APPENDIX A

MATRIX ALGEBRA

A.1 TERMINOLOGY

A **matrix** is a rectangular array of numbers, denoted

$$\mathbf{A} = [a_{ik}] = [\mathbf{A}]_{ik} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ & & \cdots & \\ a_{n1} & a_{n2} & \cdots & a_{nK} \end{bmatrix}. \qquad \textbf{(A-1)}$$

The typical element is used to denote the matrix. A subscripted element of a matrix is always read as $a_{\text{row, column}}$. An example is given in Table A.1. In these data, the rows are identified with years and the columns with particular variables.

A vector is an ordered set of numbers arranged either in a row or a column. In view of the preceding, a **row vector** is also a matrix with one row, whereas a **column vector** is a matrix with one column. Thus, in Table A.1, the five variables observed for 1972 (including the date) constitute a row vector, whereas the time series of nine values for consumption is a column vector.

A matrix can also be viewed as a set of column vectors or as a set of row vectors.[1] The **dimensions** of a matrix are the numbers of rows and columns it contains. "**A** is an $n \times K$ matrix" (read "n by K") will always mean that **A** has n rows and K columns. If n equals K, then **A** is a **square matrix**. Several particular types of square matrices occur frequently in econometrics.

- A **symmetric matrix** is one in which $a_{ik} = a_{ki}$ for all i and k.
- A **diagonal matrix** is a square matrix whose only nonzero elements appear on the **main diagonal**, that is, moving from upper left to lower right.
- A **scalar matrix** is a diagonal matrix with the same value in all diagonal elements.
- An **identity matrix** is a scalar matrix with ones on the diagonal. This matrix is always denoted **I**. A subscript is sometimes included to indicate its size, or **order**. For example, $\mathbf{I}_4$ indicates a 4×4 identity matrix.
- A **triangular matrix** is one that has only zeros either above or below the main diagonal. If the zeros are above the diagonal, the matrix is **lower triangular**.

[1] Henceforth, we shall denote a matrix by a boldfaced capital letter, as is **A** in (A-1), and a vector as a boldfaced lowercase letter, as in **a**. Unless otherwise noted, a vector will always be assumed to be a *column vector*.

TABLE A.1 Matrix of Macroeconomic Data

	Column				
		2	***3***		***5***
	1	***Consumption***	***GNP***	***4***	***Discount Rate***
Row	***Year***	***(billions of dollars)***	***(billions of dollars)***	***GNP Deflator***	***(N.Y Fed., avg.)***
1	1972	737.1	1185.9	1.0000	4.50
2	1973	812.0	1326.4	1.0575	6.44
3	1974	808.1	1434.2	1.1508	7.83
4	1975	976.4	1549.2	1.2579	6.25
5	1976	1084.3	1718.0	1.3234	5.50
6	1977	1204.4	1918.3	1.4005	5.46
7	1978	1346.5	2163.9	1.5042	7.46
8	1979	1507.2	2417.8	1.6342	10.28
9	1980	1667.2	2633.1	1.7864	11.77

Source: Data from the *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, 1983).

A.2 ALGEBRAIC MANIPULATION OF MATRICES

A.2.1 EQUALITY OF MATRICES

Matrices (or vectors) **A** and **B** are equal if and only if they have the same dimensions and each element of **A** equals the corresponding element of **B**. That is,

$$\mathbf{A} = \mathbf{B} \quad \text{if and only if } a_{ik} = b_{ik} \quad \text{for all } i \text{ and } k. \tag{A-2}$$

A.2.2 TRANSPOSITION

The **transpose** of a matrix **A**, denoted **A**′, is obtained by creating the matrix whose *kth* row is the *kth* column of the original matrix.[2] Thus, if $\mathbf{B} = \mathbf{A}'$, then each column of **A** will appear as the corresponding row of **B**. If **A** is $n \times K$, then **A**′ is $K \times n$.

An equivalent definition of the transpose of a matrix is

$$\mathbf{B} = \mathbf{A}' \Leftrightarrow b_{ik} = a_{ki} \quad \text{for all } i \text{ and } k. \tag{A-3}$$

The definition of a symmetric matrix implies that

$$\text{if (and only if) } \mathbf{A} \text{ is symmetric, then } \mathbf{A} = \mathbf{A}'. \tag{A-4}$$

It also follows from the definition that for any **A**,

$$(\mathbf{A}')' = \mathbf{A}. \tag{A-5}$$

Finally, the transpose of a column vector, **a**, is a row vector:

$$\mathbf{a}' = [a_1 \quad a_2 \quad \cdots \quad a_n].$$

[2] Authors sometimes denote the transpose of a matrix with a superscript "T," as in $\mathbf{A}^{\mathrm{T}}$ = the transpose of **A**. We will use the prime notation throughout this book .

A.2.3 VECTORIZATION

In some derivations and analyses, it is occasionally useful to reconfigure a matrix into a vector (rarely the reverse). The matrix function Vec(**A**) takes the columns of an $n \times K$ matrix and rearranges them in a long $nK \times 1$ vector. Thus, $Vec\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = [1, 2, 2, 4]'$. A related operation is the half vectorization, which collects the lower triangle of a symmetric matrix in a column vector. For example, $Vech\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$.

A.2.4 MATRIX ADDITION

The operations of addition and subtraction are extended to matrices by defining

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = [a_{ik} + b_{ik}]. \tag{A-6}$$

$$\mathbf{A} - \mathbf{B} = [a_{ik} - b_{ik}]. \tag{A-7}$$

Matrices cannot be added unless they have the same dimensions, in which case they are said to be **conformable for addition**. A **zero matrix** or **null matrix** is one whose elements are all zero. In the addition of matrices, the zero matrix plays the same role as the scalar 0 in scalar addition; that is,

$$\mathbf{A} + \mathbf{0} = \mathbf{A}. \tag{A-8}$$

It follows from (A-6) that matrix addition is commutative,

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}. \tag{A-9}$$

and associative,

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}), \tag{A-10}$$

and that

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'. \tag{A-11}$$

A.2.5 VECTOR MULTIPLICATION

Matrices are multiplied by using the **inner product**. The inner product, or **dot product**, of two vectors, **a** and **b**, is a scalar and is written

$$\mathbf{a}'\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_nb_n = \Sigma_{j=1}^{n} a_jb_j. \tag{A-12}$$

Note that the inner product is written as the transpose of vector **a** times vector **b**, a row vector times a column vector. In (A-12), each term a_jb_j equals b_ja_j; hence

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}. \tag{A-13}$$

A.2.6 A NOTATION FOR ROWS AND COLUMNS OF A MATRIX

We need a notation for the ith row of a matrix. Throughout this book, an untransposed vector will always be a column vector. However, we will often require a notation for the column vector that is the transpose of a row of a matrix. This has the potential to create some ambiguity, but the following convention based on the subscripts will suffice for our work throughout this text:

- $\mathbf{a}_k$, or $\mathbf{a}_l$ or $\mathbf{a}_m$ will denote *column k, l,* or *m* of the matrix **A**,
- $\mathbf{a}_i$, or $\mathbf{a}_j$ or $\mathbf{a}_t$ or $\mathbf{a}_s$ will denote the column vector formed by the transpose of row *i*, *j*, *t*, or *s* of matrix **A**. Thus, $\mathbf{a}_i'$ is row *i* of **A**. **(A-14)**

For example, from the data in Table A.1 it might be convenient to speak of $\mathbf{x}_i$, where $i = 1972$ as the 5×1 vector containing the five variables measured for the year 1972, that is, the transpose of the 1972 row of the matrix. In our applications, the common association of subscripts "*i*" and "*j*" with individual *i* or *j*, and "*t*" and "*s*" with time periods *t* and *s* will be natural.

A.2.7 MATRIX MULTIPLICATION AND SCALAR MULTIPLICATION

For an $n \times K$ matrix **A** and a $K \times M$ matrix **B**, the product matrix, $\mathbf{C} = \mathbf{AB}$, is an $n \times M$ matrix whose *ik*th element is the inner product of row *i* of **A** and column *k* of **B**. Thus, the product matrix **C** is

$$\mathbf{C} = \mathbf{AB} \Rightarrow c_{ik} = \mathbf{a}_i'\mathbf{b}_k. \quad \textbf{(A-15)}$$

[Note our use of (A-14) in (A-15).] To multiply two matrices, the number of columns in the first must be the same as the number of rows in the second, in which case they are **conformable for multiplication**.[3] Multiplication of matrices is generally not commutative. In some cases, **AB** may exist, but **BA** may be undefined or, if it does exist, may have different dimensions. In general, however, even if **AB** and **BA** do have the same dimensions, they will not be equal. In view of this, we define **premultiplication** and **postmultiplication** of matrices. In the product **AB**, **B** is *premultiplied* by **A**, whereas **A** is *postmultiplied* by **B**.

Scalar multiplication of a matrix is the operation of multiplying every element of the matrix by a given scalar. For scalar *c* and matrix **A**,

$$c\mathbf{A} = [ca_{ik}]. \quad \textbf{(A-16)}$$

If two matrices **A** and **B** have the same number of rows and columns, then we can compute the **direct product** (also called the **Hadamard product** or the Schur product or the entrywise product), which is a new matrix (or vector) whose *ij* element is the product of the corresponding elements of **A** and **B**. The usual symbol for this operation is "∘." Thus,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \circ \begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} 1a & 2b \\ 2b & 3c \end{bmatrix} \text{ and } \begin{pmatrix} 3 \\ 5 \end{pmatrix} \circ \begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 6 \\ 20 \end{pmatrix}.$$

The product of a matrix and a vector is written

$$\mathbf{c} = \mathbf{Ab}.$$

[3] A simple way to check the conformability of two matrices for multiplication is to write down the dimensions of the operation, for example, $(n \times K)$ times $(K \times M)$. The inner dimensions must be equal; the result has dimensions equal to the outer values.

The number of elements in **b** must equal the number of columns in **A**; the result is a vector with number of elements equal to the number of rows in **A**. For example,

$$\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

We can interpret this in two ways. First, it is a compact way of writing the three equations

$$\begin{aligned} 5 &= 4a + 2b + 1c, \\ 4 &= 2a + 6b + 1c, \\ 1 &= 1a + 1b + 0c. \end{aligned}$$

Second, by writing the set of equations as

$$\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix} = a \begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix} + b \begin{bmatrix} 2 \\ 6 \\ 1 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

we see that the right-hand side is a **linear combination** of the columns of the matrix where the coefficients are the elements of the vector. For the general case,

$$\mathbf{c} = \mathbf{A}\mathbf{b} = b_1\mathbf{a}_1 + b_2\mathbf{a}_2 + \cdots + b_K\mathbf{a}_K. \quad \textbf{(A-17)}$$

In the calculation of a matrix product **C** = **AB**, each column of **C** is a linear combination of the columns of **A**, where the coefficients are the elements in the corresponding column of **B**. That is,

$$\mathbf{C} = \mathbf{AB} \Leftrightarrow \mathbf{c}_k = \mathbf{A}\mathbf{b}_k. \quad \textbf{(A-18)}$$

Let $\mathbf{e}_k$ be a column vector that has zeros everywhere except for a one in the *kth* position. Then $\mathbf{A}\mathbf{e}_k$ is a linear combination of the columns of **A** in which the coefficient on every column but the *k*th is zero, whereas that on the *k*th is one. The result is

$$\mathbf{a}_k = \mathbf{A}\mathbf{e}_k. \quad \textbf{(A-19)}$$

Combining this result with (A-17) produces

$$(\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n) = \mathbf{A}(\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_n) = \mathbf{AI} = \mathbf{A}. \quad \textbf{(A-20)}$$

In matrix multiplication, the identity matrix is analogous to the scalar 1. For any matrix or vector **A, AI** = **A**. In addition, **IA** = **A**, although if **A** is not a square matrix, the two identity matrices are of different orders.

A conformable matrix of zeros produces the expected result: **A0** = **0**.

Some general rules for matrix multiplication are as follows:

- **Associative law:** (**AB**)**C** = **A**(**BC**). **(A-21)**
- **Distributive law:** **A**(**B** + **C**) = **AB** + **AC**. **(A-22)**
- **Transpose of a product:** $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. **(A-23)**
- **Transpose of an extended product:** $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$. **(A-24)**

A.2.8 SUMS OF VALUES

Denote by **i** a vector that contains a column of ones. Then,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n = \mathbf{i}'\mathbf{x}. \quad \textbf{(A-25)}$$

If all elements in **x** are equal to the same constant a, then $\mathbf{x} = a\mathbf{i}$ and

$$\sum_{i=1}^{n} x_i = \mathbf{i}'(a\mathbf{i}) = a(\mathbf{i}'\mathbf{i}) = na. \qquad \textbf{(A-26)}$$

For any constant a and vector **x**,

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i = a\mathbf{i}'\mathbf{x}. \qquad \textbf{(A-27)}$$

If $a = 1/n$, then we obtain the arithmetic mean,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{i}'\mathbf{x}, \qquad \textbf{(A-28)}$$

from which it follows that

$$\sum_{i=1}^{n} x_i = \mathbf{i}'\mathbf{x} = n\bar{x}.$$

The sum of squares of the elements in a vector **x** is

$$\sum_{i=1}^{n} x_i^2 = \mathbf{x}'\mathbf{x}; \qquad \textbf{(A-29)}$$

while the sum of the products of the n elements in vectors **x** and **y** is

$$\sum_{i=1}^{n} x_i y_i = \mathbf{x}'\mathbf{y}. \qquad \textbf{(A-30)}$$

By the definition of matrix multiplication,

$$[\mathbf{X}'\mathbf{X}]_{kl} = [\mathbf{x}_k'\mathbf{x}_l] \qquad \textbf{(A-31)}$$

is the inner product of the kth and lth columns of **X**. For example, for the data set given in Table A.1, if we define **X** as the 9 × 3 matrix containing (year, consumption, GNP), then

$$\begin{aligned}[\mathbf{X}'\mathbf{X}]_{23} &= \sum_{t=1972}^{1980} \text{consumption}_t\,\text{GNP}_t = 737.1(1185.9) + \cdots + 1667.2(2633.1)\\ &= 19{,}743{,}711.34.\end{aligned}$$

If **X** is $n \times K$, then [again using (A-14)]

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'.$$

This form shows that the $K \times K$ matrix $\mathbf{X}'\mathbf{X}$ is the sum of n $K \times K$ matrices, each formed from a single row (year) of **X**. For the example given earlier, this sum is of nine 3 × 3 matrices, each formed from one row (year) of the original data matrix.

A.2.9 A USEFUL IDEMPOTENT MATRIX

A fundamental matrix in statistics is the "centering matrix" that is used to transform data to deviations from their mean. First,

$$\mathbf{i}\bar{x} = \mathbf{i}\,\frac{1}{n}\mathbf{i}'\mathbf{x} = \begin{bmatrix}\bar{x}\\ \bar{x}\\ \vdots\\ \bar{x}\end{bmatrix} = \frac{1}{n}\mathbf{i}\mathbf{i}'\mathbf{x}. \qquad \textbf{(A-32)}$$

The matrix $(1/n)\mathbf{ii}'$ is an $n \times n$ matrix with every element equal to $1/n$. The set of values in deviations form is

$$\begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \cdots \\ x_n - \bar{x} \end{bmatrix} = [\mathbf{x} - \mathbf{i}\bar{x}] = \left[\mathbf{x} - \frac{1}{n}\mathbf{ii}'\mathbf{x}\right]. \tag{A-33}$$

Because $\mathbf{x} = \mathbf{Ix}$,

$$\left[\mathbf{x} - \frac{1}{n}\mathbf{ii}'\mathbf{x}\right] = \left[\mathbf{Ix} - \frac{1}{n}\mathbf{ii}'\mathbf{x}\right] = \left[\mathbf{I} - \frac{1}{n}\mathbf{ii}'\right]\mathbf{x} = \mathbf{M}^0\mathbf{x}. \tag{A-34}$$

Henceforth, the symbol $\mathbf{M}^0$ will be used only for this matrix. Its diagonal elements are all $(1 - 1/n)$, and its off-diagonal elements are $-1/n$. The matrix $\mathbf{M}^0$ is primarily useful in computing sums of squared deviations. Some computations are simplified by the result

$$\mathbf{M}^0\mathbf{i} = \left[\mathbf{I} - \frac{1}{n}\mathbf{ii}'\right]\mathbf{i} = \mathbf{i} - \frac{1}{n}\mathbf{i}(\mathbf{i}'\mathbf{i}) = \mathbf{0},$$

which implies that $\mathbf{i}'\mathbf{M}^0 = \mathbf{0}'$. The sum of deviations about the mean is then

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \mathbf{i}'[\mathbf{M}^0\mathbf{x}] = \mathbf{0}'\mathbf{x} = 0. \tag{A-35}$$

For a single variable $\mathbf{x}$, the sum of squared deviations about the mean is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\sum_{i=1}^{n}x_i^2\right) - n\bar{x}^2. \tag{A-36}$$

In matrix terms,

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\mathbf{x} - \bar{x}\mathbf{i})'(\mathbf{x} - \bar{x}\mathbf{i}) = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{x}) = \mathbf{x}'\mathbf{M}^{0\prime}\mathbf{M}^0\mathbf{x}.$$

Two properties of $\mathbf{M}^0$ are useful at this point. First, because all off-diagonal elements of $\mathbf{M}^0$ equal $-1/n$, $\mathbf{M}^0$ is symmetric. Second, as can easily be verified by multiplication, $\mathbf{M}^0$ is equal to its square; $\mathbf{M}^0\mathbf{M}^0 = \mathbf{M}^0$.

DEFINITION A.1 Idempotent Matrix
An ***idempotent*** *matrix,* $\mathbf{M}$, *is one that is equal to its square, that is,* $\mathbf{M}^2 = \mathbf{MM} = \mathbf{M}$. *If* $\mathbf{M}$ *is a symmetric idempotent matrix (all of the idempotent matrices we shall encounter are symmetric), then* $\mathbf{M}'\mathbf{M} = \mathbf{M}$ as well.

Thus, $\mathbf{M}^0$ is a symmetric idempotent matrix. Combining results, we obtain

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x}. \tag{A-37}$$

Consider constructing a matrix of sums of squares and cross products in deviations from the column means. For two vectors **x** and **y**,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (\mathbf{M}^0\mathbf{x})'(\mathbf{M}^0\mathbf{y}), \qquad \textbf{(A-38)}$$

so

$$\begin{bmatrix} \sum_{i=1}^{n}(x_i - \bar{x})^2 & \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^{n}(y_i - \bar{y})^2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\mathbf{M}^0\mathbf{x} & \mathbf{x}'\mathbf{M}^0\mathbf{y} \\ \mathbf{y}'\mathbf{M}^0\mathbf{x} & \mathbf{y}'\mathbf{M}^0\mathbf{y} \end{bmatrix}. \qquad \textbf{(A-39)}$$

If we put the two column vectors **x** and **y** in an $n \times 2$ matrix $\mathbf{Z} = [\mathbf{x}, \mathbf{y}]$, then $\mathbf{M}^0\mathbf{Z}$ is the $n \times 2$ matrix in which the two columns of data are in mean deviation form. Then

$$(\mathbf{M}^0\mathbf{Z})'(\mathbf{M}^0\mathbf{Z}) = \mathbf{Z}'\mathbf{M}^0\mathbf{M}^0\mathbf{Z} = \mathbf{Z}'\mathbf{M}^0 Z.$$

A.3 GEOMETRY OF MATRICES

A.3.1 VECTOR SPACES

The K elements of a column vector

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_K \end{bmatrix}$$

can be viewed as the coordinates of a point in a K-dimensional space, as shown in Figure A.1 for two dimensions, or as the definition of the line segment connecting the origin and the point defined by **a**.

Two basic arithmetic operations are defined for vectors, **scalar multiplication** and **addition**. A scalar multiple of a vector, **a**, is another vector, say **a***, whose coordinates are the scalar multiple of **a**'s coordinates. Thus, in Figure A.1,

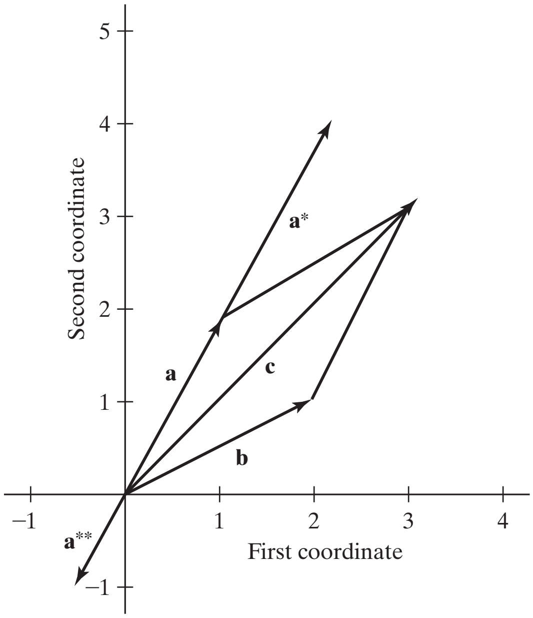
$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{a}^* = 2\mathbf{a} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \quad \mathbf{a}^{**} = -\frac{1}{2}\mathbf{a} = \begin{bmatrix} -\frac{1}{2} \\ -1 \end{bmatrix}.$$

The set of all possible scalar multiples of **a** is the line through the origin, **0** and **a**. Any scalar multiple of **a** is a segment of this line. The sum of two vectors **a** and **b** is a third vector whose coordinates are the sums of the corresponding coordinates of **a** and **b**. For example,

$$\mathbf{c} = \mathbf{a} + \mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}.$$

Geometrically, **c** is obtained by moving in the distance and direction defined by **b** from the tip of **a** or, because addition is commutative, from the tip of **b** in the distance and direction of **a**. Note that scalar multiplication and addition of vectors are special cases of (A-16) and (A-6) for matrices.

FIGURE A.1 Vector Space.



The two-dimensional plane is the set of all vectors with two real-valued coordinates. We label this set $\mathbb{R}^2$ ("R two," not "R squared"). It has two important properties.

- $\mathbb{R}^2$ *is closed under scalar multiplication;* every scalar multiple of a vector in $\mathbb{R}^2$ is also in $\mathbb{R}^2$.
- $\mathbb{R}^2$ *is closed under addition;* the sum of any two vectors in the plane is always a vector in $\mathbb{R}^2$.

DEFINITION A.2 Vector Space
*A **vector space** is any set of vectors that is closed under scalar multiplication and addition.*

Another example is the set of all real numbers, that is, $\mathbb{R}^1$, that is, the set of vectors with one real element. In general, that set of K-element vectors all of whose elements are real numbers is a K-dimensional vector space, denoted $\mathbb{R}^K$. The preceding examples are drawn in $\mathbb{R}^2$.

A.3.2 LINEAR COMBINATIONS OF VECTORS AND BASIS VECTORS

In Figure A.2, $\mathbf{c} = \mathbf{a} + \mathbf{b}$ and $\mathbf{d} = \mathbf{a}^* + \mathbf{b}$. But since $\mathbf{a}^* = 2\mathbf{a}$, $\mathbf{d} = 2\mathbf{a} + \mathbf{b}$. Also, $\mathbf{e} = \mathbf{a} + 2\mathbf{b}$ and $\mathbf{f} = \mathbf{b} + (-\mathbf{a}) = \mathbf{b} - \mathbf{a}$. As this exercise suggests, any vector in $\mathbb{R}^2$ could be obtained as a **linear combination** of **a** and **b**.

DEFINITION A.3 Basis Vectors
A set of vectors in a vector space is a ***basis*** *for that vector space if they are linearly independent and any vector in the vector space can be written as a linear combination of that set of vectors.*

As is suggested by Figure A.2, any pair of two-element vectors, including **a** and **b**, that point in different directions will form a basis for $\mathbb{R}^2$. Consider an arbitrary set of three vectors in $\mathbb{R}^2$, **a, b**, and **c**. If **a** and **b** are a basis, then we can find numbers α_1 and α_2 such that $\mathbf{c} = \alpha_1\mathbf{a} + \alpha_2\mathbf{b}$. Let

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

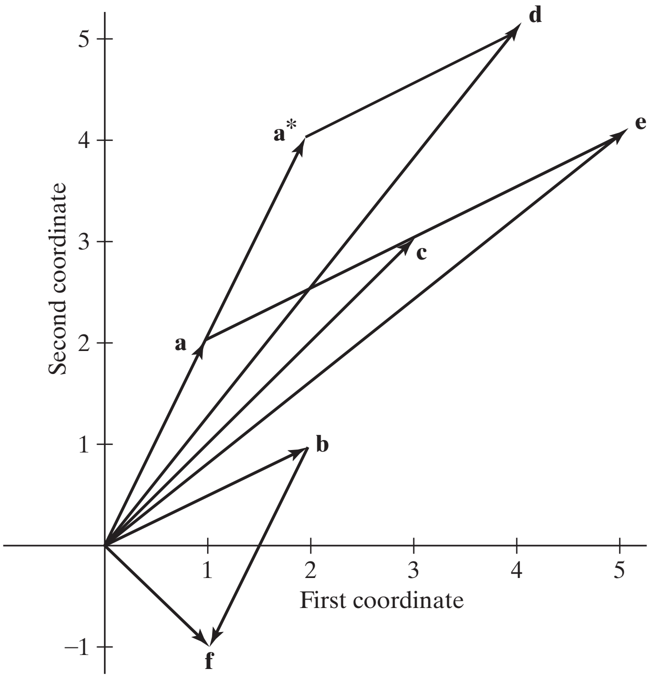
Then

$$\begin{aligned} c_1 &= \alpha_1 a_1 + \alpha_2 b_1, \\ c_2 &= \alpha_1 a_2 + \alpha_2 b_2. \end{aligned} \tag{A-40}$$

The solutions (α_1, α_2) to this pair of equations are

$$\alpha_1 = \frac{b_2 c_1 - b_1 c_2}{a_1 b_2 - b_1 a_2}, \quad \alpha_2 = \frac{a_1 c_2 - a_2 c_1}{a_1 b_2 - b_1 a_2}. \tag{A-41}$$

FIGURE A.2 Linear Combinations of Vectors.



This result gives a unique solution unless $(a_1b_2 - b_1a_2) = 0$. If $(a_1b_2 - b_1a_2) = 0$, then $a_1/a_2 = b_1/b_2$, which means that **b** is just a multiple of **a**. This returns us to our original condition, that **a** and **b** must point in different directions. The implication is that if **a** and **b** are any pair of vectors for which the denominator in (A-41) is not zero, then any other vector **c** can be formed as a *unique* linear combination of **a** and **b**. The basis of a vector space is not unique, since any set of vectors that satisfies the definition will do. But for any particular basis, only one linear combination of them will produce another particular vector in the vector space.

A.3.3 LINEAR DEPENDENCE

As the preceding should suggest, K vectors are required to form a basis for $\mathbb{R}^K$. Although the basis for a vector space is not unique, not every set of K vectors will suffice. In Figure A.2, **a** and **b** form a basis for $\mathbb{R}^2$, but **a** and **a*** do not. The difference between these two pairs is that **a** and **b** are linearly *independent,* whereas **a** and **a*** are linearly *dependent*.

DEFINITION A.4 Linear Dependence
*A set of $k \geq 2$ vectors is **linearly dependent** if at least one of the vectors in the set can be written as a linear combination of the others.*

Because **a*** is a multiple of **a**, **a** and **a*** are linearly dependent. For another example, if

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} 10 \\ 14 \end{bmatrix},$$

then

$$2\mathbf{a} + \mathbf{b} - \frac{1}{2}\mathbf{c} = \mathbf{0},$$

so **a**, **b**, and **c** are linearly dependent. Any of the three possible pairs of them, however, are linearly independent.

DEFINITION A.5 Linear Independence
*A set of vectors is **linearly independent** if and only if the only solution $(\alpha_1, \ldots, \alpha_K)$ to*

$$\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \cdots + \alpha_K\mathbf{a}_K = \mathbf{0}$$

is

$$\alpha_1 = \alpha_2 = \cdots = \alpha_K = 0.$$

The preceding implies the following equivalent definition of a basis.

DEFINITION A.6 Basis for a Vector Space
A basis for a vector space of K dimensions is any set of K linearly independent vectors in that vector space.

Because any $(K + 1)$st vector can be written as a linear combination of the K basis vectors, it follows that any set of more than K vectors in $\mathbb{R}^K$ must be linearly dependent.

A.3.4 SUBSPACES

DEFINITION A.7 Spanning Vectors
The set of all linear combinations of a set of vectors is the vector space that is ***spanned*** *by those vectors.*

For example, by definition, the space spanned by a basis for $\mathbb{R}^K$ is $\mathbb{R}^K$. An implication of this is that if **a** and **b** are a basis for $\mathbb{R}^2$ and **c** is another vector in $\mathbb{R}^2$, the space spanned by [**a, b, c**] is, again, $\mathbb{R}^2$. Of course, **c** is superfluous. Nonetheless, any vector in $\mathbb{R}^2$ *can* be expressed as a linear combination of **a, b**, and **c**. (The linear combination will not be unique. Suppose, for example, that **a** and **c** are also a basis for $\mathbb{R}^2$.)

Consider the set of three coordinate vectors whose third element is zero. In particular,

$$\mathbf{a}' = [a_1 \quad a_2 \quad 0] \quad \text{and} \quad \mathbf{b}' = [b_1 \quad b_2 \quad 0].$$

Vectors **a** and **b** do not span the three-dimensional space $\mathbb{R}^3$. Every linear combination of **a** and **b** has a third coordinate equal to zero; thus, for instance, $\mathbf{c}' = [1 \quad 2 \quad 3]$ could not be written as a linear combination of **a** and **b**. If $(a_1b_2 - a_2b_1)$ is not equal to zero [see (A-41)]; however, then *any vector whose third element is zero can be expressed as a linear combination of* **a** *and* **b**. So, although **a** and **b** do not span $\mathbb{R}^3$, they do span something; they span the set of vectors in $\mathbb{R}^3$ whose third element is zero. This area is a plane (the "floor" of the box in a three-dimensional figure). This plane in $\mathbb{R}^3$ is a **subspace**, in this instance, a two-dimensional subspace. Note that *it is not* $\mathbb{R}^2$; it is the set of vectors in $\mathbb{R}^3$ whose third coordinate is 0. Any plane in $\mathbb{R}^3$ that contains the origin, $(0, 0, 0)$, regardless of how it is oriented, forms a two-dimensional subspace. Any two independent vectors that lie in that subspace will span it. But without a third vector that points in some other direction, we cannot span any more of $\mathbb{R}^3$ than this two-dimensional part of it. By the same logic, any line in $\mathbb{R}^3$ that passes through the origin is a one-dimensional subspace, in this case, the set of all vectors in $\mathbb{R}^3$ whose coordinates are multiples of those of the vector that define the line. A subspace is a vector space in all the respects in which we have defined it. We emphasize that it is *not* a vector space of lower dimension. For example, $\mathbb{R}^2$ is not a subspace of $\mathbb{R}^3$. The essential difference is the number of dimensions in the vectors. The vectors in $\mathbb{R}^3$ that form a two-dimensional subspace are still three-element vectors; they all just happen to lie in the same plane.

The space spanned by a set of vectors in $\mathbb{R}^K$ has at most K dimensions. If this space has fewer than K dimensions, it is a subspace, or **hyperplane.** But the important point in the preceding discussion is that *every set of vectors spans some space;* it may be the entire space in which the vectors reside, or it may be some subspace of it.

A.3.5 RANK OF A MATRIX

We view a matrix as a set of column vectors. The number of columns in the matrix equals the number of vectors in the set, and the number of rows equals the number of coordinates in each column vector. If the matrix contains K rows, its column space might have K dimensions. But,

DEFINITION A.8 Column Space
*The **column space** of a matrix is the vector space that is spanned by its column vectors.*

as we have seen, it might have fewer dimensions; the column vectors might be linearly dependent, or there might be fewer than K of them. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 6 \\ 2 & 6 & 8 \\ 7 & 1 & 8 \end{bmatrix}.$$

It contains three vectors from $\mathbb{R}^3$, but the third is the sum of the first two, so the column space of this matrix cannot have three dimensions. Nor does it have only one, because the three columns are not all scalar multiples of one another. Hence, it has two, and the column space of this matrix is a two-dimensional subspace of $\mathbb{R}^3$. It follows that the column rank of a matrix is

DEFINITION A.9 Column Rank
*The **column rank** of a matrix is the dimension of the vector space that is spanned by its column vectors.*

equal to the largest number of linearly independent column vectors it contains. The column rank of $\mathbf{A}$ is 2. For another specific example, consider

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 1 & 5 \\ 6 & 4 & 5 \\ 3 & 1 & 4 \end{bmatrix}.$$

It can be shown (we shall see how later) that this matrix has a column rank equal to 3. Each column of $\mathbf{B}$ is a vector in $\mathbb{R}^4$, so the column space of $\mathbf{B}$ is a three-dimensional subspace of $\mathbb{R}^4$.

Consider, instead, the set of vectors obtained by using the *rows* of $\mathbf{B}$ instead of the columns. The new matrix would be

$$\mathbf{C} = \begin{bmatrix} 1 & 5 & 6 & 3 \\ 2 & 1 & 4 & 1 \\ 3 & 5 & 5 & 4 \end{bmatrix}.$$

This matrix is composed of four column vectors from $\mathbb{R}^3$. (Note that $\mathbf{C}$ is $\mathbf{B}'$.) The column space of $\mathbf{C}$ is at most $\mathbb{R}^3$, since four vectors in $\mathbb{R}^3$ must be linearly dependent. In fact, the column space of **C** *is* $\mathbb{R}^3$. Although this is not the same as the column space of **B**, it does have the same dimension. Thus, the column rank of **C** and the column rank of **B** are the same. But the columns of **C** are the rows of **B**. Thus, the column rank of **C** equals the **row rank** of **B**. That the column and row ranks of **B** are the same is not a coincidence. The general results (which are equivalent) are as follows:

THEOREM A.1 Equality of Row and Column Rank
The ***column rank*** *and* ***row rank*** *of a matrix are equal. By the definition of row rank and its counterpart for column rank, we obtain the corollary, the* ***row space*** *and* ***column space*** *of a matrix have the same dimension.* **(A-42)**

Theorem A.1 holds regardless of the actual row and column rank. If the column rank of a matrix happens to equal the number of columns it contains, then the matrix is said to have **full column rank**. **Full row rank** is defined likewise. Because the row and column ranks of a matrix are always equal, we can speak unambiguously of the **rank of a matrix**. For either the row rank or the column rank (and, at this point, we shall drop the distinction), it follows that

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}') \leq \min\text{ (number of rows, number of columns)}. \qquad \textbf{(A-43)}$$

In most contexts, we shall be interested in the columns of the matrices we manipulate. We shall use the term **full rank** to describe a matrix whose rank is equal to the number of columns it contains.

Of particular interest will be the distinction between full rank and **short rank matrices**. The distinction turns on the solutions to $\mathbf{Ax} = \mathbf{0}$. If a nonzero $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{0}$ exists, then **A** does not have full rank. Equivalently, if the nonzero **x** exists, then the columns of **A** are linearly dependent and at least one of them can be expressed as a linear combination of the others. For example, a nonzero set of solutions to

$$\begin{bmatrix} 1 & 3 & 10 \\ 2 & 3 & 14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

is any multiple of $\mathbf{x}' = (2, 1, -\tfrac{1}{2})$.

In a product matrix $\mathbf{C} = \mathbf{AB}$, every column of **C** is a linear combination of the columns of **A**, so each column of **C** is in the column space of **A**. It is possible that the set of columns in **C** could span this space, but it is not possible for them to span a higher-dimensional space. At best, they could be a full set of linearly independent vectors in **A**’s column space. We conclude that the column rank of **C** could not be greater than that of **A**. Now, apply the same logic to the rows of **C**, which are all linear combinations of the rows of **B**. For the same reason that the column rank of **C** cannot exceed the column rank of **A**, the row rank of **C** cannot exceed the row rank of **B**. Row and column ranks are always equal, so we can conclude that

$$\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})). \qquad \textbf{(A-44)}$$

A useful corollary to (A-44) is

If $\mathbf{A}$ is $M \times n$ and $\mathbf{B}$ is a square matrix of rank n, then $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$. **(A-45)**

Another application that plays a central role in the development of regression analysis is, for any matrix $\mathbf{A}$,

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}'). \tag{A-46}$$

A.3.6 DETERMINANT OF A MATRIX

The determinant of a square matrix—determinants are not defined for nonsquare matrices—is a function of the elements of the matrix. There are various definitions, most of which are not useful for our work. Determinants figure into our results in several ways, however, that we can enumerate before we need formally to define the computations.

PROPOSITION
The determinant of a matrix is nonzero if and only if it has full rank.

Full rank and short rank matrices can be distinguished by whether or not their determinants are nonzero. There are some settings in which the value of the determinant is also of interest, so we now consider some algebraic results.

It is most convenient to begin with a diagonal matrix

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & d_K \end{bmatrix}.$$

The column vectors of $\mathbf{D}$ define a "box" in $\mathbb{R}^K$ whose sides are all at right angles to one another.[4] Its "volume," or determinant, is simply the product of the lengths of the sides, which we denote

$$|\mathbf{D}| = d_1 d_2 \ldots d_K = \prod_{k=1}^{K} d_k. \tag{A-47}$$

A special case is the identity matrix, which has, regardless of K, $|\mathbf{I}_K| = 1$. Multiplying $\mathbf{D}$ by a scalar c is equivalent to multiplying the length of each side of the box by c, which would multiply its volume by c^K. Thus,

$$|c\mathbf{D}| = c^K|\mathbf{D}|. \tag{A-48}$$

Continuing with this admittedly special case, we suppose that only one column of $\mathbf{D}$ is multiplied by c. In two dimensions, this would make the box wider but not higher, or vice versa. Hence, the "volume" (area) would also be multiplied by c. Now, suppose that each side of the box were multiplied by a different c, the first by c_1, the second by c_2, and so

[4] Each column vector defines a segment on one of the axes.

on. The volume would, by an obvious extension, now be $c_1 c_2 \dots c_K |\mathbf{D}|$. The matrix with columns defined by $[c_1\mathbf{d}_1\ c_2\mathbf{d}_2 \dots]$ is just **DC**, where **C** is a diagonal matrix with c_i as its ith diagonal element. The computation just described is, therefore,

$$|\mathbf{DC}| = |\mathbf{D}| \cdot |\mathbf{C}|. \tag{A-49}$$

(The determinant of **C** is the product of the c_i's since **C**, like **D**, is a diagonal matrix.) In particular, note what happens to the whole thing if one of the c_i's is zero.

For 2 × 2 matrices, the computation of the determinant is

$$\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc. \tag{A-50}$$

Notice that it is a function of all the elements of the matrix. This statement will be true, in general. For more than two dimensions, the determinant can be obtained by using an **expansion by cofactors**. Using *any* row, say, i, we obtain

$$|\mathbf{A}| = \sum_{k=1}^{K} a_{ik}(-1)^{i+k}|\mathbf{A}_{(ik)}|, \quad k = 1, \dots, K, \tag{A-51}$$

where $\mathbf{A}_{(ik)}$ is the matrix obtained from **A** by deleting row i and column k. The determinant of $\mathbf{A}_{(ik)}$ is called a **minor** of **A**.[5] When the correct sign, $(-1)^{i+k}$, is added, it becomes a **cofactor**. This operation can be done using any column as well. For example, a 4 × 4 determinant becomes a sum of four 3 × 3s, whereas a 5 × 5 is a sum of five 4 × 4s, each of which is a sum of four 3 × 3s, and so on. Obviously, it is a good idea to base (A-51) on a row or column with many zeros in it, if possible. In practice, this rapidly becomes a heavy burden. It is unlikely, though, that you will ever calculate any determinants over 3 × 3 without a computer. A 3 × 3, however, might be computed on occasion; if so, the following shortcut known as Sarrus's rule will prove useful:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{31}a_{22}a_{13} - a_{21}a_{12}a_{33} - a_{11}a_{23}a_{32}.$$

Although (A-48) and (A-49) were given for diagonal matrices, they hold for general matrices **C** and **D**. One special case of (A-48) to note is that of $c = -1$. Multiplying a matrix by -1 does not necessarily change the sign of its determinant. It does so only if the order of the matrix is odd. By using the expansion by cofactors formula, an additional result can be shown:

$$|\mathbf{A}| = |\mathbf{A}'|. \tag{A-52}$$

A.3.7 A LEAST SQUARES PROBLEM

Given a vector **y** and a matrix **X**, we are interested in expressing **y** as a linear combination of the columns of **X**. There are two possibilities. If **y** lies in the column space of **X**, then we shall be able to find a vector **b** such that

$$\mathbf{y} = \mathbf{Xb}. \tag{A-53}$$

[5] If i equals k, then the determinant is a principal minor.

FIGURE A.3 Least Squares Projections.

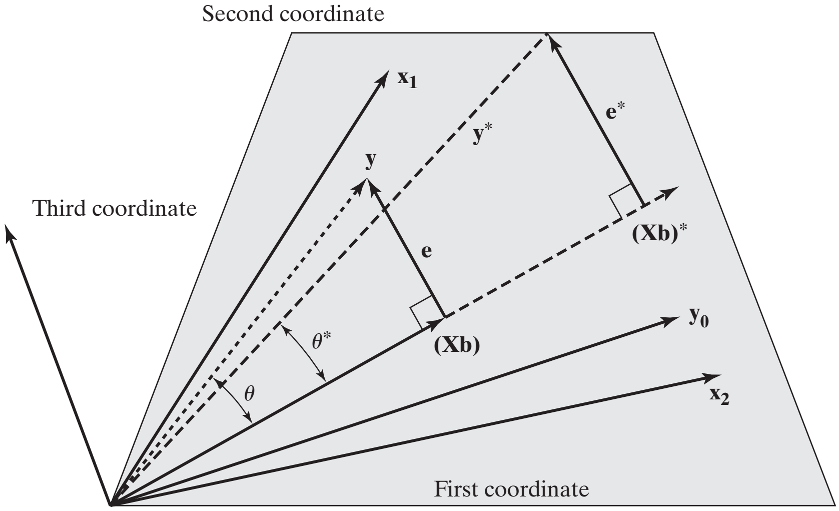


Figure A.3 illustrates such a case for three dimensions in which the two columns of **X** both have a third coordinate equal to zero. Only **y**'s whose third coordinate is zero, such as $\mathbf{y}^0$ in the figure, can be expressed as **Xb** for some **b**. For the general case, assuming that **y** is, indeed, in the column space of **X**, we can find the coefficients **b** by solving the set of equations in (A-53). The solution is discussed in the next section.

Suppose, however, that **y** is not in the column space of **X**. In the context of this example, suppose that **y**'s third component is not zero. Then there is no **b** such that (A-53) holds. We can, however, write

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \tag{A-54}$$

where **e** is the difference between **y** and **Xb**. By this construction, we find an **Xb** that is in the column space of **X**, and **e** is the difference, or "residual." Figure A.3 shows two examples, **y** and **y***. For the present, we consider only **y**. We are interested in finding the **b** such that **y** is as close as possible to **Xb** in the sense that **e** is as short as possible.

DEFINITION A.10 Length of a Vector
The length, or ***norm****, of a vector* **e** *is given by the Pythagorean theorem:*

$$\|\mathbf{e}\| = \sqrt{\mathbf{e}'\mathbf{e}}. \tag{A-55}$$

The problem is to find the **b** for which

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

is as small as possible. The solution is that **b** that makes **e** perpendicular, or *orthogonal,* to **Xb**.

DEFINITION A.11 Orthogonal Vectors

Two nonzero vectors **a** *and* **b** *are* ***orthogonal****, written* $\mathbf{a} \perp \mathbf{b}$, *if and only if*

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \mathbf{0}.$$

Returning once again to our fitting problem, we find that the **b** we seek is that for which

$$\mathbf{e} \perp \mathbf{Xb}.$$

Expanding this set of equations gives the requirement

$$\begin{aligned}(\mathbf{Xb})'\mathbf{e} &= \mathbf{0}\\ &= \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{Xb}\\ &= \mathbf{b}'[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Xb}],\end{aligned}$$

or, assuming **b** is not **0**, the set of equations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Xb}.$$

The means of solving such a set of equations is the subject of Section A.4.

In Figure A.3, the linear combination **Xb** is called the **projection** of **y** into the column space of **X**. The figure is drawn so that, although **y** and **y*** are different, they are similar in that the projection of **y** lies on top of that of **y***. The question we wish to pursue here is, Which vector, **y** or **y***, is closer to its projection in the column space of **X**? Superficially, it would appear that **y** is closer, because **e** is shorter than **e***. Yet **y*** is much more nearly parallel to its projection than **y**, so the only reason that its residual vector is longer is that **y*** is longer compared with **y**. A measure of comparison that would be unaffected by the length of the vectors is the angle between the vector and its projection (assuming that angle is not zero). By this measure, θ^* is smaller than θ, which would reverse the earlier conclusion.

THEOREM A.2 The Cosine Law

The angle θ *between two vectors* **a** *and* **b** *satisfies* $\cos\theta = \dfrac{\mathbf{a}'\mathbf{b}}{\|a\| \times \|\mathbf{b}\|}$.

The two vectors in the calculation would be **y** or **y*** and **Xb** or (**Xb**)*. A zero cosine implies that the vectors are orthogonal. If the cosine is one, then the angle is zero, which means that the vectors are the same. (They would be if **y** were in the column space of **X**.) By dividing by the lengths, we automatically compensate for the length of **y**. By this measure, we find in Figure A.3 that **y** * is closer to its projection, (**Xb**)* than **y** is to its projection, **Xb**.

A.4 SOLUTION OF A SYSTEM OF LINEAR EQUATIONS

Consider the set of n linear equations

$$\mathbf{Ax} = \mathbf{b}, \tag{A-56}$$

in which the K elements of $\mathbf{x}$ constitute the unknowns. $\mathbf{A}$ is a known matrix of coefficients, and $\mathbf{b}$ is a specified vector of values. We are interested in knowing whether a solution exists; if so, then how to obtain it; and finally, if it does exist, then whether it is unique.

A.4.1 SYSTEMS OF LINEAR EQUATIONS

For most of our applications, we shall consider only square systems of equations, that is, those in which $\mathbf{A}$ is a square matrix. In what follows, therefore, we take n to equal K. Because the number of rows in $\mathbf{A}$ is the number of equations, whereas the number of columns in $\mathbf{A}$ is the number of variables, this case is the familiar one of "n equations in n unknowns."

There are two types of systems of equations.

DEFINITION A.12 Homogeneous Equation System
A homogeneous system is of the form $\mathbf{Ax} = \mathbf{0}$.

By definition, a nonzero solution to such a system will exist if and only if $\mathbf{A}$ does not have **full rank**. If so, then for at least one column of $\mathbf{A}$, we can write the preceding as

$$\mathbf{a}_k = -\sum_{m \neq k} \frac{x_m}{x_k} \mathbf{a}_m.$$

This means, as we know, that the columns of $\mathbf{A}$ are linearly dependent and that $|\mathbf{A}| = \mathbf{0}$.

DEFINITION A.13 Nonhomogeneous Equation System
A nonhomogeneous system of equations is of the form $\mathbf{Ax} = \mathbf{b}$, *where* $\mathbf{b}$ *is a nonzero vector.*

The vector $\mathbf{b}$ is chosen arbitrarily and is to be expressed as a linear combination of the columns of $\mathbf{A}$. Because $\mathbf{b}$ has K elements, this solution will exist only if the columns of $\mathbf{A}$ span the entire K-dimensional space, $\mathbb{R}^K$.[6] Equivalently, we shall require that the columns of $\mathbf{A}$ be linearly independent or that $|\mathbf{A}|$ not be equal to zero.

A.4.2 INVERSE MATRICES

To solve the system $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x}$, something akin to division by a matrix is needed. Suppose that we could find a square matrix $\mathbf{B}$ such that $\mathbf{BA} = \mathbf{I}$. If the equation system is premultiplied by this $\mathbf{B}$, then the following would be obtained:

$$\mathbf{BAx} = \mathbf{Ix} = \mathbf{x} = \mathbf{Bb}. \tag{A-57}$$

[6] If $\mathbf{A}$ does not have full rank, then the nonhomogeneous system will have solutions for *some* vectors $\mathbf{b}$, namely, any $\mathbf{b}$ in the column space of $\mathbf{A}$. But we are interested in the case in which there are solutions for *all* nonzero vectors $\mathbf{b}$, which requires $\mathbf{A}$ to have full rank.

If the matrix **B** exists, then it is the **inverse** of **A**, denoted

$$\mathbf{B} = \mathbf{A}^{-1}.$$

From the definition,

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

In addition, by premultiplying by **A**, postmultiplying by $\mathbf{A}^{-1}$, and then canceling terms, we find

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

as well.

If the inverse exists, then it must be unique. Suppose that it is not and that **C** is a different inverse of **A**. Then **CAB** = **CAB**, but (**CA**)**B** = **IB** = **B** and **C**(**AB**) = **C**, which would be a contradiction if **C** did not equal **B**. Because, by (A-57), the solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, the solution to the equation system is unique as well.

We now consider the calculation of the inverse matrix. For a 2 × 2 matrix, **AB** = **I** implies that

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} = 1 \\ a_{11}b_{12} + a_{12}b_{22} = 0 \\ a_{21}b_{11} + a_{22}b_{21} = 0 \\ a_{21}b_{12} + a_{22}b_{22} = 1 \end{bmatrix}.$$

The solutions are

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} = \frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \quad \textbf{(A-58)}$$

Notice the presence of the reciprocal of $|\mathbf{A}|$ in $\mathbf{A}^{-1}$. This result is not specific to the 2 × 2 case. We infer from it that if the determinant is zero, then the inverse does not exist.

DEFINITION A.14 Nonsingular Matrix
A matrix is nonsingular if and only if its inverse exists.

The simplest inverse matrix to compute is that of a diagonal matrix. If

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & d_K \end{bmatrix}, \quad \text{then} \quad \mathbf{D}^{-1} = \begin{bmatrix} 1/d_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/d_2 & 0 & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1/d_K \end{bmatrix},$$

which shows, incidentally, that $\mathbf{I}^{-1} = \mathbf{I}$.

We shall use a^{ik} to indicate the *ik*th element of $\mathbf{A}^{-1}$. The general formula for computing an inverse matrix is

$$a^{ik} = \frac{|\mathbf{C}_{ki}|}{|\mathbf{A}|}, \quad \textbf{(A-59)}$$

where $|\mathbf{C}_{ki}|$ is the *ki*th cofactor of $\mathbf{A}$. [See (A-51).] It follows, therefore, that for $\mathbf{A}$ to be nonsingular, $|\mathbf{A}|$ must be nonzero. Notice the reversal of the subscripts

Some computational results involving inverses are

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}, \tag{A-60}$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}, \tag{A-61}$$

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}. \tag{A-62}$$

$$\text{If } \mathbf{A} \text{ is symmetric, then } \mathbf{A}^{-1} \text{ is symmetric.} \tag{A-63}$$

When both inverse matrices exist,

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{A-64}$$

Note the condition preceding (A-64). It may be that $\mathbf{AB}$ is a square, nonsingular matrix when neither $\mathbf{A}$ nor $\mathbf{B}$ is even square. (Consider, e.g., $\mathbf{A}'\mathbf{A}$.) Extending (A-64), we have

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}(\mathbf{AB})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{A-65}$$

Recall that for a data matrix $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$ is the sum of the *outer products* of the rows $\mathbf{X}$. Suppose that we have already computed $\mathbf{S} = (\mathbf{X}'\mathbf{X})^{-1}$ for a number of years of data, such as those given in Table A.1. The following result, which is called an **updating formula**, shows how to compute the new $\mathbf{S}$ that would result when a new row is added to $\mathbf{X}$: For symmetric, nonsingular matrix $\mathbf{A}$,

$$[\mathbf{A} \pm \mathbf{bb}']^{-1} = \mathbf{A}^{-1} \mp \left[\frac{1}{1 \pm \mathbf{b}'\mathbf{A}^{-1}\mathbf{b}}\right]\mathbf{A}^{-1}\,\mathbf{bb}'\mathbf{A}^{-1}. \tag{A-66}$$

Note the reversal of the sign in the inverse. Two more general forms of (A-66) that are occasionally useful are

$$[\mathbf{A} \pm \mathbf{bc}']^{-1} = \mathbf{A}^{-1} \mp \left[\frac{1}{1 \pm \mathbf{c}'\mathbf{A}^{-1}\mathbf{b}}\right]\mathbf{A}^{-1}\mathbf{bc}'\mathbf{A}^{-1}, \tag{A-66a}$$

$$[\mathbf{A} \pm \mathbf{BCB}']^{-1} = \mathbf{A}^{-1} \mp \mathbf{A}^{-1}\mathbf{B}[\mathbf{C}^{-1} \pm \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}]^{-1}\mathbf{B}'\mathbf{A}^{-1}. \tag{A-66b}$$

A.4.3 NONHOMOGENEOUS SYSTEMS OF EQUATIONS

For the nonhomogeneous system

$$\mathbf{Ax} = \mathbf{b},$$

if $\mathbf{A}$ is nonsingular, then the unique solution is

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

A.4.4 SOLVING THE LEAST SQUARES PROBLEM

We now have the tool needed to solve the least squares problem posed in Section A.3.7. We found the solution vector, $\mathbf{b}$ to be the solution to the nonhomogenous system

$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$. Let $\mathbf{a}$ equal the vector $\mathbf{X}'\mathbf{y}$ and let $\mathbf{A}$ equal the square matrix $\mathbf{X}'\mathbf{X}$. The equation system is then

$$\mathbf{A}\mathbf{b} = \mathbf{a}.$$

By the preceding results, if $\mathbf{A}$ is nonsingular, then

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

assuming that the matrix to be inverted is nonsingular. We have reached the irreducible minimum. If the columns of $\mathbf{X}$ are linearly independent, that is, if $\mathbf{X}$ has full rank, then this is the solution to the least squares problem. If the columns of $\mathbf{X}$ are linearly dependent, then this system has no unique solution.

A.5 PARTITIONED MATRICES

In formulating the elements of a matrix, it is sometimes useful to group some of the elements in **submatrices**. Let

$$\mathbf{A} = \left[\begin{array}{cc|c} 1 & 4 & 5 \\ 2 & 9 & 3 \\ \hline 8 & 9 & 6 \end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

$\mathbf{A}$ is a **partitioned matrix**. The subscripts of the submatrices are defined in the same fashion as those for the elements of a matrix. A common special case is the **block-diagonal matrix**:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square matrices.

A.5.1 ADDITION AND MULTIPLICATION OF PARTITIONED MATRICES

For conformably partitioned matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} \\ \mathbf{A}_{21} + \mathbf{B}_{21} & \mathbf{A}_{22} + \mathbf{B}_{22} \end{bmatrix}, \tag{A-67}$$

and

$$\mathbf{A}\mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}. \tag{A-68}$$

In all these, the matrices must be conformable for the operations involved. For addition, the dimensions of $\mathbf{A}_{ik}$ and $\mathbf{B}_{ik}$ must be the same. For multiplication, the number of columns in $\mathbf{A}_{ij}$ must equal the number of rows in $\mathbf{B}_{jl}$ for all pairs i and j. That is, all the necessary matrix products of the submatrices must be defined. Two cases frequently encountered are of the form

$$\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}'\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} = [\mathbf{A}_1' \quad \mathbf{A}_2']\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} = [\mathbf{A}_1'\mathbf{A}_1 + \mathbf{A}_2'\mathbf{A}_2], \tag{A-69}$$

and

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}' \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}'\mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}'\mathbf{A}_{22} \end{bmatrix}. \qquad \textbf{(A-70)}$$

A.5.2 DETERMINANTS OF PARTITIONED MATRICES

The determinant of a block-diagonal matrix is obtained analogously to that of a diagonal matrix:

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}| \times |\mathbf{A}_{22}|. \qquad \textbf{(A-71)}$$

The determinant of a general 2 × 2 partitioned matrix is

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{22}| \times |\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| = |\mathbf{A}_{11}| \times |\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|. \qquad \textbf{(A-72)}$$

A.5.3 INVERSES OF PARTITIONED MATRICES

The inverse of a block-diagonal matrix is

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix}, \qquad \textbf{(A-73)}$$

which can be verified by direct multiplication. For the general 2 × 2 partitioned matrix, one form of the **partitioned inverse** is

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1}(\mathbf{I} + \mathbf{A}_{12}\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1}) & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{F}_2 \\ -\mathbf{F}_2\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{F}_2 \end{bmatrix}, \qquad \textbf{(A-74)}$$

where

$$\mathbf{F}_2 = (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}.$$

The upper left block could also be written as

$$\mathbf{F}_1 = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}.$$

A.5.4 DEVIATIONS FROM MEANS

Suppose that we begin with a column vector of n values $\mathbf{x}$ and let

$$\mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{x} \\ \mathbf{x}'\mathbf{i} & \mathbf{x}'\mathbf{x} \end{bmatrix}.$$

We are interested in the lower-right-hand element of $\mathbf{A}^{-1}$. Upon using the definition of $\mathbf{F}_2$ in (A-74), this is

$$\mathbf{F}_2 = [\mathbf{x}'\mathbf{x} - (\mathbf{x}'\mathbf{i})(\mathbf{i}'\mathbf{i})^{-1}(\mathbf{i}'\mathbf{x})]^{-1} = \left\{\mathbf{x}'\left[\mathbf{Ix} - \mathbf{i}\left(\frac{1}{n}\right)\mathbf{i}'\mathbf{x}\right]\right\}^{-1}$$

$$= \left\{\mathbf{x}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{ii}'\right]\mathbf{x}\right\}^{-1} = (\mathbf{x}'\mathbf{M}^0\mathbf{x})^{-1}.$$

Therefore, the lower-right-hand value in the inverse matrix is

$$(\mathbf{x}'\mathbf{M}^0\mathbf{x})^{-1} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = a^{22}.$$

Now, suppose that we replace **x** with **X**, a matrix with several columns. We seek the lower-right block of $(\mathbf{Z}'\mathbf{Z})^{-1}$, where $\mathbf{Z} = [\mathbf{i}, \mathbf{X}]$. The analogous result is

$$(\mathbf{Z}'\mathbf{Z})^{22} = [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}'\mathbf{X}]^{-1} = (\mathbf{X}'\mathbf{M}^0\mathbf{X})^{-1},$$

which implies that the $K \times K$ matrix in the lower-right corner of $(\mathbf{Z}'\mathbf{Z})^{-1}$ is the inverse of the $K \times K$ matrix whose jkth element is $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$. Thus, when a data matrix contains a column of ones, the elements of the inverse of the matrix of sums of squares and cross products will be computed from the original data in the form of deviations from the respective column means.

A.5.5 KRONECKER PRODUCTS

A calculation that helps to condense the notation when dealing with sets of regression models(see Chapter 10) is the **Kronecker product**. For general matrices **A** and **B**,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1K}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2K}\mathbf{B} \\ & & \cdots & \\ a_{n1}\mathbf{B} & a_{n2}\mathbf{B} & \cdots & a_{nK}\mathbf{B} \end{bmatrix}. \qquad \textbf{(A-75)}$$

Notice that there is no requirement for conformability in this operation. The Kronecker product can be computed for any pair of matrices. If **A** is $K \times L$ and **B** is $m \times n$, then $\mathbf{A} \otimes \mathbf{B}$ is $(Km) \times (Ln)$.

For the Kronecker product,

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}), \qquad \textbf{(A-76)}$$

If **A** is $M \times M$ and **B** is $n \times n$, then

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^n|\mathbf{B}|^M,$$

$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}',$$

$$\text{trace}(\mathbf{A} \otimes \mathbf{B}) = \text{trace}(\mathbf{A})\,\text{trace}(\mathbf{B}).$$

(The trace of a matrix is defined in Section A.6.7.) For **A**, **B**, **C**, and **D** such that the products are defined,

$$(\mathbf{A} \otimes B)(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}.$$

A.6 CHARACTERISTIC ROOTS AND VECTORS

A useful set of results for analyzing a square matrix **A** arises from the solutions to the set of equations

$$\mathbf{Ac} = \lambda\mathbf{c}. \qquad \textbf{(A-77)}$$

The pairs of solutions $(\mathbf{c},\lambda)$ are the **characteristic vectors c** and **characteristic roots** λ. If $\mathbf{c}$ is any nonzero solution vector, then $k\mathbf{c}$ is also for any value of K. To remove the indeterminancy, $\mathbf{c}$ is **normalized** so that $\mathbf{c}'\mathbf{c} = 1$.

The solution then consists of λ and the $n - 1$ unknown elements in $\mathbf{c}$.

A.6.1 THE CHARACTERISTIC EQUATION

Solving (A-77) can, in principle, proceed as follows. First, (A-77) implies that

$$\mathbf{Ac} = \lambda\mathbf{Ic},$$

or that

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = \mathbf{0}.$$

This equation is a homogeneous system that has a nonzero solution only if the matrix $(\mathbf{A} - \lambda\mathbf{I})$ is singular or has a zero determinant. Therefore, if λ is a solution, then

$$|\mathbf{A} - \lambda\mathbf{I}| = 0. \qquad \textbf{(A-78)}$$

This polynomial in λ is the **characteristic equation** of **A**. For example, if

$$\mathbf{A} = \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix},$$

then

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 5-\lambda & 1 \\ 2 & 4-\lambda \end{vmatrix} = (5-\lambda)(4-\lambda) - 2(1) = \lambda^2 - 9\lambda + 18.$$

The two solutions are $\lambda = 6$ and $\lambda = 3$.

In solving the characteristic equation, there is no guarantee that the characteristic roots will be real. In the preceding example, if the 2 in the lower-left-hand corner of the matrix were -2 instead, then the solution would be a pair of complex values. The same result can emerge in the general $n \times n$ case. The characteristic roots of a symmetric matrix such as $\mathbf{X}'\mathbf{X}$ are real, however.[7] This result will be convenient because most of our applications will involve the characteristic roots and vectors of symmetric matrices.

For an $n \times n$ matrix, the characteristic equation is an nth-order polynomial in λ. Its solutions may be n distinct values, as in the preceding example, or may contain repeated values of λ, and may contain some zeros as well.

A.6.2 CHARACTERISTIC VECTORS

With λ in hand, the characteristic vectors are derived from the original problem,

$$\mathbf{Ac} = \lambda\mathbf{c},$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = \mathbf{0}. \qquad \textbf{(A-79)}$$

Neither pair determines the values of c_1 and c_2. But this result was to be expected; it was the reason $\mathbf{c}'\mathbf{c} = 1$ was specified at the outset. The additional equation $\mathbf{c}'\mathbf{c} = 1$, however, produces complete solutions for the vectors.

[7] A proof may be found in Theil (1971).

A.6.3 GENERAL RESULTS FOR CHARACTERISTIC ROOTS AND VECTORS

A $K \times K$ symmetric matrix has K distinct characteristic vectors, $\mathbf{c}_1, \mathbf{c}_2, \ldots \mathbf{c}_K$. The corresponding characteristic roots, $\lambda_1, \lambda_2, \ldots, \lambda_K$, although real, need not be distinct. The characteristic vectors of a symmetric matrix are orthogonal,[8] which implies that for every $i \neq j$, $\mathbf{c}_i'\mathbf{c}_j = 0$.[9] It is convenient to collect the K-characteristic vectors in a $K \times K$ matrix whose ith column is the $\mathbf{c}_i$ corresponding to λ_i,

$$\mathbf{C} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad c_K],$$

and the K-characteristic roots in the same order, in a diagonal matrix,

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_K \end{bmatrix}.$$

Then, the full set of equations

$$\mathbf{A}\mathbf{c}_k = \lambda_k \mathbf{c}_k$$

is contained in

$$\mathbf{AC} = \mathbf{C}\boldsymbol{\Lambda}. \qquad \textbf{(A-80)}$$

Because the vectors are orthogonal and $\mathbf{c}_i'\mathbf{c}_i = 1$, we have

$$\mathbf{C}'\mathbf{C} = \begin{bmatrix} \mathbf{c}_1'\mathbf{c}_1 & \mathbf{c}_1'\mathbf{c}_2 & \cdots & \mathbf{c}_1'\mathbf{c}_K \\ \mathbf{c}_2'\mathbf{c}_1 & \mathbf{c}_2'\mathbf{c}_2 & \cdots & \mathbf{c}_2'\mathbf{c}_K \\ & & \vdots & \\ \mathbf{c}_K'\mathbf{c}_1 & \mathbf{c}_K'\mathbf{c}_2 & \cdots & \mathbf{c}_K'\mathbf{c}_K \end{bmatrix} = \mathbf{I}. \qquad \textbf{(A-81)}$$

Result (A-81) implies that

$$\mathbf{C}' = \mathbf{C}^{-1}. \qquad \textbf{(A-82)}$$

Consequently,

$$\mathbf{CC}' = \mathbf{CC}^{-1} = \mathbf{I} \qquad \textbf{(A-83)}$$

as well, so the rows as well as the columns of $\mathbf{C}$ are orthogonal.

A.6.4 DIAGONALIZATION AND SPECTRAL DECOMPOSITION OF A MATRIX

By premultiplying (A-80) by $\mathbf{C}'$ and using (A-81), we can extract the characteristic roots of $\mathbf{A}$.

[8] For proofs of these propositions, see Strang (1988–2014).

[9] This statement is not true if the matrix is not symmetric. For instance, it does not hold for the characteristic vectors computed in the first example. For nonsymmetric matrices, there is also a distinction between "right" characteristic vectors, $\mathbf{Ac} = \lambda c$, and "left" characteristic vectors, $\mathbf{d}'\mathbf{A} = \lambda\mathbf{d}'$, which may not be equal.

DEFINITION A.15 Diagonalization of a Matrix
*The **diagonalization** of a matrix* **A** *is*

$$\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{C}'\mathbf{C}\boldsymbol{\Lambda} = \mathbf{I}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}. \quad \textbf{(A-84)}$$

Alternatively, by *post* multiplying (A-80) by $\mathbf{C}'$ and using (A-83), we obtain a useful representation of **A**.

DEFINITION A.16 Spectral Decomposition of a Matrix
The spectral decomposition of **A** *is*

$$\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}' = \sum_{k=1}^{K} \lambda_k \mathbf{c}_k \mathbf{c}_k'. \quad \textbf{(A-85)}$$

In this representation, the $K \times K$ matrix **A** is written as a sum of K rank one matrices. This sum is also called the **eigenvalue** (or, "own" value) decomposition of **A**. In this connection, the term *signature* of the matrix is sometimes used to describe the characteristic roots and vectors. Yet another pair of terms for the parts of this decomposition are the **latent roots** and **latent vectors** of **A**.

A.6.5 RANK OF A MATRIX

The diagonalization result enables us to obtain the rank of a matrix very easily. To do so, we can use the following result.

THEOREM A.3 Rank of a Product
For any matrix **A** *and nonsingular matrices* **B** *and* **C**, *the rank of* **BAC** *is equal to the rank of* **A**. ***Proof:*** *By (A-45),* $\text{rank}(\mathbf{BAC}) = \text{rank}[(\mathbf{BA})\mathbf{C}] = \text{rank}(\mathbf{BA})$. *By (A-43),* $\text{rank}(\mathbf{BA}) = \text{rank}(\mathbf{A}'\mathbf{B}')$, *and applying (A-45) again,* $\text{rank}(\mathbf{A}'\mathbf{B}') = \text{rank}(\mathbf{A}')$ *because* $\mathbf{B}'$ *is nonsingular if* **B** *is nonsingular [once again, by (A-43)]. Finally, applying (A-43) again to obtain* $\text{rank}(\mathbf{A}') = \text{rank}(\mathbf{A})$ *gives the result.*

Because **C** and $\mathbf{C}'$ are nonsingular, we can use them to apply this result to (A-84). By an obvious substitution,

$$\text{rank}(\mathbf{A}) = \text{rank}(\boldsymbol{\Lambda}). \quad \textbf{(A-86)}$$

Finding the rank of $\boldsymbol{\Lambda}$ is trivial. Because $\boldsymbol{\Lambda}$ is a diagonal matrix, its rank is just the number of nonzero values on its diagonal. By extending this result, we can prove the following theorems. (Proofs are brief and are left for the reader.)

THEOREM A.4 Rank of a Symmetric Matrix
The rank of a symmetric matrix is the number of nonzero characteristic roots it contains.

Note how this result enters the spectral decomposition given earlier. If any of the characteristic roots are zero, then the number of rank one matrices in the sum is reduced correspondingly. It would appear that this simple rule will not be useful if **A** is not square. But recall that

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}). \tag{A-87}$$

Because $\mathbf{A}'\mathbf{A}$ is always square, we can use it instead of **A**. Indeed, we can use it even if **A** is not square, which leads to a fully general result.

THEOREM A.5 Rank of a Matrix
The rank of any matrix **A** *equals the number of nonzero characteristic roots in* $\mathbf{A}'\mathbf{A}$.

The row rank and column rank of a matrix are equal, so we should be able to apply Theorem A.5 to $\mathbf{AA}'$ as well. This process, however, requires an additional result.

THEOREM A.6 Roots of an Outer Product Matrix
The nonzero characteristic roots of $\mathbf{AA}'$ *are the same as those of* $\mathbf{A}'\mathbf{A}$.

The proof is left as an exercise. A useful special case the reader can examine is the characteristic roots of $\mathbf{aa}'$ and $\mathbf{a}'\mathbf{a}$, where **a** is an $n \times 1$ vector.

If a characteristic root of a matrix is zero, then we have $\mathbf{Ac} = \mathbf{0}$. Thus, if the matrix has a zero root, it must be singular. Otherwise, no nonzero **c** would exist. In general, therefore, a matrix is singular; that is, it does not have full rank if and only if it has at least one zero root.

A.6.6 CONDITION NUMBER OF A MATRIX

As the preceding might suggest, there is a discrete difference between full rank and short rank matrices. In analyzing data matrices such as the one in Section A.2, however, we shall often encounter cases in which a matrix is not quite short ranked, because it has all nonzero roots, but it is close. That is, by some measure, we can come very close to being able to write one column as a linear combination of the others. This case is important; we shall examine it at length in our discussion of multicollinearity in Section 4.9.1. Our definitions of rank and determinant will fail to indicate this possibility, but an alternative measure, the **condition number**, is designed for that purpose. Formally, the condition number for a square matrix **A** is

$$\gamma = \left[\frac{\text{maximum root}}{\text{minimum root}}\right]^{1/2}. \tag{A-88}$$

For nonsquare matrices $\mathbf{X}$, such as the data matrix in the example, we use $\mathbf{A} = \mathbf{X}'\mathbf{X}$. As a further refinement, because the characteristic roots are affected by the scaling of the columns of $\mathbf{X}$, we scale the columns to have length 1 by dividing each column by its norm [see (A-55)]. For the $\mathbf{X}$ in Section A.2, the largest characteristic root of $\mathbf{A}$ is 4.9255 and the smallest is 0.0001543. Therefore, the condition number is 178.67, which is extremely large. (Values greater than 20 are large.) That the smallest root is close to zero compared with the largest means that this matrix is nearly singular. Matrices with large condition numbers are difficult to invert accurately.

A.6.7 TRACE OF A MATRIX

The **trace** of a square $K \times K$ matrix is the sum of its diagonal elements:

$$\operatorname{tr}(\mathbf{A}) = \sum_{k=1}^{K} a_{kk}.$$

Some easily proven results are

$$\operatorname{tr}(c\mathbf{A}) = c(\operatorname{tr}(\mathbf{A})), \tag{A-89}$$

$$\operatorname{tr}(\mathbf{A}') = \operatorname{tr}(\mathbf{A}), \tag{A-90}$$

$$\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}), \tag{A-91}$$

$$\operatorname{tr}(\mathbf{I}_K) = K. \tag{A-92}$$

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}). \tag{A-93}$$

$$\mathbf{a}'\mathbf{a} = \operatorname{tr}(\mathbf{a}'\mathbf{a}) = \operatorname{tr}(\mathbf{aa}')$$

$$\operatorname{tr}(\mathbf{A}'\mathbf{A}) = \sum_{k=1}^{K} \mathbf{a}_k'\mathbf{a}_k = \sum_{i=1}^{K}\sum_{k=1}^{K} a_{ik}^2.$$

The permutation rule can be extended to any *cyclic* permutation in a product:

$$\operatorname{tr}(\mathbf{ABCD}) = \operatorname{tr}(\mathbf{BCDA}) = \operatorname{tr}(\mathbf{CDAB}) = \operatorname{tr}(\mathbf{DABC}). \tag{A-94}$$

By using (A-84), we obtain

$$\operatorname{tr}(\mathbf{C}'\mathbf{AC}) = \operatorname{tr}(\mathbf{ACC}') = \operatorname{tr}(\mathbf{AI}) = \operatorname{tr}(\mathbf{A}) = \operatorname{tr}(\mathbf{\Lambda}). \tag{A-95}$$

Because $\mathbf{\Lambda}$ is diagonal with the roots of $\mathbf{A}$ on its diagonal, the general result is the following.

THEOREM A.7 Trace of a Matrix

The trace of a matrix equals the sum of its characteristic roots. **(A-96)**

A.6.8 DETERMINANT OF A MATRIX

Recalling how tedious the calculation of a determinant promised to be, we find that the following is particularly useful. Because

$$\begin{aligned} \mathbf{C}'\mathbf{A}\mathbf{C} &= \mathbf{\Lambda}, \\ |\mathbf{C}'\mathbf{A}\mathbf{C}| &= |\mathbf{\Lambda}|. \end{aligned} \tag{A-97}$$

Using a number of earlier results, we have, for orthogonal matrix **C**,

$$\begin{aligned} |\mathbf{C}'\mathbf{A}\mathbf{C}| &= |\mathbf{C}'|\cdot|\mathbf{A}|\cdot|\mathbf{C}| = |\mathbf{C}'|\cdot|\mathbf{C}|\cdot|\mathbf{A}| = |\mathbf{C}'\mathbf{C}|\cdot|\mathbf{A}| = |\mathbf{I}|\cdot|\mathbf{A}| = 1\cdot|\mathbf{A}| \\ &= |\mathbf{A}| \\ &= |\mathbf{\Lambda}|. \end{aligned} \tag{A-98}$$

Because $|\mathbf{\Lambda}|$ is just the product of its diagonal elements, the following is implied.

THEOREM A.8 Determinant of a Matrix

The determinant of a matrix equals the product of its characteristic roots. **(A-99)**

Notice that we get the expected result if any of these roots is zero. The determinant is the product of the roots, so it follows that a matrix is singular if and only if its determinant is zero and, in turn, if and only if it has at least one zero characteristic root.

A.6.9 POWERS OF A MATRIX

We often use expressions involving powers of matrices, such as $\mathbf{AA} = \mathbf{A}^2$. For positive integer powers, these expressions can be computed by repeated multiplication. But this does not show how to handle a problem such as finding a **B** such that $\mathbf{B}^2 = \mathbf{A}$, that is, the square root of a matrix. The characteristic roots and vectors provide a solution. Consider, first

$$\mathbf{AA} = \mathbf{A}^2 = (\mathbf{C\Lambda C}')(\mathbf{C\Lambda C}') = \mathbf{C\Lambda C}'\mathbf{C\Lambda C}' = \mathbf{C\Lambda I\Lambda C}' = \mathbf{C\Lambda\Lambda C}' = \mathbf{C\Lambda}^2\mathbf{C}'. \tag{A-100}$$

Two results follow. Because $\mathbf{\Lambda}^2$ is a diagonal matrix whose nonzero elements are the squares of those in $\mathbf{\Lambda}$, the following is implied.

For any symmetric matrix, the characteristic roots of $\mathbf{A}^2$ *are the squares of those of* **A**, *and the characteristic vectors are the same.* **(A-101)**

The proof is obtained by observing that the last result in (A-100) is the spectral decomposition of the matrix $\mathbf{B} = \mathbf{AA}$. Because $\mathbf{A}^3 = \mathbf{AA}^2$ and so on, (A-101) extends to any positive integer. By convention, for any **A**, $\mathbf{A}^0 = \mathbf{I}$. Thus, for any symmetric matrix **A**, $\mathbf{A}^K = \mathbf{C\Lambda}^K\mathbf{C}'$, $K = 0, 1, \ldots$. Hence, the characteristic roots of $\mathbf{A}^K$ are λ^K, whereas the characteristic vectors are the same as those of **A**. If **A** is nonsingular, so that all its roots λ_i are nonzero, then this proof can be extended to negative powers as well.

If $\mathbf{A}^{-1}$ exists, then

$$\mathbf{A}^{-1} = (\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}')^{-1} = (\mathbf{C}')^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{C}^{-1} = \mathbf{C}\boldsymbol{\Lambda}^{-1}\mathbf{C}', \qquad \textbf{(A-102)}$$

where we have used the earlier result, $\mathbf{C}' = \mathbf{C}^{-1}$. This gives an important result that is useful for analyzing inverse matrices.

THEOREM A.9 Characteristic Roots of an Inverse Matrix
If $\mathbf{A}^{-1}$ exists, then the characteristic roots of $\mathbf{A}^{-1}$ are the reciprocals of those of $\mathbf{A}$, and the characteristic vectors are the same.

By extending the notion of repeated multiplication, we now have a more general result.

THEOREM A.10 Characteristic Roots of a Matrix Power
For any nonsingular symmetric matrix $\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$, $\mathbf{A}^K = \mathbf{C}\boldsymbol{\Lambda}^K\mathbf{C}'$, $K = \ldots, -2, -1, 0, 1, 2, \ldots .$

We now turn to the general problem of how to compute the square root of a matrix. In the scalar case, the value would have to be nonnegative. The matrix analog to this requirement is that all the characteristic roots are nonnegative. Consider, then, the candidate

$$\mathbf{A}^{1/2} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}\mathbf{C} = \mathbf{C}\begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}\mathbf{C}'. \qquad \textbf{(A-103)}$$

This equation satisfies the requirement for a square root, because

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{C}\boldsymbol{\Lambda}^{1/2}\mathbf{C}'\mathbf{C}\boldsymbol{\Lambda}^{1/2}\mathbf{C}' = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}' = \mathbf{A}. \qquad \textbf{(A-104)}$$

If we continue in this fashion, we can define the nonnegative powers of a matrix more generally, still assuming that all the characteristic roots are nonnegative. For example, $\mathbf{A}^{1/3} = \mathbf{C}\boldsymbol{\Lambda}^{1/3}\mathbf{C}'$. If all the roots are strictly positive, we can go one step further and extend the result to any real power. For reasons that will be made clear in the next section, we say that a matrix with positive characteristic roots is **positive definite.** It is the matrix analog to a positive number.

DEFINITION A.17 Real Powers of a Positive Definite Matrix

*For a **positive definite** matrix $\mathbf{A}$, $\mathbf{A}^r = \mathbf{C}\boldsymbol{\Lambda}^r\mathbf{C}'$, for any real number, r.* **(A-105)**

The characteristic roots of $\mathbf{A}^r$ are the rth power of those of $\mathbf{A}$, and the characteristic vectors are the same.

If $\mathbf{A}$ is only **nonnegative definite**—that is, has roots that are either zero or positive—then (A-105) holds only for nonnegative r.

A.6.10 IDEMPOTENT MATRICES

Idempotent matrices are equal to their squares [see (A-37) to (A-39)]. In view of their importance in econometrics, we collect a few results related to idempotent matrices at this point. First, (A-101) implies that if λ is a characteristic root of an idempotent matrix, then $\lambda = \lambda^K$ for all nonnegative integers K. As such, if $\mathbf{A}$ is a symmetric idempotent matrix, then all its roots are one or zero. Assume that all the roots of $\mathbf{A}$ are one. Then $\mathbf{\Lambda} = \mathbf{I}$, and $\mathbf{A} = \mathbf{C\Lambda C}' = \mathbf{CIC}' = \mathbf{CC}' = \mathbf{I}$. If the roots are not all one, then one or more are zero. Consequently, we have the following results for symmetric idempotent matrices:[10]

- *The only full rank, symmetric idempotent matrix is the identity matrix* **I**. **(A-106)**
- *All symmetric idempotent matrices except the identity matrix are singular*. **(A-107)**

The final result on idempotent matrices is obtained by observing that the count of the nonzero roots of $\mathbf{A}$ is also equal to their sum. By combining Theorems A.5 and A.7 with the result that for an idempotent matrix, the roots are all zero or one, we obtain this result:

- *The rank of a symmetric idempotent matrix is equal to its trace*. **(A-108)**

A.6.11 FACTORING A MATRIX: THE CHOLESKY DECOMPOSITION

In some applications, we shall require a matrix $\mathbf{P}$ such that

$$\mathbf{P}'\mathbf{P} = \mathbf{A}^{-1}.$$

One choice is

$$\mathbf{P} = \mathbf{\Lambda}^{-1/2}\mathbf{C}',$$

so that

$$\mathbf{P}'\mathbf{P} = (\mathbf{C}')'(\mathbf{\Lambda}^{-1/2})'\mathbf{\Lambda}^{-1/2}\mathbf{C}' = \mathbf{C\Lambda}^{-1}\mathbf{C}',$$

as desired.[11] Thus, the **spectral decomposition** of $\mathbf{A}$, $\mathbf{A} = \mathbf{C\Lambda C}'$ is a useful result for this kind of computation.

The **Cholesky factorization** of a symmetric positive definite matrix is an alternative representation that is useful in regression analysis. Any symmetric positive definite matrix $\mathbf{A}$ may be written as the product of a **lower triangular matrix L** and its transpose (which is an **upper triangular matrix**) $\mathbf{L}' = \mathbf{U}$. Thus, $\mathbf{A} = \mathbf{LU}$. This result is the Cholesky decomposition of $\mathbf{A}$. The square roots of the diagonal elements of $\mathbf{L}$, d_i, are the **Cholesky values** of $\mathbf{A}$. By arraying these in a diagonal matrix $\mathbf{D}$, we may also write $\mathbf{A} = \mathbf{LD}^{-1}\mathbf{D}^2\mathbf{D}^{-1}\mathbf{U} = \mathbf{L}^*\mathbf{D}^2\mathbf{U}^*$, which is similar to the spectral decomposition in (A-85). The usefulness of this formulation arises when the inverse of $\mathbf{A}$ is required. Once $\mathbf{L}$ is

[10] Not all idempotent matrices are symmetric. We shall not encounter any asymmetric ones in our work, however.

[11] We say that this is "one" choice because if $\mathbf{A}$ is symmetric, as it will be in all our applications, there are other candidates. The reader can easily verify that $\mathbf{C\Lambda}^{-1/2}\mathbf{C}' = \mathbf{A}^{-1/2}$ works as well.

computed, finding $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ is also straightforward as well as extremely fast and accurate. Most recently developed econometric software packages use this technique for inverting positive definite matrices.

A.6.12 SINGULAR VALUE DECOMPOSITION

A third type of decomposition of a matrix is useful for numerical analysis when the inverse is difficult to obtain because the columns of **A** are "nearly" collinear. Any $n \times K$ matrix **A** for which $n \geq K$ can be written in the form $\mathbf{A} = \mathbf{UWV}'$, where **U** is an orthogonal $n \times K$ matrix—that is, $\mathbf{U}'\mathbf{U} = \mathbf{I}_K$—**W** is a $K \times K$ diagonal matrix such that $w_i \geq 0$, and **V** is a $K \times K$ matrix such that $\mathbf{V}'\mathbf{V} = \mathbf{I}_K$. This result is called the **singular value decomposition** (SVD) of **A**, and w_i are the singular values of **A**.[12] (Note that if **A** is square, then the spectral decomposition is a singular value decomposition.) As with the Cholesky decomposition, the usefulness of the SVD arises in inversion, in this case, of $\mathbf{A}'\mathbf{A}$. By multiplying it out, we obtain that $(\mathbf{A}'\mathbf{A})^{-1}$ is simply $\mathbf{VW}^{-2}\mathbf{V}'$. Once the SVD of **A** is computed, the inversion is trivial. The other advantage of this format is its numerical stability, which is discussed at length in Press et al. (2007).

A.6.13 QR DECOMPOSITION

Press et al. (2007) recommend the SVD approach as the method of choice for solving least squares problems because of its accuracy and numerical stability. A commonly used alternative method similar to the SVD approach is the QR decomposition. Any $n \times K$ matrix, **X**, with $n \geq K$ can be written in the form $\mathbf{X} = \mathbf{QR}$ in which the columns of **Q** are orthonormal ($\mathbf{Q}'\mathbf{Q} = \mathbf{I}$) and **R** is an upper triangular matrix. Decomposing **X** in this fashion allows an extremely accurate solution to the least squares problem that does not involve inversion or direct solution of the normal equations. Press et al. suggest that this method may have problems with rounding errors in problems when **X** is nearly of short rank, but based on other published results, this concern seems relatively minor.[13]

A.6.14 THE GENERALIZED INVERSE OF A MATRIX

Inverse matrices are fundamental in econometrics. Although we shall not require them much in our treatment in this book, there are more general forms of inverse matrices than we have considered thus far. A **generalized inverse** of a matrix **A** is another matrix $\mathbf{A}^{+}$ that satisfies the following requirements:

1. $\mathbf{AA}^{+}\mathbf{A} = \mathbf{A}$.
2. $\mathbf{A}^{+}\mathbf{AA}^{+} = \mathbf{A}^{+}$.
3. $\mathbf{A}^{+}\mathbf{A}$ is symmetric.
4. $\mathbf{AA}^{+}$ is symmetric.

[12] Discussion of the singular value decomposition (and listings of computer programs for the computations) may be found in Press et al. (1986).

[13] The National Institute of Standards and Technology (NIST) has published a suite of benchmark problems that test the accuracy of least squares computations (http://www.nist.gov/itl/div898/strd). Using these problems, which include some extremely difficult, ill-conditioned data sets, we found that the QR method would reproduce all the NIST certified solutions to 15 digits of accuracy, which suggests that the QR method should be satisfactory for all but the worst problems. NIST's benchmark for hard to solve least squares problems, the "Filipelli problem," is solved accurately to at least 9 digits with the QR method. Evidently, other methods of least squares solution fail to produce an accurate result.

A unique $\mathbf{A}^+$ can be found for any matrix, whether $\mathbf{A}$ is singular or not, or even if $\mathbf{A}$ is not square.[14] The unique matrix that satisfies all four requirements is called the **Moore–Penrose inverse** or **pseudoinverse** of $\mathbf{A}$. If $\mathbf{A}$ happens to be square and nonsingular, then the generalized inverse will be the familiar ordinary inverse. But if $\mathbf{A}^{-1}$ does not exist, then $\mathbf{A}^+$ can still be computed.

An important special case is the overdetermined system of equations

$$\mathbf{Ab} = \mathbf{y},$$

where $\mathbf{A}$ has n rows, $K < n$ columns, and column rank equal to $R \leq K$. Suppose that R equals K, so that $(\mathbf{A}'\mathbf{A})^{-1}$ exists. Then the Moore–Penrose inverse of $\mathbf{A}$ is

$$\mathbf{A}^+ = (\mathbf{A}'\mathbf{A})^{-1}\,\mathbf{A}',$$

which can be verified by multiplication. A "solution" to the system of equations can be written

$$\mathbf{b} = \mathbf{A}^+\mathbf{y}.$$

This is the vector that minimizes the length of $\mathbf{Ab} - \mathbf{y}$. Recall this was the solution to the least squares problem obtained in Section A.4.4. If $\mathbf{y}$ lies in the column space of $\mathbf{A}$, this vector will be zero, but otherwise, it will not.

Now suppose that $\mathbf{A}$ does not have full rank. The previous solution cannot be computed. An alternative solution can be obtained, however. We continue to use the matrix $\mathbf{A}'\mathbf{A}$. In the spectral decomposition of Section A.6.4, if $\mathbf{A}$ has rank R, then there are R terms in the summation in (A-85). In (A-102), the spectral decomposition using the reciprocals of the characteristic roots is used to compute the inverse. To compute the Moore–Penrose inverse, we apply this calculation to $\mathbf{A}'\mathbf{A}$, using only the nonzero roots, then postmultiply the result by $\mathbf{A}'$. Let $\mathbf{C}_1$ be the R characteristic vectors corresponding to the nonzero roots, which we array in the diagonal matrix, $\mathbf{\Lambda}_1$. Then the Moore–Penrose inverse is

$$\mathbf{A}^+ = \mathbf{C}_1\mathbf{\Lambda}_1^{-1}\mathbf{C}_1'\mathbf{A}',$$

which is very similar to the previous result.

If $\mathbf{A}$ is a symmetric matrix with rank $R \leq K$, the Moore–Penrose inverse is computed precisely as in the preceding equation without postmultiplying by $\mathbf{A}'$. Thus, for a symmetric matrix **A,**

$$\mathbf{A}^+ = \mathbf{C}_1\mathbf{\Lambda}_1^{-1}\mathbf{C}_1',$$

where $\mathbf{\Lambda}_1^{-1}$ is a diagonal matrix containing the reciprocals of the *nonzero* roots of $\mathbf{A}$.

A.7 QUADRATIC FORMS AND DEFINITE MATRICES

Many optimization problems involve double sums of the form

$$q = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}. \qquad \textbf{(A-109)}$$

[14] A proof of uniqueness, with several other results, may be found in Theil (1983).

This **quadratic form** can be written

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}$$

where **A** is a symmetric matrix. In general, q may be positive, negative, or zero; it depends on **A** and **x**. There are some matrices, however, for which q will be positive regardless of **x**, and others for which q will always be negative (or nonnegative or nonpositive). For a given matrix **A**,

1. If $\mathbf{x}'\mathbf{A}\mathbf{x} > (<) \, 0$ for all nonzero **x**, then **A** is **positive (negative) definite.**
2. If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq (\leq) \, 0$ for all nonzero **x**, then **A** is **nonnegative definite** or **positive semidefinite** (nonpositive definite).

It might seem that it would be impossible to check a matrix for definiteness, since **x** can be chosen arbitrarily. But we have already used the set of results necessary to do so. Recall that a symmetric matrix can be decomposed into

$$\mathbf{A} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'.$$

Therefore, the quadratic form can be written as

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'\mathbf{x}.$$

Let $\mathbf{y} = \mathbf{C}'\mathbf{x}$. Then

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{y}'\boldsymbol{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2. \tag{A-110}$$

If λ_i is positive for all i, then regardless of **y**—that is, regardless of **x**—q will be positive. This case was identified earlier as a positive definite matrix. Continuing this line of reasoning, we obtain the following theorem.

THEOREM A.11 Definite Matrices
Let **A** *be a symmetric matrix. If all the characteristic roots of* **A** *are positive (negative), then* **A** *is* **positive definite (negative definite)**. *If some of the roots are zero, then* **A** *is* **nonnegative (nonpositive) definite** *if the remainder are positive (negative). If* **A** *has both negative and positive roots, then* **A** *is* **indefinite.**

The preceding statements give, in each case, the "if" parts of the theorem. To establish the "only if" parts, assume that the condition on the roots does not hold. This must lead to a contradiction. For example, if some λ can be negative, then $\mathbf{y}'\boldsymbol{\Lambda}\mathbf{y}$ could be negative for some **y**, so **A** cannot be positive definite.

A.7.1 NONNEGATIVE DEFINITE MATRICES

A case of particular interest is that of nonnegative definite matrices. Theorem A.11 implies a number of related results.

- If **A** is nonnegative definite, then $|\mathbf{A}| \geq 0$. **(A-111)**

  Proof: The determinant is the product of the roots, which are nonnegative.

The converse, however, is not true. For example, a 2×2 matrix with two negative roots is clearly not positive definite, but it does have a positive determinant.

- If $\mathbf{A}$ is positive definite, so is $\mathbf{A}^{-1}$. **(A-112)**

  Proof: The roots are the reciprocals of those of $\mathbf{A}$, which are, therefore positive.

- The identity matrix $\mathbf{I}$ is positive definite. **(A-113)**

  Proof: $\mathbf{x}'\mathbf{I}\mathbf{x} = \mathbf{x}'\mathbf{x} > 0 \quad \textbf{if} \quad \mathbf{x} \neq \mathbf{0}$.

A very important result for regression analysis is

- If $\mathbf{A}$ is $n \times K$ with full column rank and $n > K$, then $\mathbf{A}'\mathbf{A}$ is positive definite and $\mathbf{A}\mathbf{A}'$ is nonnegative definite. **(A-114)**

  Proof: By assumption, $\mathbf{A}\mathbf{x} \neq \mathbf{0}$. So $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x}) = \mathbf{y}'\mathbf{y} = \sum_j y_j^2 > 0$.

A similar proof establishes the nonnegative definiteness of $\mathbf{A}\mathbf{A}'$. The difference in the latter case is that because $\mathbf{A}$ has more rows than columns there is an $\mathbf{x}$ such that $\mathbf{A}'x = 0$. Thus, in the proof, we only have $\mathbf{y}'\mathbf{y} \geq 0$. The case in which $\mathbf{A}$ does not have full column rank is the same as that of $\mathbf{A}\mathbf{A}'$.

- If $\mathbf{A}$ is positive definite and $\mathbf{B}$ is a nonsingular matrix, then $\mathbf{B}'\mathbf{A}\mathbf{B}$ is positive definite. **(A-115)**

  Proof: $\mathbf{x}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{y}'\mathbf{A}\mathbf{y} > 0$, where $\mathbf{y} = \mathbf{B}\mathbf{x}$. But $\mathbf{y}$ cannot be $\mathbf{0}$ because $\mathbf{B}$ is nonsingular.

Finally, note that for $\mathbf{A}$ to be negative definite, all $\mathbf{A}$'s characteristic roots must be negative. But, in this case, $|\mathbf{A}|$ is positive if $\mathbf{A}$ is of even order and negative if $\mathbf{A}$ is of odd order.

A.7.2 IDEMPOTENT QUADRATIC FORMS

Quadratic forms in idempotent matrices play an important role in the distributions of many test statistics. As such, we shall encounter them fairly often. Two central results are of interest.

- Every symmetric idempotent matrix is nonnegative definite. **(A-116)**

  Proof: All roots are one or zero; hence, the matrix is nonnegative definite by definition.

Combining this with some earlier results yields a result used in determining the sampling distribution of most of the standard test statistics.

- If $\mathbf{A}$ is symmetric and idempotent, $n \times n$ with rank J, then every quadratic form in $\mathbf{A}$ can be written

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{j=1}^{J} y_j^2 \tag{A-117}$$

  Proof: This result is (A-110) with $\lambda =$ one or zero.

A.7.3 COMPARING MATRICES

Derivations in econometrics often focus on whether one matrix is "larger" than another. We now consider how to make such a comparison. As a starting point, the two matrices must have the same dimensions. A useful comparison is based on

$$d = \mathbf{x}'\mathbf{A}\mathbf{x} - \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}'(\mathbf{A} - \mathbf{B})\mathbf{x}.$$

If d is always positive for any nonzero vector, $\mathbf{x}$, then by this criterion, we can say that $\mathbf{A}$ is larger than $\mathbf{B}$. The reverse would apply if d is always negative. It follows from the definition that

$$\text{if } d > 0 \text{ for all nonzero } \mathbf{x}, \text{ then } \mathbf{A} - \mathbf{B} \text{ is positive definite.} \qquad \textbf{(A-118)}$$

If d is only greater than or equal to zero, then $\mathbf{A} - \mathbf{B}$ is nonnegative definite. The ordering is not complete. For some pairs of matrices, d could have either sign, depending on $\mathbf{x}$. In this case, there is no simple comparison.

A particular case of the general result which we will encounter frequently is.

$$\text{If } \mathbf{A} \text{ is positive definite and } \mathbf{B} \text{ is nonnegative definite, then } \mathbf{A} + \mathbf{B} \geq \mathbf{A}. \qquad \textbf{(A-119)}$$

Consider, for example, the "updating formula" introduced in (A-66). This uses a matrix

$$\mathbf{A} = \mathbf{B}'\mathbf{B} + \mathbf{b}\mathbf{b}' \geq \mathbf{B}'\mathbf{B}.$$

Finally, in comparing matrices, it may be more convenient to compare their inverses. The result analogous to a familiar result for scalars is:

$$\text{If } \mathbf{A} > \mathbf{B}, \text{ then } \mathbf{B}^{-1} > \mathbf{A}^{-1}. \qquad \textbf{(A-120)}$$

To establish this intuitive result, we would make use of the following, which is proved in Goldberger (1964, Chapter 2):

THEOREM A.12 Ordering for Positive Definite Matrices
If $\mathbf{A}$ *and* $\mathbf{B}$ *are two positive definite matrices with the same dimensions and if every characteristic root of* $\mathbf{A}$ *is larger than (at least as large as) the corresponding characteristic root of* $\mathbf{B}$ *when both sets of roots are ordered from largest to smallest, then* $\mathbf{A} - \mathbf{B}$ *is positive (nonnegative) definite.*

The roots of the inverse are the reciprocals of the roots of the original matrix, so the theorem can be applied to the inverse matrices.

A.8 CALCULUS AND MATRIX ALGEBRA[15]

A.8.1 DIFFERENTIATION AND THE TAYLOR SERIES

A variable y is a function of another variable x written

$$y = f(x), \quad y = g(x), \quad y = y(x),$$

[15] For a complete exposition, see Magnus and Neudecker (2007).

and so on, if each value of x is associated with a single value of y. In this relationship, y and x are sometimes labeled the **dependent variable** and the **independent variable**, respectively. Assuming that the function $f(x)$ is continuous and differentiable, we obtain the following derivatives:

$$f'(x) = \frac{dy}{dx}, f''(x) = \frac{d^2y}{dx^2},$$

and so on.

A frequent use of the derivatives of $f(x)$ is in the **Taylor series approximation**. A Taylor series is a polynomial approximation to $f(x)$. Letting x^0 be an arbitrarily chosen expansion point

$$f(x) \approx f(x^0) + \sum_{i=1}^{P} \frac{1}{i!} \frac{d^i f(x^0)}{d(x^0)^i} (x - x^0)^i. \qquad \textbf{(A-121)}$$

The choice of P, the number of terms, is arbitrary; the more that are used, the more accurate the approximation will be. The approximation used most frequently in econometrics is the **linear approximation**,

$$f(x) \approx \alpha + \beta x, \qquad \textbf{(A-122)}$$

where, by collecting terms in (A-121), $\alpha = [f(x^0) - f'(x^0)x^0]$ and $\beta = f'(x^0)$. The superscript "0" indicates that the function is evaluated at x^0. The **quadratic approximation** is

$$f(x) \approx \alpha + \beta x + \gamma x^2, \qquad \textbf{(A-123)}$$

where $\alpha = [f^0 - f'^0 x^0 + \frac{1}{2} f''^0 (x^0)^2]$, $\beta = [f'^0 - f''^0 x^0]$ and $\gamma = \frac{1}{2} f''^0$.

We can regard a function $y = f(x_1, x_2, \dots, x_n)$ as a **scalar-valued function** of a vector; that is, $y = f(\mathbf{x})$. The vector of partial derivatives, or **gradient vector**, or simply **gradient**, is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \cdots \\ \partial y/\partial x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{bmatrix}. \qquad \textbf{(A-124)}$$

The vector $\mathbf{g}(\mathbf{x})$ or $\mathbf{g}$ is used to represent the gradient. Notice that it is a column vector. The shape of the derivative is determined by the denominator of the derivative.

A **second derivatives matrix** or **Hessian** is computed as

$$\mathbf{H} = \begin{bmatrix} \partial^2 y/\partial x_1 \partial x_1 & \partial^2 y/\partial x_1 \partial x_2 & \cdots & \partial^2 y/\partial x_1 \partial x_n \\ \partial^2 y/\partial x_2 \partial x_1 & \partial^2 y/\partial x_2 \partial x_2 & \cdots & \partial^2 y/\partial x_2 \partial x_n \\ \cdots & \cdots & \cdots & \cdots \\ \partial^2 y/\partial x_n \partial x_1 & \partial^2 y/\partial x_n \partial x_2 & \cdots & \partial^2 y/\partial x_n \partial x_n \end{bmatrix} = [f_{ij}]. \qquad \textbf{(A-125)}$$

In general, $\mathbf{H}$ is a square, symmetric matrix. (The symmetry is obtained for continuous and continuously differentiable functions from Young's theorem.) Each column of $\mathbf{H}$ is the derivative of $\mathbf{g}$ with respect to the corresponding variable in $\mathbf{x}'$. Therefore,

$$\mathbf{H} = \left[\frac{\partial(\partial y/\partial \mathbf{x})}{\partial x_1}\frac{\partial(\partial y/\partial \mathbf{x})}{\partial x_2}\cdots\frac{\partial(\partial y/\partial \mathbf{x})}{\partial x_n}\right] = \frac{\partial(\partial y/\partial \mathbf{x})}{\partial(x_1 \quad x_2\cdots x_n)} = \frac{\partial(\partial y/\partial \mathbf{x})}{\partial \mathbf{x}'} = \frac{\partial^2 y}{\partial \mathbf{x}\partial \mathbf{x}'}.$$

The first-order, or linear Taylor series approximation is

$$y \approx f(\mathbf{x}^0) + \sum_{i=1}^{n} f_i(\mathbf{x}^0)(x_i - x_i^0). \tag{A-126}$$

The right-hand side is

$$f(\mathbf{x}^0) + \left[\frac{\partial f(\mathbf{x}^0)}{\partial \mathbf{x}^0}\right]'(\mathbf{x} - \mathbf{x}^0) = [f(\mathbf{x}^0) - \mathbf{g}(\mathbf{x}^0)'\mathbf{x}^0] + \mathbf{g}(\mathbf{x}^0)'\mathbf{x} = [f^0 - \mathbf{g}^{0\prime}\mathbf{x}^0] + \mathbf{g}^{0\prime}\mathbf{x}.$$

This produces the linear approximation,

$$y \approx \alpha + \beta'\mathbf{x}.$$

The second-order, or quadratic, approximation adds the second-order terms in the expansion,

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}^0(x_i - x_i^0)(x_j - x_j^0) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^0)'\mathbf{H}^0(\mathbf{x} - \mathbf{x}^0),$$

to the preceding one. Collecting terms in the same manner as in (A-126), we have

$$y \approx \alpha + \beta'\mathbf{x} + \frac{1}{2}\mathbf{x}'\boldsymbol{\Gamma}\mathbf{x}, \tag{A-127}$$

where

$$\alpha = f^0 - \mathbf{g}^{0\prime}\mathbf{x}^0 + \frac{1}{2}\mathbf{x}^{0\prime}\mathbf{H}^0\mathbf{x}^0, \quad \beta = \mathbf{g}^0 - \mathbf{H}^0\mathbf{x}^0 \quad \text{and} \quad \boldsymbol{\Gamma} = \mathbf{H}^0.$$

A linear function can be written

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

so

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}. \tag{A-128}$$

Note, in particular, that $\partial(\mathbf{a}'\mathbf{x})/\partial\mathbf{x} = \mathbf{a}$, not $\mathbf{a}'$. In a set of linear functions

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

each element y_i of $\mathbf{y}$ is

$$y_i = \mathbf{a}_i'\mathbf{x},$$

where $\mathbf{a}_i'$ is the ith row of $\mathbf{A}$ [see (A-14)]. Therefore,

$$\frac{\partial y_i}{\partial \mathbf{x}} = \mathbf{a}_i = \text{transpose of } i\text{th row of } \mathbf{A},$$

and

$$\begin{bmatrix} \partial y_1/\partial \mathbf{x}' \\ \partial y_2/\partial \mathbf{x}' \\ \cdots \\ \partial y_n/\partial \mathbf{x}' \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \cdots \\ \mathbf{a}_n' \end{bmatrix}.$$

Collecting all terms, we find that $\partial \mathbf{Ax}/\partial \mathbf{x}' = \mathbf{A}$, whereas the more familiar form will be

$$\frac{\partial \mathbf{x}'\mathbf{A}'}{\partial \mathbf{x}} = \mathbf{A}'. \qquad \textbf{(A-129)}$$

A quadratic form is written

$$\mathbf{x}'\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}. \qquad \textbf{(A-130)}$$

For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix},$$

so that

$$\mathbf{x}'\mathbf{Ax} = 1x_1^2 + 4x_2^2 + 6x_1x_2.$$

Then

$$\frac{\partial \mathbf{x}'\mathbf{Ax}}{\partial \mathbf{x}} = \begin{bmatrix} 2x_1 + 6x_2 \\ 6x_1 + 8x_2 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 6 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{Ax}, \qquad \textbf{(A-131)}$$

which is the general result when **A** is a symmetric matrix. If **A** is not symmetric, then

$$\frac{\partial(\mathbf{x}'\mathbf{Ax})}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}. \qquad \textbf{(A-132)}$$

Referring to the preceding double summation, we find that for each term, the coefficient on a_{ij} is $x_i x_j$. Therefore,

$$\frac{\partial(\mathbf{x}'\mathbf{Ax})}{\partial a_{ij}} = x_i x_j.$$

The square matrix whose $i\,j$th element is $x_i x_j$ is $\mathbf{xx}'$, so

$$\frac{\partial(\mathbf{x}'\mathbf{Ax})}{\partial \mathbf{A}} = \mathbf{xx}'. \qquad \textbf{(A-133)}$$

Derivatives involving determinants appear in maximum likelihood estimation. From the cofactor expansion in (A-51),

$$\frac{\partial|\mathbf{A}|}{\partial a_{ij}} = (-1)^{i+j}|\mathbf{A}_{ij}| = c_{ij}$$

where $|\mathbf{C}_{ji}|$ is the jith cofactor in **A**. The inverse of **A** can be computed using

$$\mathbf{A}_{ij}^{-1} = \frac{|\mathbf{C}_{ji}|}{|\mathbf{A}|}$$

(note the reversal of the subscripts), which implies that

$$\frac{\partial \ln |\mathbf{A}|}{\partial a_{ij}} = \frac{(-1)^{i+j}\,|\mathbf{A}_{ij}|}{|\mathbf{A}|},$$

or, collecting terms,

$$\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{-1\prime}.$$

Because the matrices for which we shall make use of this calculation will be symmetric in our applications, the transposition will be unnecessary.

A.8.2 OPTIMIZATION

Consider finding the x where $f(x)$ is maximized or minimized. Because $f'(x)$ is the slope of $f(x)$, either optimum must occur where $f'(x) = 0$. Otherwise, the function will be increasing or decreasing at x. This result implies the **first-order or necessary condition for an optimum** (maximum or minimum):

$$\frac{dy}{dx} = 0. \tag{A-134}$$

For a maximum, the function must be concave; for a minimum, it must be convex. The **sufficient condition for an optimum** is.

$$\begin{aligned} \text{For a maximum, } \frac{d^2y}{dx^2} &< 0; \\ \text{for a minimum, } \frac{d^2y}{dx^2} &> 0. \end{aligned} \tag{A-135}$$

Some functions, such as the sine and cosine functions, have many **local optima**, that is, many minima and maxima. A function such as $(\cos x)/(1 + x^2)$, which is a damped cosine wave, does as well but differs in that although it has many local maxima, it has one, at $x = 0$, at which $f(x)$ is greater than it is at any other point. Thus, $x = 0$ is the **global maximum**, whereas the other maxima are only **local maxima**. Certain functions, such as a quadratic, have only a single optimum. These functions are **globally concave** if the optimum is a maximum and **globally convex** if it is a minimum.

For maximizing or minimizing a function of several variables, the first-order conditions are

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}. \tag{A-136}$$

This result is interpreted in the same manner as the necessary condition in the univariate case. At the optimum, it must be true that no small change in any variable leads to an improvement in the function value. In the single-variable case, d^2y/dx^2 must be positive for a minimum and negative for a maximum. The second-order condition for an optimum in the multivariate case is that, at the optimizing value,

$$\mathbf{H} = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}\,\partial \mathbf{x}'} \tag{A-137}$$

must be positive definite for a minimum and negative definite for a maximum.

In a single-variable problem, the second-order condition can usually be verified by inspection. This situation will not generally be true in the multivariate case. As discussed earlier, checking the definiteness of a matrix is, in general, a difficult problem. For most of the problems encountered in econometrics, however, the second-order condition will be implied by the structure of the problem. That is, the matrix **H** will usually be of such a form that it is always definite.

For an example of the preceding, consider the problem

$$\text{maximize}_{\mathbf{x}} R = \mathbf{a}'\mathbf{x} - \mathbf{x}'\mathbf{A}\mathbf{x},$$

where

$$\mathbf{a}' = (5 \quad 4 \quad 2),$$

and

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 3 & 2 \\ 3 & 2 & 5 \end{bmatrix}.$$

Using some now familiar results, we obtain

$$\frac{\partial R}{\partial \mathbf{x}} = \mathbf{a} - 2\mathbf{A}\mathbf{x} = \begin{bmatrix} 5 \\ 4 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 4 \\ 6 & 4 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{0}. \qquad \textbf{(A-138)}$$

The solutions are

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 4 \\ 6 & 4 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 11.25 \\ 1.75 \\ -7.25 \end{bmatrix}.$$

The sufficient condition is that

$$\frac{\partial^2 R(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}'} = -2\mathbf{A} = \begin{bmatrix} -4 & -2 & -6 \\ -2 & -6 & -4 \\ -6 & -4 & -10 \end{bmatrix} \qquad \textbf{(A-139)}$$

must be negative definite. The three characteristic roots of this matrix are −15.746, −4, and −0.25403. Because all three roots are negative, the matrix is negative definite, as required.

In the preceding, it was necessary to compute the characteristic roots of the Hessian to verify the sufficient condition. For a general matrix of order larger than 2, this will normally require a computer. Suppose, however, that **A** is of the form

$$\mathbf{A} = \mathbf{B}'\mathbf{B},$$

where **B** is some known matrix. Then, as shown earlier, we know that **A** will always be positive definite (assuming that **B** has full rank). In this case, it is not necessary to calculate the characteristic roots of **A** to verify the sufficient conditions.

A.8.3 CONSTRAINED OPTIMIZATION

It is often necessary to solve an optimization problem subject to some constraints on the solution. One method is merely to "solve out" the constraints. For example, in the maximization problem considered earlier, suppose that the constraint $x_1 = x_2 - x_3$ is imposed on the solution. For a single constraint such as this one, it is possible merely to substitute the right-hand side of this equation for x_1 in the objective function and solve the resulting problem as a function of the remaining two variables. For more general constraints, however, or when there is more than one constraint, the method of Lagrange multipliers provides a more straightforward method of solving the problem. We seek to

$$\begin{aligned} \text{maximize}_{\mathbf{x}}\, f(\mathbf{x}) \text{ subject to } c_1(\mathbf{x}) &= 0 \\ c_2(\mathbf{x}) &= 0, \\ &\cdots \\ c_J(\mathbf{x}) &= 0. \end{aligned} \qquad \textbf{(A-140)}$$

The Lagrangean approach to this problem is to find the stationary points—that is, the points at which the derivatives are zero—of

$$L^*(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{j=1}^{J} \lambda_j c_j(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}'\mathbf{c}(\mathbf{x}). \qquad \textbf{(A-141)}$$

The solutions satisfy the equations

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{x}} &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \frac{\partial \boldsymbol{\lambda}'\mathbf{c}(\mathbf{x})}{\partial \mathbf{x}} = \mathbf{0}(n \times 1), \\ \frac{\partial L^*}{\partial \boldsymbol{\lambda}} &= \mathbf{c}(\mathbf{x}) = \mathbf{0}\ (J \times 1). \end{aligned} \qquad \textbf{(A-142)}$$

The second term in $\partial L^*/\partial \mathbf{x}$ is

$$\frac{\partial \boldsymbol{\lambda}'\mathbf{c}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{c}(\mathbf{x})'\boldsymbol{\lambda}}{\partial \mathbf{x}} = \left[\frac{\partial \mathbf{c}(\mathbf{x})'}{\partial \mathbf{x}}\right]\boldsymbol{\lambda} = \mathbf{C}'\boldsymbol{\lambda}, \qquad \textbf{(A-143)}$$

where $\mathbf{C}$ is the matrix of derivatives of the constraints with respect to $\mathbf{x}$. The jth row of the $J \times n$ matrix $\mathbf{C}$ is the vector of derivatives of the jth constraint, $c_j(\mathbf{x})$, with respect to $\mathbf{x}'$. Upon collecting terms, the first-order conditions are

$$\begin{aligned} \frac{\partial L^*}{\partial \mathbf{x}} &= \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0}, \\ \frac{\partial L^*}{\partial \lambda} &= \mathbf{c}(\mathbf{x}) = \mathbf{0}. \end{aligned} \qquad \textbf{(A-144)}$$

There is one very important aspect of the constrained solution to consider. In the unconstrained solution, we have $\partial f(\mathbf{x})/\partial \mathbf{x} = \mathbf{0}$. From (A-144), we obtain, for a constrained solution,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = -\mathbf{C}'\boldsymbol{\lambda}, \qquad \textbf{(A-145)}$$

which will not equal $\mathbf{0}$ unless $\boldsymbol{\lambda} = \mathbf{0}$. This result has two important implications:

- The constrained solution cannot be superior to the unconstrained solution. This is implied by the nonzero gradient at the constrained solution. (That is, unless $\mathbf{C} = \mathbf{0}$ which could happen if the constraints were nonlinear. But, even if so, the solution is still not better than the unconstrained optimum.)
- If the Lagrange multipliers are zero, then the constrained solution will equal the unconstrained solution.

To continue the example begun earlier, suppose that we add the following conditions:

$$x_1 - x_2 + x_3 = 0,$$
$$x_1 + x_2 + x_3 = 0.$$

To put this in the format of the general problem, write the constraints as $\mathbf{c}(\mathbf{x}) = \mathbf{Cx} = \mathbf{0}$, where

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

The Lagrangean function is

$$R^*(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{a}'\mathbf{x} - \mathbf{x}'\mathbf{Ax} + \boldsymbol{\lambda}'\mathbf{Cx}.$$

Note the dimensions and arrangement of the various parts. In particular, $\mathbf{C}$ is a 2 × 3 matrix, with one row for each constraint and one column for each variable in the objective function. The vector of Lagrange multipliers thus has two elements, one for each constraint. The necessary conditions are

$$\mathbf{a} - 2\mathbf{Ax} + \mathbf{C}'\boldsymbol{\lambda} = \mathbf{0} \quad \text{(three equations)}, \tag{A-146}$$

and

$$\mathbf{Cx} = \mathbf{0} \quad \text{(two equations)}.$$

These may be combined in the single equation

$$\begin{bmatrix} -2\mathbf{A} & \mathbf{C}' \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\mathbf{a} \\ \mathbf{0} \end{bmatrix}.$$

Using the partitioned inverse of (A-74) produces the solutions

$$\boldsymbol{\lambda} = -[\mathbf{CA}^{-1}\mathbf{C}']^{-1}\,\mathbf{CA}^{-1}\mathbf{a} \tag{A-147}$$

and

$$\mathbf{x} = \frac{1}{2}\mathbf{A}^{-1}[\mathbf{I} - \mathbf{C}'(\mathbf{CA}^{-1}\mathbf{C}')^{-1}\mathbf{CA}^{-1}]\mathbf{a}. \tag{A-148}$$

The two results, (A-147) and (A-148), yield analytic solutions for $\boldsymbol{\lambda}$ and $\mathbf{x}$. For the specific matrices and vectors of the example, these are $\lambda = [-0.5\ -7.5]'$, and the constrained solution vector, $\mathbf{x}^* = [1.50\ -1.5]'$. Note that in computing the solution to this sort of problem, it is not necessary to use the rather cumbersome form of (A-148). Once λ is obtained from (A-147), the solution can be inserted in (A-146) for a much simpler computation. The solution

$$\mathbf{x} = \frac{1}{2}\mathbf{A}^{-1}\mathbf{a} + \frac{1}{2}\mathbf{A}^{-1}\mathbf{C}'\boldsymbol{\lambda}$$

suggests a useful result for the constrained optimum:

$$\text{constrained solution} = \text{unconstrained solution} + [2\mathbf{A}]^{-1}\,\mathbf{C}'\boldsymbol{\lambda}. \qquad \textbf{(A-149)}$$

Finally, by inserting the two solutions in the original function, we find that $R = 24.375$ and $R^* = 2.25$, which illustrates again that the constrained solution (in this *maximization* problem) is inferior to the unconstrained solution.

A.8.4 TRANSFORMATIONS

If a function is strictly monotonic, then it is a **one-to-one function**. Each y is associated with exactly one value of x, and vice versa. In this case, an **inverse function** exists, which expresses x as a function of y, written

$$y = f(x)$$

and

$$x = f^{-1}(y).$$

An example is the inverse relationship between the log and the exponential functions.

The slope of the inverse function,

$$J = \frac{dx}{dy} = \frac{df^{-1}(y)}{dy} = f^{-1'}(y),$$

is the **Jacobian** of the transformation from y to x. For example, if

$$y = a + bx,$$

then

$$x = -\frac{a}{b} + \left[\frac{1}{b}\right]y$$

is the inverse transformation and

$$J = \frac{dx}{dy} = \frac{1}{b}.$$

Looking ahead to the statistical application of this concept, we observe that if $y = f(x)$ were *vertical,* then this would no longer be a functional relationship. The same x would be associated with more than one value of y. In this case, at this value of x, we would find that $J = 0$, indicating a singularity in the function.

If $\mathbf{y}$ is a column vector of functions, $\mathbf{y} = \mathbf{f}(\mathbf{x})$, then

$$\mathbf{J} = \frac{\partial \mathbf{x}}{\partial \mathbf{y}'} = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 & \cdots & \partial x_1/\partial y_n \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 & \cdots & \partial x_2/\partial y_n \\ & & \vdots & \\ \partial x_n/\partial y_1 & \partial x_n/\partial y_2 & \cdots & \partial x_n/\partial y_n \end{bmatrix}.$$

Consider the set of linear functions $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{f}(\mathbf{x})$. The inverse transformation is $\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y})$, which will be

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y},$$

if **A** is nonsingular. If **A** is singular, then there is no inverse transformation. Let **J** be the matrix of partial derivatives of the inverse functions:

$$\mathbf{J} = \left[\frac{\partial x_i}{\partial y_j}\right].$$

The absolute value of the determinant of **J**,

$$\text{abs}(|\mathbf{J}|) = \text{abs}\left(\det\left(\left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}'}\right]\right)\right),$$

is the Jacobian determinant of the transformation from **y** to **x**. In the nonsingular case,

$$\text{abs}(|\mathbf{J}|) = \text{abs}(|\mathbf{A}^{-1}|) = \frac{1}{\text{abs}(|\mathbf{A}|)}.$$

In the singular case, the matrix of partial derivatives will be singular and the determinant of the Jacobian will be zero. In this instance, the singular Jacobian implies that **A** is singular or, equivalently, that the transformations from **x** to **y** are functionally dependent. The singular case is analogous to the single-variable case.

Clearly, if the vector **x** is given, then $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be computed from **x**. Whether **x** can be deduced from **y** is another question. Evidently, it depends on the Jacobian. If the Jacobian is not zero, then the inverse transformations exist, and we can obtain **x**. If not, then we cannot obtain **x**.

APPENDIX B

PROBABILITY AND DISTRIBUTION THEORY

B.1 INTRODUCTION

This appendix reviews the distribution theory used later in the book. A previous course in statistics is assumed, so most of the results will be stated without proof. The more advanced results in the later sections will be developed in greater detail.

B.2 RANDOM VARIABLES

We view our observation on some aspect of the economy as the **outcome** or realization of a random process that is almost never under our (the analyst's) control. In the current literature, the descriptive (and perspective laden) term **data generating process (DPG)** is often used for this underlying mechanism. The observed (measured) outcomes of the process are assigned unique numeric values. The assignment is one to one; each outcome gets one value, and no two distinct outcomes receive the same value. This outcome variable, X, is a **random variable** because, until the data are actually observed, it is uncertain what value X will take. Probabilities are associated with outcomes to quantify this uncertainty. We usually use capital letters for the "name" of a random variable and lowercase letters for the values it takes. Thus, the probability that X takes a particular value x might be denoted Prob $(X = x)$.

A random variable is **discrete** if the set of outcomes is either finite in number or countably infinite. The random variable is **continuous** if the set of outcomes is infinitely divisible and, hence, not countable. These definitions will correspond to the types of data we observe in practice. Counts of occurrences will provide observations on discrete random variables, whereas measurements such as time or income will give observations on continuous random variables.

B.2.1 PROBABILITY DISTRIBUTIONS

A listing of the values x taken by a random variable X and their associated probabilities is a **probability distribution**, $f(x)$. For a discrete random variable,

$$f(x) = \text{Prob}(X = x). \qquad \textbf{(B-1)}$$

The **axioms of probability** require that

1. $0 \leq \text{Prob}(X = x) \leq 1.$ **(B-2)**

2. $\sum_x f(x) = 1.$ **(B-3)**

For the continuous case, the probability associated with any particular point is zero, and we can only assign positive probabilities to intervals in the range (or **support**) of x. The **probability density function (pdf)**, $f(x)$, is defined so that $f(x) \geq 0$ and

1. $\text{Prob}(a \leq x \leq b) = \int_a^b f(x)\,dx \geq 0.$ **(B-4)**

This result is the area under $f(x)$ in the range from a to b. For a continuous variable,

2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1.$ **(B-5)**

If the range of x is not infinite, then it is understood that $f(x) = 0$ anywhere outside the appropriate range. Because the probability associated with any individual point is 0,

$$\begin{aligned}\text{Prob}(a \leq x \leq b) &= \text{Prob}(a \leq x < b)\\ &= \text{Prob}(a < x \leq b)\\ &= \text{Prob}(a < x < b).\end{aligned}$$

B.2.2 CUMULATIVE DISTRIBUTION FUNCTION

For any random variable X, the probability that X is less than or equal to a is denoted $F(a)$. $F(x)$ is the **cumulative density function (cdf)**, or **distribution function**. For a discrete random variable,

$$F(x) = \sum_{X \leq x} f(X) = \text{Prob}(X \leq x). \qquad \textbf{(B-6)}$$

In view of the definition of $f(x)$,

$$f(x_i) = F(x_i) - F(x_{i-1}). \tag{B-7}$$

For a continuous random variable,

$$F(x) = \int_{-\infty}^{x} f(t)\, dt, \tag{B-8}$$

and

$$f(x) = \frac{dF(x)}{dx}. \tag{B-9}$$

In both the continuous and discrete cases, $F(x)$ must satisfy the following properties:

1. $0 \le F(x) \le 1$.
2. If $x > y$, then $F(x) \ge F(y)$.
3. $F(+\infty) = 1$.
4. $F(-\infty) = 0$.

From the definition of the cdf,

$$\text{Prob}(a < x \le b) = F(b) - F(a). \tag{B-10}$$

Any valid pdf will imply a valid cdf, so there is no need to verify these conditions separately.

B.3 EXPECTATIONS OF A RANDOM VARIABLE

DEFINITION B.1 Mean of a Random Variable
*The **mean**, or **expected value**, of a random variable is*

$$E[x] = \begin{cases} \sum_x x f(x) & \text{if } x \text{ is discrete,} \\ \int_x x f(x)\, dx & \text{if } x \text{ is continuous.} \end{cases} \tag{B-11}$$

The notation $\sum_x$ or $\int_x$, used henceforth, means the sum or integral over the entire range of values of x. The mean is usually denoted μ. It is a weighted average of the values taken by x, where the weights are the respective probabilities or densities. It is not necessarily a value actually taken by the random variable. For example, the expected number of heads in one toss of a fair coin is $\frac{1}{2}$.

Other **measures of central tendency** are the **median**, which is the value m such that $\text{Prob}(X \le m) \ge \frac{1}{2}$ and $\text{Prob}(X \ge m) \ge \frac{1}{2}$, and the **mode**, which is the value of x at which $f(x)$ takes its maximum. The first of these measures is more frequently used than the second. Loosely speaking, the median corresponds more closely than the mean to the middle of a distribution. It is unaffected by extreme values. In the discrete case, the modal value of x has the highest probability of occurring. The modal value for a continuous variable will usually not be meaningful.

Let $g(x)$ be a function of x. The function that gives the expected value of $g(x)$ is denoted

$$E[g(x)] = \begin{cases} \sum_x g(x)\,\text{Prob}(X = x) & \text{if } X \text{ is discrete,} \\ \int_x g(x)f(x)\,dx & \text{if } X \text{ is continuous.} \end{cases} \tag{B-12}$$

If $g(x) = a + bx$ for constants a and b, then

$$E[a + bx] = a + bE[x].$$

An important case is the expected value of a constant a, which is just a.

DEFINITION B.2 Variance of a Random Variable

The ***variance*** *of a random variable is*

$$\text{Var}[x] = E[(x - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } x \text{ is discrete,} \\ \int_x (x - \mu)^2 f(x)\,dx & \text{if } x \text{ is continuous.} \end{cases} \tag{B-13}$$

The variance of x, Var[x], which must be positive, is usually denoted σ^2. This function is a measure of the dispersion of a distribution. Computation of the variance is simplified by using the following important result:

$$\text{Var}[x] = E[x^2] - \mu^2. \tag{B-14}$$

A convenient corollary to (B-14) is

$$E[x^2] = \sigma^2 + \mu^2. \tag{B-15}$$

By inserting $y = a + bx$ in (B-13) and expanding, we find that

$$\text{Var}[a + bx] = b^2\,\text{Var}[x], \tag{B-16}$$

which implies, for any constant a, that

$$\text{Var}[a] = 0. \tag{B-17}$$

To describe a distribution, we usually use σ, the positive square root, which is the **standard deviation** of x. The standard deviation can be interpreted as having the same units of measurement as x and μ. For any random variable x and any positive constant k, the **Chebychev inequality** states that

$$\text{Prob}(\mu - k\sigma \le x \le \mu + k\sigma) \ge 1 - \frac{1}{k^2}. \tag{B-18}$$

Two other measures often used to describe a probability distribution are

$$\text{skewness} = E[(x - \mu)^3],$$

and

$$\text{kurtosis} = E[(x - \mu)^4].$$

Skewness is a measure of the asymmetry of a distribution. For symmetric distributions,

$$f(\mu - x) = f(\mu + x),$$

and

$$\text{skewness} = 0.$$

For asymmetric distributions, the skewness will be positive if the "long tail" is in the positive direction. Kurtosis is a measure of the thickness of the tails of the distribution. A shorthand expression for other **central moments** is

$$\mu_r = E[(x - \mu)^r].$$

Because μ_r tends to explode as r grows, the normalized measure, μ_r/σ^r, is often used for description. Two common measures are

$$\text{skewness coefficient} = \frac{\mu_3}{\sigma^3},$$

and

$$\text{degree of excess} = \frac{\mu_4}{\sigma^4} - 3.$$

The second is based on the normal distribution, which has excess of zero. (The value 3 is sometimes labeled the "mesokurtotic" value.)

For any two functions $g_1(x)$ and $g_2(x)$,

$$E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]. \tag{B-19}$$

For the general case of a possibly nonlinear $g(x)$,

$$E[g(x)] = \int_x g(x)f(x)\,dx, \tag{B-20}$$

and

$$\text{Var}[g(x)] = \int_x (g(x) - E[g(x)])^2 f(x)\,dx. \tag{B-21}$$

(For convenience, we shall omit the equivalent definitions for discrete variables in the following discussion and use the integral to mean either integration or summation, whichever is appropriate.)

A device used to approximate $E[g(x)]$ and $\text{Var}[g(x)]$ is the linear Taylor series approximation:

$$g(x) \approx [g(x^0) - g'(x^0)x^0] + g'(x^0)x = \beta_1 + \beta_2 x = g^*(x). \tag{B-22}$$

If the approximation is reasonably accurate, then the mean and variance of $g^*(x)$ will be approximately equal to the mean and variance of $g(x)$. A natural choice for the expansion point is $x^0 = \mu = E(x)$. Inserting this value in (B-22) gives

$$g(x) \approx [g(\mu) - g'(\mu)\mu] + g'(\mu)x, \tag{B-23}$$

so that

$$E[g(x)] \approx g(\mu), \tag{B-24}$$

and

$$\text{Var}[g(x)] \approx [g'(\mu)]^2 \,\text{Var}[x]. \tag{B-25}$$

A point to note in view of (B-22) to (B-24) is that $E[g(x)]$ will generally not equal $g(E[x])$. For the special case in which $g(x)$ is concave—that is, where $g''(x) < 0$—we know from **Jensen's inequality** that $E[g(x)] \leq g(E[x])$. For example, $E[\log(x)] \leq \log(E[x])$. The result in (B-25) forms the basis for the **delta method**.

B.4 SOME SPECIFIC PROBABILITY DISTRIBUTIONS

Certain experimental situations naturally give rise to specific probability distributions. In the majority of cases in economics, however, the distributions used are merely models of the observed phenomena. Although the normal distribution, which we shall discuss at length, is the mainstay of econometric research, economists have used a wide variety of other distributions. A few are discussed here.[1]

B.4.1 THE NORMAL AND SKEW NORMAL DISTRIBUTIONS

The general form of the normal distribution with mean μ and standard deviation σ is

$$f(x\,|\,\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)^2/\sigma^2]}. \tag{B-26}$$

This result is usually denoted $x \sim N[\mu, \sigma^2]$. The standard notation $x \sim f(x)$ is used to state that "x has probability distribution $f(x)$." Among the most useful properties of the normal distribution is its preservation under linear transformation.

$$\text{If } x \sim N[\mu, \sigma^2], \qquad \text{then } (a + bx) \sim N[a + b\mu, b^2\sigma^2]. \tag{B-27}$$

One particularly convenient transformation is $a = -\mu/\sigma$ and $b = 1/\sigma$. The resulting variable $z = (x - \mu)/\sigma$ has the **standard normal distribution**, denoted $N[0, 1]$, with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{B-28}$$

[1] A much more complete listing appears in Maddala (1977a, Chapters 3 and 18) and in most mathematical statistics textbooks. See also Poirier (1995) and Stuart and Ord (1989). Another useful reference is Evans, Hastings, and Peacock (2010). Johnson et al. (1974, 1993, 1994, 1995, 1997) is an encyclopedic reference on the subject of statistical distributions.

The specific notation $\phi(z)$ is often used for this density and $\Phi(z)$ for its cdf. It follows from the definitions above that if $x \sim N[\mu, \sigma^2]$, then

$$f(x) = \frac{1}{\sigma}\phi\left[\frac{x - \mu}{\sigma}\right].$$

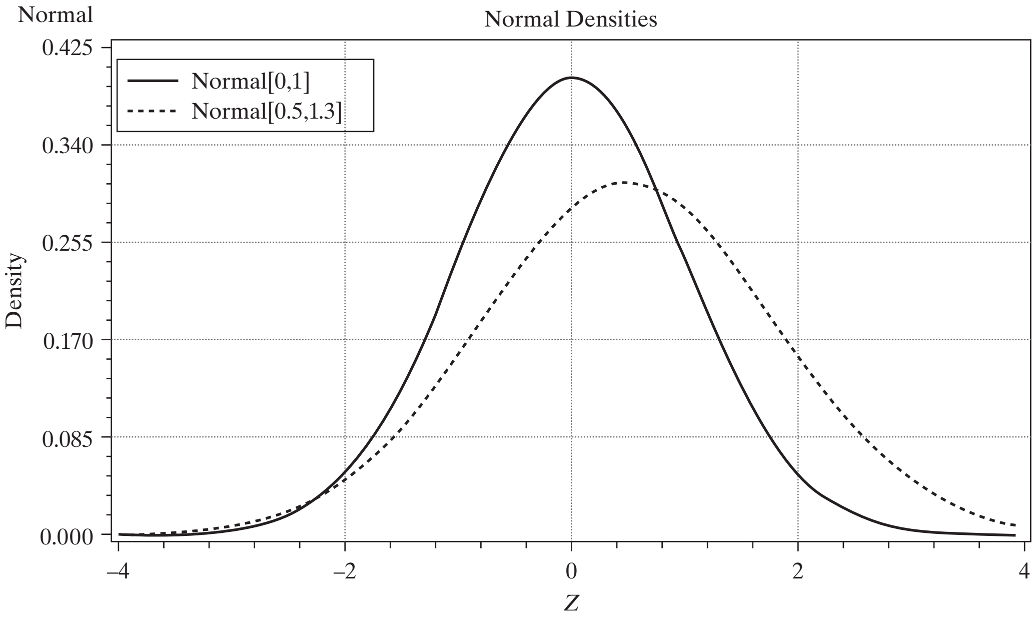
Figure B.1 shows the densities of the standard normal distribution and the normal distribution with mean 0.5, which shifts the distribution to the right, and standard deviation 1.3, which, it can be seen, scales the density so that it is shorter but wider. (The graph is a bit deceiving unless you look closely; both densities are symmetric.)

Tables of the standard normal cdf appear in most statistics and econometrics textbooks. Because the form of the distribution does not change under a linear transformation, it is not necessary to tabulate the distribution for other values of μ and σ. For any normally distributed variable,

$$\text{Prob}(a \leq x \leq b) = \text{Prob}\left(\frac{a - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right), \qquad \textbf{(B-29)}$$

which can always be read from a table of the standard normal distribution. In addition, because the distribution is symmetric, $\Phi(-z) = 1 - \Phi(z)$. Hence, it is not necessary to tabulate both the negative and positive halves of the distribution.

FIGURE B.1 The Normal Distribution.



The centerpiece of the stochastic frontier literture is the skew normal distribution. See Examples 12.2 and 14.8 and Section 19.2.4.) The density of the skew normal random variable is

$$f(x\,|\,\mu, \sigma, \lambda) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(\frac{-\lambda\varepsilon}{\sigma}\right), \varepsilon = (x - \mu).$$

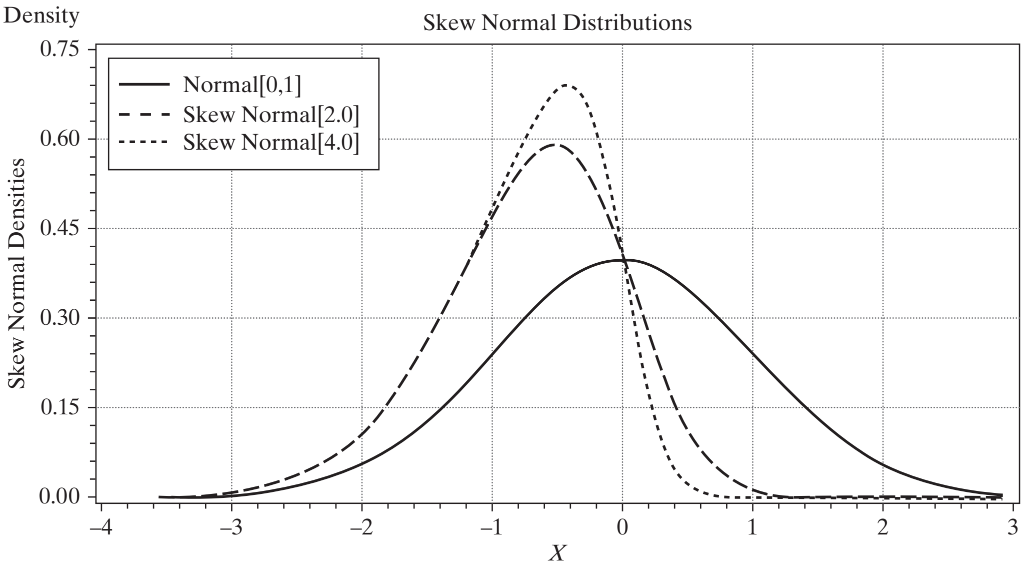
The skew normal reverts to the standard normal if $\lambda = 0$. The random variable arises as the density of $\varepsilon = \sigma_v v - \sigma_u|u|$ where u and v are standard normal variables, in which case $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{u}}^2$. (If $\sigma_u|u|$ is added, then $-\lambda$ becomes $+\lambda$ in the density. Figure B.2 shows three cases of the distribution, $\lambda = 0, 2$, and 4. This asymmetric distribution has mean $-\frac{\sigma\lambda}{\sqrt{1 + \lambda^2}}\sqrt{\frac{2}{\pi}}$ and variance $\frac{\sigma^2}{1 + \lambda^2}\left(1 + \lambda^2\left(\frac{\pi - 2}{\pi}\right)\right)$ (which revert to 0 and 1 if $\lambda = 0$).

These are $-\sigma_u(2/\pi)^{1/2}$ and $\sigma_v{}^2 + \sigma_u{}^2(\pi - 2)/\pi$ for the convolution form.

B.4.2 THE CHI-SQUARED, *T*, AND *F* DISTRIBUTIONS

The chi-squared, *t*, and *F* distributions are derived from the normal distribution. They arise in econometrics as sums of n or n_1 and n_2 other variables. These three distributions have associated with them one or two "degrees of freedom" parameters, which for our purposes will be the number of variables in the relevant sum.

FIGURE B.2 Skew Normal Densities.



The first of the essential results is

- If $z \sim N[0, 1]$, then $x = z^2 \sim$ chi-squared[1]—that is, **chi-squared** with one degree of freedom—denoted

$$z^2 \sim \chi^2[1]. \quad \textbf{(B-30)}$$

This distribution is a skewed distribution with mean 1 and variance 2. The second result is

- If $x_1, \dots, x_n$ are n *independent* chi-squared[1] variables, then

$$\sum_{i=1}^{n} x_i \sim \text{chi-squared}[n]. \quad \textbf{(B-31)}$$

The mean and variance of a chi-squared variable with n degrees of freedom are n and $2n$, respectively. A number of useful corollaries can be derived using (B-30) and (B-31).

- If $z_i, i = 1, \dots, n$, are independent $N[0, 1]$ variables, then

$$\sum_{i=1}^{n} z_i^2 \sim \chi^2[n]. \quad \textbf{(B-32)}$$

- If $z_i, i = 1, \dots, n$, are independent $N[0, \sigma^2]$ variables, then

$$\sum_{i=1}^{n} (z_i/\sigma)^2 \sim \chi^2[n]. \quad \textbf{(B-33)}$$

- If x_1 and x_2 are independent chi-squared variables with n_1 and n_2 degrees of freedom, respectively, then

$$x_1 + x_2 \sim \chi^2[n_1 + n_2]. \quad \textbf{(B-34)}$$

This result can be generalized to the sum of an arbitrary number of independent chi-squared variables.

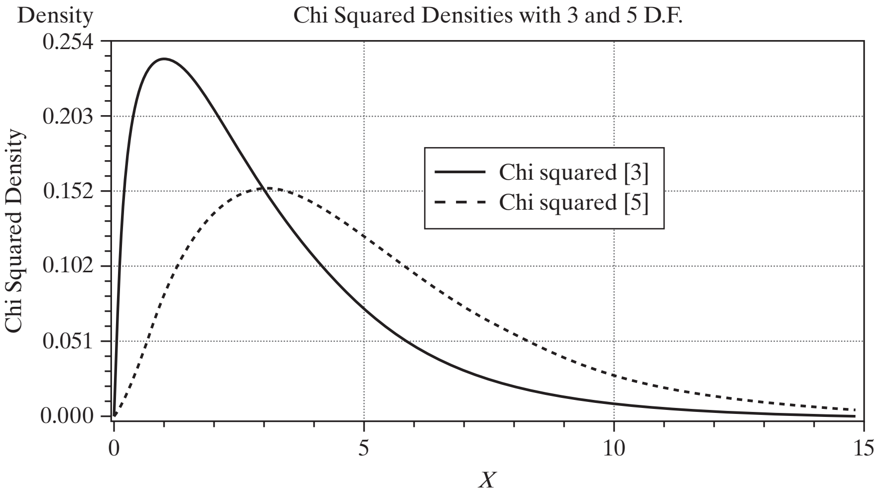
Figure B.3 shows the chi-squared densities for 3 and 5 degrees of freedom. The amount of skewness declines as the number of degrees of freedom rises. Unlike the normal distribution, a separate table is required for the chi-squared distribution for each value of n. Typically, only a few percentage points of the distribution are tabulated for each n.

- The chi-squared[n] random variable has the density of a gamma variable [See (B-39)] with
- parameters $\lambda = {}^1\!/_2$ and $P = n/2$.
- If x_1 and x_2 are two *independent* chi-squared variables with degrees of freedom parameters x_1 and x_1 respectively, then the ratio

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2} \quad \textbf{(B-35)}$$

has the ***F* distribution** with n_1 and n_2 degrees of freedom.

FIGURE B.3 The Chi-Squared[3] Distribution.



The two degrees of freedom parameters n_1 and n_2 are the "numerator and denominator degrees of freedom," respectively. Tables of the F distribution must be computed for each pair of values of (n_1, n_2). As such, only one or two specific values, such as the 95 percent and 99 percent upper tail values, are tabulated in most cases.

- If z is an $N[0, 1]$ variable and x is $\chi^2[n]$ and is independent of z, then the ratio

$$t[n] = \frac{z}{\sqrt{x/n}} \tag{B-36}$$

has the t **distribution** with n degrees of freedom.

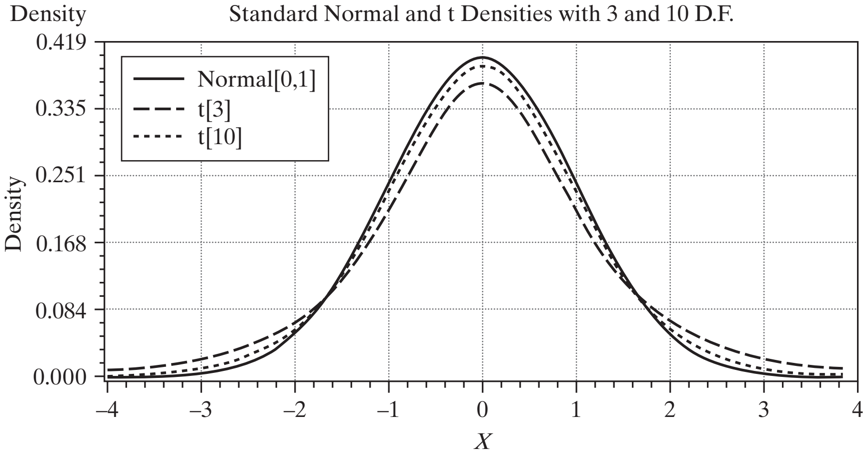
The t distribution has the same shape as the normal distribution but has thicker tails. Figure B.4 illustrates the t distributions with 3 and 10 degrees of freedom with the standard normal distribution. Two effects that can be seen in the figure are how the distribution changes as the degrees of freedom increases, and, overall, the similarity of the t distribution to the standard normal. This distribution is tabulated in the same manner as the chi-squared distribution, with several specific cutoff points corresponding to specified tail areas for various values of the degrees of freedom parameter.

Comparing (B-35) with $n_1 = 1$ and (B-36), we see the useful relationship between the t and F distributions:

- If $t \sim t[n]$, then $t^2 \sim F[1, n]$.

If the numerator in (B-36) has a nonzero mean, then the random variable in (B-36) has a noncentral t distribution and its square has a noncentral F distribution. These

FIGURE B.4 The Standard Normal, t[3], and t [10] Distributions.



distributions arise in the F tests of linear restrictions [see (5-16)] when the restrictions do not hold as follows:

1. *Noncentral chi-squared distribution*. If z has a normal distribution with mean μ and standard deviation 1, then the distribution of z^2 is *noncentral* chi-squared with parameters 1 and $\mu^2/2$.
   a. If $\mathbf{z} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ with J elements, then $\mathbf{z}'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{z}$ has a noncentral chi-squared distribution with J degrees of freedom and noncentrality parameter $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}/2$, which we denote $\boldsymbol{\chi}^2_*[J, \boldsymbol{\mu}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}/2]$.
   b. If $\mathbf{z} \sim N[\boldsymbol{\mu}, \mathbf{I}]$ and $\mathbf{M}$ is an idempotent matrix with rank J, then $\mathbf{z}'\mathbf{M}\mathbf{z} \sim \boldsymbol{\chi}^2_*[J, \boldsymbol{\mu}'\mathbf{M}\boldsymbol{\mu}/2]$.
2. *Noncentral F distribution*. If X_1 has a noncentral chi-squared distribution with noncentrality parameter λ and degrees of freedom n_1 and X_2 has a central chi-squared distribution with degrees of freedom n_2 and is independent of X_1, then

$$F_* = \frac{X_1/n_1}{X_2/n_2}$$

has a noncentral F distribution with parameters n_1, n_2, and λ. (The denominator chi-squared could also be noncentral, but we shall not use any statistics with doubly noncentral distributions.) In each of these cases, the statistic and the distribution are the familiar ones, except that the effect of the nonzero mean, which induces the noncentrality, is to push the distribution to the right.

B.4.3 DISTRIBUTIONS WITH LARGE DEGREES OF FREEDOM

The chi-squared, t, and F distributions usually arise in connection with sums of sample observations. The degrees of freedom parameter in each case grows with the number of observations. We often deal with larger degrees of freedom than are shown in the tables.

Thus, the standard tables are often inadequate. In all cases, however, there are **limiting distributions** that we can use when the degrees of freedom parameter grows large. The simplest case is the t distribution. The t distribution with infinite degrees of freedom is equivalent (identical) to the standard normal distribution. Beyond about 100 degrees of freedom, they are almost indistinguishable.

For degrees of freedom greater than 30, a reasonably good approximation for the distribution of the chi-squared variable x is

$$z = (2x)^{1/2} - (2n - 1)^{1/2}, \qquad \textbf{(B-37)}$$

which is approximately standard normally distributed. Thus,

$$\text{Prob}(\chi^2[n] \leq a) \approx \Phi[(2a)^{1/2} - (2n - 1)^{1/2}].$$

Another simple approximation that relies on the central limit theorem would be $z = (x - n)/(2n)^{1/2}$.

As used in econometrics, the F distribution with a large-denominator degrees of freedom is common. As n_2 becomes infinite, the denominator of F converges identically to one, so we can treat the variable

$$x = n_1 F \qquad \textbf{(B-38)}$$

as a chi-squared variable with n_1 degrees of freedom. The numerator degree of freedom will typically be small, so this approximation will suffice for the types of applications we are likely to encounter.[2] If not, then the approximation given earlier for the chi-squared distribution can be applied to $n_1 F$.

B.4.4 SIZE DISTRIBUTIONS: THE LOGNORMAL DISTRIBUTION

In modeling size distributions, such as the distribution of firm sizes in an industry or the distribution of income in a country, the **lognormal distribution**, denoted $LN[\mu, \sigma^2]$, has been particularly useful.[3] The density is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} e^{-1/2[(\ln x - \mu)/\sigma]^2}, \quad x > 0.$$

A lognormal variable x has

$$E[x] = e^{\mu + \sigma^2/2},$$

and

$$\text{Var}[x] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1).$$

The relation between the normal and lognormal distributions is

$$\text{If } y \sim LN[\mu, \sigma^2], \quad \ln y \sim N[\mu, \sigma^2].$$

[2] See Johnson, Kotz, and Balakrishnan (1994) for other approximations.

[3] A study of applications of the lognormal distribution appears in Aitchison and Brown (1969).

A useful result for transformations is given as follows:

If x has a lognormal distribution with mean θ and variance λ^2, then

$$\ln x \sim N(\mu, \sigma^2), \quad \text{where } \mu = \ln \theta^2 - \tfrac{1}{2}\ln(\theta^2 + \lambda^2) \quad \text{and} \quad \sigma^2 = \ln(1 + \lambda^2/\theta^2).$$

Because the normal distribution is preserved under linear transformation,

$$\text{if } y \sim LN[\mu, \sigma^2], \quad \text{then } \ln y^r \sim N[r\mu, r^2\sigma^2].$$

If y_1 and y_2 are independent lognormal variables with $y_1 \sim LN[\mu_1, \sigma_1^2]$ and $y_2 \sim LN[\mu_2, \sigma_2^2]$, then

$$y_1 y_2 \sim LN[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2].$$

B.4.5 THE GAMMA AND EXPONENTIAL DISTRIBUTIONS

The **gamma distribution** has been used in a variety of settings, including the study of income distribution[4] and production functions.[5] The general form of the distribution is

$$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}, \quad x \geq 0, \lambda > 0, P > 0. \tag{B-39}$$

Many familiar distributions are special cases, including the **exponential distribution** ($P = 1$) and chi-squared ($\lambda = \frac{1}{2}$, $P = \frac{n}{2}$). The **Erlang distribution** results if P is a positive integer. The mean is P/λ, and the variance is P/λ^2. The **inverse gamma distribution** is the distribution of $1/x$, where x has the gamma distribution. Using the change of variable, $y = 1/x$, the Jacobian is $|dx/dy| = 1/y^2$. Making the substitution and the change of variable, we find

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda/y}\, y^{-(P+1)}, y \geq 0, \lambda > 0, P > 0.$$

The density is defined for positive P. However, the mean is $\lambda/(P - 1)$ which is defined only if $P > 1$ and the variance is $\lambda^2/[(P - 1)^2(P - 2)]$ which is defined only for $P > 2$.

B.4.6 THE BETA DISTRIBUTION

Distributions for models are often chosen on the basis of the range within which the random variable is constrained to vary. The lognormal distribution, for example, is sometimes used to model a variable that is always nonnegative. For a variable constrained between 0 and $c > 0$, the **beta distribution** has proved useful. Its density is

$$f(x) = \frac{1}{c}\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{x}{c}\right)^{\alpha-1}\left(1 - \frac{x}{c}\right)^{\beta-1}. \tag{B-40}$$

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$, standard uniform if $\alpha = \beta = c = 1$, asymmetric otherwise, and

[4] Salem and Mount (1974).

[5] Greene (1980a).

can be hump-shaped or U-shaped. The mean is $c\alpha/(\alpha + \beta)$, and the variance is $c^2\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$. The beta distribution has been applied in the study of labor force participation rates.[6]

B.4.7 THE LOGISTIC DISTRIBUTION

The normal distribution is ubiquitous in econometrics. But researchers have found that for some microeconomic applications, there does not appear to be enough mass in the tails of the normal distribution; observations that a model based on normality would classify as "unusual" seem not to be very unusual at all. One approach has been to use thicker-tailed symmetric distributions. The **logistic distribution** is one candidate; the cdf for a logistic random variable is denoted

$$F(x) = \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

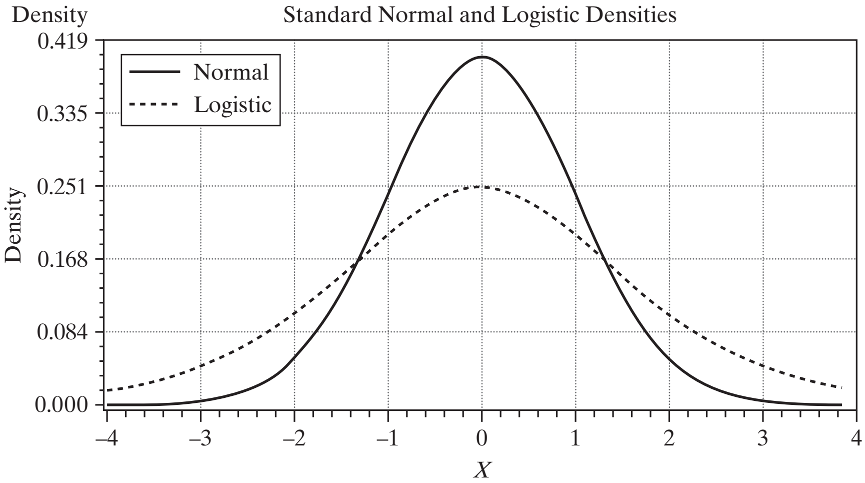
The density is $f(x) = \Lambda(x)[1 - \Lambda(x)]$. The mean and variance of this random variable are zero and $\pi^2/3$. Figure B.5 compares the logistic distribution to the standard normal. The logistic density has a greater variance and thicker tails than the normal. The standardized variable, $z/(\pi/3^{1/2})$ is very close to the t[8] variable.

B.4.8 THE WISHART DISTRIBUTION

The Wishart distribution describes the distribution of a random matrix obtained as

$$\mathbf{W} = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})',$$

FIGURE B.5 Normal and Logistic Densities.



[6] Heckman and Willis (1976).

where $\mathbf{x}_i$ is the ith of n K element random vectors from the multivariate normal distribution with mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. This is a multivariate counterpart to the chi-squared distribution. The density of the Wishart random matrix is

$$f(\mathbf{W}) = \frac{\exp\left[-\frac{1}{2}trace(\boldsymbol{\Sigma}^{-1}\mathbf{W})\right]|\mathbf{W}|^{-\frac{1}{2}(n-K-1)}}{2^{nK/2}|\boldsymbol{\Sigma}|^{K/2}\pi^{K(K-1)/4}\,\Pi_{j=1}^{K}\Gamma\left(\frac{n+1-j}{2}\right)}.$$

The mean matrix is $n\boldsymbol{\Sigma}$. For the individual pairs of elements in $\mathbf{W}$,

$$\text{Cov}[w_{ij}, w_{rs}] = n(\sigma_{ir}\sigma_{js} + \sigma_{is}\sigma_{jr}).$$

B.4.9 DISCRETE RANDOM VARIABLES

Modeling in economics frequently involves random variables that take integer values. In these cases, the distributions listed thus far only provide approximations that are sometimes quite inappropriate. We can build up a class of models for discrete random variables from the **Bernoulli distribution** for a single binomial outcome (trial)

$$\text{Prob}(x = 1) = \alpha,$$
$$\text{Prob}(x = 0) = 1 - \alpha,$$

where $0 \leq \alpha \leq 1$. The modeling aspect of this specification would be the assumptions that the success probability α is constant from one trial to the next and that successive trials are independent. If so, then the distribution for x successes in n trials is the **binomial distribution**,

$$\text{Prob}(X = x) = \binom{n}{x}\alpha^x(1 - \alpha)^{n-x}, \quad x = 0, 1, \ldots, n.$$

The mean and variance of x are $n\alpha$ and $n\alpha(1 - \alpha)$, respectively. If the number of trials becomes large at the same time that the success probability becomes small so that the mean $n\alpha$ is stable, then, the limiting form of the binomial distribution is the **Poisson distribution**,

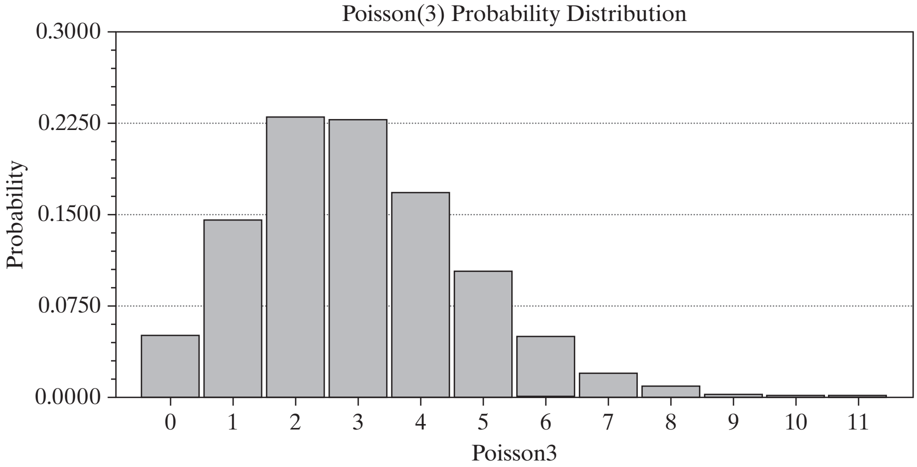
$$\text{Prob}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The Poisson distribution has seen wide use in econometrics in, for example, modeling patents, crime, recreation demand, and demand for health services. (See Chapter 18.) An example is shown in Figure B.6.

B.5 THE DISTRIBUTION OF A FUNCTION OF A RANDOM VARIABLE

We considered finding the expected value of a function of a random variable. It is fairly common to analyze the random variable itself, which results when we compute a function of some random variable. There are three types of transformation to consider. One discrete random variable may be transformed into another, a continuous variable may be transformed into a discrete one, and one continuous variable may be transformed into another.

FIGURE B.6 The Poisson[3] Distribution.



The simplest case is the first one. The probabilities associated with the new variable are computed according to the laws of probability. If y is derived from x and the function is one to one, then the probability that $Y = y(x)$ equals the probability that $X = x$. If several values of x yield the same value of y, then Prob $(Y = y)$ is the sum of the corresponding probabilities for x.

The second type of transformation is illustrated by the way individual data on income are typically obtained in a survey. Income in the population can be expected to be distributed according to some skewed, continuous distribution such as the one shown in Figure B.7.

Data are often reported categorically, as shown in the lower part of the figure. Thus, the random variable corresponding to observed income is a discrete transformation of the actual underlying continuous random variable. Suppose, for example, that the transformed variable y is the mean income in the respective interval. Then

$$
\begin{aligned}
\text{Prob}(Y = \mu_1) &= P(-\infty < X \le a),\\
\text{Prob}(Y = \mu_2) &= P(a < X \le b),\\
\text{Prob}(Y = \mu_3) &= P(b < X \le c),
\end{aligned}
$$

and so on, which illustrates the general procedure.

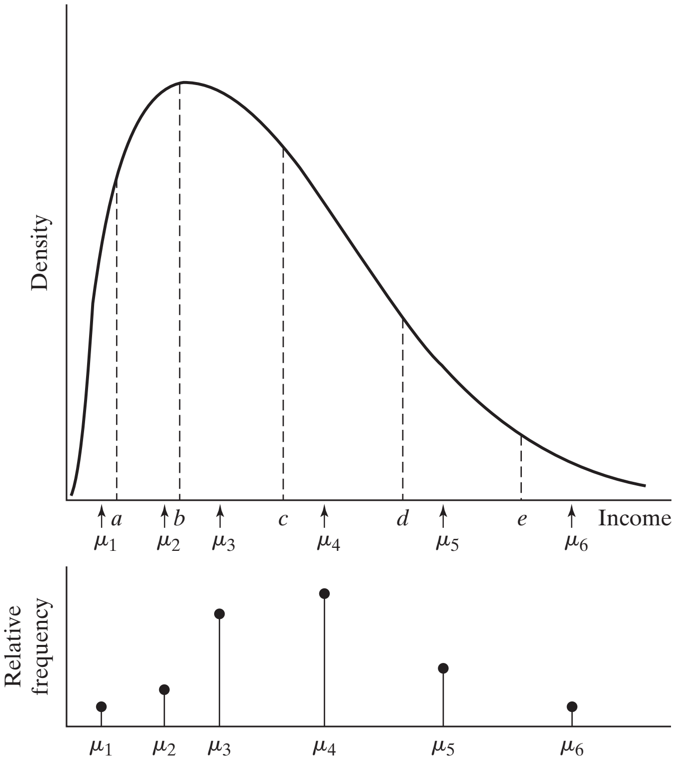
If x is a continuous random variable with pdf $f_x(x)$ and if $y = g(x)$ is a continuous monotonic function of x, then the density of y is obtained by using the change of variable technique to find the cdf of y:

$$\text{Prob}(y \le b) = \textstyle\int_{-\infty}^{b} f_x|(g^{-1}(y))|g^{-1\prime}(y)|\,dy.$$

This equation can now be written as

$$\text{Prob}(y \le b) = \textstyle\int_{-\infty}^{b} f_y|(y)\,dy.$$

FIGURE B.7 Censored Distribution.



Hence,

$$f_y(y) = f_x(g^{-1}(y))\left|g^{-1\prime}(y)\right|. \qquad \textbf{(B-41)}$$

To avoid the possibility of a negative pdf if $g(x)$ is decreasing, we use the absolute value of the derivative in the previous expression. The term $\left|g^{-1\prime}(y)\right|$ must be nonzero for the density of y to be nonzero. In words, the probabilities associated with intervals in the range of y must be associated with intervals in the range of x. If the derivative is zero, the correspondence $y = g(x)$ is vertical, and hence all values of y in the given range are associated with the same value of x. This single point must have probability zero.

One of the most useful applications of the preceding result is the linear transformation of a normally distributed variable. If $x \sim N[\mu, \sigma^2]$, then the distribution of

$$y = \frac{x - \mu}{\sigma}$$

is found using the preceding result. First, the derivative is obtained from the inverse transformation

$$y = \frac{x}{\sigma} - \frac{\mu}{\sigma} \Rightarrow x = \sigma y + \mu \Rightarrow f^{-1\prime}(y) = \frac{dx}{dy} = \sigma.$$

Therefore,

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-[(\sigma y + \mu) - \mu]^2/(2\sigma^2)}|\sigma| = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}.$$

This is the density of a normally distributed variable with mean zero and unit standard deviation one. This is the result which makes it unnecessary to have separate tables for the different normal distributions which result from different means and variances.

B.6 REPRESENTATIONS OF A PROBABILITY DISTRIBUTION

The probability density function (pdf) is a natural and familiar way to formulate the distribution of a random variable. But, there are many other functions that are used to identify or characterize a random variable, depending on the setting. In each of these cases, we can identify some other function of the random variable that has a one-to-one relationship with the density. We have already used one of these quite heavily in the preceding discussion. For a random variable which has density function $f(x)$, the distribution function, $F(x)$, is an equally informative function that identifies the distribution; the relationship between $f(x)$ and $F(x)$ is defined in (B-6) for a discrete random variable and (B-8) for a continuous one. We now consider several other related functions.

For a continuous random variable, the **survival function** is $S(x) = 1 - F(x) = \text{Prob}[X \geq x]$. This function is widely used in epidemiology, where x is time until some transition, such as recovery from a disease. The **hazard function** for a random variable is

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}.$$

The hazard function is a conditional probability;

$$h(x) = \lim_{t \downarrow 0} \text{Prob}(X \leq x \leq X + t | X \geq x).$$

Hazard functions have been used in econometrics in studying the duration of spells, or conditions, such as unemployment, strikes, time until business failures, and so on. The connection between the hazard and the other functions is $h(x) = -d \ln S(x)/dx$. As an exercise, you might want to verify the interesting special case of $h(x) = 1/\lambda$, a constant—the only distribution which has this characteristic is the exponential distribution noted in Section B.4.5.

For the random variable X, with probability density function $f(x)$, if the function

$$M(t) = E[e^{tx}]$$

exists, then it is the **moment generating function (MGF)**. Assuming the function exists, it can be shown that

$$d^r M(t)/dt^r|_{t=0} = E[x^r].$$

The moment generating function, like the survival and the hazard functions, is a unique characterization of a probability distribution. When it exists, the moment generating function has a one-to-one correspondence with the distribution. Thus, for example, if we begin with some random variable and find that a transformation of it has a particular MGF, then we may infer that the function of the random variable has the distribution associated with that MGF. A convenient application of this result is the MGF for the normal distribution. The MGF for the standard normal distribution is $M_z(t) = e^{t^2/2}$.

A useful feature of MGFs is the following:

If x and y are independent, then the MGF of $x + y$ is $M_x(t)M_y(t)$.

This result has been used to establish the **contagion** property of some distributions, that is, the property that sums of random variables with a given distribution have that same distribution. The normal distribution is a familiar example. This is usually not the case. It is for Poisson and chi-squared random variables.

One qualification of all of the preceding is that in order for these results to hold, the MGF must exist. It will for the distributions that we will encounter in our work, but in at least one important case, we cannot be sure of this. When computing sums of random variables which may have different distributions and whose specific distributions need not be so well behaved, it is likely that the MGF of the sum does not exist. However, the **characteristic function**,

$$\phi(t) = E[e^{itx}], i^2 = -1,$$

will always exist, at least for relatively small t. The characteristic function is the device used to prove that certain sums of random variables converge to a normally distributed variable—that is, the characteristic function is a fundamental tool in proofs of the central limit theorem.

B.7 JOINT DISTRIBUTIONS

The **joint density function** for two random variables X and Y denoted $f(x,y)$ is defined so that

$$\text{Prob}(a \le x \le b, c \le y \le d) = \begin{cases} \sum_{a \le x \le b} \sum_{c \le y \le d} f(x, y) & \text{if } x \text{ and } y \text{ are discrete,} \\ \int_a^b \int_c^d f(x, y)\, dy\, dx & \text{if } x \text{ and } y \text{ are continuous.} \end{cases} \quad \textbf{(B-42)}$$

The counterparts of the requirements for a univariate probability density are

$$\begin{aligned} & f(x, y) \ge 0, \\ & \sum_x \sum_y f(x, y) = 1 \qquad \text{if } x \text{ and } y \text{ are discrete,} \\ & \int_x \int_y f(x, y)\, dy\, dx = 1 \quad \text{if } x \text{ and } y \text{ are continuous.} \end{aligned} \quad \textbf{(B-43)}$$

The cumulative probability is likewise the probability of a joint event:

$$F(x, y) = \text{Prob}(X \le x, Y \le y) = \begin{cases} \sum_{X \le x} \sum_{Y \le y} f(x, y) & \text{in the discrete case} \\ \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s)\, ds\, dt & \text{in the continuous case.} \end{cases} \quad \textbf{(B-44)}$$

B.7.1 MARGINAL DISTRIBUTIONS

A **marginal probability density** or marginal probability distribution is defined with respect to an individual variable. To obtain the marginal distributions from the joint density, it is necessary to sum or integrate out the other variable:

$$f_x(x) = \begin{cases} \sum_y f(x, y) & \text{in the discrete case} \\ \int_y f(x, s)\, ds & \text{in the continuous case,} \end{cases} \tag{B-45}$$

and similarly for $f_y(y)$.

Two random variables are statistically independent if and only if their joint density is the product of the marginal densities:

$$f(x, y) = f_x(x)f_y(y) \Leftrightarrow x \text{ and } y \text{ are independent.} \tag{B-46}$$

If (and only if) x and y are independent, then the cdf factors as well as the pdf:

$$F(x, y) = F_x(x)F_y(y), \tag{B-47}$$

or

$$\text{Prob}(X \le x, Y \le y) = \text{Prob}(X \le x)\text{Prob}(Y \le y).$$

B.7.2 EXPECTATIONS IN A JOINT DISTRIBUTION

The means, variances, and higher moments of the variables in a joint distribution are defined with respect to the marginal distributions. For the mean of x in a discrete distribution,

$$\begin{aligned} E[x] &= \sum_x x f_x(x) \\ &= \sum_x x\left[\sum_y f(x, y)\right] \\ &= \sum_x \sum_y x f(x, y). \end{aligned} \tag{B-48}$$

The means of the variables in a continuous distribution are defined likewise, using integration instead of summation:

$$\begin{aligned} E[x] &= \int_x x f_x(x)\, dx \\ &= \int_x \int_y x f(x, y)\, dy\, dx. \end{aligned} \tag{B-49}$$

Variances are computed in the same manner:

$$\begin{aligned} \text{Var}[x] &= \sum_x (x - E[x])^2 f_x(x) \\ &= \sum_x \sum_y (x - E[x])^2 f(x, y). \end{aligned} \tag{B-50}$$

B.7.3 COVARIANCE AND CORRELATION

For any function $g(x, y)$,

$$E[g(x, y)] = \begin{cases} \sum_x \sum_y g(x, y)f(x, y) & \text{in the discrete case} \\ \int_x \int_y g(x, y)f(x, y)\,dy\,dx & \text{in the continuous case.} \end{cases} \tag{B-51}$$

The covariance of x and y is a special case:

$$\begin{aligned} \operatorname{Cov}[x, y] &= E[(x - \mu_x),(y - \mu_y)] \\ &= E[xy] - \mu_x\mu_y \\ &= \sigma_{xy}. \end{aligned} \tag{B-52}$$

If x and y are independent, then $f(x, y) = f_x(x)f_y(y)$ and

$$\begin{aligned} \sigma_{xy} &= \sum_x \sum_y f_x(x)f_y(y)(x - \mu_x)(y - \mu_y) \\ &= \sum_x (x - \mu_x)f_x(x) \sum_y (y - \mu_y)f_y(y) \\ &= E[x - \mu_x]E[y - \mu_y] \\ &= 0. \end{aligned}$$

The sign of the covariance will indicate the direction of covariation of X and Y. Its magnitude depends on the scales of measurement, however. In view of this fact, a preferable measure is the correlation coefficient:

$$r[x, y] = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \tag{B-53}$$

where σ_x and σ_y are the standard deviations of x and y, respectively. The correlation coefficient has the same sign as the covariance but is always between -1 and 1 and is thus unaffected by any scaling of the variables.

Variables that are uncorrelated are not necessarily independent. For example, in the discrete distribution $f(-1, 1) = f(0, 0) = f(1, 1) = \frac{1}{3}$, the correlation is zero, but $f(1, 1)$ does not equal $f_x(1)f_y(1) = (\frac{1}{3})(\frac{2}{3})$. An important exception is the joint normal distribution discussed subsequently, in which lack of correlation does imply independence.

Some general results regarding expectations in a joint distribution, which can be verified by applying the appropriate definitions, are

$$E[ax + by + c] = a\,E[x] + bE[y] + c, \tag{B-54}$$

$$\begin{aligned} \operatorname{Var}[ax + by + c] &= a^2 \operatorname{Var}[x] + b^2\operatorname{Var}[y] + 2ab \operatorname{Cov}[x, y] \\ &= \operatorname{Var}[ax + by], \end{aligned} \tag{B-55}$$

and

$$\text{Cov}[ax + by, cx + dy] = ac\,\text{Var}[x] + bd\,\text{Var}[y] + (ad + bc)\text{Cov}[x, y]. \quad \textbf{(B-56)}$$

If X and Y are uncorrelated, then

$$\begin{aligned}\text{Var}[x + y] &= \text{Var}[x - y] \\ &= \text{Var}[x] + \text{Var}[y]. \quad \textbf{(B-57)}\end{aligned}$$

For any two functions $g_1(x)$ and $g_2(y)$, if x and y are independent, then

$$E[g_1(x)g_2(y)] = E[g_1(x)]E[g_2(y)]. \quad \textbf{(B-58)}$$

B.7.4 DISTRIBUTION OF A FUNCTION OF BIVARIATE RANDOM VARIABLES

The result for a function of a random variable in (B-41) must be modified for a joint distribution. Suppose that x_1 and x_2 have a joint distribution $f_x(x_1, x_2)$ and that y_1 and y_2 are two monotonic functions of x_1 and x_2:

$$y_1 = y_1(x_1, x_2),\ y_2 = y_2(x_1, x_2).$$

Because the functions are monotonic, the inverse transformations,

$$x_1 = x_1(y_1, y_2),\ x_2 = x_2(y_1, y_2),$$

exist. The Jacobian of the transformations is the matrix of partial derivatives,

$$\mathbf{J} = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} = \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}'}\right].$$

The joint distribution of y_1 and y_2 is

$$f_y(y_1, y_2) = f_x[x_1(y_1, y_2), x_2(y_1, y_2)]\text{abs}(|\mathbf{J}|).$$

The determinant of the Jacobian must be nonzero for the transformation to exist. A zero determinant implies that the two transformations are functionally dependent.

Certainly the most common application of the preceding in econometrics is the linear transformation of a set of random variables. Suppose that x_1 and x_2 are independently distributed $N[0, 1]$, and the transformations are

$$\begin{aligned} y_1 &= \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2, \\ y_2 &= \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2. \end{aligned}$$

To obtain the joint distribution of y_1 and y_2, we first write the transformations as

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}.$$

The inverse transformation is

$$\mathbf{x} = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{a}),$$

so the absolute value of the determinant of the Jacobian is

$$\text{abs}\,|\mathbf{J}| = \text{abs}\,|\mathbf{B}^{-1}| = \frac{1}{\text{abs}|\mathbf{B}|}.$$

The joint distribution of **x** is the product of the marginal distributions since they are independent. Thus,

$$f_x(\mathbf{x}) = (2\pi)^{-1} e^{-(x_1{}^2 + x_2{}^2)/2} = (2\pi)^{-1} e^{-\mathbf{x}'\mathbf{x}/2}.$$

Inserting the results for $\mathbf{x}(\mathbf{y})$ and J into $f_y(y_1, y_2)$ gives

$$f_y(\mathbf{y}) = (2\pi)^{-1} \frac{1}{\text{abs } |\mathbf{B}|} e^{-(\mathbf{y}-\mathbf{a})'(\mathbf{BB}')^{-1}(\mathbf{y}-\mathbf{a})/2}.$$

This **bivariate normal distribution** is the subject of Section B.9. Note that by formulating it as we did earlier, we can generalize easily to the multivariate case, that is, with an arbitrary number of variables.

Perhaps the more common situation is that in which it is necessary to find the distribution of one function of two (or more) random variables. A strategy that often works in this case is to form the joint distribution of the transformed variable and one of the original variables, then integrate (or sum) the latter out of the joint distribution to obtain the marginal distribution. Thus, to find the distribution of $y_1(x_1, x_2)$, we might formulate

$$y_1 = y_1(x_1, x_2)$$
$$y_2 = x_2.$$

The absolute value of the determinant of the Jacobian would then be

$$\mathbf{J} = \text{abs}\begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ 0 & 1 \end{vmatrix} = \text{abs}\left|\frac{\partial x_1}{\partial y_1}\right|.$$

The density of y_1 would then be

$$f_{y_1}(y_1) = \int_{y_2} f_x[x_1(y_1, y_2), y_2] \text{ abs } |\mathbf{J}| \, dy_2.$$

B.8 CONDITIONING IN A BIVARIATE DISTRIBUTION

Conditioning and the use of conditional distributions play a pivotal role in econometric modeling. We consider some general results for a bivariate distribution. (All these results can be extended directly to the multivariate case.)

In a bivariate distribution, there is a **conditional distribution** over y for each value of x. The conditional densities are

$$f(y \,|\, x) = \frac{f(x, y)}{f_x(x)}, \tag{B-59}$$

and

$$f(x \,|\, y) = \frac{f(x, y)}{f_y(y)}.$$

It follows from (B-46) that.

$$\text{If } x \text{ and } y \text{ are independent, then } f(y|x) = f_y(y) \text{ and } f(x|y) = f_x(x). \qquad \textbf{(B-60)}$$

The interpretation is that if the variables are independent, the probabilities of events relating to one variable are unrelated to the other. The definition of conditional densities implies the important result

$$f(x, y) = f(y|x)f_x(x) \ = f(x|y)f_y(y). \qquad \textbf{(B-61)}$$

B.8.1 REGRESSION: THE CONDITIONAL MEAN

A **conditional mean** is the mean of the conditional distribution and is defined by

$$E[y|x] = \begin{cases} \int_y yf(y|x)dy & \text{if } y \text{ is continuous} \\ \sum_y yf(y|x) & \text{if } y \text{ is discrete.} \end{cases} \qquad \textbf{(B-62)}$$

The conditional mean function $E[y|x]$ is called the **regression** of y on x.

A random variable may always be written as

$$\begin{aligned} y &= E[y|x] + (y - E[y|x]) \\ &= E[y|x] + \varepsilon. \end{aligned}$$

B.8.2 CONDITIONAL VARIANCE

A conditional variance is the variance of the conditional distribution:

$$\begin{aligned} \text{Var}[y|x] &= E[(y - E[y|x])^2|x] \\ &= \int_y (y - E[y|x])^2 f(y|x)dy, \quad \text{if } y \text{ is continuous,} \end{aligned} \qquad \textbf{(B-63)}$$

or

$$\text{Var}[y|x] = \sum_y (y - E[y|x])^2 f(y|x) \quad \text{if } y \text{ is discrete.} \qquad \textbf{(B-64)}$$

The computation can be simplified by using

$$\text{Var}[y|x] = E[y^2|x] - (E[y|x])^2. \qquad \textbf{(B-65)}$$

The conditional variance is called the **scedastic function** and, like the regression, is generally a function of x. Unlike the conditional mean function, however, it is common for the conditional variance not to vary with x. We shall examine a particular case. This case does not imply, however, that Var[$y|x$] equals Var[y], which will usually not be true. It implies only that the conditional variance is a constant. The case in which the conditional variance does not vary with x is called **homoscedasticity** (same variance).

B.8.3 RELATIONSHIPS AMONG MARGINAL AND CONDITIONAL MOMENTS

Some useful results for the moments of a conditional distribution are given in the following theorems.

THEOREM B.1 Law of Iterated Expectations

$$E[y] = E_x[E[y|x]]. \tag{B-66}$$

The notation $E_x[.]$ indicates the expectation over the values of x. Note that $E[y|x]$ is a function of x.

THEOREM B.2 Covariance
In any bivariate distribution,

$$\text{Cov}[x, y] = \text{Cov}_x[x, E[y|x]] = \int_x (x - E[x])\, E[y|x] f_x(x)\, dx. \tag{B-67}$$

(Note that this is the covariance of x and a function of x.)

The preceding results provide an additional, extremely useful result for the special case in which the conditional mean function is linear in *x*.

THEOREM B.3 Moments in a Linear Regression
If $E[y|x] = \alpha + \beta x$, then

$$\alpha = E[y] - \beta E[x]$$

and

$$\beta = \frac{\text{Cov}[x,y]}{\text{Var}[x]}. \tag{B-68}$$

The proof follows from (*B*-66). *Whether* E[y|x] *is nonlinear or linear, the result in* (*B*-68) *is the linear projection of y on x. The linear projection is developed in Section B*.8.5.

The preceding theorems relate to the conditional mean in a bivariate distribution. The following theorems, which also appear in various forms in regression analysis, describe the conditional variance.

THEOREM B.4 Decomposition of Variance
In a joint distribution,

$$\text{Var}[y] = \text{Var}_x[E[y|x]] + E_x[\text{Var}[y|x]]. \tag{B-69}$$

The notation $\text{Var}_x[.]$ indicates the variance over the distribution of x. This equation states that in a bivariate distribution, the variance of y decomposes into the variance of the conditional mean function plus the expected variance around the conditional mean.

THEOREM B.5 Residual Variance in a Regression
In any bivariate distribution,

$$E_x[\text{Var}[y|x]] = \text{Var}[y] - \text{Var}_x[E[y|x]]. \tag{B-70}$$

On average, conditioning reduces the variance of the variable subject to the conditioning. For example, if y is homoscedastic, then we have the unambiguous result that the variance of the conditional distribution(s) is less than or equal to the unconditional variance of y. Going a step further, we have the result that appears prominently in the bivariate normal distribution (Section B.9).

THEOREM B.6 Linear Regression and Homoscedasticity
In a bivariate distribution, if $E[y|x] = \alpha + \beta x$ and if $\text{Var}[y|x]$ is a constant, then

$$\text{Var}[y|x] = \text{Var}[y](1 - \text{Corr}^2[y,x]) = \sigma_y^2(1 - \rho_{xy}^2). \tag{B-71}$$

The proof is straightforward using Theorems B.2 to B.4.

B.8.4 THE ANALYSIS OF VARIANCE

The variance decomposition result implies that in a bivariate distribution, variation in y arises from two sources:

1. Variation because $E[y|x]$ varies with x:

$$\textbf{regression variance} = \text{Var}_x[E[y|x]]. \tag{B-72}$$

2. Variation because, in each conditional distribution, y varies around the conditional mean:

$$\textbf{residual variance} = E_x[\text{Var}[y|x]]. \tag{B-73}$$

Thus,

$$\text{Var}[y] = \text{regression variance} + \text{residual variance}. \tag{B-74}$$

In analyzing a regression, we shall usually be interested in which of the two parts of the total variance, $\text{Var}[y]$, is the larger one. A natural measure is the ratio

$$\textbf{coefficient of determination} = \frac{\text{regression variance}}{\text{total variance}}. \tag{B-75}$$

In the setting of a linear regression, (B-75) arises from another relationship that emphasizes the interpretation of the correlation coefficient.

$$\text{If } E[y|x] = \alpha + \beta x, \quad \text{then the coefficient of determination} = \text{COD} = \rho^2, \qquad \textbf{(B-76)}$$

where ρ^2 is the squared correlation between x and y. We conclude that the correlation coefficient (squared) is a measure of the proportion of the variance of y accounted for by variation in the mean of y given x. It is in this sense that correlation can be interpreted as a **measure of linear association** between two variables.

B.8.5 LINEAR PROJECTION

Theorems B.3 (Moments in a Linear Regression) and B.6 (Linear Regression and Homoscedasticity) begin with an assumption that $E[y|x] = \alpha + \beta x$. If the conditional mean is not linear, then the results in THEOREM B.6 do not give the slopes in the conditional mean. However, in a bivariate distribution, we can always define the linear projection of y on x, as

$$Proj(y|x) = \gamma_0 + \gamma_1 x$$

where

$$\gamma_0 = E[y] - \gamma_1 E[x] \text{ and } \gamma_1 = \text{Cov}(x,y)/\text{Var}(x).$$

We can see immediately in THEOREM B.3 that if the conditional mean function is linear, then the conditional mean function (the regression of y on x) is also the linear projection. When the conditional mean function is not linear, then the regression and the projection functions will be different. We consider an example that bears some connection to the formulation of loglinear models. If

$$y|x \sim \text{Poisson with conditional mean function } \exp(\beta x),\ y = 0, 1, \ldots,$$

$$x \sim \text{U[0,1]};\ f(x) = 1,\ 0 \le x \le 1,$$

$$f(x,y) = f(y|x)f(x) = \exp[\text{-}\exp(\beta x)][\exp(\beta x)]^y/y! \times 1,$$

Then, as noted, the conditional mean function is nonlinear; $E[y|x] = \exp(\beta x)$. The slope in the projection of y on x is $\gamma_1 = \text{Cov}(x,y)/\text{Var}[x] = \text{Cov}(x, E[y|x])/\text{Var}[x] = \text{Cov}(x,\exp(\beta x))/\text{Var}[x]$. (THEOREM B.2.) We have $E[x] = 1/2$ and $\text{Var}[x] = 1/12$. To obtain the covariance, we require

$$E[x\exp(\beta x)] = \int_0^1 x\exp(\beta x)dx = \left[\left(\frac{x}{\beta} - \frac{1}{\beta^2}\right)\exp(\beta x)\right]_{x=0}^{x=1}$$

and

$$E[x]E[\exp(\beta x)] = \left(\frac{1}{2}\right)\int_0^1 \exp(\beta x)dx = \left(\frac{1}{2}\right)\left[\frac{\exp(\beta x)}{\beta}\right]_{x=0}^{x=1} = \left(\frac{1}{2}\right)\left[\frac{\exp(\beta) - 1}{\beta}\right].$$

After collecting terms, $\gamma_1 = h(\beta)$. The constant is $\gamma_0 = E[y] - h(\beta)(1/2)$. $E[y] = E[E[y|x]] = [\exp(\beta)\text{-}1]/\beta$. (THEOREM B.1.) Then, the projection is the linear function $\gamma_0 + \gamma_1 x$ while the regression function is the nonlinear function $\exp(\beta x)$. The projection can be viewed as a linear approximation to the conditional mean. (Note, it is not a linear Taylor series approximation.)

In similar fashion to Theorem B.5, we can define the variation around the projection,

$$Proj.Var[y|x] = E_x[\{y - Proj(y|x)\}^2|x].$$

By adding and subtracting the regression, $E[y|x]$, in the expression, we find

$$Proj.Var[y|x] = \text{Var}[y|x] + E_x[\{Proj(y|x) - E[y|x]\}^2|x].$$

This states that the variation of y around the projection consists of the regression variance plus the expected squared approximation error of the projection. As a general observation, we find, not surprisingly, that when the conditional mean is not linear, the projection does not do as well as the regression at prediction of y.

B.9 THE BIVARIATE NORMAL DISTRIBUTION

A bivariate distribution that embodies many of the features described earlier is the **bivariate normal**, which is the joint distribution of two normally distributed variables. The density is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} e^{-1/2[(\varepsilon_x{}^2 + \varepsilon_y{}^2 - 2\rho\varepsilon_x\varepsilon_y)/(1-\rho^2)]},$$

$$\varepsilon_x = \frac{x - \mu_x}{\sigma_x}, \quad \varepsilon_y = \frac{y - \mu_y}{\sigma_y}. \qquad \textbf{(B-77)}$$

The parameters μ_x, σ_x, μ_y, and σ_y are the means and standard deviations of the marginal distributions of x and y, respectively. The additional parameter ρ is the correlation between x and y. The covariance is

$$\sigma_{xy} = \rho\sigma_x\sigma_y. \qquad \textbf{(B-78)}$$

The density is defined only if ρ is not 1 or -1, which in turn requires that the two variables not be linearly related. If x and y have a bivariate normal distribution, denoted

$$(x, y) \sim N_2[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho],$$

then

- The marginal distributions are normal:

$$f_x(x) = N[\mu_x, \sigma_x^2],$$
$$f_y(y) = N[\mu_y, \sigma_y^2]. \qquad \textbf{(B-79)}$$

- The conditional distributions are normal:

$$f(y|x) = N[\alpha + \beta x, \sigma_y^2(1 - \rho^2)],$$

$$\alpha = \mu_y - \beta\mu_x \quad \beta = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad \textbf{(B-80)}$$

  and likewise for $f(x|y)$.
- x and y are independent if and only if $\rho = 0$. The density factors into the product of the two marginal normal distributions if $\rho = 0$.

Two things to note about the conditional distributions beyond their normality are their linear regression functions and their constant conditional variances. The conditional variance is less than the unconditional variance, which is consistent with the results of the previous section.

B.10 MULTIVARIATE DISTRIBUTIONS

The extension of the results for bivariate distributions to more than two variables is direct. It is made much more convenient by using matrices and vectors. The term **random vector** applies to a vector whose elements are random variables. The joint density is $f(\mathbf{x})$, whereas the cdf is

$$F(\mathbf{x}) = \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} f(\mathbf{t})dt_1\cdots dt_{n-1}\,dt_n. \qquad \textbf{(B-81)}$$

Note that the cdf is an n-fold integral. The marginal distribution of any one (or more) of the n variables is obtained by integrating or summing over the other variables.

B.10.1 MOMENTS

The expected value of a vector or matrix is the vector or matrix of expected values. A mean vector is defined as

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix} = E[\mathbf{x}]. \qquad \textbf{(B-82)}$$

Define the matrix

$$(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' = \begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1) & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1),(x_n-\mu_n) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)(x_2-\mu_2) & \cdots & (x_2-\mu_2)(x_n-\mu_n) \\ \vdots & & \vdots & \\ (x_n-\mu_n)(x_1-\mu_1) & (x_n-\mu_n)(x_2-\mu_2) & \cdots & (x_n-\mu_n)(x_n-\mu_n) \end{bmatrix}.$$

The expected value of each element in the matrix is the covariance of the two variables in the product. (The covariance of a variable with itself is its variance.) Thus,

$$E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & & \vdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = E[\mathbf{x}\mathbf{x}'] - \boldsymbol{\mu}\boldsymbol{\mu}', \qquad \textbf{(B-83)}$$

which is the **covariance matrix** of the random vector $\mathbf{x}$. Henceforth, we shall denote the covariance matrix of a random vector in boldface, as in

$$\text{Var}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

By dividing σ_{ij} by $\sigma_i\sigma_j$, we obtain the **correlation matrix**:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{bmatrix}.$$

B.10.2 SETS OF LINEAR FUNCTIONS

Our earlier results for the mean and variance of a linear function can be extended to the multivariate case. For the mean,

$$\begin{aligned} E[a_1x_1 + a_2x_2 + \cdots + a_nx_n] &= E[\mathbf{a}'\mathbf{x}] \\ &= a_1\,E[x_1] + a_2E[x_2] + \cdots + a_nE[x_n] \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \\ &= \mathbf{a}', \boldsymbol{\mu}. \qquad \textbf{(B-84)} \end{aligned}$$

For the variance,

$$\begin{aligned} \text{Var}[\mathbf{a}'\mathbf{x}] &= E[(\mathbf{a}'\mathbf{x} - E[\mathbf{a}'\mathbf{x}])^2] \\ &= E[\{\mathbf{a}'(\mathbf{x} - E[\mathbf{x}])\}^2] \\ &= E[\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\,\mathbf{a}] \end{aligned}$$

as $E[\mathbf{x}] = \boldsymbol{\mu}$ and $\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{a}$. Because $\mathbf{a}$ is a vector of constants,

$$\text{Var}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\sigma_{ij} \qquad \textbf{(B-85)}$$

It is the expected value of a square, so we know that a variance cannot be negative. As such, the preceding quadratic form is nonnegative, and the symmetric matrix $\boldsymbol{\Sigma}$ must be nonnegative definite.

In the set of linear functions $\mathbf{y} = \mathbf{A}\mathbf{x}$, the ith element of $\mathbf{y}$ is $y_i = \mathbf{a}_i\mathbf{x}$, where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ [see result (A-14)]. Therefore,

$$E[y_i] = \mathbf{a}_i\boldsymbol{\mu}.$$

Collecting the results in a vector, we have

$$E[\mathbf{A}\mathbf{x}] = \mathbf{A}\boldsymbol{\mu}. \qquad \textbf{(B-86)}$$

For two row vectors $\mathbf{a}_i$ and $\mathbf{a}_j$,

$$\text{Cov}[\mathbf{a}_i\mathbf{x}, \mathbf{a}_j\mathbf{x}] = \mathbf{a}_i, \boldsymbol{\Sigma}\mathbf{a}_j'.$$

Because $\mathbf{a}_i\,\boldsymbol{\Sigma}\mathbf{a}_j'$ is the ijth element of $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$,

$$\text{Var}[\mathbf{A}\mathbf{x}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \qquad \textbf{(B-87)}$$

This matrix will be either nonnegative definite or positive definite, depending on the column rank of $\mathbf{A}$.

B.10.3 NONLINEAR FUNCTIONS: THE DELTA METHOD

Consider a set of possibly nonlinear functions of $\mathbf{x}$, $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Each element of $\mathbf{y}$ can be approximated with a linear Taylor series. Let $\mathbf{j}^i$ be the row vector of partial derivatives of the i th function with respect to the n elements of $\mathbf{x}$:

$$\mathbf{j}^i(\mathbf{x}) = \frac{\partial g_i(\mathbf{x})}{\partial \mathbf{x}'} = \frac{\partial y_i}{\partial \mathbf{x}'}. \tag{B-88}$$

Then, proceeding in the now familiar way, we use $\boldsymbol{\mu}$, the mean vector of $\mathbf{x}$, as the expansion point, so that $\mathbf{j}^i(\boldsymbol{\mu})$ is the row vector of partial derivatives evaluated at $\boldsymbol{\mu}$. Then

$$g_i(\mathbf{x}) \approx g_i(\boldsymbol{\mu}) + \mathbf{j}^i(\boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}). \tag{B-89}$$

From this we obtain

$$E[g_i(\mathbf{x})] \approx g_i(\boldsymbol{\mu}), \tag{B-90}$$

$$\text{Var}[g_i(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{j}^i(\boldsymbol{\mu})', \tag{B-91}$$

and

$$\text{Cov}[g_i(\mathbf{x}), g_j(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{j}^j(\boldsymbol{\mu})'. \tag{B-92}$$

These results can be collected in a convenient form by arranging the row vectors $\mathbf{j}^i(\boldsymbol{\mu})$ in a matrix $\mathbf{J}(\boldsymbol{\mu})$. Then, corresponding to the preceding equations, we have

$$E[\mathbf{g}(\mathbf{x})] \simeq \mathbf{g}(\boldsymbol{\mu}), \tag{B-93}$$

$$\text{Var}[\mathbf{g}(\mathbf{x})] \simeq \mathbf{J}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{J}(\boldsymbol{\mu})'. \tag{B-94}$$

The matrix $\mathbf{J}(\boldsymbol{\mu})$ in the last preceding line is $\partial\mathbf{y}/\partial\mathbf{x}'$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$.

B.11 THE MULTIVARIATE NORMAL DISTRIBUTION

The foundation of most multivariate analysis in econometrics is the multivariate normal distribution. Let the vector $(x_1, x_2, \ldots, x_n)' = \mathbf{x}$ be the set of n random variables, $\boldsymbol{\mu}$ their mean vector, and $\boldsymbol{\Sigma}$ their covariance matrix. The general form of the joint density is

$$f(\mathbf{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}e^{(-1/2)(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{B-95}$$

If $\mathbf{R}$ is the correlation matrix of the variables and $\mathbf{R}_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$, then

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma_1\sigma_2\cdots\sigma_n)^{-1}|\mathbf{R}|^{-1/2}\, e^{(-1/2)\boldsymbol{\varepsilon}\mathbf{R}^{-1}\boldsymbol{\varepsilon}}, \tag{B-96}$$

where $\varepsilon_i = (x_i - \mu_i)/\sigma_i$.[7]

[7] This result is obtained by constructing $\boldsymbol{\Delta}$, the diagonal matrix with σ_i as its ith diagonal element. Then, $\mathbf{R} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}$, which implies that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Delta}^{-1}\mathbf{R}^{-1}\boldsymbol{\Delta}^{-1}$. Inserting this in (B-95) yields (B-96). Note that the ith element of $\boldsymbol{\Delta}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is $(x_i - \mu_i)/\sigma_i$.

Two special cases are of interest. If all the variables are uncorrelated, then $\rho_{ij} = 0$ for $i \neq j$. Thus, $\mathbf{R} = \mathbf{I}$, and the density becomes

$$\begin{aligned} f(\mathbf{x}) &= (2\pi)^{-n/2}(\sigma_1\sigma_2 \cdots \sigma_n)^{-1}e^{-\varepsilon'\varepsilon/2} \\ &= f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \end{aligned} \tag{B-97}$$

As in the bivariate case, if normally distributed variables are uncorrelated, then they are independent. If $\sigma_i = \sigma$ and $\boldsymbol{\mu} = \mathbf{0}$, then $x_i \sim N[0, \sigma^2]$ and $\varepsilon_i = x_i/\sigma$, and the density becomes

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\mathbf{x}'\mathbf{x}/(2\sigma^2)}. \tag{B-98}$$

Finally, if $\sigma = 1$,

$$f(\mathbf{x}) = (2\pi)^{-n/2}e^{-\mathbf{x}'\mathbf{x}/2}. \tag{B-99}$$

This distribution is the **multivariate standard normal**, or **spherical normal distribution**.

B.11.1 MARGINAL AND CONDITIONAL NORMAL DISTRIBUTIONS

Let $\mathbf{x}_1$ be any subset of the variables, including a single variable, and let $\mathbf{x}_2$ be the remaining variables. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ likewise so that

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then the marginal distributions are also normal. In particular, we have the following theorem.

THEOREM B.7 Marginal and Conditional Normal Distributions
If $[\mathbf{x}_1, \mathbf{x}_2]$ have a joint multivariate normal distribution, then the marginal distributions are

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \tag{B-100}$$

and

$$\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \tag{B-101}$$

The conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is normal as well:

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}), \tag{B-102}$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \tag{B-102a}$$

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{B-102b}$$

THEOREM B.7 (continued)

Proof: *We partition* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *as shown earlier and insert the parts in* (*B*-95). *To construct the density, we use* (*A*-72) *to partition the determinant,*

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|,$$

and (*A*-74) *to partition the inverse,*

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11.2}^{-1} & -\boldsymbol{\Sigma}_{11.2}^{-1}\mathbf{B} \\ -\mathbf{B}'\boldsymbol{\Sigma}_{11.2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \mathbf{B}'\boldsymbol{\Sigma}_{11.2}^{-1}\mathbf{B} \end{bmatrix}.$$

For simplicity, we let

$$\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

Inserting these in (*B*-95) *and collecting terms produces the joint density as a product of two terms:*

$$f(\mathbf{x}_1, \mathbf{x}_2) = f_{1.2}(\mathbf{x}_1|\mathbf{x}_2)f_2(\mathbf{x}_2).$$

The first of these is a normal distribution with mean $\boldsymbol{\mu}_{1.2}$ *and variance* $\boldsymbol{\Sigma}_{11.2}$, *whereas the second is the marginal distribution of* $\mathbf{x}_2$.

The conditional mean vector in the multivariate normal distribution is a linear function of the unconditional mean and the conditioning variables, and the conditional covariance matrix is constant and is smaller (in the sense discussed in Section A.7.3) than the unconditional covariance matrix. Notice that the conditional covariance matrix is the inverse of the upper left block of $\boldsymbol{\Sigma}^{-1}$; that is, this matrix is of the form shown in (A-74) for the partitioned inverse of a matrix.

B.11.2 THE CLASSICAL NORMAL LINEAR REGRESSION MODEL

An important special case of the preceding is that in which $\mathbf{x}_1$ is a single variable, y, and $\mathbf{x}_2$ is K variables, $\mathbf{x}$. Then the conditional distribution is a multivariate version of that in (B-80) with $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\sigma_{\mathbf{xy}}$, where $\sigma_{\mathbf{xy}}$ is the vector of covariances of y with $\mathbf{x}_2$. Recall that any random variable, y, can be written as its mean plus the deviation from the mean. If we apply this tautology to the multivariate normal, we obtain

$$y = E[y|\mathbf{x}] + (y - E[y|\mathbf{x}]) = \alpha + \boldsymbol{\beta}'\mathbf{x} + \varepsilon,$$

where $\boldsymbol{\beta}$ is given earlier, $\alpha = \mu_y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{x}}$, and ε has a normal distribution. We thus have, in this multivariate normal distribution, the **classical normal linear regression model**.

B.11.3 LINEAR FUNCTIONS OF A NORMAL VECTOR

Any linear function of a vector of joint normally distributed variables is also normally distributed. The mean vector and covariance matrix of $\mathbf{Ax}$, where $\mathbf{x}$ is normally distributed, follow the general pattern given earlier. Thus,

$$\text{If } \mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \quad \text{then } \mathbf{Ax} + \mathbf{b} \sim N[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}']. \qquad \textbf{(B-103)}$$

If $\mathbf{A}$ does not have full rank, then $\mathbf{A\Sigma A}'$ is singular and the density does not exist in the full dimensional space of $\mathbf{x}$ although it does exist in the subspace of dimension equal to the rank of $\mathbf{\Sigma}$. Nonetheless, the individual elements of $\mathbf{Ax} + \mathbf{b}$ will still be normally distributed, and the joint *distribution* of the full vector is still a multivariate normal.

B.11.4 QUADRATIC FORMS IN A STANDARD NORMAL VECTOR

The earlier discussion of the chi-squared distribution gives the distribution of $\mathbf{x}'\mathbf{x}$ if $\mathbf{x}$ has a standard normal distribution. It follows from (A-36) that

$$\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2. \tag{B-104}$$

We know from (B-32) that $\mathbf{x}'\mathbf{x}$ has a chi-squared distribution. It seems natural, therefore, to invoke (B-34) for the two parts on the right-hand side of (B-104). It is not yet obvious, however, that either of the two terms has a chi-squared distribution or that the two terms are independent, as required. To show these conditions, it is necessary to derive the distributions of **idempotent quadratic forms** and to show when they are independent.

To begin, the second term is the square of $\sqrt{n}\,\bar{x}$, which can easily be shown to have a standard normal distribution. Thus, the second term is the square of a standard normal variable and has chi-squared distribution with one degree of freedom. But the first term is the sum of n nonindependent variables, and it remains to be shown that the two terms are independent.

DEFINITION B.3 Orthonormal Quadratic Form
A particular case of $(B$-103) *is the following:*

If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{C}$ *is a square matrix such that* $\mathbf{C}'\mathbf{C} = \mathbf{I}$, *then* $\mathbf{C}'\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$.

Consider, then, a quadratic form in a standard normal vector $\mathbf{x}$ with symmetric matrix $\mathbf{A}$:

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}. \tag{B-105}$$

Let the characteristic roots and vectors of $\mathbf{A}$ be arranged in a diagonal matrix $\mathbf{\Lambda}$ and an orthogonal matrix $\mathbf{C}$, as in Section A.6.3. Then

$$q = \mathbf{x}'\mathbf{C\Lambda C}'\mathbf{x}. \tag{B-106}$$

By definition, $\mathbf{C}$ satisfies the requirement that $\mathbf{C}'\mathbf{C} = \mathbf{I}$. Thus, the vector $\mathbf{y} = \mathbf{C}'\mathbf{x}$ has a standard normal distribution. Consequently,

$$q = \mathbf{y}'\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2. \tag{B-107}$$

If λ_i is always one or zero, then

$$q = \sum_{j=1}^{J} y_j^2, \tag{B-108}$$

which has a chi-squared distribution. The sum is taken over the $j = 1, \ldots, J$ elements associated with the roots that are equal to one. A matrix whose characteristic roots are all zero or one is idempotent. Therefore, we have proved the next theorem.

THEOREM B.8 Distribution of an Idempotent Quadratic Form in a Standard Normal Vector
If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{A}$ *is idempotent, then* $\mathbf{x}'\mathbf{A}\mathbf{x}$ *has a chi-squared distribution with degrees of freedom equal to the number of unit roots of* $\mathbf{A}$*, which is equal to the rank of* $\mathbf{A}$.

The rank of a matrix is equal to the number of nonzero characteristic roots it has. Therefore, the degrees of freedom in the preceding chi-squared distribution equals J, the rank of $\mathbf{A}$.

We can apply this result to the earlier sum of squares. The first term is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x},$$

where $\mathbf{M}^0$ was defined in (A-34) as the matrix that transforms data to mean deviation form:

$$\mathbf{M}^0 = \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'.$$

Because $\mathbf{M}^0$ is idempotent, the sum of squared deviations from the mean has a chi-squared distribution. The degrees of freedom equals the rank $\mathbf{M}^0$, which is not obvious except for the useful result in (A-108), that

- The rank of an idempotent matrix is equal to its trace. **(B-109)**

  Each diagonal element of $\mathbf{M}^0$ is $1 - (1/n)$; hence, the trace is $n[1 - (1/n)] = n - 1$. Therefore, we have an application of Theorem B.8.

$$\text{If } \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}), \sum_{i=1}^{n}(x_i - \bar{x})^2 \sim \chi^2[n - 1]. \qquad \textbf{(B-110)}$$

We have already shown that the second term in (B-104) has a chi-squared distribution with one degree of freedom. It is instructive to set this up as a quadratic form as well:

$$n\bar{x}^2 = \mathbf{x}'\left[\frac{1}{n}\mathbf{i}\mathbf{i}'\right]\mathbf{x} = \mathbf{x}'[\mathbf{j}\mathbf{j}']\mathbf{x}, \quad \text{where } \mathbf{j} = \left(\frac{1}{\sqrt{n}}\right)\mathbf{i}. \qquad \textbf{(B-111)}$$

The matrix in brackets is the outer product of a nonzero vector, which always has rank one. You can verify that it is idempotent by multiplication. Thus, $\mathbf{x}'\mathbf{x}$ is the sum of two chi-squared variables, one with $n - 1$ degrees of freedom and the other with one. It is now necessary to show that the two terms are independent. To do so, we will use the next theorem.

THEOREM B.9 Independence of Idempotent Quadratic Forms

If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{x}'\mathbf{Ax}$ *and* $\mathbf{x}'\mathbf{Bx}$ *are two idempotent quadratic forms in* $\mathbf{x}$, *then* $\mathbf{x}'\mathbf{Ax}$ *and* $\mathbf{x}'\mathbf{Bx}$ *are independent if* $\mathbf{AB} = \mathbf{0}$. **(B-112)**

As before, we show the result for the general case and then specialize it for the example. Because both $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent, $\mathbf{A} = \mathbf{A}'\mathbf{A}$ and $\mathbf{B} = \mathbf{B}'\mathbf{B}$. The quadratic forms are therefore

$$\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{A}'\mathbf{Ax} = \mathbf{x}_1'\mathbf{x}_1, \quad \text{where } \mathbf{x}_1 = \mathbf{Ax}, \quad \text{and } \mathbf{x}'\mathbf{Bx} = \mathbf{x}_2'\mathbf{x}_2, \quad \text{where } \mathbf{x}_2 = \mathbf{Bx}. \tag{B-113}$$

Both vectors have zero mean vectors, so the covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$E(\mathbf{x}_1\mathbf{x}_2') = \mathbf{AIB}' = \mathbf{AB} = \mathbf{0}.$$

Because $\mathbf{Ax}$ and $\mathbf{Bx}$ are linear functions of a normally distributed random vector, they are, in turn, normally distributed. Their zero covariance matrix implies that they are statistically independent,[8] which establishes the independence of the two quadratic forms. For the case of $\mathbf{x}'\mathbf{x}$, the two matrices are $\mathbf{M}^0$ and $[\mathbf{I} - \mathbf{M}^0]$. You can show that $\mathbf{M}^0[\mathbf{I} - \mathbf{M}^0] = \mathbf{0}$ just by multiplying it out.

B.11.5 THE *F* DISTRIBUTION

The normal family of distributions (chi-squared, F, and t) can all be derived as functions of idempotent quadratic forms in a standard normal vector. The F distribution is the ratio of two independent chi-squared variables, each divided by its respective degrees of freedom. Let $\mathbf{A}$ and $\mathbf{B}$ be two idempotent matrices with ranks r_a and r_b, and let $\mathbf{AB} = \mathbf{0}$. Then

$$\frac{\mathbf{x}'\mathbf{Ax}/r_a}{\mathbf{x}'\mathbf{Bx}/r_b} \sim F[r_a, r_b]. \tag{B-114}$$

If $\text{Var}[\mathbf{x}] = \sigma^2\mathbf{I}$ instead, then this is modified to

$$\frac{(\mathbf{x}'\mathbf{Ax}/\sigma^2)/r_a}{(\mathbf{x}'\mathbf{Bx}/\sigma^2)/r_b} \sim F[r_a, r_b]. \tag{B-115}$$

B.11.6 A FULL RANK QUADRATIC FORM

Finally, consider the general case,

$$\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}].$$

[8] Note that both $\mathbf{x}_1 = \mathbf{Ax}$ and $\mathbf{x}_2 = \mathbf{Bx}$ have singular covariance matrices. Nonetheless, every element of $\mathbf{x}_1$ is independent of every element $\mathbf{x}_2$, so the vectors are independent.

We are interested in the distribution of

$$q = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{B-116}$$

First, the vector can be written as $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{z}$ as well as of $\mathbf{x}$. Therefore, we seek the distribution of

$$q = \mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} = \mathbf{z}'(\text{Var}[\mathbf{z}])^{-1}\mathbf{z}, \tag{B-117}$$

where $\mathbf{z}$ is normally distributed with mean $\mathbf{0}$. This equation is a quadratic form, but not necessarily in an idempotent matrix.[9] Because $\boldsymbol{\Sigma}$ is positive definite, it has a square root. Define the symmetric matrix $\boldsymbol{\Sigma}^{1/2}$ so that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$. Then

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2},$$

and

$$\begin{aligned} \mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} &= \mathbf{z}'\boldsymbol{\Sigma}^{-1/2\prime}\boldsymbol{\Sigma}^{-1/2}\mathbf{z} \\ &= (\boldsymbol{\Sigma}^{-1/2}\mathbf{z})'(\boldsymbol{\Sigma}^{-1/2}\mathbf{z}) \\ &= \mathbf{w}'\mathbf{w}. \end{aligned}$$

Now $\mathbf{w} = \mathbf{A}\mathbf{z}$, so

$$E(\mathbf{w}) = \mathbf{A}E[\mathbf{z}] = \mathbf{0},$$

and

$$\text{Var}[\mathbf{w}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{0} = \mathbf{I}.$$

This provides the following important result:

THEOREM B.10 Distribution of a Standardized Normal Vector

If $\mathbf{x} \sim N[\mu, \boldsymbol{\Sigma}]$, *then* $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N[\mathbf{0}, \mathbf{I}]$.

The simplest special case is that in which $\mathbf{x}$ has only one variable, so that the transformation is just $(x - \mu)/\sigma$. Combining this case with (B-32) concerning the sum of squares of standard normals, we have the following theorem.

THEOREM B.11 Distribution of $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ When x Is Normal

If $\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, *then* $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2[n]$.

[9] It will be idempotent only in the special case of $\boldsymbol{\Sigma} = \mathbf{I}$.

B.11.7 INDEPENDENCE OF A LINEAR AND A QUADRATIC FORM

The t distribution is used in many forms of hypothesis tests. In some situations, it arises as the ratio of a linear to a quadratic form in a normal vector. To establish the distribution of these statistics, we use the following result.

> **THEOREM B.12 Independence of a Linear and a Quadratic Form**
> *A linear function* $\mathbf{Lx}$ *and a symmetric idempotent quadratic form* $\mathbf{x}'\mathbf{Ax}$ *in a standard normal vector are statistically independent if* $\mathbf{LA} = \mathbf{0}$.

The proof follows the same logic as that for two quadratic forms. Write $\mathbf{x}'\mathbf{Ax}$ as $\mathbf{x}'\mathbf{A}'\mathbf{Ax} = (\mathbf{Ax})'(\mathbf{Ax})$. The covariance matrix of the variables $\mathbf{Lx}$ and $\mathbf{Ax}$ is $\mathbf{LA} = \mathbf{0}$, which establishes the independence of these two random vectors. The independence of the linear function and the quadratic form follows because functions of independent random vectors are also independent.

The t distribution is defined as the ratio of a standard normal variable to the square root of an independent chi-squared variable divided by its degrees of freedom:

$$t[J] = \frac{N[0, 1]}{\{\chi^2[J]/J\}^{1/2}}.$$

A particular case is

$$t[n - 1] = \frac{\sqrt{n}\,\bar{x}}{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{1/2}} = \frac{\sqrt{n}\bar{x}}{s},$$

where s is the standard deviation of the values of $\mathbf{x}$. The distribution of the two variables in $t[n - 1]$ was shown earlier; we need only show that they are independent. But

$$\sqrt{n}\bar{x} = \frac{1}{\sqrt{n}}\mathbf{i}'\mathbf{x} = \mathbf{j}'\mathbf{x},$$

and

$$s^2 = \frac{\mathbf{x}'\mathbf{M}^0\mathbf{x}}{n - 1}.$$

It suffices to show that $\mathbf{M}^0\mathbf{j} = \mathbf{0}$, which follows from

$$\mathbf{M}^0\mathbf{i} = [\mathbf{I} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}']\mathbf{i} = \mathbf{i} - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}(\mathbf{i}'\mathbf{i}) = \mathbf{0}.$$

APPENDIX C

ESTIMATION AND INFERENCE

C.1 INTRODUCTION

The probability distributions discussed in Appendix B serve as models for the underlying data generating processes that produce our observed data. The goal of statistical inference in econometrics is to use the principles of mathematical statistics to combine these theoretical distributions and the observed data into an empirical model of the economy. This analysis takes place in one of two frameworks, classical or Bayesian. The overwhelming majority of empirical study in econometrics has been done in the classical framework. Our focus, therefore, will be on classical methods of inference. Bayesian methods are discussed in Chapter 16.[1]

C.2 SAMPLES AND RANDOM SAMPLING

The classical theory of statistical inference centers on rules for using the sampled data effectively. These rules, in turn, are based on the properties of samples and sampling distributions.

A sample of n observations on one or more variables, denoted $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a **random sample** if the n observations are drawn independently from the same population, or probability distribution, $f(\mathbf{x}_i, \boldsymbol{\theta})$. The sample may be univariate if $\mathbf{x}_i$ is a single random variable or multivariate if each observation contains several variables. A random sample of observations, denoted $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ or $\{\mathbf{x}_i\}_{i=1,\ldots,n}$, is said to be **independent, identically distributed,** which we denote *i. i. d*. The vector $\boldsymbol{\theta}$ contains one or more unknown parameters. Data are generally drawn in one of two settings. A **cross section** is a sample of a number of observational units all drawn at the same point in time. A **time series** is a set of observations drawn on the same observational unit at a number of (usually evenly spaced) points in time. Many recent studies have been based on time-series cross sections, which generally consist of the same cross-sectional units observed at several points in time. Because the typical data set of this sort consists of a large number of cross-sectional units observed at a few points in time, the common term **panel data set** is usually more fitting for this sort of study.

[1] An excellent reference is Leamer (1978). A summary of the results as they apply to econometrics is contained in Zellner (1971) and in Judge et al. (1985). See, as well, Poirier (1991, 1995). Recent textbooks on Bayesian econometrics include Koop (2003), Lancaster (2004) and Geweke (2005).

C.3 DESCRIPTIVE STATISTICS

Before attempting to estimate parameters of a population or fit models to data, we normally examine the data themselves. In raw form, the sample data are a disorganized mass of information, so we will need some organizing principles to distill the information into something meaningful. Consider, first, examining the data on a single variable. In most cases, and particularly if the number of observations in the sample is large, we shall use some summary **statistics** to describe the sample data. Of most interest are measures of **location**—that is, the center of the data—and **scale**, or the dispersion of the data. A few measures of central tendency are as follows:

$$\textbf{mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\textbf{median: } M = \text{middle ranked observation},$$

$$\textbf{sample midrange: } \text{midrange} = \frac{\text{maximum} + \text{minimum}}{2}. \qquad \textbf{(C-1)}$$

The dispersion of the sample observations is usually measured by the

$$\textbf{standard deviation: } s_x = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right]^{1/2}. \qquad \textbf{(C-2)}$$

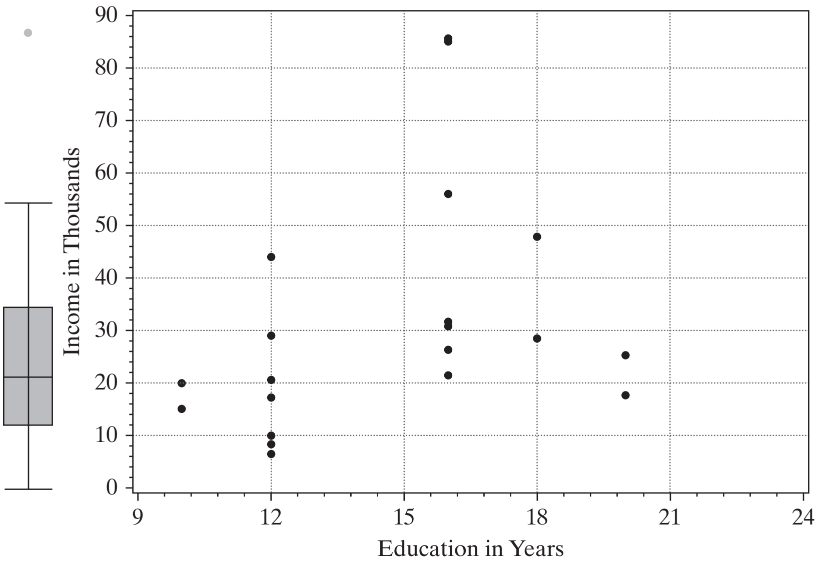
Other measures, such as the average absolute deviation from the sample mean, are also used, although less frequently than the standard deviation. The shape of the distribution of values is often of interest as well. Samples of income or expenditure data, for example, tend to be highly skewed while financial data such as asset returns and exchange rate movements are relatively more symmetrically distributed but are also more widely dispersed than other variables that might be observed. Two measures used to quantify these effects are the

$$\textbf{skewness} = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3(n-1)}\right], \quad \text{and} \quad \textbf{kurtosis} = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4(n-1)}\right].$$

(Benchmark values for these two measures are zero for a symmetric distribution, and three for one which is “normally” dispersed.) The skewness coefficient has a bit less of the intuitive appeal of the mean and standard deviation, and the kurtosis measure has very little at all. The **box and whisker plot** is a graphical device which is often used to capture a large amount of information about the sample in a simple visual display. This plot shows in a figure the median, the range of values contained in the 25th and 75th percentile, some limits that show the normal range of values expected, such as the median plus and minus two standard deviations, and in isolation values that could be viewed as outliers. A box and whisker plot is shown in Figure C.1 for the income variable in Example C.1.

If the sample contains data on more than one variable, we will also be interested in measures of association among the variables. A **scatter diagram** is useful in a bivariate sample if the sample contains a reasonable number of observations. Figure C.1 shows an

FIGURE C.1 Box and Whisker Plot for Income and Scatter Diagram for Income and Education.



example for a small data set. If the sample is a multivariate one, then the degree of linear association among the variables can be measured by the pairwise measures

$$\textbf{covariance: } s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1},$$

$$\textbf{correlation: } r_{xy} = \frac{s_{xy}}{s_x s_y}. \qquad \textbf{(C-3)}$$

If the sample contains data on several variables, then it is sometimes convenient to arrange the covariances or correlations in a

$$\textbf{covariance matrix: } \mathbf{S} = [s_{ij}], \qquad \textbf{(C-4)}$$

or

$$\textbf{correlation matrix: } \mathbf{R} = [r_{ij}].$$

Some useful algebraic results for any two variables (x_i, y_i), $i = 1, \ldots, n$, and constants a and b are

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}, \qquad \textbf{(C-5)}$$

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\,\bar{y}}{n-1}, \qquad \textbf{(C-6)}$$

$$-1 \leq r_{xy} \leq 1,$$

$$r_{ax,\,by} = \frac{ab}{|ab|} r_{xy}, \quad a, b \neq 0, \tag{C-7}$$

$$s_{ax} = |a|\; s_x, \tag{C-8}$$

$$s_{ax,\,by} = (ab) s_{xy}.$$

Note that these algebraic results parallel the theoretical results for bivariate probability distributions. [We note in passing, while the formulas in (C-2) and (C-5) are algebraically the same, (C-2) will generally be more accurate in practice, especially when the values in the sample are very widely dispersed.]

Example C.1 Descriptive Statistics for a Random Sample

Appendix Table FC.1 contains a (hypothetical) sample of observations on income and education (The observations all appear in the calculations of the means below.) A scatter diagram appears in Figure C.1. It suggests a weak positive association between income and education in these data. The box and whisker plot for income at the left of the scatter plot shows the distribution of the income data as well.

$$\textit{Means}: \bar{I} = \frac{1}{20}\begin{bmatrix} 20.5 + 31.5 + 47.7 + 26.2 + 44.0 + 8.28 + 30.8 + \\ 17.2 + 19.9 + 9.96 + 55.8 + 25.2 + 29.0 + 85.5 + \\ 15.1 + 28.5 + 21.4 + 17.7 + 6.42 + 84.9 \end{bmatrix} = 31.278,$$

$$\bar{E} = \frac{1}{20}\begin{bmatrix} 12 + 16 + 18 + 16 + 12 + 12 + 16 + 12 + 10 + 12 + \\ 16 + 20 + 12 + 16 + 10 + 18 + 16 + 20 + 12 + 16 \end{bmatrix} = 14.600.$$

Standard deviations:

$$s_I = \sqrt{\tfrac{1}{19}[(20.5 - 31.278)^2 + \cdots + (84.9 - 31.278)^2]} = 22.376,$$

$$s_E = \sqrt{\tfrac{1}{19}[(12 - 14.6)^2 + \cdots + (16 - 14.6)^2]} = 3.119.$$

Covariance: $s_{IE} = \frac{1}{19}[20.5(12) + \cdots + 84.9(16) - 20(31.28)(14.6)] = 23.597,$

Correlation: $r_{IE} = \dfrac{23.597}{(22.376)(3.119)} = 0.3382.$

The positive correlation is consistent with our observation in the scatter diagram.

The statistics just described will provide the analyst with a more concise description of the data than a raw tabulation. However, we have not, as yet, suggested that these measures correspond to some underlying characteristic of the process that generated the data. We do assume that there is an underlying mechanism, the data generating process that produces the data in hand. Thus, these serve to do more than describe the data; they characterize that process, or population. Because we have assumed that there is an underlying probability distribution, it might be useful to produce a statistic that gives a broader view of the DGP. The **histogram** is a simple graphical device that produces this result—see Examples C.3 and C.4 for applications. For small samples or widely dispersed data, however, histograms tend to be rough and difficult to make informative. A burgeoning literature[2] has demonstrated the usefulness of the **kernel density estimator** as a substitute for the histogram as a descriptive tool for the underlying distribution that produced a sample of data. The underlying theory of the kernel density estimator is fairly complicated, but the computations are surprisingly simple. The estimator is computed using

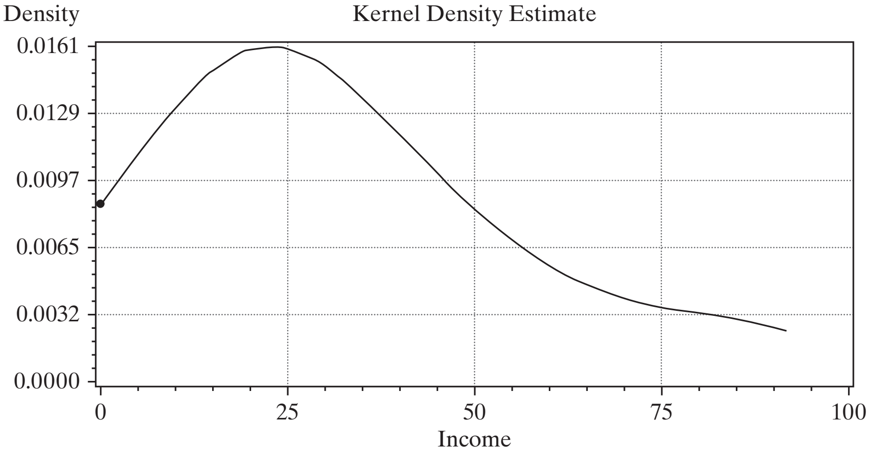
$$\hat{f}(x^*) = \frac{1}{nh}\sum_{i=1}^{n} K\left[\frac{x_i - x^*}{h}\right],$$

where $x_1, \ldots, x_n$ are the n observations in the sample, $\hat{f}(x^*)$ denotes the estimated density function, x^* is the value at which we wish to evaluate the density, and h and $K[\cdot]$ are the "bandwidth" and "kernel function" that we now consider. The density estimator is rather like a histogram, in which the bandwidth is the width of the intervals. The kernel function is a weight function which is generally chosen so that it takes large values when x^* is close to x_i and tapers off to zero in as they diverge in either direction. The weighting function used in the following example is the logistic density discussed in Section B.4.7. The bandwidth is chosen to be a function of $1/n$ so that the intervals can become narrower as the sample becomes larger (and richer). The one used for Figure C.2 is $h = 0.9 \text{ Min } (s, \text{range}/3)/n^{.2}$. (We will revisit this method of estimation in Chapter 12.) Example C.2 illustrates the computation for the income data used in Example C.1.

Example C.2 Kernel Density Estimator for the Income Data

Figure C.2 suggests the large skew in the income data that is also suggested by the box and whisker plot (and the scatter plot in Example C.1.)

FIGURE C.2 Kernel Density Estimate for Income.



[2] See for example, Pagan and Ullah (1999), Li and Racine (2007) and Henderson and Parmeter (2015).

C.4 STATISTICS AS ESTIMATORS—SAMPLING DISTRIBUTIONS

The measures described in the preceding section summarize the data in a random sample. Each measure has a counterpart in the population, that is, the distribution from which the data were drawn. Sample quantities such as the means and the correlation coefficient correspond to population expectations, whereas the kernel density estimator and the values in Table C.1 parallel the population pdf and cdf. In the setting of a random sample, we expect these quantities to mimic the population, although not perfectly. The precise manner in which these quantities reflect the population values defines the sampling distribution of a sample statistic.

DEFINITION C.1 Statistic
A statistic is any function computed from the data in a sample.

If another sample were drawn under identical conditions, different values would be obtained for the observations, as each one is a random variable. Any statistic is a function of these random values, so it is also a random variable with a probability distribution called a **sampling distribution**. For example, the following shows an exact result for the sampling behavior of a widely used statistic.

THEOREM C.1 Sampling Distribution of the Sample Mean
If $x_1, \ldots, x_n$ are a random sample from a population with mean μ and variance σ^2, then $\overline{x}$ is a random variable with mean μ and variance σ^2/n.
Proof: $\overline{x} = (1/n)\Sigma_i x_i$. $E[\overline{x}] = (1/n)\Sigma_i \mu = \mu$. *The observations are independent, so* $\text{Var}[\overline{x}] = (1/n)^2 \text{ Var}[\Sigma_i x_i] = (1/n^2)\Sigma_i \sigma^2 = \sigma^2/n$.

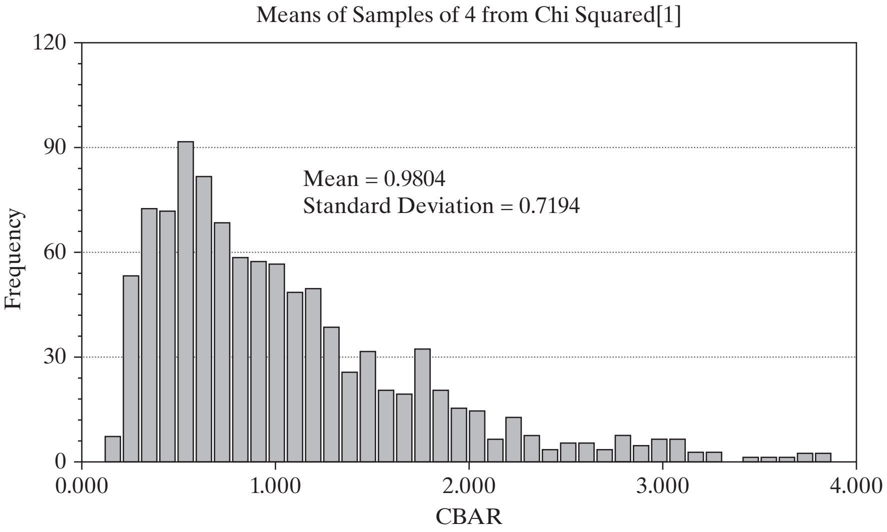
Example C.3 illustrates the behavior of the sample mean in samples of four observations drawn from a chi-squared population with one degree of freedom. The crucial concepts illustrated in this example are, first, the mean and variance results in Theorem C.1 and, second, the phenomenon of **sampling variability**.

Notice that the fundamental result in Theorem C.1 does not assume a distribution for x_i. Indeed, looking back at Section C.3, nothing we have done so far has required any assumption about a particular distribution.

TABLE C.1 Income Distribution

Range	*Relative Frequency*	*Cumulative Frequency*
<\$10,000	0.15	0.15
10,000–25,000	0.30	0.45
25,000–50,000	0.40	0.85
>50,000	0.15	1.00

FIGURE C.3 Sampling Distribution of Means of 1,000 Samples of Size 4 from Chi-Squared[1].



Example C.3 Sampling Distribution of a Sample Mean

Figure C.3 shows a frequency plot of the means of 1,000 random samples of four observations drawn from a chi-squared distribution with one degree of freedom, which has mean 1 and variance 2.

We are often interested in how a statistic behaves as the sample size increases. Example C.4 illustrates one such case. Figure C.4 shows two sampling distributions, one based on samples of three and a second, of the same statistic, but based on samples of six. The effect of increasing sample size in this figure is unmistakable. It is easy to visualize the behavior of this statistic if we extrapolate the experiment in Example C.4 to samples of, say, 100.

Example C.4 Sampling Distribution of the Sample Minimum

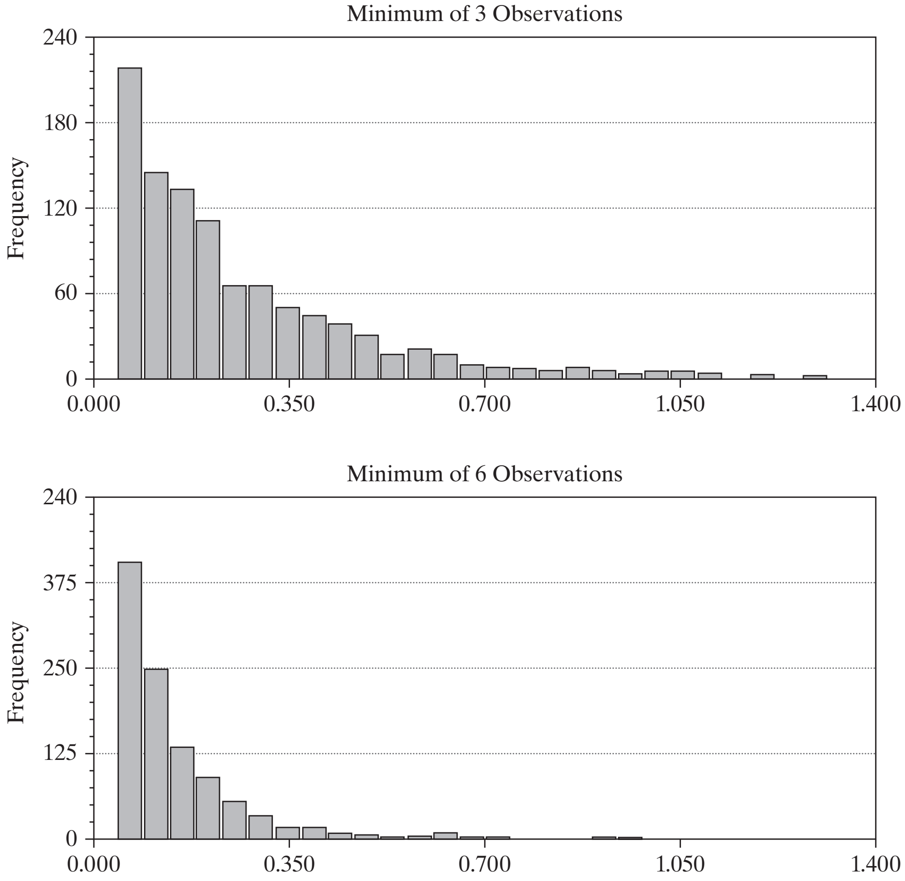
If $x_1, \ldots, x_n$ are a random sample from an exponential distribution with $f(x) = \theta e^{-\theta x}$, then the sampling distribution of the sample minimum in a sample of n observations, denoted $x_{(1)}$, is

$$f(x_{(1)}) = (n\theta)e^{-(n\theta)x_{(1)}}.$$

Because $E[x] = 1/\theta$ and $\text{Var}[x] = 1/\theta^2$, by analogy $E[x_{(1)}] = 1/(n\theta)$ and $\text{Var}[x_{(1)}] = 1/(n\theta)^2$. Thus, in increasingly larger samples, the minimum will be arbitrarily close to 0. [The Chebychev inequality in Theorem D.2 can be used to prove this intuitively appealing result.]

Figure C.4 shows the results of a simple sampling experiment you can do to demonstrate this effect. It requires software that will allow you to produce pseudorandom numbers uniformly distributed in the range zero to one and that will let you plot a histogram and control the axes. (We used NLOGIT. This can be done with Stata, Excel, or several other packages.) The experiment consists of drawing 1,000 sets of nine random values,

FIGURE C.4 Histograms of the Sample Minimum of 3 and 6 Observations.



U_{ij}, $i = 1, \ldots 1{,}000$, $j = 1, \ldots, 9$. To transform these uniform draws to exponential with parameter θ—we used $\theta = 1.5$, use the inverse probability transform—see Section E.2.3. For an exponentially distributed variable, the transformation is $z_{ij} = -(1/\theta)\log(1 - U_{ij})$. We then created $z_{(1)}|3$ from the first three draws and $z_{(1)}|6$ from the other six. The two histograms show clearly the effect on the sampling distribution of increasing sample size from just 3 to 6.

Sampling distributions are used to make inferences about the population. To consider a perhaps obvious example, because the sampling distribution of the mean of a set of normally distributed observations has mean μ, the sample mean is a natural candidate for an estimate of μ. The observation that the sample "mimics" the population is a statement about the sampling distributions of the sample statistics. Consider, for example, the sample data collected in Figure C.3. The sample mean of four observations clearly has a sampling distribution, which appears to have a mean roughly equal to the population mean. Our theory of parameter estimation departs from this point.

C.5 POINT ESTIMATION OF PARAMETERS

Our objective is to use the sample data to infer the value of a parameter or set of parameters, which we denote θ. A **point estimate** is a statistic computed from a sample that gives a single value for θ. The **standard error** of the estimate is the standard deviation of the sampling distribution of the statistic; the square of this quantity is the **sampling variance**. An **interval estimate** is a range of values that will contain the true parameter with a preassigned probability. There will be a connection between the two types of estimates; generally, if $\hat{\theta}$ is the point estimate, then the interval estimate will be $\hat{\theta} \pm$ a measure of sampling error.

An **estimator** is a rule or strategy for using the data to estimate the parameter. It is defined before the data are drawn. Obviously, some estimators are better than others. To take a simple example, your intuition should convince you that the sample mean would be a better estimator of the population mean than the sample minimum; the minimum is almost certain to underestimate the mean. Nonetheless, the minimum is not entirely without virtue; it is easy to compute, which is occasionally a relevant criterion. The search for good estimators constitutes much of econometrics. Estimators are compared on the basis of a variety of attributes. **Finite sample properties** of estimators are those attributes that can be compared regardless of the sample size. Some estimation problems involve characteristics that are not known in finite samples. In these instances, estimators are compared on the basis on their large sample, or **asymptotic properties**. We consider these in turn.

C.5.1 ESTIMATION IN A FINITE SAMPLE

The following are some finite sample estimation criteria for estimating a single parameter. The extensions to the multiparameter case are direct. We shall consider them in passing where necessary.

DEFINITION C.2 Unbiased Estimator
An estimator of a parameter θ is unbiased if the mean of its sampling distribution is θ. Formally,

$$E[\hat{\theta}] = \theta$$

or

$$E[\hat{\theta} - \theta] = \text{Bias}[\hat{\theta} | \theta] = 0$$

implies that $\hat{\theta}$ is unbiased. Note that this implies that the expected sampling error is zero. If $\boldsymbol{\theta}$ is a vector of parameters, then the estimator is unbiased if the expected value of every element of $\hat{\boldsymbol{\theta}}$ equals the corresponding element of $\boldsymbol{\theta}$.

If samples of size n are drawn repeatedly and $\hat{\theta}$ is computed for each one, then the average value of these estimates will tend to equal θ. For example, the average of the 1,000 sample means underlying Figure C.3 is 0.9804, which is reasonably close to the population mean of one. The sample minimum is clearly a biased estimator of the mean; it will almost always underestimate the mean, so it will do so on average as well.

Unbiasedness is a desirable attribute, but it is rarely used by itself as an estimation criterion. One reason is that there are many unbiased estimators that are poor uses of the data. For example, in a sample of size n, the first observation drawn is an unbiased estimator of the mean that clearly wastes a great deal of information. A second criterion used to choose among unbiased estimators is efficiency.

DEFINITION C.3 Efficient Unbiased Estimator
An unbiased estimator $\hat{\theta}_1$ is more efficient than another unbiased estimator $\hat{\theta}_2$ if the sampling variance of $\hat{\theta}_1$ is less than that of $\hat{\theta}_2$. That is,

$$\text{Var}[\hat{\theta}_1] < \text{Var}[\hat{\theta}_2].$$

In the multiparameter case, the comparison is based on the covariance matrices of the two estimators; $\hat{\boldsymbol{\theta}}_1$ is more efficient than $\hat{\boldsymbol{\theta}}_2$ if $\text{Var}[\hat{\boldsymbol{\theta}}_2] - \text{Var}[\hat{\boldsymbol{\theta}}_1]$ *is a positive definite matrix.*

By this criterion, the sample mean is obviously to be preferred to the first observation as an estimator of the population mean. If σ^2 is the population variance, then

$$\text{Var}[x_1] = \sigma^2 > \text{Var}[\bar{x}] = \frac{\sigma^2}{n}.$$

In discussing efficiency, we have restricted the discussion to unbiased estimators. Clearly, there are biased estimators that have smaller variances than the unbiased ones we have considered. Any constant has a variance of zero. Of course, using a constant as an estimator is not likely to be an effective use of the sample data. Focusing on unbiasedness may still preclude a tolerably biased estimator with a much smaller variance, however. A criterion that recognizes this possible tradeoff is the mean squared error. Figure C.5 illustrates the effect. In this example,

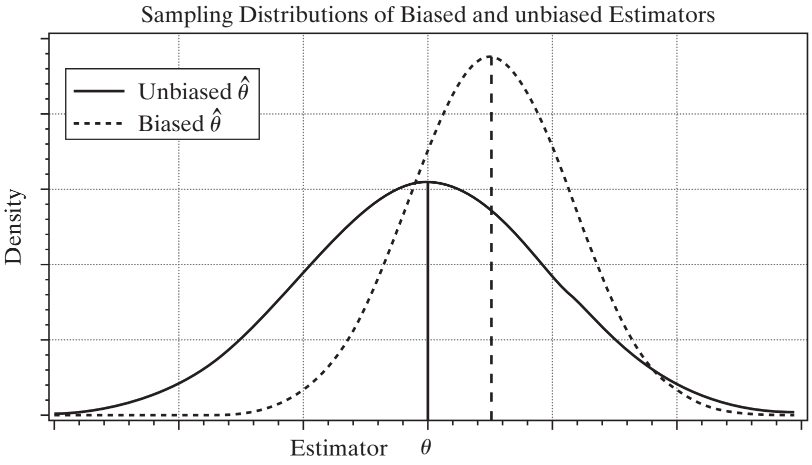
DEFINITION C.4 Mean Squared Error
The mean squared error of an estimator is

$$\begin{aligned} \text{MSE}[\hat{\theta}\,|\,\theta] &= E[(\hat{\theta} - \theta)^2] \\ &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}\,|\,\theta])^2 && \text{if } \theta \text{ is a scalar,} \\ \text{MSE}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}] &= \text{Var}[\hat{\boldsymbol{\theta}}] + \text{Bias}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}]\text{Bias}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}]' && \text{if } \boldsymbol{\theta} \text{ is a vector.} \end{aligned} \quad \textbf{(C-9)}$$

on average, the biased estimator will be closer to the true parameter than will the unbiased estimator.

Which of these criteria should be used in a given situation depends on the particulars of that setting and our objectives in the study. Unfortunately, the MSE criterion is rarely

FIGURE C.5 Sampling Distributions.



operational; minimum mean squared error estimators, when they exist at all, usually depend on unknown parameters. Thus, we are usually less demanding. A commonly used criterion is **minimum variance unbiasedness**.

Example C.5 Mean Squared Error of the Sample Variance

In sampling from a normal distribution, the most frequently used estimator for $\boldsymbol{\sigma}^2$ is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}.$$

It is straightforward to show that s^2 is unbiased, so

$$\text{Var}[s^2] = \frac{2\sigma^4}{n - 1} = \text{MSE}[s^2 \mid \sigma^2].$$

A proof is based on the distribution of the idempotent quadratic form $(\mathbf{x} - \mathbf{i}\mu)'\mathbf{M}^0(\mathbf{x} - \mathbf{i}\mu)$, which we discussed in Section B.11.4. A less frequently used estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = [(n - 1)/n]s^2.$$

This estimator is slightly biased downward:

$$E[\hat{\sigma}^2] = \frac{(n - 1)E(s^2)}{n} = \frac{(n - 1)\sigma^2}{n},$$

so its bias is

$$E[\hat{\sigma}^2 - \sigma^2] = \text{Bias}[\hat{\sigma}^2 \mid \sigma^2] = \frac{-1}{n}\sigma^2.$$

But it has a smaller variance than s^2:

$$\text{Var}[\hat{\sigma}^2] = \left[\frac{n - 1}{n}\right]^2\left[\frac{2\sigma^4}{n - 1}\right] < \text{Var}[s^2].$$

To compare the two estimators, we can use the difference in their mean squared errors:

$$\text{MSE}[\hat{\sigma}^2 | \sigma^2] - \text{MSE}[s^2 | \sigma^2] = \sigma^4 \left[\frac{2n - 1}{n^2} - \frac{2}{n - 1} \right] < 0.$$

The biased estimator is a bit more precise. The difference will be negligible in a large sample, but, for example, it is about 1.2 percent in a sample of 16.

C.5.2 EFFICIENT UNBIASED ESTIMATION

In a random sample of n observations, the density of each observation is $f(x_i, \theta)$. Because the n observations are independent, their joint density is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n, \theta) &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \\ &= \prod_{i=1}^{n} f(x_i, \theta) = L(\theta | x_1, x_2, \ldots, x_n). \end{aligned} \tag{C-10}$$

This function, denoted $L(\theta | \mathbf{X})$, is called the likelihood function for θ given the data $\mathbf{X}$. It is frequently abbreviated to $L(\theta)$. Where no ambiguity can arise, we shall abbreviate it further to L.

Example C.6 Likelihood Functions for Exponential and Normal Distributions

If $x_1, \ldots, x_n$ are a sample of n observations from an exponential distribution with parameter θ, then

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \Sigma^n{}_{i=1} x_i}.$$

If $x_1, \ldots, x_n$ are a sample of n observations from a normal distribution with mean μ and standard deviation σ, then

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-[1/(2\sigma^2)](x_i - \mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-[1/(2\sigma^2)]\Sigma_i (x_i - \mu)^2}. \end{aligned} \tag{C-11}$$

The likelihood function is the cornerstone for most of our theory of parameter estimation. An important result for efficient estimation is the following.

THEOREM C.2 Cramér–Rao Lower Bound

Assuming that the density of x satisfies certain regularity conditions, the variance of an unbiased estimator of a parameter θ will always be at least as large as

$$[I(\theta)]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] \right)^{-1} = \left(E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right] \right)^{-1}. \tag{C-12}$$

The quantity $I(\theta)$ is the information number for the sample. We will prove the result that the negative of the expected second derivative equals the expected square of the first derivative in Chapter 14. Proof of the main result of the theorem is quite involved. See, for example, Stuart and Ord (1989).

The regularity conditions are technical. (See Section 14.4.1.) Loosely, they are conditions imposed on the density of the random variable that appears in the likelihood function; these conditions will ensure that the Lindeberg–Levy central limit theorem will apply to moments of the sample of observations on the random vector $\mathbf{y} = \partial \ln f(x_i|\theta)/\partial\theta,\ i = 1, \ldots, n$. Among the conditions are finite moments of x up to order 3. An additional condition usually included in the set is that the range of the random variable be independent of the parameters.

In some cases, the second derivative of the log likelihood is a constant, so the Cramér–Rao bound is simple to obtain. For instance, in sampling from an exponential distribution, from Example C.6,

$$\ln L = n \ln \theta - \theta \sum_{i=1}^{n} x_i,$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i,$$

so $\partial^2 \ln L/\partial\theta^2 = -n/\theta^2$ and the variance bound is $[I(\theta)]^{-1} = \theta^2/n$. In many situations, the second derivative is a random variable with a distribution of its own. The following examples show two such cases.

Example C.7 Variance Bound for the Poisson Distribution

For the Poisson distribution,

$$f(x) = \frac{e^{-\theta}\theta^x}{x!},$$

$$\ln L = -n\theta + \left(\sum_{i=1}^{n} x_i\right) \ln \theta - \sum_{i=1}^{n} \ln(x_i\,!),$$

$$\frac{\partial \ln L}{\partial \theta} = -n + \frac{\sum_{i=1}^{n} x_i}{\theta},$$

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{-\sum_{i=1}^{n} x_i}{\theta^2}.$$

The sum of n identical Poisson variables has a Poisson distribution with parameter equal to n times the parameter of the individual variables. Therefore, the actual distribution of the first derivative will be that of a linear function of a Poisson distributed variable. Because $E[\sum_{i=1}^{n} x_i] = nE[x_i] = n\theta$, the variance bound for the Poisson distribution is $[I(\theta)]^{-1} = \theta/n$. (Note also that the same result implies that $E[\partial \ln L/\partial\theta] = 0$, which is a result we will use in Chapter 14. The same result holds for the exponential distribution.)

Consider, finally, a multivariate case. If $\boldsymbol{\theta}$ is a vector of parameters, then $\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix**. The Cramér–Rao theorem states that the difference between the covariance matrix of any unbiased estimator and the inverse of the information matrix,

$$[\mathbf{I}(\boldsymbol{\theta})]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right)^{-1} = \left\{E\left[\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]\right\}^{-1}, \qquad \textbf{(C-13)}$$

will be a nonnegative definite matrix.

In some settings, numerous estimators are available for the parameters of a distribution. The usefulness of the Cramér–Rao bound is that if one of these is known to attain the variance bound, then there is no need to consider any other to seek a more efficient estimator. Regarding the use of the variance bound, we emphasize that if an unbiased estimator attains it, then that estimator is efficient. If a given estimator does not attain the variance bound, however, then we do not know, except in a few special cases, whether this estimator is efficient or not. It may be that no unbiased estimator can attain the Cramér–Rao bound, which can leave the question of whether a given unbiased estimator is efficient or not unanswered.

We note, finally, that in some cases we further restrict the set of estimators to linear functions of the data.

DEFINITION C.5 Minimum Variance Linear Unbiased Estimator (MVLUE)
An estimator is the minimum variance linear unbiased estimator or best linear unbiased estimator (BLUE) if it is a linear function of the data and has minimum variance among linear unbiased estimators.

In a few instances, such as the normal mean, there will be an efficient linear unbiased estimator; $\bar{x}$ is efficient among all unbiased estimators, both linear and nonlinear. In other cases, such as the normal variance, there is no linear unbiased estimator. This criterion is useful because we can sometimes find an MVLUE without having to specify the distribution at all. Thus, by limiting ourselves to a somewhat restricted class of estimators, we free ourselves from having to assume a particular distribution.

C.6 INTERVAL ESTIMATION

Regardless of the properties of an estimator, the estimate obtained will vary from sample to sample, and there is some probability that it will be quite erroneous. A point estimate will not provide any information on the likely range of error. The logic behind an **interval estimate** is that we use the sample data to construct an interval, [lower (**X**), upper (**X**)], such that we can expect this interval to contain the true parameter in some specified proportion of samples, or equivalently, with some desired level of confidence. Clearly, the wider the interval, the more confident we can be that it will, in any given sample, contain the parameter being estimated.

The theory of interval estimation is based on a **pivotal quantity**, which is a function of both the parameter and a point estimate that has a known distribution. Consider the following examples.

Example C.8 Confidence Intervals for the Normal Mean

In sampling from a normal distribution with mean μ and standard deviation σ,

$$z = \frac{\sqrt{n}(\bar{x} - \mu)}{s} \sim t[n - 1],$$

and

$$c = \frac{(n - 1)s^2}{\sigma^2} \sim \chi^2[n - 1].$$

Given the pivotal quantity, we can make probability statements about events involving the parameter and the estimate. Let $p(g, \theta)$ be the constructed random variable, for example, z or c. Given a prespecified **confidence level**, $1 - \alpha$, we can state that

$$\text{Prob}(\text{lower} \leq p(g, \theta) \leq \text{upper}) = 1 - \alpha, \qquad \textbf{(C-14)}$$

where lower and upper are obtained from the appropriate table. This statement is then manipulated to make equivalent statements about the endpoints of the intervals. For example, the following statements are equivalent:

$$\text{Prob}(-z \leq \frac{\sqrt{n}(\bar{x} - \mu)}{s} \leq z) = 1 - \alpha,$$

$$\text{Prob}\left(\bar{x} - \frac{zs}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{zs}{\sqrt{n}}\right) = 1 - \alpha.$$

The second of these is a statement about the interval, not the parameter; that is, it is the interval that is random, not the parameter. We attach a probability, or $100(1 - \alpha)$ percent confidence level, to the interval itself; in repeated sampling, an interval constructed in this fashion will contain the true parameter $100(1 - \alpha)$ percent of the time.

In general, the interval constructed by this method will be of the form

$$\text{lower}(\mathbf{X}) = \hat{\theta} - e_1,$$
$$\text{upper}(\mathbf{X}) = \hat{\theta} + e_2,$$

where $\mathbf{X}$ is the sample data, e_1 and e_2 are sampling errors, and $\hat{\theta}$ is a point estimate of θ. It is clear from the preceding example that if the sampling distribution of the pivotal quantity is either t or standard normal, which will be true in the vast majority of cases we encounter in practice, then the confidence interval will be

$$\hat{\theta} \pm C_{1-\alpha/2}[\text{se}(\hat{\theta})], \qquad \textbf{(C-15)}$$

where se (.) is the (known or estimated) standard error of the parameter estimate and $C_{1-\alpha/2}$ is the value from the t or standard normal distribution that is exceeded with probability $1 - \alpha/2$. The usual values for α are 0.10, 0.05, or 0.01. The theory does not prescribe exactly how to choose the endpoints for the confidence interval. An obvious criterion is to minimize the width of the interval. If the sampling distribution is symmetric, then the symmetric interval is the best one. If the sampling distribution is not symmetric, however, then this procedure will not be optimal.

Example C.9 Estimated Confidence Intervals for a Normal Mean and Variance

In a sample of 25, $\bar{x} = 1.63$ and $s = 0.51$. Construct a 95 percent confidence interval for μ.

Assuming that the sample of 25 is from a normal distribution,

$$\text{Prob}\left(-2.064 \leq \frac{5(\bar{x} - \mu)}{s} \leq 2.064\right) = 0.95,$$

where 2.064 is the critical value from a t distribution with 24 degrees of freedom. Thus, the confidence interval is $1.63 \pm [2.064(0.51)/5]$ or [1.4195, 1.8405].

Remark: Had the parent distribution not been specified, it would have been natural to use the standard normal distribution instead, perhaps relying on the central limit theorem. But a sample size of 25 is small enough that the more conservative t distribution might still be preferable.

The chi-squared distribution is used to construct a confidence interval for the variance of a normal distribution. Using the data from Example C.9, we find that the usual procedure would use

$$\text{Prob}\left(12.4 \leq \frac{24s^2}{\sigma^2} \leq 39.4\right) = 0.95,$$

where 12.4 and 39.4 are the 0.025 and 0.975 cutoff points from the chi-squared (24) distribution. This procedure leads to the 95 percent confidence interval [0.1581, 0.5032]. By making use of the asymmetry of the distribution, a narrower interval can be constructed. Allocating 4 percent to the left-hand tail and 1 percent to the right instead of 2.5 percent to each, the two cutoff points are 13.4 and 42.9, and the resulting 95 percent confidence interval is [0.1455, 0.4659].

Finally, the confidence interval can be manipulated to obtain a confidence interval for a function of a parameter. For example, based on the preceding, a 95 percent confidence interval for σ would be $[\sqrt{0.1581}, \sqrt{0.5032}] = [0.3976, 0.7094]$.

C.7 HYPOTHESIS TESTING

The second major group of statistical inference procedures is hypothesis tests. The classical testing procedures are based on constructing a statistic from a random sample that will enable the analyst to decide, with reasonable confidence, whether or not the data in the sample would have been generated by a hypothesized population. The formal procedure involves a statement of the hypothesis, usually in terms of a "null" or maintained hypothesis and an "alternative," conventionally denoted H_0 and H_1, respectively. The procedure itself is a rule, stated in terms of the data, that dictates whether the null hypothesis should be rejected or not. For example, the hypothesis might state a parameter is equal to a specified value. The decision rule might state that the hypothesis should be rejected if a sample estimate of that parameter is too far away from that value (where "far" remains to be defined). The classical, or Neyman–Pearson, methodology involves partitioning the sample space into two regions. If the observed data (i.e., the test statistic) fall in the **rejection region** (sometimes called the **critical region**), then the null hypothesis is rejected; if they fall in the **acceptance region**, then it is not.

C.7.1 CLASSICAL TESTING PROCEDURES

Because the sample is random, the test statistic, however defined, is also random. The same test procedure can lead to different conclusions in different samples. As such, there are two ways such a procedure can be in error:

1. **Type I error.** The procedure may lead to rejection of the null hypothesis when it is true.
2. **Type II error.** The procedure may fail to reject the null hypothesis when it is false.

To continue the previous example, there is some probability that the estimate of the parameter will be quite far from the hypothesized value, even if the hypothesis is true. This outcome might cause a type I error.

DEFINITION C.6 Size of a Test
The probability of a type I error is the **size** *of the test. This is conventionally denoted α and is also called the* **significance level**.

The size of the test is under the control of the analyst. It can be changed just by changing the decision rule. Indeed, the type I error could be eliminated altogether just by making the rejection region very small, but this would come at a cost. By eliminating the probability of a type I error—that is, by making it unlikely that the hypothesis is rejected—we must increase the probability of a type II error. Ideally, we would like both probabilities to be as small as possible. It is clear, however, that there is a tradeoff between the two. The best we can hope for is that for a given probability of type I error, the procedure we choose will have as small a probability of type II error as possible.

DEFINITION C.7 Power of a Test
The ***power*** *of a test is the probability that it will correctly lead to rejection of a false null hypothesis:*

$$\text{power} = 1 - \beta = 1 - \text{Prob(type II error)}. \quad \textbf{(C-16)}$$

For a given significance level α, we would like β to be as small as possible. Because β is defined in terms of the alternative hypothesis, it depends on the value of the parameter.

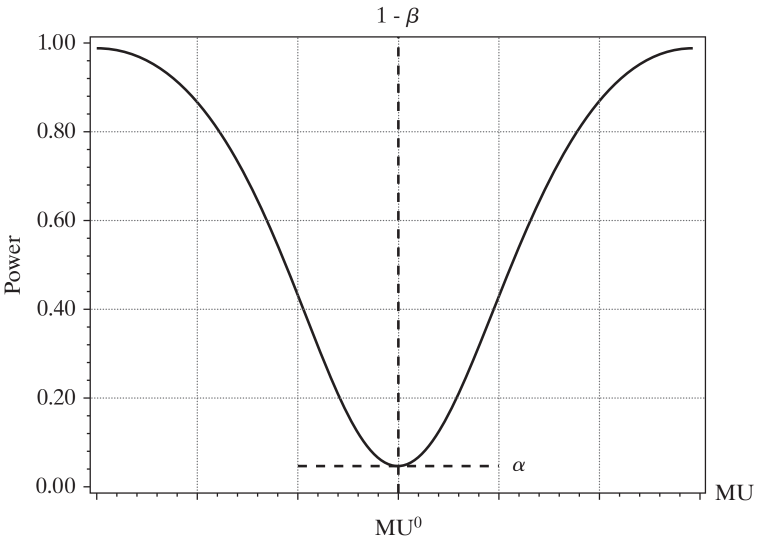
Example C.10 Testing a Hypothesis About a Mean

For testing H_0: $\mu = \mu^0$ in a normal distribution with known variance σ^2, the decision rule is to reject the hypothesis if the absolute value of the z statistic, $\sqrt{n}(\bar{x} - \mu^0)/\sigma$, exceeds the predetermined critical value. For a test at the 5 percent significance level, we set the critical value at 1.96. The power of the test, therefore, is the probability that the absolute value of the test statistic will exceed 1.96 given that the true value of μ is, in fact, not μ^0. This value depends on the alternative value of μ, as shown in Figure C.6. Notice that for this test the power is equal to the size at the point where μ equals μ^0. As might be expected, the test becomes more powerful the farther the true mean is from the hypothesized value.

Testing procedures, like estimators, can be compared using a number of criteria.

DEFINITION C.8 Most Powerful Test
A test is ***most powerful*** *if it has greater power than any other test of the same size.*

FIGURE C.6 Power Function for a Test.



This requirement is very strong. Because the power depends on the alternative hypothesis, we might require that the test be **uniformly most powerful (UMP)**, that is, have greater power than any other test of the same size for all admissible values of the parameter. There are few situations in which a UMP test is available. We usually must be less stringent in our requirements. Nonetheless, the criteria for comparing hypothesis testing procedures are generally based on their respective power functions. A common and very modest requirement is that the test be unbiased.

DEFINITION C.9 Unbiased Test
A test is ***unbiased*** *if its power* $(1 - \beta)$ *is greater than or equal to its size* α *for all values of the parameter.*

If a test is **biased**, then, for some values of the parameter, we are more likely to retain the null hypothesis when it is false than when it is true.

The use of the term *unbiased* here is unrelated to the concept of an unbiased estimator. Fortunately, there is little chance of confusion. Tests and estimators are clearly connected, however. The following criterion derives, in general, from the corresponding attribute of a parameter estimate.

DEFINITION C.10 Consistent Test
A test is ***consistent*** *if its power goes to one as the sample size grows to infinity.*

Example C.11 Consistent Test About a Mean

A confidence interval for the mean of a normal distribution is $\bar{x} \pm t_{1-\alpha/2}(s/\sqrt{n})$, where $\bar{x}$ and s are the usual consistent estimators for μ and σ (see Section D.2.1), n is the sample size, and $t_{1-\alpha/2}$ is the correct critical value from the t distribution with $n - 1$ degrees of freedom. For testing H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, let the procedure be to reject H_0 if the confidence interval does not contain μ_0. Because $\bar{x}$ is consistent for μ, one can discern if H_0 is false as $n \rightarrow \infty$, with probability 1, because $\bar{x}$ will be arbitrarily close to the true μ. Therefore, this test is consistent.

As a general rule, a test will be consistent if it is based on a consistent estimator of the parameter.

C.7.2 TESTS BASED ON CONFIDENCE INTERVALS

There is an obvious link between interval estimation and the sorts of hypothesis tests we have been discussing here. The confidence interval gives a range of plausible values for the parameter. Therefore, it stands to reason that if a hypothesized value of the parameter does not fall in this range of plausible values, then the data are not consistent with the hypothesis, and it should be rejected. Consider, then, testing

$$H_0\colon \theta = \theta_0,\ H_1\colon \theta \neq \theta_0.$$

We form a confidence interval based on $\hat{\theta}$ as described earlier:

$$\hat{\theta} - C_{1-\alpha/2}[\mathrm{se}(\hat{\theta})] < \theta < \hat{\theta} + C_{1-\alpha/2}[\mathrm{se}(\hat{\theta})].$$

H_0 is rejected if θ_0 exceeds the upper limit or is less than the lower limit. Equivalently, H_0 is rejected if

$$\left|\frac{\hat{\theta} - \theta_0}{\mathrm{se}(\hat{\theta})}\right| > C_{1-\alpha/2}.$$

In words, the hypothesis is rejected if the estimate is too far from θ_0, where the distance is measured in standard error units. The critical value is taken from the t or standard normal distribution, whichever is appropriate.

Example C.12 Testing a Hypothesis About a Mean with a Confidence Interval

For the results in Example C.8, test H_0: $\mu = 1.98$ versus H_1: $\mu \neq 1.98$, assuming sampling from a normal distribution:

$$t = \left|\frac{\bar{x} - 1.98}{s/\sqrt{n}}\right| = \left|\frac{1.63 - 1.98}{0.102}\right| = 3.43.$$

The 95 percent critical value for $t(24)$ is 2.064. Therefore, reject H_0. If the critical value for the standard normal table of 1.96 is used instead, then the same result is obtained.

If the test is one-sided, as in

$$H_0\colon \theta \geq \theta_0,$$
$$H_1\colon \theta < \theta_0,$$

then the critical region must be adjusted. Thus, for this test, H_0 will be rejected if a point estimate of θ falls sufficiently below θ_0. (Tests can usually be set up by departing from the decision criterion, "What sample results are inconsistent with the hypothesis?")

Example C.13 One-Sided Test About a Mean

A sample of 25 from a normal distribution yields $\bar{x} = 1.63$ and $s = 0.51$. Test

$$H_0\colon \mu \leq 1.5,$$
$$H_1\colon \mu > 1.5.$$

Clearly, no observed $\bar{x}$ less than or equal to 1.5 will lead to rejection of H_0. Using the borderline value of 1.5 for μ, we obtain

$$\text{Prob}\left(\frac{\sqrt{n}(\bar{x} - 1.5)}{s} > \frac{5(1.63 - 1.5)}{0.51}\right) = \text{Prob}(t_{24} > 1.27).$$

This is approximately 0.11. This value is not unlikely by the usual standards. Hence, at a significant level of 0.11, we would not reject the hypothesis.

C.7.3 SPECIFICATION TESTS

The hypothesis testing procedures just described are known as classical testing procedures. In each case, the null hypothesis tested came in the form of a restriction on the alternative. You can verify that in each application we examined, the parameter space assumed under the null hypothesis is a subspace of that described by the alternative. For that reason, the models implied are said to be *nested*. The null hypothesis is contained within the alternative. This approach suffices for most of the testing situations encountered in practice, but there are common situations in which two competing models cannot be viewed in these terms. For example, consider a case in which there are two completely different, competing theories to explain the same observed data. Many models for censoring and truncation discussed in Chapter 19 rest upon a fragile assumption of normality, for example. Testing of this nature requires a different approach from the classical procedures discussed here. These are discussed at various points throughout the book, for example, in Chapter 19, where we study the difference between fixed and random effects models.

APPENDIX D

LARGE-SAMPLE DISTRIBUTION THEORY

D.1 INTRODUCTION

Most of this book is about parameter estimation. In studying that subject, we will usually be interested in determining how best to use the observed data when choosing among competing estimators. That, in turn, requires us to examine the sampling behavior of estimators. In a few cases, such as those presented in Appendix C and the least squares estimator considered in Chapter 4, we can make broad statements about sampling distributions that will apply regardless of the size of the sample. But, in most situations, it will only be possible to make approximate statements about estimators, such as whether they improve as the sample size increases and what can be said about their sampling distributions in large samples as an approximation to the finite samples we actually observe. This appendix will collect most of the formal, fundamental theorems and results needed for this analysis. A few additional results will be developed in the discussion of time-series analysis later in the book.

D.2 LARGE-SAMPLE DISTRIBUTION THEORY[1]

In most cases, whether an estimator is exactly unbiased or what its exact sampling variance is in samples of a given size will be unknown. But we may be able to obtain approximate results about the behavior of the distribution of an estimator as the sample becomes large. For example, it is well known that the distribution of the mean of a sample tends to approximate normality as the sample size grows, regardless of the distribution of the individual observations. Knowledge about the limiting behavior of the distribution of an estimator can be used to infer an approximate distribution for the estimator in a finite sample. To describe how this is done, it is necessary, first, to present some results on convergence of random variables.

D.2.1 CONVERGENCE IN PROBABILITY

Limiting arguments in this discussion will be with respect to the sample size n. Let x_n be a sequence random variable indexed by the sample size.

DEFINITION D.1 Convergence in Probability
The random variable x_n **converges in probability** *to a constant* c *if* $\lim_{n\to\infty}\text{Prob}(|x_n - c| > \varepsilon) = 0$ *for any positive* ε.

Convergence in probability implies that the values that the variable may take that are not close to c become increasingly unlikely as n increases. To consider one example, suppose that the random variable x_n takes two values, zero and n, with probabilities $1 - (1/n)$ and $(1/n)$, respectively. As n increases, the second point will become ever more remote from any constant but, at the same time, will become increasingly less probable. In this example, x_n converges in probability to zero. The crux of this form of convergence is that all the mass of the probability distribution becomes concentrated at points close to c. If x_n converges in probability to c, then we write

$$\text{plim}\, x_n = c. \tag{D-1}$$

[1] A comprehensive summary of many results in large-sample theory appears in White (2001). The results discussed here will apply to samples of independent observations. Time-series cases in which observations are correlated are analyzed in Chapters 20 and 21.

We will make frequent use of a special case of convergence in probability, **convergence in mean square** or **convergence in quadratic mean**.

THEOREM D.1 Convergence in Quadratic Mean
If x_n has mean μ_n and variance σ_n^2 such that the ordinary limits of μ_n and σ_n^2 are c and 0, respectively, then x_n converges in mean square to c , and

$$\text{plim } x_n = c.$$

A proof of Theorem D.1 can be based on another useful theorem.

THEOREM D.2 Chebychev's Inequality
If x_n is a random variable and c and ε are constants, then $\text{Prob}(|x_n - c| > \varepsilon) \leq E[(x_n - c)^2]/\varepsilon^2$.

To establish the Chebychev inequality, we use another result [see Goldberger (1991, p. 31)].

THEOREM D.3 Markov's Inequality
If y_n is a nonnegative random variable and δ is a positive constant, then $\text{Prob}[y_n \geq \delta] \leq E[y_n]/\delta$.
Proof: $E[y_n] = \text{Prob}[y_n < \delta]E[y_n | y_n < \delta] + \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$.
Because y_n is non-negative, both terms must be nonnegative, so $E[y_n] \geq \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$. *Because $E[y_n | y_n \geq \delta]$ must be greater than or equal to δ,* $E[y_n] \geq \text{Prob}[y_n \geq \delta]\delta$, *which is the result.*

Now, to prove Theorem D.1, let y_n be $(x_n - c)^2$ and δ be ε^2 in Theorem D.3. Then, $(x_n - c)^2 > \delta$ implies that $|x_n - c| > \varepsilon$. Finally, we will use a special case of the Chebychev inequality, where $c = \mu_n$, so that we have

$$\text{Prob}(|x_n - \mu_n| > \varepsilon) \leq \sigma_n^2/\varepsilon^2. \quad \textbf{(D-2)}$$

Taking the limits of μ_n and σ_n^2 in (D-2), we see that if

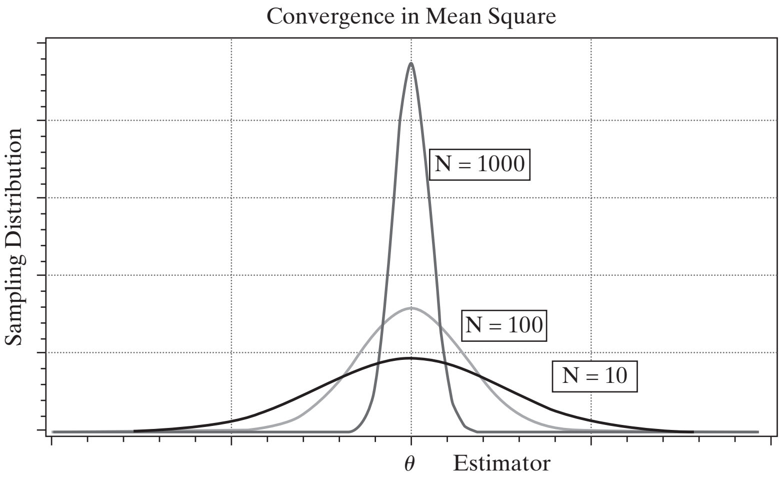
$$\lim_{n\to\infty} E[x_n] = c, \quad \text{and} \quad \lim_{n\to\infty} \text{Var}[x_n] = 0, \quad \textbf{(D-3)}$$

then

$$\text{plim } x_n = c.$$

We have shown that convergence in mean square implies convergence in probability. Mean-square convergence implies that the distribution of x_n collapses to a spike at plim x_n, as shown in Figure D.1.

FIGURE D.1 Quadratic Convergence to a Constant, θ.



Example D.1 Mean Square Convergence of the Sample Minimum in Exponential Sampling

As noted in Example C.4, in sampling of n observations from an exponential distribution, for the sample minimum $x_{(1)}$,

$$\lim_{n\to\infty} E[x_{(1)}] = \lim_{n\to\infty} \frac{1}{n\theta} = 0$$

and

$$\lim_{n\to\infty} \operatorname{Var}[x_{(1)}] = \lim_{n\to\infty} \frac{1}{(n\theta)^2} = 0.$$

Therefore,

$$\operatorname{plim} x_{(1)} = 0.$$

Note, in particular, that the variance is divided by n^2. This estimator converges very rapidly to 0.

Convergence in probability does not imply convergence in mean square. Consider the simple example given earlier in which x_n equals either zero or n with probabilities $1 - (1/n)$ and $(1/n)$. The exact expected value of x_n is 1 for all n, which is not the probability limit. Indeed, if we let $\text{Prob}(x_n = n^2) = (1/n)$ instead, the mean of the distribution explodes, but the probability limit is still zero. Again, the point $x_n = n^2$ becomes ever more extreme but, at the same time, becomes ever less likely.

The conditions for convergence in mean square are usually easier to verify than those for the more general form. Fortunately, we shall rarely encounter circumstances in which it will be necessary to show convergence in probability in which we cannot rely upon convergence in mean square. Our most frequent use of this concept will be in formulating consistent estimators.

DEFINITION D.2 Consistent Estimator
An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if

$$\operatorname{plim} \hat{\theta}_n = \theta. \tag{D-4}$$

THEOREM D.4 Consistency of the Sample Mean
The mean of a random sample from any population with finite mean μ and finite variance σ^2 is a consistent estimator of μ.
Proof: *$E[\bar{x}_n] = \mu$ and $\text{Var}[\bar{x}_n] = \sigma^2/n$. Therefore, $\bar{x}_n$ converges in mean square to μ, or plim $\bar{x}_n = \mu$.*

Theorem D.4 is broader than it might appear at first.

COROLLARY TO THEOREM D.4 Consistency of a Mean of Functions
In random sampling, for any function $g(x)$, if $E[g(x)]$ and $\text{Var}[g(x)]$ are finite constants, then

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} g(x_i) = E[g(x)]. \tag{D-5}$$

Proof: *Define $y_i = g(x_i)$ and use Theorem D.4.*

Example D.2 Estimating a Function of the Mean

In sampling from a normal distribution with mean μ and variance 1, $E[e^x] = e^{\mu+1/2}$ and $\text{Var}[e^x] = e^{2\mu+2} - e^{2\mu+1}$. (See Section B.4.4 on the lognormal distribution.) Hence,

$$\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} e^{x_i} = e^{\mu+1/2}.$$

D.2.2 OTHER FORMS OF CONVERGENCE AND LAWS OF LARGE NUMBERS

Theorem D.4 and the corollary just given are particularly narrow forms of a set of results known as **laws of large numbers** that are fundamental to the theory of parameter estimation. Laws of large numbers come in two forms depending on the type of convergence considered. The simpler of these are “weak laws of large numbers” which rely on convergence in probability as we defined it above. “Strong laws” rely on a broader type of convergence called **almost sure convergence**. Overall, the law of large numbers is a statement about the behavior of an average of a large number of random variables.

THEOREM D.5 Khinchine's Weak Law of Large Numbers
If x_i, $i = 1, \ldots, n$ is a random (i.i.d.) sample from a distribution with finite mean $E[x_i] = \mu$, then

$$\text{plim}\, \overline{x}_n = \mu.$$

Proofs of this and the theorem below are fairly intricate. Rao (1973) provides one.

Notice that this is already broader than Theorem D.4, as it does not require that the variance of the distribution be finite. On the other hand, it is not broad enough, because most of the situations we encounter where we will need a result such as this will not involve i.i.d. random sampling. A broader result is

THEOREM D.6 Chebychev's Weak Law of Large Numbers
If $x_i, i = 1, \ldots, n$ is a sample of observations such that $E[x_i] = \mu_i < \infty$ and $\text{Var}[x_i] = \sigma_i^2 < \infty$ such that $\overline{\sigma}_n^2/n = (1/n^2)\Sigma_i \sigma_i^2 \rightarrow 0$ as $n \rightarrow \infty$, then $\text{plim}(\overline{x}_n - \overline{\mu}_n) = 0$.

There is a subtle distinction between these two theorems that you should notice. The Chebychev theorem does not state that $\overline{x}_n$ converges to $\overline{\mu}_n$, or even that it converges to a constant at all. That would require a precise statement about the behavior of $\overline{\mu}_n$. The theorem states that as n increases without bound, these two quantities will be arbitrarily close to each other—that is, the difference between them converges to a constant, zero. This is an important notion that enters the derivation when we consider statistics that converge to random variables, instead of to constants. What we do have with these two theorems are extremely broad conditions under which a sample mean will converge in probability to its population counterpart. The more important difference between the Khinchine and Chebychev theorems is that the second allows for heterogeneity in the distributions of the random variables that enter the mean.

In analyzing time-series data, the sequence of outcomes is itself viewed as a random event. Consider, then, the sample mean, $\overline{x}_n$. The preceding results concern the behavior of this statistic as $n \rightarrow \infty$ for a particular realization of the sequence $\overline{x}_1, \ldots, \overline{x}_n$. But, if the sequence, itself, is viewed as a random event, then the limit to which $\overline{x}_n$ converges may be also. The stronger notion of almost sure convergence relates to this possibility.

DEFINITION D.3 Almost Sure Convergence
The random variable x_n converges almost surely to the constant c if and only if

$$\text{Prob}\left(\lim_{n \rightarrow \infty} x_n = c\right) = 1.$$

This is denoted $x_n \xrightarrow{a.s.} c$. It states that the probability of observing a sequence that does not converge to c ultimately vanishes. Intuitively, it states that once the sequence x_n becomes close to c, it stays close to c.

Almost sure convergence is used in a stronger form of the law of large numbers:

THEOREM D.7 Kolmogorov's Strong Law of Large Numbers
If $x_i, i = 1, \ldots, n$ is a sequence of independently distributed random variables such that $E[x_i] = \mu_i < \infty$ and $\text{Var}[x_i] = \sigma_i^2 < \infty$ such that $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$ as $n \rightarrow \infty$ then $\bar{x}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$.

THEOREM D.8 Markov's Strong Law of Large Numbers
If $\{z_i\}$ is a sequence of independent random variables with $E[z_i] = \mu_i < \infty$ and if for some $0 < \delta < 1$, $\sum_{i=1}^{\infty} E[|z_i - \mu_i|^{1+\delta}]/i^{1+\delta} < \infty$, then $\bar{z}_n - \bar{\mu}_n$ converges almost surely to 0, which we denote $\bar{z}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$.[2]

The variance condition is satisfied if every variance in the sequence is finite, but this is not strictly required; it only requires that the variances in the sequence increase at a slow enough rate that the sequence of variances as defined is bounded. The theorem allows for heterogeneity in the means and variances. If we return to the conditions of the Khinchine theorem, i.i.d. sampling, we have a corollary:

COROLLARY TO THEOREM D.8 (Kolmogorov)
If $x_i, i = 1, \ldots, n$ is a sequence of independent and identically distributed random variables such that $E[x_i] = \mu < \infty$ and $E[|x_i|] < \infty$, then $\bar{x}_n - \mu \xrightarrow{a.s.} 0$.

Note that the corollary requires identically distributed observations while the theorem only requires independence. Finally, another form of convergence encountered in the analysis of time-series data is convergence in rth mean:

[2] The use of the expected absolute deviation differs a bit from the expected squared deviation that we have used heretofore to characterize the spread of a distribution. Consider two examples. If $z \sim N[0, \sigma^2]$, then $E[|z|] = \text{Prob}[z < 0]E[-z|z < 0] + \text{Prob}[z \geq 0]E[z|z \geq 0] = 0.7979\sigma$. (See Theorem 18.2.) So, finite expected absolute value is the same as finite second moment for the normal distribution. But if z takes values $[0, n]$ with probabilities $[1 - 1/n, 1/n]$, then the variance of z is $(n - 1)$, but $E[|z - \mu_z|]$ is $2 - 2/n$. For this case, finite expected absolute value occurs without finite expected second moment. These are different characterizations of the spread of the distribution.

DEFINITION D.4 Convergence in rth Mean
If x_n is a sequence of random variables such that $E[|x_n|^r] < \infty$ and $\lim_{n\to\infty} E[|x_n - c|^r] = 0$, then x_n converges in rth mean to c. This is denoted $x_n \xrightarrow{r.m.} c$.

Surely the most common application is the one we met earlier, convergence in means square, which is convergence in the second mean. Some useful results follow from this definition:

THEOREM D.9 Convergence in Lower Powers
If x_n converges in rth mean to c, then x_n converges in sth mean to c for any $s < r$. The proof uses Jensen's Inequality, Theorem D.13. Write $E[|x_n - c|^s] = E[(|x_n - c|^r)^{s/r}] \le E[(|x_n - c|^r)]\}^{s/r}$ and the inner term converges to zero so the full function must also.

THEOREM D.10 Generalized Chebychev's Inequality
If x_n is a random variable and c is a constant such that with $E[|x_n - c|^r] < \infty$ and ε is a positive constant, then $\text{Prob}(|x_n - c| > \varepsilon) \le E[|x_n - c|^r]/\varepsilon^r$.

We have considered two cases of this result already, when $r = 1$ which is the Markov inequality, Theorem D.3, and when $r = 2$, which is the Chebychev inequality we looked at first in Theorem D.2.

THEOREM D.11 Convergence in rth mean and Convergence in Probability
If $x_n \xrightarrow{r.m.} c$, for some $r > 0$, then $x_n \xrightarrow{p} c$. The proof relies on Theorem D.10. By assumption, $\lim_{n\to\infty} E[|x_n - c|^r] = 0$ so for some n sufficiently large, $E[|x_n - c|^r] < \infty$. By Theorem D.10, then, $\text{Prob}(|x_n - c| > \varepsilon) \le E[|x_n - c|^r]/\varepsilon^r$ for any $\varepsilon > 0$. The denominator of the fraction is a fixed constant and the numerator converges to zero by our initial assumption, so $\lim_{n\to\infty}\text{Prob}(|x_n - c| > \varepsilon) = 0$, which completes the proof.

One implication of Theorem D.11 is that although convergence in mean square is a convenient way to prove convergence in probability, it is actually stronger than necessary, as we get the same result for any positive *r*.

Finally, we note that we have now shown that both almost sure convergence and convergence in *r*th mean are stronger than convergence in probability; each implies the latter. But they, themselves, are different notions of convergence, and neither implies the other.

DEFINITION D.5 Convergence of a Random Vector or Matrix
Let $\mathbf{x}_n$ *denote a random vector and* $\mathbf{X}_n$ *a random matrix, and* **c** *and* **C** *denote a vector and matrix of constants with the same dimensions as* $\mathbf{x}_n$ *and* $\mathbf{X}_n$*, respectively. All of the preceding notions of convergence can be extended to* $(\mathbf{x}_n, \mathbf{c})$ *and* $(\mathbf{X}_n, \mathbf{C})$ *by applying the results to the respective corresponding elements.*

D.2.3 CONVERGENCE OF FUNCTIONS

A particularly convenient result is the following.

THEOREM D.12 Slutsky Theorem
For a continuous function $g(x_n)$ *that is not a function of* n*,*

$$\text{plim } g(x_n) = g(\text{plim } x_n). \qquad \textbf{(D-6)}$$

The generalization of Theorem D.12 to a function of several random variables is direct, as illustrated in the next example.

Example D.3 Probability Limit of a Function of $\bar{x}$ *and* s^2

In random sampling from a population with mean μ and variance σ^2, the exact expected value of $\bar{x}_n^2/s_n^2$ will be difficult, if not impossible, to derive. But, by the Slutsky theorem,

$$\text{plim } \frac{\bar{x}_n^2}{s_n^2} = \frac{\mu^2}{\sigma^2}.$$

An application that highlights the difference between expectation and probability limit is suggested by the following useful relationships.

THEOREM D.13 Inequalities for Expectations
Jensen's Inequality. *If* $g(x_n)$ *is a concave function of* x_n*, then* $g(E[x_n]) \geq E[g(x_n)]$. **Cauchy–Schwarz Inequality**. *For two random variables,* $E[|xy|] \leq \{E[x^2]\}^{1/2}\{E[y^2]\}^{1/2}$.

Although the expected value of a function of x_n may not equal the function of the expected value—it exceeds it if the function is concave—the probability limit of the function *is* equal to the function of the probability limit.

The Slutsky theorem highlights a comparison between the expectation of a random variable and its probability limit. Theorem D.12 extends directly in two important directions. First, though stated in terms of convergence in probability, the same set of results applies to convergence in *r*th mean and almost sure convergence. Second, so long as the functions are continuous, the Slutsky theorem can be extended to vector or matrix valued functions of random scalars, vectors, or matrices. The following describe some specific applications. Some implications of the Slutsky theorem are now summarized.

THEOREM D.14 Rules for Probability Limits
If x_n and y_n are random variables with plim $x_n = c$ *and* plim $y_n = d$, *then*

$$\text{plim}(x_n + y_n) = c + d, \quad \textbf{(sum rule)} \tag{D-7}$$

$$\text{plim}\, x_n y_n = cd, \quad \textbf{(product rule)} \tag{D-8}$$

$$\text{plim}\, x_n/y_n = c/d \quad \text{if} \quad d \neq 0. \quad \textbf{(ratio rule)} \tag{D-9}$$

If $\mathbf{W}_n$ is a matrix whose elements are random variables and if plim $\mathbf{W}_n = \boldsymbol{\Omega}$, *then*

$$\text{plim}\, \mathbf{W}_n^{-1} = \boldsymbol{\Omega}^{-1}. \quad \textbf{(matrix inverse rule)} \tag{D-10}$$

If $\mathbf{X}_n$ and $\mathbf{Y}_n$ are random matrices with plim $\mathbf{X}_n = \mathbf{A}$ *and* plim $\mathbf{Y}_n = \mathbf{B}$, *then*

$$\text{plim}\, \mathbf{X}_n\mathbf{Y}_n = \mathbf{AB}. \quad \textbf{(matrix product rule)} \tag{D-11}$$

D.2.4 CONVERGENCE TO A RANDOM VARIABLE

The preceding has dealt with conditions under which a random variable converges to a constant, for example, the way that a sample mean converges to the population mean. To develop a theory for the behavior of estimators, as a prelude to the discussion of limiting distributions, we now consider cases in which a random variable converges not to a constant, but to another random variable. These results will actually subsume those in the preceding section, as a constant may always be viewed as a degenerate random variable, that is one with zero variance.

DEFINITION D.6 Convergence in Probability to a Random Variable
The random variable x_n converges in probability to the random variable x if $\lim_{n\to\infty} \text{Prob}(|x_n - x| > \varepsilon) = 0$ *for any positive* ε.

As before, we write plim $x_n = x$ to denote this case. The interpretation (at least the intuition) of this type of convergence is different when x is a random variable. The notion of closeness defined here relates not to the concentration of the mass of the probability mechanism generating x_n at a point c, but to the closeness of that probability mechanism to that of x. One can think of this as a convergence of the CDF of x_n to that of x.

DEFINITION D.7 Almost Sure Convergence to a Random Variable
The random variable x_n converges almost surely to the random variable x if and only if $\lim_{n\to\infty}\text{Prob}(|x_i - x| > \varepsilon \text{ for all } i \geq n) = 0$ *for all* $\varepsilon > 0$.

DEFINITION D.8 Convergence in rth Mean to a Random Variable
The random variable x_n converges in rth mean to the random variable x if and only if $\lim_{n\to\infty}E[|x_n - x|^r] = 0$. *This is labeled* $x_n \xrightarrow{r.m.} x$. *As before, the case $r = 2$ is labeled convergence in mean square.*

Once again, we have to revise our understanding of convergence when convergence is to a random variable.

THEOREM D.15 Convergence of Moments
Suppose $x_n \xrightarrow{r.m.} x$ and $E[|x|^r]$ is finite. Then, $\lim_{n\to\infty}E[|x_n|^r] = E[|x|^r]$.

Theorem D.15 raises an interesting question. Suppose we let r grow, and suppose that $x_n \xrightarrow{r.m.} x$ and, in addition, all moments are finite. If this holds for any r, do we conclude that these random variables have the same distribution? The answer to this longstanding problem in probability theory—the problem of the sequence of moments—is no. The sequence of moments does not uniquely determine the distribution. Although convergence in rth mean and almost surely still both imply convergence in probability, it remains true, even with convergence to a random variable instead of a constant, that these are different forms of convergence.

D.2.5 CONVERGENCE IN DISTRIBUTION: LIMITING DISTRIBUTIONS

A second form of convergence is **convergence in distribution**. Let x_n be a sequence of random variables indexed by the sample size, and assume that x_n has cdf $F_n(x_n)$.

DEFINITION D.9 Convergence in Distribution
x_n converges in distribution to a random variable x with CDF $F(x)$ if $\lim_{n\to\infty}|F_n(x_n) - F(x)| = 0$ *at all continuity points of* $F(x)$.

This statement is about the probability distribution associated with x_n; it does not imply that x_n converges at all. To take a trivial example, suppose that the exact distribution of the random variable x_n is

$$\text{Prob}(x_n = 1) = \frac{1}{2} + \frac{1}{n+1}, \quad \text{Prob}(x_n = 2) = \frac{1}{2} - \frac{1}{n+1}.$$

As n increases without bound, the two probabilities converge to $\frac{1}{2}$, but x_n does not converge to a constant.

DEFINITION D.10 Limiting Distribution
If x_n converges in distribution to x, where $F_n(x_n)$ is the CDF of x_n, then $F(x)$ is the **limiting distribution** *of x_n. This is written $x_n \xrightarrow{d} x$.*

The limiting distribution is often given in terms of the pdf, or simply the parametric family. For example, "the limiting distribution of x_n is standard normal."

Convergence in distribution can be extended to random vectors and matrices, although not in the element by element manner that we extended the earlier convergence forms. The reason is that convergence in distribution is a property of the CDF of the random variable, not the variable itself. Thus, we can obtain a convergence result analogous to that in Definition D.9 for vectors or matrices by applying definition to the joint CDF for the elements of the vector or matrices. Thus, $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ if $\lim_{n\to\infty}|F_n(\mathbf{x}_n) - F(\mathbf{x})| = 0$ and likewise for a random matrix.

Example D.4 Limiting Distribution of t_{n-1}

Consider a sample of size n from a standard normal distribution. A familiar inference problem is the test of the hypothesis that the population mean is zero. The test statistic usually used is the t statistic:

$$t_{n-1} = \frac{\overline{x}_n}{s_n/\sqrt{n}},$$

where

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{n-1}.$$

The exact distribution of the random variable t_{n-1} is t with $n - 1$ degrees of freedom. The density is different for every n:

$$f(t_{n-1}) = \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}[(n-1)\pi]^{-1/2}\left[1 + \frac{t_{n-1}^2}{n-1}\right]^{-n/2}, \qquad \textbf{(D-12)}$$

as is the CDF, $F_{n-1}(t) = \int_{-\infty}^{t} f_{n-1}(x)\,dx$. This distribution has mean zero and variance $(n-1)/(n-3)$. As n grows to infinity, t_{n-1} converges to the standard normal, which is written

$$t_{n-1} \xrightarrow{d} N[0, 1].$$

DEFINITION D.11 Limiting Mean and Variance
The **limiting mean** *and* **variance** *of a random variable are the mean and variance of the limiting distribution, assuming that the limiting distribution and its moments exist.*

For the random variable with $t[n]$ distribution, the exact mean and variance are zero and $n/(n - 2)$, whereas the limiting mean and variance are zero and one. The example might suggest that the limiting mean and variance are zero and one; that is, that the moments of the limiting distribution are the ordinary limits of the moments of the finite sample distributions. This situation is almost always true, but it need not be. It is possible to construct examples in which the exact moments do not even exist, even though the moments of the limiting distribution are well defined.[3] Even in such cases, we can usually derive the mean and variance of the limiting distribution.

Limiting distributions, like probability limits, can greatly simplify the analysis of a problem. Some results that combine the two concepts are as follows.[4]

THEOREM D.16 Rules for Limiting Distributions

1. *If* $x_n \xrightarrow{d} x$ *and* plim $y_n = c$, *then*

$$x_n y_n \xrightarrow{d} cx, \qquad \textbf{(D-13)}$$

which means that the limiting distribution of $x_n y_n$ *is the distribution of* cx. *Also,*

$$x_n + y_n \xrightarrow{d} x + c, \qquad \textbf{(D-14)}$$

$$x_n/y_n \xrightarrow{d} x/c, \quad \text{if } c \neq 0. \qquad \textbf{(D-15)}$$

2. *If* $x_n \xrightarrow{d} x$ *and* $g(x_n)$ *is a continuous function, then*

$$g(x_n) \xrightarrow{d} g(x). \qquad \textbf{(D-16)}$$

This result is analogous to the Slutsky theorem for probability limits. For an example, consider the t_n *random variable discussed earlier. The exact distribution of* t_n^2 *is* $F[1, n]$. *But as* $n \longrightarrow \infty$, t_n *converges to a standard normal variable. According to this result, the limiting distribution of* t_n^2 *will be that of the square of a standard normal, which is chi-squared with one degree of freedom. We conclude, therefore, that*

$$F[1, n] \xrightarrow{d} \textit{chi-squared}[1]. \qquad \textbf{(D-17)}$$

We encountered this result in our earlier discussion of limiting forms of the standard normal family of distributions.

3. *If* y_n *has a limiting distribution and* plim $(x_n - y_n) = 0$, *then* x_n *has the same limiting distribution as* y_n.

[3] See, for example, Maddala (1977a, p. 150).

[4] For proofs and further discussion, see, for example, Greenberg and Webster (1983).

The third result in Theorem D.16 combines convergence in distribution and in probability. The second result can be extended to vectors and matrices.

Example D.5 The F Distribution

Suppose that $\boldsymbol{t}_{1,n}$ and $\boldsymbol{t}_{2,n}$ are a $K\times1$ and an $M\times1$ random vector of variables whose components are independent with each distributed as t with n degrees of freedom. Then, as we saw in the preceding, for any component in either random vector, the limiting distribution is standard normal, so for the entire vector, $\boldsymbol{t}_{j,n} \xrightarrow{d} \boldsymbol{z}_j$, a vector of independent standard normally distributed variables. The results so far show that $\dfrac{(\boldsymbol{t}'_{1,n}\,\boldsymbol{t}_{1,n})/K}{(\boldsymbol{t}'_{2,n}\,t_{2,n})/M} \xrightarrow{d} F[K, M]$.

Finally, a specific case of result 2 in Theorem D.16 produces a tool known as the Cramér–Wold device.

THEOREM D.17 Cramer–Wold Device
If $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, *then* $\mathbf{c}'\mathbf{x}_n \xrightarrow{d} \mathbf{c}'\mathbf{x}$ *for all conformable vectors* $\mathbf{c}$ *with real valued elements.*

By allowing **c** to be a vector with just a one in a particular position and zeros elsewhere, we see that convergence in distribution of a random vector $\mathbf{x}_n$ to $\mathbf{x}$ does imply that each component does likewise.

D.2.6 CENTRAL LIMIT THEOREMS

We are ultimately interested in finding a way to describe the statistical properties of estimators when their exact distributions are unknown. The concepts of consistency and convergence in probability are important. But the theory of limiting distributions given earlier is not yet adequate. We rarely deal with estimators that are not consistent for something, though perhaps not always the parameter we are trying to estimate. As such,

$$\text{if plim } \hat{\theta}_n = \theta, \quad \text{then } \hat{\theta}_n \xrightarrow{d} \theta.$$

That is, the limiting distribution of $\hat{\theta}_n$ is a spike. This is not very informative, nor is it at all what we have in mind when we speak of the statistical properties of an estimator. (To endow our finite sample estimator $\hat{\theta}_n$ with the zero sampling variance of the spike at θ would be optimistic in the extreme.)

As an intermediate step, then, to a more reasonable description of the statistical properties of an estimator, we use a **stabilizing transformation** of the random variable to one that does have a well-defined limiting distribution. To jump to the most common application, whereas

$$\text{plim } \hat{\theta}_n = \theta,$$

we often find that

$$z_n = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} f(z),$$

where $f(z)$ is a well-defined distribution with a mean and a positive variance. An estimator which has this property is said to be **root-*n* consistent**. The single most important theorem in econometrics provides an application of this proposition. A basic form of the theorem is as follows.

THEOREM D.18 Lindeberg–Levy Central Limit Theorem (Univariate)
If $x_1, \ldots, x_n$ are a random sample from a probability distribution with finite mean μ and finite variance σ^2 and $\overline{x}_n = (1/n)\sum_{i=1}^{n} x_i$, then $\sqrt{n}\,(\overline{x}_n - \mu) \xrightarrow{d} N[0, \sigma^2]$. A proof appears in Rao (1973, p. 127*).*

The result is quite remarkable as it holds regardless of the form of the parent distribution. For a striking example, return to Figure C.3. The distribution from which the data were drawn in that figure does not even remotely resemble a normal distribution. In samples of only four observations the force of the central limit theorem is clearly visible in the sampling distribution of the means. The sampling experiment Example D.6 shows the effect in a systematic demonstration of the result.

The Lindeberg–Levy theorem is one of several forms of this extremely powerful result. For our purposes, an important extension allows us to relax the assumption of equal variances. The Lindeberg–Feller form of the central limit theorem is the centerpiece of most of our analysis in econometrics.

THEOREM D.19 Lindeberg–Feller Central Limit Theorem (with Unequal Variances)
Suppose that $\{x_i\}, i = 1, \ldots, n$, is a sequence of independent random variables with finite means μ_i and finite positive variances σ_i^2. Let

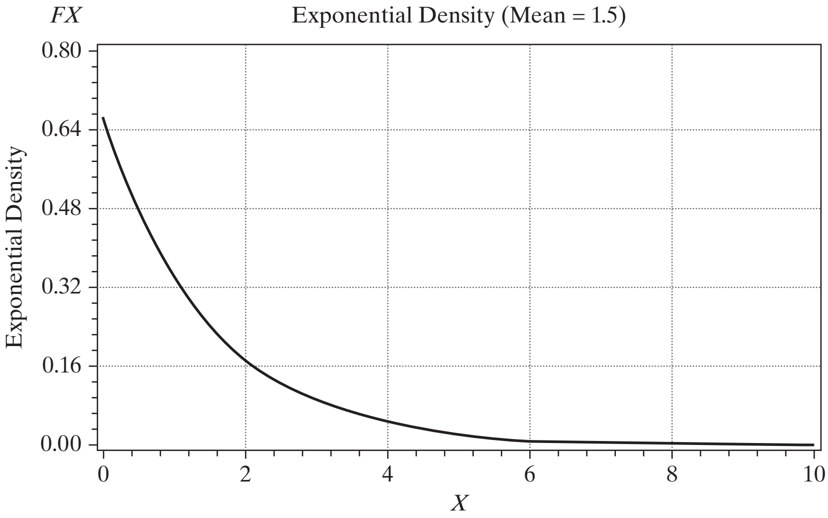
$$\overline{\mu}_n = \frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n), \quad \text{and} \quad \overline{\sigma}_n^2 = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 + \cdots, \sigma_n^2).$$

If no single term dominates this average variance, which we could state as $\lim_{n\to\infty}\max(\sigma_i)/(\sqrt{n}\overline{\sigma}_n) = 0$, *and if the average variance converges to a finite constant, $\overline{\sigma}^2 = \lim_{n\to\infty}\overline{\sigma}_n^2$, then $\sqrt{n}\,(\overline{x}_n - \overline{\mu}_n) \xrightarrow{d} N[0, \overline{\sigma}^2]$.*

In practical terms, the theorem states that sums of random variables, regardless of their form, will tend to be normally distributed. The result is yet more remarkable in that *it does not require the variables in the sum to come from the same underlying distribution. It requires, essentially, only that the mean be a mixture of many random variables, none of which is large compared with their sum.* Because nearly all the estimators we construct in econometrics fall under the purview of the central limit theorem, it is obviously an important result.

Proof of the Lindeberg–Feller theorem requires some quite intricate mathematics [see, e.g., Loeve (1977)] that are well beyond the scope of our work here. We do note an important consideration in this theorem. The result rests on a condition known as the *Lindeberg condition*. The sample mean computed in the theorem is a mixture of random

FIGURE D.2 The Exponential Distribution.



variables from possibly different distributions. The Lindeberg condition, in words, states that the contribution of the tail areas of these underlying distributions to the variance of the sum must be negligible in the limit. The condition formalizes the assumption in Theorem D.19 that the average variance be positive and not be dominated by any single term. [For an intuitively crafted mathematical discussion of this condition, see White (2001, pp. 117–118).] The condition is essentially impossible to verify in practice, so it is useful to have a simpler version of the theorem that encompasses it.

Example D.6 The Lindeberg–Levy Central Limit Theorem

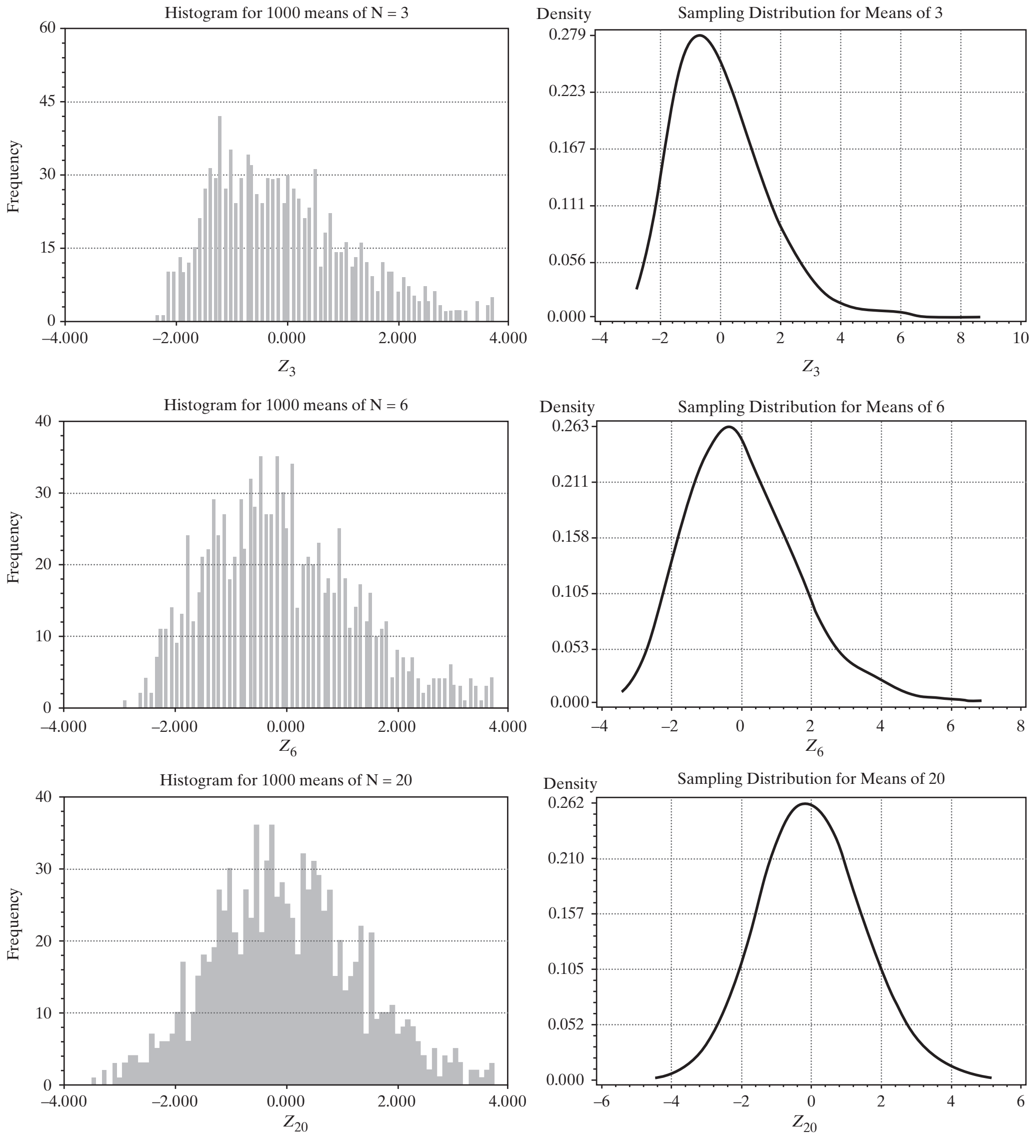
We'll use a sampling experiment to demonstrate the operation of the central limit theorem. Consider random sampling from the exponential distribution with mean 1.5—this is the setting used in Example C.4. The density is shown in Figure D.2.

We've drawn 1,000 samples of 3, 6, and 20 observations from this population and computed the sample means for each. For each mean, we then computed $z_{in} = \sqrt{n}(\bar{x}_{in} - \mu)$, where $i = 1, \ldots, 1{,}000$ and n is 3, 6, or 20. The three rows of figures in Figure D.3 show histograms of the observed samples of sample means and kernel density estimates of the underlying distributions for the three samples of transformed means. The force of the central limit is clearly visible in the shapes of the distributions.

THEOREM D.20 Liapounov Central Limit Theorem

Suppose that $\{x_i\}$ is a sequence of independent random variables with finite means μ_i and finite positive variances σ_i^2 such that $E[|x_i - \mu_i|^{2+\delta}]$ is finite for some $\delta > 0$. If $\bar{\sigma}_n$ is positive and finite for all n sufficiently large, then $\sqrt{n}(\bar{x}_n - \bar{\mu}_n)/\bar{\sigma}_n \xrightarrow{d} N[0, 1]$.

FIGURE D.3 The Central Limit Theorem.



Histogram for 1000 means of N = 3

Sampling Distribution for Means of 3

Histogram for 1000 means of N = 6

Sampling Distribution for Means of 6

Histogram for 1000 means of N = 20

Sampling Distribution for Means of 20

This version of the central limit theorem requires only that moments slightly larger than two be finite.

Note the distinction between the laws of large numbers in Theorems D.5 and D.6 and the central limit theorems. Neither asserts that sample means tend to normality. Sample means (i.e., the distributions of them) converge to spikes at the true mean. It is the transformation of the mean, $\sqrt{n}(\bar{x}_n - \mu)/\sigma$, that converges to standard normality. To see this at work, if you have access to the necessary software, you might try reproducing Example D.6 using the raw means, $\bar{x}_{in}$. What do you expect to observe?

For later purposes, we will require multivariate versions of these theorems. Proofs of the following may be found, for example, in Greenberg and Webster (1983) or Rao (1973) and references cited there.

THEOREM D.18A Multivariate Lindeberg–Levy Central Limit Theorem
If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a random sample from a multivariate distribution with finite mean vector $\boldsymbol{\mu}$ and finite positive definite covariance matrix $\mathbf{Q}$, then

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}],$$

where

$$\bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

To get from D.18 to D.18A (and D.19 to D.19A) we need to add a step. Theorem D.18 applies to the individual elements of the vector. A vector has a multivariate normal distribution if the individual elements are normally distributed and if every linear combination is normally distributed. We can use Theorem D.18 (D.19) for the individual terms and Theorem D.17 to establish that linear combinations behave likewise. This establishes the extensions.

The extension of the Lindeberg–Feller theorem to unequal covariance matrices requires some intricate mathematics. The following is an informal statement of the relevant conditions. Further discussion and references appear in Fomby, Hill, and Johnson (1984) and Greenberg and Webster (1983).

THEOREM D.19A Multivariate Lindeberg–Feller Central Limit Theorem
Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a sample of random vectors such that $E[\mathbf{x}_i] = \boldsymbol{\mu}_i$, $\text{Var}[\mathbf{x}_i] = \mathbf{Q}_i$, and all mixed third moments of the multivariate distribution are finite. Let

$$\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i \text{ and } \overline{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i.$$

We assume that

$$\lim_{n\to\infty}\overline{\mathbf{Q}}_n = \mathbf{Q},$$

where $\mathbf{Q}$ is a finite, positive definite matrix, and that for every i,

$$\lim_{n\to\infty}(n\overline{\mathbf{Q}}_n)^{-1}\mathbf{Q}_i = \lim_{n\to\infty}\left(\sum_{i=1}^{n}\mathbf{Q}_i\right)^{-1}\mathbf{Q}_i = \mathbf{0}.$$

We allow the means of the random vectors to differ, although in the cases that we will analyze, they will generally be identical. The second assumption states that individual components of the sum must be finite and diminish in significance. There is also an implicit assumption that the sum of matrices is nonsingular. Because the limiting matrix is nonsingular, the assumption must hold for large enough n, which is all that concerns us here. With these in place, the result is

$$\sqrt{n}(\bar{\mathbf{x}}_n - \bar{\mu}_n) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}].$$

D.2.7 THE DELTA METHOD

At several points in Appendix C, we used a linear Taylor series approximation to analyze the distribution and moments of a random variable. We are now able to justify this usage. We complete the development of Theorem D.12 (probability limit of a function of a random variable), Theorem D.16 (2) (limiting distribution of a function of a random variable), and the central limit theorems, with a useful result that is known as the **delta method**. For a single random variable (sample mean or otherwise), we have the following theorem.

THEOREM D.21 Limiting Normal Distribution of a Function
If $\sqrt{n}(z_n - \mu) \xrightarrow{d} N[0, \sigma^2]$ and if $g(z_n)$ is a continuous and continuously differentiable function with $g'(\mu)$ not equal to zero and not involving n, then

$$\sqrt{n}[g(z_n) - g(\mu)] \xrightarrow{d} N[0, \{g'(\mu)\}^2 \sigma^2]. \qquad \textbf{(D-18)}$$

Notice that the mean and variance of the limiting distribution are the mean and variance of the linear Taylor series approximation:

$$g(z_n) \simeq g(\mu) + g'(\mu)(z_n - \mu).$$

The multivariate version of this theorem will be used at many points in the text.

THEOREM D.21A Limiting Normal Distribution of a Set of Functions
If $\mathbf{z}_n$ is a $K \times 1$ sequence of vector-valued random variables such that $\sqrt{n}(\mathbf{z}_n - \boldsymbol{\mu}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$ and if $\mathbf{c}(\mathbf{z}_n)$ is a set of J continuous and continuously differentiable functions of $\mathbf{z}_n$ with $\mathbf{C}(\boldsymbol{\mu})$ not equal to zero, not involving n, then

$$\sqrt{n}[\mathbf{c}(\mathbf{z}_n) - \mathbf{c}(\boldsymbol{\mu})] \xrightarrow{d} N[\mathbf{0}, \mathbf{C}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{C}(\boldsymbol{\mu})'], \qquad \textbf{(D-19)}$$

where $\mathbf{C}(\boldsymbol{\mu})$ is the $J \times K$ matrix $\partial\mathbf{c}(\boldsymbol{\mu})/\partial\boldsymbol{\mu}'$. The jth row of $\mathbf{C}(\boldsymbol{\mu})$ is the vector of partial derivatives of the jth function with respect to $\boldsymbol{\mu}'$.

D.3 ASYMPTOTIC DISTRIBUTIONS

The theory of limiting distributions is only a means to an end. We are interested in the behavior of the estimators themselves. The limiting distributions obtained through the central limit theorem all involve unknown parameters, generally the ones we are trying to estimate. Moreover, our samples are always finite. Thus, we depart from the limiting distributions to derive the asymptotic distributions of the estimators.

DEFINITION D.12 Asymptotic Distribution
An asymptotic distribution is a distribution that is used to approximate the true finite sample distribution of a random variable.[5]

By far the most common means of formulating an asymptotic distribution (at least by econometricians) is to construct it from the known limiting distribution of a function of the random variable. If

$$\sqrt{n}[(\bar{x}_n - \mu)/\sigma] \xrightarrow{d} N[0, 1],$$

then approximately, or asymptotically, $\bar{x}_n \sim N[\mu, \sigma^2/n]$, which we write as

$$\bar{x}_n \overset{a}{\sim} N[\mu, \sigma^2/n].$$

The statement "$\bar{x}_n$ is asymptotically normally distributed with mean μ and variance σ^2/n" says only that this normal distribution provides an approximation to the true distribution, not that the true distribution is exactly normal.

Example D.7 Asymptotic Distribution of the Mean of an Exponential Sample

In sampling from an exponential distribution with parameter θ, the *exact* distribution of $\bar{x}_n$ is that of $\theta/(2n)$ times a chi-squared variable with $2n$ degrees of freedom. The *asymptotic* distribution is $N[\theta, \theta^2/n]$. The exact and asymptotic distributions are shown in Figure D.4 for the case of $\theta = 1$ and $n = 16$.

Extending the definition, suppose that $\hat{\boldsymbol{\theta}}_n$ is an estimator of the parameter vector $\boldsymbol{\theta}$. The asymptotic distribution of the vector $\hat{\boldsymbol{\theta}}_n$ is obtained from the limiting distribution:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \mathbf{V}] \tag{D-20}$$

implies that

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\left[\boldsymbol{\theta}, \frac{1}{n}\mathbf{V}\right]. \tag{D-21}$$

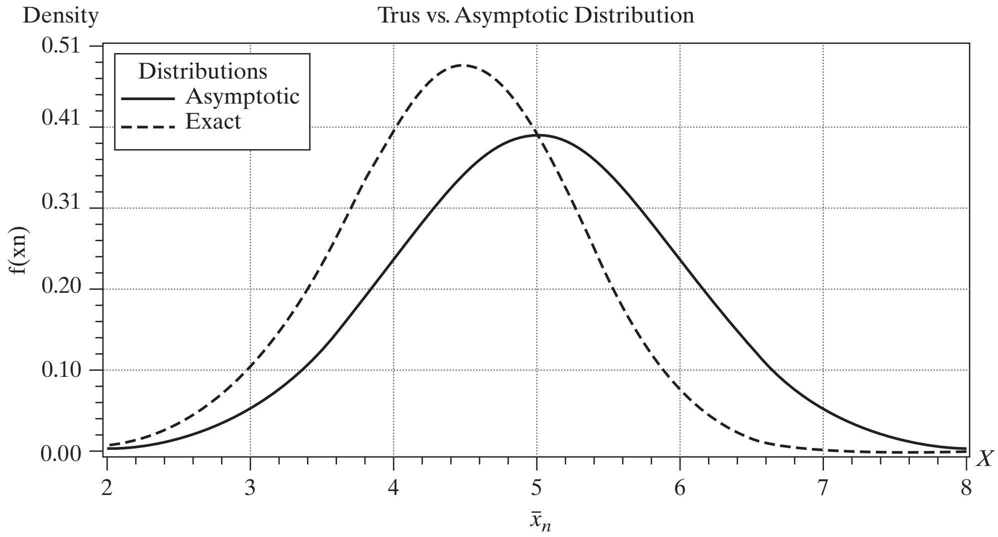
This notation is read "$\hat{\boldsymbol{\theta}}_n$ is asymptotically normally distributed, with mean vector $\boldsymbol{\theta}$ and covariance matrix $(1/n)\mathbf{V}$." The covariance matrix of the asymptotic distribution is the **asymptotic covariance matrix** and is denoted

$$\text{Asy. Var}[\hat{\boldsymbol{\theta}}_n] = \frac{1}{n}\mathbf{V}.$$

Note, once again, the logic used to reach the result; (D-20) holds exactly as $n \to \infty$. We assume that it holds approximately for finite n, which leads to (D-21).

[5] We differ a bit from some other treatments—for example, White (2001), Hayashi (2000, p. 90) —at this point, because they make no distinction between an asymptotic distribution and the limiting distribution, although the treatments are largely along the lines discussed here. In the interest of maintaining consistency of the discussion, we prefer to retain the sharp distinction and derive the asymptotic distribution of an estimator, **t** by first obtaining the *limiting* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$. By our construction, the *limiting* distribution of **t** is degenerate, whereas the *asymptotic* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$ is not useful.

FIGURE D.4 True Versus Asymptotic Distribution.



> **DEFINITION D.13 Asymptotic Normality and Asymptotic Efficiency**
> *An estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically normal if (D-20) holds. The estimator is asymptotically efficient if the covariance matrix of any other consistent, asymptotically normally distributed estimator exceeds $(1/n)\mathbf{V}$ by a nonnegative definite matrix.*

For most estimation problems, these are the criteria used to choose an estimator.

Example D.8 Asymptotic Inefficiency of the Median in Normal Sampling

In sampling from a normal distribution with mean μ and variance σ^2, both the mean $\bar{x}_n$ and the median M_n of the sample are consistent estimators of μ. The limiting distributions of both estimators are spikes at μ, so they can only be compared on the basis of their asymptotic properties. The necessary results are

$$\bar{x}_n \overset{a}{\sim} N[\mu, \sigma^2/n], \quad \text{and} \quad M_n \overset{a}{\sim} N[\mu, (\pi/2)\sigma^2/n]. \tag{D-22}$$

Therefore, the mean is more efficient by a factor of $\pi/2$. (But, see Example 15.7 for a finite sample result.)

D.3.1 ASYMPTOTIC DISTRIBUTION OF A NONLINEAR FUNCTION

Theorems D.12 and D.14 for functions of a random variable have counterparts in asymptotic distributions.

THEOREM D.22 Asymptotic Distribution of a Nonlinear Function
If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, \sigma^2]$ and if $g(\theta)$ is a continuous and continuously differentiable function with $g'(\theta)$ not equal to zero and not involving n, then $g(\hat{\theta}_n) \overset{a}{\sim} N[g(\theta), (1/n)\{g'(\theta)\}^2 \sigma^2]$. If $\hat{\boldsymbol{\theta}}_n$ is a vector of parameter estimators such that $\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N[\boldsymbol{\theta}, (1/n)\mathbf{V}]$ and if $\mathbf{c}(\boldsymbol{\theta})$ is a set of J continuous functions not involving n, then $\mathbf{c}(\hat{\boldsymbol{\theta}}_n) \overset{a}{\sim} N[\mathbf{c}(\boldsymbol{\theta}), (1/n)\mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})']$, where $\mathbf{C}(\boldsymbol{\theta}) = \partial\mathbf{c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$.

Example D.9 Asymptotic Distribution of a Function of Two Estimators

Suppose that b_n and t_n are estimators of parameters β and θ such that

$$\begin{bmatrix} b_n \\ t_n \end{bmatrix} \overset{a}{\sim} N\left[\begin{pmatrix} \beta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_{\beta\beta} & \sigma_{\beta\theta} \\ \sigma_{\theta\beta} & \sigma_{\theta\theta} \end{pmatrix}\right].$$

Find the asymptotic distribution of $c_n = b_n/(1 - t_n)$. Let $\gamma = \beta/(1 - \theta)$. By the Slutsky theorem, c_n is consistent for γ. We shall require

$$\frac{\partial\gamma}{\partial\beta} = \frac{1}{1 - \theta} = \gamma_\beta, \quad \frac{\partial\gamma}{\partial\theta} = \frac{\beta}{(1 - \theta)^2} = \gamma_\theta.$$

Let $\boldsymbol{\Sigma}$ be the 2×2 asymptotic covariance matrix given previously. Then the asymptotic variance of c_n is

$$\text{Asy. Var}[c_n] = (\gamma_\beta \ \gamma_\theta)\boldsymbol{\Sigma}\begin{pmatrix} \gamma_\beta \\ \gamma_\theta \end{pmatrix} = \gamma_\beta^2\sigma_{\beta\beta} + \gamma_\theta^2\sigma_{\theta\theta} + 2\gamma_\beta\gamma_\theta\sigma_{\beta\theta},$$

which is the variance of the linear Taylor series approximation:

$$\hat{\gamma}_n \simeq \gamma + \gamma_\beta(b_n - \beta) + \gamma_\theta(t_n - \theta).$$

D.3.2 ASYMPTOTIC EXPECTATIONS

The asymptotic mean and variance of a random variable are usually the mean and variance of the asymptotic distribution. Thus, for an estimator with the limiting distribution defined in

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \mathbf{V}],$$

the asymptotic expectation is $\boldsymbol{\theta}$ and the asymptotic variance is $(1/n)\,\mathbf{V}$. This statement implies, among other things, that the estimator is "asymptotically unbiased."

At the risk of clouding the issue a bit, it is necessary to reconsider one aspect of the previous description. We have deliberately avoided the use of consistency even though, in most instances, that is what we have in mind. The description thus far might suggest that consistency and asymptotic unbiasedness are the same. Unfortunately (because it is a source of some confusion), they are not. They are if the estimator is consistent and asymptotically normally distributed, or CAN. They may differ in other settings, however. There are at least three possible definitions of asymptotic unbiasedness:

1. The mean of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is 0.

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta. \tag{D-23}$$

2. plim $\theta_n = \theta$.

In most cases encountered in practice, the estimator in hand will have all three properties, so there is no ambiguity. It is not difficult to construct cases in which the left-hand sides of all three definitions are different, however.[6] There is no general agreement among authors as to the precise meaning of asymptotic unbiasedness, perhaps because the term is misleading at the outset; *asymptotic* refers to an approximation, whereas *unbiasedness* is an exact result.[7] Nonetheless, the majority view seems to be that (2) is the proper definition of asymptotic unbiasedness.[8] Note, though, that this definition relies on quantities that are generally unknown and that may not exist.

A similar problem arises in the definition of the asymptotic variance of an estimator. One common definition is[9]

$$\text{Asy. Var}[\hat{\theta}_n] = \frac{1}{n}\lim_{n\to\infty} E\,[\{\sqrt{n}(\hat{\theta}_n - \lim_{n\to\infty} E[\hat{\theta}_n])\}^2]. \tag{D-24}$$

This result is a **leading term approximation**, and it will be sufficient for nearly all applications. Note, however, that like definition 2 of asymptotic unbiasedness, it relies on unknown and possibly nonexistent quantities.

Example D.10 Asymptotic Moments of the Normal Sample Variance

The exact expected value and variance of the variance estimator in a normal sample

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{D-25}$$

are

$$E[m_2] = \frac{(n-1)\sigma^2}{n}, \tag{D-26}$$

and

$$\text{Var}[m_2] = \frac{\mu_4 - \sigma^4}{n} - \frac{2(\mu_4 - 2\sigma^4)}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}, \tag{D-27}$$

where $\mu_4 = E[(x - \mu)^4]$. [See Goldberger (1964, pp. 97–99).] The leading term approximation would be

$$\text{Asy. Var}[m_2] = \frac{1}{n}(\mu_4 - \sigma^4).$$

[6] See, for example, Maddala (1977a, p. 150).

[7] See, for example, Theil (1971, p. 377).

[8] Many studies of estimators analyze the "asymptotic bias" of, say, $\hat{\theta}_n$ as an estimator of a parameter θ. In most cases, the quantity of interest is actually plim $[\hat{\theta}_n - \theta]$. See, for example, Greene (1980b) and another example in Johnston (1984, p. 312).

[9] Kmenta (1986, p.165).

D.4 SEQUENCES AND THE ORDER OF A SEQUENCE

This section has been concerned with sequences of constants, denoted, for example, c_n, and random variables, such as x_n, that are indexed by a sample size, n. An important characteristic of a sequence is the rate at which it converges (or diverges). For example, as we have seen, the mean of a random sample of n observations from a distribution with finite mean, μ, and finite variance, σ^2, is itself a random variable with variance $\gamma_n^2 = \sigma^2/n$. We see that as long as σ^2 is a finite constant, γ_n^2 is a sequence of constants that converges to zero. Another example is the random variable $x_{(1),n}$, the minimum value in a random sample of n observations from the exponential distribution with mean $1/\theta$ defined in Example C.4. It turns out that $x_{(1),n}$ has variance $1/(n\theta)^2$. Clearly, this variance also converges to zero, but, intuition suggests, faster than σ^2/n does. On the other hand, the sum of the integers from one to n, $S_n = n(n + 1)/2$, obviously diverges as $n \to \infty$, albeit faster (one might expect) than the log of the likelihood function for the exponential distribution in Example C.6, which is $\ln L(\theta) = n(\ln \theta - \theta\bar{x}_n)$. As a final example, consider the downward bias of the maximum likelihood estimator of the variance of the normal distribution, $c_n = (n - 1)/n$, which is a constant that converges to one. (See Example C.5.)

We will define the rate at which a sequence converges or diverges in terms of the **order of the sequence**.

DEFINITION D.14 Order n^δ
A sequence c_n is of order n^δ, denoted $O(n^\delta)$, if and only if $\text{plim}(1/n^\delta)c_n$ *is a finite nonzero constant.*

DEFINITION D.15 Order less than n^δ
A sequence c_n, is of order less than n^δ, denoted $o(n^\delta)$, if and only if $\text{plim}(1/n^\delta)c_n$ *equals zero.*

Thus, in our examples, γ_n^2 is $O(n^{-1})$, $\text{Var}[x_{(1),n}]$ is $O(n^{-2})$ and $o(n^{-1})$, S_n is $O(n^2)$ (δ equals $+2$ in this case), $\ln L(\theta)$ is $\text{O}(n)$ (δ equals $+1$), and c_n is $O(1)(\delta = 0)$. Important particular cases that we will encounter repeatedly in our work are sequences for which $\delta = 1$ or -1.

The notion of order of a sequence is often of interest in econometrics in the context of the variance of an estimator. Thus, we see in Section D.3 that an important element of our strategy for forming an asymptotic distribution is that the variance of the limiting distribution of $\sqrt{n}(\bar{x}_n - \mu)/\sigma$ is $O(1)$. In Example D.10 the variance of m_2 is the sum of three terms that are $O(n^{-1})$, $O(n^{-2})$, and $O(n^{-3})$. The sum is $O(n^{-1})$, because $n\,\text{Var}[m_2]$ converges to $\mu_4 - \sigma^4$, the numerator of the first, or *leading term,* whereas the second and third terms converge to zero. This term is also the *dominant term* of the sequence. Finally, consider the two divergent examples in the preceding list. S_n is simply a deterministic function of n that explodes. However, $\ln L(\theta) = n \ln \theta - \theta \Sigma_i x_i$ is the sum of a constant that is $O(n)$ and a random variable with variance equal to n/θ. The random variable "diverges" in the sense that its variance grows without bound as n increases.

APPENDIX E

COMPUTATION AND OPTIMIZATION

E.1 INTRODUCTION

The computation of empirical estimates by econometricians involves using digital computers and software written either by the researchers themselves or by others.[1] It is also a surprisingly balanced mix of art and science. It is important for software users to be aware of how results are obtained, not only to understand routine computations, but also to be able to explain the occasional strange and contradictory results that do arise. This appendix will describe some of the basic elements of computing and a number of tools that are used by econometricians.[2] Section E.2 describes some techniques for computing certain integrals and derivatives that are recurrent in econometric applications. Section E.3 presents methods of optimization of functions. Some examples are given in Section E.4.

E.2 COMPUTATION IN ECONOMETRICS

This section will discuss some methods of computing integrals that appear frequently in econometrics.

[1] It is one of the interesting aspects of the development of econometric methodology that the adoption of certain classes of techniques has proceeded in discrete jumps with the development of software. Noteworthy examples include the appearance, both around 1970, of G. K. Joreskog's LISREL [Joreskog and Sorbom (1981)] program, which spawned a still-growing industry in linear structural modeling, and TSP [Hall (1982, 1984)], which was among the first computer programs to accept symbolic representations of econometric models and which provided a significant advance in econometric practice with its LSQ procedure for systems of equations. An extensive survey of the evolution of econometric software is given in Renfro (2007, 2009).

[2] This discussion is not intended to teach the reader how to write computer programs. For those who expect to do so, there are whole libraries of useful sources. Three very useful works are Kennedy and Gentle (1980), Abramovitz and Stegun (1971), and especially Press et al. (2007). The third of these provides a wealth of expertly written programs and a large amount of information about how to do computation efficiently and accurately. A recent survey of many areas of computation is Judd (1998).

E.2.1 COMPUTING INTEGRALS

One advantage of computers is their ability rapidly to compute approximations to complex functions such as logs and exponents. The basic functions, such as these, trigonometric functions, and so forth, are standard parts of the libraries of programs that accompany all scientific computing installations.[3] But one of the very common applications that often requires some high-level creativity by econometricians is the evaluation of integrals that do not have simple closed forms and that do not typically exist in "system libraries." We will consider several of these in this section. We will not go into detail on the nuts and bolts of how to compute integrals with a computer; rather, we will turn directly to the most common applications in econometrics.

E.2.2 THE STANDARD NORMAL CUMULATIVE DISTRIBUTION FUNCTION

The standard normal cumulative distribution function (cdf) is ubiquitous in econometric models. Yet this most homely of applications must be computed by approximation. There are a number of ways to do so.[4] Recall that what we desire is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt, \quad \text{where } \phi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

One way to proceed is to use a Taylor series:

$$\Phi(x) \approx \sum_{i=0}^{M} \frac{1}{i!}\frac{d^i\Phi(x_0)}{dx_0^i}(x - x_0)^i.$$

The normal cdf has some advantages for this approach. First, the derivatives are simple and not integrals. Second, the function is **analytic**; as $M \rightarrow \infty$, the approximation converges to the true value. Third, the derivatives have a simple form; they are the **Hermite polynomials** and they can be computed by a simple recursion. The 0*th* term in the preceding expansion is $\Phi(x)$ evaluated at the expansion point. The first derivative of the cdf is the pdf, so the terms from 2 onward are the derivatives of $\phi(x)$, once again evaluated at x_0. The derivatives of the standard normal pdf obey the recursion

$$\phi^i/\phi(x) = -x\phi^{i-1}/\phi(x) - (i-1)\phi^{i-2}/\phi(x),$$

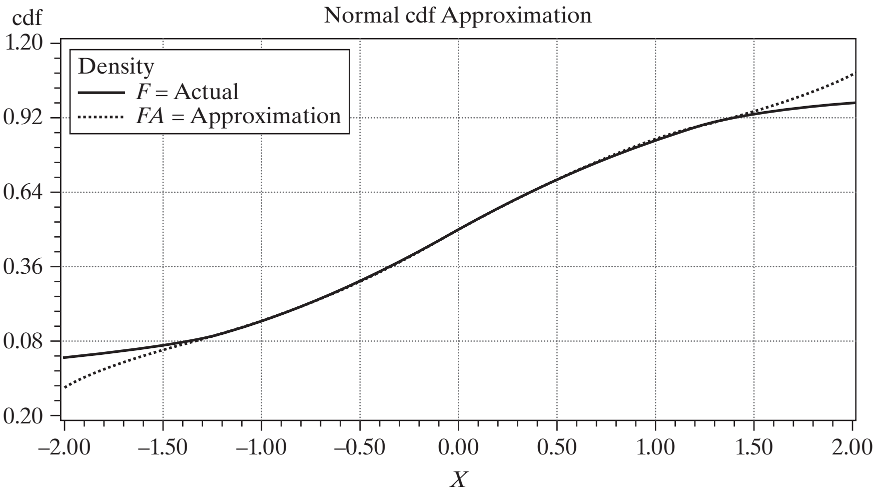
where ϕ^i is $d^i\phi(x)/dx^i$. The zero and one terms in the sequence are one and $-x$. The next term is $x^2 - 1$, followed by $3x - x^3$ and $x^4 - 6x^2 + 3$, and so on. The approximation can be made more accurate by adding terms. Consider using a fifth-order Taylor series approximation around the point $x = 0$, where $\Phi(0) = 0.5$ and $\phi(0) = 0.3989423$. Evaluating the derivatives at zero and assembling the terms produces the approximation $\Phi(x) \approx 12 + 0.3989423[x - x^3/6 + x^5/40]$.

[Some of the terms (every other one, in fact) will conveniently drop out.] Figure E.1 shows the actual values (F) and approximate values (FA) over the range -2 to 2. The figure shows two important points. First, the approximation is remarkably good over

[3] Of course, at some level, these must have been programmed as approximations by someone.

[4] Many system libraries provide a related function, the *error function,* $\text{erf}(x) = (2/\sqrt{\pi})\int_0^x e^{-t^2}\,dt$. If this is available, then the normal cdf can be obtained from $\Phi(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(x/\sqrt{2}), x \geq 0$ and $\Phi(x) = 1 - \Phi(-x), x \leq 0$.

FIGURE E.1 Approximation to Normal cdf.



most of the range. Second, as is usually true for Taylor series approximations, the quality of the approximation deteriorates as one gets far from the expansion point.

Unfortunately, it is the tail areas of the standard normal distribution that are usually of interest, so the preceding is likely to be problematic. An alternative approach that is used much more often is a polynomial approximation[5]:

$$\Phi(-|x|) = \phi(x)\sum_{i=1}^{5} a_i t^i + \varepsilon(x), \quad \text{where } t = 1/[1 + a_0|x|].$$

(The complement is taken if x is positive.) The error of approximation is less than $\pm 7.5 \times 10^{-8}$ for all x. (Note that the error exceeds the function value at $|x| > 5.7$, so this is the operational limit of this approximation.)

E.2.3 THE GAMMA AND RELATED FUNCTIONS

The standard normal cdf is probably the most common application of numerical integration of a function in econometrics. Another very common application is the class of gamma functions. For positive constant P, the gamma function is

$$\Gamma(P) = \int_0^\infty t^{P-1}e^{-t}\,dt.$$

The gamma function obeys the recursion $\Gamma(P) = (P-1)\Gamma(P-1)$, so for integer values of P, $\Gamma(P) = (P-1)!$ This result suggests that the gamma function can be viewed as a generalization of the factorial function for noninteger values. Another convenient value is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. By making a change of variable, it can be shown that for positive constants a, c, and P,

[5] Reported by Abramovitz and Stegun (1971, p. 932).

$$\int_0^\infty t^{P-1}e^{-at^c}\,dt = \int_0^\infty t^{-(P+1)}e^{-a/t^c}dt = \left(\frac{1}{c}\right)a^{-P/c}\Gamma\left(\frac{P}{c}\right). \tag{E-1}$$

As a generalization of the factorial function, the gamma function will usually overflow for the sorts of values of P that normally appear in applications. The log of the function should normally be used instead. The function $\ln \Gamma(P)$ can be approximated remarkably accurately with only a handful of terms and is very easy to program. A number of approximations appear in the literature; they are generally modifications of **Stirling's approximation** to the factorial function $P! \approx (2\pi P)^{1/2}P^P e^{-P}$, so

$$\ln \Gamma(P) \approx (P - 0.5)\ln P - P + 0.5\ln(2\pi) + C + \varepsilon(P),$$

where C is the correction term[6] and $\varepsilon(P)$ is the approximation error.[7]

The derivatives of the gamma function are

$$\frac{d^r\Gamma(P)}{dP^r} = \int_0^\infty (\ln t)^r t^{P-1}e^{-t}\,dt.$$

The first two derivatives of $\ln \Gamma(P)$ are denoted $\Psi(P) = \Gamma'/\Gamma$ and $\Psi'(P) = (\Gamma\Gamma'' - \Gamma'^2)/\Gamma^2$ and are known as the **digamma** and **trigamma** functions.[8] The **beta function**, denoted $\beta(a, b)$,

$$\beta(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

is related.

E.2.4 APPROXIMATING INTEGRALS BY QUADRATURE

The digamma and trigamma functions, and the gamma function for noninteger values of P and values that are not integers plus $\frac{1}{2}$, do not exist in closed form and must be approximated. Most other applications will also involve integrals for which no simple computing function exists. The simplest approach to approximating

$$F(x) = \int_{L(x)}^{U(x)} f(t)\,dt$$

is likely to be a variant of Simpson's rule, or the trapezoid rule. For example, one approximation is[9]

$$F(x) \approx \Delta[\tfrac{1}{3}f_1 + \tfrac{4}{3}f_2 + \tfrac{2}{3}f_3 + \tfrac{4}{3}f_4 + \cdots + \tfrac{2}{3}f_{N-2} + \tfrac{4}{3}f_{N-1} + \tfrac{1}{3}f_N],$$

[6] See, for example, Abramovitz and Stegun (1971, p. 257), Press et al. (2007), or Rao (1973, p. 59).

[7] For example, one widely used formula is $C = z^{-1}/12 - z^{-3}/360 - z^{-5}/1260 + z^{-7}/1680 - q$, where $z = P$ and $q = 0$ if $P > 18$, or $z = P + J$ and $q = \ln[P(P+1)(P+2)\ldots(P+J-1)]$, where $J = 18 - \text{INT}(P)$, if not. Note, in the approximation, we write $\Gamma(P) = (P!)/P +$ a correction.

[8] Tables of specific values for the gamma, digamma, and trigamma functions appear in Abramovitz and Stegun (1971). Most contemporary econometric programs have built-in functions for these common integrals, so the tables are not generally needed.

[9] See Press et al. (2007).

where f_j is the function evaluated at N equally spaced points in $[L, U]$ including the endpoints and $\Delta = (L - U)/(N - 1)$. There are a number of problems with this method, most notably that it is difficult to obtain satisfactory accuracy with a moderate number of points.

Gaussian quadrature is a popular method of computing integrals. The general approach is to use an approximation of the form

$$\int_L^U W(x)f(x)\,dx \approx \sum_{j=1}^{M} w_j f(a_j),$$

where $W(x)$ is viewed as a "weighting" function for integrating $f(x)$, w_j is the **quadrature weight**, and a_j is the **quadrature abscissa**. Different weights and abscissas have been derived for several weighting functions. Two weighting functions common in econometrics are

$$W(x) = x^c e^{-x}, \quad x \in [0, \infty),$$

for which the computation is called **Gauss–Laguerre quadrature**, and

$$W(x) = e^{-x^2}, \quad x \in (-\infty, \infty),$$

for which the computation is called **Gauss–Hermite quadrature.** The theory for deriving weights and abscissas is given in Press et al. (2007). Tables of weights and abscissas for many values of M are given by Abramovitz and Stegun (1971). Applications of the technique appear in Chapters 14 and 17.

E.3 OPTIMIZATION

Nonlinear optimization (e.g., maximizing log-likelihood functions) is an intriguing practical problem. Theory provides few hard and fast rules, and there are relatively few cases in which it is obvious how to proceed. This section introduces some of the terminology and underlying theory of nonlinear optimization.[10] We begin with a general discussion on how to search for a solution to a nonlinear optimization problem and describe some specific commonly used methods. We then consider some practical problems that arise in optimization. An example is given in the final section.

Consider maximizing the quadratic function

$$F(\boldsymbol{\theta}) = a + \mathbf{b}'\boldsymbol{\theta} - \tfrac{1}{2}\boldsymbol{\theta}'\mathbf{C}\boldsymbol{\theta},$$

b is a vector and **C** is a positive definite matrix. The first-order condition for a maximum is

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{b} - \mathbf{C}\boldsymbol{\theta} = \mathbf{0}. \tag{E-2}$$

[10] There are numerous excellent references that offer a more complete exposition. Among these are Quandt (1983), Bazaraa and Shetty (1979), Fletcher (1980), and Judd (1998). We note, modern econometric computer packages such as *Stata*, *SAS*, *NLOGIT*, *MATLAB*, *R*, and *GAUSS* all provide a "Maximize" (or "Minimize") "command" that allows a user to define a function to be maximized symbolically, and that put these details behind the curtain.

This set of *linear* equations has the unique solution

$$\boldsymbol{\theta} = \mathbf{C}^{-1}\mathbf{b}. \tag{E-3}$$

This is a linear optimization problem. Note that it has a **closed-form solution**; for any a, $\mathbf{b}$, and $\mathbf{C}$, the solution can be computed directly.[11] In the more typical situation,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{E-4}$$

is a set of nonlinear equations that cannot be solved explicitly for $\boldsymbol{\theta}$.[12] The techniques considered in this section provide systematic means of searching for a solution.

We now consider the general problem of maximizing a function of several variables:

$$\text{maximize}_{\theta}\, F(\boldsymbol{\theta}), \tag{E-5}$$

where $F(\boldsymbol{\theta})$ may be a log-likelihood or some other function. Minimization of $F(\boldsymbol{\theta})$ is handled by maximizing $-F(\boldsymbol{\theta})$. Two special cases are

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_i(\boldsymbol{\theta}), \tag{E-6}$$

which is typical for maximum likelihood problems, and the **least squares problem**,[13]

$$f_i(\boldsymbol{\theta}) = -(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2. \tag{E-7}$$

We treated the nonlinear least squares problem in detail in Chapter 7. An obvious way to search for the $\boldsymbol{\theta}$ that maximizes $F(\boldsymbol{\theta})$ is by trial and error. If $\boldsymbol{\theta}$ has only a single element and it is known approximately where the optimum will be found, then a **grid search** will be a feasible strategy. An example is a common time-series problem in which a one-dimensional search for a correlation coefficient is made in the interval $(-1, 1)$. The grid search can proceed in the obvious fashion—that is, $\ldots, -0.1, 0, 0.1, 0.2, \ldots$, then $\hat{\theta}_{\max} - 0.1$ to $\hat{\theta}_{\max} + 0.1$ in increments of 0.01, and so on—until the desired precision is achieved.[14] If $\boldsymbol{\theta}$ contains more than one parameter, then a grid search is likely to be extremely costly, particularly if little is known about the parameter vector at the outset. Nonetheless, relatively efficient methods have been devised.[15]

There are also systematic, derivative-free methods of searching for a function optimum that resemble in some respects the algorithms that we will examine in the next section. The **downhill simplex** (and other simplex) methods[16] have been found to be very fast and effective for some problems. A recent entry in the econometrics literature is the method of **simulated annealing**.[17] These derivative-free methods, particularly the latter, are often very effective in problems with many variables in the objective function, but they usually require far more function evaluations than the methods based on derivatives that are

[11] Notice that the constant a is irrelevant to the solution. Many maximum likelihood problems are presented with the preface "neglecting an irrelevant constant." For example, the log-likelihood for the normal linear regression model contains a term, $(-n/2)\ln(2\pi)$, that can be discarded.

[12] See, for example, the normal equations for the nonlinear least squares estimators of Chapter 7.

[13] Least squares is, of course, a minimization problem. The negative of the criterion is used to maintain consistency with the general formulation.

[14] There are more efficient methods of carrying out a one-dimensional search, for example, the golden section method. See Press et al. (2007).

[15] Quandt (1983) and Fletcher (1980) contain further details.

considered below. Because the problems typically analyzed in econometrics involve relatively few parameters but often quite complex functions involving large numbers of terms in a summation, on balance, the gradient methods are usually going to be preferable.[18]

E.3.1 ALGORITHMS

A more effective means of solving most nonlinear maximization problems is by an **iterative algorithm:**

> Beginning from initial value $\boldsymbol{\theta}_0$, at entry to iteration t, if $\boldsymbol{\theta}_t$ is not the optimal value for $\boldsymbol{\theta}$, compute direction vector $\boldsymbol{\Delta}_t$, step size λ_t, then

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t. \tag{E-8}$$

Figure E.2 illustrates the structure of an iteration for a hypothetical function of two variables. The direction vector $\boldsymbol{\Delta}_t$ is shown in the figure with $\boldsymbol{\theta}_t$. The dashed line is the set of points $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$. Different values of λ_t lead to different contours; for this $\boldsymbol{\theta}_t$ and $\boldsymbol{\Delta}_t$, the best value of λ_t is about 0.5.

Notice in Figure E.2 that for a given direction vector $\boldsymbol{\Delta}_t$ and current parameter vector $\boldsymbol{\theta}_t$, a secondary optimization is required to find the best λ_t. Translating from Figure E.2, we obtain the form of this problem as shown in Figure E.3. This subsidiary search is called a **line search**, as we search along the line $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$ for the optimal value of $F(.)$. The formal solution to the line search problem would be the λ_t that satisfies

$$\frac{\partial F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)}{\partial \lambda_t} = \mathbf{g}(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)' \boldsymbol{\Delta}_t = 0, \tag{E-9}$$

where $\mathbf{g}$ is the vector of partial derivatives of $F(.)$ evaluated at $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$. In general, this problem will also be a nonlinear one. In most cases, adding a formal search for λ_t will be too expensive, as well as unnecessary. Some approximate or ad hoc method will usually be chosen.

It is worth emphasizing that finding the λ_t that maximizes $F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)$ at a given iteration does not generally lead to the overall solution in that iteration. This situation is clear in Figure E.3, where the optimal value of λ_t leads to $F(.) = 2.0$, at which point we reenter the iteration.

E.3.2 COMPUTING DERIVATIVES

For certain functions, the programming of derivatives may be quite difficult. Numeric approximations can be used, although it should be borne in mind that analytic derivatives obtained by formally differentiating the functions involved are to be preferred. First derivatives can be approximated by using

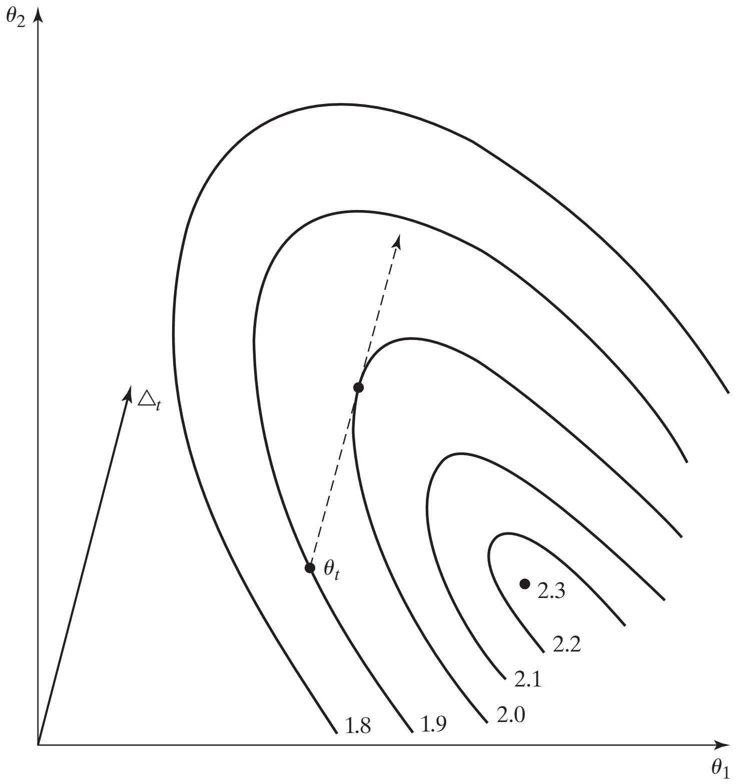
$$\frac{\partial F(\boldsymbol{\theta})}{\partial \theta_i} \approx \frac{F(\cdots \theta_i + \varepsilon \cdots) - F(\cdots \theta_i - \varepsilon \cdots)}{2\varepsilon}.$$

[16] See Nelder and Mead (1965) and Press et al. (2007).

[17] See Goffe, Ferrier, and Rodgers (1994) and Press et al. (2007).

[18] Goffe, Ferrier, and Rodgers (1994) did find that the method of simulated annealing was quite adept at finding the best among multiple solutions. This problem is common for derivative-based methods, because they usually have no method of distinguishing between a local optimum and a global one.

FIGURE E.2 Iteration.



The choice of ε is a remaining problem.[19]

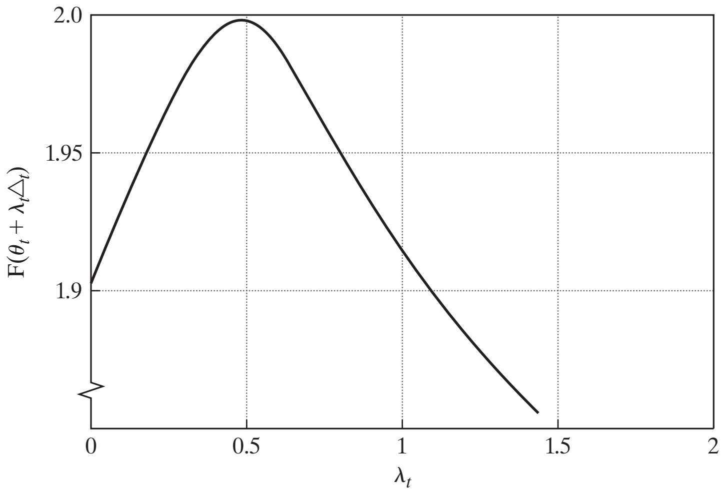
There are three drawbacks to this means of computing derivatives compared with using the analytic derivatives. A possible major consideration is that it may substantially increase the amount of computation needed to obtain a function and its gradient. In particular, $K + 1$ function evaluations (the criterion and K derivatives) are replaced with $2K + 1$ functions. The latter may be more burdensome than the former, depending on the complexity of the partial derivatives compared with the function itself. The comparison will depend on the application. But in most settings, careful programming that avoids superfluous or redundant calculation can make the advantage of the analytic derivatives substantial. Second, the choice of ε can be problematic. If it is chosen too large, then the approximation will be inaccurate. If it is chosen too small, then there may be insufficient variation in the function to produce a good estimate of the derivative. A compromise that is likely to be effective is to compute ε_i separately for each parameter, as in[20]

$$\varepsilon_i = \text{Max}[\alpha|\theta_i|, \gamma].$$

[19] Extensive discussion may be found in Quandt (1983).

[20] See Goldfeld and Quandt (1983).

FIGURE E.3 Line Search.



The values α and γ should be relatively small, such as 10^{-5}. Third, although numeric derivatives computed in this fashion are likely to be reasonably accurate, in a sum of a large number of terms, say, several thousand, enough approximation error can accumulate to cause the numerical derivatives to differ significantly from their analytic counterparts. Second derivatives can also be computed numerically. In addition to the preceding problems, however, it is generally not possible to ensure negative definiteness of a Hessian computed in this manner. Unless the choice of ε is made extremely carefully, an indefinite matrix is a possibility. In general, the use of numeric derivatives should be avoided if the analytic derivatives are available.

E.3.3 GRADIENT METHODS

The most commonly used algorithms are **gradient methods**, in which

$$\boldsymbol{\Delta}_t = \mathbf{W}_t\mathbf{g}_t, \tag{E-10}$$

where $\mathbf{W}_t$ is a positive definite matrix and $\mathbf{g}_t$ is the **gradient** of $F(\boldsymbol{\theta}_t)$:

$$\mathbf{g}_t = \mathbf{g}(\boldsymbol{\theta}_t) = \frac{\partial F(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}. \tag{E-11}$$

These methods are motivated partly by the following. Consider a linear Taylor series approximation to $F(\boldsymbol{\theta}_t + \lambda_t\boldsymbol{\Delta}_t)$ around $\lambda_t = 0$:

$$F(\boldsymbol{\theta}_t + \lambda_t\boldsymbol{\Delta}_t) \simeq F(\boldsymbol{\theta}_t) + \lambda_t\mathbf{g}(\boldsymbol{\theta}_t)'\boldsymbol{\Delta}_t. \tag{E-12}$$

Let $F(\boldsymbol{\theta}_t + \lambda_t\boldsymbol{\Delta}_t)$ equal F_{t+1}. Then,

$$F_{t+1} - F_t \simeq \lambda_t\mathbf{g}_t'\boldsymbol{\Delta}_t.$$

If $\boldsymbol{\Delta}_t = \mathbf{W}_t\mathbf{g}_t$, then

$$F_{t+1} - F_t \simeq \lambda_t\mathbf{g}_t'\mathbf{W}_t\mathbf{g}_t.$$

If $\mathbf{g}_t$ is not $\mathbf{0}$ and λ_t is small enough, then $F_{t+1} - F_t$ must be positive. Thus, if $F(\boldsymbol{\theta})$ is not already at its maximum, then we can always find a step size such that a gradient-type iteration will lead to an increase in the function. (Recall that $\mathbf{W}_t$ is assumed to be positive definite.)

In the following, we will omit the iteration index t, except where it is necessary to distinguish one vector from another. The following are some commonly used algorithms.[21]

Steepest Ascent, The simplest algorithm to employ is the **steepest ascent** method, which uses

$$\mathbf{W} = \mathbf{I} \text{ so that } \boldsymbol{\Delta} = \mathbf{g}. \quad \textbf{(E-13)}$$

As its name implies, the direction is the one of greatest increase of $F(.)$. Another virtue is that the line search has a straightforward solution; at least near the maximum, the optimal λ is

$$\lambda = \frac{-\mathbf{g}'\mathbf{g}}{\mathbf{g}'\mathbf{H}\mathbf{g}}, \quad \textbf{(E-14)}$$

where

$$\mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}.$$

Therefore, the steepest ascent iteration is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\mathbf{g}_t'\mathbf{g}_t}{\mathbf{g}_t'\mathbf{H}_t\mathbf{g}_t}\mathbf{g}_t. \quad \textbf{(E-15)}$$

Computation of the second derivatives matrix may be extremely burdensome. Also, if $\mathbf{H}_t$ is not negative definite, which is likely if $\boldsymbol{\theta}_t$ is far from the maximum, the iteration may diverge. A systematic line search can bypass this problem. This algorithm usually converges very slowly, however, so other techniques are usually used.

Newton's Method The template for most gradient methods in common use is Newton's method. The basis for **Newton's method** is a linear Taylor series approximation. Expanding the first-order conditions,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0},$$

equation by equation, in a linear Taylor series around an arbitrary $\boldsymbol{\theta}^0$ yields

$$\frac{\partial F(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \mathbf{0}, \quad \textbf{(E-16)}$$

where the superscript indicates that the term is evaluated at $\boldsymbol{\theta}^0$. Solving for $\boldsymbol{\theta}$ and then equating $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{t+1}$ and $\boldsymbol{\theta}^0$ to $\boldsymbol{\theta}_t$, we obtain the iteration

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1}\mathbf{g}_t. \quad \textbf{(E-17)}$$

[21] A more extensive catalog may be found in Judge et al. (1985, Appendix B). Those mentioned here are some of the more commonly used ones and are chosen primarily because they illustrate many of the important aspects of nonlinear optimization.

Thus, for Newton's method,

$$\mathbf{W} = -\mathbf{H}^{-1}, \qquad \boldsymbol{\Delta} = -\mathbf{H}^{-1}\mathbf{g}, \qquad \lambda = 1. \qquad \textbf{(E-18)}$$

Newton's method will converge very rapidly in many problems. If the function is quadratic, then this method will reach the optimum in one iteration from any starting point. If the criterion function is globally concave, as it is in a number of problems that we shall examine in this text, then it is probably the best algorithm available. This method is very well suited to maximum likelihood estimation.

Alternatives to Newton's Method Newton's method is very effective in some settings, but it can perform very poorly in others. If the function is not approximately quadratic or if the current estimate is very far from the maximum, then it can cause wide swings in the estimates and even fail to converge at all. A number of algorithms have been devised to improve upon Newton's method. An obvious one is to include a line search at each iteration rather than use $\lambda = 1$. Two problems remain, however. At points distant from the optimum, the second derivatives matrix may not be negative definite, and, in any event, the computational burden of computing **H** may be excessive.

The **quadratic hill-climbing method** proposed by Goldfeld, Quandt, and Trotter (1966) deals directly with the first of these problems. In any iteration, if **H** is not negative definite, then it is replaced with

$$\mathbf{H}_\alpha = \mathbf{H} - \alpha\mathbf{I}, \qquad \textbf{(E-19)}$$

where α is a positive number chosen large enough to ensure the negative definiteness of $\mathbf{H}_\alpha$. Another suggestion is that of Greenstadt (1967), which uses, at every iteration,

$$\mathbf{H}_\pi = -\sum_{i=1}^{n} |\pi_i| \mathbf{c}_i\mathbf{c}_i', \qquad \textbf{(E-20)}$$

where π_i is the ith characteristic root of **H** and $\mathbf{c}_i$ is its associated characteristic vector. Other proposals have been made to ensure the negative definiteness of the required matrix at each iteration.[22]

Quasi-Newton Methods: Davidon–Fletcher–Powell A very effective class of algorithms has been developed that eliminates second derivatives altogether and has excellent convergence properties, even for ill-behaved problems. These are the **quasi-Newton methods**, which form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{E}_t,$$

where $\mathbf{E}_t$ is a positive definite matrix.[23] As long as $\mathbf{W}_0$ is positive definite—**I** is commonly used—$\mathbf{W}_t$ will be positive definite at every iteration. In the **Davidon–Fletcher–Powell (DFP) method**, after a sufficient number of iterations, $\mathbf{W}_{t+1}$ will be an approximation to $-\mathbf{H}^{-1}$. Let

$$\boldsymbol{\delta}_t = \lambda_t\Delta_t, \quad \text{and} \quad \gamma_t = \mathbf{g}(\boldsymbol{\theta}_{t+1}) - \mathbf{g}(\boldsymbol{\theta}_t). \qquad \textbf{(E-21)}$$

[22] See, for example, Goldfeld and Quandt (1983).

[23] See Fletcher (1980).

The DFP **variable metric algorithm** uses

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\delta_t\delta_t'}{\delta_t'\gamma_t} + \frac{\mathbf{W}_t\gamma_t\gamma_t'\mathbf{W}_t}{\gamma_t'\mathbf{W}_t\gamma_t}. \qquad \textbf{(E-22)}$$

Notice that in the DFP algorithm, the change in the first derivative vector is used in **W**; an estimate of the inverse of the second derivatives matrix is being accumulated.

The variable metric algorithms are those that update **W** at each iteration while preserving its definiteness. For the DFP method, the accumulation of $\mathbf{W}_{t+1}$ is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{aa}' + \mathbf{bb}' = \mathbf{W}_t + [\mathbf{a} \quad \mathbf{b}][\mathbf{a} \quad \mathbf{b}]'.$$

The two-column matrix [**a b**] will have rank two; hence, DFP is called a **rank two update** or **rank two correction**. The **Broyden–Fletcher–Goldfarb–Shanno (BFGS)** method is a rank three correction that subtracts $v\mathbf{dd}'$ from the **DFP** update, where $v = (\gamma_t'\mathbf{W}_t\gamma_t)$ and

$$\mathbf{d}_t = \left(\frac{1}{\delta_t'\gamma_t}\right)\delta_t - \left(\frac{1}{\gamma_t'\mathbf{W}_t\gamma_t}\right)\mathbf{W}_t\gamma_t.$$

There is some evidence that this method is more efficient than DFP. Other methods, such as **Broyden's method**, involve a rank one correction instead. Any method that is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{QQ}'$$

will preserve the definiteness of **W** regardless of the number of columns in **Q**.

The DFP and BFGS algorithms are extremely effective and are among the most widely used of the gradient methods. An important practical consideration to keep in mind is that although $\mathbf{W}_t$ accumulates an estimate of the negative inverse of the second derivatives matrix for both algorithms, in maximum likelihood problems it rarely converges to a very good estimate of the covariance matrix of the estimator and should generally not be used as one.

E.3.4 ASPECTS OF MAXIMUM LIKELIHOOD ESTIMATION

Newton's method is often used for maximum likelihood problems. For solving a maximum likelihood problem, the **method of scoring** replaces **H** with

$$\overline{\mathbf{H}} = E[\mathbf{H}(\boldsymbol{\theta})], \qquad \textbf{(E-23)}$$

which will be recognized as the asymptotic covariance of the maximum likelihood estimator. There is some evidence that where it can be used, this method performs better than Newton's method. The exact form of the expectation of the Hessian of the log likelihood is rarely known, however.[24] Newton's method, which uses actual instead of expected second derivatives, is generally used instead.

One-Step Estimation A convenient variant of Newton's method is the **one-step maximum likelihood estimator**. It has been shown that if $\boldsymbol{\theta}^0$ is *any* consistent initial

[24] Amemiya (1981) provides a number of examples.

estimator of $\boldsymbol{\theta}$ and is $\mathbf{H}$, $\overline{\mathbf{H}}$, or any other asymptotically equivalent estimator of $\text{Var}[\mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MLE}})]$, then

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - (\mathbf{H}^*)^{-1}\mathbf{g}^0 \tag{E-24}$$

is an estimator of $\boldsymbol{\theta}$ that has the same asymptotic properties as the maximum likelihood estimator.[25] (Note that it is *not* the maximum likelihood estimator. As such, for example, it should not be used as the basis for likelihood ratio tests.)

Covariance Matrix Estimation In computing maximum likelihood estimators, a commonly used method of estimating $\mathbf{H}$ simultaneously simplifies the calculation of $\mathbf{W}$ and solves the occasional problem of indefiniteness of the Hessian. The method of Berndt et al. (1974) replaces $\mathbf{W}$ with

$$\hat{\mathbf{W}} = \left[\sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i'\right]^{-1} = (\mathbf{G}'\mathbf{G})^{-1}, \tag{E-25}$$

where

$$\mathbf{g}_i = \frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{E-26}$$

Then, $\mathbf{G}$ is the $n \times K$ matrix with ith row equal to $\mathbf{g}_i'$. Although $\hat{\mathbf{W}}$ and other suggested estimators of $(-\mathbf{H})^{-1}$ are asymptotically equivalent, $\hat{\mathbf{W}}$ has the additional virtues that it is always nonnegative definite, and it is only necessary to differentiate the log-likelihood once to compute it.

The Lagrange Multiplier Statistic The use of $\hat{\mathbf{W}}$ as an estimator of $(-\mathbf{H})^{-1}$ brings another intriguing convenience in maximum likelihood estimation. When testing restrictions on parameters estimated by maximum likelihood, one approach is to use the **Lagrange multiplier** statistic. We will examine this test at length at various points in this book, so we need only sketch it briefly here. The logic of the LM test is as follows. The gradient $\mathbf{g}(\boldsymbol{\theta})$ of the log-likelihood function equals $\mathbf{0}$ at the unrestricted maximum likelihood estimators (that is, at least to within the precision of the computer program in use). If $\hat{\boldsymbol{\theta}}_r$ is an MLE that is computed subject to some restrictions on $\boldsymbol{\theta}$, then we know that $\mathbf{g}(\hat{\boldsymbol{\theta}}_r) \neq \mathbf{0}$. The LM test is used to test whether, at $\hat{\boldsymbol{\theta}}_r$, $\mathbf{g}_r$ is *significantly* different from $\mathbf{0}$ or whether the deviation of $\mathbf{g}_r$ from $\mathbf{0}$ can be viewed as sampling variation. The covariance matrix of the gradient of the log-likelihood is $-\mathbf{H}$, so the Wald statistic for testing this hypothesis is $W = \mathbf{g}'(-\mathbf{H})^{-1}\mathbf{g}$. Now, suppose that we use $\hat{\mathbf{W}}$ to estimate $-\mathbf{H}^{-1}$. Let $\mathbf{G}$ be the $n \times K$ matrix with ith row equal to $\mathbf{g}_i'$, and let $\mathbf{i}$ denote an $n \times 1$ column of ones. Then the LM statistic can be computed as

$$\text{LM} = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}.$$

Because $\mathbf{i}'\mathbf{i} = n$,

$$\text{LM} = n[\mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}/n] = nR_i^2,$$

where R_i^2 is the *uncentered* R^2 in a regression of a column of ones on the derivatives of the log-likelihood function.

The Concentrated Log-Likelihood Many problems in maximum likelihood estimation can be formulated in terms of a partitioning of the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$ such

[25] See, for example, Rao (1973).

that at the solution to the optimization problem, $\boldsymbol{\theta}_{2,\,\mathrm{ML}}$, can be written as an explicit function of $\boldsymbol{\theta}_{1,\,\mathrm{ML}}$. When the solution to the likelihood equation for $\boldsymbol{\theta}_2$ produces

$$\boldsymbol{\theta}_{2,\,\mathrm{ML}} = \mathbf{t}(\boldsymbol{\theta}_{1,\,\mathrm{ML}}),$$

then, if it is convenient, we may "concentrate" the log-likelihood function by writing

$$F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = F[\boldsymbol{\theta}_1, \mathbf{t}(\boldsymbol{\theta}_1)] = F_c(\boldsymbol{\theta}_1).$$

The unrestricted solution to the problem $\mathrm{Max}_{\boldsymbol{\theta}_1} F_c(\boldsymbol{\theta}_1)$ provides the full solution to the optimization problem. Once the optimizing value of $\boldsymbol{\theta}_1$ is obtained, the optimizing value of $\boldsymbol{\theta}_2$ is simply $\mathbf{t}(\hat{\boldsymbol{\theta}}_{1,\,\mathrm{ML}})$. Note that $F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a subset of the set of values of the log-likelihood function, namely those values at which the second parameter vector satisfies the first-order conditions.[26]

E.3.5 OPTIMIZATION WITH CONSTRAINTS

Occasionally, some of or all the parameters of a model are constrained, for example, to be positive in the case of a variance or to be in a certain range, such as a correlation coefficient. Optimization subject to constraints is often yet another art form. The elaborate literature on the general problem provides some guidance—see, for example, Appendix B in Judge et al. (1985)—but applications still, as often as not, require some creativity on the part of the analyst. In this section, we will examine a few of the most common forms of constrained optimization as they arise in econometrics.

Parametric constraints typically come in two forms, which may occur simultaneously in a problem. Equality constraints can be written $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$, where $c_j(\boldsymbol{\theta})$ is a continuous and differentiable function. Typical applications include linear constraints on slope vectors, such as a requirement that a set of elasticities in a log-linear model add to one; exclusion restrictions, which are often cast in the form of interesting hypotheses about whether or not a variable should appear in a model (i.e., whether a coefficient is zero or not); and equality restrictions, such as the symmetry restrictions in a translog model, which require that parameters in two different equations be equal to each other. Inequality constraints, in general, will be of the form $a_j \le c_j(\boldsymbol{\theta}) \le b_j$, where a_j and b_j are known constants (either of which may be infinite). Once again, the typical application in econometrics involves a restriction on a single parameter, such as $\sigma > 0$ for a variance parameter, $-1 \le \rho \le 1$ for a correlation coefficient, or $\beta_j \ge 0$ for a particular slope coefficient in a model. We will consider the two cases separately.

In the case of equality constraints, for practical purposes of optimization, there are usually two strategies available. One can use a Lagrangean multiplier approach. The new optimization problem is

$$\mathrm{Max}_{\boldsymbol{\theta},\boldsymbol{\lambda}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\theta}).$$

The necessary conditions for an optimum are

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \mathbf{g}(\boldsymbol{\theta}) + \mathbf{C}(\boldsymbol{\theta})'\boldsymbol{\lambda} = \mathbf{0},$$

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0},$$

[26] A formal proof that this is a valid way to proceed is given by Amemiya (1985, pp. 125–127).

where $\mathbf{g}(\boldsymbol{\theta})$ is the familiar gradient of $F(\boldsymbol{\theta})$ and $\mathbf{C}(\boldsymbol{\theta})$ is a $J \times K$ matrix of derivatives with jth row equal to $\partial c_j/\partial\boldsymbol{\theta}'$. The joint solution will provide the constrained optimizer, as well as the Lagrange multipliers, which are often interesting in their own right. The disadvantage of this approach is that it increases the dimensionality of the optimization problem. An alternative strategy is to eliminate some of the parameters by either imposing the constraints directly on the function or by solving out the constraints. For exclusion restrictions, which are usually of the form $\theta_j = 0$, this step usually means dropping a variable from a model. Other restrictions can often be imposed just by building them into the model. For example, in a function of θ_1, θ_2, and θ_3, if the restriction is of the form $\theta_3 = \theta_1\theta_2$, then θ_3 can be eliminated from the model by a direct substitution.

Inequality constraints are more difficult. For the general case, one suggestion is to transform the constrained problem into an unconstrained one by imposing some sort of penalty function into the optimization criterion that will cause a parameter vector that violates the constraints, or nearly does so, to be an unattractive choice. For example, to force a parameter θ_j to be nonzero, one might maximize the augmented function $F(\boldsymbol{\theta}) - |1/\theta_j|$. This approach is feasible, but it has the disadvantage that because the penalty is a function of the parameters, different penalty functions will lead to different solutions of the optimization problem. For the most common problems in econometrics, a simpler approach will usually suffice. One can often reparameterize a function so that the new parameter is unconstrained. For example, the "method of squaring" is sometimes used to force a parameter to be positive. If we require θ_j to be positive, then we can define $\theta_j = \alpha^2$ and substitute α^2 for θ_j wherever it appears in the model. Then an unconstrained solution for α is obtained. An alternative reparameterization for a parameter that must be positive that is often used is $\theta_j = \exp(\alpha)$. To force a parameter to be between zero and one, we can use the function $\theta_j = 1/[1 + \exp(\alpha)]$. The range of α is now unrestricted. Experience suggests that a third, less orthodox approach works very well for many problems. When the constrained optimization is begun, there is a starting value $\boldsymbol{\theta}^0$ that begins the iterations. Presumably, $\boldsymbol{\theta}^0$ obeys the restrictions. (If not, and none can be found, then the optimization process must be terminated immediately.) The next iterate, $\boldsymbol{\theta}^1$, is a step away from $\boldsymbol{\theta}^0$, by $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \lambda_0\boldsymbol{\delta}^0$. Suppose that $\boldsymbol{\theta}^1$ violates the constraints. By construction, we know that there is some value $\boldsymbol{\theta}^1_*$ between $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^1$ that does not violate the constraint, where "between" means only that a shorter step is taken. Therefore, the next value for the iteration can be $\boldsymbol{\theta}^1_*$. The logic is true at every iteration, so a way to proceed is to alter the iteration so that the step length is shortened when necessary when a parameter violates the constraints.

E.3.6 SOME PRACTICAL CONSIDERATIONS

Different algorithms may perform differently in given settings. Indeed, for some problems, one algorithm may fail to converge whereas another will succeed in finding a solution without great difficulty. In view of this, computer programs such as *Gauss*, and *MatLab* that offer a menu of different preprogrammed algorithms can be particularly useful. It is sometimes worth the effort to try more than one algorithm on a given problem.

Step Sizes Except for the steepest ascent case, an optimal line search is likely to be infeasible or to require more effort than it is worth in view of the potentially large number of function evaluations required. In most cases, the choice of a step size is likely to be rather ad hoc. But within limits, the most widely used algorithms appear to be robust to inaccurate line searches. For example, one method employed by the widely used TSP computer program[27] is the method of *squeezing,* which tries $\lambda = 1, \frac{1}{2}, \frac{1}{4}$, and so on until an improvement in the function results. Although this approach is obviously a bit unorthodox, it appears to be quite effective when used with the Gauss–Newton method for nonlinear least squares problems. (See Chapter 7.) A somewhat more elaborate rule is suggested by Berndt et al. (1974). Choose an ε between 0 and $\frac{1}{2}$, and then find a λ such that

$$\varepsilon < \frac{F(\boldsymbol{\theta} + \lambda\boldsymbol{\Delta}) - F(\boldsymbol{\theta})}{\lambda\mathbf{g}'\Delta} < 1 - \varepsilon. \tag{E-27}$$

Of course, which value of ε to choose is still open, so the choice of λ remains ad hoc. Moreover, in neither of these cases is there any optimality to the choice; we merely find a λ that leads to a function improvement. Other authors have devised relatively efficient means of searching for a step size without doing the full optimization at each iteration.[28]

Assessing Convergence Ideally, the iterative procedure should terminate when the gradient is zero. In practice, this step will not be possible, primarily because of accumulated rounding error in the computation of the function and its derivatives. Therefore, a number of alternative convergence criteria are used. Most of them are based on the relative changes in the function or the parameters. There is some variation in those used in different computer programs, and there are some pitfalls that should be avoided. A critical absolute value for the elements of the gradient or its norm will be affected by any scaling of the function, such as normalizing it by the sample size. Similarly, stopping on the basis of small absolute changes in the parameters can lead to premature convergence when the parameter vector approaches the maximizer. It is probably best to use several criteria simultaneously, such as the proportional change in both the function and the parameters. Belsley (1980) discusses a number of possible stopping rules. One that has proved useful and is immune to the scaling problem is to base convergence on $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$.

Multiple Solutions It is possible for a function to have several local extrema. It is difficult to know a priori whether this is true of the one at hand. But if the function is not globally concave, then it may be a good idea to attempt to maximize it from several starting points to ensure that the maximum obtained is the global one. Ideally, a starting value near the optimum can facilitate matters; in some settings, this can be obtained by using a consistent estimate of the parameter for the starting point. The method of moments, if available, is sometimes a convenient device for doing so.

No Solution Finally, it should be noted that in a nonlinear setting the iterative algorithm can break down, even in the absence of constraints, for at least two reasons. The first possibility is that the problem being solved may be so numerically complex as to defy solution. The second possibility, which is often neglected, is that the proposed model may simply be inappropriate for the data. In a linear setting, a low R^2 or some other diagnostic test may suggest that the model and data are mismatched, but as long as the full rank condition is met by the regressor matrix, a linear regression can *always* be

[27] Hall (1982, p. 147).

[28] See, for example, Joreskog and Gruvaeus (1970), Powell (1964), Quandt (1983), and Hall (1982).

computed. Nonlinear models are not so forgiving. The failure of an iterative algorithm to find a maximum of the criterion function may be a warning that the model is not appropriate for this body of data.

E.3.7 THE EM ALGORITHM

The latent class model can be characterized as a **missing data model**. Consider the mixture model we used for DocVis in Chapter 14, which we will now generalize to allow more than two classes:

$$f(y_{it}|\mathbf{x}_{it}, class_i = j) = \theta_{it,j}(1-\theta_{it,j})^{y_{it}}, \theta_{it,j} = 1/(1+\lambda_{it,j}), \lambda_{it,j} = \exp(\mathbf{x}_{it}'\beta_j), y_{it} = 0, 1, \ldots.$$

$$\text{Prob}(class_i = j|\mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\alpha_j)}{\sum_{j=1}^{j} exp(z_i'\alpha_j)}, j = 1,2, \ldots, J.$$

With all parts incorporated, the log-likelihood for this latent class model is

$$\begin{aligned}\ln L_M &= \sum_{i=1}^{n} \ln L_{i,M}\\ &= \sum_{i=1}^{n} \ln\left\{\sum_{j=1}^{J} \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}{\sum_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\alpha}_m)} \prod_{t=1}^{T_i}\left(\frac{1}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)\left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)^{y_{it}}\right\}. \quad \textbf{(E-28)}\end{aligned}$$

Suppose the actual class memberships were known (i.e., observed). Then, the class probabilities in ln L_M would be unnecessary. The appropriate **complete data log-likelihood** for this case would be

$$\begin{aligned}\ln L_C &= \sum_{i=1}^{n} \ln L_{i,C}\\ &= \sum_{i=1}^{n} \ln\left\{\sum_{j=1}^{J} D_{ij} \prod_{t=1}^{T_i}\left(\frac{1}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)\left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)^{y_{it}}\right\}, \quad \textbf{(E-29)}\end{aligned}$$

where D_{ij} is an observed dummy variable that equals one if individual i is from class j, and zero otherwise. With this specification, the log-likelihood breaks into J separate log-likelihoods, one for each (now known) class. The maximum likelihood estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ would be obtained simply by separating the sample into the respective subgroups and estimating the appropriate model for each group using maximum likelihood. The method we have used to estimate the parameters of the full model is to replace the D_{ij} variables with their unconditional espectations, Prob($class_i = j|\mathbf{z}_i$), then maximize the resulting log-likelihood function. This is the essential logic of the **EM (expectation–maximization) algorithm**[29]; however, the method uses the conditional (posterior) class probabilities instead of the unconditional probabilities. The iterative steps of the EM algorithm are

(E step) Form the expectation of the missing data log-likelihood, conditional on the previous parameter estimates and the data in the sample;

[29] Dempster et al. (1977).

(M step) Maximize the expected log-likelihood function. Then either return to the E step or exit if the estimates have converged.

The EM algorithm can be used in a variety of settings.[30] It has a particularly appealing form for estimating latent class models. The iterative steps for the latent class model are as follows:

(E step) Form the conditional (posterior) class probabilities, $\pi_{ij}|\mathbf{z}_i$, based on the current estimates. These are based on the likelihood function.

(M step) For each class, estimate the class-specific parameters by maximizing a weighted log-likelihood,

$$\ln L_{M\,step,\,j} = \sum_{i=1}^{n_c} \pi_{ij} \ln L_i|class = j.$$

The parameters of the class probability model are also reestimated, as shown later, when there are variables in $\mathbf{z}_i$ other than a constant term.

This amounts to a simple weighted estimation. For example, in the latent class linear regression model, the M step would amount to nothing more than weighted least squares. For nonlinear models such as the geometric model above, the M step involves maximizing a weighted log-likelihood function.

For the preceding geometric model, the precise steps are as follows: First, obtain starting values for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J$. Recall, $\boldsymbol{\alpha}_J = \mathbf{0}$. Then;

1. Form the contributions to the likelihood function using (E-28),

$$\begin{aligned} L_i &= \sum_{j=1}^{J} \pi_{ij} \prod_{t=1}^{T_i} f(y_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}_j, class_i = j) \\ &= \sum_{j=1}^{J} L_i|class = j. \end{aligned} \qquad \textbf{(E-30)}$$

2. Form the conditional probabilities, $w_{ij} = \dfrac{L_i|class = j}{\sum_{m=1}^{J} L_i|class = m}$. **(E-31)**

3. For each j, now maximize the weighted log likelihood functions (one at a time),

$$\ln L_{j,\,M}(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} w_{ij} \ln \prod_{t=1}^{T_i} \left(\frac{1}{1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)\left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}\right)^{y_{it}} \qquad \textbf{(E-32)}$$

4. To update the $\boldsymbol{\alpha}_j$ parameters, maximize the following log-likelihood function

$$\ln L(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J) = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij} \ln \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}{\sum_{j=1}^{J} \exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}, \quad \boldsymbol{\alpha}_J = \mathbf{0}. \qquad \textbf{(E-33)}$$

[30] See McLachlan and Krishnan (1997).

Step 4 defines a multinomial logit model (with "grouped") data. If the class probability model does not contain any variables in $\mathbf{z}_i$, other than a constant, then the solutions to this optimization will be

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}}, \text{ then } \hat{\alpha}_j = \ln \frac{\hat{\pi}_j}{\hat{\pi}_J}. \tag{E-34}$$

(Note that this preserves the restriction $\hat{\alpha}_J = 0$.) With these in hand, we return to steps 1 and 2 to rebuild the weights, then perform steps 3 and 4. The process is iterated until the estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ converge. Step 1 is constructed in a generic form. For a different model, it is necessary only to change the density that appears at the end of the expresssion in (E-32). For a cross section instead of a panel, the product term in step 1 becomes simply the log of the single term.

The EM algorithm has an intuitive appeal in this (and other) settings. In practical terms, it is often found to be a very slow algorithm. It can take many iterations to converge. (The estimates in Example 14.17 were computed using a gradient method, not the EM algorithm.) In its favor, the EM method is very stable. It has been shown that the algorithm always climbs uphill.[31] The log-likelihood improves with each iteration. Applications differ widely in the methods used to estimate latent class models. Adding to the variety are the very many Bayesian applications, none of which use either of the methods discussed here.

E.4 EXAMPLES

To illustrate the use of gradient methods, we consider some simple problems.

E.4.1 FUNCTION OF ONE PARAMETER

First, consider maximizing a function of a single variable, $f(\theta) = \ln(\theta) - 0.1\theta^2$. The function is shown in Figure E.4. The first and second derivatives are

$$f'(\theta) = \frac{1}{\theta} - 0.2\theta,$$

$$f''(\theta) = \frac{-1}{\theta^2} - 0.2.$$

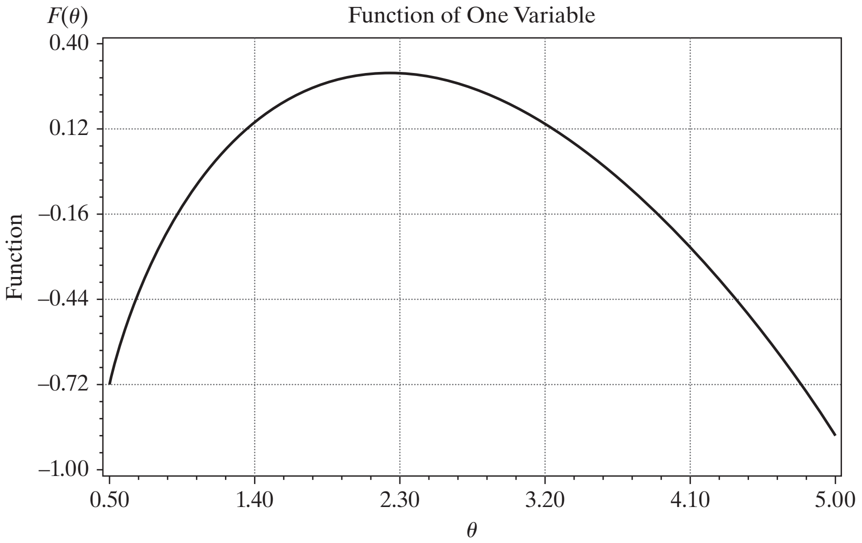
Equating f' to zero yields the solution $\theta = \sqrt{5} = 2.236$. At the solution, $f'' = -0.4$, so this solution is indeed a maximum. To demonstrate the use of an iterative method, we solve this problem using Newton's method. Observe, first, that the second derivative is always negative for any admissible (positive) θ.[32] Therefore, it should not matter where we start the iterations; we shall eventually find the maximum. For a single parameter, Newton's method is

$$\theta_{t+1} = \theta_t - [f'_t / f''_t].$$

[31] Dempster, Laird, and Rubin (1977).

[32] In this problem, an inequality restriction, $\theta > 0$, is required. As is common, however, for our first attempt we shall neglect the constraint.

FIGURE E.4 Function of One Variable Parameter.



The sequence of values that results when 5 is used as the starting value is given in Table E.1. The path of the iterations is also shown in the table.

E.4.2 FUNCTION OF TWO PARAMETERS: THE GAMMA DISTRIBUTION

For random sampling from the gamma distribution, the density is

$$f(y_i, \beta, \rho) = \frac{\beta^\rho}{\Gamma(\rho)} e^{-\beta y_i} y_i^{\rho-1}.$$

The log-likelihood is $\ln L(\beta, \rho) = n\rho \ln \beta - n \ln \Gamma(\rho) - \beta \sum_{i=1}^{n} y_i + (\rho - 1) \sum_{i=1}^{n} \ln y_i$.

(See Section 14.6.4 and Example 13.5.) It is often convenient to scale the log-likelihood by the sample size. Suppose, as well, that we have a sample with $\overline{y} = 3$ and $\overline{\ln y} = 1$. Then the function to be maximized is $F(\beta, \rho) = \rho \ln \beta - \ln \Gamma(\rho) - 3\beta + \rho - 1$. The derivatives are

$$\frac{\partial F}{\partial \beta} = \frac{\rho}{\beta} - 3, \qquad \frac{\partial F}{\partial \rho} = \ln \beta - \frac{\Gamma'}{\Gamma} + 1 = \ln \beta - \Psi(\rho) + 1,$$

$$\frac{\partial^2 F}{\partial \beta^2} = \frac{-\rho}{\beta^2}, \qquad \frac{\partial^2 F}{\partial \rho^2} = \frac{-(\Gamma\Gamma'' - \Gamma'^2)}{\Gamma^2} = -\Psi'(\rho), \qquad \frac{\partial^2 F}{\partial \beta\, \partial \rho} = \frac{1}{\beta}.$$

Finding a good set of starting values is often a difficult problem. Here we choose three starting points somewhat arbitrarily: $(\rho^0, \beta^0) = (4, 1), (8, 3)$, and $(2, 7)$. The solution to the problem is (5.233, 1.7438). We used Newton's method and DFP with a line search to maximize this function.[33] For Newton's method, $\lambda = 1$. The results are shown in

TABLE E.1 Iterations for Newton's Method

Iteration	θ	f	f'	f''
0	5.00000	−0.890562	−0.800000	−0.240000
1	1.66667	0.233048	0.266667	−0.560000
2	2.14286	0.302956	0.030952	−0.417778
3	2.23404	0.304718	0.000811	−0.400363
4	2.23607	0.304719	0.0000004	−0.400000

Table E.2. The two methods were essentially the same when starting from a good starting point (trial 1), but they differed substantially when starting from a poorer one (trial 2). Note that DFP and Newton approached the solution from different directions in trial 2. The third starting point shows the value of a line search. At this starting value, the Hessian is extremely large, and the second value for the parameter vector with Newton's method is (−47.671, −233.35), at which point F cannot be computed and this method must be abandoned. Beginning with $\mathbf{H} = \mathbf{I}$ and using a line search, DFP reaches the point (6.63, 2.03) at the first iteration, after which convergence occurs routinely in three more iterations. At the solution, the Hessian is $[(-1.72038, 0.191153)', (0.191153, -0.210579)']$. The diagonal elements of the Hessian are negative and its determinant is 0.32574, so it is negative definite. (The two characteristic roots are −1.7442 and −0.18675). Therefore, this result is indeed the maximizer of the function.

E.4.3 A CONCENTRATED LOG-LIKELIHOOD FUNCTION

There is another way that the preceding problem might have been solved. The first of the necessary conditions implies that at the joint solution for (β, ρ), β will equal $\rho/3$. Suppose that we impose this requirement on the function we are maximizing. The **concentrated** (over β) **log-likelihood function** is then produced:

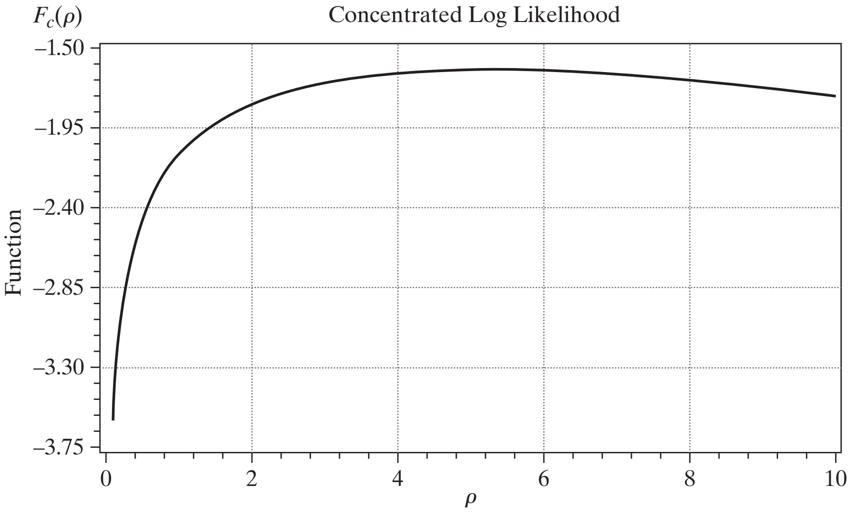
$$\begin{aligned} F_c(\rho) &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 3(\rho/3) + \rho - 1 \\ &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 1. \end{aligned}$$

TABLE E.2 Iterative Solutions to Max Max$(\rho, \beta)\rho \ln \beta - \ln \mathbf{\Gamma}(\rho) - 3\beta + \rho - \mathbf{1}$

	Trial 1				***Trial 2***				***Trial 3***			
	DFP		***Newton***		***DFP***		***Newton***		***DFP***		***Newton***	
Iteration	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β
0	4.000	1.000	4.000	1.000	8.000	3.000	8.000	3.000	2.000	7.000	2.000	7.000
1	3.981	1.345	3.812	1.203	7.117	2.518	2.640	0.615	6.663	2.027	−47.7	−233.
2	4.005	1.324	4.795	1.577	7.144	2.372	3.203	0.931	6.195	2.075	—	—
3	5.217	1.743	5.190	1.728	7.045	2.389	4.257	1.357	5.239	1.731	—	—
4	5.233	1.744	5.231	1.744	5.114	1.710	5.011	1.656	5.251	1.754	—	—
5	—	—	—	—	5.239	1.747	5.219	1.740	5.233	1.744	—	—
6	—	—	—	—	5.233	1.744	5.233	1.744	—	—	—	—

[33] The one used is described in Joreskog and Gruvaeus (1970).

FIGURE E.5 Concentrated Log-Likelihood.



This function could be maximized by an iterative search or by a simple one-dimensional grid search. Figure E.5 shows the behavior of the function. As expected, the maximum occurs at $\rho = 5.233$. The value of β is found as $5.23/3 = 1.743$.

The concentrated log-likelihood is a useful device in many problems. (See Section 14.9.3 for an application.) Note the interpretation of the function plotted in Figure E.5. The original function of ρ and β is a surface in three dimensions. The curve in Figure E.5 is a projection of that function; it is a plot of the function values above the line $\beta = \rho/3$. By virtue of the first-order condition, we know that one of these points will be the maximizer of the function. Therefore, we may restrict our search for the overall maximum of $F(\beta, \rho)$ to the points on this line.