \cdots + a_r^2 = \sum_{i=1}^{r} a_i^2 \tag{A.13}$$

Thus, $\mathbf{a}'\mathbf{a}$ provides a compact notation for the sum of the squares of the elements of a vector $\mathbf{a}$. The square root of this quantity $(\mathbf{a}'\mathbf{a})^{1/2}$ is called the *norm* or *length* of the vector $\mathbf{a}$. Similarly, if $\mathbf{a}$ and $\mathbf{b}$ are both $r \times 1$ vectors, then we obtain

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^{r} a_i b_i = \sum_{i=1}^{r} b_i a_i = \mathbf{b}'\mathbf{a}$$

The fact that $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}$ is often quite useful in manipulating the vectors used in regression calculations.

Another useful formula in regression calculations is obtained by applying the distributive law

$$(\mathbf{a} - \mathbf{b})'(\mathbf{a} - \mathbf{b}) = \mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - 2\mathbf{a}'\mathbf{b} \tag{A.14}$$

A.6.5 Inverse of a Matrix

For any real number $c \neq 0$, there is another number called the *inverse* of c, say d, such that the product $cd = 1$. For example, if $c = 3$, then $d = 1/c = 1/3$, and the inverse of 3 is 1/3. Similarly, the inverse of 1/3 is 3. The number 0 does not have an inverse because there is no other number d such that $0 \times d = 1$.

Square matrices can also have an inverse. We will say that the inverse of a matrix $\mathbf{C}$ is another matrix $\mathbf{D}$, such that $\mathbf{CD} = \mathbf{I}$, and we write $\mathbf{D} = \mathbf{C}^{-1}$. Not all square matrices have an inverse. The collection of matrices that have an inverse are called full rank, invertible, or nonsingular. A square matrix that is not invertible is of less than full rank, or singular. If a matrix has an inverse, it has a unique inverse.

The inverse is easy to compute only in special cases, and its computation in general can require a very tedious calculation that is best done on a computer. High-level matrix and statistical languages such as `Matlab`, `Maple`, `Mathematica` and `R` include functions for inverting matrices, or returning an appropriate message if the inverse does not exist.

The identity matrix **I** is its own inverse. If **C** is a diagonal matrix, say

$$\mathbf{C} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

then $\mathbf{C}^{-1}$ is the diagonal matrix

$$\mathbf{C} = \begin{pmatrix} \frac{1}{3} & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

as can be verified by direct multiplication. For any diagonal matrix with nonzero diagonal elements, the inverse is obtained by inverting the diagonal elements. If any of the diagonal elements are 0, then no inverse exists.

A.6.6 Orthogonality

Two vectors **a** and **b** of the same length are *orthogonal* if $\mathbf{a}'\mathbf{b} = 0$. An $r \times c$ matrix **Q** has *orthonormal columns* if its columns, viewed as a set of $c \leq r$ different $r \times 1$ vectors, are orthogonal and in addition have length 1. This is equivalent to requiring that $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, the $r \times r$ identity matrix. A square matrix **A** is *orthogonal* if $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}$, and so $\mathbf{A}^{-1} = \mathbf{A}'$. For example, the matrix

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

can be shown to be orthogonal by showing that $\mathbf{A}'\mathbf{A} = \mathbf{I}$, and therefore

$$\mathbf{A}^{-1} = \mathbf{A}' = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix}$$

A.6.7 Linear Dependence and Rank of a Matrix

Suppose we have a $n \times p$ matrix $\mathbf{X}$ with columns given by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_p$; we consider only the case $p \leq n$. We will say that $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are *linearly dependent* if we can find multipliers $a_1, \ldots, a_p$, not all of which are 0, such that

$$\sum_{i=1}^{p} a_i \mathbf{x}_i = \mathbf{0} \tag{A.15}$$

If no such multipliers exist, then we say that the vectors are *linearly independent*, and the matrix is *full rank*. In general, the *rank* of a matrix is the maximum number of $\mathbf{x}_i$ that form a linearly independent set.

For example, the matrix $\mathbf{X}$ given at (A.12) can be shown to have linearly independent columns because no a_i not all equal to zero can be found that satisfy (A.15). On the other hand, the matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 5 \\ 1 & 1 & 4 \\ 1 & 3 & 6 \\ 1 & 8 & 11 \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \tag{A.16}$$

has linearly dependent columns and is singular because $\mathbf{x}_3 = 3\mathbf{x}_1 + \mathbf{x}_2$. The matrix has rank 2, because the linearly independent subset of the columns with the most elements has two elements.

The matrix $\mathbf{X}'\mathbf{X}$ is a $p \times p$ matrix. If $\mathbf{X}$ has rank p, so does $\mathbf{X}'\mathbf{X}$. Full-rank square matrices always have an inverse. Square matrices of less than full rank never have an inverse.

A.7 RANDOM VECTORS

An $n \times 1$ vector $\mathbf{Y}$ is a *random vector* if each of its elements is a random variable. The mean of an $n \times 1$ random vector $\mathbf{Y}$ is also an $n \times 1$ vector whose elements are the means of the elements of $\mathbf{Y}$. The variance of an $n \times 1$ vector $\mathbf{Y}$ is an $n \times n$ square symmetric matrix, often called a *covariance matrix*, written $\mathrm{Var}(\mathbf{Y})$ with $\mathrm{Var}(y_i)$ as its (i, i) element and $\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_j, y_i)$ as both the (i, j) and (j, i) element.

The rules for means and variances of random vectors are matrix equivalents of the scalar versions in Appendix A.2. If $\mathbf{a}_0$ is a vector of constants, and $\mathbf{A}$ is a matrix of constants,

$$\mathrm{E}(\mathbf{a}_0 + \mathbf{A}\mathbf{Y}) = \mathbf{a}_0 + \mathbf{A}\mathrm{E}(\mathbf{Y}) \tag{A.17}$$

$$\mathrm{Var}(\mathbf{a}_0 + \mathbf{A}\mathbf{Y}) = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}' \tag{A.18}$$

A.8 LEAST SQUARES USING MATRICES

The multiple linear regression model can be written as

$$\mathrm{E}(Y|X = \mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \quad \mathrm{Var}(Y|X = \mathbf{x}) = \sigma^2$$

The matrix version is

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of response values and $\mathbf{X}$ is a $n \times p'$ matrix. If the mean function includes an intercept, then the first column of $\mathbf{X}$ is a vector of ones, and $p' = p + 1$. If the mean function does not include an intercept, then the column of one is not included in $\mathbf{X}$ and $p' = p$. The ith row of the $n \times p'$ matrix $\mathbf{X}$ is $\mathbf{x}_i'$, $\boldsymbol{\beta}$ is a $p' \times 1$ vector of parameters for the mean function.

The OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by the arguments that minimize the residual sum of squares function,

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Using (A.14)

$$\mathrm{RSS}(\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} \tag{A.19}$$

$\mathrm{RSS}(\boldsymbol{\beta})$ depends on only three functions of the data: $\mathbf{Y}'\mathbf{Y}$, $\mathbf{X}'\mathbf{X}$, and $\mathbf{Y}'\mathbf{X}$. Any two data sets that have the same values of these three quantities will have the same least squares estimates. Using (A.8), the information in these quantities is equivalent to the information contained in the sample means of the regressors plus the sample covariances of the regressors and the response.

To minimize (A.19), differentiate with respect to $\boldsymbol{\beta}$ and set the result equal to 0. This leads to the matrix version of the normal equations,

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y} \tag{A.20}$$

The OLS estimates are any solution to these equations. If the inverse of $(\mathbf{X}'\mathbf{X})$ exists, as it will if the columns of $\mathbf{X}$ are linearly independent, the OLS estimates are unique and are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{A.21}$$

If the inverse does not exist, then the matrix $(\mathbf{X}'\mathbf{X})$ is of less than full rank, and the OLS estimate is not unique. In this case, most computer programs will use a linearly independent subset of the columns of $\mathbf{X}$ in fitting the model, so that the reduced model matrix does have full rank. This is discussed in Section 4.1.4.

A.8.1 Properties of Estimates

Using the rules for means and variances of random vectors, (A.17) and (A.18), we find

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}\right] \\ &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathrm{E}(\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned} \tag{A.22}$$

so $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, as long as the mean function that was fit is the true mean function. The variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathrm{Var}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathrm{Var}(\mathbf{Y}|\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\sigma^2\mathbf{I}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{A.23}$$

The variances and covariances are compactly determined as σ^2 times a matrix whose elements are determined only by $\mathbf{X}$ and not by $\mathbf{Y}$.

A.8.2 The Residual Sum of Squares

Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be the $n \times 1$ vector of fitted values corresponding to the n cases in the data, and $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the vector of residuals. One representation of the residual sum of squares, which is the residual sum of squares function evaluated at $\hat{\boldsymbol{\beta}}$, is

$$\mathtt{RSS} = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \sum_{i=1}^{n} \hat{e}_i^2$$

which suggests that the residual sum of squares can be computed by squaring the residuals and adding them up. In multiple linear regression, it can also be computed more efficiently on the basis of summary statistics. Using (A.19) and the summary statistics $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$, and $\mathbf{Y}'\mathbf{Y}$, we write

$$\mathtt{RSS} = \mathtt{RSS}(\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

We will first show that $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$. Substituting for one of the $\hat{\boldsymbol{\beta}}$s, we get

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

The last result follows because taking the transpose of a 1×1 matrix does not change its value. The residual sum of squares function can now be rewritten as

$$\begin{aligned}\texttt{RSS} &= \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}\end{aligned}$$

where $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are the fitted values. The residual sum of squares is the difference in the squares of the lengths of the two vectors $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. Another useful form for the residual sum of squares is

$$\texttt{RSS} = \texttt{SYY}(1 - R^2)$$

where R^2 is the square of the sample correlation between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$.

A.8.3 Estimate of Variance

Under the assumption of constant variance, the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\texttt{RSS}}{d} \tag{A.24}$$

with d *df*, where d is equal to the number of cases n minus the number of regressors with estimated coefficients in the model. If the matrix $\mathbf{X}$ is of full rank, then $d = n - p'$, where $p' = p$ for mean functions without an intercept, and $p' = p + 1$ for mean functions with an intercept. The number of estimated coefficients will be less than p' if $\mathbf{X}$ is not of full rank.

A.8.4 Weighted Least Squares

From Section 7.1, the WLS model can be written in matrix notation as

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{W}^{-1} \tag{A.25}$$

To distinguish OLS and WLS results, we will use a subscript W on several quantities. In practice, there is no need to distinguish between OLS and WLS, and this subscript is dropped elsewhere in the book.

- The WLS estimator $\hat{\boldsymbol{\beta}}_W$ of $\boldsymbol{\beta}$ is given by the arguments that minimize the residual sum of squares function,

$$\begin{aligned}\texttt{RSS}_W(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{W}\mathbf{Y} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta} - 2\mathbf{Y}'\mathbf{W}\mathbf{X}\boldsymbol{\beta}\end{aligned}$$